Sélection objective de variables à l'aide d'algorithmes génétiques ensachés

Mémoire présenté par

Maxime Larocque

Comme exigence partielle de la

Maîtrise ès sciences de la gestion,

spécialisation

Intelligence d'affaires

À l'attention distinguée de

Jean-François Plante, Ph.D. [1]

et

Michel Adès, Ph.D. [2]

(1). Département des Sciences de la décision, HEC Montréal,
(2). Département de Mathématiques, Université du Québec à Montréal.

À Louis & Ève

Table des matières

Cette page est intentionnellement vide.

I Préface

Lecteurs, amis, collègues scientifiques, vous vous apprêtez à cet instant à découvrir cette invraisemblable et indescriptible *chose*, cet insondable mystère et cet improbable rêve que les hautes institutions académiques de ce monde sont convenues d'appeler *mémoire*. Ce dernier s'avère, soulignons-le, le fruit de nombreuses et solennelles années de labeur acharné, à défricher le dos courbé les étendues à peine explorées de la science. Cette recherche, cette vertigineuse quête, dis-je bien, se révéla bien rapidement pour moi un exercice des plus personnel, voir identitaire, où s'entremêla savamment la froide rigueur de la méthode scientifique aux tumultes orageux de l'intuition.

Laissez-moi ainsi, je vous en prie, vous esquisser l'implausible concours de circonstances qui a pu donner lieu à cet ardent désir qui me possédait de poursuivre ces études supérieures.

À l'apogée des années deux-mille, alors à l'emploi d'auxiliaire d'enseignement et au service d'un insolite professeur du nom de Matthieu Dufour [(3)], l'occasion me fut donnée de diriger ce qui d'ordinaire n'aurait été qu'un banal travail pratique. Or, cette excentricité mathématique s'avéra comme la chute de petites pierres qui sans bruit déclenchent l'avalanche en montagne. Bien malin aurais-je alors été de me douter que cette fugace épiphanie m'élancerait sur une quasi-décennie d'envoûtantes explorations mathématiques.

À l'occasion de ce travail noté, M. Dufour exigea donc de ses étudiants qu'ils résolvent, avec pour seuls outils l'ordinateur et leurs facultés mentales, mais autrement libre de toute contrainte, un problème qu'il m'affectionne particulièrement de surnommer le *problème de* 30π. Le défi consistait à déterminer le sous-ensemble de la séquence de nombres $\sqrt{1}, \sqrt{2}, \ldots, \sqrt{50}$ dont la somme serait la plus près de l'éponyme valeur de 30π. Alors qu'il s'avéra relativement simple de résoudre le problème de manière satisfaisante, la vérité est qu'il me fallut y consacrer de nombreuses années d'épisodiques offensives avant d'enfin réussir à le résoudre de manière complète et *absolue*. De péripéties en péripéties, il me fallut ainsi emprunter une myriade de chemins détournés au travers desquels je fus exposé pour la toute première fois aux algorithmes génétiques [11][17], lesquels constituent, et c'est bien là le point, une des pièces centrales de l'article autour duquel ce mémoire s'articule [15]. Ainsi, de ce frivole problème, de prime abord sans but, put germer des années plus tard l'idée d'approcher sous un angle novateur un problème tout ce qu'il y a de plus classique — la sélection de variables. C'est bien là d'ailleurs la morale de cette histoire : *un exercice en apparence sans lien saura parfois mettre en défaut les mécanismes naturels de la pensée causant en quelque sorte un effondrement intellectuel, lequel*

(3). Matthieu Dufour, professeur régulier à la section Actuariat du Département de Mathématiques de l'Université du Québec à Montréal.

éveille un aspect non-prémédité et non-causal de la raison, et par l'entremise de brusques et saccadés sauts logiques, nous amène à un angle d'approche autrefois hors d'atteinte. N'est-ce pas là une des manifestations la plus pure et authentique de l'intelligence qui soit ? (4)

Tout comme j'ai tenté de le démontrer ici, cette histoire, mon histoire, est inextricablement redevable aux contributions préalables de tout un assortiment de pionniers, grands et petits, savants, mathématiciens, enseignants ou encore simples étudiants. C'est ainsi, pleinement conscient d'être à certains égards l'héritier de cette prestigieuse lignée, qu'il me faut tout d'abord remercier enseignants et professeurs, notamment Alec Coulombe, Robert Guay, feu Normand Guillet, Matthieu Dufour, Serge Alalouf, Pierre Hansen et Denis Larocque, pour ne nommer que ceux-là ; en outre, je remercie bien entendu famille et amis, lesquels ont su croire en moi en dépit de la trajectoire parfois sinueuse de mon parcours ; je remercie haut et fort ma conjointe pour son inconditionnel support, mais surtout en raison de son incommensurable compréhension face à l'accaparante présence qu'occupe les mathématiques dans ma vie ; enfin, il me faut vivement et profondément remercier mes très nobles et très honorables directeurs de thèse ; MERCI à Michel Adès, chevalier des lettres, pour son amour de la rigueur, mais surtout pour le bienveillant aplomb de son moral, placide et serein, face au caractère obstiné et à l'occasion incendiaire de mes interventions ; MERCI à Jean-François Plante pour la lucidité grâce à laquelle son regard jamais ne s'écarte de l'horizon, qu'importe l'abysse mathématique au fond de laquelle nous nous étions embourbés, et pour la sagacité et la constante vigilance de son esprit. Merci donc à tous ces gens qui ont contribué à faire un scientifique de l'artiste en moi, mais aussi un artiste du scientifique. Sans vous, il ne m'aurait jamais été donné de me remettre en question, certes face aux autres, mais avant toute chose face à moi-même.

De même, au sein de cet ouvrage, j'ai tâché humblement, au meilleur de mes connaissances et de mes capacités, de consigner une vérité « *sub specie æternitatis* » (5), c'est-à-dire une vérité objective et intemporelle qui a valeur au-delà des confins de la culture, de la langue et du contexte d'usage. Pour reprendre les paroles de Moriarty (6), j'espère ainsi de tout cœur que ce mémoire saura vous apparaître franc et sans réserve, riche et généreux, et dont la valeur n'est point dissimulée à l'arrière-plan, mais se fond plutôt librement à l'avant-scène.

ENFIN, laissons-nous sur une courte citation de Nassim Nicolas Taleb [22] (7) et tâchons alors d'apprécier, autant que faire se peut, cette infinie beauté mathématique qui sans le sacrifice des hommes de sciences jamais n'aurait pu aujourd'hui nous être révélée : « *If you study everyday, you expect to learn something in proportion to your studies. If you feel that you are*

(4). Aussi, que cela fusse prémédité ou pas et malgré son caractère inorthodoxe, nous nous devons de reconnaître le génie sans pareil de la pédagogie dont sut faire preuve M. Dufour en confrontant ses étudiants à ce problème.
(5). Expression attribuée à Spinoza [21] peut être traduite par *sous l'aspect de l'éternité.*
(6). Brian Moriarty, concepteur de jeux vidéos américain et professeur au Worcester Polytechnic Institute au Massachusetts.
(7). Nicolas Taleb, professeur d'ingénierie du risque et épistémologue des probabilités à la *New York University* (NYU).

not going anywhere, your emotions will cause you to become demoralized. But modern reality rarely gives us the privilege of a satisfying, linear, positive progression: you may think about a problem for a year and learn nothing; then, unless you are disheartened by the emptiness of the results and give up, something will come to you in a flash. [...] There are routes to success that are nonrandom, but few, very few, people have the mental stamina to follow them... Most people give up before the rewards. » (8)

Maxime Larocque,
Laval, décembre 2018

(8). Un corollaire découlant naturellement de cette proposition et on ne peut plus applicable à ces recherches est, pour reprendre les mots de Dorian Chandelier (dans sa série ϕ disponible sur le site web `http://www.nesblog.com`) que « *parfois pour trouver, il faut arrêter de chercher, ce qui est en soit une position paradoxale* ».

II Introduction

En cette ère de la révolution technologique, de nombreux milieux où l'accès aux données fut jadis limité se voient aujourd'hui, bien au contraire, submergés au beau milieu d'un océan numérique, plus que jamais assoiffés de connaissances.

De ce nouveau contexte put naître la *sélection de variables*, un pilier central de la modélisation prédictive. En effet, la spécification du modèle se compose généralement de deux parties : le choix d'une forme fonctionnelle permettant de lier variables explicatives à variable réponse, d'une part, ainsi que le choix de ces mêmes variables explicatives, d'autre part. Ainsi, dans le contexte d'un ensemble de variables explicatives toujours grandissant et où l'expression *embarras du choix* prend tout son sens, que ce soit pour le chercheur ou le praticien, on peut comprendre que la sélection des variables s'est rapidement vue devenir un enjeu d'envergure, au point de se voir affubler l'épithète de *problème de la sélection des variables* [8].

La prémisse des recherches ayant certes culminées en ce mémoire, mais plus encore en un article intitulé *Bagged genetic algorithms for objective model selection* [15], fut ainsi celle d'explorer la possibilité de mieux répondre à ce problème d'une manière automatisable, et ce, notamment en considération des difficultés auxquelles sont confrontées les méthodes de résolution actuelles. À l'instar de cet article que le mémoire présente et met justement en valeur, une relative familiarité avec les concepts élémentaires de la statistique, de la régression et plus spécifiquement de la régression linéaire sera pour ainsi dire tenue pour acquise.

À cet égard, nous commencerons tout d'abord par introduire la question sur la base de la régression linéaire. Comme stipulé, nous nous contenterons de décrire cette dernière par l'entremise de ces quelques mots : *la régression linéaire est une mouture particulière de la modélisation où la tendance conditionnelle d'une variable réponse* $\boldsymbol{Y}$ (*l'espérance ou la moyenne,* $\mathrm{E}\left[\boldsymbol{Y} \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right]$) *est caractérisée par une combinaison linéaire de variables exogènes, où le proverbial bruit,* ε*, est supposé indépendant, homoscédastique et gaussien.* Fort de ce contexte, il devrait désormais nous être plus aisé, avant même d'aborder le comment de la question, d'en discuter le pourquoi. Nous entendons par là qu'il nous apparaît nécessaire et judicieux de mieux définir les raisons justifiant le besoin même de procéder à la sélection des variables. Considérons ainsi les éléments de réponses suivantes :

i. **Surparamétrisation** : L'inclusion de paramètres superflus entraîne la perte de degrés de liberté, ce qui a pour conséquence de déformer le portrait que dresse toute mesure d'ajustement évaluée *sur-échantillon*[9] — par opposition à une évaluation *hors-échantillon*[10]

(9). De l'anglais, « in-sample ».
(10). De l'anglais, « out-of-sample ».

qui, elle, consisterait plutôt à mesurer le pouvoir prédictif du modèle à l'aide d'observations non-contaminées par le procédé d'ajustement ;

ii. **Biais pour variables omises** : L'omission de variables explicatives utiles (ainsi que l'inclusion de variables excédentaires) donne lieu à un phénomène appelé *biais pour variables omises* [4] sous lequel l'erreur quadratique moyenne des coefficients de la régression est exacerbée par l'introduction d'un biais statistique et d'une variance accrue ;

iii. **Effet de substitution** : Un ensemble de variables explicatives multicolinéaires donnera lieu à un effet dit de *substitution* ou encore de *proxy* où certains groupes de variables fortement corrélées entre-elles (rapportant en quelque sorte une seule et même information) agiront à titre de substitut les unes pour les autres, remplaçant essentiellement une variable hautement significative par un plus grand nombre de variables de moindre significativité ;

iv. **Interprétabilité** : un modèle excessivement complexe sera plus difficile à interpréter et l'influence d'une variable spécifique plus difficile à cerner et justifier.

Historiquement, l'exercice de la sélection des variables s'est accompli au moyen de méthodes dites *heuristiques* (nommément les méthodes ascendante, descendante et bidirectionnelle) initialement développées au milieu du siècle dernier par Efroymson [18] des laboratoires de recherche Esso, sous l'éminente assistance de Tukey. Alors que ces techniques représentaient pour l'époque une innovation sans précédent, celles-ci étant d'ailleurs encore souvent enseignées et utilisées de nos jours, il n'en demeure pas moins qu'elles souffrent de nombreuses faiblesses lesquelles émanent principalement de deux caractéristiques :

— **Déterminisme** : Une méthode sera qualifiée de déterministe si son extrant se trouve entièrement déterminé par ses intrants, c'est-à-dire dont le fonctionnement ne repose en aucune partie sur l'aléatoire. Cela a pour conséquence de figer les aléas naturellement présents dans l'échantillon et d'ainsi créer un biais systématique, lequel rend beaucoup plus difficile le départage des signaux du bruit ;

— **Gloutonnerie** : Un algorithme d'optimisation itérative opérant sous la contrainte d'optimiser à tout coup son objectif, c'est-à-dire d'*itération en itération*, est qualifié de *glouton*. Cela est susceptible de porter atteinte à sa performance en faisant de chaque optimum local un état absorbant, c'est-à-dire se faisant prisonnier des optimums locaux de par son refus de sacrifier à court terme (par là, on entend de s'éloigner de son objectif d'optimisation, de procéder à des mouvements sous-optimaux) au profit d'un gain à long terme.

Dans cette optique, une alternative naturelle fut pour nous de faire volte-face pour ainsi focaliser notre attention sur des méthodes plutôt *stochastiques*. Notre choix s'arrêta sur la gamme de

techniques évolutives que l'on appelle *algorithmes génétiques*. S'inscrivant dans le domaine du biomimétisme, ces dernières consistent à générer au hasard une population de solutions candidates pour ensuite leur permettre d'évoluer. Cette évolution se déroule en séquence, c'est-à-dire de génération en génération, et ce, au moyen d'opérateurs dits *génétiques* dont le fonctionnement est étroitement lié à une vision darwinienne de l'évolution des espèces, nommément la *sélection*, le *croisement* et la *mutation*.

Sous le contexte du problème de la sélection des variables, une solution candidate correspondra à un modèle candidat, où l'inclusion de chacune des variables explicatives candidates sera encodée au moyen d'un indicateur binaire, un bit, formant ainsi de manière conjointe ce que l'on appelle une *chaîne binaire* ou encore *vecteur binaire* [(11)].

De cette population de modèles candidats $\left\{ B_{igu} \,\middle|\, 1 \leq i \leq I \right\}$, la population des *parents*, deux individus ou parents, B_{i_1gu} et B_{i_2gu}, sont alors sélectionnés de manière proportionnelle à leur *score*, c'est-à-dire $\Pr[B_{igu} = B] \propto f_u(B)$, lequel correspond à une mesure fixée par l'utilisateur — la fonction d'ajustement f — de son pouvoir prédictif ou ajustement. Dans le but de créer une nouvelle population prédestinée à supplanter celle des parents, la population des *enfants*, les parents ainsi sélectionnés sont alors recombinés au hasard de manière à engendrer un nouveau modèle, c'est-à-dire l'embryon $B^*_{i(g+1)u}$. Une fois muté de manière aléatoire, celui-ci pourra devenir enfant. Enfants et parents se voient alors retournés à leurs populations respectives, et le processus répété jusqu'à l'obtention d'une population d'enfants et de parents de même taille. La population des parents peut alors être remplacée par celle des enfants, marquant ainsi le passage d'une génération, et permettant par le fait même au bassin de population d'évoluer. Après le passage d'un nombre fixe et prédéterminé de générations, le *meilleur* individu, c'est-à-dire l'individu le mieux ajusté, possédant le plus grand pouvoir prédictif, est retourné par l'algorithme en guise de solution.

On comprend ainsi d'où ces algorithmes tirent leur nom et en quoi ces derniers peuvent être qualifiés d'évolutifs : l'ensemble de leur fonctionnement tente aussi bien que faire se peut de reproduire les mécanismes de la sélection naturelle énoncés par Darwin dans son illustre ouvrage *The Origins of Species* [5].

Draper et Fouskakis [7] furent parmi les premiers à explorer l'applicabilité des algorithmes génétiques au problème de la sélection des variables. Alors que leur constat en fut un d'échec, une lueur d'espoir surgit à l'horizon avec la publication d'un article intitulé *Darwinian Evolution in Parallel Universes*: *A Parallel Genetical Algorithm for Variable Selection* [25]. Dans ce dernier,

(11). Par exemple, si un total de trois variables explicatives candidates étaient disponibles, et que seules la première et la troisième étaient incluses à un modèle, ce dernier serait décrit par la chaîne binaire $B = \begin{bmatrix} 1 & | & 1 & 0 & 1 \end{bmatrix}^T$, en raison de la colonne pour interception, laquelle est toujours incluse et figure au premier rang.

les auteurs surent adroitement développer et exploiter une stratégie novatrice selon laquelle une multitude d'algorithmes génétiques seraient évalués en parallèle et leurs résultats recombinés en un unique modèle, élu par la majorité, à l'instar des méthodes d'ensemble. Bien que prometteuse, cette nouvelle méthodologie n'en demeurait pas moins en proie à plusieurs difficultés :

1. **Subjectivité** : La méthode, reposant sur l'appréciation visuelle d'un graphique, est assujettie à la subjectivité de l'expérimentateur et, à cet effet, sans fondement statistique quel qu'il soit ;
2. **Non-automatisation** : La méthode, reposant sur l'intervention humaine de l'expérimentateur, ne pouvait être automatisée ;
3. **Non-calibration** : La méthode, n'explicitant pas de manière quantifiable et analytique l'effet des paramètres d'entrée sur sa sortie, fait de son calibrage (choix du taux d'activation, de la probabilité de mutation, de la taille des populations, du nombre d'univers, et à certains égards le choix de la fonction d'ajustement, etc.) une entreprise difficile, subtile et, encore une fois, assujettie à la subjectivité de l'expérimentateur.

Une fois de plus, les méthodes heuristiques nous étaient révélées comme n'étant guère mieux que d'alambiquées *règles du pouce* dont l'avantage premier était celui de rendre les choses plus faciles à comprendre et implanter. Leurs utilisateurs ne peuvent ainsi y trouver utilité qu'en reconnaissant l'existence des limites qu'engendrent et imposent leurs imperfections et prenant bien garde d'être bernés par le pouvoir que ces dernières apparaissent conférer. Leur utilisation risque alors de s'avérer capricieuse, voir dangereuse, lorsqu'on ignore cette réalité ou que l'on choisit délibérément d'en faire fi.

Était-ce là la fin ? Loin de nous laisser freiner par ces obstacles, nos recherches au contraire se donnèrent précisément comme objectif de pallier à ces problèmes et de corriger la classe de méthodes en l'inscrivant fermement au sein d'un rigoureux cadre statistique. De même, nos recherches nous menèrent à explorer les questions suivantes : quelle est précisément l'hypothèse nulle ? Sur la base de quelle statistique de test devrions-nous tester cette dernière ? Comment dériver la distribution sous H_0 de cette même statistique du test et comment alors fixer le niveau de son erreur de type I [(12)] ? Comment étendre nos résultats à une gamme plus variée de modèles tels que les modèles linéaires généralisés ? Chacune de ces questions trouve réponse dans l'article, lequel est présenté à la prochaine et troisième section du mémoire.

Sur une note davantage récapitulative, la section IV et la section V approfondissent et détaillent certains des aspects plus subtils de l'article. Enfin, la section VI conclut la dissertation,

(12). L'erreur de type I correspond à la probabilité de faux positif, tandis que l'erreur de type II correspond à la probabilité de faux négatif.

fait la synthèse des différentes idées abordées tout au long du mémoire et esquisse quelques ouvertures et réalisations découlant naturellement de l'article. Une liste des notations est donnée à la section VII.

III Algorithmes génétiques ensachés

Le lecteur est avisé que le présent chapitre se trouve être, exception faite de cette note éditoriale, une reproduction quasi-exacte de l'article dont fait référence la section précédente et la culmination même de ces années de recherche : « *Bagged parallel genetic algorithms for objective variable selection* » [15] [(13)]. Ledit article est, à l'heure où ces lignes sont écrites, soumis à la revue scientifique *Journal of the American Statistical Association*. À cet égard, il convient de noter que le contenu de ce dernier :

(a) est intégralement rédigé en anglais ;

(b) comporte sa propre liste de références (incluse à sa fin), lesquelles peuvent s'entrecouper avec celles présentées en fin du présent ouvrage ;

(c) contient une notation pouvant à l'occasion diverger de celle utilisée dans le détail des preuves présentées à la section IV, c'est-à-dire là où la fluidité des explications s'y portait ;

(d) possède un format (police, marges, etc.) se distinguant du présent ouvrage, et n'ayant pas été altéré dans le seul but de reproduire le plus fidèlement possible son format d'origine.

(13). Le terme « *ensaché* » se trouve être une traduction libre du nom de la technique statistique dite du « *bagging* ».

Bagged parallel genetic algorithms for objective model selection

Maxime Larocque
HEC Montréal, Montréal, Canada.
Jean-François Plante
HEC Montréal, Montréal, Canada.
E-mail: jfplante@hec.ca
Michel Adès
Université du Québec à Montréal, Montréal, Canada.

Abstract. Genetic algorithms are used for feature selection through a fitness function that drives the evolution of populations. With parallel universes, an importance score may be produced for each feature to determine from a plot which to retain. The authors derive the distribution of those importance scores under the null hypothesis that none of the features have predictive power and they determine an objective threshold for feature selection. The authors discuss the parameters for which the theoretical results hold. They illustrate their method on real data and run simulation studies to describe its performance.

Keywords: Feature selection; generalized linear models; genetic algorithms; linear models; machine learning; model selection; predictions.

1. Introduction

In predictive models, feature selection can boost performance and help avoid overfitting. Classical approaches such as stepwise regression, and all subset selection based on penalized statistics such as AIC and BIC are routinely taught (see e.g. Draper and Smith, 2014). Penalized regression methods such as the lasso select features and fit a model simultaneously (see e.g. Tibshirani, 1996). Genetic algorithms have also been proposed (see e.g. Chatterjee et al., 1996), including a parallel version by Zhu and Chipman (2006).

Biomimetics consist in imitating nature to solve complex human problems. For instance, genetic algorithms are inspired from natural selection. For feature selection, models are seen as individuals whose DNA is a vector of binary markers that indicate whether or not each feature is used in the model. Successive generations of models are generated by selecting among the fittest parents of the current population, and creating offsprings by selecting genes randomly from both parents, a step called crossover. To improve the richness of the population, models may undergo mutation where genes are subject to being flipped. While the general steps of a genetic algorithm are clear, the actual recipe for each is chosen from numerous options.

Genetic algorithms often involve a single population that evolves until the fittest individual is found. A generalization proposed by Zhu and Chipman (2006) considers a fixed number of parallel populations that evolve independently of one another.

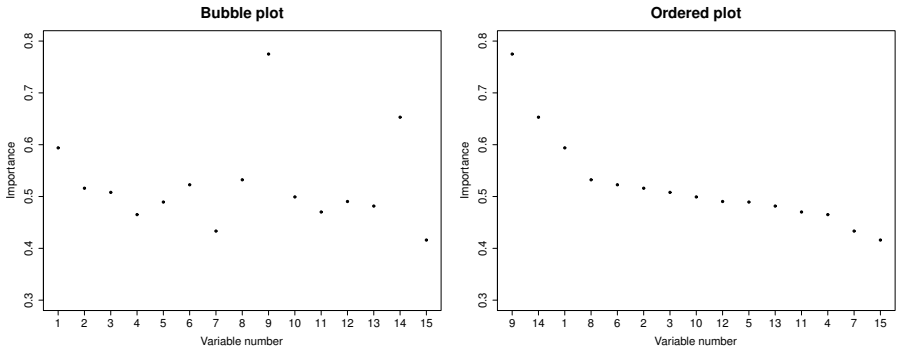

Figure 1. Example of a bubble plot and its associated ordered plot as suggested by Zhu and Chipman (2006). The importance of each feature is measured by its prevalence among the individuals of the last generation in a parallel genetic algorithm.

After a number of steps that could be predetermined, the frequency of a feature in the last generation of all populations is used as a measure of its importance. A bubble plot allows to visually identify where the importance drops to separate important features from those that should be dismissed.

To illustrate the bubble plot, let us consider the Pollution dataset of McDonald and Schwing (1973) where a measure of mortality is predicted with linear regression from 15 features on 60 data points. The left panel of Figure 1 shows a bubble plot, a tool proposed by Zhu and Chipman (2006). The associated ordered plot is shown on the right panel. The ordered plot is identical to the bubble plot, except that the features are ranked in order of decreasing importance. The figure suggests the selection of features x_1, x_9 and x_{14}. Although Zhu and Chipman (2006) analyzed the Pollution dataset, Figure 1 was generated with a different genetic algorithm hence our figure differs from theirs. While the bubble plot is interesting and intuitive, it requires a visual inspection, which makes the method subjective and limits its automation as a human intervention is required to complete the selection.

To make parallel genetic algorithms more objective, we derive the distribution of the importance score of every feature under the null hypothesis that none of the features are good predictors of the response. To increase the richness of the parallel populations and their ability to detect truly important features, we also add a step of bagging, providing every parallel universe with a bootstrap sample of the data. The benefits of adding noise for feature selection with genetic algorithms have already been explored by Zhang et al. (2015) and Wang and Zhang (2015). We use the null distribution of the importance to determine an objective threshold at the global level α. Drawing the bubble plot then becomes optional since feature selection can be automatized by comparing the importance of each feature against a given threshold.

Section 2 of the paper introduces the basics of genetic algorithms for model selection as well as some notation. Our method which we call Bagged Genetic Algorithm (BGA) is described in Section 3 along with the mathematical results that support it. Case studies and Monte Carlo simulations are analyzed in Section 4. Section 5 discusses the choice of parameters in a bagged genetic algorithm. Section 6

offers a short conclusion.

2. Genetic algorithms for model selection

Genetic algorithms are meta-heuristic optimization techniques mimicking a Darwinian vision of evolution in order to solve complex, numerically intractable problems (see e.g. Shonkwiler and Mendivil, 2009). The general functioning of genetic algorithms consists of the successive application of three genetic operators – selection, crossover and mutation – to an initial population of candidate solutions or individuals. For model selection, each individual is encoded using a binary string which indicates whether or not each feature is present in the model. At every generation, the likelihood of an individual generating an offspring depends on its fitness, and the process is iterated a number of times.

We first describe the genetic operator in the more classic setting of a single-thread genetic algorithm, where the evolution of only one population is generated. We then extend the notation to the parallel universes as proposed by Zhu and Chipman (2006).

2.1. Single-thread genetic algorithm (SGA)

In a single-thread algorithm, individual $i \in \{1, \ldots, I\}$ of generation $g \in \{1, \ldots, G\}$ is represented by the binary vector $B_{ig} = [b_{1ig}, \ldots, b_{Dig}]$ where b_{dig} is one when feature $d \in \{1, \ldots, D\}$ is active for this individual and zero otherwise. The number of individuals in the population is usually held fixed from generation to generation. A typical application for model selection is to consider linear regression where each explanatory variable is included or not in the model as encoded by b_{dig}.

2.1.1. Initial population

The initial population of individuals is generated randomly. The b_{di0} are hence drawn from independent Bernoullis with probability of success π_0, the activation rate.

2.1.2. Fitness

In a general genetic algorithm, the fitness function $f(B_{ig})$ measures the ability of an individual to "solve the problem". In a model selection setting, f should measure the predictive ability of the corresponding model. An ideal measure of fitness should not be biased by the number of features present in a model. Zhu and Chipman (2006) used a leave-one-out cross-validated residual sum of squares, but we prefer to use a validation subset to evaluate a properly scaled residuals sum of squares (RSS). Since better individuals should have a larger fitness, we use a negative power of our rescaled RSS.

2.1.3. Selection

At generation g, I individuals are available to become parents. To create an offspring, two individuals are selected randomly, with replacement (parthenogenesis is allowed). Zhu and Chipman (2006) determine a survival pool among which all

individuals have the same probability of being selected as parents. The survivors also make it directly to the following generation, an operation called elitism. We rather implement another popular selection operator where the probability of selection of an individual is proportional to its fitness. We do not implement elitism: all members of a generation are offsprings born of the previous generation. Namely, for g fixed, the probability of selection of individual i' is $f(B_{i'g})/\{\sum_{i=1}^{I} f(B_{ig})\}$.

2.1.4. Crossover

Crossover determines how the genes of the two selected parents are combined to generate one offspring. Zhu and Chipman (2006) used one point crossover where an integer is chosen randomly between 1 and $D-1$. The genes of the father are used up to that integer, and the genes of the mother are used for the rest. One disadvantage of this approach is that the arbitrary order in which the variables are coded has an influence on the models that are possible to generate. We have a preference for uniform crossover where each gene of the offspring is taken randomly from its mother or father. Let $\mathbf{C}_{ig}$ be a diagonal matrix where the diagonal entries are independent Bernoullis with probability of success 1/2. Then if individuals i_1 and i_2 of generation g are selected as the parents of individual i in generation $g+1$, the crossover will first generate the embryo

$$B^*_{ig} = \mathbf{C}_{ig} B_{i_1 g} + (\mathbf{I} - \mathbf{C}_{ig}) B_{i_2 g}$$

where $\mathbf{I}$ is the $D \times D$ identity matrix. The embryo B^* will become an individual in generation $g+1$ once the mutation step is complete.

2.1.5. Mutation

Mutation helps maintaining genetic diversity by making each gene of the embryo subject to a random flip. Let $\mathbf{M}_{ig}$ be a diagonal matrix whose diagonal is filled with independent Bernoulli random variables with parameter θ_g, the probability of a mutation, which could vary as generations evolve or be held fix. A decreasing θ_g, for instance, is akin to simulated annealing (see e.g. Kirkpatrick et al., 1983). Individual i of generation $g+1$ may then be obtained as

$$B_{i(g+1)} = (\mathbf{I} - \mathbf{M}_{ig}) B^*_{ig} + \mathbf{M}_{ig}(\mathbf{1} - B^*_{ig}).$$

2.1.6. Evolution

The process described is repeated recursively. While it is possible to iterate until a convergence criterion is met, we suppose a fixed number of generations, G. For a single-thread algorithm, the fittest individual of the last generation, $\arg\max f(B_{iG})$, is usually outputted as the solution.

As a meta-heuristic algorithm, there exists a very large number of variants for the genetic operators to which additional parameters may also be added. Our description of the operators is focused towards the method that we develop, however readers who are interested in learning more about the numerous uses of genetic algorithms and many variants of the genetic operators may consult, e.g. Mitchell (1998), Cantú-Paz (2000), Haupt and Haupt (2004) or Poli et al. (2008).

2.2. Parallel genetic algorithms (PGA)

Zhu and Chipman (2006) suggest to create U universes in which populations evolve in parallel. New indices need to be added to the notation introduced previously to account for the universe. Namely,

Genes: $B_{igu} = [b_{1igu}, \ldots, b_{Digu}]$ is a binary vector with the genetic code of individual i of generation g in universe $u \in \{1, \ldots, U\}$.

Fitness: Since we use out-of-sample validation, the fitness function in parallel universes will be based on different hold-out datasets. The notation f_u shows this dependence of the fitness function on the universe.

Selection, crossover and mutation: The genetic operators are identical in the parallel universes. They are applied independently in each of the parallel populations, from initial generations that are created then evolved in the same fashion previously described.

With parallel worlds, the output is not a single fittest individual, but an importance score based on the frequency of each gene in the final generation of all populations. For feature d, the importance score is the proportion

$$\hat{\pi}_d = \frac{1}{IU} \sum_{i=1}^{I} \sum_{u=1}^{U} b_{diGu}$$

and we note $\hat{\boldsymbol{\pi}} = [\hat{\pi}_1, \ldots, \hat{\pi}_D]$ the vector of those values for all features. The bubble and ordered plots of Figure 1 are a graphical representation of $\hat{\boldsymbol{\pi}}$.

In the next section, we describe how to derive a threshold for the values of $\hat{\boldsymbol{\pi}}$ to determine which features should be retained, and which should be dismissed.

3. Bagged genetic algorithms (BGA)

Using PGA for feature selection yields importance score for the variables rather than a single solution as does SGA. The final decision of which variables to include is however based on the visual inspection of a figure. The purpose of this paper is to offer an objective and automatizable way to make that final decision.

Let us focus on a linear model setting where predictor d has a parameter β_d and where $\beta_d = 0$ means that the feature has no effect on the target. To establish a threshold value, we derive the distribution of $\hat{\pi}_d$ for $d \in \{1, \ldots, D\}$ under a null hypothesis that implies some symmetry between individuals. Namely, under H_0 : $\beta_1 = \cdots = \beta_D = 0$, the fitness of any individual is assumed to be approximately equal since they all have an equally low ability to predict the target variable. At the selection step, this means that all individuals are equally fit, hence have an equal probability of being selected. This hypothesis of symmetry is reinforced through careful choices in the design of the genetic algorithm which are described next.

3.1. Hold out sample and bootstrap

In each universe, the set of N data is split into a training and a validation set. The $\tilde{n}$ data from the training set are stored in the vector of response $\tilde{\mathbf{Y}}_u$, and the predictors in the $\tilde{n} \times (D + 1)$ design matrix $\tilde{\mathbf{X}}_u$ whose first column is filled with

ones to account for the intercept. The rest of the data becomes the validation set and we denote the corresponding values n, $\mathbf{Y}_u$ and $\mathbf{X}_u$. When working with a very large dataset, sample sizes may be chosen so that $n + \tilde{n} < N$, with each universe drawing possibly different data. For smaller samples, $n + \tilde{n}$ is likely equal to N, so we may use bootstrap instead to increase the diversity between universes. Sampling with replacement within training and validation sets then generates bootstrapped versions of the data (of the same sizes $\tilde{n}$ and n) that are also denoted by $\tilde{\mathbf{Y}}_u$, $\tilde{\mathbf{X}}_u$, $\mathbf{Y}_u$ and $\mathbf{X}_u$. The numerical investigation in Section 4 adopts this approach. Since we combine the output of all universes in the end, this is a form of bagging (see e.g. Breiman, 1996b) and is likely to boost the performance of the algorithm. This seems to occur for Zhang et al. (2015) and Wang and Zhang (2015) who add noise to the parallel universes in a PGA.

An additional benefit of bagging is the dilution of the spurious correlations that could occur by mere chance. Such correlations could be amplified through the generations of the genetic algorithm, but the bootstrapping makes it less likely to occur simultaneously for the same variables in multiple parallel universes.

3.2. Choice of fitness function

The derivation of the distribution of the importance scores in the next section is based on the assumption that the fitness of all models is approximately equal under the null hypothesis. Using out-of-sample validation helps yielding a fitness that is not influenced by the number of active features. The fitness we use is based on a rescaled sum of residuals arising from the validation sample, but other choices would also be acceptable.

Consider an arbitrary individual B with p active features, i.e. $||B||^2 = p$. Let $\mathbf{B}$ be a $(D+1)\times(p+1)$ matrix generated by removing some columns from the $(D+1)\times(D+1)$ identity matrix. Namely, column one always remains for the intercept, and the following are matched to the binary values in B, all columns associated to a zero being removed. The $\mathbf{B}$ matrix allows to select appropriate columns from the design matrices and the usual matrix notation for regression estimates may be adjusted by adding a $\mathbf{B}$ besides every design matrix. In universe u, the estimate of the regression slope parameters are then given by $(\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{X}}_u\mathbf{B})^{-1}\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{Y}}_u$ that may then be used to calculate out-of-sample predictions

$$\hat{\mathbf{Y}}_u(B) = \mathbf{X}_u\mathbf{B}(\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{X}}_u\mathbf{B})^{-1}\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{Y}}_u,$$

yielding the unscaled residuals $\hat{\boldsymbol{\varepsilon}}(B) = \mathbf{Y}_u - \hat{\mathbf{Y}}_u(B)$. By the independence of the error in the training and the validation datasets, this vector of residuals is a multivariate normal with mean zero and covariance

$$\Sigma_u(B) = \sigma^2\left\{\mathbf{I} + \mathbf{X}_u\mathbf{B}(\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{X}}_u\mathbf{B})^{-1}\mathbf{B}^\intercal\mathbf{X}_u^\intercal\right\} \tag{1}$$

where $\mathbf{I}$ is the $n\times n$ identity matrix. Since the selection is based on the relative value of the fitness, the actual value of σ^2 is irrelevant in the calculation of the fitness function

$$f_u(B) \propto \left[\hat{\boldsymbol{\varepsilon}}(B)^\intercal\{\Sigma_u(B)\}^{-1}\hat{\boldsymbol{\varepsilon}}(B)\right]^{-\gamma} \tag{2}$$

where γ is a positive scalar that can enhance the peakedness of the function. We found empirically that $\gamma = 1.5$ seems to work well.

Since the residuals are asymptotically normal, the distribution of the expression inside the brackets in Equation 2 up to a multiplicative constant is chi-square with n degrees of freedom. Most importantly, the degrees of freedom do not depend on p, the number of active features.

REMARK 1. *Using the Woodbury matrix identity and some algebra, Equation 1 may be written under the equivalent form*

$$\Sigma_u(B) = \sigma^2 \left\{ \mathbf{I} - \mathbf{X}_u \mathbf{B} (\mathbf{B}^\intercal \tilde{\mathbf{X}}_u^\intercal \tilde{\mathbf{X}}_u \mathbf{B} + \mathbf{B}^\intercal \mathbf{X}_u^\intercal \mathbf{X}_u \mathbf{B})^{-1} \mathbf{B}^\intercal \mathbf{X}_u^\intercal \right\}^{-1},$$

which may be used to evaluate Equation 2 with one less matrix inversion. This result is similar in nature to Problem 7.3a of Friedman et al. (2001).

There are numerous other possible fitness functions in the literature. As long as a fitness function is sufficiently unbiased to offer an approximately equal fitness to all models under the null hypothesis, then the results of the next section should hold.

3.3. Distribution under the null hypothesis

In generation $g = 0$, all genes are generated at random, hence yielding DIU independent Bernoulli variates with probability of success π_0. As the populations evolve, these genes are recombined into new offsprings. We look into the distribution of those genes at generation G, when the evolution of all populations is stopped.

Note that under the null hypothesis, the expected value of the genes of an embryo B^* is equal to that of its parents since all generated genes have the same probability of being selected. The mutation step, however, has an effect on the expected proportion of ones. Fixing d and u, then conditioning on the presence of a mutation, we get the recurrence

$$\pi_g = \mathrm{E}(b_{digu}) = (1 - \pi_{g-1})\theta_g + \pi_{g-1}(1 - \theta_g)$$

that may be written alternatively as $\Delta_{g+1} = (1-2\theta_g)\Delta_g$ if we define $\Delta_g = 1/2 - \pi_g$. Iterating yields $\Delta_G = \Delta_0 \prod_{g=1}^{G} (1 - 2\theta_g)$, or

$$\pi_G = \frac{1}{2} \left\{ 1 - (1 - 2\pi_0) \prod_{g=1}^{G} (1 - 2\theta_g) \right\} \tag{3}$$

which tends to $1/2$ as $G \to \infty$ for most choices of sequence $\theta_g \in (0, 1/2)$. In practice, the values of θ_g are likely to be kept away from 0 or $1/2$, yielding a product that diverges to 0 as $G \to \infty$. Other limits are however possible if θ_g tends to zero fast enough to make the product converge.

REMARK 2. *If the probability of mutation is fixed for all generations such that $\theta_g = \theta$, then equation 3 simplifies into $\pi_G = \{1 - (1 - 2\pi_0)(1 - 2\theta)^G\}/2$ and does indeed converge to $1/2$ as $G \to \infty$ if $\theta > 0$.*

Since there is no exchange between parallel universes, bits b_{diGu_1} and b_{diGu_2} will always be independent when they come from different universes. By the virtues of uniform crossover as well as the assumption that all models are equally likely to be selected, genes in different positions, b_{d_1iGu} and b_{d_2iGu} should also be uncorrelated. For fixed d and u, two different individuals may however be correlated as they are

likely to share common ancestors. In the process of creating generation $g = 1$, we may condition on having inherited gene d from a common parent to get $c_0^* = \text{cov}(b^*_{di_1 0u}, b^*_{di_2 0u}) = \pi_0(1-\pi_0)/I$ for the embryo, or more generally, when creating embryos from parents of generation g,

$$c_g^* = \text{cov}(b^*_{di_1 gu}, b^*_{di_2 gu}) = \frac{1}{I} v_g + \frac{I-1}{I} c_g \tag{4}$$

where $v_g = \text{var}(b_{digu}) = \pi_g(1-\pi_g)$ and $c_g = \text{cov}(b_{di_1 gu}, b_{di_2 gu})$ for arbitrary i, d and u, and with $c_0 = 0$ for the initial population. Let m_{digu} identify the diagonal elements of $\mathbf{M}_{igu}$, and note for simplicity $m_1 = m_{di_1 gu}$, $b_1^* = b^*_{di_1 gu}$ and similarly for m_2 and b_2^* of individual i_2. We can then write explicitely

$$\begin{aligned} c_{g+1} &= \text{cov}\{(1-m_1)b_1^* + m_1(1-b_1^*), (1-m_2)b_2^* + m_2(1-b_2^*)\} \\ &= (1-2\theta_g)^2 c_g^* \end{aligned} \tag{5}$$

after simplifications due to the independence of the mutation binary markers and properties of the covariance. Substituting Equation 4 in 5 yields the recurrence

$$c_{g+1} = (1-2\theta_g)^2 \left\{ \frac{1}{I}\pi_g(1-\pi_g) + \frac{I-1}{I} c_g \right\}.$$

Despite their unwieldy expressions, the sequences c_g and π_g are easy to determine numerically with a simple loop.

The decision of which features to retain is made from the importance scores, $\hat{\pi}_d$ which can be seen as the average of U independent and identically distributed random variables. As the number of parallel universes U increases, the central limit theorem (see e.g. Casella and Berger, 2002) guarantees the convergence of $\sqrt{U}(\hat{\pi}_d - \pi_G)$ to a normal distribution with mean 0 and variance

$$\sigma^2_{\hat{\pi}} = \frac{1}{I}\pi_G(1-\pi_G) + \frac{I-1}{I} c_G.$$

The decision to retain any given variable may therefore be made from a normal quantile. Since all variables are simultaneously tested, and since their individual tests are uncorrelated, a Šidák correction (see Šidák, 1967) is applied to account for multiple comparisons. After setting a global level α, the user should thus keep those variables for which

$$\hat{\pi}_d > \pi_G + z_{1-\alpha_S}\sigma_{\hat{\pi}}/\sqrt{U}$$

where $\alpha_S = 1-(1-\alpha)^{1/D}$ is the Šidák corrected level.

Except for $\hat{\pi}_d$ itself, none of those values need to be estimated. They are deterministic functions of the parameters of the genetic algorithm. This result does not depend on the fitness function either, as long as it has the ability to make all models approximately equally fit under the null hypothesis. In particular, a constant fitness function will yield this result, but would display no ability for detecting relevant variables when they are present. A more useful extension is to determine a fitness function with appropriate properties for generalized linear models (GLM), which we do next.

3.4. Extension to generalized linear models

Let us consider GLM as described in McCullagh and Nelder (1989) for a response variable Y that follows an exponential family whose density may be written as

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

where θ is called the canonical or location parameter, ϕ the dispersion parameter, and the functions a, b and c are distribution-specific known functions. Following simple calculations (see e.g. Section 2.2.2 of McCullagh and Nelder, 1989), the following expressions for the mean and variance may be derived, namely, $\mathrm{E}(Y) = \mu = b'(\theta)$ and $\mathrm{var}(Y) = b''(\theta)a(\phi)$. These equations are used to find the variance function, $V(\mu) = \mathrm{var}(Y)/a(\phi)$ which must be expressed as a function of μ. Actual values of those functions for known distributions may be found in different references, including Table 2.1 of McCullagh and Nelder (1989).

In the definition of GLM, a linear combination η of the predictors is linked to an independent observation Y from an exponential family through a link function. The expression can also be reversed to have $\mu = \ell(\eta)$ where ℓ is the inverse link function. With the convention that ℓ is applied componentwise to a vector, we can write $\boldsymbol{\mu} = \ell(\mathbf{X}\boldsymbol{\beta})$ for a model with all features, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_D)$ is the vector of parameters of the model. For some families, ϕ may be a known constant, but in other cases, it is a nuisance parameter.

For exponential family GLM, the estimate of $\boldsymbol{\beta}$ may be found through maximum likelihood and usual properties thereof are retained. In the context of this paper, we only consider canonical link functions that provide additional simplifications, including the fact that $\theta = \eta$.

Moving to the context of BGA with the notation previously introduced, universe u has the bootstrapped samples $\tilde{\mathbf{Y}}_u$ and $\tilde{\mathbf{X}}_u$ and the p active features are indicated by B. The maximum likelihood equation is then based on the relation $\mathrm{E}(\tilde{\mathbf{Y}}_u) = \ell(\tilde{\mathbf{X}}_u\mathbf{B}\mathbf{B}^\intercal\boldsymbol{\beta})$ which depends only on a subset of $p+1$ elements of $\boldsymbol{\beta}$, namely $\mathbf{B}^\intercal\boldsymbol{\beta}$. The MLE $\hat{\boldsymbol{\beta}}_u(B)$ is therefore an asymptotically normal vector of $p+1$ elements with limiting mean $\mathbf{B}^\intercal\boldsymbol{\beta}$ and variance $\{\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{V}}_u(B)\tilde{\mathbf{X}}_u\mathbf{B}\}^{-1}$ where $\tilde{\mathbf{V}}_u(B)$ is a diagonal matrix with the variance function V applied componentwise to $\tilde{\boldsymbol{\mu}}_u(B) = \ell(\tilde{\mathbf{X}}_u\mathbf{B}\mathbf{B}^\intercal\boldsymbol{\beta})$ on its diagonal. The dependence on B does not change the dimension of $\tilde{\mathbf{V}}_u(B)$, but it affects the values therein. The same applies to $\mathbf{V}_u(B)$ which is based on $\boldsymbol{\mu}_u(B) = \ell(\mathbf{X}_u\mathbf{B}\mathbf{B}^\intercal\boldsymbol{\beta})$. The distributional result follows from the properties of the MLE and some algebra to determine the Fisher information for $\mathbf{B}^\intercal\boldsymbol{\beta}$. Applying the chain rule for second-order derivatives helps those calculations that are further simplified by properties emerging from the choice of the canonical link. In practice, $\tilde{\boldsymbol{\mu}}$ may need to be estimated by replacing $\mathbf{B}^\intercal\boldsymbol{\beta}$ with its MLE $\hat{\boldsymbol{\beta}}_u(B)$.

Building towards a fitness function based on a rescaled residual sum of squares, we may write out of sample predictions as

$$\hat{\mathbf{Y}}_u(B) = \ell\left\{X_u\mathbf{B}\hat{\boldsymbol{\beta}}_u(B)\right\}.$$

From an application of the multivariate delta method (see e.g. Section 3.4 of Das-Gupta, 2008), $\hat{\mathbf{Y}}_u(B)$ is asymptotically multivariate normal with mean $\boldsymbol{\mu}_u(B)$ and variance

$$a(\phi)\mathbf{V}_u(B)\mathbf{X}_u\mathbf{B}\{\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{V}}_u(B)\tilde{\mathbf{X}}_u\mathbf{B}\}^{-1}\mathbf{B}^\intercal\mathbf{X}_u^\intercal\mathbf{V}_u(B).$$

The Jacobian of the link function appears as $\mathbf{V}_u(B)$ due to their equality resulting from the use of the canonical link. The unscaled residuals $\hat{\boldsymbol{\varepsilon}}(B) = \mathbf{Y}_u - \hat{\mathbf{Y}}_u(B)$ then have mean 0 and variance

$$\Sigma_u(B) = a(\phi)\left[\mathbf{V}_u(B) + \mathbf{V}_u(B)\mathbf{X}_u\mathbf{B}\left\{\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{V}}_u(B)\tilde{\mathbf{X}}_u\mathbf{B}\right\}^{-1}\mathbf{B}^\intercal\mathbf{X}_u^\intercal\mathbf{V}_u(B)\right] \quad (6)$$

which will be required up to a multiplicative constant, hence the nuisance parameter (or constant) ϕ may be ignored. Pearsons X^2 statistic uses a sum of squared rescaled residuals to measure the goodness-of-fit. With out-of-sample validation, Equation 6 provides an appropriate factor. For GLM, we therefore use the same fitness function shown in Equation 2, but with modified values for $\hat{\boldsymbol{\varepsilon}}(B)$ and $\Sigma_u(B)$ which must be calculated up to a multiplicative constant.

REMARK 3. *The identity in Remark 1 and some algebra may be used to express the term in brackets in Equation 6 as*

$$\left[\{\mathbf{V}_u(B)\}^{-1} - \mathbf{X}_u\mathbf{B}\left\{\mathbf{B}^\intercal\tilde{\mathbf{X}}_u^\intercal\tilde{\mathbf{V}}_u(B)\tilde{\mathbf{X}}_u\mathbf{B} + \mathbf{B}^\intercal\mathbf{X}_u^\intercal\mathbf{V}_u(B)\mathbf{X}_u\mathbf{B}\right\}^{-1}\mathbf{B}^\intercal\mathbf{X}_u^\intercal\right]^{-1}$$

which allows to evaluate the fitness with one less matrix inversion when considering that inversing the diagonal matrix $\mathbf{V}_u(B)$ is much less computationally intensive.

The results of Section 3.3 do not depend on the actual models, but only on the assumption that the fitness of any model is approximately equal under the null hypothesis of no predictive power. No modifications need be made to the threshold once an appropriate fitness function has been developped. Similarly, it would be possible to use other goodness-of-fit methods as fitness functions, as long as they assess fairly models that have different number of variables.

3.5. Notable differences with PGA

While we adopt the parallel populations of Zhu and Chipman (2006), some of the choices they make in the parameters of their genetic algorithms are not suitable for BGA as they violate the assumptions that we use to derive the null distribution of the importance scores. Namely, the following elements are different.

Crossover: Zhu and Chipman (2006) use one-point crossover, but as we mentioned in the description of crossover, our preference goes to uniform crossover to make the arbitrary order of the features irrelevant.

Early-stopping: Using a finite predetermined number of generations yields importance scores that are averages of U independent values of equal variance. Early-stopping means that the number of generations depends on their diversity. While the theoretical complications are significant, the benefits of early-stopping are less clear.

Elitism: Zhu and Chipman (2006) copy the fittest half of a generation directly to the following generation. The derivation of the null distribution for the importance scores assumes no such elitism.

Selection: At each generation of Zhu and Chipman (2006), half the population survives, and all parents are picked at random from that survival pool. The expected value and variance of the importance score are based on a different mechanism where all individuals may be parents and their probability of selection is proportional to their fitness.

Fitness function: To enhance the symmetry between all possible models under the null, we use out-of-sample validation. One challenge with in-sample measures of fitness is to ensure that it does not display systematic preference (e.g. a bias toward models with more active features).

Bootstrap: We bootstrap the datasets to enhance the symmetry between features under the null by diluting spurious links that could appear out of pure chance.

In the next section, case studies and Monte Carlo methods are used to explore the behaviour of the importance scores under the null hypothesis.

4. Empirical exploration

We use real datasets and Monte Carlo studies to explore the behaviour of BGA in practice. The definition of "best" model is subject to debates and could be based, for instance, on the predictive abilities of the selected model according to different metrics. We rather adopt the same view as Zhu and Chipman (2006), and look at the ability of the models to detect the true features, which we call activated features in the context of the simulations. Part of the exploration is also designed to verify that the distribution of $\hat{\boldsymbol{\pi}}$ under the null hypothesis derived in Section 3.3 is observed empirically.

4.1. Illustrative data and harder problem

In the illustrative example of Zhu and Chipman (2006), 20 features labeled x_1 to x_{20} are simulated as independent normal variates with mean zero and variance one. We took the liberty to increase the sample size from 40 to $N = 200$ points that were generated along with the response

$$Y = x_5 + 2x_{10} + 3x_{15} + \varepsilon \tag{7}$$

where ε are independent normal variables with mean zero and variance one. We used the same parameters as Zhu and Chipman (2006) for the genetic algorithm, namely: $I = 20$, $G = 8$, $U = 25$, and $\theta_g = 1/D$ for all g. The value of the activation rate does not seem to be reported in their paper; we used $\pi_0 = 0.3$. The validation set have $n = 40$ which leaves $\tilde{n} = 160$ for the training. Figure 2 displays the boxplots of the importance scores obtained by one run of BGA on each of 1000 datasets generated following Equation 7. The horizontal plain line shows π_G, the expected importance. The plain dots are for the empirical 95$^{\text{th}}$ quantile of the importance scores for each feature and can be compared to the dashed line which shows their theoretical values under the null on the left panel. On the right panel, the correction for multiple comparisons is applied so the dashed line is the threshold to determine which features to select. The null model was obtained by setting all regression coefficients to zero in the same equation when simulating Y.

Under the null, both the expected value and the 95$^{\text{th}}$ quantile of the importance scores match their theoretical values remarkably well. When a signal is present, we note that the correct relevant variables are always detected, but also, that the values of the empirical 95$^{\text{th}}$ quantiles are reasonably close to their expected values under the null.

Let us now consider the hard problem that Zhu and Chipman (2006) attribute to George and McCulloch (1993). We preserved the sample size of 120 for the training

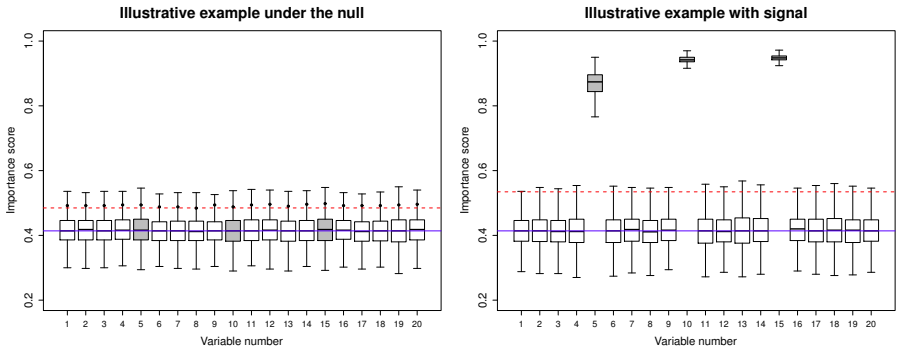

Figure 2. Importance scores obtained from the BGA of 1000 different samples generated from the same "Illustrative example" scenario. The left panel shows the importance scores under the null hypothesis, where Y is independent of the features. The right panel shows the importance scores when the signal described in Equation 7 is present. While the plain line shows the expected importance π_G, the dashed line displays the threshold $\pi_G + z_{0.95}\sigma_{\hat{\pi}}/\sqrt{U}$ on the left panel, and its Šidák adjusted equivalent on the right panel. The dots show the empirical 95^{th} quantile of the importance scores for each variable to compare against that theoretical bound. The boxplots of the activated features are coloured in gray.

sets by simulating samples of 150 data of which a proportion of 20% is dedicated to validation. A total of 60 features of variance 2 with a compound symmetry correlation structure are simulated. A correlation of 0.5 is present between any two predictors. The error term has mean 0 and variance 4. The regression parameters have four different values with $\beta_i = \lfloor (i-1)/15 \rfloor$, making 45 of the 60 features activated. Figure 3 displays the boxplots of the importance scores of 1000 samples, under the null and with a signal. The parameters of the genetic algorithm were set to $I = 60$, $G = 15$, $U = 100$, $\pi_0 = 0.9$ and $\theta_g = 1/D$ for all g.

Under the null, the values of the 95^{th} quantile and their expected value match fairly well as do the expected importance. On the right panel, the 45 activated features are found all the time with very few exceptions where the smallest signals are very occasionally left out – less than 4% of the time. The inactivated variables sometimes get picked up, but not more often than under the null, with a false positive rate of about 3% when using the Šidák corrected level. These results are as good as Figure 3 of Zhu and Chipman (2006). Interestingly, those figures require a large π_0. Smaller values show excellent performances as well, but they are less stellar for variables 1 to 30 which see their rates of errors increase. The ideal π_0 seems to be linked with the unknown true number of features. If the initial population has too few or too many active features, the genetic algorithm is more likely to miss its target. In absence of information about the likely number of features, comparing the results obtained with large and small values of π_0 may be a good option. We ran the PGA of Zhu and Chipman (2006) using their crossover and selection mechanisms, but with smaller π_0. Intuitively, their pointwise crossover could be thought to help in detecting the unactivated features that all come first. Empirically, this did not seem to be a driving factor as we observed a similar decrease in performance for smaller π_0.

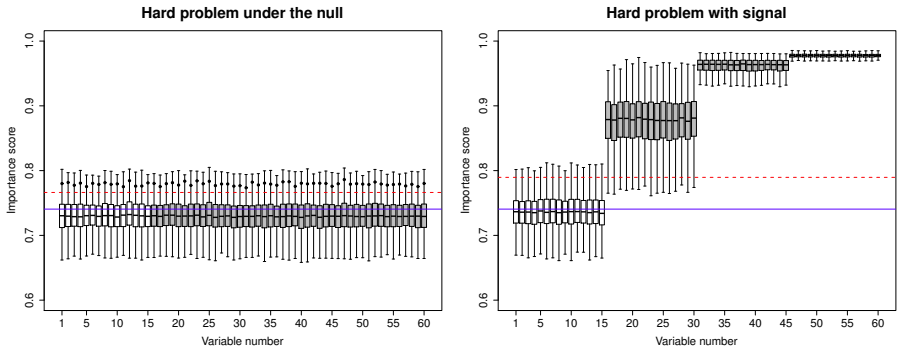

Figure 3. Importance scores obtained from the BGA of 1000 different samples generated from the same "Hard problem" scenario. The left panel shows the importance scores under the null hypothesis, where Y is independent of the features. The right panel shows the importance scores when $\beta_i = \lfloor (i-1)/15 \rfloor$ for feature i. While the plain line shows the expected importance π_G, the dashed line displays the threshold $\pi_G + z_{0.95}\sigma_{\hat{\pi}}/\sqrt{U}$ on the left panel, and its Šidák adjusted equivalent on the right panel. The plain dots show the empirical 95^{th} quantile of the importance scores for each variable to compare against that theoretical bound. The boxplots of the activated features are coloured in gray.

4.2. Pollution data

Let us consider the pollution data used in Example 4.1 of Zhu and Chipman (2006) and based on data used by Miller (2002) but initially published by McDonald and Schwing (1973). Let us assume that the three-variable model suggested in Figure 1 is correct. The same genetic parameters are used as in Figure 1, namely $I = 50$, $G = 20$, $U = 75$, $\pi_0 = 0.3$ and $\theta_g = 1/D$ for all g. The right panel of Figure 4 displays the boxplot of the importance scores of each feature obtained from 1000 datasets with the original values of the covariates and a new Y generated from a linear regression model where all regression coefficients equal 0 except for $\beta_0 = 796.5$, $\beta_1 = 2.347$, $\beta_9 = 2.961$, and $\beta_{14} = 0.3911$. The standard deviation of the error term is set to 38.58. Those parameters were determined from a fit on the entire original dataset, then deemed true values for the simulation. The lines and solid points have the same meaning as in Figure 2 presented in Section 4.1. The left panel under the null is obtained by randomly permuting the values of Y.

The theoretical mean under the null matches the empirical median closely, but contrarily to the illustrative example, we observe discrepancies between the theoretical and empirical values of the 95^{th} quantiles. While the correlation between the features and the presence of some outliers may play a role, an investigation led us to conclude that the small sample size was the real culprit here. We first ran the same simulation, but with features simulated as multivariate normal with the same mean and covariance as the original data. The ensuing plot was qualitatively identical to Figure 4. However, with a sample size of 300 instead of 60, the empirical 95^{th} quantiles are aligned on their theoretical values similarly as Figure 2. Digging further, we observed that the small sample size caused even smaller training and validation sets that are a lot more prone to spurious correlations, making the assumption of equal fitness fail strongly within some universes. Although the variables are equally likely to be favoured by chance, they are globally picked more often in their respec-

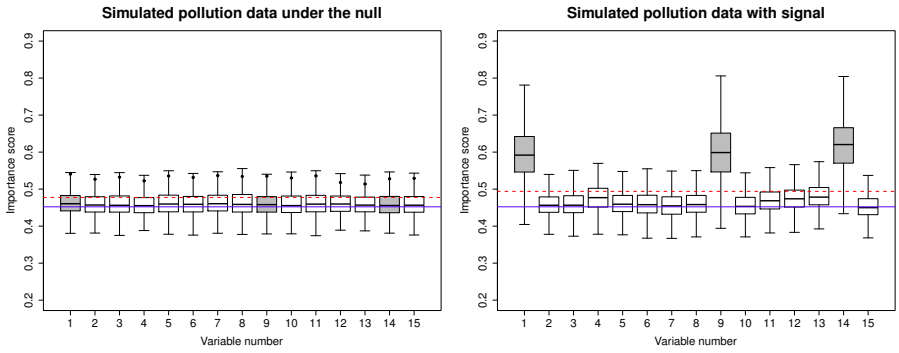

Figure 4. Importance scores obtained from the BGA of 1000 different samples generated from the same scenario based on the Pollution dataset. The left panel shows the importance scores under the null hypothesis, where Y is independent of the features. The right panel shows the importance scores when the signal depends on covariates 1, 9 and 14. While the plain line shows the expected importance π_G, the dashed line displays the threshold $\pi_G + z_{0.95}\sigma_{\hat{\pi}}/\sqrt{U}$ on the left, and the Šidák adjusted equivalent on the right. The plain dots show the empirical 95^{th} quantile of the importance scores for each variable to compare against that theoretical bound. The boxplots of the activated features are coloured in gray.

tive wrongly favourable universes, hence the increased importance scores. In cases where the sample is small, it may be advisable to consider a fitness function that does not rely on a validation set or to avoid bootstrapping those small samples. Our distributional results depend on the approximate equality of the fitness under the null, not on the specific choice that we propose.

Looking at the right panel of Figure 4, we note that the BGA detects the activated features, rightly selecting them over 90% of the time. Other variables however get wrongly picked between 12% and 37% of the time as a consequence of the artificially increased importance.

4.3. Wisconsin breast cancer dataset

Consider now the Wisconsin breast cancer dataset from Street et al. (1993) that we obtained from Dheeru and Karra Taniskidou (2017). A binary response variable indicating that the tumour is malignant (rather than bening) is explained with logistic regression using 30 features derived from medical imaging. In this sample of 569 patients, 212 have a malignant tumour. The parameter α of BGA can be used to control the number of features that are selected. To simulate a simpler model, we chose a conservative value of $\alpha = 0.005$. The other parameters of the BGA were $I = 20$, $G = 15$, $U = 50$, $\pi_0 = 0.3$ and $\theta_g = 1/D$ for all g. Figure 5 shows the bubble plot of this BGA with the theoretical mean importance and threshold as plain and dashed lines respectively. Variables 8 and 21 are selected since they are above the objective threshold, but we may also note that there is a gap between them and the next best variable, so the subjective decision from a bubble plot would have been the same in this case.

To assess the ability of BGA to detect the activated features in a realistic logistic

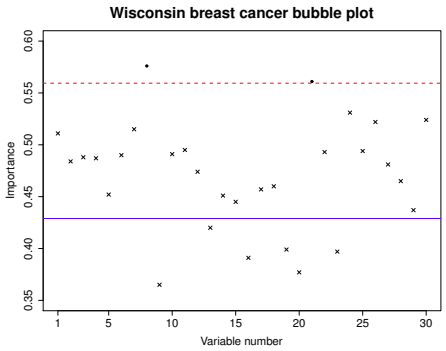

Figure 5. Bubble plot for the Wisconsin breast cancer dataset obtained from a BGA with parameters $I = 20$, $G = 15$, $U = 50$, $\pi_0 = 0.3$ and $\theta_g = 1/D$ for all g. The global level was set to $\alpha = 0.005$. The plain line shows π_G, the theoretical mean of the importance under the null hypothesis. The dashed line shows the threshold that was derived to decide which features to retain. The retained factors are shown as dots, the dismissed factors as crosses.

regression setting, we used the original predictors of the Wisconsin breast cancer dataset, but we generated new responses that follow a logistic regression model with parameters $\beta_i = 0$, except for $\beta_0 = -18.38$, $\beta_8 = 75.25$ and $\beta_{21} = 0.8797$ that we obtained from a fit on the original dataset. We created 1000 simulated dataset reusing the same features, but generating a new response. We used the same genetic parameters as for Figure 5. The right panel in Figure 6 shows the boxplots of the importance scores obtained. The left panel was obtained under the null by randomly permuting the binary values in the simulated response, hence breaking any predictive power of the features.

The left panel shows that the importance scores behave as expected under the null, even in the case of a GLM. On the right panel, the two activated features are the ones that get picked up most often, for 79% and 32% of the simulated repetitions. Variable 24 wrongly gets picked up 29% of the time, but its correlation is 0.83 with X_8 and 0.984 with X_{21}. All other features are wrongly selected less than 9% of the time although some have correlations exceeding 0.96 with the two activated features. This is undoubtedly a hard problem: The covariates are very highly correlated, and the response of this GLM is binary. Yet, the BGA that we proposed shows a good performance.

To confirm that high correlations was a major challenge in this example, we ran another simulation where the correlation in the predictors was removed by independently shuffling values of each feature once, yielding an uncorrelated sample that we used to simulate the target value in the same fashion as before. Figure 7 shows the results obtained from 1000 runs. With uncorrelated features, BGA works clearly very well for logistic regression.

5. Tuning of parameters

Genetic algorithms in general depend on many parameters that need to be somewhat arbitrarily fixed and so does BGA. Experience helps in choosing appropriate values,

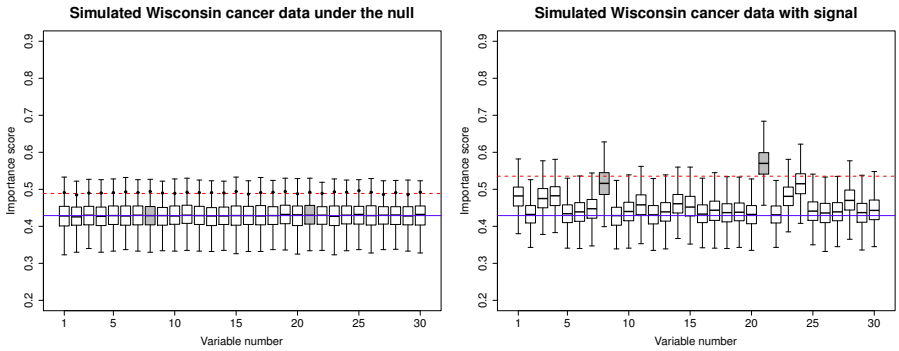

Figure 6. Importance scores obtained from the BGA of 1000 different samples generated from a scenario based on the Wisconsin breast cancer dataset. The left panel shows the importance scores under the null hypothesis, where the malingness of the tumour is independent of the features. The right panel shows the importance score when the signal depends on covariates 8 and 21. While the plain line shows the expected importance π_G, the dashed line displays the threshold $\pi_G + z_{0.95}\sigma_{\hat{\pi}}/\sqrt{U}$ on the left, and the Šidák adjusted equivalent on the right. The plain dots show the empirical 95^{th} quantile of the importance scores for each variable to compare against that theoretical bound. The boxplots of the activated features are coloured in gray.

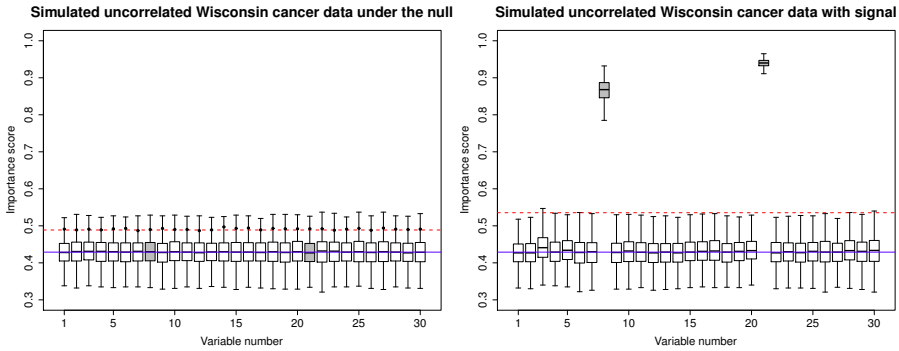

Figure 7. Importance scores obtained from the BGA of 1000 different samples generated from a scenario based on the Wisconsin breast cancer dataset where correlation between the predictors was removed. The left panel shows the importance scores under the null hypothesis, where the malingness of the tumour is independent of the features. The right panel shows the importance score when the signal depends on covariates 8 and 21. While the plain line shows the expected importance π_G, the dashed line displays the threshold $\pi_G + z_{0.95}\sigma_{\hat{\pi}}/\sqrt{U}$ on the left, and the Šidák adjusted equivalent on the right. The plain dots show the empirical 95^{th} quantile of the importance scores for each variable to compare against that theoretical bound. The boxplots of the active features are coloured in gray.

as does occasional simulations to explore the behaviour of the method. There are however general guidelines that may help.

5.1. Fitness function

Under the null, the fitness function must yield approximately equal probability of selection to all possible models. The construction of Equation 2 was designed with this property in mind and the Monte Carlo studies in Section 4 verified that the assumption is reasonably valid in realistic settings. A similar empirical endeavour could contribute to validate newly proposed fitness functions. As long as this property is preserved, the choice of fitness has no bearings on the distribution of selection ratios under the null hypothesis. Of course, the ability of the fitness to identify desirable models outside of the null hypothesis is also key to good performances.

The peakedness parameter (γ) highlights fitter individuals further, thus helping to speed up evolution. As such, it may help when computational resources are limited. A large γ may however also highlight naturally occurring fortuitous fluctuations, so larger values are not systematically better. We mostly used values of γ in the neighbourhood of 1.5, although we experimented with values as high as $\gamma = 4$ which reminded us of the ARCX4 algorithm of Breiman (1996a), but a fourth power did not have the same beneficial effect here.

5.2. Rates

The generation and evolution of populations depend on random elements that appear at some rates.

Activation rate (π_0): The value of π_0 has an effect on the expected value of the importance scores. Values close to 0 or 1 should be avoided as they lower the diversity of the initial population which then relies on mutation to explore numerous models. A rate of $\pi_0 = 0.5$ appears to be a good choice that maximizes initial diversity, but, in practice, best results are obtained when π_G is close to the true proportion of activated features in the models. If prior information is known about the expected number of active features, it would make most sense to start π_0 in the neighbourhood of that value. Since models are often expected to be sparse, smaller values of π_0 appear to be more appropriate.

Mutation rates (θ_g, $g \in \{1, \ldots, G\}$): Mutation breaks stagnation by infusing diversity in each generation and helps to explore the space of all candidate models, while it also precludes the convergence of a population. A lack of convergence is detrimental for single-thread genetic algorithms, but it helps BGA which seeks diversity between the universes and looks for a signal in the aggregation of all individuals. With a finite number of generations, a large mutation is approriate, but it should stay far below $\theta_g = 1/2$ which would make the next generation random, without genetic memory. As Zhu and Chipman (2006) pointed out, many instances in the literature suggest that $\theta_g = 1/D$, which we also found to work properly. Although we allow for a changing rate of mutation, this appears to be especially appropriate for single-thread genetic algorithm where diversity has value for the initial generations, but convergence eventually requires low mutation rates. Hence for BGA, we typically kept the mutation rate constant.

5.3. Number of individuals, universes and generations

The computational cost of a BGA depends highly on the number of individuals that will be generated, namely IGU.

Number of parallel universes (U): Universes are conditionally independent by design of the algorithm, and the accuracy of the selection ratios is improved by a larger number of universes. To justify the central limit theorem, using no less than 30 universes is advised, but that number could be far greater if the computational resources are available. Note also that the computation of the universes is embarrassingly parallel, so it provides a trivial way to scale the algorithm on multi-core platforms.

Number of generations (G): A large number of generations may be detrimental as the covariance between two individuals (c_g) will tend to increase as g increases, thus yielding a larger threshold. On the other hand, too few generations might not allow sufficient time for the genetic operators to detect the truly best models under the alternative when some features are good predictors. To complicate things, the speed of convergence to fit individuals also depends on the probability of activation and the mutation rates, especially if the mutation varies with g. Values of 10 to 15 appeared to work well for G.

Population size (I): Contrary to the intuition, larger population sizes do not yield better results. In a very large population, the probability of selection of an individual twice as fit as anybody else may be so diluted that his genes will not be passed on to the next generation. In a smaller population, his relative competitive advantage is much larger. The population size (I) should thus be as small as possible, but not to the point of being detrimental to the exploration of the possible models. Zhu and Chipman (2006) suggest having a population size (I) of the same magnitude as the dimension of the data (D) and we found that guideline worked well, although $I < D$ also has merits when the number of dimensions gets large.

With the total computational cost of BGA being $\mathcal{O}(IGU)$, it appears advisable to find a moderate value of I, then make U as large as possible.

6. Conclusion

The parallel genetic algorithm of Zhu and Chipman (2006) is an embarrassingly parallel approach to feature selection. It computes importance scores for each variable from a summary of all models in a number of universes that are purposefully not fully evolved to enhance diversity.

We derive the distribution of the importance score when none of the covariates are good predictors, which yields an objective criterion for variable selection. Simulations show that the distributional results hold well under the null hypothesis, both for linear regression and GLM. BGA performs well even in cases where predictors are highly correlated with one another. In such circumstances, predictors may play the role of a proxy for one another: An unactivated variable will be important when it is compensating for a correlated covariate that is absent from the model but should not. In less challenging settings, BGA performs exceptionally well.

The distribution of the importance scores under the null does not depend on the actual fitness function, but on its ability to assess models fairly. The out-of-sample RSS that we proposed works well for linear regression and GLM. Other fitness functions are certainly possible and could be the topic of future research. More importantly, our results apply to feature selection for any model, linear or not, as long as a fair fitness function can be established.

Ackowledgement

For partial support of this work through research grants, thanks are due to the *Natural Sciences and Engineering Research Council of Canada*.

References

Breiman, L. (1996a). Arcing classifiers. *Annals of Statistics 26*, 801–849.

Breiman, L. (1996b). Bagging predictors. *Machine learning 24*(2), 123–140.

Cantú-Paz, E. (2000). *Efficient and accurate parallel genetic algorithms*, Volume 1. Springer.

Casella, G. and R. L. Berger (2002). *Statistical inference*, Volume 2. Duxbury Pacific Grove, CA.

Chatterjee, S., M. Laudato, and L. A. Lynch (1996). Genetic algorithms and their statistical applications: an introduction. *Computational Statistics & Data Analysis 22*(6), 633–651.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.

Dheeru, D. and E. Karra Taniskidou (2017). UCI machine learning repository.

Draper, N. R. and H. Smith (2014). *Applied regression analysis*, Volume 326. John Wiley & Sons.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*. Springer series in statistics New York, NY, USA:.

George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Haupt, R. L. and S. E. Haupt (2004). *Practical genetic algorithms*. John Wiley & Sons.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by simulated annealing. *science 220*(4598), 671–680.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*, Volume 37. CRC press.

McDonald, G. C. and R. C. Schwing (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics 15*(3), 463–481.

Miller, A. (2002). *Subset selection in regression*. Chapman and Hall/CRC.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press.

Poli, R., W. W. B. Langdon, N. F. McPhee, and J. R. Koza (2008). *A field guide to genetic programming*. Lulu.com.

Shonkwiler, R. W. and F. Mendivil (2009). *Explorations in Monte Carlo Methods*. Springer Science & Business Media.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

Street, W. N., W. H. Wolberg, and O. L. Mangasarian (1993). Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, Volume 1905, pp. 861–871. International Society for Optics and Photonics.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wang, G.-W. and C.-X. Zhang (2015). Building variable selection ensembles for linear regression models by adding noise. In *Machine Learning and Cybernetics (ICMLC), 2015 International Conference on*, Volume 2, pp. 554–559. IEEE.

Zhang, C.-X., G.-W. Wang, and J.-M. Liu (2015). Randga: injecting randomness into parallel genetic algorithm for variable selection. *Journal of Applied Statistics 42*(3), 630–647.

Zhu, M. and H. A. Chipman (2006). Darwinian evolution in parallel universes: a parallel genetic algorithm for variable selection. *Technometrics 48*(4), 491–502.

IV Détails des preuves

La présente section est constituée de trois volets, lesquels ont tous pour objectif d'étoffer certains passages plus subtils et alambiqués de l'ouvrage. La sous-section IV.1 se contente de détailler les étapes permettant ultimement de poser le seuil critique $C = \pi_G + q_\alpha \sigma_{\hat{\pi}_{Gu}} / \sqrt{U}$ du test d'hypothèse introduit à la section 3.3 de l'article. La sous-section IV.2 explicite de manière rigoureuse les motivations sous-tendant le choix de notre fonction d'ajustement, et à cet égard, détaille sa capacité à noter équitablement, en moyenne, tout modèle qui soit sous l'hypothèse nulle, et ce, tant pour les modèles linéaires que linéaires généralisés.

IV.1 Distribution des proportions de rétention globales sous H_0

Cette section se subdivise également en trois volets. Nous nous attarderons dans un premier temps à déterminer la distribution marginale des gènes d'une même population, ainsi que la relation d'interdépendance qui les caractérise. Nous nous intéresserons alors aux moments des proportions de rétention intra-populations. Nous concluons par la distribution des proportions de rétention globales et la sélection du seuil critique des tests d'hypothèse.

IV.1.1 Distribution des gènes intra-univers

IV.1.1.1 Espérance

Puisque le croisement ne consiste qu'à recombiner les gènes de deux parents [(14)], il n'y a rien de surprenant à affirmer que l'espérance des gènes de l'embryon issu d'une telle union, laquelle est définie à la section 2.1.4 de l'article, soit à son tour identique à celle des gènes de n'importe quel parent candidat (c'est-à-dire $\mathrm{E}[b^*_{di(g+1)u}] = \mathrm{E}[b_{digu}]$). Pour toutes ces raisons, l'espérance de l'individu que deviendra cet embryon est donnée par l'expression que voici

$$\begin{aligned} \mathrm{E}[b_{digu}] &= \theta_g \left(1 - \mathrm{E}\left[b^*_{digu}\right]\right) + (1-\theta_g)\,\mathrm{E}\left[b^*_{digu}\right] \\ &= \theta_g \left(1 - \mathrm{E}\left[b_{di(g-1)u}\right]\right) + (1-\theta_g)\,\mathrm{E}\left[b_{di(g-1)u}\right] \\ &= \theta_g + (1-2\theta_g)\,\mathrm{E}\left[b_{di(g-1)u}\right], \end{aligned} \tag{8}$$

laquelle, substituant π_g à $\mathrm{E}[b_{digu}]$, est présentée à la section 3.3 de l'article sous la forme

$$\pi_g = \theta_g + (1-2\theta_g)\,\pi_{g-1}. \tag{9}$$

Après simplification, et, par souci de compacité d'écriture, nous montrons, sous la convention

(14). Dont les gènes, en vertu de l'argument de symétrie invoqué à la section 3 de l'article, sont de même distribution sous l'hypothèse nulle, c'est-à-dire $B_{i_1gu} \sim B_{i_2gu}$, $\forall i_1, i_2 \in \{1, \ldots, I\}$.

additionnelle stipulant que $\theta_0 = \pi_0$, que le terme général de cette séquence est

$$\pi_g = \frac{1}{2}\left\{1 - \prod_{j=0}^{g}(1-2\theta_j)\right\}. \tag{10}$$

Un cas particulier de ce résultat pour θ_g fixe, c'est-à-dire pour lequel $\boldsymbol{\theta} = \begin{bmatrix}\theta & \cdots & \theta\end{bmatrix}^T$, est

$$\pi_g = \frac{1-(1-2\pi_0)(1-2\theta)^g}{2}. \tag{11}$$

Dans tous les cas, nous démontrerons ce résultat au moyen d'un simple raisonnement par récurrence. Tout d'abord, nous montrons que cette expression est correctement évaluée lorsque $g = 0$:

$$\frac{1}{2}\left\{1 - \prod_{j=0}^{0}(1-2\theta_j)\right\} = \theta_0 = \pi_0,$$

Nous montrons alors que si ce résultat est vrai pour g, il l'est également pour $g+1$:

$$\theta_{g+1} + (1-2\theta_{g+1})\frac{1}{2}\left\{1 - \prod_{j=0}^{g}(1-2\theta_j)\right\} = \frac{1}{2}\left\{1 - \prod_{j=0}^{g+1}(1-2\theta_j)\right\} = \pi_{g+1}.$$

Fait intéressant, en posant $\Delta_g = \prod_{j=0}^{g}(1-2\theta_j)$, il est possible de récrire π_g comme $\frac{1}{2}(1-\Delta_g)$. On remarque alors que pour plusieurs structures de mutation $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2 & \cdots\end{bmatrix}^T$ plausibles, $\pi_g \to \frac{1}{2}$, et ce, étant donné la divergence de Δ_g vers 0 [15]. Par plausible, nous entendons bien là une séquence dont les valeurs sont intuitivement éléments de $\left(0, \frac{1}{2}\right)$. En effet, l'utilité de l'opérateur de mutation résidant en sa capacité à briser la stagnation des populations pour ainsi mieux explorer l'espace solutions Ω, il serait contre-productif de s'en priver par l'utilisation d'une valeur nulle ou même quasi-nulle de θ_g. À l'inverse, un taux de mutation de valeur $\frac{1}{2}$ aurait pour conséquence d'annihiler toute mémoire génétique des populations, réduisant ainsi l'application des opérateurs génétiques à une simple recherche primitive et aléatoire sur Ω.

Afin de mieux illustrer la chose, considérons seulement l'exemple d'un taux de mutation $\theta_g = \frac{1}{2(g+1)}$ pour $g > 0$ et π_0 de valeur quelconque. De tendance décroissante, mais non pas si faibles à en devenir négligeables, ce dernier engendre les probabilités de rétention $\pi_g = \frac{1-(1-2\pi_0)/(G+1)}{2}$, lesquelles tendent bel et bien vers $\frac{1}{2}$ pour $G \to \infty$. Au contraire, le taux de mutation $\theta_g = \frac{\theta^2}{2\pi^2 g^2}$ (pour $|\theta| \leq \sqrt{2}\pi$), également décroissant mais cette fois selon une cadence ô combien plus accélérée, génère la séquence $\pi_g = \frac{1}{2}\left(1-(1-2\pi_0)\prod_{g=1}^{G}\left(1-\frac{\theta^2}{\pi^2 g^2}\right)\right)$, laquelle tend plutôt vers $\frac{1-(1-2\pi_0)\sin(\theta)/\theta}{2} \neq \frac{1}{2}$. Nous observons ainsi que deux séquences, l'une plausible et l'autre pas, posséderont à leur tour la capacité d'amener les probabilités de rétention espérées

(15). Dans le cas où tous les termes d'une suite sont non-nuls, on dit d'un produit infini qu'il *converge* quand la limite des produits finis existe et est non-nulle ; sinon, on dit du produit infini qu'il *diverge* [24].

vers la vraisemblable limite de $\frac{1}{2}$, ou ailleurs.

IV.1.1.2 Covariance

La covariance des gènes encodant la même variable d chez deux embryons distincts d'une même génération et même population, ici désignés par $b^*_{di_1gu}$ et $b^*_{di_2gu}$, est donnée par l'expression

$$\begin{aligned} \mathrm{Cov}\Big[b^*_{di_1gu}, b^*_{di_2gu}\Big] &= \tfrac{1}{I}\mathrm{Var}\Big[b_{di_1(g-1)u}\Big] + \Big(1-\tfrac{1}{I}\Big)\,\mathrm{Cov}\Big[b_{di_1(g-1)u}, b_{di_2(g-1)u}\Big] \\ &= \mathrm{Var}\Big[\widehat{\pi}_{d(g-1)u}\Big]\,. \end{aligned} \tag{12}$$

En effet, de par le mécanisme de sélection et de croisement, ces deux gènes ont respectivement une chance sur I de correspondre au seul et même gène tiré du patrimoine génétique de la génération précédente (leurs parents étant en fait un seul et même individu) et $I-1$ chances sur I de correspondre à deux gènes différents. À son tour, la covariance des gènes encodant la même variable d chez deux individus d'une même génération et même population, ici désignés par b_{di_1gu} et b_{di_2gu}, est donnée par l'expression

$$\begin{aligned} \mathrm{Cov}[b_{di_1gu}, b_{di_2gu}] &= \mathrm{E}\Big[\mathrm{Cov}\Big[b_{di_1gu}, b_{di_2gu}\,\Big|\, M_{di_1gu}, M_{di_2gu}\Big]\Big] \\ &+ \mathrm{Cov}\Big[\mathrm{E}\Big[b_{di_1gu}\,\Big|\, M_{di_1gu}, M_{di_2gu}\Big], \mathrm{E}\Big[b_{di_1gu}\,\Big|\, M_{di_1gu}, M_{di_2gu}\Big]\Big]\,, \end{aligned} \tag{13}$$

en vertu du théorème de la covariance totale [(16)]. Puisque

$$\begin{aligned} &\mathrm{E}\Big[\mathrm{Cov}\Big[b_{di_1gu}, b_{di_2gu}\,\Big|\, M_{di_1gu}, M_{di_2gu}\Big]\Big] \\ &\quad= \mathrm{Cov}\Big[b^*_{di_1gu}, b^*_{di_2gu}\Big]\,\Pr[M_{di_1gu}=0, M_{di_2gu}=0] \\ &\quad+ \mathrm{Cov}\Big[1-b^*_{di_1gu}, b^*_{di_2gu}\Big]\,\Pr[M_{di_1gu}=1, M_{di_2gu}=0] \\ &\quad+ \mathrm{Cov}\Big[b^*_{di_1gu}, 1-b^*_{di_2gu}\Big]\,\Pr[M_{di_1gu}=0, M_{di_2gu}=1] \\ &\quad+ \mathrm{Cov}\Big[1-b^*_{di_1gu}, 1-b^*_{di_2gu}\Big]\,\Pr[M_{di_1gu}=1, M_{di_2gu}=1] \\ &\quad+ \mathrm{Cov}\Big[b^*_{di_1gu}, b^*_{di_2gu}\Big]\,(1-\theta_g)^2 + \mathrm{Cov}\Big[1-b^*_{di_1gu}, b^*_{di_2gu}\Big]\,\theta_g\,(1-\theta_g) \\ &\quad+ \mathrm{Cov}\Big[b^*_{di_1gu}, 1-b^*_{di_2gu}\Big]\,(1-\theta_g)\,\theta_g + \mathrm{Cov}\Big[1-b^*_{di_1gu}, 1-b^*_{di_2gu}\Big]\,\theta_g^2 \\ &\quad= (1-2\theta_g)^2\mathrm{Cov}\Big[b^*_{di_1gu}, b^*_{di_2gu}\Big]\,, \end{aligned} \tag{14}$$

et

$$\begin{aligned} &\mathrm{Cov}\Big[\mathrm{E}\Big[b_{di_1gu}\,\Big|\, M_{di_1gu}, M_{di_2gu}\Big], \mathrm{E}\Big[b_{di_1gu}\,\Big|\, M_{di_1gu}, M_{di_2gu}\Big]\Big] \\ &\quad= \mathrm{Cov}[M_{di_1gu} + (1-2M_{di_1gu})\,\pi_{g-1}, M_{di_2gu} + (1-2M_{di_2gu})\,\pi_{g-1}] \\ &\quad= 0, \end{aligned} \tag{15}$$

par l'indépendance de l'opérateur de mutation, tel que posé à la section 2.1.5 de l'article, nous

(16). Lequel est ici appliqué en conditionnant sur les variables M_{di_1gu} et M_{di_2gu} qui, telles que définies à la section 2.1.15 de l'article, encodent la mutation des gènes d des embryons $b^*_{di_1gu}$ et $b^*_{di_2gu}$ lors du passage de la génération $g-1$ à g.

concluons que

$$\mathrm{Cov}[b_{di_1gu}, b_{di_2gu}] = (1-2\theta_g)^2 \mathrm{Cov}\left[b^*_{di_1gu}, b^*_{di_2gu}\right] = (1-2\theta_g)^2 \mathrm{Var}\left[\widehat{\pi}_{d(g-1)u}\right]. \tag{16}$$

Nous pourrons alors mettre à profit cette expression dans notre dérivation de la variance des proportions de rétention sous l'hypothèse nulle décrite à la section suivante.

IV.1.2 Distribution des proportions de rétention intra-univers

Si b_{digu} encode la rétention de la variable d au modèle i de la génération g de l'univers u, $\widehat{\pi}_{dgu} = \frac{1}{I}\sum_{i=1}^{I} b_{digu}$ symbolise alors la proportion de rétention de cette même variable et pour ce même univers.

De par la propriété de linéarité de l'espérance, on montre que l'espérance des proportions de rétention intra-univers est la même que celle des gènes que celles-ci agrègent, soit

$$\mathrm{E}[\widehat{\pi}_{dgu}] = \mathrm{E}\left[\frac{1}{I}\sum_{i=1}^{I} b_{digu}\right] = \frac{1}{I}\sum_{i=1}^{I} \mathrm{E}[b_{digu}] = \pi_{dgu}. \tag{17}$$

De même, par les propriétés de la variance, on montre que

$$\mathrm{Var}[\widehat{\pi}_{dgu}] = \mathrm{Var}\left[\frac{1}{I}\sum_{i=1}^{I} b_{digu}\right] = \frac{1}{I}\mathrm{Var}[b_{digu}] + \left(1-\frac{1}{I}\right)\mathrm{Cov}[b_{di_1gu}, b_{di_2gu}], \tag{18}$$

telle que donnée à la section 3.3 de l'article. Il est alors possible de substituer la covariance $\mathrm{Cov}[b_{di_1gu}, b_{di_2gu}]$ par son expression définie à l'équation 16, pour ainsi obtenir :

$$\begin{aligned} \mathrm{Var}[\widehat{\pi}_{dgu}] &= \frac{1}{I}\mathrm{Var}[b_{digu}] + \left(1-\frac{1}{I}\right)(1-2\theta_g)^2 \mathrm{Var}\left[\widehat{\pi}_{d(g-1)u}\right] \\ &= \frac{1}{I}\pi_g(1-\pi_g) + (1-2\theta_g)^2 \mathrm{Var}\left[\widehat{\pi}_{d(g-1)u}\right]. \end{aligned} \tag{19}$$

En désignant alors par $\sigma^2_{\widehat{\pi}_{gu}}$ la variance $\mathrm{Var}[\widehat{\pi}_{dgu}]$, il est possible de récrire cette récurrence comme suit

$$\sigma^2_{\widehat{\pi}_{gu}} = \frac{1}{I}\pi_g(1-\pi_g) + \left(1-\frac{1}{I}\right)(1-2\theta_g)^2 \sigma^2_{\widehat{\pi}_{(g-1)u}} = \frac{1}{4I}\left(1-\Delta_g^2\right) + \left(1-\frac{1}{I}\right)\frac{\Delta_g^2}{\Delta_{g-1}^2}\sigma^2_{\widehat{\pi}_{(g-1)u}}. \tag{20}$$

Cette même représentation nous permet alors de proposer le terme général que voici [17]

$$\sigma^2_{\widehat{\pi}_{gu}} = \frac{1}{I}\sum_{j=0}^{g} \pi_j(1-\pi_j)\left(1-\frac{1}{I}\right)^{g-j} \prod_{k=j+1}^{g}(1-2\theta_k)^2 = \frac{\Delta_g^2}{4I}\sum_{j=0}^{g}\frac{1-\Delta_j^2}{\Delta_j^2}\left(1-\frac{1}{I}\right)^{g-j}, \tag{21}$$

laquelle ne figurait pas à l'article, par manque d'espace, mais n'étant pas moins vraie pour autant. Nous démontrerons ce résultat au moyen d'un simple raisonnement par récurrence.

(17). Notons que ce résultat, même sous l'hypothèse d'un taux de mutation constant, ne se simplifie pas davantage.

Nous montrons dans un premier temps que $\sigma^2_{\hat{\pi}_{gu}}$ est correctement évaluée lorsque $g = 0$:

$$\frac{\Delta_0^2}{4I}\sum_{j=0}^{0}\frac{1-\Delta_j^2}{\Delta_j^2}\Big(1-\frac{1}{I}\Big)^{0-j} = \frac{1-\Delta_0^2}{4I} = \frac{\theta_0(1-\theta_0)}{I} = \frac{\pi_0(1-\pi_0)}{I} = \sigma^2_{\hat{\pi}_{0u}}.$$

Montrons à présent que si ce résultat est vrai pour g, il l'est également pour $g+1$:

$$\begin{aligned}
\tfrac{1}{4I}\left(1-\Delta_{g+1}^2\right) \;+\; & \Big(1-\tfrac{1}{I}\Big)\tfrac{\Delta_{g+1}^2}{\Delta_g^2}\cdot\tfrac{\Delta_g^2}{4I}\sum_{j=0}^{g}\tfrac{1-\Delta_j^2}{\Delta_j^2}\Big(1-\tfrac{1}{I}\Big)^{g-j} \\
= \;& \tfrac{\Delta_{g+1}^2}{4I}\left\{\tfrac{1-\Delta_{g+1}^2}{\Delta_{g+1}^2}\Big(1-\tfrac{1}{I}\Big)^{g+1-(g+1)} + \sum_{j=0}^{g}\tfrac{1-\Delta_j^2}{\Delta_j^2}\Big(1-\tfrac{1}{I}\Big)^{g+1-j}\right\} \\
= \;& \tfrac{\Delta_{g+1}^2}{4I}\sum_{j=0}^{g+1}\tfrac{1-\Delta_j^2}{\Delta_j^2}\Big(1-\tfrac{1}{I}\Big)^{g+1-j} \\
= \;& \sigma^2_{\hat{\pi}_{(g+1)u}}.
\end{aligned}$$

Disposant ainsi des premiers moments des proportions de rétention intra-populations, il nous est possible de procéder à la section suivante où la distribution limite des proportions de rétention est donnée, et le seuil critique du test est posé.

IV.1.3 Seuil critique des tests

Les proportions de rétention *globales*, ou simplement *proportions de rétention*, sont définies comme la *grande* moyenne des proportions de rétention intra-populations, soit

$$\widehat{\pi}_{dg} = \tfrac{1}{U}\sum_{u=1}^{U}\widehat{\pi}_{dgu} = \tfrac{1}{IU}\sum_{u=1}^{U}\sum_{i=1}^{I} b_{digu}, \tag{22}$$

sous la convention que

$$\widehat{\boldsymbol{\pi}}_g = \begin{bmatrix}\widehat{\pi}_{1,g} & \cdots & \widehat{\pi}_{d,g}\end{bmatrix}^T, \tag{23}$$

excluant ainsi la variable explicative $d = 0$, laquelle est entièrement constituée de 1 et utilisée aux seules fins de procéder à l'estimation du paramètre d'interception des modèles, β_0.

Selon le fonctionnement des différents univers qu'invoque l'exécution d'un algorithme génétique ensaché, ainsi qu'en vertu du rééchantillonnage bootstrap sur lesquels ceux-ci sont évalués, et du partitionnement aléatoire apprentissage et validation, il semble raisonnable de supposer l'applicabilité du théorème central limite [6] aux proportions de rétention $\widehat{\boldsymbol{\pi}}_g$ (comportement également confirmé empiriquement à la section 4 de l'article). Il s'ensuit alors un comportement asymptotiquement normal, c'est-à-dire

$$\sqrt{U}\left(\widehat{\pi}_{dg} - \pi_g\right) \xrightarrow{\mathscr{L}} \mathscr{N}\Big(0, \sigma^2_{\hat{\pi}_g}\Big),$$

ou encore

$$\widehat{\pi}_{dg} \approx \mathscr{N}\left(\pi_g, \sigma^2_{\hat{\pi}_g}/U\right) \sim \mathscr{N}\left\{\frac{1-\Delta_g}{2},\ \frac{\Delta_g^2}{4IU}\sum_{j=0}^{g}\frac{1-\Delta_j^2}{\Delta_j^2}\left(1-\frac{1}{I}\right)^{g-j}\right\}, \tag{24}$$

pour U grand. Tester une hypothèse composée comme la nôtre, H_0: $\beta_1 = \ldots = \beta_D = 0$, implique alors l'évaluation d'un certain nombre de sous-tests, ici D, donnant ainsi lieu au *problème des comparaisons multiples* [1] pour lequel il était adéquat de pallier par l'entremise de la correction de Šidák [20]. Ainsi, pour un niveau de test global α, cette dernière rehausse le niveau individuel des tests, comme toute correction est portée à le faire, de telle sorte que la probabilité d'observer *au moins* un faux positif soit de α. Mathématiquement, cela se traduit par l'utilisation d'un niveau corrigé correspondant à $\alpha_S = 1 - (1-\alpha)^{1/D}$, lequel nous permet à son tour de poser le seuil critique du test comme la borne supérieure d'un intervalle de confiance unilatéral à droite, soit

$$C = \pi_G + q_{1-\alpha_S}\frac{\sigma_{\hat{\pi}_{Gu}}}{\sqrt{U}}, \tag{25}$$

tel que donné à la section 3.3 de l'article.

IV.2 Choix de la fonction d'ajustement

À l'intérieur de cette sous-section, nous ferons la démonstration des propriétés désirables des fonctions d'ajustements présentées à l'équation 2 de l'article, c'est-à-dire que nous montrerons en quoi ces dernières confèrent sous H_0 une chance équiprobable aux modèles d'être sélectionnés pour la reproduction.

IV.2.1 Régression linéaire multiple

IV.2.1.1 Fondations du modèle

Le modèle linéaire assume une variable réponse dont la tendance s'exprime comme une combinaison linéaire des variables explicatives candidates, $\tilde{\boldsymbol{X}}_u\boldsymbol{\beta}$, et dont les erreurs, $\tilde{\boldsymbol{\varepsilon}}_u$, gaussiennes, indépendantes et homoscédastiques, donnent lieu à la distribution ci-dessous

$$\tilde{\boldsymbol{Y}}_u \,\Big|\, \tilde{\boldsymbol{X}}_u \sim \mathscr{N}\left(\tilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right), \tag{26}$$

où $\boldsymbol{I}$ est la matrice identité d'ordre $\tilde{n}$, et $\boldsymbol{\beta}$ un vecteur de coefficients de dimension $D+1$ (D pour le coefficient de chacune des variables explicatives ainsi qu'un de plus pour le paramètre d'interception). En ce sens, il nous apparaît crucial de souligner que l'utilisation d'un tel modèle invoque la reconnaissance implicite du possible effet (du coefficient possiblement non-nul) de chacune des variables explicatives susmentionnées. De même, en vertu des propres hypothèses sur lesquelles le modèle se fonde, l'omission de n'importe laquelle d'entre elles aurait pour conséquence *potentielle* d'engendrer un biais selon une modalité particulière du phénomène

appelé *biais pour variable omise* [18]. Tandis que ce phénomène se rapporte d'ordinaire à l'effet de variables omises de par leur absence dans l'ensemble des régresseurs candidats (les coefficients des unes tentant de *compenser* pour l'absence des autres), et donc inconnues, nous faisons ici plus spécifiquement référence à l'effet de variables sciemment omises *alors* qu'elles se trouvaient disponibles. Nous montrons que ce biais s'annule sous l'hypothèse nulle et nous permet alors de procéder, pourvu que l'on prenne soin de corriger pour la variance.

IV.2.1.2 Distribution échantillonnale de l'estimateur des coefficients

Alors qu'un modèle incluant chacune des variables explicatives candidates produirait l'estimateur des coefficients de la régression suivant

$$\widehat{\beta}_u = \left(\tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{X}}_u\right)^{-1} \tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{Y}}_u \sim \mathscr{N}\left\{\beta, \sigma^2 \left(\tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{X}}_u\right)^{-1}\right\}, \tag{27}$$

celui se basant plutôt sur un sous-ensemble $\tilde{\boldsymbol{X}}_u(B) = \tilde{\boldsymbol{X}}_u \boldsymbol{B}$ de ces mêmes variables explicatives (où $\boldsymbol{B}$ correspondrait à une matrice identité $\boldsymbol{I}$ d'ordre $D+1$ à laquelle on aurait retiré les colonnes correspondant aux variables inactives de B) serait plutôt défini par l'expression

$$\begin{aligned} \widehat{\beta}_u(B) &= \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{Y}}_u \\ &\sim \mathscr{N}\left\{\left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u \beta, \sigma^2 \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B)\right)^{-1}\right\}, \end{aligned} \tag{28}$$

un vecteur de coefficients estimés de dimensions $D(B) \times 1$, où $D(B) = B^T B$ est un scalaire correspondant au nombre de paramètres inclus au modèle B.

IV.2.1.3 Distribution hors-échantillon des résidus

Définissons tout d'abord une matrice $\boldsymbol{H}_u(B)$ qu'il nous conviendra de nommer *matrice chapeau* hors-échantillon

$$\boldsymbol{H}_u(B) = \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \tilde{\boldsymbol{X}}_u^T(B)\,. \tag{29}$$

N'étant pas un projecteur orthogonal, cette matrice n'est également pas idempotente, c'est-à-dire [19]

$$\boldsymbol{H}_u(B)\, \boldsymbol{H}_u^T(B) = \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \boldsymbol{X}_u^T(B) \neq \boldsymbol{H}_u(B)\,.$$

Ces dernières sont ainsi nommées parce que multipliées à $\tilde{\boldsymbol{Y}}_u$, permettent de lui apposer son

(18). La conséquence est dite potentielle en ce sens où la valeur du coefficient β d'une variable donnée pourrait être de 0, rendant son omission pour ainsi dire sans conséquence en matière de biais statistique, et même souhaitable en terme de réduction de la variance.

(19). Contrairement à la matrice chapeau sur-échantillon, $\tilde{\boldsymbol{H}}_u(B) = \tilde{\boldsymbol{X}}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \tilde{\boldsymbol{X}}_u^T(B)$.

proverbial chapeau et ainsi d'engendrer les prévisions $\widehat{\boldsymbol{Y}}_u(B)$ (ainsi que $\widehat{\widetilde{\boldsymbol{Y}}}_u(B)$). En effet,

$$\widehat{\boldsymbol{Y}}_u(B) = \boldsymbol{X}_u(B)\,\widehat{\boldsymbol{\beta}}_u(B) = \boldsymbol{H}_u(B)\,\widetilde{\boldsymbol{Y}}_u \sim \mathscr{N}\Big(\boldsymbol{H}_u(B)\,\tilde{\boldsymbol{X}}_u\beta, \sigma^2\boldsymbol{H}_u(B)\,\boldsymbol{H}_u^T(B)\Big). \tag{30}$$

La distribution des résidus s'énonce alors comme

$$\widehat{\varepsilon}_u(B) = \boldsymbol{Y}_u - \widehat{\boldsymbol{Y}}_u(B) \sim \mathscr{N}\Big\{\Big(\boldsymbol{X}_u - \boldsymbol{H}_u(B)\,\tilde{\boldsymbol{X}}_u\Big)\,\beta, \boldsymbol{\Sigma}_u(B)\Big\}, \tag{31}$$

où la variance de ces derniers est ici donnée par l'expression

$$\begin{aligned}\boldsymbol{\Sigma}_u(B) &= \mathrm{Var}[\boldsymbol{Y}_u] + \mathrm{Var}\Big[\widehat{\boldsymbol{Y}}_u(B)\Big] \\ &= \sigma^2\Big(\boldsymbol{I} + \boldsymbol{H}_u(B)\,\boldsymbol{H}_u^T(B)\Big) \\ &= \sigma^2\left(\boldsymbol{I} + \boldsymbol{X}_u(B)\left(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{X}}_u(B)\right)^{-1}\boldsymbol{X}_u^T(B)\right).\end{aligned} \tag{32}$$

Observons que les résidus ne sont pas centrés en $\mathbf{0}$, les prévisions étant biaisées, mais qu'ils le deviennent sous l'hypothèse nulle (l'inclusion ou l'omission d'une variable sans effet n'impactant pas leur centre). En effet, nous avons bien

$$\mathrm{E}[\widehat{\varepsilon}_u(B)] = \Big(\boldsymbol{X}_u - \boldsymbol{H}_u(B)\,\tilde{\boldsymbol{X}}_u\Big)\,\beta \overset{H_0}{=} \mathbf{0},$$

il s'ensuit donc que la somme réduite des carrés des résidus, de laquelle s'inspire largement la fonction d'ajustement, est distribuée comme

$$\widehat{\varepsilon}_u^T(B)\,\boldsymbol{\Sigma}_u^{-1}(B)\,\widehat{\varepsilon}_u(B) \sim \chi^2_n, \tag{33}$$

alors que, bien au contraire, le cas sur-échantillon nous aurait lui procuré l'expression suivante [20] [21]

$$\frac{1}{\sigma^2}\widehat{\widetilde{\boldsymbol{\varepsilon}}}_u^T(B)\,\widehat{\widetilde{\boldsymbol{\varepsilon}}}_u(B) \sim \chi^2_{\tilde{n}-D(B)}. \tag{34}$$

On remarque en premier lieu qu'aucun degré de liberté ne s'avère perdu à l'équation 33, et que l'incertitude caractérisant les résidus hors-échantillons semble en quelque sorte amplifiée par les différences en matière de tailles, mais de distributions également qui existent entre $\tilde{\boldsymbol{X}}_u(B)$ et $\boldsymbol{X}_u(B)$ [22]. Comme le stipule la section 3.2 de l'article, cette *intégrité des degrés de liberté* s'avère primordiale à nos efforts. En effet, cette statistique ne dépendant d'aucune manière du modèle B, nous sommes assurés qu'une fonction d'ajustement qui se base sur celle-ci

(20). Rappelons qu'en vertu du partitionnement de type « holdout method » appliqué de manière systématique et indépendante à chacun des rééchantillonnages bootstrap de taille N (lesquels sont assignés à chacune des populations de l'algorithme génétique ensaché), deux partitions portant les noms d'*ensemble d'apprentissage* et d'*ensemble de validation* sont créées de manière aléatoire et de telle sorte que leur taille respective soit de $\tilde{n}$ et $n = N - \tilde{n}$.

(21). Notons que $\Big(\sigma^2\left(\boldsymbol{I} - \tilde{\boldsymbol{H}}_u(B)\right)\Big)^+ = \frac{1}{\sigma^2}\boldsymbol{I}$ (où $\boldsymbol{A}^+$ correspond au pseudo-inverse de $\boldsymbol{A}$), et donc que l'équation 34 s'énonce elle aussi sous la forme $\widehat{\widetilde{\boldsymbol{\varepsilon}}}_u^T(B)\,\tilde{\boldsymbol{\Sigma}}_u^{-1}(B)\,\widehat{\widetilde{\boldsymbol{\varepsilon}}}_u(B)$.

(22). On montre d'ailleurs qu'une grossière approximation de l'équation 32 est $\boldsymbol{\Sigma}_u(B) \approx \sigma^2\left(1 + \frac{n}{\tilde{n}}\right)\boldsymbol{I}$, sous l'hypothèse relativement plausible que $\boldsymbol{X}_u \sim \tilde{\boldsymbol{X}}_u$. Ainsi, selon ce résultat, prédire *beaucoup* avec *peu* ($n \gg \tilde{n}$) sera sujet à un plus grand degré d'incertitude que l'inverse, ce qui se trouve de manière très réconfortante conforme à l'intuition.

offrira, *en moyenne*, une chance égale à tout modèle sous l'hypothèse nulle, et surtout pas de les discriminer sur une base arbitraire comme, par exemple, le nombre de prédicteurs que ces derniers incluent, à la manière du critère AIC [2] ou BIC [19]. Nous définissons ainsi la fonction d'ajustement comme une simple transformation de cette statistique, nommément

$$f_u(B) \propto \left\{ \widehat{\boldsymbol{\varepsilon}}_u^T(B)\, \boldsymbol{\Sigma}_u^{-1}(B)\, \widehat{\boldsymbol{\varepsilon}}_u(B) \right\}^{-\gamma}, \tag{35}$$

laquelle ne requiert pas que le paramètre de nuisance, σ^2, soit estimé, étant donné le caractère proportionnel de l'opérateur génétique de sélection (décrit à la section 2.1.3 de l'article). Le paramètre γ permet alors à l'utilisateur de calibrer l'intensité de la dentelure de la fonction d'ajustement et, indirectement, d'influencer le compromis entre erreur de type I et erreur de type II inhérent à tout test d'hypothèse.

IV.2.1.4 Avancées computationnelles

Alors que la statistique définie à l'équation 33 nous a, à toute fin pratique, permis de définir une fonction d'ajustement respectant les critères spécifiés à l'article, la nécessité d'inverser à répétition [(23)] la matrice de variance définie à l'équation 32 peut facilement venir compliquer l'exécution d'un vaste nombre d'univers parallèles, ou encore même d'un nombre plus restreint sur une machine dont les capacités de calculs seraient plus modestes. Or, il se trouve qu'il est possible de simplifier grandement l'évaluation de cet inverse par l'utilisation de l'identité matricielle de Woodbury [10] [(24)]. Posons tout d'abord

$$\boldsymbol{S}_u(B) = \tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B) + \boldsymbol{X}_u^T(B)\, \boldsymbol{X}_u(B)\,, \qquad \boldsymbol{H}_u^*(B) = \boldsymbol{X}_u(B)\, \boldsymbol{S}_u^{-1}(B)\, \boldsymbol{X}_u^T(B)\,,$$

on peut alors récrire l'expression $\boldsymbol{I} + \boldsymbol{H}_u(B)\, \boldsymbol{H}_u^T(B)$ comme

$$\begin{aligned}
\boldsymbol{I} \;+\; & \boldsymbol{H}_u(B)\, \boldsymbol{H}_u^T(B) \\
= \; & \boldsymbol{I} + \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B) \right)^{-1} \boldsymbol{X}_u^T(B) \\
= \; & \boldsymbol{I} + \boldsymbol{X}_u(B) \left(\boldsymbol{S}_u(B) - \boldsymbol{X}_u^T(B)\, \boldsymbol{X}_u(B) \right)^{-1} \boldsymbol{X}_u^T(B) \\
= \; & \boldsymbol{I} + \boldsymbol{X}_u(B) \left(\boldsymbol{S}_u^{-1}(B) + \boldsymbol{S}_u^{-1}(B)\, \boldsymbol{X}_u^T(B)\, (\boldsymbol{I} - \boldsymbol{H}_u^*(B))^{-1} \boldsymbol{X}_u(B)\, \boldsymbol{S}_u^{-1}(B) \right) \boldsymbol{X}_u^T(B) \\
= \; & \boldsymbol{I} + \boldsymbol{H}_u^*(B) + \boldsymbol{H}_u^*(B)\, (\boldsymbol{I} - \boldsymbol{H}_u^*(B))^{-1} \boldsymbol{H}_u^*(B) \\
= \; & \boldsymbol{I} + \boldsymbol{H}_u^*(B)\, (\boldsymbol{I} - \boldsymbol{H}_u^*(B))^{-1}\, (\boldsymbol{I} - \boldsymbol{H}_u^*(B) + \boldsymbol{H}_u^*(B)) \\
= \; & (\boldsymbol{I} - \boldsymbol{H}_u^*(B) + \boldsymbol{H}_u^*(B))\, (\boldsymbol{I} - \boldsymbol{H}_u^*(B))^{-1} \\
= \; & \left\{ \boldsymbol{I} - \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{X}}_u(B) + \boldsymbol{X}_u^T(B)\, \boldsymbol{X}(B) \right)^{-1} \boldsymbol{X}_u^T(B) \right\}^{-1}.
\end{aligned} \tag{36}$$

(23). C'est-à-dire à chaque évaluation de la fonction d'ajustement, d'où la complexité algorithmique $\mathcal{O}(IUG)$ mentionnée à la section 5.3 de l'article.

(24). Cette dernière généralise une identité dans $\mathbb{R}$ selon laquelle $(1+x)^{-1} = 1 - x(1+x)^{-1}$, et nous permet à notre tour de généraliser l'identité suivante : $1 + \frac{x^2}{\tilde{x}^2} = \left(1 - \frac{x^2}{\tilde{x}^2 + x^2}\right)^{-1}$.

Ainsi, l'inversion de la matrice de variance à l'équation 35 peut être complètement évitée en substituant à cette dernière

$$\boldsymbol{\Sigma}_u^{-1}(B) \propto \boldsymbol{I} - \boldsymbol{X}_u(B)\left(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{X}}_u(B) + \boldsymbol{X}_u^T(B)\,\boldsymbol{X}_u(B)\right)^{-1}\boldsymbol{X}_u^T(B)\,. \tag{37}$$

IV.2.2 Extension aux modèles linéaires généralisés

Nous avons démontré à la sous-section précédente en quoi la fonction d'ajustement que nous avions retenue fait sens au sein de son contexte d'origine (les modèles linéaires), c'est-à-dire en quoi le respect des critères énoncés à la section 3.2 de l'article, cette dernière garantissait aux probabilités de rétention de posséder la distribution introduite à son tour à la section 3.3 de l'article. Dans un même ordre d'idées, nous détaillerons ici comment et pourquoi cette dernière s'accommode aisément aux modèles linéaires généralisés, lesquels incluent la loi normale mise en vedette chez les modèles *simplement* linéaires. Pour ce faire, nous déterminerons dans un premier temps la distribution de l'estimateur des coefficients lorsque toutes les variables se trouvent incluses au modèle. Nous lui appliquerons ensuite les corrections nécessaires afin qu'il tienne compte d'un sous-ensemble des prédicteurs seulement et de l'intarissable biais pour variables omises. Enfin, nous nous intéresserons à la distribution des résidus hors-échantillon. Diverses percées computationnelles seront présentées.

IV.2.2.1 Fondations du modèle

Réintroduisons brièvement le concept de modèles linéaires généralisés, ou GLM [(25)], sous l'hypothèse du choix de la fonction de lien canonique, introduite à la section 3.4 de l'article. À tout coup, la fonction de vraisemblance du GLM appartiendra à la famille exponentielle linéaire [16], laquelle est définie comme suit

$$L\Big(\tilde{\boldsymbol{Y}}_u \,\Big|\, \tilde{\boldsymbol{X}}_u\Big) = \exp\left\{\frac{1}{a(\phi)}\left(\tilde{\boldsymbol{\eta}}_u^T\tilde{\boldsymbol{Y}}_u - b(\tilde{\boldsymbol{\eta}}_u)\right) + c\Big(\tilde{\boldsymbol{Y}}_u, \phi\Big)\right\}, \tag{38}$$

où $\tilde{\boldsymbol{\eta}}_u$ symbolise la combinaison linéaire des variables explicatives candidates, c'est-à-dire

$$\tilde{\boldsymbol{\eta}}_u = \tilde{\boldsymbol{X}}_u\beta = \Big[\tilde{\boldsymbol{x}}_{u,1}^T\beta \quad \cdots \quad \tilde{\boldsymbol{x}}_{u,\tilde{n}}^T\beta\Big]^T = \Big[\tilde{\eta}_{u,1} \quad \cdots \quad \tilde{\eta}_{u,\tilde{n}}\Big]^T. \tag{39}$$

Sous cette formulation, l'espérance de la variable réponse est donnée par l'expression

$$\mathrm{E}\Big[\tilde{\boldsymbol{Y}}_u\Big] = \tilde{\boldsymbol{\mu}}_u = \ell(\tilde{\boldsymbol{\eta}}_u) = \Big[\ell(\tilde{\eta}_1) \quad \cdots \quad \ell(\tilde{\eta}_{\tilde{n}})\Big]^T, \tag{40}$$

où la fonction ℓ est dite *fonction de lien inverse* ou encore *fonction de moyenne*, laquelle respecte

(25). De l'anglais, « Generalized Linear Model ».

l'égalité $\ell(\tilde{\nu}_{u,i}) = b'(\tilde{\nu}_{u,i})$. À son tour, la variance est donnée par l'expression

$$\mathrm{Var}\Big[\tilde{\boldsymbol{Y}}_u\Big] = a(\phi)\,\tilde{\boldsymbol{V}}_u = a(\phi)\,v(\tilde{\boldsymbol{\mu}}_u) = \begin{bmatrix} a(\phi)\,v(\tilde{\mu}_{u,1}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a(\phi)\,v(\tilde{\mu}_{u,\tilde{n}}) \end{bmatrix}, \tag{41}$$

où la fonction de variance, permettant en outre d'exprimer la variance comme une simple transformation de l'espérance, respecte la relation d'équivalence $v(\tilde{\mu}_{u,i}) = b''(\tilde{\eta}_{u,i})$.

IV.2.2.2 Distribution échantillonnale de l'estimateur des coefficients

L'estimateur des coefficients de la régression étant issu de la méthode du maximum de vraisemblance, ce dernier doit être de distribution asymptotiquement normale. Nous montrerons dans un premier temps quels en sont les paramètres lorsque toutes les variables explicatives sont incluses au modèle. Nous procéderons alors, à l'instar des travaux de la sous-section IV.2.1, à en déterminer les paramètres pour tout sous-ensemble B des variables données, c'est-à-dire en tenant spécifiquement compte de l'effet du biais d'omission de variables.

Par les propriétés des estimateurs du maximum de vraisemblance, nous savons que la distribution de $\widehat{\boldsymbol{\beta}}_u$ est asymptotiquement normale, c'est-à-dire [(26)]

$$\sqrt{\tilde{n}}\left(\widehat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \overline{\mathcal{I}}_{\tilde{n}}^{-1}(\boldsymbol{\beta})\right), \tag{42}$$

où $\overline{\mathcal{I}}_{\tilde{n}}(\boldsymbol{\beta})$ est l'*information de Fisher moyenne*, ou simplement *information de Fisher*, laquelle est définie comme la quantité d'information contenue dans une seule observation. Disposant ici d'un échantillon au sein duquel la contribution d'une observation se trouve être variable, cette quantité devient alors la quantité *moyenne* d'information que véhicule les observations. Ainsi, en vertu de la propriété d'additivité de l'information de Fisher, sa valeur moyenne est alors donnée par [6]

$$\mathcal{I}_{\tilde{n}}(\boldsymbol{\beta}) = \tilde{n}\overline{\mathcal{I}}_{\tilde{n}}(\boldsymbol{\beta}) = -\mathrm{E}\Big[\tfrac{\partial^2}{\partial \boldsymbol{\beta}^2} \log\Big(L\Big(\boldsymbol{\beta} \,\Big|\, \tilde{\boldsymbol{Y}}_u, \tilde{\boldsymbol{X}}_u\Big)\Big)\Big]. \tag{43}$$

Suite à quoi, il nous sera possible d'isoler $\widehat{\boldsymbol{\beta}}_u$ de l'équation 42, pour ainsi obtenir

$$\widehat{\boldsymbol{\beta}}_u \xrightarrow{\mathscr{L}} \mathscr{N}\left(\boldsymbol{\beta}, \tilde{n}\overline{\mathcal{I}}_{\tilde{n}}(\boldsymbol{\beta})\right) \sim \mathscr{N}\left(\boldsymbol{\beta}, \mathcal{I}_{\tilde{n}}(\boldsymbol{\beta})\right). \tag{44}$$

Or, il s'avère préférable de scinder le calcul de cette quantité en plusieurs sous-étapes. Par la règle de dérivation en chaîne, et plus spécifiquement la formule de Faà di Bruno [3], nous

(26). Nous faisons référence à la taille d'échantillon $\tilde{n}$, car l'estimation des coefficients de la régression est réalisée, par définition, sur-échantillon, par opposition aux prévisions qui dans notre cas sont réalisées hors-échantillon.

montrons d'abord que

$$\begin{array}{rcl} \frac{\partial \tilde{\boldsymbol{\eta}}_u}{\partial \boldsymbol{\beta}} & = & \boldsymbol{X}_u, \\ \frac{\partial^2 \tilde{\boldsymbol{\eta}}_u}{\partial \boldsymbol{\beta}^2} & = & \mathbf{0}, \\ \frac{\partial b}{\partial \tilde{\boldsymbol{\beta}}} & = & \left(\frac{\partial \tilde{\boldsymbol{\eta}}_u}{\partial \boldsymbol{\beta}}\right)^T \frac{\partial b}{\partial \tilde{\boldsymbol{\eta}}_u} = \tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{\mu}}_u, \\ \frac{\partial^2 b}{\partial \boldsymbol{\beta}^2} & = & \left(\frac{\partial \tilde{\boldsymbol{\eta}}_u}{\partial \boldsymbol{\beta}}\right)^T \frac{\partial^2 b}{\partial \tilde{\boldsymbol{\eta}}_u^2} \frac{\partial \tilde{\boldsymbol{\eta}}_u}{\partial \boldsymbol{\beta}} + \left(\frac{\partial^2 \tilde{\boldsymbol{\eta}}_u}{\partial \boldsymbol{\beta}^2}\right)^T \frac{\partial b}{\partial \tilde{\boldsymbol{\eta}}_u} = \tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u + \mathbf{0}^T \tilde{\boldsymbol{\mu}}_u = \tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u. \end{array} \tag{45}$$

De l'équation 38, nous obtenons la fonction de log-vraisemblance

$$\log\Big(L\Big(\beta \,\Big|\, \tilde{\boldsymbol{Y}}_u, \tilde{\boldsymbol{X}}_u\Big)\Big) = \frac{1}{a(\phi)} \left(\tilde{\eta}^T \tilde{\boldsymbol{Y}}_u - b(\tilde{\boldsymbol{\eta}}_u)\right) + c\Big(\tilde{\boldsymbol{Y}}_u, \phi\Big). \tag{46}$$

La fonction de score est alors définie comme la dérivée première de cette dernière

$$\frac{\partial \log(L)}{\partial \beta} = \frac{1}{a(\phi)} \tilde{\boldsymbol{X}}_u^T \left(\tilde{\boldsymbol{Y}}_u - \tilde{\boldsymbol{\mu}}_u\right), \tag{47}$$

et la matrice hessienne à son tour est définie comme sa dérivée seconde

$$\frac{\partial^2 \log(L)}{\partial \beta^2} = -\frac{1}{a(\phi)} \tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u. \tag{48}$$

On conclut donc que la distribution échantillonnale des coefficients de la régression est

$$\widehat{\boldsymbol{\beta}}_u \xrightarrow{\mathscr{L}} \mathscr{N}\left(\beta, a(\phi) \left(\tilde{\boldsymbol{X}}_u^T \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u\right)^{-1}\right). \tag{49}$$

IV.2.2.3 Distribution hors-échantillon des résidus

Notre utilisation des algorithmes génétiques nous amenant, par construction, à nous limiter à un sous-ensemble B des colonnes de $\tilde{\boldsymbol{X}}_u$, nous sommes contraints de nous intéresser à la distribution de $\widehat{\boldsymbol{\beta}}_u(B)$, laquelle est donnée, à l'instar des calculs présentés à l'équation 28, par [(27)]

$$\widehat{\boldsymbol{\beta}}_u(B) \xrightarrow{\mathscr{L}} \mathscr{N}\left\{\left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u \beta,\ a(\phi) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u(B)\right)^{-1}\right\}, \tag{50}$$

laquelle se simplifie à l'équation 49 lorsque toutes les variables sont incluses au modèle (c'est-à-dire lorsque $\boldsymbol{X}_u(B) = \boldsymbol{X}_u$). On remarque ainsi que l'estimateur des coefficients de la régression s'en trouve à nouveau en proie au biais pour variables omises, tout comme c'était le cas avec

(27). Une illustration encore plus convaincante de ce résultat intermédaire est obtenue en dérivant la distribution échantillonnale de l'estimateur des coefficients d'un modèle linéaire hétéroscédastique. En effet, toute chose étant égale par ailleurs (quant aux hypothèses du modèle linéaire), si $\tilde{\boldsymbol{Y}}_u \,\big|\, \tilde{\boldsymbol{X}}_u \sim \mathscr{N}\left(\tilde{\boldsymbol{X}}_u \boldsymbol{\beta}, \tilde{\boldsymbol{V}}_u\right)$, où $\tilde{\boldsymbol{V}}_u$ est une matrice de variance diagonale, fixe et connue d'avance, alors $\widehat{\boldsymbol{\beta}}_u(B) \sim \mathscr{N}\left\{\left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{V}}_u^{-1} \tilde{\boldsymbol{X}}_u(B)\right)^{-1} \tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{V}}_u^{-1} \tilde{\boldsymbol{X}}_u \boldsymbol{\beta},\ a(\phi) \left(\tilde{\boldsymbol{X}}_u^T(B)\, \tilde{\boldsymbol{V}}_u^{-1} \tilde{\boldsymbol{X}}_u(B)\right)^{-1}\right\}$.

la régression linéaire multiple, c'est-à-dire

$$\text{Biais}\Big[\widehat{\boldsymbol{\beta}}_u(B)\Big] = \Big\{\Big(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u(B)\Big)^{-1}\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u - \boldsymbol{I}\Big\}\,\boldsymbol{\beta} \neq \mathbf{0}. \tag{51}$$

Cela n'a cependant pas pour conséquence d'invalider un seul de nos résultats, pour la simple et bonne raison que ce biais s'annule sous l'hypothèse nulle, en effet, $\text{Biais}\Big[\widehat{\boldsymbol{\beta}}_u(B)\Big] \overset{H_0}{=} \mathbf{0}$. L'estimateur $\widehat{\boldsymbol{\eta}}_u(B) = \boldsymbol{X}_u(B)\,\widehat{\boldsymbol{\beta}}_u(B)$ de la combinaison linéaire $\boldsymbol{\eta}_u(B) = \boldsymbol{X}_u(B)\,\boldsymbol{\beta}$ étant à son tour linéaire en $\widehat{\boldsymbol{\beta}}_u(B)$, lequel est gaussien, il n'en est lui-même pas moins gaussien, et ce, avec espérance

$$\text{E}[\widehat{\boldsymbol{\eta}}_u(B)] = \boldsymbol{X}_u(B)\,\Big(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u(B)\Big)^{-1}\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \tag{52}$$

ainsi que variance

$$\text{Var}[\widehat{\boldsymbol{\eta}}_u(B)] = a(\phi)\,\boldsymbol{X}_u(B)\,\Big(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u(B)\Big)^{-1}\boldsymbol{X}_u^T(B)\,. \tag{53}$$

Ainsi, force est de constater que le biais qui affligeait $\widehat{\boldsymbol{\beta}}_u(B)$ s'est pour ainsi dire propagé à $\widehat{\boldsymbol{\eta}}_u(B)$. En effet,

$$\text{Bais}[\widehat{\boldsymbol{\eta}}_u(B)] = \boldsymbol{X}_u(B)\,\Big\{\Big(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u(B)\Big)^{-1}\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u - \boldsymbol{I}\Big\}\,\boldsymbol{\beta} \neq \mathbf{0}.$$

Une hypothèse nettement moins forte que H_0 et malgré tout imbriquée à cette dernière est toutefois de supposer les coefficients des variables non-incluses au modèle décrit par le vecteur B sont tous de 0. Annihilant du coup le biais précédemment observé, cela nous permet alors, en vertu de l'équation 42, d'appliquer la méthode δ [23] et ainsi d'affirmer que le vecteur des prévisions $\widehat{\boldsymbol{Y}}_u(B) = \widehat{\boldsymbol{\mu}}_u(B) = \ell(\widehat{\boldsymbol{\eta}}_u(B))$ est asymptotiquement normal de moyenne $\text{E}\Big[\widehat{\boldsymbol{Y}}_u(B)\Big] = \boldsymbol{\mu}_u$, certes, mais de variance [(28)]

$$\text{Var}\Big[\widehat{\boldsymbol{Y}}_u(B)\Big] = a(\phi)\,\boldsymbol{V}_u\boldsymbol{X}_u(B)\,\Big(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u(B)\Big)^{-1}\boldsymbol{X}_u^T(B)\,\boldsymbol{V}_u. \tag{54}$$

Cela engendre dès lors des résidus hors-échantillon $\widehat{\boldsymbol{\varepsilon}}_u(B) = \boldsymbol{Y}_u - \widehat{\boldsymbol{Y}}_u(B)$ d'espérance nulle $\text{E}[\widehat{\boldsymbol{\varepsilon}}_u(B)] = \mathbf{0}$ et de variance

$$\begin{aligned}\boldsymbol{\Sigma}_u(B) &= \text{Var}[\widehat{\boldsymbol{\varepsilon}}_u(B)] \\ &= \text{Var}[\boldsymbol{Y}_u] + \text{Var}\Big[\widehat{\boldsymbol{Y}}_u(B)\Big] \\ &\propto \boldsymbol{V}_u + \boldsymbol{V}_u\boldsymbol{X}_u(B)\,\Big(\tilde{\boldsymbol{X}}_u^T(B)\,\tilde{\boldsymbol{V}}_u\tilde{\boldsymbol{X}}_u(B)\Big)^{-1}\boldsymbol{X}_u^T(B)\,\boldsymbol{V}_u,\end{aligned} \tag{55}$$

de par l'indépendance entre $\boldsymbol{Y}_u$ et $\widehat{\boldsymbol{Y}}_u(B)$ que nous confère l'utilisation d'un ensemble distinct de validation. Notons également que la non-connaissance de $\boldsymbol{V}_u$ (en effet, cette dernière dépendant des véritables paramètres, et non pas de leurs estimés) nous contraint à la remplacer par sa meilleure estimation sous le modèle courant, B, donnant ainsi lieu à la matrice estimée de la

(28). Additionnellement, on remarque que la matrice $\boldsymbol{V}_u$ correspond au Jacobien de la transformation.

variance suivante

$$\widehat{\boldsymbol{\Sigma}}_u(B) \propto \widehat{\boldsymbol{V}}_u(B) + \widehat{\boldsymbol{V}}_u(B)\,\boldsymbol{X}_u(B)\left(\tilde{\boldsymbol{X}}_u^T(B)\,\hat{\tilde{\boldsymbol{V}}}_u(B)\,\tilde{\boldsymbol{X}}_u(B)\right)^{-1}\boldsymbol{X}_u^T(B)\,\widehat{\boldsymbol{V}}_u(B)\,, \tag{56}$$

où les matrices $\widehat{\boldsymbol{V}}_u(B)$ et $\hat{\tilde{\boldsymbol{V}}}_u(B)$ sont estimées à l'aide des prévisions générées par le modèle, c'est-à-dire $\widehat{\boldsymbol{V}}_u(B) = v\big(\widehat{\boldsymbol{Y}}_u(B)\big)$ et $\hat{\tilde{\boldsymbol{V}}}_u(B) = v\big(\hat{\tilde{\boldsymbol{Y}}}_u(B)\big)$. S'il n'avait pas été nécessaire de procéder ainsi avec les modèles linéaires, c'est bien parce que de manière unique au sein de la famille exponentielle linéaire, la variance de la loi normale ne dépend pas de son espérance. C'est d'ailleurs pourquoi sa fonction de lien est dite *fonction identité*.

Dans tous les cas, la fonction d'ajustement peut alors s'énoncer comme

$$f_u(B) \propto \left\{\widehat{\boldsymbol{\varepsilon}}_u^T(B)\,\widehat{\boldsymbol{\Sigma}}_u^{-1}(B)\,\widehat{\boldsymbol{\varepsilon}}_u(B)\right\}^{-\gamma}, \tag{57}$$

en effet, le paramètre de nuisance ϕ s'appliquant de manière multiplicative, celui-ci se verra à tout coup annulé, encore une fois de par notre choix d'un opérateur génétique de sélection proportionnelle, il devient ainsi superflu de l'estimer.

Alors qu'il n'a été possible ici d'établir analytiquement la distribution, même asymptotique, de la somme réduite des carrés des résidus, nous avons tout de même montré que sous l'hypothèse nulle, cette statistique consiste somme toute à déterminer la somme d'une séquence de variables aléatoires d'espérance nulle et de variance unitaire. Il convient donc d'accepter comme raisonnablement valide la présomption d'une fonction d'ajustement juste et équitable, et ce, spécialement en ce qui a trait à de faibles valeurs du paramètre de dentelure γ, pour lequel une valeur de 1.5 est justement recommandée. En effet, nonobstant la non-normalité de $\boldsymbol{Y}_u$, à la manière du test de Wald ou du test du χ^2 de Pearson, nous ne croyons pas déraisonnable de prétendre que sous H_0

$$\widehat{\boldsymbol{\varepsilon}}_u^T(B)\,\boldsymbol{\Sigma}_u^{-1}(B)\,\widehat{\boldsymbol{\varepsilon}}_u(B) \approx \chi_n^2, \tag{58}$$

et, par extension, d'affirmer que

$$\widehat{\boldsymbol{\varepsilon}}_u^T(B)\,\widehat{\boldsymbol{\Sigma}}_u^{-1}(B)\,\widehat{\boldsymbol{\varepsilon}}_u(B) \approx \chi_n^2. \tag{59}$$

IV.2.2.4 Avancées computationnelles

Enfin, à l'instar des résultats présentés à l'équation 36, il est également possible ici de simplifier l'inverse de la variance obtenue à l'équation 56. En effet, si l'on pose $\tilde{\boldsymbol{S}}_u(B) = \tilde{\boldsymbol{V}}_u^{1/2}\tilde{\boldsymbol{X}}_u(B)$ et $\boldsymbol{S}_u(B) = \boldsymbol{V}_u^{1/2}\boldsymbol{X}_u(B)$, il devient alors possible de réorganiser les termes de l'équation 55 comme

suit [29]

$$\begin{aligned} \boldsymbol{V}_u &+ \boldsymbol{V}_u \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B) \, \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u(B) \right)^{-1} \boldsymbol{X}_u^T(B) \, \boldsymbol{V}_u \\ &= \boldsymbol{V}_u^{1/2} \left\{ \boldsymbol{I} + \boldsymbol{S}_u(B) \left(\tilde{\boldsymbol{S}}_u^T(B) \, \tilde{\boldsymbol{S}}_u(B) \right)^{-1} \boldsymbol{S}_u^T(B) \right\} \boldsymbol{V}_u^{1/2} \\ &= \boldsymbol{V}_u^{1/2} \left\{ \boldsymbol{I} - \boldsymbol{S}_u(B) \left(\tilde{\boldsymbol{S}}_u^T(B) \, \tilde{\boldsymbol{S}}_u(B) + \boldsymbol{S}_u^T(B) \, \boldsymbol{S}_u(B) \right)^{-1} \boldsymbol{S}_u^T(B) \right\}^{-1} \boldsymbol{V}_u^{1/2} \\ &= \left\{ \boldsymbol{V}_u^{-1} - \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B) \, \tilde{\boldsymbol{V}}_u \tilde{\boldsymbol{X}}_u(B) + \boldsymbol{X}_u^T(B) \, \boldsymbol{V}_u \boldsymbol{X}_u(B) \right)^{-1} \boldsymbol{X}_u^T(B) \right\}^{-1}. \end{aligned} \tag{60}$$

Il convient donc de remplacer l'inverse matriciel par l'expression que voici

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_u^{-1}(B) &\propto \widehat{\boldsymbol{V}}_u^{-1}(B) \\ &- \boldsymbol{X}_u(B) \left(\tilde{\boldsymbol{X}}_u^T(B) \, \widehat{\tilde{\boldsymbol{V}}}_u(B) \, \tilde{\boldsymbol{X}}_u(B) + \boldsymbol{X}_u^T(B) \, \widehat{\boldsymbol{V}}_u(B) \, \boldsymbol{X}_u(B) \right)^{-1} \boldsymbol{X}_u^T(B) \, . \end{aligned} \tag{61}$$

(29). Cette opération généralise ainsi l'identité $\sigma^2 \left(1 + \frac{\sigma^2 x^2}{\tilde{\sigma}^2 \tilde{x}^2} \right) = \left(\frac{1}{\sigma^2} - \frac{x^2}{\tilde{\sigma}^2 \tilde{x}^2 + \sigma^2 x^2} \right)^{-1}$.

V Éléments de discussion additionnels

Les algorithmes génétiques ont tantôt été décrits comme des heuristiques, tantôt comme des métaheuristiques, et ce, en raison de leur modularité, laquelle leur confère une considérable versatilité et capacité d'adaptation. Cette classe d'algorithmes est en réalité une technique d'optimisation à large spectre que nous, et d'autres avant nous, ont su façonner en un outil bien adapté aux réalités du problème de la sélection des variables. C'est bien à cette étape critique que la modélisation se révèle être un art au même titre qu'une science. En effet, là où nous avons cru bon et justifié de privilégier certains mécanismes d'actions (ou opérateurs génétiques, pour récupérer le jargon du milieu), d'autres ont préféré miser sur leurs alternatives. Le but de ce court segment est de mieux illustrer les pours et les contres des opérateurs (les classiques, tout comme ceux introduits dans ces recherches) et ainsi jeter une lumière nouvelle sur la richesse des algorithmes génétiques.

V.1 Sélection

Le rôle de la sélection est avant toute chose d'assurer la convergence de la population (vers un optimum local susceptible de changer d'une population à l'autre) en retenant les individus forts des faibles. À chaque génération, les individus de la population des parents se voient attribuer un score lequel est mesuré au moyen d'une fonction d'ajustement. Les individus sont alors typiquement sélectionnés pour la reproduction avec probabilité proportionnelle à leur score d'ajustement relatif à l'ensemble de la population, à l'instar de la méthodologie employée dans l'article. Une fonction d'ajustement trop *pointue*, c'est-à-dire qui attribuerait des valeurs démesurément grandes aux meilleurs individus de la population aura tôt fait de créer un goulot d'étranglement, annihilant par le fait même toute diversité génétique et réduisant sévèrement la capacité de l'algorithme à explorer l'espace solution. Au contraire, une fonction d'ajustement trop *plate*, c'est-à-dire qui attribuerait des scores relativement similaires peut indépendamment du pouvoir prédictif du modèle, réduire l'algorithme à une simple et primitive recherche aléatoire. Il convient également d'observer que la justesse d'une fonction d'ajustement se trouve immanquablement influencée par la taille des populations à l'œuvre. En effet, qu'importe leur score d'ajustement, les meilleurs individus auront peine à se démarquer d'un vaste bassin de population et ainsi le praticien aura tendance à favoriser des fonctions d'ajustement plus pointue ; conséquemment, la probabilité d'une erreur de type II s'en trouvera indubitablement affectée.

V.2 Élitisme

En marge de cette approche qui vient d'être décrite se trouve l'approche employée par Zhu & Chipman [25]. Plutôt que de sélectionner les individus de manière proportionnelle à leur valeur de score, les individus sont ordonnés selon leur valeur de score et le meilleur $100\alpha\%$ est

directement recopié dans la population des enfants. Le reste de la population est alors comblé de manière usuelle, mais en sélectionnant les parents de manière totalement aléatoire. Bien qu'inorthodoxe, il est possible de s'imaginer comment la population se dirigera lentement, mais sûrement vers un état de convergence en retenant systématiquement les meilleurs individus. Bien qu'analytiquement trop complexe à traiter, il convient malgré tout de souligner que cette alternative a pour avantage de quasi éliminer le risque de goulot d'étranglement [(30)] et, surtout, de ne pas dépendre de la taille de la population.

V.3 Croisement

Le croisement consiste à combiner les gènes de deux individus de manière à produire un nouvel individu, un enfant, qui, retenant les caractéristiques communes aux deux parents et misant sur le hasard pour les autres, possédera le potentiel d'atteindre un état supérieur. Alors que Zhu & Chipman [25] ont opté pour le *croisement en un point*, lequel consiste à permuter les sous-chaînes binaires des parents, c'est-à-dire les sous-chaînes inférieures et supérieures à un point de coupe sélectionné au hasard, nous avons cru plus judicieux d'opter pour un *croisement uniforme* où chaque gène de l'enfant de manière indépendante a 50% de chance de correspondre à celui du père ou de la mère. Notre préférence repose entièrement sur le principe selon lequel l'ordre d'encodage des variables explicatives au sein du modèle (ou de manière plus générale, d'une solution candidate) ne devrait pas avoir d'incidence sur la capacité de l'algorithme à produire de bons modèles. On peut facilement imaginer, par exemple, qu'un problème à 60 variables où seules les 45 dernières seraient utiles sera facilement résolu par l'emploi d'un croisement à un point et d'une haute probabilité d'activation, à l'instar du « *hard problem* » décrit par Zhu & Chipman [25], mais originellement énoncé par George & McCulloch [9].

V.4 Mutation

Le dernier opérateur génétique classique est l'opérateur de mutation. Ce dernier ne possède virtuellement pas d'alternative dans le cadre de problèmes encodés au moyen de chaînes binaires, et, conséquemment, il n'est pas surprenant de retrouver le même opérateur tant chez Zhu & Chipman [25] que chez Larocque *et al* [15].

V.5 Recuit simulé

L'utilisation d'une séquence de taux de mutation $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \cdots & \theta_G\end{bmatrix}^T$ non-constante s'inscrit au sein d'une gamme de techniques que l'on appelle *recuit simulé* (de l'anglais « *simulated annealing* » [14]), lesquelles s'inspirent d'un procédé métallurgique consistant à refroidir et puis réchauffer un métal en séquence de manière à positionner les molécules dans une conformation

(30). Par *risque de goulot d'étranglement*, nous faisons référence à une population qui aurait *convergée* (dont la diversité génétique serait très faible) avant que le compte des G générations n'ait été atteint, et donc ne pourrait mettre à profit les itérations résiduelles.

de moindre énergie (et donc plus stable). Si grands et faibles taux de mutation sont respectivement conçus comme une alternance d'épisodes de réchauffement et de refroidissement, on peut s'imaginer en quoi sous ces circonstances l'algorithme génétique appartient également à la classe du recuit simulé. Alors que sous le couvert des algorithmes génétiques ensachés nous n'avons pas réussi à justifier l'utilisation d'une séquence non-constante de taux de mutation, il nous apparaît évident que sous le couvert des algorithmes génétiques *simples*, une telle séquence serait certes en mesure d'établir une certaine forme de *convergence*.

V.6 Ensachage

L'algorithme mis de l'avant dans ce mémoire et l'article y étant contenu fait intervenir le concept d'ensachage, lequel, loin de constituer ce que l'on pourrait qualifier d'opérateur génétique *classique*, n'en demeure pas moins un pilier fondateur en permettant aux résultats provenant de chaque univers (les proportions de rétention intra-populations *ultimes*, $\widehat{\boldsymbol{\pi}}_{Gu} = \begin{bmatrix}\widehat{\pi}_{1Gu} & \cdots & \widehat{\pi}_{DGu}\end{bmatrix}^T$, $u = 1, \ldots, U.$) d'être analysés sous la présomption d'indépendance. Les jeux de données de vastes tailles seront souvent difficiles à analyser, en raison de temps d'exécution impraticables, ou de par l'impossibilité de charger simultanément les données en mémoire, ou encore de par leur répartition sur une architecture distribuée. En cette ère des mégadonnées, cela constitue un phénomène non pas seulement plausible, mais dont la présence est portée à s'intensifier au courant des prochaines années. À cet effet, ayant nous même été confrontés à celui-ci au courant de nos batteries de tests empiriques, il nous apparut que ce *fléau* pouvait en fait être tourné en un avantage, et ce, tant au niveau computationnel que statistique. En effet, alors que chacune des populations de l'algorithme opère sur la base d'un rééchantillonnage bootstrap des données d'origine (c'est-à-dire que la fonction d'ajustement, f_u, est évaluée sur la base d'un échantillon bootstrap), lequel implique en soi un certain coût computationnel, il se trouve qu'il serait plutôt possible de scinder ces dernières en U partitions disjointes (pourvu que les données soient disponibles en quantité suffisante, et c'est bien là le cas des mégadonnées), garantissant ainsi par construction l'indépendance statistique des partitions et permettant également de réduire le volume simultané de données à traiter (estimation des coefficients de la régression, calculs des résidus, etc.). Par exemple, un jeu de données comprenant 100 000 observations (un nombre relativement restreint, mais suffisant pour les besoins de cet exemple) pourrait être assigné de manière aléatoire et sans remise en lots de 1000 à 100 univers, suite à quoi l'algorithme génétique *ensaché* pourrait être lancé comme tel, sans que jamais il ne soit nécessaire de rééchantillonner les données d'origine. Conséquemment, par l'entremise de cet artifice, nos études empiriques ont pu générer de bons résultats, aussi probants que ceux n'employant pas ce dernier, et ce, à une fraction du temps d'exécution.

VI Conclusion

La sélection des variables est un aspect fondamental de l'analyse prédictive portée à croître en importance au fur et à mesure que la dimensionnalité des problèmes à résoudre continuera à prendre de l'expansion. Un modèle prédictif mal spécifié comporte un biais inhérent et court ainsi le risque de déformer le portrait que tente de dépeindre les variables explicatives incluses sous celui-ci. Ce problème peine toutefois à s'exprimer sous une mesure sur-échantillon. Une évaluation hors-échantillon permet de rectifier la situation. Les algorithmes génétiques ensachés sont présentés comme un raffinement naturel et nécessaire des algorithmes génétiques parallèles de Zhu & Chipman [25] en imposant à ces derniers un cadre statistique rigoureux et objectif. L'effet de substitution ou proxy complexifie grandement le travail de la sélection des variables, mais le BGA sait s'en affranchir. Loin de se limiter aux modèles linéaires, ce dernier s'est montré très adaptable en travaillant de concert avec les modèles linéaires généralisés par l'entremise d'une simple modification de la méthode de réduction des résidus prédictifs. Cela laisse entrevoir la possibilité bien réelle d'étendre la méthode à davantage de modèles de prévisions, comme la régression Tweedie [12], afin de couvrir d'autres champs d'application de la discipline étendue des statistiques.

À travers nos explorations, nous nous sommes interrogés quant à la meilleure façon de recombiner les probabilités de rétention intra-populations que produisent chacun des algorithmes génétiques, lesquels sont exécutés de manière indépendante et parallèle. Telle que la méthodologie décrite à la section 2.2 de l'article le stipule, et elle-même mise de l'avant par Zhu & Chipman avant nous, cette recombinaison consiste à calculer la *grande moyenne* de ces dernières. Cela a pour conséquence de permettre à l'effet de *grands aléas* de se propager avec aisance vers les probabilités de sélection finales, et donc, d'impacter de manière importante la conclusion des sous-tests d'hypothèse. Alternativement, nous avons considéré la possibilité de tester de manière intermédiaire l'hypothèse nulle sur la base des proportions de rétention intra-univers. La moyenne des décisions (une décision étant symbolisée à l'aide d'un indicateur binaire), et non pas des probabilités de rétention spécifiques, serait alors calculée de manière à obtenir ce qu'il nous convient d'appeler *proportion de rétention corrigée*. Ce sont alors ces dernières proportions qui feraient office de statistique du test et permettraient de déterminer l'inclusion ou l'exclusion finale des variables explicatives. Bien que ces proportions corrigées jouissent d'une puissance statistique réduite, leur très grande robustesse, spécialement face à l'effet proxy, suffit largement à les rendre dignes de mention. Il nous apparaît évident que si ces deux façons de procéder pouvaient être hybridées, à l'instar de la *médienne* [13] [(31)], de manière à produire le

(31). La *médienne*, $\overline{m}$, est une mesure robuste de la valeur centrale initialement introduite en 1818 par Laplace et définie comme $\overline{m} = \left(\frac{\overline{x}}{\sigma^2_{\overline{x}}} + \frac{m}{\sigma^2_m}\right) / \left(\frac{1}{\sigma^2_{\overline{x}}} + \frac{1}{\sigma^2_m}\right)$, où m correspond ici à la médiane échantillonnale. Cette dernière, se balançant savamment entre L_1 et L_2, permet d'allier réactivité et stabilité de manière à grandement amortir l'effet d'observations extrêmes ou aberrantes dans l'échantillon.

compromis idéal entre robustesse et puissance, stabilité et réactivité, nous disposerions alors d'un outil dont le regard serait à même de percer les jeux de données les plus menus.

VII Liste des notations

VII.1 Notations de portée générale

Expression	Définition
N	Taille du jeu de données initiales.
$\tilde{n} \quad n$	Les tailles des ensembles d'apprentissage et de validation, respectivement, lesquelles respectent l'égalité $N = \widetilde{n} + n$.
$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_D \end{bmatrix}^T$	Le vecteur des coefficients de la régression (incluant le paramètre d'interception).
$\widetilde{\boldsymbol{Y}}_u \quad \boldsymbol{Y}_u$	Les observations de la variable réponse (sur-échantillon et hors-échantillon, respectivement) de l'univers u (obtenues par bootstrap sur les données d'origine).
$\widetilde{\boldsymbol{X}}_u \quad \boldsymbol{X}_u$	Les observations des variables explicatives (sur-échantillon et hors-échantillon, respectivement) de l'univers u (obtenues par bootstrap sur les données d'origines, et incluant une première colonne de 1 afin d'estimer le paramètre d'interception β_0).
D	Le nombre de variables explicatives candidates ($1 \leq d \leq D$).
$H_0 : \beta_1 = \ldots = \beta_D = 0$	L'hypothèse nulle, laquelle avance l'absence de signal (excluant, bien entendu, l'interception de cette affirmation).

VII.2 Notations relatives aux modèles linéaires

Expression	Définition
$\widetilde{\boldsymbol{X}}_{u(B)} = \widetilde{\boldsymbol{X}}_u \boldsymbol{B} \quad \boldsymbol{X}_{u(B)} = \boldsymbol{X}_u \boldsymbol{B}$	Le sous-ensemble colonnes B des variables explicatives $\tilde{\boldsymbol{X}}_u$ et $\boldsymbol{X}_u$, respectivement.
$D(B) = \|B\| = B^T B$	Le nombre de paramètres du modèle décrit par le sous-ensemble colonnes B.
$\widehat{\boldsymbol{\beta}}_{u(B)} = \left(\widetilde{\boldsymbol{X}}_{u(B)}^T \widetilde{\boldsymbol{X}}_{u(B)} \right)^{-1} \widetilde{\boldsymbol{X}}_{u(B)}^T \widetilde{\boldsymbol{Y}}_u$	L'estimateur des coefficients de la régression basé sur le sous-ensemble colonnes B.
$\widehat{\boldsymbol{Y}}_u(B) = \boldsymbol{X}_u(B) \widehat{\boldsymbol{\beta}}_u(B)$	Les prévisions hors-échantillon du modèle décrit par le sous-ensemble colonnes B.
$\widehat{\widetilde{\boldsymbol{\varepsilon}}}_u(B) = \boldsymbol{Y}_u - \widehat{\boldsymbol{Y}}_u(B)$	Les résidus sur-échantillon du modèle décrit par le sous-ensemble colonnes B.
$\boldsymbol{\Sigma}_{u(B)} \propto \boldsymbol{I} + \boldsymbol{X}_{u(B)} \left(\widetilde{\boldsymbol{X}}_{u(B)}^T \widetilde{\boldsymbol{X}}_{u(B)} \right)^{-1} \boldsymbol{X}_{u(B)}^T$	La matrice de variance des résidus hors-échantillon du modèle décrit par le sous-ensemble colonnes B.
$f_{u(B)} = \left\{ \widehat{\boldsymbol{\varepsilon}}_{u(B)}^T \widehat{\boldsymbol{\Sigma}}_{u(B)}^{-1} \widehat{\boldsymbol{\varepsilon}}_{u(B)} \right\}^{-\gamma}$	La fonction d'ajustement de la population u, sous modèles linéaires.

VII.3 Notations relatives aux modèles linéaires généralisés

Expression	Définition
$\ell(\eta) \quad v(\mu)$	La fonction de lien canonique inverse et la fonction de variance, respectivement.
$\boldsymbol{\mu}_u = \ell(\boldsymbol{\eta}_u)$	La moyenne théorique hors-échantillon de la variable réponse.
$\tilde{\boldsymbol{V}}_u = v(\tilde{\boldsymbol{\mu}}_u) \quad \boldsymbol{V}_u = v(\boldsymbol{\mu}_u)$	Les matrices de variance théoriques (à un facteur près) des observations sur-échantillon et hors-échantillon, respectivement.
$\widehat{\boldsymbol{\eta}}_u(B) = \boldsymbol{X}_u(B)\,\widehat{\boldsymbol{\beta}}_u(B)$	La combinaison linéaire estimée hors-échantillon du modèle décrit par le sous-ensemble colonne B.
$\widehat{\boldsymbol{Y}}_u(B) = \widehat{\boldsymbol{\mu}}_u(B) = \ell(\widehat{\boldsymbol{\eta}}_u(B))$	Les prévisions hors-échantillon du modèle décrit par le sous-ensemble colonne B.
$\widehat{\boldsymbol{\varepsilon}}_u(B) = \boldsymbol{Y}_u - \widehat{\boldsymbol{Y}}_u(B)$	Les résidus hors-échantillon du modèle décrit par le sous-ensemble colonnes B.
$\boldsymbol{\Sigma}_{u(B)} \propto \boldsymbol{V}_u + \boldsymbol{V}_u \boldsymbol{X}_{u(B)} \left(\widetilde{\boldsymbol{X}}^T_{u(B)} \widetilde{\boldsymbol{V}}_u \widetilde{\boldsymbol{X}}_{u(B)} \right)^{-1} \boldsymbol{X}^T_{u(B)} \boldsymbol{V}_u$	La matrice de variance des résidus hors-échantillon du modèle décrit par le sous-ensemble colonnes B.
$\widehat{\boldsymbol{V}}_{u(B)} = v\left(\widehat{\boldsymbol{Y}}_{u(B)}\right) \quad \widehat{\boldsymbol{V}}_{u(B)} = v\left(\widehat{\boldsymbol{Y}}_{u(B)}\right)$	Les matrices de variance estimées sur-échantillon et hors-échantillon, respectivement.
$\widehat{\boldsymbol{\Sigma}}_{u(B)} \propto \widehat{\boldsymbol{V}}_{u(B)} + \widehat{\boldsymbol{V}}_{u(B)} \boldsymbol{X}_{u(B)} \left(\widetilde{\boldsymbol{X}}^T_{u(B)} \widehat{\widetilde{\boldsymbol{V}}}_{u(B)} \widetilde{\boldsymbol{X}}_{u(B)} \right)^{-1} \boldsymbol{X}^T_{u(B)} \widehat{\boldsymbol{V}}_{u(B)}$	La matrice de variance des résidus hors-échantillon du modèle décrit par le sous-ensemble colonnes B.
$f_{u(B)} = \left\{ \widehat{\boldsymbol{\varepsilon}}^T_{u(B)} \widehat{\boldsymbol{\Sigma}}^{-1}_{u(B)} \widehat{\boldsymbol{\varepsilon}}_{u(B)} \right\}^{-\gamma}$	La fonction d'ajustement de la population u, sous modèles linéaires généralisés.

VII.4 Notations relatives aux algorithmes génétiques

Expression	Définition
$I \quad G \quad U$	La taille des populations ($1 \leq i \leq I$), le nombre de générations ($1 \leq g \leq G$), et le nombre de populations ($1 \leq u \leq U$), respective.
$f_1, f_2, \ldots, f_U$	Les fonctions d'ajustement des U univers.
γ	Le paramètre de dentelure des fonctions d'ajustement ($\gamma > 0$).
$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \cdots & \theta_G \end{bmatrix}^T$	La séquence des taux de mutation ($0 \leq \theta_g \leq 1$).
π_0	Le taux d'activation initial des gènes ($0 \leq \pi_0 \leq 1$), lequel est parfois noté θ_0, par souci de compacité.
$B_{igu} = \begin{bmatrix} b_{1igu} & \cdots & b_{Digu} \end{bmatrix}^T$	Individu : modèle (sous-ensemble des variables explicatives) encodé au moyen d'une chaîne binaire.
$b_{digu} \sim \mathcal{B}(\pi_g)$	Gène : variable binaire encodant l'inclusion de la variable explicative d à la génération g du modèle i de la population u.
$B^*_{igu} = \begin{bmatrix} b^*_{1igu} & \cdots & b^*_{Digu} \end{bmatrix}^T$	Embryon : proto-individu représentant l'état par lequel ce dernier transite avant que ne lui soit appliqué l'opérateur de mutation.
$b^*_{digu} \sim \mathcal{B}(\pi_{g-1})$	Le gène d de l'embryon B^*_{igu}.

VII.5 Notations relatives aux proportions de rétention

Expression	Définition
$\widehat{\pi}_{dgu} = \frac{1}{I}\sum_{i=1}^{I} b_{digu}$	L'estimateur des proportions de rétention intra-populations.
$\widehat{\boldsymbol{\pi}}_{gu} = \begin{bmatrix}\widehat{\pi}_{1gu} & \cdots & \widehat{\pi}_{Dgu}\end{bmatrix}^T$	Le vecteur des estimateurs des proportions de rétention intra-populations.
$\widehat{\pi}_{dg} = \frac{1}{IU}\sum_{u=1}^{U}\sum_{i=1}^{I} b_{digu}$	L'estimateur des proportions de rétention globales.
$\widehat{\boldsymbol{\pi}}_{g} = \begin{bmatrix}\widehat{\pi}_{1g} & \cdots & \widehat{\pi}_{Dg}\end{bmatrix}^T$	Le vecteur des estimateurs des proportions de rétention globales.
$\Delta_g = \prod_{j=0}^{g}(1-2\theta_j)$	Construction arbitraire ayant pour vertu de simplifier l'écriture des premiers moments des estimateurs des proportions de rétention, nommément π_g et $\sigma^2_{\hat{\pi}_{gu}}$.
$\pi_g = \frac{1}{2}(1-\Delta_g)$	La proportion de rétention théorique sous l'hypothèse nulle. Espérance des estimateurs des proportions de rétentions intra-populations et globales.
$\sigma^2_{\hat{\pi}_g} = \frac{\Delta_g^2}{4I}\sum_{j=0}^{g}\frac{1-\Delta_j^2}{\Delta_j^2}\left(1-\frac{1}{I}\right)^{g-j}$	La variance des estimateurs des proportions de rétention intra-populations sous l'hypothèse nulle.
$\alpha_S = 1-(1-\alpha)^{1/D}$	Le niveau spécifique des tests, dont le niveau global est α, lesquels tiennent compte du problème des comparaisons multiples par l'entremise de la correction de Šidák.
$q_{1-\alpha_S} = \Phi^{-1}(1-\alpha_S)$	Le quantile $1-\alpha_S$ de la loi normale centrée réduite.
$C = \pi_G + q_{1-\alpha_S}\sigma_{\hat{\pi}_{Gu}}/\sqrt{U}$	Le seuil critique des tests.

Références

[1] M.H. Aickin and H. Gensler. Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American journal of public health*, 86 :726–8, 06 1996.

[2] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiado.

[3] Louis F.A. Arbogast. *DU CALCUL DES DÉRIVATIONS*. 1800.

[4] Kevin A. Clarke. The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, 22(4) :341–352, 2005.

[5] Charles R. Darwin. *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. J. Murray, 1859.

[6] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Addison-Wesley, 2012.

[7] David Draper and Dimitris Fouskakis. A Case Study of Stochastic Optimization in Health Policy: Problem Formulation and Preliminary Results. *Journal of Global Optimization*, 18 :399–416, 12 2000.

[8] George Edward I. The Variable Selection Problem. *Journal of the American Statistical Association*, 95(452) :1304–1308, 2000.

[9] Edward I. George and Robert McCulloch. Variable Selection Via Gibbs Sampling. *Journal of The American Statistical Association*, 88 :881–889, 09 1993.

[10] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2002.

[11] John H. Holland. *Adaptation in natural and artificial systems*: *an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, 1975.

[12] Bent Jørgensen. Exponential Dispersion Models. *Journal of the Royal Statistical Society*: *Series B (Methodological)*, 49 :127–145, 01 1987.

[13] Didier Josselin. Une piste pour la recherche de la « valeur centrale optimale ». Discussion autour de la robustesse et du comportement de la « médienne », combinaison de normes L_p. *Cybergeo*, 2004.

[14] Scott Kirkpatrick, C.D. Gelatt, and Mario P. Vecchi. Optimization by Simulated Annealing. *Science*, 220 :671–680, 01 1983.

[15] Maxime Larocque, Jean-François Plante, and Michel Adès. Bagged parallel genetic algorithms for objective model selection. *Les Cahiers du GERAD*, (G-2018-70), 2018.

[16] Peter McCullagh and John A. Nelder. *Generalized Linear Models*, volume 37. CRC press, 1989.

[17] Melanie Mitchell. *An Introduction to Genetic Algorithms*. A Bradford book. Bradford Books, 1998.

[18] A. Ralston and H.S. Wilf. *Mathematical Methods for Digital Computers*. Number vol. 1 in Mathematical Methods for Digital Computers. Wiley, 1960.

[19] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461–464, 03 1978.

[20] Zbynek Šidák. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62, 1967.

[21] B. Spinoza. *Ethics*. Classics of World Literature Series. 1677.

[22] N.N. Taleb. *Fooled by Randomness*: *The Hidden Role of Chance in Life and in the Markets*. Incerto. Random House Publishing Group, 2008.

[23] Jay M. Ver Hoef. Who Invented the Delta Method?. *The American Statistician*, 66(2) :124–127, 2012.

[24] Wikipédia. Produit infini — Wikipédia, l'encyclopédie libre, 2018.

[25] Mu Zhu and Hugh A. Chipman. Darwinian Evolution in Parallel Universes: A Parallel Genetic Algorithm for Variable Selection. *Technometrics*, 48(4) :491–502, 2006.