New methods for inference on demographic history from genetic data

Coralie Merle

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S - Information, Structures, Systèmes**
Et de l'unité de recherche **UMR 5149 - IMAG - Institut Montpelliérain Alexander Grothendieck**

Spécialité: **Biostatistique**

Présentée par **Coralie Merle**

Nouvelles méthodes d'inférence de l'histoire démographique à partir de données génétiques

Soutenue le 12 décembre 2016 devant le jury composé de

Gilles CELEUX	Directeur de recherche	Inria Saclay-Île-de-France	Rapporteur
Sophie DONNET	Chargée de recherche	INRA-AgroParisTech	Examinatrice
Andreas FUTSCHIK	Professeur	Johannes Kepler Universität Linz	Rapporteur
Raphaël LEBLOIS	Chargé de recherche	INRA Montpellier	Co-encadrant
Jean-Michel MARIN	Professeur	CNRS & Université de Montpellier	Directeur
François ROUSSET	Directeur de recherche	CNRS & Université de Montpellier	Co-Directeur

Jury présidé par Gilles CELEUX

À la mémoire de
Bon Papa,
Tatie Marie,
Christophe.

Le recherche scientifique ressemble beaucoup à celle de l'or : vous sortez et vous chassez, armé de cartes et d'instruments, mais en fin de compte, ni vos préparatifs, ni même votre intuition ne comptent. Vous avez besoin de votre chance...

La Variété Andromède (1969) de Michael Crichton

REMERCIEMENTS

Raphaël Leblois, Jean-Michel Marin, Pierre Pudlo et François Rousset sont évidemment les premiers que je remercie pour m'avoir initiée aux thèmes de recherche traités dans cette thèse. Merci pour tous vos conseils, que ce soit en probabilité, en statistique, en génétique des populations, dans l'écriture d'un code R ou C++, dans la rédaction d'un manuscrit et bien d'autres. Vous m'avez permis d'acquérir des connaissances et compétences dans tous ces domaines qui m'étaient étrangers auparavant et m'avez accordé une indépendance croissante. Jean-Michel pour ton franc(et fort)-parlé en toutes circonstances, Pierre pour tes visites spontanées et régulières dans mon bureau, François pour m'inciter à plus de rigueur et de clarté, Raphaël pour te soucier de mon moral. Je finirai en vous remerciant pour les multiples relectures attentives et porteuses de ce manuscrit.

Je tiens ensuite à exprimer ma gratitude à Gilles Celeux et à Andreas Futschik [1] pour avoir accepté sans hésitation de rapporter ma thèse. Je les remercie pour leurs relectures attentives et pour leurs remarques pertinentes. Je remercie également Sophie Donnet d'avoir accepté de se joindre à eux pour constituer mon jury de thèse.

Un grand merci à Simon Boitard et Stéphane Robin pour avoir participé activement à mes comités de suivi de thèse. Soucieux que ma thèse se déroule pour le mieux, vous avez toujours été de bon conseil. Votre implication a joué un rôle non négligeable dans cet aboutissement.

Je remercie aussi André Mas et Jean-Noël Bacro pour s'être assuré du bon déroulement de cette thèse ainsi que Damien Calaque qui a toujours les informations dont on a besoin. Globalement je remercie les acteurs de l'école doctorale I2S notamment pour la formation complémentaire qu'elle nous offre. Je remercie plus particulièrement Martha Boeglin formatrice "Writing a scientific paper step by step" et Christophe Fiorio formateur "Apprendre et maîtriser LaTeX : Rédaction de documents" dont les formations étaient pédagogiques et stimulantes. Elles m'ont permis de progesser dans ces deux domaines et elles m'ont aidée à rédiger le présent manuscrit.

Je n'oublie pas Christophe Giraud et Pascal Massart sans qui cette thèse n'aurait tout simplement jamais exister, merci de m'avoir encouragée dans cette direction et de m'avoir connectée à mes directeurs par le biais du stage de master 2 MathSV d'Orsay. Je suis reconnaissante de la bourse au mérite que la Fondation Mathématiques Jacques Hadamard m'a accordée pour mener à bien mon année de master 2.

Je remercie les labex Numev et Cemeb d'avoir cofinancé cette thèse ainsi que l'IBC pour son financement du projet qui a donné lieu au chapitre 4 de cette thèse et en particulier Geneviève Carrière et Caroline Benoist pour la gestion de mes nombreuses missions.

1. Ich möchte Adreas Futschik meinen Dank dafür ausdrücken, dass er mit Freude eingewilligt hat mein Berichterstatter zu sein. Weiterhin bedanke ich mich, dass er die Reise nach Montpellier gemacht hat um ein Mitglied der Kommission zu sein.

J'ai aussi une pensée pour tous les enseignants qui ont su cultiver mon goût pour l'apprentissage en particulier des mathématiques et de la biologie. Mes deux plus belles années d'études sont celles que j'ai passées en Maths sup Maths spé grâce au cadre extraodinaire des CPGE au lycée Notre dame de la Merci mais surtout grâce à l'équipe enseignante : Monique Letourmy, Arnaud de Saint Julien, Jean-Marc Reverdy et Bruno Harington. Ces années ont été le point de départ de mon parcours supérieur, me permettant de m'épanouïr scientifiquement et personnellement. Je remercie aussi tous les enseignants du département de mathématiques d'Orsay pour la qualité de la formation dispensée et pour leurs qualités humaines. Je souligne aussi la valeur ajoutée apportée par le magistère de mathématiques d'Orsay et en particulier par l'encadrement et les conseils de Frédéric Paulin.

L'accueil chaleureux des membres de l'IMAG, et plus particulièrement des membres de l'équipe probabilité et statistiques, a rendu ma présence au laboratoire toujours plus agréable : Ali, Alice, André, Baptiste, Benjamin, Benoîte, Catherine, Christian, Christophe, Cyrille, Elodie, Fabien, Gilles, Gwladys, Irène, Jean-Noël, Ludovic, Mathieu, Nicolas, Sophie L., Sophie C., Vanessa, Véronique, Xavier. Un mention particulière pour ceux avec qui j'ai pu échanger au sujet des enseignements : Jean Malgoire, Benjamin Charlier, Elsa Ibanez, Julien Stoehr, Élodie Brunel, Hanen Ben Salah, Jean Peyhardi, Véronique Ladret, Nicolas Molinari...

Pour leur disponibilité et leur soutien constant dans nos démarches administratives, un grand merci à Sophie Cazanave-Pin, Myriam Debus, Éric Hugounenq, Bernadette Lacan, Carmela Madonia, Nathalie Quintin et Laurence Roux. Merci aussi à Gemma Bellal pour ses petits mots d'encouragement.

Je remercie aussi les membres du CBGP, d'autant plus contents de mes visites qu'elles étaient clairsemées. Je pense plus particulièrement à Alexandre Dehne Garcia pour ton aide constante autour du cluster, Arnaud Estoup pour la poésie de tes coccinnelles, Mathieu Gautier pour toute ton aide et tes explications d'ordre génétique mais pas que, Miguel Navascués pour nos différentes discussions, Renaud Vitalis pour coalesceR, Nathalie Vieira pour ton aide informatique et avec Thibaud pour les matchs de volley ou de beach-volley.

Quelques mots pour tous les doctorants et post-doctorants que j'ai rencontrés durant ces trois années ; merci pour l'entraide générale et sans faille ainsi que pour toutes les discussions quelquefois scientifiques et très souvent farfelues que j'ai pu avoir avec chacun d'entre vous : Thomas, Angelina, Jean, Pierre, Etienne, Christophe, Arnaud, Alaaeddine, Benjamin, Elsa, Mickaël, Tutu , Hanene, Walid, Antoine, Myriam, Wafa, Guillaume, Joubine, Boushra, Julien Sig., Anis, Francesco, Samuel, Davis, Tito, Jérémie, Wenran, Nejib, Stéphanie, Paul-Marie, Yann, Théo, Quentin, Mike, Alexandre, Rodrigo, Michele, Vianney, Florent, Ridha, Jocelyn, Alexandra, Paul, Abel, Jérémy, Mario, May, Maud, Valentin, Florian, Elodie, Emma, Manuel, Ch. Julian, Raphaël. Mention particulière pour mes co-bureaux modèles Claudia, Karine, Amina, Abdessamad, Emmanuelle, Paul, Guillain, Martin et ceux qui m'ont moins vue ! Merci Mickaël pour nos discussions toujours contradictoires, pour tes blagues merveilleuses et surtout pour tes danses endiablées, Christian pour discuter de sujets toujours plus chauds et bien sûr pour toutes les activités extra-scolaires ! Un grand merci à toi Gautier pour ton efficacité légendaire et ton calme durant l'organisation des doctiss, pour ta culture, pour tes

danses, pour les poèmes et surtout pour ta bonne humeur à toute épreuve.

Immanquables remerciements à mes (demi)-frères de thèse : Mohammed pour tes conseils avisés et rassurants de grand frère, Julien pour tes conseils, ton soutien, tes bêtises, les danses, les bières, les fous rires et tout le reste, Paul-Marie pour tous les sursauts quand tu cognes la porte, Louis pour ton sens du partage du directeur. Je pense aussi à Champak toujours disponible pour discuter de génétique des pop ou de l'Inde.

Une pensée pour tous les amis qui ont accompagné mon parcours universitaire de près ou de loin : Juliette, Guillaume, Robin et Marion, Thomas, Lucie, Lucile, Charlotte, Julien et Nathalie, David et Fanny, Anne et Steph, Olivier et Phyllis, Sébastien et Alicia, Rémi, Loïc, Alex, Yvan.

Pour finir, je remercie ma famille de me supporter chaque jour : me supporter car en tant que matheuse je ne vis pas toujours dans ce monde et que quand j'y vis c'est bruyant ! Vous croyez toujours plus que moi en ma réussite dans quoi que ce soit ! Et pourtant dans les moments de découragement, vous me rappelez qu'un tas de gens vivent heureux sans agreg ni doctorat ! J'ajoute mes oncles, tantes, cousins et cousines et surtout mes grands-parents toujours intéréssés, curieux, confiants et fiers. Merci Papa pour tes conseils philosophiques si sages (des fois !) mais si difficiles à appliquer et pour ta curiosité scientifique à laquelle je n'ai jamais de réponse ! Merci Maman pour les innombrables relectures de mes différents mémoires universitaires et pour ton attention lors de mes répétitions en pyjama ! Un énorme merci à Amandine, Pierre et Benjamin pour n'être jamais tous d'accord mais toujours frères et soeurs quoi qu'il arrive et même quand je blague maths... Audrey merci pour tes questions déstabilisantes mais toujours pertinentes ;-).

Je remercie également les membres de ma chaleureuse belle-famille de se soucier de moi comme ma propre famille et de m'avoir soutenue et encouragée durant cette longue aventure. Mon dernier remerciement s'adresse naturellement à Tédy. Toi qui a su t'éclipser au moment clé où je manquais de temps, pour ne pas voir ma mauvaise mine des derniers jours de rédaction et surtout pour éviter l'interminable soutenance de thèse. Plus sérieusement, merci pour ton soutien moral tout au long de mon parcours universitaire, malgré la distance toujours plus grande.

À vous tous et à ceux que je n'ai pas mentionnés dans ma distraction légendaire, tout simplement merci.

Table des matières

Table des figures

1 Introduction

Le polymorphisme génétique désigne la coexistence de plusieurs allèles pour un gène ou locus donné, dans une population animale, végétale, fongique, bactérienne... Il explique qu'une espèce présente des individus aux caractères phénotypiques différents au sein d'une même population. On considère que cet élément de la diversité génétique participe à l'adaptation des populations à leur environnement plus ou moins changeant.

Bien avant que James Watson et Francis Crick ne présentent la structure en double hélice de l'ADN [Watson et al., 1953], les statisticiens du siècle dernier avaient posé les bases conceptuelles et une formalisation mathématique de l'évolution de la variation génétique dans les populations. Les travaux pionniers de Sewall Wright, John B. S. Haldane et Ronald A. Fisher, généralement considérés comme les fondateurs de la génétique des populations, ont à présent été étendus et les prédictions théoriques de leurs modèles ont récemment été largement utilisées sur des milliers de séquences génomiques de différentes populations d'espèces diverses.

La masse de données disponibles dans le cadre de l'étude du polymorphisme génétique s'est considérablement accrue avec le développement récent des technologies de biologie moléculaire. Conjointement, la génétique des populations théorique a fourni un modèle aléatoire, le coalescent de Kingman [Kingman, 1982a] décrivant le polymorphisme génétique d'un échantillon à partir d'un petit nombre de paramètres génétiques (taux de mutations, ...) et au niveau de la population (taille de la population, ...). À partir du coalescent de Kingman, il est désormais possible de modéliser des scénarios démographiques complexes expliquant le polymorphisme génétique. On peut alors envisager l'inférence statistique des paramètres démographiques de ces scénarios ou de choisir entre différents scénarios évolutifs (histoire d'une pandémie, ...).

Les modèles probabilistes de coalescent sur la généalogie de gènes ne fournissent généralement pas de vraisemblances explicites. En effet, la présence de structures latentes complexes telles que des généalogies de gènes dans le modèle complique fortement le calcul de la vraisemblance de ce modèle pour un jeu de données (voir Stephens and Donnelly [2000], Beaumont et al. [2002], De Iorio et al. [2005]). Des schémas d'inférence comme le maximum de vraisemblance nécessitent alors l'emploi de méthodologies statistiques innovantes pour contourner ce problème (méthodes de Monte-Carlo

par chaînes de Markov (MCMC, voir Robert and Casella [2013]), méthodes d'échantillonnage préférentiel ou importance sampling (IS, voir Stephens and Donnelly [2000], De Iorio and Griffiths [2004a], De Iorio and Griffiths [2004b], De Iorio et al. [2005], Leblois et al. [2014]). Loin d'être routinière, l'utilisation de ces techniques pour analyser des scénarios démographiques complexes sur des jeux de données de grande dimension nécessite au minimum d'optimiser les méthodes d'inférences actuelles voire de recourir à des méthodes basées sur des pseudo-vraisemblances. En outre, si différentes procédures de choix de modèles existent dans le paradigme bayésien, peu de méthodologies ont été proposées en dehors de ce cadre.

L'objectif de cette thèse est de développer des outils statistiques adaptés à des modèles stochastiques relativement complexes de génétique des populations. Plus particulièrement, l'accélération d'un algorithme d'estimation de la vraisemblance sous un modèle avec un changement de la taille de la population au cours du temps constitue notre première contribution, présentée dans le paragraphe suivant. Par ailleurs, l'arrivée de nouvelles technologies de séquençage (Next Generation Sequencing) a permi d'augmenter le nombre de marqueurs dans les jeux de données et d'avoir accès à des données de type séquences ou génomes complets. Pour tirer parti de l'information contenue dans ces nouvelles données, notre deuxième contribution, présentée ensuite, repose sur le développement d'une procédure de choix de modèles pour des données de type longueurs de séquences conservées entre deux fragments d'ADN.

Pour un modèle paramétrique basé sur le coalescent, étant donné un échantillon observé, la vraisemblance en un point de l'espace des paramètres s'écrit comme la somme des probabilités de toutes les histoires (généalogies munies de mutations) qui peuvent avoir généré cet échantillon. À l'heure actuelle, les meilleures méthodes pour l'inférence des paramètres de ce type de modèles sont d'une part les méthodes de recherche du maximum de vraisemblance (Monte Carlo Markov Chains (MCMC) [Robert and Casella, 2013], Important Sampling (IS) [Griffiths and Tavaré, 1994b]) et d'autre part les méthodes bayésiennes approchées (Approximate Bayesian Computation (ABC) [Beaumont et al., 2002] ; [Marin et al., 2012]). L'algorithme d'échantillonnage préférentiel séquentiel (Sequential Important Sampling : SIS) estime la vraisemblance, en parcourant de manière habile et efficace l'espace latent des histoires possibles pour l'échantillon observé (Stephens and Donnelly [2000], De Iorio et al. [2005], Rousset and Leblois [2012]). Dans ce schéma, la distribution d'importance propose les histoires ayant pu générer l'échantillon observé les plus probables possible. Cette technique est malheureusement lente mais fournit des estimations par maximum de vraisemblance d'une grande précision. Lorsque la taille de la population est constante au cours du temps, on dit que le modèle est homogène ou à l'équilibre. Au contraire, les modèles stochastiques que nous souhaitons inférer incluent des variations de taille de la population. Ces méthodes d'échantillonnage préférentiel ne sont pas efficaces pour des modèles de populations en déséquilibre car les distributions d'importance ont été développées pour une population de taille constante au cours du temps. Le temps de calcul augmente fortement pour la même précision de l'estimation de la vraisemblance. La première contribution de cette thèse a consisté à explorer l'algorithme SIS avec rééchantillonnage (Sequential Important Sampling and Resampling : SISR). L'idée est de rééchantillonner de façon à apprendre quelles sont les généalogies proposées par la distribution d'importance qui seront les plus probables avant d'avoir terminé leur simulation et ainsi diminuer le temps de calcul. Par ailleurs, nous avons proposé une nouvelle

distribution de rééchantillonnage, tirant profit de l'information contenue dans la vraisemblance composite par paire de l'échantillon.

Le développement récent des technologies de séquençage à haut débit a révolutionné la génération de données de polymorphisme chez de nombreux organismes, augmentant le nombre de marqueurs disponibles dans les jeux de données. Les méthodes d'inférence classiques par maximum de vraisemblance (MCMC, IS) ou basées sur le Sites Frequency Spectrum (SFS), adaptées à des jeux de données de polymorphisme génétique de quelques loci, supposent que les généalogies de différents loci sont indépendantes. Pour tirer parti de données beaucoup plus denses sur le génome, constituées de génomes entiers ou de fragments de génome, il nous faut considérer la dépendance des généalogies sur des positions voisines du génome et modéliser la recombinaison génétique. Ainsi, lorsque l'on modélise la recombinaison, la vraisemblance prend la forme d'une intégrale sur tous les graphes de recombinaison ancestraux (ARG, de l'anglais Ancestral Recombination Graph, voir Griffiths and Tavaré [1994a]) possibles pour les séquences échantillonnées. Cet espace est de bien plus grande dimension que l'espace des généalogies, au point que les méthodes d'inférence basée sur la vraisemblance ne peuvent plus être utilisées sans plus d'approximations. De nombreuses méthodes infèrent les changements historiques de la taille efficace de la population mais ne considèrent pas la complexité du modèle ajusté. Même si certaines proposent un contrôle d'un potentiel surajustement du modèle, à notre connaissance, aucune procédure de choix de modèle entre des modèles démographiques de complexité différente n'a été proposée à partir de longueurs de segments identiques. Nous nous concentrons sur un modèle de taille de population constante et un modèle de population ayant subi un unique changement de taille dans le passé. Puisque ces modèles sont emboîtés, nous avons développé un critère de choix de modèle pénalisé basé sur la comparaison d'homozygoties haplotypiques observée et théorique. Notre pénalisation repose sur des indices de sensibilité de Sobol. C'est une forme de pénalité liée à la complexité du modèle.

La suite de cette partie introductive est consacrée à une brève présentation de nos contributions aux problèmes d'inférence démographique dans des modèles de taille de population variable au cours du temps. La première concerne l'amélioration des méthodes d'échantillonnage préférentiel pour des modèles de population de taille variable grâce à une technique de rééchantillonnage conjuguée à l'utilisation d'une vraisemblance composite par paire. La seconde s'intéresse à la détection de changements passés de la taille de la population grâce à l'information contenue dans l'homozygotie haplotypique. Pour celà nous proposons un critère de choix de modèle pénalisé par des indices de sensibilité. Ces contributions seront précédées de quelques définitions de génétique utilisées dans toute la suite du manuscrit.

1.1 Définitions de génétique

Commençons par définir les objets de génétique qui seront utilisés dans la suite de ce manuscrit. Quel que soit l'organisme considéré, l'ADN est transmis au cours de la reproduction : il est le support de l'hérédité. Les molécules d'ADN sont formées de

deux brins enroulés l'un autour de l'autre pour former une double hélice. De façon simplifiée, chacun des deux brins de cette double hélice est constitué de nucléotides formés d'une base azotée parmi l'adénine (A), la cytosine (C), la guanine (G) ou la thymine (T). Chaque couple de ces bases avec son complémentaire (A et T, C et G) forme une paire de bases.

L'ordre dans lequel se succèdent les nucléotides le long d'un brin d'ADN constitue la séquence de ce brin. C'est cette séquence qui porte l'information génétique. On appelera gène la copie d'une information génétique. Un organisme diploïde présente deux copies de la même information génétique, une héritée de chaque parent, alors qu'un organisme haploïde n'en contient qu'une.

Le génome d'un grand nombre d'espèces n'est pas entièrement inclus dans le noyau de chaque cellule. Un brin d'ADN se trouve dans la mitochondrie. Cette partie du génome est généralement héritée uniquement de la mère. C'est pourquoi l'ADN mitochondrial a été largement utilisé dans les premières études de génétique des populations basées sur des modèles, voir par exemple le concept d'Ève mitochondriale (Griffiths and Tavaré [1994a], Marjoram and Donnelly [1997]). L'analogue de la mitochondrie chez les végétaux est le chloroplaste et le mécanisme de transmission du génome qu'il contient est encore un peu différent. Ce travail ne traite pas ce type d'ADN.

On appellera locus (ou site) une région spécifique donnée du génome. À cause de l'occurence de mutations, différentes versions d'un locus (ou d'un gène) peuvent exister, d'où la notion de variation génétique (ou de polymorphisme). Ces différentes versions sont appelées des allèles (ou états alléliques) : deux gènes sont donc dans le même état allélique si l'information qu'ils portent est codée par la même séquence d'ADN.

Plusieurs portions d'un même chromosome considérées simultanément constituent un haplotype. Les segments constituant un haplotype ne sont pas nécessairement adjacents mais peuvent former un ensemble clairsemé de loci de régions génomiques différentes.

Chaque division cellulaire est précédée d'une réplication de l'ADN conduisant à une réplication des chromosomes. Ce processus conserve normalement l'information génétique de la cellule, chacune des deux cellules filles héritant d'un patrimoine génétique complet identique à celui de la cellule mère. Il arrive cependant que ce processus ne se déroule pas normalement et que l'information génétique de la cellule s'en trouve modifiée. On parle dans ce cas de mutation génétique. Cette modification du génotype peut être sans conséquence ou, au contraire, changer également le phénotype résultant de l'expression des gènes concernés. Lorsqu'une mutation influence la valeur sélective de l'individu la portant, on parle de sélection. Une hypothèse simplificatrice importante pour modéliser la mutation sous le coalescent considère que les variations génétiques n'influent pas sur la valeur sélective des individus et donc sur la reproduction dans la population. Sous cette hypothèse, la démographie de la population n'est donc pas impactée par les mutations et la structure de la généalogie de l'échantillon est préservée. Ainsi le processus généalogique et le processus de mutation peuvent être séparés et chaque mutation peut en principe être appliquée en tenant compte des changements le long des branches de la généalogie. Ce manuscrit se place intégralement sous l'hypothèse de neutralité sélective.

De façon générale, les différents chromosomes sont physiquement séparés ; cela garantit une forme de stabilité et de pérennité de l'information génétique. Cependant les deux chromosomes parentaux ont la possibilité d'échanger des fragments d'ADN lors du processus de formation des gamètes, appelé la méïose. Ce phénomène est appelé la recombinaison génétique. Elle permet aux chromosomes d'échanger du matériel génétique et de produire de nouvelles combinaisons de gènes, ce qui accroît l'efficacité de la sélection naturelle.

Dans la suite, nous nous intéressons à l'histoire démographique d'une population au travers d'un échantillon de gènes extrait de cette population. On considèrera que la reproduction est panmictique au sein de la population, c'est-à-dire que les individus se reproduisent aléatoirement et de façon équiprobable les uns avec les autres.

Données Nous présentons ici les données génétiques utilisées pour conduire nos analyses. Un jeu de données est constitué d'un échantillon de gènes provenant d'une unique population supposée isolée. Certaines espèces sont diploïdes ; les individus portent deux copies de l'information génétique : une héritée de leur mère et une de leur père. Si la population est panmictique, cette correspondance n'apporte pas d'information et un individu diploïde peut être assimilé à deux individus haploïdes. Dans la suite, nous faisons donc plutôt référence à des copies de gènes qu'à des individus. La taille de la population fera ainsi référence à une taille de population dénombrée en nombre de copies de gènes.

Nous détaillons à présent l'information génétique contenue dans différents types de marqueurs. Il existe de nombreux types de marqueurs, les trois types les plus utilisés aujourd'hui sont les microsatellites, les Single Nucleotide Polymorphism (SNPs) et les séquences.

Dans le chapitre 3, nous nous concentrons sur des marqueurs microsatellites mais la méthode est assez générale pour être adaptée à des marqueurs de type SNP ou séquences (sans recombinaison). Un microsatellite est une portion d'ADN constituée de nombreuses répétitions d'un court motif, typiquement de longueur 1 à 6 paires de bases. L'information qui constitue le jeu de données est la longueur totale de cette suite (en nombre de répétitions). Un génome comporte des milliers de marqueurs microsatellites, cette abondance permet d'en trouver facilement. Ils ont été largement utilisés également car ils sont faciles à génotyper et très variables. Nous décrivons un modèle mutationnel classique pour ce type de locus.

Un SNP (prononcé snip) est l'occurrence d'une variation d'un unique nucléotide à une position spécifique du génome. Par exemple, une position spécifique est occupée par une base nucléotidique C chez la plupart des individus mais une minorité des individus présente une base A. Il existe donc un SNP à cette position précise et les deux variants possibles, C ou A, sont appelés allèles pour cette position. Bien que dans cet exemple uniquement deux allèles sont possibles, il existe aussi des SNPs pour lesquels trois ou quatre variants coexistent dans la population.

Les techniques de séquençage de nouvelle génération nous donnent accès à des données de plus en plus larges, de types séquences. Ainsi nous n'observons plus seule-

ment les positions présentant différents variants mais toute la séquence des nucléotides avec leur ordre dans le génome. Ces séquences portent donc de l'information sur la recombinaison, ce que nous n'avons pas en général avec de simples SNP (sauf pour les espèces pour lesquelles une carte génétique est disponible). Dans cette thèse, nous n'avons pas traité de données de type SNPs. Le chapitre 4 présente une méthode de détection d'une contraction passée de la taille de la population basée sur des données de type séquence pour une espèce dont le génome est connu. Elle exploite l'information du déséquilibre de liaison entre loci en modélisant la recombinaison.

Dans ce manuscrit et sauf mention contraire, nous étudierons l'histoire démographique d'une unique population supposée isolée et en particulier aucun événement de migration ne sera pris en compte.

1.2 Estimation de la vraisemblance de modèles de taille de population variable par échantillonnage préférentiel

La méthode du maximum de vraisemblance est une méthode d'inférence classique des paramètres de la distribution de probabilité d'un échantillon. Dans le cadre de la première partie de cette thèse, nous souhaitons estimer les paramètres d'un modèle démographique par maximum de vraisemblance. Pour ce faire, nous devons calculer la vraisemblance des paramètres du modèle, pour un échantillon de copies de gènes donné. Dans le cas de variables aléatoires indépendantes et identiquement distribuées, la fonction de vraisemblance du paramètre ϕ associée à l'échantillon observé $\mathbf{x} = x_1, \ldots, x_n$ s'écrit :

$$L(x_1, \ldots, x_n|\phi) = \Pi_{i=1}^{n} f(x_i|\phi),$$

où $f(x_i|\phi)$ est la densité de probabilité de x_i sachant ϕ. On considère à présent un problème à variable latente ou cachée, c'est-à-dire qu'une partie du modèle n'est pas observée. La densité de probabilité de l'échantillon dépend ainsi d'une variable h non observée dite variable latente. La vraisemblance s'écrit alors comme une somme sur toutes les valeurs possibles de la variable latente h :

$$L(\mathbf{x}|\phi) = \int_{h \in \mathcal{H}} f(\mathbf{x}, h|\phi) \mathrm{d}h. \tag{1.1}$$

Il n'est pas toujours possible de calculer cette intégrale explicitement. En particulier, lorsque l'espace $\mathcal{H}$ est de très grande dimension et de structure complexe, il est exclu de parcourir tous les éléments qu'il contient. Le problème d'inférence statistique dans des modèles à variable latente dont l'espace est complexe et de grande dimension a donné naissance à de nombreuses techniques de probabilités numériques et d'optimisation. On distingue les algorithmes de recherche du maximum de vraisemblance dans le cadre de l'approche fréquentielle et les schémas MCMC pour l'approximation de la loi a posteriori pour l'approche bayésienne. Dans cette thèse, nous considérons le cas où l'espace de la variable latente h est de très grande dimension et comporte des directions de différentes natures (discrètes et continues) ce qui rend l'intégrale de

l'Eq. 1.1 incalculable et on s'intéresse à une méthode fréquentielle d'inférence dans ce cas.

Le champ d'application privilégié de cette thèse est l'inférence dans les modèles de génétique des populations. Un modèle de génétique des populations est contraint par un scénario démographique qui décrit le fonctionnement de la démographie actuelle et passée. Les généticiens des populations s'intéressent à l'inférence de ces quantités. En effet, obtenir des informations sur les paramètres qui agissent sur la diversité génétique permet de mieux gérer les espèces, en assurant par exemple la biodiversité. Pour mener leurs études, les généticiens des populations se basent sur l'information génétique extraite de la population le plus souvent au temps présent. Cette information est représentée par la variable observée $\mathbf{x}$. L'information incluant l'histoire temporelle et généalogique de la population considérée est inaccessible en général et constitue la variable latente h du modèle. La prise en compte de modèles relativement complexes rend l'exploration de l'espace des valeurs prises par le paramètre et de l'espace de la variable latente plus difficile.

Pour maximiser la vraisemblance d'un modèle dépendant de variables latentes comme c'est le cas ici, il est possible d'appliquer un algorithme Espérance-Maximisation (EM) de Dempster et al. [1977]. Il comporte une étape E d'évaluation de l'espérance de la vraisemblance complétée puis une étape M de maximisation de la fonctionnelle déterminée à l'étape E. On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance et l'on itère.

Si l'on considère un échantillon i.i.d. $\mathbf{x} = (x_1, \ldots, x_n)$ de loi $f(x_i|\phi)$ paramétrée par ϕ, on cherche à déterminer le maximum de la vraisemblance donnée par :

$$L(\mathbf{x}; \phi) = \prod_{i=1}^{n} f(x_i|\phi).$$

La vraisemblance complétée d'un modèle dépendant d'une variable latente h est :

$$L(\mathbf{x}, \mathbf{h}; \phi) = \prod f(x_i, h_i|\phi)$$

L'étape E consiste alors à évaluer l'espérance de la vraisemblance complétée :

$$E_{\mathbf{h}|\mathbf{x},\phi^{(c)}} \left[L(\mathbf{x}, \mathbf{h}; \phi)\right] \tag{1.2}$$

où $\phi^{(c)}$ est la valeur courante du paramètre. Pour des modèles de coalescent en génétique des populations, les calculs analytiques de l'étape E ne sont pas possibles en général car nous n'avons pas accès à la loi de $\mathbf{h}$ sachant $\mathbf{x}$. Des variantes stochastiques de l'EM ont été développées pour remplacer ces calculs par des approximations (voir par exemple SEM de Celeux and Diebolt [1985], SAEM de Delyon et al. [1999] et MCEM Wei and Tanner [1990]). Cependant la difficulté d'exploration de l'espace de la variable latente pour des modèles complexes rend les estimations de l'espérance de l'Eq. 1.2 trop variables.

Un estimateur Monte-Carlo naïf consisterait à tirer des réalisations de la variable latente h suivant la loi $p(.)$ du processus et à estimer la vraisemblance par :

$$\widehat{L(\mathbf{x}|\phi)} = \frac{1}{N} \sum_{j=1}^{N} p(\mathbf{x}, H^{(j)}|\phi), \quad \text{où } H^{(j)} \sim p(.). \tag{1.3}$$

En pratique, les réalisations d'histoires $H^{(j)}$ sont simulées d'après le coalescent de Kingman en partant d'un échantillon de lignées génétiques (autant que dans l'échantillon observé $\mathbf{x}$ mais sans information sur les types alléliques) et en remontant jusqu'à obtenir une unique lignée ancestrale. On parle de simulation *backward* de la généalogie de gènes, illustrée par la figure 1.1. À chaque instant, deux des lignées présentes dans l'état courant de l'histoire généalogique peuvent coalescer en une seule. Puis lorsque cette histoire ne présente plus qu'une lignée ancestrale, la construction de la généalogie est terminée (on a ainsi simulé la topologie de l'arbre de gène). On ajoute ensuite aléatoirement les événements de mutations sur les branches. On dérive finalement les génotypes composant l'échantillon final que l'on vient de simuler à partir du génotype ancestral (tiré uniformément au hasard). Si la composition génétique de cet échantillon simulé ne correspond pas à l'échantillon observé alors $p(\mathbf{x}, H^{(j)}|\phi) = 0$ et on dira que la réalisation $H^{(j)}$ n'est pas compatible avec les données observées. En pratique, la probabilité que la réalisation $H^{(j)}$ soit compatible avec les données observées $\mathbf{x}$ (et ainsi que $p(\mathbf{x}, H^{(j)}|\phi)$ soit non nulle) est tellement faible que cet estimateur est fortement variable.

Griffiths and Tavaré [1994a] proposent d'approcher cette vraisemblance grâce à une technique d'échantillonnage préférentiel. Pour estimer la vraisemblance du paramètre ϕ conditionnellement à l'échantillon observé $\mathbf{x}$, on échantillonne l'espace latent $\mathcal{H}$ suivant une loi auxiliaire, absolument continue par rapport à $p(\mathbf{x}, .|\phi)$. Elle sera notée $q(.)$ et appelée loi d'importance.

$$L(\mathbf{x}|\phi) = \int_{h\in\mathcal{H}} \frac{p(\mathbf{x}, h|\phi)}{q(h)} q(h)\mathrm{d}h = \mathbb{E}_{H\sim q}\left(\frac{p(\mathbf{x}, H|\phi)}{q(H)}\right) \tag{1.4}$$

L'intégrale est alors vue comme l'espérance du rapport des vraisemblances sous la loi d'importance q. À ce titre, nous pouvons l'approcher par un estimateur de type Monte-Carlo :

$$\widehat{L(\mathbf{x}|\phi)} = \frac{1}{N}\sum_{j=1}^{N} \frac{p(\mathbf{x}, H^{(j)}|\phi)}{q(H^{(j)})}, \quad \text{où } H^{(j)} \sim q.$$

Cette fois-ci, la loi $q(.)$ est choisie de sorte que toutes les réalisations de la variable latente sont compatibles avec les données observées et ainsi $p(\mathbf{x}, H^{(j)}|\phi)$ est toujours non nul. En effet, la loi d'importance nous permet de simuler des histoires en partant de la composition génétique de l'échantillon observé et en tirant ensuite aléatoirement les événements de coalescence et de mutations jusqu'à obtenir une seule lignée ancestrale. Cette simulation *backward* de la généalogie en partant de l'échantillon observé est illustrée par la figure 1.2.

La première difficulté classiquement rencontrée avec cette technique est de calibrer la loi d'importance $q(.)$ de sorte que la variance du rapport $p(\mathbf{x}, .|\phi)/q(.)$, et donc de l'estimateur Monte-Carlo, soit faible. Déterminer la loi $q(.)$ optimale est aussi complexe que le problème initial de calcul de la vraisemblance. En revanche, nous disposons de lois d'importance bien calibrées pour une classe de modèles de génétique des populations (voir Stephens and Donnelly [2000], [De Iorio and Griffiths, 2004a], [De Iorio and Griffiths, 2004b], [De Iorio et al., 2005]). Elles sont idéales au sens où elles coïncident avec la loi optimale dans le cas d'un modèle mutationnel et démographique très simples. Cependant, pour le cas d'une population dont la taille varie au cours

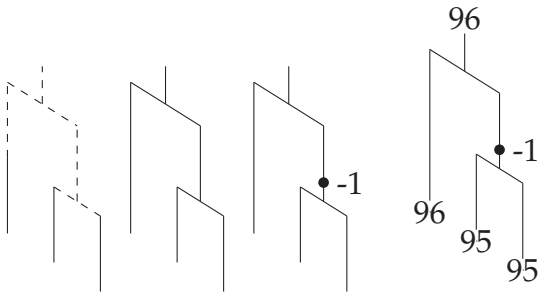


FIGURE 1.1: *Simulation* backward *d'une généalogie de gènes, sans tenir compte d'un échantillon observé. Les différentes étapes de la simulation sont (de gauche à droite) la construction de la topologie de l'arbre, l'ajout des mutations uniformément sur l'arbre, le tirage du type allélique ancestral et la description des types alléliques présents dans l'échantillon.*

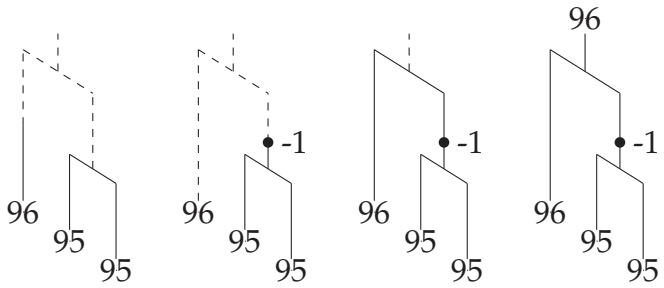


FIGURE 1.2: *Simulation* backward *d'une généalogie de gènes, en partant d'un échantillon observé. Les différentes étapes de la simulation sont (de gauche à droite) la construction de la topologie de l'arbre et l'ajout des mutations en même temps, la description du type allélique ancestral.*

du temps, nous ne disposons pas d'une loi d'importance bien calibrée. Les arguments théoriques ayant permi d'obtenir les lois d'importance pour les modèles homogènes dans le temps ne sont plus valables lorsque la taille de la population varie. Nous utilisons donc la loi d'importance calibrée pour le modèle homogène et l'adaptons au calcul de notre estimateur Monte-Carlo. L'inhomogénéité en temps du modèle entraîne la dégénérescence des poids d'importance : un grand nombre de ces poids est proche de zéro et quelques un seulement prennent des valeurs plus importantes. La variance du rapport de vraisemblance tend alors vers l'infini exponentiellement vite.

Ce constat a motivé la première partie de cette thèse. Elle vise à limiter la dégénérescence des poids en appliquant une technique de rééchantillonnage mettant à profit l'information portée par la vraisemblance composite par paires de l'échantillon observé. Dans le chapitre 3 de cette thèse, nous mettons en évidence la réduction du coût de calcul obtenu grâce à cette technique de rééchantillonnage sur des jeux de données simulés et sur un jeu de données de chauve-souris. Non seulement la présence de cette vraisemblance composite dans la distribution de rééchantillonnage permet un gain plus important en précision d'estimation de la vraisemblance mais c'est aussi une nouveauté qui a une portée plus large que notre cadre de travail. L'idée est d'utiliser la vraisemblance composite par paires pour capter le devenir de la probabilité d'une histoire en cours de construction sans avoir besoin de la simuler effectivement entièrement.

1.3 Choix de modèle pour la détection de contraction passée de la taille de la population

Le séquençage des génomes complets d'organismes de plus en plus divers (bactéries, animaux, humains...) ouvre des perspectives nouvelles dans le domaine de la génétique des populations. Dans ce travail, nous montrons comment on peut tirer profit de longues portions d'ADN pour estimer des tailles de populations. Pour ce faire, nous utilisons la distribution empirique des longueurs de segments identiques entre deux haplotypes, appelée homozygotie haplotypique. Ces quantités portent beaucoup d'information sur l'histoire démographique de la population et en particulier sa taille efficace.

Dans le cas d'un organisme eucaryote diploïde, le génome d'un seul individu, comme mosaïque des génomes de tous ses ancêtres, reflète l'histoire démographique de la population. Avec cet unique génome comme donnée pour inférer les tailles efficaces historiques de la population, nous proposons une méthodologie basée sur la distribution des longueurs de séquences conservées entre deux haplotypes, appelée Homozygotie Haplotypique (notée *HH*).

Les méthodes de maximum de vraisemblance (MCMC, IS) et les méthodes bayesiennes approchées (ABC) sont adaptées pour des jeux de données de polymorphisme de quelques loci, celles basées sur Sites Frequency Spectrum (SFS) peuvent traiter un grand nombre de SNPs. Cependant, les unes comme les autres supposent l'indépendance des généalogies des différents loci. Pour tirer parti de l'information portée par la recombinaison, nous devons considérer la dépendance des généalogies des positions adjacentes dans le génome.

Une famille de méthodes basées sur la structure des haplotypes considère le déséquilibre de liaison (LD, de l'anglais Linkage Disequilibrium) c'est-à-dire la corrélation entre les généalogies des sites voisins. Elles tirent donc parti de l'information de la recombinaison présente dans des données denses de type génome complet. Ainsi, les loci ne sont plus supposés indépendants et la vraisemblance des données ne s'écrit plus comme le produit des vraisemblances pour chaque locus. Contrairement à la mutation, la recombinaison affecte la structure de l'arbre généalogique de l'échantillon. Lorsque l'on modélise la recombinaison, la vraisemblance prend alors la forme d'une intégrale sur l'espace de tous les graphes de recombinaison ancestraux (ARG, de l'anglais Ancestral Recombination Graph) possibles pour l'échantillon de séquences (Griffiths and Tavaré [1994a]). Cet espace est de dimension bien plus grande que celle de l'espace des généalogies de gènes, et il devient plus difficile encore de mener l'inférence en utilisant les méthodes d'inférence basées sur la vraisemblance, sans plus de simplifications. Une analyse des données sous un modèle de coalescent complet n'est pas envisageable en pratique pour un grand nombre de marqueurs en considérant la recombinaison. Deux principales approches ont été développées, d'une part des méthodes basées sur l'approximation du coalescent avec recombinaison : le coalescent séquentiel avec modèles de Markov cachés (PSMC et coal-HMM) et d'autre part des méthodes basées sur les longueurs de segments identiques entre deux haplotypes.

La méthode Pairwise Sequential Markovian Coalescent (PSMC) de Li and Durbin

[2011] utilise un modèle de Markov caché pour approximer la dépendance des temps de coalescence des deux haplotypes ente des loci voisins. Elle a été étendue par Mailund et al. [2011] et des variantes de cette approche ont été développées (Burgess and Yang [2008], Schiffels and Durbin [2014]). Au lieu d'échantillonner parmi les généalogies de gènes, la méthode Coalescent Hidden Markov Model (Coal-HMM) de Hobolth et al. [2007] considère la généalogie non observée de chaque position génomique comme un état latent d'un modèle de Markov caché. Les temps de calcul deviennent importants pour ces méthodes lorsque l'on inclut de multiples séquences.

L'approche alternative modélisant la recombinaison est basée sur la distribution des longueurs de segments identiques. Deux segments d'ADN sont dits identiques par état (IBS, de l'anglais Identical By State) s'ils présentent les mêmes suites de bases dans ce segment alors que deux segments sont dits identiques par descendance (IBD, de l'anglais Identical By Descent) s'ils sont dérivés d'un même ancêtre commun sans événement de recombinaison, c'est-à-dire que ces deux segments ont la même origine ancestrale. Une définition alternative considère que deux segments IBD sont dérivés d'un même ancêtre commun sans événement de recombination. Comme détaillé plus bas, les segments courts sont principalement affectés par le passé ancien alors que les segments longs sont principalement affectés par le passé récent. À cause de la difficulté à détecter de courts segments IBD, la distribution des longueurs de segments IBD est plutôt utilisée pour inférer la démographie passée récente. Palamara et al. [2012] expriment cette distribution comme une fonction de la démographie de la population et dérivent une procédure d'inférence pour reconstituer cette histoire démographique. Ils testent des modèles paramétriques de plus en plus flexibles et utilisent un critère d'information d'Akaike (AIC, de l'anglais An Information Criterion, Akaike [1974]) pour comparer leurs modèles en tenant compte de leurs degrés de liberté. Browning and Browning [2015] proposent une méthode d'estimation non paramétrique de la taille efficace récente de la population en utilisant aussi des longs segments IBD inférés. Ils utilisent une procédure généralisée d'espérance-maximisation (EM) pour ajuster les trajectoires des tailles de populations historiques. Ringbauer et al. [2016] ont récemment utilisé une approximation de la diffusion pour remonter les lignées génétiques ancestrales dans le passé et ont dérivé des formules analytiques pour des patrons de longs blocs de segment IBD isolés par la distance, pouvant tenir compte des changements récents de la taille de la population. Leur schéma d'inférence repose sur une approche de vraisemblance pour ajuster ces formules aux quantités observées.

Bien que le statut IBD d'un segment ne soit pas observable directement, on peut observer si deux haplotypes présentent des bases identiques pour un segment d'ADN donné. Tout d'abord, Hayes et al. [2003] et MacLeod et al. [2009] ont introduit des formules analytiques de la probabilité pour une paire d'haplotypes de partager un nombre donné de positions adjacentes IBS. Modélisant la taille de la population comme une fonction constante par morceaux, leurs formules sont basées sur un coalescent approché tenant compte de la mutation et de la recombinaison. Malheureusement, les évaluations de ces formules sont coûteuses. Ces outils ont été affinés par MacLeod et al. [2013] pour inférer l'histoire de la population et ont été appliqués à des données de type génome complet. Ils ont estimé les paramètres démographiques en ajustant leur HH_{th} théorique au $\widehat{HH}$ observé sur une vache donnée. Pour valider le modèle démographique inféré, ils ont ensuite simulé des séquences sous ce modèle et les ont

comparées aux séquences observées pour une autre vache puis inversé les rôles. Harris and Nielsen [2013] présentent une méthode d'estimation conjointe des dates et intensités d'événements passés d'*admixture* ainsi que des dates de divergence de populations et de changement de taille efficace de la population. Ils ont estimé une taille de population constante par morceaux à partir de la distribution des longueurs de segments IBS en maximisant une vraisemblance composite dérivée d'une approximation markovienne du coalescent. Ces méthodes basées sur l'identité par état plutôt que par descendance s'affranchissent donc de la détection de segments IBD. Ainsi, nous choisissons ici de considérer le LD et de nous baser sur la distribution des longueurs de segments IBS entre deux haplotypes pour détecter des changements passés de la taille de la population.

Les patrons de LD entre des marqueurs polymorphiques sont façonnés par l'histoire ancestrale de la population. On peut donc inférer l'histoire passée de la population à partir de prédicteurs multiloci de LD. Plusieurs méthodes existent pour mesurer le déséquilibre de liaison entre deux haplotypes, basées par exemple sur des fréquences alléliques (voir Zhao et al. [2007]). Nous avons choisi de mesurer le LD grâce à l'homozygotie haplotypique utilisée par MacLeod et al. [2009], notée HH. Lors d'un tirage aléatoire d'un segment de longueur donnée dans la séquence complète observée, HH est la probabilité que les marqueurs de ce segment soient tous homozygotes entre les deux brins d'ADN considérés, ici les deux brins d'un même individu diploïde.

Les patrons d'homozygotie haplotypique observés sont affectés par des changements de taille de la population : plus la taille de la population est grande et plus l'homozygotie haplotypique est faible. De plus, la théorie prédit que le LD pour un segment de longueur c Morgans est principalement affecté par la taille de la population approximativement $1/2c$ générations auparavant, si l'on suppose un changement linéaire de la taille de la population. Ainsi, Hayes et al. [2003] ont montré sur des simulations que l'on peut inférer des tailles efficaces de la population à différentes dates dans le passé en utilisant des segments de chromosomes de longueurs comprises dans une large plage de valeurs.

Les méthodes précédentes se concentrent principalement sur l'inférence des changements historiques de la taille efficace de la population. Elles ne considèrent pas la complexité du modèle démographique ajusté et peuvent souffrir d'un surajustement aux données : dans de nombreux cas, un modèle démographique plus simple pourrait aussi s'ajuster correctement aux données en ne retenant que les événements les plus marquants de l'histoire de la population considérée. Même si certaines d'entre elles proposent un contrôle d'un potentiel surajustement, à notre connaissance, aucune procédure de choix de modèle entre des modèles démographiques de complexité différentes n'a été proposée à partir de longueurs de segments IBS.

Nous proposons de combler ce manque en ciblant une question simple : existe-t-il un changement dans la taille de la population au cours du temps ou bien est-elle constante ? Nous introduisons une procédure de choix de modèle entre deux modèles emboîtés : un modèle de taille de population constante et un modèle présentant un changement historique de la taille de la population. Pour ce faire, nous définissons un critère des moindres carrés pénalisé, basé sur la comparaison d'homozygoties haplotypiques observée et attendue.

Notre pénalisation repose sur des calculs d'indices de sensibilité de Sobol. Le principe est de tenir compte uniquement de la part d'erreur que le modèle a la capacité d'expliquer. C'est une forme de pénalité liée à la complexité du modèle puisqu'une erreur d'ajustement donnée du HH_{th} attendu au $\widehat{HH}$ observé pénalisera plus fortement un modèle complexe qu'un modèle plus simple. En effet, une des préoccupations majeures de l'analyse de sensibilité est d'identifier les paramètres les moins influents pour réduire la dimension du modèle et de quantifier l'erreur commise (Sobol et al. [2007], Saltelli et al. [2010]). À cet effet, elle décompose la variance de la sortie du modèle en fractions qui peuvent être attribuées aux paramètres d'entrée au travers d'indices de sensibilité de Sobol. Dans notre cas, on considère l'homozygotie haplotypique attendue comme une fonction des paramètres démographiques, supposés aléatoires pour le calcul. Les indices de sensibilité de Sobol mesurent alors la part de la variabilité de l'homozygotie haplotypique théorique qui est expliquée par chaque paramètre du modèle le plus complexe considéré. Ce critère pénalisé nous a permis de détecter des tailles de population constantes ainsi que des contractions passées de la taille de la population sur des jeux de données simulés et sur un jeu de données de vache.

Plan

Le chapitre 2 présente les modèles de génétique des populations qui constituent le socle de cette thèse. Il rappelle le modèle de Wright-Fisher ainsi que le coalescent de Kingman et ses extensions permettant notamment de modéliser les tailles de population variables au cours du temps et la recombinaison génétique. Ce chapitre introduit finalement les questions d'inférence qui nous intéressent dans la suite de la thèse à savoir l'estimation de paramètres démographiques par maximum de vraisemblance et un choix de modèle pénalisé pour détecter une contraction passée de la taille de la population.

La première contribution de cette thèse est développée dans le chapitre 3. Elle repose sur l'ajout d'une technique de rééchantillonnage dans l'algorithme d'échantillonnage préférentiel séquentiel pour estimer les paramètres démographiques d'un modèle de taille de population variable au cours du temps. En utilisant l'information portée par la vraisemblance composite par paire de l'échantillon observé, cette technique permet de limiter la dégénérescence des poids et ainsi de réduire le coût de calcul. Nous mettons en évidence cette amélioration de l'inférence sur des jeux de données simulés ainsi que sur un jeu de données de chauve-souris. Cette contribution a fait l'objet d'une publication dans *Theoretical Population Biology*, disponible en ligne.

Le chapitre 4 expose la seconde contribution de cette thèse. Elle vise à déterminer si la taille de la population est restée constante par le passé ou bien si elle a subi un changement. Cette question est abordée grâce à une procédure de choix de modèle. Le critère de choix de modèle est basé sur la comparaison d'homozygoties haplotypiques théorique et observée. L'homozygotie haplotypique permet de modéliser la recombinaison génétique et donc de tenir compte de la dépendance entre les généalogies des positions adjacentes d'une séquence d'ADN. Les indices de sensibilité de Sobol sont utilisés pour construire une pénalité tenant compte de la complexité des

modèles mis en jeu. Ce critère pénalisé nous a permis de détecter des tailles de population constantes ainsi que des contractions passées de la taille de la population sur des jeux de données simulés et sur un jeu de données de vache.

Un bref chapitre termine ce manuscrit en discutant les limites de ces travaux et en proposant des perspectives.

2 Population genetics models and variable population size

Contents

2.1 Introduction

The theoretical field underlying the analysis of population data is population genetics. This field and many of its major theoretical results are quite old. It was pioneered by three major founding fathers: Sewall Wright (1889-1988), Ronald A. Fisher (1890-1962) and J. B. S. Haldane (1892-1964) during the 1920s and 1930s. The genome of a species and the observed variations are the result of interaction among evolutionary forces such as random genetic drift, mutation, recombination, selection and migration. This thesis focuses on a single isolated population and do not consider spatial dimensions. They set the outlines for studying the genetic variation under these evolutionary forces.

Later, the contributions of Motoo Kimura (1924 - 1994) made the theory more rigorous by advanced use of diffusion theory. He caused a major shift in the biological world view introducing the neutral theory. The neutral theory postulates that most of the genetic variation observed is selectively neutral. This does not imply that new mutations cannot be selected for or against, but that such variation is normally rapidly

fixed or eliminated from the population by selection. The neutral theory does imply that only a very small fraction of new mutations are selectively advantageous. It provoked debate because it offered a much smaller role to natural selection than the most prevalent contemporary view; instead it emphasized the importance of stochastic factors such as variation in the frequency of an allele by random genetic drift.

The perspective in the field shifted further in the 1970s an 1980s when the emphasis changed from prospective (looking forward in time) to retrospective (looking backward in time) methods of analysis. This development was a natural consequence of the availability of genetic data sampled at the present time but shaped by past processes. These processes were of interest to make inference about.

The inferential analysis is retrospective: we aim to understand aspects of the sample's evolutionary past and especially of the population's evolutionary past through analysis of the present sample. Such data are collected from one or several present populations of a single species and from this sample we want to infer details about the evolutionary process that created the data.

The central approach of coalescent-based analysis is a stochastic characterization of the histories (genealogies with mutations) that relate the sequences. Evaluating the probability of a given data set then consists of two steps. First, modeling genetic drift in the population which leads to a probabilistic description of the genealogical relationship of the sampled data. Second, each genealogy will generate the data with a specific probability when combined with a model for the mutation processes.

In this chapter, we give a state of the art of mathematical models existing to describe the process generating the genetic data, with a particular interest in the models of a single population and varying size during its past history.

2.2 The Wright-Fisher model

The distribution of the genetic variability that we observe in current populations is highly influenced by the demographic history. Events such as migrations or population size changes determine how allelic frequency changes, which may differ substantially from a population to another.

Many different models of idealized populations have been developed in order to study the distribution of the genetic variation. In all cases, the goal is to simplify the relevant biological process to reach mathematical solutions while maintaining the highest possible level of realism.

The Wright-Fisher model Fisher [1930], Wright [1931] is the most important and most widely used demographic model. This simple model describes the reproductive behaviour in an isolated panmictic population. It also provides a dynamic description of the evolution of an idealized population and the transmission of genes from one generation to the next. It allows for example to verify that without mutation, the biological diversity vanishes because of drift.

To facilitate comparison of haploid and diploid models, we may assume a population size of $2N$ gene copies, corresponding to N diploid or $2N$ haploid individuals. Thus, haploid reproduction is modeled assuming $2N$ individuals. Note that other treatments of the Wright-Fisher model may assume N genes instead of $2N$ and that results therefore may differ by a factor of two reflecting this.

In the haploid model, each gene of generation t has a random number of offspring, one in average, in the generation $t+1$ which are copies of their parent gene. Each gene in generation $t+1$ will thus have one parent gene in generation t. A gene in generation t might not have any descendants in generation $t+1$ and consequently its lineage dies out.

Assumptions of the Wright-Fisher model Some assumptions are made in the Wright-Fisher model:

1. *Discrete and non-overlapping generations.* All individuals constituting the population are replaced at the same time by a new generation.
2. *The population is panmictic.* The individuals reproduce uniformly at random with each other. Population structure of any kind may greatly affect genealogies and this assumption is therefore important for analysis of many real data sets.
3. *The population size is finite and constant over time.* At each generation, the number of individuals is the same as in the previous generation.
4. *All individuals are equally fit.* At a given generation, each individual reproduce to the next generation with the same probability as others. The average number of offspring of each individual is equal to one.

A real physical population is not likely to behave reproductively as the Wright-Fisher model. First, in many species, like human, generations are overlapping. Fortunately, it turns out that this assumption is of little practical consequence. Models assuming overlapping generations (not all genes give birth or die at the same time) have been developed, for example the Moran model in which a unique individual changes (i.e. replace a random parent) at each generation. These models give probabilistically similar genealogies. Second, most real populations are not panmictic and show some form of reproductive structure, either due to geographical proximity of individuals or due to social constraints. Finally, a finite and constant population size over time is a genuine biological assumption. Important quantities of the model will be different if the population is growing, shrinking, oscillating or has gone through a transient very small size, termed a population bottleneck. In this thesis, we focus on varying population size models and thus do not respect this Wright-Fisher assumption. The size of a population under the haploid Wright-Fisher model which in some sense best approximates a given real population is called the effective population size. The effective population size reflects the number of individuals that contribute offspring to the descendant generation and is almost always smaller than the census population size. For now, we assume that the genes in the population are not recombining. This is an important assumption that needs to be relaxed when analyzing real data sets.

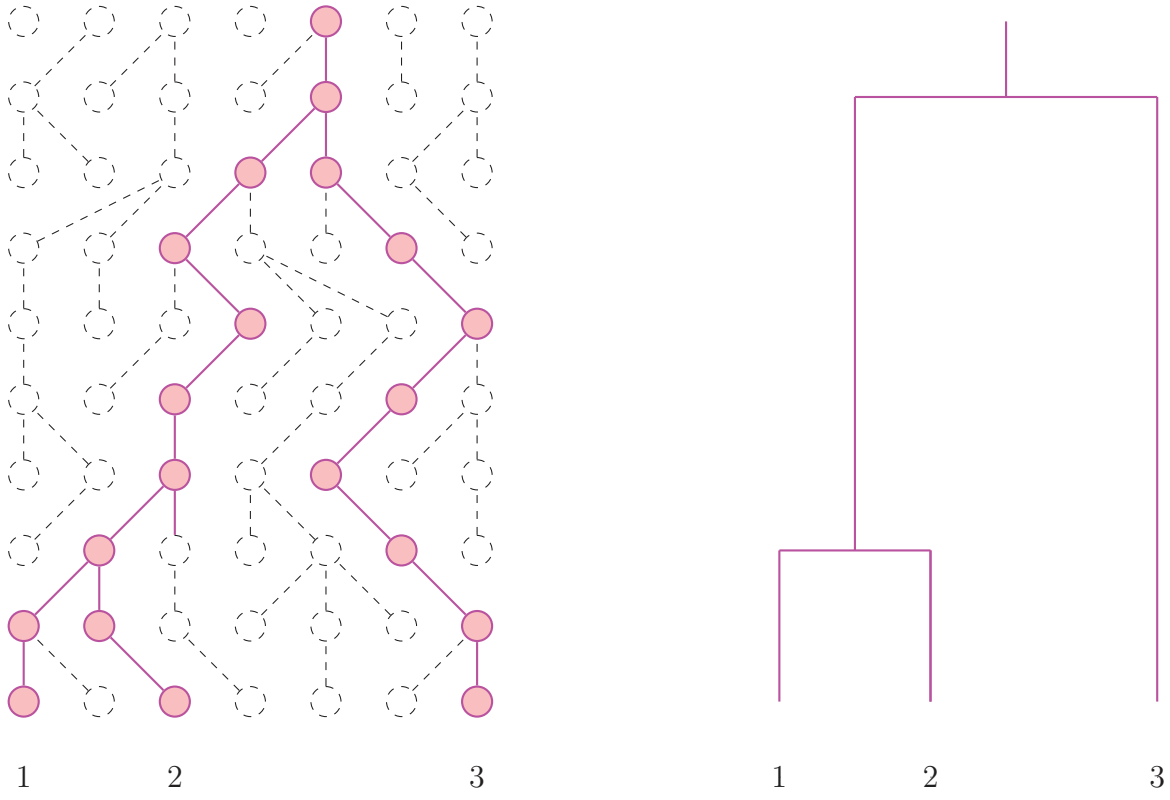


FIGURE 2.1: *The genealogy of a three randomly sampled lineages named* $1, 2, 3$ *of a Wright-Fisher population of size* 7 *(left) and the corresponding coalescent tree (right).*

2.3 The basic coalescent

Figure 2.1 represents an example of the reproduction of a Wright-Fisher population of size seven for ten generations. Each gene copy is linked to its ancestor gene copy in the previous generation. This diagram completely describes the genealogical relationships of all gene copies during ten generations. As a consequence of the Wright-Fisher model, any set of gene copies will find a common ancestor in a finit number of generations. The concept of a MRCA of the whole population of gene copies occurring at some time is a back in time statement that has a forward in time consequence: at a certain generation all the lineages starting from the $2N$ gene copies will die out except for one. To see this, sample the whole population and trace their ancestry back until their MRCA has been found. All the other genes in that generation do not have descendants in the most recent generation.

From the point of view of data analysis, only a sample of n gene copy (typically n is much smaller than $2N$) is taken from the present population and the genealogical ancestry of this sample is of interest. In Figure 2.1 three gene copies (denoted $1, 2$ and 3) have been sampled randomly in the present population. Edges back in time tracking the ancestors of these three sequences are highlighted. Two generations back in time, genes 1 and 2 find a common ancestor and this lineage might be labeled $(1, 2)$ to reflect this fact. Five generations further back in time $(1, 2)$ finds a common ancestor with 3 and the genealogical relationships of the three genes are now fully described.

The genealogy of the Wright-Fisher model can be described by the coalescent when

the population size is large.

2.3.1 Discrete time coalescent

In [Tajima, 1983], [Kingman, 1982a] and [Kingman, 1982b], Tajima and Kingman have independently formulated a stochastic process describing the genealogical dynamics emerging from several idealized population models, including the Wright-Fisher model. In the coalescent, the ancestral lineages of a sample of gene copies from a Wright-Fisher population are traced back in time, allowing a description of key genealogical events.

If we trace the ancestral lineages of two gene copies from a Wright-Fisher population back in time, repeatedly sampling a random ancestor from the previous generation, a common ancestor will be found when both lineages share the same parent. We say that these two lineages coalesce or that we observe a coalescence event.

The probability that two lineages find a common ancestor at each generation is $1/(2N)$. Indeed, the first lineage can choose any parent freely and the second one must choose the same parent as the first gene copy, which is one out of $2N$ possibilities. Thus, there is a probability of $1 - 1/(2N)$ that the two gene copies have different ancestors in the previous generation. Since sampling in different generations is independent of each other, the probability that two gene copies find a common ancestor j generations back in time is:

$$\left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}$$

Indeed, in the first $j - 1$ generations they choose a different parent and in generation j they choose the same parent. Thus, G_2, the coalescent time (in generations) for two gene copies to find their MRCA, is distributed as:

$$P(G_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}, \quad j = 1, 2, \ldots$$

which implies that G_2 is geometrically distributed with parameter $1/(2N)$. The mean of G_2 is therefore $E[G_2] = 1/(1/(2N)) = 2N$ generations. Thus, the expected time until a MRCA is the same as the number of gene copies in the population.

The waiting time for a coalescent event in a sample of k gene copies of a Wright-Fisher population can also be calculated. The probability that no coalescence occurs in the previous generation is

$$\begin{aligned}
\frac{2N-1}{2N}\,\frac{2N-2}{2N}\cdots\frac{2N-k+1}{2N} &= \prod_{i=1}^{k-1}\left(1 - \frac{i}{2N}\right) \\
&= 1 - \sum_{i=1}^{k-1}\frac{i}{2N} + O\left(\frac{1}{N^2}\right) \\
&= 1 - \binom{k}{2}\frac{1}{2N} + O\left(\frac{1}{N^2}\right),
\end{aligned}$$

where $\binom{k}{2} = \frac{k!}{2!(k-2)!}$ is the binomial coefficient. Since we assume that k is much smaller than N, $O(1/N^2)$ is negligible and can be ignored. This approximation is equivalent to ignore the possibility that more than one pair of gene copies find a common ancestor in the same generation.

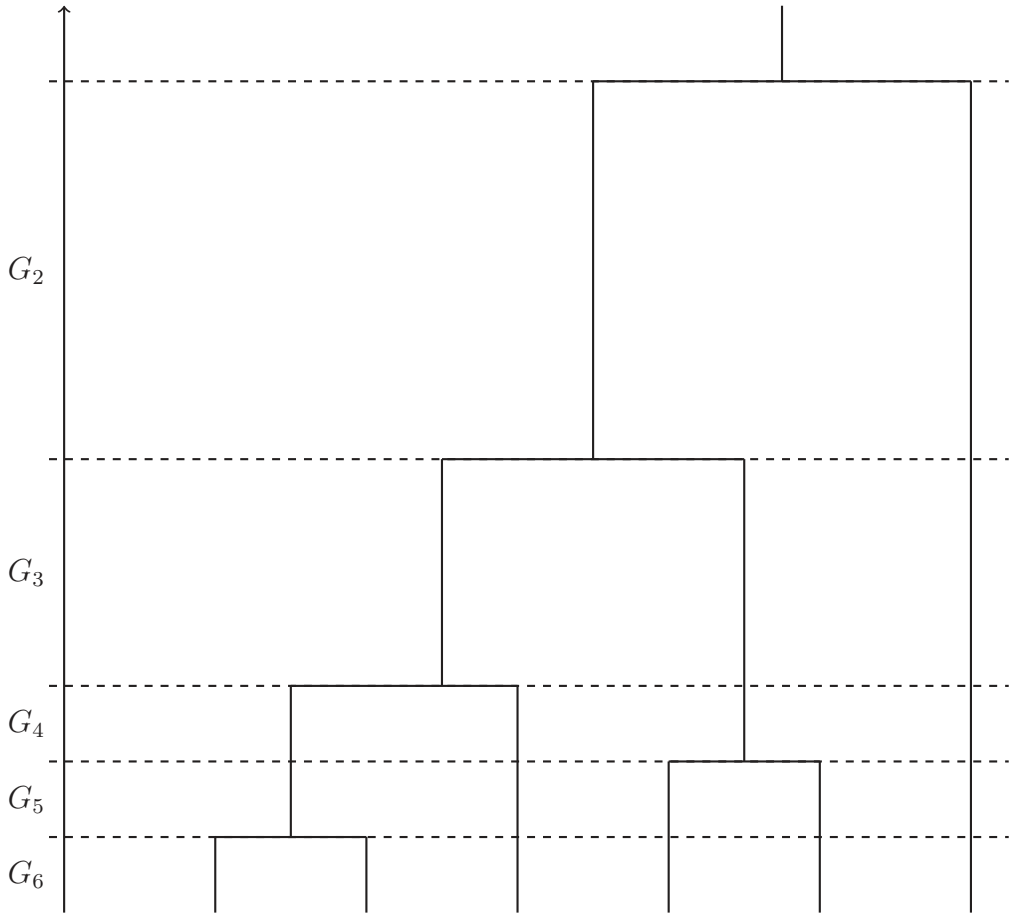


FIGURE 2.2: *An example of coalescent tree from the MRCA leading to a sample of six gene copies at time* 0. $G_k, k = 2, 3, 4, 5, 6$ *is the time while there are* k *ancestors to the sampled four gene copies.*

In consequence the probability that two gene copies out of the k gene copies find a common ancestor $G_k = j, j = 1, 2, \ldots$ generations ago is

$$P(G_k = j) \approx \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{j-1} \binom{k}{2}\frac{1}{2N}. \tag{2.1}$$

Times G_k are independent and distributed approximately as a geometric distribution with parameter $\binom{k}{2}/(2N)$. Because all pairs of gene copies are equally likely to find a common ancestor, the pair that finds a common ancestor is chosen uniformly at random among the $\binom{k}{2}$ possible pairs. Figure 2.2 represents a possible coalescent tree for a sample of six gene copies.

2.3.2 Continuous time coalescent

In the Wright-Fisher model time is measured in discrete units that is in generations. It is conceptually and computationally advantageous to consider continuous time approximations. A natural choice for the coalescent has been to scale in continuous time so that the unit of time corresponds to the average time for two genes to find a common ancestor, which was shown above to be $2N$ generations. Note that other treatments of the coalescent process prefer scaling time by N (or occasionally $4N$) rather than 2, leading to results differing by a factor of two. Using any of these transformations of time, the coalescent becomes independent of the constant population size. It will only be used if we want to translate time back into generations. This emphasizes that the structure of the coalescent process is the same for any population as long as the sample size n is small compared to the population size $2N$; only the time scale differs between populations when $2N$ differs.

To derive the continuous coalescent process we let $t = j/(2N)$, where j is time measured in generations. It follows that $j = 2Nt$ translates continuous time t back into generations j, G_k^c, the waiting time for k gene copies to have $k-1$ ancestors, geometrically distributed in discrete time, is exponentially distributed in the continuous representation, denoted $G_k^c \sim \text{Exp}\left(\binom{k}{2}\right)$, that is,

$$P(G_k^c \leq t) = 1 - e^{-\binom{k}{2}t}.$$

A continuous time realization of the coalescent process is shown in Figure 2.3, with time scaled in generations of the left and in units of $2N$ generations on the right.

Simulating the genealogy for a sample of n individuals in a population of constant size $2N$ gene copies is easy using this formulation. Given the current sample size k of gene copies at time t, we draw a coalescent time according to an exponential distribution with parameter $\binom{k}{2}\frac{1}{2N}$ and choose uniformly at random a pair of individuals to coalesce. This reflects the decreasing number of ancestral lineages as pairs of individuals find a common ancestors. The discrete version is analogous to this continuous version.

A computational advantage of going back in time when simulating the history of a present sample relative to an alternative forward approach is that in the former one only needs to keep track of the ancestry of the sample of interest, while in a forward approach the whole population needs to be traced. Referring back to Figure 2.1, a forward approach would have to calculate all edges in the whole illustration and additionally make sure that sufficiently many generations had been simulated, such that the MRCA to all extant gene copies had been found. A backward approach would only have to find the highlighted edges.

From now on we will only refer to the continuous model unless stated otherwise and the superscript c is dropped. Whenever we think of a particular population with a certain size $2N$, time can be translated back into generations by multiplication by $2N$.

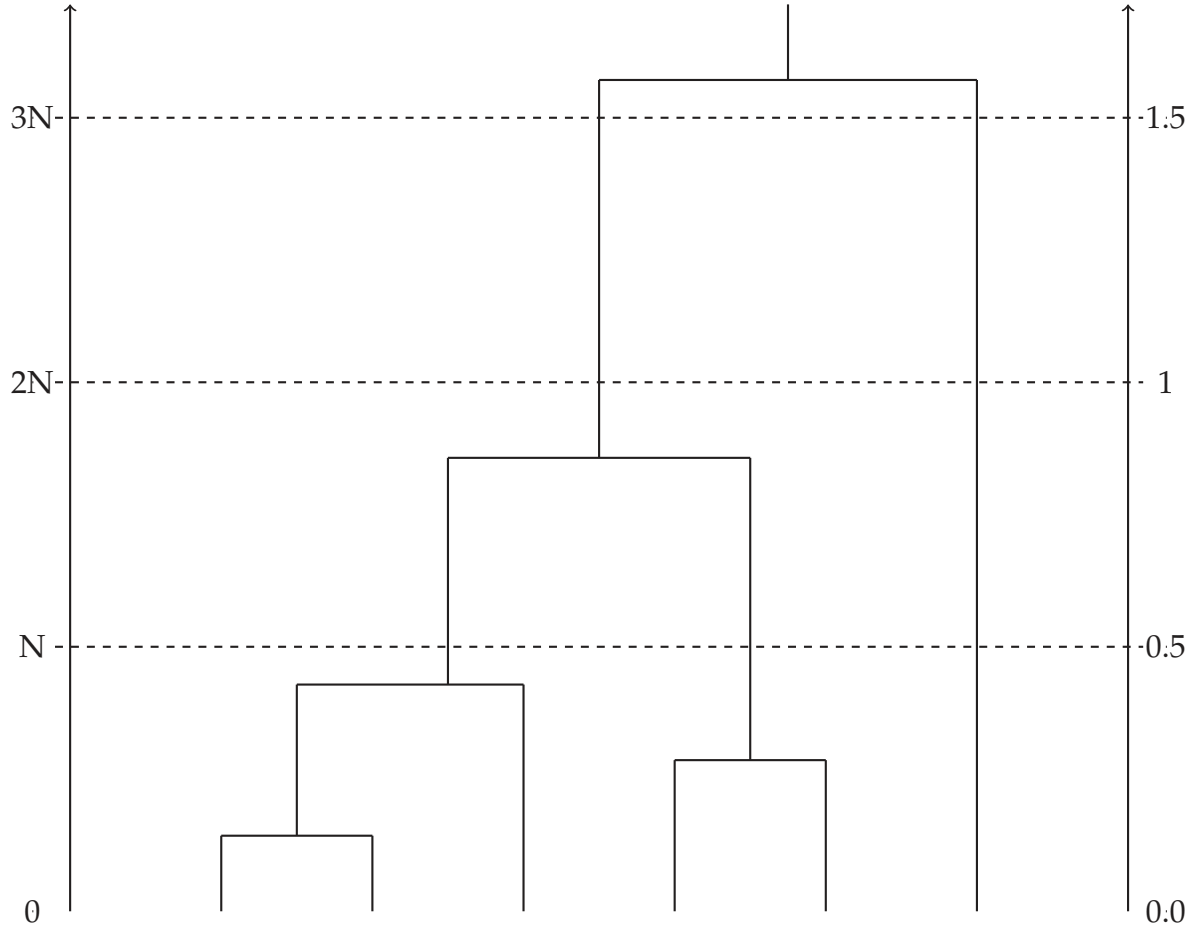


FIGURE 2.3: *A continuous time genealogy of six gene copies with time measured in units of generations (left) and in units of $2N$ generations (right).*

Height of a tree The height H_n of the coalescent tree of a sample of size n is the sum of time epochs, G_j, while there are $j = n, n-1, \ldots, 2$ ancestors. The distribution of H_n is obtained as a convolution of the exponential variables,

$$P(H_n \leq t) = \sum_{k=1}^{n} e^{-\binom{k}{2}t} \frac{(-1)^{k-1}(2k-1)n(n-1)\ldots(n-k+1)}{n(n+1)\ldots(n+k-1)}.$$

The mean of H_n is thus

$$E(H_n) = \sum_{j=2}^{n} E(G_j) = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2(1-\frac{1}{n}). \tag{2.2}$$

An infinite sample finds a common ancestors in expected time $2(1-1/\infty) = 2$, thus in finite time. Hence, in the continuous time coalescent, we can at least formally consider a sample of infinite size.

The expected waiting time for n gene copies to find their MRCA is twice that of the expected waiting time for two genes to find their common ancestor ($E(H_2) = E(G_2) = 1$). Further, for any sample size n, the variability in the time G_2 for which there are two branches in the tree accounts for most of the variability in the depth of the whole tree.

The effect of sampling more sequences It is possible to follow the MRCA of any sample of genes or the whole population further back in time. However there will never be any information in a genetic lineages sample on what happened long before there MRCA. In particular, we will hardly be able to infer past changes in the population size much more ancient than the mean Time to the MRCA (TMRCA) of the sample. We do not know the TMRCA of a given sample but the coalescent process allows to compute an expectation.

The sum in equation 2.2 grows towards 2 (scaled in $2N$ generations) as the number of gene copies increases. When the sample size is large enough, the expected tree height of the sample is 2. This indicates that the MRCA of the sample may not have undergone demographic events that occurred long before $2 \times 2N$ generations before present. This implies that inference of ancient demographic processes are limited, in particular for small population size.

Increasing the sample size has a weak impact on the power of the demographic inference. Figure 2.4 represents a continuous time genealogy of four gene copies (top) and two possible continuous time genealogies when adding a gene copy (bottom). The TMRA of the five gene copies may be the same as for only four gene copies (Figure 2.4 botttom left) or larger than for only four gene copies (Figure 2.4 botttom right). In the first previous case, most of the deep branches (those near to the root) are already in the tree with initial sample of gene copies do not bring information about the more ancient past. Indeed, the expected height for a tree for fifty gene copies is 1.96 while the expected height of a tree for ten gene copies is 1.80. Thus a fivefold increase in the number of gene copies on the average leads to less than 10% increase in tree height. Furthermore, Saunders et al. [1984] showed that the probability that the MRCA of a sample of size n is the MRCA of the whole population is $(n-1)/(n+1)$. A large sample is not required to go far back in the past.

2.4 Mutational processes

In the previous section, we discussed how genes or sequences are related in a population through their common ancestry. However, modeling genetic polymorphism requires to model how mutations cause changes in the DNA. The coalescent process is suitable to include mutation events, and therefore the distribution of genetic variation in Wright-Fisher populations.

In this section we describe how to obtain the distribution of the alleles in the sample knowing the genealogy. We describe the positions of the mutations as a realization of a Poisson process on the tree branches. Then we introduce mutation models for microsatellite loci and mutation models on sequence data.

Mutation is the bridge from genealogies to genetic data because the structure of the genealogy, in which each branch divides the sample into two groups, is revealed only if polymorphisms exist among the sampled gene copies. A series of better technologies have refined our picture of genetic variation. The first DNA sequence data set in population genetics was published by Kreitman (Kreitman [1983]) coincident with the

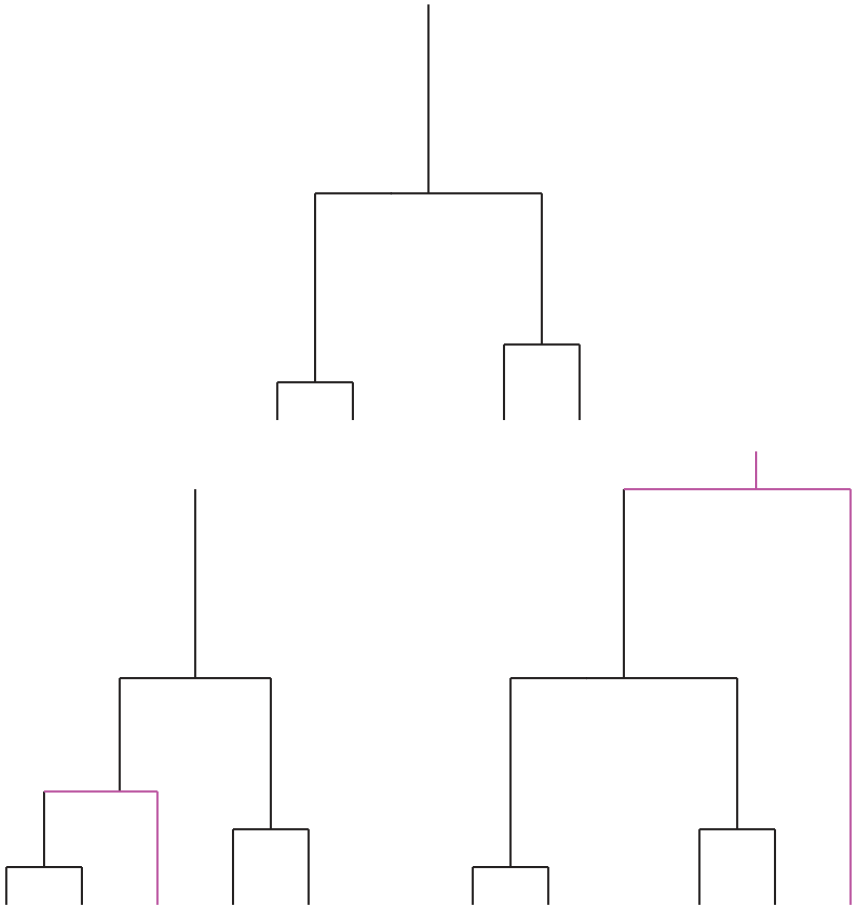


FIGURE 2.4: *A continuous time genealogy of four gene copies (top) and two possible continuous time genealogies when adding a lineage in pink line (bottom): on the left the time to the MRCA of the augmented sample is the same as for only four gene copies, on the right the time to the MRCA of the augmented sample is larger than for only four gene copies.*

development of coalescent theory.

Different mutation models have been designed for the different genetic markers. We can divide these roughly into two groups: allele-based models and DNA-sequence models. An important distinction between the two is that models for nucleotide sequences bring more information about the historical relationships among sites in the sample in the patterns of polymorphism. Here we will focus on three particular models: the stepwise mutation model and the generalized mutation model (SMM and GSM, Ohta and Kimura [1973]) and the for microsatellites and the infinite-alleles (Malécot [1946]; Kimura and Crow [1964]) and infinite-sites models (Kimura [1969]; Watterson [1975]). However, any mutation model can be implemented under the coalescent.

Remember that the important simplifying assumption in modeling mutations within the coalescent is that all variation does not provide any selective advantage or disadvantage. Under neutrality, variation by definition does not affect the reproductive success of organisms and has no influence on the structure of the genealogy of a sample. Thus, the genealogical process and the mutation process can be separated and any mutation model can in principle be applied by considering changes along the branches of each genealogy.

Mutation positions on the tree The coalescent process occurs on a time scale of $2N$ generations. In population genetics, we often consider the scaled mutation rate $\theta = 2N\mu$, where μ is the mutation rate per generation per lineage. Although the mutation rates per generation are small, assuming that N is large, the gene genealogy of a sample will offer a great number of opportunities for mutations to occur. Thus, most populations are genetically variable and polymorphims are observed even in modest samples. However, the levels of polymorphism within many organisms are low enough that multiple mutations per generation seem unlikely on a single lineage.

Conditionally on a genealogy, the positions of the mutations are approximately distributed according to a Poisson process of intensity $\mu/2$ on the genealogy. In other words, the number of mutations on a branch of length L follow a Binomial distribution of parameter $L, \mu/2$. Since μ is low and L is large, this distribution of the number of mutations on a branch of length L is approximately Poisson distribution with mean $\mu L/2$ and these mutations are uniformly distributed on this branch.

Neutral mutations create the polymorphisms that reveal underlying genealogies. However, the precision with which they do this depends on how mutations occur, or on the kind of genetic data under consideration. The kind of genetic data considered conditions the faithfulness of the unveiling of the underlying genealogies by the mutations.

Microsatellite loci Alleles at microsatellite loci are characterized by the number of repeated short motifs (1 to 6 bp) of DNA and are very common in eukaryotic genomes. They are highly mutable, with the primary mutational mechanism believed to be replication slippage [Ellegren, 2000]: during replication of the microsatellite region, the strands may be displaced and then realign incorrectly, leading to insertion or deletion of a number of repeat units. Their abundance and high variability has led to wide use in a variety of genetic analyses. The simplest model of microsatellite mutation, known as the Stepwise Mutation Model (SMM, Ohta and Kimura [1973]), assumes that length, measured as number of repeat units, changes by only 1 unit per mutation, with loss and gain equally likely. A more elaborate model is the Generalized Stepwise Model (GSM) in which the locus length is modified by a gain or a loss of X units, where X is a random variable geometrically distributed with parameter p and loss and gain equally likely.

The microsatellite data analysis of Chapter 3 is based on the SMM and can be done under the GSM.

Sequence data Sequence data contain the precise order of nucleotides within a DNA strand. These sequences reveal the recombination information that is not available with SNPs, except for species for which a reference genome is known. The Infinite Sites Model (ISM) assumes that each mutation occurs on a given sequence at a previously unmutated site. Thus, when a mutation happens under the infinite sites model without intra-locus recombination, it necessarily creates a new allele. The infinite sites model is equivalent to consider an infinitely large number of possible positions, each with a very small mutation rate. This assumption is often appropriate for DNA sequences, in

which the rate of mutation per nucleotide site is low. Since the chance of two mutations affecting the same site is inversely proportional to the number of available sites, assuming an extremely large genome results in an infinitesimal probability for this event. The Infinite Alleles Model (IAM) assumes that every time a mutation occurs it introduces a new allele into the population. Under this model, IBD and IBS segments are indistinguishable. Historically, work on the infinite alleles model was associated with investigations of identity by descent (IBD). The concept of identity by descent, which Malécot [1946] introduced, posits that two or more gene copies are descended from a common ancestor without mutation or recombination. That is, their identity is a direct reflection of their common ancestry, as opposed to the more general case of identity by state (IBS), which includes the possibility that gene copies are identical due to multiple, convergent mutations. The infinite sites model is also an infinite-alleles model, with the additional genealogical information between haplotypes.

In addition to their applicability to DNA data, the infinite sites and infinite alleles models are attractive because of the relative ease with which predictions can be generated. There is an inherent connection under these models between gene genealogy, mutation, and polymorphism, which fosters understanding about the forces that produce and maintain genetic variation in natural populations and about the ways in which genealogical structure influences patterns of polymorphism.

The work presented in Chapter 4 rely on the infinite sites model associated with the concept of haplotypic identity by state.

2.5 Varying population size

The basic coalescent process is based on the Wright-Fisher model, assuming constant population size, random mating, non-overlapping generations and absence of selection. These are strict conditions, which are rarely met in nature. Fortunately, the coalescent framework can be extended to include simple deviations from these assumptions, resulting in a process that is believed to model the true demography more closely. In this section we consider in turn how to deal with varying population size. This section also point out some major differences between genealogies generated under the basic coalescent and under varying population size models. Throughout this section, we assume that there is no recombination, unless stated otherwise.

Populations commonly fluctuate in size over time. Changes can either be extrinsic (due to the environment) or intrinsic (due to competitive ability of the species). From a modeling perspective, we distinguish between stochastic and systematic changes. Systematic changes are trends over generations in the population size to change in a certain way. Systematic trends are modeled deterministically such that the population size at time t is given by $N(t)$, a function of t only. In addition, a stochastic term might be added to $N(t)$ to reflect random deviations from $N(t)$ caused by environmental factors. Random deviations from the trend are assumed to be of minor importance compared to the trend itself. In this case, the effective population size is computed as the harmonic mean of different past population sizes: $\frac{1}{Ne} = \sum_t \frac{1}{N(t)}$. Because it is do-

minated by smaller terms, it provides the best means for summarizing fluctuations in population size when events characterized by a small value are of great biological significance. These deviations might therefore be ignored and we make this assumption throughout the remainder of this thesis.

Genealogical effect of population size contractions We only consider fully deterministic changes, that is the effective population size at time t is given by the function $N(t)$ and set $N = N(0)$. Recall that the natural time scale of the basic coalescent process is in units of $2N$ generations, that is one generation counts $1/(2N)$ units in the continuous time analogue. This is also the probability by which two gene copies find a common ancestor in the previous generation in the discrete time coalescent. If the population size changes over time, things are different. The probability that two gene copies coalesce in the previous generation is $p(t) = 1/(2N(t))$ which might change over time, and the natural scale of the coalescent process is $2N(t)$ locally. To simulate and describe genealogies probabilistically under scenarios with changing population sizes, changes in $p(t)$ have to be taken into account. When $N(t)$ is larger than N, for example if the size of the population is increasing backward in time, the probability of coalescence event decreases backward in time and a MRCA is found less rapidly than if $N(t)$ is constant over time equal to $N(0)$. This implies that a typical coalescent genealogy under a contraction has relatively longer branches closer to the root or shorter terminal branches than a genealogy under constant population size. The shape of the genealogy is affected by the past population sizes.

Figure 2.5 illustrates two demographic models: (a) the population size is constant, (b) the population size was constant and has quickly jumped from being very large to being very small at some time in the past. Given a sample of 100 gene copies of this population, we simulated a possible genealogy using the R package `coalesceR` developped by Renaud Vitalis and available online at https://r-forge.r-project.org/projects/coalescer/. Figure 2.6 represents six realizations from the coalescent for 20 gene copies sampled in a population under both previous models. More precisely, we simulate three genealogies under (a) a constant population size equal to 500 gene copies represented and (b) a sudden decrease of the population size 1500 generations before present from 5000 gene copies to 500 gene copies in present day represented. In the constant demographic scenario, the expected height of a coalescent tree is $2N = 1000$, as observed on Figure 2.6a. By contrast in the contraction scenario the height of the tree is expected to be larger because of the larger ancestral population size. If the TMRCA of the 20 gene copies sampled is found more recently than the size change, then the coalescent tree is not affected by the contraction, as we can see on the coalescent tree in the middle of Figure 2.6b. But if their TMRCA is older than the contraction, the branches near to the root tend to be longer as we can see on the coalescent trees in the left and right of Figure 2.6b. It is not surprising since the expected height of a coalescent tree of a constant population size equal to 5000 is $2N = 10000$ generations.

More generally, the conclusions we can draw about the ancestry of the sample of genes depends on the relative sizes of the population before and after the contraction (from the past towards the present), and on when the contraction happened. With a contracting population, the terminal branches tend to be shorter and branches close to the root longer than under a standard coalescent, with a growing population, the

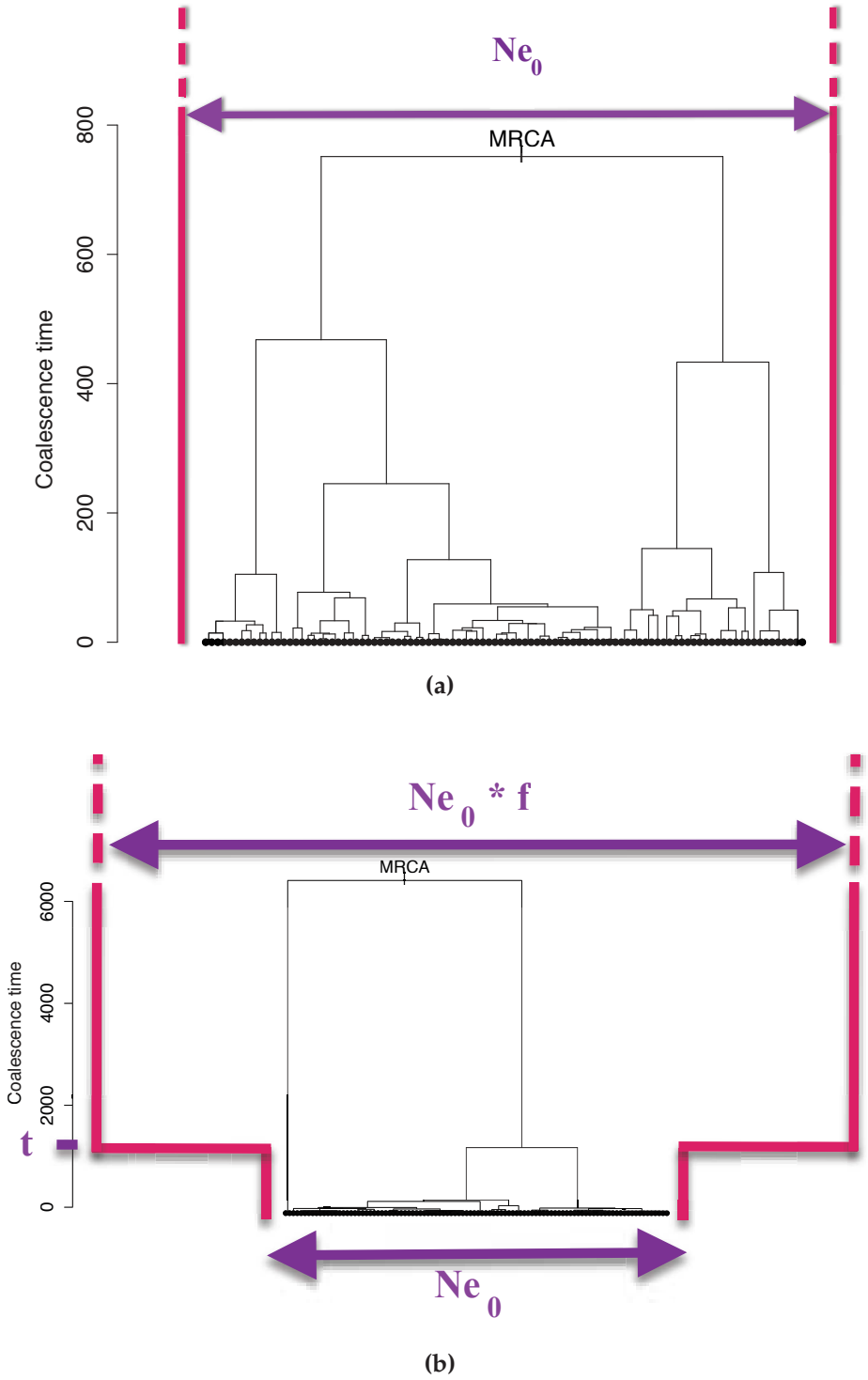


FIGURE 2.5: *A population with a (a) constant size, (b) sudden large decrease of size (contraction). The width between the pink lines represents the size of the population at a given time in the past. The gene tree between the pink lines represents the genealogy, strongly influenced by the demography, of a given sample of* 100 *individual among the whole population.*

opposite holds. If the contraction happened too far back in time, the genealogy of a sample will look like the genealogy in the first case because the most recent ancestor of all genes in the sample will be found more recent than the contraction. At the other extreme, if the contraction happened in recent times and the population size is rela-

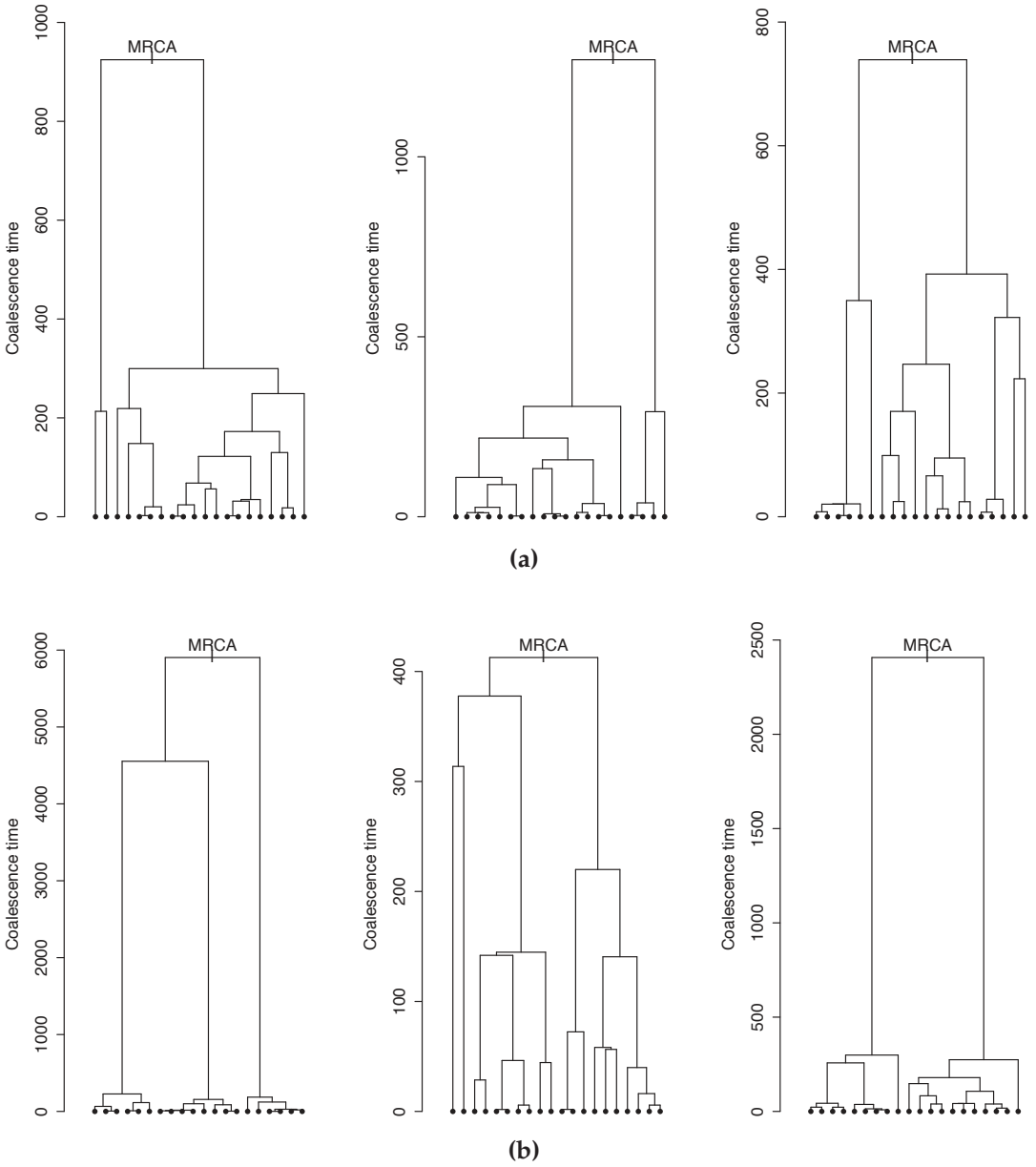


FIGURE 2.6: *A sample of six realizations from the coalescent relating twenty gene copies sampled in a population with a (a) constant population size, (b) sudden large decrease of the population size (contraction). The Y-axis represents coalescent time in the unit of generations before present.*

tively small after the contraction, most lineages would collapse rapidly just after the contraction.

The lesson from these simple examples is that the demography affects the tree shape and then the mutations that occur on the branches and consequently the genetic polymorphism observed in a sample. Given an observed data set, mathematical models will give probabilities as a function of underlying parameters describing population history. Using coalescent methods it is possible to develop statistical methods for inferring ancestral processes and to test hypotheses based on the analysis of real data.

The relevant questions are both qualitative and quantitative in nature. Examples of

qualitative questions are: does the data show sign of population structure, recombination, population growth or decline, or selection ? Quantitative questions aim at estimating parameters, such as current and ancestral population sizes, recombination or migration rate. Several studies have demonstrated that population structure can leave genomic signatures similar to those of population size changes. Existing approaches infer population size histories jointly in multiple populations, while accounting for the history of divergences and migrations in these populations. ABC represents a perfect framework for developing such approaches, because of the flexibility offered by the simulation procedure. It is already widely used in population genetics for estimating parameters in multiple population models including for instance admixture events and some population size changes Cornuet et al. [2008]. In this thesis, we focus on detection of past changes in population size and demographic parameter inference in particular under contracting population size models for a single isolated population.

2.6 The coalescent with recombination

Recombination potentially occurs in most sequences and carries a lot of demographic information. To conclude this overview of the coalescent process, we thus include the modeling of recombination events along the sequences during the genealogical process, introduced in Hudson [1983]. While mutations modify a given site, recombination events may occur inbetween any pair of sites at any transmission of the genetic material (provided the recombination rate between these sites is non zero). As mutation does not affect the tree structure of the genealogy, however, recombination does. Unfortunately, relaxing the assumption makes analysis much more mathematically complex, mainly because the sequence sample is no longer related by a genealogical tree but rather a graph (the ancestral recombination graph see Section 2.6) or a collection of trees.

Genetic recombination occurs in most organisms on earth, including eucaryotes, bacteria, and viruses, by quite different mechanisms in these three different types of organisms. No generally accepted theory is available to explain why genetic recombination is so common, but in its favor are arguments that genetic recombination ensures that favorable variants are brought together in the same sequence and that recombination can maintain more variation which may enable survival over an evolutionary time scale. The simple coalescence process, including various extensions like varying population size, population structure and different kind of selection, assumes no recombination and this theory can therefore only be used on non-recombining sequences. Fortunately, it is possible to incorporate recombination into the process as shown by Hudson [1983] very shortly after the coalescent process was first formulated. However, recombination adds much more complexity than any of the extensions mentioned above, mainly because no single coalescent tree can describe a sample of recombining sequences.

Consider n sequences of length s sites. Assume recombination occurs at the same rate along the sequence, if a recombination event occurs, the exact location can be randomly sampled along the sequence. It is possible that, for long chromosomal regions

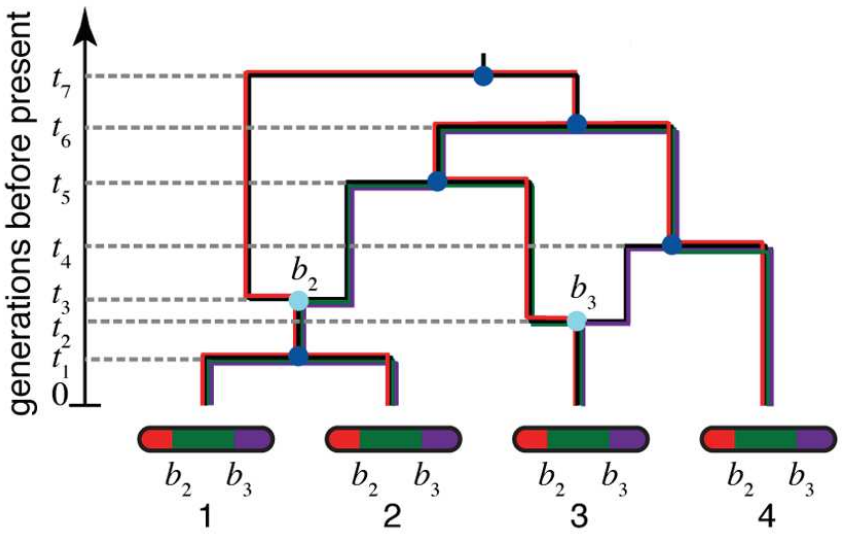


FIGURE 2.7: *Going backwards in time (from bottom to top), the graph shows how lineages that lead to modern-day chromosomes (bottom) either coalesce into common ancestral lineages (dark blue circles), or split into the distinct parental chromosomes that were joined (in forward time) by recombination events (light blue circles). Each coalescence and recombination event is associated with a specific time (dashed lines), and each recombination event is also associated with a specific breakpoint along the chromosomes (here b_2 and b_3). Each non-recombining interval of the sequences (shown in red, green, and purple) corresponds to a local tree embedded in the ARG (shown in matching colors). Recombinations cause these trees to change along the length of the sequences, making the correlation structure of the data set highly complex.*

that have high recombination rates, more than one recombination occurs during one generation. Again, however, we measure time in the continuous space, so that only one recombination event is allowed to occur at a time. The effect of a recombination event occurring between the sites s_i and s_{i+1} is to break the ancestral lineages that we are tracing backwards in time. This creates two lineages, one harboring the ancestral material in the range $[1, s_i]$, and the other carrying the ancestral material in $[s_{i+1}, s]$. After a recombination event occurs, the number of ancestral lineages being traced increases by one. This turns the genealogical structure representing the sample's genetic history from a tree into a graph, as shown in Figure 2.7 (Figure extracted from Rasmussen et al. [2014]). This graph structure introduced in Griffiths and Marjoram [1997] is called the ancestral recombination graph (ARG). It can take very complex forms for large sample sizes and long genomic regions. While some quantities may still be derived analytically, the ARG is a fairly complex mathematical object, and it often requires the use of numerical sampling for its use in quantitative analyses. As for the case with no recombination, sequences can be generated after having sampled an ancestral recombination graph, by introducing mutations over the graph edges.

The presence of recombination makes some estimators of evolutionary parameters more accurate that when applied on non-recombining sequences, as we will show in Chapter 4.

As the major consequence of recombination is to decouple the ancestries of sites at different places on a chromosome, the effects of recombination are usually inferred

from patterns of statistical association between polymorphisms at different loci. Associations of this sort are referred to as linkage disequilibrium (LD) and have been the subject of extensive study in population genetics. With the field now framed by coalescent theory, linkage disequilibrium can be viewed in terms of statistical associations of gene genealogies at different loci. The patterns of LD between polymorphic markers are shaped by the ancestral population history. It is therefore possible to use multilocus predictors of LD to infer past population history. A variety of methods exists to measure pairwise LD based on genotype frequencies at two loci [Zhao et al., 2007]. However, it has been pointed out that these measures are very diverse and likely not as informative for inferring population history compared to using data from multiple markers along a segment taking into account recombination [Nordborg and Tavaré, 2002]. In Chapter 4, we choose to quantify the LD with a multilocus haplotype homozygosity denoted HH.

2.7 Calculation of the likelihood and inference

Kingman's coalescent allows to model complex demographic scenarios describing the genetic polymorphism. Thus, we can consider the statistical inference of demographic parameters of these scenarios and model choice procedures between different evolutionary scenarios.

When the population size is constant and finite, Wright-Fisher models describe these evolutionary scenarios and the coalescent theory approximates these models when the population size is large. In this context, the history (genealogy with mutations) of the observed sample is a latent process. One of the major challenges to conduct a parametric inference analysis with these models is computing the likelihood at any point ϕ of the parametric space for a given observed data. Indeed, probabilistic models of coalescent on gene genealogies do not provide explicit likelihoods since these models imply complex latent structures (see Stephens and Donnelly [2000], Beaumont et al. [2002], De Iorio et al. [2005]). Classical inference methods such as maximum likelihood of the data require innovative statistical tools. Currently, the best methods to infer parameters of coalescent-based models are on the one hand the research of the maximum likelihood in the frequentist approaches (for example IS, see Griffiths and Tavaré [1994b]) and on the other hand approximation of the posterior distribution for the bayesian approach (ABC, see Beaumont et al. [2002] and Marin et al. [2012]).

For given values of parameters, this technique estimates the likelihood as sum of the probabilities of each possible histories (genealogies with mutations) of the data set by exploring efficiently the space of histories. In this scheme, the importance sampling distribution proposes histories among those who contribute the most to the sum of the likelihood. It provides maximum likelihood estimates with great accuracy but present a high computation cost. Stochastic models that we aim to infer include variations of the population size. They have been rarely addressed with IS. Stephens and Donnelly [2000, Theorem 1] characterized the optimal proposal distribution under a simple model. This distribution is very efficient for a class of time homogeneous models, but not for varying population size models. In most cases (even in time-homogeneous models)

the optimal distribution cannot be practically computed and has to be approximated. Under disequilibrium models these approximate importance distributions are not efficient because they have been developed under constant population size models. For a given likelihood estimation accuracy, the computation time increases strongly against time-homogeneous models. The first contribution of this thesis consists in exploring the sequencial importance sampling (SIS) with resampling (SISR). The idea is to resample so that we learn which histories proposed by the importance sampling distribution contribute the most to the sum defining the likelihood and thus reduce the computation cost. Furthermore, we proposed a new resampling distribution, taking advantage of the information contained in the pairwise composite likelihood.

The recent development of high-throughput sequencing technologies has revolutionized the generation of genetic data for many organisms: genome wide sequence data are now available. Classical inference methods (maximum likelihood methods (MCMC, IS), approximate bayesian methods (ABC)) suitable for genetic polymorphism data sets consisting of a few tens of loci, do not exploit the genetic recombination, assuming that the genealogies of different loci are independent. Taking advantage of the recombination information present in genome wide sequence data, require to consider the dependency between the genealogies of adjacent positions in the genome. Thus the likelihood for a given data set is no longer the product of the likelihoods for each independent locus. While mutation does not affect the tree structure of the genealogy of the sample, recombination does. When we model recombination, the likelihood takes the form of an integral over all possible ancestral recombination graphs [Griffiths and Tavaré, 1994a] for the sampled sequences. This space is of much larger dimension, to the extent that we cannot handle likelihood-based inference while modeling recombination without further simplifications. Approximations of the coalescent with recombination have been developed recently, as the sequential coalescent with Hidden Markov Models (PSMC, coalHMM). We want to adapt and test new methods to infer demographic parameters modeling the evolution of the population size over time from next generation sequencing data (NGS). Assuming that we observe whole genomes, we choose to summarize the polymorphism information contained in two (haploid) genomes or within a (diploid) individual by the haplotype homozygosity in the pairwise alignment. Our inference approach of the demographic history is based on the comparison of this summary of the data to a theoretical predictor. The biological question of interest sometimes takes the form of a model choice question. In particular, has the population undergone a size change in the past? If the approximate Bayesian methods provide answer to these questions with their limits (number and choice of summary statistics), they have been less addressed with frequentist methods. They represent the second contribution of this thesis. To achieve this, we used a model choice procedure between a simple model of constant population size and a slightly more complex model with a single past change in population size. Since these models are embedded, in order to avoid choosing always the more complex model, we developed a penalized model choice criterion. The penalization weights are derived from a sensitivity analysis.

2.8 Main contributions of this thesis

The first contribution of this thesis seek to estimate parameters of a demographic model by maximizing the likelihood. For this purpose, we have to compute the likelihood of a sample of gene copies under a given parametric demographic model. In this framework, the history of the sample (genealogy with mutations) is not observed, it is a latent variable. The probability density of the sample depends on this variable h not observed. The likelihood is then the sum over all possible realizations of the latent variable h:

$$L(x|\phi) = \int_{h\in\mathcal{H}} f(x,h|\phi)\mathrm{d}h. \tag{2.3}$$

This integral is not always tractable. Particularly, when the space $\mathcal{H}$ of the latent process is of large dimension and of a complex structure. It is then excluded to browse all the elements of $\mathcal{H}$. The statistical inference in latent variable models whose space is large and complex led to several numerical probability techniques and optimization methods. A distinction is drawn between algorithms of research of the maximum likelihood in the frequentist approaches and MCMC schemes for the approximation of the posterior distribution for the Bayesian approach. In this thesis, we consider a space of the latent variable h of very large dimension and different types of direction (discrete and continuous) what makes the exact calculation of the integral impossible. We consider a frequentist inference method in this case.

The application field of this thesis is the inference in population genetics models. These models are constraints by demographic scenarios describing the current and past demographic mechanisms and aim to infer their characteristic quantities. In this context, the observed variable x represents the genetic information extracted from the population, usually at present time. The information concerning the temporal and genealogical history of the considered population is generally not available and corresponds to the latent variable h of the model. The exploration of the demographic parameter space and of the latent space becomes harder when the model complexity of the model increases.

In a frequentist approach, the maximization in the variable ϕ of the likelihood of a model depending on a latent process has been the focus of many developments. One of the most used is the Expectation-Maximisation (EM) algorithm of Dempster et al. [1977]. The E step evaluates the completed likelihood expectation and is followed by the M step of mamization of the functional obtained at the E step. The parameter found in the M step are then used as start point of a E step and the process is iterated.

Consider a sample $\mathbf{x} = (x_1, \ldots, x_n)$ i.i.d following the distribution $f(x_i|\phi)$, we want to determine the maximum of the following likelihood function

$$L(\mathbf{x};\phi) = \prod_{i=1}^{n} f(x_i|\phi).$$

The completed likelihood of a model depending on a latent variable h is:

$$L(\mathbf{x},\mathbf{h};\phi) = \prod f(x_i,h_i|\phi).$$

The E step consists in evaluating the following expectation

$$E_{\mathbf{h}|\mathbf{x},\phi^{(c)}}\left[L(\mathbf{x},\mathbf{h};\phi)\right] \tag{2.4}$$

where $\phi^{(c)}$ is the current value of the parameter. For coalescent models in population genetics, the analytic computations of the E step are not possible in general since the distribution of the latent variable h conditionally to the observed sample $\mathbf{x}$ is not known. Stochastic versions of the EM algorithm have been developed to replace the analytical computations of step E, (see for example SEM of Celeux and Diebolt [1985], SAEM of Delyon et al. [1999] and MCEM Wei and Tanner [1990]). However, the difficulty in the exploration of the latent space for complex models makes the estimates of the expectation of Eq. 2.4 too variables.

A naive Monte Carlo estimate of the likelihood, obtained by drawing realizations $H^{(j)}$ of the latent variable according to the process distribution p, is:

$$\widehat{L(\mathbf{x}|\phi)} = \frac{1}{N}\sum_{j=1}^{N} p(\mathbf{x}, H^{(j)}|\phi). \tag{2.5}$$

The histories $H^{(j)}$ are simulated according to the Kingman coalescent, starting from a sample of genetic lineages (as many as in the observed sample $\mathbf{x}$ but without information on the allelic types) and going back to a unique ancestral lineage. This backward simulation of a genealogy is illustrated on Figure 1.1. At each time, two lineages present is the sample can coalesce, that is find a common ancestral lineage until the genealogy involves a single lineage. At this step, we say that the topology of the gene tree is fixed and we draw at random mutation positions of the branches of the genealogy. The ancestral allelic type is drawn uniformly at random and the allelic types at the leaves are derived from this ancestral allelic type according to the coalescent and mutations event on the history. If the allelic composition of the sample at the leaves does not correspond to the observed sample, then $p(\mathbf{x}, H^{(j)}|\phi) = 0$ and we say that the simulated history $H^{(j)}$ is not compatible with the observed sample $\mathbf{x}$. In practice, the probability that $H^{(j)}$ is compatible with the observed data $\mathbf{x}$ (and that $p(\mathbf{x}, H^{(j)}|\phi)$ is not zero) is so low that this estimator is highly variable.

Griffiths and Tavaré [1994a] propose to approximate the likelihood with an importance sampling method. To estimate the likelihood of the parameter ϕ conditionally on the observed sample $\mathbf{x}$, we sample the latent space $\mathcal{H}$ according to a probability distribution, denoted $q(.)$ and called the importance distribution, an absolutely continuous distribution with respect to $p(\mathbf{x}, .|\phi)$.

$$L(\mathbf{x}|\phi) = \int_{h\in\mathcal{H}} \frac{p(\mathbf{x},h|\phi)}{q(h)} q(h)\mathrm{d}h = \mathbb{E}_{H\sim q}\left(\frac{p(\mathbf{x},H|\phi)}{q(H)}\right) \tag{2.6}$$

The integral is the expectation of the likelihood ratio under the probability distribution q. Thus we can approximate it with a Monte Carlo estimator:

$$\widehat{L(\mathbf{x}|\phi)} = \frac{1}{N}\sum_{j=1}^{N} \frac{p(\mathbf{x}, H^{(j)}|\phi)}{q(H^{(j)})}, \quad \text{where } H^{(j)} \sim q.$$

The IS weight of the history number j is $p(\mathbf{x}, H^{(j)}|\phi)q(H^{(j)})$. This time, the probability q is chosen such that all the realizations of the latent variable are compatible with the observed data and thus $p(\mathbf{x}, H^{(j)}|\phi)$ is always non-zero. Indeed, the importance distribution allows to simulate histories starting from the genetic composition of the observed sample and then to draw randomly coalescent of mutation events at the same time until there is a unique ancestral lineage in the history, as illustrated by Figure 1.2.

The first classical difficulty is to calibrate the importance distribution so that the variance of the ratio $p(\mathbf{x}, .|\phi)/q(.)$ and thus the variance of the Monte Carlo estimator is low. Determine the optimal distribution q is as complicated as the initial computation of the likelihood. Stephens and Donnelly [2000] characterized the optimal proposal distribution for a class of time homogeneous models. However, in most cases (even in time-homogeneous models) the optimal distribution is not known and has to be approximated. When the distribution of population allelic type frequencies follows a diffusion process, De Iorio and Griffiths [2004a] characterized the importance sampling proposal distribution. This characterization led to a method for constructing efficient approximate IS proposal distributions by approximating the generator of the process. Such approximate importance distributions are available for a class of population genetics models (see Stephens and Donnelly [2000], [De Iorio and Griffiths, 2004a], [De Iorio and Griffiths, 2004b], [De Iorio et al., 2005]). They are well calibrated in the sense that they match the optimal distribution for the simple model of a single isolated population of constant size with parent independent mutation model.

The latent process, namely the history of the data, of models with past changes in population size exhibits inhomogeneity in time. Thus the previous theoretical arguments which derive efficient proposal distributions are no longer applicable in this context. Nevertheless, as shown by Griffiths and Tavaré [1994b] and Leblois et al. [2014], we can adapt an importance distribution from the importance distributions of models with constant size populations. But their simulation tests demonstrated some limits of the algorithm. Indeed, when disequilibrium is strong, the variance of the likelihood ratio goes to infinity exponentially fast, due to some the degeneration of the IS weights: just a few of the histories present a weight higher than most of the others, close to zero.

This issue has motivated the first part of the present work. It aims to limit the degeneration of the IS weights by using a resampling technique. In chapter 3 of this thesis, we add a resampling procedure according to a new distribution to improve the efficiency of the likelihood estimation. The novelty is to incorporate the information carried by the pairwise composite likelihood of the observed sample in the resampling distribution. This additional information allows for an even more accurate estimation of the likelihood. The idea is to use the pairwise composite likelihood to predict the probability of a sample along a history partially constructed, to have information on final likelihood based on the probability of the current sample. We present the reduction of the computation cost thus obtained on simulated data sets and on a bat data set.

The second contribution of this thesis takes advantage of wide sequence data through IBS segment length between two haplotypes. Inference methods of the population history using genome wide data have been developed using different aspects of genomic polymorphism, see the review Chen [2015] on population genetic studies in the geno-

mic sequencing era. They are mostly based on (1) the Allele Frequency Spectrum (AFS, alternatively Site Frequency Spectrum SFS) and (2) linkage disequilibrium (LD) or haplotype structure. On the one hand, the AFS is a sampling distribution of alleles in a finite sample Chen [2012] and focuses on the allele frequency distribution of a single locus, ignoring the correlation among nearby loci. Such approximation greatly simplifies theory and methodology development. AFS theory was developed in two parallel frameworks: the diffusion Kimura [1955] and coalescent processes Fu [1995]. On the other hand, a group of methods consider linkage disequilibrium, that is the correlation of genealogies of adjacent positions. They take advantage of the recombination information present in genome wide sequence data. Two main approaches have been developed, either based on the approximation of the coalescent with recombination: the sequential coalescent with Hidden Markov Models (PSMC, coalHMM), or based on identical segment lengths.

The Pairwise Sequential Markovian Coalescent method (PSMC, Li and Durbin [2011]) uses a hidden Markov model to approximate the dependency of the coalescent times of two haplotypes between adjacent loci. It has been further extended by Mailund et al. [2011] to allow two sequences from two populations. They infer the detailed ancient population size from coalescent times. Burgess and Yang [2008] developed a Markov chain Monte Carlo approach (MCMCCoal) to sample gene genealogies. They inferred ancient population size by analyzing multiple sequences, with each sequence representing one population. Their method has been further extended by Gronau et al. [2011] to allow two sequences from each sampled population. Schiffels and Durbin [2014] extended the PSMC to the Multiple Sequential Markovian Coalescent (MSMC), allowing to analyze multiple diploid genomes by focusing only on some summary statistics of the genealogies such as first coalescent time of any two sequences and total length of all singleton branches of the genealogy. The Coal-HMM method Hobolth et al. [2007] can also analyze multiple genomes from several populations. Instead of sampling over gene genealogies, Coal-HMM treats the unobserved gene genealogy at each genomic position as latent states in a hidden Markov model. Boitard et al. [2016] introduced an approximate Bayesian computation approach named PopSizeABC that allows to estimate the evolution of the effective population size through time. In this method, observed genomes are summarized using a small number of statistics related to allele frequencies and linkage disequilibrium. They assume that the considered population has evolved forever as an isolated population, as well as other SMC presented before (Li and Durbin [2011], Schiffels and Durbin [2014]) or IBS-based methods (MacLeod et al. [2013]) presented in the following paragraph.

The alternative approach with recombination is based on the distribution of identical segment lengths. Two DNA segments are Identical By State (IBS) if they have identical nucleotide sequences in this segment whereas two DNA segments are Identical By Descent (IBD) if they are inherited from a common ancestor without recombination nor mutation. As detailed below, short segments are mostly affected by the ancient past and long segments mostly by the recent past. Since the detection of short IBD segments is difficult, the distribution of IBD segment lengths is rather used to infer recent past demography. Palamara et al. [2012] expressed the distribution of IBD segment lengths across pairs of individuals as a function of the population's demography and derived an inference procedure to reconstruct such demographic history. They tested increasingly flexible parametric models to infer the demographic history. In order to control

for potential overfitting, they evaluated the parameters obtained for different models by using a likelihood approach and the Akaike's Information Criterion (AIC, Akaike [1974]) to compare models while controlling for their respective degrees of freedom. Browning and Browning [2015] in turn presented a nonparametric method for accurately estimating recent effective population size by using inferred long segments of IBD. They proposed a generalized expectation-maximization (EM) procedure for fitting the trajectory of the historical population size. Ringbauer et al. [2016] recently used a diffusion approximation to trace genetic ancestry back in time, and derived analytical formulas for patterns of isolation by distance of long IBD-blocks, which can also incorporate recent population density changes. Their inference scheme uses a composite likelihood approach to fit observed block sharing to these formulas.

Although we cannot observe DNA segment IBD status, we can observe whether or not two haplotypes contain identical sites along a given segment. That is, they are observed to be identical by state and these segments of conserved segment lengths may occur through recombination. First Hayes et al. [2003] and MacLeod et al. [2009] introduced explicit formulas for the probability for a haplotype pair to share a given number of adjacent positions being IBS. Modeling the population size as a constant piecewise function, their derivation accounts for recombination and mutation. It is based on approximate coalescence theory, assuming that only one recombination may occur per segment within one generation. However computations of these coalescent based formulas are time consuming. These tools have been further developed by MacLeod et al. [2013] to infer ancestral population history and applied to whole-genome sequence data. They used their theoretical HH_{th} to determine population parameters that best match the empirical $\widehat{HH}$ of a given Holstein cattle sequence. To validate the inferred demographic model, they simulated sequence data under the demographic model inferred on one animal and compared these data to the observed sequence in a second animal, then switch roles. Harris and Nielsen [2013] presented a method for using sequence data to jointly estimate the timing and magnitude of past admixture events, along with population divergence times and changes in effective population size. They inferred a piecewise constant effective population size from a collection of pairwise sequence alignments by summarizing their length distribution of tracts of IBS and maximizing an analytic composite likelihood derived from a Markovian coalescent approximation. In contrast, these methods skip detection of IBD segments and instead work directly with IBS haplotypes. An advantage of this approach is that one can examine shorter segments and hence look further back into the past. The present work aims to detect past changes in the population size by considering the LD and relying on IBS segment lengths between two haplotypes.

A variety of methods exist to measure pairwise LD based on genotype frequencies at two loci [Zhao et al., 2007]. However, it has been pointed out that these measures are very diverse and likely not as informative for inferring population history compared to using data from multiple markers along a segment [Nordborg and Tavaré, 2002]. We chose to quantify the LD with a multilocus haplotype homozygosity, denoted HH. Given a pairwise comparison of two haplotypes, HH is the probability for a given number of adjacent positions drawn at random in the genome to be homozygotes between the two haplotypes.

The haplotype homozygosity pattern is affected by the demography, notably by the

changes of the effective population sizes over time. Indeed, the theory predicts that LD over shorter distances reflects parameters of more ancestral population than LD at larger distances. LD on a segment size of c Morgans is mostly affected by the population size approximately $1/2c$ generations in the past, assuming a linearly changing population size [Hayes et al., 2003]. Early studies on HH showed with simulations that estimates of segment homozygosity for a wide range of chromosome segment lengths can be used to estimate the effective population size at multiple times in the past [Hayes et al., 2003].

The previous methods focus mainly on the inference of the historical changes in the effective population size. In general, they do not consider the complexity of the demographic model fitted and may suffer from an overfitting problem: in many cases, a much simpler demographic model could record the key events of the considered population history. Even if some of them propose a control for potential overfitting, to the best of our knowledge, no model choice procedure between demographic models of different complexity have been proposed based on the IBS segment lengths. Our aim is to overcome this lack by proposing a model choice procedure between demographic models of different complexity. In the present work, we focus on a simple model of constant population size and a slightly more complex model with a single past change in the population size. Since these models are nested and to avoid choosing always the more complex model, we developed a penalized model choice criterion. It is based on the comparison of observed $\widehat{HH}$ and predicted HH_{th} haplotype homozygosity. Our penalization relies on the computation of Sobol's sensitivity indices. The idea is to take into account only the part of the error that the model had the ability to explain. It is a form of penalty related to the complexity of the model since a given adjustment error (between observed $\widehat{HH}$ and theoretical HH_{th}) will penalize more a complex model than a simpler one. Indeed, one of the major challenges of the sensitivity analysis is to identify the less contributing parameters to reduce the dimension of the model and quantify the committed error (Sobol et al. [2007], Saltelli et al. [2010]). For this purpose, the sensitivity analysis decomposes the output variance of the model in fractions attributed to each input parameter through Sobol's sensitivity indices. In our case, considering the theoretical haplotype homozygosity as a function of the demographic parameters, assumed to be random variables for the computation, these indices measure the part of the variance of the theoretical HH_{th} explained by each parameter of the more complex model considered.

3 Resampling: an improvement of Importance Sampling in varying population size models

Contents

Abstract Sequential importance sampling algorithms have been defined to estimate likelihoods in models of ancestral population processes. However, these algorithms are based on features of the models with constant population size, and become inefficient when the population size varies in time, making likelihood-based inferences difficult in many demographic situations. In this work, we modify a previous sequential importance sampling algorithm to improve the efficiency of the likelihood estimation. Our procedure is still based on features of the model with constant size, but uses a resampling technique with a new resampling probability distribution depending on the pairwise composite likelihood. We tested our algorithm, called sequential importance sampling with resampling (SISR) on simulated data sets under different demographic cases. In most cases, we divided the computational cost by two for the same accuracy of inference, in some cases even by one hundred. This study provides the first assessment of the impact of such resampling techniques on parameter inference using sequential

importance sampling, and extends the range of situations where likelihood inferences can be easily performed.

Key words Importance Sampling, resampling, jump Markov process, population genetics, demographic inference, coalescent.

3.1 Introduction

Under genetic neutrality, the distribution of the genetic polymorphism in a sample of individuals depends on the evolution of the population size through unobserved stochastic processes. Typically, these stochastic processes describe the evolution of the alleles at a given locus of the individuals from a population backward to their Most Recent Common Ancestor (MRCA). When the population size is constant and finite, Wright-Fisher models describe this evolution and the coalescent theory approximates these models when the population size is large. In this context, the history (genealogy with mutations) of the observed sample is a latent process. One of the major challenges to conduct a parametric inference analysis with these models is computing the likelihood of the data at any point ϕ of the parametric space. Indeed, the likelihood at ϕ is the integral of the probabilities of each possible realization of the latent process. In population genetics, the likelihood for an observed sample is an integral over the distribution of ancestral histories that may have led to this sample. In this work, we consider a class of Monte Carlo methods based on Sequential Importance Sampling (SIS) which provides an estimate of the integral. In this scheme, the importance sampling distribution proposes paths of the process among those who contribute the most to the sum defining the likelihood.

For models of panmictic population with constant size, Griffiths and Tavaré [1994b] described an algorithm wherein a proposal distribution suggests histories of the sample by stepwise reduction of the data set, either by coalescence of two identical genes or by removal of a mutation on a single gene lineage. Stephens and Donnelly [2000, Theorem 1] characterized the optimal proposal distribution for a large class of time homogeneous models, but not for varying population size models. However, in most cases (even in time-homogeneous models) the optimal distribution cannot be practically computed and has to be approximated. De Iorio and Griffiths [2004a] developed a method for constructing such approximations for any model where the mutation process can be described as a Markov chain on gene types and De Iorio and Griffiths [2004b] extended this to subdivided population models. These methods have been further elaborated for stepwise mutation models in a subdivided population by De Iorio et al. [2005]. The latent process, namely the history of the data, of models with past changes in population size exhibits inhomogeneity in time. Thus the previous theoretical arguments which derive efficient proposal distributions are no longer applicable in this context. Nevertheless, as shown by Leblois et al. [2014], we can adapt an importance distribution from the importance distributions of models with constant size populations. But their simulation tests demonstrated some limits of the algorithm. Most importantly, under demographic scenarios with strong changes, they face large com-

putation times due to the large variance of the Monte Carlo estimate of the likelihood.

Our aim in this work is to improve the accuracy of the likelihood estimation for a given computational cost. One direction could be to derive a new importance sampling proposal distribution like Hobolth et al. [2008] did for the infinite site model. In this paper we chose another direction which consists of resampling among the paths proposed by the importance distribution. Under a demographic scenario of constant population size, the likelihood is a sum of products of probabilities of each event whatever their time because of the time homogeneity of the latent process. By contrast, under a varying population size model, these probabilities depend on the times of occurrences of the events. Thus we have to integrate over the possible times of occurrences. It follows that the whole integral (namely the likelihood of the data) we estimate with importance sampling in the general demographic case is an integral over a space of much larger dimension. Because the efficiency of importance sampling decreases with the dimension of the integral, even if the importance distribution adapted by Leblois et al. [2014] were the most efficient distribution among a certain class of distributions, we expect more variance of the likelihood estimate.

We took the opportunity of the paper to present rigorously the stochastic model for a rather general demographic scenario in Section 3.2. The time inhomogeneous latent process is part of the folklore in the neutral population genetic literature, but has never been written down explicitly. In particular, Eq. (3.6) at the end of Section 3.2.2 shows that the integral defining likelihood is of much larger dimension than in the constant demographic case. Indeed, when the population size varies over time, the integrals over the random times cannot be removed, as explained in Section 3.2.4, after a presentation of the sequential importance sampling (SIS) algorithm.

The major contribution of the paper is the addition and the calibration of the resampling procedure of Section 3.3 in the SIS algorithm, based on Liu et al. [2001] and Liu [2008]. The novelty is mostly in the choice of the resampling distribution that we propose in Section 3.3.3, which depends on both the current weight of the latent path and the pairwise composite likelihood (PCL) of the current state of the latent process. Section 3.4 presents numerical results on the likelihood estimates of simulated data sets. These results highlight the benefit due to the proposed resampling distributions in the likelihood estimates. We then plug the likelihood estimates in an inference method presented in Section 3.5.1. The remainder of Section 3.5 highlights how our proposals improve the estimate of the parameter, and the likelihood surface around the maximum likelihood estimate to compute confidence intervals (CIs). We can thus confirm that the gain due to resampling also benefits to the demographic parameter estimation. We end Section 3.5 with a discussion on cases where the data do not hold much information regarding the parameter of interest, leading to flat likelihood surfaces. Finally we show the relevance of our methods by presenting numerical results on a bat data set where strong evidence for population contraction had been already provided by Storz and Beaumont [2002]. All computations for this work were performed using an updated version of the Migraine software, available at http://kimura.univ-montp2.fr/~rousset/Migraine.htm [Rousset and Leblois, 2007, 2012, Leblois et al., 2014].

3.2 The stochastic model and its likelihood

To illustrate our method we consider genetic data from individuals of a single population sampled at time $t = 0$. Let $N(t)$ be the population size, expressed in number of gene copies in the whole paper, t generations away from the sampling time $t = 0$. We assume that $N(t)$ is a parametric function of t, see Section 3.4.1 for examples. In this Section we focus only on data from a given locus.

3.2.1 Stochastic model

Kingman [1982a]'s coalescent process is the usual model to describe ancestral relationships between gene copies of the sample under neutrality in a population of constant, but relatively large size. We superimpose a mutation model on the coalescent process to describe gene modifications along lineages. Since the evolution is neutral, the coalescent is independent of the mutation process. To describe the resulting process, we introduce a random vector $\mathbf{H}_t$, indexed by the set of possible types of genes (possible alleles) E: if $A \in E$, the component $\mathbf{H}_t(A)$ counts the number of genes of type A at time t (i.e., t generations away, backward in time, from the sampling time) in the genealogy of the sample. The likelihood of the genetic data is given by the distribution of $\mathbf{H}_0$, which cannot be written as an explicit function of the parameter of interest, that we note ϕ in this chapter.

3.2.2 Markovian description of the evolution

Actually, we only have at our disposal the following description of the process $\mathbf{H}_t$ forward in time.

Let $\mathbf{e}_A$ denote the vector indexed by E whose components are all equal to 0, except the A-component which is equal to one. On one hand, the probability of a new lineage of type A in the genealogy at time $t - \delta$, knowing that $\mathbf{h}$ is the value of the process at time t, is

$$\mathbb{P}\Big(\mathbf{H}_{t-\delta} = \mathbf{h} + \mathbf{e}_A \Big| \mathbf{H}_t = \mathbf{h}\Big) = \frac{(|\mathbf{h}| + 1)\,\mathbf{h}(A)}{2\,N(t)}\delta + o(\delta) \tag{3.1}$$

where $|\mathbf{h}| = \sum_{A\in E} \mathbf{h}(A)$ is the total number of lineages at time t in the genealogy and $o(\delta)$ a quantity that is negligible in front of δ when $\delta \to 0$. On the other hand the probability of a mutation of a gene of type B at time t to a gene of type A at time $t - \delta$, knowing that $\mathbf{h}$ is the value of the process at time t, is

$$\mathbb{P}\Big(\mathbf{H}_{t-\delta} = \mathbf{h} + \mathbf{e}_A - \mathbf{e}_B \Big| \mathbf{H}_t = \mathbf{h}\Big) = \mu\,\mathbf{h}(B)\,p_{B,\,A}\,\delta + o(\delta) \tag{3.2}$$

where $p_{B,A}$ is the mutation probability from allele B to allele A forward in time, μ is the mutation rate per generation per lineage, and $o(\delta)$ is a quantity that is negligible in front of δ when $\delta \to 0$. Additionally, when there is only one lineage in the genealogy, the

distribution of the gene type (the allele) is supposed to be the stationary distribution $\psi(\cdot)$ of the transition matrix $p = \{p_{B,A}; B, A \in E\}$:

$$\mathbb{P}\Big(\mathbf{H}_t = \mathbf{e}_A \Big| |\mathbf{H}_t| = 1\Big) = \psi(A).$$

Hence, forward in time (i.e., when t decreases), the coalescent based model $\mathbf{H}_t$ is a pure jump, continuous time Markov process taking values in the set of integer vectors indexed by E, namely $\mathbb{N}^E$. But the process is time inhomogeneous because the coalescence rate of Eq. (3.1) depends on the current date t through the function $N(t)$. Note that Eq. (3.1) and (3.2) can both be written

$$\mathbb{P}\Big(\mathbf{H}_{t-\delta} = \mathbf{h}' \Big| \mathbf{H}_t = \mathbf{h}\Big) = \Lambda_t(\mathbf{h}'|\mathbf{h})\delta + o(\delta)$$

for any $\mathbf{h} \neq \mathbf{h}'$ in $\mathbb{N}^E$ where

$$\Lambda_t(\mathbf{h}'|\mathbf{h}) = \begin{cases} (|\mathbf{h}| + 1)\, \mathbf{h}(A) \Big/ \big(2\, N(t)\big) & \text{if } \mathbf{h}' = \mathbf{h} + \mathbf{e}_A \\ \mu\, \mathbf{h}(B)\, p_{B,\,A} & \text{if } \mathbf{h}' = \mathbf{h} + \mathbf{e}_A - \mathbf{e}_B \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

defines the intensity matrix of the Markov process.

We can give a more explicit description of the process since it is a pure jump process, or, in other words, $\mathbf{H}_t$ is a piecewise constant function of t. We denote $\mathbf{X}_0 = \mathbf{H}_0$ the first value of the process at time $T_0 = 0$. The process $\mathbf{H}_t$ remains constant until (random) time T_1, where it takes another value $\mathbf{X}_1 = \mathbf{H}_{T_1}$. After T_1, the process $\mathbf{H}_t$ stays equal to $\mathbf{X}_1$ until time T_2 where it jumps to a another value $\mathbf{X}_2$, and so on.

Fig. 3.10 in the supplementary material represents a possible path of the process $\mathbf{H}$. We refer to such paths as (possible) histories of the sample (composed of three genes of type a and one gene of type b in Fig. 1 in the supplementary material). Note that many genealogies correspond to a possible history since, at time T_1 we have chosen a genealogy with a coalescence between the two left-handmost genes, but we could have chosen another genealogy joining any pair of genes of type a at time T_1. Likewise at time T_3 we could have chosen any of the three possible pairs. Actually each genealogy leading to the same path of the process $\mathbf{H}$ has the same probability because of the exchangeability of the lineages carrying the same allele.

We set $\Delta_i = T_{i+1} - T_i$ which is usually named the holding time at value $\mathbf{X}_i$ for any integer number i. The distribution of the process is given by the density

$$\mathbb{P}\Big(\mathbf{X}_i = \mathbf{h}_i \text{ for all } i = 0, \ldots, n \text{ and } \Delta_i \in (\delta_i; \delta_i + d\delta_i) \text{ for all } i = 0, \ldots, n-1 \Big| \mathbf{X}_n = \mathbf{h}_n\Big) =$$
$$\prod_{i=1}^{n} P_{t_i}(\mathbf{h}_{i-1}|\mathbf{h}_i) \times \prod_{i=0}^{n-1} \lambda_{t_i+\delta_i}(\mathbf{h}_i) \exp\left(-\int_0^{\delta_i} \lambda_{t_i+u}(\mathbf{h}_i) du\right) d\delta_i \tag{3.4}$$

for any $\delta_0, \ldots, \delta_{n-1} > 0$, $\mathbf{h}_0, \ldots, \mathbf{h}_n$ in $\mathbb{N}^E$, where $t_0 = 0$, $t_i = \delta_0 + \cdots + \delta_{i-1}$,

$$\lambda_t(\mathbf{h}) = \sum_{\mathbf{h}' \neq \mathbf{h}} \Lambda_t(\mathbf{h}'|\mathbf{h}) \tag{3.5}$$

$$\text{and} \quad P_t(\mathbf{h}'|\mathbf{h}) = \Lambda_t(\mathbf{h}'|\mathbf{h}) \Big/ \lambda_t(\mathbf{h}).$$

The term $\lambda_t(\mathbf{h})$ is interpreted as the (infinitesimal) jump rate of the process $\mathbf{H}$ at time t, knowing that the process takes value $\mathbf{h}$ at time t, namely

$$\mathbb{P}\Big(\mathbf{H}_{t-\delta} \neq \mathbf{h} \Big| \mathbf{H}_t = \mathbf{h}\Big) = \lambda_t(\mathbf{h})\, \delta + o(\delta)$$

when $\delta > 0$ and $o(\delta)$ is negligible in front of δ when $\delta \to 0$. With Eq. (3.3), it can be computed as

$$\lambda_t(\mathbf{h}) = \frac{|\mathbf{h}|(|\mathbf{h}|+1)}{2N(t)} + \mu|\mathbf{h}| = \frac{|\mathbf{h}|}{2N(t)} \left((|\mathbf{h}|+1) + \theta(t)\right),$$

where $\theta(t) = 2\mu N(t)$ is the usual composite mutation rate parameter of population genetics. On the other hand, $P_t(\mathbf{h}'|\mathbf{h})$ is interpreted as a transition matrix (forward in time), and gives the probability that the new value (forward in time) of $\mathbf{H}_{t'}$ is $\mathbf{h}'$ for $t' < t$ knowing that the process jumps at time t' and that $\mathbf{H}_t = \mathbf{h}$.

Set $\tau = \inf\big\{t > 0 : \ |\mathbf{H}_t| = 1\big\}$, which is the age of the most recent common ancestor (MRCA) of the sample. Since $\mathbf{H}$ is a pure jump process, $\tau = T_\sigma$ where $\sigma = \inf\big\{n > 0 : |\mathbf{X}_n| = 1\big\}$. And note that the distribution of $\mathbf{H}_\tau = \mathbf{X}_\sigma$ is known explicitly as

$$\mathbb{P}(\mathbf{H}_\tau = \mathbf{e}_A) = \psi(A).$$

To recover the likelihood of the observed genetic data, that is to say the distribution of $\mathbf{H}_0$ from this knowledge and the (forward in time) transition mechanism of Eq. (3.4), we have to integrate over all possible histories from the MRCA, that is to say all possible values n of σ, $\mathbf{h}_1, \ldots, \mathbf{h}_n$ in $\mathbb{N}^E$ and $\delta_0, \ldots, \delta_{n-1} > 0$. Hence

$$\mathbb{P}\big(\mathbf{H}_0 = \mathbf{h}_0\big) = \sum_{n=1}^{\infty} \int \cdots \int \sum_{\mathbf{h}_1, \cdots, \mathbf{h}_n} \sum_{A} \psi(A)\, \mathbf{1}_{\{\mathbf{h}_n = \mathbf{e}_A, |\mathbf{h}_{n-1}|>1\}} \times$$
$$\prod_{i=1}^{n} P_{\delta_0+\cdots+\delta_{i-1}}(\mathbf{h}_{i-1}|\mathbf{h}_i) \times \prod_{i=0}^{n-1} \lambda_{\delta_0+\cdots+\delta_i}(\mathbf{h}_i) \exp\left(-\int_0^{\delta_i} \lambda_{\delta_0+\cdots+\delta_{i-1}+u}(\mathbf{h}_i) du\right) d\delta_i. \quad (3.6)$$

The challenge to conduct a likelihood based inference in population size varying models is in computing the multidimensional integral of Eq. (3.6), which cannot be computed formally. Note that, to alleviate notation, we have dropped the dependency in the parameter of interest ϕ of the quantities arising in Eq. (3.6) but both $P_t(\mathbf{h}'|\mathbf{h})$ and $\lambda_t(\mathbf{h})$ are functions of ϕ, and even sometimes the stationary distribution $\psi(A)$.

3.2.3 Evaluating the likelihood with importance sampling

The trick to compute the likelihood defined in Eq. (3.6) is to rely on another transition matrix $Q_t(\mathbf{h}|\mathbf{h}')$ backward in time and set

$$W_n(\mathbf{h}_{0:n}; t_{0:n}) = \prod_{i=1}^{n} P_{t_i}(\mathbf{h}_{i-1}|\mathbf{h}_i) \Big/ \prod_{i=1}^{n} Q_{t_i}(\mathbf{h}_i|\mathbf{h}_{i-1}). \quad (3.7)$$

Then Eq. (3.6) leads to

$$\mathbb{P}\big(\mathbf{H}_0 = \mathbf{h}_0\big) = \sum_{n=1}^{\infty} \int \cdots \int \sum_{\mathbf{h}_1,\cdots,\mathbf{h}_n} \sum_{A} \psi(A)\mathbf{1}_{\{\mathbf{h}_n=\mathbf{e}_A,|\mathbf{h}_{n-1}|>1\}} \times$$
$$W_n(\mathbf{h}_{0:n}; t_{0:n}) \times \prod_{i=1}^{n} Q_{t_i}(\mathbf{h}_i|\mathbf{h}_{i-1}) \times \prod_{i=0}^{n-1} \lambda_{t_i+\delta_i}(\mathbf{h}_i) \exp\left(-\int_0^{\delta_i} \lambda_{t_i+u}(\mathbf{h}_i)du\right) d\delta_i, \quad (3.8)$$

where $t_i = \delta_0 + \cdots + \delta_{i-1}$ are implicit functions of the δ_i's. The right hand side of Eq. (3.8) can be interpreted as the expected value of

$$W_\sigma = \sum_n \sum_A \psi(A)\mathbf{1}_{\{\mathbf{h}_n=\mathbf{e}_A,|\mathbf{h}_{n-1}|>1\}} \times W_n(\mathbf{h}_{0:n}; t_{0:n})$$

when $\mathbf{h}_0, \mathbf{h}_1, \ldots$ and $t_0, t_1, \ldots$ are the realizations of the embedded chain and the jump times of an inhomogeneous Markov process $\widetilde{\mathbf{H}}$ whose intensity matrix (backward in time) is given by

$$\widetilde{\Lambda}_t(\mathbf{h}|\mathbf{h}') = \lambda_t(\mathbf{h}')Q_t(\mathbf{h}|\mathbf{h}')$$

and which starts from $\widetilde{H}_0 = \mathbf{h}_0$, the observed data. In other words,

$$\mathbb{P}(\mathbf{H}_0 = \mathbf{h}_0) = \widetilde{\mathbb{E}}(W_\sigma).$$

Interpreting the likelihood of the data as an expected value over another distribution is the first step toward an importance sampling estimate of the likelihood. Indeed, the Monte Carlo estimation of the likelihood is the empirical average of W_σ computed on simulated replicates $\widetilde{\mathbf{H}}^{(j)}$ (for $j = 1, \ldots, n_H$) of the process $\widetilde{\mathbf{H}}$:

$$\widehat{\mathbb{P}}(\mathbf{H}_0 = \mathbf{h}_0) = \frac{1}{n_H} \sum_{j=1}^{n_H} W_\sigma^{(j)}. \quad (3.9)$$

3.2.4 Practical aspects and efficiency

To sum up the above, the idea of importance sampling is to interpret the likelihood as an expected value over a sampling distribution of histories (the distribution of the process $\widetilde{\mathbf{H}}$) which may differ from the distribution of the genuine model described in Section 3.2.2. We can choose this importance distribution freely, as long as any path from the data $\mathbf{h_0}$ to the MRCA with positive density under the distribution of the latent process $\mathbf{H}$ has also a positive density under the distribution of the importance process $\widetilde{\mathbf{H}}$. But this choice has a major impact on the efficiency of the approximation in Eq. (3.9) since the variance of W_σ depends on the distribution of $\widetilde{\mathbf{H}}$.

When the population size $N(t)$ is fixed to N_0 for all t, we have at our disposal an efficient sampling distribution from the literature [Stephens and Donnelly, 2000, De Iorio and Griffiths, 2004a, De Iorio et al., 2005] which represents a major improvement over the first proposal of Griffiths and Tavaré [1994a]. In this simple demographic scenario, the importance sampling estimate is the most efficient under a parent independent mutation model (which means that p_{BA} does not depend on B) and Eq. (3.8) provides

an exact evaluation of the likelihood for $n_H = 1$ replicate. In other words, the variance of W_σ under the efficient importance sampling distribution is zero. For other mutation models and a constant population size, the variance of W_σ is no longer zero, but the number n_H of replicates required to get a sharp estimation of the likelihood with Eq. (3.9) is much smaller when relying on the efficient importance distribution rather than on the proposal of Griffiths and Tavaré [1994a]. In the scenario where the population size varies over time, we resort to the proposal of Leblois et al. [2014] to define the transition matrix of $\widetilde{\mathbf{H}}$. That is to say that for any value of $t > 0$, $Q_t(\mathbf{h}, \mathbf{h}')$ is the transition matrix of the efficient importance distribution as if the population size were constant over time and fixed to the current value of $N(t)$. But the efficiency of this importance distribution depends on the variation of $N(t)$, as explained in Leblois et al. [2014].

Nevertheless, the choice of importance distribution leads to a process $\widetilde{\mathbf{H}}$ which is defined explicitly as an inhomogeneous Markov process, backward in time, starting from the observed data $\mathbf{h}_0$. The computation of a simulated path of $\widetilde{\mathbf{H}}$, as well as of W_σ is performed with the sequential method of Algorithm 1.

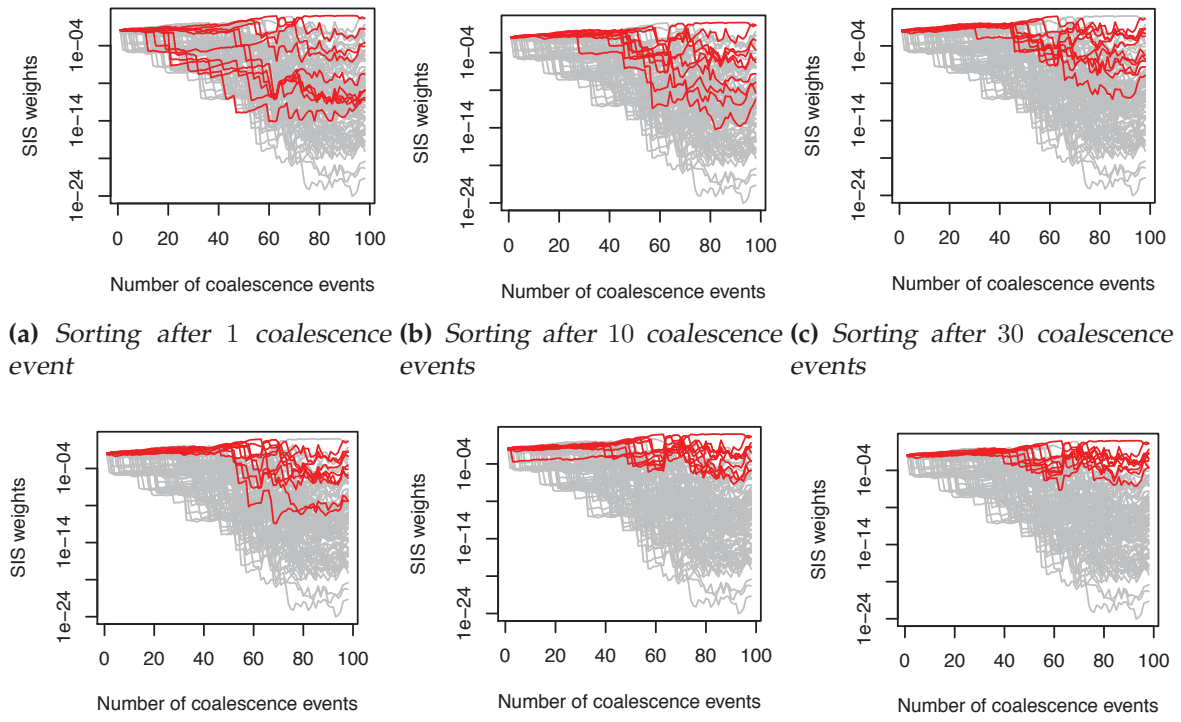
Algorithm 1: Sequential importance sampling

1 Initialization: set $\mathbf{h} = \mathbf{h}_0$, $w = 1$ and $t = 0$;
2 **while** $|\mathbf{h}| > 1$ **do**
3 Draw the holding time δ according to its density $\lambda_{t+\delta}(\mathbf{h}) \exp\left(-\int_0^\delta \lambda_{t+\delta}(\mathbf{h})\right)$ and update $t = t + \delta$;
4 Draw $\mathbf{h}'$ according to $q_t(\mathbf{h}'|\mathbf{h})$ and update $w = w \times p_t(\mathbf{h}|\mathbf{h}')/q_t(\mathbf{h}'|\mathbf{h})$;
5 Update $\mathbf{h} = \mathbf{h}'$;
6 Set $W_\sigma = w \times \psi(\mathbf{h})$ and return W_σ.

Note that, since W_σ is a product along the path of the process $\widetilde{\mathbf{H}}$, its value is computed sequentially at step 4 of Algorithm 1. The update depends only on the current time t, the value of $\widetilde{\mathbf{H}}$ just before time t, namely $\mathbf{h}$, and the new value $\mathbf{h}'$. Hence, Algorithm 1 does not have to keep track of the whole path of $\widetilde{\mathbf{H}}$. Moreover, drawing the holding time δ can be done either with a rejection algorithm, see Appendix 3.9.2 in the Supplementary Materials, or by inverting the cumulative distribution function when possible, e.g., see Griffiths and Tavaré [1994b]. Finally, the above algorithm is run n_H times to approximate the likelihood by the average in Eq. (3.9).

Unfortunately, the decrease of efficiency of the importance sampling scheme from the constant population size scenario to the general case (where $N(t)$ varies over time) can be drastic, see Leblois et al. [2014]. To understand the major difference, we recall that, when the population size $N(t)$ is constant, neither $\Lambda_t(\mathbf{h}'|\mathbf{h})$ nor $P_t(\mathbf{h}'|\mathbf{h})$ depend on t. In this simpler case we choose an importance distribution characterized by a $Q_t(\mathbf{h}|\mathbf{h}')$ which does not depend on t, and W_σ does not depend on the random jump times $T_1, T_2, \ldots$ of the process $\widetilde{\mathbf{H}}$. But in the varying size scenario, these random times do contribute to the variance of W_σ.

Fig. 3.1 follows w, computed at line 4 of Algorithm 1, over successive coalescence for one hundred replicates of the SIS algorithm. It shows first that the variance of the final SIS weights is quite large, second that replicates that lead to the highest final W_σ also tend to have high w throughout the sequence of coalescence events. Moreover, Fig. 3.1 shows that, as the number of coalescence events undergone increases, the range of values of w increases exponentially. This exponential increase is evidence that importance sampling becomes inefficient in spaces of high dimension, and that each random jump times $T_1, T_2, \ldots$ has a multiplicative contribution to the overall range of W_σ.



(a) *Sorting after* 1 *coalescence event* **(b)** *Sorting after* 10 *coalescence events* **(c)** *Sorting after* 30 *coalescence events*

(d) *Sorting after* 50 *coalescence events* **(e)** *Sorting after* 70 *coalescence events* **(f)** *Sorting after* 98 *coalescence events*

FIGURE 3.1: ***Evolution of the current partial SIS weights as a function of the number of coalescence events undergone by the sample****, for* 100 *histories drawn according to the importance distribution. At each step, the current weights are normalized by the sum of the current weights of all the histories and represented on a logarithmic scale. In red: the* 10 *histories corresponding to the* 10 *highest partial SIS weights computed after (a)* 1*, (b)* 10*, (c)* 30*, (d)* 50*, (e)* 70 *and (f)* 98 *coalescence events.*

The final approximation of the likelihood is the average of these one hundred replicates of W_σ, see Eq. (3.9). The difference between the values of the final SIS weights W_σ is so huge that most of them are negligible compared to the top ten highest ones and do not contribute to the average in Eq. (3.9). If we look at Fig. 3.1, considerations of their trajectories further back in time show that these weights do not rise back much. Hence, we can reduce the computing effort and increase the accuracy of the method by ignoring the histories which give low values of w after some times through the sequence of coalescence events and by replacing them by histories with higher values of SIS weight w. The above idea can be seen as a correction of the proposal distribution, and the estimate will remain unbiased if the SIS weights are accordingly corrected. In order to

implement it, we rely on the resampling procedure explained in the next Section.

3.3 Resampling

We derive a resampling procedure from that of Liu [2008, Section 4.1.2] that consists of pruning away the partial histories associated with very small weights and reusing those associated with high weights instead of restarting from scratch.

3.3.1 Sequential importance sampling with resampling: the algorithm

Algorithm 2: Sequential importance sampling with resampling (SISR)

1 Initialization: **for** $j = 1$ **to** $j = n_H$ **do**
2 Set $w^{(j)} = 1$, $t^{(j)} = 0$ and $\mathbf{h}^{(j)} = \mathbf{h}_0$;
3 Set $\text{ESS}_- = n_H$;
4 **while** $|\mathbf{h}^{(j)}|$ *are not all equal to* 1 **do**
5 **for** $j = 1$ **to** $j = n_H$ **do**
6 Set $n_{control} = 0$;
7 **while** $|\mathbf{h}^{(j)}| > 1$ ***and*** $n_{control} < k$ **do**
8 Draw the holding time δ according to its distribution
9 and update $t^{(j)} = t^{(j)} + \delta$;
10 Draw $\mathbf{h}'$ according to $q_{t^{(j)}}(\mathbf{h}'|\mathbf{h}^{(j)})$
11 and update $w^{(j)} = w^{(j)} \times p_{t^{(j)}}(\mathbf{h}^{(j)}|\mathbf{h}')/q_{t^{(j)}}(\mathbf{h}'|\mathbf{h}^{(j)})$;
12 Update n_{control} according to its definition and update $\mathbf{h}^{(j)} = \mathbf{h}'$;
13 Compute $\text{ESS}_+ = \left(\sum_{j=1}^{n_H} w^{(j)}\right)^2 \Big/ \sum_{j=1}^{n_H} \left(w^{(j)}\right)^2$;
14 **if** $ESS_+ < ESS_-/10$ **then**
15 Resample according to Algorithm 3;
16 Update ESS_- to the ESS of the resampled collection;
17 **for** $j = 1$ **to** $j = n_H$ **do**
18 Update $w^{(j)} = w^{(j)} \times \psi(\mathbf{h}^{(j)})$;
19 Return the average $n_H^{-1} \sum_{j=1}^{n_H} w^{(j)}$.

The sequential importance sampling with resampling (SISR) is as follows. We initiate n_H independent runs of Algorithm 1, hence n_H draws of W_σ from the importance distribution. But we stop the repeat-until loop at step 2 of Algorithm 1 before hitting the MRCA. Indeed each run of Algorithm 1 is stopped at checkpoints (stopping times of the Markov process $\widetilde{H}_t$); at these checkpoints we evaluate the quality of the collection of n_H partial histories and, if necessary, we resample these histories. To that aim, we propose a new resampling distribution detailed in Section 3.3.2.

The checkpoints at which we test whether to resample correspond either to a given number k of events (coalescences and mutations) undergone after the previous checkpoint or to a given number k of coalescence events undergone. Once a checkpoint is reached, we evaluate the quality of the collection of partial histories to assess whether it is necessary to resample. The quality test is based on the Effective Sample Size (ESS) relative to the collection of partial histories, namely

$$\text{ESS} = \left(\sum_{j=1}^{n_H} w^{(j)}\right)^2 \Big/ \sum_{j=1}^{n_H} \left(w^{(j)}\right)^2 \tag{3.10}$$

where $w^{(j)}$ is the current value of w of the j-th partial history at the checkpoint. The largest value of ESS is n_H and occurs when $w^{(1)} = \cdots = w^{(n_H)}$; the ESS decreases when the range of values of $w^{(j)}$ expands. When a single weight, $w^{(1)}$ say, is much larger than the other ones, the ESS is approximately equal to 1 since both numerator and denominator of Eq. (3.10) are approximately equal to $\left(w^{(1)}\right)^2$. Actually, the ESS assesses how the random holding times and events of the partial histories between two successive checkpoints contribute to the variance of the estimate in Eq. (3.9). Thus we resample the collection of partial histories whenever the ESS falls below a threshold value, for instance $\text{ESS}_-/10$, where ESS_- is the value of the ESS after the last resampling.

The checkpoints are frequent when $k = 1$, but computing the ESS, as well as resampling the whole collection is time consuming, so that higher values of k might be more pragmatic.

The SISR is presented in Algorithm 2, where $w^{(j)}$, $t^{(j)}$ and $\mathbf{h}^{(j)}$, for $j = 1, \ldots, n_H$, are arrays which keep tracks of the values of current weight w, time t and state $\mathbf{h}$ of each run of Algorithm 1. Moreover, we define n_{control} which stores the number of undergone events (either the total number of undergone events, or the number of undergone coalescences) in-between checkpoints.

3.3.2 The resampling procedure

Assume that SISR has reached a checkpoint and that the ESS is low enough to resample. At this time, we create a new collection of n_H simulated histories by drawing at random in the previous collection of histories according to a multinomial distribution $\mathcal{M}(\mathbf{v}, n_H)$ where $\mathbf{v} = (v^{(1)}, \ldots, v^{(n_H)})$ is a resampling probability distribution on the collection of n_H partial histories, see Section 3.3.3 for examples of such distributions.

Resampling is equivalent to applying a second importance sampling algorithm within the SIS. Indeed,

$$\sum_{j=1}^{n_H} w^{(j)} = \sum_{j=1}^{n_H} \frac{w^{(j)}}{v^{(j)}} v^{(j)} \tag{3.11}$$

and the right hand side can be interpreted as the expected value of $w^{(j)}/v^{(j)}$ over the distribution of a random j drawn from $\mathbf{v}$. Thus, in order not to bias the procedure, the new weight associated to the j-th history is $w^{(j)}/v^{(j)}$ whenever this history appears in

the resampled collection of histories. Algorithm 3 summarizes the procedure, with the notations of Algorithm 2.

Algorithm 3: Resampling procedure

for $j = 1$ **to** $j = n_H$ **do**
  Draw J' from distribution $\mathbf{v}$;
  Set $\widetilde{\mathbf{h}}^{(j)} = \mathbf{h}^{(J')}$ and $\widetilde{t}^{(j)} = t^{(J')}$;
  Set $\widetilde{w}^{(j)} = w^{(J')}/v^{(J')}$;
Replace the old collection $\left\{(\mathbf{h}^{(j)}, t^{(j)}, w^{(j)})\right\}_{j=1}^{n_H}$ with the new collection $\left\{(\widetilde{\mathbf{h}}^{(j)}, \widetilde{t}^{(j)}, \widetilde{w}^{(j)})\right\}_{j=1}^{n_H}$.

3.3.3 The resampling distribution

Resampling introduces a new possible cause of variance, since we replace the sum of Eq. (3.11) with a Monte Carlo estimate. The resampling probability distribution $\mathbf{v} = (v_1, \ldots, v_{n_H})$ on the collection of histories could be any distribution. To achieve efficiency we should pick a distribution $\mathbf{v}$ that reflects the future trend of the partial histories, more precisely that sets relatively high probabilities on the partial histories that will correspond to the highest W_σ. If the resampling distribution brings helpful information about histories, then the mean square error (MSE) of Eq. (3.9) should decrease. On the contrary if the resampling probability distribution brings useless information or no information, then the resampling just adds noise and the MSE increases. For example a uniform $\mathbf{v}$ distribution that clearly does not bring any information about the histories introduces only an additional variance in the estimation of the likelihood.

The resampling algorithm could face two difficulties: first it does not always choose the replicates with the highest W_σ; second, regardless of the optimal replicate, it is not perfectly predicted from an intermediary w. However, as explained at the end of Section 3.2.4, the current value of w helps predicting the contribution of the final value W_σ to the average in Eq. (3.9). One might thus be tempted to resample the j-th partial history with a probability proportional to $w^{(j)}$, i.e. $v_j \propto w^{(j)}$. But this resampling distribution $\mathbf{v}$ might not well choose the optimal W_σ because of the spread of the $w^{(j)}$'s. Thus Liu et al. [2001] proposed to rely on $v_j \propto \left[w^{(j)}\right]^\alpha$ for some $\alpha \in [0; 1]$, with an arbitrary preference on $\alpha = 1/2$. Another possible predictor of the contribution of the j-th partial history to the average in Eq. (3.9) is its current state $\mathbf{h}^{(j)}$. Note that the probability of $\mathbf{h}^{(j)}$ is the expected value of $W_\sigma/v^{(j)}$. Of course the probability of $\mathbf{h}^{(j)}$ is intractable, but we might replace it by an easily computed pseudo-likelihood. We propose here to rely on the pairwise composite likelihood $L_2(\mathbf{h}^{(j)})$ defined in the Appendix 3.9.1 in the Supplementary Materials since we can compute it very easily. But such pseudo-likelihoods are much more contrasted than the true one and thus should be tempered by some exponent $\beta \ll 1$. Hence we advocate the following resampling distribution

$$v^{(j)} \propto \left(w^{(j)}\right)^\alpha \left(L_2(\mathbf{h}^{(j)})\right)^\beta, \quad \alpha, \beta \in [0, 1], \beta \ll 1. \tag{3.12}$$

The tuning parameters α and β are used to balance the effect of the information provi-

ded by the SIS weight and by the composite likelihood. Section 3.4 provides numerical examples showing the efficiency of the above resampling distribution for a large range of values of the tuning parameters α and β.

3.4 Improvements on the likelihood estimate: numerical results

3.4.1 The simulated demographic model

We use the model described in Sections 3.2.1 and 3.2.2 to analyze microsatellite markers under a Stepwise Mutational Model (SMM). The set E of allele types corresponding to microsatellite markers is the set $\mathbb{N}$ of non negative integers. When a mutation occurs under a SMM, the size of the allele is either increased by 1 or decreased by 1 with the same probability [Kimura and Ohta, 1978].

As in Leblois et al. [2014] we consider a single isolated population whose size has undergone past changes. We denote $N(t)$ the population size expressed as the number of gene copies, t generations away from the sampling time $t = 0$. The haploid population size at sampling time is $N(0) = N$. Then, going backward in time, the population size changes according to a deterministic function until reaching an ancestral population size N_{anc} at time $t = T$. Then, $N(t) = N_{\text{anc}}$ for all $t > T$. To illustrate our method we consider an exponentially contracting population size, represented in Fig. 3.11 in the supplementary material, which can be written

$$N(t) = \begin{cases} N \times (N_{\text{anc}}/N)^{t/T} & \text{if } t \in [0; T], \\ N_{\text{anc}} & \text{if } t \geq T. \end{cases}$$

To ensure identifiability the demographic parameters are scaled as $\theta = 2\mu N$, $D = T/2N$ and $\theta_{\text{anc}} = 2\mu N_{\text{anc}}$, where μ is the mutation rate per locus per generation. The parameter space of the model is thus the set of vectors $\phi = (\theta, D, \theta_{\text{anc}})$. Additionally we set $\theta(t) = 2\mu N(t)$, and $N_{\text{ratio}} = \theta/\theta_{\text{anc}} = N/N_{\text{anc}}$, the latter being useful to characterize the strength of the contraction.

All the data sets simulated in this work where produced with the IBDSIM software available at http://www1.montpellier.inra.fr/CBGP/software/ibdsim/index.html [Leblois et al., 2009]. Monomorphic loci were not excluded a priori but were very unlikely given the chosen values of population sizes and mutation rates.

3.4.2 Reduction of the MSE between the true value of the likelihood and its estimate

Numerical results of Section 3.4 are presented on data sets simulated under the demographic scenario where $\phi = (\theta, D, \theta_{\text{anc}}) = (0.4, 0.25, 400)$ which models a recent and

strong past contraction in population size. This scenario is the most representative of our conclusions although we studied more moderate and/or older contractions. To assess the efficiency of estimates of the likelihood at a given point of the parameter space, we compared these estimates with a reference value L that is a sharp approximation of the true likelihood. We have computed the reference value L by estimating the likelihood with SIS with a huge collection of $n_H = 20,000,000$ independent histories.

To evaluate the variability of the likelihood estimate returned by a given algorithm, with given values of its tuning parameters, we first plot the empirical distribution of 100 runs of the same algorithm with a boxplot (Fig. 3.2–3.3). We also measured the mean square error of 100 estimates $\widehat{L_i}$ around the reference value L, and averaged this MSE over 100 simulated data sets to remove the dependency on the simulated data set (Fig. 3.4). Note that this last method computes an empirical mean square error that approximates the MSE of the likelihood estimate.

To compare the efficiency of both SIS and SISR algorithms with the same computational effort of simulating a collection of $n_H = 2000$ histories, we considered the ratio of the empirical mean square error of SISR over that of SIS. The MSE ratio brings out the reduction of the mean square error in likelihood estimate due to the resampling technique.

For Fig. 3.2–3.3 discussed below, we note that all the SISR calibrations lead to a smaller variance in likelihood estimation than SIS and that the average is much closer to the reference value. It means that resampling improves the accuracy of the likelihood estimate and that even if the resampling procedure is not much calibrated, the likelihood inference is already better than without resampling. Our method is thus not too sensitive to the tuning of the resampling parameters α, β described in Section 3.3.3. This is an important feature of our proposal: it will ease its use and facilitate its dissemination.

Type and frequency of checkpoints As said in Section 3.3.1, the checkpoints might be defined either in term of a fixed number of events (coalescences and mutations) undergone on each partial history or in term of a fixed number of coalescence events. Note that the total number of gene copies in the current state $\mathbf{h}$ is determined by the size of the observed data $\mathbf{h}_0$ and the number of coalescence events undergone from time $t = 0$ to state $\mathbf{h}$. Hence the second definition of a checkpoint proposes to compare partial histories that lead to current states $\mathbf{h}$ with the same number of lineages. Thus Liu [2008] advocated the second type of checkpoints. Fig. 3.2 shows that resampling among histories with the same number of lineages (that is by the number of coalescence events) is slightly more efficient than the alternative condition. Indeed, when relying on the alternative checkpoints, the departure of the likelihood estimate from the reference value is sometimes much larger (see the number of points outside the whiskers in Fig. 3.2). This large departure can be particularly misleading when the true likelihood surface is almost flat. In this situation, it can bring out a false pick in the estimated likelihood surface which shift the position of the maximum likelihood estimate (see the end of Section 3.5.4). In general, for a given number of events, histories with fewer lineages (more coalescences) often correspond to a current weight which is

lower than the others but which might have a high final weight.

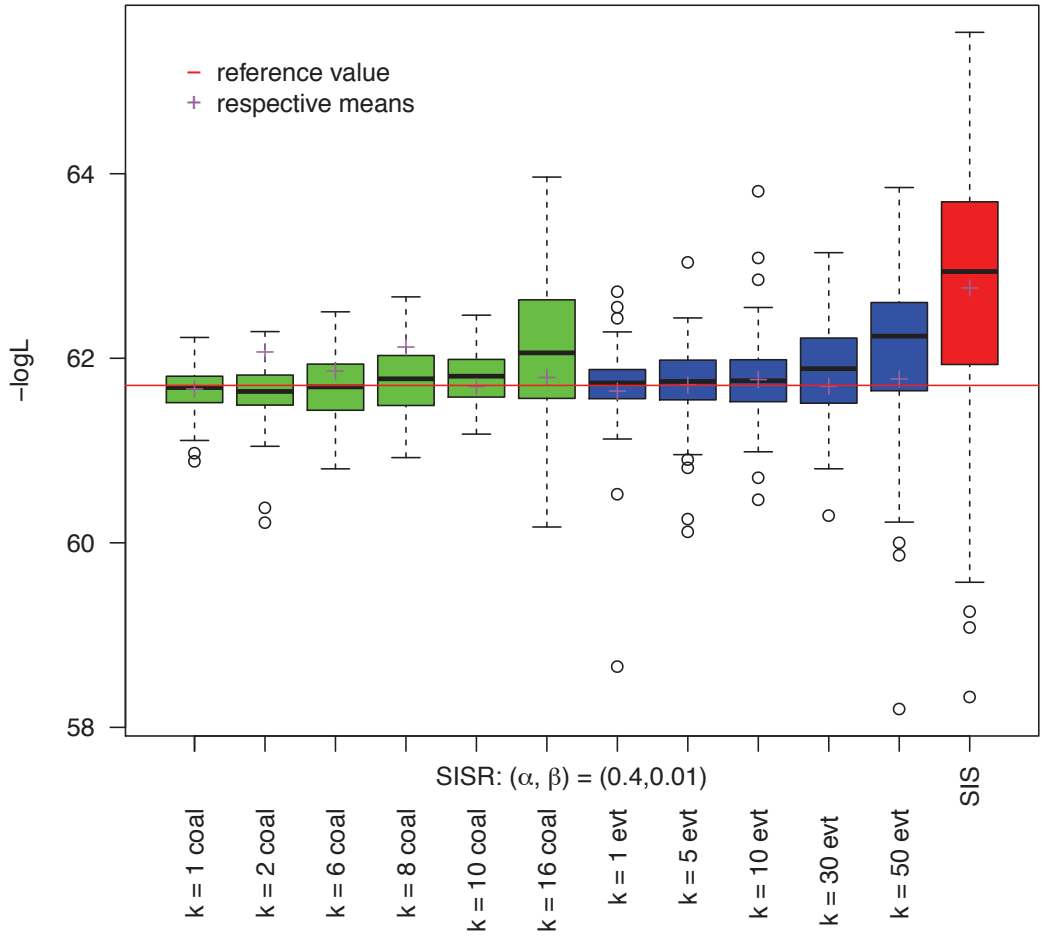


FIGURE 3.2: ***Boxplots of*** 100 ***estimates of the likelihood in a given parameter point with different inference algorithms.*** *The green boxes were obtained by resampling among histories with the same number of coalescences undergone while the blue boxes were obtained by resampling among histories with the same number of events undergone and the red box corresponds to SIS estimates.*

In Fig. 3.3a and 3.3c, we propose to resample every $k = 8, 6, 2$ or 1 coalescence events, which corresponds respectively to resampling 12, 16, 49 or 98 times during Algorithm 3 since the simulated data set is composed of 100 genes. We observe that the variation in likelihood estimation decreases with the frequency of the checkpoints, indicating that we should propose a resampling step as often as possible. Fig. 3.4a and 3.4b leads to the same conclusion regarding the best calibration of k: MSE ratio comparing SISR with $k = 1$ to the SIS estimate (represented by a $*$) are often the lowest MSE ratios when compared to other MSE ratios with the same tuning parameters α and β. Both Fig. 3.4a and 3.4b also indicate that the MSE ratio always decreases with the frequency of checkpoints when the resampling distribution depends on the pairwise composite likelihood ($\beta = 0.01$).

Calibration of the powers α and β of the resampling distribution The efficiency of the resampling distribution of Eq. (3.12) greatly differs between different values of the resampling parameters α and β. For this reason, they must be adequately chosen to

balance the effect of the information provided by the SIS weights and by the composite likelihood. Note that, when relying on SISR, inaccurate likelihood estimates can be obtained only with very poor choices of α and β but here we present only the choices of α and β that provide an improvement of the likelihood estimation. First of all, we notice that the MSE Ratio is substantially below one for all sixty (α, β) couples considered (see Fig. 3.4a and Fig. 3.4b) meaning again that resampling improves the inference of the likelihood in terms of MSE, even when the resampling parameters are not much calibrated. Then we observe that when the resampling distribution only depends on the SIS weights ($\beta = 0$), the MSE Ratio globally decreases when α increases (see Fig. 3.4b). When resampling is performed after each coalescence, as recommended above, we find that any α between 0.5 and 1 is a reasonably good choice (Fig. 3.4c). This is further supported by Fig. 3.3b which represents the boxplots obtained with $k = 1$, $\alpha = 0.4$, 0.5, 0.6, 0.7 and 1 for three different values $\beta = 0$, 0.001 and 0.01. Indeed the variation in likelihood estimates corresponding to $\alpha = 0.4$ is larger than for other values of α whereas the variation in likelihood estimates corresponding to $\alpha = 0.7$ is slightly reduced compared to other α values. In the following we therefore set $\alpha = 0.7$, which is higher than the arbitrary calibration $\alpha = 0.5$ proposed by Liu et al. [2001]. Likewise with $k = 1$, Fig. 3.4c represents the MSE ratios obtained with $\beta = 0$ and 0.01, and different values of α. First we note that the ten values of MSE ratios are between 0.05 and 0.20, which means that the resampling reduces the MSE by a factor 5 to 20. Second the horizontal blue line indicates that the lowest MSE ratio obtained with $\beta = 0$ is higher than the MSE ratio obtained with five out of seven values of α when $\beta = 0.01$, that is when using the composite likelihood. Moreover, this choice of β is supported by Fig. 3.3d which sets $\alpha = 0.7$ and shows that the span of the distribution of the likelihood estimate is smaller when $\beta = 0.01$ (and $k = 1$) than with other values of β.

In conclusion we can choose $k = 1$, $\alpha = 0.7$ and $\beta = 0.01$ although other choices of tuning parameters of the SISR also lead to an improvement of the likelihood estimation. The MSE ratio for the likelihood estimates using SIS vs. SISR is lower than 60% for most choices of the resampling parameters, and below 10% when the procedure proposes to resample after every coalescence event and when using the PCL, which allows a strong improvement of the likelihood estimation in a parameter point.

3.5 Improvements in the likelihood based inference of demographic parameters

The final aim of inference from a data set is not to compute the likelihood at a given point in parameter space. Rather, it is to provide an estimate of the parameter $\phi = (\theta, D, \theta_{\text{anc}})$ and confidence intervals (CI) around each coordinates of ϕ, i.e, marginal CIs. Thus, we will quantify the impact of our procedures on the performance of such inferences.

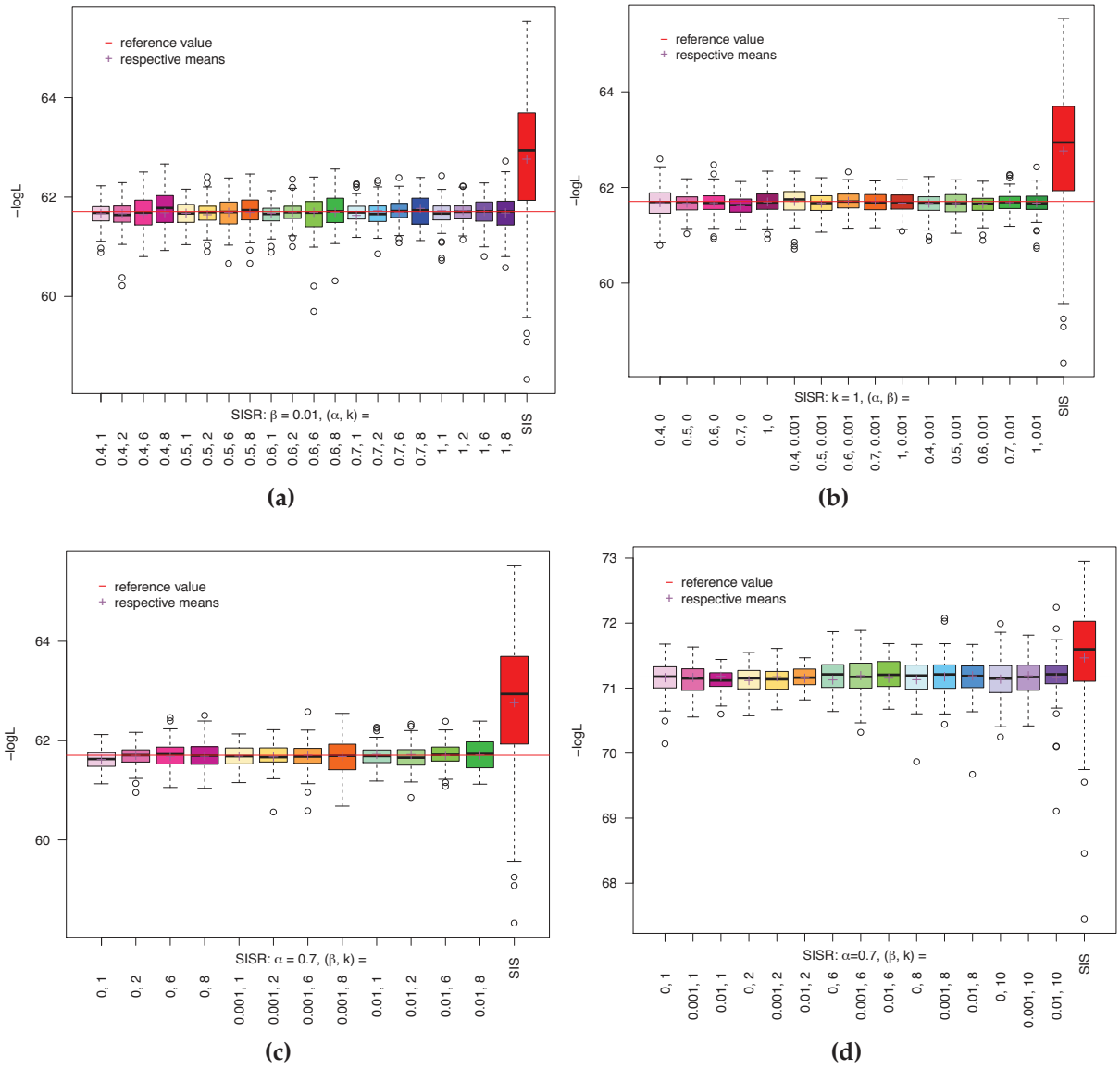


FIGURE 3.3: ***Boxplots of*** 100 ***estimates of the likelihood in a given parameter point with different inference algorithms.*** *The red box corresponds to SIS estimates. The red horizontal line represents the reference value and the cross on each box represents the mean of the* 100 *estimates of this box. Each block of three to five boxes with similar colors (pink, orange, green, blue and mauve) corresponds to SISR with a fixed value: (a) of* α *for four different values of* k *when* $\beta = 0.01$ *is fixed for all the SISR estimations, (b) of* β *for five different values of* α *when* $k = 1$ *is fixed for all the SISR estimations, (c) of* β *for four different values of* k *when* $\alpha = 0.7$ *is fixed for all the SISR estimations, and (d) of* k *for three different values of* β *when* $\alpha = 0.7$ *is fixed for all the SISR estimations. See main text for details about* k*,* α *and* β*.*

3.5.1 Inference method

In the numerical results below, we conduct a maximum likelihood analysis of the data. An estimate of the parameter ϕ is given by the maximum likelihood estimate (MLE) $\widehat{\phi}(\mathbf{x})$ of a multilocus data set $\mathbf{x}$. The multilocus likelihood is the product of the

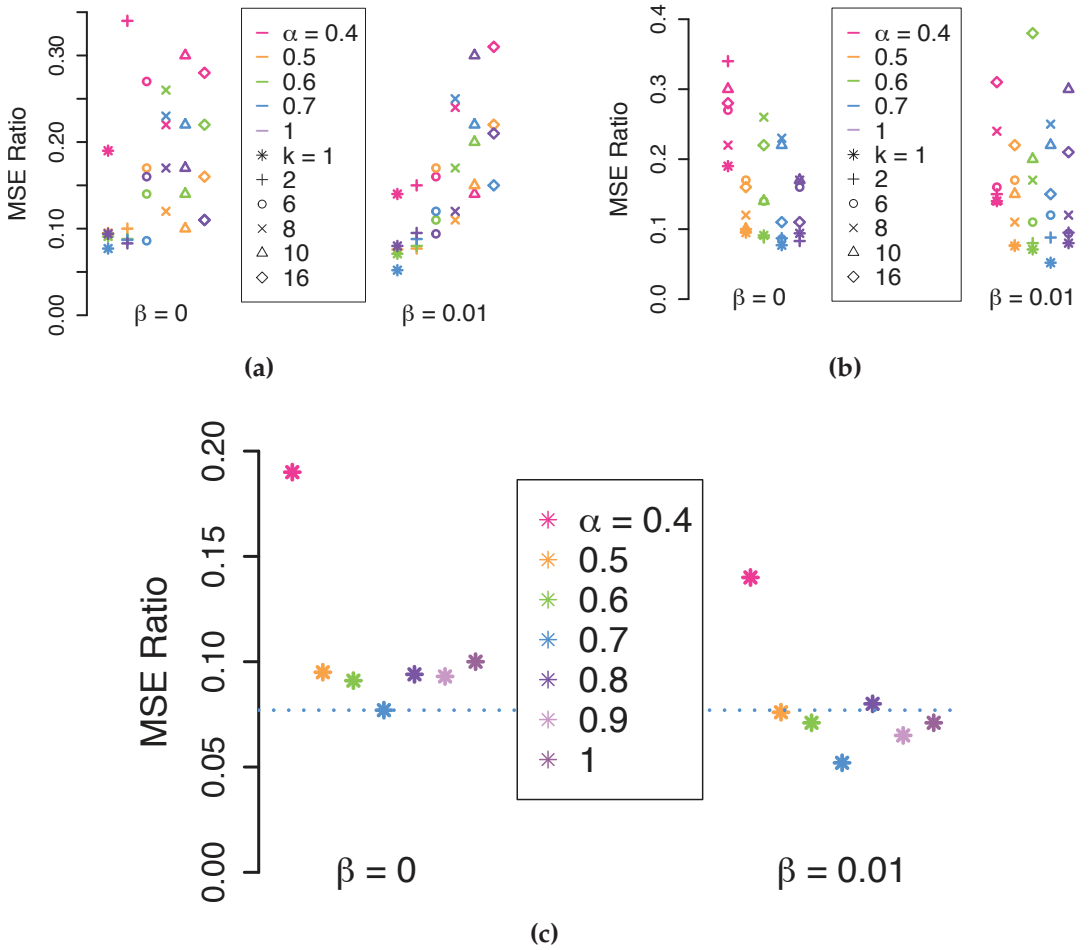


FIGURE 3.4: ***MSE ratios obtained with*** $\beta = 0$ ***and*** 0.01. *(a) and (b) both represent the same MSE ratios but differ on the arrangement of the points. Each color corresponds to a value of* α *and each shape corresponds to a value of* k*, (a) each vertical alignment of five points corresponds to a fixed value of* k *for five different values of* α*, (b) each vertical alignment of six points corresponds to a fixed value of* k *for six different values of* α*. (c) MSE ratios obtained with* $k = 1$ *and different values of* α*. The horizontal dotted blue line represents the lower MSE ratio obtained with* $k = 1$ *and* $\beta = 0$ *among different values of* α*. See main text for details about* k*,* α *and* β*.*

likelihoods for each locus x_ℓ, $\ell = 1, \ldots, d$:

$$L_d(\phi, \mathbf{x}) = \prod_{\ell=1}^{d} L(\phi, x_\ell),$$

where d is the number of loci in the sample, and each term of the product can be estimated with a SIS or a SISR algorithm. The biological assumption that permits the above writing is that the loci are distant enough in the genome to have independent past histories.

Then we derived marginal CIs on each coordinate, from likelihood-ratio test based on the profile likelihood [see Davison, 2003, for details]. To obtain a numerical value of the marginal CIs, we rely on asymptotic theory which states that under H_0 profile log likelihood-ratio is approximately χ^2-distributed (e.g., Severini, 2000), here with a

degree of freedom equal to 1 when the size of the data set is large, i.e., when there is enough information on the parameter in the data.

3.5.2 Inference algorithm and its evaluation

As in Rousset and Leblois [2007], Rousset and Leblois [2012] and Leblois et al. [2014] we can conduct the inference process as follows. We first define a set of parameter points through a stratified random sample within a range of the parameter space provided by the user. Then, at each parameter point, the multilocus likelihood is the product of the likelihoods for each locus, which are estimated through the SIS or SISR algorithm. The likelihoods inferred at the different parameter points are then smoothed by a Kriging scheme. After a first analysis of the smoothed likelihood surface, the algorithm can be repeated a second time to increase the density of parameter points in the neighborhood of the first MLE. The Kriging step removes part of the estimation error of likelihood in any given parameter point by assuming that the likelihood is a smooth function of the parameter. In this Section, we thus conducted numerical experiments to show that the gain of SISR over SIS in accuracy of likelihood estimates is retained through the Kriging step.

A convenient way to evaluate numerically the whole inference procedure, and in particular the coverage of the marginal CIs, is to check that distribution of p-value of the likelihood-ratio test of $H_0 : \phi_1 = \phi_1^*$ against $H_1 : \phi_1 \neq \phi_1^*$ is uniform on the interval $[0;\ 1]$ when the data set $\mathbf{x}$ is simulated from the model with $\phi_1 = \phi_1^*$. To this end we represent the empirical cumulative distribution function (ECDF) of the p-value on many simulated data sets $\mathbf{x}$, which should be closed to the $1 : 1$ diagonal. Deviation from a uniform distribution can occur: either because the likelihood is poorly estimated or because the exact profile log likelihood ratios do not follow the asymptotic χ^2 distribution. We perform a Kolmogorov-Smirnov test to assess the uniform distribution of the sample of p-values.

We have also computed other measures of the performance of the inference method of ϕ_1 on many simulated data sets $\mathbf{x}_i$ from $\phi_1 = \phi_1^*$, namely

- the mean relative bias of the MLE , computed as

$$\frac{(\text{observed bias on } \phi_1^*)}{\phi_1^*} = \frac{1}{\phi_1^*}\left(\text{mean}_i\, \widehat{\phi}_1(\mathbf{x}_i) - \phi_1^*\right),$$

- and the relative root mean square error (relative RMSE) of the MLE , computed as

$$\sqrt{\frac{\text{MSE on } \phi_1^*}{\phi_1^{*2}}} = \frac{1}{\phi_1}\sqrt{\text{mean}_i\left(\widehat{\phi}_1(\mathbf{x}_i) - \phi_1^*\right)^2}.$$

3.5.3 Numerical experiment cases and previous results

We performed numerical experiments on four different demographic scenarios, all modeled according to the exponential contraction of Section 3.4.1, which are as follows.

(i) $\phi = (\theta, D, \theta_{\text{anc}}) = (0.4, 1.25, 40)$, which is the baseline scenario of Leblois et al. [2014]: the population size has undergone a contraction of strength $N_{\text{ratio}} = 0.01$ and $D = T/2N = 1.25$, where N is the size of the population at time $t = 0$,

(ii) $\phi = (\theta, D, \theta_{\text{anc}}) = (0.4, 1.25, 400)$ which differs from the baseline scenario in the strength $N_{\text{ratio}} = 0.001$ of the contraction,

(iii) $\phi = (\theta, D, \theta_{\text{anc}}) = (0.4, 0.25, 40)$ which differs from the baseline scenario in the speed of the contraction, since it corresponds to a contraction of strength $N_{\text{ratio}} = 0.01$ but $D = T/2N = 0.25$ which is five times smaller than in the baseline scenario,

(iv) $\phi = (\theta, D, \theta_{\text{anc}}) = (0.4, 0.25, 400)$ which represents the strongest and recentest contraction of the four scenarios. Numerical results on the estimation of the likelihood in this last demographic scenario have already been presented in Section 3.4 above.

The baseline scenario is a case where the inference procedure performs well when the likelihood of each locus at each point of the parameter space is estimated with the sequential importance sampling on $n_H = 2,000$ histories, but $n_H = 100$ gives satisfactory results. A more careful look at the results shows that the profile likelihoods exhibit clear peaks around the MLE for all parameters. This is a first evidence that the data contain information regarding all parameters. And, indeed, the ECDF of the p-values are almost aligned on the $1:1$ diagonal (Fig. 3.5).

3.5.4 Numerical results

To analyze the performances of the inference procedure we have simulated 500 data sets for each scenario, composed of 10 independent loci and 100 genes per locus.

Gain in accuracy of the MLE The numerical results regarding the relative bias and relative RMSE are given in Table 3.1. We find that the resampling technique allows overall to reduce the relative RMSE and the relative bias, mostly for the parameter θ.

(i) For the baseline scenario $(\theta = 0.4, D = 1.25, \theta_{\text{anc}} = 40)$, the resampling procedure allows to reduce the relative bias on θ by 30%, the relative bias on D by 20% and the relative bias on θ_{anc} by 80% and also the relative RMSE on θ_{anc} by 10%, other values being similar.

(ii) In the situation $(\theta = 0.4, D = 1.25, \theta_{\text{anc}} = 400)$ of a stronger contraction but not too recent, the resampling procedure allows to reduce the relative bias on θ by 40%, and also the relative RMSE on θ by 35%, other values being similar.

(iii) Concerning the scenario $(\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 40)$ of a more recent contraction than the baseline situation but with the same strength, the resampling procedure allows to reduce the relative bias on θ by 79% and the relative RMSE on θ by 10%. However, the relative bias and relative RMSE on D increase by a factor of 2.5 and 1.8 respectively.

(iv) We observe the same trend with the situation $(\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 400)$ of a more recent and stronger contraction than the baseline situation. Indeed, when using resampling, the relative bias and relative RMSE on θ decrease by a factor 4.5 and 3 respectively while the relative bias and relative RMSE on the parameter D increase by a factor 2.5 approximately.

TABLE 3.1: ***Accuracy of the MLE with SIS and SISR algorithms***

$(\theta, D, \theta_{anc}) =$		$(0.4, 1.25, 40)$		$(0.4, 1.25, 400)$		$(0.4, 0.25, 40)$		$(0.4, 0.25, 400)$	
Algorithm		SIS	SISR	SIS	SISR	SIS	SISR	SIS	SISR
with $n_H =$		100		100		$2,000$		$2,000$	
rel. bias	θ	0.17	0.12	0.56	0.34	1.07	0.23	4.9	1.1
	D	0.063	0.051	-0.02	-0.017	0.23	0.571	-0.06	0.15
	θ_{anc}	0.076	0.016	0.048	-0.042	0.032	0.023	0.044	-0.053
rel. RMSE	θ	0.47	0.46	0.71	0.53	2	1.8	5.2	1.7
	D	0.28	0.28	0.14	0.14	0.45	0.8	0.14	0.37
	θ_{anc}	0.53	0.48	0.37	0.38	0.29	0.27	0.25	0.23

Reducing the size of the collection of histories thanks to the resampling We compare here the inference method relying on estimates of the likelihood either based on SIS or on SISR with a smaller collection of histories per locus and per point of the parameter space.

We first compare both methods on the baseline situation $(\theta = 0.4, D = 1.25, \theta_{\text{anc}} = 40)$ of a relatively weak and not too recent contraction. In this case, the SIS algorithm performs well due to the large amount of information in the genetic data (Leblois et al., 2014).

We find out that SISR with 50 ancestral histories produces comparable results to SIS with 100 ancestral histories. With the same number of ancestral histories, the relative bias and relative RMSE are lower with SISR than with SIS as explained above, and the ECDF of the p-values are closer to the diagonal as shown in Fig. 3.5. We conclude that resampling improves the parameter estimation in a situation where the SIS performs well, dividing by 2 the required number of ancestral histories.

We then compare both methods on a more difficult situation $(\theta = 0.4, D = 1.25, \theta_{\text{anc}} = 400)$ of a stronger contraction. In this situation, the SIS procedure performs less well but we obtain satisfying results with $n_H = 2000$ sampled histories (Leblois et al., 2014),

which leads to reasonable computation time. Indeed, for a single data set with one hundred gene copies and ten loci, analyses are carried out in 8 hours in C++ process time on average. Here we decrease the number of ancestral histories in both SIS and SISR, in order to show that the SISR performs well when the SIS does not. For both procedures we explored $n_H = 100, 200$ and 400 ancestral histories per parameter point. With 400 ancestral histories, the analyses are carried out in 1.7 hours in C++ process time on average, for a single data set with one hundred gene copies and ten loci.

Fig. 3.6 shows that with the same number of explored ancestral histories, we obtain lower relative bias and relative RMSE and also better ECDF, which means better coverage properties of the CIs. By comparing Fig. 3.6c with 3.6b, we also find that the SISR procedure provides comparable results (relative bias, relative RMSE and ECDF) to the SIS procedure with half of the number of explored ancestral histories, as in the baseline scenario. The comparison of Fig. 3.6e with Fig. 3.6d shows the same results.

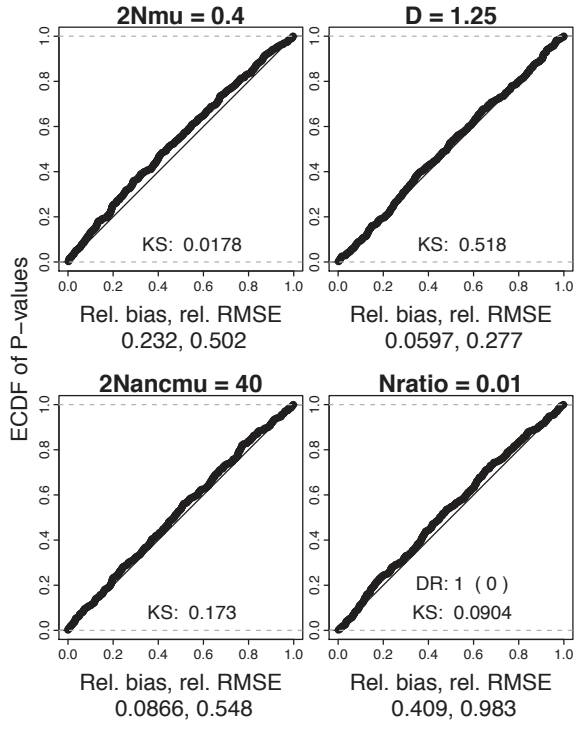
Improvements in more difficult situations We compare both SIS and SISR procedure on two more difficult situations $(\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 40)$ and $(\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 400)$ of very recent contraction, very strong for the second case. Thus in these two cases, the proposal distribution is inefficient. In the extreme case of $(\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 400)$, the SIS algorithm does not provide satisfactory CI coverage properties, even with $200,000$ sampled histories.

We first consider the situation ($\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 40$) in which the magnitude of the contraction is the same as the baseline case but here it occurred much more recently. In this situation, the SIS procedure performs less well and we do not obtain satisfactory results with 2000 sampled histories (Leblois et al., 2014). With the same number of explored ancestral histories, applying the resampling technique, we obtain lower relative bias and relative RMSE for θ but higher for D. We also obtain less good CI coverage properties for θ and D (see Fig. 3.7). We conclude that the method detects the contraction, estimates the date and the ancestral size of the population but does not find enough information in the data about the current size of the population. Indeed, the likelihood surface is quite flat.

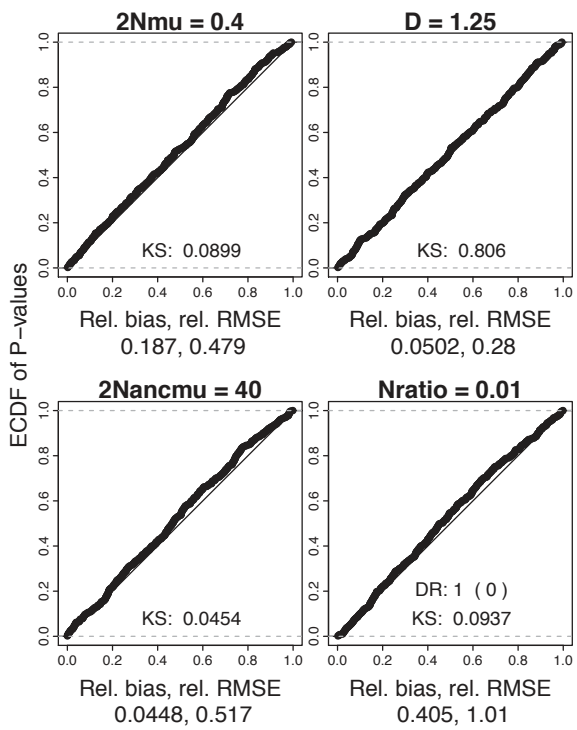
We then consider the extreme situation with a very recent and stronger past contraction ($\theta = 0.4, D = 0.25, \theta_{\text{anc}} = 400$). Increasing greatly the number of ancestral histories sampled per parameter point up to $200,000$ and consequently the computation times by a factor 100, decreases relative bias and relative RMSE on θ but does not provide satisfactory CI coverage (Leblois et al., 2014). The resampling technique allows us to decrease relative bias and relative RMSE of θ for a fixed number of ancestral histories sampled and also to provide better CI coverage (see Fig. 3.8). Comparing Fig. 3.8c and 3.8d, we obtain approximately the same relative bias and relative RMSE on θ and close CI coverage performance with SISR with 100 times less histories sampled than with SIS. In this scenario as in the previous one there is not enough information in the data and hence the likelihood surface is flat. We also face an issue due to the strength of the contraction and the inefficiency of the proposal distribution in this situation, in addition to the previous difficulty. The resampling technique allows a global gain, since it corrects the inefficiency of the proposal distribution in disequilibrium situations. However, in this situation as in the previous one, the resampling technique does not bring

an improvement to the lack of information in the data about the parameter θ when the contraction is too recent.

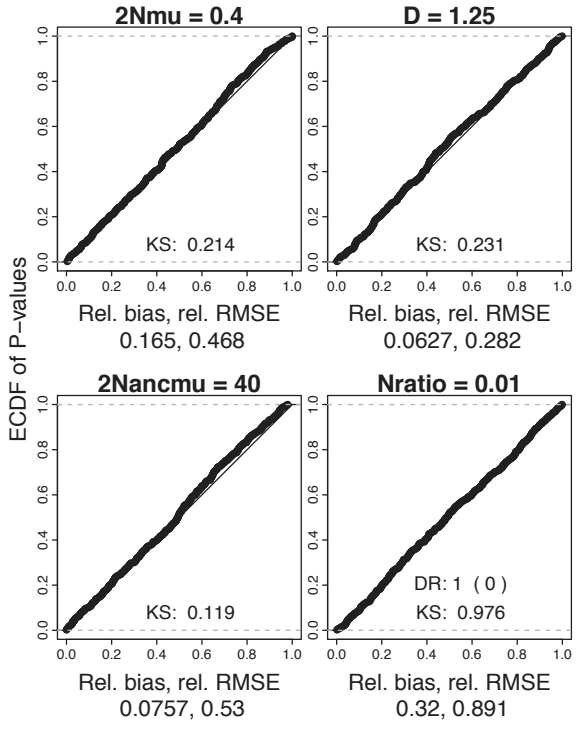
About flat likelihood surfaces The inference method of Section 3.5.1 suffers from two major defects when the likelihood surface is flat. First any error regarding the likelihood estimate at some point of the parameter space can lead to artificial local maxima or minima in the smoothed surface, depending on how the Kriging method performs. Second even if we recover the true flat likelihood surface, the distribution of the likelihood-ratio statistic is not correctly approximated by the χ^2 distribution because of the lack of information on the parameter in the data. Both defects can be seen in our numerical studies. First the median of the SIS likelihood estimates is much lower than the reference value in Fig. 3.2 to 3.3d. It indicates that the SIS likelihood estimates are very often below the true value of the likelihood which can lead to artificial local minimum in the smoothed likelihood surface obtained with the Kriging algorithm. When comparing Fig. 3.9 (a) and (b), the SIS likelihood estimates introduce a local minimum at $\theta = 2N\mu \approx 0.001$ while the more reliable SISR likelihood estimates manage to recover the flat likelihood surface. Likewise on $D = T/2N$, the SIS likelihood estimates introduce a local minimum around $D \approx 0.4$, see Fig. 3.9 (c), while the SISR likelihood estimates recover the flat likelihood surface, see Fig. 3.9 (d). The second defect can be seen on the p-value's ECDF in Fig. 3.8. Even with the SISR likelihood estimates which manage to recover the flat likelihood surface, the Figure exhibits a departure from the uniform distribution, meaning that the χ^2 approximation is not accurate. Flat likelihood surface as in Fig. 3.9 and very large (even if untrustworthy) CIs, for instance the 95% CI with SISR likelihood estimates is approximately $[0.00016; 2.5]$ on $\theta = 2N\mu$ and $[0.14; 0.72]$ on $D = T/2N$, should act as warnings that the data do not carry much information about the parameters of interest.



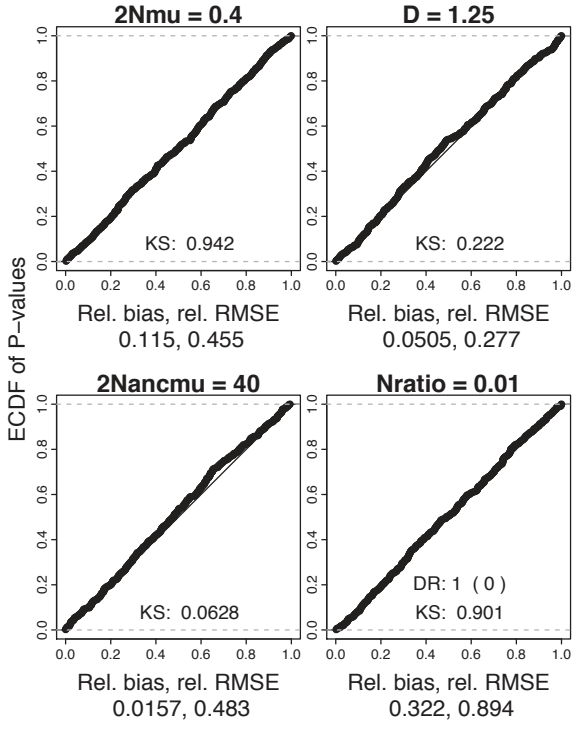
(a) ***SIS*** $n_H = 50$



(b) ***SISR*** $n_H = 50$

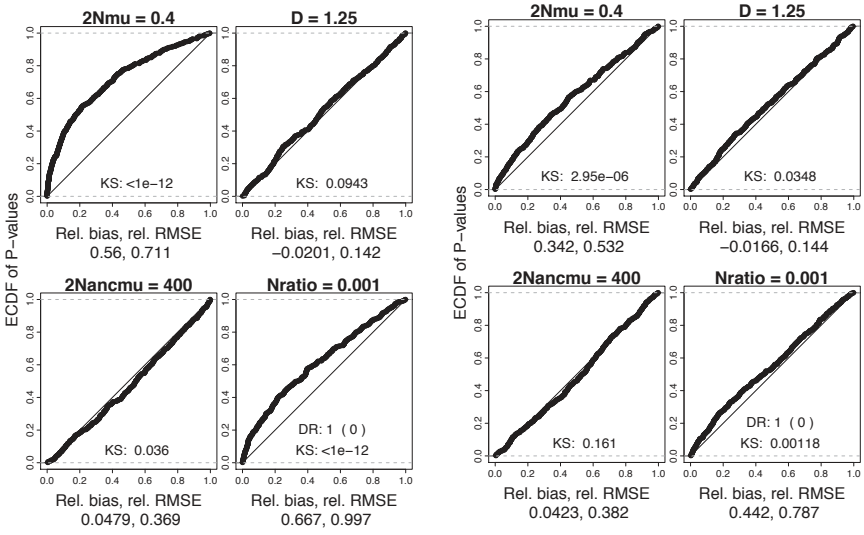


(c) ***SIS*** $n_H = 100$



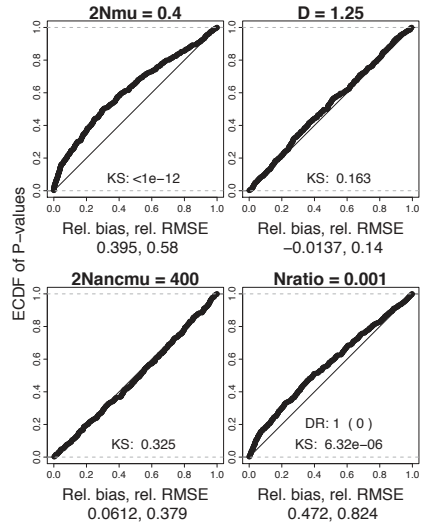
(d) ***SISR*** $n_H = 100$

FIGURE 3.5: ***Empirical Cumulative Distribution Functions (ECDF) of p-values of likelihood ratio tests for the scenario*** $\theta = 0.4$, $D = 1.25$ ***and*** $\theta_{\textbf{anc}} = 40$***.*** *Inference (a) with the SIS procedure with* $n_H = 50$ *sampled histories (b) with the SISR procedure with* $n_H = 50$ *(c) SIS with* $n_H = 100$ *(d) SISR with* $n_H = 100$*, on* 500 *simulated data sets. Relative bias and relative RMSE are also reported, and KS indicate the p-value of the Kolmogorov-Smirnov test for departure of LRT p-values distributions from uniformity.*

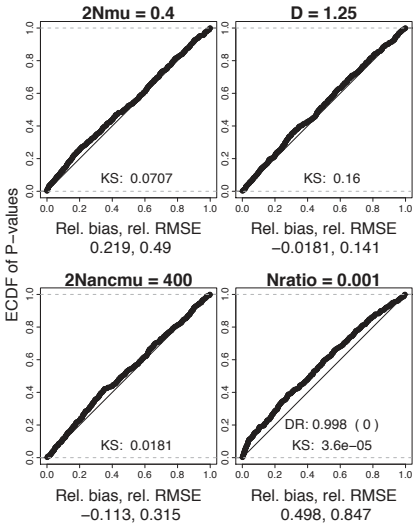


(a) ***SIS*** $n_H = 100$

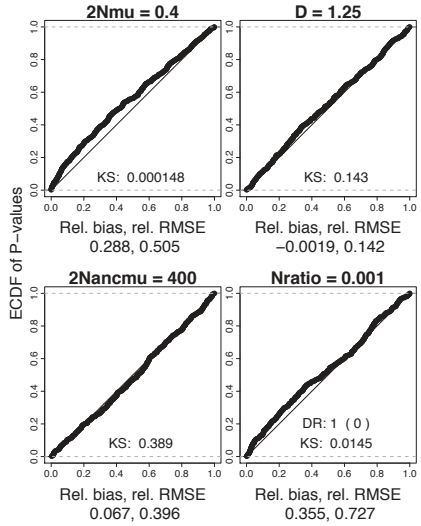
(b) ***SISR*** $n_H = 100$



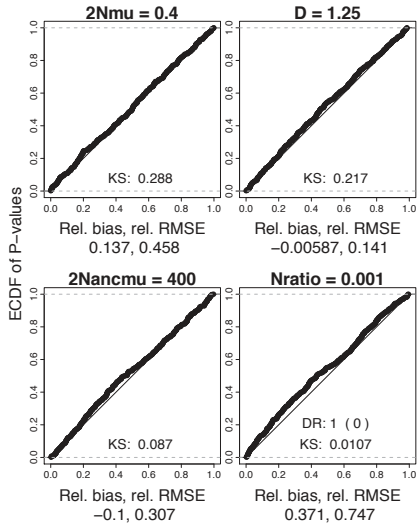
(c) ***SIS*** $n_H = 200$



(d) ***SISR*** $n_H = 200$

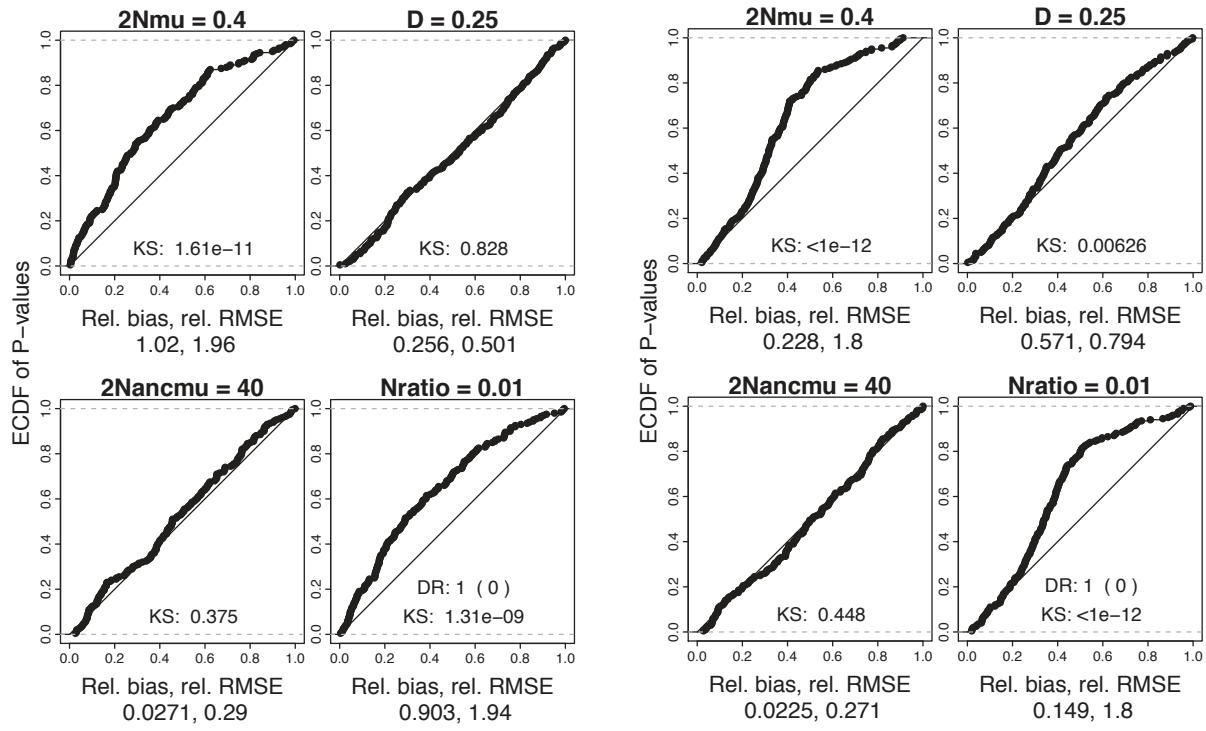


(e) ***SIS*** $n_H = 400$



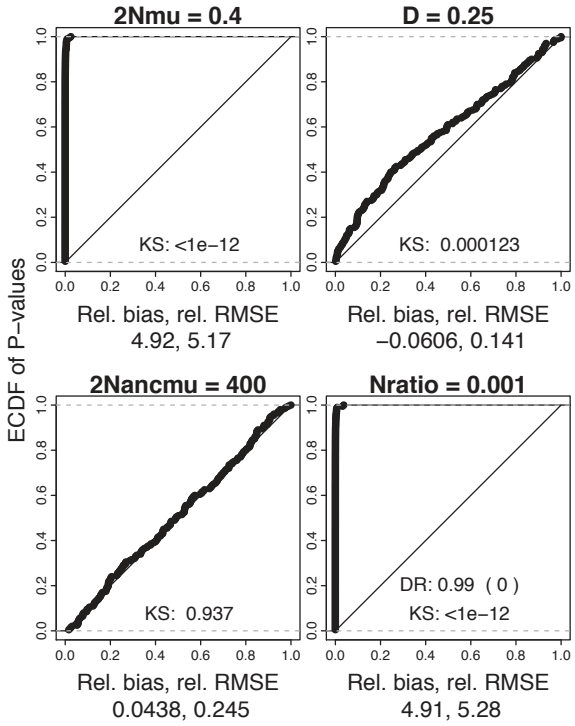
(f) ***SISR*** $n_H = 400$

FIGURE 3.6: ***ECDF of p-values of likelihood ratio tests for the scenario*** $\theta = 0.4$, $D = 1.25$ ***and*** $\theta_{\textbf{anc}} = 400$. *(a) and (b) with* $n_H = 100$ *(c) and (d) with* $n_H = 200$ *and (e) and (f) with* $n_H = 400$, *on* 500 *data sets. See Fig. 3.5 for details.*

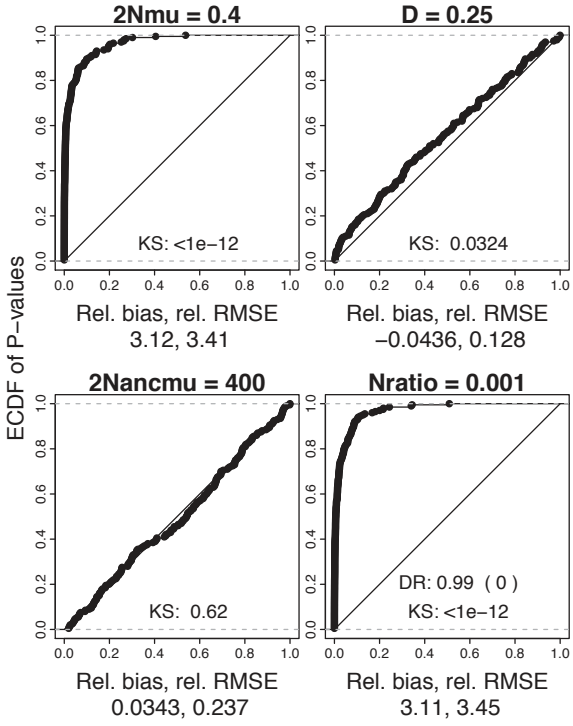


(a) ***SIS*** $n_H = 2000$ **(b)** ***SISR*** $n_H = 2000$

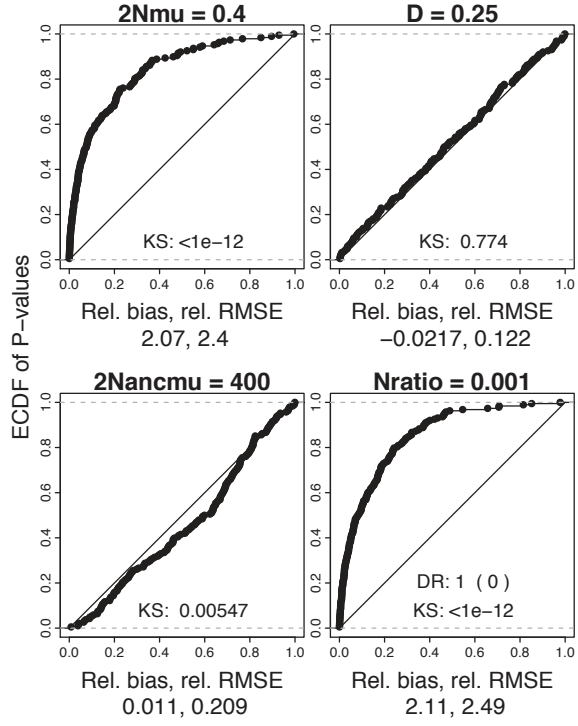
FIGURE 3.7: ***ECDF of p-values of Likelihood ratio tests for the scenario*** $\theta = 0.4$***,*** $D = 0.25$ ***and*** $\theta_{\mathbf{anc}} = 40$***, with*** $n_H = 2000$ *sampled histories, on* 200 *simulated data sets. See Fig. 3.5 for details.*



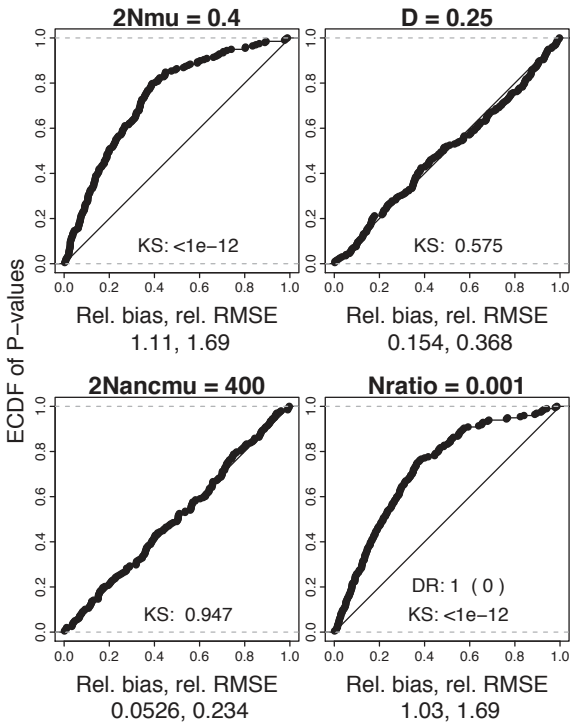
(a) ***SIS*** $n_H = 2000$



(b) ***SIS*** $n_H = 20000$



(c) ***SIS*** $n_H = 200000$



(d) ***SISR*** $n_H = 2000$

FIGURE 3.8: ***ECDF of p-values of Likelihood ratio tests for the scenario*** $\theta = 0.4$***,*** $D = 0.25$ ***and*** $\theta_{\mathbf{anc}} = 400$***.*** *(a) SIS with* $n_H = 2,000$ *sampled histories (b) SIS with* $n_H = 20,000$ *sampled histories (c) SIS with* $n_H = 200,000$ *sampled histories (d) SISR with* $n_H = 2,000$ *sampled histories, on* 200 *simulated data sets. See Fig. 3.5 for details.*

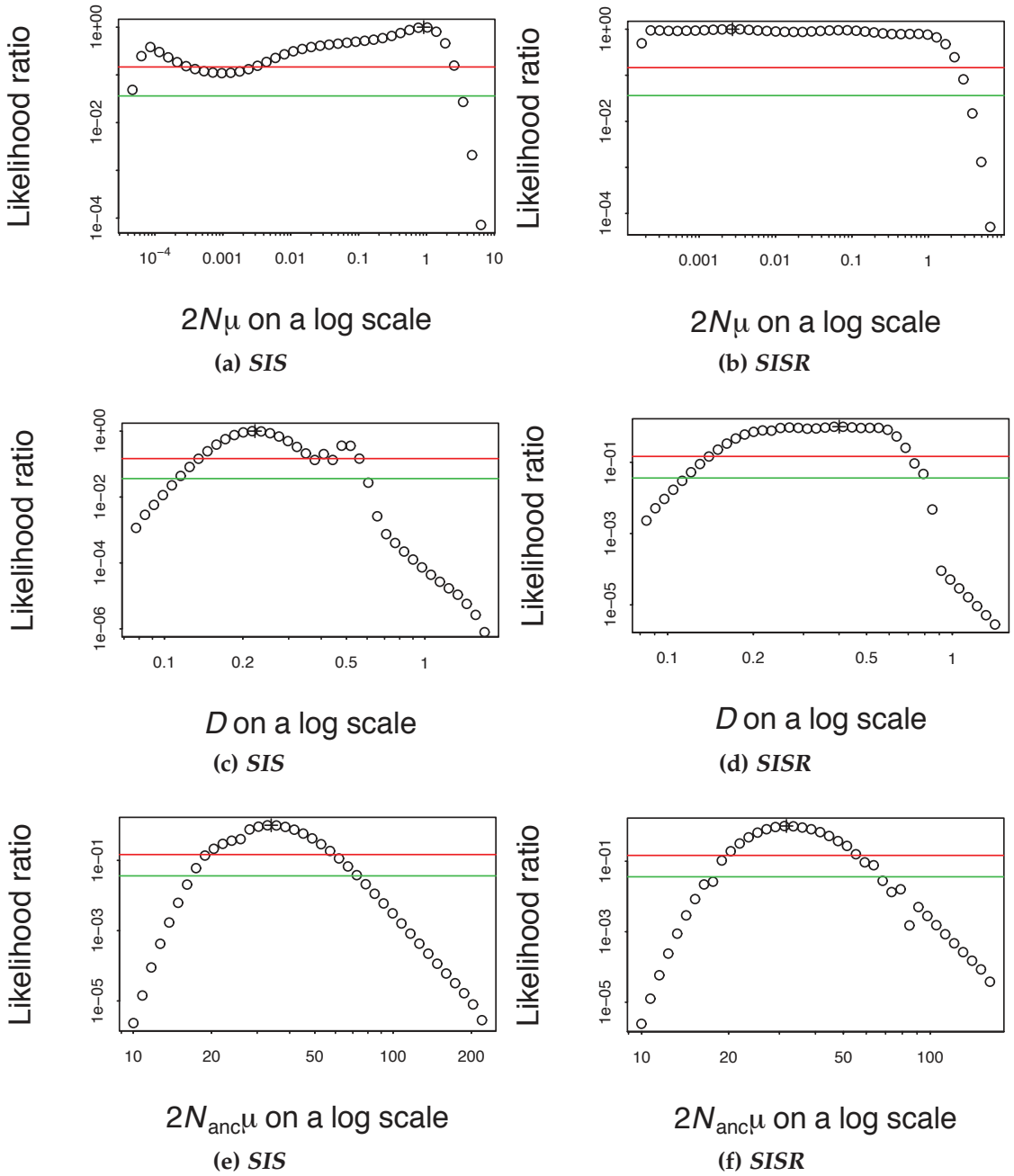


FIGURE 3.9: ***One-parameter profile likelihood ratios***. *(a) and (b) represent the likelihood profile function divided by its maximum value* $\theta \mapsto \hat{L}_\theta(\hat{D}_\theta, \hat{\theta}_{\text{anc}\,\theta}) / \hat{L}(\hat{\theta}, \hat{D}_\theta, \hat{\theta}_{\text{anc}\,\theta})$*, (c) and (d) represent in the same way the profile likelihood ratio for* D *and (e) and (f) represent the profile likelihood ratio for* θ_{anc}*, estimated respectively with SIS (left) or with SISR (right). See Appendix. 3.9.3 for details on profile likelihood.*

3.6 Cynopterus sphinx data set

We applied the inference method of Section 3.5.1 on the fruit bat *Cynopterus sphinx* data set presented in Storz and Beaumont [2002], consisting of allelic frequencies computed from a sample of 246 individuals, hence 492 genes, genotyped at 8 microsatellite loci. Using a coalescent-based Monte Carlo Markov Chain (MCMC) algorithm, Storz and Beaumont [2002] found a strong evidence for a pronounced population contraction with (i) a flat posterior distribution for the strength N_{ratio} of the contraction, characterized by large 50% and 90% highest probability density (HPD) intervals of $[2\cdot 10^{-5} - 0.05]$ and $[< 10^{-6} - 0.13]$, respectively; and (ii) a more peaked posterior distribution for the scaled time of occurrence, $D = T/2N$, with a mode around 0.3 and 90% HPD interval of $[0.10 - 0.65]$. (Note that the parameter t_f of Storz and Beaumont [2002] is defined as T/N rather than $T/2N$ as in this work).

Here, we compare their results with those obtained with our methods, with an emphasis on the gain in precision and in computation cost using the SISR algorithm with different resampling probabilities vs. the SIS algorithm. To assess the accuracy of the different algorithms on this real data for which we do not have concrete expectations, we rely on the following steps.

We know from the results presented in the previous sections that both SIS ans SISR algorithms may give inaccurate point estimates and CIs for some parameters if the number of simulated histories is too low to infer the likelihood at each parameter point with enough precision. For this reason, we first compared parameter estimates obtained by analyzing the data with the SIS algorithm using $n_H = 1,000; 10,000; 100,000$ and $1,000,000$ sampled histories per point. Moreover, to evaluate the IS variance of the log-likelihood estimate by the importance sampling algorithm, we also considered the likelihood RMSE estimate from kriged duplicate points. It is computed from independent pairs of likelihood estimates at the same parameter point for the points retained at the Kriging step.

TABLE 3.2: ***Evolution of the MLE with SIS when increasing the number of sampled histories per point, on the bat data set***

n_H	$\hat{\theta}$	$\hat{D}$	$\widehat{\theta_{\text{anc}}}$	$\widehat{N_{\text{ratio}}}$	$\ln(\hat{L}(\hat{\phi}))$	lik-RMSE
$1,000$	1.1	0.44	320	0.0034	-663.74	9.2
	$[0.76 - 1.5]$	$[0.35 - 0.55]$	$[180 - 560]$	$[0.0018 - 0.0062]$		
$10,000$	0.80	0.45	340	0.0024	-639.08	8.3
	$[0.53 - 1.16]$	$[0.37 - 0.55]$	$[205 - 553]$	$[0.0013 - 0.0045]$		
$100,000$	0.62	0.48	460	0.0014	-620.68	5.7
	$[0.38 - 0.94]$	$[0.40 - 0.57]$	$[254 - 492]$	$[0.00078 - 0.0027]$		
$1,000,000$	0.42	0.49	380	0.0011	-606.59	4.4
	$[0.24 - 0.68]$	$[0.41 - 0.59]$	$[240 - 589]$	$[0.00056 - 0.0022]$		

Our results, presented in Table 3.2, show a strong decrease in the MLEs of θ et N_{ratio}, from 1.1 to 0.42 and 0.0034 to 0.00111 respectively , with an associated shift and narrowing of CIs. An opposite and weaker trend of increase is also observed for $\hat{D}$, from 0.44 to 0.49, and for $\widehat{\theta_{\text{anc}}}$, from 320 to 380, with a slight shift of CIs for both parameters.

As expected, increasing the number of simulated histories leads to a decrease in the variance of the likelihood estimate at each parameter point. But more interestingly, it also leads to an important increase of the log-likelihood value at the MLE $\ln(\hat{L}(\hat{\phi}))$, from -664 to -607. This first step implied a very high computational burden but allowed us to see a clear trend towards an improvement of the SIS inference for each analyses with more histories. We probably did not reach a plateau in the precision even when considering $n_H = 1,000,000$ histories (which took approximately 15000 hours in C++ process time), but considering more histories is merely unfeasible. We thus consider the results obtained with $1,000,000$ histories as the best result we can get with the SIS algorithm, irrespectively of computation costs.

In a second step, we compared estimations obtained with the SIS and various versions of SISR algorithms by playing with different resampling strategies, for a fixed and reasonable computation cost. To this end, we fixed the number of sampled histories per parameter point n_H to 1,000, leading to an average of 15 hours in C++ process time for each analysis. Table 3.3 presents, for each analysis, the point estimates and CIs for each parameter of interest, including the N_{ratio}, as well as the log-likelihood value estimated at the MLE $\ln(\hat{L}(\hat{\phi}))$ and the RMSE estimate for the log-likelihood at each parameter point (i.e. the IS log-likelihood variance).

A first global conclusion is that all SISR analyses with different resampling probability distributions lead to very similar results both in terms of parameter estimates and maximum log-likelihood value over the surface. Moreover, results from SISR analyses differ from those obtained with the SIS algorithm with the same number of sampled histories, but are closer to the best results we get with the SIS algorithm when considering 1,000,000 histories. Our results thus show that SISR analyses on this data set lead to more accurate inferences than SIS for similar computation costs. According to Table 3.3, the most significant improvement due to resampling concerns the parameter θ, which is the most poorly estimated by the SIS algorithm with $n_H = 1,000$. To a lesser extent, resampling also improves inference of the parameter D compared to the SIS algorithm, but conclusions about the parameter θ_{anc} are less obvious, even if almost all analyses lead to values closer to the best estimate we have than the one obtained without resampling.

Tackling precisely the effects of the different resampling strategies on this real data set analysis is more difficult as we do not have in hand the true values of the different parameters and also because our results show subtle differences among parameters for different resampling strategies. Thus, we can only globally conclude that resampling, especially when using PCL, always lowers the likelihood estimation variance and increases the maximum likelihood value on this data set compared to the analysis without resampling. It should also be noted that resampling according to a distribution which does not depend on the IS weights but only on the PCL also improves the inference precision compared to SIS in terms of likelihood estimation variance, maximum likelihood value and parameter estimates (results not shown).

Overall, our results on this fruit bat data set qualitatively agree with those of Storz and Beaumont [2002] showing a strong evidence for a past population contraction. However, a comparison between their results and our own inferences shows slight quantitative differences in point estimates and associated CIs. Our analyses globally show

greater precision for all parameter estimates with CIs that are much narrower than their credibility intervals (see above). This is especially true for the strength N_{ratio} of the contraction, for which we found good inference precision using our method compared to the flat posterior distributions obtained by Storz and Beaumont [2002]. Note that all results from Storz and Beaumont [2002] reported here so far were obtained using a non-hierarchical Bayesian model in which the model parameters cannot differ between loci, similarly to what is done in our method. Considering a hierarchical model with variability among loci, they found slightly different results. They obtained a more peaked posterior distribution for N_{ratio} and conversely, a flatter posterior distribution for the timing. Compared to our analyses, this latter analysis of Storz and Beaumont [2002] principally suggests a much weaker past contraction (e.g. $\widehat{N_{\text{ratio}}}$ around 0.03 for their hierarchical analysis vs. 0.0005 for our method) that occurred slightly further in the past with our $\hat{D}$ being about 0.55 vs. 0.18. Moreover, despite Storz and Beaumont [2002] do not present their results in terms of θ and θ_{anc}, we found that our scaled estimates of current population size (i.e. θ) are congruent with their unscaled estimates of current population sizes and mutation rates. The large difference observed under the hierarchical model in the amplitude of the past contraction inferred by Storz and Beaumont [2002] thus comes mostly from differences in past population size estimates. Our SIS and SISR analyses inferred ancestral population sizes that are hundred times larger than their estimates. This latter result does not seem to be due (1) to a potential bias in our methods as ancestral population size estimates do not decrease with increasing precision (low RMSE) of estimates of likelihood (the opposite may actually occur); nor (2) to a great variability among loci that would have been best modeled by the hierarchical model of Storz and Beaumont [2002] because a leave one out inference procedure with our method did not find any loci strongly differing from the others (results not shown).

TABLE 3.3: ***Accuracy of the MLE with SIS and SISR algorithms on the bat data set***

(α, β, k)	$\hat{\theta}$	$\hat{D}$	$\widehat{\theta_{\text{anc}}}$	$\widehat{N_{\text{ratio}}}$	$\ln(\hat{L}(\hat{\phi}))$	lik-RMSE
SIS	1.08 $[0.76 - 1.5]$	0.44 $[0.35 - 0.55]$	320 $[180 - 560]$	0.0034 $[0.0018 - 0.0062]$	-663.74	9.2
SISR						
$(1, 0, 50)$	0.20 $[0.085 - 0.4]$	0.54 $[0.46 - 0.62]$	490 $[270 - 490]$	0.00041 $[0.00017 - 0.00098]$	-586.81	6.7
$(1, 0.0005, 50)$	0.18 $[0.073 - 0.37]$	0.53 $[0.45 - 0.63]$	380 $[240 - 500]$	0.00047 $[0.00018 - 0.0011]$	-587.38	5.8
$(1, 0.001, 50)$	0.21 $[0.091 - 0.39]$	0.53 $[0.45 - 0.62]$	390 $[250 - 490]$	0.00053 $[0.00022 - 0.0011]$	-586.78	5.3
$(0.7, 0, 50)$	0.19 $[0.075 - 0.40]$	0.53 $[0.45 - 0.63]$	410 $[230 - 510]$	0.00046 $[0.00016 - 0.0012]$	-588.25	6.5
$(0.7, 0.0005, 50)$	0.20 $[0.084 - 0.43]$	0.54 $[0.45 - 0.63]$	450 $[250 - 490]$	0.00046 $[0.00018 - 0.0011]$	-588.11	6.2
$(0.7, 0.001, 50)$	0.22 $[0.093 - 0.43]$	0.53 $[0.45 - 0.63]$	440 $[272 - 581]$	0.00050 $[0.00019 - 0.0011]$	-588.25	8.0
$(0.5, 0, 50)$	0.21 $[0.088 - 0.40]$	0.53 $[0.44 - 0.61]$	400 $[240 - NA]$	0.00053 $[0.00020 - 0.0012]$	-588.43	6.4
$(0.5, 0.0005, 50)$	0.17 $[0.066 - 0.38]$	0.52 $[0.44 - 0.61]$	360 $[230 - 490]$	0.00048 $[0.00017 - 0.0012]$	-589.07	5.0
$(0.5, 0.001, 50)$	0.21 $[0.091 - 0.40]$	0.54 $[0.46 - 0.61]$	490 $[330 - 490]$	0.00042 $[0.00018 - 0.00089]$	-588.99	7.4

3.7 Limits and perspectives

A first limitation of our method concerns the size of the data sets to be analyzed. Here, we considered a sample size of 100 haploid individuals genotyped at 10 independent loci for all our simulations, and about 250 diploid individuals genotyped at 8 loci for the fruit bat data set. Under similar demographic models of past changes in population size, Leblois et al. [2014] considered up to 500 haploid individuals genotyped at 50 loci. Data sets of such size can be analyzed on a desktop computer within a day or two depending on the number of histories explored. Larger sample sizes are feasible but then the analyses must be parallelized over large number of cores, e.g. 10 or more, on a computer grid for processing a single data set in a single day. Note however that simulation tests considering hundreds of simulated data sets for a single scenario, as done in this chapter, needs much more computation time and may thus be quickly intractable for very large data sets.

A second important limitation concerns the type of genetic markers that can be analyzed. All tests presented in this study considered allelic data analyzed under a strict stepwise mutation model (SMM, typical for microsatellite makers) but different mutation models have been implemented and tested in our Migraine software (e.g. the parent independent mutation model, PIM, and the generalized stepwise mutation model, GSM, for allelic data; and the infinite site model, ISM, for DNA sequences). The resampling technique developed here can thus easily be extended to all those models, notably allowing the analysis of short non-recombining DNA sequences in addition to microsatellite data. Interestingly, under the ISM without recombination, the topology

of trees that are compatible with the observed data is almost fixed, which strongly reduces the space of histories to explore. Thus, under simple demographic models such as a single population of constant size or with a single continuous past change in population size, and for sufficiently small data sets, dynamic programming can be used to sum over all ancestral histories and thus the exact likelihood can be computed as in Faisal et al. [2015]. Under those conditions, the exact likelihood calculation will lead to more accurate (and sometimes faster) estimates than approximate likelihood approaches such as importance sampling algorithms. However, the work of Faisal et al. [2015] also showed that, when increasing the scaled mutation rate and sample size, the number of histories to be considered severely increases and the algorithm rapidly becomes too slow for practical inferences. For example, for $\theta = 6$ and 10 sampled haploid individuals sequenced at 10 independent loci, the maximum likelihood estimates of θ is calculated in 254 seconds but the same calculation takes 6967 seconds when considering only 20 sampled haploid individuals, showing a non linear increase with sample size. Moreover, when the mutation model allows for reversible mutations on a finite number of sites or alleles, as for the SMM considered in this study, the number of possible histories explodes to the extent that the exact likelihood computation is no longer possible and only approximate algorithms such as importance sampling allow the inference.

Moreover, and more importantly, our method can only consider non-recombining independent loci. Given the current explosion of new sequencing techniques that provide longer and longer DNA sequences, analyzing DNA sequences without a model for recombination is an important limit. However, there is no method based on the class of importance sampling algorithms used here that can so far handle realistic recombination models (e.g. multiple recombination points within DNA sequences). Existing IS algorithms dealing with recombination only consider a single recombination point, i.e. recombination between independent pairs of non-recombining loci (e.g. Jenkins and Griffiths, 2011), and considering more recombination points seems computationally infeasible for the moment. Finally, given the burst of single nucleotide polymorphism (SNP) data, it would be interesting to develop IS algorithms specifically adapted to SNPs, but we are not aware of any development, application or test of IS algorithms for SNP data.

A third limitation concerns the maximum number of model parameters that our method can deal with, i.e. the level of complexity of the demographic and mutational models used. This limit is mostly due to (1) the way we extrapolate the likelihood surface on the explored parameter space using a kriging technique; and (2) the computation of profile likelihoods used for confidence interval calculations and plots of likelihood profile surfaces. Here, we considered models with three parameters, but more complex models can be analyzed. Simulated and real data sets have been routinely analyzed under the same demographic model as the one used here but with an additional parameter for the GSM mutation model, thus with four parameters (e.g. Leblois et al., 2014; Vignaud et al., 2014a,b; Lalis et al., 2016; Zenboudji et al., 2016). More recently, models considering two past population size changes and a generalized stepwise mutation model, with up to five parameters, have been treated successfully (e.g. Rousset et al., submitted). However, models with more than eight parameters would clearly be difficult to handle, due in particular to limitations in some libraries on which our R code depends, as well as to a general increase in time of all other explorations of

the likelihood surface for profile likelihood computations.

3.8 Conclusions

In this study, we proposed to improve sequential importance sampling algorithms for likelihood inference of demographic parameters by adding a resampling procedure based on Liu et al. [2001]. The new resampling probability distribution we considered depends on the SIS weights, as proposed in Liu et al. [2001] and Liu [2008], but also on the pairwise composite likelihood of the sample, providing additional information about the future trend of each sampled history. To evaluate the gain in efficiency due to this new strategy, we focused on varying population size models for which no efficient proposal distribution is available.

We first showed using simulations that resampling allows to reduce the variance and the bias in the likelihood estimate at a parameter point. This step allowed us to show that resampling is more efficient when the checkpoints correspond to a number of coalescent events undergone by each history rather than considering coalescent and mutations events together, and that checkpoints must be frequent, more precisely after each coalescence event. This step also showed that the information provided by the composite likelihood allows a stronger decrease in variance of the likelihood estimates than the information provided only by the SIS weights.

Then we showed that the increased precision in estimation of likelihood also improved the likelihood-based inferences. According to numerical results based on the analysis of simulated data sets under various scenarios of past population size contraction with different strength and timing, we conclude that the resampling procedure helps to correct the inefficiency of the SIS proposal distribution. Under those time-inhomogeneous models, the stronger the population size contraction is, the less efficient the proposal distribution is because it is computed under equilibrium hypothesis. For a similar precision of inference, SISR provides at least a two-fold, and sometimes a much higher gain in computational efficiency for these models. However, when the contraction is very recent, the genetic data does not contain much information about the parameter θ, resulting in flat likelihood surfaces. Consequently, our simulation results show that (i) when the contraction of the population size is not very recent, the resampling procedure divides the computation cost by at least a factor 2; (ii) when the contraction is not very strong but very recent, it is not obvious how the resampling improves the inference because we face an important lack of information for the θ parameter; and finally (iii) in extreme cases of both very recent and very strong contractions, the resampling procedure divides the computation cost by a factor 100, partially correcting the inefficiency of the SIS proposal distribution but not the lack of information in the data, and thus leads to potential bias in θ estimates and associated incorrect CI coverage properties.

Finally, we analyzed a fruit bat microsatellite data set previously analyzed by Storz and Beaumont [2002] using coalescent-based MCMC algorithms and found that the SISR method allows to infer the past demographic history of this population with computation times reduced by at least a factor one hundred compared to SIS. Our results

also shows a slightly more pronounced past contraction of population size, with larger ancestral size, than previously found.

3.9 Appendix

3.9.1 Pairwise Composite Likelihood

In this work, we propose a new resampling probability distribution $\mathbf{v}$ on the collection of histories. To achieve efficiency we should pick a distribution $\mathbf{v}$ that reflects the future trend of the partial histories, more precisely that sets relatively high probabilities on the partial histories that will be the most likely at the end of the simulation. Since the information is mainly in the dependency between individuals of the sample, we propose to substitute the Pairwise Composite Likelihood (PCL), noted $L_2(\mathbf{h})$, for the likelihood of the end of the history $\mathbf{h}$. Indeed, it is the product of the likelihoods of each pair of individuals remaining in the sample:

$$L_2(\mathbf{h};\phi) = \prod_{A\in E:\,\mathbf{h}(A)\geq 2} L_2((A,A);\phi)^{\mathbf{h}(A)(\mathbf{h}(A)-1)/2} \times \prod_{A<B\in E:\,\mathbf{h}(A)\geq 1,\,\mathbf{h}(B)\geq 1} L_2((A,B);\phi)^{\mathbf{h}(A)\mathbf{h}(B)},$$

where $L_2(x,y|\phi)$ is the likelihood of a sample of two alleles in the stepwise mutational model, known explicitly when the scaled population size is constant equal to θ_{anc}:

$$\forall x,y\in E,\ L_2((x,y);\phi) = \frac{1}{\sqrt{1+2\theta_{\text{anc}}}}\rho(\theta_{\text{anc}})^{|y-x|} \text{ where: } \rho(\theta_{\text{anc}}) = \frac{\theta_{\text{anc}}}{1+\theta_{\text{anc}}+\sqrt{1+2\theta_{\text{anc}}}}.$$

The PCL does not approximate the likelihood but it behaves the same way. In particular, it reflects the behavior of the remaining lineages in the history $\mathbf{h}$ when the population size is equal to the ancestral size N_{anc}. Thus we chose to compute the PCL as an expression depending on ϕ only through θ_{anc}. A judicious idea could be to derive an expression of the PCL with $\theta(t)$ depending on t. However, we are concerned that its calculation would become too complex with respect to the expected gain.

3.9.2 Simulating holding times of H and $\widetilde{\mathbf{H}}$

The importance sampling scheme assumes that we are always able to draw simulations from the holding time distribution. When the process $\widetilde{\mathbf{H}}$ has just jumped to state $\mathbf{h}\in\mathbb{N}^E$ at time t, the distribution of the holding time Δ has the following density

$$\mathbb{P}\big(\Delta\in(\delta;\delta+d\delta)\big|T_i=t,\mathbf{X}_i=\mathbf{h}\big) = \lambda_{t+\delta}(\mathbf{h})\exp\left(-\int_0^\delta \lambda_{t+u}(\mathbf{h})du\right)d\delta$$

and the following probability distribution function

$$\mathbb{P}\big(\Delta\leq\delta\big|T_i=t,\mathbf{X}_i=\mathbf{h}\big) = 1-\exp\left(-\int_0^\delta \lambda_{t+u}(\mathbf{h})du\right).$$

The intensity $\lambda_t(\mathbf{h})$ is defined by Eq. (3.5) and depends on the parametric model we set on the population size $N(t)$. When the population size is constant, $\lambda_t(\mathbf{h})$ does not depend on t and this distribution boils down to the simple exponential distribution with rate $\lambda(\mathbf{h})$. When we face an exponentially changing population size, the above

probability distribution function can be computed explicitly, see Griffiths and Tavaré [1994b], and we can rely on the inverse of the probability distribution function to draw simulations.

Algorithm 4: Simulation of the holding time

```
Initialization: set t' = t;
while u > λ_{t'}(h)/M do
    Compute some M greater or equal to sup_{s≥t'} λ_s(h);
    Draw δ_0 from the exponential distribution with rate M and Set t' = t' + δ_0;
    Draw u uniformly in [0, 1);
Return t' − t
```

In the general case where $N(t)$ can be any parametric function of t, Algorithm 4 might be interesting. It follows the well-known algorithm of Gillespie [1977] which aims at simulating pure jump Markov processes. There is many ways to show that Algorithm 4 is correct. The simplest one is to claim that the holding time is exactly the first point after t of a Poisson point process [Kingman, 1992] with intensity $\lambda_s(\mathbf{h})$ at any $s \geq t$. And such Poisson point processes can be simulated from a first Poisson point process with intensity M, constant over time, bounding the intensity $\lambda_s(\mathbf{h})$ for any $s \geq t$. To this end, we simply takes any point T' from the first process with probability $\lambda_{T'}(\mathbf{h})/M$. Thus, Algorithm 4 simulates the first point t' of Poisson point process with intensity M at step 2.(b) and rejects it until an event of probability $\lambda_{t'}(\mathbf{h})/M$ occurs, namely that $u \leq \lambda_{t'}(\mathbf{h})/M$. Finally, note that we can easily bound from above the intensity $\lambda_s(\mathbf{h})$ at any $s \geq t$ by bounding from below the population size after time t, see Eq. (3.5).

3.9.3 One-parameter profile likelihood ratios

The likelihood function depends here on three parameter $(\theta, D, \theta_{\text{anc}})$. For each fixed value of the parameter θ, we rewrite the likelihood estimate surface as $\hat{L}_\theta(D, \theta_{\text{anc}}) = \hat{L}(\theta, D, \theta_{\text{anc}})$ that is θ is fixed and D and θ_{anc} vary. We estimate D and θ_{anc} by maximizing $\hat{L}_\theta(D, \theta_{\text{anc}})$ with respect to D and θ_{anc}, *i.e.*

$$(\hat{D}_\theta, \hat{\theta}_{\text{anc}\,\theta}) = \text{argmax}_{D,\theta_{\text{anc}}} \hat{L}_\theta(D, \theta_{\text{anc}}).$$

As θ is unknown, we evaluate $(\hat{D}_\theta, \hat{\theta}_{\text{anc}\,\theta})$ for each θ. Then we estimate θ by maximizing $\hat{L}_\theta(\hat{D}_\theta, \hat{\theta}_{\text{anc},\theta})$:

$$\hat{\theta} = \text{argmax}_\theta \hat{L}(\theta, \hat{D}_\theta, \hat{\theta}_{\text{anc},\theta}).$$

We have profiled out the parameters $(\hat{D}, \hat{\theta}_{\text{anc}})$ and the likelihood profile $\hat{L}_\theta(\hat{D}_\theta, \hat{\theta}_{\text{anc},\theta})$ is completely in terms of the parameter θ. Then, we represent the likelihood profile function divided by its maximum value:

$$\theta \mapsto \hat{L}_\theta(\hat{D}_\theta, \hat{\theta}_{\text{anc}\,\theta}) / \hat{L}(\hat{\theta}, \hat{D}_\theta, \hat{\theta}_{\text{anc},\theta}),$$

We also represent in the same way the profile likelihood ratio for D and the profile likelihood ratio for θ_{anc}, estimated here with an IS algorithm.

3.9.4 Figures

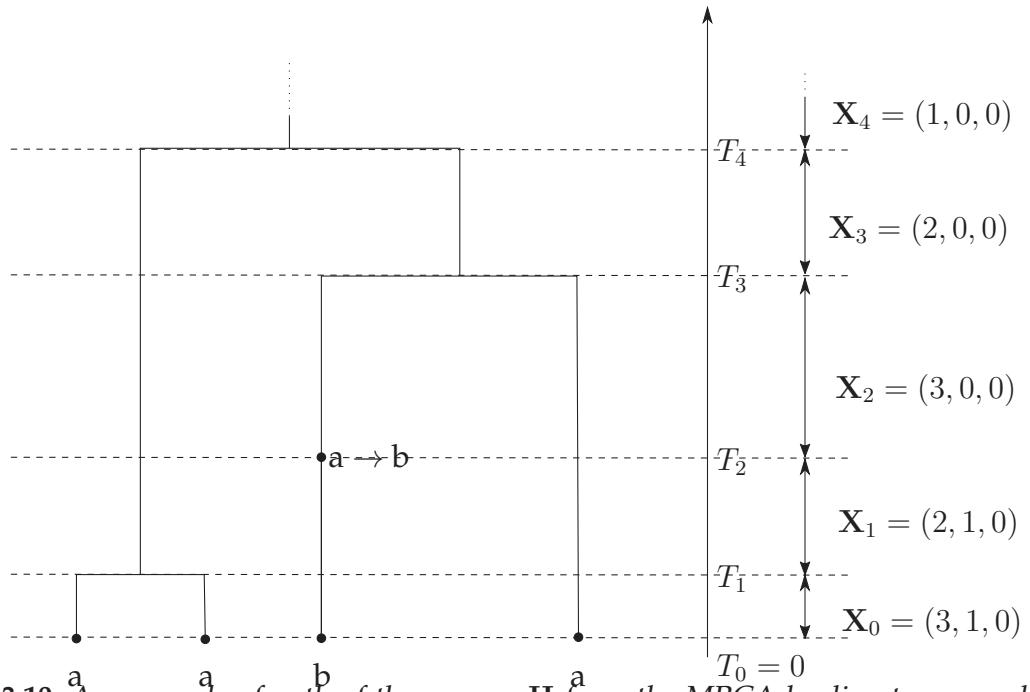


FIGURE 3.10: *An example of path of the process* $\mathbf{H}$ *from the MRCA leading to a sample of 4 genes at time* 0*, when the set of gene types* $E = \{a, b, c\}$ *is composed of three possible alleles.*

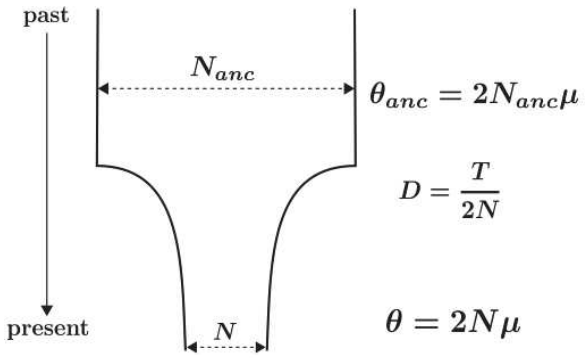


FIGURE 3.11: ***Demographic model.*** *Representation of the exponentially contracting population size model used in the study.* N *is the current population size,* N_{anc} *is the ancestral population size (before the demographic change),* T *is the time measured in generation since present, and* μ *is the mutation rate of the marker used. Those four parameters are the canonical parameters of the model.* θ*,* D*, and* θ_{anc} *are the inferred scaled parameters (Leblois et al., 2014).*

4 Detecting past changes in population size using haplotype homozygosity

Contents

Abstract The recent development of high-throughput sequencing technologies has revolutionized the generation of genetic data for many organisms: genome wide sequence data are now available. Classical inference methods (maximum likelihood methods (MCMC, IS), methods based on the Sites Frequency Spectrum (SFS)) suitable for polymorphism data sets of some loci assume that the genealogies of the loci are independent. To take advantage of genome wide sequence data, we need to consider the dependency of genealogies of adjacent positions in the genome. Thus, when we model recombination, the likelihood takes the form of an integral over all possible ancestral recombination graphs for the sampled sequences. Then the likelihood-based inference then require further simplifications. Several methods modeling recombination, based either on identical by state or identical by descent segment lengths, already exist to infer historical changes in the effective population size and do not require the likelihood computation. Even if some of them propose a control for potential overfitting, they do not consider the complexity of the demographic model fitted. To the best of our knowledge, no model choice procedure between demographic models of different complexity have been proposed based on identical by state segment lengths. The aim of the present work is to overcome this lack by proposing a model choice procedure between demographic models of different complexity. We focus on a simple model of constant population size and a slightly more complex model with a single past change in the population size. Since these models are embedded, we developed a penalized model choice criterion based on the comparison of observed and predicted haplotype homozygosity. Our penalization relies on the computation of Sobol's sensitivity indices and is a form of penalty related to the complexity of the model. This penalized

model choice criterion allowed us to choose between a population of constant size and a population size with a past change on simulated data sets and also on a cattle data set.

4.1 Introduction

The demographic history of a population affects the present-day genetic polymorphism. Understanding the mechanisms of how factors, such as random drift and mutation, drive the evolutionary process is a central task for population geneticists. The latter address these questions quantitatively by constructing mathematical models, developing statistical methods for inferring ancestral processes and testing hypotheses based on the analysis of real data. With the recent development of next generation sequencing (NGS) technology, genome wide sequence data are now available for individuals from different organisms (bacteria, human, animals, ...) (Mardis [2008]). The knowledge of a population's history allows to examine for example the factors driving past population dynamics (Finlay et al. [2007], Atkinson et al. [2008], Stiller et al. [2010]) and trace the transmission and spread of viruses (Kitchen et al. [2008], Magiorkinis et al. [2009]).

One main barrier to apply existing population genetic methods to large genomic data is intensive computation. The likelihood function of most population genetic methods present a latent variable: the gene genealogy. These methods evaluate the likelihood function by adopting Markov Chain Monte Carlo (MCMC) or importance sampling (IS) to integrate over the gene genealogies space. It is computationally very intensive, and only works for a small sample of haplotypes (Griffiths and Tavare [1994]). Such methods cannot be directly scaled up to large-sample genomic data even with high performance computers. Developing efficient computing algorithms is necessary. A range of methods are available for estimating demographic patterns using nucleotide sequence data (see Emerson et al. [2001] and the review of Chen [2015] on recent methodological developments in the field of population genetics). Here we present the most widely used.

Skyline-plot methods constitute a first class of non parametric and semi-parametric methods to infer demographic history from sequence data (see the technical review of Ho and Shapiro [2011]). The skyline-plot family methods, methodological extensions of the classical skyline-plot introduced by Pybus et al. [2000] and the generalized skyline-plot of Strimmer and Pybus [2001], allow to estimate the historical population size patterns without the need for possible demographic models. They are based on the coalescent theory which quantifies the relationship between the genealogy of the sequences and the demographic history of the population (see Kingman [1982a], Kingman [1982b], Hein et al. [2004] and Wakeley [2005] for details). The skyline-plot framework assumes that the genealogy of the sequences is obtained independently and known without error. Their application to real data has shown the potential of these methods for elucidating complex patterns of the demographic history. However they are subject to a number of significant limitations. Notably, they assume non-recombining loci and thus cannot deal with genome wide sequence data but only with

a few tens of independent loci.

A second class of inference methods is based on another aspect of the genomic polymorphism: the Allele Frequency Spectrum (AFS, alternatively Site Frequency Spectrum, SFS). The AFS is a sampling distribution of alleles in a finite sample (Chen [2012]) and focuses on the allele frequency distribution of a single locus, ignoring the correlation among nearby loci. AFS theory was developed in two parallel frameworks: the diffusion (Kimura [1955]) and coalescent processes (Fu [1995]). Parametric or nonparametric models of the demographic history can be considered (Polanski and Kimmel [2003], Marth et al. [2004], Evans et al. [2007], Gutenkunst et al. [2009] and Excoffier et al. [2013]). These methods can manage a large number of loci, assumed independent and hence do not use the information provided by recombination.

Azzou et al. [2015] introduced an importance sampling approach, the skywis plot, for estimating the demographic history of a sample of DNA sequences with no recombination further extended in Azzou et al. [2016]. More precisely, they proposed a nonparametric maximum likelihood estimate of a population size that changes over time. They approximate the effective population size by a piecewise constant function.

Finally, we present methods based on haplotype structure, which consider the linkage disequilibrium (LD), or the correlation of gene genealogies of adjacent sites. They take advantage of the recombination information present in genome wide sequence data. Thus the likelihood of the data set is no longer the product of the likelihoods for each independent locus. While mutation does not affect the tree structure of the genealogy of the sample, recombination does. When we model recombination, the likelihood takes the form of an integral over all possible ancestral recombination graphs for the sampled sequences (Griffiths and Tavaré [1994a]). This space is of much larger dimension than the genealogies space, to the extent that we cannot handle likelihood-based inference while modeling recombination without further simplifications. Two main approaches have been developed, on the one hand methods based on the approximation of the coalescent with recombination: the sequential coalescent with Hidden Markov Models (PSMC, coalHMM), and on the other methods based on identical segment lengths between two haplotypes.

The Pairwise Sequential Markovian Coalescent method (PSMC, Li and Durbin [2011]) uses a hidden Markov model to approximate the dependency of the coalescent times of two haplotypes between adjacent loci. It has been further extended by Mailund et al. [2011] to allow for two sequences from two population. They infer the detailed ancient population size from coalescent times. Burgess and Yang [2008] developed a Markov chain Monte Carlo approach (MCMCCoal) to sample gene genealogies. They inferred ancient population size by analyzing multiple sequences, with each sequence representing one population. Their method has been further extended by Gronau et al. [2011] to allow for two sequences from each sampled population. Schiffels and Durbin [2014] extended the PSMC to the Multiple Sequential Markovian Coalescent (MSMC), allowing to analyze multiple diploid genomes, by focusing only on some summary statistics of the genealogies, such as first coalescent time of any two sequences and total length of all singleton branches of the genealogy. The Coal-HMM method Hobolth et al. [2007] can also analyze multiple genomes from several populations. Instead of sampling over gene genealogies, Coal-HMM treats the unobserved gene genealogy at each genomic

position as latent states in a hidden Markov model. The possible number of gene genealogies increases dramatically when multiple sequences are included, and computation again becomes intensive for the above methods. Boitard et al. [2016] introduced an approximate Bayesian computation approach, named PopSizeABC, that allows to estimate the evolution of the effective population size through time. In this method, observed genomes are summarized using a small number of statistics related to allele frequencies and linkage disequilibrium. They assume that the considered population has evolved forever as an isolated population, as well as other SMC presented before (Li and Durbin [2011], Schiffels and Durbin [2014]) or IBS-based methods (MacLeod et al. [2013]) presented in the following paragraph.

The alternative approach with recombination is based on the distribution of identical segment lengths. Two DNA segments are Identical By State (IBS) if they have identical nucleotide sequences in this segment whereas two DNA segments are Identical By Descent (IBD) if they are inherited from a common ancestor without recombination nor mutation. As detailed below, short segments are mostly affected by the ancient past and long segments mostly by the recent past. Since the detection of short IBD segments is difficult, the distribution of IBD segment lengths is rather used to infer recent past demography. Palamara et al. [2012] expressed the distribution of IBD segment lengths across pairs of individuals as a function of the population's demography and derived an inference procedure to reconstruct such demographic history. They tested increasingly flexible parametric models to infer the demographic history. In order to control for potential overfitting, they evaluated the parameters obtained for different models by using a likelihood approach, using the Akaike Information Criterion (AIC, Akaike [1974]) to compare models while controlling for their respective degrees of freedom. Browning and Browning [2015] in turn presented a nonparametric method for accurately estimating recent effective population size by using inferred long segments of IBD. They proposed a generalized expectation-maximization (EM) procedure for fitting the trajectory of the historical population size. Ringbauer et al. [2016] recently used a diffusion approximation to trace genetic ancestry back in time, and derived analytical formulas for patterns of isolation by distance of long IBD-blocks, which can also incorporate recent population density changes. Their inference scheme uses a composite likelihood approach to fit observed block sharing to these formulas.

Although we cannot observe DNA segment IBD status, we can observe whether or not two haplotypes contain identical sites along a given segment. That is, they are observed to be identical by state and these segments of conserved segment lengths may occur through recombination. First Hayes et al. [2003] and MacLeod et al. [2009] introduced explicit formulas for the probability for a haplotype pair to share at least a given number of adjacent positions being IBS. Modeling the population size as a constant piecewise function, their derivation accounts for recombination and mutation. It is based on approximate coalescence theory, assuming that only one recombination may occur per segment within one generation. However computations of these coalescent based formulas are time consuming. These tools have been further developed by MacLeod et al. [2013] to infer ancestral population history and applied to whole-genome sequence data. They used their theoretical HH_{th} to determine population parameters that best match the empirical $\widehat{HH}$ of a given Holstein cattle sequence. To validate the inferred demographic model, they simulated sequence data under the demographic

model inferred on one animal and compared these data to the observed sequence in a second animal, then switch roles. Harris and Nielsen [2013] presented a method for using sequence data to jointly estimate the timing and magnitude of past admixture events, along with population divergence times and changes in effective population size. They inferred a piecewise constant effective population size from a collection of pairwise sequence alignments by summarizing their length distribution of tracts of IBS and maximizing an analytic composite likelihood derived from a Markovian coalescent approximation. In contrast, these methods skip detection of IBD segments and instead work directly with IBS haplotypes. An advantage of that approach is that one can examine shorter segments and hence look further back into the past. The present work aims to detect past changes in the population size by considering the LD and relying on IBS segment lengths between two haplotypes.

A variety of methods exist to measure pairwise LD based on genotype frequencies at two loci [Zhao et al., 2007]. However, it has been pointed out that these measures are very diverse and likely not as informative for inferring population history compared to using data from multiple markers along a segment [Nordborg and Tavaré, 2002]. We chose to quantify the LD with a multilocus haplotype homozygosity, denoted HH. Given a pairwise comparison of two haplotypes, HH is the probability for a given number of adjacent positions drawn at random in the genome to be homozygotes between the two haplotypes.

The haplotype homozygosity pattern is affected by the demography, notably by the changes of the effective population sizes over time as shown of Figure 4.1. Intuitively, the larger the effective population size, the lower the haplotype homozygosity. Moreover, the theory predicts that LD over shorter distances reflects parameters of more ancestral population than LD at larger distances. LD on a segment size of c Morgans is mostly affected by the population size approximately $1/2c$ generations in the past, assuming a linearly changing population size [Hayes et al., 2003]. Early studies on HH showed with simulations that estimates of segment homozygosity for a wide range of chromosome segment lengths can be used to estimate the effective population size at multiple times in the past [Hayes et al., 2003].

The previous methods focus mainly on the inference of the historical changes in the effective population size. In general, they do not consider the complexity of the demographic model fitted and may suffer from an overfitting problem: in many cases, a much simpler demographic model could record the key events of the considered population history. Even if some of them propose a control for potential overfitting, to the best of our knowledge, no model choice procedure between demographic models of different complexity have been proposed based on the IBS segment lengths. Our aim is to overcome this lack by proposing a model choice procedure between demographic models of different complexity. In the present work, we focus on a simple model of constant population size and a slightly more complex model with a single past change in the population size. Since these models are nested and to avoid choosing always the more complex model, we developed a penalized model choice criterion. It is based on the comparison of observed $\widehat{HH}$ and predicted HH_{th} haplotype homozygosity. Our penalization relies on the computation of Sobol's sensitivity indices. The idea is to take into account only the part of the error that the model had the ability to explain. It is a form of penalty related to the complexity of the model since a given adjustment

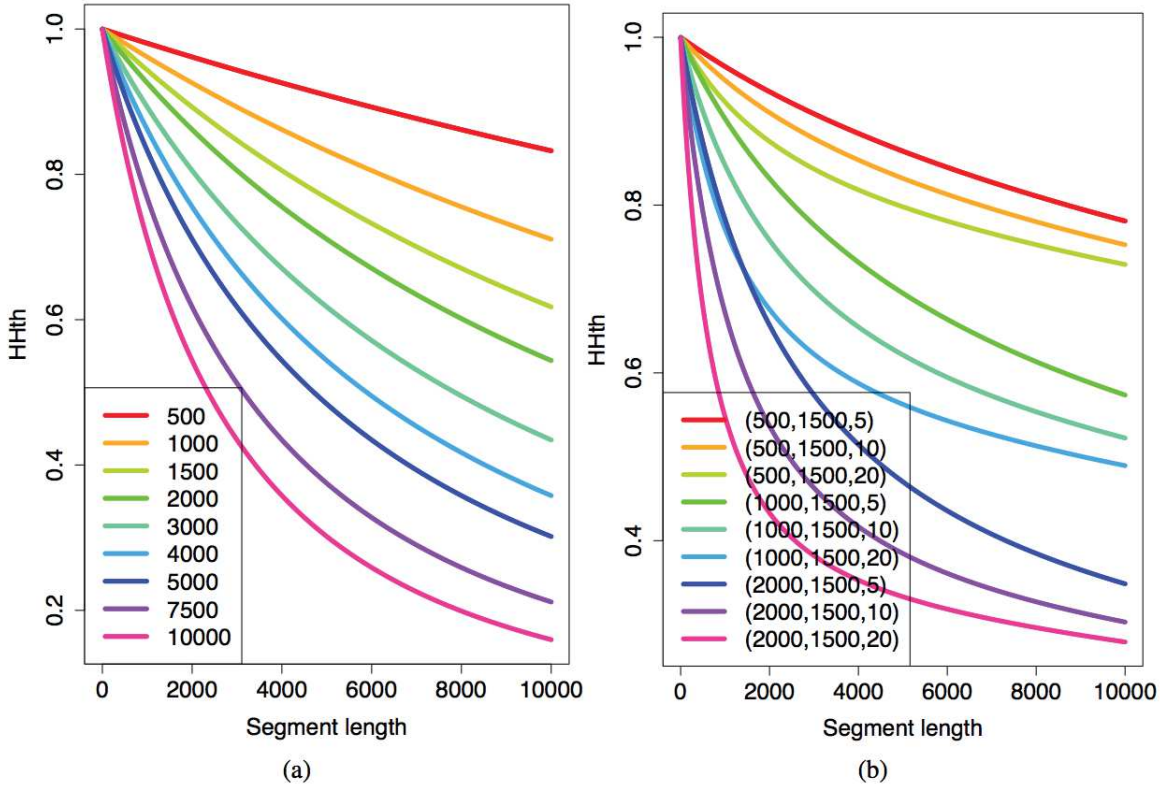


FIGURE 4.1: *HH_{th} corresponding to (a) constant and (b) contracting population size models with different values of the parameter (a) $\theta_0 = Ne_0$ and (b) $\theta_1 = (Ne_0, t, f)$.*

error (between observed $\widehat{HH}$ and theoretical HH_{th}) will penalize more a complex model than a simpler one. Indeed, one of the major challenges of the sensitivity analysis is to identify the less contributing parameters to reduce the dimension of the model and quantify the committed error (Sobol et al. [2007], Saltelli et al. [2010]). For this purpose, the sensitivity analysis decomposes the output variance of the model in fractions attributed to each input parameter through Sobol's sensitivity indices. In our case, considering the theoretical haplotype homozygosity as a function of the demographic parameters, assumed to be random variables for the computation, these indices measure the part of the variance of the theoretical HH_{th} explained by each parameter of the more complex model considered.

4.2 Demographic inference

As we mentioned in the introduction and illustrated on Figure 4.1, the HH curve characterizes the demographic history. Two different demographic histories produce two different HH curves, that is we can distinguish two different demographic histories by comparing there respective HH curves.

The demographic model of the population history consists of stepwise changes in the effective population size. Each phase of constant population size is constituted of several generations. We aim to be able to determine the number of significant past

changes in the population size. A first stage is to discriminate between a constant population size and a single past change in the population size. In the present work, we chose to focus on this simple case. Consequently, the remainder of the present work focuses on the model of a constant population size $\theta_0 = Ne_0$, denoted $\mathcal{M}_0$ and on the model denoted $\mathcal{M}_1$ of a unique past change in the population size parametrized by $\theta_1 = (Ne_0, t, f)$, where Ne_0 is the present effective population size, t the date of the change in population size and f the intensity of the change such that $f \times Ne_0$ is the ancestral population size before t generations before present. Both models are represented in Figure 2.5. The parameter estimation is not the objective of this work. It is a step towards the model choice.

4.2.1 Empirical $\widehat{HH}$ and theoretical HH_{th}

To infer information about population parameters from multiple marker haplotype homozygosity data, we need a model to predict HH from population demographic history.

MacLeod et al. [2009] introduced explicit formulas of the probability HH_{th} for a haplotypes pair to share a given number of adjacent markers being identical by state. Their derivation is based on simplified coalescence theory and accounts for recombination and mutation, under assumptions of a unique diploid panmictic population with no selection, no migration. The mutation rate and recombination rate are assumed to be constant along the genome. In this work, we chose to predict the HH of a given demographic history with MacLeod et al. [2009]'s HH_{th}.

In the following, we note n_{pos} the number of adjacent positions on considered haplotypes and $n \mapsto HH_{\text{th}}(\theta, n)$ the theoretical haplotype homozygosity of $n \in \{1, \ldots, n_{\text{pos}}\}$ adjacent positions, for a population history parametrized by θ. To compute $HH_{\text{th}}(n)$ for a position number n, MacLeod et al. [2009] assume first a phase of constant population size equal to N. This phase is split in two periods:

1. No event for s generations with probability

$$\alpha^s = \left[(1 - \frac{1}{2N})(1-c)^2(1-\mu)^{2n}\right]^s,$$

   where μ is the mutation rate and c is the recombinant distance of the segment.

2. At generation $s+1$, an event of coalescence or recombination takes place. The probability of a coalescence event at generation s is: $\beta = (\frac{1}{2N})(1-c)^2(1-\mu)^{2n}$. The probability τ of a recombination event in the segment of length $n \in \{1, \ldots, n_{\text{pos}}\}$ is the sum or the probabilities of recombination between each pair of adjacent positions k and $k+1$ and include the joint probability that all positions in both recombined segments will then coalesce or recombine without mutation:

$$\tau \approx \sum_{k=1}^{n-1} \left[(1-(1-r_{k,k+1})^2)(1-\mu)^{2n}\left(1-\frac{1}{2N}\right) HH_{\text{th},1 \text{ to } k}\ HH_{\text{th},k+1 \text{ to } n}\right],$$

where $r_{k,k+1}$ is the recombinant distance between position k and $k+1$. MacLeod et al. [2009] ignored the possibility of more than one recombination per segment within one generation.

Combining the two previous periods, they obtain

$$HH_{\text{th}}(n,\theta) = \left[\sum_{s=0}^{T-1} \alpha^s\right](\beta+\tau).$$

To accommodate varying population size, they model the effective population size by a piecewise constant function and calculate the probabilities of no event followed by an event within each of the different constant population size phases.

The probability of no event occurring for any given number of T_1 generations followed by an event within phase 1 is as before:

$$HH_{\text{th,Phase 1}}(n,\theta) = \left[\sum_{s=0}^{T_1-1} \alpha_1^s\right](\beta_1+\tau_1) = \left(\frac{1-\alpha_1^{T_1}}{1-\alpha_1}\right)(\beta_1+\tau_1),$$

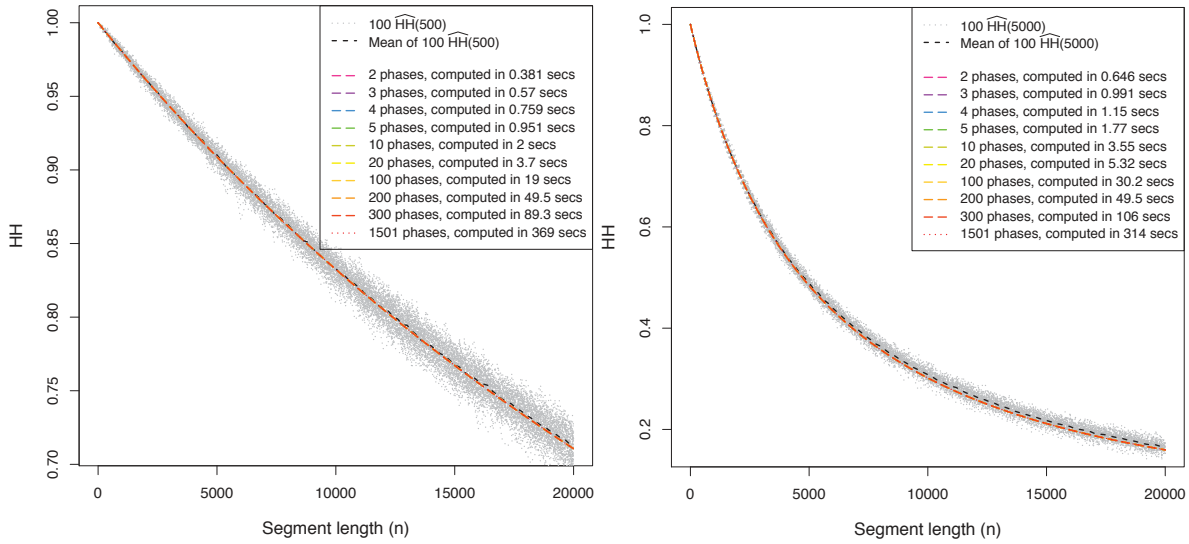
subscript 1 referring to phase 1, the most recent. For each pth phase going back in time, they compute the probability of no event for s generations and the probability that no event had occured in any of the more recent phases:

$$HH_{\text{th}}(n,\theta) = HH_{\text{th,Phase 1}}(n,\theta) + \sum_{p=2}^{\text{most ancestral Phase}} \left[\prod_{h=1}^{p-1} \alpha_h^{T_h}\right]\left(\frac{1-(\alpha_p)^{T_p}}{1-\alpha_p}\right)(\beta_p+\tau_p),$$

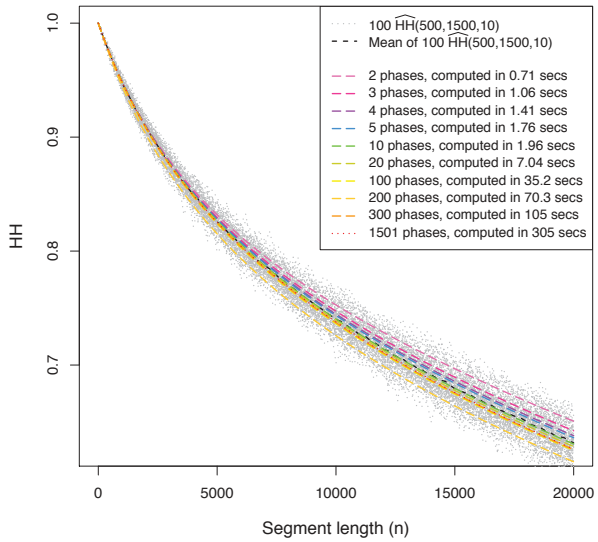
subscript p referring to phase number p, T_p being the total number of generations in phase p of constant population size equal to N_p and $\theta = (T_1, \ldots, T_{\text{most ancestral phase }-1}, N_1, \ldots, N_{\text{most ancestral phase }-1})$. The joint HH_{th} of the two recombined segments in τ_p, should be traced back in time from the generation in which recombination took place within a given phase.

To simplify this computationally, MacLeod et al. [2009] approximate by assuming recombination always takes place in the first generation of a given phase. When the population size is constant in time, this approximation do not lead to any error since the duration of the phase is not limited. In contrast, when the model presents at least one past change in the population size, that is two phases of different population sizes, the duration of the first phase is limited to a given number of generations and the time of the recombination event affects the probability HH_{th}. Any associated error can be minimized by splitting a long phase into a number of shorter phases. To avoid this approximation, each phase can be reduced to a single generation, although computing time may ultimately enforce practical limitations as we can see on Figure 4.2.

A second assumption has been mentioned before: this computation ignores the possibility of more than one recombination per segment within one generation. For short segments, that is for small values of n, this assumption has negligible consequences but for longer segments, that is when increasing the number of adjacent positions considered this assumption is more and more far from the reality. On a long segment, more than one recombination may occur in practice. The impact of this approximation is



(a) $Ne_0 = 500$ **(b)** $Ne_0 = 5000$

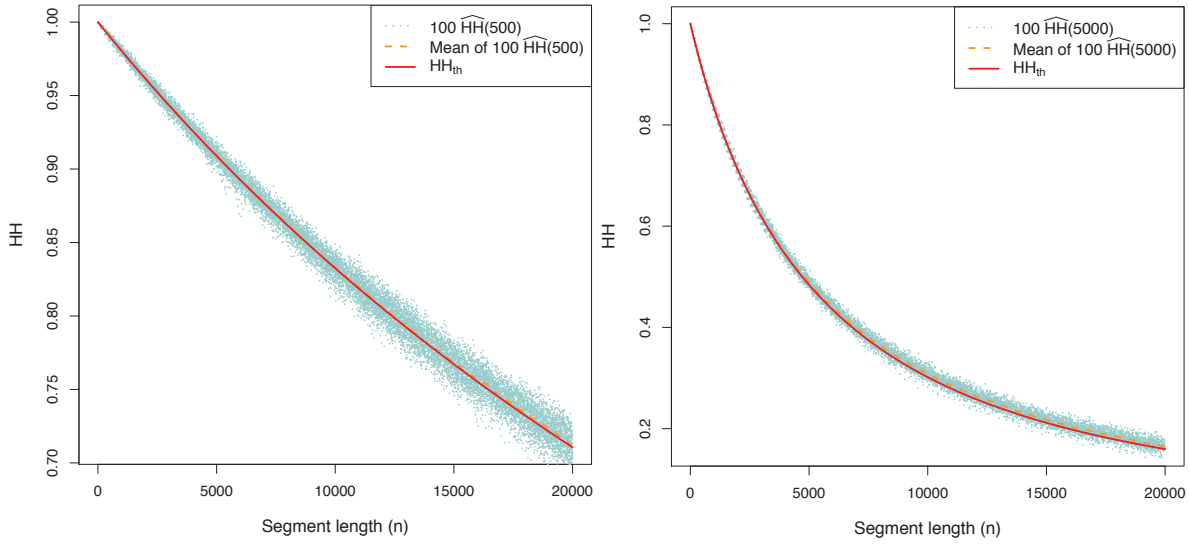


(c) $(Ne_0, t, f) = (500, 10, 1500)$

FIGURE 4.2: *Representation of* 100 $\widehat{HH}$ *curves (grey dotted lines) simulated under (a)* $Ne_0 = 500$*, (b)* $Ne_0 = 5000$ *and (c)* $(Ne_0, t, f) = (500, 10, 1500)$*, the mean of these* 100 *empirical curves (orange dashed line), the theoretical* HH_{th} *evaluated at the parameter of simulation (dotted red line) and different approximations of* HH_{th} *(dashed colored line) when splitting the recent constant population size phase into several sub-phases.*

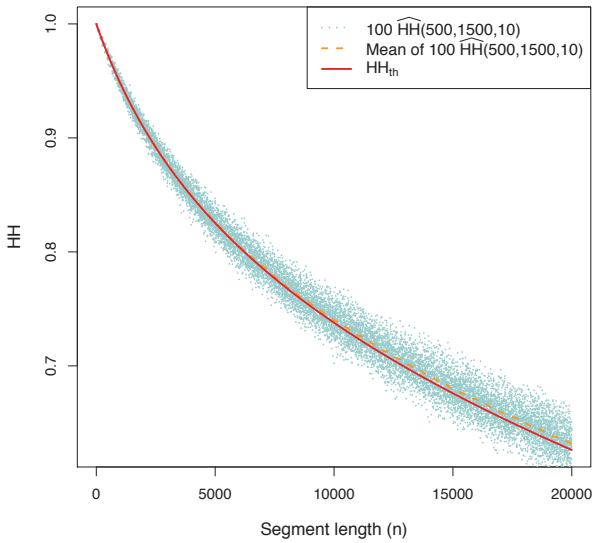
even greater when the population size is large since the probability of coalescence event is lower compared to the probability of recombination events. Figure 4.3 illustrate these statements: first for small values of n, the HH_{th} curve is centered among the 100 $\widehat{HH}$ computed on simulated data sets and close the observed mean curve of these 100 $\widehat{HH}$, whether $Ne_0 = 500$ or 5000. When n increases, the HH_{th} curve depart from the 100 $\widehat{HH}$ and from the observed mean curve, the more remote as the population size is large. This assumption will lead to a bias in the parameter estimate discussed in Section 4.4.

In an other hand, we need to compute the empirical HH, denoted $\widehat{HH}$, from the



(a) $Ne_0 = 500$

(b) $Ne_0 = 5000$



(c) $(Ne_0, t, f) = (500, 10, 1500)$

FIGURE 4.3: *Representation of* 100 $\widehat{HH}$ *curves (blue dotted lines) simulated under (a)* $Ne_0 = 500$*, (b)* $Ne_0 = 5000$ *and (c)* $(Ne_0, t, f) = (500, 10, 1500)$*, the mean of these* 100 *empirical curves (orange dashed line) and the theoretical* HH_{th} *evaluated at the parameter of simulation (solid red line).*

observed data. MacLeod et al. [2009] calculated $\widehat{HH}$ assuming equal genetic distance between markers genome wide:

$$\widehat{HH}(n) = \sum_{w=1}^{S} \mathbf{1}_{\mathcal{H}(w)}/S,$$

where n is a given number of markers on a homozygous haplotype for which $\widehat{HH}(n)$ probability is being calculated, w is every possible overlapping segment of n positions along the haplotype, numbered 1 to S and $\mathcal{H}(w)$ is the event: "All positions in window number w are homozygotes". Since we do not suppose that the positions right before and after the considered window are heterozygotes, $\widehat{HH}(n)$ is not the proportion of

segments of exactly n adjacent homozygotes markers but the probability for n adjacent positions drawn at random over the genome to be homozygotes. The computation of the segment proportion of at least n IBS adjacent positions corresponds to the theoretical $HH_{\text{th}}(n)$ proposed by MacLeod et al. [2009].

However, exploring every overlapping windows of n adjacent positions introduces a dependency between the considered windows. Indeed some of them differ just in one position. When translating the window by a single marker position, the probability of the new window of n adjacent positions to be homozygote is the probability that the last position is homozygote between the two haplotypes (knowing that the previous $n-1$ positions are identical). This complete scan computation of $\widehat{HH}$ leads to a high computation time. For example, the computation of $\widehat{HH}(n)$ for $n = 10000$ on a unique locus of 40000 base pairs takes account of 30000 overlapping segments.

Here, we propose a different way to compute $\widehat{HH}(n)$ for a given number n of adjacent positions to be considered. We want to explore a fewer number of windows while accessing to as much information on the recombination. Our idea is simply to spread uniformly the windows on the pairwise alignment of the two haplotypes. Idealy, we aim to explore the whole range of the haplotype but just once. We choose a number of windows $W(n)$ of n adjacent positions to explore. Then we draw uniformly at random in the genome a window of n adjacent markers and check if it contains a polymorphism between the two haplotypes or not. Finally, $\widehat{HH}(n)$ is the observed proportion of homozygotes segments among all randomly visited segments:

$$\widehat{HH}(n) = \sum_{w=1}^{W(n)} \mathbf{1}_{\mathcal{H}(w)}/W(n).$$

The number of windows $W(n)$ could vary with n for $n \in \{1, \ldots, n_{\text{pos}}\}$ but all the results presented in section 4.4 were obtained with a constant number of 5000 windows. It is much lower than the number of windows explored by the complete scan but leads to a comparable variability of the $\widehat{HH}$ curves since it explores well the whole genome. The random scan with 5000 windows per $\widehat{HH}$ curve save computation time compared to complete scan.

As we mentioned before, effective population sizes may change over time (contractions and expansions) affecting the observed homozygosity patterns. For example, we notice that a population having undergone a contraction in the past leads to a more convex $n \mapsto HH(n)$ curve than a constant population size. Figure 4.1 illustrates this difference of HH patterns between constant and contracting population size scenario.

This observation underlies the present work, justifying that we can expect to detect contractions using runs of homozygosity.

4.2.2 Parameter inference

Figure 4.1 illustrates the characterization of a demographic history by the summary statistic HH_{th}. Figure 4.3 illustrates the ability of HH_{th} to fit a sample of 100 $\widehat{HH}$ com-

puted on 100 independent data sets. These two observations indicate that we should be able to infer demographic parameters by numerically minimizing the distance between observed and expected HH. However, as mentioned in Section 4.2.1, the HH_{th} curve deviate from the observed $\widehat{HH}$ for large population size and/or large values of segment length because HH_{th} ignores the possibility of more than one recombination event per segment within one generation. This approximation will lead to a bias in the parametric estimations in some cases.

We choose to estimate θ_j under model $\mathcal{M}_j$ by minimizing a mean square relative error criterion between $HH_{\text{th}}(\theta_j)$ and $\widehat{HH}$:

$$\widehat{\theta}_0 \in \arg\min_{\theta_0} \sum_{n\in\mathcal{I}} \left(\frac{HH_{\text{th}}(\theta_0, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2$$

$$\widehat{\theta}_1 \in \arg\min_{\theta_1} \sum_{n\in\mathcal{I}} \left(\frac{HH_{\text{th}}(\theta_1, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2 .$$

Minimizing these two criteria requires to explore the values of $\theta \mapsto HH_{\text{th}}(\theta)$ for varying in the parameter space. However, the computation time of $HH_{\text{th}}(\theta)$ for a given value of θ in the parameter space is so high that we cannot evaluate HH_{th} on a fine grid of the parameter space. For example the computation of $HH_{\text{th}}(500, 1500, 10)$ from 0.71 seconds to 5 minutes, depending on the chosen number of artificial phases (see section 4.2.1 and Figure 4.2) Therefore, we use the R package `blackbox`, available on the CRAN, which performs prediction of a response function from simulated response values using a krigring technique. To refine the surface of the blackbox function HH_{th}, it samples cleverly new points in the parameter space. Instead of evaluating the blackbox function HH_{th} on this new grid, it performs predictions of the function and save computation time. After initial attempts with other optimization R functions (for example with the function `optim()` or the package `DiceOptim`), we found no significant differences for the optimization when the model is of dimension one but a more accurate optimization of the above criterion with the package `blackbox` when the dimension of the model is three.

4.3 Model choice

As mentioned before, demographic inference methods have been developed to estimate the historical effective population size from patterns of identical segment lengths (MacLeod et al. [2009], MacLeod et al. [2013], Palamara et al. [2012], Harris and Nielsen [2013], Browning and Browning [2015]). Some of them propose a control for potential overfitting but to the best of our knowledge, there exists no model choice procedure based on identical segment lengths to detect past changes in population size.

We propose to address the question: "Has the population size undergone a past change?" To achieve this, we used a model choice procedure between the simple model $\mathcal{M}_0$ of constant population size and the more complex model $\mathcal{M}_1$ with a past change in the population size.

Model $\mathcal{M}_0$ is nested in model $\mathcal{M}_1$. Thus, if we just compare the quality of the fit of each model to the data, aside from numerical considerations, we will always choose $\mathcal{M}_1$. Intuitively, when $\mathcal{M}_1$ produces an adjustment equal or only slightly better than $\mathcal{M}_0$, it is the sign that a simpler model captures the essential history. In this case, we aim to choose model $\mathcal{M}_0$ and avoid overfitting by a too complex model. This motivates our following choice of a penalized criterion. Our penalization relies on the computation of Sobol's sensitivity indices. The principle is to only take into account the part of the error that the model had the ability to explain. It is a form of penalty of the model related to its complexity, since a given adjustment error between observed and theoretical HH penalizes more a complex model than a simpler one.

Indeed, one of the major challenges of the sensitivity analysis is to identify the less contributing parameters to reduce the dimension of the model and quantify the error (Sobol et al. [2007], Saltelli et al. [2010]). In this purpose, the sensitivity analysis decomposes the output variance of the model in fractions attributed to each input parameter through Sobol's sensitivity indices. In our case, considering the theoretical haplotype homozygosity as a function of the parameters assumed to be random variables for the computation, these indices measure the part of the variance of the theoretical HH explained by each parameter of the more complex model considered.

The following algorithm summarizes the different algorithm steps of the model choice procedure.

Algorithm 5: Model Choice procedure

1 Computation of $\widehat{HH}$ on the data;
2 Parameters estimation under $\mathcal{M}_0$ and $\mathcal{M}_1$ with the R package `blackbox` numerically solving:

$$\widehat{\theta}_0 \in \arg\min_{\theta_0} \sum_{n\in\mathcal{I}} \left(\frac{HH_{\text{th}}(\theta_0, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2, \widehat{\theta}_1 \in \arg\min_{\theta_1} \sum_{n\in\mathcal{I}} \left(\frac{HH_{\text{th}}(\theta_1, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2;$$

3 Model choice according to:

$$\arg\min_{j\in\{0,1\}} \sum_{n\in\mathcal{I}} w_j(n) \left(\frac{HH_{\text{th}}(\widehat{\theta}_j, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2,$$

where $w_j(n), j \in \{0, 1\}$ are the sensitivity weights.

4.3.1 Sensitivity analysis

Variance-based sensitivity analysis decomposes the output variance of a function into fractions which can be attributed to input parameters. These methods measure the sensitivity across the whole parameter space, they can deal with nonlinear responses and measure the effect of interactions in non-additive systems. In this purpose, we consider the input parameters θ as random variables and $HH_{\text{th}}(., n)$ as the function to be evaluated on these parameters and apply the following analysis to $HH_{\text{th}}(n, .)$ for every $n \in \{1, \ldots, n_{\text{pos}}\}$ to compute the sensitivity weights $w_j(n)$, for $j \in \{0, 1\}$, and

$n \in \{1, \ldots, n_{\text{pos}}\}$.

Variance decomposition Consider a function f of a vector X of d input parameters $\{X_1, X_2, \ldots, X_d\}$ and denote $Y = f(X)$ the function output. Here we focus on scalar function and we have: $Y = HH_{\text{th}}(n, \theta)$, θ playing the role of X and $HH_{\text{th}}(n, .)$ the role of f for $n \in \{1, \ldots, n_{\text{pos}}\}$. Furthermore, we assume that the parameters X_i are random, independently and uniformly distributed in $[0, 1]$, $i \in \{1, \ldots, d\}$. This incurs no loss of generality because any input space can be transformed onto this unit hypercube. The random variable $Y = f(X)$ may be decomposed in the following way,

$$f(X) = f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{i<j}^{d} f_{ij}(X_i, X_j) + \ldots + f_{1,2,\ldots,d}(X_1, X_2, \ldots, X_d)$$

where f_0 is a constant and f_i is a function of X_i, f_{ij} a function of X_i and X_j and so on. Assuming that,

$$\int_0^1 f_{i_1,i_2,\ldots,i_s}(X_{i_1}, X_{i_2}, \ldots, X_{i_s}) \mathrm{d}X_k = 0, \text{ for } k = i_1, \ldots, i_s$$

leads to definitions of the terms of the functional decomposition in terms of conditional expected values,

$$f_0 = E(Y), f_i(X_i) = E(Y|X_i) - f_0,$$

$$f_{ij}(X_i, X_j) = E(Y|X_i, X_j) - f_0 - f_i - f_j.$$

Thus, f_i is the effect of varying X_i alone, that is the main effect of X_i, and f_{ij} is the effect of varying X_i ad X_j simultaneously, additional to the effect of their individual variations, that is a second order interaction. High-order terms have analogous definition.

Further assuming that $f(X)$ has a second order moment, the functional decomposition may be squared and integrated to give,

$$\int_0^1 f^2(X) \mathrm{d}X - f_0^2 = \sum_{s=1}^{d} \sum_{1 \le i_1 < i_2 < \ldots < i_s \le d} \int_0^1 f_{i_1 \ldots i_s}^2, \mathrm{d}X_{i_1} \ldots \mathrm{d}X_{i_s}.$$

If we remark that the left hand side is the variance of Y and the terms of the right hand side are variance terms decomposed with respect to sets of the X_i, we obtain the decomposition of variance expression:

$$\text{Var}(Y) = \sum_{i=1}^{d} V_i + \sum_{i<j}^{d} V_{ij} + \ldots + V_{12\ldots d},$$

where

$$V_i = \text{Var}_{X_i}(E_{X_{\sim i}}(Y|X_i)), V_{ij} = \text{Var}_{X_{ij}}(E_{X_{\sim ij}}(Y|X_i, X_j))) - V_i - V_j \text{ and so on.}$$

First-order indices A direct variance-based measure of sensitivity is the first-order Sobol index of a given parameter X_i denoted S_{X_i} in this work, defined as:

$$S_{X_i} = \frac{V_i}{\text{Var}(Y)}.$$

It measures the contribution of the main effect of X_i to the output variance of $f(X)$, that is the effect of varying X_i alone but averaged over variations in other input parameters. It is divided by the total variance to provide a fractional contribution. Higher-order interaction indices S_{X_i,X_j}, S_{X_i,X_j,X_k} and so on can be defined by dividing other terms in the variance decomposition by $\text{Var}(Y)$. This implies that:

$$\sum_{i=1}^{d} S_{X_i} + \sum_{i<j}^{d} S_{X_i,X_j} + \sum_{i<j<k}^{d} S_{X_i,X_j,X_k} + \ldots + S_{X_1,X_2,\ldots,X_d} = 1.$$

The first order indices measure the importance of each input variable in determining the output variance, the second and higher order indices $S_{X_i,X_j}, \ldots, S_{X_1,X_2,\ldots,X_d}$ measure the output variance due to the interaction of two or more parameters. However the present work use simply involves the first order indices of each parameter of the demographic model.

Calculation of indices For analytically tractable functions, the indices above may be calculated analytically by evaluating the integrals in the decomposition. However, in the vast majority of cases, and in this work, they are usually estimated by the Monte Carlo method.

The Monte Carlo approach involves generating a sequence of randomly distributed points inside the unit hypercube (strictly speaking these will be pseudorandom). In practice, it is common to substitute random sequences with low-discrepancy sequences to improve the efficiency of the estimators. This is then known as the Quasi-Monte Carlo method. Some low-discrepancy sequences commonly used in sensitivity analysis include the Sobol sequence and the Latin hypercube design. In this work we used the R package `lhs` to create Latin Hypercube Samples.

To calculate the sensitivity indices using the Monte Carlo method, we applied the following steps:

1. Generate a $N_{\text{simu}} \times 2d$ sample matrix with the `randomlhs()` function of the R package `lhs`. Each of its row is a sample point in the hyperspace of $2d$ dimensions.

2. Use the first d columns of this matrix as the matrix A and the remaining d columns as matrix B. This provides two independent samples of N_{simu} points in the d-dimensional unit hypercube.

3. Build d further $N_{\text{simu}} \times d$ matrices A_B^i, for $i = 1, \ldots, d$, such that the ith column of A_B^i is equal to the ith column of B and the remaining columns are from A.

4. The A, B and the d A_B^i matrices in total specify $N_{\text{simu}}(d+2)$ points in the input parameter space. Evaluate the function f at each design point in those matrices, giving a total of $N_{\text{simu}}(d+2)$ model evaluations.

5. Calculate the sensitivity indices using the estimators below.

The accuracy of the estimators is of course dependent on N_{simu}.

Estimators There are a number of possible Monte Carlo estimators available for both indices. Following Saltelli et al. [2010], we used the following estimator, computed by the `soboljansen()` function of the R package `sensitivity`:

$$\widehat{V_i} = \widehat{\text{Var}_{X_i}(E_{X_{\sim i}}(Y|X_i))} = \frac{1}{N}\sum_{j=1}^{N_{\text{simu}}} f(B)_j \left(f(A_B^i)_j - f(A)_j\right)$$

Computation cost For the estimation of the S_{X_i} for all input variables, $N_{\text{simu}}(d+2)$ model evaluations are required. Since N_{simu} is often of the order of hundreds or thousands of runs, computational expense can quickly become a problem when the model takes a significant amount of time for a single run. In our case it is not an issue since we compute these estimators once for all the numerical results presented after in Section 4.4.

4.3.2 Model choice penalty with sensitivity weights

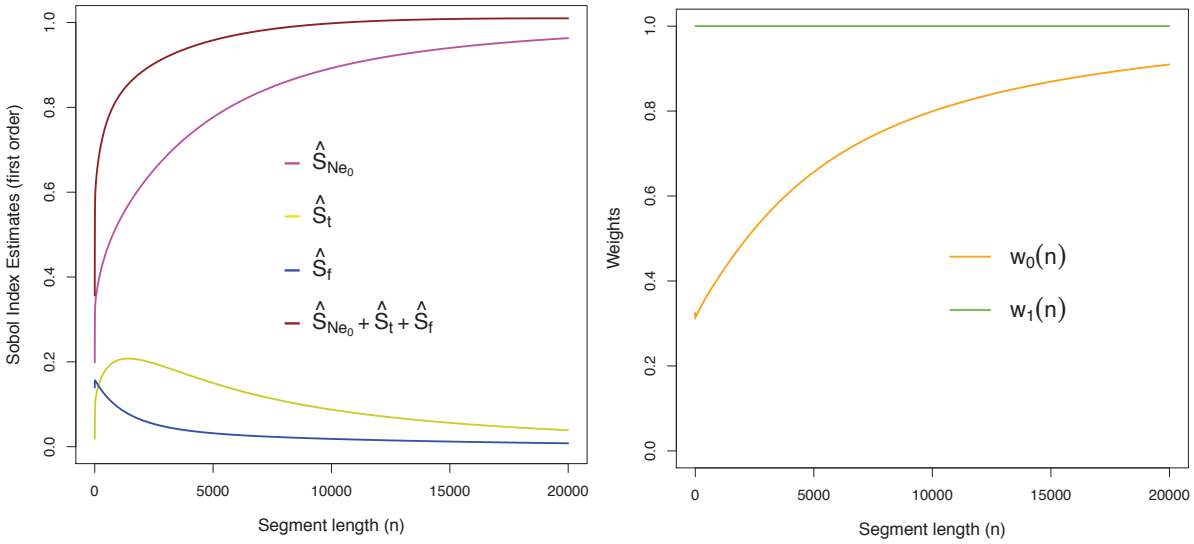
Our model choice criterion is based on a relative mean square formula in which, for every $n \in \mathcal{I}$, the difference between the prediction $HH_{\text{th}}(\widehat{\theta_j}, n)$ and the empirical $\widehat{HH}(n)$ is multiplied by a sensitivity weight $w_j(n)$ derived from the sensitivity analysis presented in section 4.3.1.

$$\arg\min_{j\in\{0,1\}} \sum_{i\in\mathcal{I}} w_j(n) \left(\frac{HH_{\text{th}}(\widehat{\theta}_j, n) - \widehat{HH}(n)}{\widehat{HH}(n)}\right)^2.$$

We only use the Sobol's indices of order 1 associated to each parameter, namely $S_{Ne_0}(n), S_t(n)$ and $S_f(n)$, represented on Figure 4.4a. These indices are estimated under the more complex model $\mathcal{M}_1$ with the R package `sensitivity`, available on the CRAN. For model $\mathcal{M}_0$, the weight $w_0(n)$ is proportional to the Sobol's index of order 1 relative to the parameter Ne_0, instead for $\mathcal{M}_1$, the weight $w_1(n)$ is proportionnal to the sum of the Sobol's indices or order one of the tree parameters (Ne_0, t, f). We choose to define $w_j(n)$ as the ratio between the sum of the first order Sobol's indices of the parameter of the model $\mathcal{M}_j$ and the sum of the first order Sobol's indices of the parameter of the more complex model:

$$w_0(n) = \left(\frac{\widehat{S}_{Ne_0}(n)}{\widehat{S}_{Ne_0}(n) + \widehat{S}_t(n) + \widehat{S}_f(n)} \right)^2, \quad w_1(n) = 1,$$

where $\widehat{S}_{Ne_0}(n), \widehat{S}_t(n)$ and $\widehat{S}_f(n)$ are the estimates of Sobol's indices of order 1 associated to each parameter. Figure 4.4b represents $w_0(n)$ as a function of the segment length n.



(a) *Sobol indices estimates (first order)* **(b)** *Sensitivity weight* $w_0(n)$

FIGURE 4.4: *Representation of (a) the Sobol indices estimates of first order* $\widehat{S}_{Ne_0}(n)$, $\widehat{S}_t(n)$, $\widehat{S}_f(n)$ *and (b) the sensitivity weights* $w_0(n)$ *and* $w_1(n)$ *as functions of the segment length* n.

The following section presents the estimates of the demographic parameter and the model choice true positive rate obtained when analyzing simulated data sets under constant and contracting population size. We also applied our procedure on a Holstein data set and found evidences of a past contraction in population size.

4.4 Numerical results

4.4.1 Simulated data sets

We repeat the inference procedure describes in algorithm 5 on 100 data sets simulated with the software `ms` of Hudson [2002], under a given model. For each data set, we simulated $L = 2000$ independent locus of 40000 adjacent positions. The calculation of $\widehat{HH}(n)$ for each segment length n is based on $W = 5000$ windows drawn uniformly at random among the L locus. To save computation time, we do not compute $\widehat{HH}(n)$ for every $n \in \{1, \ldots, n_{\text{pos}}\}$ but for a subset $\mathcal{I} \subset \{1, \ldots, n_{\text{pos}}\}$ of homogeneously distributed values throughout $\{1, \ldots, n_{\text{pos}}\}$. To compute HH_{th} relatively fast with enough precision under a contracting population size model, we divide every constant population size

phase in 10 artificial phases of the same duration (in number of generations). A unique phase under a constant population size model is sufficient as explained in section 4.2.1.

We consider the non penalized and penalized relative Mean Square Error (MSE) model choice criterion:

$$\arg\min_{j\in\{0,1\}} \sum_{n\in\mathcal{I}} \left(\frac{HH_{\text{th}}(\widehat{\theta}_j, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2, \ \arg\min_{j\in\{0,1\}} \sum_{n\in\mathcal{I}} w_j(n) \left(\frac{HH_{\text{th}}(\widehat{\theta}_j, n) - \widehat{HH}(n)}{\widehat{HH}(n)} \right)^2.$$

To assess the performance of the model choice procedure, we compute the True Positive Rate (TPR), that is the proportion of data sets for which the procedure chose the true model (under which we simulated the data sets). The model choice procedure is efficient if the TPR is close to 1 for both following cases: when the data sets are simulated under a constant population size (avoiding overfitting and not systematically choosing the more complex model) and when the data sets are simulated under a contracting population size (detecting a contraction event in the population history). If it is the case, we conclude that the procedure efficiently recognizes a constant versus a contracting population size.

To evaluate the accuracy of the demographic parameter estimates, we compute the mean over 100 estimations and symetric confidence intervals as:

$$\overline{\widehat{\theta_0}} = \overline{\widehat{Ne_0}} = \frac{1}{100} \sum_{k=1}^{100} \widehat{Ne_0}^{(k)}, \quad \overline{\widehat{\theta_1}} = (\overline{\widehat{Ne_0}}, \bar{\hat{t}}, \bar{\hat{f}})$$

$$CI_{0.95}(Ne_0) = \left[\overline{\widehat{Ne_0}} - q_{0.975} * \frac{\text{sd}(\widehat{Ne_0})}{100}; \overline{\widehat{Ne_0}} + q_{0.975} * \frac{\text{sd}(\widehat{Ne_0})}{100} \right],$$

where $q_{0.975}$ is the quantile of order 0.975 of the standard normal distribution $\mathcal{N}(0,1)$ and $\text{sd}(\widehat{Ne_0})$ the empirical standard deviation of the estimates $\widehat{Ne_0}^{(1)}, \ldots, \widehat{Ne_0}^{(100)}$. Analogous definitions are used to compute $CI_{0.95}(t)$ and $CI_{0.95}(f)$.

We successively analyzed the data sets considering IBS segments of length of $n_{\text{pos}} = 5000$, 10000 and 15000 positions, mostly affected respectively by the population size 10000, 5000, 3000 generations in the past and more ancient ($c = n_{\text{pos}} * r$ and $r = 10^{-8}$). These are relatively short segments and thus carry information mainly on ancient population size. We decide to use short segments to avoid the impact of the assumptions of a unique recombination event per segment per generation. The following results show that it still allows to infer past changes about 1500 generations before present.

Avoiding overfitting Theoretically, when analysing a given simulated data set with constant population size, the non penalized mean square criterion is expected to choose systematically $\mathcal{M}_1$ since, aside from maximization imprecision, both can exactly infer the simulated model, with a potential advantage for model $\mathcal{M}_1$ more flexible. In contrast and ideally, to goal is that when analyzing this simulated data set with constant population size, the penalized mean square criterion choose $\mathcal{M}_0$, since $\mathcal{M}_0$ and $\mathcal{M}_1$ will both infer the model with a comparable accuracy.

Figure 4.5 illustrates this fact, on a chosen simulated data set with constant population size $\theta_0 = 500$. Indeed, on this data set, the non penalized mean square criterion choose $\mathcal{M}_1$ whereas the penalized mean square criterion choose $\mathcal{M}_0$.

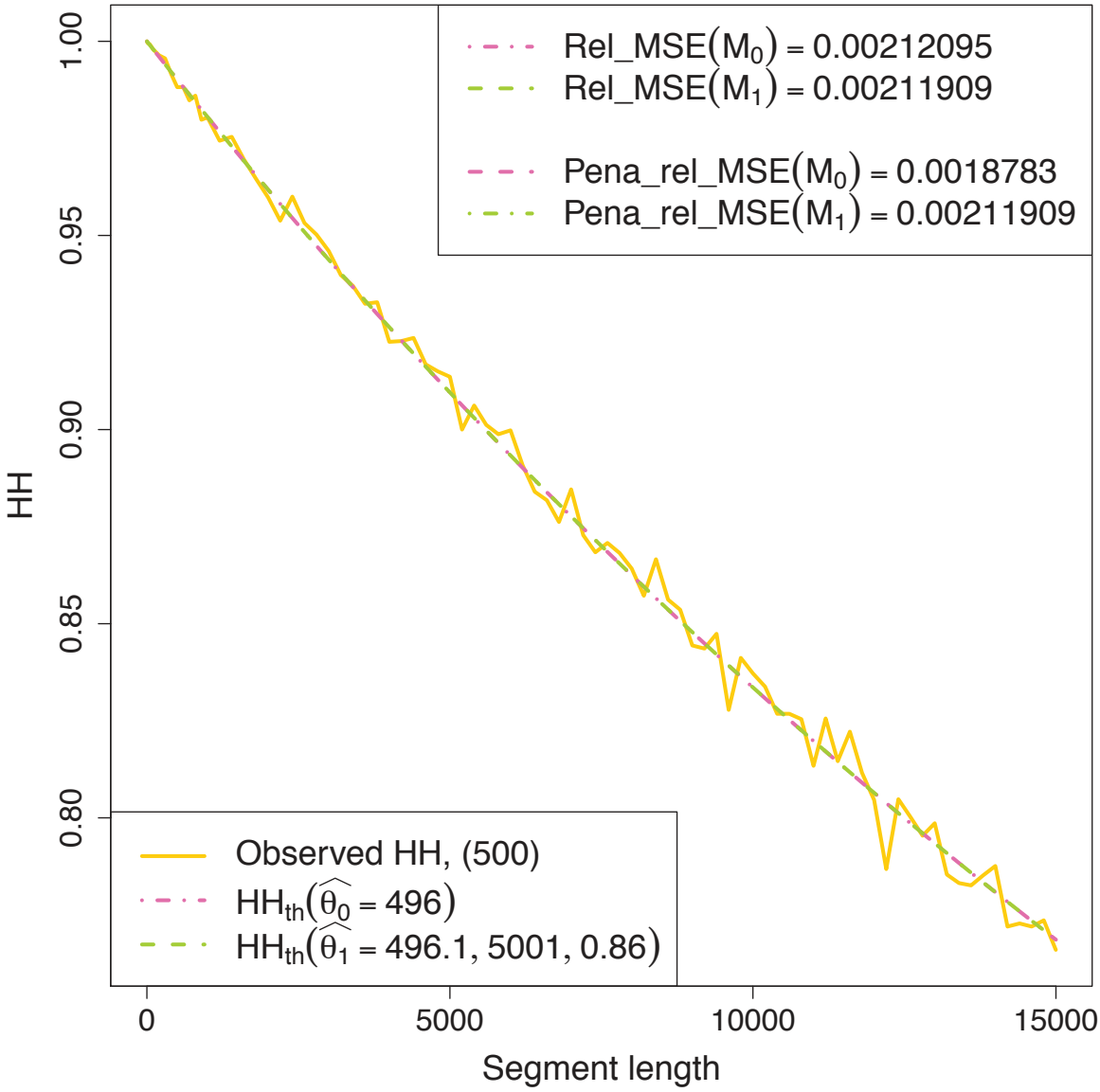


FIGURE 4.5: *Analyze following algorithm 5, of a given simulated data set under contracting population size $\theta_0 = (500)$. The observed $\widehat{HH}$ is represented in yellow line while the theoretical $HH_{th}(\widehat{\theta}_0)$ and $HH_{th}(\widehat{\theta}_1)$ evaluated on the parameter estimates under each model are represented respectively in pink and green solid lines while. Parameter estimates are presented in the bottom-left corner and values of the non penalized and penalized relative MSE criterion are presented in the top-right corner.*

Detecting a contraction On the other hand, when analyzing a given simulated data set with contracting population size, the non penalized mean square criterion is expected to choose $\mathcal{M}_1$ since $\mathcal{M}_0$ will poorly infer this model. Although the sensitivity weights penalize model $\mathcal{M}_1$ against $\mathcal{M}_0$. If $\mathcal{M}_1$ fit sufficiently more accurately the data than $\mathcal{M}_0$ it is chosen by the penalized criterion. Figure 4.6 illustrates this statement,

on a given simulated data set with contracting population size $\theta_1 = (500, 1500, 10)$. Indeed, the non penalized and the penalized mean square criterion choose $\mathcal{M}_1$.

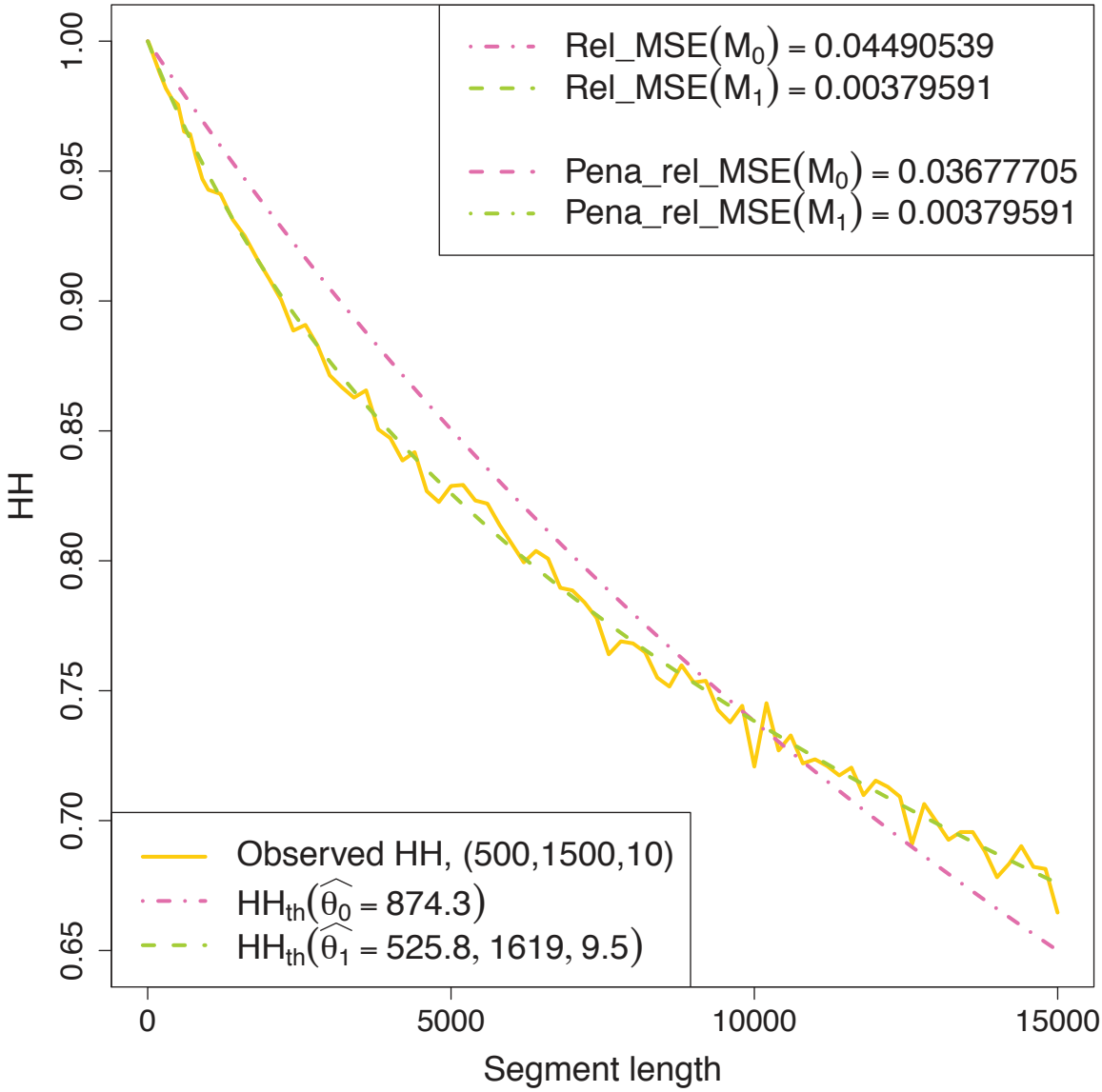


FIGURE 4.6: *Analyze following algorithm 5, of a given simulated data set under contracting population size $\theta_1 = (500, 1500, 10)$. The observed $\widehat{HH}$ is represented in yellow line while the theoretical $HH_{th}(\widehat{\theta}_0)$ and $HH_{th}(\widehat{\theta}_1)$ evaluated on the parameter estimates under each model are represented respectively in pink and green solid lines while. Parameter estimates are presented in the bottom-left corner and values of the non penalized and penalized relative MSE criterion are presented in the top-right corner.*

Simulated data sets under a constant population size model Tables 4.1 and 4.2 present the results obtained on simulated data sets under a constant population size respectively equal to 500 and 5000. Our model choice criterion penalized with sensitivity weights lead to TPR between 98% and 100% when $\theta_0 = 500$. When $\theta_0 = 5000$, our penalized model choice criterion lead to high TPR when analyzing only short IBS segments length. The TPR decreases down to 82% when analyzing IBS segment length

up to 15000 adjacent positions. The larger the constant population size, the closest the HH patterns to a contracting population size HH pattern. When zooming on $HH_{\text{th}}(n)$ for $n \in \{1, \ldots, 5000\}$, the curve is closed to a linear trend line but when considering $n \in \{1, \ldots, 15000\}$, the curve is more convex and not close to a line. The flexibility of model $\mathcal{M}_1$ allows it to adapt on different curve types, even more with the range of values of n. In this case, we may need more contrasted penalization weights to balance the flexibility of $\mathcal{M}_1$.

The sensitivity weights allow to recognize constant population size data sets with a high level of confidence and avoid overfitting when considering short IBS segment lengths.

With regard to parameter estimations, we notice that estimates under $\mathcal{M}_0$ of the recent population size are very close to the true value of the simulation. The confidence intervals are narrow in both cases $\theta_0 = 500$ and $\theta_0 = 5000$.

TABLE 4.1: ***Model choice procedure applied on*** 100 ***data sets simulated under*** $\mathcal{M}_0$***, with*** $\theta_0 = Ne_0 = 500$***. True Positive Rate (TPR) obtained with the penalized relative MSE criterion are presented as well as the mean of the*** 100 ***population size estimates and the confidence interval based on Gaussian quantiles.***

	n_{pos}	5000	10000	15000
TPR	Penalized rel. MSE	1.00	0.98	0.98
$\widehat{\overline{\theta_0}}$	$\widehat{\overline{Ne_0}}$	499	500	499
	$CI_{0.95}(Ne_0)$	$[496 - 503]$	$[496 - 503]$	$[495 - 502]$

TABLE 4.2: ***Model choice procedure applied on*** 100 ***data sets simulated under*** $\mathcal{M}_0$***, with*** $\theta_0 = Ne_0 = 5000$***. True Positive Rate (TPR) obtained with the penalized relative MSE criterion are presented as well as the mean of the*** 100 ***population size estimates and the confidence interval based on Gaussian quantiles.***

	n_{pos}	5000	10000	15000
TPR	Penalized Adj.	1.00	1.00	0.82
$\widehat{\overline{\theta_0}}$	$\widehat{\overline{Ne_0}}$	4955	4915	4894
	$CI_{0.95}(Ne_0)$	$[4935 - 4976]$	$[4891 - 4939]$	$[4866 - 4923]$

Simulated data sets under a contracting population size model Table 4.3 presents the results obtained on simulated data sets under a contracting population size model, currently equal to 500 with a past contraction of a factor 10, 1500 generations before present. The sensitivity weights applied to penalized criterion are always higher for $\mathcal{M}_1$ than for $\mathcal{M}_0$. It is then satisfying that, despite the penalty applied to model $\mathcal{M}_1$ compared to model $\mathcal{M}_0$ in the analysis of simulated data sets under $\mathcal{M}_1$, the TPR of the penalized criterion increase with n_{pos} from 79% to 0.96% for the scenario $\theta_1 = (500, 1500, 10)$.

The penalization with sensitivity weights does not prevent the detection of contraction. Overall, the sensitivity weights allowed us to recognize both constant population size and contracting population size data sets.

With regard to parameter estimations of the data sets under a contraction $\theta_1 = (500, 1500, 10)$, the estimations of Ne_0, t and f under $\mathcal{M}_1$ tend to be higher than the true value but move closer to the true value when the segment length increases. The confidence intervals are large, reflecting the numerical optimization difficulty in a parameter space of dimension 3 instead of 1 previously.

TABLE 4.3: ***Model choice procedure applied on*** 100 ***data sets simulated under*** $\mathcal{M}_1$***, with*** $\theta_1 = (Ne_0, t, f) = (500, 1500, 10)$***. True Positive Rate (TPR) obtained with the penalized relative MSE criterion are presented as well as the mean of the*** 100 ***estimates of each parameter and the confidence intervals based on Gaussian quantiles.***

	n_{pos}	5000	10000	15000
TPR	Penalized Adj.	0.79	0.85	0.96
$\widehat{\overline{\theta_1}}$	$\widehat{\overline{Ne_0}}$	731	627	542
	$CI_{0.95}(Ne_0)$	$[708 - 755]$	$[611 - 643]$	$[526 - 557]$
	$\widehat{\overline{t}}$	3370	2485	1886
	$CI_{0.95}(t)$	$[3143 - 3596]$	$[2354 - 2616]$	$[1767 - 2005]$
	$\widehat{\overline{f}}$	11.7	12.3	12.0
	$CI_{0.95}(f)$	$[11.1 - 12.2]$	$[11.8 - 12.7]$	$[11.6 - 12.3]$

4.4.2 Cow data set

This data set consists of the positions of the heterozygotes sites for the whole genome of an Holstein cow. We build a data set by drawing 2000 locus uniformly at random on the whole genome. To obtain a contraction detection rate (CDR), we analyze 100 such data sets, following algorithm 5.

We obtained contraction detection rates around 90% with our penalized relative MSE criterion when considering segments of 5000, 10000 and 15000 adjacent positions. We conclude that this data set reflects a past change in the population size.

The parameter estimates suggest a present effective population size of 6983 individuals that have undergone a contraction 7652 generations before present of intensity 4.25. The confidence intervals are still large, again due to numerical issues in the optimization.

TABLE 4.4: ***Model choice procedure applied on the Holstein data set. Contraction Detection Rate (CDR) obtained with the penalized relative MSE criterion are presented as well as the mean of the*** 100 ***estimates of each parameter and the confidence intervals based on Gaussian quantiles.***

	n_{pos}	5000	10000	15000
CDR	Penalized Adj.	0.89	0.90	0.89
$\widehat{\overline{\theta_1}}$	$\widehat{\overline{Ne_0}}$	8913	7191	6983
	$CI_{0.95}(Ne_0)$	$[8775 - 9050]$	$[7050 - 7333]$	$[6877 - 7089]$
	$\widehat{\overline{t}}$	15059	7921	7652
	$CI_{0.95}(t)$	$[14214 - 15904]$	$[7279 - 8563]$	$[6877 - 8428]$
	$\widehat{\overline{f}}$	4.74	4.41	4.25
	$CI_{0.95}(f)$	$[4.21 - 5.27]$	$[3.96 - 4.86]$	$[3.83 - 4.67]$

We compared our demographic inference with those of MacLeod et al. [2013] and Li and Durbin [2011] presented on Fig. 5 in MacLeod et al. [2013]. Figure 4.7 is taken from Fig. 5 of MacLeod et al. [2013] and we add our inferred history (bold yellow line) with confidence intervals (dashed yellow lines) to those of MacLeod et al. [2013] and Li and Durbin [2011]. Our inferred demographic history is simpler than the inferred demography of MacLeod et al. [2013] and Li and Durbin [2011], because it allows for a unique past change. Since we used relatively short IBS segment lengths, we have few information about the recent past history. The recent population size that we inferred is in the range of the very different estimations from a method to another presented in Figure. We also capture the more ancient population history. Indeed the yellow line representing our result is close to the green one of MacLeod et al. [2013] and the pink one of Li and Durbin [2011]. We found that our results are consistent with those of MacLeod et al. [2013] and Li and Durbin [2011].

Conclusion - Discussion

In this work, we proposed a model choice procedure between demographic models of different complexity based on the length of IBS segments between two genomes. To take advantage of genome wide sequence data, we considered the dependency of genealogies of adjacent positions in the genome by modeling the recombination. Since the two considered models are nested, we developed a penalized model choice criterion based on the comparison of observed and predicted haplotype homozygosity. Our penalization relies on the computation of Sobol's sensitivity indices. It is a form of

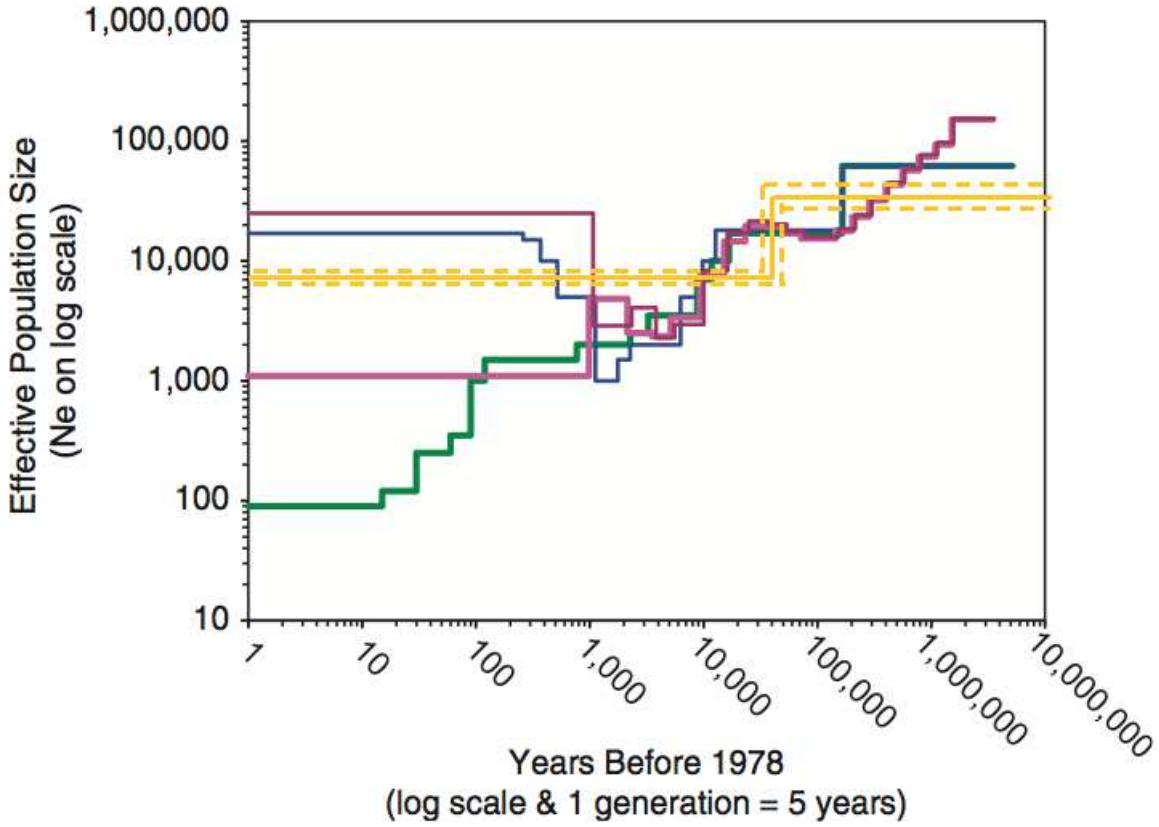


FIGURE 4.7: *Inferred demographic history (bold yellow line) with confidence intervals (dashed yellow lines) with our method for the Holstein population. Inferred demography for a given cow corrected sequence using both MacLeod et al. [2013] method (bold green line) and the Li and Durbin [2011] PSMC method (bold pink line). Also shown is the inferred demography from the cow filtered sequence (with residual false-positive errors) using both MacLeod et al. [2013] method (blue) and the PSMC method (maroon).*

penalty related to the complexity of the model since a given adjustment error (between observed and theoretical HH) penalize more a complex model than a simpler one.

We first applied our model choice procedure on simulated data sets. On the one hand, when the data sets were simulated under a contracting population size model, our model choice criterion allows to detect the contraction. The data sets not recognized as coming from a contracting population size mostly face numerical issues in the optimization under the contracting model. It is due to the large parameter space to explore and the computation cost of the evaluation of HH_{th} on a given parameter value in this space. The solution will be in a finer usage of the `blackbox` package. At the same time, we obtained accurate parameter estimates when considering large segment lengths, in accordance with the date of the contraction considered.

On the other hand, we showed that when the data sets were simulated under a constant population size model, our model choice criterion allows to avoid overfitting and to detect a constant population size model. At the same time, we obtained accurate parameter estimates for most cases. When the population size is relatively large (equal to 5000 for example), the population size estimate is biased. It is a consequence of the fact that HH_{th} ignores the possibility of more than one recombination per segment

per generation. A natural perspective is thus to account for multiple recombination events in the computation of HH_{th}. In this case, the detection of the constant population size is weaker than in other cases. In our opinion, the solution can be to choose a penalty, based on sensitivity indices but more contrasted between the two models to balance to flexibility of $\mathcal{M}_1$. Setting aside the numerical issues in the optimization under the contracting model, a stronger penalty could not prevent the contraction detection when the data set comes actually from a contracting population size.

Finally, we analyzed a cattle data set consisting of the whole genome of an Holstein cow. We obtained contraction detection rates about 90% with penalized criterion when using IBS segment lengths up to 15kb. We conclude that this data set reflects a past change in the population size.

The main perspective of this work is to extend the model choice procedure to more complex models. We could consider models of piecewise constant population size with more changes or exponential growing or declining population size.

5 Discussion and perspectives

5.1 Perspectives about sequential important sampling with resampling

The first contribution of this work improved sequential importance sampling algorithms for likelihood inference of demographic parameters by adding a resampling procedure based on Liu et al. [2001]. The new resampling probability distribution that we considered depends on the SIS weights, as proposed in Liu et al. [2001] and Liu [2008], but also on the pairwise composite likelihood of the sample, providing additional information about the future trend of each sampled history. We first showed through simulations that resampling allows to reduce the variance and the bias in the likelihood estimate at a parameter point. Then we showed that the increased precision in estimation of likelihood also improve the likelihood-based inferences. According to numerical results based on the analysis of simulated data sets under various scenarios of past population size contraction with different strength and timing, we conclude that the resampling procedure helps to correct the inefficiency of the SIS proposal distribution.

A first extension of this work is to compute the composite likelihood under a varying population size, for example an exponentially contracting model, to use a more faithful information on the future trend of each sampled history. We do not have analytically formulas for this computation as under a constant population size but we can approximate it by numerical solutions.

Another possible further work is to extend this resampling procedure to other mutational models, for example to analyse SNPs data with the SISR algorithm. It will require the choice of an efficient proposal distribution and of a resampling distribution. The resampling procedure may also allow to treat more complex demographic parameters thanks to gain in computation time.

Finally, a perspective to be explored is to use our resampling technique with a distribution depending on the pairwise composite likelihood in the step E of an EM algorithm to calculate the expected value of the likelihood function, with respect to the conditional distribution of the latent variable given the observed variable under the

current estimate of the parameters.

5.2 Perspectives about model choice between demographic models based on IBS segment lengths

In the second contribution of this thesis, we proposed a model choice procedure between demographic models of different complexity based on the length of IBS segments between two haplotypes. To take advantage of genome wide sequence data with known genome, we considered the dependency of genealogies of adjacent positions in the genome by modeling the recombination. We chose to focus on a constant population size model a demographic model with a unique past change in the population size. Since the two considered models are nested we developed a penalized model choice criterion based on the comparison of observed and predicted haplotype homozygosity. Our penalization relies on the computation of Sobol's sensitivity indices and is a form of penalty related to the complexity of the model. On the one hand, when the data sets were simulated under a contracting population size model, our model choice criterion allows to detect the contraction. On the other hand, we showed that when the data sets were simulated under a constant population size model, our model choice criterion allows to avoid overfitting and to detect a constant population size model. Overall we obtain accurate parameter estimates on simulated data sets under constant and contracting population size.

A first possibility to be explored is the approximate coalescent used to derive theoretical HH_{th}. Indeed, the present computation of HH_{th} ignores the possibility of more than one recombination events per segment per generation. When the population size is relatively large and the segment considered quite long, this approximation leads to a bias in the parameter estimates. The second assumption to be overcome consider that the recombination event always takes place in the first generation of a constant population size phase. This assumption has no effect when the population size is constant far in the past but leads to approximate HH_{th} when the population size has undergone past changes and thus requires to find a trade-off between accuracy and cost of this computation.

A second aspect of a further work is to extend the model choice procedure to more complex models. It would be interesting to consider models of piecewise constant population size with more changes or exponential growing or declining population size. We expect that we will face computation issues, due to a computation cost of the theoretical haplotype homozygosity combined to larger parameter spaces. A previous effort on numerical considerations, that is computation of the theoretical haplotype homozygosity and a refine use of the `blackbox` package for optimizing is required.

Finally our model choice procedure, with a penalty based on Sobol's sensitivity indices, is general enough to be applied on the comparison of different statistics based of the IBS segment lengths distribution. In particular, we could choose a statistic faster to compute than the theoretical haplotype homozygosity. This will enable us to consider more complex demographic models.

RÉSUMÉ

Cette thèse consiste à améliorer les outils statistiques adaptés à des modèles stochastiques de génétiques des populations et à développer des méthodes statistiques adaptées à des données génétiques de nouvelle génération.

Pour un modèle paramétrique basé sur le coalescent, la vraisemblance en un point de l'espace des paramètres s'écrit comme la somme des probabilités de toutes les histoires (généalogies munies de mutations) possibles de l'échantillon observé. À l'heure actuelle, les meilleures méthodes d'inférence des paramètres de ce type de modèles sont les méthodes bayésiennes approchées et l'approximation de la fonction de vraisemblance. L'algorithme d'échantillonnage préférentiel séquentiel (SIS) estime la vraisemblance, en parcourant de manière efficace l'espace latent de ces histoires. Dans ce schéma, la distribution d'importance propose les histoires de l'échantillon observé les plus probables possible. Cette technique est lente mais fournit des estimations par maximum de vraisemblance d'une grande précision. Les modèles que nous souhaitons inférer incluent des variations de la taille de la population. Les méthodes d'IS ne sont pas efficaces pour des modèles en déséquilibre car les distributions d'importance ont été développées pour une population de taille constante au cours du temps. Le temps de calcul augmente fortement pour la même précision de l'estimation de la vraisemblance. La première contribution de cette thèse a consisté à explorer l'algorithme SIS avec rééchantillonnage (SISR). L'idée est de rééchantillonner de façon à apprendre quelles sont les histoires proposées par la distribution d'importance qui seront les plus probables avant d'avoir terminé leur simulation et diminuer le temps de calcul. Par ailleurs, nous avons proposé une nouvelle distribution de rééchantillonnage, tirant profit de l'information contenue dans la vraisemblance composite par paire de l'échantillon.

Le développement récent des technologies de séquençage à haut débit a révolutionné la génération de données de polymorphisme chez de nombreux organismes. Les méthodes d'inférence classiques de maximum de vraisemblance ou basées sur le *Sites Frequency Spectrum*, adaptées à des jeux de données de polymorphisme génétique de quelques loci, supposent l'indépendance des généalogies de différents loci. Pour tirer parti de données beaucoup plus denses sur le génome, nous considérons la dépendance des généalogies sur des positions voisines du génome et nous modélisons la recombinaison génétique. Alors, la vraisemblance prend la forme d'une intégrale sur tous les graphes de recombinaison ancestraux possibles pour les séquences échantillonnées, un espace de bien plus grande dimension que l'espace des généalogies. Les méthodes d'inférence basées sur la vraisemblance ne peuvent plus être utilisées sans plus d'approximations. De nombreuses méthodes infèrent les changements historiques de la taille de la population mais ne considèrent pas la complexité du modèle ajusté. Même si certaines proposent un contrôle d'un potentiel surajustement du modèle, à notre connaissance, aucune procédure de choix de modèle entre des modèles démographiques de complexité différente n'a été proposée à partir de longueurs de segments identiques. Nous nous concentrons sur un modèle de taille de population constante et un modèle de population ayant subi un unique changement de taille dans le passé. Puisque ces modèles sont emboîtés, la deuxième contribution de cette thèse a consisté à développer un critère de choix de modèle pénalisé basé sur la comparaison d'homozy-

gotie haplotypique observée et théorique. Notre pénalisation, reposant sur des indices de sensibilité de Sobol, est liée à la complexité du modèle. Ce critère pénalisé de choix de modèle nous a permis de choisir entre un modèle de taille de population constante et un modèle présentant un changement passé de la taille de la population sur des jeux de données simulés et sur un jeux de données de vaches.

Mots clefs : Statistiques, algorithmes stochastiques, inférence démographique, coalescent, génétique des populations, échantillonnage préférentiel, rééchantillonnage, choix de modèle pénalisé, homozygotie haplotypique, indices de sensibilité.

ABSTRACT

This thesis aims to improve statistical methods suitable for stochastic models of population genetics and to develop statistical methods adapted to next generation sequencing data.

Sequential importance sampling algorithms have been defined to estimate likelihoods in models of ancestral population processes. However, these algorithms are based on features of the models with constant population size, and become inefficient when the population size varies in time, making likelihood-based inferences difficult in many demographic situations. In the first contribution of this thesis, we modify a previous sequential importance sampling algorithm to improve the efficiency of the likelihood estimation. Our procedure is still based on features of the model with constant size, but uses a resampling technique with a new resampling probability distribution depending on the pairwise composite likelihood. We tested our algorithm, called sequential importance sampling with resampling (SISR) on simulated data sets under different demographic cases. In most cases, we divided the computational cost by two for the same accuracy of inference, in some cases even by one hundred. This work provides the first assessment of the impact of such resampling techniques on parameter inference using sequential importance sampling, and extends the range of situations where likelihood inferences can be easily performed.

The recent development of high-throughput sequencing technologies has revolutionized the generation of genetic data for many organisms : genome wide sequence data are now available. Classical inference methods (maximum likelihood methods (MCMC, IS), methods based on the Sites Frequency Spectrum (SFS)) suitable for polymorphism data sets of some loci assume that the genealogies of the loci are independent. To take advantage of genome wide sequence data with known genome, we need to consider the dependency of genealogies of adjacent positions in the genome. Thus, when we model recombination, the likelihood takes the form of an integral over all possible ancestral recombination graph for the sampled sequences. This space is of much larger dimension than the genealogies space, to the extent that we cannot handle likelihood-based inference while modeling recombination without further approximations. Several methods infer the historical changes in the effective population size but do not consider the complexity of the demographic model fitted. Even if some of them propose a control for potential overfitting, to the best of our knowledge, no

model choice procedure between demographic models of different complexity have been proposed based on IBS segment lengths. The aim of the second contribution of this thesis is to overcome this lack by proposing a model choice procedure between demographic models of different complexity. We focus on a simple model of constant population size and a slightly more complex model with a single past change in the population size. Since these models are embedded, we developed a penalized model choice criterion based on the comparison of observed and predicted haplotype homozygosity. Our penalization relies on Sobol's sensitivity indices and is a form of penalty related to the complexity of the model. This penalized model choice criterion allowed us to choose between a population of constant size and a population size with a past change on simulated data sets and also on a cattle data set.

Keywords: Statistics, stochastics algorithms, demographic inference, coalescent, population genetics, importance sampling, resampling, penalized model choice, haplotype homozygosity, sensitivity indices.

Bibliographie

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Quentin D Atkinson, Russell D Gray, and Alexei J Drummond. mtdna variation predicts population size in humans and reveals a major southern asian chapter in human prehistory. *Molecular biology and evolution*, 25(2):468–474, 2008.

Sadoune Ait Kaci Azzou, Fabrice Larribe, and Sorana Froda. A new method for estimating the demographic history from dna sequences: an importance sampling approach. *Frontiers in genetics*, 6, 2015.

Sadoune Ait Kaci Azzou, F Larribe, and S Froda. Inferring the demographic history from dna sequences: An importance sampling approach based on non-homogeneous processes. *Theoretical population biology*, 2016.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Simon Boitard, Willy Rodriguez, Flora Jay, Stefano Mona, and Frédéric Austerlitz. Inferring population size history from large samples of genome-wide molecular data-an approximate bayesian computation approach. *PLoS Genet*, 12(3):e1005877, 2016.

Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418, 2015.

Ralph Burgess and Ziheng Yang. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular biology and evolution*, 25(9):1979–1994, 2008.

Gilles Celeux and Jean Diebolt. The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2(1):73–82, 1985.

Hua Chen. The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theoretical population biology*, 81(2):179–195, 2012.

Hua Chen. Population genetic studies in the genomic sequencing era. *Dongwuxue Yanjiu*, 36(4):223, 2015.

Jean-Marie Cornuet, Filipe Santos, Mark A Beaumont, Christian P Robert, Jean-Michel Marin, David J Balding, Thomas Guillemaud, and Arnaud Estoup. Inferring population history with diy abc: a user-friendly approach to approximate bayesian computation. *Bioinformatics*, 24(23):2713–2719, 2008.

Anthony C Davison. *Statistical Models*. Cambridge University Press, 2003.

Maria De Iorio and Robert C Griffiths. Importance sampling on coalescent histories. i. *Advances in Applied Probability*, 36:417–433, 2004a.

Maria De Iorio and Robert C Griffiths. Importance sampling on coalescent histories. ii: Subdivided population models. *Advances in Applied Probability*, 36(2):434–454, 2004b.

Maria De Iorio, Robert C Griffiths, Raphael Leblois, and François Rousset. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical population biology*, 68(1):41–53, 2005.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128, 1999.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Hans Ellegren. Heterogeneous mutation processes in human microsatellite dna sequences. *Nature genetics*, 24(4):400–402, 2000.

Brent C Emerson, Emmanuel Paradis, and Christophe Thébaud. Revealing the demographic histories of species using dna sequences. *Trends in Ecology & Evolution*, 16 (12):707–716, 2001.

Steven N Evans, Yelena Shvets, and Montgomery Slatkin. Non-equilibrium theory of the allele frequency spectrum. *Theoretical population biology*, 71(1):109–119, 2007.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLoS Genet*, 9(10):e1003905, 2013.

Muhammad Faisal, Andreas Futschik, and Claus Vogl. Exact likelihood calculation under the infinite sites model. *Computation*, 3(4):701–713, 2015.

Emma K Finlay, C Gaillard, SMF Vahidi, SZ Mirhoseini, H Jianlin, XB Qi, MAA El-Barody, JF Baird, BC Healy, and Daniel G Bradley. Bayesian inference of population expansions in domestic bovines. *Biology Letters*, 3(4):449–452, 2007.

Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

Yun-Xin Fu. Statistical properties of segregating sites. *Theoretical population biology*, 48 (2):172–197, 1995.

Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

Robert C Griffiths and Simon Tavaré. Ancestral inference in population genetics. *Statistical Science*, 9:307–319, 1994a.

Robert C Griffiths and Simon Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310):403–410, 1994b.

Robert C Griffiths and Simon Tavare. Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46(2):131–159, 1994.

Ilan Gronau, Melissa J Hubisz, Brad Gulko, Charles G Danko, and Adam Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10):1031–1034, 2011.

Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10):e1000695, 2009.

Kelley Harris and Rasmus Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, 9(6):e1003521, 2013.

Ben J Hayes, Peter M Visscher, Helen C McPartlan, and Mike E Goddard. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, 13(4):635–643, 2003.

Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA, 2004.

Simon YW Ho and Beth Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular ecology resources*, 11(3):423–434, 2011.

Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3(2):e7, 2007.

Asger Hobolth, Marcy K Uyenoyama, and Carsten Wiuf. Importance sampling for the infinite sites model. *Statistical applications in genetics and molecular biology*, 7(1): Article32, 2008.

Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

Paul A. Jenkins and Robert C. Griffiths. Inference from samples of dna sequences using a two-locus model. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 18(1):109–127, JAN 2011. ISSN 1066-5277. doi: 10.1089/cmb.2009.0231.

Motoo Kimura. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences*, 41(3):144–150, 1955.

Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

Motoo Kimura and James F Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725–738, 1964.

Motoo Kimura and Tomoko Ohta. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences*, 75 (6):2868–2872, 1978.

John FC Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982a.

John FC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982b.

John Frank Charles Kingman. *Poisson processes*. Oxford university press, 1992.

Andrew Kitchen, Michael M Miyamoto, and Connie J Mulligan. Utility of dna viruses for studying human host history: case study of jc virus. *Molecular phylogenetics and evolution*, 46(2):673–682, 2008.

Martin Kreitman. locus of drosophila melanogaster. *Nature*, 304:4, 1983.

Aude Lalis, Raphael Leblois, Emmanuelle Stoetzel, Touria Benazzou, Karim Souttou, Christiane Denys, and Violaine Nicolas. Phylogeography and demographic history of Shaw's Jird (Meriones shawii complex) in North Africa. *BIOLOGICAL JOURNAL OF THE LINNEAN SOCIETY*, 118(2):262–279, JUN 2016. ISSN 0024-4066. doi: {10.1111/bij.12725}.

Raphaël Leblois, Arnaud Estoup, and Francois Rousset. Ibdsim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, 9(1):107–109, 2009.

Raphaël Leblois, Pierre Pudlo, Joseph Néron, François Bertaux, Champak Reddy Beeravolu, Renaud Vitalis, and François Rousset. Maximum likelihood inference of population size contractions from microsatellite data. *Molecular biology and evolution*, 31:2805–2823, 2014.

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Jun S Liu, Rong Chen, and Tanya Logvinenko. A theoretical framework for sequential importance sampling with resampling. In Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors, *Sequential Monte Carlo methods in practice*, Statistics for Engineering and Information Science, pages 225–246. Springer New York, 2001. ISBN 978-1-4419-2887-0.

I. MacLeod, T. Meuwissen, B. Hayes, and M. Goddard. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics research*, 91 (6):413–426, 2009.

I. MacLeod, D. Larkin, H. Lewin, B. Hayes, and M. Goddard. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. *Molecular Biology and Evolution*, 30(9):2209–2223, 2013.

Gkikas Magiorkinis, Emmanouil Magiorkinis, Dimitrios Paraskevis, Simon YW Ho, Beth Shapiro, Oliver G Pybus, Jean-Pierre Allain, and Angelos Hatzakis. The global spread of hepatitis c virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Med*, 6(12):e1000198, 2009.

Thomas Mailund, Julien Y Dutheil, Asger Hobolth, Gerton Lunter, and Mikkel H Schierup. Estimating divergence time and ancestral effective population size of bornean and sumatran orangutan subspecies using a coalescent hidden markov model. *PLoS Genet*, 7(3):e1001319, 2011.

Gustave Malécot. La consanguinité dans une population limitée. *Compt. Rend. Acad. Sci. Paris*, 222:841–843, 1946.

Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

Paul Marjoram and Peter Donnelly. Human demography and the time since mitochondrial eve. *Institute for Mathematics and Its Applications*, 87:107, 1997.

Gabor T Marth, Eva Czabarka, Janos Murvai, and Stephen T Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372, 2004.

Magnus Nordborg and Simon Tavaré. Linkage disequilibrium: what history has to tell us. *TRENDS in Genetics*, 18(2):83–90, 2002.

Tomoko Ohta and Motoo Kimura. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical research*, 22(02):201–204, 1973.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Peer. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822, 2012.

A Polanski and M Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, 2003.

Oliver G Pybus, Andrew Rambaut, and Paul H Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437, 2000.

Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genet*, 10(5):e1004342, 2014.

Harald Ringbauer, Graham Coop, and Nick Hamilton Barton. Inferring recent demography from isolation by distance of long shared sequence blocks. *bioRxiv*, page 076810, 2016.

Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

François Rousset and Raphaël Leblois. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Molecular biology and evolution*, 24(12):2730–2745, 2007.

François Rousset and Raphaël Leblois. Likelihood-based inferences under isolation by distance: two-dimensional habitats and confidence intervals. *Molecular biology and evolution*, 29(3):957–973, 2012.

François Rousset, Champak Reddy Beeravolu, and Raphaël Leblois. Likelihood analysis of population genetic data under coalescent models: computational and inferential aspects. *Journal de la société Française de Statistiques*, submitted.

A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270, 2010.

Ian W Saunders, Simon Tavaré, and GA Watterson. On the genealogy of nested subsamples from a haploid population. *Advances in Applied probability*, pages 471–491, 1984.

Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

Thomas A Severini. *Likelihood methods in statistics*. Oxford Univ. Press, 2000.

M. Sobol, S. Tarantola, D. Gatelli, S. Kucherenkoc, and W. Mauntz. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering & System Safety*, 92(7):957–960, 2007.

Matthew Stephens and Peter Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635, 2000.

Mathias Stiller, Gennady Baryshnikov, Hervé Bocherens, Aurora Grandal d'Anglade, Brigitte Hilpert, Susanne C Münzel, Ron Pinhasi, Gernot Rabeder, Wilfried Rosendahl, Erik Trinkaus, et al. Withering away—25,000 years of genetic decline preceded cave bear extinction. *Molecular biology and evolution*, 27(5):975–978, 2010.

Jay F Storz and Mark A Beaumont. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite dna variation using a hierarchical bayesian model. *Evolution*, 56(1):154–166, 2002.

Korbinian Strimmer and Oliver G Pybus. Exploring the demographic history of dna sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18(12): 2298–2305, 2001.

Fumio Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460, 1983.

Thomas M. Vignaud, Jeffrey A. Maynard, Raphael Leblois, Mark G. Meekan, Ricardo Vazquez-Juarez, Deni Ramirez-Macias, Simon J. Pierce, David Rowat, Michael L. Berumen, Champak Beeravolu, Sandra Baksay, and Serge Planes. Genetic structure of populations of whale sharks among ocean basins and evidence for their historic rise and recent decline. *MOLECULAR ECOLOGY*, 23(10):2590–2601, MAY 2014a. ISSN 0962-1083. doi: {10.1111/mec.12754}.

Thomas M. Vignaud, Johann Mourier, Jeffrey A. Maynard, Raphael Leblois, Julia Spaet, Eric Clua, Valentina Neglia, and Serge Planes. Blacktip reef sharks, Carcharhinus melanopterus, have high genetic structure and varying demographic histories in their Indo-Pacific range. *MOLECULAR ECOLOGY*, 23(21):5193–5207, NOV 2014b. ISSN 0962-1083. doi: {10.1111/mec.12936}.

John Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, 2005.

James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

GA Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975.

Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931.

Saliha Zenboudji, Marc Cheylan, Veronique Arnal, Albert Bertolero, Raphael Leblois, Guillelme Astruc, Giorgio Bertorelle, Joan Ll. Pretus, Mario Lo Valvo, Giuseppe Sotgiu, and Claudine Montgelard. Conservation of the endangered Mediterranean tortoise Testudo hermanni hermanni: The contribution of population genetics and historical demography. *BIOLOGICAL CONSERVATION*, 195:279–291, MAR 2016. ISSN 0006-3207. doi: {10.1016/j.biocon.2016.01.007}.

Honghua Zhao, Daniel S. Nettleton, and Jack C. M. Dekkers. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. *Genetical research*, 89(01): 1–6, 2007.