Statistique bayésienne et applications en génétique des populations

Michael G B Blum

Université de Grenoble

Statistique bayésienne

et applications

en génétique des populations

Mémoire présenté pour l'obtention du diplôme
d'**H**abilitation à **D**iriger les **R**echerches

par **Michael Blum**

Soutenance le lundi 3 décembre 2012, à Grenoble,
devant le jury composé de

Florence Forbes (DR INRIA, Grenoble), président du jury
Oscar Gaggiotti (Professeur UJF, Grenoble), membre du jury
Manolo Gouy (DR CNRS, Lyon), membre du jury
Didier Piau (Professeur UJF, Grenoble), membre du jury
Amaury Lambert (Professeur UPMC, Paris), rapporteur
Jean-Michel Marin (Professeur Montpellier 2, Montpellier), rapporteur
Xavier Vekemans (Professeur Lille 1, Lille), rapporteur

Année : 2012

Résumé

Résumé

Les approches statistiques en génétique des populations visent deux objectifs distincts qui sont la description des données et la possibilité d'inférer les processus évolutifs qui ont généré les patrons observés. Le premier chapitre de ce manuscrit décrit nos apports théoriques et méthodologiques concernant le calcul bayésien approché (« Approximate Bayesian Computation ») qui permet de réaliser l'objectif d'inférence des processus évolutifs. Je décris des résultats asymptotiques qui permettent de décrire des propriétés statistiques du calcul bayésien approché. Ces résultats mettent en évidence à la fois l'intérêt des méthodes dites « avec ajustement »qui reposent sur des équations de régression et aussi l'intérêt de réduire la dimension des descripteurs statistiques utilisés dans le calcul bayésien approché. Je présente ensuite une méthode originale de calcul bayésien approché qui permet de manière conjointe d'effectuer des ajustements et de réduire la dimension des descripteurs statistiques. Une comparaison des différentes méthodes de réduction de dimension clos le premier chapitre. Le deuxième chapitre est consacré à l'objectif de description des données et se place plus particulièrement dans un cadre spatial. Les méthodes statistiques proposées reposent sur le concept d'isolement par la distance qui est une forme particulière de l'autocorrélation spatiale où la corrélation entre individus décroit avec la distance. Une approche originale de krigeage nous permet de caractériser des patrons d'isolement par la distance non-stationnaire où la manière avec laquelle la corrélation entre individus décroit avec la distance dépend de l'espace. Une deuxième extension que nous proposons est celle d'isolement par la distance anisotrope que nous caractérisons et testons à partir d'une équation de régression. La conclusion de ce manuscrit met l'accent sur les problèmes d'interprétation des résultats statistiques, l'importance de l'échantillonnage et la nécessité de tester l'adéquation des modèles aux données. Je conclus par des perspectives qui se proposent de faire passer l'analyse statistique bayésienne à l'échelle des données massives produites en génétique.

Mots-clefs

Statistique bayésienne, génétique des populations, calcul bayésien approché, coalescent, processus stochastiques en biologie, krigeage

Abstract

Statistical approaches in population genetics have two distinct objectives, which consist of describing the data and of inferring the evolutionary processes that generated the observed patterns. The first chapter of this thesis describes my contributions to Approximate Bayesian Computation (ABC), which allows to compare and to infer the evolutionary processes that shaped genetic variation. First, I describe asymptotic results, which provide biases and variances of posterior estimates obtained with approximate Bayesian Computation. The results highlight what are the benefits of using regression-adjustment methods and of reducing the dimension of the descriptive statistics used in ABC. Then, I present an original method for ABC that both performs regression-adjustment and dimension reduction. An analysis where we compare different methods of dimension reduction ends the first chapter. The second chapter of the thesis is devoted to the goal of describing the data in a spatial context. The statistical methods we propose are based on the concept of isolation by distance (IBD), which is a particular form of spatial autocorrelation where correlation decays with distance. With a Kriging approach, we can characterize non-stationary patterns of isolation by distance where the decay of correlation with distance varies over the sampling range. We also propose an anisotropic extension of the concept of isolation by distance and we provide a characterization and a test for anisotropy using a regression equation. The conclusion of this thesis deals with some important caveats: the difficulty of interpreting statistical results, the robustness of the results with respect to the sampling scheme and the too often neglected goodness-of-fit. The thesis ends with some perspectives about how Bayesian methods could scale with the massive dimension of the data produced in genetics.

Keywords

Bayesian statistics, population genetics, approximate Bayesian Computation, coalescent, stochastic processes in biology, kriging

Table des matières

Introduction

> The average human has one breast and one testicle.
>
> Des McHale

Les populations d'organismes vivants ont des histoires démographiques complexes : leurs tailles et leurs aires de répartition changent au fil du temps, conduisant à des processus de fission et de fusion qui laissent des signatures sur le patrimoine génétique de ces populations. Un des objectifs de la biologie évolutive est de reconstruire cette histoire démographique en utilisant des données moléculaires obtenues chez les individus des populations concernées.

Cette « quête du passé » a été nourrie par les progrès de la biologie moléculaire et l'explosion de la dimension des données moléculaires. Les puces à SNPs sont un exemple de ces données moléculaires modernes qui permettent de génotyper un individu pour des dizaines de milliers voire des millions de paires de base. L'explosion des données a été accompagnée d'un développement de méthodes statistiques de plus en plus sophistiquées.

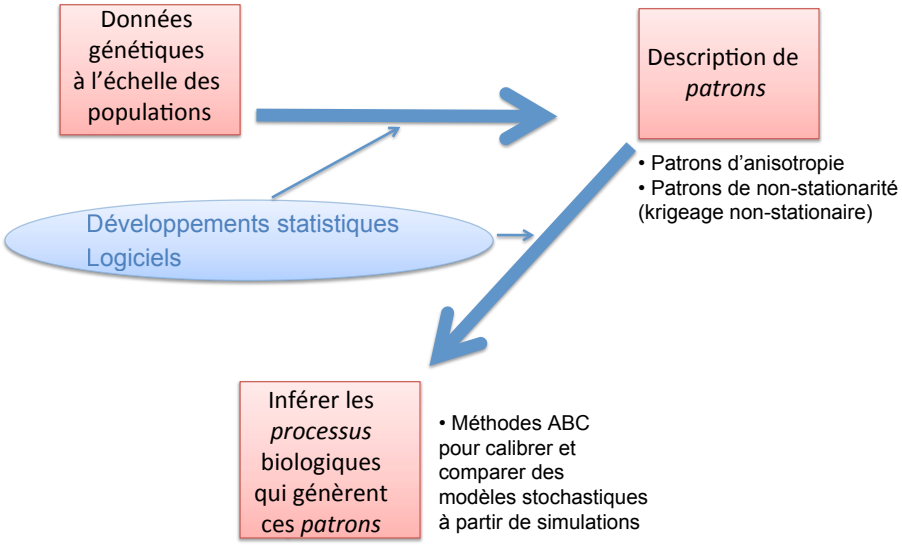

FIGURE 1 – Apport des statistiques computationnelles pour analyser des données moléculaires à l'échelle des populations. Pour chacun des objectifs (description des patrons et inférence des processus biologiques), je précise quelles sont mes contributions que je présente dans le manuscrit.

Ces approches statistiques réalisent 2 objectifs distincts (figure 1). Le premier objectif concerne la description des données qui sont de grande dimension et qu'il faut donc pouvoir résumer par un certain nombre de descripteurs statistiques. On cherchera à isoler des

patrons dans les données, par exemple à regrouper les populations entre elles (*clustering*). C'est cette approche de *statistique descriptive* qui a prévalu en génétique des populations. Traditionnellement, l'unité de travail était celle de la population et l'analyse statistique permettait de regrouper les populations qui étaient similaires en termes de fréquences d'allèles. Comme exemples emblématiques de méthodes de statistique descriptive, on peut citer la méthode *neighbor joining* (Saitou & Nei, 1987), ainsi que le positionnement multidimensionnel beaucoup plus connu sous le nom anglais de *muldimensional scaling* (MDS). Ces méthodes permettent de visualiser soit sous la forme d'un graphe (*neighbor joining*) soit dans un espace en général à deux dimensions (MDS) quels sont les groupes de populations similaires (voir figure 2 pour une illustration en génétique humaine). Bien qu'encore couramment utilisées, ces méthodes ne renseignent pas de manière explicite sur les processus qui ont conduit aux similarités et dissimilarités entre populations. Ces processus comprennent typiquement des événements de divergence ou de séparation entre populations (fission) ainsi que des évènements de métissage (fusion).

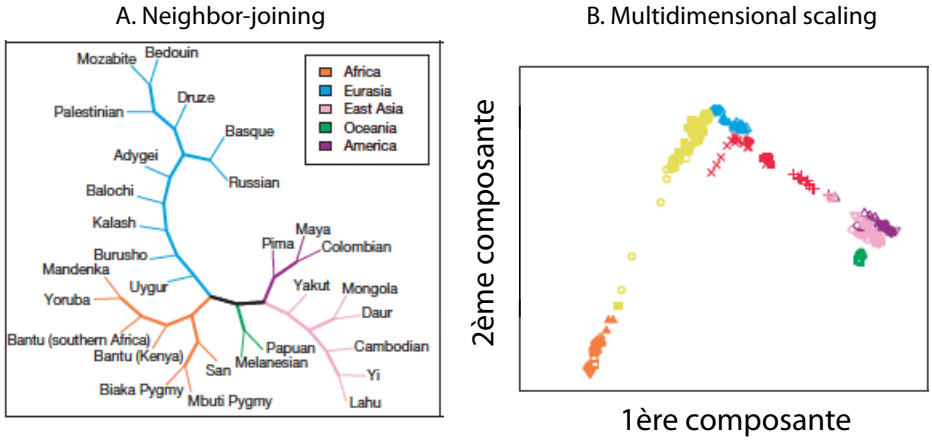

FIGURE 2 – Exemple de l'utilisation de la statistique descriptive en génétique humaine : *neighbor joining* (NJ) et méthode de réduction de dimension *multidimensional scaling* (MDS). Les données sont issues de puces à SNPs et chaque point (MDS) ou chaque feuille (NJ) correspond à une populations pour laquelle une vingtaine d'individus ont été génotypés. Tirée de (Jakobsson et al., 2008).

Le deuxième objectif se propose justement de calibrer et de comparer les différents processus évolutifs (divergence entre populations, métissage, expansion spatiale) qui ont pu générer les patrons observés. Cette étape se fait souvent en utilisant la méthode appelée Approximate Bayesian Computation (Beaumont et al., 2002; Csilléry et al., 2010; Beaumont, 2010; Marin et al., 2011). La méthode ABC est particulièrement intéressante pour des modèles statistiques où il est impossible de calculer la vraisemblance mais où il est facile de simuler des données. Ce n'est pas un hasard si la méthode ABC est apparue en génétique des populations, discipline qui a une longue tradition de simulations numériques (Cavalli-Sforza & Zei, 1967) et qui a connu plus récemment un fort développement du nombre de logiciels dédiés à la simulation (Hoban et al., 2012).

Dans ce manuscrit, je présente mes contributions statistiques qui concernent ces deux objectifs statistiques : description de patrons (chapitre 2) et inférence des processus évolutifs (chapitre 1). Dans le chapitre 1, je présente mes travaux qui concernent la méthode « Approximate Bayesian Computation ». Ces travaux sont de nature théorique puisqu'ils concernent la convergence des estimateurs des lois a posteriori pour la méthode ABC et ils

sont aussi de nature méthodologique avec pour objectif de proposer des méthodes plus performantes dans un contexte de grande dimension. Dans le deuxième chapitre, je présente certaines de mes contributions statistiques qui visent à décrire les patrons de structuration génétique des populations dans un contexte spatial.

Chapitre 1

Calcul bayésien approché ou « Approximate Bayesian Computation »

Simuler ou exagérer

www.femina.fr

Dans ce chapitre, je présente quelles sont mes contributions statistiques concernant la méthode « Approximate Bayesian Computation »(ABC). Afin de présenter le contexte d'utilisation des méthodes ABC, je commence le chapitre par un exemple d'application des méthodes ABC. Cet exemple de génétique humaine montre le type de modèles que l'on cherche à calibrer et à comparer avec les algorithmes ABC (Blum & Jakobsson, 2011). Après cet exemple introductif, je décris des résultats mathématiques qui décrivent les propriétés statistiques des méthodes ABC (Blum, 2010). Ces résultats quantifient le biais et la variance des estimateurs de la loi a posteriori de manière asymptotique. Les résultats mettent en évidence deux aspects : l'intérêt de réduire la dimension des descripteurs statistiques utilisés dans l'ABC et l'apport des méthodes avec correction qui reposent sur des modèles de régression. Les méthodes avec correction (on dit aussi « avec ajustement ») ont été initialement proposées par Beaumont et al. (2002) et nous avons proposé une extension de ces méthodes, que je présente ici, dans un cadre plus flexible de non linéarité et d'hétéroscédasticité (Blum & François, 2010). Un avantage de l'ajustement non-linéaire et héteroscédastique, plus flexible, est que l'échantillonnage de la loi a posteriori devient moins sensible au choix du pourcentage de simulations acceptées et nous illustrons cette propriété pour un modèle de coalescent. Je conclus ce chapitre par une analyse de simulations où nous avons comparé les propriétés de différentes méthodes de réduction de dimension proposées pour le calcul bayésien approché (Blum et al., 2012).

1.1 De la génétique des populations aux modèles statistiques implicites

1.1.1 Un exemple introductif

Pour illustrer l'intérêt de la méthode « Approximate Bayesian Computation », je donne comme exemple l'analyse de différents modèles d'évolution humaine que j'ai effectuée avec

Mattias Jakobsson (Blum & Jakobsson, 2011). A partir de données de séquences d'ADN (données de résequencage de 20kbp pour 20 loci), l'objectif était de discriminer entre différents modèles d'évolution humaine (*model selection*), estimer les paramètres des ou du modèle sélectionné (*parameter inference*), et vérifier que le modèle sélectionné est en adéquation raisonnable avec les données (*Goodness-of-fit*). Les différents modèles envisagés font écho au résultat retentissant obtenu avec le séquençage du Néandertal (Green et al., 2010). La figure 1.1 montre deux modèles qui correspondent soit au scénario sans métissage entre Néandertal et les hommes modernes soit au scénario avec métissage. L'aspect clé pour l'apprentissage statistique provient du fait qu'il existe des logiciels, comme *ms* (Hudson, 2002), qui permettent de simuler des données de séquence d'ADN dans chacun de ces modèles. Ces simulateurs reposent sur le modèle de la coalescence, un processus stochastique qui permet de décrire la généalogie d'un échantillon au sein d'une population (Tavaré, 2004). Je ne donnerai pas les détails de la théorie de la coalescence, mais le calcul de la vraisemblance est en général très difficile dans ces modèles ce qui a motivé l'utilisation de simulations numériques pour faire de l'inférence statistique.

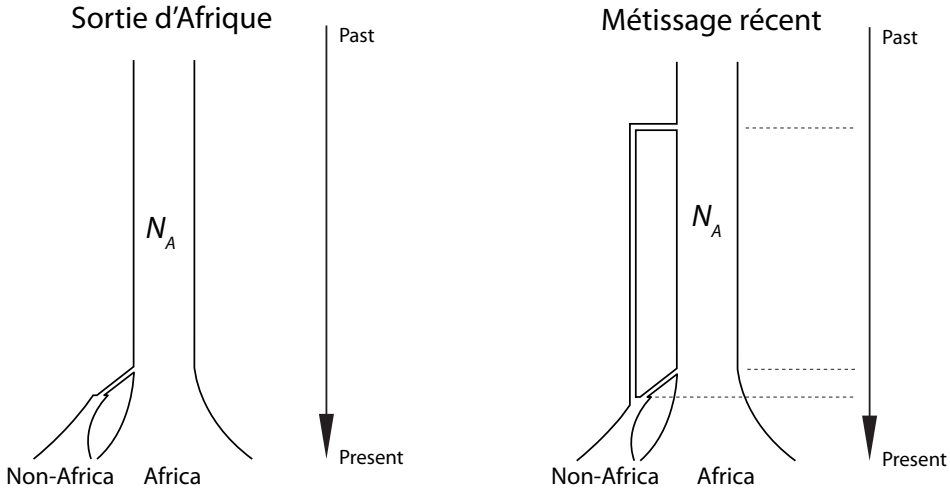

FIGURE 1.1 – Deux modèles de coalescence qui sont comparés et calibrés avec une méthode ABC par Blum & Jakobsson (2011). Dans ces modèles de coalescence (Tavaré, 2004), il est difficile de calculer la vraisemblance des paramètres mais la simulation de séquences d'ADN se fait d'autant plus facilement qu'il existe des logiciels dédiés (Hudson, 2002). Le premier modèle dit de « la sortie d'Afrique »suppose que les populations non africaines et africaines sont toutes issues de la même population ancestrale africaine qui comptait N_A individus efficaces (un des paramètres du modèle). Il y a un certain temps dans le passé, ces deux populations se sont séparées lors d'un événement de fission. Dans le deuxième modèle avec métissage récent, on suppose que les populations non africaines se sont métissées avec une population d'hommes archaïques (Néandertal) dont la divergence avec la lignée africaine est très ancienne. Ce métissage qui se fait uniquement dans la lignée non africaine correspond à un événement de fusion.

1.1.2 Modèles statistiques implicites et critiques de ces modèles

Dans la littérature statistique, ces modèles stochastiques pour lesquels on ne peut pas calculer la vraisemblance mais qu'il est facile de simuler ont été appelés *implicit statistical*

models par Diggle & Gratton (1984). Dans la discussion qui fait suite à leur papier, le fait de faire de l'inférence statistique pour ce type de modèle a été critiqué. L'argument principal étant que le rôle de l'inférence statistique n'était pas de calibrer des modèles *mécanistiques* qui sont censés mimer, forcément mal, le processus qui ont engendrés les observations. Les auteurs de la critique suggéraient que les modèles statistiques devaient au contraire être suffisamment simples pour que la vraisemblance soit calculable et ces modèles devaient par ailleurs répondre à des objectifs assez techniques comme réduire la dimension des données (analyse factorielle, analyse en composantes principales) et non pas donner la fausse impression que l'on puisse retrouver les processus qui ont engendrés les données à partir d'inférence statistique. Plus récemment, la méthode ABC a elle aussi reçu de nombreuses critiques, en particulier les algorithmes ABC qui visent à faire de la sélection de modèles. En revanche les critiques furent de natures différentes puisqu'elles ont porté soit sur la méthode bayésienne en général (Templeton, 2010) soit sur le biais que peut engendrer la perte d'information liée à l'utilisation d'un certain nombre de descripteurs statistiques en lieu et place de l'ensemble des données (Robert et al., 2011).

1.2 Méthodes avec ajustement et théorie

Les résultats théoriques que je présente concernent exclusivement l'inférence des paramètres. L'inférence bayésienne repose sur la loi a posteriori définie par

$$p(\boldsymbol{\Theta}|D) = \frac{p(D|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})}{p(D)} \tag{1.1}$$

où $\boldsymbol{\Theta} \in \mathbb{R}^p$ représente le vecteur des paramètres, et D représente les données. L'expression donnée dans l'équation (1.1) dépend de la loi a priori $\pi(\boldsymbol{\Theta})$, de la fonction de vraisemblance $p(D|\boldsymbol{\Theta})$, et de la probabilité des données $p(D) = \int_{\boldsymbol{\Theta}} p(D|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})\, d\boldsymbol{\Theta}$. Dans le contexte des méthodes ABC, l'inférence ne repose plus sur la loi a posteriori $p(\boldsymbol{\Theta}|D)$ mais sur une loi a posteriori partielle $p(\boldsymbol{\Theta}|\mathbf{s}_{obs})$ où $\mathbf{s}_{obs}$ représente un vecteur de dimension d de statistiques descriptives. La loi a posteriori partielle est définie de la manière suivante (Doksum & Lo, 1990)

$$p(\boldsymbol{\Theta}|\mathbf{s}_{obs}) = \frac{p(\mathbf{s}_{obs}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})}{p(\mathbf{s}_{obs})}. \tag{1.2}$$

Bien évidemment, la loi a posteriori partielle est égale à la loi a posteriori si les statistiques descriptives $\mathbf{s}_{obs}$ sont suffisantes pour le paramètre $\boldsymbol{\Theta}$.

1.2.1 L'algorithme de rejet

Pour simuler un échantillon suivant la loi a posteriori partielle $p(\boldsymbol{\Theta}|\mathbf{s}_{obs})$, l'algorithme de rejet fonctionne de la manière suivante

1. Simuler n valeurs $\boldsymbol{\Theta}_i$, $i = 1, \ldots, n$, suivant la loi a priori π.
2. Simuler les statistiques descriptives $\mathbf{s}_i$ suivant le modèle génératif $p(\mathbf{s}_i|\boldsymbol{\Theta}_i)$.
3. Associer à chaque couple $(\boldsymbol{\Theta}_i, \mathbf{s}_i)$ un poids $W_i \propto K(\|\mathbf{s}_i - \mathbf{s}_{obs})\|/b)$ où $\|\cdot - \cdot\|$ est une distance, K est un noyau statistique unidimensionnel et b est un seuil d'acceptation fixé.

Pour simplifier l'écriture des estimateurs, on supposera par la suite que les poids ont été renormalisés de sorte à ce que $\sum W_i = 1$. Ce qui peut sembler curieux à première vue, c'est l'absence de rejet dans l'algorithme de rejet ! C'est bien évidemment dans la troisième

étape de l'algorithme qu'a lieu le rejet. On choisit en général un noyau K à support dans $[-1, 1]$ de telle sorte que toutes les simulations pour lesquelles $\|\mathbf{s}_i - \mathbf{s}_{obs}\| > b$ soient rejetées. Dans les cas suffisamment simples où le choix $b = 0$ n'entrainent pas le rejet de toutes les simulations, alors l'algorithme de rejet produit exactement des échantillons suivant la loi conditionnelle $p(\mathbf{\Theta}|\mathbf{s}_{obs})$ (Rubin, 1984). En plus de l'algorithme de rejet, il existe aussi des algorithmes séquentiels où au lieu de ne faire qu'une passe de simulations en simulant toujours les paramètres $\mathbf{\Theta}$ suivant la loi a priori, on effectue plusieurs passes de simulations (Sisson et al., 2007; Beaumont et al., 2009; Del Moral et al., 2012).

Le premier choix à faire dans l'algorithme présenté concerne le choix du noyau K. Les choix habituels pour K sont le noyau uniforme qui revient à donner un poids de 1 à toutes les simulations acceptées (Pritchard et al., 1999) ainsi que le noyau d'Epanechnikov (Beaumont et al., 2002). Le deuxième choix concerne le seuil b. Je note ce seuil b parce que ce paramètre correspond à la taille de la fenêtre au sein de laquelle on accepte les simulations et ce paramètre se dit *bandwidth* en anglais. Pour établir le théorème de la section suivante, je suppose que b a été choisi dans un premier temps sans prendre en comptes les simulations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Cette hypothèse technique ne correspond cependant pas à ce qui se fait dans la pratique où l'on prend b égale à un quantile des distances $\|\mathbf{s}_i - \mathbf{s}_{obs}\|$. On choisit par exemple le premier percentile des distances $\|\mathbf{s}_i - \mathbf{s}_{obs}\|$ de sorte à n'accepter que 1% des simulations. Des résultats théoriques dans la situation où b dépend des simulations ont récemment été établis (Biau et al., 2012). Par souci de simplicité et afin d'alléger les notations, on suppose dorénavant que le paramètre d'intérêt est unidimensionnel et on le note θ.

Après l'utilisation d'une méthode d'échantillonnage bayésien, on reporte les résultats en reportant typiquement la moyenne empirique des échantillons ainsi que les intervalles de crédibilité à 95%. Si l'on s'intéresse à l'ensemble de la loi a posteriori, on peut utiliser une méthode à noyau pour estimer la loi a posteriori à partir de l'échantillon pondéré (θ_i, W_i)

$$\hat{p}_0(\theta|\mathbf{s}_{obs}) = \sum_{i=1}^{n} \tilde{K}_{b'}(\theta_i - \theta) W_i, \tag{1.3}$$

où la taille de la fenêtre correspondant à $\tilde{K}$ est noté b' ($b' > 0$) et nous utilisons la notation $\tilde{K}_{b'}(\cdot) = \tilde{K}(\cdot/b')/b'$ avec $\tilde{K}$ noyau statistique en dimension 1. Dans la section 1.2.3, nous allons étudier les propriétés statistiques de cet estimateur de la loi a posteriori.

1.2.2 Les méthodes avec ajustement

Le principe de ces méthodes est d'ajuster les valeurs des paramètres simulés θ_i qui ont un poids W_i non nul de sorte à prendre en compte la différence qui existe entre les statistiques simulées $\mathbf{s}_i$ et les statistiques observées $\mathbf{s}_{obs}$ (Beaumont et al., 2002). Le principe est de calibrer un modèle de régression dans le voisinage de $\mathbf{s}_{obs}$

$$\theta = m(\mathbf{s}) + \varepsilon \tag{1.4}$$

où $m(\mathbf{s})$ représente l'espérance conditionnelle de θ et ε un résidu. Si l'on fait l'hypothèse d'*homoscedasticité*, c'est à dire si l'on suppose que la loi des résidus ne dépend pas de $\mathbf{s}$, alors pour produire des échantillons tirés suivant la loi conditionnelle $p(\theta|\mathbf{s}_{obs})$, il suffit d'ajuster les θ_i acceptés de la manière suivante

$$\begin{aligned} \theta_i^* &= \hat{m}(\mathbf{s}_{obs}) + \hat{\varepsilon}_i \\ &= \hat{m}(\mathbf{s}_{obs}) + (\mathbf{s}_i - \hat{m}(\mathbf{s}_i)), \end{aligned} \tag{1.5}$$

où $\hat{m}$ représente l'estimateur de l'espérance conditionnelle obtenu en général par moindre carrés pondérés (Beaumont et al., 2002). Une illustration graphique de l'ajustement est donnée dans la figure 1.2. Après ajustement, l'estimateur de la loi a posteriori est obtenu en remplaçant les θ_i par les θ_i* dans l'équation (1.3)

$$\hat{p}_j(\theta|\mathbf{s}_{obs}) = \sum_{i=1}^{n} \tilde{K}_{b'}(\theta_i^* - \theta)W_i,\ j = 1, 2 \tag{1.6}$$

où $j = 1$ quand l'ajustement (i.e. la fonction m) est linéaire et $j = 2$ quand l'ajustement est quadratique. Dans la littérature de statistique non-paramétrique, des estimateurs avec des ajustements similaires à celui de l'équation (1.5) ont été proposés pour faire de l'estimation de densité conditionnelle (Hyndman et al., 1996; Hansen, 2004).

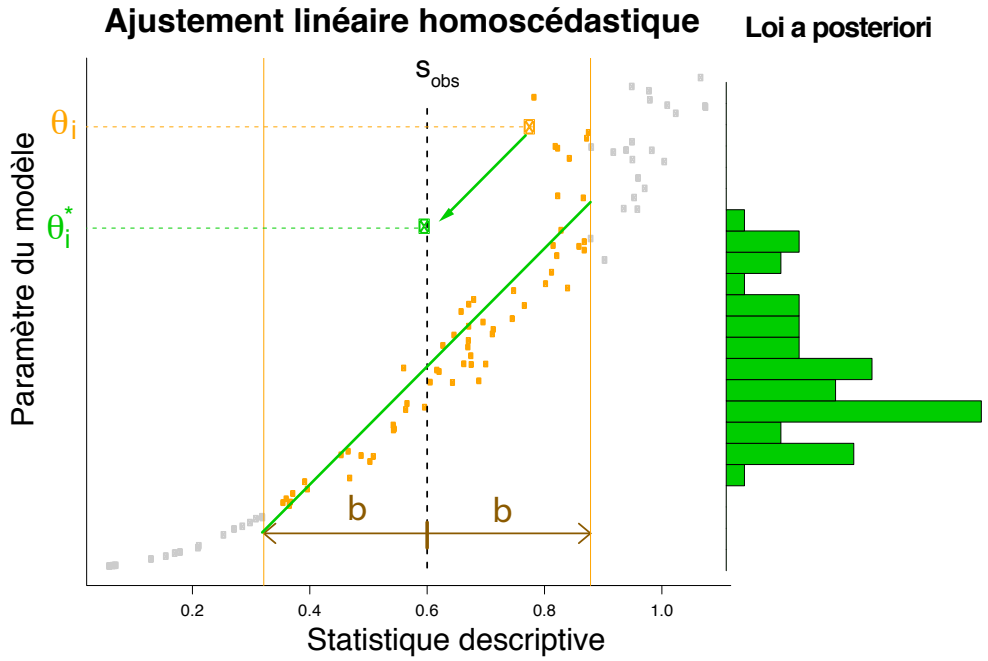

Figure 1.2 – Méthode d'ajustement linéaire et homoscédastique (Beaumont et al., 2002). Le paramètre b mesure la taille de la fenêtre au sein de laquelle les simulations sont acceptées. Les paramètres θ_i représentent les valeurs des paramètres issus de l'algorithme de rejet et les θ_i^* correspondent aux valeurs des paramètres après ajustement (équation (1.5)). Cette figure est tirée du papier de revue de Csilléry et al. (2010).

1.2.3 Biais et variance des estimateurs

Dans cette section, nous donnons le théorème principal qui décrit les propriétés statistiques des estimateurs de la loi a posteriori obtenus par la méthode de rejet (équation (1.3)) et par les méthodes avec ajustement (équation (1.6)). Pour énoncer le théorème, nous introduisons les notations suivantes : si X_n est une suite de variables aléatoires et a_n est une suite déterministe, la notation $X_n = o_P(a_n)$ signifie que X_n/a_n converge vers zéro en probabilité et $X_n = O_P(a_n)$ signifie que le ratio X_n/a_n reste borné à la limite en probabilité. Les hypothèses assez techniques du théorème sont données en appendice de (Blum, 2010).

Théorème 1. On suppose que les conditions (A1)-(A5) de l'appendice de (Blum, 2010) sont vérifiées. Le biais et la variance des estimateurs $\hat{p}_j(\theta|\mathbf{s}_{obs})$, $j = 0, 1, 2$ sont donnés par les expressions suivantes

$$E[\hat{p}_j(\theta|\mathbf{s}_{obs}) - p(\theta|\mathbf{s}_{obs})] = C_1 b'^2 + C_{2,j} b^2 + O_P((b^2 + b'^2)^2) + O_P(\frac{1}{nb^d}), \tag{1.7}$$

$$\text{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{C_3}{nb^d b'}(1 + o_P(1)), \tag{1.8}$$

ou d est la dimension du vecteur de statistiques descriptives $\mathbf{s}_{obs}$ et les constantes C_1, $C_{2,j}$ et C_3 sont données par Blum (2010).

Preuve : Voir (Blum, 2010).

Remarque 1. Compromis biais-variance Le biais des estimateurs augmente avec le carré de la taille de la fenêtre b. Pour minimiser le biais, il faut donc prendre la plus petite fenêtre possible. En revanche la variance croit comme $1/(nb^d)$ où d représente le nombre de statistiques descriptives. Ce terme provient du fait que la variance est inversement proportionnelle au nombre de simulations acceptées et que le volume de la sphère au sein de laquelle les simulations sont acceptées est asymptotiquement proportionnel à b^d. Pour minimiser le terme de variance, il faut donc chercher à augmenter b, i.e. à augmenter le nombre de simulations acceptées. L'erreur que l'on cherche à minimiser est l'erreur quadratique moyenne qui est égale à la somme du carré du biais et de la variance. Pour minimiser cette erreur, il faut donc réaliser un compromis entre le terme de biais et le terme de variance.

Remarque 2. Fléau de la dimension Avec de l'algèbre élémentaire, on peut montrer que pour les 3 estimateurs, l'erreur quadratique moyenne est de l'ordre de $n^{-1/(d+5)}$ quand les tailles de fenêtre sont choisies de manière optimale. La vitesse avec laquelle l'erreur tend vers 0 diminue donc de manière drastique quand la dimension des statistiques descriptives augmente. Cet théorème met donc en évidence (de manière compliquée certes) l'importance de réduire la dimension des statistiques (voir section 1.4). Néanmoins, les conclusions issues de ces théorèmes asymptotiques, classiques en statistique non-paramétrique, sont souvent beaucoup plus pessimistes que les résultats observés en pratique ; en particulier parce que ces théorèmes asymptotiques ne prennent pas en compte les corrélations entre les statistiques (Scott, 1992).

Remarque 3. Biais des estimateurs avec et sans ajustement Il ne semble pas possible de donner des inégalités entre les constantes $C_{2,j}$, $j = 0, 1, 2$, qui soient vraies pour n'importe quel modèle statistique. En revanche, si l'on suppose que la loi des résidus ε dans l'équation (1.4) ne dépend pas de $\mathbf{s}$, alors on peut montrer que la constante $C_{2,2}$ est égale à 0. Lorsque l'on fait cette hypothèse dite d'homoscedasticité, l'estimateur qui réalise (de manière asymptotique) la plus petite erreur quadratique moyenne est celui avec ajustement quadratique $\hat{p}_2(\theta|\mathbf{s}_{obs})$. Si l'on suppose en plus que l'espérance conditionnelle m est linéaire en $\mathbf{s}$, alors à la fois $\hat{p}_1(\theta|\mathbf{s}_{obs})$ et $\hat{p}_2(\theta|\mathbf{s}_{obs})$ ont une erreur inférieure à celle de l'estimateur sans ajustement.

Dans la prochaine section, nous présentons une méthode qui permet à la fois de réduire la dimension des statistiques descriptives (cf. remarque 2), et de faire des ajustements non-linéaires et hétéroscédastiques (cf. remarque 3).

1.3 Ajustement non-linéaire et hétéroscédastique

Le principe de l'ajustement hétéroscédastique est de prendre en compte que la variance des résidus ε dans l'équation (1.4) peut dépendre de $\mathbf{s}$. L'équation de régression prend désormais la forme suivante (Blum & François, 2010)

$$\theta = m(\mathbf{s}) + \sigma(\mathbf{s})\zeta, \tag{1.9}$$

où $\sigma(\mathbf{s})$ représente la racine carrée de la variance conditionnelle et ζ est le résidu de l'équation de régression. L'ajustement hétéroscédastique est donnée par (voir aussi figure 1.3)

$$\begin{aligned} \theta_i^{**} &= \hat{m}(\mathbf{s}_{obs}) + \hat{\sigma}(\mathbf{s}_{obs})\hat{\zeta}_i \\ &= \hat{m}(\mathbf{s}_{obs}) + \frac{\hat{\sigma}(\mathbf{s}_{obs})}{\hat{\sigma}(\mathbf{s}_i)}(\theta_i - \hat{m}(\mathbf{s}_i)) \end{aligned} \tag{1.10}$$

où $\hat{m}$ et $\hat{\sigma}$ sont des estimateurs de la moyenne conditionnelle et de l'écart-type conditionnel.

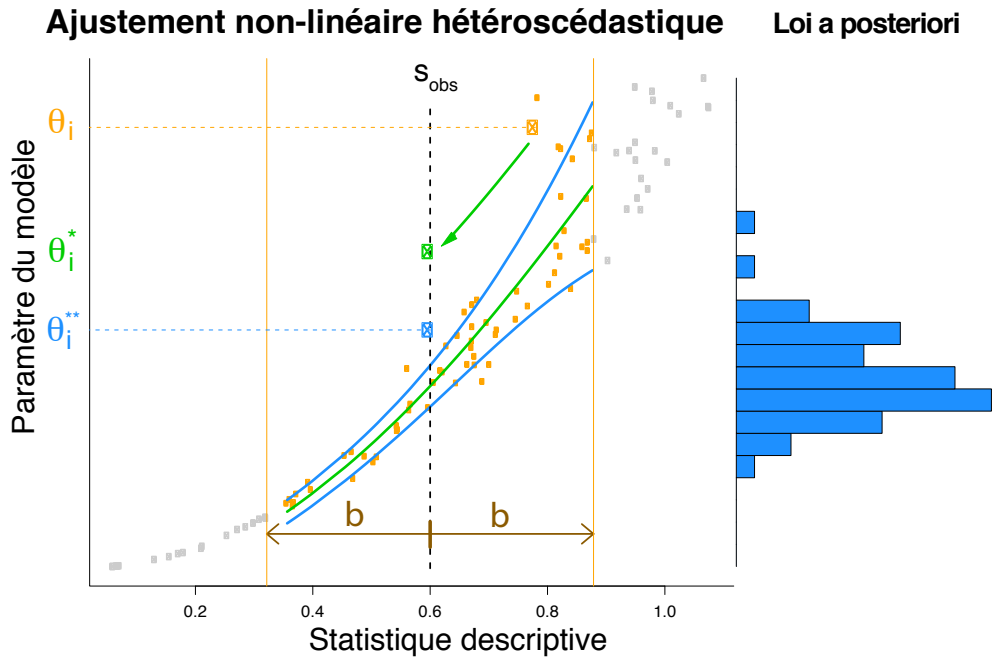

Figure 1.3 – Méthode d'ajustement non-linéaire et hétéroscédastique (Blum & François, 2010). Les paramètres θ_i correspondent aux valeurs issues de l'algorithme de rejet, les θ_i^* (en vert) correspondent aux valeurs des paramètres après l'ajustement non-linéaire pour la moyenne (ajustement homoscédastique de l'équation (1.5)) et les θ_i^{**} (en bleue) correspondent aux valeurs des paramètres après l'ajustement non-linéaire pour la variance (ajustement hétéroscédastique de l'équation (1.10)).

L'ajustement homoscédastique (équation (1.5)) entraine systématiquement un « rétrécissement »(*shrinkage* en anglais) de la loi a posteriori (voir l'exemple de la figure 1.4). Cette propriété provient du fait que la variance empirique des échantillons θ_i^* après ajustement est égale à celle des résidus empirique $\hat{\varepsilon}_i$ qui est forcément inférieure à la variance initiale des θ_i. En revanche l'ajustement hétéroscédastique n'entraîne pas forcément un rétrécissement supplémentaire par rapport à l'ajustement homoscédastique. Par exemple, si

$\hat{\sigma}(\mathbf{s}_{obs}) > \hat{\sigma}(\mathbf{s}_i), i = 1, \dots n$, on aura un élargissement de la loi a posteriori après ajustement hétéroscédastique.

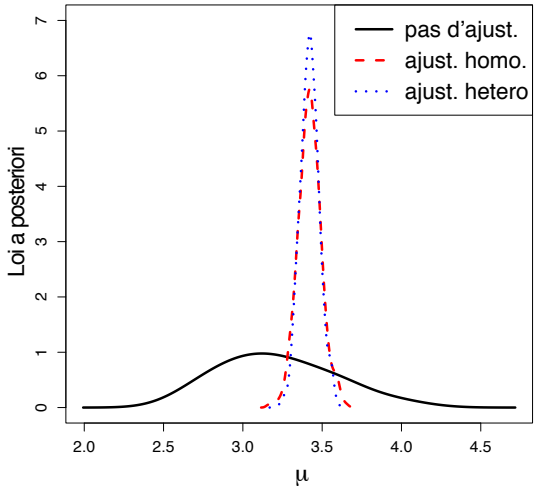

FIGURE 1.4 – Illustration du rétrécissement de la loi a posteriori engendré par les ajustements. Dans cet exemple, l'ajustement hétéroscédastique entraine un rétrécissement supplémentaire mais faible par rapport à l'ajustement homoscédastique. La loi a posteriori correspond à la loi de la moyenne μ d'un échantillon gaussien $\mathcal{N}(\mu, v)$ en prenant comme loi a priori une loi uniforme entre 0 et 10 pour μ et une loi inverse chi-deux à un degré de liberté pour v. Les statistiques observées $\mathbf{s}_{obs}$ valent 3.428 pour la moyenne empirique et 0.1436 pour la variance empirique. L'abréviation ajust. signifie ajustement, homo. signifie homoscédastique et hetero. signifie hétéroscédastique.

1.3.1 Réseaux de neurones

L'équation (1.10) d'ajustement hétéroscédastique dépend du choix de l'estimateur de la moyenne conditionnelle $\hat{m}$ et de l'écart type conditionnel $\hat{\sigma}$. Dans le papier (Blum & François, 2010), nous proposions d'utiliser les réseaux de neurones pour estimer ces deux quantités. Ce choix était motivé par la possibilité offerte par les réseaux de neurones (*feedforward neural network* en anglais) de réduire la dimension des statistiques descriptives via une projection interne sur un espace de dimension plus faible (Ripley, 1994).

En général, les hypothèses de l'ajustement linéaire homoscédastique (équation (1.5)) sont d'autant plus fausses que le pourcentage de simulation acceptées est grand. En revanche, puisque l'ajustement de l'équation (1.10) est plus flexible, en prenant en compte la non-linéarité de l'espérance conditionnelle et l'hétéroscédasticté, les résultats obtenus seront moins sensibles aux choix du pourcentage de simulations acceptées. Dans un modèle de coalescent où l'on cherchait à estimer le taux de mutations nous avons mis en évidence que l'ajustement hétéroscédastique avec réseaux de neurones était en effet moins sensible au choix du pourcentage de simulations acceptées (figure 1.5). Le paquetage `R` *abc* que Katalin Csilléry a développé pendant son postdoc au laboratoire implémente entre autres l'ajustement hétéroscédastique (équation (1.10)) avec réseaux de neurones (Csilléry et al., 2012).

Le choix des réseaux de neurones était motivé par la possibilité de réduire la dimension

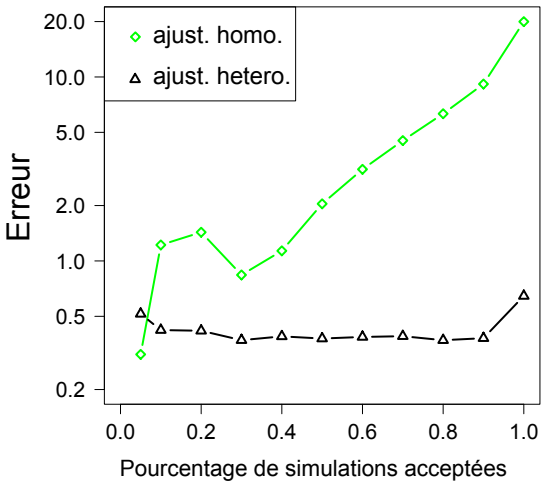

Figure 1.5 – Erreur en fonction du pourcentage de simulations acceptées pour les deux types d'ajustement. Les simulations ont été faites dans un modèle de coalescent et le paramètre d'intérêt est les taux de mutation. Dans cet exemple, comme il est possible de simuler exactement la loi a posteriori (en prenant $b = 0$ dans l'algorithme de rejet), nous utilisons comme erreur la somme pour 5 quantiles différents de la différence en valeur absolue entre le vrai quantile de la loi a posteriori et le quantile estimé (Blum & François, 2010). L'abréviation ajust. signifie ajustement, homo. signifie homoscédastique et hetero. signifie hétéroscédastique.

des statistiques descriptives via la projection sur une couche cachée. D'autres méthodes ont été proposées pour réduire la dimension des statistiques dans le cadre des méthodes bayésiennes approchées et la prochaine section présente une analyse comparative de ces différentes méthodes.

1.4 Analyse comparative des méthodes de réduction de dimension

Comme on l'a vu dans la sous-section 1.2.3, réduire la dimension des statistiques descriptives permet de réduire l'erreur due à l'estimation de la loi a posteriori par une méthode de statistique non paramétrique (équations (1.3) et (1.6)) (Blum, 2010; Fearnhead & Prangle, 2012). En revanche, l'autre approximation qui consiste à remplacer la loi a posteriori $p(\theta|D)$ par la loi a posteriori partielle $p(\theta|\mathbf{s}_{obs})$ va avoir tendance à se dégrader lorsque le nombre de statistiques descriptives diminue. Il existe donc un compromis à effectuer pour réduire de manière optimale la dimension des statistiques descriptives. Dans notre article de revue (Blum et al., 2012), nous avons séparé les méthodes de réduction de dimension en trois grandes classes.

1. Les méthodes qui consistent à choisir le **meilleur sous-ensemble de statistiques descriptives**. Cet ensemble optimal est ensuite utilisé à la fois lors de l'algorithme de rejet et des ajustements basés sur les équations de régression (équations (1.5) et (1.10)). Pour déterminer cet ensemble optimal, il faut effectuer un certain nombre de simulations du couple $(\theta, \mathbf{s})$ suivant le modèle génératif et calculer ensuite un score pour chacun des sous-ensemble de statistiques descriptives. Comme l'espace de ces sous-ensembles est trop vaste à explorer (il y a $2^d - 1$ sous-ensemble de statistiques

descriptives si l'on exclue l'ensemble vide), on fait souvent appel à des algorithmes gloutons pour trouver le meilleur sous-ensemble de statistiques. Les différents scores utilisés dans le cadre de l'ABC sont : des mesures de suffisance du sous-ensemble de statistique (Joyce & Marjoram, 2008), une mesure de l'entropie de la loi a posteriori (Nunes & Balding, 2010) ainsi que des critères du type AIC/BIC (Sedki & Pudlo, 2012).

2. Les **méthodes de projection** qui consistent à projeter l'ensemble de statistiques descriptives dans un espace de plus petite dimension. Ces méthodes comprennent l'approche « partial-least squares regression »(PLS) (Wegmann et al., 2009) et l'approche « posterior loss »(Fearnhead & Prangle, 2012). Dans la première approche, on construit K (K est à déterminer) statistiques descriptives qui sont des combinaisons linéaires des statistiques initiales tandis que dans la seconde approche on utilise une seule combinaison linéaire des statistiques initiales pour chaque paramètre à estimer. Dans l'approche « posterior loss », l'idée est de dire que la statistique optimale (au sens d'une erreur quadratique) est l'espérance conditionnelle $E[\theta|\mathbf{s}_{obs}]$ et que pour estimer cette statistique, on calibre le modèle de régression suivant

$$\theta = \alpha + \sum_{i=1}^{d} \beta_i f(\mathbf{s}^i) + \eta; \tag{1.11}$$

où α et les β_i, $i = 1, \ldots, d$, sont les coefficients de régression, f est une fonction que l'on se donne et $\mathbf{s}^i$ est la $i^{ème}$ composante de la statistique descriptive $\mathbf{s}$. La nouvelle et unique statistique descriptive qui sera utilisée pour estimer θ sera donc $\hat{\alpha} + \sum \hat{\beta}_i f(\mathbf{s}^i)$ avec cette approche. Dans les approches « posterior loss »et « partial-least squares regression », les nouvelles statistiques sont utilisées à la fois lors de l'algorithme de rejet et lors des ajustements de régression. L'approche de régression avec réseaux de neurones vue dans la section 1.3 fait aussi partie de ces méthodes de projection puisque la dimension des statistiques descriptives est réduite via la projection sur la couche cachée du réseau de neurones. En revanche, lorsque l'on utilise cette approche, les poids W_i, $i = 1, \ldots, n$, de l'algorithme de rejet sont calculés avec la totalité des statistiques descriptives initiales. La réduction de dimension n'intervient que dans la partie ajustement avec régression.

3. Les **méthodes de régularisation** cherchent à éviter que l'on fasse des sur-ajustements dans les directions où les statistiques descriptives sont peu corrélées au paramètre d'intérêt. Le principe est d'estimer la moyenne conditionnelle m (équation (1.5)) avec une méthode des moindres carrés régularisés (Hoerl & Kennard, 1970) ou l'on pénalise les coefficients de régression qui ont des valeurs trop importantes. Dans le contexte de l'ABC, la régularisation a été proposée pour faire des ajustements linéaires en utilisant une approche de type « ridge regression »(Hoerl & Kennard, 1970) pour estimer les paramètres de l'équation de régression (1.5) (Blum et al., 2012). L'approche par réseaux de neurones comprend aussi une étape de régularisation puisque les poids du réseau de neurones sont estimés en minimisant un critère de moindres carrés régularisés. Pour toutes ces méthodes, les poids W_i, $i = 1, \ldots, n$, de l'algorithme de rejet sont calculés avec la totalité des statistiques descriptives initiales.

Nous avons considéré 3 exemples différents de modèles stochastiques (coalescent, processus de naissance et de mort, modèle de valeurs extrêmes) qui comprennent respectivement 6, 11 et 113 statistiques descriptives. Pour chaque méthode de réduction de

dimension, nous utilisons l'algorithme de rejet suivi d'un ajustement hétéroscédastique (équation (1.10)).

En comparant l'erreur obtenue avec chacune des méthodes à celle obtenue avec l'algorithme de rejet, nous avons pu comparer les différentes méthodes de dimension de réduction (figure 1.6, voir Blum et al. (2012) pour le détail du calcul de l'erreur). Si l'on moyenne les résultats obtenus, la plus mauvaise méthode est celle qui calcule un score de suffisance (Joyce & Marjoram, 2008) et elle obtient un résultat beaucoup moins bon (baisse de l'erreur de 8% par rapport au rejet) que lorsque l'on utilise l'ensemble des statistiques descriptives (baisse de l'erreur de 17% par rapport au rejet). Les deux meilleures méthodes qui ont un coût de calcul modéré (i.e. parmi les méthodes de projection et de régularisation) sont l'ajustement par réseaux de neurones (baisse de l'erreur de 17% par rapport au rejet) et l'approche « posterior loss »(Fearnhead & Prangle, 2012) (baisse de l'erreur de 25% par rapport au rejet). Cette dernière approche donne des résultats bien meilleurs que les autres dans le troisième exemple où elle implique une réduction de dimension drastique puisque les 113 statistiques initiales sont résumées par une seule et unique combinaison linéaire de statistiques pour chaque paramètre estimé.

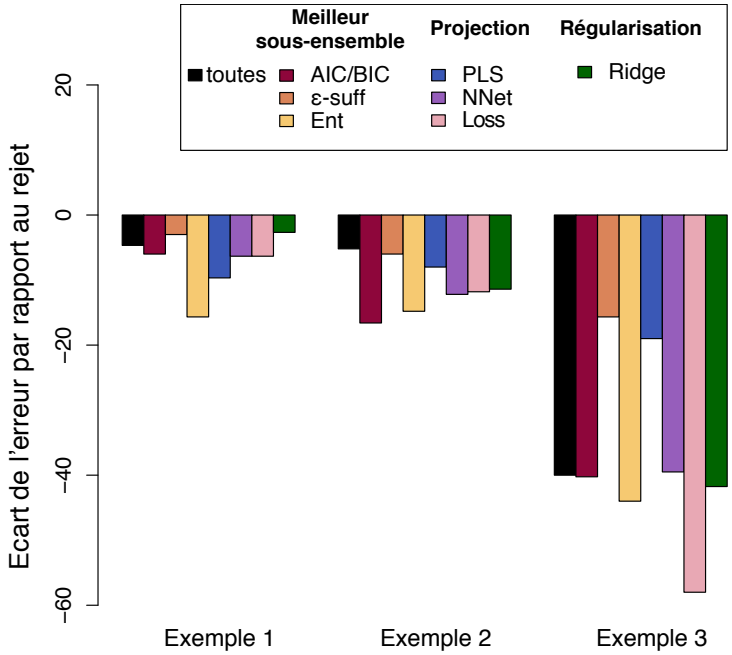

Figure 1.6 – Erreur relative par rapport à l'algorithme de rejet. Les détails du calcul de l'erreur sont donnés dans (Blum et al., 2012). Les méthodes comprennent : celle qui utilisent toutes les statistiques (toutes), AIC/BIC, le critère de suffisance (ε-suff.), le critère d'entropie (Ent), « partial least squares »(PLS), les réseaux de neurones (NNet), « posterior loss »(Loss) et « ridge regression »(Ridge).

Chapitre 2

Méthodes descriptives dans un contexte spatial

Traduttore, traditore

Proverbe italien

Lorsque l'on fait appel aux méthodes ABC vues dans le chapitre 1, il faut déjà pouvoir modéliser quels sont les processus évolutifs qui ont eu lieu. Pour ce faire, il faut déjà avoir une idée même grossière des processus évolutifs en jeu. Une manière de procéder pour cette étape initiale qui précède la modélisation est de recourir à une « observation »préalable des données. Par exemple si deux groupes d'individus émergent lors d'une analyse de classification (*clustering*), on pourra supposer qu'il y a deux populations dans l'échantillon étudié et utiliser un modèle de divergence paramétré par un temps de divergence que l'on cherchera à estimer. Si au contraire, on n'observe pas de groupes bien distincts, il n'est pas forcément opportun de chercher à calibrer un modèle de divergence. Il existe donc souvent une étape d'observation qui précède la modélisation de processus évolutifs complexes et le calibrage de ces modèles.

Toute la difficulté réside dans la possibilité d'observer ou de décrire les données. En fait, vu la dimension des données, plus d'un million de marqueurs moléculaires pour les puces à SNPs récentes, il n'est pas possible d'observer directement les données et l'on fait appel à des techniques de statistiques descriptives. Même pour des données de plus petite dimension, par exemple avec une dizaine de marqueurs microsatellites, on cherchera à résumer l'information de manière suffisamment synthétique. Cependant, décrire c'est traduire et comme traduire c'est trahir, on ne peut appréhender les données que via le prisme forcément imparfait de méthodes descriptives.

Parmi ces méthodes descriptives, le clustering joue un rôle de premier plan pour analyser des données moléculaires à l'échelle des populations. En plus du clustering hiérarchique et du positionnement multidimensionnel évoqués dans l'introduction (figure 2), on peut citer l'analyse en composantes principales récemment remis au gout du jour (Patterson et al., 2006) ainsi que les modèles de mélange (logiciel *structure*, Pritchard et al., 2000) qui jouent un rôle particulier puisque ce ne sont pas des méthodes descriptives *stricto sensu* même si elles sont souvent utilisées comme telles. Il existe un grand nombre de modèle bayésien dans le cadre de ces modèles de mélange et nous avons aussi contribué à ce développement (Jay et al., 2011) mais dans ce chapitre, je n'évoquerai plus le clustering (ou classification automatique en français) qui constitue un sujet en soi.

Dans ce chapitre, je vais exclusivement présenter des méthodes qui permettent de dé-

crire les données dans un cadre spatial et qui reposent sur le concept d'isolement par la distance expliqué plus loin. Je commence par décrire quelles sont les principales méthodes statistiques qui ont jusqu'à présent permis de décrire comment se structure spatialement la variation génétique. Un aspect central de la statistique spatiale est celui d'autocorrélation spatiale, définie comme la propriété qu'a une variable d'être corrélée aux variables des voisins géographiques (Sokal & Oden, 1978). En génétique des populations, c'est une forme particulière de l'autocorrélation spatiale, l'isolement par la distance qui joue une rôle clé. Ce concept stipule que la corrélation entre individus ou populations diminue au fur et à mesure que la distance entre ces entités augmente (Wright, 1943). Avec mes étudiants en thèse (Flora Jay, Nicolas Duforet-Frebourg), nous avons proposé deux extensions du concept d'isolement par la distance pour mieux appréhender la structuration génétique dans un cadre spatial. Tout d'abord, nous avons caractérisé les patrons d'isolement par la distance *non-stationnaire* où la manière avec laquelle la corrélation entre deux populations (ou individus) diminue avec la distance n'est pas la même sur l'ensemble de l'aire d'échantillonnage. Pour estimer ces variations spatiales, nous avons développé un modèle de *krigeage* (interpolation spatiale) bayésien (Handcock & Stein, 1993). L'originalité de notre approche provient du fait que nous n'utilisons pas le krigeage de manière classique pour interpoler spatialement une variable (fréquence d'allèle par exemple) mais le modèle du krigeage nous permet d'estimer comment la fonction de covariance (ou de corrélation) varie spatialement (Duforet-Frebourg & Blum, 2012). La deuxième extension que nous proposons est celle d'isolement par la distance anisotrope qui va nous permettre d'estimer la direction suivant laquelle la différentiation génétique augmente le plus vite. Nous avons développé un modèle de régression qui permet en plus de tester si le patron d'anisotropie est significatif (Jay et al., 2013). Nous illustrons ces deux extensions du patron d'isolement par la distance avec des données de SNPs chez l'homme ainsi qu'avec des marqueurs moléculaires obtenus pour des plantes alpines.

2.1 Décrire la différentiation génétique dans un cadre spatial

2.1.1 Cartes synthétiques

Un des précurseurs des analyses spatiales en génétique des populations est Luigi Cavalli-Sforza. Il a proposé d'établir des *cartes synthétiques* pour chaque continent afin de décrire la variation génétique humaine (Cavalli-Sforza et al., 1994). Pour l'essentiel, Cavalli-Sforza et al. (1994) ont collecté des données de comptage pour de nombreux variants génétiques. Ces comptages ayant été faits pour des populations réparties spatialement, ils ont ensuite interpolé spatialement les fréquences d'allèles. En utilisant une analyse en composante principale, ils ont résumé ce grand nombre de cartes de fréquences d'allèles en un ensemble de cartes synthétiques. Chacune de ces cartes représente la variation des scores de l'ACP pour chacune des composantes principales (voir figure 2.1). Du fait des propriétés de l'ACP, chaque carte de fréquence d'allèle peut être approchée par une combinaison linéaire de cartes synthétiques. Les cartes synthétiques produisaient en général des clines, et pour l'espèce humaine, ces clines ont été interprétés avec des phénomènes d'expansion spatiale (expansion néolithique). Néanmoins, l'interprétation de ces cartes synthétiques est loin d'être évidente. Il a été montré que les clines obtenus était en fait une simple conséquence de l'autocorrélation spatiale dans les données (Novembre & Stephens, 2008). De plus, dans des modèles d'expansion spatiale, la directions des clines dépend en effet de la vague d'expansion mais les clines obtenus sont perpendiculaires à la

vague d'expansion (François et al., 2010) et non dans la même direction comme suggéré par Cavalli-Sforza et al. (1994). Tous ces problèmes d'interprétation limitent bien évidemment l'utilisation de ces cartes synthétiques.

2.1.2 Détection de barrières

En écologie moléculaire, un problème récurrent est celui de la détection de barrières qui sont des zones où les fréquences d'allèles varient de manière abrupte (Storfer et al., 2010). Ce problème de la détection de frontières est aussi récurrent en géographie (Jacquez et al., 2000). On peut distinguer deux méthodes principales pour détecter les barrières génétiques : l'**algorithme de Monmonier** (Monmonier, 1973), et le « **Wombling** »(Womble, 1951). L'algorithme de Monmonier repose sur un réseau (triangulation de Delauney par exemple) qui relie les populations ou individus échantillonnés. A chaque arête de ce réseau est associée une mesure de la distance génétique entre les noeuds du réseau. L'algorithme de Monmonier est un algorithme glouton qui commence par l'arête de plus forte distance et construit des barrières en ajoutant de manière itérative les arêtes adjacentes avec les distances le plus fortes. Après avoir construit une barrière de la sorte, l'algorithme peut recommencer à partir d'une autre arête pour construire une deuxième barrière et ainsi de suite. Cette approche comporte néanmoins plusieurs défauts : une barrière arbitraire sera toujours construite même s'il n'existe pas de telle barrière (pas de test statistique), il faut choisir le nombre de barrières que retourne l'algorithme et l'approche gloutonne de l'algorithme ne permet pas de détecter tous les types de barrières.

La deuxième approche pour trouver les barrière est celle du Wombling (Womble, 1951; Barbujani et al., 1989; Bocquet-Appel & Bacro, 1994). Dans le contexte de la génétique, l'objectif est d'estimer les gradient des fréquences d'allèles et de moyenner la norme de ces gradients pour former une fonction S dite systémique

$$S(x, y) = \sum_j \|\nabla f_j(x, y)\|$$

où (x, y) est un point de l'espace et ∇f_j est le gradient de la fréquence d'allèle f_j. Les zones de variations génétiques abruptes correspondent aux zones où la valeur de la fonction systémique est élevée (Cercueil et al., 2007) (voir figure 2.1).

2.2 Isolement par la distance

C'est un concept central en génétique des populations et il existe une certaine confusion à son sujet. Le *modèle* d'isolement par la distance a été introduit par Wright (1943) et il stipule que la reproduction sexuée se fait préférentiellement entre individus séparés par de faibles distances géographiques en raison du caractère local de la dispersion. Le modèle dit du « stepping-stone »est un exemple du modèle d'isolement par la distance ; ce modèle discret suppose que seules les populations voisines peuvent échanger des migrants (Kimura & Weiss, 1964). Les modèles d'isolement par la distance vont générer ce que j'appelle le *patron* d'isolement par la distance : la corrélation entre les populations en termes de fréquences d'allèles décroit avec la distance (Sokal & Wartenberg, 1983; Hardy & Vekemans, 1999). Plutôt que de travailler avec les corrélations, il est aussi très standard en génétique des population de mettre en évidence le patron d'isolement par la distance en montrant que la différentiation génétique entre populations (F_{ST}) ou une transformation de cette différentiation augmente avec la distance (Slatkin, 1993; Rousset, 1997). Le patron d'isolement par la distance peut aussi se voir à l'échelle individuelle en montrant

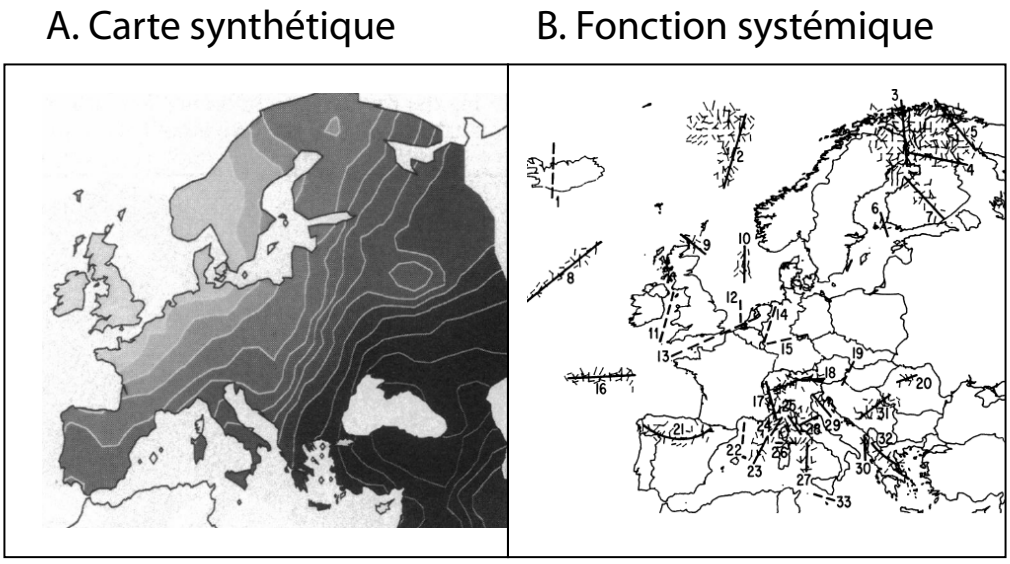

FIGURE 2.1 – Deux méthodes pour visualiser la différentiation génétique dans un cadre spatial. A. Carte synthétique : Interpolation spatiale de la première composante principale obtenue à partir d'une matrice de 95 fréquences de gènes disponibles pour différentes populations européennes (Cavalli-Sforza et al., 1994). B. Fonction systémique obtenue avec 60 fréquences de gènes. Les barrières correspondent aux 5% des valeurs les plus élevées de la fonction systémique. Les longueurs des barrière sont proportionnelles aux valeurs de la fonction systémique et les barrières sont orientées suivant la direction de plus grande variation (Barbujani & Sokal, 1990).

que la corrélation entre individus décroit avec la distance qui les sépare (figure 2.2). La confusion provient du fait que l'on utilise l'expression isolement par la distance pour décrire indistinctement le patron et le modèle d'isolement par la distance. C'est un raccourci regrettable puisque les deux notions ne sont pas équivalentes ; des processus complexes de divergence et d'expansion autre que la dispersion locale peuvent aussi générer des patrons d'isolement par la distance (Marko & Hart, 2011). Dorénavant, par isolement par la distance, nous entendrons le patron d'isolement par la distance. L'isolement par la distance est un concept qui dépasse la génétique des populations. En géographie, il est connu sous le nom de la première loi de la géographie de Tobler et stipule que « everything is related to everything else, but near things are more related than distant things »(Tobler, 1970). Dans les section 2.3 et 2.4, je montre comment nous avons étendu le concept d'isolement par la distance afin de caractériser de manière plus fine la différentiation génétique entre les populations ou les individus.

2.3 Isolement par la distance non-stationnaire

Le patron d'isolement par la distance (figure 2.2, panel A) peut masquer des variations complexes des paramètres démographiques, et ces variations peuvent entraîner des vitesses de décorrélation différentes pour différentes régions de l'aire de répartition de l'organisme étudié. Ce genre de situation peut arriver lorsque la densité de population ou le taux de migration varie dans l'espace. Avec l'avènement de la génétique du paysage (Manel et al., 2003), la variation spatiale des paramètres démographiques est devenue un sujet important en particulier parce que l'hétérogénéité spatiale (i.e. le paysage) est maintenant

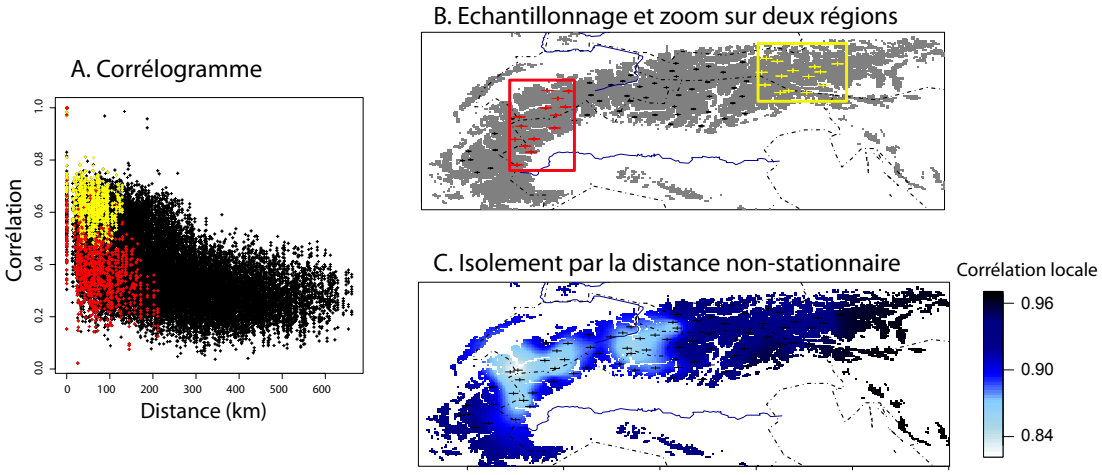

Figure 2.2 – Caractérisation du patron d'isolement par la distance pour la plante alpine *Phyteuma hemisphaericum* à partir de marqueurs AFLP (Gugerli et al., 2008). A. Corrélation génétique entre individus en fonction de la distance géographique. Les points en rouge correspondent aux comparaisons faites entre individus qui vivent dans la zone encadrée en rouge à l'ouest des Alpes tandis que les points en jaune correspondent aux comparaisons faites entre individus qui vivent dans la zone encadrée en jaune située à l'est des Alpes (voir B.) La corrélation à distance égale est plus faible dans la zone rouge que dans la zone jaune. C. Caractérisation du patron d'isolement par la distance non stationnaire. Les corrélations correspondent aux corrélations entre individus échantillonnés et voisins fictifs qui vivent à 10km. Conformément au patron A., les corrélations à distance fixée sont plus fortes dans l'est des Alpes.

reconnue comme étant un facteur clé pour expliquer la différentiation génétique entre les populations (McRae & Beier, 2007).

D'un point de vue statistique, l'isolement par la distance non-stationnaire est défini par le fait que la manière avec laquelle la corrélation entre individus ou entre populations décroit avec la distance dépend de l'endroit où l'on se trouve. Pour quantifier ces variations, nous proposons d'estimer en chaque point échantillonné quelle est la corrélation entre l'individu ou la population au point d'échantillonnage et un voisin fictif qui habite à une distance fixée que l'on se donne (Duforet-Frebourg & Blum, 2012). Notre méthode statistique utilise la matrice de corrélation S entre les sites échantillonnés comme donnée initiale et renvoie une prédiction de la corrélation locale avec les voisins fictifs pour chaque site échantillonné. Pour pouvoir estimer ces corrélations locales, nous allons utiliser des techniques dites de krigeage (Cressie, 1992).

2.3.1 Krigeage

Le krigeage est une technique de géostatistique qui consiste à estimer la valeur d'une variable en un certain nombre de sites non échantillonnés à partir des valeurs mesurées aux sites échantillonnés. Cette estimation se fait en utilisant un modèle du corrélogramme qui décrit la vitesse avec laquelle la corrélation décroit avec la distance.

Une manière de voir le krigeage fait appel au processus gaussiens et consiste à supposer que la loi jointe de la variable d'intérêt aux sites échantillonnés et non échantillonnés est

une gaussienne multivariée (Bishop, 2006)

$$(X, Y) \rightsquigarrow \mathcal{N}(m, \Psi), \tag{2.1}$$

avec

$$\Psi = \begin{pmatrix} \Psi_{xx} & \Psi_{xy} \\ \Psi_{xy} & \Psi_{yy} \end{pmatrix},$$

où X (resp. Y) est le vecteur des valeurs aux sites échantillonnés (resp. non échantillonnés), m est une moyenne constante (krigeage ordinaire), Ψ_{xx} (resp. Ψ_{yy}) modélise la matrice de covariance de X (resp. Y) et Ψ_{xy} correspond à la matrice de covariance entre les éléments de X et ceux de Y. Dans notre approche, on choisira une matrice de corrélation pour Ψ.

L'interpolation de la variable aux sites non échantillonnés est obtenue à partir de la loi conditionnelle de Y sachant X qui peut être écrite sous la forme de l'équation de régression suivante

$$Y - m = \tau(X - m) + \epsilon, \tag{2.2}$$

où $\tau = \Psi_{xy}\Psi_{xx}^{-1}$ et ϵ est un résidu indépendant de X et de variance connue (Le & Zidek, 1992).

L'originalité de notre approche consiste à ne pas utiliser le krigeage de manière classique qui consisterait par exemple à estimer quelles sont les fréquences d'allèles pour un ensemble de sites non échantillonnés. En revanche, nous allons utiliser le krigeage pour estimer la covariance (ou la corrélation) qui existe entre les sites échantillonnés et les sites voisins (non échantillonnés). En utilisant l'équation de régression (2.2) qui relie les variables aux sites échantillonnés et non échantillonnés, on peut prédire quelle est la matrice de covariance entre X and Y

$$\Sigma_{xy} = \tau\Sigma_{xx}. \tag{2.3}$$

où Σ_{xx} est la matrice de covariance pour les sites échantillonnés que l'on estime avec la matrice de covariance empirique S dans les calculs. La corrélation s'obtient après renormalisation de la matrice de covariance de l'équation (2.3) (Duforet-Frebourg & Blum, 2012). La prédiction de Σ_{xy} se fait en intégrant sur la loi a posteriori du paramètre τ qui est une fonction de la matrice de covariance Ψ. Le modèle paramétrique pour Ψ est décrit dans la sous-section 2.3.2 qui suit.

2.3.2 Modèle du corrélogramme

Nous considérons le modèle standard du krigeage stationnaire qui suppose que la corrélation entre deux points ne dépend que de la distance entre ces deux points. En utilisant cette hypothèse, il ne reste plus qu'à spécifier la fonction C qui décrit comment la corrélation décroit avec la distance. Nous supposons que cette fonction C, appelée le corrélogramme, décroît de manière exponentielle

$$C(d) = (\alpha + (1 - \alpha)e^{-d/r} + \lambda\mathbb{1}_{d=0})/(1 + \lambda), \tag{2.4}$$

où d est la distance entre deux points, $\mathbb{1}$ est la fonction indicatrice, α est le palier qui détermine la valeur limite de la corrélation pour de grandes distances, r est la portée qui décrit la vitesse avec laquelle décroit la corrélation et λ est un paramètre de *régularisation* qui permet principalement d'inverser la matrice Ψ_{xx}. Nous échantillonnons (α, λ, r) en utilisant un échantillonneur de Gibbs (Handcock & Stein, 1993).

Il peut sembler paradoxal de considérer un modèle de corrélogramme stationnaire afin de caractériser le patron non-stationaire qui peut exister dans les données. Cependant,

le modèle stationnaire de l'équation (2.4) ne fournit que les valeurs de la matrice de régression τ et ce sont les données contenues dans la matrice de covariance empirique qui contiennent les informations à propos de la non-stationnarité. L'information contenue dans le modèle du corrélogramme et celle contenue dans les données se combinent via l'équation (2.3) qui fournit une prédiction de la matrice de covariance. Il existe aussi des modèles paramétriques de corrélogramme non-stationnaire (Paciorek & Schervish, 2006) mais nous avons trouvé que ces modèles ne sont pas assez flexibles pour reproduire les patrons simulés dans des modèles d'isolement par la distance (Duforet-Frebourg & Blum, 2012).

2.3.3 Etude de simulations et application à des données moléculaires

Je montre un exemple d'une situation où la détection de barrières ainsi qu'une analyse multivariée classique comme le positionnement multidimensionnel (MDS) ne produisent pas de résultats interprétables à la différence d'un estimation de la corrélation locale (figure 2.3). Dans un modèle de « stepping-stone »en deux dimensions (Kimura & Weiss, 1964), on suppose que le flux de gène est maximal dans le coin inférieur gauche de l'espace et décroit de manière proportionnelle à la distance à ce coin là de sorte à ce que le flux de gène soit minimal dans le coin supérieur droit de l'espace. La carte représentée dans la figure 2.3 représente la dissimilarité génétique locale définie comme 1 moins la corrélation génétique locale (i.e. la corrélation entre populations échantillonnées et des populations voisines fictives vivant à une distance de 0.1). Comme attendu, cette dissimilarité locale est maximale en haut à droite de l'espace et est minimale en bas à gauche de l'espace là où le flux de gène est maximal. L'algorithme de Monmonier, implémentée dans le logiciel *barrier* (Manni et al., 2004), trouve une barrière en haut à droite ce qui est compatible avec le fait que les flux de gènes sont plus faibles dans cette région mais qui est néanmoins difficilement interprétable du fait qu'il n'existe pas de discontinuité dans cette région. Le résultat obtenu avec MDS est lui aussi parfaitement compatible avec les simulations ; les points (clairs) situés dans les zones de faible flux de gènes et qui sont donc peu corrélés avec le reste sont loin du principal nuage de points (points noirs). En revanche, il peut être très difficile d'interpréter cette figure obtenue avec MDS dans le cas réaliste où on ne connait pas les processus évolutifs qui ont façonné la variation génétique.

Nous avons aussi appliqué notre méthode à des données de marqueurs AFLP obtenus pour une vingtaine de plantes alpines échantillonnées le long des alpes (Gugerli et al., 2008; Jay et al., 2012). Ici, nous choisissons de montrer les résultats pour une seule plante alpine *Phyteuma hemispaericum* qui a un patron de non-stationnarité particulièrement simple (figure 2.2). En effet, la dissimilarité locale (un moins la corrélation locale) est plus forte dans une zone qui se trouve dans la partie ouest et centrale des Alpes ce qui est conforme avec le fait que la corrélation est plus faible à distance fixée dans cette zone (rectangle rouge de la figure 2.2). La figure 2.2 montre aussi les limites des approches qui recherchent une barrière sous la forme d'une étroite bande (algorithme de Monmonier) puisque pour cette plante la dissimilarité locale est élevée sur une grande région des Alpes. En revanche, il n'est pas immédiat d'inférer quels sont les processus évolutifs qui ont généré ce patron ; barrière au flux de gène ou zone de contact secondaire qui résulte de recolonisations postglaciaires sont deux explications possibles.

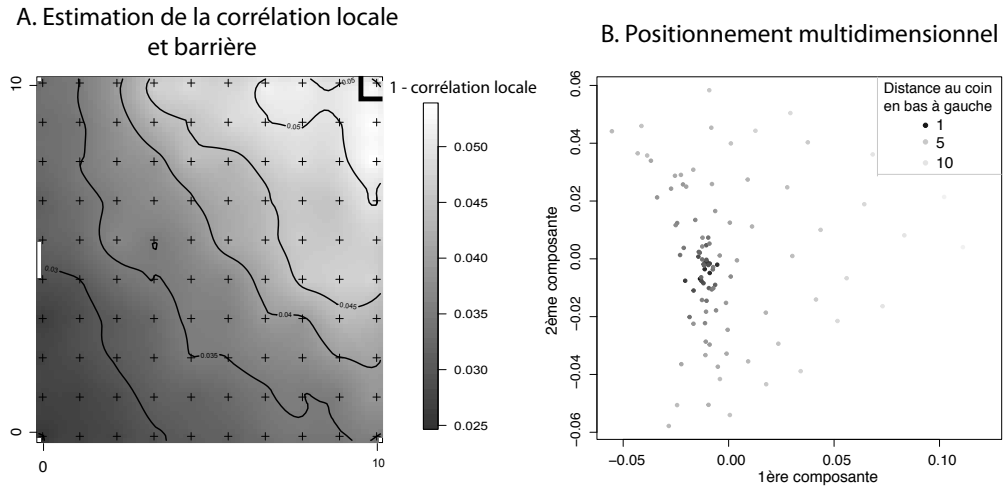

FIGURE 2.3 – Etude du patron de différentiation génétique dans le cas d'un gradient de flux de gènes dans un espace à deux dimensions. Le flux de gène est maximal dans le coin inférieur gauche et diminue proportionnellement à la distance au coin inférieur gauche. A. Estimation de 1 moins la corrélation locale pouvant aussi s'interpréter comme une mesure de dissimilarité génétique locale. La corrélation est calculée entre sites échantillonnés (croix) et voisins qui se trouvent à une distance de 0.1. La première barrière génétique trouvée avec l'algorithme de Monmonier est indiquée par une ligne noire épaisse. B. Positionnement multidimensionnel avec une palette de couleurs en niveaux de gris pour représenter la distance de chaque population au coin inférieur gauche. Les points sombres correspondent aux populations qui sont proches du coin inférieur gauche, tandis que les points les plus clairs sont les plus éloignés du coin inférieur gauche.

2.4 Isolement par la distance anisotrope

La deuxième extension que nous proposons prend en compte l'anisotropie, c'est à dire le fait que la corrélation entre populations (calculée à partir des fréquences d'allèles) peut diminuer avec des vitesses différentes suivant la direction géographique (nord-sud, est-ouest, etc) (Jay et al., 2013). Plutôt que de travailler avec la mesure de corrélation, nous travaillerons dans cette section avec la mesure de différentiation génétique F_{ST} qui est une mesure de dissimilarité entre populations (Weir & Cockerham, 1984).

2.4.1 Un modèle de régression

Pour prendre en compte l'anisotropie, nous allons supposer que la vitesse avec laquelle la F_{ST} croit avec la distance est une fonction β de la direction entre les deux populations étudiées

$$F_{ST} = \alpha + \beta(\theta)d + \varepsilon, \tag{2.5}$$

où ε est le résidu, d est la distance entre les deux populations et θ est l'azimut entre ces deux populations quand on suit une ligne de direction fixée (loxodrome). On suppose que 0° correspond à la direction nord-sud et on l'on tourne dans le sens des aiguilles d'une montre.

La fonction β de l'équation (2.5) est périodique de période π puisque les directions que nous considérons ne sont pas orientées de sorte à ce que nous ne faisons pas de différence

par exemple entre les directions nord-sud et sud-nord. En considérant une approximation de Fourier au premier ordre pour la fonction β de l'équation (2.5), nous pouvons considérer le modèle paramétrique suivant pour prendre en compte l'anisotropie

$$F_{ST} = \alpha + \beta_0 d + \beta_1 d\cos(2\theta) + \beta_2 d\sin(2\theta) + \varepsilon. \tag{2.6}$$

Le modèle de régression paramétrique de l'équation (2.6) a l'avantage qu'il contient le modèle isotrope classique de l'isolement par la distance (Slatkin, 1993; Rousset, 1997) lorsque $\beta_1 = \beta_2 = 0$. Nous pouvons effectuer un test d'hypothèse pour tester s'il y a une anisotropie significative de la différentiation génétique entre populations. La difficulté avec un test dans le cadre de comparaison entre paires de populations ou d'individus provient du fait que les résidus ne sont pas indépendants. On va donc utiliser une procédure classique dans ce contexte qui est le test partiel de Mantel. Pour calculer une P-valeur, nous régressons tout d'abord les F_{ST} avec la distance pour obtenir une matrice de résidus. Ensuite, nous régressons ces résidus avec $\cos(2)d$ et $\sin(2)d$ et on calcule la statistique R^2 correspondante que l'on utilisera comme statistique de test. Pour trouver la distribution de la statistique de test sous l'hypothèse nulle d'isotropie, nous permutons aléatoirement lignes et colonnes de la matrice de résidus (Legendre, 2000). Cette section 2.4 sur l'isolement par la distance anisotrope est la seule partie de ce manuscrit où les méthodes que nous développons ne se font pas dans un cadre bayésien ; en raison des critiques soulevées par les tests de Mantel et les tests de Mantel partiels (voir Raufaste & Rousset, 2001; Guillot & Rousset, 2011), il serait néanmoins souhaitable de développer des alternatives bayésiennes à l'avenir.

2.4.2 Application à des données humaines

A partir de l'équation de régression (2.6), nous avons testé l'anisotropie de la différentiation génétique chez l'homme à l'échelle de plusieurs continents (Jay et al., 2013). Le seul continent ou le test partiel de Mantel n'est pas significatif est le continent américain pour lequel nous ne détectons même pas de patron d'isolement par la distance. Dans les autres continents, nous détectons un patron anisotrope d'isolement par la distance ($P < 0.02$) et les directions principales de différentiation génétiques sont N-S en Afrique, E-W en Asie et SSE-NNW en Europe. Une autre approche, géométrique, a donné des résultats similaires (Jay et al., 2013).

Nous avons aussi poursuivi un autre objectif où nous avons cherché à prédire la localisation géographique des individus à partir des puces à SNPs. Afin de tester la pertinence des résultats obtenus avec la régression (2.6) qui prend en compte l'anisotropie, nous avons évalué l'erreur de localisation dans les différentes directions de l'espace et nous avons pu montrer que l'erreur minimale est bien obtenue dans la direction de plus forte différentiation. La localisation des individus s'est faite en apprenant le modèle suivant basé sur une régression à partir des scores de l'ACP (Jolliffe, 2005)

$$L = \delta_0 + \sum_{i=1}^{K} \delta_i \mathrm{PC}_i, \tag{2.7}$$

où L représente soit la latitude soit la longitude correspondant à un individu, PC_i représente le score de la $i^{ème}$ composante principale calculée à partir des données de SNP, $\delta_0, \ldots, \delta_K$ sont les coefficients du modèle de régression et K représente le nombres de composantes de l'ACP qui est utilisé. Le choix de K était effectué avec une méthode de validation croisée.

Conclusions et perspectives

Data scientist : the sexiest job of the 21st Century

Thomas H. Davenport and D.J. Patil
in the Harvard Business Review

Interprétation des résultats

Les statistiques qu'elles soient bayésiennes ou pas ont joué et jouent un rôle de plus en plus central en génétique des populations (Beaumont & Rannala, 2004). C'est bien évidemment une chance pour ceux qui développent ces méthodes puisque leur position devient désormais centrale. En revanche, l'utilisation massive des statistiques doit s'accompagner d'un certain nombre de garde-fous et notre rôle en tant que statisticien est d'en faire la promotion. Le piège le plus classique est bien évidemment celui qui conduit à surinterpréter les résultats obtenus après une analyse statistique. La surinterprétation des résultats d'une analyse descriptive n'est bien évidemment pas une chose nouvelle et lorsque l'on voit la figure 2.1 et les 33 barrière génétiques trouvées par Barbujani & Sokal (1990), on est en droit de s'interroger sur la pertinence statistique de toutes ces barrières, imprudence d'autant plus surprenante qu'elle émane d'un auteur qui a écrit un livre qui fait référence en biométrie (Sokal & Rohlf, 1995).

Dans la suite, je donne deux des raisons qui peuvent conduire à surinterpréter ou à mal interpréter les résultats. Tout d'abord, les patrons observés peuvent être générés en partie par l'échantillonnage des données. Dans le cas de l'ACP, il a été montré que les patrons obtenus dépendent fortement du nombre d'individus échantillonnés par population (McVean, 2009). Il a aussi été montré que les échantillonnages non réguliers peuvent tromper les algorithmes de clustering qui détectent plusieurs populations alors qu'une seule population est générée dans les simulations (Schwartz & McKelvey, 2009). Pour la méthode anisotrope de la section 2.4, des simulations nous ont permis de montrer que les résultats obtenus avec l'équation de régression (2.6) était en revanche robuste vis à vis de l'échantillonnage. La seconde raison qui peut conduire à une mauvaise interprétation des résultats provient de la possible mauvaise adéquation des modèles statistiques que l'on essaye de calibrer (chapitre 1). Bien sur, ça peut sembler une évidence qu'il faille vérifier l'adéquation des modèles aux données, mais ce n'est que récemment que l'importance du *goodness-of-fit* est devenue prégnante dans le cadre de l'ABC et nous avons oeuvré pour en faire la promotion (Cornuet et al., 2010; Csilléry et al., 2010). En revanche, il existe encore un certain nombre de logiciels très populaires en génétique des populations qui ne proposent pas de tester l'adéquation des modèles démographiques calibrés avec les données (par exemple BEAST ; Drummond & Rambaut, 2007). C'est bien évidemment

regrettable puisque la diffusion de logiciels et de méthodes statistiques doit s'accompagner d'une perspective critique qui ne cache pas quelles sont les difficultés d'interprétation.

Perspectives

Lorsque j'ai commencé ma thèse en 2002, les écologues et généticiens travaillaient typiquement avec un unique marqueur moléculaire (ADN mitochondrial) voir des dizaines de marqueurs génétiques (AFLP, microsatellites). Chez l'homme, on dispose aujourd'hui de données qui comptent des millions de marqueurs voire des dizaines de millions de marqueurs génétiques (projet 1000 génomes chez l'homme, Altshuler et al., 2010). Du fait de l'avènement des techniques de NGS (Next Generation Sequencing), une grande quantité de variants variables au sein d'une espèce ou d'une population sera bientôt disponible pour un grand nombre d'espèces. Les méthodes statistiques développées dans les années 90-200 pour répondre aux questions d'intérêt en génétique des populations ne pourront pas passer à l'échelle avec les données générées par les NGS. Les méthodes reposent typiquement sur des algorithmes MCMC qui sont en général assez lents. Un des défis de l'analyse statistique bayésienne en génétique est donc de pouvoir passer à l'échelle des données massives qui sont en train d'être produite (« Scalable Bayesian Computation »). Dans d'autres domaines de la science comme l'imagerie ou le web sémantique, des techniques de régression et de classification qui peuvent passer à l'échelle des données massives ont été développées. Il existe donc un important corpus méthodologique qui n'a pas encore été exploité dans le domaine de la génétique des populations et plus généralement pour la génétique médicale et l'écologie. Les applications potentielles sont vastes : détection de la structure des populations, détection des gènes impliqués dans des processus d'adaptation biologique, analyse comparative en phylogénie et dans un cadre plus médical, l'étude de l'association entre gènes et maladies.

Je donne ci-dessous 3 exemples de perspectives qui concernent l'analyse de données massives dans un contexte biologique et médical. Le premier exemple a trait à la détection des gènes qui sont impliqués dans les processus de sélection naturelle. Ces méthodes sont souvent dénommées « genome scan »puisque l'on effectue un balayage statistique de l'ensemble du génome pour trouver quelles sont les zones atypiques (*outliers*) qui correspondraient aux zones sous sélection darwinienne. Une approche classique est celle des scans à F_{st} (indice de différentiation génétique) où l'on cherche les marqueurs très différenciés entre les populations étudiées (fort F_{st}) (Holsinger & Weir, 2009). En effet, cette forte différentiation peut s'interpréter par des processus d'adaptation locale qui accentuent les différences de fréquences d'allèles entre les populations. Cette approche comporte plusieurs défauts principalement le fait que ce soit une approche supervisée puisqu'il faut définir quelles sont les populations à l'avance. Mon objectif est de développer une approche non supervisée, sans définir de populations, qui permette à la fois de détecter les marqueurs candidats pour l'adaptation locale ainsi que les régions géographiques où ont eu lieu ces processus d'adaptation locale. Notre approche qui doit passer à l'échelle des données massives sera basée sur le principe de l'analyse factorielle dont l'utilité pour la génétique des populations a été reconnue récemment (Engelhardt & Stephens, 2010; Frichot et al., 2012). L'idée est d'utiliser un mélange de régression avec un seul et unique facteur (Tipping & Bishop, 1999). Plus précisément, si l'on note x_i le vecteur qui contient le $i^{ème}$ marqueur génétique pour l'ensemble des n individus, alors ce vecteur pourra s'écrire sous la forme

$$x_i = \alpha_i u^k + \epsilon_i, \tag{2.8}$$

où u^k, $k = 1, \dots K$ est un vecteur de taille n (le facteur) à choisir parmi les K facteurs

qui sont eux mêmes des paramètres du modèle. Le terme ϵ_i est un résidu qui prendra en compte l'autocorrélation spatiale. Le paramètre K du modèle est un hyper paramètre du modèle et devra être choisi à partir d'un critère de sélection de modèle. Pour chaque marqueur génétique, on pourra calculer sa corrélation avec le facteur u^k auxquels il est associé et pour les marqueurs les plus corrélés, on pourra caractériser le processus d'adaptation dans lequel il est impliqué en étudiant le facteur u^k. Dans un cadre spatial par exemple, les facteurs pourront avoir un gradient directionnel (est-ouest, nord-sud) ou simplement contraster une région géographique avec toutes les autres. Une attention particulière devra être portée aux algorithmes qui permettent d'estimer les paramètres du modèle en raison des dimensions importantes des données (dizaine de millions de marqueurs, milliers d'individus). Des algorithmes performants ont été développé pour des modèles statistiques proches (non-negative matrix factorization, low-rank matrix factorization) fournissant un panel d'algorithmes dont on pourra s'inspirer (Seung & Lee, 2001; Achlioptas & Mcsherry, 2007).

Le deuxième et troisième exemple concernent le modèle linéaire à effet mixte qui joue un rôle clé dans de nombreuses analyses en biostatistique (McCulloch & Neuhaus, 2005). Dans une forme simplifié, ce modèle peut s'écrire sous la forme suivante

$$y = X\beta + u + \epsilon, \tag{2.9}$$

avec y la réponse à expliquer, X la matrice des variables explicatives, u l'effet aléatoire modélisée par une gaussienne de matrice de covariance donnée ou à estimer, et ε le résidu de l'équation de régression. La deuxième exemple a trait aux études d'association entre le phénotype y et le génotype X et dans ce cadre l'effet aléatoire u permet de prendre en compte la corrélation qui existe entre individus soit en raison de leur apparentement soit en raison de la structure génétique des populations (Yu et al., 2005). Les approches classiques pour déterminer quelles sont les marqueurs génétiques qui expliquent un trait phénotypique d'intérêt (présence/absence d'une maladie, trait quantitatif comme la taille) analysent les marqueurs génétiques indépendamment les uns des autres et renvoient typiquement une P-valeur pour chacun de ces marqueurs. En revanche, on sait désormais que si plusieurs gènes ont un effet fort, ces approches marqueur par marqueur augmente le nombre de faux positif en raison de la corrélation entre marqueurs (Segura et al., 2012). Des méthodes multigéniques sont donc en train d'être développées, et pour l'instant elles ont le défaut d'être assez couteuses en temps de calcul. Ces approches comprennent soit des approches naïves où l'on essaye de rajouter ou d'enlever les marqueurs un à un dans le modèle de régression (Segura et al., 2012), soit des approches bayésienne de régularisation qui pénalisent de manière drastique les β de l'équation (2.9) (Carbonetto & Stephens, 2012). Mon objectif est double. Tout d'abord, je souhaite proposer des méthodes bayésiennes qui passent à l'échelle des données génétique haut-débit. Ensuite, je compte proposer un cadre cohérent au sein duquel le terme de covariance de l'effet aléatoire u soit estimé conjointement avec les termes de régression. Au lieu d'estimer le terme de covariance lors d'une analyse préliminaire comme c'est le cas en général (Yu et al., 2005) cette estimation pourra se faire en utilisant l'analyse factorielle (Carvalho et al., 2008).

Le dernier exemple concerne toujours le modèle de régression de l'équation (2.9) mais cette fois ci dans le cadre de l'analyse comparative en phylogénie. Dans ce cadre, on cherche à expliquer une variable typique d'une espèce donnée (poids moyen d'un individu au sein de cette espèce par exemple) à partir d'autres variables tout en prenant en compte la corrélation entre les variables. Cette corrélation provient de l'histoire évolutive qui relie les espèces entre elles et qui est représentée sous la forme d'une phylogénie (Pagel, 1994). La corrélation entre espèces est prise en compte par le terme d'effet aléatoire u dans l'équation

de régression en faisant en sorte que la matrice de covariance reflète la phylogénie entre espèces. Une nouvelle fois l'objectif est de faire passer le modèle de l'équation (2.9) à l'échelle des données massives sans utiliser d'algorithmes MCMC (Hadfield, 2010), et de pouvoir estimer conjointement le signal phylogénétique dans les données qui est jusqu'à présent estimé indépendamment des variables explicatives ce qui biaise son estimation (Blomberg et al., 2003). Dans ces deux exemples d'application du modèle mixte (étude d'association, analyse comparative), j'ai des objectifs communs qui sont de développer des algorithmes bayésiens efficaces dans le contexte des données massives et de pouvoir estimer le terme de covariance et les coefficients de régression de manière conjointe afin de produire des estimations moins biaisées et ainsi diminuer le nombre de faux positifs.

Bibliographie

Achlioptas D. and Mcsherry F. Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)*, 54(2) :9, 2007.

Altshuler D., Lander E., Ambrogio L., Bloom T., Cibulskis K., Fennell T., Gabriel S., Jaffe D., Shefler E., Sougnez C., and others . A map of human genome variation from population scale sequencing. *Nature*, 467(7319) :1061–1073, 2010.

Barbujani G. and Sokal R. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences*, 87(5) :1816–1819, 1990.

Barbujani G., Oden N., and Sokal R. Detecting regions of abrupt change in maps of biological variables. *Systematic Biology*, 38(4) :376–389, 1989.

Beaumont M. and Rannala B. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4) : 251–261, 2004.

Beaumont M. A. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41 :379–406, 2010.

Beaumont M. A., Zhang W., and Balding D. J. Approximate Bayesian computation in population genetics. *Genetics*, 162 :2025–2035, 2002.

Beaumont M. A., Cornuet J.-M., Marin J.-M., and Robert C. P. Adaptive approximate Bayesian computation. *Biometrika*, 96(4) :983–990, 2009.

Biau G., Cérou F., and Guyader A. New insights into approximate Bayesian computation. *arXiv :1207.6461*, 2012.

Bishop C. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

Blomberg S., Garland Jr T., and Ives A. Testing for phylogenetic signal in comparative data : behavioral traits are more labile. *Evolution*, 57(4) :717–745, 2003.

Blum M. G. B. Approximate Bayesian computation : A nonparametric perspective. *Journal of the American Statistical Association*, 105 :1178–1187, 2010.

Blum M. G. B. and François O. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20 :63–73, 2010.

Blum M. G. B. and Jakobsson M. Deep divergences of human gene trees and models of human origins. *Molecular biology and evolution*, 28(2) :889–898, 2011.

Blum M. G. B., Nunes M., Prangle D., and Sisson S. A comparative review of dimension reduction methods in approximate Bayesian computation, 2012. *Statistical Science*, in press, 2012.

Bocquet-Appel J.-P. and Bacro J.-N. Generalized Wombling. *Systematic Biology*, 43(3) :442–448, 1994.

Carbonetto P. and Stephens M. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1) :73–108, 2012.

Carvalho C. M., Chang J., Lucas J. E., Nevins J. R., Wang Q., and West M. High-dimensional sparse factor modeling : applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484) :1438–1456, 2008.

Cavalli-Sforza L. and Zei G. Experiments with an artificial population. In *Proceedings of the Third International Congress of Human Genetics*, pages 473–478. John Hopkins Press Chicago, 1967.

Cavalli-Sforza L., Menozzi P., and Piazza A. *The history and geography of human genes*. Princeton university press, 1994.

Cercueil A., François O., and Manel S. The genetical bandwidth mapping : a spatial and graphical representation of population genetic structure based on the wombling method. *Theoretical population biology*, 71(3) :332–341, 2007.

Cornuet J., Ravigné V., and Estoup A. Inference on population history and model checking using dna sequence and microsatellite data with the software diyabc (v1. 0). *BMC Bioinformatics*, 11(1) :401, 2010.

Cressie N. *Statistics for spatial data*. Wiley Online Library, 1992.

Csilléry K., Blum M. G. B., Gaggiotti O., and François O. Approximate Bayesian computation in practice. *Trends in Ecology & Evolution*, 25 :410–418, 2010.

Csilléry K., François O., and Blum M. G. B. abc : an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*, 3 :475–479, 2012.

Del Moral P., Doucet A., and Jasra A. An adaptive sequential monte carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5) :1009–1020, 2012.

Diggle P. J. and Gratton R. J. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

Doksum K. and Lo A. Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, 18(1) :443–453, 1990.

Drummond A. and Rambaut A. Beast : Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1) :214, 2007.

Duforet-Frebourg N. and Blum M. G. B. Non-stationary patterns of isolation-by-distance : inferring measures of genetic friction. *arXiv :1209.5242*, 2012.

Engelhardt B. and Stephens M. Analysis of population structure : a unifying framework and novel methods based on sparse factor analysis. *PLoS genetics*, 6(9) :e1001117, 2010.

Fearnhead P. and Prangle D. Constructing summary statistics for approximate Bayesian computation : semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(3) :419–474, 2012.

François O., Currat M., Ray N., Han E., Excoffier L., and Novembre J. Principal component analysis under population genetic models of range expansion and admixture. *Molecular biology and evolution*, 27(6) :1257–1268, 2010.

Frichot E., Schoville S., Bouchard G., and François O. Landscape genomic tests for associations between loci and environmental gradients. *arXiv :1205.3347*, 2012.

Green R., Krause J., Briggs A., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M., and others . A draft sequence of the Neandertal genome. *Science*, 328(5979) :710–722, 2010.

Gugerli F., Englisch T., Niklfeld H., Tribsch A., Mirek Z., Ronikier M., Zimmermann N., Holderegger R., and Taberlet P. Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation–a project synopsis. *Perspectives in Plant Ecology, Evolution and Systematics*, 10(4) : 259–281, 2008.

Guillot G. and Rousset F. On the use of the simple and partial Mantel tests in presence of spatial autocorrelation. *arXiv :1112.0651*, 2011.

Hadfield J. MCMC methods for multi-response generalized linear mixed models : the MCMCglmm R package. *Journal of Statistical Software*, 33(2) :1–22, 2010.

Handcock M. and Stein M. A Bayesian analysis of kriging. *Technometrics*, 35(4) :403–410, 1993.

Hansen B. E. Nonparametric conditional density estimation. Working paper available at `http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf`, 2004.

Hardy O. and Vekemans X. Isolation by distance in a continuous population : reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, 83(2) :145–154, 1999.

Hoban S., Bertorelle G., and Gaggiotti O. Computer simulations : tools for population and evolutionary genetics. *Nature Reviews Genetics*, 2012.

Hoerl A. and Kennard R. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.

Holsinger K. and Weir B. Genetics in geographically structured populations : defining, estimating and interpreting fst. *Nature Reviews Genetics*, 10(9) :639–650, 2009.

Hudson R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18 :337–338, 2002.

Hyndman R. J., Bashtannyk D. M., and Grunwald G. K. Estimating and visualizing conditional densities. *Journal of Computing and Graphical Statistics*, 5 :315–336, Dec. 1996. ISSN 1061-8600.

Jacquez G., Maruca S., and Fortin M. From fields to objects : a review of geographic boundary analysis. *Journal of Geographical Systems*, 2(3) :221–241, 2000.

Jakobsson M., Scholz S., Scheet P., Gibbs J., VanLiere J., Fung H., Szpiech Z., Degnan J., Wang K., Guerreiro R., and others . Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181) :998–1003, 2008.

Jay F., François O., and Blum M. G. B. Predictions of Native American population structure using linguistic covariates in a hidden regression framework. *PloS One*, 6(1) :e16227, 2011.

Jay F., Manel S., Alvarez N., Durand E. Y., Thuiller W., Holderegger R., Taberlet P., and François O. Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, 2012.

Jay F., Sjodïn P., Jakobsson M., and Blum M. G. B. Anisotropic isolation by distance : the main orientations of human genetic differentiation. *Molecular Biology and Evolution*, Advance Access, 2013.

Jolliffe I. *Principal component analysis*. Wiley Online Library, 2005.

Joyce P. and Marjoram P. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7, 2008. Article 26.

Kimura M. and Weiss G. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4) :561, 1964.

Le N. and Zidek J. Interpolation with uncertain spatial covariances : A Bayesian alternative to kriging. *Journal of Multivariate Analysis*, 43(2) :351–374, 1992.

Legendre P. Comparison of permutation methods for the partial correlation and partial mantel tests. *Journal of Statistical Computation and Simulation*, 67(1) :37–73, 2000.

Manel S., Schwartz M., Luikart G., and Taberlet P. Landscape genetics : combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4) :189–197, 2003.

Manni F., Guerard E., and Heyer E. Geographic patterns of (genetic, morphologic, linguistic) variation : how barriers can be detected by using Monmonier's algorithm. *Human Biology*, 76(2) :173–190, 2004.

Marin J.-M., Pudlo P., Robert C. P., and Ryder R. J. Approximate Bayesian computational methods. *Statistics and Computing*, pages 1–14, 2011.

Marko P. and Hart M. The complex analytical landscape of gene flow inference. *Trends in ecology & evolution*, 26(9) :448–456, 2011.

McCulloch C. and Neuhaus J. *Generalized linear mixed models*. Wiley Online Library, 2005.

McRae B. and Beier P. Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, 104(50) :19885–19890, 2007.

McVean G. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10) :e1000686, 2009.

Monmonier M. Maximum-difference barriers : An alternative numerical regionalization method. *Geographical Analysis*, 5(3) :245–261, 1973.

Novembre J. and Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5) :646–649, 2008.

Nunes M. A. and Balding D. J. On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

Paciorek C. and Schervish M. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5) :483–506, 2006.

Pagel M. Detecting correlated evolution on phylogenies : a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 255 (1342) :37–45, 1994.

Patterson N., Price A., and Reich D. Population structure and eigenanalysis. *PLoS genetics*, 2(12) :e190, 2006.

Pritchard J., Stephens M., and Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–959, 2000.

Pritchard J. K., Seielstad M. T., Perez-Lezaun A., and Feldman M. W. Population growth of human Y chromosomes : a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16 : 1791–1798, 1999.

Raufaste N. and Rousset F. Are partial Mantel tests adequate ? *Evolution*, 55(8) :1703–1705, 2001.

Ripley B. Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 409–456, 1994.

Robert C. P., Cornuet J.-M., Marin J.-M., and Pillai N. S. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37) :15112–15117, 2011.

Rousset F. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145(4) :1219–1228, 1997.

Rubin D. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

Saitou N. and Nei M. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4) :406–425, 1987.

Schwartz M. and McKelvey K. Why sampling scheme matters : the effect of sampling scheme on landscape genetic results. *Conservation Genetics*, 10(2) :441–452, 2009.

Scott D. W. *Multivariate density estimation*. Wiley, New York, 1992.

Sedki M. A. and Pudlo P. Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation : Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society : Series B*, 74 :466–467, 2012.

Segura V., Vilhjálmsson B., Platt A., Korte A., Seren Ü., Long Q., and Nordborg M. An efficient multilocus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 2012.

Seung D. and Lee L. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13 :556–562, 2001.

Sisson S., Fan Y., and Tanaka M. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6) :1760–1765, 2007.

Slatkin M. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, pages 264–279, 1993.

Sokal R. and Oden N. Spatial autocorrelation in biology : 1. methodology. *Biological Journal of the Linnean Society*, 10(2) :199–228, 1978.

Sokal R. and Rohlf F. Biometry (3rd edn). *WH Freman and company : New York*, 1995.

Sokal R. and Wartenberg D. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics*, 105(1) :219–237, 1983.

Storfer A., Murphy M., Spear S., Holderegger R., and Waits L. Landscape genetics : where are we now ? *Molecular Ecology*, 19(17) :3496–3514, 2010.

Tavaré S. Ancestral inference in population genetics. *Lectures on probability theory and statistics*, pages 1931–1931, 2004.

Templeton A. Coherent and incoherent inference in phylogeography and human evolution. *Proceedings of the National Academy of Sciences*, 107(14) :6376, 2010.

Tipping M. and Bishop C. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2) :443–482, 1999.

Tobler W. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46 : 234–240, 1970.

Wegmann D., Leuenberger C., and Excoffier L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182 :1207–1218, 2009.

Weir B. and Cockerham C. Estimating F-statistics for the analysis of population structure. *Evolution*, pages 1358–1370, 1984.

Womble W. Differential systematics. *Science*, 114(2961) :315–322, 1951.

Wright S. Isolation by distance. *Genetics*, 28(2) :114, 1943.

Yu J., Pressoir G., Briggs W., Bi I., Yamasaki M., Doebley J., McMullen M., Gaut B., Nielsen D., Holland J., and others . A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2) :203–208, 2005.