Bayesian statistical inference for intractable likelihood models

Louis Raynal

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique

École doctorale I2S - Information, Structures, Systèmes

Unité de recherche UMR 5149 - IMAG - Institut Montpelliérain Alexander Grothendieck

Inférence statistique bayésienne pour les modélisations donnant lieu à un calcul de vraisemblance impossible

Présentée par Louis RAYNAL
Le 10 septembre 2019

Sous la direction de Jean-Michel MARIN

Devant le jury composé de

Mark BEAUMONT	Professeur	University of Bristol	Rapporteur
Michael BLUM	Directeur de recherche	Université de Grenoble Alpes	Rapporteur
Alice CLEYNEN	Chargée de recherche	Université de Montpellier	Examinatrice
Raphael LEBLOIS	Chargé de recherche	CBGP - INRA Montpellier	Examinateur
Jean-Michel MARIN	Professeur	Université de Montpellier	Directeur de thèse
Anne PHILIPPE	Professeure	Université de Nantes	Présidente du jury

UNIVERSITÉ DE MONTPELLIER

Remerciements

Je remercie mon directeur de thèse, Jean-Michel, pour son encadrement lors de ces trois années, mais également pour les nombreux moments de rire et de travail ! Un grand merci pour avoir réussi à me garder motivé même quand rien n'allait ! C'est un plaisir d'avoir travaillé avec vous !

Merci à Mark Beaumont et Michael Blum pour avoir accepté de rapporter ce manuscrit, ainsi qu'à Anne Philippe, Alice Cleynen et Raphael Leblois pour avoir accepté de faire partie du jury.

Merci à mes nombreux collaborateurs : Arnaud qui a attisé mon intérêt pour la génétique des populations, ainsi que Marie-Pierre. J'ai beaucoup apprécié nos nombreux échanges qui ont été très bénéfiques. Erwan et Gérard pour les nombreuses tentatives sur les forêts locales. Mathieu, Christian et surtout Pierre, sans qui je n'aurai jamais commencé cette thèse si tu ne m'avais pas proposé ce fameux stage un été, portant sur l'ABC. Finalement, à nouveau Alice, merci pour ton soutien constant mais aussi et surtout lors de la dernière ligne droite !

Un merci aux nombreux doctorants et post-doctorants qui ont contribué à la bonne ambiance du laboratoire, sans prise de tête : Matthieu, Amina, Robert, Pascal... pour n'en citer que quelques uns, mais aussi aux nombreux chercheurs/ingénieurs, notamment Jean-Noël et Sophie lors des comités de suivi individuel, ainsi que François-David avec qui j'ai beaucoup aimé échanger et Elodie pour cette très bonne expérience d'enseignement. Merci à Nathalie, Sophie, Laurence, Carmela, Eric et Bernadette pour leur travail administratif. Mention spéciale à Natalie, bien que tu ne sois pas là depuis très longtemps j'ai beaucoup aimé discuter avec toi.

Ismaël, depuis le temps qu'on se connaît, tu as toujours eu pleinement confiance en moi ! Merci beaucoup ! Merci aussi à Alexia, pour les nombreuses soirées et moments de fun, bonne continuation à toi ! Merci à Matthieu, Lily et Florent pour ces années de master et d'avoir maintenu le contact ensuite, ces nombreux moments de retrouvailles et de rires. Je vous souhaite le meilleur pour la suite !

Merci à Diana, Laure et Elena de l'atelier des pelotes, pour tous ces ateliers et activités qui m'ont changés les idées, ainsi qu'à Marie-Carmen, Christine, Françoise et Kaori avec qui j'ai bien ri en faisant du crochet et en dégustant des smoothies.

Évidement, un immense merci à ma famille, pour votre soutien constant, maman et papa, Adèle, Léon, Jules, Pauline, Delphine, Bernard, Brigitte, Martine, Géralde et Jaqueline. Une pensée pour Solange et Kenneth. Vous avez tous toujours été à

mes côtés (même si ce que je fais doit vous paraître un peu obscur haha), merci encore !

Finalement, merci aux co-bureaux (source très importante de motivation). Mario pour sa bonne humeur et son humour décalé, Benjamin pour son accompagnement lors de toutes ces conf' sans qui elles auraient été bien moins drôles ! Un immense merci à Jocelyn, tu m'as accompagné pendant plus de trois ans, nous avons bien ri, tu m'as remonté le moral dans les pires moments et, dans les hauts comme dans les bas, ta présence a été essentielle au bon déroulement de la rédaction et de la thèse tout simplement !

Résumé long de la thèse

Les méthodes d'inférence statistique les plus courantes se basent bien souvent sur le calcul de la fonction de vraisemblance associée aux données observées. C'est par exemple le cas du maximum de vraisemblance ou encore de stratégies bayésiennes. Cependant, la complexité grandissante des modèles statistiques peut donner lieu à une vraisemblance incalculable. En effet, pour de nombreux domaines d'application, la modélisation de structures de données complexes nécessite l'utilisation de modèles pour lesquels l'expression de la vraisemblance n'est pas forcément disponible sous forme analytique, ou bien son évaluation n'est pas faisable en un temps raisonnable. Les méthodes classiques sont alors inutilisables.

Étant donné un modèle paramétré par un vecteur θ, il y a deux situations notables pour lesquelles la vraisemblance $f(\mathbf{y} \mid \theta)$ est non disponible.
La première survient lorsque le modèle contient des données non-observées, il s'agit de variables latentes $\mathbf{u}$. Il est alors nécessaire d'intégrer sur toutes ses valeurs possibles pour obtenir $f(\mathbf{y} \mid \theta)$, ce qui peut être infaisable lorsque $\mathbf{u}$ est de grande dimension. Ce cas arrive par exemple pour des problèmes de génétique des populations, sous le modèle de coalescence de Kingman (1982a), où les données génétiques au temps présent dépendent de l'histoire non-observée des individus échantillonnés. La deuxième situation que l'on peut évoquer est lorsque la vraisemblance $f(\mathbf{y} \mid \theta)$ nécessite le calcul d'une constante de normalisation qui est incalculable. En effet, la grande dimension de $\mathbf{y}$ peut rendre son calcul difficile car il est nécessaire d'intégrer/sommer sur toutes ses valeurs possibles. Cela arrive par exemple lorsque des données de réseaux sont modélisées par un modèle de graphe aléatoire exponentiel, ou encore pour les champs aléatoires de Markov. Durant cette thèse, nous nous concentrons sur le premier cas de figure, avec des applications à la génétique des populations.

De nombreuses stratégies ont été développées en réponse au problème de vraisemblance incalculable. On retrouve par exemple l'algorithme Espérance-Maximisation (EM, Dempster et al., 1977), les méthodes de vraisemblances composites (Lindsay, 1988 ; Varin et al., 2011), d'inférence indirectes (Gourieroux et al., 1993 ; Smith, 1993) ou encore les approches de Calcul Bayésien Approché (ABC, Beaumont, Zhang et al., 2002 ; Marin, Pudlo, Robert et al., 2012 ; Sisson, Fan et Beaumont, 2018). Cette thèse s'intéresse à cette dernière solution qui est très flexible car elle contourne le calcul de la vraisemblance en utilisant uniquement des simulations selon le modèle considéré. En effet, bien que l'expression de $f(\mathbf{y} \mid \theta)$ soit non calculable

ou difficile à approximer, il est souvent plus aisé de générer des données selon le modèle, conditionnellement à des valeurs de paramètres fixées. Pour cette raison, un tel modèle est aussi mentionné sous le terme "génératif".

C'est à travers les travaux de Tavaré et al. (1997), Weiss et von Haeseler (1998), Pritchard et al. (1999) et Beaumont, Zhang et al. (2002) que l'ABC est apparu. Initialement développé pour des problèmes de génétique des populations, il est maintenant utilisé dans de nombreux autres domaines tels que l'épidémiologie (Rodrigues, Francis et al., 2018), l'écologie (Fasiolo et Wood, 2018), la médecine (Fan et al., 2018). L'ABC se base sur la comparaison entre données simulées et observées. Dans un premier temps, un paramètre θ' est simulé selon une loi *a priori*. Conditionnellement à θ', une pseudo-donnée $\mathbf{x}$ est générée selon le modèle génératif, et comparée à l'observation $\mathbf{y}$. Si la distance entre $\mathbf{x}$ et $\mathbf{y}$ est assez faible, le paramètre θ' est conservé. Ce processus de simulation-comparaison est répété un grand nombre de fois. Alors que l'objectif de nombreuses stratégies bayésiennes est d'obtenir un échantillon selon la loi *a posteriori* $\pi(\theta \mid \mathbf{y})$, l'ABC génère un échantillon selon une approximation de celle-ci, et ce pour deux raisons. Tout d'abord, pour faciliter leurs comparaisons, les données sont projetées dans un espace de plus faible dimension au travers d'un ensemble de résumés statistiques, qui ne sont pas forcément exhaustifs. Ensuite, un seuil de tolérance est utilisé pour spécifier si oui ou non la distance entre les résumés simulés et observés est suffisamment faible. Ainsi, la majorité des méthodes ABC dépend du choix judicieux d'une distance, d'un seuil de similarité et d'un ensemble de résumés statistiques.

Une grande partie de la littérature s'intéresse à ces choix, en particulier concernant les résumés statistiques qui doivent être peu nombreux tout en étant suffisamment informatifs, dans le but d'éviter le fléau de la dimension. De nombreux efforts ont été fournis face à cette difficulté (voir par exemple Blum, Nunes et al., 2013 ; Prangle, 2018). De plus, des approches ABC ont été développées dans le but de pallier ces problèmes de réglage. On retrouve notamment les méthodes d'ajustement par régression où les valeurs de paramètres simulés sont corrigées pour réduire l'influence du seuil d'acceptation (Beaumont, Zhang et al., 2002 ; Blum et François, 2010), ainsi que des approches ABC itératives pour pouvoir considérer des seuils plus faibles comparées à la version de base, sans altérer leurs performances (Marjoram et al., 2003 ; Beaumont, Cornuet et al., 2009 ; Sisson, Fan et Tanaka, 2009 ; Del Moral et al., 2012).

Des développements récents se basent sur l'association de simulations selon le modèle génératif et l'entraînement d'algorithmes d'apprentissage automatique. On retrouve des stratégies utilisant des réseaux de neurones profonds qui présentent des qualités prédictives notables (Sheehan et Song, 2016 ; Papamakarios et Murray, 2016 ; Lueckmann et al., 2017). Dans cette idée, pour des problèmes de choix de modèle, une approche intéressante introduite par Pudlo et al. (2016) est basée sur les forêts aléatoires de Breiman (2001).

L'algorithme des forêts aléatoires est un outil d'apprentissage automatique supervisé, constitué d'un ensemble d'arbres CART (Breiman, Friedman et al., 1984) construits à partir d'échantillons *bootstrap* dont l'entraînement est randomisé. Une forêt aléatoire présente de nombreux atouts. Tout d'abord, sa performance est assez remarquable en ne nécessitant quasiment aucun réglage (voir par exemple Fernández-Delgado et al., 2014). De plus, bien que disponibles sous certaines sim-

plifications, les forêts bénéficient de résultats de convergence (Biau, 2012 ; Scornet et al., 2015), soulignant le fait que la vitesse de convergence des estimateurs associés dépend du nombre de variables explicatives pertinentes et non du nombre total de covariables. Une forêt se montre ainsi très robuste face aux variables de bruit et tend à utiliser lors de sa construction majoritairement les variables informatives. De plus, elle fournit des outils interprétatifs comme des mesures d'importance des variables. Enfin, il est possible d'obtenir une mesure de la qualité de la forêt en termes de prédictions, sans forcément nécessiter de données de test additionnelles, en utilisant l'information *out-of-bag*. Ces avantages justifient bien l'utilisation de forêts en association avec des tables d'entraînement simulées par ABC. En effet, elles évitent une étape de présélection des résumés statistiques, réduisent l'impact du fléau de la dimension, et il n'y a ni distance, ni seuil de tolérance à spécifier. Une partie de ce manuscrit est ainsi consacrée à l'adaptation de l'approche de Pudlo et al. (2016) pour des problèmes d'inférence de paramètres. Dans la suite, ce genre de stratégies mêlant ABC et forêts aléatoires est dénoté par ABC-RF (pour *approximate Bayesian computation - random forest*).

Nous proposons ainsi une approche ABC basée sur les forêts aléatoires dans un contexte de régression. Dans un premier temps, nous générons une table de référence de taille N par ABC, c'est-à-dire un ensemble de paramètres simulés selon la loi *a priori*, pour lesquels des données sont produites via le modèle génératif puis résumées. Nous créons ainsi artificiellement une table d'entraînement constituée de paramètres θ et de résumés statistiques associés. Le principe de notre approche est d'entraîner une forêt en régression par dimension de l'espace des paramètres, ainsi nous ne visons pas toute la distribution *a posteriori* mais seulement les dimensions d'intérêt. Par exemple, si θ_j nous intéresse, une forêt en régression va être construite avec comme variable réponse le vecteur des valeurs simulées $\theta_j^{(1)}, \ldots, \theta_j^{(N)}$, et comme variables explicatives les résumés statistiques. Inspirés par les travaux de Meinshausen (2006), à partir de cette forêt, nous déduisons un vecteur de poids associé à $\mathbf{y}$, $\mathbf{w_y} = (w_{\mathbf{y}}^{(1)}, \ldots, w_{\mathbf{y}}^{(N)})$, et proposons différentes stratégies dans le but d'estimer les quantités *a posteriori* telles que l'espérance, la variance et des quantiles. Nous nous intéressons de plus à l'estimation de covariances *a posteriori* grâce à l'utilisation de plusieurs forêts et des informations *out-of-bag* résultantes. Selon les cibles *a posteriori* à estimer, les stratégies sont les suivantes.

- **Espérance** : somme pondérée des θ_j par $\mathbf{w_y}$.
- **Variance** : utilisation de $\mathbf{w_y}$ et des erreurs *out-of-bag* au carré.
- **Quantiles** : estimer la fonction de répartition grâce aux poids $\mathbf{w_y}$.
- **Covariance** : construction d'une nouvelle forêt utilisant le produit des erreurs *out-of-bag*.

Sur des exemples simulés ou réels, comparée à des approches classiques ABC (par ajustement ou itératives) ainsi qu'aux derniers développements adaptatifs, notre approche montre des résultats convaincants pour un nombre de simulations égal. On observe un bon compromis entre la qualité des estimations de paramètres et celle des intervalles de crédibilités, ainsi qu'une insensibilité relative à l'ajout de variables de bruits, ce qui permet l'incorporation de nombreux résumés statistiques

sans avoir à les présélectionner. Les résultats pour l'estimation de covariances sont encourageants mais nécessitent de plus amples améliorations. Ces stratégies pour l'inférence de paramètres ont été implémentées dans la libraire R `abcrf`.

Les méthodologies ABC-RF pour le choix de modèle et l'inférence de paramètres (Pudlo et al., 2016 ; Raynal et al., 2019, respectivement), et de manière plus générale les forêts sous-jacentes, permettent d'obtenir des prédictions dans tout l'espace des variables explicatives, peu importe la donnée à prédire. Dans le contexte ABC, nous disposons généralement d'une seule donnée à prédire, la donnée observée, et les approches ABC tirent particulièrement bien profit d'approches dites locales, c'est-à-dire qui utilisent l'information fournie par cette observation. C'est le cas de l'algorithme ABC de base qui peut être vu comme une méthode de k plus proches voisins (Biau, Cérou et al., 2015), ou encore des méthodes ABC par ajustements qui peuvent tirer profit de régression locales (Beaumont, Zhang et al., 2002). Il est ainsi naturel de se demander s'il est possible de construire une forêt qui tire profit de la donnée observée pour obtenir une meilleure prédiction associée à celle-ci. Cela pourrait être particulièrement bénéfique pour les récents développements par ABC-RF, ou même mener à d'éventuelles approches de réajustement par forêts. En nous écartant du cadre de l'ABC, nous nous intéressons donc au problème plus général de la construction de ce type de forêts, que l'on mentionne comme "locales", dans un contexte de classification. Nous passons en revue les méthodes d'arbres/forêts locaux existants, et proposons de nouvelles approches.

Les forêts aléatoires présentent plusieurs particularités sur lesquelles il est possible d'agir dans le but de prendre en compte la donnée observée. La plus naturelle concerne le critère de coupure des arbres permettant de partitionner l'espace des covariables. Dans un contexte de classification, nous proposons ainsi d'utiliser un critère prenant en compte l'observation cible au travers de noyaux (unidimensionnels ou multidimensionnels), donnant ainsi plus de poids aux données voisines. Rappelons que chaque arbre est entraîné sur un échantillon *bootstrap*, ce qui nous donne un deuxième champs d'action possible, qui a été abordé notamment au travers des travaux de Fulton et al. (1996) ou encore Xu et al. (2016). L'idée est ainsi d'effectuer un tirage des données d'entraînement (avec remise ou non) pondéré par leur proximité avec l'observation d'intérêt. Alors que ces approches se basent sur la pondération d'individus, nous en proposons une se basant sur la pondération des variables explicatives. Grâce aux résultats d'une première forêt aléatoire, nous déterminons une importance des variables propre à l'observation, que nous utilisons lors de la construction des arbres. Enfin, une forêt en classification agrège un ensemble d'arbres et utilise comme prédiction finale les votes fournis par chacun. Une dernière optique est donc d'agir sur ces votes, dans le but de donner plus de poids aux arbres capables de prédire correctement les données similaires à celle qui nous intéresse (voir par exemple Robnik-Šikonja, 2004).

Grâce à plusieurs exemples pour lesquels des stratégies locales peuvent présenter un avantage face à l'algorithme classique de Breiman, nous comparons un ensemble d'adaptations locales. Bien que notre proposition d'approche par importance de variables locales présente de bons résultats dans certains cas, il n'y a pas de consensus qui se dégage en terme de qualité de prédiction quant à la meilleure approche. De plus, alors que ces nouvelles stratégies dépendent bien souvent d'un paramètre

additionnel de réglage dirigeant leur caractère local (nombre de voisins, fenêtres de noyaux...), les forêts classiques restent très difficiles à surpasser.

Cette thèse s'intéresse majoritairement à des applications en génétique des populations. À partir d'informations génétiques obtenues au temps présent, un objectif est de reconstruire l'histoire de populations naturelles, dans le but de mieux comprendre les forces évolutives expliquant la présente diversité. Cela peut permettre, par exemple, de lutter plus efficacement contre certaines espèces invasives et ainsi faciliter la préservation d'autres espèces.

Un modèle fréquemment utilisé est le coalescent de Kingman (1982a). Il permet de reconstruire l'histoire d'échantillons, du présent vers le passé, en générant dans un premier temps une généalogie sur laquelle sont ensuite ajoutées des mutations. Lorsqu'un scénario évolutif liant des populations est considéré, c'est-à-dire qu'un ensemble d'événements démographiques doit être respecté pour refléter la réalité, le coalescent s'adapte facilement. Le coalescent est un parfait exemple de modèle pour lequel la vraisemblance est incalculable à cause des variables latentes. En effet, l'histoire de l'échantillon est non-observée, la vraisemblance s'écrit alors

$$f(\mathbf{y} \mid \theta) = \int f(\mathbf{y}, \mathcal{H} \mid \theta) d\mathcal{H}.$$

En d'autres termes, son calcul nécessite d'intégrer sur toutes les histoires possibles $\mathcal{H}$ menant à l'observation $\mathbf{y}$, ce qui est difficilement faisable en grande dimension car un très grand nombre d'histoires est envisageable. Cependant, il est très facile de générer des histoires et données possibles selon ce modèle, ainsi les méthodes ABC sont particulièrement bien appropriées pour traiter des données de génétique des populations sous un modèle de coalescent. Il existe notamment un large choix de programmes de simulation de données génétiques tels que `ms` (Hudson, 2002) ou encore DIYABC (Cornuet, Pudlo et al., 2014) que nous employons en plusieurs occasions.

Deux collaborations avec les biologistes Arnaud Estoup et Marie-Pierre Chapuis, ont mené à l'amélioration des méthodes ABC-RF pour le choix de modèles et l'inférence de paramètres, et par la même occasion de la librairie `R abcrf`. Les deux nouveautés introduites respectivement dans Estoup, Raynal et al. (2018) et Chapuis, Raynal et al. (2019) sont mentionnées ci-dessous.

Face à un problème de choix de modèles, plus particulièrement de sélection de scénarios évolutifs en génétique des populations, il est parfois difficile de discerner avec forte certitude si un évènement démographique (par exemple un changement de taille de population, un mélange de populations,...) est important ou non. Pour pallier ce problème, nous proposons de former des groupes de scénarios selon la présence ou non de l'évènement en question. Le choix de modèle par ABC-RF est ainsi effectué sur ces groupements. Une telle approche permet d'évaluer la difficulté avec laquelle est identifié chaque évènement et ainsi de mieux comprendre le scénario final sélectionné.

Pour l'estimation des paramètres, les forêts permettent d'obtenir des erreurs *out-of-bag*, donnant ainsi une mesure globale de sa qualité de prédiction (sur tout l'espace des covariables). Cependant cette mesure d'erreur n'est pas liée à l'observation d'intérêt, alors que la qualité de prédiction peut dépendre de la zone où elle se

situe. Nous proposons ainsi des mesures d'erreurs *a posteriori*, évaluées exactement au point d'intérêt, c'est-à-dire conditionnellement à l'observation. Nous calculons ces erreurs grâce aux forêts aléatoires, en faisant usage des poids associés à $\mathbf{y}$ déjà évoqués dans le développement ABC-RF pour l'inférence de paramètres.

Ces deux améliorations ont bénéficié à deux études de cas. La première porte sur un problème de choix de modèles pour sélectionner le meilleur scénario évolutif liant des populations Pygmées d'Afrique à leurs voisines non-Pygmées. La seconde étude s'intéresse à la reconstruction de l'histoire évolutive de deux sous-espèces de criquets pèlerins au nord et sud de l'Afrique (*Schistocerca gregaria*). Cette analyse tire profit de toutes les méthodes ABC-RF développées jusqu'à présent, pour le choix de modèle et l'inférence de paramètres, et emploie les deux améliorations mentionnées ci-dessus. Cette analyse souligne un temps de divergence très récent entre les deux sous-espèces et une colonisation du nord vers le sud par un faible nombre d'individus. Cela est mis en relation avec des évènements passés connus pouvant l'expliquer, notamment des migrations de très longue distance ou encore l'expansion de certaines populations humaines et avec elles de terrains propices à la vie du criquet.

Plan de la thèse

Cette thèse porte ainsi sur le développement de méthodologies ABC par forêts aléatoires, avec des applications à la génétique des populations. Nous nous intéressons de plus à l'élaboration d'algorithmes de forêts aléatoires locales dans un contexte de classification.

Les trois premiers chapitres sont consacrés à des rappels.
Le Chapitre 1 donne un aperçu de méthodes existantes face au problème de vraisemblance incalculable, et plus particulièrement des méthodes ABC. Le Chapitre 2 décrit l'algorithme des forêts aléatoires de Breiman ainsi que les avantages qu'elles peuvent présenter pour l'ABC. Le Chapitre 3 expose les bases du coalescent et présente des résumés statistiques communément utilisés.

Les trois derniers chapitres sont les contributions de cette thèse.
Le Chapitre 4 présente la méthodologie ABC-RF pour l'estimation des paramètres ainsi que de nombreuses comparaisons aux approches existantes. Le Chapitre 5 étudie le potentiel d'approches locales par forêts aléatoires pour la classification. Enfin, le Chapitre 6 introduit des améliorations pour les approches ABC par forêts aléatoires et expose deux études de cas, en génétique des populations, les mettant en oeuvre.

Contents

Introduction

Most statistical methods rely on the computation of the likelihood. This is the case of maximum likelihood analyses or some Bayesian strategies for example. However, as statistical models and data structures get increasingly complex, managing the likelihood function becomes a more and more frequent issue. In various application fields, we now face many fully parametric situations where the likelihood function cannot be computed in a reasonable time or simply is unavailable analytically. As a result, while the corresponding parametric model is well-defined, with unknown parameter θ, standard solutions based on the density function $f(\mathbf{y} \mid \theta)$ are prohibitive to implement.

There are two common situations for which the likelihood is intractable.
The first one occurs when the model contains some unobserved information, a.k.a. latent variables $\mathbf{u}$. Recovering the likelihood $f(\mathbf{y} \mid \theta)$ thus implies integrating over all possible $\mathbf{u}$ values, but this may not be possible, especially when $\mathbf{u}$ is of high dimension. This is typically the case for population genetics problems, when the Kingman (1982a)'s coalescent model is used, and genetic data are observed at the present time.
The second situation arises when the likelihood $f(\mathbf{y} \mid \theta)$ is expressed depending on a normalising constant which is difficult to calculate. Indeed, the high dimensionality of $\mathbf{y}$ might prevent its evaluation as it requires to integrate over all possible values of $\mathbf{y}$. It happens for many data structures, of which Markov random fields and network data under an exponential random graph model are good examples.
In both cases, handling a high dimensional integral is at the core of the likelihood intractability. This thesis focuses on the first situation, and is interested in applications to population genetics.

To bypass this hurdle, the last decades witnessed different inferential strategies, among which composite likelihoods (Lindsay, 1988; Varin et al., 2011), indirect inference (Gourieroux et al., 1993; Smith, 1993) and likelihood-free methods such as approximate Bayesian computation (ABC, Beaumont, Zhang et al., 2002; Marin, Pudlo, Robert et al., 2012; Sisson, Fan and Beaumont, 2018) became popular options. Here, we focus on ABC approaches, which are very flexible as they circumvent the likelihood computation by solely relying on simulations generated from the model. Indeed, even though the likelihood is unavailable or cannot be evaluated in a reasonable time, it is often simpler to simulate data according to the model for a given parameter value. For this reason such a model is also mentioned as

"generative" model.

ABC methods appeared thanks to the work of Tavaré et al. (1997), Weiss and von Haeseler (1998), Pritchard et al. (1999) and Beaumont, Zhang et al. (2002). Initially for population genetics problems, it is now widely used in many application fields such as epidemiology (Rodrigues, Francis et al., 2018), ecology (Fasiolo and Wood, 2018), nuclear imaging (Fan et al., 2018). Its principle is very simple and based on comparisons between observed and simulated data. It consists in generating a simulated parameter value θ' from a prior distribution. Then, conditionally on θ', an artificial data $\mathbf{x}$ is simulated according to the generative model, which is compared to the observation of interest $\mathbf{y}$. Provided that the distance between $\mathbf{x}$ and $\mathbf{y}$ is low enough, the initial θ' value is stored. This simulation-comparison process is repeated a large number of times. While the objective of Bayesian strategies is often to recover a sample from the posterior distribution $\pi(\theta \mid \mathbf{y})$, the basic ABC approach obtains a sample from an approximation of it, for two reasons. First, raw data are projected in a lower dimensional space thanks to a set of summary statistics, which are not necessarily sufficient. Then, a threshold is used to indicate whether or not the distance between the summarised simulated and observed data is small enough. Thus, most ABC approaches depend heavily on the careful choice of a distance, a similarity threshold and a set of summary statistics.

A large part of the literature addresses these choices, especially concerning the set of summary statistics that should be few in number but still informative to avoid the curse of dimensionality. This difficulty can be handled thanks to selection procedures in a predefined set of statistics (e.g. Joyce and Marjoram, 2008; Nunes and Balding, 2010), using projection techniques (e.g. Wegmann et al., 2009; Fearnhead and Prangle, 2012) or even by means of regularisation approaches (e.g. Blum, Nunes et al., 2013; Saulnier et al., 2017). Blum, Nunes et al. (2013) and Prangle (2018) review strategies regarding this subject. Moreover, many ABC approaches were developed to lower the influence of this tuning issue. Let us mention the regression adjustment methods that aim at correcting the discrepancy between the simulated and observed data, lowering the threshold influence (Beaumont, Zhang et al., 2002; Blum and François, 2010), or again some iterative schemes allow the use of a smaller value without compromising the algorithm efficiency (Marjoram et al., 2003; Beaumont, Cornuet et al., 2009; Del Moral et al., 2012).

Some recent developments are based on the association of ABC simulations and machine learning algorithms. Neural networks are being used more and more, often under the term "deep learning", and when properly calibrated they can achieve great performance. It motivates their use with ABC simulations (Sheehan and Song, 2016; Mondal et al., 2019), and more sophisticated schemes are possible (Papamakarios and Murray, 2016; Lueckmann et al., 2017). In this framework, for model choice problems a recent combination based on Breiman (2001)'s random forests has been proposed by Pudlo et al. (2016), that avoids some tuning parameters mentioned above thanks to the random forest assets.

A random forest is a supervised ensemble learning algorithm made of trees (CARTs, Breiman, Friedman et al., 1984) whose construction is randomised and performed on bootstrap samples. It presents several advantages that could benefit to ABC strategies. Firstly, its performance is quite good when no tuning is performed (e.g. Fernández-Delgado et al., 2014). Moreover, while available under

simplifying assumptions, some convergence results exist (Biau, 2012; Scornet et al., 2015) highlighting convergence rates for the associated estimators depending only on the number of relevant explanatory variables. Thus, a forest is robust toward noise variables and aims at using covariates useful for the problem at hand during its construction. It provides interesting interpretative tools as variable importance measures that can be computed during the training phase. Finally, its quality in terms of prediction accuracy can be assessed without requiring any additional test data, thanks to the so called out-of-bag errors. These perks motivate the use of forests trained on reference tables generated with the generative model. Indeed, a preliminary selection of a small number of summary statistics is not needed, they reduce the influence of the curse of dimensionality, and no distance or acceptance threshold needs to be specified.

ABC-RF for parameter inference

A large part of this manuscript is devoted to extending the model choice approach of Pudlo et al. (2016) to the parameter inference framework. In the following this type of strategy combining ABC and random forests is mentioned as ABC-RF (for approximate Bayesian computation - random forest). We propose to generate a reference table of large size, i.e. a set of parameters generated from a prior distribution for which data are simulated from the generative model and summarised. This table is used as an artificial training set for some regression random forests, aiming at learning the relationship between parameters and summary statistics. Because the whole posterior distribution of θ is difficult to approximate, while only some few parameters might be of interest, our proposal consists in training a regression forest per dimension of the parameter space to deduce posterior quantities of interest. Based on the work of Meinshausen (2006), we propose different strategies to approximate posterior expectation, variance, quantiles and we make use of additional forests when covariances need to be estimated. On simulated and applied examples, we compare our ABC-RF approach with a set of earlier ABC methods including regression adjustment and last adaptive developments. Our proposal is described in Raynal et al. (2019) and implemented in the R package `abcrf`.

Local random forest

The ABC-RF methodologies for model choice and parameter inference (Pudlo et al., 2016; Raynal et al., 2019, respectively), and more generally the underlying random forests, are an eager algorithm in the sense that predictions on the whole space of explanatory variables can be recovered, no matter the data we would like to predict. This is done through a separate training and prediction phase. However, in the ABC framework, in most cases only one observation $\mathbf{y}$ is of interest for prediction, and ABC approaches use this data through local techniques. For example, this is the case of the basic ABC algorithm that can be perceived as a k-nearest neighbours method (Biau, Cérou et al., 2015), or regression adjustment ABC that takes profit

of local regression (Beaumont, Zhang et al., 2002). This is why, in the second part of this thesis, we are interested in random forests strategies that take into account the additional information provided by this observed data $\mathbf{y}$. This could be useful to improve the ABC-RF strategies but also lead to regression adjustment techniques by means of forest. We therefore deviate from the ABC framework, for the more general development of some so-called "local" forests. For classification problems we review some existing strategies and propose new ones.

A random forest presents different characteristics on which we can act to take into account the observed data. The most intuitive is the splitting criterion involved in the tree construction to partition the predictor space. In the classification framework, we take into account the observation thanks to kernels (unidimensional or multidimensional), to propose a weighted Gini index that gives more weights to training data close $\mathbf{y}$. Moreover, each tree is built on a bootstrap sample, this is the second action field. This was tackled in the work of Fulton et al. (1996) or more recently Xu et al. (2016). The idea is to perform data sampling according to their proximity with the observed data. While this is based on weighting individuals, we propose a weighting scheme of the explanatory variables instead. Because some explanatory variables might be more or less important depending on the data to predict, we compute some local importance measures that we use in the covariate sampling scheme of the forest. Finally, a usual classification forest uses a set of tree votes to provide a prediction. The last strategy consists in weighting the tree votes depending on how well they are able to predict data similar to $\mathbf{y}$ (see e.g. Robnik-Šikonja, 2004).

Applications in population genetics

This manuscript mainly focuses on applications to population genetics. Thanks to genetic data obtained on individuals in natural populations at the present time, an objective is to reconstruct the evolutionary history of the sampled genes, in order to understand the evolutionary forces explaining the present diversity. It can, for instance, help to control invasive species.

A statistical model commonly used in population genetics is the Kingman's coalescent (Kingman, 1982a). It allows the reconstruction of the sample history, from the present to the past, by first generating a genealogy on which are added mutations in a second step. This construction is constrained by an evolutionary scenario, i.e. a set of time ordered demographic events linking populations, and the coalescent process can easily mimic most of them. This model is a perfect example of intractable likelihood due to latent variables. Indeed, the gene history is a latent variable and the density is hence expressed as

$$f(\mathbf{y} \mid \theta) = \int f(\mathbf{y}, \mathcal{H} \mid \theta) d\mathcal{H}.$$

In other terms, to recover the likelihood function we need to integrate over all possible histories $\mathcal{H}$ that can lead to the observation $\mathbf{y}$, this is rarely possible especially when $\mathbf{y}$ is high dimensional as too many histories can lead to $\mathbf{y}$. However, it is fairly

simple to generate data thanks to the coalescent model, making ABC methods well suited to analyse population genetics data. Many simulators exist to generate them, the most known is probably `ms` (Hudson, 2002). In our experiments we employ the DIYABC software (Cornuet, Pudlo et al., 2014) for this task.

Two collaborations with the population geneticists Arnaud Estoup and Marie-Pierre Chapuis led to some improvements on the ABC-RF methodologies for model choice and parameter inference, as well as the R package `abcrf`. Two novelties are introduced in Estoup, Raynal et al. (2018), Chapuis, Raynal et al. (2019) and in this manuscript.

In the model choice framework, more precisely to select the best evolutionary scenario using present genetic information, it can be difficult to disentangle with trust whether or not a demographic event (as a change of population size, an admixture between populations,...) is important or not. This statement becomes even more true when a high number of populations and events are considered. To better understand the scenario events, we propose to study groups of scenarios depending on the presence or the absence of some key events. It allows to identify and quantify which events are hard to discriminate and which ones are not.

For parameter estimation, the random forest algorithm provides some out-of-bag error measurements, giving insights regarding its predictive performance (over the entire covariate space). However, this type of error is not related to the observation we are interested in, when the prediction accuracy may depend on the area of the predictor space it is located in. For this reason, we propose some posterior measures of error computed conditionally on the observed data to predict, thanks to regression random forests.

These improvements are introduced in the papers Estoup, Raynal et al. (2018) and Chapuis, Raynal et al. (2019), on which two population genetics case studies are displayed. The first one is a model choice problem to select the best scenario linking African Pygmy populations to their non-Pygmy neighbours. The second study aims at reconstructing the past of two desert locust sub-species (*Schistocerca gregaria*) in North and South Africa. It takes profit of all the ABC-RF approaches, for model choice and parameter inference, as well as the model grouping strategy and the computation of posterior measures of error.

Thesis outline

This thesis focuses on the development of ABC random forest methodologies, with some applications to population genetics. Moreover, we are also interested in the creation of local random forests in a classification setting.

The first three chapters are reminders.
Chapter 1 gives an overview of existing inferential techniques when the likelihood is intractable due to latent variables, we especially insist on the ABC methodologies. Chapter 2 describes Breiman's random forest algorithm and its advantages for ABC. Chapter 3 presents the basis of the coalescent model and gives some common associated summary statistics.

The last three chapters are the contributions of this thesis.
Chapter 4 introduces the ABC-RF methodology for parameter inference as well as various comparisons with earlier ABC approaches. Chapter 5 studies how to construct local random forests in a classification setting. Finally, Chapter 6 proposes two improvements for the ABC-RF methodologies, which are successfully applied in two population genetics case studies.

Part I

Approximate Bayesian Computation, Random Forests and Population Genetics

Chapter 1

Likelihood-free inference

Contents

1.1 Introduction

We consider an observed data $\mathbf{y} \in \mathcal{Y}$, whose generation process can be described by a statistical model, $\mathcal{M}$, parameterized by an unknown vector $\theta \in \Theta \subseteq \mathbb{R}^p$. The likelihood of the observation is denoted $f(\mathbf{y} \mid \theta)$.

To find the best suited θ value for $\mathbf{y}$, the most known frequentist solution is the maximum likelihood estimator (MLE), which consists in finding θ that provides the maximal likelihood value for $\mathbf{y}$:

$$\hat{\theta}_{MLE} \in \arg\max_{\theta \in \Theta} f(\mathbf{y} \mid \theta).$$

Bayesian paradigm The Bayesian paradigm introduces prior knowledge on parameters thanks to a prior distribution denoted $\pi(\theta)$. This prior is updated by the observed data $\mathbf{y}$ through its likelihood $f(\mathbf{y} \mid \theta)$. A Bayesian analysis hence relies on the posterior distribution $\pi(\theta \mid \mathbf{y})$, derived from the Bayes' theorem

$$\pi(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta} \propto f(\mathbf{y} \mid \theta)\pi(\theta).$$

The posterior distribution is the probability distribution of θ given $\mathbf{y}$. It provides all the necessary information for parameter inference: posterior expectations, variances and credible sets. When its expression is explicitly available, an estimator of θ analogous to the MLE, is the maximum a posteriori (MAP):

$$\hat{\theta}_{MAP} \in \arg\max_{\theta \in \Theta} \pi(\theta \mid \mathbf{y}) = \arg\max_{\theta \in \Theta} f(\mathbf{y} \mid \theta)\pi(\theta).$$

Note that when the prior is uniform on the support of θ, the MAP is equal to the MLE. Moreover, the MAP is not associated to any loss function for continuous parameter values. When the L_2-loss is considered, the Bayesian estimator is the posterior expectation

$$\hat{\theta} = \mathbb{E}(\theta \mid \mathbf{y}) = \int \theta\pi(\theta \mid \mathbf{y})d\theta.$$

The exact expression of $\pi(\theta \mid \mathbf{y})$ is rarely available and some strategies to sample from the posterior are used. The class of Markov chain Monte Carlo methods (MCMC, Robert and Casella, 2005) are very common. The Metropolis-Hastings and the Gibbs algorithms are well known MCMC methods, to obtain realisations according to the posterior distribution. The former, for example, creates a Markov chain of $\theta^{(i)}$ values with stationary distribution $\pi(\theta \mid \mathbf{y})$, a new state $\theta^{(i+1)}$ is sampled according to a suited proposal distribution $q(\cdot \mid \theta^{(i)})$ and accepted with probability

$$\min\left\{1, \frac{f(\mathbf{y} \mid \theta^{(i+1)})}{f(\mathbf{y} \mid \theta^{(i)})} \frac{q(\theta^{(i)} \mid \theta^{(i+1)})}{q(\theta^{(i+1)} \mid \theta^{(i)})} \frac{\pi(\theta^{(i+1)})}{\pi(\theta^{(i)})}\right\}. \tag{1.1}$$

The acceptance rate is expressed as the ratio of the product between the target ($\pi(\theta \mid \mathbf{y})$ here) and the proposal.

Intractable likelihood

As statistical models and data structures get increasingly complex, managing the likelihood function becomes a more and more frequent issue. We now face many realistic fully parametric situations where the likelihood function cannot be computed in a reasonable time or simply is unavailable. In this case, the above mentioned strategies that rely on the likelihood are not directly appliable. There are two main reasons that can explain such intractability of the likelihood, we describe them below.

Latent variables The first reason is the presence of latent variables $\mathbf{u} \in \mathcal{U}$, so that the likelihood expression is deduced by integrating (or summing for discrete $\mathbf{u}$) the completed likelihood $f(\mathbf{y}, \mathbf{u} \mid \theta)$ over all possible $\mathbf{u}$ values:

$$f(\mathbf{y} \mid \theta) = \int f(\mathbf{y}, \mathbf{u} \mid \theta) d\mathbf{u}. \tag{1.2}$$

There are two types of missing information. The first one is the case of missing data, when random variables are unobserved even though they should be. For example when people refuse to answer questions in surveys. The second case is when important random variables for the model are never observed, referred as hidden/latent variables (or again auxiliary variables when they are introduced artificially). Examples are state-space models or coalescent models. We present the latter in Chapter 3, for which $\mathbf{u}$ is the unobserved gene trees that lead to the present observed genetic information $\mathbf{y}$. When $\mathbf{u}$ is of high dimension, the calculation of the integral (1.2) becomes difficult or even impossible.

Normalising constant The second source of intractability is the presence of an intractable normalising constant. In this case, the likelihood of the model is written as

$$f(\mathbf{y} \mid \theta) = \frac{\widetilde{f}(\mathbf{y} \mid \theta)}{Z(\theta)},$$

where $\widetilde{f}(\mathbf{y} \mid \theta)$ is the unnormalised likelihood, and $Z(\theta) = \int \widetilde{f}(\mathbf{y} \mid \theta) d\mathbf{y}$ its normalising constant. This value $Z(\theta)$ is hard to calculate when the dimension of $\mathbf{y}$ is very large. Hence, a very high dimensional integral is again at the core of the likelihood intractability. The Ising model, in Markov random field, presents such an issue, as well as the exponential random graph model for network data (Robins et al., 2007). In both of them, we have $Z(\theta) = \sum_{\mathbf{y}} \exp(\theta S(\mathbf{y}))$, where $S(\cdot)$ is a set of sufficient statistics. This sum operates over all possible values of $\mathbf{y}$. For networks with n nodes, this sum is made of $2^{n(n-1)/2}$ terms, which cannot be computed easily when n is large.

Some models can present a normalising constant and latent variables at the same time, this is a case of doubly-intractable likelihood. In the remaining we focus on the intractability issue due to latent variables, so that the likelihood is expressed as the integral of the completed $f(\mathbf{y}, \mathbf{u} \mid \theta)$ (Equation (1.2)).

Moreover, in this thesis, the main application field of interest is population genetics, when the model is the Kingman's coalescent, (see Chapter 3). In brief, the

data $\mathbf{y}$ is the genetic information observed at the present time, and the unobserved data $\mathbf{u}$ is the associated past gene history (denoted $\mathcal{H}$). In this case, θ are demographic, historical or mutational parameters, and performing simulations from $\mathbf{u} \mid \theta$ is feasible and presented in Chapter 3. Note that the likelihood of $\mathbf{y}$ is intractable but it is possible to evaluate $\mathbf{y} \mid \mathbf{u}, \theta$ (Felsenstein, 1981; RoyChoudhury, Felsenstein et al., 2008). In this application framework, it is common to have $\mathbf{y}$ that is a very high dimensional vector.

In general, we hence make the assumption that even though the likelihood is intractable, performing simulations according to the generative model is easier. In other terms, for a given θ value, artificial data $\mathbf{y}$ can be generated (which often rely on a preliminary simulation of $\mathbf{u} \mid \theta$).

Furthermore, a major problem we face is that the analytical expression of the density $f(\mathbf{u} \mid \mathbf{y}, \theta)$ is unknown (because it would be equivalent to the knowledge of the likelihood), and it is not possible to simulate from it either (at least for large $\mathbf{y}$). These constraints prevent the use of most strategies as we will see below.

We provide in the next section a non-exhaustive overview of solutions and we especially develop the case of Approximate Bayesian Computation techniques (ABC), at the core of this thesis. Next section is divided in three parts depending on how the parameter inference is performed: by either maximising an approximated likelihood, circumventing the calculation of likelihood, or relying on the posterior distribution.

1.2 Intractable likelihood solutions

1.2.1 Approximating the likelihood

A very natural strategy to perform parameter inference when the likelihood expression is not available consists in approximating the likelihood for a fixed value of θ, and then choosing the parameter value that provides the maximal approximated likelihood. We present below some existing strategies for this purpose.

1.2.1.1 Monte Carlo estimators

By examining the likelihood expression (Equation (1.2)), an intuitive approach would be to approximate the likelihood thanks to simulations, using a **Monte-Carlo estimator**. From the completed likelihood decomposition

$$f(\mathbf{y} \mid \theta) = \int f(\mathbf{y}, \mathbf{u} \mid \theta) d\mathbf{u} = \int f(\mathbf{y} \mid \mathbf{u}, \theta) f(\mathbf{u} \mid \theta) d\mathbf{u} = \mathbb{E}_{f(\mathbf{u}|\theta)} \left[f(\mathbf{y} \mid \mathbf{u}, \theta) \right],$$

we can independently generate N latent variables $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(N)}$ from $f(\cdot \mid \theta)$, and derive the estimated likelihood

$$\hat{f}(\mathbf{y} \mid \theta) = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{y} \mid \mathbf{u}^{(i)}, \theta).$$

Unfortunately, when $\mathbf{u}$ is of high dimension, simulating $\mathbf{u}^{(i)}$ without taking into account $\mathbf{y}$, leads in most cases to probability $f(\mathbf{y} \mid \mathbf{u}^{(i)}, \theta) = 0$. That is to say, the

observed data cannot be retrieved with the simulated configuration of the latent variable, $\mathbf{y}$ and $\mathbf{u}^{(i)}$ are incompatible. The corresponding estimator has very high variance, and the number of simulations N should be extremely large to counterbalance this issue, which is computational unfeasible in practice.

An alternative to prevent this issue, and approximating the likelihood, is to use importance sampling (IS). When considering an IS proposal distribution $q(\cdot \mid \theta)$ on $\mathbf{u}$, such that $\{\mathbf{u} : q(\mathbf{u} \mid \theta) > 0\} \supset \{\mathbf{u} : f(\mathbf{u} \mid \theta) > 0\}$, from

$$f(\mathbf{y} \mid \theta) = \int f(\mathbf{y}, \mathbf{u} \mid \theta) d\mathbf{u} = \int \frac{f(\mathbf{y}, \mathbf{u} \mid \theta)}{q(\mathbf{u} \mid \theta)} q(\mathbf{u} \mid \theta) d\mathbf{u} = \mathbb{E}_{q(\mathbf{u}|\theta)} \left[\frac{f(\mathbf{y}, \mathbf{u} \mid \theta)}{q(\mathbf{u} \mid \theta)} \right],$$

we have the importance sampling estimate

$$\hat{f}(\mathbf{y} \mid \theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\mathbf{y}, \mathbf{u}^{(i)} \mid \theta)}{q(\mathbf{u}^{(i)} \mid \theta)}, \tag{1.3}$$

where the $\mathbf{u}^{(i)}$ values are drawn from $q(\cdot \mid \theta)$. This distribution must be chosen so that the importance weights $f(\mathbf{y}, \mathbf{u}^{(i)} \mid \theta)/q(\mathbf{u}^{(i)} \mid \theta)$ have small variance, to achieve small variance estimator. Its optimal choice for the proposal is $f(\mathbf{u} \mid \mathbf{y}, \theta)$. Indeed, from basic calculations we have:

$$\frac{f(\mathbf{y}, \mathbf{u} \mid \theta)}{q(\mathbf{u} \mid \theta)} = \frac{f(\mathbf{y}, \mathbf{u} \mid \theta)}{f(\mathbf{u} \mid \mathbf{y}, \theta)} = \frac{f(\mathbf{u} \mid \mathbf{y}, \theta) f(\mathbf{y} \mid \theta)}{f(\mathbf{u} \mid \mathbf{y}, \theta)} = f(\mathbf{y} \mid \theta),$$

and the importance sampling estimator is a zero variance estimator. Unfortunately, $f(\mathbf{u} \mid \mathbf{y}, \theta)$ is unknown. Indeed, outside of the importance sampling aspect, if $f(\mathbf{u} \mid \mathbf{y}, \theta)$ is known in addition of $f(\mathbf{y}, \mathbf{u} \mid \theta)$, this is equivalent to the knowledge of the likelihood. Thus, the main drawback of this estimator lies in the choice of $q(\mathbf{u} \mid \theta)$, which is very challenging. Usually, it must be chosen with a dependence on $\mathbf{y}$, to always produce latent variables consistent with the observation, that is to say $f(\mathbf{y}, \mathbf{u}^{(i)} \mid \theta) \neq 0$. However, when the dimension of $\mathbf{u}$ grows, the variance of the weights and thus the importance sampling estimator can be very large, leading to poor likelihood approximations. For a large review of efficient importance sampling techniques we recommend Owen (2000) and Liu (2004).

For the coalescent setting, Griffiths and Tavaré (1994a) obtained a Monte Carlo approximation of the likelihood thanks to Markov chain recursion. This was then largely applied and extended by Griffiths and Tavaré (1994b,c), Nielsen (1997), Griffiths and Tavaré (1999) and Bahlo and Griffiths (2000). It was then pointed out by Felsenstein et al. (1999) that the Griffiths and Tavaré's approach is a version of IS with a particular proposal distribution. Importance sampling is thus largely used to approximate the likelihood of genetic samples, improvements and adaptation of the proposal is recurrent, for example, Stephens and Donnelly (2000) use as importance distribution an approximation of the optimal choice $f(\mathbf{u} \mid \mathbf{y}, \theta)$. See also De Iorio and Griffiths (2004a,b), De Iorio, Griffiths et al. (2005), Merle et al. (2017) and Rousset et al. (2018). These methods can be very computer intensive because of the large state space of the genealogies. Note that Cornuet, Marin et al. (2012) proposed the AMIS algorithm which reuses the importance weights to decrease the computational cost. In parallel of Griffiths and Tavaré's work, from a more MCMC perspective, Kuhner et al. (1995, 1998) propose to construct a Metropolis-Hastings

algorithm with stationary distribution $f(\mathbf{u} \mid \mathbf{y}, \theta)$, for a fixed θ value, in order to approximate a relative likelihood surface.

Moreover, adaptive/sequential importance sampling or population/sequential Monte-Carlo (PMC/SMC) are more evolved techniques to approximate the likelihood, but also computationally more expensive, see e.g. Cappé et al. (2004) and Del Moral et al. (2006) or the recent survey paper of Bugallo et al. (2017).

The Monte-Carlo estimators of the likelihood presented here are unbiased, in the sense that their expectations yield the exact likelihood. For this reason, they are often used in pair with MCMC algorithms, as we will see below.

1.2.1.2 Approximated model

Another strategy to estimate the likelihood is to relax the assumptions on the data. When the likelihood computation for the whole data is intractable, but still feasible on subsets, some **composite likelihood** (Lindsay, 1988) (or pseudolikelihood, Besag, 1974, 1975) approaches can be useful. The principle of composite likelihood is to split the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ in subsets, on which a likelihood (marginal or conditional) can be calculated for each of them. Then, these likelihoods are multiplied to form the composite likelihood, as if the subsets were independent. For example, the pairwise composite likelihood compute the marginal likelihood of pair data, and is hence expressed as

$$f_{pair}(\mathbf{y} \mid \theta) = \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} f(y_i, y_j \mid \theta).$$

For the coalescent model, the pairwise version has for example been used by Hudson (2001), and Fearnhead and Donnelly (2002) when non-overlapping segments of data are assumed independent. Note that the site/allele frequency spectrum computed on genomic data sets are often used in order to determine an associated composite likelihood estimator (see e.g. Gutenkunst et al., 2009; Excoffier, Dupanloup et al., 2013). In a more general context, a large amount of composite likelihood versions exists. See the review papers of Varin (2008) and Varin et al. (2011) for a large overview of these methods, and Larribe and Fearnhead (2011) for a special focus on genetic applications.

Another approach is the PAC-likelihood introduced by Li and Stephens (2003), it assumes that the likelihood of $\mathbf{y} = (y_1, \ldots, y_n)$ can be decomposed into a so called product of approximate conditionals (PAC). Hence, it uses the decomposition

$$f(\mathbf{y} \mid \theta) = f(y_1 \mid \theta) f(y_2 \mid y_1, \theta) \ldots f(y_N \mid y_1, \ldots, y_{N-1}, \theta).$$

All the conditional likelihoods on the right-hand side are then approximated and multiplied to produce the PAC version. Even though, this likelihood is approximated, it can achieve good performance on genetic problems (Smith and Fearnhead, 2005; RoyChoudhury and Stephens, 2007; Cornuet and Beaumont, 2007).

1.2.2 Frequentist solutions to circumvent the likelihood

Another strategy consists in bypassing the likelihood calculation, by using moments or auxiliary models.

1.2.2.1 Expectation-Maximisation

To find the MLE in presence of latent variables, the most popular approach is probably the expectation-maximisation algorithm (EM, Dempster et al., 1977). Its principle consists in building a sequence of parameter values $\theta^{(i)}$, so that the log-likelihood increases through the algorithm iterations. Given an initial parameter value, this algorithm is divided in two steps:

- The E-step (for expectation) consists in computing the conditional expectation

$$\mathbb{E}_{f(\mathbf{u}|\mathbf{y},\theta^{(i)})}\left[\log(f(\mathbf{y},\mathbf{u} \mid \theta))\right].$$

- The M-step (for maximisation) maximises the previous expectation with respect to θ, to obtain the next iteration $\theta^{(i+1)}$.

This is repeated until a convergence criterion is reached. Dempster et al. (1977) prove that each algorithm iteration increases the log-likelihood value. However, the final $\theta^{(i)}$ might be a local optimum. Moreover, this algorithm requires the knowledge of the distribution of $\mathbf{u}$ given $\mathbf{y}$ and $\theta^{(i)}$, at the current iteration. The calculation of the expectation can be replaced by stochastic approximations, see for example the SEM (Celeux and Diebolt, 1985), the SAEM (Delyon et al., 1999) or the MCEM (Wei and Tanner, 1990). For the coalescent model, $f(\mathbf{u} \mid \mathbf{y}, \theta)$ is unknown and performing simulations according to it is very difficult. Moreover, exploring efficiently the high dimensional latent space (the gene histories) is also a complicated task, for these reasons these EM algorithms are not appropriate.

1.2.2.2 Indirect inference

When the likelihood of the model is not available in closed form, a solution is to use a simpler one, with a tractable likelihood, able to capture information about the observation $\mathbf{y}$. This auxiliary model is parameterized by $\phi \in \Phi$, and $f_{aux}(\mathbf{y} \mid \phi)$ denotes its likelihood. In this section we present the indirect inference introduced by Gourieroux et al. (1993) and Smith (1993), which relies on such an artificial model. Its principle is to deduce the relationship between the original and auxiliary models, thanks to their respective parameters θ and ϕ. The idea is (1) to compute an estimate $\hat{\phi}_{\mathbf{y}}$ associated to $\mathbf{y}$, (2) to deduce the link between ϕ and θ, (3) in order to find the value of θ that produced $\hat{\phi}_{\mathbf{y}}$. When the link between ϕ and θ is known, this task is easy. Indeed, this relationship is described in the literature through the so called binding or mapping function: $\phi(\cdot)$, providing the value of ϕ given a value of θ. When this function is known and injective, its inverse, denoted $\theta(\cdot)$, provides the estimate $\hat{\theta} = \theta(\hat{\phi}_{\mathbf{y}})$ we are looking for.

The estimate $\hat{\phi}_{\mathbf{y}}$ is deduced thanks to an estimating function, $Q(\mathbf{y}; \phi)$, for example the log-likelihood of the auxiliary model. Hence

$$\hat{\phi}_{\mathbf{y}} \in \arg\max_{\phi \in \Phi} Q(\mathbf{y}; \phi).$$

When the binding function is unknown, we assume that simulations thanks to the original model can still be performed. Hence, for a given parameter value θ, indirect inference uses a set of simulated data of equal dimension to $\mathbf{y}$: $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, generated from the generative model $f(\cdot \mid \theta)$. The auxiliary parameter for each of these artificial data is deduced as

$$\hat{\phi}_{\mathbf{y}_i}(\theta) \in \arg\max_{\phi \in \Phi} Q(\mathbf{y}_i; \phi),$$

and an estimate of the binding function $\phi_N(\theta)$ is expressed as:

$$\phi_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \hat{\phi}_{\mathbf{y}_i}(\theta).$$

The final step to obtain an approximation of $\hat{\theta}$ proposed by Gourieroux et al. (1993) consists in selecting the parameter θ that generates the closest $\phi_N(\theta)$ value to $\hat{\phi}_{\mathbf{y}}$:

$$\hat{\theta}_N \in \arg\min_{\theta \in \Theta} \left\{ (\phi_N(\theta) - \hat{\phi}_{\mathbf{y}})^\top W (\phi_N(\theta) - \hat{\phi}_{\mathbf{y}}) \right\},$$

where the superscript $^\top$ designates the transpose function and W a positive definite weight matrix. Some discussions concerning the optimal choice of W can be found in Gourieroux et al. (1993) and Monfardini (1998).

Other alternatives exists for this final step, for example the simulated quasi-maximum likelihood estimator, proposed by Smith (1993), consists in maximising the auxiliary likelihood of the observation:

$$\hat{\theta}_N \in \arg\max_{\theta \in \Theta} f_{aux}(\mathbf{y} \mid \phi_N(\theta)).$$

More improvements followed, Gallant and Tauchen (1996), Monfardini (1998) and Heggland and Frigessi (2004) are examples. This auxiliary model strategy is also encountered in approximate Bayesian computation (Section 1.3.2), it justifies the extensive details provided in this part.

1.2.2.3 Moment based methods

Other methods relies on moment conditions $m(\mathbf{y} \mid \theta)$. These are statements involving the observed data and the parameters, with respect that $\mathbb{E}_{f(\mathbf{y}|\theta)}(m(\mathbf{y} \mid \theta^*)) = 0$, where θ^* is the parameter value we are searching for. In our likelihood based inference framework, as this expectation is not available in closed form, a Monte-Carlo estimator can be used for a given value of θ. The generalized method of moments (GMM, Hansen, 1982; Hansen and Singleton, 1982) searches for the parameter value that provides the approximated moments as close to zero as possible (in the sense of a weighted quadratic distance). When no closed form for the moment conditions

exists, difference between simulated and observed summary statistics can be used as moments. This simulated method of moments (SMM, McFadden, 1989) searches for the parameter value θ minimising the quadratic distance between the summaries. Let us mention that it can be seen as a special case of indirect inference (Heggland and Frigessi, 2004).

1.2.2.4 Gaussian assumptions

Finally, assumptions on data summary statistics can be considered, which lead to the **synthetic likelihood** approach proposed by Wood (2010). Its proposal does not try to replace the likelihood of $\mathbf{y}$, but the likelihood of a set of informative summary statistics $s = s(\mathbf{y})$. For a fixed θ value, the key assumption on which is built the synthetic likelihood is that, conditional to θ, the summary statistics s follow a multivariate normal distribution, with mean vector and covariance matrix parameters, respectively $\mu(\theta)$ and $\sigma(\theta)$. These parameters are estimated through N simulations from the generative model. We thus have

$$f_N(s \mid \theta) = \mathcal{N}(s \mid \mu_N(\theta), \sigma_N(\theta)),$$

where $\mathcal{N}$ is the density of the multivariate normal distribution and $\mu_N(\theta)$, $\sigma_N(\theta)$ the estimated parameters. Synthetic likelihood approaches can also be seen as a specific case of indirect inference, where the auxiliary model used is multivariate Gaussian, the estimating function its likelihood, and focusing on summary statistics instead of raw data. Some Bayesian improvements exists, see Drovandi, Pettitt and Lee (2015) and Price et al. (2018) for more information.

1.2.3 Bayesian inference

Finally, in the Bayesian framework, some strategies exist to deal with latent variables.

Metropolis-Hastings As we noticed earlier, a Metropolis-Hastings algorithm with stationary distribution $\pi(\theta \mid \mathbf{y})$ cannot be implemented as it requires the intractable likelihood expression twice. Toward this issue, many Metropolis-Hastings adaptations developed.

Even though the likelihood $f(\mathbf{y} \mid \theta)$ is unavailable, the completed likelihood $f(\mathbf{y}, \mathbf{u} \mid \theta)$ is often easier to compute. As a result, a version of Metropolis-Hastings whose target distribution is $\pi(\theta, \mathbf{u} \mid \mathbf{y})$ emerged (see Wilson and Balding, 1998; Beaumont, 1999, for coalescent context). Algorithm 1.1 presents this strategy, which yields MCMC samples from the joint $\pi(\theta, \mathbf{u} \mid \mathbf{y})$, discarding the unwanted dimensions results in marginalising over them, and an approximation of the posterior can be deduced.

The main drawback of Algorithm 1.1 is its low effectiveness when the joint distribution is hard to explore due to its high dimension, as well as the specification of the transition distribution. For the coalescent model, choosing a proposal on $\mathbf{u}$, (i.e. on the gene history), implies performing small modifications on the current

Algorithm 1.1 : Metropolis-Hastings with target $\pi(\theta, \mathbf{u} \mid \mathbf{y})$

Input : A starting value $\mathbf{u}^{(1)}$ and $\theta^{(1)}$, the number of iterations N

$i \leftarrow 1$;
while $i \leq N$ **do**
  Generate $\mathbf{u}', \theta'$ from a proposal distribution $q(\cdot, \cdot \mid \mathbf{u}^{(i)}, \theta^{(i)})$;
  Compute the acceptance ratio

$$A = \min \left\{ \frac{f(\mathbf{y}, \mathbf{u}' \mid \theta')}{f(\mathbf{y}, \mathbf{u}^{(i)} \mid \theta^{(i)})} \frac{q(\mathbf{u}^{(i)}, \theta^{(i)} \mid \mathbf{u}', \theta')}{q(\mathbf{u}', \theta' \mid \mathbf{u}^{(i)}, \theta^{(i)})} \frac{\pi(\theta')}{\pi(\theta^{(i)})} \right\};$$

  Set $\theta^{(i+1)} \leftarrow \theta'$ and $\mathbf{u}^{(i+1)} \leftarrow \mathbf{u}'$ with probability A,
   otherwise $\theta^{(i+1)} \leftarrow \theta^{(i)}$ and $\mathbf{u}^{(i+1)} \leftarrow \mathbf{u}^{(i)}$;
  $i \leftarrow i + 1$;
end

state. Too small changes involve slow exploration of the history space, and bigger changes results in low acceptance rate, (see Wilson and Balding, 1998; Beerli and Felsenstein, 1999; Felsenstein et al., 1999; Beaumont, 1999; Nielsen, 2000, for more information).

Another natural idea consists in using the usual Metropolis-Hastings algorithm with target $\pi(\theta \mid \mathbf{y})$, for which the intractable likelihoods are estimated and plugged-into each acceptance ratio (1.1). When the importance sampling estimator of the likelihood is used (Equation (1.3)), such a strategy is referred as Monte Carlo within Metropolis (O'Ryan et al., 1998; O'Neill et al., 2000; Chikhi et al., 2001; Berthier et al., 2002; Beaumont, 2003). For each iteration of the Metropolis-Hastings algorithm, two new independent likelihood estimates are required and the resulting samples do not come from $\pi(\theta \mid \mathbf{y})$, but from an approximation of it. Beaumont (2003) proposed to keep track of the likelihood estimates from one iteration to the next. Called the grouped independent Metropolis-Hastings, his algorithm uses an unbiased estimator of the likelihood (Equation (1.3)) and proved to converge toward the exact posterior $\pi(\theta \mid \mathbf{y})$. This idea was then extended and generalised by Andrieu and Roberts (2009), under the well known pseudo-marginal Metropolis-Hastings (PMMH) methods. Recent adaptations are the blockwise PMMH (Tran, Kohn et al., 2017a) and the correlated PMMH (Deligiannidis et al., 2018).

(Metropolis-within-)Gibbs Another class of MCMC algorithm to sample from the posterior distribution is the Gibbs sampler. A basic two stage Gibbs sampler can be used to draw samples from the joint posterior distribution $\pi(\theta, \mathbf{u} \mid \mathbf{y})$. Successive sampling from the distributions of $\theta \mid \mathbf{u}, \mathbf{y}$ and $\mathbf{u} \mid \theta, \mathbf{y}$ are performed. This strategy was initially proposed by Tanner and Wong (1987) under the name data augmentation, outside of the Gibbs framework. Sampling from the conditional $\pi(\theta \mid \mathbf{u}, \mathbf{y})$ is usually easy, however from $f(\mathbf{u} \mid \theta, \mathbf{y})$ is difficult and it limits this approach, especially when the dimension of $\mathbf{u}$ is large. To solve this problem, one might think about a Metropolis-within-Gibbs algorithm – i.e. obtaining a simulation from $f(\mathbf{u} \mid \theta, \mathbf{y})$ by building a Metropolis-Hastings chain targeting this distribution – but its expression is unknown, preventing such approach.

Importance sampling squared An unbiased estimator of the likelihood can also be used to make feasible an importance sampling strategy to approximate the posterior distribution. Indeed, a classic importance sampling method can be used to approximate any posterior expectation for some function of parameters $h(\theta)$:

$$\begin{aligned}\mathbb{E}(h(\theta) \mid \mathbf{y}) = \int h(\theta)\pi(\theta \mid \mathbf{y})d\theta &= \int h(\theta)\frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{Z}d\theta \\ &\propto \int h(\theta)f(\mathbf{y} \mid \theta)\pi(\theta)d\theta, \end{aligned} \tag{1.4}$$

where Z is the normalising constant $\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta$. Considering an importance distribution $q(\theta)$ to generate N independent and identically distributed (i.i.d.) θ values, we have the respective approximations for Equation (1.4) and Z:

$$\frac{1}{N}\sum_{i=1}^{N} w(\theta^{(i)})h(\theta^{(i)}) \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N} w(\theta^{(i)}), \quad \text{where} \quad w(\theta^{(i)}) = \frac{f(\mathbf{y} \mid \theta^{(i)})\pi(\theta^{(i)})}{q(\theta^{(i)})}.$$

The resulting approximated posterior expectation is

$$\hat{\mathbb{E}}(h(\theta) \mid \mathbf{y}) = \sum_{i=1}^{N} \frac{w(\theta^{(i)})}{\sum_{i=1}^{N} w(\theta^{(i)})} h(\theta^{(i)}).$$

However, the importance weights cannot be computed due to the presence of the intractable likelihood expression. To circumvent this issue, Tran, Scharth et al. (2013) propose to replace its expression by an unbiased estimation, thanks to importance sampling. They name this approach "importance sampling squared" (IS^2) due to the double use of importance sampling. Unfortunately, the difficulties to approximate the likelihood mentioned in Section 1.2.1 still remain, (due to the high dimensionality of $\mathbf{u}$).

Variational inference To approximate the posterior distribution $\pi(\theta \mid \mathbf{y})$ one deterministic approach is the variation inference (a.k.a. variational Bayes in this setting) (Attias, 2000; Bishop, 2006). Its principle is to use a set of tractable distributions on θ, denoted $q_{\mathbf{v}}(\theta)$, to approximate the posterior of interest. $q_{\mathbf{v}}(\theta)$ is the so called variational distribution, and $\mathbf{v}$ its variational parameters. To determine the closest distribution to $\pi(\theta \mid \mathbf{y})$, the Kullback-Leiber (KL) divergence between $q_{\mathbf{v}}(\theta)$ and $\pi(\theta \mid \mathbf{y})$ is minimised:

$$\begin{aligned} KL(q_{\mathbf{v}}(\theta) \,||\, \pi(\theta \mid \mathbf{y})) &= -\int q_{\mathbf{v}}(\theta) \log \frac{\pi(\theta \mid \mathbf{y})}{q_{\mathbf{v}}(\theta)} d\theta \\ &= -\int q_{\mathbf{v}}(\theta) \log \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{f(\mathbf{y})} d\theta + \int q_{\mathbf{v}}(\theta) \log q_{\mathbf{v}}(\theta) d\theta \\ &= -\int q_{\mathbf{v}}(\theta) \log f(\mathbf{y} \mid \theta)\pi(\theta) d\theta + \int q_{\mathbf{v}}(\theta) \log q_{\mathbf{v}}(\theta) d\theta + \log f(\mathbf{y}) \end{aligned}$$

Note that the logarithm of the evidence, $\log f(\mathbf{y})$, is not involved in the minimisation toward $q_{\mathbf{v}}(\theta)$, moreover, the two left hand terms are in fact the negation of the evidence lower bound (denoted ELBO in the literature). The likelihood expression is again required, preventing such an approach. An alternative is to target the joint posterior distribution $\pi(\theta, \mathbf{u} \mid \mathbf{y})$. In this case, the mean-field variational inference

can be used and the variational function is decomposed into $q_{\mathbf{v}}(\theta, \mathbf{u}) = q(\theta)q(\mathbf{u})$. This mean-field variational Bayes method alternates between optimisation according to $\mathbf{u}$ and θ, to finally consider the optimal $q(\theta)$ as an approximation of the posterior $\pi(\theta \mid \mathbf{y})$. The dependence between θ and $\mathbf{u}$ is hence broken. Again relying on an unbiased estimator of the likelihood, Tran, Kohn et al. (2017b) and Gunawan et al. (2017) recently extended the variational Bayes approach to approximate the posterior on θ when the likelihood is intractable. Still in the approximate Bayesian setting, let finally mention the integrated nested Laplace approximation method (INLA, Rue et al., 2009), that is increasingly used for Bayesian inference.

The above presented strategies are usually hard to apply due to the large data dimension we are facing (observed or latent): MCMC have difficulties to explore such a high dimensional space, IS techniques suffer from the high dimensionality of the latent space inducing a large variance for its weights, and both can be computationally intensive. Moreover, mean-field variational inference can produce underestimated posterior variances (Consonni and Marin, 2007). In a general manner, a limited range of model structures is allowed, and simplifying assumptions might be required.

1.3 Approximate Bayesian Computation

The solution we focus on to face the intractability of the likelihood, is called approximate Bayesian computation (ABC). ABC relies on the assumption that, for a given parameter value, even if the likelihood cannot be evaluated, it is still possible to generate artificial data from the model. It is a likelihood-free method relying only on simulations and is thus very flexible. Initially introduced in population genetics problems (Weiss and von Haeseler, 1998; Pritchard et al., 1999; Beaumont, Zhang et al., 2002), ABC methods have been used in an ever increasing range of applications, corresponding to different types of complex models in diverse scientific fields as epidemiology (e.g. Rodrigues, Francis et al., 2018), systems biology (e.g. Liepe and Stumpf, 2018), climatism (e.g. Holden et al., 2018), ecology (e.g. Fasiolo and Wood, 2018), nuclear imaging (e.g. Fan et al., 2018), population linguistics (e.g. Thouzeau et al., 2017).

In the following, we focus on ABC methods to perform parameter inference: we present the basics of ABC, emphasise the tuning aspects involved in ABC, and introduce the methods used in the following chapters. We do not present the asymptotic and theoretical properties of ABC, see Fernhead (2018) for a recent review.

1.3.1 Foundation of ABC

As in many Bayesian techniques, the goal of ABC is to obtain simulations from the posterior distribution $\pi(\theta \mid \mathbf{y})$. To do so, the idea is to generate a parameter value θ' from the prior $\pi(\cdot)$, and to accept it as coming from the posterior if a data $\mathbf{x}$ simulated from the generative model $f(\cdot \mid \theta')$, is similar enough to the observation $\mathbf{y}$. As $\mathbf{y}$ is a high dimensional vector, projections in a simpler space are required to ease comparisons, and one of the main questions of ABC is the meaning of "similar".

The earliest precursor of ABC is Rubin (1984). Outside of the intractability problem we are facing, but simply to illustrate the concept of posterior distribution in an intuitive way, Rubin (1984), proposed the rejection sampling algorithm described in Algorithm 1.2. It generates an N sample of parameter values from exactly $\pi(\theta \mid \mathbf{y})$, by iteratively drawing a parameter value from $\pi(\theta)$, generating a pseudo-data $\mathbf{x}$ from the generative model conditioned on this parameter, and accepting the simulated parameters if there is an exact match between $\mathbf{x}$ and $\mathbf{y}$. This

Algorithm 1.2 : Basic rejection sampling

for $i \leftarrow 1$ **to** N **do**
  repeat
    Simulate $\theta^{(i)} \sim \pi(\cdot)$;
    Simulate $\mathbf{x}^{(i)} \sim f(\cdot \mid \theta^{(i)})$;
  until $\mathbf{x}^{(i)} = \mathbf{y}$;
  Accept $\theta^{(i)}$;
end

rejection sampler is likelihood-free but only efficient when $\mathbf{y}$ is a low dimensional discrete vector. In the opposite case, simulating a pseudo-data $\mathbf{x}$ exactly equal to $\mathbf{y}$ is a zero probability event.

To avoid this issue, a solution that emerged is to project $\mathbf{y}$ into a lower dimensional space, thanks to a set of summary statistics $\eta(\cdot) \in \mathcal{S} \subseteq \mathbb{R}^d$, where d is the number of summary statistics. For convenience, we denote in this chapter $\eta_{\mathbf{x}} := \eta(\mathbf{x})$. Tavaré et al. (1997) proposed a method to estimate the time of the most recent common ancestor in a coalescent tree, where raw data are summarised. The final acceptance probability of their rejection algorithm depends on a probability proportional to the likelihood of the summarised observation, thus this is still not a likelihood-free algorithm.

Independently of Rubin's work, Weiss and von Haeseler (1998) introduce a similar algorithm. Based on a grid of parameter values, and not on prior simulations (even though a uniform grid can be seen as a uniform prior), they propose to approximate the likelihood for each grid coordinate, by the proportion of times simulated summaries are close enough to the observed ones, in the sense that their distance being lower than a threshold. This way of comparing data is a cornerstone of most ABC methods.

The incorporation of prior and of what will later be called basic ABC algorithm (Algorithm 1.3) is achieved by Pritchard et al. (1999). The terms "approximate Bayesian computation" are given by Beaumont, Zhang et al. (2002). The similarity between the simulated and observed data is evaluated thanks to a distance ρ, a threshold parameter ϵ, also called tolerance, and a set of summary statistics η.

The approximation aspects of ABC are twofold (outside the standard Monte-Carlo approximation):

- the data are projected into a lower dimensional space thanks to a set of summary statistics. Unless these are sufficient, an approximation occurs as the

Algorithm 1.3 : Basic ABC rejection sampling

for $i \leftarrow 1$ **to** N **do**
  repeat
    Simulate $\theta^{(i)} \sim \pi(\cdot)$;
    Simulate $\mathbf{x}^{(i)} \sim f(\cdot \mid \theta^{(i)})$;
    Compute $\eta_{\mathbf{x}^{(i)}}$;
  until $\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon$;
  Accept $(\theta^{(i)}, \eta_{\mathbf{x}^{(i)}})$;
end

target posterior is thus $\pi(\theta \mid \eta_{\mathbf{y}})$. This posterior is often mentioned as partial posterior distribution,

- the similarity is measured thanks to a well chosen distance $\rho(\cdot, \cdot)$ and its associated tolerance level ϵ, leading to an approximated posterior $\pi_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y}})$.

The sampling procedure described in Algorithm 1.3 draws $(\theta, \eta_{\mathbf{x}})$ values from the joint posterior

$$\pi_{\rho,\epsilon}(\theta, \eta_{\mathbf{x}} \mid \eta_{\mathbf{y}}) \propto \mathbb{1}\left\{\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}}) \leq \epsilon\right\} f(\eta_{\mathbf{x}} \mid \theta)\pi(\theta), \tag{1.5}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, $f(\eta_{\mathbf{x}} \mid \theta)$ is the likelihood of a summarised data, and intuitively sampling from it consists in generating $\mathbf{x}$ from the model, and computing $\eta_{\mathbf{x}}$. Note also that the proportional symbol stands for the omission of the normalisation constant $Z = \int\int \mathbb{1}\left\{\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}}) \leq \epsilon\right\} f(\eta_{\mathbf{x}} \mid \theta)\pi(\theta)d\theta d\eta_{\mathbf{x}}$.

The joint empirical measure from such an N sample is given by

$$\pi_{\rho,\epsilon}(\theta, \eta_{\mathbf{x}} \mid \eta_{\mathbf{y}}) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{(\theta^{(i)}, \eta_{\mathbf{x}^{(i)}})}(\theta, \eta_{\mathbf{x}}),$$

where δ denotes the Dirac measure.

When integrating over $\eta_{\mathbf{x}}$ it returns N parameter values drawn not from the exact posterior $\pi(\theta \mid \mathbf{y})$, but from the approximated distribution

$$\pi_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y}}) \propto \int \mathbb{1}\left\{\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}}) \leq \epsilon\right\} f(\eta_{\mathbf{x}} \mid \theta)\pi(\theta)d\eta_{\mathbf{x}}.$$

Intuitively, if statistics are sufficient and $\epsilon = 0$, we retrieve Algorithm 1.2 as well as the exact posterior. However, summary statistics are rarely sufficient in practice, and the threshold value should be as small as possible without rendering the algorithm computationally unfeasible. We can also remark that if $\epsilon \to \infty$, we get simulations from the prior.

Let mention that instead of sampling from the joint posterior distribution (1.5), one can choose to sample from

$$\pi_{\rho,\epsilon}(\theta, \eta_{\mathbf{x}(1:B)} \mid \eta_{\mathbf{y}}) \propto \left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\rho(\eta_{\mathbf{x}(b)}, \eta_{\mathbf{y}}) \leq \epsilon\}\right)\left(\prod_{b=1}^{B} f(\eta_{\mathbf{x}(b)} \mid \theta)\right)\pi(\theta), \tag{1.6}$$

where $\eta_{\mathbf{x}(1:B)}$ denotes the quantities $\eta_{\mathbf{x}(1)}, \ldots, \eta_{\mathbf{x}(B)}$. These are obtained by drawing data B times from the generative model using a fixed value of θ. Equation (1.6) still admits as marginal in θ the desired posterior $\pi_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y}})$, moreover we retrieve (1.5) if $B = 1$. This more costly version is notably used in importance sampling ABC strategies to reduce the importance weight variance, or again in sequential approaches to lower the variance of the Metropolis-Hastings ratio (see e.g. Toni et al., 2009; Del Moral et al., 2012).

ABC rejection in practice In practice, Algorithm 1.3 can be extremely expensive in time if the threshold is too small. A more practical generalisation was proposed by Beaumont, Zhang et al. (2002) and presented in Algorithm 1.4. It consists in generating an N sample according to $f(\eta_{\mathbf{x}} \mid \theta)\pi(\theta)$ and assigning to each simulated couple $(\theta^{(i)}, \eta_{\mathbf{x}^{(i)}})$ a weight $w^{(i)} \propto K_\epsilon(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}))$, where $K_\epsilon(\cdot)$ is a unidimensional smoothing kernel with bandwidth parameter ϵ, such that $K_\epsilon(\cdot) = K(\cdot/\epsilon)/\epsilon$. This kernel gives higher weight to simulated data closer to $\mathbf{y}$, and vice-versa. This formulation has the advantage to retrieve the regular ABC acceptance condition when considering a uniform kernel with bandwidth ϵ, so that $K_\epsilon(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}})) \propto \mathbb{1}\{\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon\}$. Moreover, smoother weights can be used, as the Gaussian or Epanechnikov kernel, this latter gives zero weights to data in its tails, and smooth weights otherwise. The choice of K presents little influence on final inference and the Epanechnikov kernel is often used for smooth weighting. In practice, the bandwidth ϵ is taken so that a proportion p_ϵ of the simulations carries a non-zero weight, which is equivalent to choosing a uniform kernel with bandwidth equal to the p_ϵ-quantile of the distances $\{\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}})\}_{i=1,\ldots,N}$. The weighted sample is used to obtain an estimator of the posterior distribution, and ϵ can be seen as a trade-off parameter between bias and variance of the estimator: a large value of ϵ results in larger sample size, reducing the estimator variance at the cost of a higher bias, and small ϵ leads to less bias (as $\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})$ are small), at the price of a higher estimator variance.

Algorithm 1.4 : Weighted ABC sampler

for $i \leftarrow 1$ **to** N **do**
- Simulate $\theta^{(i)} \sim \pi(\cdot)$;
- Simulate $\mathbf{x}^{(i)} \sim f(\cdot \mid \theta^{(i)})$;
- Compute $\eta_{\mathbf{x}^{(i)}}$ and $\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}})$;
- Assign to $(\theta^{(i)}, \eta_{\mathbf{x}^{(i)}})$ a weight $w^{(i)} \propto K_\epsilon(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}))$;

end

Curse of dimensionality One important phenomenon that affects most ABC methods is the curse of dimensionality, in the sense that the number of summary statistics d greatly deteriorates the ABC algorithms performances as it gets larger. Indeed, the basic ABC rejection sampling (Algorithm 1.3) tends to reject more and more simulations as the dimension of η increases, and it quickly gets computationally inefficient because the number of artificial data generated must be extremely large to achieve decent performances, (see the discussion in Beaumont, 2010). Algorithm 1.4 is not spared by this problem as the kernel weights are directly influenced by d.

More precisely, in a non-parametric perspective, an estimate of the posterior density at θ can be obtained, using the weighted sample and a (Nadaraya–Watson) kernel density estimation, it results

$$\hat{\pi}(\theta \mid \eta_{\mathbf{y}}) = \frac{\sum_{i=1}^{N} K_b(\theta^{(i)} - \theta) K_\epsilon(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}))}{\sum_{i=1}^{N} K_\epsilon(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}))}, \tag{1.7}$$

where K_b is a density-estimation kernel with bandwidth b, usually different from K_ϵ, (Beaumont, Zhang et al., 2002; Blum, 2010). Blum (2010) showed that the mean squared error of such estimator directly depends on the number of summary statistics, and the larger it is the slower this error decreases toward zero. Similar results have been deduced by Fearnhead and Prangle (2012), Barber et al. (2015) and Biau, Cérou et al. (2015). It highlights the importance of the summary statistics choice.

ABC interpretations As noticed in Equation (1.5), a weighted sample is drawn from the joint distribution

$$K_\epsilon(\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})) f(\eta_{\mathbf{x}} \mid \theta) \pi(\theta).$$

The retained θ are hence drawn from the approximate (ABC) posterior distribution

$$\pi_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y}}) \propto \int K_\epsilon(\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})) f(\eta_{\mathbf{x}} \mid \theta) \pi(\theta) d\eta_{\mathbf{x}},$$

which becomes exact as ϵ tends to zero. As the above integrate does not depend on θ, we have

$$\pi_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y}}) \propto \underbrace{\int K_\epsilon(\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})) f(\eta_{\mathbf{x}} \mid \theta) d\eta_{\mathbf{x}}}_{f_{ABC}(\eta_{\mathbf{y}} \mid \theta)} \pi(\theta),$$

and ABC can be interpreted as approximating a likelihood $f_{ABC}(\eta_{\mathbf{y}} \mid \theta)$, which is used to perform regular Bayesian analysis.

As we mentioned earlier, the number of accepted particles can be pre-defined as a percentile of simulated distances. In this way, Algorithm 1.4 can be seen as a k_N-nearest neighbours algorithm (when K_ϵ is uniform), in which choosing ϵ means choosing the number of neighbours. This is the viewpoint provided by Biau, Cérou et al. (2015), for which they studied properties, and provided a posterior density estimate similar to (1.7).

Another insight, highlighted by Wilkinson (2013) is that the basic ABC Algorithm 1.3 (without consideration of summary statistics) provides exact draws from $\pi(\theta \mid \mathbf{y})$ under the assumption of model error. Indeed, it assumes that the observation $\mathbf{y}$ cannot be exactly retrieved by a realisation of the considered model $f(\cdot \mid \theta)$, due to either model and/or data measurement errors. In this way, the error assumption is

$$\mathbf{y} = \mathbf{x} + \delta,$$

where $\mathbf{x}$ is a realisation from $f(\cdot \mid \theta)$, and δ is the error term. When δ follows a uniform distribution $\mathcal{U}\{\mathbf{x} : \rho(\mathbf{x}, \mathbf{y}) \leq \epsilon\}$ Algorithm 1.3 is exact. Considering such a

uniform distribution on δ means the error (of the model or data) is 1 in a ball of radius ϵ, centred in $\mathbf{y}$, and 0 otherwise. Wilkinson (2013) provides a more general perspective of ABC, and depending on the density of δ, multiples ABC strategies can be retrieved.

1.3.2 Tuning in ABC

The main drawback of classical ABC algorithm is its tuning aspects. It requires to specify a distance ρ, summary statistics η, tolerance/bandwidth ϵ. Radical differences can be observed depending on their choices. In this section we present some of the approaches and results to facilitate the tuning of ABC.

1.3.2.1 Summary statistics

As mentioned earlier, in order to make feasible the comparison between artificial data and the observation, their dimension is reduced using a set of d summary statistics. These are pivotal to ensure good performance of the ABC algorithm, and in practice they must capture enough information from the observed and simulated data, without being too numerous (and ideally sufficient), as the ABC suffers from the curse of dimensionality. Indeed, among the many visions of ABC, one of them is seeing ABC as a nearest neighbour approach, however finding neighbours for a data when the space is extremely large is a quite complicated task. We present below a brief overview of existing solutions for this choice, and we recommend Beaumont (2010), Blum, Nunes et al. (2013) and Prangle (2018) for more information concerning this subject.

Selection The first strategy to determine a set of summary statistics consists in performing selection among a set of d user-prespecified statistics. When θ is univariate, Joyce and Marjoram (2008) use a stepwise selection scheme to deduce whether a summary is important or not for the posterior estimation. This approach relies on scoring the inclusion of a summary statistic $\eta_{\mathbf{y},k+1}$ among a set $\eta_{\mathbf{y},1}, \ldots, \eta_{\mathbf{y},k}$ to access whether or not $\eta_{\mathbf{y},k+1}$ is informative for the approximation of the posterior. This is done by measuring the difference between the ratio of approximated posteriors $\hat{\pi}_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y},1}, \ldots, \eta_{\mathbf{y},k-1}, \eta_{\mathbf{y},k}) / \hat{\pi}_{\rho,\epsilon}(\theta \mid \eta_{\mathbf{y},1}, \ldots, \eta_{\mathbf{y},k-1})$ and 1, for inferring that $\eta_{\mathbf{y},k}$ is informative if the difference is greater than a threshold $T(\theta)$. The proposal of Nunes and Balding (2010) is to select the subset of the summary statistics that minimises the entropy of the approximated posterior. In the frame of a regression explaining a dimension of θ thanks to the statistics, Sedki and Pudlo (2012) and Blum, Nunes et al. (2013) suggested the use of variable selection thanks to Akaike information criterion (AIC) or Bayesian information criterion (BIC). An approach based on the Kullback-Leibler divergence is proposed by Barnes et al. (2012) and Filippi et al. (2012) to assess of the sufficiency of statistics.

The principle of selection is very advantageous in terms of interpretability, as summary statistics are usually chosen by practitioners with in mind the meaning associated to each summary. However, this advantage also becomes its defect when the informative summaries are not in the proposed set. Moreover, greedy search of

the best set quickly becomes unfeasible when d gets large, and less effective stepwise methods must be used.

Projection The relevant summary statistics might be combinations of existent ones. Wegmann et al. (2009) propose to use the partial least squares (PLS) components which are linear combinations of statistics, decorrelated, that maximise the covariance with the response θ. This approach is advantageous because it can deal with collinearity and dependence between statistics. These are trained on a set of pilot ABC simulations, and the number of components retained is usually chosen by cross-validation. Some regression adjustment techniques (Section 1.3.3) make use of feed-forward neural networks as they internally perform projection of the statistics on a lower dimensional space (the number of hidden units of the multilayer perceptron, see Blum and François, 2010). Fearnhead and Prangle (2012) proved that the optimal choice for the summary statistics should be the posterior expectation $\mathbb{E}(\theta \mid \mathbf{y})$, thus η should have the same dimension as the parameters. This expectation is unknown and Fearnhead and Prangle (2012) propose to use its approximation as summary statistics. To do so, thanks to ABC simulations, for each parameter θ_j they fit a linear regression explaining θ_j thanks to a set of preliminary summary statistics, and use the approximated posterior means as new summary statistics for a classic ABC run. Finally, Aeschbacher et al. (2012) develop an approach based on boosting for choosing summary statistics. The disadvantage of these two last strategies is their lack of interpretability.

Indirect inference summaries A major strategy to choose the summary statistics, is to rely on a simpler and more tractable likelihood in order to extract information and statistics. This is related to the indirect inference method we detailed in Section 1.2.2. Under this idea, Drovandi and Pettitt (2011) propose to use as summary statistics for an ABC analysis, the estimated parameter values of the tractable auxiliary model, for the observed data $\mathbf{y}$ and each generated $\mathbf{x}$ from $f(\cdot \mid \theta)$. The new summaries are thus the estimates $\hat{\phi}_{\mathbf{y}}$ and $\hat{\phi}_{\mathbf{x}}$. This approach – later called ABC-IP by Gleim and Pigorsch (2013) (for indirect parameter) – uses as distance a weighted quadratic one. Gleim and Pigorsch (2013) propose to replace this distance by the log-likelihood difference:

$$\log\left[f_{aux}(\mathbf{y} \mid \hat{\phi}_{\mathbf{y}})\right] - \log\left[f_{aux}(\mathbf{y} \mid \hat{\phi}_{\mathbf{x}})\right],$$

giving the so called ABC-IL (for indirect likelihood). Another proposal of Gleim and Pigorsch (2013) is the ABC-IS (for indirect score), in which the summary statistics of a data $\mathbf{x}$ are the scores

$$\left(\frac{\partial \log\left[f_{aux}(\mathbf{x} \mid \phi)\right]}{\partial \phi_1}, \ldots, \frac{\partial \log\left[f_{aux}(\mathbf{x} \mid \phi)\right]}{\partial \phi_{dim(\phi)}}\right),$$

evaluated at the estimate $\hat{\phi}_{\mathbf{y}}$. Comparisons between these ABC indirect inference approaches (ABC-II) can be found in Drovandi, Pettitt and Lee (2015). Here, the choice of summary statistics for the generative model is transformed into the choice of an auxiliary model able to provide information on the first one. For a discussion and large review of indirect inference for ABC we suggest the book chapter of Drovandi (2018).

Regularisation In the frame of regression adjustment techniques, regularization is an approach proposed by Blum, Nunes et al. (2013) and Saulnier et al. (2017), (see Section 1.3.3).

1.3.2.2 Distance choice

The distance ρ to measure the discrepancy between the observed and simulated summary statistics is critical too. Because the components of η can present different spread as well as correlations, scaling weights should be used for ρ to avoid that the most variable statistics dominate in the distance calculation. A common choice is to use a weighted Euclidean distance, normalised by the empirical mean absolute deviation (Csilléry et al., 2012) or standard deviation (Beaumont, Zhang et al., 2002) of each summary, (these are Mehalanobis distance, Vo et al., 2015), i.e.

$$\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}}) = \left[\sum_{i=1}^{d} \left(\frac{\eta_{\mathbf{x},i} - \eta_{\mathbf{y},i}}{\sigma_i} \right)^2 \right]^{\frac{1}{2}},$$

where $\eta_{\cdot,i}$ denotes the i-th dimension of $\eta_{\cdot}$ and the σ_i are the scaling quantities, which are often deduced thanks to a pilot ABC run. Prangle (2017) largely addresses this question and proposes an adaptive distance for sequential ABC methods (see section 1.3.4), as these quantities might vary through iterations of such techniques. In his paper, the distance is updated at each step of the algorithm. Very recently, instead of relying on summary statistics, Bernton et al. (2017) proposed to use the Wasserstein distance to compare empirical distributions of raw data.

1.3.2.3 Threshold

The threshold specification depends on the considered ABC approach. For example, regarding the basic rejection ABC (Algorithm 1.3), ϵ controls the trade-off between precision of the algorithm, and computational cost. As we mentioned earlier, the lower the better in terms of accuracy, however the computational cost is drastically increased when ϵ gets close to zero, because artificial data are rejected more and more. This parameter is chosen depending on how long the user is willing to wait before obtaining some results and of the desired quality. Moreover, a preliminary ABC run is usually necessary to obtain some insights concerting the ϵ value to use. For the ABC version of Beaumont, Zhang et al. (2002) (Algorithm 1.4), in practice, a pre-specified percentage of the simulations providing the lowest distances are retained with a non-zero weight (Beaumont, Zhang et al., 2002).

Some strategies specifically developed to either reduce the influence of ϵ, or either allow the use of a smaller ϵ value without compromising the efficiency of the algorithm. The first situation refers to regression adjustment methods we present in Section 1.3.3, the second one is a reference, for example, to ABC-MCMC (Marjoram et al., 2003) or even ABC-PMC and ABC-SMC we present in Section 1.3.4.

1.3.3 Regression adjustment

Once a set of parameters has been simulated and weighted (Algorithm 1.4), some post-treatment can be achieved on them to reduce the discrepancy between simulated and observed summary statistics. Hence, another valuable idea initiated by Beaumont, Zhang et al. (2002) is the so called regression adjustment we present here. The idea is to assume a certain relationship between parameters and summary statistics, in order to recalibrate simulations toward the observed data.

Linear adjustment For more simplicity, we consider in this part that θ is univariate. Beaumont, Zhang et al. (2002) assumed that the relation between the parameters $\theta^{(i)}$ and the summary statistics $\eta_{\mathbf{x}^{(i)}}$ can be modelled by a local linear regression:

$$\theta^{(i)} = \alpha + (\eta_{\mathbf{x}^{(i)}} - \eta_{\mathbf{y}})^\top \beta + \zeta^{(i)}, \quad i = 1, \dots, N, \tag{1.8}$$

where α is the intercept, β is a vector of regression coefficients, and $\zeta^{(i)}$ are the i.i.d. residuals with zero mean and common variance. The conditional expectation is $\mathbb{E}(\theta \mid \eta_{\mathbf{x}^{(i)}}) = \alpha + (\eta_{\mathbf{x}^{(i)}} - \eta_{\mathbf{y}})^\top \beta$. When evaluated at $\eta_{\mathbf{y}}$ we have

$$\theta = \alpha + \zeta. \tag{1.9}$$

The idea of adjustment is to obtain an empirical distribution for the residuals, that can be plugged into Equation (1.9) and correct the parameter values toward $\mathbf{y}$.

The unknowns α and β are obtained by minimising the weighted least squares criterion

$$\sum_{i=1}^{N} w^{(i)} \left(\theta^{(i)} - \alpha - (\eta_{\mathbf{x}^{(i)}} - \eta_{\mathbf{y}})^\top \beta\right)^2. \tag{1.10}$$

The empirical distribution provided by $\zeta^{(i)}$'s is then plugged into Equation (1.9), to obtain corrected $\theta^{(i)}$ values, denoted $\theta_c^{(i)}$:

$$\begin{aligned} \theta_c^{(i)} &= \hat{\alpha} + \hat{\zeta}^{(i)} \\ &= \hat{\alpha} + (\theta^{(i)} - \hat{\alpha} - (\eta_{\mathbf{x}^{(i)}} - \eta_{\mathbf{y}})^\top \hat{\beta}) \\ &= \theta^{(i)} - (\eta_{\mathbf{x}^{(i)}} - \eta_{\mathbf{y}})^\top \hat{\beta}. \end{aligned}$$

These $\theta_c^{(i)}$ values, weighted by $w^{(i)}$, with $i = 1, \dots, N$, yield a sample from the approximated posterior distribution. Note that these corrected values might end outside of the prior limits, to prevent this issue transformations can be necessary (it holds for adjustment methods in general).

When θ is multivariate, a linear adjustment can be performed on each component separately, or a multivariate regression can be adopted. This approach is valid when the assumed relationship (1.8) is exact. However, θ and $\eta_{\mathbf{y}}$ rarely present a linear relation, and the homoscedastic assumption is often violated. In response to this issue more complicated relationships can be assumed.

Non-linear adjustment Blum and François (2010) proposed the more flexible non-linear conditional heteroscedastic model

$$\theta^{(i)} = \mu(\eta_{\mathbf{x}^{(i)}}) + \sigma(\eta_{\mathbf{x}^{(i)}})\zeta^{(i)}, \quad i = 1, \dots, N,$$

where $\mu(\eta_{\mathbf{x}^{(i)}})$ is the conditional expectation $\mathbb{E}(\theta \mid \eta_{\mathbf{x}^{(i)}})$, $\sigma^2(\eta_{\mathbf{x}^{(i)}})$ is the conditional variance $\mathbb{V}(\theta \mid \eta_{\mathbf{x}^{(i)}})$, and $\zeta^{(i)}$ is the residual, still i.i.d. centred with common variance. These posterior quantities are estimated thanks to neural networks (Blum and François, 2010) (or thanks to another regression approach taking benefit of $w^{(i)}$). Once $\hat{\mu}$ and $\hat{\sigma}$ are deduced, the adjustment is done as follow

$$\begin{aligned}\theta_c^{(i)} &= \hat{\mu}(\eta_{\mathbf{y}}) + \hat{\sigma}(\eta_{\mathbf{y}})\hat{\zeta}^{(i)} \\ &= \hat{\mu}(\eta_{\mathbf{y}}) + \hat{\sigma}(\eta_{\mathbf{y}}) \left\{ \frac{1}{\hat{\sigma}(\eta_{\mathbf{x}^{(i)}})} \left(\theta^{(i)} - \hat{\mu}(\eta_{\mathbf{x}^{(i)}})\right) \right\}.\end{aligned}$$

This non-linear adjustment presents great performance, and is even less sensitive to the proportion of accepted simulations (i.e. with non-zero weight) compared to linear approaches. This is particularly advantageous when the cost to obtain a pseudo-data is high, as it reduces the importance of the threshold choice ϵ. In addition, the main motivation to the use of neural networks is their ability to reduce the summary statistics space internally thanks to projection on a lower dimensional one, this technique hence lowers the importance of the choice of summary statistics.

Many more correction methods were developed assuming different regression models. A quadratic one was proposed by Blum (2010) or again a generalised linear one (Leuenberger and Wegmann, 2010). Still assuming a linear relationship, regularised regression techniques have been studied as ridge (Blum, Nunes et al., 2013) or LASSO (Saulnier et al., 2017). Instead of the criterion (1.10), the regularised weighted least squares is used:

$$\sum_{i=1}^{N} w^{(i)} \left(\theta^{(i)} - \alpha - (\eta_{\mathbf{x}^{(i)}} - \eta_{\mathbf{y}})^\top \beta\right)^2 + \lambda ||\beta||,$$

where λ is the positive regularisation parameter, $||\beta||$ is either the L_2 norm for ridge or L_1 for LASSO. These regularisation techniques have the advantage of decreasing the importance of the summary statistics choice, as they shrink the regression coefficients toward zero it reduces the contribution of uninformative summary statistics. They also prevent singularity problems that can occur when α and β are estimated using the classical least squares. The R package `abc` (Csilléry et al., 2012) offers the implementations of local linear, neural networks or ridge adjustments that we will encounter in Chapter 4.

Finally, note that the posterior density estimation recalled in (1.7), can be applied with adjusted parameters, by replacing $\theta^{(i)}$ with $\theta_c^{(i)}$.

1.3.4 Iterative improvements

In this section we present some existing improvements of ABC, through iterative techniques, called population Monte Carlo (PMC) and sequential Monte Carlo (SMC). Both are special cases of sequential importance sampling (SIS) (Liu, 2004, Chapter 2). The motivation behind these approaches is to sample from a smarter distribution than the prior, a distribution closer to the shape of the posterior. This is also the goal of MCMC-ABC techniques not presented here (Marjoram et al., 2003).

1.3.4.1 ABC importance sampling

As mentioned earlier, the basic ABC rejection approach (Algorithm 1.3) draws $(\theta, \eta_{\mathbf{x}})$ values from the target joint distribution

$$\pi_{\rho,\epsilon}(\theta, \eta_{\mathbf{x}} \mid \eta_{\mathbf{y}}) = \mathbb{1}\left\{\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}}) \leq \epsilon\right\} f(\eta_{\mathbf{x}} \mid \theta)\pi(\theta)/Z, \tag{1.11}$$

where Z is the normalising constant.

This algorithm can be redesigned into a rejection IS one, thanks to the introduction of an importance distribution $q(\theta, \eta_{\mathbf{x}})$. A practical choice is $q(\theta, \eta_{\mathbf{x}}) = f(\eta_{\mathbf{x}} \mid \theta)q(\theta)$, so that it only requires to choose a proposal on the parameters $q(\theta)$, and an (unnormalised) importance weight simplifies as follows:

$$w^{(i)} = \frac{\mathbb{1}\left\{\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon\right\} f(\eta_{\mathbf{x}^{(i)}} \mid \theta^{(i)})\pi(\theta^{(i)})}{f(\eta_{\mathbf{x}^{(i)}} \mid \theta^{(i)})q(\theta^{(i)})} = \frac{\mathbb{1}\left\{\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon\right\} \pi(\theta^{(i)})}{q(\theta^{(i)})}.$$

The resulting ABC rejection IS adaptation is presented in Algorithm 1.5. We can notice that an accepted simulation verifies $\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon$, resulting in the omission of the indicator function in the importance weights.

Algorithm 1.5 : Basic ABC rejection importance sampling

```
for i ← 1 to N do
    repeat
        Simulate θ^(i) ~ q(·);
        Simulate x^(i) ~ f(· | θ^(i));
        Compute η_{x^(i)};
    until ρ(η_{x^(i)}, η_y) ≤ ε;
    Compute the weight w^(i) = π(θ^(i)) / q(θ^(i));
    Accept (θ^(i), η_{x^(i)}) and weight w^(i);
end
```

Note that the normalising constant Z can be approximated by the average of the $w^{(i)}$ weights. The normalised version of the weights, denoted $W^{(i)} = w^{(i)} / \sum_{j=1}^{N} w^{(j)}$, leads to an approximation of $\pi_{\rho,\epsilon}(\theta, \eta_{\mathbf{x}} \mid \eta_{\mathbf{y}})$, which is

$$\pi_{\rho,\epsilon}(\theta, \eta_{\mathbf{x}} \mid \eta_{\mathbf{y}}) \approx \sum_{i=1}^{N} W^{(i)} \delta_{(\theta^{(i)}, \eta_{\mathbf{x}^{(i)}})}.$$

In the remaining of this section, w denotes the unnormalised weight, and W its normalised version.

The main drawback of importance sampling is again the choice of the importance distribution $q(\theta, \eta_{\mathbf{x}})$ (or $q(\theta)$ here), from which we must be able to sample in high density regions of the target.

1.3.4.2 Sequential importance sampling

Sequential importance sampling (SIS) addresses this problem. Its principle is to consider a sequence of T intermediate distributions

$$f_1(z), \ldots, f_T(z),$$

and perform IS for each of them starting from 1 to T, in such a way that $f_T(z)$ is the final target distribution we would like to sample from. Here, z denotes the target variables, for which value is called a particle and is denoted, at time t, by z_t.

To extend SIS to ABC, the natural target density at time t is

$$\begin{aligned} f_t(z) &= \pi_{\rho,\epsilon_t}(\theta, \eta_{\mathbf{x}} \mid \eta_{\mathbf{y}}) \\ &\propto \mathbb{1}\left\{\rho(\eta_{\mathbf{x}}, \eta_{\mathbf{y}}) \leq \epsilon_t\right\} f(\eta_{\mathbf{x}} \mid \theta)\pi(\theta) := \tilde{f}_t(z), \end{aligned}$$

where $z = (\theta, \eta_{\mathbf{x}})$ and $\tilde{f}_t$ denotes the unnormalised version of f_t. Moreover, a decreasing sequence of tolerance levels $\epsilon_1 > \ldots > \epsilon_T$ must be considered to specify all the intermediate distributions, in such a way that $\epsilon_T = \epsilon$ to retrieve Equation (1.11) as final target. A main problem of iterative ABC methods concerns the specification of this sequence. Note that we could also consider the augmented target $\pi_{\rho,\epsilon_t}(\theta, \eta_{\mathbf{x}(1:B)} \mid \eta_{\mathbf{y}})$ (Equation (1.6)), however for convenience we only focus on the intuitive one with $B = 1$.

At each iteration t, the idea consists in performing IS to obtain a set of N weighted particles that provides an empirical distribution for the corresponding target f_t. The specificity of SIS is that the importance distribution $q_t(z)$ depends on the past particles. The intuition is that if there is only few changes between two successive targets f_{t-1} and f_t, then a correct proposal at time t can be obtained by small changes in the particles at time $t-1$.

The specification of the proposal (or importance) distribution is pivotal, and in the following we present two ABC iterative Monte Carlo approaches: the ABC-PMC (Beaumont, Cornuet et al., 2009) and ABC-SMC (Del Moral et al., 2012). These are two methods we compare to in Chapter 4 and both rely on different importance proposal distributions.

1.3.4.3 Population Monte Carlo ABC

The ABC-PMC method is introduced by Beaumont, Cornuet et al. (2009) and is based on the population Monte Carlo approach (Cappé et al., 2004). Algorithm 1.6 presents the ABC-PMC. Its structure is very similar to the previously derived ABC rejection IS (Algorithm 1.5) and we describe and explain below its principle.

The first iteration of the ABC-PMC algorithm is a classic importance sampling step as described in Section 1.3.4.1, for which the importance distribution on θ is the prior $\pi(\theta)$.

For $t > 1$, ABC-PMC performs importance sampling at each algorithm iteration in order to approximate the corresponding target where the importance distribution is derived from the previous particle layer. Indeed, at iteration t, a parameter value is sampled in the previous particle population with probability proportional to its weight. The drawn θ_{t-1} value is then perturbed into θ_t thanks to a transition kernel $\tilde{K}_t(\theta_t \mid \theta_{t-1})$. An associated summary statistics is then simulated from the generative model using this disrupted parameter value. It means the proposal distribution used by ABC-PMC on $(\theta_t, \eta_{\mathbf{x}_t})$ is

$$q_t(\theta_t, \eta_{\mathbf{x}_t}) = f(\eta_{\mathbf{x}_t} \mid \theta_t) q_t(\theta_t),$$

Algorithm 1.6 : ABC-PMC

$t \leftarrow 1$;
for $i \leftarrow 1$ **to** N **do**
  repeat
    Simulate $\theta_t^{(i)} \sim \pi(\cdot)$;
    Simulate $\mathbf{x}_t^{(i)} \sim f(\cdot \mid \theta_t^{(i)})$ and compute $\eta_{\mathbf{x}_t^{(i)}}$;
  until $\rho(\eta_{\mathbf{x}_t^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon_t$;
  Assign a weight $w_t^{(i)} = 1$;
  Accept $(\theta_t^{(i)}, \eta_{\mathbf{x}_t^{(i)}})$ and weight $w_t^{(i)}$;
end
$t \leftarrow t + 1$;
while $t \leq T$ **do**
  for $i \leftarrow 1$ **to** N **do**
    repeat
      Sample $\theta_t^{(i)}$ from the empirical importance density

$$q_t(\cdot) = \sum_{j=1}^{N} \frac{w_{t-1}^{(j)}}{\sum_{j=1}^{N} w_{t-1}^{(j)}} \tilde{K}_t(\cdot \mid \theta_{t-1}^{(j)});$$

      Simulate $\mathbf{x}_t^{(i)} \sim f(\cdot \mid \theta_t^{(i)})$ and compute $\eta_{\mathbf{x}_t^{(i)}}$;
    until $\rho(\eta_{\mathbf{x}_t^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon_t$;
    Compute $w_t^{(i)} = \pi(\theta_t^{(i)}) / q_t(\theta_t^{(i)})$;
    Accept $(\theta_t^{(i)}, \eta_{\mathbf{x}_t^{(i)}})$ and weight $w_t^{(i)}$;
  end
  $t \leftarrow t + 1$;
end

where

$$q_t(\theta_t) = \sum_{j=1}^{N} \frac{w_{t-1}^{(j)}}{\sum_{j=1}^{N} w_{t-1}^{(j)}} \tilde{K}_t(\theta_t \mid \theta_{t-1}^{(j)}). \tag{1.12}$$

This proposal provides as unnormalised importance weight for an accepted particle:

$$w_t := w_t(\theta_t, \eta_{\mathbf{x}_t}) = \frac{\tilde{f}_t(\theta_t, \eta_{\mathbf{x}_t})}{q_t(\theta_t, \eta_{\mathbf{x}_t})} = \frac{\mathbb{1}\left\{\rho(\eta_{\mathbf{x}_t}, \eta_{\mathbf{y}}) \leq \epsilon_t\right\} f(\eta_{\mathbf{x}_t} \mid \theta_t)\pi(\theta_t)}{f(\eta_{\mathbf{x}_t} \mid \theta_t) q_t(\theta_t)} = \frac{\pi(\theta_t)}{q_t(\theta_t)},$$

which is again likelihood-free.

In Beaumont, Cornuet et al. (2009), the transition kernel $\tilde{K}_t(\theta_t \mid \theta_{t-1})$ is taken as a Gaussian density with mean θ_{t-1} and variance-covariance matrix twice the empirical one computed using the values $\{\theta_{t-1}^{(i)}\}_{i=1,\dots,N}$ and weights $\{w_{t-1}^{(i)}\}_{i=1,\dots,N}$ normalised. The covariance matrix is hence updated after each iteration.

As a final remark, let us point that the proposal density on θ_t is in fact

$$\int\int f_{t-1}(\theta_{t-1}, \eta_{\mathbf{x}_{t-1}}) \tilde{K}_t(\theta_t \mid \theta_{t-1}) d\theta_{t-1} d\eta_{\mathbf{x}_{t-1}}, \tag{1.13}$$

and its empirical version is used when needed. Indeed, as the weighted samples $\{(\theta_{t-1}^{(i)}, \eta_{t-1}^{(i)}), W_{t-1}^{(i)}\}_{i=1,\dots,N}$ provide an empirical distribution for f_{t-1}, we retrieve $q_t(\theta_t)$ (Equation (1.12)). In other terms, ABC-PMC aims at building a proposal thanks to the previous target f_{t-1} which is slightly transformed by the transition kernel $\tilde{K}_t(\theta_t \mid \theta_{t-1})$. The resulting importance weights are hence deduced conditionally on the whole past particles.

Very similar algorithms exist (Sisson, Fan and Tanaka, 2007; Sisson, Fan and Tanaka, 2009; Toni et al., 2009), and overcome some improvements, in particular concerning the thresholds specification $\epsilon_1, \dots, \epsilon_T$ which are hard to select without a preliminary ABC run. An adaptive version was proposed by Drovandi and Pettitt (2011), by using for iteration t the α-quantile of the previous accepted distances. This version requires a stopping rule, for example a total number of simulations or a final ϵ below which the algorithm stops. Still for iterative ABC methods, Prangle (2017) proposed two adaptive updates of the distance ρ. When a weighted Euclidean distance is used, one of the proposition made by Prangle is to update the scaling weights thanks to some accepted and discarded simulations of the previous iteration.

1.3.4.4 Sequential Monte Carlo ABC

While ABC-PMC is fully based on IS reasoning, the ABC-SMC approach we present now, from Del Moral et al. (2012), is based on the sequential Monte Carlo sampler (Del Moral et al., 2006) and involves MCMC elements. Algorithm 1.7 describes the full method.

The original ABC-SMC algorithm of Del Moral et al. (2012) considers as sequence of distributions the posterior joints $\pi_{\rho,\epsilon_t}(\theta, \eta_{\mathbf{x}(1:B)} \mid \eta_{\mathbf{y}})$, however, in the remaining we consider the simplifying case where $B = 1$, so that ABC-PMC and ABC-SMC have the same target.

As we noticed before, the ABC-PMC recovers some importance weights conditionally on the past particles. These weights are computed thanks to an empirical

version of the proposal, this can be expensive and increase the algorithmic time. Del Moral et al. (2006) introduce the SMC sampler to circumvent this approximation thanks to an artificial backward kernel $L_t(z_t \mid z_{t+1})$, which needs to be specified. Indeed, this kernel aims at increasing the dimension of the target, to keep track of all the past of a particle. The target f_t can thus be expressed in terms of the whole sequence of past particles $z_1, \dots, z_t$ and has the expression:

$$f_t(z_{1:t}) = \frac{\tilde{f}_t(z_t)}{Z_t} \prod_{k=1}^{t-1} L_k(z_k \mid z_{k+1}) \propto \tilde{f}_t(z_{1:t}),$$

where $z_{1:t} = (z_1, \dots, z_t)$, and Z_t is the normalising constant.

In this way, the unnormalised importance weight is a joint weight on the whole path of the particles and is expressed as

$$w_t(z_{1:t}) = \frac{\tilde{f}_t(z_{1:t})}{q_t(z_{1:t})}, \quad \text{where} \quad q_t(z_{1:t}) = q_1(z_1) \prod_{k=2}^{t} K_k(z_k \mid z_{k-1}).$$

When a transition kernel K_t is used to move the particles from iteration $t-1$ to iteration t, weights can be more practically expressed by the update:

$$w_t(z_{1:t}) = w_t(z_{1:(t-1)}) \frac{\tilde{f}_t(z_t) L_{t-1}(z_{t-1} \mid z_t)}{\tilde{f}_{t-1}(z_{t-1}) K_t(z_t \mid z_{t-1})},$$

where the ratio is called the unnormalised incremental weight (Liu, 2004, chapter 3). The ABC-SMC approach of Del Moral et al. (2012) is based on this representation. They consider for backward kernel, the reversal kernel defined by

$$L_{t-1}(z_{t-1} \mid z_t) = \frac{\tilde{f}_t(z_{t-1}) K_t(z_t \mid z_{t-1})}{\tilde{f}_t(z_t)}.$$

In this way the incremental weight simplifies into $\tilde{f}_t(z_{t-1}) / \tilde{f}_{t-1}(z_{t-1})$. Combined with the target expression, it leads to the weight update present in Algorithm 1.7.

Contrary to PMC where a weighted resampling of the particles is performed at each iteration, this ABC-SMC approach uses an MCMC step to determine the next value of each particle. So the transition kernel used here is an MCMC kernel of invariant density $f_t(z_{1:t})$. Each particle $z_{t-1}^{(i)}$, for which weight $W_{t-1}^{(i)}$ is non-zero, overcomes an MCMC transition step in the following way:

(1) propose a potential parameter value θ_t' using a transition kernel $\tilde{K}_t(\cdot \mid \theta_{t-1}^{(i)})$;

(2) simulate $\mathbf{x}_t'$ using the generative model $f(\cdot \mid \theta_t')$ and compute $\eta_{\mathbf{x}_t'}$;

(3) accept $(\theta_t', \eta_{\mathbf{x}_t'})$ with probability

$$\min \left\{ 1, \frac{\mathbb{1}\{\rho(\eta_{\mathbf{x}_t'}, \eta_{\mathbf{y}}) \leq \epsilon_t\} \tilde{K}_t(\theta_{t-1}^{(i)} \mid \theta_t') \pi(\theta_t')}{\mathbb{1}\{\rho(\eta_{\mathbf{x}_{t-1}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon_t\} \tilde{K}_t(\theta_t' \mid \theta_{t-1}^{(i)}) \pi(\theta_{t-1}^{(i)})} \right\}. \tag{1.14}$$

The first two steps indicate that a potential particle $(\theta'_t, \eta_{\mathbf{x}'_t})$ is generated from

$$K(\theta'_t, \eta_{\mathbf{x}'_t} \mid \theta^{(i)}_{t-1}, \eta_{\mathbf{x}^{(i)}_{t-1}}) = \tilde{K}(\theta'_t \mid \theta^{(i)}_{t-1}) f(\eta_{\mathbf{x}'_t} \mid \theta'_t).$$

The acceptance rate (1.14) is derived from the usual Metropolis-Hastings one, after simplification of the likelihoods. As for Algorithm 1.6, the transition kernel $\tilde{K}_t$ can be an adaptive Gaussian distribution with a variance-covariance matrix updated after each iteration.

The main point that distinguishes the ABC-PMC and ABC-SMC approach is their importance distributions. ABC-PMC builds its proposal on the previous particle layer thanks to importance sampling arguments, resulting in conditional weights $w_t(z_t)$, while ABC-SMC builds its proposal on the whole past of the particles and its importance weights stand for this complete past, $w_t(z_{1:t})$.

Particle degeneracy A major problem of sequential importance sampling is the particle degeneracy issue. It refers to the situation where a small number of particles carry very high weights and the remaining very low weights. Strategies to reduce the weights variance must be used. ABC-PMC avoids this issue thanks to the resampling that occurs at each iteration, so that low weighted particles are discarded. To control this degeneracy, the ABC-SMC (Algorithm 1.7) relies on the effective sample size (ESS) to measure the degeneracy (Liu, 2004). For a weighted sample of size N, the ESS is equal to $\text{ESS}(\{W^{(i)}\}_{i=1,\dots,N}) = (\sum_{i=1}^{N}(W^{(i)})^2)^{-1} \in [1, N]$, and can be seen as the amount of information contained in the weighted sample, in the sense that making an inference using the weighted sample is equivalent to making inference using $\text{ESS}(\{W^{(i)}\}_{i=1,\dots,N})$ particles truly sampled from the target. The strategy used by Del Moral et al. (2012) is to resample the particles if the ESS falls below a threshold N_T, after what the weights are all set to $1/N$. This mechanism has its roots in particle filter (or bootstrap filter, Gordon et al., 1993).

Moreover, Algorithm 1.7 does not rely on a pre-specified sequence of thresholds. Instead, from one iteration to the next, it deduces the next value so that the ESS – that depends on the corresponding threshold value – is decreased approximately by a factor $\alpha \in]0, 1[$. And the Algorithm stops when the threshold value drops below a final ϵ value. This "quality index" α determines how fast we move toward the final target, if α is close to one, the distributions are slowly moving toward it, and vice-versa for α close to zero.

As we mentioned earlier, the iterative ABC techniques have evolved to include adaptive distances, thresholds, perturbation kernels. One of the remaining adaptive improvements lies into the population size N, which should usually be quite large to ensure a sufficient degree of reliability. With this goal, Klinger and Hasenauer (2017) propose such an adaptation based on the uncertainty of kernel density estimates to automatise the number of particles during the algorithm iterations. This approach is available in the recent Python module `pyABC` (Klinger, Rickert et al., 2018).

Algorithm 1.7 : ABC-SMC

$t \leftarrow 1$;
for $i \leftarrow 1$ **to** N **do**
  Simulate $\theta_t^{(i)} \sim \pi(\cdot)$;
  Simulate $\mathbf{x}_t^{(i)} \sim f(\cdot \mid \theta_t^{(i)})$ and compute $\eta_{\mathbf{x}_t^{(i)}}$;
  Set $W_t^{(i)} \leftarrow 1/N$;
end
$t \leftarrow t + 1$;
while $\epsilon_{t-1} > \epsilon$ **do**
  Determine ϵ_t by solving
  $\mathrm{ESS}(\{W_t^{(i)}, \epsilon_t\}_{i=1,\dots,N}) = \alpha \mathrm{ESS}(\{W_{t-1}^{(i)}, \epsilon_{t-1}\}_{i=1,\dots,N})$ where

$$W_t^{(i)} \propto W_{t-1}^{(i)} \frac{\mathbb{1}\{\rho(\eta_{\mathbf{x}_{t-1}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon_t\}}{\mathbb{1}\{\rho(\eta_{\mathbf{x}_{t-1}^{(i)}}, \eta_{\mathbf{y}}) \leq \epsilon_{t-1}\}};$$

  if $\epsilon_t < \epsilon$ **then** $\epsilon_t \leftarrow \epsilon$;
  if $ESS(\{W_t^{(i)}\}_{i=1,\dots,N}) < N_T$ **then**
    Resample N particles with probability proportional to $W_t^{(i)}$;
    $W_t^{(i)} \leftarrow 1/N$ for $i \leftarrow 1$ **to** N;
  end
  for $i \leftarrow 1$ **to** N **do**
    if $W_t^{(i)} > 0$ **then**
      Sample $(\theta_t^{(i)}, \eta_{\mathbf{x}_t^{(i)}})$ from the MCMC kernel $K_t(\cdot, \cdot \mid \theta_{t-1}^{(i)}, \eta_{\mathbf{x}_{t-1}^{(i)}})$;
    end
  end
end

1.4 Machine learning and reference table

An increasing number of methods are based on the use of machine learning tools on a training set produced in an ABC style, in the sense that Bayesian framework allows an easy way to generate artificial data, on which a machine learning algorithm can be trained and then used to provide prediction for the observed data $\mathbf{y}$.

A training set for such approaches can be generated by sampling from the joint distribution $\pi(\theta)f(\eta_{\mathbf{x}} \mid \theta)$. Algorithm 1.8 presents how to generate a set of N data, which is called reference table in the following (Pudlo et al., 2016).

Algorithm 1.8 : Generation of a reference table of size N

for $i \leftarrow 1$ **to** N **do**
  Simulate $\theta^{(i)} \sim \pi(\cdot)$;
  Simulate $\mathbf{x}^{(i)} \sim f(\cdot \mid \theta^{(i)})$;
  Compute $\eta_{\mathbf{x}^{(i)}}$;
end

Neural networks are being used more and more, often under the term "deep learning". We can mention the work of Mondal et al. (2019) or again Sheehan and Song (2016). This latter makes use of multilayer neural networks learned to explain population genetics parameters of interest, thanks to hundreds of correlated summary statistics. However, neural networks are often not tuning-free and require meticulous calibrations.

An interesting approach has been provided by Papamakarios and Murray (2016) that we describe here. Its goal is to directly approximate the posterior distribution $\pi(\theta \mid \eta_{\mathbf{y}})$ thanks to a family of conditional densities $q_{\mathbf{v}}(\theta \mid \eta_{\mathbf{y}})$, parameterized by $\mathbf{v}$. They proposed to use $q_{\mathbf{v}}$ as a mixture density network (Bishop, 1994): $q_{\mathbf{v}}$ is a mixture of K Gaussian distributions, i.e.

$$q_{\mathbf{v}}(\theta \mid \eta_{\mathbf{x}}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\theta \mid \mu_k, \Sigma_k),$$

where α_k, μ_k and Σ_k are respectively the mixing coefficients, means and covariance matrices, and these parameters are calculated by a feed-forward neural network. The parameters $\mathbf{v}$ are in fact the networks parameters.

Two elements are required to obtain the approximation of the posterior distribution, denoted $\hat{\pi}(\theta \mid \eta_{\mathbf{y}})$. The first is the generation of a reference table of size N from the joint distribution $\tilde{\pi}(\theta)f(\eta_{\mathbf{y}} \mid \theta)$, where $\tilde{\pi}(\theta)$ is a so called "proposal prior". The second element is the property demonstrated in Papamakarios and Murray (2016), saying that the quantity

$$\frac{1}{N} \sum_{i=1}^{N} \log q_{\mathbf{v}}(\theta^{(i)} \mid \eta_{\mathbf{x}^{(i)}}) \tag{1.15}$$

is maximised with respect to $\mathbf{v}$, if and only if, (when $N \to \infty$),

$$q_{\mathbf{v}}(\theta \mid \eta_{\mathbf{x}}) \propto \frac{\tilde{\pi}(\theta)}{\pi(\theta)} \pi(\theta \mid \eta_{\mathbf{x}}).$$

Using this relationship and the reference table of $\{\theta^{(i)}, \eta_{\mathbf{x}^{(i)}}\}$ values, three key steps are performed:

(1) find $\mathbf{v}$ maximising Equation (1.15),

(2) evaluate the trained $q_{\mathbf{v}}$ at $\eta_{\mathbf{y}}$,

(3) reweight $q_{\mathbf{v}}(\theta \mid \eta_{\mathbf{y}})$ by $\pi(\theta)/\tilde{\pi}(\theta)$ and normalise it to recover the desired posterior distribution expression.

Moreover, in the spirit of sequential ABC techniques, they propose to iteratively learn as proposal prior $\tilde{\pi}(\theta)$ a distribution that gets closer and closer to the posterior thanks to the generation of small training data set.

The major drawback of this approach is that $\pi(\theta)q_{\mathbf{v}}(\theta \mid \eta_{\mathbf{y}})/\tilde{\pi}(\theta)$ must be easy to evaluate and to normalise, in order to retrieve the expression of the approximated posterior $\hat{\pi}(\theta \mid \eta_{\mathbf{y}})$. This limits the choice of possible prior distributions to normal or uniform priors. Based on this work and to overcome this difficulty, Lueckmann et al. (2017) consider a similar criterion to approximate the posterior $\pi(\theta \mid \eta_{\mathbf{y}})$. By including the importance weights inside the criterion to maximise, w.r.t. $\mathbf{v}$, the expression

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\pi(\theta^{(i)})}{\tilde{\pi}(\theta^{(i)})}K_{\epsilon}(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}}))\log q_{\mathbf{v}}(\theta^{(i)} \mid \eta_{\mathbf{x}^{(i)}}),$$

where K_{ϵ} is a kernel of bandwidth ϵ, the posterior density $q_{\mathbf{v}}(\theta \mid \eta_{\mathbf{y}})$ directly approximates the desired posterior.

1.5 ABC model choice

ABC can be adapted for model choice problems. This is widely used in population genetics, to select the best evolutionary scenario among a set of competing ones. To fit such a problem into the ABC framework, the unknown model index $\mathcal{M}$ is considered as an additional parameter, with a prior distribution. As for parameter inference, the standard ABC-model choice procedure can be again perceived as a k_N-nearest neighbors algorithm, in which a set of N artificial data is simulated in three steps:

1. simulate m from the model prior,
2. simulate θ_m from $\pi_m(\cdot)$,
3. simulate $\mathbf{x}$ from $f_m(\cdot \mid \theta_m)$ and compute $\eta_{\mathbf{x}}$.

The distances with the observation are computed, then the k_N lowest distances indicate the simulations to retain, and the selected model index values provide an approximate sample from the model posterior probability. The frequencies of model index can be used to obtain their approximate posterior probabilities (Grelaud et al., 2009; Toni et al., 2009). Another solution consists in using some local logistic

regression techniques (Fagundes et al., 2007; Cornuet, Ravigné et al., 2010) or more evolved strategies as expectation propagation ABC (Barthelmé and Chopin, 2014).

Finally, a methodology used for model choice by Pudlo et al. (2016), relies on random forests (RF, Breiman, 2001) trained on a reference table for model choice. This ABC-RF model choice idea predicts the model index and also derives the posterior probability of the selected model thanks to RFs. The two very profitable advantages of RFs are their robustness toward noise variables and the very few tuning parameters required (see Chapter 2).

1.6 Conclusion

In this chapter we gave a brief overview of likelihood-free techniques, especially ABC, their main disadvantage being the tuning parameters that need to be calibrated by the user, (distance to compare summary statistics, threshold of acceptance, set of pertinent summary statistics). As we exposed in Section 1.3.2 a large panel of strategies have been developed to guide these choices. The methods with regression adjustment (Section 1.3.3) are designed to reduce the impact of tuning in ABC. The iterative ABC techniques (Section 1.3.4) also improved to be as automatised as possible. Furthermore, the union between ABC and machine learning is growing, mainly for the high performances they can provide but also to lower the calibration required for ABC.

Our first contribution lies into this framework. We extend the ABC-RF model choice approach of Pudlo et al. (2016) to perform parameter inference, and provide easy to use tools for practitioners (Chapter 4) thanks to the R package `abcrf`. The next chapter is a reminder on the random forest algorithm, its principle and focus on its benefits for ABC methods.

Chapter 2

Random forest

Contents

2.1 Introduction

In this chapter we present the principle of Breiman (2001)'s random forests (RF) for regression and classification. We especially emphasise on their well known properties that can be beneficial to ABC analysis, namely their robustness to noise variables, their good performances, the practical information they offer and their quasi-absence of tuning parameters.

In a general manner, a decision tree here refers to a binary tree structure made of allocation rules with the aim of taking a decision regarding a problem at hand after following these different rules. In our case, we are either interested in classification or regression tasks and we would like to provide a prediction for an observed data. Briefly, a decision forest can be seen as an ensemble of decision trees that tries to improve the prediction accuracy by aggregating the tree predictions.

More precisely, a random forest is a non-parametric ensemble learning technique (Dietterich, 2000) for which a base learner is a random tree, i.e. a decision tree for which randomness has been introduced during its construction. The version of Breiman (2001) is at the core of this chapter. As presented below, its base tree structure is a CART (for classification and regression tree, Breiman, Friedman et al., 1984) and each one is different thanks to randomisation on data and covariates incorporated for the tree construction.

For this chapter we denote by Y the response variable, and $X = (X_1, \ldots, X_d)$ the associated set of d explanatory variables. We assume that we have a set of N independent and identically distributed response-predictors data $\{(y^{(i)}, x^{(i)})\}_{i=1,\ldots,N}$, which are used as training data set. We would like to predict the response variable for a test explanatory variable x^*. In regression, the response Y is continuous and the forest provides an approximation of $\mathbb{E}(Y \mid X = x^*)$, while in classification Y is categorical with K modalities.

2.2 Classification and regression tree (CART)

Principle The cornerstone of a random forest is the CART algorithm (Breiman, Friedman et al., 1984). A CART (or simply tree when there is no possible confusion), is a machine learning algorithm whose principle is to partition the predictor space into disjoint subspaces, in an iterative manner, and each one is assigned a prediction value which will be used for test data falling in this subspace.

A binary tree starts from the complete predictor space, a binary rule then divides it in two parts. Each resulting subspace is divided in two by a different condition and so on until a certain stopping rule is reached. A split involves a covariate j and a splitting value s. Such an iterative process can be visualised by a binary tree, see Figure 2.1 for an example, where each subspace is a tree node.

A tree is thus a structure made of internal nodes and terminal nodes. Each internal node carries a condition to partition the predictor space in a binary fashion. The first internal node, for which the predictor space is untouched, is the root. A terminal node is called a leaf, after which the subspace is not developed anymore. Each one carries a prediction value.

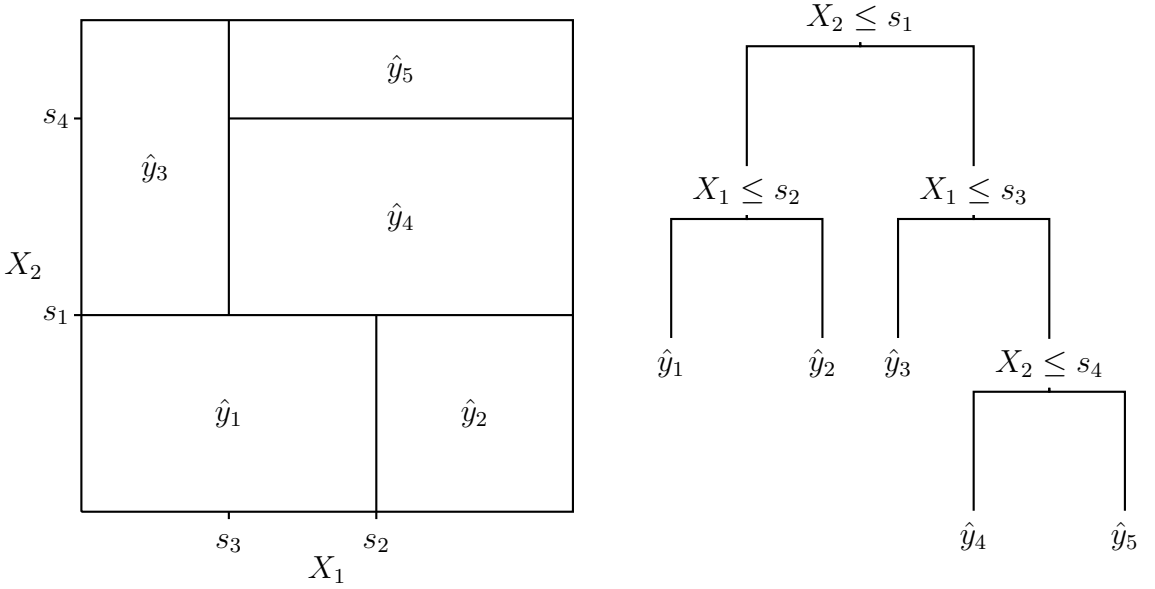


Figure 2.1 – An example of CART and the associated disjoint partition of the two dimensional predictor space. Each splitting condition takes the form $X_j \leq s$ and the prediction at a leaf is denoted $\hat{y}_\ell$.

Splitting criterion The way the space is partitioned is pivotal for a tree algorithm. For continuous covariates, at each internal node, the splitting condition takes the form $X_j \leq s$, where the pair (j, s) has to be determined. Data verifying this condition fall in the left resulting node, the right one otherwise. The subspace is hence partitioned according to the binary rule $\{X_j \leq s\}$ versus $\{X_j > s\}$. For each covariate X_j and possible bound s a non-negative gain criterion is calculated, j and s are selected has providing the maximal gain. This is done for each internal node, and the same covariate may be used for the choice of j at different levels of the tree construction.

Because of the shape of a split condition, it can be noticed that affine transformations on numeric covariates do not alter the criterion value and thus the tree construction. Moreover, a tree can handle categorical predictors, in this case the split concerns the belonging or not to a subset of its modalities. For the remaining we focus on numeric ones, to avoid switching between notations, even though the principles presented below holds for the categorical case.

The gain criterion can be seen as the decrease of information $I(\cdot)$ measured at the mother node t (before split) and at the two resulting left and right daughter nodes t_L and t_R (after split) weighted by the proportion of instances they contain:

$$G(j, s) = I(t) - \left(\frac{\#t_L}{\#t} I(t_L) + \frac{\#t_R}{\#t} I(t_R) \right), \tag{2.1}$$

where # designates the number of data at the corresponding node.

- For regression, the node information is measured using the empirical L_2 loss

of the response variable

$$I(t) = \frac{1}{\#t} \sum_{i:x^{(i)} \in t} (y^{(i)} - \bar{y}_t)^2,$$

where $\bar{y}_t$ is the mean response value at the node t.

- For classification, the node purity is measured thanks to the Gini index

$$I(t) = \sum_{k=1}^{K} p_{k,t}(1 - p_{k,t}),$$

where $p_{k,t}$ refers to the proportion of instances with class k in node t. The entropy measure is also often encountered and its expression is

$$I(t) = -\sum_{k=1}^{K} p_{k,t} \log(p_{k,t}).$$

Both, Gini and entropy, are alike and provide similar results.

In practice not all s values are tried for potential split values, but only the middle of consecutive points. The criterion performs a greedy search to find the best sub-optimal partition. Sub-optimal is in the sense that a tree is built sequentially, and that the search for the best split is done for each node without taking into account the past unselected possible splits. Moreover, let us mention that splits near the root of the tree are often the more informative (higher decrease of information) compared to splits near the leaves, because based on more data.

A tree is built recursively until either

- all individuals of the data set at a given node have the same response value (the node is pure);
- all individuals have the same covariate values;
- an internal node has less than $N_{\min}$ instances, it then becomes a leaf. $N_{\min}$ is an user-defined integer value, typically set to 1 for classification and 5 for regression.

Other conditions can be encountered such as a maximal tree depth, or getting below an information gain threshold. Usually, the tree is built as deep as possible in order to perform pruning in a second step (see below).

Prediction During a tree construction, the internal node encountering a terminal condition becomes a leaf. A value is assigned to each leaf, corresponding to the prediction provided to data falling in it.

- For regression, the assigned value is the average of the training data responses at the leaf (denoted t):

$$\frac{1}{\#t} \sum_{i:x^{(i)} \in t} y^{(i)}. \tag{2.2}$$

- For classification, the prediction for a tree is performed by majority voting. Each leaf carries the majority class of the training instances in this subspace, that is to say:

$$\underset{1\leq k\leq K}{\arg\max} \sum_{i:x^{(i)}\in t} \mathbb{1}\{y^{(i)} = k\}. \tag{2.3}$$

For a given observed data, that corresponds to a new covariate $X = x^*$, predicting the associated value of the response implies following the path of the binary rules. For a tree, the outcome of the prediction is the allocated value of the leaf where this data set ends after following this path.

Pruning Building a single CART whose purpose is to provide predictions, usually consists in two steps: (1) a maximal tree is trained, where each leaf contains a small number of instances, (2) then a pruning step is performed, where the optimal intermediate tree is searched among all possible sub-trees of the maximal one, by cutting its branches. Indeed, the maximal tree is a predictor with very low bias but high variance (subject to over-fitting), in contrast to the minimal tree (i.e. only the root) that gives an estimator with high bias but low variance as its prediction is a constant (subject to under-fitting) (Dietterich and Kong, 1995). The goal of pruning is to find the intermediate tree T that provides the lowest penalised error, where the regularisation term depends on the number of leaves $|T|$ of the tree. The penalised error criterion is hence

$$err(T) + \lambda|T|,$$

where $err(T)$ is either the average quadratic error or the misclassification error over all leaves, and λ is the penalisation parameter. A high value of λ provides more penalisation toward deep trees (with large $|T|$). It is usually selected by cross-validation. We do not detail pruning further as no pruning is done when training random forest.

Pros and cons of CART A single CART is naturally easily interpretable, in the sense that examining the splitting conditions provide useful information concerning the link between the response and covariates. In addition, the intrinsic splitting procedure – covariates after covariates – makes it suitable for high dimensional problems, where the number of covariates is much larger than the sample size. Nevertheless, to achieve good performance it requires pruning which implies cross-validation procedures. But its major drawback is its instability. Indeed, a small perturbation in the training data set can lead to totally different tree structures. This instability is however beneficial for random forests and justifies their development.

2.3 Random forest construction

Constructing a Breiman's random forest consists in aggregating an ensemble of CARTs whose learning phase contains two additional random aspects detailed in this section.

From a more general point of view, a random forest is composed of an ensemble of B random trees. Each tree is a predictor $h(\cdot, \vartheta)$, where ϑ is a random variable modelling the random aspect of the tree construction. Given a set of B trees, $h(\cdot, \vartheta_1), \ldots, h(\cdot, \vartheta_B)$, where $\vartheta_1, \ldots, \vartheta_B$ are i.i.d., a random forest predictor is deduced in the following way:

- for regression, the random forest predictor is taken as the average of the random tree predictors:
$$h_{RF}(\cdot) = \frac{1}{B} \sum_{b=1}^{B} h(\cdot, \vartheta_b). \tag{2.4}$$
- for classification, it corresponds to the majority class:
$$h_{RF}(\cdot) = \underset{1 \leq k \leq K}{\arg\max} \sum_{b=1}^{B} \mathbb{1}\{h(\cdot, \vartheta_b) = k\}.$$

Note that $h(x^*, \vartheta_b)$ refers to the prediction associated to x^* provided by the b-th tree (Equation 2.2 or 2.3) built with the randomisation aspects described below.

Randomisation As we mentioned above, a random tree is subject to instability. Ensemble learning aims at aggregating a large amount of unstable base learners (trees here), to reduce the final predictor variance and obtain better predictions. A random forest is hence an ensemble learning algorithm. The variability of a CART is strengthen in two ways:

1. each tree is trained on a bootstrap sample (sampling with replacement) of the training set. This approach was originally proposed by Breiman (1996) under the name bagging (for **b**ootstrap **agg**regat**ing**);
2. during a tree construction, at each internal node, a set of m_{try} covariates is uniformly drawn among the d available without replacement, and the splitting criterion (2.1) is maximised using these m_{try} covariates.

In this case, the ϑ_i variables encapsulate the alea in the bootstrap and random sampling of covariates.

Diversity and precision According to Breiman (2001) there are two key ingredients to make a useful random forest: tree diversity, and tree precision. Indeed, when B grows, Breiman deduced an upper bound for the generalisation error of the forest, that depends on the correlation between tree errors (residuals for regression, or margin for classification) and the variance of an individual tree. Thus the error is small when the correlation between the trees is low, which is induced by the tree diversity, and when the predictive performance of each tree is good. Moreover, Hastie, Tibshirani et al. (2009) express the variance of the random forest as the product of the correlations between any pair of trees and individual tree variance. Building a forest thus consists in decorrelating trees without increasing their individual variance too much, in order to obtain a forest with low variance. This is the purpose of using

randomised CART, randomness increases the tree diversity. The parameter m_{try} has for effect to reduce the correlation between trees, so has the bootstrap. Earlier randomisation strategies can be found in the literature, see for example Ho (1998), Dietterich (2000) or again Breiman (2000).

2.4 Performance and convergence

2.4.1 Out-of-bag information

Because each tree is trained on a different bootstrap sample, per tree there is about 37% of the data that remain unused. These left aside data are called out-of-bag data (or out-of-bootstrap), OOB for short, and form a free test sample. In this way, each data can be used as test instance on the trees for which it is OOB. The OOB forest prediction $\hat{y}_{\text{oob}}^{(i)}$ for a data $x^{(i)}$ is hence provided by aggregating all the trees in which it was OOB. An OOB error can hence be computed using these OOB predictions, and in consequence the RF provides a direct measure of its quality.

2.4.2 Tuning in random forests

Some RF parameters values are often considered as default. For m_{try} it is $d/3$ for regression and $\sqrt{d}$ for classification problems. The number of trees, n_{tree}, is 500, and the general rule of thumb is: the higher, the better. The minimal number of data per leaf N_{min} is 5 for regression and 1 for classification. These parameters achieve very satisfying results in practice (e.g. Fernández-Delgado et al., 2014), however we provide below some comments on their tuning if it is of interest.

Number of trees The number of trees n_{tree} should be as high as possible, and its choice is a trade-off between computational time and gain in accuracy. In practice the OOB error can be used to plot this error depending on the number of trees. n_{tree} should be chosen as the value after which the error remains steady, (or a negligible improvement is observed). According to Breiman (2001), increasing the number of trees does not over-fit as the generalised error tends to a limiting value.

Covariate sampling Empirically, m_{try} should not be very large (Breiman, 2001), as its purpose is to reduce the trees correlation, which is accentuated for low values. Genuer, Poggi and Tuleau-Malot (2008) empirically study the influence of this parameter for regression and classification in standard setting ($n >> p$) and high dimensional ($n << p$). The default parameters seem to be better suited for high dimensional problems, even though in classification better results can sometime be obtained for higher m_{try} values.

Minimal number of observations N_{min} should be low, 5 is very close to build maximal tree and should not influence much the prediction quality (Breiman, 2001; Genuer, Poggi and Tuleau-Malot, 2008).

2.4.3 Convergence

Unfortunately, there are very few theoretical results on the Breiman's RF, especially because the random variable ϑ implied in a tree construction depends on the original training data, making demonstrations a lot harder. This is why demonstrations are often provided on simpler tree structures, for which ϑ is independent of the training sample. For example, Biau (2012) demonstrates the consistency of the centred forest (a simplified version of regression RF). Nevertheless, the major breakthrough, in an additive regression setting, brought by Scornet et al. (2015) is the L_2 consistency of the random forest algorithm when subsampling (without replacement) is assumed instead of bootstrap, maximal trees (each leaf has one data) and certain hypothesis. Moreover, both Biau (2012) and Scornet et al. (2015) show that the convergence rates of the associated estimators only depend on the number of strong covariates and not on the total dimension of the covariate space. This makes RF particularly relevant in sparse settings. By sparse we refer to the situation where a large number of covariates (summary statistics for ABC) are considered, without true knowledge of which one are important, but only few are really useful in terms of prediction accuracy. Still assuming subsampling instead of bootstrap the asymptotic normality of forest predictions (when $N \to \infty$) has been shown by Wager and Athey (2018) and Mentch and Hooker (2016), for simplified versions of RF, respectively when $B \to \infty$ or not. For a larger theoretical-oriented review of the random forests, we suggest the well known paper Biau and Scornet (2016).

2.5 Robustness

A valuable characteristic of random forest is its robustness to noise covariates, and its capacity to detect relevant ones. This can be especially useful in the ABC framework.

Detection of relevant variables By construction, a RF sequentially searches for the best covariate to cut the predictor space. It is thus natural to think that irrelevant covariates are ignored, making RF especially well suited for sparse problems. This is supported by Biau (2012), for centred forest, that demonstrates that the convergence rate only depends on relevant variables, and by Scornet et al. (2015) for a slightly modified version of Breiman's algorithm with $m_{\text{try}} = d$ (i.e. bagged trees), showing that splits are selected mostly along informative predictors.

End-cut preference Another useful feature of random forests is the so called end-cut preference of the splitting criterion used for CART construction (Breiman, Friedman et al., 1984). It states that during a tree construction, the best split value (above denoted s) associated to a non-informative explanatory variable is likely to be near its edges. That is to say, s is located near one of the two boundaries of the explanatory variable values. According to Ishwaran (2015), this property is beneficial for random forest. Indeed, when the random sampling of covariates only draws uninformative ones, splitting at an edge reduces the importance of this "bad" split. Moreover, this is also beneficial when an informative variable is selected but

not necessary pertinent in this subspace to explain the response. Splitting near an edge will ensure that the resulting branch with few instances quickly becomes a leaf, allowing the RF to recover from this bad split. It reinforces the robustness of RF to noise variables.

The above comments and the inherent way of maximising the gain criterion make RF able to deal with high dimensional problems, where $n << d$. Moreover, RF is easily parallelizable as the trees are built independently of each other.

2.6 Variable importance

Another attractive feature of RF is the fact that it provides covariate measures of importance of the variables, with respect to the prediction of the response variable. Indeed, it is possible to retrieve for each covariate X_j a measure of its importance giving useful insights concerning their use to explain the response, and even allows the implementation of variable selection methods. There is two commonly used measures: the mean decreased impurity and the mean decreased accuracy.

Mean decreased impurity (MDI) Given the information gain described in equation (2.1), a natural variable importance measure for a given covariate X_j, is to count the decrease of impurity induced by a split using this covariate, weighted by the fraction of the total data at the mother node ($\#t/N$). The variable importance (VI) resulting is hence

$$\mathrm{VI}_{\mathrm{MDI}}(X_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{\substack{t\in T_b \\ j_t^*=j}} \frac{\#t}{N} G(j_t^*, s^*),$$

where T_b refers to the b-th tree, and j_t^* the optimal covariate selected for the node t, with s^* the associated split value. In the extreme case where a covariate is never selected across all trees, its importance is zero. This is lowest achievable variable importance, because the gain is always non-negative. The MDI measure was originally proposed by Breiman (2001), and a CART version can be obtained.

For forests based on totally randomised trees, (an alternative version of CART), Louppe et al. (2013) provide some theoretical results concerning the MDI. In the simplifying case of categorical response and covariates, they derive a decomposition of this measure and pointed out that the importance of a covariate is exactly zero if and only if it is irrelevant. However, this property does not extend to the Breiman's random forest.

Mean decreased accuracy (MDA) Another VI measure is the so called mean decreased accuracy, also known as permutation importance. For a covariate X_j, the idea is to quantify the impact on prediction error when random permutations are performed on the covariate values. The objective is to break the link between X_j and Y and measure its effect. Note that it also breaks the link with other covariates. In this vein, the OOB error is computed for each tree and compared to the OOB

error for the randomised sample, the error differences are then averaged over the B trees. If the covariate is useful, the differences should be large. The MDA has for expression

$$\mathrm{VI}_{\mathrm{MDA}}(X_j) = \frac{1}{B}\sum_{b=1}^{B}\left(\mathrm{err}_{b,j} - \mathrm{err}_b\right),$$

where err_b and $\mathrm{err}_{b,j}$ refer respectively to the OOB error for the b-th tree before and after permutations. Again, if a covariate is never used for data partitioning across all the trees, its importance is zero. Note that this importance measure can very rarely be negative when $\mathrm{err}_{b,j} < \mathrm{err}_b$, meaning that the random permutations have no negative effect and thus that the variable j does not have a role in the prediction.

In general, variable importance measures are affected by the random forest parameters. Naturally, m_{try} has the most impact, as it increases the diversity of VI compared to bagging. When m_{try} is low, greater importance is provided to not necessary important covariates, and when m_{try} increases, a masking effect appears where important covariates are often used and less important ones see their importance decreased (Louppe et al., 2013). A higher number of trees ensures less VI variance from one forest to another.

Variable importance measures were analysed mostly from an empirical point of view, especially the effect of correlated covariates on them. From a theoretical point of view and on simplified models (Gaussian), Gregorutti et al. (2017) highlighted the influence of correlated covariates on the MDA measure. Correlation between covariates is important, as well as the size of the correlated group. They showed that their importance decreases when their correlation increases, the same holds for the size of the group. In this way, a highly correlated informative covariate might present similar importance with a less informative but also not correlated one. This was previously noticed by Archer and Kimes (2008), Auret and Aldrich (2011), Strobl, Boulesteix, Kneib et al. (2008) and Genuer, Poggi and Tuleau-Malot (2010) and especially Toloşi and Lengauer (2011). The two former also relate similar behaviours for MDI. Moreover, Strobl, Boulesteix, Zeileis et al. (2007) pointed out that MDI can be biased when categorical covariates are considered.

Finally, in a regression setting, Ishwaran (2007) aims at providing theoretical results regarding the MDA criterion, thanks to a similar but also simpler variable importance measure, in which prediction of permuted covariates is replaced with noise in the trees (random left-right daughter node assignment).

Note that the MDA is of course more costly than the MDI, moreover this latter is less sensitive to data perturbations compared to the MDA (Calle and Urrea, 2011). Note also that correlation between covariates influences the importance measure, not the predictive performance. Strobl, Boulesteix, Kneib et al. (2008) proposed a variable importance measure avoiding this correlation issues.

Variable selection Let us finally mention that variable importance measures obtained from RF can be used to perform selection of covariates. There are two aims for variable selection: interpretability, by searching the most correlated covariates with the response; prediction, by selecting a small amount of covariates providing the best performance in terms of prediction accuracy. A large battery of approaches

developed, see for example Hapfelmeier and Ulm (2013), Díaz-Uriarte and Alvarez de Andrés (2006), Genuer, Poggi and Tuleau-Malot (2010), Gregorutti et al. (2017), Janitza et al. (2018) and Nembrini et al. (2018) and references therein for more information.

2.7 Implementations

There are many implementations of Breiman's random forest. The original code of Leo Breiman and Adele Cutler was written in the Fortran language, and is freely available at `www.stat.berkeley.edu/~breiman/RandomForests/`. In the following we mostly focus on R versions.

The most known is the R package `randomForest`, adapted from the mentioned Fortran code, by Liaw and Wiener (Liaw and Wiener, 2002). A large panel of packages now exist, for example `Rborist` (Seligman, 2019), `randomForestSRC` (Ishwaran and Kogalur, 2019) or again `ranger` (Wright and Ziegler, 2017). This latter is based on C++ and also provides a standalone version in this language. `ranger` is very fast compared to other packages, as shown in the comparison paper Wright and Ziegler (2017). It notably uses two different strategies to determine the best decision rule, depending on the node size, during a tree construction to avoid runtime bottleneck and memory overflow. Moreover, it provides all the usual information returned by `randomForest` and is continuously optimised and updated with new features. In the remaining when forest training and predictions are needed, we rely on `ranger` for this task.

2.8 Conclusion

Random forest is a very efficient machine learning tool to solve classification and regression problems, and it tends more and more to lose its black-box flavor. Moreover, its scope goes much wider than regression and classification settings. Indeed, it can be used to deal with missing data thanks to proximities between instances computed by RF (Breiman and Cutler, 2003), but also survival analysis (Ishwaran, Kogalur et al., 2008), multivariate regression (Kocev et al., 2007; Segal and Xiao, 2011), quantile computation (Meinshausen, 2006) or handle unbalanced classes (Chen et al., 2004). Recently, Athey et al. (2019) developed the R package `grf` to approximate any solution of a local moment condition thanks to RF. Finally, toward the treatment of "Big Data", in their survey paper, Genuer, Poggi, Tuleau-Malot and Villa-Vialaneix (2017) present different approaches to scale RF to the increasing amount of data. These rely on bootstrap alternatives, subsampling, data partitioning or again parallel computation.

In this chapter we gave a presentation of the RF algorithm, with an emphasised on its qualities, namely its performance with very few tuning parameters, their robustness to irrelevant covariates and their interpretability tools thanks to variable importance measure. These advantages can be greatly beneficial for approximate Bayesian computation approaches as they require tuning parameters and suffer from the curse of dimensionality. This was pointed out in the classification setting by

Pudlo et al. (2016) and successful results were provided. One of our contribution, presented in Chapter 4, is the extension to perform parameter inference.

Chapter 3

Population genetics

Contents

3.1 Introduction

The main application field of this thesis is population genetics, which is concerned with the study of the genetic diversity in populations, their causes and effects. The transmission and evolution of genes over time are of prime interest as they provide insights regarding the evolutionary mechanisms involved. The main goal of population genetics is hence to recover some elements concerning the history of populations. This history is described by a succession of time-ordered inter- or intra-demographic events, called evolutionary scenario. It thus summarises the evolution of an ancestral population until the distribution we today observe.

From present genetic information we would like to fit a model that explains the generating process of such data, while being constrained to a scenario. Gene transmission trees are particularly well suited for this task. In this chapter we especially focus on the Kingman's coalescent (Kingman, 1982a,b,c).

We are working under two assumptions that ease the model construction. We assume that the observed genetic diversity is **only the result of mutations**, i.e. random modifications in the DNA sequence (insertion, deletion –indel– or substitution of nucleotides), that can occur from one generation to another at very low rate. Furthermore, we are working under the **hypothesis of neutrality**, it means a mutation has no effect on the survival of an individual, in other terms there is no selection. Under these hypotheses, the genealogy of genes is independent of the mutational process. It means the construction of a model can be decomposed into two steps. First, build the ancestor-descendent relationships (genealogy) between genes through time, then, add mutations on the genealogy thanks to a mutational process.

Recombination between gene locations (a.k.a. loci) is an important factor that impacts the genealogy of multiple loci. Indeed, when there is no recombination between two loci we can consider they share the same genealogy. When their rate of recombination is very large, their genealogies can be considered as independent and we can build one genealogy per locus. The intermediate case, which is the difficult one, considers that genealogies are correlated (Hudson, 1991). In the remaining we assume that all loci are far apart from each other on the genome so that the **loci are considered independent**, in this case each locus has its own genealogy.

This chapter aims at introducing basic notions for genetic data modelling and simulation, thanks to the Kingman's coalescent, when constrained to an evolutionary scenario. We first give some information regarding the genetic data and scenario linking natural populations (Section 3.2). We explain how to generate genealogies thanks to the coalescent process in an isolated population (Section 3.3), and how it extends when constrained to demographic events (Section 3.4). We then detail the mutational process involved in the derivation of genetic data (Section 3.5) and highlight the inferential difficulties brought by the coalescent, that make it particularly well suited for ABC analysis (Section 3.6). We end by presenting some of the summary statistics that must be computed in the ABC setting for our applications during this manuscript (Section 3.7).

3.2 Data and scenario

3.2.1 Data

For the study of species populations, the data we observe are composed of genetic information obtained from individuals in the populations, at the present time. In each population i, a number n_i of individuals are sampled in the total population size N_i. Each is genotyped to determine its allelic states at some very specific identifiable DNA locations, called DNA markers. A large common set of highly polymorphic markers is of interest to characterise the genetic diversity in populations. Many types of genetic markers exist, during this chapter we focus on **microsatellite** and **single nucleotide polymorphism** (SNP, pronounced *snip*).

Microsatellite or Short Tandem Repeats (STRs) Short tandem repeated sequences consisting of 2-5 base pairs can be observed on DNA sequences. We are interested in the number of repeated motifs, that is called a microsatellite marker. As this number may fluctuate depending on individuals, it exists as many alleles as the number of observable repetitions. It is thus a good indicator of genetic diversity.

Single Nucleotide Polymorphism A SNP is a location on the genome where the nucleotide type change depending on individuals. The DNA sequence of two individuals might differ at some very specific positions in terms of nucleotide, by one base. When more than 1% of the population present the same variation at a given location, this can be considered as a SNP. For haploid individuals, a SNP is encoded by a 0 or a 1 depending on whether or not the allelic state we observe differs from the ancestral one.

Mutational models associated to such types of data are detailed in Section 3.5.2. The data we study thus comes from observations of a large number of DNA markers on individuals. When diploid populations are of interest, each individual carries genes from two parents. For more simplicity, we assimilate a diploid individual to two haploid individuals. A population of N diploids is thus treated as a population of $2N$ haploids.

3.2.2 Evolutionary scenario

The populations we study are linked by their common unknown history we would like to reconstruct. More precisely, some evolutionary scenarios are formulated to summarise the succession of possible demographic events that affected the common ancestral population and led to the present distribution of individuals. An example of scenario involving four present populations (Asian, European, Afro-American and African) is illustrated in Figure 3.1.

We introduce below three types of demographic events we later encounter, all these are occurring instantaneously and more detailed in Section 3.4.

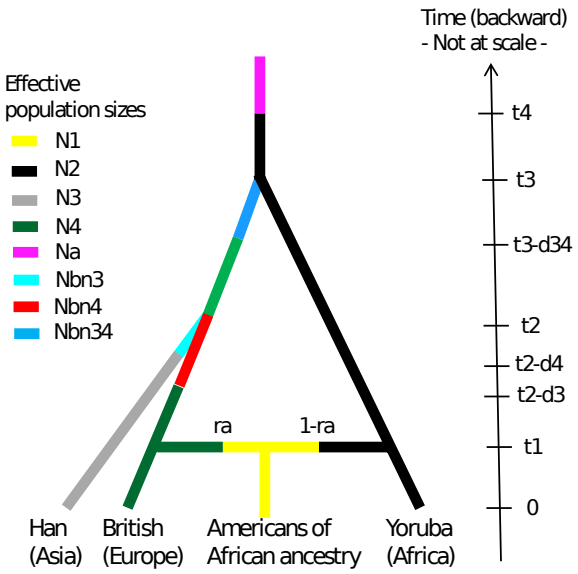


Figure 3.1 – An evolutionary scenario shared by four present populations: Asian, European, Americans of African ancestry and African. See Chapter 4 where this scenario is studied for more information.

- **Change of population size**: the population size changes suddenly at a given time.
- **Divergence**: an ancestral population divides into (at least) two more recent sub-populations.
- **Admixture**: two distantly related populations merge together in some proportions r and $1-r$ to form a new one. r is called the admixture rate.

Given a set of formulated scenarios and corresponding genetic data, there are two main biological problems we would like to deal with:

– **model choice problems**, for which we determine the most likely scenario able to explain the present genetic diversity;

– **parameter inference problems**, for which we estimate parameters as divergence times, admixture rates, population sizes,...

The modelling of the observation must be constrained to reproduce the past demographic events at the proper times, in other terms it must respect the considered evolutionary scenario. As mentioned in the introduction, the first step of the model construction is the generation of a locus genealogy. We explain in the following section how this is done on the simplifying case of an isolated population, we then present how it spreads to mimic the scenario events.

3.3 Genealogy sampling and coalescent

A genealogy is represented by a dendrogram that depicts the evolution of the sample most recent common ancestor (MRCA), through time, until the present. Figure 3.2

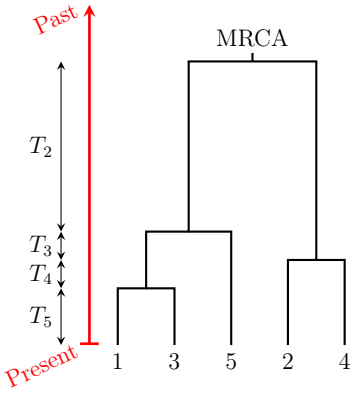


Figure 3.2 – Example of a genealogy for a sample of 5 genes. "MRCA" denotes the most recent common ancestor of the total sample, and the numbers at the bottom of the tree indicate the gene labels. T_k denotes the time to switch from k lineages to $k-1$.

displays an example of genealogy for a sample of 5 genes. Each branch is a gene lineage that can split in two, meaning the gene is duplicated and transmitted to the corresponding offsprings. This perspective from past to present is often mentioned as time-forward, in contrast to the time-backward perspective where the dendrogram is read from present up to the past, until the MRCA. In this case, the event where two lineages find their common ancestor and merge is called a coalescence, and we can notice that a genealogy of an n-sample is made of $n-1$ coalescent events. Note that the genealogy does not carry any information concerning the allelic states. This is explained by the neutral assumption we made, where genetic variations (allelic states) do not impact the survival of the lineage or the number of offsprings, and thus the genealogy structure.

We consider an n-sample, drawn from an haploid population of effective size N, without any geographical or social structure. The effective population size is the number of individuals that effectively participates in producing the next generation. Usually, N is very large compared to n, in the following we assume that this is the case.

Wright-Fisher model The most known model to generate a genealogy is the Wright-Fisher model, (Wright, 1931; Fisher, 1930). It is built on the following assumptions. First, it assumes that time is measured in terms of non-overlapping generations, so that at each generation all individuals in the population die and are replaced by their offsprings. Moreover, the population of size N is constant and finite over generations. Finally, all individuals reproduce at random (the population is panmictic), so that some can produce multiple offsprings, while others can have none. In a time-forward perspective, the whole population evolution is simulated starting from the past. Each new generation is obtained by copying genes from the previous one, a number of times equal to their number of descendants. Then, the genealogy of n genes is obtained by drawing them in the most recent simulated generation, and by going back in time to retrieve their most recent common ancestors (in a backward perspective).

Coalescent model Based on the same assumptions as the Wright-Fisher model, a very simple time-backward approach is given by the coalescent theory (Kingman, 1982a,b,c). The Kingman's coalescent model reconstructs the genes genealogy from the present sample back to the MRCA. It thus consists in generating coalescent events between pairs of lineages, to obtain the wanted dendrogram. We denote T_k the length of time during which we have k lineages, i.e. the waiting time of a coalescence when k lineages are observed.

The probability distribution of the genealogy for n genes is characterised by the choice of the lineages to coalesce and the distribution of the time between coalescent events $T_n, \ldots, T_2$. To derive the waiting time between two coalescences, we start at n lineages and, under the Wright-Fisher assumptions, we investigate the probability to observe a coalescent event in the previous generation. Note that more than two lineages can coalesce one generation in the past under the Wright-Fisher model. Let us denote $P_{i,j}$ the probability that i lineages are descended from j ancestors one generation ago (with $i \geq j$). Then $P_{n,n}$ is the probability of non-coalescence, i.e. the probability that the n lineages have distinct ancestors one generation in the past among the N of the population. The first lineage finds its ancestor with probability 1, the second lineage has $N-1$ out of N changes to have an ancestor different from the first one. The third, has $N-2$ out of N changes to have an ancestor different from the first two, and so on. From such basic reasoning it results that

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

where $\binom{k}{k'}$ is the binomial coefficient, equal to $\frac{k!}{(k-k')!k'!}$ if $k \geq k'$, and 0 otherwise. Similarly, we can calculate the probability to obtain one coalescence between two lineages, $P_{n,n-1}$. There is $\binom{n}{2}$ possible pairs that can coalesce one generation ago, which choose its ancestor among N possible ones. From this, the third lineage has $N-1$ possibilities out of N to have a different ancestor from the two first, and so on. The resulting probability is

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

From the $P_{n,n}$ and $P_{n,n-1}$ expressions, we deduce that the sum of the remaining probabilities $P_{n,k}$ with $k < n-1$ (i.e. the probabilities of more than one coalescence in one generation) is equal to $\mathcal{O}(\frac{1}{N^2})$. Thus, when N is large, we can suppose that observing more that one coalescence in the direct previous generation is a zero

probability event. It results that observing a coalescence in the previous generation can be perceived as a Bernoulli trial with success probability $\binom{n}{2}\frac{1}{N}$.

Under the above approximation, and because coalescent events are independent at each generation, a geometric distribution with parameter $\binom{n}{2}\frac{1}{N}$ describes the waiting time before a coalescence event for an n-sample:

$$T_n \sim \text{Geo}\left(\binom{n}{2}\frac{1}{N}\right).$$

This time is given in number of generations. However, a standard practice is to consider the coalescent time in continuous units, that we denote below by T_n^c. Still under the assumption that N is very large, because $T_n \sim \text{Geo}\left(\binom{n}{2}\frac{1}{N}\right)$ we have the approximation that $T_n^c = \frac{T_n}{N} \sim \text{Exp}(\binom{n}{2})$. In this way, one unit of continuous time is equal to N generations. For more simplicity, we consider the unit of time for which one unit of continuous time is equal to one generation. It results from a change of variable that

$$T_n^c \sim \text{Exp}\left(\binom{n}{2}\frac{1}{N}\right).$$

In the remaining only the continuous coalescent process is used and we discard the c exponent. We can remark that the expected time of coalescence when n genes are studied is equal to $\frac{2N}{n(n-1)}$. Thus, the higher N, the longer the time of coalescence. Moreover, the smaller n, the longer the coalescence time. This explains why a genealogy presents longer branches near its root.

Algorithm 3.1 : Genealogy generation for a single population

Input : the sample size n, the effective population size N

$k \leftarrow n$;

while $k > 1$ **do**

  Simulate the coalescence time T_k from $\text{Exp}\left(\binom{k}{2}\frac{1}{N}\right)$;

  Increase all lineage branches of length T_k;

  Coalesce a random pair of lineages among the $\binom{k}{2}$ possible pairs;

  Decrease the number of lineages k by one: $k \leftarrow k-1$;

end

Algorithm 3.1 describes the generation of a genealogy for n lineages. Branches are increased time-backward by lengths equal to the simulate waiting time of coalescence, and a random pair of lineages coalesce. This is repeated until the sample MRCA.

3.4 Coalescent under demographic events

A species can present demographic structures (geographical or social), different subpopulations with different effective sizes over time, as well as interactions between

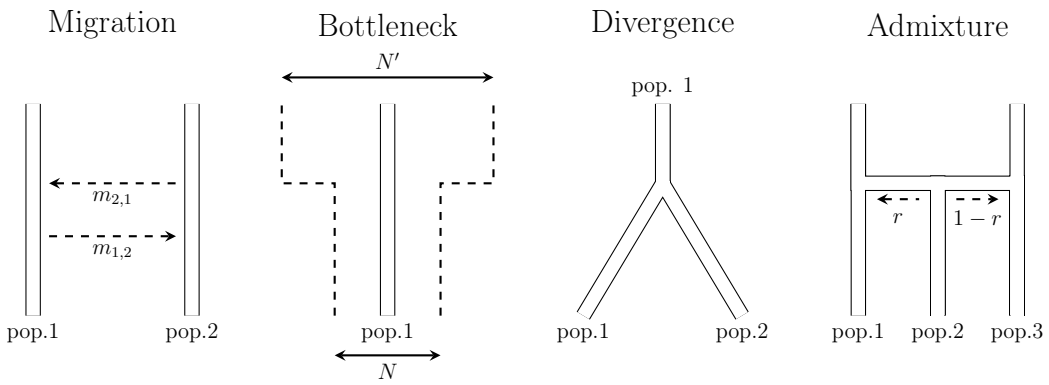


Figure 3.3 – The four principal demographic events: migration, bottleneck, divergence, admixture. $m_{i,j}$ denotes the mutation rate from population i toward j. N and N' are effective population sizes and r is the admixture rate.

them. The coalescent process must take this into account and be modified accordingly to reflect the evolutionary scenario it represents. We consider as possible events: a stepwise change in effective population size, an asymmetric admixture and a divergence illustrated in Figure 3.3, and explain how to generate genealogies mimicking them. These are instantaneous events, and we assume they occur at time t'. Continuous gene flow can also occur, as migration or introgression, where populations give or receive genes over a fixed time period, but we do not present them. Note that an instantaneous introgression event can be seen as a special case of an admixture event, where a population receives a certain percentage of genes from another one.

Change of population size (bottleneck) A natural population, due to a climate disaster or the apparition of a disease in a population (for example), might see its effective population size drop very quickly, barely instantaneously, from a size N' to N. As we mentioned earlier, the smaller the size, the higher the coalescence rate because lineages are more likely to find their common ancestors when they are few of them. This stepwise change of population size hence drives the shape of the genealogy. More evolved approaches consider that the size is changing over time, however we do not present them in this chapter.

Divergence The phenomenon where an ancestral population divides itself into two more recent sub-populations is named divergence. It can model in particular a colonisation, where a small number of individuals leave the main population. This event is often coupled with a bottleneck event. In a time backward perspective, at time t' two populations fuse to generate an ancestral one. Divergence can of course involve more that two populations.

Admixture Two distantly-related populations can merge together in some proportions and form a new one. This gene flow between distant populations is called admixture. A recent genetic admixture example is the African-Americans which result from a mixture between African and Europeans. From present to past, the

new born population is made of a certain percentage r of the first population and $1-r$ of the second, where r is the admixture rate.

When considering only instantaneous demographic events, without continuous gene flux (migration/introgression) between populations, introducing such events in the genealogy construction first requires to build independent genealogies for each population, until time t'. This is done similarly to Algorithm 3.1, for which the genealogy construction is interrupted before t' is exceeded, in this case all branches are increased to this event time. Then the instantaneous event is applied in the following way.

- **Change of population size** The effective population size N is replaced by N', and the subsequent coalescent times are generating using this new size.
- **Divergence** If population i diverged from population j, then move all lineages from population i toward j.
- **Admixture** If population i resulted from an admixture of populations j and k with respective rates r and $1-r$, each lineage in deme i is moved with probability r and $1-r$ toward population j and k respectively.

This is repeated until all scenario events are met and one lineage is obtained (the sample MRCA).

Finally, if continuous migration between populations is of interest, it consists in moving lineages from one population to another over a fixed period of time. In this case, coalescence and migration are competing events, both exponentially distributed. Generating a genealogy under migration consists in simulating exponential realisations corresponding to each event. The shortest simulated time determines which event occurs first. Figure 3.4 illustrates the genealogy shared by 10 genes sampled in two populations in presence of migration. Two migrations events occurred and are represented by horizontal dashed lines.

3.5 Mutational process and data derivation

We now are interested in adding mutations according to a mutational process on the previously simulated genealogy. To derive the genotypes of a sample at a given locus, the MRCA genotype is modified along the genealogy when it encounters a mutation.

3.5.1 Mutation location

Let μ be the rate of mutation of a gene per generation and per locus, i.e. a given gene can overcome a mutation with probability μ in a time period of one generation. This mutation event can be taken into account during the genealogy construction. However, as mutations are considered neutral, we can toss mutations on the genealogy branches depending on their length. To do so, instead of focusing on the dates of mutations, we focus on the number of mutations M_L arising in a time period

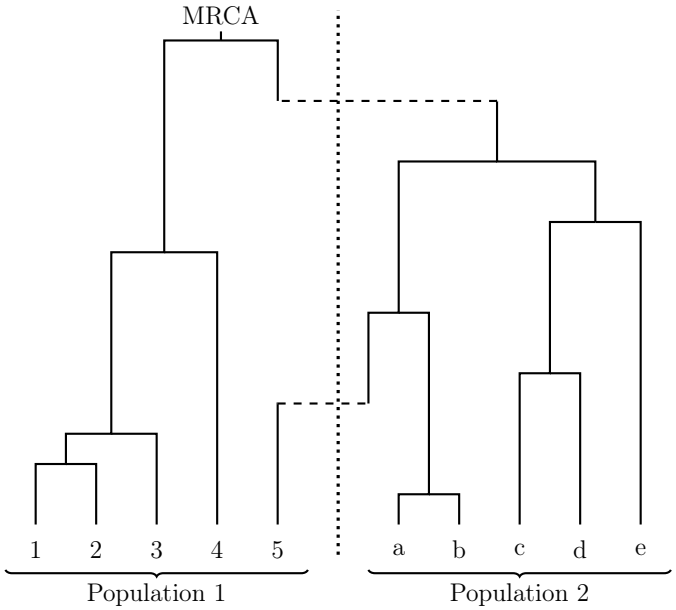


Figure 3.4 – An example of coalescent with migration for two populations. Five genes are samples in each population and migration events are represented by horizontal dashed lines.

of L units of time, that is a branch of length L. Because mutations on different lineages and at different generations arise independently, the Poisson distribution is tailored to model M_L and $M_L \sim \text{Poisson}(L\mu)$. For each branch we can therefore draw the number of mutations from this distribution and then uniformly distribute them along the branch. This describes a Poisson point process of rate μ.

3.5.2 Mutation model

The effect of a mutation depends on the mutational model that is considered for the different DNA markers we study. We present the case of microsatellite and SNP markers (introduced in Section 3.2.1), and how is chosen the ancestral allelic state.

Microsatellite or Short Tandem Repeats (STRs) There are many models to explain how a mutation acts on a microsatellite locus. The simplest is the stepwise mutation model (SMM, Ohta and Kimura, 1973), in which one repeat unit is either gained or lost with equal probability. A generalisation is the generalised mutation model (GSM, Estoup, Jarne et al., 2002). Under this model, a mutation increases or decreases (still with equal probability) the repeated motif number according to a geometric distribution of parameter P. This model better reflects reality as mutations have been observed to change the repeat length by more than one unit. The DIYABC software (Cornuet, Pudlo et al., 2014), sample the ancestral allelic state at random into the allelic range of the observation (or user-specified bounds). Moreover, DIYABC also allows the consideration of an additional mutational process that inserts or deletes a single nucleotide in the microsatellite sequence with equal probability, depending on a mutation parameter denoted μ_{SNI} (see Cornuet, Pudlo et al., 2014).

Single-Nucleotide Polymorphism (SNP) For haploid populations, a SNP data is arbitrary coded by 0 for the ancestral allelic state, which is changed into a 1, for derived state, after a mutation is met. Even though the mutation location can be added to the genealogy as mentioned in Section 3.5.1, the DIYABC software bends this approach a little (following the `ms` algorithm of Hudson (2002)). As a SNP is assumed to always be polymorphic (if not the SNP is discarded from the data) and as only two allelic states are possible (ancestral and derived), it assumes only one mutation occurs in the SNP locus genealogy. The branch carrying the mutation is drawn with probability proportional to its length, the longer it is, the more likely it carries a mutation. Thus this mutational model for SNP does not require parameterisation. This has the advantage of being fast in terms of data generation. However, adding a single mutation on every generated genealogy can be discussed. Indeed, as mentioned in Cornuet, Pudlo et al. (2014) (Appendix S1), a genealogy carrying a mutation should have longer branches compared to one without, hence the genealogy generation should be performed conditionally on having a single mutation on it (by keeping the usual Poisson point process and retaining genealogies for which only one mutation is generated).

3.5.3 Data-derivation

Once a genealogy has been constructed for the sample thanks to the coalescent process, and the mutations have been simulated, the final step of the data-generating process is the deduction of the associated present data. The ancestral allelic state corresponding to the genotype of the MRCA in the genealogy is sampled (if needed), and then passed through the tree until the leaves:

- if a mutation is met, it is applied on the sequence, modifying it according to the mutational process,
- if a coalescence is met, the genotype is duplicated and passed to the two daughter branches.

The simulated sample is observed at the present time. Figure 3.5 illustrates the generation of data for a sample of five lineages and one SNP marker.

3.6 Inferential difficulties

An essential biological interest is to characterise the parameters, denoted θ. From the previous sections we can distinguish three types of parameters:

- historical (admixture times, divergence times,...);
- demographic (admixture rates, effective population sizes,...);
- genetic (mutation rate,...).

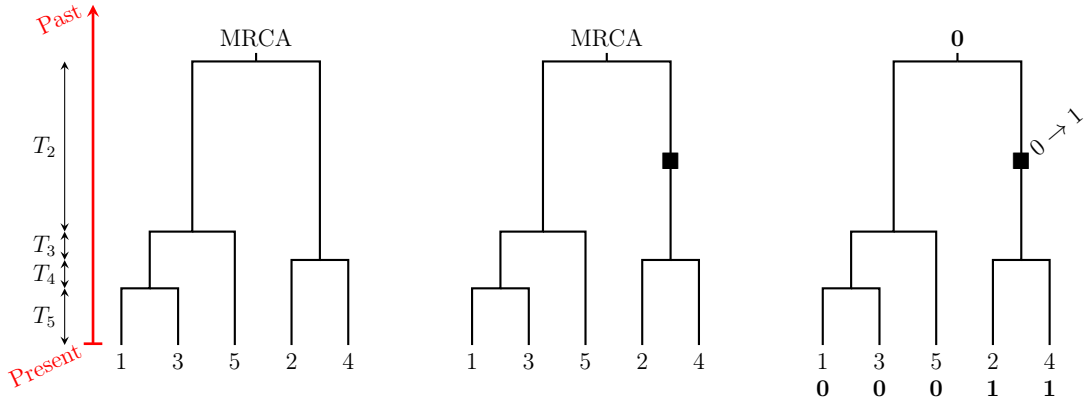


Figure 3.5 – The data-generating process of present genetic data divided in three steps: (1) the construction of a genealogy (left), (2) the addition of mutations on the genealogy (middle), (3) the derivation of the present sample allelic states thanks to the ancestral one (right). This example includes one SNP marker common to five individuals, with a unique mutation on the genealogy.

The objective is to infer such parameters from a polymorphic data observed at the present time, denoted $\mathbf{y}$.

Unfortunately, the likelihood of the coalescent model cannot be computed easily, especially for large sample size. This is due to the presence of latent variables which are the unknown histories that produced the observation. Indeed, let $\mathcal{H}$ be an history, the likelihood is thus expressed by

$$f(\mathbf{y} \mid \theta) = \int f(\mathbf{y}, \mathcal{H} \mid \theta) d\mathcal{H},$$

and the likelihood calculation requires an integration over all possible histories yielding $\mathbf{y}$.

More precisely, we worked under the assumption of loci independence, and for each locus ℓ, a genealogy is built on which mutations are added. The likelihood of the observed data can thus be rewritten as

$$f(\mathbf{y} \mid \theta) = \prod_{\ell} f(\mathbf{y}_\ell \mid \theta) = \prod_{\ell} \int \int f(\mathbf{y}_\ell, \mathcal{M}_\ell, \mathcal{G}_\ell \mid \theta) d\mathcal{M}_\ell d\mathcal{G}_\ell,$$

where $\mathbf{y}_\ell$, $\mathcal{G}_\ell$ and $\mathcal{M}_\ell$ denote respectively the observations, the genealogy and mutation setting at locus ℓ. It means that for each locus ℓ, we need to integrate over all possible genealogies (configurations and branch length) and for each one we need to integrate over all possible mutation placements and natures. For large data sets this is unfeasible as there are too many possibilities that can lead to $\mathbf{y}$.

As mentioned in Chapter 1, simulation techniques have been developed to deal with this intractable likelihood issue, in particular approximate Bayesian computation (ABC). We end this chapter with some details on two key elements for ABC: simulators to generate pseudo-data and summary statistics used to reduce the sample size.

3.7 Simulators and summary statistics

3.7.1 Data simulators

A large choice of softwares and packages exists to simulate genetic data from the coalescent process. The most known is probably the `ms` software of Hudson (2002). It generates DNA segments from the coalescent model, constrained to a large variety of demographic events, in particular migrations and continuous change of population size, even for loci subject to recombination. A recent reimplementation of `ms` is `msprime` (Kelleher et al., 2016) which is more efficient and easier to use thanks to its Python interface.

In the remaining of this manuscript we focus on the DIYABC software (version 2.1.0, Cornuet, Pudlo et al., 2014). It is freely available with a detailed user-manual and example projects for academic and teaching purposes at `www1.montpellier.inra.fr/CBGP/diyabc`. Cornuet, Santos et al. (2008), Cornuet, Ravigné et al. (2010) and Cornuet, Pudlo et al. (2014) developed DIYABC to provide a user-friendly interface, which allows biologists with little background in programming to perform ABC model choice and parameter inference studies. DIYABC is a coalescent-based program which can consider complex population histories through instantaneous demographic events including any number of divergence, admixture (i.e. punctual genetic introgression), and population size variation events. Continuous gene flow (introgression/migration of genes) are not implemented but an instantaneous version can be treated by seeing it has a special case of admixture. The software accepts various types of molecular data (microsatellites, DNA sequences, and SNP) evolving under various mutation models and located on various chromosome types (autosomal, X or Y chromosomes, and mitochondrial DNA). DIYABC also provides a large amount of summary statistics specific to each marker type, we now describe some of them.

3.7.2 Summary statistics computed by DIYABC

We here present some summary statistics for microsatellite and SNP markers, returned by the software DIYABC. These are proposed by biologist experts and computed for a single, a pair or a trio of population samples to allow a rough description of their allelic spectrum. Tables 3.1 and 3.2 provide the complete list of summary statistics with their associated acronyms and references.

Single population Let us describe some summaries of allele diversity in a single population sample. A common measure is the Nei's gene diversity (a.k.a. heterozygosity, Nei, 1987). It represents the probability that two randomly drawn alleles are non-identical.

For microsatellite markers, the gene diversity is averaged over all loci (HET), so is the number of alleles (NAL), as well as their size variance (VAR). The M index at a locus (Garza and Williamson, 2001; Excoffier, Estoup et al., 2005) is computed to characterise a drop in population size. It is the number of alleles divided by the microsatellite range (the difference between the maximal and minimal microsatellite

size). This index is averaged over all loci in the population sample (MWG).

For SNP, the proportion of monomorphic loci is also employed as a summary statistic (HP0), in addition to the mean and variance of heterozygosity across polymorphic loci (HM1 and HV1 respectively).

Two populations When two populations i and j are studied, treating the two populations as one allows to use single population summaries on this joint population. This is the case of the average number of alleles (N2P) or mean gene diversity (H2P), for example. However, genetic distances between two populations are more commonly used to measure the degree of genetic difference between two populations.

For microsatellite, a distance returned by DIYABC (among others) is the so called shared allele distance between the two population samples (DAS, Jin and Chakraborty, 1994). It is expressed as

$$1 - \frac{2n_b}{n_{wi} + n_{wj}},$$

where n_{wk} and n_b denote respectively the average number of alleles shared by pairs of individuals within the same population sample k or two distinct populations. It is one if there is no common allele between the two population samples and zero if $(n_{wi} + n_{wj})/2 = n_b$.

For SNP markers, some DIYABC summary statistics involve the computation of the Nei's distance (Nei, 1972) between two populations i and j at a locus (NP0, NM1, NV1 and NMO).

A common dissimilarity measure for both types of markers at a given locus is the F_{ST} (involved in the statistics FST, FP0, FM1, FV1, FMO). Denoted θ in Weir and Cockerham (1984), it quantifies the degree of genetic diversity due to difference of allele frequencies between populations.

Three populations Finally, to characterise an admixture event including three populations, the admixture rate can be estimated and used as summary statistics (Choisy et al., 2004). Indeed, let us consider that a population i results from an admixture between populations j and k in proportions r and $1 - r$. To infer this admixture rate r, it is possible to compute the likelihood of the genotypes of the admixed samples. This likelihood results from the relationship between the proportion of alleles u, (p_u), in the three populations, and the admixture rate:

$$p_{u,i} = rp_{u,j} + (1 - r)p_{u,k}.$$

From this equation, the likelihood of an hybrid individual genotype at a locus ℓ can be derived. If loci are independent, their product provides the likelihood for a multilocus genotype. The likelihood for the whole hybrid sample is then deduced by the product of the individual ones. DIYABC returns the admixture MLE for all possible combinations of three populations admixture event (AML) or is involved in other summary statistics computation (AP0, AM1, AV1, AMO).

Note that DIYABC allows the consideration of groups of loci, to define different mutational process for each one. The loci group is hence introduced in the summary

Type of statistics	Summary statistics	Acronym	References
Single sample statistics across loci	Mean number of alleles	NAL	
	Mean gene diversity	HET	Nei, 1987
	Mean allele size variance	VAR	
	Mean M index	MWG	Garza and Williamson, 2001, Excoffier, Estoup et al., 2005
Two sample statistics across loci pooling two samples	Mean number of alleles	N2P	
	Mean gene diversity	H2P	
	Mean allele size variance	V2P	
Two sample statistics	F_{ST} between two samples	FST	Weir and Cockerham, 1984
	Mean index of classification	LIK	Rannala and Mountain, 1997, Pascual et al., 2007
	Shared allele distance	DAS	Jin and Chakraborty, 1994
	$(\delta\mu)^2$ distance	DM2	Goldstein et al., 1995
Three sample statistics	Maximum likelihood coefficient of admixture	AML	Choisy et al., 2004

Table 3.1 – Summary statistics available in the DIYABC software for microsatellite markers, computed on one, two or three population samples, with the corresponding acronym and references.

statistic acronym (Tables 3.1 and 3.2). For instance, for group 1, the maximum likelihood coefficient of admixture, when population i comes from an admixture event between populations j and k, is denoted AML_1_$i.j.k$. In the following, no more than 1 group is studied. To get an idea of the number of summary statistics computed by DIYABC, when five populations are studied, for microsatellite markers a total of 130 summary statistics can be obtained, compared to 220 for SNP markers.

3.8 Conclusion

To handle genetic data and resolve population genetics problems, the Kingman's coalescent is widely used. In this chapter we presented its principle and how to constrain it to respect an evolutionary scenario. While direct inferences by means of the likelihood are unfeasible in much cases, simulations can be performed easily and ABC methods are a more suitable solution.

The content of this chapter is essential for the better understanding of the examples, displayed in the different chapters of this manuscript. Indeed, population genetics applications are a major part of our work. The data simulation process thanks to the Kingman's coalescent and how to summarise the information are important aspects.
Chapter 4 contains parameter inference analyses performed using $50,000$ SNP markers genotyped in four human populations (African, East Asian, European, North American, The 1000 genomes project consortium, 2012). In Chapter 6, we present two improvements for the ABC random forest methodologies, both are supported by complex case-studies using sets of microsatellite markers. The first one concerns a scenario choice problem, involving four African Pygmy populations and their non-Pygmy neighbours (Estoup, Raynal et al., 2018). The second study is about the reconstruction of the evolutionary past of two desert locust subspecies (*Schistocerca gregaria*). It displays ABC-RF scenario choice and parameter inference analyses,

Type of statistics	Summary statistics	Acronym	References
Single sample statistics across loci	Proportion of monomorphic loci	HP0	
	Mean gene diversity (on polymorphic loci)	HM1	Nei, 1987
	Variance of gene diversity (on polymorphic loci)	HV1	
	Mean gene diversity	HMO	
Two sample statistics across loci	Proportion of loci with null F_{ST} distances	FP0	Weir and Cockerham, 1984
	Mean of non null F_{ST} distances	FM1	
	Variance of non null F_{ST} distances	FV1	
	Mean of F_{ST} distances	FMO	
	Proportion of loci with null Nei's distances	NP0	Nei, 1972
	Mean of non null Nei's distances	NM1	
	Variance of non null Nei's distances	NV1	
	Mean of Nei's distances	NMO	
Three sample statistics across loci	Proportion of loci with null admixture estimates	AP0	Choisy et al., 2004
	Mean of non null admixture estimates	AM1	
	Variance of non null admixture estimates	AV1	
	Mean of admixture estimates	AMO	

Table 3.2 – Summary statistics available in the DIYABC software for SNP markers, computed on one, two or three population samples, with the corresponding acronym and references.

results regarding the two subsequent developments, to finally propose interesting hypothesis linking our results to known past events (Chapuis, Raynal et al., 2019).

Part II

Contributions

Chapter 4

ABC random forests for Bayesian parameter inference

This chapter is based on our article Raynal et al. (2019), published in *Bioinformatics*.

Contents

4.1 Introduction

Posterior distributions are the cornerstone of any Bayesian analysis as they constitute both a sufficient summary of the data and a means to deliver all aspects of inference, from point estimators to predictions and uncertainty quantification. However, it is rather common that practitioners and users of Bayesian inference are not directly interested in the posterior distribution *per se*, but rather in some summary aspects, like posterior mean, posterior variance or posterior quantiles, since these are easier to interpret and report. With this motivation, we consider a version of ABC focusing on the approximation of unidimensional transforms of interest like the above, instead of resorting to the classical ABC approach that aims at approximating the entire posterior distribution and then handling it as in regular Bayesian inference. The approach we study here is based on random forests (RF, Breiman, 2001), which produces non-parametric regressions on an arbitrary set of potential regressors. We recall that the calibration side of RF (i.e. the choice of the RF parameters: typically the number of trees and the number of summary statistics sampled at each node) was successfully exploited in Pudlo et al. (2016) for conducting ABC model choice.

Let $\{f(\mathbf{y} \mid \theta)\colon \mathbf{y} \in \mathcal{Y}, \theta \in \Theta\}$, $\mathcal{Y} \subseteq \mathbb{R}^n$, $\theta \subseteq \mathbb{R}^p, p, n \geq 1$ be a parametric statistical model and $\pi(\theta)$ be a prior distribution on the parameter θ. Given an observation (or sample) $\mathbf{y}$ issued from this model, Bayesian parameter inference is based on the posterior distribution $\pi(\theta \mid \mathbf{y}) \propto \pi(\theta) f(\mathbf{y} \mid \theta)$. The computational difficulty addressed by ABC techniques is that a numerical evaluation of the density (a.k.a., likelihood) $f(\mathbf{y} \mid \theta)$ is impossible or at least very costly, hence preventing the derivation of the posterior $\pi(\theta \mid \mathbf{y})$, even by techniques like MCMC (Marin and Robert, 2014).

In Chapter 1 we exposed the basic ABC principles, and the calibration aspects these methods require, as well as the random forest algorithms and its advantages in Chapter 2. We now explain how to fuse both methodologies towards Bayesian inference about parameters of interest. We then illustrate the performance of our proposal and compare it with earlier ABC methods on a normal toy example and a population genetics example dealing with human population evolution.

We first recall in Algorithm 4.1 how a reference table is generated. η still denotes the set of d summary statistics and $\mathbf{x}^{(i)}$ is a data simulated from the generative model. Such a reference table will later be used as a training data set for the different RF methods explained below.

Algorithm 4.1 : Generation of a reference table of size N

for $i \leftarrow 1$ **to** N **do**
  Simulate $\theta^{(i)} \sim \pi(\cdot)$;
  Simulate $\mathbf{x}^{(i)} \sim f\left(\cdot \mid \theta^{(i)}\right)$;
  Compute $\eta_{\mathbf{x}^{(i)}} = \{\eta_{\mathbf{x}^{(i)},1}, \ldots, \eta_{\mathbf{x}^{(i)},d}\}$;
end

4.2 ABC parameter inference using random forest

4.2.1 Motivations and main principles

The particular choice of RF as a (non-parametric) estimation method in a regression setting is justified by the robustness of both random forests to "noise", that is, to the presence of irrelevant predictors, even when the proportion of such covariates amongst the entire set of proposed predictors is substantial (Marin, Pudlo, Estoup et al., 2018). By comparison, the method of k-nearest neighbour classifiers lacks such characteristics (Biau, Cérou et al., 2015). In the setting of building an ABC algorithm without preliminary selection of some summary statistics, our conjecture is that RF allows for the inclusion of an arbitrary and potentially large number of summary statistics in the derivation of the forest and therefore that it does not require the usual preliminary selection of summary statistics. When implementing this approach, we hence bypass the selection of summary statistics and include a large collection of summary statistics, some or many of which being potentially poorly informative if not irrelevant for the regression.

A regression RF produces an expected predicted value for an arbitrary transform of θ, conditional on an observed data set. This prediction is the output of a piecewise constant function of the summary statistics. RF aggregates trees, partitions the feature space (here the space of summary statistics) in a way tuned to the forecast of a scalar output, i.e. a one dimensional functional of the parameter. This partition and prediction are done without requiring the definition of a particular distance on the feature space and is hence not dependant of any type of tolerance level. From an ABC perspective, each tree of a RF provides a partition of the covariate space, in our case the d-dimensional space of summary statistics, adapted for the forecasting of the response variable, corresponding to a scalar transformation $h(\theta)$ of the parameter θ. In the following section we present how to compute quantities of interest in a context of parameter inference, thanks to the calculation of weights.

4.2.2 Calculation of weights and approximation of the posterior expectation

Assume we have now grown a RF made of B trees that predicts $\tau = h(\theta) \in \mathbb{R}$ using the summarised observed data set $\eta_{\mathbf{y}}$ and the training sample $\{(\tau^{(i)}, \eta_{\mathbf{x}^{(i)}})\}_{i=1,\ldots,N}$, where $\tau^{(i)} = h(\theta^{(i)})$. In the examples below, we will consider the case where h is the projection on a given coordinate of θ. To sum up, we are training a RF using simulated data sets from the reference table, where the covariates are the summary statistics and the response variable is a unidimensional parameter of interest. Each of these B trees produces a partition of the space of summary statistics, with a constant prediction of the expected value of τ on each set of the partition. More precisely, given the b-th tree in the forest, let us denote $n_b^{(i)}$ the number of times the pair $(\tau^{(i)}, \eta_{\mathbf{x}^{(i)}})$ is repeated in the bootstrap sample that is used for building the b-th tree. Note that $n_b^{(i)}$ is equal to zero when the pair does not belong to the bootstrap sample. These pairs form the out-of-bag sample of the b-th tree. Now, let $L_b(\eta_{\mathbf{y}})$ denote the leaf reached after following the path of binary choices given by the tree,

which depends on the value of $\eta_{\mathbf{y}}$. The number of items of the bootstrap sample that fall in that leaf is

$$\left|L_b(\eta_{\mathbf{y}})\right| = \sum_{i=1}^{N} n_b^{(i)} \mathbb{1}\left\{\eta_{\mathbf{x}^{(i)}} \in L_b(\eta_{\mathbf{y}})\right\},$$

where $\mathbb{1}$ denotes the indicator function, and the mean value of τ of that leaf of the b-th tree is

$$\frac{1}{\left|L_b(\eta_{\mathbf{y}})\right|} \sum_{i=1}^{N} n_b^{(i)} \mathbb{1}\left\{\eta_{\mathbf{x}^{(i)}} \in L_b(\eta_{\mathbf{y}})\right\} \tau^{(i)}.$$

Averaging these B predictions of τ leads to an approximation of the posterior expected value of τ, also denoted mean value of τ, which can be written as follows:

$$\widetilde{\mathbb{E}}\big(\tau\big|\eta_{\mathbf{y}}\big) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{N} \frac{1}{\left|L_b(\eta_{\mathbf{y}})\right|} n_b^{(i)} \mathbb{1}\left\{\eta_{\mathbf{x}^{(i)}} \in L_b(\eta_{\mathbf{y}})\right\} \tau^{(i)}.$$

This is the classic random forest prediction in the regression setting (Chapter 2, Equation 2.4).

As exhibited by Meinshausen (2006), the above can be seen as a weighted average of τ along the whole training sample of size N made by the reference table. In fact, the weight of the i-th pair $(\tau^{(i)}, \eta_{\mathbf{x}^{(i)}})$ given $\eta_{\mathbf{y}}$ is

$$w_i(\eta_{\mathbf{y}}) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{\left|L_b(\eta_{\mathbf{y}})\right|} n_b^{(i)} \mathbb{1}\left\{\eta_{\mathbf{x}^{(i)}} \in L_b(\eta_{\mathbf{y}})\right\}.$$

4.2.3 Approximation of the posterior quantile and variance

The weights $w_i(\eta_{\mathbf{y}})$ provide an approximation of the posterior cumulative distribution function (c.d.f.) of τ given $\eta_{\mathbf{y}}$ as

$$\widetilde{F}(\tau \mid \eta_{\mathbf{y}}) = \sum_{i=1}^{N} w_i(\eta_{\mathbf{y}}) \mathbb{1}\left\{\tau^{(i)} < \tau\right\}.$$

Posterior quantiles, and hence credible intervals, are then derived by inverting this empirical c.d.f., that is by plugging $\widetilde{F}$ in the regular quantile definition

$$\widetilde{\mathbb{Q}}_\alpha(\tau \mid \eta_{\mathbf{y}}) = \inf\left\{\tau : \widetilde{F}(\tau \mid \eta_{\mathbf{y}}) \geq \alpha\right\}.$$

This derivation of a quantile approximation is implemented in the R package `quantregForest` and the consistency of $\widetilde{F}$ is established in Meinshausen (2006).

An approximation of $\mathbb{V}(\tau \mid \mathbf{y})$ can be derived in a natural way from $\widetilde{F}$, leading to

$$\widehat{\mathbb{V}}(\tau \mid \eta_{\mathbf{y}}) = \sum_{i=1}^{N} w_i(\eta_{\mathbf{y}}) \left(\tau^{(i)} - \sum_{u=1}^{N} w_u(\eta_{\mathbf{y}}) \tau^{(u)}\right)^2.$$

4.2.4 Alternative variance approximation

Regarding the specific case of the posterior variance of τ, we propose a slightly more involved albeit manageable version of a variance estimate. Recall that, in any given tree b, some entries from the reference table are not included since each tree relies on a bootstrap sample of the training data set. The out-of-bag simulations, i.e. unused in a bootstrap sample, can be exploited toward returning an approximation of $\mathbb{E}(\tau \mid \eta_{\mathbf{x}^{(i)}})$, denoted $\hat{\tau}^{(i)}_{\text{oob}}$. Indeed, given a vector of summary statistics $\eta_{\mathbf{x}^{(i)}}$ of the training data set, passing this vector down the ensemble of trees where it has not been used and mean averaging the associated predictions provide such an approximation. Since

$$\mathbb{V}(\tau \mid \eta_{\mathbf{y}}) = \mathbb{E}\left([\tau - \mathbb{E}(\tau \mid \eta_{\mathbf{y}})]^2 \mid \eta_{\mathbf{y}} \right) ,$$

we advocate applying the original RF weights $w_i(\eta_{\mathbf{y}})$ to the out-of-bag square residuals $(\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})^2$, which results in the alternative approximation

$$\widetilde{\mathbb{V}}(\tau \mid \eta_{\mathbf{y}}) = \sum_{i=1}^{N} w_i(\eta_{\mathbf{y}})(\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})^2.$$

Indeed, as presented below in Section 4.3.1.1, an estimator of the posterior variance can be obtained by training a new forest on the out-of-bag squared residuals, resulting into a weighted sum of these quantities. This estimator consists in replacing these new weights by the one from the first forest, assuming these are good enough to approximate any posterior expectations: $\mathbb{E}(g(\theta_j) \mid \eta_{\mathbf{y}})$, where g is a function of θ_j. A comparison between different variance estimators is detailed in Section 4.3.1.1. Owing to the results of this comparative study, we choose to use the above alternative variance estimator when presenting the results from two examples.

4.2.5 Approximation of posterior covariances

We are here interested in another estimate that is frequently produced in a Bayesian analysis, that is the posterior covariance between two univariate transforms of the parameter, $\tau = h(\theta)$ and $\sigma = g(\theta)$, say $\text{Cov}(\tau, \sigma \mid \eta_{\mathbf{y}})$. Since we cannot derive this quantity from the approximations to the marginal posteriors of τ and σ, we propose to construct a specific RF for this purpose. We denote approximations of posterior expectations for τ and σ, produced by out-of-bag information, by $\hat{\tau}^{(i)}_{\text{oob}}$ and $\hat{\sigma}^{(i)}_{\text{oob}}$. We use the product of out-of-bag errors for τ and σ in the empirical covariance, and consider $(\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})(\sigma^{(i)} - \hat{\sigma}^{(i)}_{\text{oob}})$ as the response variable. With the previously introduced notations, the corresponding RF estimator is

$$\widetilde{\text{Cov}}(\tau, \sigma \mid \eta_{\mathbf{y}}) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{\left| L_b(\eta_{\mathbf{y}}) \right|} \sum_{i : \eta_{\mathbf{x}^{(i)}} \in L_b(\eta_{\mathbf{y}})} n_b^{(i)} (\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})(\sigma^{(i)} - \hat{\sigma}^{(i)}_{\text{oob}}).$$

This posterior covariance approximation requires a total of three regression RFs: one for each parameters and one for the covariance approximation.

4.2.6 A new R package for conducting parameter inferences using ABC-RF

When several parameters are jointly of interest, our recommended global strategy consists in constructing one independent RF for each parameter of interest and estimate from each RF several summary measurements of the posterior distribution (i.e. posterior expectation, quantiles and variance) of each parameter. However, if one is interested in estimating the posterior covariance between pair of parameters, an additional RF is required. Our R library `abcrf` was initially developed for Bayesian model choice using ABC-RF as in Pudlo et al. (2016). The version 1.7.1 of `abcrf` includes all the methods proposed in this chapter to estimate posterior expectations, quantiles, variances (and covariances) of parameter(s). `abcrf` version 1.7.1 is available on CRAN. We provide in Appendix A.1, a commented R code that will allow non-expert users to run random forest inferences about parameters using the `abcrf` package.

4.3 Results

We illustrate the performances of our ABC-RF method for Bayesian parameter inference on a normal toy example and on a realistic population genetics example. In the first case, approximations of posterior quantities can be compared with their exact counterpart. For both examples, we further compare the performances of our methodology with those of earlier ABC methods based on solely rejection, adjusted local linear (Beaumont, Zhang et al., 2002), ridge regression (Blum, Nunes et al., 2013), adjusted neural networks (Blum and François, 2010), and adaptive PMC (ABC-PMC, Beaumont, Cornuet et al., 2009; Prangle, 2017). Moreover, we carried out additional comparisons between ABC-RF, adaptive ABC-PMC (Beaumont, Cornuet et al., 2009; Prangle, 2017), ABC-SMC (Del Moral et al., 2012) and adaptive ABC-SMC (Klinger and Hasenauer, 2017) methods for various tuning parameters. Due to excessive computational heaviness and in agreement with the content of the results obtained on the normal toy example, we did not extended the latter comparisons to the population genetics example. Normalised mean absolute errors (NMAE) are used to measure performance on test data sets, the normalisation being done by dividing the absolute error by the true value of the target, these are then averaged to provide the NMAE. A normalised version offers the advantage of being hardly impacted when only a few observations get poorly predicted.

For both illustrations, RFs were trained based on the functions of the R package `ranger` (Wright and Ziegler, 2017) with forests made of $B = 500$ trees, with $m_{\text{try}} = d/3$ selected covariates (i.e. summary statistics) for split-point selection at each node, and with a minimum node size equals to 5 (Breiman, 2001, and see Section 4.3.3, Practical recommendations regarding the implementation of the ABC-RF algorithm). The other ABC methods in the comparison were based on the same reference tables, calling the corresponding functions in the R package `abc` (Csilléry et al., 2012; Nunes and Prangle, 2015) with its default parameters. ABC with neural network adjustment uses one hidden layer, but multiples networks are trained so that the final prediction is the median of the neural networks predictions, their

number is here equal to 10, the default value in the R package `abc`. A correction for heteroscedasticity is applied by default when considering regression adjustment approaches. Note that regression corrections are univariate for local linear and ridge regression as well as for RF, whereas neural network – by construction – performs multivariate corrections.

4.3.1 Normal toy example

We consider the hierarchical normal mean model

$$\begin{aligned} y_j \mid \theta_1, \theta_2 &\sim \mathcal{N}(\theta_1, \theta_2), \\ \theta_1 \mid \theta_2 &\sim \mathcal{N}(0, \theta_2), \\ \theta_2 &\sim I\mathcal{G}(4, 3), \end{aligned}$$

where $I\mathcal{G}(\kappa, \lambda)$ denotes an inverse Gamma distribution with shape parameter κ and scale parameter λ. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a n-sample from the above model. Given these conjugate distributions, the marginal posterior distributions of the parameters θ_1 and θ_2 are closed-forms:

$$\begin{aligned} \theta_1 \mid \mathbf{y} &\sim \mathcal{T}\left(n+8, \frac{n\bar{\mathbf{y}}}{n+1}, \frac{2\left(3 + s^2/2 + n\bar{\mathbf{y}}^2/(2n+2)\right)}{(n+1)(n+8)}\right) \\ \theta_2 \mid \mathbf{y} &\sim I\mathcal{G}\left(\frac{n}{2} + 4, 3 + \frac{s^2}{2} + \frac{n\bar{\mathbf{y}}^2}{2n+2}\right), \end{aligned}$$

where $\bar{\mathbf{y}}$ is the sample mean and $s^2 = \sum_{j=1}^{n}(y_j - \bar{\mathbf{y}})^2$ the sum of squared deviations. $\mathcal{T}(\nu, a, b)$ denotes the general t distribution with ν degrees of freedom (Marin and Robert, 2014).

From the above expressions and for a given sample $\mathbf{y}$, it is straightforward to derive the exact values of $\mathbb{E}(\theta_1 \mid \mathbf{y})$, $\mathbb{V}(\theta_1 \mid \mathbf{y})$, $\mathbb{E}(\theta_2 \mid \mathbf{y})$, $\mathbb{V}(\theta_2 \mid \mathbf{y})$ and posterior quantiles for the two parameters. This provides us with a benchmark on which to assess the performances of ABC-RF. For the present simulation study, we opted for a reference table made of $N = 10^4$ replicates of a sample of size $n = 10$ and $d = 61$ summary statistics. Those statistics included the sample mean, the sample variance, the sample median absolute deviation (MAD), all possible sums and products with these three elements resulting in eight new summary statistics and 50 additional independent (pure) noise variables that were generated from a uniform $\mathcal{U}_{[0;1]}$ distribution. The performances of our method were evaluated on an independent test table of size $N_{\text{pred}} = 100$, produced in the same way as the reference table. Current ABC methods (rejection, adjusted local linear, ridge and neural network) all depend on the choice of a tolerance level p_ϵ corresponding to the proportion of selected simulated parameters with lowest distances between simulated and observed summary statistics. On this example we consider a tolerance level of $p_\epsilon = 0.01$ for ABC with rejection, and $p_\epsilon = 0.1$ for the ABC methods with adjustment. We also compare estimation results obtained from the adaptive ABC-PMC algorithm described in Prangle (2017) (Algorithm 5 of his paper). We implement two designs of this scheme with both $2,000$ simulated particles per iteration, $1,000$ accepted particles, schemes iterate until we get approximately 10^4 (a-PMC-1) and 10^5 (a-PMC-2) simulated particles. Finally, we carry out additional comparisons

with two sequential ABC methods: ABC-PMC based on the algorithm of Beaumont, Cornuet et al. (2009) and ABC-SMC based on the algorithm of Del Moral et al. (2012) (both presented in Chapter 1). Two different implementations of ABC-PMC (named PMC-1 and PMC-2) are considered. PMC-1 and PMC-2 include $1,000$ and 100 simulated particles per iteration, 100 and 10 accepted particles and 10 and 100 iterations, respectively, resulting in 10^4 simulated particles. Two different implementations of ABC-SMC (named SMC-1 and SMC-2) are considered. Both SMC-1 and SMC-2 include $1,000$ simulated particles per iteration and a stopping rule based on two pre-computed quantiles of the distances between the observed summary statistics and simulated ones. For SMC-1, we use a quantile of 10% and for SMC-2 a quantile of 1%. At the end, we simulate approximately $40,000$ particles for SMC-1 and $360,000$ for SMC-2. Furthermore, we also compare with an ABC-SMC scheme (named py-SMC-1 and py-SMC-2) using adaptive population sizes described in Klinger and Hasenauer (2017), using the `pyABC` Python module (Klinger, Rickert et al., 2018). An initial population size equal to $1,000$ is used and two different values of a target density variation. An additional version of ABC-PMC (named py-PMC-1 and py-PMC-2) is implemented using this module to mimic the PMC-1 and PMC-2 designs. In some comparisons, we consider two different situations by including or not a large number of noise variables (50 or 500 noise variables drawn into uniform distributions on $[0;1]$) as explanatory variables. The R codes are available at https://github.com/jmm34/abc-rf-param.

Figure 4.1 compares the exact values $\psi_1(\mathbf{y}) = \mathbb{E}(\theta_1 \mid \mathbf{y})$, $\psi_2(\mathbf{y}) = \mathbb{E}(\theta_2 \mid \mathbf{y})$, $\psi_3(\mathbf{y}) = \mathbb{V}(\theta_1 \mid \mathbf{y})$ and $\psi_4(\mathbf{y}) = \mathbb{V}(\theta_2 \mid \mathbf{y})$ with the estimates obtained from the ABC-RF approach. It shows that the proposed estimators have good overall performances for both $\psi_1(\mathbf{y})$ and $\psi_2(\mathbf{y})$. Our estimators perform less satisfactorily for both $\psi_3(\mathbf{y})$ and $\psi_4(\mathbf{y})$ but remain acceptable. Figure 4.2 shows furthermore that the quantile estimation are good for θ_1 if less accurate for θ_2.

We then run an experiment to evaluate the precision of the marginal posterior approximation provided by ABC-RF for the parameter θ_1, using two different test data sets and 40 independent reference tables. As exhibited in Figure 4.3, results are mixed. For one data set, the fit is quite satisfactory, with the RF approximation showing only slightly fatter tails than the true posterior density distribution function (Figure 4.3; upper panel). For the other data set, we obtain stronger divergence both in location and precision of the posterior density distribution function (Figure 4.3; lower panel).

Using the same reference table, we now compare our ABC-RF results with a set of five earlier ABC methods, namely, ABC methods based on straightforward rejection, adjusted local linear, ridge regression, adjusted neural networks and adaptive ABC-PMC. Table 4.1 shows that the ABC-RF approach leads to results better than all other ABC methods for all quantities of interest. Expectations and quantiles are noticeably more accurately predicted. Figure 4.4 compares differences between estimated and true values of the posterior variances $\psi_3(\mathbf{y})$, $\psi_4(\mathbf{y})$. It shows the global underestimation associated with rejection and ABC methods with adjustment, when compared to ABC-RF, the latter only slightly overestimating the posterior variance. The adaptive ABC-SMC method performs very decently for $\psi_4(\mathbf{y})$, however highly overestimates $\psi_3(\mathbf{y})$. Finally, by looking at the width of the boxplots of Figure 4.4, we deduce that our ABC-RF estimations exhibit a lower estimation variability.

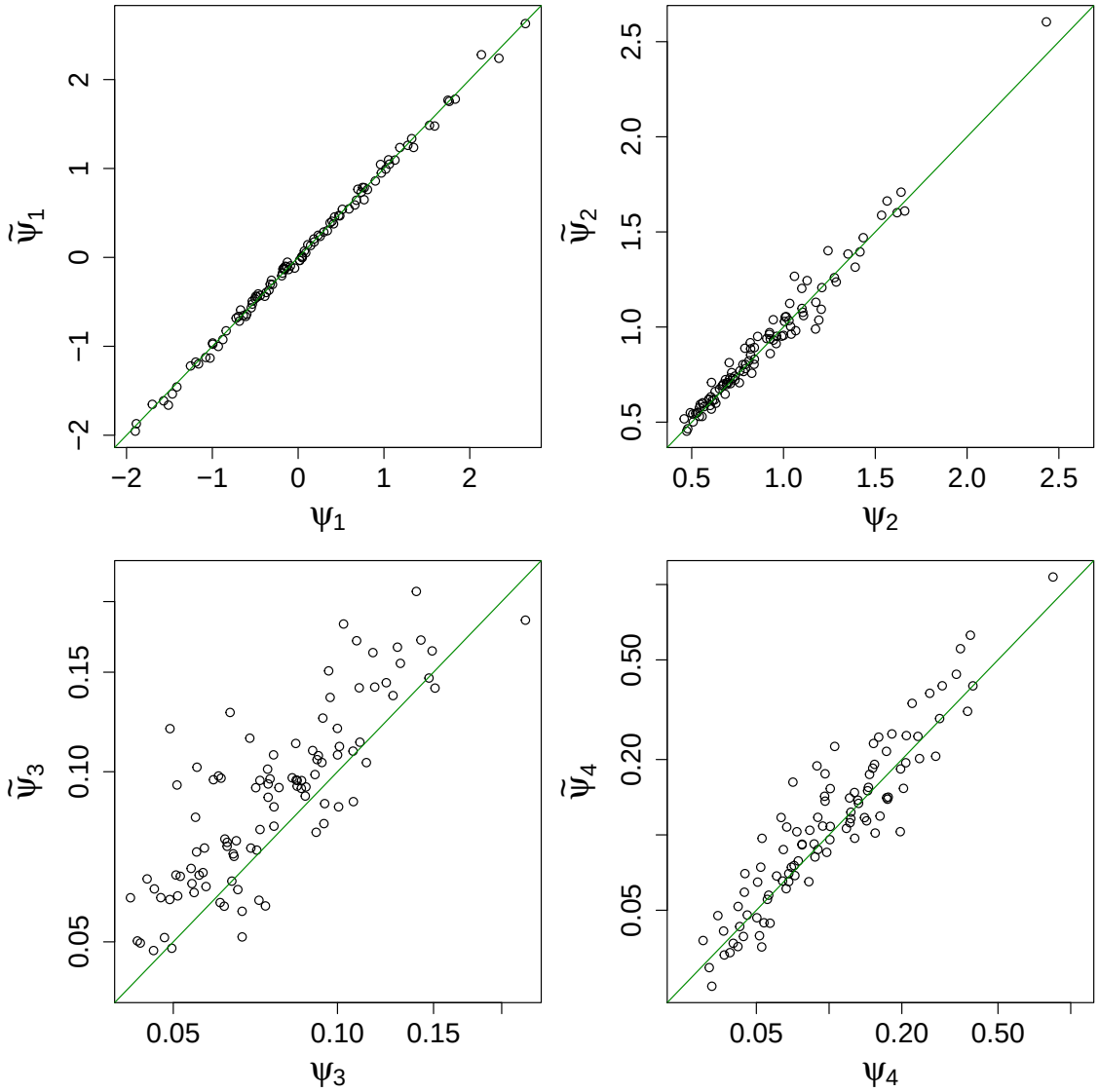


Figure 4.1 – Scatterplot of the theoretical values $\psi_1(\mathbf{y}) = \mathbb{E}(\theta_1 \mid \mathbf{y})$, $\psi_2(\mathbf{y}) = \mathbb{E}(\theta_2 \mid \mathbf{y})$, $\psi_3(\mathbf{y}) = \mathbb{V}(\theta_1 \mid \mathbf{y})$ and $\psi_4(\mathbf{y}) = \mathbb{V}(\theta_2 \mid \mathbf{y})$ for the normal model with their corresponding estimates $\widetilde{\psi}_1$, $\widetilde{\psi}_2$, $\widetilde{\psi}_3$, $\widetilde{\psi}_4$ obtained using ABC-RF. Variances are represented on a log scale.

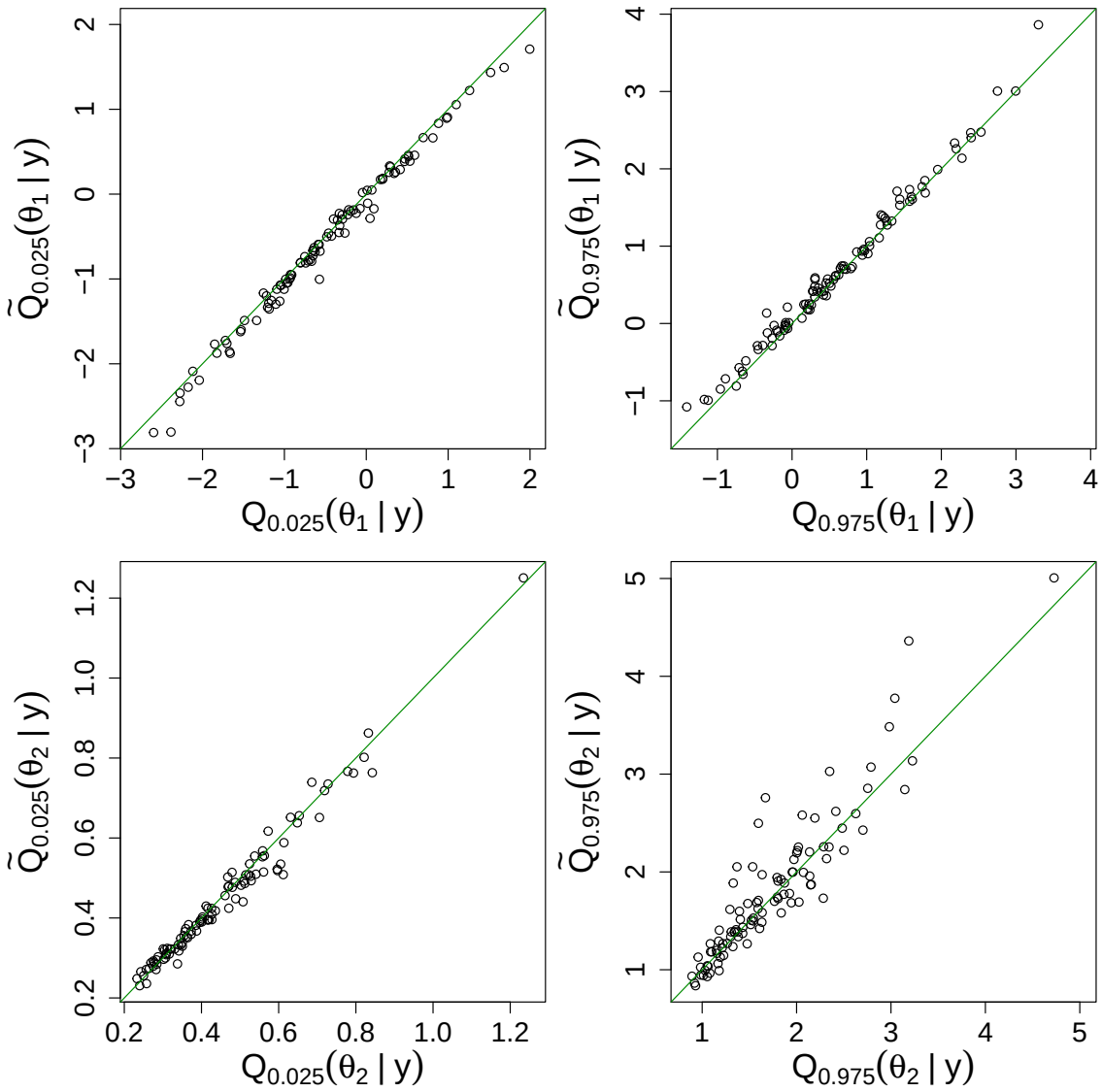


Figure 4.2 – Scatterplot of the theoretical values of 2.5% and 97.5% posterior quantiles for θ_1 and θ_2, for the normal model with their corresponding estimates obtained using ABC-RF.

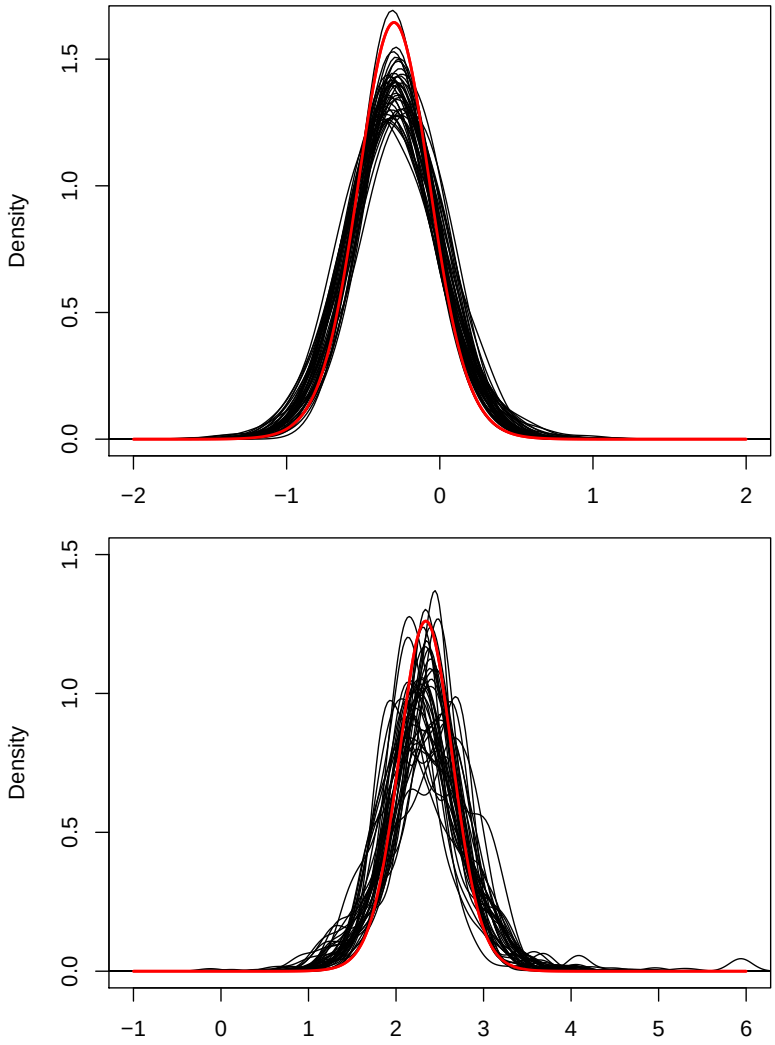


Figure 4.3 – Comparisons of the true posterior density distribution function of θ_1 in the normal model with a sample of 40 ABC-RF approximations of the posterior density (using RF weights), based on 40 independent reference tables and for two different test data sets (upper and lower panels). True posterior densities are represented by red lines.

	RF	Reject	ALL	ARR	ANN	py-SMC-1
$\psi_1(\mathbf{y}) = \mathbb{E}(\theta_1 \mid \mathbf{y})$	**0.18**	0.32	0.34	0.31	0.42	0.24
$\psi_2(\mathbf{y}) = \mathbb{E}(\theta_2 \mid \mathbf{y})$	**0.05**	0.10	0.14	0.17	0.17	0.09
$\psi_3(\mathbf{y}) = \mathbb{V}(\theta_1 \mid \mathbf{y})$	**0.25**	2.21	0.70	0.69	0.48	0.79
$\psi_4(\mathbf{y}) = \mathbb{V}(\theta_2 \mid \mathbf{y})$	**0.25**	0.43	0.66	0.70	0.97	0.55
$\mathbb{Q}_{0.025}(\theta_1 \mid \mathbf{y})$	**0.34**	1.61	0.69	0.84	0.50	0.36
$\mathbb{Q}_{0.025}(\theta_2 \mid \mathbf{y})$	**0.04**	0.13	0.34	0.55	0.80	0.17
$\mathbb{Q}_{0.975}(\theta_1 \mid \mathbf{y})$	**0.25**	1.35	0.53	0.70	0.60	0.45
$\mathbb{Q}_{0.975}(\theta_2 \mid \mathbf{y})$	**0.10**	0.14	0.20	0.20	0.42	0.21

Table 4.1 – Comparison of normalised mean absolute errors (NMAE) of estimated quantities of interest obtained with ABC-RF and other ABC methodologies. RF, Reject, ALL, ARR, ANN and py-SMC-1 stand for random forest (ABC-RF), rejection, adjusted local linear, adjusted ridge regression, adjusted neural network methods and ABC-SMC with adaptive population size from Klinger and Hasenauer (2017), respectively. The smallest NMAE values are in bold characters.

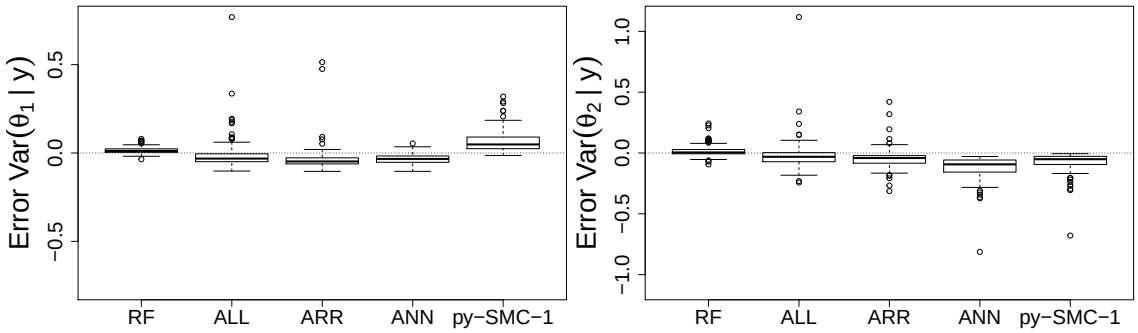


Figure 4.4 – Boxplot comparison of the differences between our predictions for $\mathbb{V}(\theta_1 \mid \mathbf{y})$ and $\mathbb{V}(\theta_2 \mid \mathbf{y})$ and the corresponding true values, using ABC-RF and other ABC methods. (RF, ALL, ARR and ANN notations as in the legend of Table 4.1). py-SMC-1 refers to the adaptive ABC-SMC algorithm of Klinger and Hasenauer (2017) with the same tuning parameters as in Table 4.1. The closer to the $y = 0$ axis, the better the predictions. Boxplots above this axis imply overestimation of the predictions and below underestimation.

	RF	a-PMC-1	a-PMC-2	PMC-1	PMC-2	SMC-1	SMC-2	py-PMC-1	py-PMC-2	py-SMC-1	py-SMC-2
$\mathbb{E}(\theta_1 \mid \mathbf{y})$	**0.18**	0.53	0.34	0.50	0.96	0.63	0.47	0.46	0.32	0.24	0.24
$\mathbb{E}(\theta_2 \mid \mathbf{y})$	**0.05**	0.14	0.10	0.13	0.14	0.13	0.11	0.11	0.09	0.09	0.08
$\mathbb{V}(\theta_1 \mid \mathbf{y})$	**0.25**	3.50	2.11	2.92	2.43	3.00	1.55	2.04	2.11	0.79	1.23
$\mathbb{V}(\theta_2 \mid \mathbf{y})$	**0.25**	0.36	0.28	0.39	0.48	0.33	0.43	0.52	0.44	0.55	0.50
$\mathbb{Q}_{0.025}(\theta_1 \mid \mathbf{y})$	**0.34**	2.82	1.77	2.72	1.30	2.65	1.58	1.71	1.62	0.36	1.03
$\mathbb{Q}_{0.025}(\theta_2 \mid \mathbf{y})$	**0.04**	0.21	0.16	0.14	0.25	0.18	0.13	0.16	0.13	0.17	0.13
$\mathbb{Q}_{0.975}(\theta_1 \mid \mathbf{y})$	**0.25**	2.19	1.28	1.54	1.04	1.75	1.03	1.32	1.41	0.45	0.75
$\mathbb{Q}_{0.975}(\theta_2 \mid \mathbf{y})$	**0.10**	0.14	0.11	0.14	0.26	0.15	0.17	0.21	0.18	0.21	0.18

Table 4.2 – Comparison of Normalised Mean Absolute Errors (NMAE) obtained with ABC-RF, adaptive ABC-PMC and sequential ABC methods for various tuning parameters based on 100 test data sets. RF stands for the ABC-RF scheme on a reference table of size 10^4. a-PMC-1 and a-PMC-2 stand for two designs of the adaptive ABC-PMC method of Prangle (2017) (Algorithm 5) with $2,000$ simulated particles per iteration, $1,000$ accepted particles, schemes iterate until we get approximately 10^4 (a-PMC-1) and 10^5 (a-PMC-2) simulated particles, respectively. PMC-1 and PMC-2 stand for two designs of the sequential ABC-PMC algorithm of Beaumont, Cornuet et al. (2009): PMC-1 and PMC-2 include $1,000$ and 100 simulated particles per iteration, 100 and 10 accepted particles during 10 and 100 iterations, respectively. For PMC-1 and PMC-2, we simulate 10^4 particles. SMC-1 and SMC-2 stand for two designs of the sequential ABC-SMC algorithm of Del Moral et al. (2012). Both SMC-1 and SMC-2 include $1,000$ simulated particles per iteration and a stopping rule based on two pre-computed quantiles of the distances between the observed summary statistics and simulated ones. For SMC-1, we use a quantile of 10% and for SMC-2 a quantile of 1%. We also use the `pyABC` Python module (Klinger, Rickert et al., 2018) to mimic the PMC algorithm as well as a full adaptive SMC method described in Klinger and Hasenauer (2017). The main drawback of this module is the absence of a stopping rule concerning the maximal number of simulations, thus the total number of simulations is higher than $10,000$ (number of simulations used for ABC-RF). py-PMC-1 and py-PMC-2 is a PMC adaptation obtained thanks to this module, using respectively 100 and $1,000$ accepted particles per iteration and a stopping rule based on a 1% pre-computed quantile of the distances between observed and simulated data, ($13,000$ and $132,000$ simulations are respectively needed in average). py-SMC-1 and py-SMC-2 designate the `pyABC` all adaptive version of ABC-SMC, including an adaptive population size, with a target density variation parameter respectively equal to 0.15 and 0.1 (see Klinger and Hasenauer (2017) for more details), and with 7 and 5 maximal iterations as stopping rule.

	RF	a-PMC-1	a-PMC-2	PMC-1	PMC-2	SMC-1	SMC-2	py-PMC-1	py-PMC-2	py-SMC-1	py-SMC-2
$\mathbb{E}(\theta_1 \mid \mathbf{y})$	**0.14**	0.78	0.68	0.64	1.28	0.81	0.81	0.72	0.57	0.45	0.45
$\mathbb{E}(\theta_2 \mid \mathbf{y})$	**0.05**	0.21	0.18	0.18	0.19	0.18	0.16	0.15	0.13	0.12	0.13
$\mathbb{V}(\theta_1 \mid \mathbf{y})$	**0.22**	6.42	5.31	5.06	5.34	4.76	3.42	4.17	4.37	2.26	2.82
$\mathbb{V}(\theta_2 \mid \mathbf{y})$	**0.28**	0.71	0.59	0.53	0.54	0.59	0.53	0.43	0.37	0.46	0.38
$\mathbb{Q}_{0.025}(\theta_1 \mid \mathbf{y})$	**0.30**	4.82	4.10	4.48	3.11	3.50	2.35	2.80	3.20	1.60	2.19
$\mathbb{Q}_{0.025}(\theta_2 \mid \mathbf{y})$	**0.05**	0.25	0.24	0.23	0.31	0.22	0.20	0.16	0.15	0.14	0.15
$\mathbb{Q}_{0.975}(\theta_1 \mid \mathbf{y})$	**0.23**	3.49	3.03	2.67	1.37	2.85	1.96	2.37	2.65	1.50	1.83
$\mathbb{Q}_{0.975}(\theta_2 \mid \mathbf{y})$	**0.10**	0.23	0.19	0.18	0.24	0.22	0.20	0.21	0.18	0.20	0.18

Table 4.3 – Same as Table 4.2, except that 500 independent noise summary statistics were added as explanatory variables instead of 50. Noise variables were simulated by randomly drawing into uniform distributions on $[0; 1]$. There are thus $11+500 = 511$ summary statistics in total (versus 61 summary statistics in Table 4.2).

We provide further comparisons obtained using ABC-RF, adaptive ABC-PMC and sequential ABC methods for various tuning parameters (Tables 4.2 and 4.3). We considered two different situations by including 50 (Table 4.2) or 500 (Table 4.3) noise variables (drawn into uniform distributions on $[0; 1]$) as explanatory variables. We found that the ABC-RF algorithm outperforms all adaptive and sequential methods (and designs) considered, and this adding or not a high amount of noise variables. Note that, in contrast to other methods, ABC-RF was only weakly affected by the presence of a large number of noise variables.

Details on tuning parameters used for ABC sequential methods:

- For all non-adaptive ABC methods, we use an Euclidean distance normalised by the MAD calculated on a precomputed reference table of size $10,000$. For adaptive methods, a similar distance is used where the standardisation is performed with iteratively updated MAD values.
- Concerning the transition kernel, a Gaussian distribution is considered. The variance-covariance matrix is taken as the weighted empirical one computed on the accepted parameters of the previous algorithm iteration, multiplied by 2 for the methods we implemented (a-PMC, PMC and SMC) and multiplied by a scaling factor involving the Silverman's rule of thumb (see Klinger and Hasenauer (2017) and Klinger, Rickert et al. (2018) for more details) for the remaining ones using `pyABC`.
- For the ABC-SMC algorithm of Del Moral et al. (2012), we do not change the tuning parameters described in the original paper (i.e. $\alpha = 0.90$, $N_T = N/2$ where N is the population size.)
- The adaptive ABC-SMC algorithm of Prangle (2017) requires a value (also denoted α) indicating the proportion of accepted simulations per iteration, here chosen at 0.5.
- To mimic the ABC-PMC strategy we use `pyABC` with an adaptive threshold equal to the median of the previous iteration distances, with a constant population size (100 or $1,000$) and with a minimum threshold value equal to the 1% quantile of a precomputed reference table of size $10,000$.

- Finally, for the full adaptive method of Klinger and Hasenauer (2017), we use an adaptive population size depending on a desired target density variation value (E_{cv}) equal to 0.15 or 0.1 in our experimentations and an initial population size equal to $1,000$. Note that the default value 0.05 induced a change in the population size from $1,000$ to $10,000$ in only one iteration, hence that is not relevant in our comparison with ABC-RF due to its high simulation cost. The threshold is taken as the median of the previous iteration distances (as in Klinger and Hasenauer, 2017). We do not use a minimum threshold value but a maximal number of iteration equal to 7 and 5 respectively. Note that this method requires about $24,000$ and $30,000$ simulations when 50 noise summary statistics are considered, and about $39,000$ and $41,000$ when 500 is added.

4.3.1.1 Comparing three methods of variance estimation of parameters

Finally, for this normal example, we here compare three methods to estimate posterior variance of a parameter transformation of interest τ using ABC-RF. Two of them have already been explained in Section 4.2 (i.e. methods 1 and 3 below).

- **Method 1**: One reuses the original random forest (RF) weights $w_i(\eta_{\mathbf{y}})$ to the out-of-bag square residuals $(\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})^2$, giving the variance estimator

$$\widetilde{\mathbb{V}}(\tau \mid \eta_{\mathbf{y}}) = \sum_{i=1}^{N} w_i(\eta_{\mathbf{y}})(\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})^2.$$

- **Method 2**: A similar estimator can be obtained by building a new RF thanks to the training sample $\left\{\left((\tau^{(i)} - \tau^{(i)}_{\text{oob}})^2, \eta_{\mathbf{x}^{(i)}}\right)\right\}_{i=1,\ldots,N}$, resulting in the estimator

$$\mathbb{V}^{\#}(\tau \mid \eta_{\mathbf{y}}) = \sum_{i=1}^{N} \widetilde{w}_i(\eta_{\mathbf{y}})(\tau^{(i)} - \hat{\tau}^{(i)}_{\text{oob}})^2,$$

  where $\widetilde{w}_i(\eta_{\mathbf{y}})$ is the computed weights of this newly trained RF. This estimator is based on the expression of the posterior variance as a conditional expectation:

$$\mathbb{V}(\tau \mid \eta_{\mathbf{y}}) = \mathbb{E}\left([\tau - \mathbb{E}(\tau \mid \eta_{\mathbf{y}})]^2 \mid \eta_{\mathbf{y}}\right)$$

  and the fact that such a RF is able to estimate this posterior expectation. This approach is more expensive due to the additional RF requirement.

- **Method 3**: The variance estimator is based on the cumulative distribution function (c.d.f.) approximation,

$$\widehat{\mathbb{V}}(\tau \mid \eta_{\mathbf{y}}) = \sum_{i=1}^{N} w_i(\eta_{\mathbf{y}}) \left(\tau^{(i)} - \sum_{u=1}^{N} w_u(\eta_{\mathbf{y}})\tau^{(u)}\right)^2.$$

We here compare these three estimators on the normal toy example detailed above with h the projection on both coordinates of the parameter vector θ. We

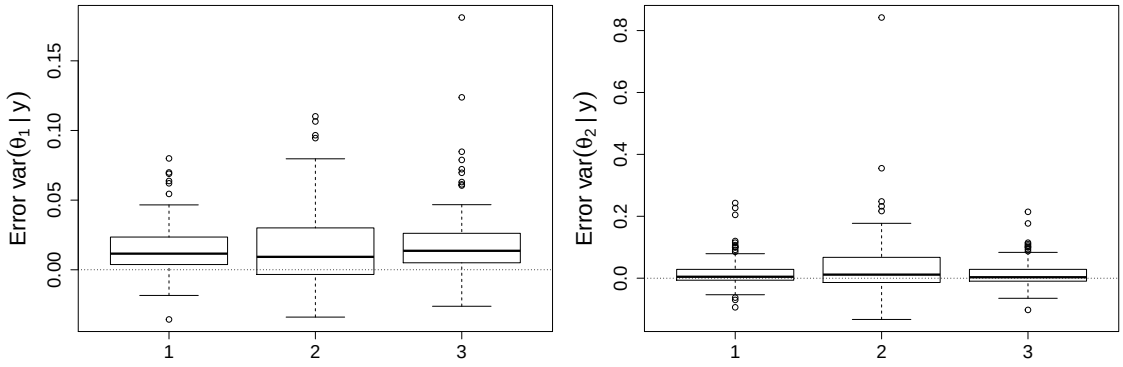


Figure 4.5 – Boxplot comparison of differences between our predictions for $\mathbb{V}(\theta_i \mid \mathbf{y})$ and the true values with the three methods of variance estimation: by reusing weights (method 1, boxplot 1), by building a new RF on square residuals (method 2, boxplot 2) and by using the estimation of the cumulative distribution function (method 3, boxplot 3).

Method	1	2	3
$\mathbb{V}(\theta_1 \mid \mathbf{y})$	**0.25**	0.28	0.30
$\mathbb{V}(\theta_2 \mid \mathbf{y})$	**0.25**	0.47	**0.25**

Table 4.4 – Comparison of normalised mean absolute errors (NMAE) of estimate variances when using three methods, (see legend of Figure 4.5). The smallest NMAE values are in bold characters.

find that the three estimators behave similarly: boxplots are alike and all tend to overestimate posterior variances (Figure 4.5). Results summarised in Table 4.4 also support this similarity in NMAE terms. Because the estimator 1 appears to show slightly lower errors for both parameter θ_1 and θ_2, we decided to use it in the normal and human population genetics examples.

4.3.2 Human population genetics example

We illustrate our methodological findings with the study of a population genetics data set including 50,000 single nucleotide polymorphic (SNP) markers genotyped in four human population samples (The 1000 genomes project consortium, 2012; see details in Pudlo et al., 2016). The four populations include Yoruba (Africa; YRI), Han (East Asia; CHB), British (Europe; GBR) and American individuals of African ancestry (North America; ASW). The considered evolutionary model is represented in Figure 4.6. It includes a single out-of-Africa event with a secondarily split into one European and one East Asian population lineage and a recent genetic admixture of Afro-Americans with their African ancestors and with Europeans. The model was robustly chosen as most appropriate among a set of eight evolutionary models, when compared using ABC-RF for model choice in Pudlo et al. (2016).

We here focused our investigations on two parameters of interest in this model: (i) the admixture rate ra (i.e. the proportion of genes with a non-African origin) that describes the genetic admixture between individual of British and African ancestry in Afro-Americans individuals; and (ii) the ratio N2/Na between the ancestral effective population size Na and African N2 (in number of diploid individuals), roughly describing the increase of African population size in the past. Considering ratios of effective population sizes allows preventing identifiability issues of the model.

We used the software DIYABC v.2.1.0 (Cornuet, Santos et al., 2008; Cornuet, Pudlo et al., 2014) to generate a reference table of size 200,000, with $N = 199,000$ data sets being used as training data set and $N_{\text{pred}} = 1,000$ remaining as test data sets. RFs are built in the same way as for our normal example and make use of the $d = 112$ summary statistics provided for SNP markers by DIYABC, (see Pudlo et al. (2016), and Chapter 3, Table 3.2).

Due to the complexity of this model, the exact calculation of any posterior quantity of interest is unfeasible. To bypass this difficulty we compute NMAE using simulated parameters from the test table, rather than targeted posterior expectations; in this case the normalisation is performed by dividing by simulated parameter values. Here, 95% credible intervals (CI) are deduced from posterior quantile estimate of order 2.5% and 97.5%. Performances are measured via mean range and coverage, with coverage corresponding to the percentage of rightly bounded parameters. For example a 95% CI should provide coverage equal to 95% of the test table.

Figure 4.7, Figure 4.8 and Table 4.5 illustrate the quality of the ABC-RF method when compared with ABC with either rejection, local linear, ridge or neural network adjustment (with logit transforms of the parameters for non rejection methods) using different tolerance levels (i.e., with tolerance proportion ranging from 0.005 to 1). We recall that considering the ABC rejection method with a tolerance equals to 1 is equivalent to work with the prior. Note that, due to memory allocation issues when using ABC method with adjusted ridge regression and a tolerance level of 1 on large reference table, we did not manage to recover results in this specific case.

Interesting methodological features can be observed in association with this example. ABC with rejection performs poorly in terms of NMAE and provides conservative and hence wide CIs (i.e., with coverage higher than the formal level). For ABC with adjustment, the lower the tolerance the lower the error (Table 4.5). The

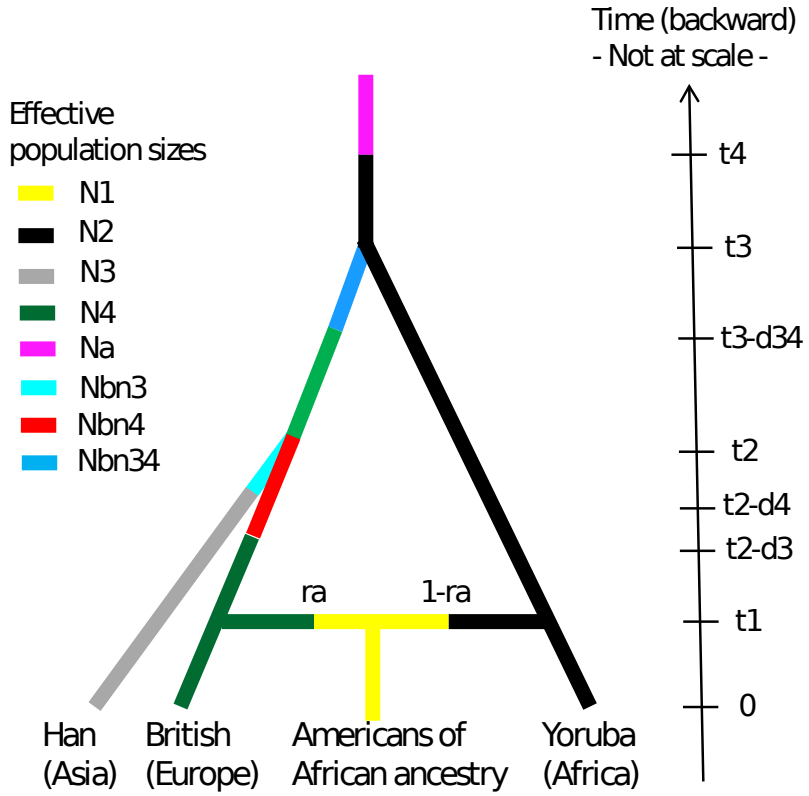


Figure 4.6 – Evolutionary model of four human populations considered for Bayesian parameter inference using ABC-RF. The prior distributions of the demographic and historical parameters used to simulate SNP data sets are as followed: Uniform $\mathcal{U}_{[100;10,000]}$ for the split times t2 and t3 (in number of generations), $\mathcal{U}_{[1;30]}$ for the admixture time t1, $\mathcal{U}_{[0.05;0.95]}$ for the admixture rate ra (proportion of genes with a non-African origin), $\mathcal{U}_{[1,000;100,000]}$ for the stable effective population sizes N1, N2, N3, N4 and N34 (in number of diploid individuals), $\mathcal{U}_{[5;500]}$ for the bottleneck effective population sizes Nbn3, Nbn4, and Nbn34, $\mathcal{U}_{[5;500]}$ for the bottleneck durations d3, d4, and d34, $\mathcal{U}_{[100;10,000]}$ for both the ancestral effective population size Na and t4 the time of change to Na. Conditions on time events were t4>t3>t2. See Pudlo et al. (2016) for details. Regarding the genetic model, we simulated biallelic polymorphic SNP data sets using the algorithm proposed by Hudson (2002) (cf "-s 1" option in the program `ms` associated to Hudson (2002)). This coalescent-based algorithm provides the simulation efficiency and speed necessary in the context of ABC, where large numbers of simulated data sets including numerous (statistically independent) SNP loci have to be generated (see Supplementary Appendix S1 of Cornuet, Pudlo et al. (2014) for additional comments on Hudson's algorithm).

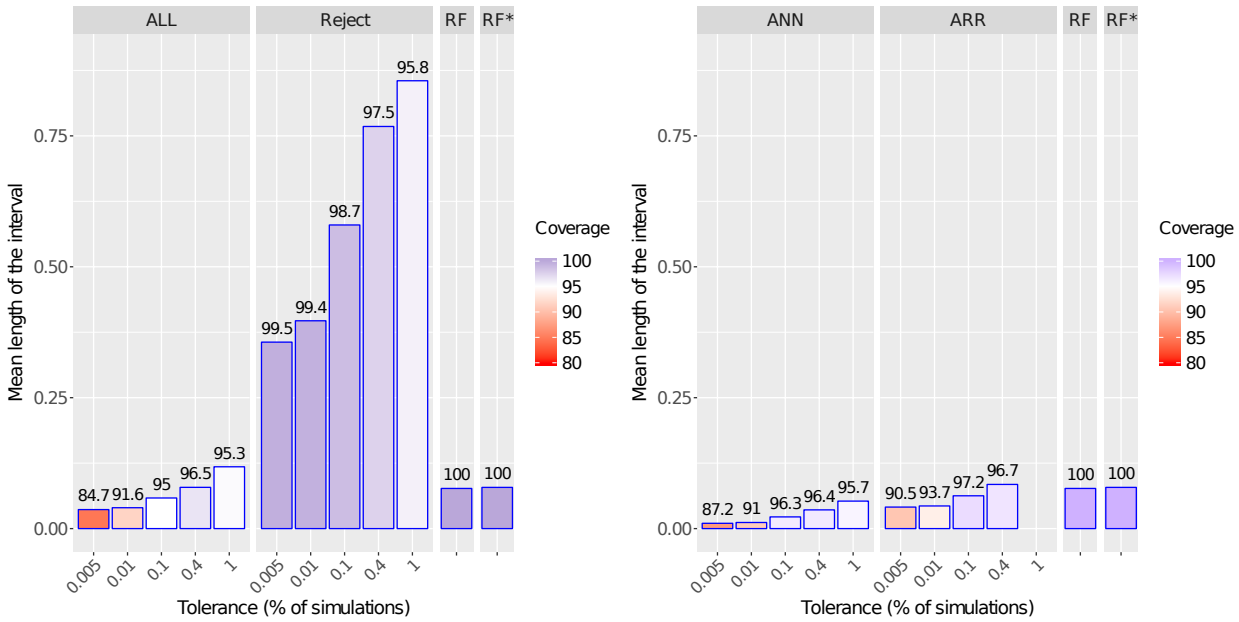


Figure 4.7 – Range and coverage comparison of approximate 95% credible intervals on the admixture parameter ra (Figure 4.6) obtained with ABC-RF (RF) and with earlier ABC methods: rejection (Reject), adjusted local linear (ALL) or ridge regression (ARR) or neural network (ANN) with various tolerance levels for Reject, ALL, ARR and ANN. Coverage values are specified by bar colours and superimposed values. Heights indicate CI mean lengths. Results for ALL, Reject and RF are presented in the left figure whereas those for ANN, ARR and RF are in the right figure. RF* refers to results obtained using ABC-RF when adding 20 additional independent noise variables generated from a uniform $\mathcal{U}_{[0;1]}$ distribution. RF refers to results without noise variables.

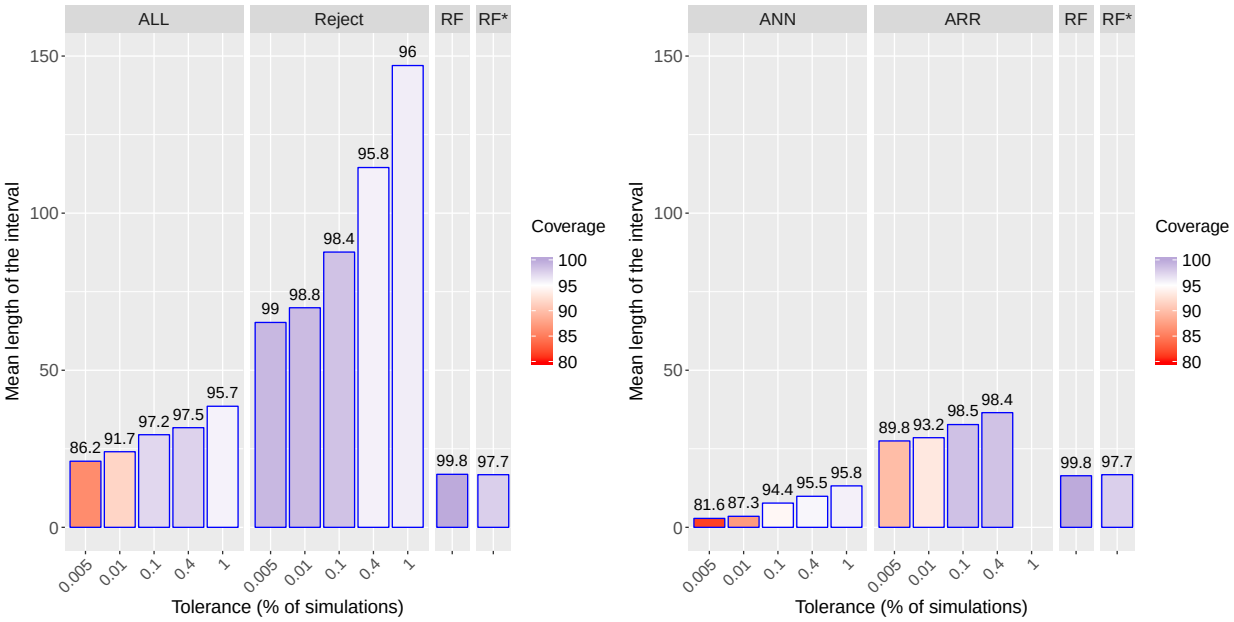


Figure 4.8 – Range and coverage comparison of approximate 95% credible intervals on the ratio N2/Na of the Human population genetics example, obtained with ABC-RF (RF) and with earlier ABC methods: rejection (Reject), adjusted local linear (ALL) or ridge regression (ARR) or neural network (ANN) with various tolerance levels for Reject, ALL, ARR and ANN. Coverage values are specified by bar colours and superimposed values. Heights indicate CI mean lengths. Results for ALL, Reject and RF are presented in the left figure whereas those for ANN, ARR and RF are in the right figure. Na is the ancestral African effective population size before the population size change event and N2 the African effective population size after the population size change event (going backward in time). RF* refers to results obtained using ABC-RF when adding 20 additional independent noise variables generated from a uniform $\mathcal{U}_{[0;1]}$ distribution. RF refers to results without noise variables.

Method	Tolerance level	ra NMAE	N2/Na NMAE
RF	NA	0.018	0.053
RF*	NA	0.019	0.053
Reject	0.005	0.151	0.355
Reject	0.01	0.178	0.454
Reject	0.1	0.322	1.223
Reject	0.4	0.574	2.025
Reject	1	0.856	4.108
ALL	0.005	0.028	0.166
ALL	0.01	0.028	0.249
ALL	0.1	0.035	0.139
ALL	0.4	0.044	0.170
ALL	1	0.062	0.209
ARR	0.005	0.027	0.220
ARR	0.01	0.027	0.317
ARR	0.1	0.035	0.140
ARR	0.4	0.044	0.163
ARR	1	–	–
ANN	0.005	**0.007**	**0.037**
ANN	0.01	0.007	0.038
ANN	0.1	0.013	0.064
ANN	0.4	0.016	0.123
ANN	1	0.025	0.095

Table 4.5 – Comparison of normalised mean absolute errors (NMAE) for the estimation of the parameters ra and N2/Na using ABC-RF (RF) and ABC with rejection (Reject), adjusted local linear (ALL) or ridge regression (ARR) or neural network (ANN) with various tolerance levels for Reject, ALL, ARR and ANN. NA stands for not appropriate. The smallest NMAE values are in bold characters. NA stands for not appropriate. RF* refers to results obtained using ABC-RF when adding 20 additional independent noise variables generated from a uniform $\mathcal{U}_{[0;1]}$ distribution. RF refers to results without noise variables.

CI quality however highly suffers from low tolerance, with underestimated coverage (Figures 4.7 and 4.8). The smaller the tolerance value, the narrower the CI. Results for the ABC method with adjusted ridge regression seems however to be unstable for the parameter N2/Na depending on the considered level of tolerance. The ABC method using neural network and a tolerance level of 0.005 provides the lowest NMAE for both parameters of interest. The corresponding coverages are however underestimated, equal to 87.2% for ra and 81.6% for N2/Na, when 95% is expected (second part of Figures 4.7 and 4.8). Note that results with this method can be very time consuming to obtain when the tolerance level and the number of repetitions are large. The ABC-RF method provides an appealing trade-off between parameter estimation quality (ABC-RF is the method with the second lowest NMAE values in Table 4.5) and slightly conservative CIs (Figures 4.7 and 4.8). Similar results and methodological features were observed when focusing on the 90% CI (results not shown). It is also worth stressing that not any calibration of any kind of a tolerance level parameter are needed with ABC-RF, which is an important plus for this method. On the opposite, earlier ABC methods require calibration to optimise their use, such calibration being time consuming when different levels of tolerance are used.

For the observed data set used in this study, posterior expectations and quantiles of the parameters of interest ra and N2/Na are reported in Tables 4.6 and 4.7. Expectation and CI values substantially vary for both parameters, depending on the method used. The impact of the tolerance levels is noteworthy for both the rejection and local linear adjustment ABC methods. The posterior expectation of ra obtained using ABC-RF was equal to 0.221 with a relatively narrow associated 95% CI of $[0.112; 0.287]$. The latter estimation lays well within previous estimates of the mean proportion of genes of European ancestry within African-American individuals, which typically ranged from 0.070 to 0.270 − with most estimates around 0.200 −, depending on individual exclusions, the population samples and sets of genetic markers considered, as well as the evolutionary models assumed and inferential methods used (reviewed in Bryc et al., 2015). Interestingly, a recent genomic analysis using a conditional random field parameterized by random forests trained on reference panels (Maples et al., 2013) and $500,000$ SNPs provided a similar expectation value of ra for the same African American population ASW (i.e. ra $= 0.213$), with a somewhat smaller 95% CI (i.e. $[0.195; 0.232]$), probably due to the ten times larger number of SNPs in their data set (Baharian et al., 2016).

The posterior expectation of N2/Na obtained using ABC-RF was equal to 4.508 with a narrow associated 95% CI of $[3.831; 5.424]$. Such values point to the occurrence of the substantial ancestral demographic and geographic expansion that is widely illustrated in previous Human population genetics studies, including African populations (e.g. Henn et al., 2012). Although our modelling setting assumes a naive abrupt change in effective population sizes in the ancestral African population, the equivalent of N2/Na values inferred from different methods and modelling settings fit rather well with our own posterior expectations and quantiles for this parameter (e.g. Schiffels and Durbin, 2014).

In contrast to earlier ABC methods, the RF approach is deemed to be mostly insensitive to the presence of covariates whose the distributions does not depend on the parameter values (i.e. ancillary covariates) (e.g. Breiman, 2001; Marin, Pudlo,

ra

Method	Tol. level	Expectation	$\mathbb{Q}_{0.025}$	$\mathbb{Q}_{0.05}$	$\mathbb{Q}_{0.95}$	$\mathbb{Q}_{0.975}$
RF	NA	0.221	0.112	0.134	0.279	0.287
RF*	NA	0.225	0.112	0.142	0.282	0.290
Reject	0.005	0.223	0.061	0.069	0.364	0.389
Reject	0.01	0.220	0.060	0.070	0.389	0.418
Reject	0.1	0.276	0.062	0.074	0.511	0.543
Reject	0.4	0.388	0.068	0.086	0.739	0.791
Reject	1	0.502	0.073	0.095	0.906	0.928
ALL	0.005	0.278	0.219	0.229	0.322	0.337
ALL	0.01	0.257	0.232	0.238	0.274	0.278
ALL	0.1	0.207	0.170	0.171	0.233	0.237
ALL	0.4	0.194	0.144	0.152	0.233	0.241
ALL	1	0.196	0.115	0.126	0.278	0.299
ARR	0.005	0.260	0.252	0.254	0.265	0.266
ARR	0.01	0.252	0.239	0.242	0.260	0.262
ARR	0.1	0.211	0.171	0.178	0.239	0.244
ARR	0.4	0.196	0.140	0.149	0.241	0.251
ARR	1	–	–	–	–	–
ANN	0.005	0.227	0.221	0.223	0.232	0.234
ANN	0.01	0.226	0.219	0.221	0.231	0.233
ANN	0.1	0.228	0.217	0.220	0.236	0.239
ANN	0.4	0.232	0.216	0.221	0.242	0.248
ANN	1	0.206	0.183	0.187	0.227	0.233

Table 4.6 – Estimation of the parameter ra and N2/Na for the observed human population genetics data set using ABC-RF (RF), and ABC with rejection (Reject), adjusted local linear (ALL) or ridge regression (ARR) or neural network (ANN) with various tolerance levels (Tol. level) for Reject, ALL, ARR and ANN. NA stands for not appropriate. RF* refers to results obtained using ABC-RF when adding 20 additional independent noise variables generated from a uniform $\mathcal{U}_{[0;1]}$ distribution. RF refers to results without noise variables.

Robert et al., 2012). To illustrate this feature, we have added 20 additional independent noise variables generated from a uniform $\mathcal{U}_{[0;1]}$ distribution (results designated by RF*) in the reference table generated for the present Human population genetics example. We found that the presence of such noise covariates do not impact the results in terms of NMAE, coverage and only slightly on parameter estimation for the observed data set (Tables 4.5, 4.6 and 4.7, and Figures 4.7 and 4.8). For the rest of this chapter, no noise variables were used.

N2/Na

Method	Tol. level	Expectation	$\mathbb{Q}_{0.025}$	$\mathbb{Q}_{0.05}$	$\mathbb{Q}_{0.95}$	$\mathbb{Q}_{0.975}$
RF	NA	4.508	3.831	3.959	5.153	5.424
RF*	NA	4.594	3.821	3.910	5.241	6.552
Reject	0.005	6.282	2.937	3.223	10.086	11.337
Reject	0.01	6.542	2.746	3.116	10.837	11.852
Reject	0.1	8.001	2.131	2.574	15.690	18.531
Reject	0.4	11.605	1.795	2.331	28.011	38.532
Reject	1	23.483	0.672	1.185	84.649	147.657
ALL	0.005	30.041	1.256	1.879	83.369	174.340
ALL	0.01	9.289	3.946	4.586	16.686	20.361
ALL	0.1	8.235	5.736	5.995	11.573	12.719
ALL	0.4	10.752	4.588	4.996	21.656	27.300
ALL	1	7.222	5.684	5.829	9.631	10.475
ARR	0.005	10.528	4.395	5.677	19.224	22.722
ARR	0.01	8.264	5.020	5.485	12.544	13.313
ARR	0.1	8.394	5.643	5.948	12.075	13.313
ARR	0.4	10.802	6.113	6.505	17.487	20.511
ARR	1	–	–	–	–	–
ANN	0.005	5.746	5.512	5.563	5.937	5.982
ANN	0.01	6.148	5.883	5.934	6.353	6.420
ANN	0.1	25.921	23.857	24.250	27.672	28.133
ANN	0.4	8.515	7.652	7.810	9.147	9.436
ANN	1	7.021	5.692	5.856	8.677	9.370

Table 4.7 – Same as Table 4.6 for the parameter N2/Na.

4.3.2.1 Contribution of summary statistics in ABC-RF estimation of the parameters ra and N2/Na of the Human population genetics example

In the same spirit than in Pudlo et al. (2016), a by-product of our ABC-RF-based approach is to automatically determine the (most) relevant statistics for the estimation of each parameter by computing a criterion of variable importance (here a variable is a summary statistic). We consider here the mean decrease of impurity (Chapter 2).

Figure 4.9 shows the contributions of the 30 most important summary statistics (among the 112 statistics proposed by DIYABC) for the ABC-RF estimation of the parameters ra and N2/Na of the Human population genetics example. The most informative summary statistics are clearly different depending on the parameter of interest. For the admixture rate between two sources populations (ra), all ten most informative statistics correspond to statistics characterising a pair or a trio of populations (e.g. AV1 or FMO statistics; see Chapter 3, Table 3.2). Moreover, all those "best" statistics include the populations ASW, GBP and YRI which correspond to the target and the two source populations respectively. On the contrary, for the effective population size ratio N2/Na, seven of the ten most most informative statistics correspond to statistics characterising within population genetic variation (e.g. HV1 or HMO; see Chapter 3, Table 3.2). In this case, all those "best" statistics include the African population, which makes sense since N2 is the effective population size in the studied African population and Na in the population ancestral to all studied populations. It is worth stressing that, although the most informative summary statistics make sense in relation to the studied parameters it was difficult if not impossible to a priori and objectively select those statistics. This is not an issue when using the ABC-RF approach as the method automatically extracts the maximum of information from the entire set of proposed statistics.

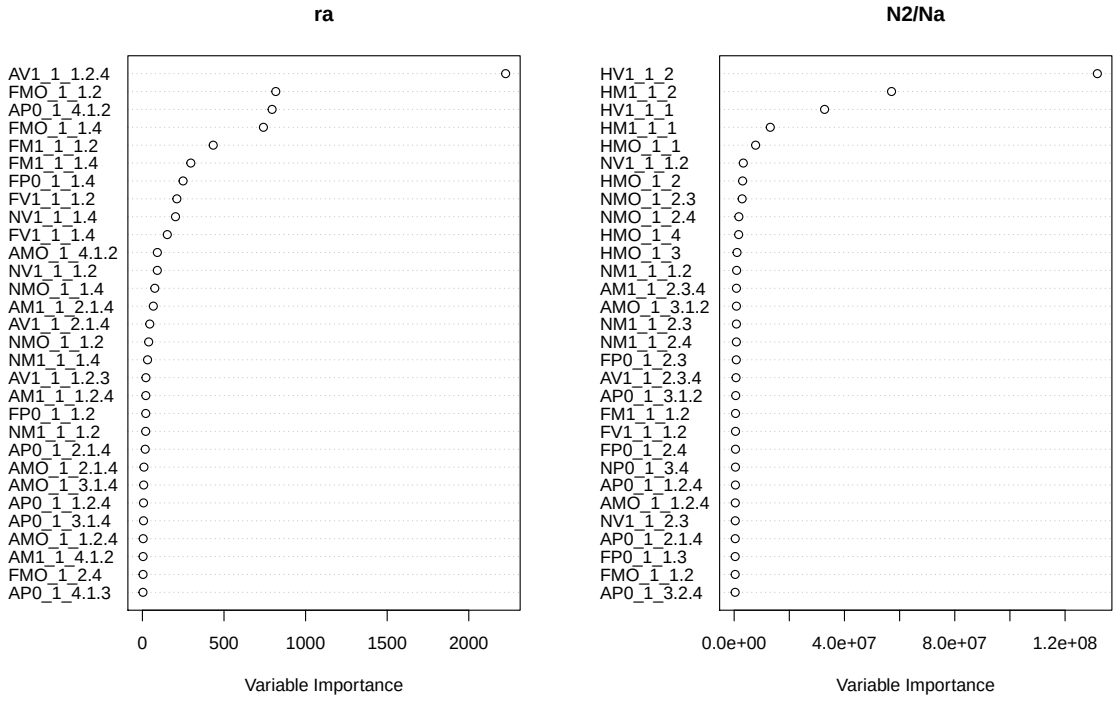


Figure 4.9 – Contributions of the 30 most important summary statistics for the ABC-RF estimation of the parameters ra and N2/Na of the Human population genetics example. The contribution of each statistic is evaluated the mean decrease of impurity, for each of the 112 used summary statistics provided for SNP markers by DIYABC. The higher the variable importance the more informative the statistics. The population index(s) is indicated at the end of each statistic. 1 = pop ASW (Americans of African ancestry), 2 = pop YRI (Yoruba, Africa), 3 = pop CHB (Han, Asia) and 4 = pop GBP (British, Europe). For instance FMO_1_2.4 = mean across loci of F_{ST} distance between the populations 2 and 4, the 1 corresponds to the first loci group (in our study there is only one group). See also Chapter 3, Table 3.2. Note the difference of scale for the importance criterion for parameters ra and N2/Na. This difference can be explained by the difference of scale in parameter values. Indeed, it directly influences the residual sum of squares. The parameter ra being bounded in $[0; 1]$ contrary to N2/Na, a higher decrease can be expected for the ratio N2/Na than ra.

Method	Tol. level	CPU time (in minutes)
RF	NA	16.64
Reject	0.005	7.54
Reject	0.01	7.54
Reject	0.1	7.78
Reject	0.4	7.98
Reject	1	9.14
ALL	0.005	7.66
ALL	0.01	7.70
ALL	0.1	8.81
ALL	0.4	9.21
ALL	1	11.32
ARR	0.005	6.71
ARR	0.01	6.97
ARR	0.1	40.57
ARR	0.4	560.39
ARR	1	–
ANN	0.005	22.31
ANN	0.01	33.60
ANN	0.1	216.61
ANN	0.4	1160.67
ANN	1	4028.63

Table 4.8 – Comparison of the computation time (in minutes) required – after the generation of the reference table – for the estimation of the parameter of interest ra on a data set test table, using ABC-RF (RF), ABC with rejection (Reject), adjusted local linear (ALL), ridge regression (ARR) and neural network (ANN), with various tolerance levels for Reject, ALL, ARR and ANN. The test table included $1,000$ pseudo-observed data sets and the reference table included $199,000$ simulated data sets summarised with 112 statistics. Results were computed on a cluster with 28 CPU cores of 2.4 GHz. NA stands for not appropriate.

4.3.2.2 Computation times required by the statistical treatments of the studied methods processed following the generation of the reference table

We here present a comparison of the computation time requirement for the different methods studied in this chapter, when predicting estimations of the admixture rate ra in the human population genetics example. ABC methods with rejection or adjusted with local linear regression provide the best results in terms of CPU time even when the tolerance level is equal to 1. The ABC-RF strategy requires moderately higher computing time. The calculation of the RF weights is the most expensive computation part (i.e. 3/4 of computation time). ABC methods using ridge regression or neural network correction become very time consuming when the tolerance level is high.

4.3.3 Practical recommendations regarding the implementation of the ABC-RF algorithm

We mainly consider in this section two important practical issues, namely the choice of the number of simulations (N) in the reference table and of the number of trees (B) in the random forest. For sake of simplicity and concision, we focus our recommendations on the above human population genetics example (Section 4.3.2). We stress here that, although not generic, our recommendations fit well with other examples of complex model settings that we have analysed so far (results not shown). We also stress that for simpler model settings substantially smaller N and B values were sufficient to obtain good results. Finally, we provide practical comments about the main sources of variabilities in inferences typical of the ABC-RF methodology.

Reference table size – We consider a reference table made of $N = 199,000$ simulated data sets. However, Table 4.9 shows a negligible decrease of NMAE when using $N = 100,000$ to $N = 199,000$ data sets. Table 4.10 also exhibits small variations between predictions on the observed data set, especially for $N \geq 75,000$. The level of variation thus seems to be compatible with the random variability of the RF themselves. Altogether, using a reference table including $100,000$ data sets seems to be a reasonable default choice. It is worth stressing that the out-of-bag mean squared error can be easily retrieved without requiring the simulation of a (small size) secondary test table. It provides a good indicator of the quality of the RF at a low computational cost (Tables 4.9 and 4.11).

Number of trees – A forest including 500 trees is a default choice when building RFs, as this provides a good trade-off between computation efficiency and statistical precision (Breiman, 2001; Pudlo et al., 2016). To evaluate whether or not this number is sufficient, we recommend to compute the out-of-bag mean squared error (or another type of error) depending on the number of trees in the forest for a given reference table. If 500 trees is a satisfactory calibration, one should observe a stabilisation of the error around this value. Figure 4.10 illustrates this representation on the human population genetics example and points to a negligible decrease of the error after 500 trees. This graphical representation is produced via our R package `abcrf`.

Minimum node size (maximum leaf size) – We recall that splitting events during a tree construction stop when a node has less than $N_{\min}$ observations, in that case, the node becomes a leaf. Note that the higher $N_{\min}$ the quicker RF treatments. In all RF treatments presented here, we used the default size $N_{\min} = 5$. Table 4.11 illustrates the influence of $N_{\min}$ on the human population genetics example and highlights a negligible decrease of the error for $N_{\min}$ lower than 5.

Finally, we see no reason to change the number of summary statistics sampled at each split m_{try} within a tree, which is traditionally chosen as $d/3$ for regression when d is the total number of predictors (Breiman, 2001).

NMAE

N ($\times 10^3$)	10	25	50	75	100	150	199
ra	0.028	0.023	0.021	0.020	0.019	0.018	0.018
N2/Na	0.080	0.067	0.059	0.057	0.055	0.053	0.053

OOB MSE

N ($\times 10^3$)	10	25	50	75	100	150	199
ra ($\times 10^{-4}$)	1.670	1.176	0.914	0.823	0.745	0.695	0.664
N2/Na ($\times 10^3$)	0.194	0.179	0.143	0.125	0.115	0.111	0.110

Table 4.9 – Comparison of normalised mean absolute errors (NMAE) and out-of-bag mean squared errors (OOB MSE) for the estimation of the parameters ra and N2/Na obtained with ABC-RF, using different reference table sizes (N). We use the test table mentioned in Section 4.3.3. The number of trees in the RF is 500.

N ($\times 10^3$)	10	25	50	75	100	150	199
ra expectation	0.231	0.222	0.224	0.223	0.222	0.223	0.221
ra $\mathbb{Q}_{0.025}$	0.097	0.095	0.102	0.104	0.106	0.109	0.112
ra $\mathbb{Q}_{0.975}$	0.317	0.309	0.305	0.305	0.289	0.292	0.287
N2/Na expectation	4.538	4.588	4.652	4.530	4.475	4.483	4.508
N2/Na $\mathbb{Q}_{0.025}$	3.651	3.679	3.782	3.802	3.751	3.840	3.831
N2/Na $\mathbb{Q}_{0.975}$	6.621	6.221	6.621	5.611	5.555	5.315	5.424

Table 4.10 – Estimation of the parameters ra and N2/Na for the observed population genetics data set with ABC-RF, using different reference table sizes (N). The number of trees in the RF is 500.

Variability in the ABC-RF methodology – The ABC-RF methodology is associated with different sources of variabilities the user should be aware of. Using a simulated reference table is the main source, RF being the second. Indeed, predicting quantities of interest for the same test data set with two different reference tables of equal size N will result in slightly different estimates. This variation has been previously highlighted in Figure 4.3 dealing with the analysis of the normal toy example. We recall that RFs are composed of trees trained on bootstrap samples, each one considering m_{try} covariates randomly selected amongst the d available at each split. This random aspects of RF results in variability. In practice, a good user habit should be to run ABC-RF more than once on different training data sets to ensure that the previously mentioned variabilities are negligible. If this variability is significant, we recommend considering a reference table of higher size.

NMAE

N_{min}	1	2	3	4	5	10	20	50	100	200	500
ra	0.019	0.019	0.019	0.019	0.019	0.019	0.020	0.021	0.023	0.027	0.033
N2/Na	0.054	0.055	0.054	0.055	0.055	0.055	0.055	0.058	0.062	0.068	0.082

OOB MSE

N_{min}	1	2	3	4	5	10	20	50	100	200	500
ra ($\times 10^{-4}$)	0.745	0.739	0.744	0.739	0.745	0.760	0.783	0.925	1.129	1.480	2.280
N2/Na ($\times 10^{3}$)	0.114	0.116	0.115	0.115	0.115	0.116	0.119	0.131	0.153	0.183	0.252

Table 4.11 – Comparison of normalised mean absolute errors (NMAE) and out-of-bag mean squared errors (OOB MSE) for the estimation of the parameters ra and N2/Na obtained with ABC-RF, using different minimum node sizes (N_{min}). We use the reference table of size $N = 100,000$ and the test table mentioned in Section 4.3.3. The number of trees in the RF is 500.

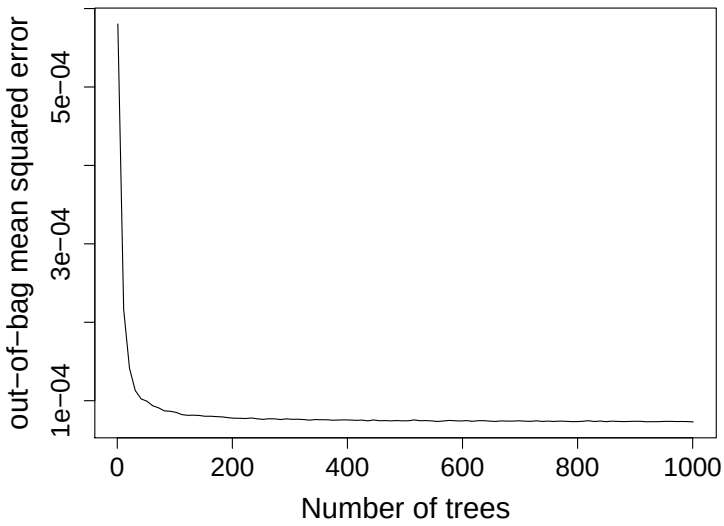


Figure 4.10 – Relations between the number of trees in the forest and the ABC-RF out-of-bag mean squared errors, for a reference table of size $N = 100,000$ in the human population genetics example, for the parameter ra.

4.3.4 Study of covariances of parameters using random forests on a regression toy example

Finally, we now apply our RF methodology to a toy regression example for which its non-zero covariance between parameters is the main quantity of interest, hence we consider the case where g and h (see Section 4.2.5) are the projections on a given coordinate of the parameter vector θ. For a simulated $n \times 2$ design matrix $X = [x_1; x_2]$, we consider the Zellner's hierarchical model (Marin and Robert, 2014, chapter 3):

$$\begin{aligned}(y_1, \ldots, y_n) \mid \beta_1, \beta_2, \sigma^2 &\sim \mathcal{N}_n(X\beta, \sigma^2 Id),\\ \beta_1, \beta_2 \mid \sigma^2 &\sim \mathcal{N}_2(0, n\sigma^2 (X^\top X)^{-1}),\\ \sigma^2 &\sim I\mathcal{G}(4, 3),\end{aligned}$$

where $\mathcal{N}_k(\mu, \Sigma)$ denotes the multivariate normal distribution of dimension k with mean vector μ and covariance matrix Σ, and $I\mathcal{G}(\kappa, \lambda)$ an inverse Gamma distribution with shape parameter κ and scale parameter λ. Provided $X^\top X$ is invertible, this conjugate model leads to closed-form marginal posteriors (Marin and Robert, 2014)

$$\begin{aligned}\beta_1, \beta_2 \mid \mathbf{y} \sim \mathcal{T}_2\Big(&\frac{n}{n+1}(X^\top X)^{-1} X^\top \mathbf{y},\\ &\frac{3 + \mathbf{y}^\top (Id - X(X^\top X)^{-1} X^\top)\mathbf{y}/2}{4 + n/2} \frac{n}{n+1} (X^\top X)^{-1}, 8 + n\Big),\\ \sigma^2 \mid \mathbf{y} \sim I\mathcal{G}&\left(4 + \frac{n}{2}, 3 + \frac{1}{2}\mathbf{y}^\top (Id - X(X^\top X)^{-1} X^\top)\mathbf{y}\right),\end{aligned}$$

where $\mathcal{T}_k(\mu, \Sigma, \nu)$ is the multivariate Student distribution of dimension k, with location parameter μ, scale matrix Σ and degree of freedom ν.

In our simulation experiment, we concentrate on the non zero covariance of the posterior distribution namely $\mathrm{Cov}(\beta_1, \beta_2 \mid \mathbf{y})$. A reference table of $N = 10,000$ replicates of a n-sample with $n = 100$ is generated. We then create $d = 60$ summary statistics: the maximum likelihood estimates of β_1, β_2, the residual sum of squares, the empirical covariance and correlation between $\mathbf{y}$ and x_1, covariance and correlation between $\mathbf{y}$ and x_2, the sample mean, the sample variance, the sample median, and 50 independent noise variables simulated from a uniform distribution $\mathcal{U}_{[0;1]}$. These noise variables were introduced to be in a sparse context.

Similarly to the normal example, we assess the performance of our approach using an independent (Monte Carlo) test data set of size $N_{\text{pred}} = 100$ and compare estimation accuracy with the ABC-RF approach from the ones with adjusted ridge regression and neural network ABC methodologies. RF are once again built with $B = 500$ trees, $m_{\text{try}} = d/3$ and minimum node size equals to 5 and ABC methods rely on the R package `abc` with a tolerance parameter equals to 0.1 for ABC methods with adjustment. ABC with neural network adjustment again makes use of 10 independent runs of the neural network. For local linear or ridge regression the corrections are univariate. That is not the case for neural networks which, by construction, perform multivariate correction.

Covariance estimation is a novel feature in this example, Table 4.12 shows that the ABC-RF approach does better in NMAE terms. As exhibited in Figure 4.11,

	RF	ARR	ANN
$\mathrm{Cov}(\beta_1, \beta_2 \mid \mathbf{y})$	**0.26**	0.85	0.64

Table 4.12 – Comparison of normalised mean absolute errors (NMAE) of estimate posterior covariances between β_1 and β_2 using random forest (RF), adjusted ridge regression (ARR) and adjusted neural network (ANN) ABC methods. The smallest NMAE value is in bold characters.

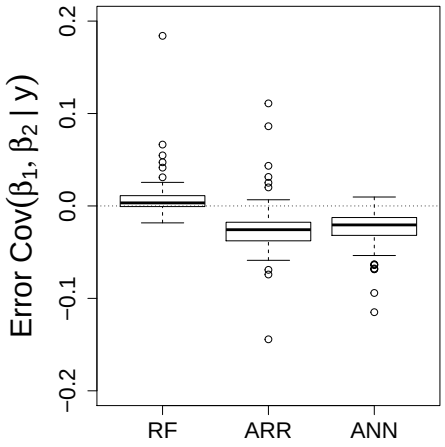


Figure 4.11 – Boxplot comparison of differences between prediction and true values for $\mathrm{Cov}(\beta_1, \beta_2 \mid \mathbf{y})$ using random forest (RF), adjusted ridge regression (ARR) and adjusted neural network (ANN) ABC methods.

ABC-RF overestimates covariances when earlier ABC methods underestimate it. Results are quite encouraging even though we believe the method might still be improved.

4.4 Conclusion

This chapter introduces a novel approach to parameter estimation in likelihood-free problems, relying on the machine-learning tool of regression RF to automate the inclusion of summary statistics in ABC algorithms. Our simulation experiments demonstrate several advantages of our methodological proposal compared with earlier ABC methods.

When using the same reference table and test data set for all compared methods, our RF approach appears to be more accurate than previous ABC solutions. Approximations of expectations are quite accurate, while posterior variances are only slightly overestimated, which is an improvement compared with other approaches that typically underestimate these posterior variances. The performances for covariance approximation are quite encouraging as well, although the method is still incomplete and need further developments on this particular point. We found that quantile estimations depend on the corresponding probability and we believe this must be related to the approximation error of the posterior cumulative function $F(x \mid \eta_{\mathbf{y}})$. More specifically, we observed that upper quantiles may be overestimated, whereas lower quantiles may be underestimated (Figure 4.2), indicating fatter

tails in the approximation. Hence, credible intervals produced by the RF solution may be larger than the exact ones. However from a risk assessment point of view, this overestimation aspect clearly presents less drawbacks than underestimation of credible intervals. Altogether, owing to the various models and data sets we analysed, we argue that ABC-RF provides a good trade-off in terms of quality between parameter estimation of point estimators (e.g. expectation, median or variance) and credible interval coverage, and its computing time is also very decent.

Throughout our experiments, we found that, contrary to earlier ABC methods, the RF approach is mostly insensitive to the presence of covariates whose the distributions do not depend on the parameter values (ancillary covariates). Therefore, we argue that the RF method can deal with a very large number of summary statistics, bypassing any form of pre-selection of those summaries. Interestingly, the property of ABC-RF to extract and adaptively weight information carried by each of the numerous summary statistics proposed as explanatory variables can be represented by graphs, showing the relative contribution of summary statistics in ABC-RF estimation for each studied parameter.

In population genetics, which historically corresponds to the field of introduction of ABC methods, next generation sequencing technologies result in large genome-wide data sets that can be quite informative about the demographic history of the genotyped populations. Several recently developed inferential methods relying on the observed site frequency spectrum appear particularly well suited to accurately characterising the complex evolutionary history of invasive populations (Gutenkunst et al., 2009; Excoffier, Dupanloup et al., 2013). Because of the reduced computational resources demanded by ABC-RF and the above-mentioned properties of the method, we believe that ABC-RF can efficiently contribute to the analysis of massive SNP data sets, including both model choice (Pudlo et al., 2016) and Bayesian inference about parameters of interest. We present in Chapter 6 two applications of the ABC-RF methodologies on population genetics case studies, and we introduce at the same time some improvements to this ABC-RF approach for model choice and parameter inference. More generally, the method should appeal to all scientific fields in which large data sets and complex models are analysed using simulation-based methods such as ABC (e.g. Beaumont, 2010; Sisson, Fan and Beaumont, 2018).

Chapter 5

Local tree-based methods for classification

This chapter is based on a survey paper we wrote, dealing with local tree-based methods. We plan to submit it to the *Machine Learning* journal.

Contents

5.1 Introduction

The machine learning field of local/lazy/instance-based/case-specific learning (Aha et al., 1991) aims at taking into account a particular instance x^* to produce a prediction thanks to its similarity to the training data set. It is opposed to eager learning, where the prediction is divided in two parts: a training phase where a global model is fitted and then a prediction phase. The local approach, in contrast, fits a model taking into account the information provided by x^*.

Two closely related learning fields need to be mentioned: semi-supervised learning (Chapelle et al., 2010) and transductive learning (Gammerman et al., 1998). Semi-supervised learning introduces unlabelled data (whose response is unknown) in addition to labelled ones to build a general model within the training phase. Then, in the testing phase this model is used to predict the response value of a new unlabelled data (different from the first ones). Transductive learning takes profit of a set of labelled and unlabelled data to avoid the construction of a general model and directly predicts the response values of those same unlabelled data. To our knowledge, semi-supervised and transductive learning require a high number of test/unlabelled instances. In our case only one is provided, making those approaches unsuitable.

The main drawback of local learning approaches is their high computational cost, because for each new test data a model has to be constructed. However, it can be very useful in domains where only one test instance is provided.

As presented in Chapter 1, when the likelihood is intractable, a statistical approach was developed: the approximate Bayesian computation (ABC, Weiss and von Haeseler, 1998; Pritchard et al., 1999). This strategy relies on simulations according to Bayesian hierarchical models to generate pseudo-data. These artificial data are then compared to the test/observed one. The most basic algorithm is based on nearest neighbours (NN). Recently, in what we name ABC-RF, Breiman's machine learning algorithm of random forests (RF) proved to bring a meaningful improvement to the ABC paradigm in both a context of model choice (Pudlo et al., 2016) and parameter inference (Raynal et al., 2019, Chapter 4). Here, we focus on the model choice problem and thus the classification setting. Unlike some ABC techniques that take advantage of local methods, such as local adjustment (Beaumont, Zhang et al., 2002; Blum and François, 2010; Blum, Nunes et al., 2013), ABC-RF trains an eager RF to predict, later on, the observed data. It seems sub-optimal because in the ABC framework only the observed data is of interest for prediction. The ABC-RF strategy might therefore greatly benefit from local versions of RF.

Here, we focus on reviewing and proposing tree-based method to predict at best a specific data of interest. We start with some reminders on Breiman's RF algorithm. We then study local tree-based approaches depending on the way the localisation process is performed. In Section 5.3, we present/introduce internal modifications of the RF concerning the splitting rule. Then, we take an interest on modifying the random aspects of RF to turn them into local ones. We focus on modifying the sampling of individuals in Section 5.4, and the sampling of predictors in Section 5.5. Local weighting of votes is finally presented in Section 5.6. We empirically compare these strategies with the original, eager one in three examples where a local approach

might be of interest.

5.2 Reminders on Breiman's random forest

In the following we consider a classification problem. We use a set of d explanatory variables $X = (X_1, \ldots, X_d)$ to predict the categorical/discrete response Y.

The training data set is composed of N realisations $\left\{(y^{(i)}, x^{(i)})\right\}_{i=1,\ldots,N}$. We consider Breiman's random forest as the reference method we try to improve.

A classification RF is a set of randomised trees (Breiman, Friedman et al., 1984), each one partitioning the covariates space thanks to a series of allocation rules and assigning a class label as prediction to each partition. A binary tree is composed of internal and terminal nodes (a.k.a. leaves). For each internal node, a splitting rule on an explanatory variable is determined by maximising an information gain, dividing the training set in two parts. This process is recursively iterated until a stopping rule is achieved. The internal node encountering a stopping rule becomes terminal. For continuous covariates, a splitting rule compares a covariate X_j to a bound s, allocating to the left branch the data verifying the rule $X_j \leq s$, and to the right all others. For categorical covariates, the splitting rule is chosen among all the possible two way splits of the covariate categories.

The covariate index j and the bound s are chosen to maximise the decrease of impurity between the mother, denoted t, and the two resulting left and right daughter nodes, denoted t_L and t_R, (weighted by the number of data at each node). This gain associated to a covariate j and split value s is always non negative and is written as

$$G(j, s) = \mathcal{I}(t) - \left(\frac{\#t_L}{\#t} \mathcal{I}(t_L) + \frac{\#t_R}{\#t} \mathcal{I}(t_R) \right), \tag{5.1}$$

where # refers to the number of data in the associated node, and $\mathcal{I}(\cdot)$ is the impurity. The impurity, i.e. the heterogeneity at a given node, is measured with either the Gini index or the entropy. The objective is to select the allocation rule that reduces the impurity the most, in other terms that produces the highest gain.

Splitting events stop when one of the three following situation is reached:

- all individuals of the data set at a given node have the same response value (the node is pure),
- all individuals have the same covariate values,
- a node has less than $N_{\min}$ instances, $N_{\min}$ being an user-defined integer value, typically set to 1 for classification.

Once the tree construction is complete, each leaf predicts a model index, corresponding to the majority class of its instances. For a new set of explanatory variables x^*, predicting its model index implies passing x^* through the tree, following the path of binary rules, and the predicted value is the value associated to the leaf where it falls.

The RF method consists in bootstrap aggregating (bagging, Breiman, 1996) randomised (classification) trees. A large number of trees is trained on bootstrap samples of the training data set and m_{try} covariates are randomly selected at each internal node, on which the splitting rule will be defined. m_{try} is usually set at $\lfloor\sqrt{d}\rfloor$, where $\lfloor\cdot\rfloor$ denotes the floor function. The predicted value for a data x^* is the majority vote across all tree predictions. RF methods have some theoretical guarantees for sparse problems (Biau, 2012; Scornet et al., 2015). Moreover, it is well-known that their performances are quite good even when no tuning is made.

5.3 Local splitting rules

A first option to localise the tree construction is to change the information gain to the benefit of a local one. The idea is to use the test instance x^* to drive the splits and thus the tree construction.

Indeed, because the best split is selected on average, an eager tree may lead to many irrelevant splits to predict x^*, potentially putting aside interesting data. This behaviour results from data fragmentation (Fulton et al., 1996), i.e. from the recursive partitioning of the explanatory variables space to achieve good global performances. In the following we mention this phenomenon as the fragmentation problem. The very simple 2-class classification problem presented in Figure 5.1 illustrates this issue. The distribution of the training data set will induce, when possible, an initial cut for the tree construction in $X_1 \approx 0.5$, however, the unlabelled instance (represented by a black star) is in a region where a lot of relevant instances will be discarded after this first data split. A more pertinent first cut should occur in $X_2 \approx 0.25$. This problem, called fragmentation problem, also leads to less significant splitting rules at deeper levels of the tree construction since based on fewer instances. It is thus interesting to consider a local approach taking x^* into account.

It is interesting to note that building a local tree by modifying its internal construction results in building a path. Indeed, once a splitting rule is determined, this recursive process is only applied on the branch where x^* falls. Thus, a local random forest might be much faster for its construction compared to the eager version, especially if only one instance is of interest.

In this section we present the approach of Friedman et al. (1997) to build local decision trees, called lazy decision trees, and expand it for RF. We also present our attempts at using unidimensional or multidimensional kernels to give more weights to training samples closer to x^*.

5.3.1 Lazy decision trees

The lazy decision tree algorithm (LazyDT) is introduced in Friedman et al. (1997). Its objective is to take into account x^* during the tree construction. To do so, the information gain – depending on j and s – to maximise at each node is modified compared to criterion (5.1). Only the difference of impurity between the mother node t and the daughter node where x^* ends, denoted t^*, is considered. The resulting

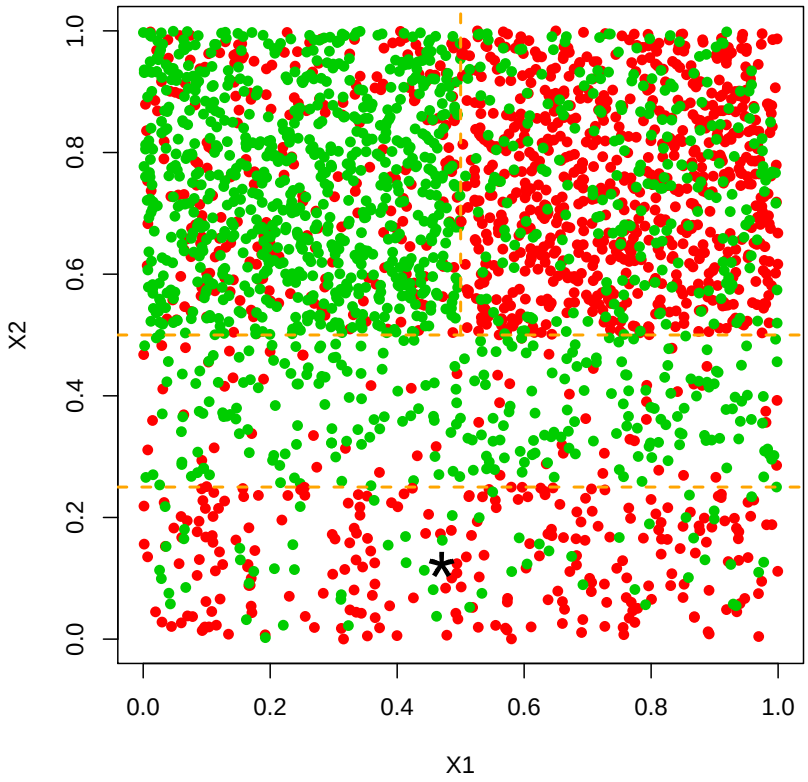


Figure 5.1 – An illustrative classification problem with 2 classes (red and green), containing two covariates describing four distinguishable regions (delimited by orange dashed lines) and an unlabelled data to classify (black star). This case will give rise to a fragmentation problem.

local information gain is defined by

$$G_w(j,s) = \mathcal{I}_w(t) - \mathcal{I}_w(t^*), \tag{5.2}$$

where $\mathcal{I}_w$ is the information gain computed with data at the node, weighted by a weight vector $w = (w^{(1)}, \ldots, w^{(N)})$ (described below). Note the absence of the proportion of individuals $\#t_L/\#t$ or $\#t_R/\#t$ compared to gain (5.1).

To ensure that this gain is always non-negative, to each instance $(y^{(i)}, x^{(i)})$ is assigned a weight $w^{(i)} = \frac{1}{n_k K}$ when $y^{(i)} = k$ and where n_k is the number of data labelled k at the mother node. Indeed, this weight ensures that all the weighted class frequencies are equal at the mother node, hence the weighted mother node impurity $\mathcal{I}_w(t)$ is maximal and the resulting gain always non-negative. The value of $\mathcal{I}_w(t)$ is equal to $\frac{K-1}{K}$ for the Gini index, and to $\log(K)$ for the entropy. Due to this constant value, the maximisation of (5.2) is equivalent to the minimisation of $\mathcal{I}_w(t^*)$. Note that the weights used at t^* and t are the same (limited to the sub-sample induced by the potential cut depending on j and s for t^*), but are recomputed after each accepted tree partition.

Moreover, those weights also avoid the problem that the impurity measures only use the classes proportions, without distinction of their associated class labels. Indeed, let us take the example of a two-class classification problem (1 and 2), where

the mother node contains 80% of data labelled 1 and 20% labelled 2. A splitting rule computed on unweighted data might induce, at the daughter node where x^* falls, 20% and 80% as proportions of 1 and 2, respectively. In this way, the non-weighted gain (5.2) would be zero, even though the discriminatory power of this cut is clearly non-null.

LazyDT provides three other major features: the use of discretised explanatory variables, the use of options and a condition on allowed split events.

- This algorithm only handles discretised explanatory variables. A preliminary discretisation is thus necessary, using for example the minimum description length principle (Fayyad and Irani, 1995). This was initially introduced to enhance the algorithm speed. According to our experiments this might also be useful when noise variables are considered as features. Indeed, for continuous covariates the construction might stop early due to some noise variables. For example, for a given covariate j, x_j^* can unfortunately be localised on one of the two covariate borders with few data all carrying some identical labels. The next splitting rule will isolate them with x_j^* because the resulting node will be pure and hence provide the maximum gain even though highly uninformative. The discretisation will be an asset in such situations since pure noise variables are more likely to be discretised into a unique category.

- The use of *options* is introduced. Indeed, because features can induce very similar information gains, Friedman et al. (1997) advise to develop all the paths – induced by splitting rules – achieving at least 90% of the maximal possible gain. The prediction associated to a tree for x^* becomes the prediction of the leaf with the maximal number of individuals in its majority class. We tried values different from 90% and it did not provide better results. Moreover, we studied an alternative to this method of prediction: because each option provides a prediction for x^*, we considered taking as final prediction the majority vote of these option predictions, but again results were not more conclusive.

- Finally, LazyDT only considers split values exactly equal to the values of x^* as potential cuts.

The LazyDT algorithm has undergone some developments. First, a bagged version to deduce class probabilities is presented in Margineantu and Dietterich (2003). A boosted version is then introduced in Fern and Brodley (2003), (however we will not compare to this one because an implementation is hard to find). Friedman et al. (1997) mention as main drawback for this method its inability to allow pruning. Fern and Brodley (2003) propose a heuristic to overcome this drawback, but their algorithm is not guaranteed to improve the classifier accuracy. Considering trees-ensemble overcomes this weakness, this is why in the following we consider a randomised bootstrapped LazyDT version (denoted LazyDRF).

5.3.2 Unidimensional kernel approach

Most local methods are based on weights depending on the proximity to x^*. This is the case of locally weighted regression (Cleveland, 1979; Cleveland and Devlin, 1988; Fan, 1993; Hastie and Loader, 1993). There are different ways to use weights in the context of tree methods. One can think of taking into account these weights to define the training sets on which trees are built. Such type of strategy is described in Section 5.4. In this section, we examine the possibility of using the weights during the tree construction, inside the tree splitting criterion.

In the wake of locally weighted regression, we set to each training individual and per covariate j, a weight depending on its proximity to x_j^*: the closer the higher. To do so, we consider for each covariate j a Gaussian kernel centred in x_j^*, providing weights

$$K_{h_j}(x_j^{(i)} - x_j^*), \quad \text{for} \quad i \in \{1, \dots, N\}.$$

We focus on a Gaussian kernel due to its smoothness and to avoid giving exactly zero weights to some individuals.

The choice of the bandwidth h_j is tricky. We consider as bandwidth value h_j the quantile of order α, $\mathbb{Q}_\alpha\left(\mid x_j^{(i)} - x_j^* \mid_{i=1,\dots,N}\right)$. The parameter α determines the shape of the kernel. For low α values, a higher weight is given to data close to x^*, and vice-versa. In our numerical experiments, we clearly observed that low values of α again result in cuts too close to x_j^*. We set $\alpha = 1$. Moreover, the bandwidth can eventually be recalculated at each internal node or kept constant during the tree construction. We observed very few differences and h_j is set as constant in the following.

For a given class label k, at the mother node t, this approach transforms the usual class frequencies (giving uniform weights among data) into some weighted class frequencies in the following way:

$$p_k = \frac{\sum_{i:x^{(i)}\in t} \mathbb{1}\{y^{(i)} = k\}}{\#t} \quad \Rightarrow \quad \widetilde{p}_k = \frac{\sum_{i:x^{(i)}\in t} \mathbb{1}\{y^{(i)} = k\} K_{h_j}(x_j^{(i)} - x_j^*)}{\sum_{\ell:x^{(\ell)}\in t} K_{h_j}(x_j^{(\ell)} - x_j^*)},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Moreover, the proportion of individuals, for example, at the left daughter node t_L implied by a cut $X_j \leq s$ is transformed from

$$\frac{\#t_L}{\#t} = \frac{\sum_{i:x^{(i)}\in t} \mathbb{1}\{x_j^{(i)} \leq s\}}{\#t}$$

into

$$\frac{\widetilde{\#t_L}}{\widetilde{\#t}} = \frac{\sum_{i:x^{(i)}\in t} \mathbb{1}\{x_j^{(i)} \leq s\} K_{h_j}(x_j^{(i)} - x_j^*)}{\sum_{\ell:x^{(\ell)}\in t} K_{h_j}(x_j^{(\ell)} - x_j^*)}. \tag{5.3}$$

The information gain to maximise (based on the Gini index) thus becomes

$$\underbrace{\sum_{k=1}^{K} \widetilde{p}_k(1-\widetilde{p}_k)}_{\widetilde{\mathcal{I}}(t)} - \Big(\frac{\widetilde{\#t_L}}{\widetilde{\#t}} \underbrace{\sum_{k=1}^{K} \widetilde{p}_{k,L}(1-\widetilde{p}_{k,L})}_{\widetilde{\mathcal{I}}(t_L)} + \frac{\widetilde{\#t_R}}{\widetilde{\#t}} \underbrace{\sum_{k=1}^{K} \widetilde{p}_{k,R}(1-\widetilde{p}_{k,R})}_{\widetilde{\mathcal{I}}(t_R)} \Big), \tag{5.4}$$

where $\widetilde{p}_{k,L}$ and $\widetilde{p}_{k,R}$ are the weighted proportions of class k at the left and right daughter nodes, respectively. The first term $\widetilde{\mathcal{I}}(t)$ is important and cannot be omitted contrary to the eager version, because it depends on the covariate index.

We use this local Gini index during the tree construction and do not modify the default values for the RF parameters m_{try} and $N_{\min}$. For each tree, the associated prediction is the usual majority vote at the leaf.

Our local slitting rule is similar to the one used in the recent method of Armano and Tamponi (2018). In this work, an improvement to RF is introduced by using an ensemble of local trees. Each tree is trained giving more weights to training data around a centroid, which is sampled among the training instances, and different centroids are considered to map the whole predictor space. Although using a local Gini index, this approach is more of an eager one than a local one. Indeed, no test instance is involved during the forest construction. Moreover, per tree a multidimensional kernel is used.

5.3.3 Multidimensional kernel approach

In the spirit of Armano and Tamponi (2018), it is natural to extend the approach introduced in Section 5.3.2 with a multidimensional kernel centred in x^*. We assign to each data $(y^{(i)}, x^{(i)})$ a weight

$$K_V(x^{(i)} - x^*) = \exp\left(-\frac{1}{2}(x^{(i)} - x^*)^\top V (x^{(i)} - x^*)\right),$$

where V is a scaling matrix of the Gaussian kernel. Similarly to Section 5.3.2 we consider for V the diagonal matrix made of the α quantiles, i.e.

$$V = \text{diag}\left(\mathbb{Q}_\alpha\left\{\mid x_1^{(i)} - x_1^* \mid_{i=1,\dots,N}\right\}, \dots, \mathbb{Q}_\alpha\left\{\mid x_d^{(i)} - x_d^* \mid_{i=1,\dots,N}\right\}\right).$$

As for the unidimensional kernel approach, using extensive numerical experiments, we observed that low values of α result in cuts too close to x_j^* and we set $\alpha = 1$. Also, the weights are fixed during the tree construction.

The weighted frequency for a given class label k becomes

$$\widetilde{p}_k = \frac{\sum_{i=1}^N \mathbb{1}\{y^{(i)} = k\} K_V(x^{(i)} - x^*)}{\sum_{\ell=1}^N K_V(x^{(\ell)} - x^*)}.$$

The weighted proportions of individual at the node (5.3) are transformed in the same way, resulting in a gain criterion analogous to (5.4).

The major benefit of such weights is that they do not depend on the covariate index, thus the usual tree prediction, i.e. the majority class at the leaf where x^* falls, can be replaced by a more coherent strategy with the tree construction, using as prediction the class with the maximal weighted class proportion at the leaf. Thus, the prediction for x^* provided by the b-th tree is

$$\hat{y}_b^* = \underset{1 \le k \le K}{\arg\max}\, \widetilde{p}_k.$$

The forest prediction for x^* is the usual majority vote of the tree predictions.

5.4 Local weighting of individuals

To avoid the fragmentation problem, instead of modifying the way the predictor space is partitioned, one can consider directly targeting the region of interest, i.e. samples similar to x^*. In this part, we focus on strategies acting on the individuals sampling schemes involved at the first step of a tree construction, replacing the usual bootstrap sampling with a local one.

5.4.1 Weighted bootstrap

Xu et al. (2016) propose to perform weighted bootstrap sampling, where a measure of proximity between x^* and the training data is used to compute the weights. This algorithm is entitled Case-Specific Random Forest (CSRF, Algorithm 5.1).

An individual closer to x^* will have higher weight and will more likely be picked in the bootstrap sampling. However, such weights depend heavily on the choice of the proximity measure, especially in a high dimensional setting and with many irrelevant explanatory variables. This is why in this framework the proximity measure will be automatically computed thanks to a bagged tree-ensemble (i.e. with $m_{\text{try}} = d$).

Indeed, for a given tree, x^* ends in a leaf with some training data. For each $x^{(i)}$, counting the number of trees where x^* and $x^{(i)}$ end in the same leaf allows to compute the contribution of $x^{(i)}$ to predict x^*, denoted $\omega^{(i)}$ in Algorithm 5.1. The deduced weights are then used to perform weighted bootstrap sampling during the training of a new RF. This process can be seen as a nearest neighbours strategy: per tree, a leaf provides a certain amount of neighbours to x^*, those are then accumulated over all the trees to deduce instance weights.

This algorithm highly depends on the depth of the first RF trees, hence a pivotal parameter for this strategy is $N_{\min}$, the minimal number of observations at an internal node. The higher the $N_{\min}$, the shallower the trees will be. Hence, low values of $N_{\min}$ result in putting more weights on the closest individuals to x^*, and vice-versa. We tried various values of $N_{\min}$ in our experiments.

Algorithm 5.1 : CSRF – local weighting of individuals

1. Grow B_1 bootstrapped trees with $m_{\text{try}} = d$ and a given $N_{\min}$ value;
2. For each training data $(y^{(i)}, x^{(i)})$, count $c^{(i)}$ the number of times $x^{(i)}$ and x^* ends in the same leaf;
3. Compute the resampling probability of the training individual i relative to x^* as $\omega^{(i)} = \frac{c^{(i)}}{\sum_{\ell=1}^{N} c^{(\ell)}}$, for $i \in \{1, \ldots, N\}$;
4. Train a usual RF of size B_2 with bootstrap resampling probabilities $\omega^{(1)}, \ldots, \omega^{(N)}$ and deduce the prediction for x^*.

5.4.2 Nearest neighbours: 0/1 weights

A more intuitive idea is based on the deduction of κ nearest neighbours to x^*, which are then used to train a RF. Fulton et al. (1996) propose several methods

to extract data local to x^* – the best one being based on NN – in order to build decision trees on this restricted training set. Galván et al. (2009) also mention the possibility of pre-selecting closest observations to x^* (possibly with replicates) at first and applying any machine learning algorithm on these data set. This kind of strategy is more recently applied in a text classification framework by Salles et al. (2018), and shows good improvements in terms of classification errors compared to RF (and other ones).

Those approaches are closely related to CSRF (Section 5.4.1) since considering NN during a preliminary step is equivalent to giving 0/1 sampling weights (with or without replacement).

In Section 5.7, we compare the use of a preliminary selection of nearest neighbours to x^* followed by a usual RF training, this strategy is denoted in the remaining by NN-RF, for nearest neighbours - random forest. The main issue of such approaches (and local ones in general) is the difficulty to choose this neighbourhood.

5.5 Local weighting of covariates

Instead of acting on the bootstrap resampling of RF, we propose to operate on the covariates subsampling which occurs at each internal node. In the wake of Section 5.4.1 we propose to weight covariates during the RF trees construction depending on their importance to predict x^*. In the following we mention it as LVI-RF (for local variable importance - random forest).

We study the influence of considering sampling probability weights on explanatory variables. The principle is detailed in Algorithm 5.2 and is very similar to Algorithm 5.1.

We take profit of a first RF construction with default parameters to deduce covariate importance to predict x^*: in a very intuitive way we pass x^* through each tree of the RF, and count the number of times each covariate is involved in a splitting rule to allocate x^*. We can then easily deduce some predictor weights, and we propose to introduce them into the usual RF covariate sampling, so that a covariate with high weight is more likely to be drawn in the m_{try}-sample.

Our thought is that using such weights might improve the prediction accuracy of the RF, especially in a sparse framework, by avoiding useless data fragmentation according to irrelevant predictors and potential loss of useful training data for the prediction of x^*. Moreover, a different set of explanatory variables might be useful to predict different test instances, thus thanks to a local measure of variable importance we also try to ensure that interesting covariates are more likely to be sampled during the tree construction. Finally, in the case of a huge number of noise covariates, even though RF can handle a large number of features, useful ones are very unlikely to be drawn during the tree construction, deteriorating the algorithm performance. In counterpart, weighted covariate sampling might increase the prediction correlation between the RF trees and alter the performance of the global tree ensemble.

Some approaches dealing with covariate weighting have been studied in a non-local framework. Amaratunga et al. (2008) propose the enriched random forests

Algorithm 5.2 : Local weighting of covariates

1 Grow B_1 randomised trees with $m_{\text{try}} = \lfloor \sqrt{d} \rfloor$ and $N_{\min} = 1$;
2 For each covariate $j \in \{1, \dots, d\}$, count v_j the number of times X_j has been used during the paths followed by x^*;
3 Compute the resampling probability of the covariate j relative to x^* as $p_j = \frac{v_j}{\sum_{\ell=1}^{d} v_\ell}$, for $j \in \{1, \dots, d\}$;
4 Train a usual RF of size B_2 with covariate resampling probabilities $p_1, \dots, p_d$ at each internal node and deduce the prediction for x^*.

in an extremely noisy feature space, where covariate sampling is modified using global weights. Maudes et al. (2012), with their random feature weights approach, investigate the use of non-uniform sampling of covariates, changing for each tree.

5.6 Local weighting of votes

The final prediction of a classical RF is the majority vote of all trees, hence they all have equal weight. However a given tree might provide very good predictions on some test instances, but perform very poorly on others. This is why a strategy for building local random forests is based on weighting tree predictions depending on their ability to correctly predict instances similar to x^*. Majority vote is hence replaced with locally weighted vote.

In the instance-based framework, Robnik-Šikonja (2004), Tsymbal et al. (2006) and then Zhang et al. (2013) investigate this idea. Given a test instance x^*, κ neighbours are selected based on the proximity measure introduced in Breiman (2001), (i.e. the average number of times two data end in the same leaf) to compute a per-tree error score. These scores are further used to select and weight trees and to provide a final weighted-vote prediction.

5.6.1 Dynamic voting and selection

This section describes the methodology of Tsymbal et al. (2006), called Dynamic Voting with Selection Random Forest (DVSRF). A first RF is trained thanks to which κ nearest neighbours to x^* are selected. The quality of the b-th tree toward x^* is then evaluated as the average margins of the out-of-bag κ instances, weighted by proximities, i.e.

$$w_b(x^*) = \frac{\sum_{i=1}^{\kappa} \mathbb{1}\{x^{(i)} \in \text{OOB}_b\}\, \sigma(x^*, x^{(i)})\, \text{mr}_b(x^{(i)})}{\sum_{\ell=1}^{\kappa} \mathbb{1}\{x^{(\ell)} \in \text{OOB}_b\}\, \sigma(x^*, x^{(\ell)})}, \tag{5.5}$$

where OOB_b is the set of out-of-bag data for the b-th tree, $\sigma(x^*, x^{(i)})$ is the proximity measure provided by the RF, to the power of 3, and the margin function $\text{mr}_b(x^{(i)})$ is equal to 1 if the b-th tree predicts $y^{(i)}$ correctly, -1 otherwise. Weights (5.5) are then normalised to be positive and to sum to one. Finally, the prediction for x^* is

computed using the majority class of the weighted tree vote proportions

$$\hat{y}^* = \underset{1\leq k\leq K}{\arg\max}\, p_{\text{DVS},k} \qquad \text{where} \qquad p_{\text{DVS},k} = \frac{\sum_{b=1}^{B} \mathbb{1}\{\hat{y}_b^* = k\} w_b(x^*)}{\sum_{\ell=1}^{B} w_\ell(x^*)}, \tag{5.6}$$

where $\hat{y}_b^*$ denotes the original prediction of the b-th tree for x^*.
A predefined number of trees denoted B_{sel} (usually half of B), carrying the highest weights, can be selected and used for the final prediction, modifying weighted predictions (5.6) accordingly.

5.6.2 Kernel weighted voting

In the same spirit, we investigate the use of a multidimensional kernel as similarity measure (presented in Section 5.3.3) and we replace the margin function by the simpler alternative $\mathbb{1}\{\hat{y}_b^{(i)} = y^{(i)}\}$ indicating whether the b-th tree prediction for $x^{(i)}$, denoted $\hat{y}_b^{(i)}$, is correct or not.

Using the same notations as above, the b-th tree weight is hence computed in the following way:

$$w_b(x^*) = \frac{\sum_{i=1}^{N} \mathbb{1}\{x^{(i)} \in \text{OOB}_b\}\, K_V(x^{(i)} - x^*)\, \mathbb{1}\{\hat{y}_b^{(i)} = y^{(i)}\}}{\sum_{\ell=1}^{N} \mathbb{1}\{x^{(\ell)} \in \text{OOB}_b\}\, K_V(x^{(\ell)} - x^*)}. \tag{5.7}$$

All N labelled data are used for the weight computation, their importance being measured by the kernel. α is again set to 1 and tree selection is not performed. In the following this proposal is denoted as KV-RF (for kernel voting - random forest).

5.7 Numerical experiments

In this section, we compare the previously presented methods – summarised in Table 5.1 – on three examples: two Gaussian mixtures and a population genetics example.

Acronym	Method	Section
LazyDRF	Lazy decision RF	5.3.1
UK-RF	Unidimentional kernel RF	5.3.2
MK-RF	Multidimentional kernel RF	5.3.3
CSRF	Case-specific RF	5.4.1
NN-RF	Nearest-neighbours RF	5.4.2
LVI-RF	Local variable importance RF	5.5
DVSRF	Dynamic voting with selection RF	5.6.1
KV-RF	Kernel voting RF	5.6.2

Table 5.1 – Summary of the compared methods, as well as their acronyms and the sections where they are presented.

Methods are run ten times on the same test data set. The average and standard deviation of the ten resulting misclassification error rates, per method, are reported as a measure of performance. Note that in order to recover the predictions for the

whole test table, each local algorithm is reapplied to each test data. The two first Gaussian examples have the advantage of being simple enough to compute the Bayes classifier which gives the optimal error rate.

When not specified, the random forests are built using the default parameters, i.e. trees are maximal ($N_{\min} = 1$), and the covariate sampling parameter is $m_{\text{try}} = \lfloor \sqrt{d} \rfloor$. Moreover, each forest is made of 100 trees, meaning CSRF and LVI-RF use a total of 200 trees. Additional/different tuning parameters are specified in the displayed result tables.

The methods involving classic RF (CSRF, NN-RF, LVI-RF, DVSRF, KV-RF) use the R package `ranger` for their construction. The remaining were programmed by myself from scratch. The R codes for the different algorithms and how to run the examples presented below are available at https://github.com/LouisRaynal/local-rf.

5.7.1 Balanced Gaussian mixture example

We consider 40-dimensional data from four classes $(1, 2, 3, 4)$. The classes have equal prior probabilities: $p_1 = p_2 = p_3 = p_4 = 1/4$. The data are generated from 20-dimensional Gaussian distributions and 20 noise explanatory variables are added, simulated according to a uniform distribution $\mathcal{U}_{[0;10,000]}$.

The training data set is of size $3,000$ and sampled among the 4 classes with equal probabilities. 500 simulations are used as testing data set, also sampled equally among the 4 models.

The parameters associated to the 20-multidimensional Gaussian distribution are

$$\begin{aligned}
&\mu_1 = (0.8, 3, 1, 2.5, \ldots, 1, 2.5)^\top, && \mu_2 = (3.2, 3, 2.5, 2.5, \ldots, 2.5, 2.5)^\top,\\
&\mu_3 = (2, 1, 2, 2.3, \ldots, 2, 2.3)^\top, && \mu_4 = (2, 0, 2, 1.8, \ldots, 2, 1.8)^\top,\\
&\Sigma_1 = \text{diag}(3, 3, 3, 1, \ldots, 3, 1), && \Sigma_2 = \text{diag}(3, 3, 3, 5, \ldots, 3, 5),\\
&\Sigma_3 = \text{diag}(4, 1, 4, 1, \ldots, 4, 1), && \Sigma_4 = \text{diag}(2.5, 1, 2.5, 1, \ldots, 2.5, 1).
\end{aligned}$$

The two first dimensions, represented in Figure 5.2, are the most relevant for discriminating between the four classes. Indeed, although the remaining ones can provide information to identify the class labels, they are more overlapping with each others and hence less informative.

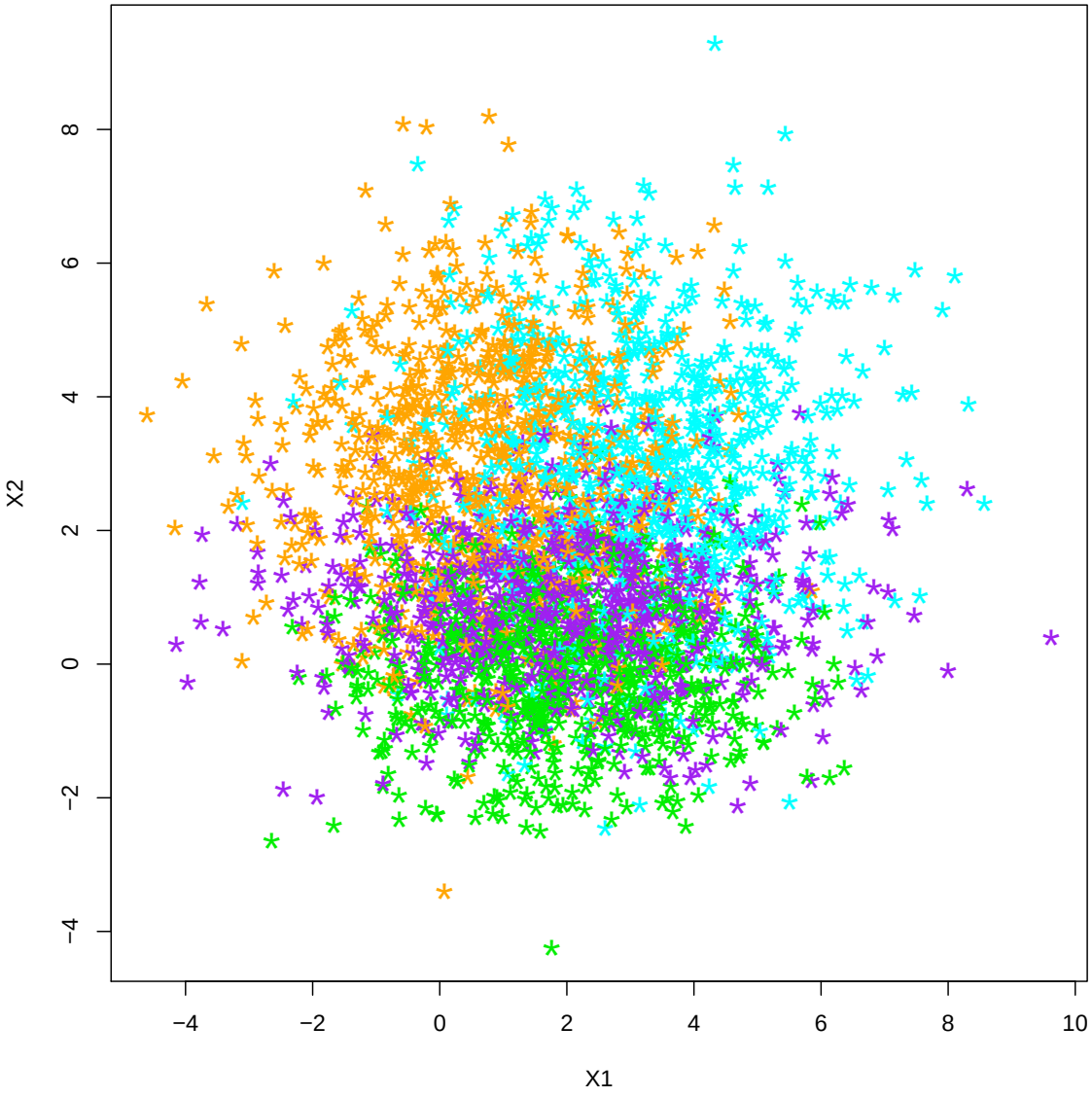


Figure 5.2 – First Gaussian example: two first explanatory variables X_1 and X_2 ; colours indicate the class labels (1-orange, 2-cyan, 3-purple, 4-green).

	Setting	Error rate (standard deviation)
Bayes classifier		**12.6**
Bagged CARTs		27.02 (0.936)
Random forest		**22.52** (0.648)
LazyDRF		24.64 (0.821)
UK-RF		23.8 (0.706)
MK-RF		**22.32** (0.880)
CSRF	$N_{\min} = 5$ (in 1st forest)	24.62 (1.680)
CSRF	$N_{\min} = 10$ (in 1st forest)	25.42 (1.397)
CSRF	$N_{\min} = 50$ (in 1st forest)	23.72 (1.259)
CSRF	$N_{\min} = 150$ (in 1st forest)	23.7 (1.055)
CSRF	$N_{\min} = 250$ (in 1st forest)	23.12 (0.527)
CSRF	$N_{\min} = 350$ (in 1st forest)	23.8 (0.693)
NN-RF	$\kappa = 1,000$	23.14 (1.170)
NN-RF	$\kappa = 1,500$	23.42 (0.643)
NN-RF	$\kappa = 2,500$	22.84 (1.028)
LVI-RF	$m_{\text{try}} = d$ (in 1st forest)	27.68 (0.985)
LVI-RF	$m_{\text{try}} = \lfloor \sqrt{d} \rfloor$ (in 1st forest)	23.76 (0.997)
DVSRF	$\kappa = 3,000$, $B_{\text{sel}} = 100$	23.02 (1.047)
DVSRF	$\kappa = 3,000$, $B_{\text{sel}} = 50$	23.48 (1.297)
KV-RF		**22.8** (0.947)

Table 5.2 – First Gaussian example: prediction error rate (in %) ; the four lowest errors are displayed in bold characters.

The results are presented in Table 5.2. The optimal Bayes classifier provides an error rate of 12.6%, all tree-based methods studied here are very far from such performance. The method that provides the lowest error rate is the local splitting rule based on a multidimensional kernel. The kernel weighted voting strategy also provides low error rate. However, the results obtained with local methods are very similar to the ones obtained with a standard RF.

5.7.2 Unbalanced Gaussian mixture example

We introduce some modifications to the previous Gaussian example.

We still consider four classes but their model prior probabilities are $p_1 = p_2 = 0.4$ and $p_3 = p_4 = 0.1$. The training data set is made of $3,000$ samples drawn among the four classes according to these probabilities. The testing set considers 500 data equally sampled among the two classes 3 and 4, the least frequent ones. In this example we therefore measure the prediction accuracy of low-frequency data.

The two first covariates, represented in Figure 5.3, are still the most important ones, however we slightly modified the Gaussian parameters (the two first diagonal terms for Σ_1 and Σ_2 are now 2 and 1) to induce as best split rule for a CART: $X_1 \approx 2$. This example hence becomes an illustration of the fragmentation problem we mentioned earlier (Figure 5.1). Indeed, the first cut produced by the eager RF algorithm – if this covariate is sampled – will split the elements labelled 3 and 4 in half (at $X_1 \approx 2$). It implies the loss of some potentially relevant training data to predict those two classes. We hope local approaches can handle such an example

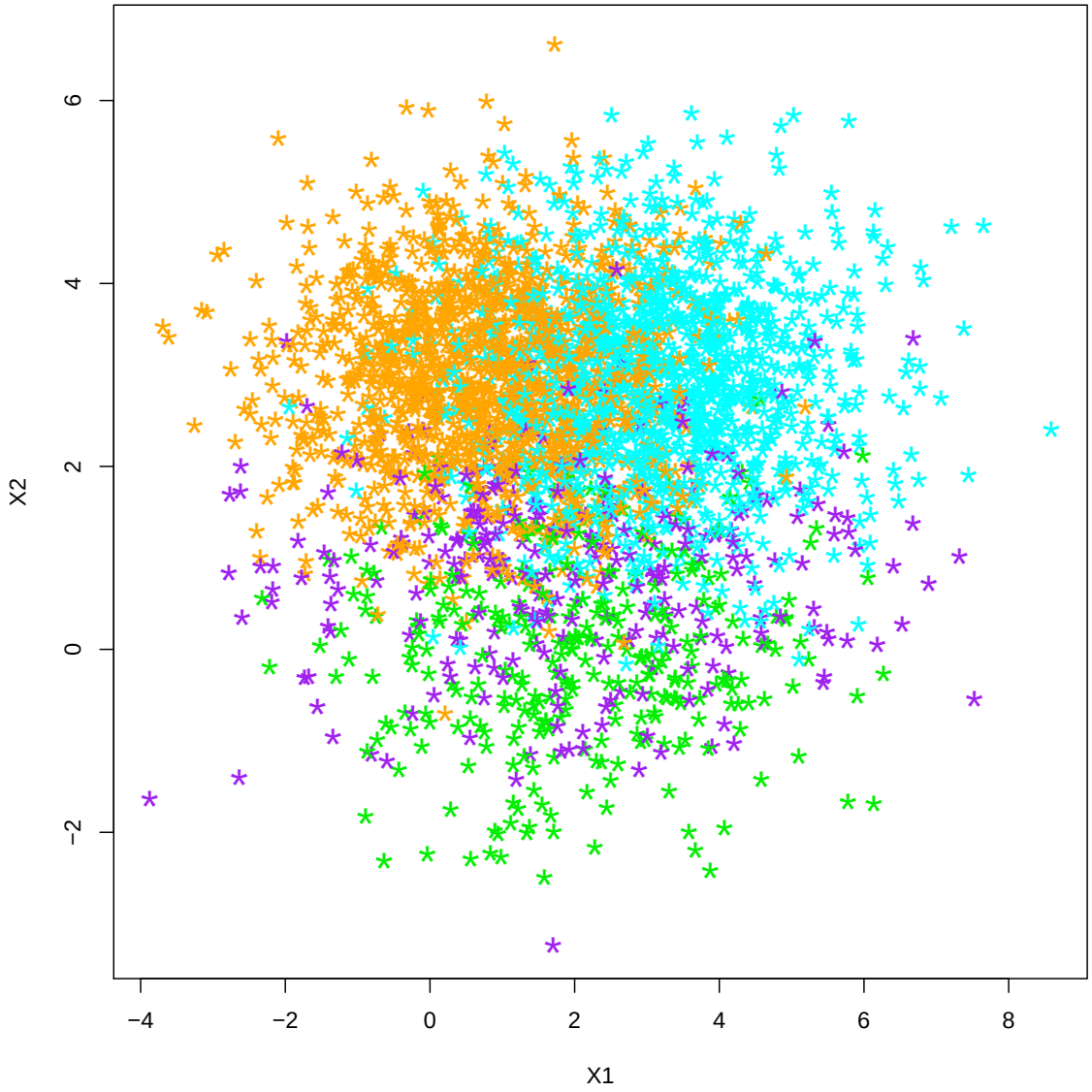


Figure 5.3 – Second Gaussian example: two first explanatory variables X_1 and X_2 ; colours indicate the classes (1-orange, 2-cyan, 3-purple, 4-green).

which also contains very unbalanced classes proportions.

The results are presented in Table 5.3. As for the previous example, all tree-based methods studied here are very far from the performance of the optimal Bayes classifier. The method that provides the lowest error rate is the local weighting of covariates. The nearest neighbour strategy also provides low error rate. However, conditionally to the fact that this example was chosen to favour them, the results obtained with local methods are disappointing and not significantly better than those obtained with eager strategies.

5.7.3 Population genetics example

We now compare a set of local strategies on a basic population genetics example introduced in Pudlo et al. (2016). The historical link between three populations of a given species is of interest. More precisely, we are interested in studying whether a

	Setting	Error rate (standard deviation)
Bayes classifier		**25**
Bagged CARTs		45 (1.215)
Random forest		46.16 (0.826)
LazyDRF		75.4 (0.894)
UK-RF		48.18 (0.877)
MK-RF		48 (0.800)
CSRF	$N_{\min} = 5$ (in 1st forest)	47.84 (0.923)
CSRF	$N_{\min} = 10$ (in 1st forest)	47.44 (1.173)
CSRF	$N_{\min} = 50$ (in 1st forest)	45.9 (1.300)
CSRF	$N_{\min} = 150$ (in 1st forest)	45.92 (1.700)
CSRF	$N_{\min} = 250$ (in 1st forest)	46.3 (1.175)
CSRF	$N_{\min} = 350$ (in 1st forest)	47.68 (1.612)
NN-RF	$\kappa = 1,000$	44.56 (1.336)
NN-RF	$\kappa = 1,500$	**43.96** (1.289)
NN-RF	$\kappa = 2,500$	46.56 (0.913)
LVI-RF	$m_{\text{try}} = d$ (in 1st forest)	**43.02** (1.291)
LVI-RF	$m_{\text{try}} = \lfloor \sqrt{d} \rfloor$ (in 1st forest)	**41.66** (0.833)
DVSRF	$\kappa = 3,000$, $B_{\text{sel}} = 100$	46.34 (1.340)
DVSRF	$\kappa = 3,000$, $B_{\text{sel}} = 50$	47.02 (0.791)
KV-RF		45.7 (0.976)

Table 5.3 – Second Gaussian example: prediction error rate (in %) ; the four lowest errors are displayed in bold characters.

third population emerged from a first or a second population, or whether it emerged from a mixture between the two first ones. This problem is hence a three classes classification question. The data is made of $1,000$ autosomal single-nucleotide polymorphisms (SNPs). We assume that the distances between these loci on the genome are large enough to neglect linkage disequilibrium, we hence consider them as having independent ancestral genealogies.

The data is summarised thanks to $d = 48$ summary statistics available within the DIYABC software for SNP markers (Cornuet, Pudlo et al., 2014), which is also used to simulate training and test sets respectively of size $10,000$ and 500, equally distributed among the three scenarios. Moreover, the data are constrained to be drawn in the $[-1; 1]^2$ square on the LDA axes projections graph, which is a region where scenarios are hard to discriminate, see Figure 5.4.

We compare on this example the RF method, and some local approaches: CSRF, NN-RF, LVI-RF and KV-RF.

In the example, again, RF does not provide better results than bagging. Using local covariates importance with a first bagged RF results ($m_{\text{try}} = d$) allows an improvement of the error, its standard deviation is notably reduced. However, the local methods do not outperform the bagging approach.

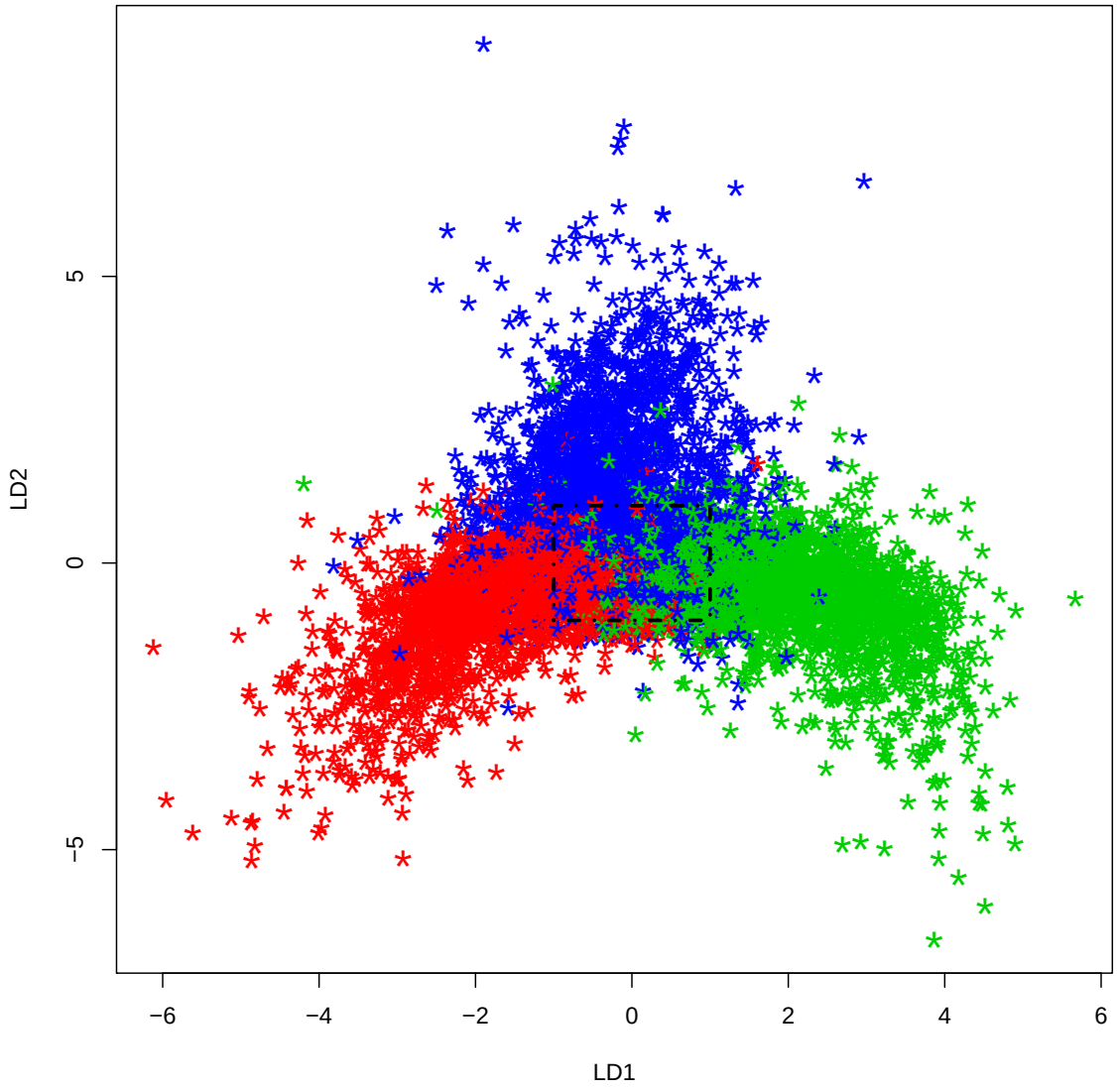


Figure 5.4 – Population genetics example: projections on the LDA axes of the $10,000$ training instances ; colours represent scenario indices: red for model 1, blue for model 2 and blue for model 3 ; the hard to discriminate $[-1;1]^2$ region is represented by black dashed lines.

	Setting	Error rate (standard deviation)
Bagged CARTs		**38.86** (0.985)
Random forest		**40.40** (0.806)
CSRF	$N_{\min} = 5$ (in 1st forest)	42.28 (1.248)
CSRF	$N_{\min} = 10$ (in 1st forest)	42.62 (1.565)
CSRF	$N_{\min} = 50$ (in 1st forest)	41.18 (1.216)
CSRF	$N_{\min} = 150$ (in 1st forest)	41.12 (1.163)
CSRF	$N_{\min} = 250$ (in 1st forest)	**40.40** (1.020)
CSRF	$N_{\min} = 350$ (in 1st forest)	41.14 (1.458)
NN-RF	$\kappa = 1,000$	42.34 (0.844)
NN-RF	$\kappa = 1,500$	41.6 (0.838)
NN-RF	$\kappa = 2,500$	41.3 (1.042)
LVI-RF	$m_{\text{try}} = d$ (in 1st forest)	**38.88** (0.620)
LVI-RF	$m_{\text{try}} = \lfloor \sqrt{d} \rfloor$ (in 1st forest)	40.94 (0.989)
KV-RF		40.64 (1.184)

Table 5.4 – Population genetics example: prediction error rate (in %) ; the four lowest errors are displayed in bold characters.

5.8 Conclusion

In this chapter, we review, discuss and propose local tree-based methods, strategies taking into account a specific test data during the learning process. We focus on classification problems. The results are not up to our expectations. We considered three examples where local methods seemed useful but we did not obtained conclusive results.

Our proposal to introduce weights in the splitting criterion is problematic. Putting too high weights around x^* results in irrelevant cut-points, closer to x^* compared to RF. It induces large correlations between the trees in the forest, and the quality of prediction is impacted negatively. This is why $\alpha = 1$ is preferred. With this choice, even if it localised the trees, we obtained results very similar to the ones of RF. This is also the case of the kernel voting RF strategy.

The CSRF of Xu et al., 2016, the nearest neighbour weights and the local weighting of covariates strategies can give good performance but depend on tuning parameters. For instance, the CSRF brings better performance when the tree depth is low, i.e. high $N_{\min}$. However, generally, results provided by these local methods are very similar to eager ones, and no great benefit is observed on our three examples. When looking at the very small benefits in terms of prediction error rate compared to the non-local approaches, we can say that local strategies are clearly not worth the additional computational cost. Especially since most of them require the choice of a tuning parameter, characterising the weights given to instances surrounding x^*.

Chapter 6

Applications in population genetics

This chapter is based on two collaborative papers in the field of population genetics. The first one is Estoup, Raynal et al. (2018) published in *Journal de la Société Française de Statistique*, the second one is Chapuis, Raynal et al. (2019) submitted to *PCI Evol Biol*. My main contributions in these papers were to develop the required methodological and statistical elements, and of course to participate in their writing. We provided the statistical methodologies as well as the computational tools, thanks to the `abcrf` R package we developed and improved accordingly.

Contents

6.1 Introduction

This chapter focuses on population genetics applications of the ABC-RF methodologies: for model choice introduced by Pudlo et al. (2016) and for parameter inference presented in Chapter 4 and in Raynal et al. (2019). Moreover, we propose some improvements to these ABC-RF methodologies which are available in our R package `abcrf`.

- In the model choice setting, more precisely to select the best evolutionary scenario using present genetic information, it can be difficult to disentangle with trust whether or not a demographic event (as a change of population size, an admixture between populations,...) is important or not. This statement becomes even more true when a high number of populations and events are considered. To obtain a better understanding of the scenario and identifying which events are hard to discriminate and which ones are not, we add to the ABC-RF model choice strategy (Pudlo et al., 2016) the possibility to study groups of scenarios instead of individuals ones.
- For parameter estimation, the random forest algorithm provides some out-of-bag error measurements, giving insights regarding its predictive performance (over the entire covariate space). However, this type of error is not related to the observation we are interested in, when the prediction accuracy may depend on the area of the predictor space it is located in. For this reason we propose some posterior measures of error (computed conditionally on the observed data to predict) thanks to regression random forests. These errors are computed and compared in Section 6.4 for the time of divergence between two desert locust populations in Africa, which include or not some informed mutational prior distributions.

In Section 6.2 we recall the ABC-RF approach for model choice (Pudlo et al., 2016), to present the grouping strategy as well as posterior measures of error. We then present in Sections 6.3 and 6.4 two applications for population genetics problems. The first one focuses only on model choice, the second exposes for the first time a full ABC-RF analysis, including model choice and parameter inference.

6.2 Statistical improvements in the ABC-RF methodologies

6.2.1 ABC-RF model choice on groups of models

We here start by bringing some recalls on the ABC-RF strategy for model choice, to then introduce the model grouping approach.

6.2.1.1 Recalls on ABC-RF model choice

Let us consider M Bayesian parametric models. For a given model index $m \in \{1, \ldots, M\}$, a prior probability $\mathbb{P}(\mathcal{M} = m)$ is defined, with θ_m its associated para-

meters and $f_m(\mathbf{y} \mid \theta_m)$ its likelihood for the observation. Our targets are the model posterior probabilities

$$\mathbb{P}(\mathcal{M} = m \mid \mathbf{y}) \propto \mathbb{P}(\mathcal{M} = m) \int f_m(\mathbf{y} \mid \theta_m)\pi_m(\theta_m)d\theta_m.$$

Considering the standard 0-1 symmetric loss function, the selected model is the one with the maximum of the model posterior probabilities

$$\arg\max_{1\leq m\leq M} \left\{ \mathbb{P}(\mathcal{M} = m) \int f_m(\mathbf{y} \mid \theta_m)\pi_m(\theta_m)d\theta_m \right\}.$$

The likelihood expressions $f_m(\mathbf{y} \mid \theta_m)$ are not available for each model in competition, therefore it is not possible to calculate $\int f_m(\mathbf{y} \mid \theta_m)\pi_m(\theta_m)d\theta_m$. To avoid these difficulties, Grelaud et al. (2009) have introduced a model choice version of the nearest-neighbours ABC scheme, see Algorithm 6.1. The problem is viewed as

Algorithm 6.1 : Nearest-neighbours ABC model choice scheme

for $i \leftarrow 1$ **to** N **do**
  Generate $m^{(i)}$ from the prior $\mathbb{P}(\mathcal{M} = m)$;
  Generate $\theta'_{m^{(i)}}$ from the prior $\pi_{m^{(i)}}(\cdot)$;
  Generate $\mathbf{x}^{(i)}$ from the model $f_{m^{(i)}}(\cdot \mid \theta'_{m^{(i)}})$;
  Calculate $\rho^{(i)} = \rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}})$;
end
Order the distances $\rho^{(1)}, \ldots, \rho^{(N)}$;
Select the model using the majority rule among the k-smallest distances index set;

a classification question and is solved using nearest-neighbours classifiers. Due to the curse of dimensionality, the methodology associated to Algorithm 6.1 has major difficulties. Typically, to ensure reliability of the method, the number of simulations should be large and the number of summary statistics small.

However, exploiting a large number of summary statistics is not an issue for some machine learning methods. The idea of Pudlo et al. (2016) is to train random forests (Breiman, 2001) on a set of simulated data, called reference table. The use of random forests is motivated by some theoretical guarantees for sparse problems they exhibit (Biau, 2012; Scornet et al., 2015) as well as other advantages described in Chapter 2. The generation process of a reference table made of N elements is recalled in Algorithm 6.2. The output takes the form of a matrix containing simulated model indexes, parameters and summary statistics, as below:

$$\begin{bmatrix} m^{(1)} & \theta_{m^{(1)}} & \eta_{\mathbf{x}^{(1)},1} & \eta_{\mathbf{x}^{(1)},2} & \cdots & \eta_{\mathbf{x}^{(1)},d} \\ m^{(2)} & \theta_{m^{(2)}} & \eta_{\mathbf{x}^{(2)},1} & \eta_{\mathbf{x}^{(2)},2} & \cdots & \eta_{\mathbf{x}^{(2)},d} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m^{(N)} & \theta_{m^{(N)}} & \eta_{\mathbf{x}^{(N)},1} & \eta_{\mathbf{x}^{(N)},2} & \cdots & \eta_{\mathbf{x}^{(N)},d} \end{bmatrix}.$$

The ABC-RF strategy for model choice is described in Algorithm 6.3. The output is the affectation of $\mathbf{y}$ to a model (scenario), this decision being made based on the majority class of the RF tree votes.

Algorithm 6.2 : Generation of a reference table with N elements

for $i \leftarrow 1$ **to** N **do**
  Generate $m^{(i)}$ from the prior $\mathbb{P}(\mathcal{M} = m)$;
  Generate $\theta_{m^{(i)}}$ from the prior $\pi_{m^{(i)}}(\cdot)$;
  Generate $\mathbf{x}^{(i)}$ from the model $f_{m^{(i)}}(\cdot \mid \theta_{m^{(i)}})$;
  Compute $\eta_{\mathbf{x}^{(i)}} = \{\eta_{\mathbf{x}^{(i)},1}, \ldots, \eta_{\mathbf{x}^{(i)},d}\}$;
end

Algorithm 6.3 : ABC-RF model choice

Input : a reference table used as learning set, made of N elements, each one composed of a model index $m^{(i)}$ and d summary statistics. A possibly large collection of summary statistics can be used, including some obtained by machine-learning techniques, but also by scientific theory and practitioner knowledge

Learning : construct a classification random forest $\hat{m}(\cdot)$ to infer model indexes

Output : apply the random forest classifier to the observed data $\eta_{\mathbf{y}}$ to obtain $\hat{m}(\eta_{\mathbf{y}})$

For the observed data $\mathbf{y}$, the random forest classifier predicts the MAP model index. The predictor is good enough to select the most likely model but not to derive directly the associated posterior probabilities. Indeed, the frequency of trees associated with the majority model is not a proper substitute to the true posterior probability.

However, we have

$$\mathbb{P}\left(\mathcal{M} = \hat{m}(\eta_{\mathbf{y}}) \mid \eta_{\mathbf{y}}\right) = 1 - \mathbb{E}\left(\mathbb{1}\{\mathcal{M} \neq \hat{m}(\eta_{\mathbf{y}})\} \mid \eta_{\mathbf{y}}\right).$$

Therefore, as explained in Pudlo et al. (2016), this justifies using a second random forest in regression to estimate the posterior probability of the selected model (Algorithm 6.4). Thus, in addition to the majority vote, this posterior probability provides a confidence measure of the previous prediction at the point of interest $\eta_{\mathbf{y}}$. This probability is approximated thanks to a regression random forest for which the explanatory variables are the summary statistics of the reference table, and the response is the vector of the indicator values $\mathbb{1}\{m^{(i)} \neq \hat{m}^{\text{oob}}(\eta_{\mathbf{x}^{(i)}})\}$. These values use the out-of-bag predictions of the training data set, hereafter denoted by the "oob" exponent. Predicting the observed data thanks to this forest allows the derivation of the posterior probability of the selected model (as described in Algorithm 6.4). Note that using the out-of-bag classifiers prevents over-fitting issues and is computationally parsimonious as it avoids the generation of a second reference table for the regression random forest training.

6.2.1.2 Grouped model choice

A very useful development for ABC-RF, we implemented in our R package `abcrf` (version 1.7.1), is the model grouping approach, where pre-defined (disjoint) groups

Algorithm 6.4 : ABC-RF computation of the posterior probability of the selected scenario

Input : the values of $\mathbb{1}\{m^{(i)} \neq \hat{m}^{\text{oob}}(\eta_{\mathbf{x}^{(i)}})\}$ for the trained random forest and corresponding summary statistics of the reference table, using the out-of-bag classifiers

Learning : construct a regression random forest $\hat{\mathbb{E}}(.)$ to infer $\mathbb{E}\left(\mathbb{1}\{\mathcal{M} \neq \hat{m}(\eta_{\mathbf{y}})\} \mid \eta_{\mathbf{y}}\right)$

Output : an estimate of the posterior probability of the selected model $\hat{m}(\eta_{\mathbf{y}})$

$$\hat{\mathbb{P}}\left(\mathcal{M} = \hat{m}(\eta_{\mathbf{y}}) \mid \eta_{\mathbf{y}}\right) = 1 - \hat{\mathbb{E}}\left(\mathbb{1}\left\{\mathcal{M} \neq \hat{m}(\eta_{\mathbf{y}})\right\} \mid \eta_{\mathbf{y}}\right)$$

of models (scenarios) are analysed instead of individual ones thanks to Algorithms 6.3 and 6.4. The model indexes used in input are modified in a preliminary step to match the corresponding groups, which are then used during the learning phase. When appropriate, unused models are discarded from the reference table. Actually, this strategy considers deterministic mixture models. Each model index is drawn from a prior distribution, and a mixture results from the formation of a group. In population genetics, this improvement is particularly useful when a high number of individual scenarios are considered and have been formalised through the absence or presence of some key demographic events (e.g. admixture, bottleneck, ...). Groups of scenarios are therefore formed depending on the inclusion or not of a certain evolutionary event, and the model choice procedure aims at deciphering whether or not this event needs to be considered. Thanks to associated errors, it can be used to assess the strength with which an event is properly identified and which ones are easy/hard to discriminate using the chosen method and data. This approach is applied in Sections 6.3 and 6.4 in two applications.

6.2.2 ABC-RF prior vs posterior errors

In this section, we present how to compute posterior measures of error for parameter inference, i.e. errors computed conditionally on the observed summaries $\eta_{\mathbf{y}}$. We oppose it with prior errors which are computed on training data simulated from the prior distribution (a training data is an ABC reference table element). Hereafter, we also denote this prior error as "global", and the posterior error as "local" as it is measured at exactly one point, the desired observation.

What we present in the following is mostly based on the ABC-RF strategy for parameter inference (Chapter 4). Algorithm 6.5 recalls its principle. For a fixed model, we try to infer on the k-th dimension of θ, denoted θ_k, thanks to a newly generated reference table. The idea is to train a regression random forest per dimension of the parameter space of interest. The output of the algorithm is a vector of weights $\mathbf{w}(\eta_{\mathbf{y}})$ that can be used to compute posterior quantities of interest such as expectation, variance and quantiles. $\mathbf{w}(\eta_{\mathbf{y}})$ provides an empirical posteriori distribution for θ_k.

Algorithm 6.5 : ABC-RF for parameter estimation

Input : a vector of $\theta_k^{(i)}$ values and d summary statistics
Learning : construct a regression random forest to infer parameter values
Output : apply the random forest to the observed data $\eta_{\mathbf{y}}$, to deduce a vector of weights $\mathbf{w}(\eta_{\mathbf{y}}) = \{w_1(\eta_{\mathbf{y}}), \ldots, w_N(\eta_{\mathbf{y}})\}$, which provides an empirical posterior distribution for θ_k.
$\mathbf{w}(\eta_{\mathbf{y}})$ is used to compute the estimators of the mean, the variance and the quantiles of the parameter of interest:

$$\hat{\mathbb{E}}(\theta_k \mid \eta_{\mathbf{y}}), \quad \hat{\mathbb{V}}(\theta_k \mid \eta_{\mathbf{y}}), \quad \hat{\mathbb{Q}}_\alpha(\theta_k \mid \eta_{\mathbf{y}})$$

Global prior errors

In both contexts, model choice or parameter inference, a global quality of the predictor can be computed, which does not take the observed data $\mathbf{y}$ into account. Indeed, a random forest makes possible the computation of errors on the training reference table, using the out-of-bag predictions (see Chapter 2).

For model choice, this type of error is called the prior error rate, which is the misclassification error rate computed over the entire prior space. It is expressed as

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left\{m^{(i)} \neq \hat{m}^{\text{oob}}(\eta_{\mathbf{x}^{(i)}})\right\}.$$

For parameter estimation, the equivalent is the prior mean squared error (MSE) or the normalised mean absolute error (NMAE) for example, the latter being less sensitive to extreme values. These errors are computed as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\theta_k^{(i)} - \hat{\theta}_k^{\text{oob},(i)}\right)^2,$$
$$\text{NMAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\theta_k^{(i)} - \hat{\theta}_k^{\text{oob},(i)}}{\theta_k^{(i)}}\right|.$$

They can be perceived as a Monte Carlo approximation of expectations with respect to the prior distribution. This type of error does not take into account $\mathbf{y}$ and we introduce below the posterior analogues.

Local posterior errors

We propose some posterior versions of error, that target the quality of prediction with respect to the posterior distribution.

For model choice, the posterior probability provided by Algorithm 6.4 is a confidence measure of the selected scenario given the observation. Therefore

$$1 - \hat{\mathbb{P}}\left(\mathcal{M} = \hat{m}(\eta_{\mathbf{y}}) \mid \eta_{\mathbf{y}}\right)$$

directly yields the posterior error associated to $\eta_{\mathbf{y}}$: $\hat{\mathbb{P}}\left(\mathcal{M} \neq \hat{m}(\eta_{\mathbf{y}}) \mid \eta_{\mathbf{y}}\right)$.

For parameter estimation, when trying to infer on θ_k, a point-wise analogous measure of a local error can be computed as the posterior expectations

$$\mathbb{E}\left(\left(\theta_k - \hat{\theta}_k\right)^2 \mid \eta_{\mathbf{y}}\right) \quad \text{and} \quad \mathbb{E}\left(\left|\frac{\theta_k - \hat{\theta}_k}{\theta_k}\right| \mid \eta_{\mathbf{y}}\right). \tag{6.1}$$

We approximate these expectations by

$$\sum_{i=1}^{N} w_i(\eta_{\mathbf{y}}) \left(\theta_k^{(i)} - \hat{\theta}_k^{\text{oob},(i)}\right)^2 \quad \text{and} \quad \sum_{i=1}^{N} w_i(\eta_{\mathbf{y}}) \left|\frac{\theta_k^{(i)} - \hat{\theta}_k^{\text{oob},(i)}}{\theta_k^{(i)}}\right|.$$

We again use the out-of-bag information to compute $\hat{\theta}_k^{(i)}$, hence avoiding the (time consuming) production of a second reference table, and we assume that the weights $\mathbf{w}(\eta_{\mathbf{y}})$ from the regression random forest (Algorithm 6.5) are good enough to approximate any posterior expectations: $\mathbb{E}(g(\theta_k) \mid \eta_{\mathbf{y}})$.

Another more expensive strategy to evaluate the posterior expectations (6.1) is to construct a new regression random forest using for response variable the out-of-bag vector of values

$$\left(\theta_k^{(i)} - \hat{\theta}_k^{\text{oob},(i)}\right)^2 \quad \text{or} \quad \left|\frac{\theta_k^{(i)} - \hat{\theta}_k^{\text{oob},(i)}}{\theta_k^{(i)}}\right|,$$

depending on the targeted error. The observation $\eta_{\mathbf{y}}$ is then given to the forest to provide the approximation of the wanted expectation (6.1).

Note that the $\hat{\theta}_k^{\text{oob},(i)}$ values in the previous formulas can be replaced by either the approximated posterior expectations $\hat{\mathbb{E}}(\theta_k^{(i)} \mid \eta_{\mathbf{y}})$ or the posterior medians $\hat{\mathbb{Q}}_{50\%}(\theta_k^{(i)} \mid \eta_{\mathbf{y}})$, still using the out-of-bag information, to compute the local posterior errors. We found in the present chapter (Section 6.4) that the posterior median provides a better accuracy of parameter estimation than the posterior expectation (a.k.a. posterior mean). This trend also holds for global prior errors that can be computed using either the mean or the median as point estimates.

As a final comment, it is worth noting that so far a common practice consisted in evaluating the quality of prediction (for model choice or parameter estimation) in the neighbourhood of the observed data, that is around $\eta_{\mathbf{y}}$ and not exactly $\eta_{\mathbf{y}}$. For model choice, in Estoup, Raynal et al. (2018) prior to the above posterior error development, we use the so called posterior predictive error rate which is an error of this type. In this case, some simulated data sets of the reference table close to the observation are selected thanks to an Euclidean distance, then new data (also denoted pseudo-observed data in the applications below) are simulated using similar parameters, on which is computed the error (see also Lippens et al., 2017, for a similar approach in a standard ABC framework). However, the main problem of processing this way is the difficulty to specify the size of the area around the observation, especially when the number of summary statistics is large. Even though we still report it in Section 6.3, we therefore do not recommend the use of such a "neighborhood" error anymore but rather to compute the local posterior errors detailed above as they measure prediction quality exactly at the position of interest $\eta_{\mathbf{y}}$.

We finish this chapter with the presentation of the two population genetics applications.

6.3 ABC-RF grouped model choice: genetic history of Pygmy human populations

6.3.1 Problem

In this study, we present a set of statistical analyses using ABC-RF applied on a molecular (DNA) data set obtained from Western Central African Pygmy and non-Pygmy populations. Central Africa and the Congo Basin are currently peopled by the largest group of forest hunter-gatherer populations worldwide, which have been historically called "Pygmies" in reference to the mythical population of short stature described by the ancient Greek poet Homer (Hewlett, 2014). Each Central African Pygmy group is in the neighbourhood of several sedentary agricultural populations (hereafter called "non-Pygmies") with whom they share complex sociocultural and economic relationships, including social rules regulating intermarriages between communities (Verdu, Becker et al., 2013; Hewlett, 2014). Due to the lack of ancient human remains in the equatorial forest, the origins of Pygmies and neighbouring non-Pygmies remain largely unknown (Cavalli-Sforza, Menozzi et al., 1994; Cavalli-Sforza and Feldman, 2003). Moreover, Western colonisers from the 19th century somewhat arbitrarily collapsed into a single "Pygmy" group more than 20 populations that were, and still are, culturally and geographically isolated in reality, which further clouded our understanding of evolutionary relationships among these populations. Thus, the questions of (i) whether all Central African Pygmy populations have a common or an independent origin, and (ii) whether they exchange genes through introgression/migration among one another and from neighbouring non-Pygmies, were still largely debated in the anthropology and ethnology communities (Cavalli-Sforza, 1986; Hewlett, 2014; Verdu, Austerlitz et al., 2009).

To tackle these questions, Verdu, Austerlitz et al. (2009) genotype strongly variable genetic markers (namely microsatellite DNA loci; Estoup, Jarne et al., 2002) in a dense sample of non-Pygmy and neighbouring Pygmy populations from Western Central Africa, and use standard ABC methods (Beaumont, Zhang et al., 2002; Estoup, Lombaert et al., 2012) to make statistical inferences. In the present study, we consider the data set of Verdu, Austerlitz et al. (2009) and reanalyse it using ABC-RF. A noticeable novelty of the statistical analyses presented here includes the application of ABC-RF algorithms to make scenario choice on predefined groups of models (i.e. on deterministic mixture models), in addition to standard analyses on the whole set of (separated) scenarios to be compared. As a matter of fact, genetic markers such as microsatellites are informative for deciphering key evolutionary events that shape genetic variation in natural populations, such as a common or an independent origin of a given set of populations, the presence or absence of genetic introgression/migration among populations, as well as major changes in effective population size, the latter feature being strongly suspected in non-Pygmy African populations (e.g. Lombaert et al., 2010; Verdu, Austerlitz et al., 2009). Under an ABC framework, such events can be modelled explicitly hence defining different

scenarios that can be grouped based on the type(s) of evolutionary events that have been incorporated into them. We show here that groups of scenarios (for instance scenarios including a common origin of Pygmy populations versus scenarios including an independent origin of those populations) can be formally and advantageously compared using ABC-RF, in addition of considering all scenarios separately. Such grouping approach in scenario choice is useful to identify main evolutionary events characterising the history of natural populations, and to determine the strength with which each one is discriminated by the considered method and data.

6.3.2 Inferential setting

6.3.2.1 Observed data set

The analysed data set includes the genotyping at 28 microsatellite loci of 400 unrelated individuals from four Pygmy groups (i.e. the Baka, Bezan, Kola and Koya; 29 to 32 individuals per group), neighbouring non-Pygmy individuals (194 individuals) from Cameroon and Gabon (Western Central Africa) (see Figure 1 and Table S1 in Verdu, Austerlitz et al., 2009, for details about geographic location of population samples and their genetic grouping). The exact data set used in the present study is available upon request to Paul Verdu or Arnaud Estoup, following ethical, informed consent and IRB appropriateness.

6.3.2.2 Models, groups of models, and parameters

We consider the same set of eight complex evolutionary scenarios, as in Verdu, Austerlitz et al. (2009). These scenarios with their historical and demographic parameters are represented in Figure 6.1, following the notation of Verdu, Austerlitz et al. (2009). See also Appendix B.1 for a detailed description of the model parameters and their prior distributions.

These eight scenarios include different combinations of three main types of evolutionary events debated in the anthropology community, and groups of scenarios are formed depending on on their presence or absence.

First evolutionary event: The scenario group G1A (scenarios 1, 2, 3 and 4; labelled noIND) corresponds to a common origin of Pygmy populations that diversified from a single ancestral Pygmy population at time t_p. The ancestral Pygmy population itself diverged at time t_{pnp} from the non-Pygmy African population. The group G1B (scenarios 5, 6, 7 and 8; labelled IND) describes an independent origin of Pygmy groups that independently diverged from the non-Pygmy African population at times t_{pnpi}. For this group, divergence times are drawn independently for each Pygmy lineage and thus, the order in which these lineages split is not predefined.

Second evolutionary event: the group G2A (scenarios 1, 3, 5 and 7; labelled MIG) includes both a recent and an ancient event of introgression/migration (cf. parameters tr_i and r_i) from the non-Pygmy African population into each Pygmy lineage independently. It was already suggested by previous anthropological and

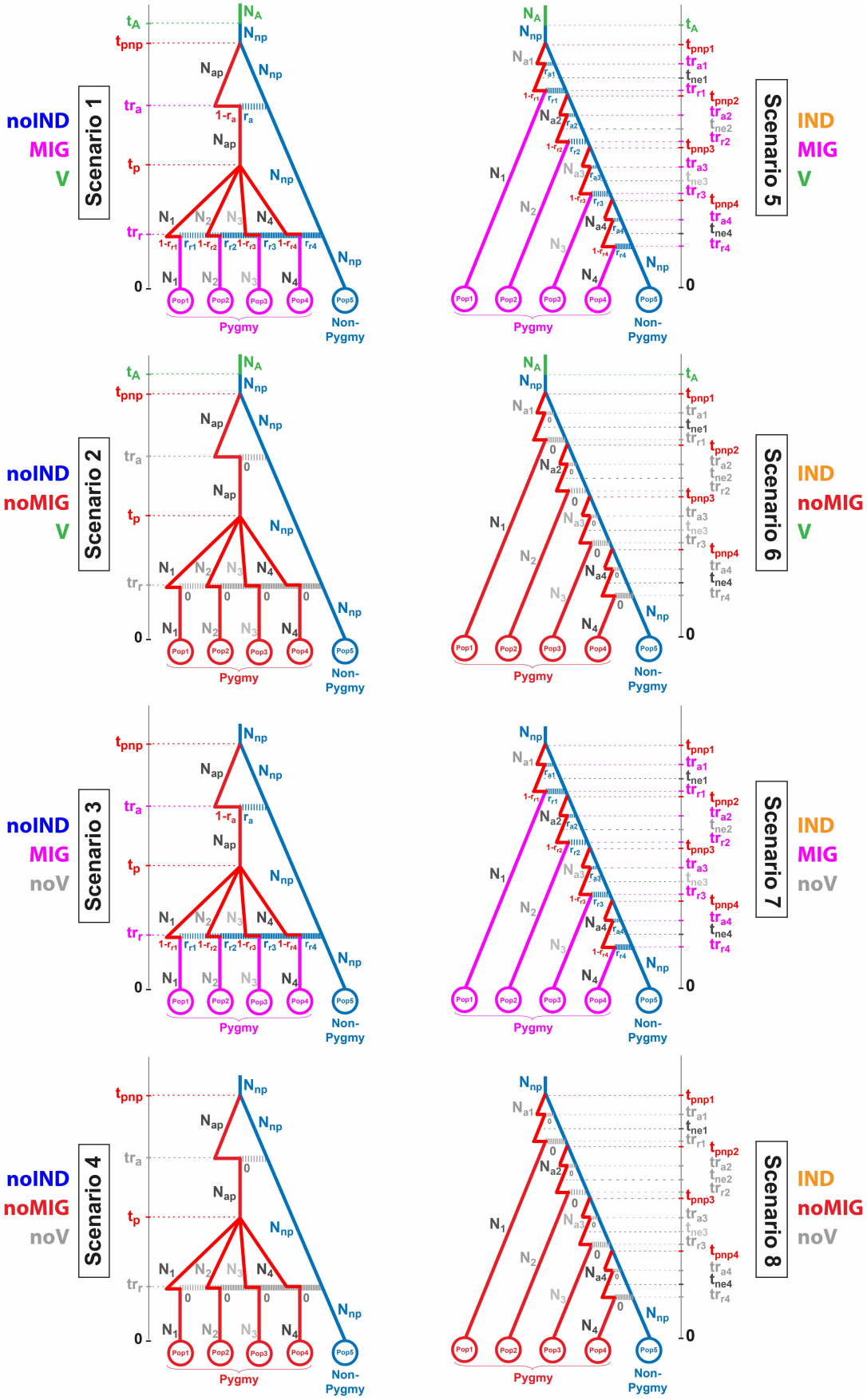


Figure 6.1 – Eight complex competing scenarios of origin and diversification of Pygmy populations from Western Central Africa.

genetic studies (e.g. Cavalli-Sforza, 1986; Hewlett, 1996; Destro-Bisol et al., 2004). The group G2B (scenarios 2, 4, 6 and 8; labelled noMIG) excludes this event by setting all introgression rates (r_i) to zero.

Third evolutionary event: the scenario group G3A (scenarios 1, 2, 5 and 6; labelled V) includes a potential stepwise change of effective population size that occurred in the non-Pygmy African population at time t_A, while not considered in G3B (scenarios 3, 4, 7 and 8; labelled noV).

For all scenarios, N_i corresponds to the effective population size of population i. It is worth stressing that there is obviously a number of other possible scenarios that might possibly fit the data, just as well as if not better than the best scenario found among the present finite set of compared scenarios.

6.3.2.3 Priors

We choose an equiprobable prior for the compared scenarios. Similarly to Verdu, Austerlitz et al. (2009) and references therein, we use flat prior distributions for all demographic parameters specified in Figure 6.1 (see Appendix B.1 for details): uniform distributions bounded between 100 and 10,000 diploid individuals for all Pygmy populations and ancestral population sizes (N_i, N_{ap}, N_{ai}, and N_A, with i between 1 and 4), between 1,000 and 100,000 for the African non-Pygmy population (N_{np}). Prior values are drawn from uniform distributions between 1 and 5,000 generations for all divergence times (t_p, t_{pnp}, t_{pnpi}, with i between 1 and 4), for the population size variation times (t_{nei}, with i between 1 and 4) and for the times of "ancient" introgression of non-Pygmy genes into ancestral Pygmy lineages (tr_a, and tr_{ai}, with i between 1 and 4). For the time of change in effective population size in the non-Pygmy population (t_A), considered only in the scenario group G3A (i.e. scenarios 1, 2, 5 and 6), we sample our prior values thanks to a uniform distribution bounded between 1 and 10,000 generations. For the "recent" introgression times from non-Pygmies into the Pygmy lineages, (tr_r, and tr_{ri}, with i between 1 and 4), we use log-uniform prior distributions bounded between 1 and 5,000 generations. For genetic markers (i.e. microsatellite loci), the parameters and associated prior distributions of mutation models and rates are the same as in Verdu, Austerlitz et al. (2009) and Estoup, Verdu et al. (2018).

6.3.2.4 ABC-RF: analyses conducted on the observed data set

Following Pudlo et al. (2016), ABC-RF treatments are processed on a reference table including 100,000 simulated data sets (i.e. 12,500 per scenario). Data sets are summarised using the whole set of summary statistics proposed by DIYABC (Cornuet, Pudlo et al., 2014) for microsatellite markers, describing genetic variation per population (e.g. number of alleles), per pair (e.g. genetic distance), or per triplet (e.g. coefficient of admixture) of populations, averaged over the 26 loci (see Chapter 3, Table 3.1 for details about such statistics), plus the linear discriminant analysis (LDA) axes as additional summary statistics. The total number of summary statistics is 130 plus a single discriminant (LDA) axis when analysing pairwise

groups of scenarios or seven LDA axes when analysing the eight models considered separately, as additional summary statistics. We checked that the number of simulated data sets of the reference table was sufficient by evaluating the stability of prior error rates (i.e. misclassification error computed when drawing model index and parameter values into priors) and posterior probabilities estimations on $80,000$, $90,000$ and $100,000$ simulated data sets (results not shown). The number of trees in the constructed random forests is fixed to $B = 1,000$; see Appendix B.2 for a justification of choosing such a number of trees per forest.

For each ABC-RF analysis, we predict the best group of scenarios or individual scenario (based on the number of votes), estimate its posterior probabilities, but also the prior error rate as well as a proximal measure of the posterior predictive error rate. Both types of error are computed from $10,000$ simulated pseudo-observed data sets (pods), for which the true scenario identity (ID) is known. The proximal measure of the posterior predictive error rate is determined conditionally on the observed data set by selecting the ID model and the evolutionary parameter values within the 100 best simulations (i.e. those closest to the observed data set as deduced by computing standardised Euclidean distances between the vectors of observed and simulated summary statistics) among a total of $800,000$ simulated data sets generated from priors. Using the parameters associated to the closest data to the observation, we generate $10,000$ simulated data on which the misclassification error rate is computed. It is worth stressing, that when pods are drawn randomly into prior distributions for both the scenario ID and the parameter values, one estimates global error levels computed over the whole (and usually huge) data space defined by the prior distributions. The levels of error may be substantially different depending on the location of an observed or pseudo-observed data set in the prior data space. Indeed, some peculiar combination of parameter values may correspond to situations of strong (weak) discrimination among the compared scenarios. Aside from their use to select the best classifier and set of summary statistics, prior-based indicators are hence relatively poorly relevant since, for a given data set, the only point of importance in the data space is the observed data set itself. Computing error indicators conditionally on the observed data set (i.e., focusing around the observed data set by using a posterior distribution) is hence clearly more relevant than blindly computing indicators over the whole prior data space.

6.3.2.5 Computer programs and computer times

For the simulation of data following the above model-prior design, we use the software DIYABC v.2.1.0 (Cornuet, Pudlo et al., 2014). Regarding ABC-RF treatments which follow the generation of the reference table using DIYABC, computations are performed with the R package `abcrf` (version 1.7.1) available on CRAN.

In the present study, all analyses were processed on a 16 cores Intel Xeon E5-2650 computer (Linux Debian platform, 64 bits system, with a maximum of 20 Gb of RAM used for the heaviest treatments). The production of a reference table including $100,000$ simulated data sets (and summary statistics) took 40 minutes with 30% of the running time devoted to the computation of the 130 summary statistics for each simulated data set. ABC-RF treatments, following the generation of the reference table and based on the R package `abcrf`, took four and eight minutes

for scenarios grouping and individual scenarios configurations, respectively.

6.3.3 Results

We first conduct ABC-RF treatments to make model choice on predefined groups of scenarios (group G1A vs group G1B; group G2A vs. group G2B; and group G3A vs. group G3B). We then carry out ABC-RF treatments on the eight scenarios considered separately.

The projection of the microsatellite population data sets from the reference table on a single (when analysing pairwise groups of scenarios) or on the first two LDA axes (when analysing the eight scenarios considered separately) provides a first visual indication about our capacity to discriminate among the compared scenarios (Figure 6.2). Simulations under the different pairwise groups of scenarios weakly overlap indicating a strong power to discriminate among the pairwise groups interest. When considering the whole set of eight scenarios individually, the projected points substantially overlap for at least some of the scenarios suggesting an overall lower power to discriminate among scenarios considered separately than when considering pairwise groups of scenarios. As a first inferential clue, one can note that the location of the observed data set (indicated by vertical line or a star symbol in Figure 6.2) suggests, albeit without any formal quantification, a marked association with the scenario groups G1A, G2A and G3A, and, to a lower extent with the scenario 1.

A quantitative measure of the power to discriminate among groups of scenarios (scenarios) is obtained by estimating the probability to choose a wrong group of scenarios (scenario) when drawing index and parameter values of group of scenarios (scenario) into priors (i.e. prior error rates). Table 6.1 indicates substantially lower prior error rates when discriminating among groups of scenarios (i.e. 8.85% for G1A vs. G1B, 2.65% for G2A vs. G2B, and 10.54% for G3A vs. G3B) than among scenarios considered individually (prior error rate equal to 20.67%). This is very interesting although not surprising because the grouping strategy simplifies the classification problem by reducing the number of models to discriminate to only two. Because for a given data set, the only point of importance in the data space is the observed data set, we conduct a second quantitative estimation of error rates corresponding to a proximal measure of the posterior predictive error rate computed conditionally on the observed data set (Table 6.1). We observe that posterior predictive error rates are substantially lower than prior error rates (i.e. posterior predictive error rates equal to 4.99 % for G1A vs. G1B, 0.91 % for G2A vs. G2B, 0.20 % for G3A vs. G3B, and 6.17 % for the scenarios considered separately), indicating that the observed data set belongs to a region of the data space where the power to discriminate among groups of scenarios (individual scenarios) is higher than the global power computed over the whole prior data space.

Figure 6.3 shows that RF analysis is able to automatically determine the (most) relevant statistics for model comparison. A typical feature of ABC-RF analysis is that LDA axes always correspond to the most informative statistics, which makes sense knowing their intrinsic construction structure. Interestingly, many of the most informative population genetics summary statistics are not selected by the experts in Verdu, Austerlitz et al. (2009), especially some crude estimates of admixture

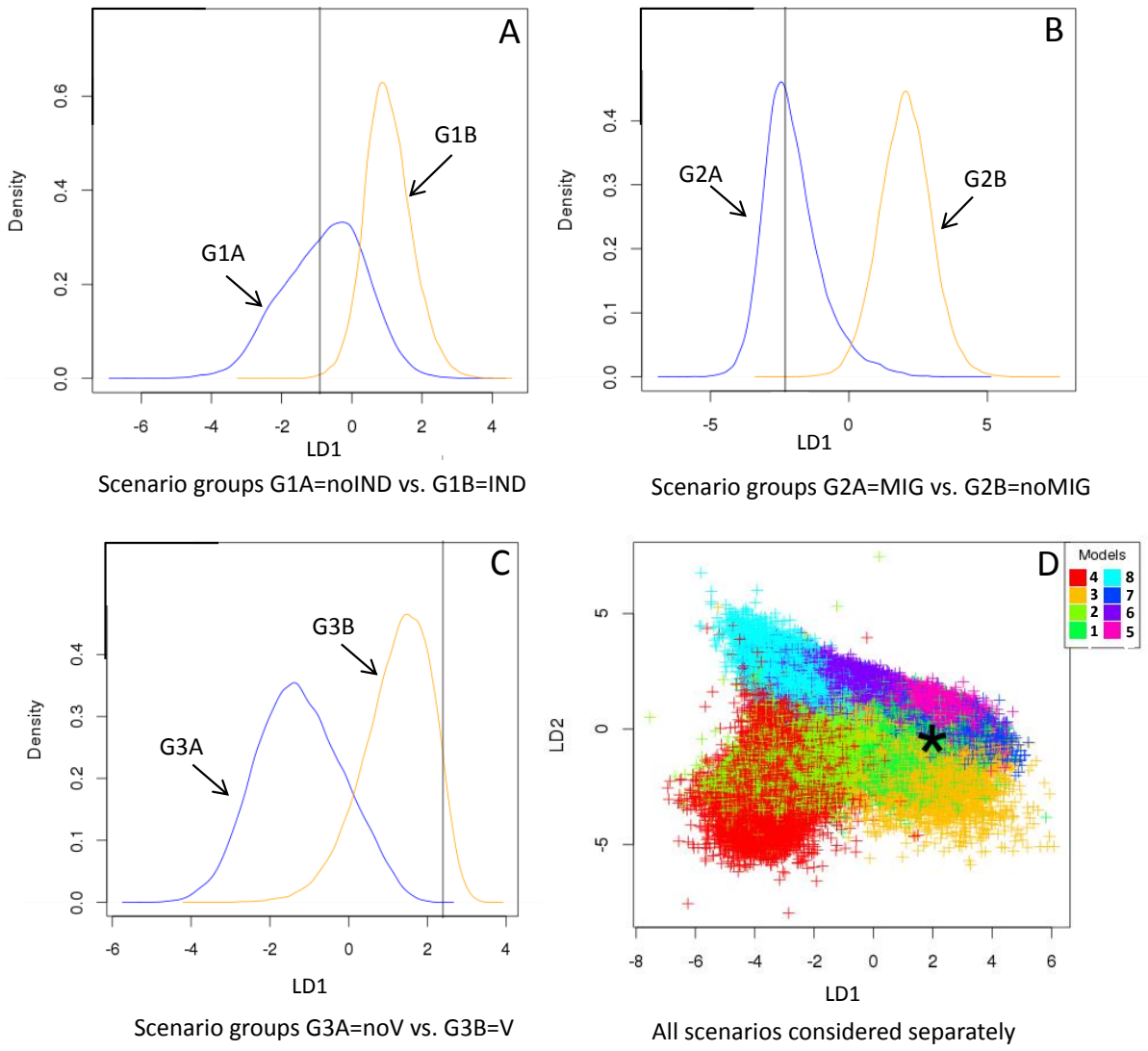


Figure 6.2 – Projection of the reference table and the observed data on a single (when analysing pairwise groups of scenarios) or on the first two LDA axes (when analysing the eight scenarios considered separately). The location of the observed data is indicated by a vertical line in panels A, B and C, and a large black star in panel D. Curves in A-C are estimated kernel densities.

	Groups of scenarios			Individual scenarios
	G1A vs. G1B	G2A vs. G2B	G3A vs. G3B	
Prior error rate	8.85%	2.65%	10.54%	20.67%
Posterior predictive error rate	4.99%	0.91%	0.20%	6.17%
Posterior probability of the selected group of scenarios or scenario	0.923 (G1A)	0.987 (G2A)	0.955 (G3A)	0.851 (Scenario 1)

Table 6.1 – Error rates on scenario group (individual scenarios) choice and posterior probabilities of the selected scenario group (scenario) when discriminating among evolutionary scenario groups (scenarios) of Pygmy human populations using Algorithms 6.3 and 6.4.

rates based on population triplets (i.e. AML statistics; see Chapter 3, Table 3.1). A possible explanation is that experts in population genetics are biased towards choosing summary statistics that are informative for parameter estimation under a given model. However, according to our own experience on this issue, the most informative statistics for model choice are often different than those that are informative for parameter estimation (Raynal et al., 2019; Robert, Cornuet et al., 2011). It is worth stressing that the most informative statistics differ depending on the model choice design. AML, FST and LIK statistics are among the most informative when discriminating among groups of scenarios dealing with independence/dependence of divergence events (G1A-B and individual scenarios) and introgresion/migration events (G2A-B and individual scenarios), whereas intra-population statistics such as VAR, V2P and MWG; see Chapter 3, Table 3.1) are the most informative ones when discriminating among groups of scenarios dealing with population size variation events (G3A-B). These differences are easy to interpret intuitively as divergence and introgression/migration events strongly impact the branching pattern of the tree topology summarising the relationships among populations, which are informed by two and three sample statistics measuring the amount of genetic variation shared between populations (e.g. AML and FST), whereas population size variation events mainly impact the level of genetic variation within populations, which corresponds to the type of variation targeted by single sample statistics (e.g. VAR, V2P and MWG). Thus, the model grouping strategy also allows to determine the relevant summary statistics to identify the presence or absence of a specific event.

The outcome of the first step of the ABC-RF statistical treatment applied to a given target data set is a classification vote for each scenario groups (or individual scenarios) which represents the number of times a given scenario group (or single scenario) is selected in a forest of B trees. The group of scenarios (or single scenario) with the highest classification vote corresponds to the (set of) model(s) best suited to the target data set among the set of compared groups of scenarios or individual scenarios. In our case study, the classification vote estimated for the observed human microsatellite data set is by far the highest for the scenario group G1A (i.e. ensemble of scenarios in which the four Pygmy populations originate non-independently from a common ancestral Pygmy population; cf. 940 of the $B = 1,000$ RF-trees selected the scenario group G1A), the scenario group G2A (i.e. scenarios including asymmetrical introgression/migration events between each Pygmy population and the non-Pygmy

one; cf. 984 of the 1,000 trees), and the scenario group G3A (i.e. scenarios including a change of population size in the non-Pygmy African population; cf. 959 of the 1,000 trees). When considering the eight scenarios separately, the classification vote estimated for the observed human microsatellite data set is the highest for the scenario 1 which congruently includes all three above-selected evolutionary events (877 of the 1,000 trees).

It is worth stressing that there is no direct connection between the frequencies of the allocation of the data of the groups of scenarios (or individual scenarios) among the tree classifiers (i.e. the classification vote) and the posterior probabilities of the competing groups of scenarios (individual scenarios) (see Figure S2 in Pudlo et al., 2016). We therefore conduct the second RF analytical step corresponding to the algorithm 3 in Pudlo et al. (2016) (recalled in Algorithm 6.4) to obtain a reliable estimation of posterior probability of the best group of scenarios (or individual scenario) (Table 6.1). The high posterior probability value provides a strong confidence in selecting the scenario groups G1A, G2A and G3A (probability equal to 0.923, 0.987 and 0.955, respectively). When considering all scenarios separately, the selected scenario 1 is associated to a moderately high posterior value (at least lower compared to those for groups of scenarios) of 0.851. We notice that the introgression/migration event is selected with the highest posterior probability.

As for any Bayesian inference, the shape of the priors used for data set simulations may affect both the posterior probabilities of scenarios and the posterior parameter estimation under ABC inference (e.g. Sunnåker et al., 2013). To empirically evaluate the influence of prior shape on our inferences, we conducted all ABC-RF analyses assuming a set of alternative non-flat priors for the simulations corresponding to the prior set 2 in Verdu, Austerlitz et al. (2009). We found that such alternative statistical treatment did not change model choice results and that error rates and posterior probabilities were only moderately affected (results not shown).

6.3.4 Practical recommendations regarding the implementation of the random forests algorithms

We develop here several points, formalised as questions, which should help users seeking to apply our methodology on their data set for statistical model choice.

Are my models and/or associated priors compatible with the observed data set?

This question is of prime interest and applies to any type of ABC treatment, including both standard ABC treatments and treatments based on ABC-RF. This issue is particularly crucial knowing that with complex models and high dimensional data sets (i.e. big an hence very informative data sets), as more and more encountered in population genomics, "all models are wrong...". Basically, if none of the proposed model - prior combinations produces some simulated data sets in a reasonable vicinity of the observed data set, this is a signal of incompatibility, and we consider that it is then useless to attempt model choice inference. In such situations, we strongly advise reformulating the compared models and/or the associated prior distributions

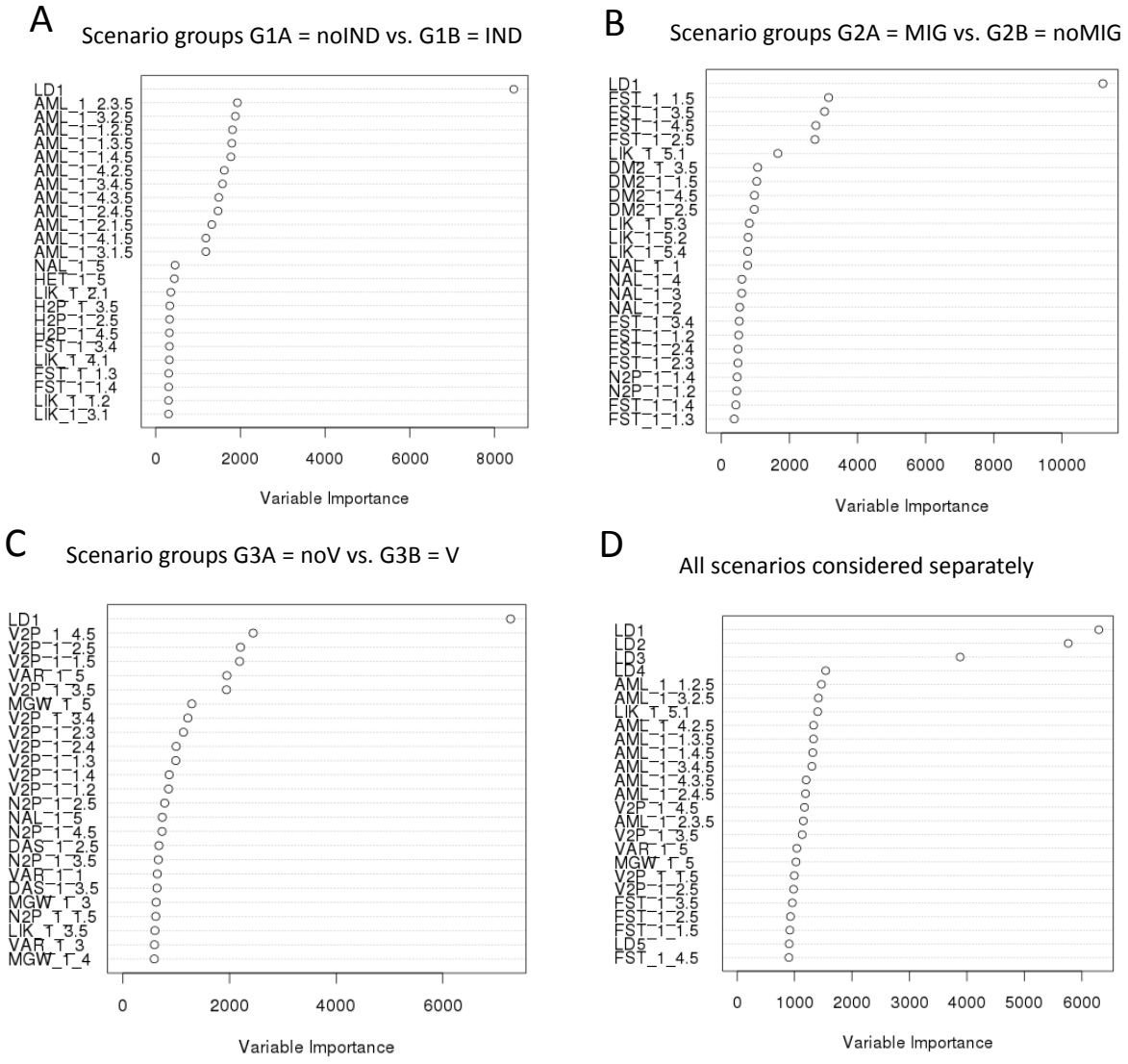


Figure 6.3 – Contributions of the 25 most informative statistics evaluated thanks to the RF mean decreased impurity.

in order to achieve some compatibility in the above sense. We propose here a visual way to address this issue, namely through the simultaneous projection of the simulated reference table data sets and of the observed data set on the first LDA axes. Such a graphical assessment can be achieved using our R package `abcrf` version 1.7.1. In the LDA projection, the observed data set has to be located reasonably within the clouds of simulated data sets (see Figure 6.2 as an illustration). Note that visual representations of a similar type (although based on PCA) as well as computation for each summary statistics and for each model of the probabilities of the observed values in the prior distributions have been proposed by Cornuet, Ravigné et al. (2010) and are already automatically provided by the DIYABC software.

Frazier et al. (2018) very recently analysed the behaviour of approximate Bayesian computation (ABC) when the model generating the simulated data differs from the actual data generating process; i.e., when the data simulator in ABC is misspecified. They demonstrate that when the model is misspecified different versions of ABC can lead to substantially different results and they suggest approaches to diagnose model misspecification in ABC.

Did I simulate enough data sets for my reference table?

A rule of thumb is to simulate between $5,000$ and $20,000$ data sets per model among those compared. In the present example we simulated $12,500$ data sets for each of the eight compared scenarios (for a total of $100,000$ data sets in the reference table). To evaluate whether or not this number is sufficient for random forests analysis, we recommend to compute global prior error rates from both the entire reference table and a subset of the reference table (for instance from a subset of $80,000$ simulated data sets if the reference table includes a total of $100,000$ simulated data sets). If the prior error rate value obtained from the subset of the reference table is similar, or only slightly higher, than the value obtained from the entire reference table, one can consider that the reference table contains enough simulated data sets. If a substantial difference is observed between both values, then we recommend an increase in the number of data sets in the reference table.

Did my forest grow enough trees?

According to our experience, a forest made of 500 trees often constitutes an interesting trade-off between computation efficiency and statistical precision (Breiman, 2001). To evaluate whether or not this number is sufficient, we recommend plotting the estimated values of the prior error rate and/or the posterior probability of the best model as a function of the number of trees in the forest. The shapes of the curves provide a visual diagnostic of whether such key quantities stabilise when the number of trees tends to a given value ($1,000$ trees in the present study).We provide illustrations of such procedure and visual representations in the case of inferences about Human Pygmy population history (see Appendix B.2 in which graphical representation have been produced by our R package `abcrf` version 1.7.1).

6.3.5 Conclusion

Choosing among a group of models (individual scenarios) is a crucial inferential issue as it allows the identification of major historical and evolutionary events formalised into a set of compared scenarios formalised as a combination of such evolutionary events. We illustrate this issue through ABC-RF analyses to make inferences about the genetic history of Pygmy human populations. The eight formalised complex scenarios incorporate (or not) three main evolutionary events: (i) whether there is an independent or non-independent origin of Pygmy groups, (ii) the possibility of introgression/migration events between Pygmy and non-Pygmy African populations, and (iii) the possibility of a change in effective size in the past in the non-Pygmy African population. We found that our scenario grouping approach allows to identify with confidence (i.e. low error rates and high posterior probabilities) the different events in the scenario, and it emphasises that the event of introgression/migration is discriminated with the most accuracy. The final selected scenario (when comparing all eight scenarios separately) corresponds to a common origin of all Western Central African Pygmy groups considered, with the ancestral Pygmy populations having diverged from the non-Pygmy African population in a more remote past. Furthermore, it encompasses both recent and ancient asymmetrical introgression events from the non-Pygmy African gene-pool into each Pygmy population considered, and a change of population size in the non-Pygmy African population. Our ABC-RF analyses confirm and strengthen the initial historical interpretation of Verdu, Austerlitz et al. (2009). We inferred a probable common origin of all Western Central African populations categorised as Pygmies by Western explorers, despite the vast cultural, morphological, and genetic diversity observed today among these populations (Hewlett, 2014). We also confirmed recent asymmetrical and heterogeneous genetic introgressions from non-Pygmies into each Pygmy population. Altogether, these results are in agreement with the ethno-historical scenario proposed by Verdu, Austerlitz et al. (2009) in which the relatively recent expansion of non-Pygmy agriculturalist populations in Western Central Africa which occurred $2,000 - 5,000$ years before present may have modified the pre-existing social relationships in the ancestral Pygmy population, in turn resulting in its fragmentation into isolated groups. Since then, enhanced genetic drift in isolated populations with small effective sizes, and different levels of genetic introgression from non-Pygmies into each Pygmy population led to the rapid genetic diversification of the various Western Central African Pygmy populations observed today.

6.4 Full ABC-RF analysis: reconstructing the evolutionary past of the desert locust

6.4.1 Problem

This section presents some ABC-RF analyses carried on an African insect species: the desert locust, also known as *Schistocerca gregaria* (*S. g.* for short). This desert locust can be found in arid grasslands and deserts in both northern and southern Africa (Figure 6.4A). In its northern range, the desert locust is one of the most wide-

spread and harmful pest species with a huge potential outbreaking area, spanning from West Africa to Southwest Asia. The desert locust is also present in the south-western arid zone (SWA) of Africa, which includes South-Africa, Namibia, Botswana and south-western Angola (Figure 6.4A). The southern populations of the desert locust are termed *S. g. flaviventris* and are geographically separated by nearly $2,500$ km from the northern Africa populations, *S. g. gregaria*.

Interestingly, the desert locust can exist as two phases: solitarious or gregarious. The first one traduces by lone-living, while the second by swarming. The phase can be switched depending on the locust density, and this phenomenon is mentioned as density-dependent phase polyphenism. However, *S. g. flaviventris* appears to lack, at least partly, the capacity to mount some of the phase polyphenism responses associated with swarming (reviewed in Chapuis, Foucart et al., 2017).

Such differential evolution of traits associated with density-dependent phase polyphenism between populations of closely related subspecies offers a hypotheses-driven framework to understand phase polyphenism, and identify candidate genes for this trait.

A promising investigation axis to reveal molecular determinants under phase polyphenism inheritance in the desert locust is to identify key genes (or transcripts) using comparative genomics (or transcriptomics) approaches between highly polyphenetic *S. g. gregaria* populations and less polyphenetic *S. g. flaviventris* populations. In particular, genomics studies based on genome scans (reviewed in Vitti et al., 2013) use population samples to measure genetic diversity and differentiation at many loci, with the goal of detecting loci under divergent selection. Genome scan data can lead to misleading signals of selection if the global effect of the demographic forces (e.g. genetic drift since divergence) is not accounted for.

Hence, for an accurate estimation of the local effect of selection, future genomic studies of the desert locust would require historical knowledge on the considered populations and in particular on the time scale of the processes that led to their phenotypic differences. The independent evolutionary history of *S. g. flaviventris* and *S. g. gregaria* subspecies was recently confirmed, by distinctive mitochondrial DNA haplotypes and male genitalia morphologies (Chapuis, Bazalet et al., 2016). Yet, the historical events, and their timing, related to the divergence of the two desert locust lineages remain unknown.

The main objective of the present study is to unravel the historical and evolutionary processes that have shaped the present disjoint geographical distribution of the desert locust and their genetic variation. To this aim, we first employ paleo-vegetation maps to construct evolutionary scenarios relevant to the desert locust. We then use molecular data obtained from microsatellite markers for which we could obtain independent information on allele size constraints and evolutionary rates in the species of interest from direct observation of germline mutations (Chapuis, Plantamp, Streiff et al., 2015). We apply the ABC-RF methodologies for both model choice and parameter inference, on microsatellite data set to compare a set of evolutionary scenarios and estimate the divergence time between *S. g. gregaria* and *S. g. flaviventris* under the most likely of our scenarios. Finally, we interpret the results in the light of the paleo-vegetation information we compiled and various biological features of the desert locust. This study contains model grouping analysis, meas-

ures of performance thanks to prior and posterior errors, and also comparisons when using previous mutational information for specification of prior distributions or not.

6.4.2 Formalisation of evolutionary scenarios

To help formalising the evolutionary scenarios to be compared, we rely on maps of vegetation cover in Africa from the Quaternary Environment Network Atlas (Adams and Faure, 1997), considering more specifically the periods representative of arid maximums (LGM and YD; Figures 6.4E and 6.4F), humid maximums (HCO; Figure 6.4D) and present-day arid conditions (see Figure 6.4C). Desert and xeric shrubland covers fit well with the present-day species range (Figures 6.4A and 6.4B) during remission periods. Tropical and Mediterranean grasslands were added separately to the desert locust predicted range since the species inhabits such environments during outbreak periods only. The coherence between present maps of species distribution (Figure 6.4A) and of open vegetation habitats (Figure 6.4C) suggests that vegetation maps for more ancient periods could be considered as good approximations of the potential range of the desert locust in the past. Maps of vegetation cover for ice ages (Figures 6.4E and 6.4F) show an expansion of open vegetation habitats (i.e. grasslands in the tropics and deserts in both the North and South of Africa) sufficient to make the potential range of the species continuous from the Horn of Africa in North-West to the Cape of Good Hope in the South. Based on the above climatic and paleo-vegetation map reconstructions, we consider a set of alternative bio-geographic hypotheses formulated into different types of evolutionary scenarios.

First, we consider scenarios involving a more or less continuous colonisation of southern Africa by the ancestral population from a northern origin. In such type of scenarios, effective population sizes are allowed to change after the divergence event, without requiring any bottleneck event (i.e. without any abrupt and strong reduction of population size) right after divergence.

Second, we consider the situation where colonisation of Southern Africa occurred through a single (or a few) long-distance migration event(s) of a small fraction of the ancestral population. This situation is formalised through scenarios which differ from the former by the occurrence of a **bottleneck event in the newly founded population**. The bottleneck event occurs into *S. g. flaviventris* right after divergence and is modelled through a limited number of founders during a short period.

Because the last Quaternary cycle includes several arid climatic periods, including the intense punctuation of the Younger Dryas (YD) and the Last Glacial Maximum (LGM), we also consider scenarios that incorporate the possibility of secondary contact with **asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris***. Since previous tests based on simulated data showed a poor power to discriminate between a single versus several admixture events (results not shown), we consider only models including a single admixture event.

Finally, at interglacial humid maximums, the map of vegetation cover shows a severe contraction of deserts. It was nearly completely vegetated with annual grasses and shrubs and supported numerous perennial lakes (see Figure 6.4D; deMenocal et al., 2000). We thus envisage the possibility that climatic-induced contractions

of population sizes have pre-dated the separation of the two subspecies. Hence, whereas so far scenarios involved a constant effective population size in the ancestral population, we formalise alternative scenarios in which we assume that a long **population size contraction event** occurred **into the ancestral population** at a time t_{ca}, with an effective population size Nc_a during a duration dc_a.

The three above-mentioned key evolutionary events (a bottleneck in *S. g. flaviventris*, an asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*, and a population size contraction in the ancestral population) define a total of eight scenarios that we compare using ABC-RF. The eight scenarios with their historical and demographic parameters are graphically depicted in Figure 6.5. All scenarios assume a northern origin for the common ancestor of the two subspecies and a subsequent southern colonisation of Africa. This assumption is supported by recent mitochondrial DNA data presented in Chapuis, Bazalet et al. (2016).

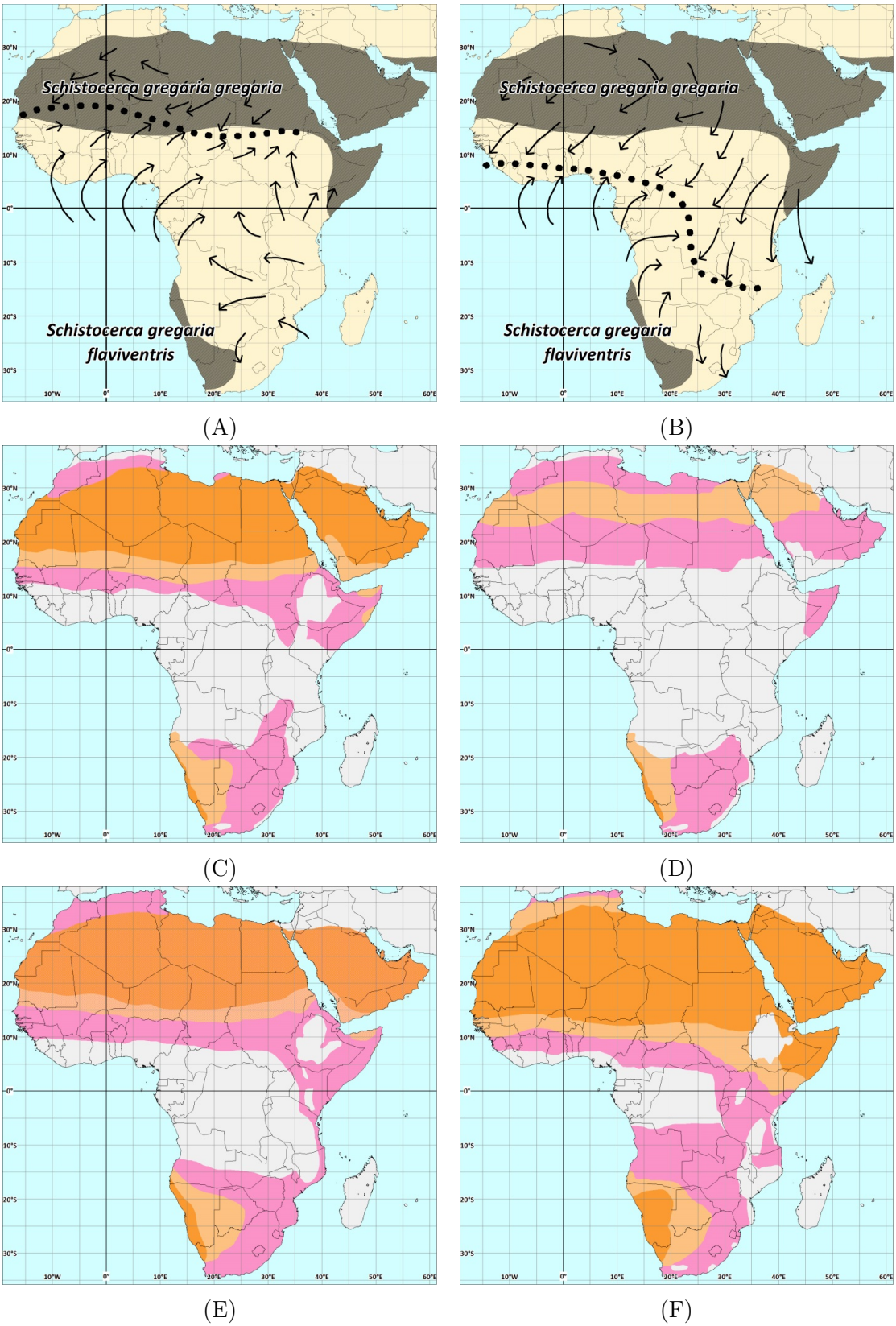
All scenarios consider three populations of current effective population sizes N_f for *S. g. flaviventris*, N_g for *S. g. gregaria*, and N_a for the ancestral population, with *S. g. flaviventris* and *S. g. gregaria* diverging t_{div} generations ago from the ancestral population. The bottleneck event which potentially occurred into *S. g. flaviventris* is modelled through a limited number of founders Nb_f during a short period db_f. The potential population size contraction event occurs into the ancestral population at a time t_{ca}, with an effective population size Nc_a during a duration dc_a. The potential asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris* occurs at a time t_{sc}, with an effective population size Nc_a and a proportion r_g of genes of *S. g. gregaria* origin.

Note for Figure 6.4: (A-B) The distribution range are adapted from Sword et al. (2010). Winds (black arrows) are adapted from Nicholson (1996) with dotted lines representing the intertropical convergence zone (ITCZ). In northern Africa, at least since 2.7 Ky, the strong northeast trade winds bring desert locust swarms equatorward in the moist ITCZ (Kröpelin et al., 2008). Most transports are westward, with records of windborne locusts in the Atlantic Ocean during plague events, including the exceptional trans-Atlantic crossing from West Africa to the Caribbean in 1988 (Lorenz, 2009). Nevertheless, at least in northern winter (January), easterly winds flow more parallel to the eastern coast of Africa. (C-F) Vegetation habitats are adapted from Adams and Faure (1997). Open vegetation habitats suitable for the desert locust correspond to deserts (light orange), xeric shrublands (dark orange) and tropical - Mediterranean grasslands (pink). Other unsuitable habitat classes (white) are forests, woodlands and temperate shrublands and savannas.

6.4.3 Inferential setting

6.4.3.1 Observed data set

We carry out our statistical inferences on the microsatellite data sets previously published in Chapuis, Bazalet et al. (2016). The 23 microsatellite loci genotyped in such data sets are derived from either genomic DNA (14 loci) or messenger RNA (9

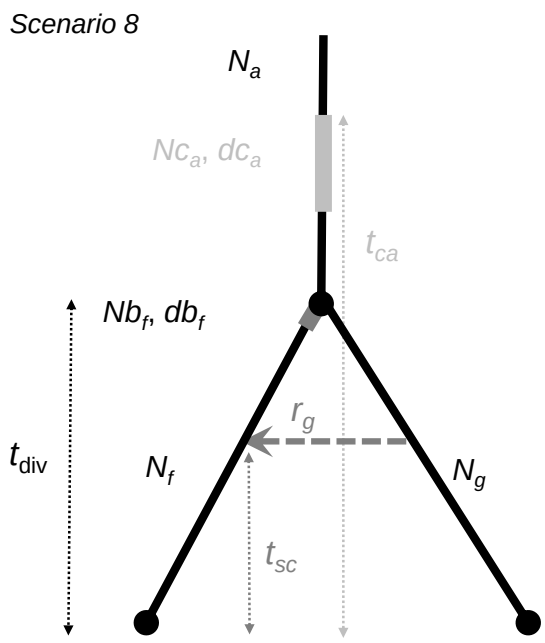


(A) (B)

(C) (D)

(E) (F)

Figure 6.4 – Present time distribution range of *Schistocerca gregaria* in Africa under remission periods with winds in August (A) and January (B), and vegetation habitats suitable for the species during the present period (C), the Holocene Climatic Optimum (HCO, 9 to 6 Ky ago) (D), the Younger Dryas (YD, 12.9 to 11.7 Ky ago) (E) and the Last Glacial Maximum (LGM, 26 to 14.8 Ky ago) (F).



	Evolutionary event		
Scenario	c_a	b	sc
1	no	no	no
2	no	yes	no
3	yes	no	no
4	yes	yes	no
5	no	no	yes
6	no	yes	yes
7	yes	no	yes
8	yes	yes	yes

Figure 6.5 – The eight formulated evolutionary scenarios. The subscripts g, f and a indicate the subspecies *S. g. gregaria*, *S. g. flaviventris* and their unsampled common ancestor, respectively.

loci) resources, and are hereafter referred to as untranscribed and transcribed microsatellite markers. These microsatellites are shown to be genetically independent, free of null alleles (i.e. alleles are functional) and at selective neutrality (Chapuis, Bazalet et al., 2016). This observed data is obtained from 170 genotypes individuals (80 and 90 individuals for *S. g. gregaria* and *S. g. flaviventris*, respectively).

6.4.3.2 Priors on demographic parameters

Prior values for time periods between sampling and secondary contact, divergence and/or ancestral population size contraction events (t_{ca}, t_{div} and t_{sc}, respectively) are drawn from log-uniform distributions bounded between 100 and 500, 000 generations, with $t_{ca} > t_{div} > t_{sc}$. Assuming an average of three generations per year (Roffey and Magor, 2003), this prior setting corresponds to a time period that goes back to the second-to-latest glacial maximum (150 Ky ago).

We use uniform prior distributions bounded between 10^4 and 10^6 diploid individuals for effective population sizes N (Chapuis, Plantamp, Blondin et al., 2014). The admixture rate (r_g; i.e. the proportion of *S. g. gregaria* genes entering into the *S. g. flaviventris* population), is sampled from a uniform prior distribution bounded between 0.05 and 0.5. We use uniform prior distributions bounded between 2 and 100 for both the numbers of founders (in diploid individuals) and duration of bottleneck events (in number of generations). For the contraction event, we consider uniform prior distributions bounded between 100 and 10, 000 for both population

size (in diploid individuals) and duration (in number of generations). Assuming an average of three generations per year (Roffey and Magor, 2003), such prior choice allows a reduction in population size for a short to a relatively long period, similar for instance to the whole duration of the HCO (from 9 to 5.5 Ky ago) which was characterised by a severe contraction of deserts.

6.4.3.3 Priors on mutational parameters

Mutations occurring in the repeat region of each microsatellite locus are assumed to follow a symmetric generalised stepwise mutation model (GSM, Estoup, Jarne et al., 2002). Prior values for any mutation model settings are drawn independently for untranscribed and transcribed microsatellites in specific distributions. The informed mutational setting is defined as follows. Because allele size constraints exist at microsatellite markers, we inform for each microsatellite locus their lower and upper allele size bounds using values estimated in Chapuis, Plantamp, Streiff et al. (2015). Prior values for the mean mutation rates (μ_{R}) are set to the empirical estimates inferred from observation of germline mutations in Chapuis, Plantamp, Streiff et al. (2015), i.e. 2.8×10^{-4} and 9.1×10^{-5} for untranscribed and transcribed microsatellites, respectively. Because each microsatellite locus can have its own mutation rate, we sample each one from a gamma distribution with mean parameter μ_{R} and a shape parameter equal to 0.7 for both types of microsatellites. For each locus, the number of added or deleted microsatellite motifs induced by a mutation is drawn from a geometric distribution with parameter P, where P follows a gamma distribution with mean parameter $\grave{P}$ and a shape parameter equal to 2. This $\grave{P}$ is set to the proportions of multistep germline mutations observed in Chapuis, Plantamp, Streiff et al. (2015), i.e. 0.14 and 0.67 for untranscribed and transcribed microsatellites, respectively. We also consider mutations that insert or delete a single nucleotide to the microsatellite sequence. To model this mutational feature, we use the DIYABC default setting values, i.e. a uniform distribution bounded between $[10^{-8}, 10^{-5}]$ for the mean parameter μ_{SNI}, and a gamma distribution (mean equal to μ_{SNI} and shape equal to 2) for individual loci parameters (see Cornuet, Ravigné et al., 2010).

We evaluate how the incorporation of independent information on prior distributions for mutational parameters affect both the posterior probabilities of scenarios and the posterior parameter estimation under our inferential framework. To this aim, we re-process our inferences using a naive mutational setting as prior, often used in many ABC microsatellite studies (e.g. Estoup, Jarne et al., 2002). In this case, prior values for mean mutation parameters are drawn from uniform distributions instead of being set to a fixed value as in the informed mutational setting. For each set of untranscribed or transcribed microsatellites, all loci are free of allele size constraints (i.e. allele size bounds are fixed to very different values such as 2 and 500 for the lower and upper bounds, respectively). Prior values for μ_{R} are drawn from a uniform distribution bounded between 10^{-5} and 10^{-3}. $\grave{P}$ values are sampled in a uniform distribution bounded between 0.1 and 0.3. Finally, the mean rate of single nucleotide indel mutations and all parameters for individual loci are set to the DIYABC default values (Chapuis, Plantamp, Blondin et al., 2014; Chapuis, Plantamp, Streiff et al., 2015).

6.4.3.4 ABC-RF model choice setting

We use the software DIYABC v.2.1.0 (Cornuet, Pudlo et al., 2014) to simulate data sets constituting the reference tables. ABC-RF statistical analysis are performed using a new version of the R library `abcrf` (version 1.8) available on CRAN. This version includes all ABC-RF strategies, as well as new posterior error measurements for parameter inference detailed in Section 6.2. Observed and simulated data are summarised thanks to a set of 32 summary statistics available from one and two population samples (Chapter 3, Table 3.1) and the one LDA axis or the seven LDA axes (i.e. number of scenarios minus 1, Pudlo et al., 2016) computed when considering pairwise groups of scenarios or individual scenarios, respectively. We process ABC-RF treatments on reference tables including $100,000$ simulated data sets (i.e. $12,500$ per scenario). The number of trees in the constructed random forests is fixed to $B = 3,000$.

We predict the best scenario/group of scenarios and estimate its posterior probability over ten replicate analyses based on ten different reference tables, that we average. In order to understand and decipher the main evolutionary events that occurred during the evolutionary history of the two desert locust subspecies, we performed an ABC-RF analysis thanks to the model grouping strategy, on three pairwise groups of scenarios (with four scenarios per group):

- groups of scenarios with versus without a bottleneck in *S. g. flaviventris*,
- groups with versus without a population size contraction in the ancestral population,
- groups with versus without a secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris*.

We also conduct ABC-RF treatments on the eight individual scenarios considered separately.

6.4.3.5 ABC-RF parameter inference setting

For parameter estimation, we also construct ten independent replicate ABC-RF treatments based on ten different reference tables to infer the time since divergence between the two subspecies (Raynal et al., 2019, and Chapter 4). For each RF, we simulate a total of $100,000$ data sets for the selected scenario (drawing parameter values into the prior distributions described in Sections 6.4.3.2, 6.4.3.3 and using the same 32 summary statistics). The number of trees in the constructed random forests is fixed to $B = 2,000$. We estimate the parameter posterior median as well as 5% and 95% quantiles of the posterior distributions thanks to ABC-RF. Accuracy of time divergence estimation is measured using out-of-bag predictions and the normalised mean absolute error (NMAE). NMAE values are computed and averaged over the ten different replicate analyses. It is worth noting that we focus all over this work on posterior medians rather than posterior expectations as point estimates. More accurate estimations were indeed obtained (according to out-of-bag predictions) when using median rather than mean values.

6.4.4 Results

6.4.4.1 Model choice

We now report the results regarding the model choice procedure thanks to ABC-RF when analysing groups of scenarios or individuals ones. For the informed mutational setting, Table 6.2 indicates lower prior error rates of ABC-RF treatments when discriminating among groups of scenarios (i.e. mean of prior error rates, over the ten replicate runs, lower than 25%) than among the eight scenarios considered separately (47.9%).

ABC-RF analyses support the same best scenario or group of scenarios for all ten replicate analyses. The classification vote for the observed desert locust microsatellite data set is the highest for the group of scenarios in which (i) *S. g. flaviventris* was subject to a bottleneck event at the time of the split (average of 2,890 votes out of 3,000 RF-trees), (ii) the ancestral population experienced a population size contraction (average of 2,245 votes out of 3,000 RF-trees), and (iii) not any admixture event occurred between populations after the split (average of 2,370 votes out of 3,000 RF-trees). When considering the eight scenarios separately, the highest classification vote concerns the scenario 4, which congruently excludes secondary contact and includes a population size contraction in the ancestral population and a bottleneck event at the time of divergence in the *S. g. flaviventris* subspecies (average of 1,777 votes out of 3,000 RF-trees). The other scenarios that obtained at least 5% of the votes are: a scenario including only a single bottleneck event in *S. g. flaviventris* (scenario 2; average of 537 votes) and a scenario with a bottleneck event in *S. g. flaviventris*, a population size contraction in the ancestral population and a secondary contact with admixture from *S. g. gregaria* into *S. g. flaviventris* (scenario 8; mean of 380 votes). All other scenarios obtained less than 5% of the votes and are hence even more weakly supported. The scenario 4 also obtained the highest number of votes for analyses based on a naive mutational setting (results not shown here but available in the complete paper Chapuis, Raynal et al., 2019).

Posterior error rates (i.e. 1 minus the posterior probabilities) are lower than prior error rates for the group of scenarios based on the presence (or not) of a bottleneck in *S. g. flaviventris* (i.e. 3.5% versus 10.2%) and for the scenarios considered separately (i.e. 41.6% versus 47.9%), while, for other groups of scenarios, the discrimination power is similar at both global and local scales (i.e. from 23.5% to 25.8%, Table 6.2). Altogether, these results indicate that the observed data set belongs to a region of the data space where the power to discriminate among scenarios is higher than the global power computed over the whole prior data space, and that the presence or absence of a bottleneck in *S. g. flaviventris* is the demographic event the best predicted by our ABC-RF treatments. Posterior probability values for the scenario 4 and for the best groups of scenarios are slightly lower when using a naive mutational setting, except for the group without any admixture event (Table 6.2).

6.4.4.2 Parameter inference

Table 6.3 reports point estimates with 90% credible intervals for the divergence time between the two subspecies under the best supported scenario 4, as well as prior and

	Analyses of groups of scenarios			Analysis of scenarios separately
	g1= no c_a vs g2 = c_a	g1 = no b vs g2 = b	g1 = no sc vs g2 = sc	
Informed mutational prior				
Prior error rate	24.9%	10.2%	23.5%	47.9%
Posterior probability	0.746 (c_a)	0.965 (b)	0.742 (no sc)	0.584 (scenario 4)
Naive mutational prior				
Prior error rate	26.7%	11.1%	24.4%	50.2%
Posterior probability	0.704 (c_a)	0.950 (b)	0.775 (no sc)	0.547 (scenario 4)

Table 6.2 – Scenario choice prior error rates and posterior probabilities (averaged over ten replicate analyses). The groups are formed depending on the presence (or not) of a bottleneck (b) in *S. g. flaviventris*, a population size contraction in ancestor (c_a) and a secondary contact with asymmetrical genetic admixture from *S. g. gregaria* into *S. g. flaviventris* (sc).

	Informed mutational prior	Naive mutational prior
prior NMAE	0.359	0.542
posterior NMAE	0.369	0.571
t_{div} (G)	$7,723$	$5,235$
90% CI	$[2,785; 19,708]$	$[1,224; 23,845]$

Table 6.3 – Prior and posterior errors for divergence time estimation, as well as divergence time estimations and 90% credible intervals (averaged over then replicate ABC-RF analyses), under the best supported scenario (scenario 4). Time is measured in terms of generations (G).

posterior NMAE. Our estimations point to a young age of subspecies divergence, with a median divergence time of 2.6 Ky and a 90% credible interval of 0.9 to 6.6 Ky using informed mutational priors. The naive mutational setting leads to a median estimate of 1.7 Ky with a 90% credible interval of 0.4 to 7.9 Ky.

The computed prediction errors highlight that the incorporation of independent information into prior distributions of mutational parameters allows a more accurate estimation of the median divergence time, with about 35% lower prior and posterior normalised mean absolute errors (Table 6.3).

Moreover, for both mutational setting, when comparing prior and posterior NMAE, we observe that the posterior error is slightly higher than the prior version (0.369 compared to 0.359 or 0.571 compared to 0.542). It means the observed data lies in a region of the summary statistics space where the predictive power to infer the divergence time t_{div} is lower compared to the global power achieved when computing the error on all prior values (i.e. prior NMAE).

6.4.5 Interpretations

A young age of subspecific divergence in the desert locust

With a 90% credible interval of the posterior density distribution of the divergence time of 0.9 to 6.6 Ky, our ABC-RF analyses clearly point to a divergence of the two desert locust subspecies occurring during the present Holocene geological epoch (0 to 11.7 Ky ago). The posterior median estimate (2.6 Ky) and interquartile range (1.8 to 3.7 Ky) postdated the middle-late Holocene boundary (4.2 Ky). Two possible reasons where proposed to explain this early divergence time, that we now describe.

Recent geological and palynological research has shown that a brief fragmentation of the African primary forest occurred during the Holocene interglacial period from 2.5 Ky to 2.0 Ky ago (reviewed in Maley et al., 2018). This forest fragmentation event is characterised by relatively warm temperatures and a lengthening of the dry season rather than an arid climate. Although this period does not correspond to a phase of general expansion of savannas and grasslands, with *Poaceae* pollen never exceeding 40% of the total pollen count, it led to the opening of the Sangha River Interval (SRI). The SRI corresponds to a 400 km wide ($14 - 18°$ E) open strip composed of savannas and grassland passing through the rainforest in a North-South direction. The SRI corridor is thought to have facilitated the southern migration of Bantu-speaking pastoralists, along with cultivation of the semi-arid sub-Saharan cereal, pearl millet, *Pennisetum glaucum* (Schwartz, 1992; Bostoen et al., 2015). The Bantu expansion took place between approximately 5 and 1.5 Ky ago and reached the southern range of the desert locust, including northern Namibia for the Western Bantu branch and southern Botswana and eastern South Africa for the Eastern Bantu branch (Vansina, 1995). Therefore, we cannot exclude that the recent subspecific distribution of the desert locust can have been mediated by this recent climatic disturbance associated with a **north-south corridor of open vegetation habitats** and the diffusion of agricultural landscapes through the Bantu expansion. The progressive reappearance of forest vegetation 2 Ky ago would have then led to the present-day isolation and subsequent genetic differentiation of such new southern populations from the more northern parental populations. Our ABC-RF results also indicated that a severe bottleneck (i.e. a strong transitory reduction of effective population size) occurred in the nascent southern subspecies of the desert locust. The very high posterior probability value (96.5%) and the low posterior error rate (3.5%) show that this evolutionary event was inferred with strong confidence. This result can be explained by the above mentioned colonisation hypothesis if the proportion of suitable habitats for the desert locust in the SRI corridor was low, strongly limiting the carrying capacity during the time for range expansion.

Alternatively, the bottleneck event in *S. g. flaviventris* can be explained by a southern colonisation of Africa through a **long-distance migration event**. Long-distance migrations are possible in the gregarious phase of the desert locust, with swarms of winged adults that regularly travel up to 100 km in a day (Roffey and Magor, 2003). However, since effective displacements are mostly downwind in this species, the likelihood of a southwestern transport of locusts depends on the dynamics of winds and pressure over Africa (Nicholson, 1996; Waloff and Pedgley, 1986). Because in southern Africa, winds blow mostly from the north-east toward the ex-

tant south-western distribution of the desert locust (at least in southern winter, i.e. August; Figure 6.4A), only exceptional conditions of a major plague event may have brought a single or a few swarm(s) in East Africa (see Figure 6.4B) and sourced the colonisation of south-western Africa. In agreement with this, rare southward movements of desert locus have been documented along the eastern coast of Africa, for instance in Mozambique in January 1945 during the peak of the major plague of 1941 – 1947 (Waloff, 1966).

Gain in statistical inferences when incorporating independent information into the mutational prior setting

The mutational rate and spectrum at molecular markers are critical parameters for model-based population genetics inferences. We found that the specification into prior distributions of previous estimations of microsatellite mutation rates and allele size constraints improved the accuracy of the divergence time estimation. Using the naive mutational prior setting, the credible interval was larger. It is worth stressing, however, that the latter credible interval did not nevertheless include another transition to a dry climatic period, such as the Younger Dryas (YD, 12.9 to 11.7 Ky) or the Last Glacial Maximum (LGM, 21.1 to 17.2 Ky), two periods with a more continuous potential ecological range for the desert locust. It also resulted in a downward bias in median estimate, which could have altered our historical interpretations. This down-biased estimate (1.7 Ky) agrees less with the timing of the aridification associated with the SRI opening, from 2.5 Ky to 2 Ky. For scenario choice, the inferential gain in incorporating independent information in mutational prior setting was much more moderate, with error rates decreasing by only a few tens of percent. This highlights, once again, the well-known potential impact of the prior settings assumed in Bayesian analyses, and calls for processing various error and accuracy analyses using different prior settings as realised in the present study. It is therefore natural to question the influence of the demographic priors for which a prior sensitivity analysis should be performed. Some first attempts were carried out by Marie-Pierre Chapuis regarding the use of uniform priors on temporal parameters, instead of log-uniform. A bias and larger credible intervals were observed, similarly to the use of the naive mutational priors.

Implications for the evolution of phase polyphenism

The recent divergence time between populations of *S. g. gregaria* and *S. g. flaviventris* give some insights concerning the evolutionary mechanism responsible for the relative loss of phase polyphenism in the southern subspecies. Given that the *S. g. flaviventris* subspecies arose about 7.7 K generations ago, it seems unlikely that recent mutations are responsible for its phenotypic divergence.

It is more likely that selection explains such a rapid evolution. This requires that alleles associated with the reduction of phase polyphenism in *S. g. flaviventris* were present in past *S. g. gregaria* populations at relatively high frequencies, which may have been favoured through prior adaptation. The southern colonisation was preceded by a prolonged and severe contraction of northern deserts, providing ecological conditions favourable for the evolution of a solitarious phase in the native

environment that may have facilitated adaptation in the novel southern range of the species. Genetic drift might also explain the spread of the reduction in phase polyphenism among *S. g. flaviventris*, intensified by its small effective population size.

Conclusion and perspectives

Most ABC methods require the choice of a distance, a threshold and a set of low dimensional summary statistics, choices which can be difficult and impact the algorithm performances. To address these tuning issues we introduced an ABC approach mixed with Breiman's regression random forests, denoted ABC-RF, in order to infer parameters of interest. This method achieves good prediction accuracy while being mostly tuning-free and suffers less of the curse of dimensionality compared to earlier ABC approaches. Indeed, the robustness of forests toward noise variables allows the inclusion of a large number of explanatory variables and avoids the preliminary selection of a small set of summary statistics. However, some of the summary statistics should still be informative for the task at hand. Furthermore, the forests provide useful interpretability tools as prior measures of errors, as well as summary statistics importance. Note that other random forest benefits could be incorporated, such as variable selection, and could be the subject of future works.

Based on the work of Pudlo et al. (2016) on ABC-RF for model choice, we included the possibility to study groups of models instead of individual ones. On two population genetics case studies (Estoup, Raynal et al., 2018; Chapuis, Raynal et al., 2019), this grouped models strategy gives some insights regarding the difficulty to discriminate a specific evolutionary event thanks to the considered method and data. This is therefore a good complementary approach to the usual analyses of individual models/scenarios, especially when the formation of groups is justified.

In a parameter inference framework, to assess the prediction accuracy of ABC-RF for a specific observation, we proposed to compute some posterior measures of error. Contrary to the out-of-bag prior measures of error returned by ABC-RF, this error describes the prediction accuracy at the observed data of interest by estimating a posterior expectation for it.

ABC-RF methodologies for parameter inference and model choice, as well as the above mentioned developments, are unified in the R package `abcrf`. It has been updated to take advantage of the fast random forest implementation offered by the R package `ranger` (Wright and Ziegler, 2017). For greater ease of use, we plan to write a documentation vignette for our library, and an analogous software to DIYABC is currently under development.

Most ABC frameworks rely only on one observed data. However, the random forests involved in the ABC-RF strategies do not take into account this unique data we would like to infer on. From this statement, we proposed an overview of existing

local random forest methods in the classification setting, as well as new proposals. We compared different strategies to take profit of the observation during the tree construction. There are four main ideas to localise the forest, either by modifying the tree splitting rule, tree aggregation scheme (i.e. how the trees vote), by acting on the sampling of individuals or explanatory variables. Our proposal relying on the local variable importance scheme seems to provide an advantage in terms of prediction, however at least for classification and on the studied examples, we concluded that such local approaches hardly outperform the original forest algorithm. Moreover, their implementations require additional tuning parameters driving the local nature of the methods. For these reasons, the local strategies turned out less relevant than we expected, at least in the classification setting. However, as mentioned below, in a regression framework some leads could be provided by the recent work of Friedberg et al. (2018).

The ABC-RF method we proposed for parameter inference is based on regression random forests built on each dimension of the parameter space. A significant future development would be to extend it to the multivariate case, without breaking the dependence between parameters. Prior to our work, some first attempts were carried by Jean-Michel Marin to use multivariate random forests (Kocev et al., 2007; Segal and Xiao, 2011) coupled with ABC simulations. While unfruitful, we think smarter adaptations need to be investigated, starting with two dimensional parameters. An alternative approach could be based on using the random forest strategies to approximate some conditional distributions, in order to recover the joint posterior using either a Gibbs sampler (based on approximated full conditionals) or Russian rule decompositions to which a product of embedded full conditionals is associated. Very recently, two Gibbs algorithms were independently proposed by Rodrigues, Nott et al. (2019) and Clarté et al. (2019) where each intractable full conditional distribution is approximated thanks to ABC. The first approach fits a regression model on ABC simulations and samples from it. The second employs a classic ABC algorithm and provides some theoretical guaranties.

ABC-RF strategies can now be applied for model choice and parameter inference problems, however a development is still missing, namely a model checking strategy. Indeed, if an inferred model fits the data, then replicated data generated from it should look similar to the observation. In other terms, data generated thanks to the posterior predictive distribution are compared to the observation. This is usually done thanks to statistical tests on data summary statistics (Gelman et al., 2013; Cornuet, Ravigné et al., 2010). Drawing parameters from the whole posterior distribution is thus required to sample from the posterior predictive. Here, some multivariate ABC-RF strategies could be very profitable. However, this checking process implies using the same observed data twice, which is subject to criticism. Some alternatives avoiding this issue could be valuable and interesting leads of research.

So far, the reference tables for the training of ABC-RF methods are simulated from simple prior distributions. As mentioned many times in Chapter 1, ABC sequential methods aim at generating parameters from a distribution closer to the desired posterior. An important aspect that could be beneficial for ABC-RF is the possibility to generate samples from a smarter proposal, in a sequential fashion. A strategy could be to use some preliminary ABC-RF runs to build the reference

table more efficiently, for example by determining the support of the parameter of interest, similarly to the iterated importance sampling strategy of Blum (2010).

Another perspective falls into the frame of regression adjustment methods, presented in Chapter 1. Because a random forest does not take into account $\mathbf{y}$ during its construction, (in other terms it is not local), using it to approximate the desired posterior expectations, and naively introducing them into a regression adjustment technique, will not work. For this reason, a random forest in a regression setting, designed to specifically predict $\mathbf{y}$ would help for this task. Related to this subject but outside the ABC framework, Bloniarz et al. (2016) propose to fit a weighted local linear regression, where weights are deduced thanks to a random forest (instead of using $K_\epsilon(\rho(\eta_{\mathbf{x}^{(i)}}, \eta_{\mathbf{y}})))$, and showed improvements in terms of prediction accuracy compared to the classical random forest method, in addition to the consistency demonstration. This approach could be used to perform local linear adjustment in the same way as presented in Chapter 1, while avoiding the specification of a kernel and distance ρ. Finally, let us mention that Friedberg et al. (2018) propose a similar strategy to the one of Bloniarz et al. (2016). Instead of the usual weighted least squares, they employ the L_2-regularised version. Moreover, the random forest splitting criterion is modified to take into account the fact that a local linear regression will be fitted in a second step. Note that it involves the inclusion of additional tuning parameters, that need to be selected by cross-validation. This work could also give some insights concerning a possible local adaptation of random forests in the regression setting.

Appendices

Appendix A

Supplementary material for Chapter 4

A.1 A basic R code to use the abcrf package version 1.7.1

We provide some basic R lines of code to use the R package `abcrf` and conduct RF inference about parameters. There are two possibilities to read simulated data: the user wants to use a reference table simulated from the software DIYABC v.2.1.0 (Cornuet, Pudlo et al., 2014) recorded within a `.bin` file associated with its `.txt` header file, or the simulated reference table is only contained within a `.txt` file. Of course if the model is simple enough, the user can simulate the reference table himself using its own simulator program. In the following, we assume θ is a vector of `p` parameters and `k` summary statistics are considered. The # symbol means the text on its right is a comment and ignored by R. We here focus on a single parameter of interest labelled "`poi`".

Installing and loading the R package abcrf

```
install.packages("abcrf") # To install the abcrf package (v. 1.7.1)
library(abcrf) # To load the package.
```

Reading data: option 1 - using a .bin and .text files obtained using DIYABC

We assume the reference table is recorded within the `reftable.bin` file and its corresponding header in the `header.txt` file. The function `readRefTable` is used to recover the data.

```
data <- readRefTable(filename = "reftable.bin",
                     header = "header.txt")
# data is a list containing the scenarios (or models) indices,
# the matrix with the parameters, the summary statistics and
```

```
# other information.

# We are here interested in the simulated data of the scenario 1.
index1 <- data$scenarios == 1 # To store the model 1 indexes.

# We then create a data frame composed of the parameter of interest
# poi and the summary statistics of the scenario 1.
data.poi <- data.frame(poi = data$params[index1, "poi"],
                       data$stats[index1, ])
```

Reading data: option 2 - using a `.txt` file

We assume that the reference table is recorded within `yourTxtFile.txt` file, composed of a first column corresponding to the scenario indices, `p` columns of parameters and `k` columns of summary statistics, the first row is the column labels. The field separator character being a white space.

```
data <- read.table(file = "youTxtFile.txt", header = TRUE, sep = "")
# data is a matrix. The first column is the model indices, the next
# p are the parameters, the last k are the summary statistics.

index1 <- data[ , 1] == 1 # To store the model 1 indexes.

# We then create a data frame composed of the parameter of interest
# poi and the summary statistics of model 1.
# p and k have to be defined.
data.poi <- data.frame(poi = data[index1, "poi"],
                       data[index1, (p+2):(p+k+1)])
```

Subsetting your data set

If required, subsetting your data sets stored in `data.poi` can be easily done with the following line.

```
data.poi <- data.poi[1:10000, ]
# If you are interest in the 10,000 first data sets.
```

Training a random forest

The random forest of the ABC-RF method is built thanks to the `regAbcrf` function, its principle arguments being a R formula and the corresponding data frame as training data set. Additional arguments are available, especially the number of trees (`ntree`, with default values `ntree` = 500), the minimum node size (`min.node.size`, with default value `min.node.size` = 5), and the number of covariates randomly considered at each split (`mtry`). See the `regAbcrf` help for further details.

```
model.poi <- regAbcrf(formula = poi~., data = data.poi, ntree = 500,
                      min.node.size = 5, paral = TRUE)
# The used formula means that we are interested in explaining the
# parameter poi thanks to all the remaining columns of data.poi
# (i.e. all the summary statistics).
# The paral argument determines if parallel computing will be
# activated or not.
```

Appendix B

Supplementary material for Chapter 6

B.1 Description of the historical and demographic model-parameters and their prior distributions, for the eight competing scenarios considered for the origin and diversification of Pygmy populations from Western Africa

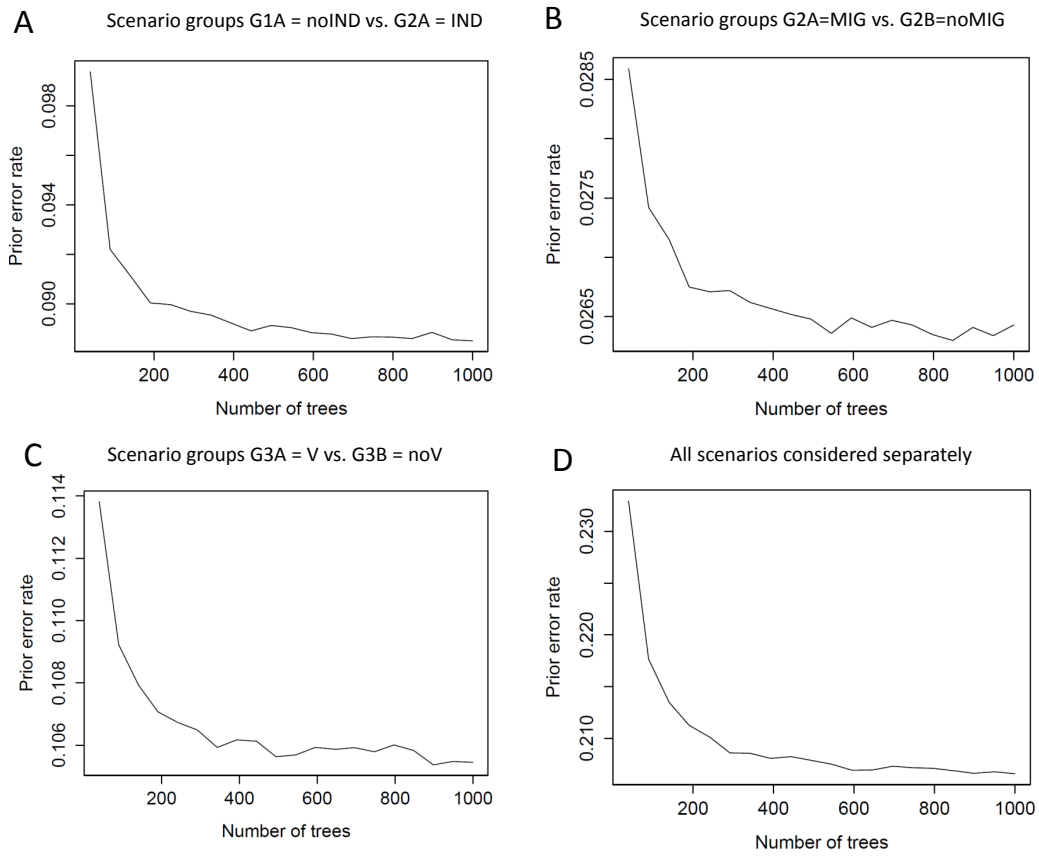
The eight scenarios with their historical and demographic parameters are represented in Chapter 6, Figure 6.1. The column "Scenarios" indicates in which scenario each model parameter appears. The column "Group" indicates in which group of scenarios each model parameter appears when processing a model-grouping approach. The index i indicated for some parameters corresponds to population index (1 - 4 = Pygmy populations and 5 = non-Pygmy African population). Scenarios and associated model parameters follow the same notation as in Verdu, Austerlitz et al. (2009).

Parameter type	**Parameter name**	**Prior distribution**	**Scenarios**	**Group**
Divergence times [(a)]	t_{pnp}; t_p	$\mathcal{U}_{[1;5,000]}$	1, 2, 3, 4	G1A
	t_{pnpi}, with $i \in \{1, \ldots, 4\}$	$\mathcal{U}_{[1;5,000]}$	5, 6, 7, 8	G1B
Admixture times [(a)]	tr_a	$\mathcal{U}_{[1;5,000]}$	1, 2, 3, 4	G1A
	tr_r	Log-$\mathcal{U}_{[1;5,000]}$	1, 2, 3, 4	G1A
	tr_{ai}, with $i \in \{1, \ldots, 4\}$	$\mathcal{U}_{[1;5,000]}$	5, 6, 7, 8	G1B
	tr_{ri}, with $i \in \{1, \ldots, 4\}$	Log-$\mathcal{U}_{[1;5,000]}$	5, 6, 7, 8	G1B
Times of effective population size changes [(a)]	t_A	$\mathcal{U}_{[1;10,000]}$	1, 2, 5, 6	G3A
	no t_A		3, 4, 7, 8	G3B
	t_{nei}, with $i \in \{1, \ldots, 4\}$	$\mathcal{U}_{[1;5,000]}$	5, 6, 7, 8	G1B
Admixture rates	r_a; r_{ai}; r_{ri}, with $i \in \{1, \ldots, 4\}$	$\mathcal{U}_{[0;1]}$	1, 3, 5, 7	G2A
	r_a; r_{ai}; r_{ri}, with $i \in \{1, \ldots, 4\}$	0 (no admixture)	2, 4, 6, 8	G2B
Effective population sizes [(b)]	N_A	$\mathcal{U}_{[100;10,000]}$	1, 2, 5, 6	G3A
	no N_A		3, 4, 7, 8	G3B
	N_{ap}	$\mathcal{U}_{[100;10,000]}$	1, 2, 3, 4	G1A
	N_{ai}, with $i \in \{1, \ldots, 4\}$	$\mathcal{U}_{[100;10,000]}$	5, 6, 7, 8	G1B
	N_{np}	$\mathcal{U}_{[1,000;100,000]}$	1, 2, 3, 4, 5, 6, 7, 8	G1A, G1B, G2A, G2B, G3A, G3B
	N_i, with $i \in \{1, \ldots, 4\}$	$\mathcal{U}_{[100;10,000]}$	1, 2, 3, 4, 5, 6, 7, 8	G1A, G1B, G2A, G2B, G3A, G3B

[(a)] In number of generations (assuming a generation time of 25 years, Verdu, Austerlitz et al., 2009)
[(b)] In number of (reproductively effective) diploid individuals

B.2 Evolution of ABC-RF prior error rates with respect to the number of trees in the forest

The following graphs represent the decrease of the ABC-RF prior error rate with the number of trees in the forest for the four RF analyses conducted using a reference table including $100,000$ simulated data sets. For all analyses the gain of increasing the number of trees becomes limited for a number of trees greater than 800, hence our final choice of building forests from $1,000$ trees.



Bibliography

Adams, J. M. and H. Faure (1997). *Review and atlas of palaeovegetation : preliminary land ecosystem maps of the world since the Last Glacial Maximum*. Oak Ridge National Laboratory, TN, USA.

Aeschbacher, S., M. A. Beaumont and A. Futschik (2012). 'A novel approach for choosing summary statistics in approximate Bayesian computation'. In: *Genetics* 192.3, pp. 1027–1047.

Aha, D. W., D. Kibler and M. K. Albert (1991). 'Instance-based learning algorithms'. In: *Machine Learning* 6 (1), pp. 37–66.

Amaratunga, D., J. Cabrera and Y.-S. Lee (2008). 'Enriched random forests'. In: *Bioinformatics* 24.18, pp. 2010–2014.

Andrieu, C. and G. O. Roberts (2009). 'The pseudo-marginal approach for efficient Monte Carlo computations'. In: *The Annals of Statistics* 37.2, pp. 697–725.

Archer, K. J. and R. V. Kimes (2008). 'Empirical characterization of random forest variable importance measures'. In: *Computational Statistics & Data Analysis* 52.4, pp. 2249–2260.

Armano, G. and E. Tamponi (2018). 'Building forests of local trees'. In: *Pattern Recognition* 76, pp. 380–390.

Athey, S., J. Tibshirani and S. Wager (2019). 'Generalized random forests'. In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Attias, H. (2000). 'A variational Baysian framework for graphical models'. In: *Advances in Neural Information Processing Systems 12*. Ed. by S. A. Solla, T. K. Leen and K. Müller. MIT Press, pp. 209–215.

Auret, L. and C. Aldrich (2011). 'Empirical comparison of tree ensemble variable importance measures'. In: *Chemometrics and Intelligent Laboratory Systems* 105.2, pp. 157–170.

Baharian, S., M. Barakatt, C. R. Gignoux, S. Shringarpure, J. Errington, W. J. Blot, C. D. Bustamante, E. E. Kenny, S. M. Williams, M. C. Aldrich and S. Gravel (2016). 'The great migration and African-American genomic diversity'. In: *PLOS Genetics* 12.5, e1006059.

Bahlo, M. and R. C. Griffiths (2000). 'Inference from gene trees in a subdivided population'. In: *Theoretical Population Biology* 57.2, pp. 79–95.

Barber, S., J. Voss and M. Webster (2015). 'The rate of convergence for approximate Bayesian computation'. In: *Electronic Journal of Statistics* 9, pp. 80–105.

Barnes, C. P., S. Filippi and M. P. H. Stumpf (2012). 'Considerate approaches to constructing summary statistics for ABC model selection'. In: *Statistics and Computing* 22, pp. 1181–1197.

Barthelmé, S. and N. Chopin (2014). 'Expectation propagation for likelihood-free inference'. In: *Journal of the American Statistical Association* 109.505, pp. 315–333.

Beaumont, M. A. (1999). 'Detecting population expansion and decline using microsatellites'. In: *Genetics* 153.4, pp. 2013–2029.

— (2003). 'Estimation of population growth or decline in genetically monitored populations'. In: *Genetics* 164.3, pp. 1139–1160.

— (2010). 'Approximate Bayesian computation in evolution and ecology'. In: *Annual Review of Ecology, Evolution, and Systematics* 41.1, pp. 379–406.

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin and C. P. Robert (2009). 'Adaptive approximate Bayesian computation'. In: *Biometrika* 96.4, pp. 983–990.

Beaumont, M. A., W. Zhang and D. Balding (2002). 'Approximate Bayesian computation in population genetics'. In: *Genetics* 162.4, pp. 2025–2035.

Beerli, P. and J. Felsenstein (1999). 'Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach'. In: *Genetics* 152.2, pp. 763–773.

Bernton, E., P. Jacob, M. Gerber and C. P. Robert (2017). 'Inference in generative models using the Wasserstein distance'. In: *ArXiv e-prints* 1701.05146.

Berthier, P., M. A. Beaumont, J.-M. Cornuet and G. Luikart (2002). 'Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach'. In: *Genetics* 160.2, pp. 741–751.

Besag, J. (1974). 'Spatial interaction and the statistical analysis of lattice systems'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 36.2, pp. 192–236.

— (1975). 'Statistical analysis of non-lattice data'. In: *Journal of the Royal Statistical Society. D (The Statistician)* 24.3, pp. 179–195.

Biau, G. (2012). 'Analysis of a random forest model'. In: *Journal of Machine Learning Research* 13, pp. 1063–1095.

Biau, G., F. Cérou and A. Guyader (2015). 'New insights into approximate Bayesian computation'. In: *Annales de l'Institut Henri Poincaré B, Probability and Statistics* 51.1, pp. 376–403.

Biau, G. and E. Scornet (2016). 'A random forest guided tour'. In: *TEST* 25.2, pp. 197–227.

Bishop, C. M. (1994). *Mixture density networks*. Tech. rep. Neural Computing Research Group, Aston University.

— (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag: New York.

Bloniarz, A., C. Wu, B. Yu and A. Talwalkar (2016). 'Supervised Neighborhoods for Distributed Nonparametric Regression'. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* 51, pp. 1450–1459.

Blum, M. G. B. (2010). 'Approximate Bayesian computation: a nonparametric perspective'. In: *Journal of the American Statistical Association* 105.491, pp. 1178–1187.

Blum, M. G. B. and O. François (2010). 'Non-linear regression models for approximate Bayesian computation'. In: *Statistics and Computing* 20, pp. 63–73.

Blum, M. G. B., M. Nunes, D. Prangle and S. A. Sisson (2013). 'A comparative review of dimension reduction methods in approximate Bayesian computation'. In: *Statistical Science* 28.2, pp. 189–208.

Bostoen, K., B. Clist, C. Doumenge, R. Grollemund, J.-M. Hombert, J. K. Muluwa and J. Maley (2015). 'Middle to late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of Western Central Africa'. In: *Current Anthropology* 56.3.

Breiman, L. (1996). 'Bagging predictors'. In: *Machine Learning* 24.2, pp. 123–140.

— (2000). 'Randomizing outputs to increase prediction accuracy'. In: *Machine Learning* 40.3, pp. 229–242.

— (2001). 'Random Forests'. In: *Machine Learning* 45.1, pp. 5–32.

Breiman, L. and A. Cutler (2003). *Setting up, using, and understanding random forests v4.0*. manual. URL: www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Bryc, K., E. Y. Durand, J. M. Macpherson, D. Reich and J. L. Mountain (2015). 'The genetic ancestry of African Americans, Latinos, and European Americans across the United States'. In: *The American Journal of Human Genetics* 96.1, pp. 37–53.

Bugallo, M. F., V. Elvira, L. Martino, D. Luengo, J. Miguez and P. M. Djuric (2017). 'Adaptive importance sampling: the past, the present, and the future'. In: *IEEE Signal Processing Magazine* 34.4, pp. 60–79.

Calle, M. L. and V. Urrea (2011). 'Letter to the editor: stability of random forest importance measures'. In: *Briefings in Bioinformatics* 12.1, pp. 86–89.

Cappé, O., A. Guillin, J.-M. Marin and C. P. Robert (2004). 'Population Monte Carlo'. In: *Journal of Computational and Graphical Statistics* 13.4, pp. 907–929.

Cavalli-Sforza, L. L. (1986). 'African pygmies'. In: ed. by L. L. Cavalli-Sforza. Orlando Academic Press. Chap. African Pygmies: an evaluation of the state of research, pp. 361–426.

Cavalli-Sforza, L. L. and M. W. Feldman (2003). 'The application of molecular genetic approaches to the study of human evolution'. In: *Nature Genetics* 33, pp. 266–275.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza (1994). *The History and Geography of Human Genes*. Princeton University Press.

Celeux, G. and J. Diebolt (1985). 'The SEM Algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem'. In: *Computational Statistics Quarterly* 2, pp. 73–82.

Chapelle, O., B. Schölkopf and A. Zien (2010). *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press.

Chapuis, M.-P., C. S. Bazalet, L. Blondin, A. Foucart, R. Vitalis and M. J. Samways (2016). 'Subspecific taxonomy of the desert locust, *Schistocerca gregaria* (Orthoptera: Acrididae), based on molecular and morphological characters'. In: *Systematic Entomology* 41.3, pp. 516–530.

Chapuis, M.-P., A. Foucart, C. Plantamp, L. Blondin, N. Leménager, L. Benoit, P.-E. Gay and C. S. Bazalet (2017). 'Genetic and morphological variation in non-polyphenic southern African populations of the desert locust'. In: *African Entomology* 25.1, pp. 13–25.

Chapuis, M.-P., C. Plantamp, L. Blondin, C. Pagès, J. M. Vassal and M. Lecoq (2014). 'Demographic processes shaping genetic variation of the solitarious phase of the desert locust'. In: *Molecular Ecology* 23.7, pp. 1749–1763.

Chapuis, M.-P., C. Plantamp, R. Streiff, L. Blondin and C. Piou (2015). 'Microsatellite evolutionary rate and pattern in *Schistocerca gregaria* inferred from direct observation of germline mutations'. In: *Molecular Ecology* 24.24, pp. 6107–6119.

Chapuis, M.-P., L. Raynal, C. Plantamp, L. Blondin, J.-M. Marin and A. Estoup (2019). 'A young age of subspecific divergence in the desert locust *Schistocerca gregaria*'. In: *bioRxiv*. DOI: 10.1101/671867.

Chen, C., A. Liaw and L. Breiman (2004). 'Using random forest to learn imbalanced data'. In: *preprint, University of California, Berkeley* 110, pp. 1–12.

Chikhi, L., M. W. Bruford and M. A. Beaumont (2001). 'Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo'. In: *Genetics* 158.3, pp. 1347–1362.

Choisy, M., P. Franck and J.-M. Cornuet (2004). 'Estimating admixture proportions with microsatellites: comparison of methods based on simulated data'. In: *Molecular Ecology* 13, pp. 955–968.

Clarté, G., C. P. Robert, R. Ryder and J. Stoehr (2019). 'Component-wise approximate Bayesian computation via Gibbs-like steps'. In: *ArXiv e-prints* 1905.13599.

Cleveland, W. S. (1979). 'Robust locally weighted regression and smoothing scatterplots'. In: *Journal of the American Statistical Association* 74.368, pp. 829–836.

Cleveland, W. S. and S. J. Devlin (1988). 'Locally weighted regression: an Approach to regression analysis by local fitting'. In: *Journal of the American Statistical Association* 83.403, pp. 596–610.

Consonni, G. and J.-M. Marin (2007). 'Mean-field variational approximate Bayesian inference for latent variable models'. In: *Computational Statistics & Data Analysis* 52.2, pp. 790–798.

Cornuet, J.-M. and M. A. Beaumont (2007). 'A note on the accuracy of PAC-likelihood inference with microsatellite data'. In: *Theoretical Population Biology* 71.1, pp. 12–19.

Cornuet, J.-M., J.-M. Marin, A. Mira and C. P. Robert (2012). 'Adaptive multiple importance sampling'. In: *Scandinavian Journal of Statistics* 39.4.

Cornuet, J.-M., P. Pudlo, J. Veyssier, A. Dehne-Garcia, M. Gautier, R. Leblois, J.-M. Marin and A. Estoup (2014). 'DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data'. In: *Bioinformatics* 30.8, pp. 1187–1189.

Cornuet, J.-M., V. Ravigné and A. Estoup (2010). 'Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0)'. In: *BMC Bioinformatics* 11.1, p. 401.

Cornuet, J.-M., F. Santos, M. A. Beaumont, C. P. Robert, J.-M. Marin, D. J. Balding, T. Guillemaud and A. Estoup (2008). 'Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation'. In: *Bioinformatics* 24.23, pp. 2713–2719.

Csilléry, K., O. François and M. G. B. Blum (2012). 'abc: an R package for approximate Bayesian computation (ABC)'. In: *Methods in Ecology and Evolution* 3.3, pp. 475–479.

De Iorio, M. and R. Griffiths (2004a). 'Importance sampling on coalescent histories. I'. In: *Advances in Applied Probability* 36.2, pp. 417–433.

— (2004b). 'Importance sampling on coalescent histories. II: Subdivided population models'. In: *Advances in Applied Probability* 36.2, pp. 434–454.

De Iorio, M., R. Griffiths, R. Leblois and F. Rousset (2005). 'Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models'. In: *Theoretical Population Biology* 68.1, pp. 41–53.

Del Moral, P., A. Doucet and A. Jasra (2006). 'Sequential Monte Carlo samplers'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 68.3, pp. 411–436.

— (2012). 'An adaptive sequential Monte Carlo method for approximate Bayesian computation'. In: *Statistics and Computing* 22.5, pp. 1009–1020.

Deligiannidis, G., A. Doucet and M. K. Pitt (2018). 'The correlated pseudomarginal method'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 80.5.

Delyon, B., M. Lavielle and E. Moulines (1999). 'Convergence of a stochastic approximation version of the EM algorithm'. In: *The Annals of Statistics* 27.1, pp. 94–128.

deMenocal, P., J. Ortiz, T. Guilderson, J. Adkins, M. Sarnthein, L. Baker and M. Yarusinsky (2000). 'Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing'. In: *Quaternary Science Reviews* 1-5, pp. 347–361.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). 'Maximum likelihood from incomplete data via the EM algorithm'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 39.1, pp. 1–38.

Destro-Bisol, G., F. Donati, V. Coia, I. Boschi, F. Verginelli, A. Caglià, S. Tofanelli, G. Spedini and C. Capelli (2004). 'Variation of female and male lineages in Sub-Saharan populations: the importance of sociocultural factors'. In: *Molecular Biology and Evolution* 21.9, pp. 1673–1682.

Díaz-Uriarte, R. and S. Alvarez de Andrés (2006). 'Gene selection and classification of microarray data using random forest'. In: *BMC Bioinformatics* 7, p. 3.

Dietterich, T. G. (2000). 'Ensemble methods in machine learning'. In: *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*. Vol. 1857. 1-15. Springer, Berlin, Heidelberg.

Dietterich, T. G. and E. B. Kong (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Tech. rep. Technical Report, Oregon State University.

Drovandi, C. C. (2018). 'ABC and indirect inference'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S.A. Sisson, Y. Fan and M. A. Beaumont. 179-210. Chapman and Hall/CRC.

Drovandi, C. C. and A. N. Pettitt (2011). 'Estimation of parameters for macroparasite population evolution using approximate Bayesian computation'. In: *Biometrics* 67.1, pp. 225–233.

Drovandi, C. C., A. N. Pettitt and A. Lee (2015). 'Bayesian indirect inference using a parametric auxiliary model'. In: *Statistical Science* 30.1, pp. 72–95.

Estoup, A., P. Jarne and J.-M. Cornuet (2002). 'Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis'. In: *Molecular Ecology* 11.9, pp. 1591–1604.

Estoup, A., E. Lombaert, J.-M. Marin, C. P. Robert, T. Guillemaud, P. Pudlo and J.-M. Cornuet (2012). 'Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics'. In: *Molecular Ecology Resources* 12.5, pp. 846–855.

Estoup, A., L. Raynal, P. Verdu and J.-M. Marin (2018). 'Model choice using approximate Bayesian computation and random forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations'. In: *Journal de la Société Française de Statistique* 159.3, pp. 167–190.

Estoup, A., P. Verdu, J.-M. Marin, C. P. Robert, A. Dehne-Garcia, J.-M. Cornuet and P. Pudlo (2018). 'Application of approximate Bayesian computation to infer the genetic history of Pygmy hunter-gatherers populations from Western Central Africa'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S. A. Sisson, Y. Fan and M. A. Beaumont. 541-568. Chapman and Hall/CRC.

Excoffier, L., I. Dupanloup, E. Huerta-Sanchez, V. C. Sousa and M. Foll (2013). 'Robust demographic inference from genomic and SNP data'. In: *PLOS Genetics* 9.10, pp. 1–17.

Excoffier, L., A. Estoup and J.-M. Cornuet (2005). 'Bayesian analysis of an admixture model with mutations and arbitrarily linked markers'. In: *Genetics* 169, pp. 1727–1738.

Fagundes, N. J. R., N. Ray, M. A. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto and L. Excoffier (2007). 'Statistical evaluation of alternative models of human evolution'. In: *Proceedings of the National Academy of Sciences, USA* 104.45, pp. 17614–17619.

Fan, J. (1993). 'Local linear regression smoothers and their minimax efficiencies'. In: *The Annals of Statistics* 21.1, pp. 196–216.

Fan, Y., S. R. Meikle, G. I. Angelis and A. Sitek (2018). 'ABC in nuclear imaging'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S. A. Sisson, Y. Fan and M. A. Beaumont. 623-647. Chapman and Hall/CRC.

Fasiolo, M. and S. N. Wood (2018). 'ABC in ecological modelling'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S. A. Sisson, Y. Fan and M. A. Beaumont. 597-622. Chapman and Hall/CRC.

Fayyad, U. M. and K. B. Irani (1995). 'Multi-interval discretization of continuous-valued attributes for classification learning'. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence* 2, pp. 1022–1027.

Fearnhead, P. and P. Donnelly (2002). 'Approximate likelihood methods for estimating local recombination rates'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 64.4, pp. 657–680.

Fearnhead, P. and D. Prangle (2012). 'Constructing summary statistics for Approximate Bayesian computation: semi-automatic Approximate Bayesian computation'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 74.3, pp. 419–474.

Felsenstein, J. (1981). 'Evolutionary trees from DNA sequences: a maximum likelihood approach'. In: *Journal of Molecular Evolution* 17.6, pp. 368–376.

Felsenstein, J., M. K. Kuhner, J. Yamato and P. Beerli (1999). 'Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data'. In: *Statistics in Molecular Biology and Genetics*. Ed.

by F. Seillier-Moiseiwitsch. Vol. 33 of *IMS Lecture Notes-Monograph Series*. 163-185. Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA.

Fern, X. Z and C. E Brodley (2003). 'Boosting lazy decision trees'. In: *Proceedings of the Twentieth International Conference on Machine Learning* 20.1, pp. 178–185.

Fernández-Delgado, M., E. Cernadas, S. Barro and D. Amorim (2014). 'Do we need hundreds of classifiers to solve real world classification problems'. In: *Journal of Machine Learning Research* 15, pp. 3133–3181.

Fernhead, P. (2018). 'Asymptotics of ABC'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S. A. Sisson, Y. Fan and M. A. Beaumont. 269-288. Chapman and Hall/CRC.

Filippi, S., C. P. Barnes and M. P. H. Stumpf (2012). 'Contribution to the discussion of Fearnhead and Prangle (2012)'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 74.3, pp. 459–460.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. 1st. Clarendon Press.

Frazier, D. T., G. M. Martin, C. P. Robert and J. Rousseau (2018). 'Asymptotic properties of approximate Bayesian computation'. In: *Biometrika* 105.3, pp. 593–607.

Friedberg, R., J. Tibshirani, S. Athey and S. Wager (2018). 'Local linear forests'. In: *ArXiv e-prints* 1807.11408.

Friedman, J. H., R. Kohavi and Y. Yun (1997). 'Lazy decision trees'. In: *Proceedings of the 13th National Conference on AAAI*, pp. 717–724.

Fulton, T., S. Kasif, S. Salzberg and D. L. Waltz (1996). 'Local induction of decision trees: towards interactive data mining'. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, pp. 14–19.

Gallant, A. R. and G. Tauchen (1996). 'Which moments to match?' In: *Econometric Theory* 12.4, pp. 657–681.

Galván, I. M., J. M. Valls, N. Lecomte and P. Isasi (2009). 'A lazy approach for machine learning algorithms'. In: *IFIP International Federation for Information Processing* 296, pp. 517–522.

Gammerman, A., V. Vovk and V. Vapnik (1998). 'Learning by transduction'. In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. Madison, Wisconsin: Morgan Kaufmann Publishers Inc., pp. 148–155.

Garza, J. and E. Williamson (2001). 'Detection of reduction in population size using data from microsatellite DNA'. In: *Molecular Ecology* 10, pp. 305–318.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (2013). *Bayesian Data Analysis*. 3nd ed. Chapman and Hall/CRC.

Genuer, R., J.-M. Poggi and C. Tuleau-Malot (2008). *Random forests: some methodological insights*. [Research Report] RR-6729, INRIA.

— (2010). 'Variable selection using random forests'. In: *Pattern Recognition Letters* 31.14, pp. 2225–2236.

Genuer, R., J.-M. Poggi, C. Tuleau-Malot and N. Villa-Vialaneix (2017). 'Random forests for big data'. In: *Big Data Research* 9, pp. 28–46.

Gleim, E. and C. Pigorsch (2013). *Approximate Bayesian computation with indirect summary statistics*. Tech. rep. University of Bonn, Bonn, Germany.

Goldstein, D., A. Linares, L. Cavalli-Sforza and N. Feldman (1995). 'Genetic absolute dating based on microsatellites and the origin of modern humans'. In: *Proceedings of the National Academy of Sciences, USA* 92.15, pp. 6723–6727.

Gordon, N., J. Salmond and A. Smith (1993). 'A novel approach to non-linear/non-Gaussian Bayesian state estimation'. In: *IEEE Proceedings on Radar and Signal Processing* 140, pp. 107–113.

Gourieroux, C., A. Monfort and E. Renault (1993). 'Indirect inference'. In: *Journal of Applied Econometrics* 8.S1, pp. 85–118.

Gregorutti, B., B. Michel and P. Saint-Pierre (2017). 'Correlation and variable importance in random forests'. In: *Statistics and Computing* 27.3, pp. 659–678.

Grelaud, A., J.-M. Marin, C.P. Robert, F. Rodolphe and F. Tally (2009). 'Likelihood-free methods for model choice in Gibbs random fields'. In: *Bayesian Analysis* 3.2, pp. 427–442.

Griffiths, R. C. and S. Tavaré (1994a). 'Ancestral inference in population genetics'. In: *Statistical Science* 9.3, pp. 307–319.

— (1994b). 'Sampling theory for neutral alleles in a varying environment'. In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences.* 344.1310, pp. 403–410.

— (1994c). 'Simulating probability distributions in the coalescent'. In: *Theoretical population biology* 46.2, pp. 131–159.

— (1999). 'The ages of mutations in gene trees'. In: *Annals of Applied Probability* 9.3, pp. 567–590.

Gunawan, D., M.-N. Tran and R. Kohn (2017). 'Fast inference for intractable likelihood problems using variational Bayes'. In: *ArXiv e-prints* 1705.06679.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson and C. D. Bustamante (2009). 'Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data'. In: *PLOS Genetics* 5.10.

Hansen, L. P. (1982). 'Large sample properties of generalized method of moments estimators'. In: *Econometrica* 50.4, pp. 1029–1054.

Hansen, L. P. and K. J. Singleton (1982). 'Generalized instrumental variables estimation of nonlinear rational expectations models'. In: *Econometrica* 50.5, pp. 1269–1286.

Hapfelmeier, A. and K. Ulm (2013). 'A new variable selection approach using random forests'. In: *Computational Statistics & Data Analysis* 60, pp. 50–69.

Hastie, T. and C. Loader (1993). 'Local regression: automatic kernel carpentry (with discussion)'. In: *Statistical Science* 8.2, pp. 120–129.

Hastie, T., R. Tibshirani and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd. Springer series in statistics. Springer.

Heggland, K. and A. Frigessi (2004). 'Estimating functions in indirect inference'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 66.2, pp. 447–462.

Henn, B. M., L. L. Cavalli-Sforza and M. W. Feldman (2012). 'The great human expansion'. In: *Proceedings of the National Academy of Sciences, USA* 109.44, pp. 17758–17764.

Hewlett, B. S (2014). *Hunter-gatherers of the Congo Basin: cultures, histories, and biology of African Pygmies*. New Brunswick: Transactions Publishers.

Hewlett, B.S. (1996). 'Cultural Diveristy among Twentieth-Century Foragers: An African Perspective'. In: ed. by S. Kent. Cambridge: Cambridge University Press. Chap. Cultural diversity among African pygmies, pp. 361–426.

Ho, T. K. (1998). 'The random subspace method for constructing decision forests'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8, pp. 832–844.

Holden, P. B., N. R. Edwards, J. Hensman and R. D. Wilkinson (2018). 'ABC for climate: dealing with expensive simulators'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S.A. Sisson, Y. Fan and M. A. Beaumont. 569-595. Chapman and Hall/CRC.

Hudson, R. R. (1991). 'Gene genealogies and the coalescent process'. In: *Oxford Surveys in Evolutionary Biology* 7, pp. 1–44.

— (2001). 'Two-locus sampling distributions and their application'. In: *Genetics* 159.4, pp. 1805–1817.

— (2002). 'Generating samples under a Wright-Fisher neutral model of genetic variation'. In: *Bioinformatics* 18.2, pp. 337–338.

Ishwaran, H. (2007). 'Variable importance in binary regression trees and forests'. In: *Electronic Journal of Statistics* 1, pp. 519–537.

— (2015). 'The effect of splitting on random forests'. In: *Machine Learning* 99.1, pp. 75–118.

Ishwaran, H. and U. B. Kogalur (2019). *Fast unified random forests for survival, regression, and classification (RF-SRC)*. R package version 2.9.0. manual. URL: https://cran.r-project.org/package=randomForestSRC.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone and M. S. Lauer (2008). 'Random survival forests'. In: *The Annals of Applied Statistics* 2.3, pp. 841–860.

Janitza, S., E. Celik and A.-L. Boulesteix (2018). 'A computationally fast variable importance test for random forests for high-dimensional data'. In: *Advances in Data Analysis and Classification* 12.4, pp. 885–915.

Jin, L. and R. Chakraborty (1994). 'Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data'. In: *Molecular Biology and Evolution* 11.1, pp. 120–127.

Joyce, P. and P. Marjoram (2008). 'Approximately sufficient statistics and Bayesian computation'. In: *Statistical Application in Genetics and Molecular Biology* 7.1, pp. 1544–6115.

Kelleher, J., A. M. Etheridge and G. McVean (2016). 'Efficient coalescent simulation and genealogical analysis for large sample sizes'. In: *PLOS Computational Biology* 12.5, pp. 1–22.

Kingman, J. F. C. (1982a). 'Exchangeability and the evolution of large populations'. In: *Exchangeability in Probability and Statistics*. North-Holland, Amsterdam, pp. 97–112.

— (1982b). 'On the genealogy of large populations'. In: *Journal of Applied Probability* 19.A, pp. 27–43.

— (1982c). 'The coalescent'. In: *Stochastic Processes and their Applications* 13, pp. 235–248.

Klinger, E. and J. Hasenauer (2017). 'A scheme for adaptive selection of population sizes in approximate Bayesian computation - sequential Monte Carlo'. In: *Computational Methods in Systems Biology. Lecture Notes in Computer Science*. Ed. by J. Feret and H. Koeppl. Vol. 10545. CMSB 2017. Springer, Cham, pp. 128–144.

Klinger, E., D. Rickert and J. Hasenauer (2018). 'pyABC: distributed, likelihood-free inference'. In: *Bioinformatics* 34.20, pp. 3591–3593.

Kocev, D., C. Vens, J. Struyf and S. Džeroski (2007). 'Ensembles of multi-objective decision trees'. In: *Machine Learning: ECML 2007. Lecture Notes in Computer Science*. Ed. by J. N. Kok, J. Koronacki, R. L. Mantazas, S. Matwin, D. MladeniČ and A. Skowron. Vol. 4701. ECML 2007. Springer, Berlin, Heidelberg, pp. 624–631.

Kröpelin, S., D. Verschuren and A.-M. et al. Lézine (2008). 'Climate-driven ecosystem succession in the Sahara: the past 6000 years'. In: *Science* 130.5877, pp. 765–768.

Kuhner, M. K., J. Yamato and J. Felsenstein (1995). 'Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling'. In: *Genetics* 140.4, pp. 1421–1430.

— (1998). 'Maximum likelihood estimation of population growth rates based on the coalescent'. In: *Genetics* 149.1, pp. 429–434.

Larribe, F. and P. Fearnhead (2011). 'On composite likelihoods in statistical genetics'. In: *Statistical Sinica* 21, pp. 43–69.

Leuenberger, C. and D. Wegmann (2010). 'Bayesian computation and model selection without likelihoods'. In: *Genetics* 184.1, pp. 243–252.

Li, N. and M. Stephens (2003). 'Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data'. In: *Genetics* 165.4, pp. 2213–2233.

Liaw, A. and M. Wiener (2002). 'Classification and regression by randomForest'. In: *R news* 2.3, pp. 18–22.

Liepe, J. and M. P. H. Stumpf (2018). 'ABC in systems biology'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S.A. Sisson, Y. Fan and M. A. Beaumont. 513-539. Chapman and Hall/CRC.

Lindsay, B. G. (1988). 'Composite likelihood methods'. In: *Contemporary Mathematics* 80, pp. 221–239.

Lippens, C., A. Estoup, M. K. Hima, A. Loiseau, C. Tatard, A. Dalecky, K. Bâ, M. Kane, M. Diallo, A. Sow, Y. Niang, S. Piry, K. Berthier, R. Leblois, J. M. Duplantier and C. Brouat (2017). 'Genetic structure and invasion history of the house mouse (Mus musculus domesticus) in Senegal, West Africa: a legacy of colonial and contemporary times'. In: *Heredity* 119.2, pp. 64–75.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing*. second. New York, NY: Springer-Verlag.

Lombaert, E., T. Guillemaud, J.-M. Cornuet, T. Malausa, B. Facon and A. Estoup (2010). 'Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird'. In: *PLOS one* 5.3, e9743.

Lorenz, M. W. (2009). 'Migration and trans-Atlantic flight of locusts'. In: *Quaternary International* 196.1-2, pp. 4–16.

Louppe, G., L. Wehenkel, A. Sutera and P. Geurts (2013). 'Understanding variable importances in forests of randomized trees'. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger. Curran Associates, Inc., pp. 431–439.

Lueckmann, J.-M., P. J. Gonçalves, G. Bassetto, K. Öcal, M. Nonnenmacher and J. H. Macke (2017). 'Flexible statistical inference for mechanistic models of neural dynamics'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. USA: Curran Associates Inc., pp. 1289–1299.

Maley, J., C. Doumenge, P. Giresse, G. Mahé and N. et al. Philippon (2018). 'Late Holocene forest contraction and fragmentation in central Africa'. In: *Quaternary Research* 89.1, pp. 43–59.

Maples, B. K., S. Gravel, E. E. Kenny and C. D. Bustamante (2013). 'RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference'. In: *The American Journal of Human Genetics* 93.2, pp. 278–288.

Margineantu, D. D. and T. G. Dietterich (2003). 'Improved class probability estimates from decision tree models'. In: *Lecture Notes in Statistics - Nonlinear Estimation and Classification* 171, pp. 173–188.

Marin, J.-M., P. Pudlo, A. Estoup and C. P. Robert (2018). 'Likelihood-free model choice'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S.A. Sisson, Y. Fan and M. A. Beaumont. 153-178. Chapman and Hall/CRC.

Marin, J.-M., P. Pudlo, C. P. Robert and R. J. Ryder (2012). 'Approximate Bayesian computational methods'. In: *Statistics and Computing*, pp. 1–14.

Marin, J.-M. and C. P. Robert (2014). *Bayesian Essentials with R*. Springer Texts in Statistics.

Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré (2003). 'Markov chain Monte Carlo without likelihoods'. In: *Proceedings of the National Academy of Sciences, USA* 100.26, pp. 15324–15328.

Maudes, J., J. J. Rodríguez, C. Carcía-Osorio and N. Garcá-Pedrajas (2012). 'Random feature weights for decision tree ensemble construction'. In: *Information Fusion* 13.1, pp. 20–30.

McFadden, D. (1989). 'A method of simulated moments for estimation of the discrete response models without numeric integration'. In: *Econometrica* 57.5, pp. 995–1026.

Meinshausen, N. (2006). 'Quantile regression forests'. In: *Journal of Machine Learning Research* 7, pp. 983–999.

Mentch, L. and G. Hooker (2016). 'Quantifying uncertainty in random forests via confidence intervals and hypothesis tests'. In: *Journal of Machine Learning Research* 17, pp. 1–41.

Merle, C., R. Leblois, F. Rousset and P. Pudlo (2017). 'Resampling: an improvement of importance sampling in varying population size models'. In: *Theoretical Population Biology* 114, pp. 70–87.

Mondal, M., J. Bertranpetit and O. Lao (2019). 'Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania'. In: *Nature Communications* 10.246.

Monfardini, C. (1998). 'Estimating stochastic volatility models through indirect inference'. In: *The Econometrics Journal* 1.1, pp. C113–C128.

Nei, M. (1972). 'Genetic distance between populations'. In: *The American Naturalist* 106.949, pp. 283–292.

— (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York, USA.

Nembrini, S., I. R. König and M. N. Wright (2018). 'The revival of the Gini importance?' In: *Bioinformatics* 34.12, pp. 3711–3718.

Nicholson, S. E. (1996). 'A review of climate dynamics and climate variability in Eastern Africa'. In: *Limnology, Climatology and Paleoclimatology of the East*

African Lakes. Ed. by T. C. Johnson and E. O. Odada. 25-56. Gordon and Breach, Amsterdam.

Nielsen, R. (1997). 'A likelihood approach to populations samples of microsatellite alleles'. In: *Genetics* 146.2, pp. 711–716.

— (2000). 'Estimation of population parameters and recombination rates from single nucleotide polymorphisms'. In: *Genetics* 154.2, pp. 931–942.

Nunes, M. A. and D. J. Balding (2010). 'On optimal selection of summary statistics for approximate Bayesian computation'. In: *Statistical Application in Genetics and Molecular Biology* 9.1, Article 34.

Nunes, M. A. and D. Prangle (2015). 'abctools: an R package for tuning approximate Bayesian computation analysis'. In: *The R Joural* 7.2, pp. 189–205.

Ohta, T. and M. Kimura (1973). 'A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population'. In: *Genetical Research* 22.2, pp. 201–204.

O'Neill, P. D., D. J. Balding, N. G. Becker, M. Eerola and D. Mollison (2000). 'Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods'. In: *Journal of the Royal Statistical Society. C (Applied Statistics)* 49.4, pp. 517–542.

O'Ryan, C., E. H. Harley, M. W. Bruford and M. Beaumont (1998). 'Microsatellite analysis of genetic diversity in fragmented South African buffalo populations'. In: *Animal Conservation* 1.2, pp. 85–94.

Owen, A. B. (2000). 'Safe and effective importance sampling'. In: *Journal of the American Statistical Association* 95.44, pp. 135–143.

Papamakarios, G. and I. Murray (2016). 'Fast ϵ-free inference of simulation models with Bayesian conditional density estimation'. In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and Garnett R. Curran Associates, Inc., pp. 1028–1036.

Pascual, M., M. Chapuis, F. Mestres, J. Balanyà, R. Huey, G. Gilchrist and A. Estoup (2007). 'Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods'. In: *Molecular Ecology* 19, pp. 3069–3083.

Prangle, D. (2017). 'Adapting the ABC distance function'. In: *Bayesian Analysis* 12.1, pp. 289–309.

— (2018). 'Summary statistics'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S. A. Sisson, Y. Fan and M. A. Beaumont. 125-152. Chapman and Hall/CRC.

Price, L. F., C. C. Drovandi, A. Lee and D. J. Nott (2018). 'Bayesian synthetic likelihood'. In: *Journal of Computational and Graphical Statistics* 27.1, pp. 1–11.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman (1999). 'Population growth of human Y chromosomes: a study of Y chromosome microsatellites'. In: *Molecular Biology and Evolution* 16, pp. 1791–1798.

Pudlo, P., J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier and C. P. Robert (2016). 'Reliable ABC model choice via random forests'. In: *Bioinformatics* 32.6, pp. 859–866.

Rannala, B. and J. Mountain (1997). 'Detecting immigration by using multilocus genotypes'. In: *Proceedings of the National Academy of Sciences, USA* 94, pp. 9197–9201.

Raynal, L., J.-M. Marin, P. Pudlo, M. Ribatet, C. P. Robert and A. Estoup (2019). 'ABC random forests for Bayesian parameter inference'. In: *Bioinformatics* 35.10, pp. 1720–1728.

Robert, C. P. and G. Casella (2005). *Monte Carlo Statistical Methods*. second. Berlin, Heidelberg: Springer-Verlag.

Robert, C. P., J.-M. Cornuet, J.-M. Marin and N. S. Pillai (2011). 'Lack of confidence in approximate Bayesian computation model choice'. In: *Proceedings of the National Academy of Sciences, USA* 108.37, pp. 15112–15117.

Robins, G., P. Pattison, Kalish. Y. and D. Lusher (2007). 'An introduction to exponential random graph (p^*) models for social networks'. In: *Social Networks* 29.2, pp. 173–191.

Robnik-Šikonja, M. (2004). 'Improving Random Forests'. In: *Machine Learning: ECML 2004*. Ed. by Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti and Dino Pedreschi. Berlin, Heidelberg: Springer Berlin, Heidelberg, pp. 359–370.

Rodrigues, G. S., A. R. Francis, S. A. Sisson and M. M. Tanaka (2018). 'Inference on the acquisition of multi-drug resistance in mycobacterium tuberculosis using molecular epidemiological data'. In: *Handbook of Approximate Bayesian Computation*. Ed. by S.A. Sisson, Y. Fan and M. A. Beaumont. 482-511. Chapman and Hall/CRC.

Rodrigues, G. S., D. J. Nott and S. A. Sisson (2019). 'Likelihood-free approximate Gibbs sampling'. In: *ArXiv e-prints* 1906.04347.

Roffey, J. and J. I. Magor (2003). *Desert locust population parameters*. Tech. rep. 30. FAO, Rome, Italy: Desert Locust Field Research Stations, Technical Series, p. 29.

Rousset, F., C. R. Beeravolu and R. Leblois (2018). 'Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations'. In: *Journal de la Société Française de Statistique* 159.3, pp. 142–166.

RoyChoudhury, A., J. Felsenstein and E. A. Thompson (2008). 'A two-stage pruning algorithm for likelihood computation for a population tree'. In: *Genetics* 180.2, pp. 1095–1105.

RoyChoudhury, A. and M. Stephens (2007). 'Fast and accurate estimation of the population-scaled mutation rate, θ, From microsatellite genotype data'. In: *Genetics* 176.2, pp. 1363–1366.

Rubin, D. (1984). 'Bayesianly justifiable and relevant frequency calculations for the applied statistician'. In: *The Annals of Statistics* 12, pp. 1151–1172.

Rue, H., S. Martino and N. Chopin (2009). 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 71.2, pp. 319–392.

Salles, T., M. Gonçalves, V. Rodrigues and L. Rocha (2018). 'Improving random forests by neighborhood projection for effective text classification'. In: *Information Systems* 77, pp. 1–21.

Saulnier, E., O. Gascuel and S. Alizon (2017). 'Inferring epidemiological parameters from phylogenies using regression-ABC: a comparative study'. In: *PLOS Computational Biology* 13.3, e1005416.

Schiffels, S. and R. Durbin (2014). 'Inferring human population size and separation history from multiple genome sequences'. In: *PLOS Computational Biology* 46, pp. 919–925.

Schwartz, D. (1992). 'Assèchement climatique vers 3000 B.P. et expansion Bantu en Afrique centrale atlantique : quelques réflexions'. In: *Bulletin de la Société Géologique de France* 163.3, pp. 353–361.

Scornet, E., G. Biau and J.-P. Vert (2015). 'Consistency of random forests'. In: *Annals of Statistics* 43.4, pp. 1716–1741.

Sedki, M. A. and P. Pudlo (2012). 'Contribution to the discussion of Fearnhead and Prangle (2012)'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 74.3, pp. 466–467.

Segal, M. and Y. Xiao (2011). 'Multivariate random forests'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 80–87.

Seligman, M. (2019). *Rborist: extensible, parallelizable implementation of the random forest algorithm*. R package version 0.1-17. URL: https://cran.r-project.org/package=Rborist.

Sheehan, S. and Y. S. Song (2016). 'Deep learning for population genetic inference'. In: *PLOS Computational Biology* 12.3.

Sisson, S. A., Y. Fan and M. A. Beaumont (2018). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC.

Sisson, S. A., Y. Fan and Mark M. Tanaka (2007). 'Sequential Monte Carlo without likelihoods'. In: *Proceedings of the National Academy of Sciences, USA* 104.6, pp. 1760–1765.

Sisson, S.A., Y. Fan and M.M. Tanaka (2009). 'Sequential Monte Carlo without likelihoods: Errata'. In: *Proceedings of the National Academy of Sciences, USA* 106.39, p. 16889.

Smith, J. A. A. (1993). 'Estimating nonlinear time-series models using simulated vector autoregressions'. In: *Journal of Applied Econometrics* 8.S1, S63–S84.

Smith, N. G. C. and P. Fearnhead (2005). 'A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness'. In: *Genetics* 171.4, pp. 2051–2062.

Stephens, M. and P. Donnelly (2000). 'Inference in molecular population genetics'. In: *Journal of the Royal Statistical Society. B (Statistical Methodology)* 62.4, pp. 605–655.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis (2008). 'Conditional variable importance for random forests'. In: *BMC Bioinformatics* 9, p. 307.

Strobl, C., A.-L. Boulesteix, A. Zeileis and T. Hothorn (2007). 'Bias in random forest variable importance measures: illustrations, sources and a solution'. In: *BMC Bioinformatics* 8, p. 25.

Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll and C. Dessimoz (2013). 'Approximate Bayesian computation'. In: *PLOS Computational Biology* 9.1, e1002803.

Sword, G. A., Lecoq M. and Simpson S. J. (2010). 'Phase polyphenism and preventative locust management'. In: *Journal of Insect Physiology* 56.8, pp. 949–957.

Tanner, M. A. and W. H. Wong (1987). 'The calculation of posterior distributions by data augmentation'. In: *Journal of the American Statistical Association* 82.398, pp. 528–540.

Tavaré, S., D. Balding, R. Griffiths and P. Donnelly (1997). 'Inferring coalescence times from DNA sequence data'. In: *Genetics* 145.2, pp. 505–518.

The 1000 genomes project consortium (2012). 'An integrated map of genetic variation from 1,092 human genomes'. In: *Nature* 491, pp. 56–65.

Thouzeau, V., P. Mennecier, P. Verdu and F. Austerlitz (2017). 'Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations'. In: *Proceedings of the Royal Society of London B: Biological Sciences* 284.1861.

Toloşi, L. and T. Lengauer (2011). 'Classification with correlated features: unreliability of feature ranking and solutions'. In: *Bioinformatics* 27.14, pp. 1986–1994.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen and M. P.H. Stumpf (2009). 'Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems'. In: *Journal of the Royal Society Interface* 6.31, pp. 187–202.

Tran, M.-N., R. Kohn, M. Quiroz and M. Villani (2017a). 'The block pseudo-marginal sampler'. In: *ArXiv e-prints* 1603.02485.

— (2017b). 'Variational Bayes with intractable likelihood'. In: *Journal of Computational and Graphical Statistics* 26.4, pp. 873–882.

Tran, M.-N., M. Scharth, M. K. Pitt and R. Kohn (2013). 'Importance sampling squared for Bayesian inference in latent variable models'. In: *ArXiv e-prints* 1309.3339.

Tsymbal, A., M. Pechenizkiy and Pádraig Cunningham (2006). 'Dynamic integration with random forests'. In: *Machine Learning: ECML 2006. ECML 2006. Lecture Notes in Computer Science*. Ed. by J. Fürnkranz, T. Scheffer and M. Spiliopoulou. Berlin, Heidelberg: Springer Berlin, Heidelberg, pp. 801–808.

Vansina, J. (1995). 'New linguistic evidence and 'the Bantu expansion''. In: *The Journal of African History* 36.2, pp. 173–195.

Varin, C. (2008). 'On composite marginal likelihoods'. In: *Advances in Statistical Analysis* 92, pp. 1–28.

Varin, C., N. Reid and D. Firth (2011). 'An overview of composite likelihood methods'. In: *Statistica Sinica* 21.1, pp. 5–42.

Verdu, P., N. S. Becker, A. Froment, M. Georges, V. Grugni, L. Quintana-Murci, J.-M. Hombert, L. Van der Veen, S. Le Bomin, S. Bahuchet, E. Heyer and F. Austerlitz (2013). 'Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies'. In: *Molecular Biology and Evolution* 30.4, pp. 918–937.

Verdu, Paul, Frederic Austerlitz, Arnaud Estoup, Renaud Vitalis, Myriam Georges, Sylvain Théry, Alain Froment, Sylvie Le Bomin, Antoine Gessain, Jean-Marie Hombert et al. (2009). 'Origins and genetic diversity of Pygmy hunter-gatherers from Western Central Africa'. In: *Current Biology* 19.4, pp. 312–318.

Vitti, J. J., S. R. Grossman and P. C. Sabeti (2013). 'Detecting natural selection in genomic data'. In: *Annual Review of Genetics* 47, pp. 97–120.

Vo, B. N., C. C. Drovandi, A. N. Pettitt and G. J. Pettet (2015). 'Melanoma cell colony expansion parameters revealed by approximate Bayesian computation'. In: *PLOS computational biology* 11.12.

Wager, S. and S. Athey (2018). 'Estimation and inference of heterogeneous treatment effects using random forests'. In: *Journal of the American Statistical Association* 113.523.

Waloff, Z. (1966). *The Upsurges and Recessions of the Desert Locust Plague: An Historical Survey*. Anti-Locust Research Centre, London.

Waloff, Z. and D. E. Pedgley (1986). 'Comparative biogeography and biology of the South American locust, *Schistocerca cancellata* (Seville), and the South African desert locust, *S. gregaria flaviventris* (Burmeister) (Orthoptera: Acrididae): a review'. In: *Bulletin of Entomological Research* 8.

Wegmann, D., C. Leuenberger and L. Excoffier (2009). 'Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood'. In: *Genetics* 182.4, pp. 1207–1218.

Wei, G. C. G. and M. A. Tanner (1990). 'A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms'. In: *Journal of the American Statistical Association* 85.411, pp. 699–704.

Weir, B. and C. Cockerham (1984). 'Estimating F-statistics for the analysis of population structure'. In: *Evolution* 38.6, pp. 1358–1370.

Weiss, G. and A. von Haeseler (1998). 'Inference of population history using a likelihood approach'. In: *Genetics* 149.3, pp. 1539–1546.

Wilkinson, R. (2013). 'Approximate Bayesian computation (ABC) gives exact results under the assumption of model error'. In: *Statistical Applications in Genetics and Molecular Biology* 12.2, pp. 129–141.

Wilson, I. J. and D. J. Balding (1998). 'Genealogical inference from microsatellite data'. In: *Genetics* 150.1, pp. 499–510.

Wood, S. N. (2010). 'Statistical inference for noisy nonlinear ecological dynamic systems'. In: *Nature* 466, pp. 1102–1104.

Wright, M. N. and A. Ziegler (2017). 'ranger: a fast implementation of random forests for high dimensional data in C++ and R'. In: *Journal of Statistical Software* 77.1, pp. 1–17.

Wright, S. (1931). 'Evolution in Mendelian populations'. In: *Genetics* 16, pp. 97–159.

Xu, R., D. Nettleton and D. J. Nordman (2016). 'Case-specific random forests'. In: *Journal of Computational and Graphical Statistics* 25.1, pp. 49–65.

Zhang, L., Y. Ren and P. N. Suganthan (2013). 'Instance based random forest with rotated feature space'. In: *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, pp. 31–35.

Résumé. Dans un processus d'inférence statistique, lorsque le calcul de la fonction de vraisemblance associée aux données observées n'est pas possible, il est nécessaire de recourir à des approximations. C'est un cas que l'on rencontre très fréquemment dans certains champs d'application, notamment pour des modèles de génétique des populations. Face à cette difficulté, nous nous intéressons aux méthodes de calcul bayésien approché (ABC, *Approximate Bayesian Computation*) qui se basent uniquement sur la simulation de données, qui sont ensuite résumées et comparées aux données observées. Ces comparaisons nécessitent le choix judicieux d'une distance, d'un seuil de similarité et d'un ensemble de résumés statistiques pertinents et de faible dimension.
Dans un contexte d'inférence de paramètres, nous proposons une approche mêlant des simulations ABC et les méthodes d'apprentissage automatique que sont les forêts aléatoires. Nous utilisons diverses stratégies pour approximer des quantités *a posteriori* d'intérêts sur les paramètres. Notre proposition permet d'éviter les problèmes de réglage liés à l'ABC, tout en fournissant de bons résultats ainsi que des outils d'interprétation pour les praticiens. Nous introduisons de plus des mesures d'erreurs de prédiction *a posteriori* (c'est-à-dire conditionnellement à la donnée observée d'intérêt) calculées grâce aux forêts. Pour des problèmes de choix de modèles, nous présentons une stratégie basée sur des groupements de modèles qui permet, en génétique des populations, de déterminer dans un scénario évolutif les évènements plus ou moins bien identifiés le constituant. Toutes ces approches sont implémentées dans la bibliothèque R `abcrf`. Par ailleurs, nous explorons des manières de construire des forêts aléatoires dites locales, qui prennent en compte l'observation à prédire lors de leur phase d'entraînement pour fournir une meilleure prédiction. Enfin, nous présentons deux études de cas ayant bénéficié de nos développements, portant sur la reconstruction de l'histoire évolutive de population pygmées, ainsi que de deux sous-espèces du criquet pèlerin *Schistocerca gregaria*.

Mots clés. Calcul bayésien approché, forêts aléatoires, inférence bayésienne, génétique des populations, méthodes locales.

Abstract. In a statistical inferential process, when the calculation of the likelihood function is not possible, approximations need to be used. This is a fairly common case in some application fields, especially for population genetics models. Toward this issue, we are interested in approximate Bayesian computation (ABC) methods. These are solely based on simulated data, which are then summarised and compared to the observed ones. The comparisons are performed depending on a distance, a similarity threshold and a set of low dimensional summary statistics, which must be carefully chosen.
In a parameter inference framework, we propose an approach combining ABC simulations and the random forest machine learning algorithm. We use different strategies depending on the parameter posterior quantity we would like to approximate. Our proposal avoids the usual ABC difficulties in terms of tuning, while providing good results and interpretation tools for practitioners. In addition, we introduce posterior measures of error (i.e., conditionally on the observed data of interest) computed by means of forests. In a model choice setting, we present a strategy based on groups of models to determine, in population genetics, which events of an evolutionary scenario are more or less well identified. All these approaches are implemented in the R package `abcrf`. In addition, we investigate how to build local random forests, taking into account the observation to predict during their learning phase to improve the prediction accuracy. Finally, using our previous developments, we present two case studies dealing with the reconstruction of the evolutionary history of Pygmy populations, as well as of two subspecies of the desert locust *Schistocerca gregaria*.

Keywords. Approximate Bayesian computation, random forests, Bayesian inference, population genetics, local methods.