Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Méthode d’estimation de l’effet des erreurs de classification sur les statistiques de deux domaines

par Yanzhe Li, Sander Scholtus et Arnout van Delden

Date de diffusion : le 3 janvier 2024

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2023
Vol. 49, nº 2, pp. 617-656
Statistique Canada, nº 12-001-X au catalogue

Méthode d'estimation de l'effet des erreurs de classification sur les statistiques de deux domaines

Yanzhe Li, Sander Scholtus et Arnout van Delden[1]

Résumé

Il est essentiel de pouvoir quantifier l'exactitude (biais, variance) des résultats publiés dans les statistiques officielles. Dans ces dernières, les résultats sont presque toujours divisés en sous-populations selon une variable de classification, comme le revenu moyen par catégorie de niveau de scolarité. Ces résultats sont également appelés « statistiques de domaine ». Dans le présent article, nous nous limitons aux variables de classification binaire. En pratique, des erreurs de classification se produisent et contribuent au biais et à la variance des statistiques de domaine. Les méthodes analytiques et numériques servant actuellement à estimer cet effet présentent deux inconvénients. Le premier inconvénient est qu'elles exigent que les probabilités de classification erronée soient connues au préalable et le deuxième est que les estimations du biais et de la variance sont elles-mêmes biaisées. Dans le présent article, nous présentons une nouvelle méthode, un modèle de mélange gaussien estimé par un algorithme espérance-maximisation (EM) combiné à un bootstrap, appelé « méthode bootstrap EM ». Cette nouvelle méthode n'exige pas que les probabilités de classification erronée soient connues au préalable, bien qu'elle soit plus efficace quand on utilise un petit échantillon de vérification qui donne une valeur de départ pour les probabilités de classification erronée dans l'algorithme EM. Nous avons comparé le rendement de la nouvelle méthode et celui des méthodes numériques actuellement disponibles, à savoir la méthode bootstrap et la méthode SIMEX. Des études antérieures ont démontré que pour les paramètres non linéaires, le bootstrap donne de meilleurs résultats que les expressions analytiques. Pour presque toutes les conditions mises à l'essai, les estimations du biais et de la variance obtenues par la méthode bootstrap EM sont plus proches de leurs vraies valeurs que celles obtenues par les méthodes bootstrap et SIMEX. Nous terminons l'article par une discussion sur les résultats et d'éventuels prolongements de la méthode.

Mots-clés : Biais; variance; classification erronée; classificateur binaire; modèle de mélange gaussien; algorithme EM.

1. Introduction

L'exactitude des résultats publiés est cruciale, surtout dans les statistiques officielles (Eurostat, 2009, page 32), dans lesquelles la plupart des résultats servent à l'élaboration de politiques. Les erreurs de mesure dans la variable servant à regrouper les résultats dans des sous-populations font partie des erreurs importantes nuisant à l'exactitude des résultats. Le type d'estimateur le plus simple pour lequel l'effet des classifications erronées a été étudié est celui des proportions dans les tableaux de contingence. Bross (1954) a donné une expression pour le biais d'une proportion estimée de variable binaire en cas de classifications erronées, qui est un cas particulier de l'expression (C.1) de l'annexe C. La situation est plus compliquée pour ce qui est de l'exactitude des estimateurs de niveau (totaux, moyennes) et les ratios de ceux-ci touchés par des classifications erronées. Par exemple, on peut s'intéresser au risque moyen de pauvreté pour les travailleurs par catégorie de niveau de scolarité (Eurostat, 2022), alors qu'une partie des valeurs relatives à l'éducation sont fondées sur des classifications erronées. L'estimation des paramètres de population d'une variable continue dans chaque classe est appelée « statistiques de domaine ». Notre étude porte sur l'estimation du biais et de la variance de totaux ou de moyennes touchés par des classifications erronées. La

1. Yanzhe Li, Statistics Netherlands, PO Box 24500, 2490 HA La Haye, Pays-Bas; Sander Scholtus, Statistics Netherlands, PO Box 24500, 2490 HA La Haye, Pays-Bas. Courriel : s.scholtus@cbs.nl; Arnout van Delden, Statistics Netherlands, PO Box 24500, 2490 HA La Haye, Pays-Bas.

littérature sur la classification erronée est abondante. Citons notamment Buonaccorsi (2010), Keogh, Shaw, Gustafson, Carroll, Deffner, Dodd, Küchenhoff, Tooze, Wallace, Kipnis et Freedman (2020) et Shaw, Gustafson, Carroll, Deffner, Dodd, Keogh, Kipnis, Tooze, Wallace, Küchenhoff et Freedman (2020). Van den Hout et Van der Heijden (2002) établissent une revue de la littérature exhaustive sur le biais et la variance en présence de classifications erronées. Les références citées soulignent que la quantification de l'effet des classifications erronées sur les estimations de la population est pertinente dans un large éventail de disciplines, y compris la médecine, l'épidémiologie, l'astronomie statistique, la sociologie, la cartographie de la couverture terrestre, les études par méthodes fondées sur la réponse aléatoire et la confidentialité des données. Les deux derniers cas sont particuliers en ce sens que leurs classifications erronées sont délibérément ajoutées aux données par un mécanisme connu.

Les statistiques officielles s'appuient souvent sur des registres statistiques pour déterminer leur population cible et la diviser en sous-populations. Il peut s'agir par exemple d'un registre statistique des entreprises contenant une liste d'unités statistiques au fil du temps ainsi que des variables de contexte comme l'activité économique et la classe de taille pour les stratifier en sous-populations (Nations Unies, 2015). Un autre exemple serait un registre de la population, qui comporterait des variables de contexte comme la date de naissance, le sexe, le lieu de résidence et le plus haut niveau de scolarité atteint (Bakker, Van Rooijen et Van Toor, 2014). Ces registres sont souvent créés à partir d'une ou de plusieurs sources administratives. Des erreurs dans les variables de classification peuvent se produire en raison d'erreurs pendant l'enregistrement dans ces sources administratives, de retards administratifs, d'erreurs de couplage des sources administratives ou de changements non déclarés de la situation des unités. Les erreurs sont aussi attribuables à la différence entre les concepts utilisés par les propriétaires de registre et ceux utilisés dans les statistiques officielles (Magnusson, Palm, Branden et Mörner, 2017). Parfois, les variables de classification sont dérivées par l'application d'algorithmes d'apprentissage automatique (voir, par exemple Meertens, Diks, Van den Herik et Takes (2020)) et les erreurs faites par ces algorithmes entraînent des classifications erronées.

Dans les situations pratiques où des données de sortie sont produites aux fins de statistiques officielles, on tente de réduire le nombre de classifications erronées, par exemple par une vérification automatique ou manuelle. Dans le cas de statistiques sur les entreprises, on sait que les codes d'activité économique sont sujets à des erreurs. Dans ce cas, on applique une vérification manuelle, qui est habituellement limitée aux unités les plus influentes, à savoir les plus grandes entreprises, tandis que les autres unités ne sont pas corrigées. Souvent, on espère que les erreurs se produisant dans les petites unités n'aient pas un effet important sur les chiffres publiés. Malheureusement, cela n'est pas nécessairement vrai. Van Delden, Scholtus et Burger (2016) ont démontré que dans une étude de cas en particulier, les niveaux de classification erronée observés dans les petites et moyennes entreprises entraînaient un biais considérable du chiffre d'affaires total pour certaines cellules de publication. Ce biais s'explique par le fait que les entrées de chiffre d'affaires des unités incorrectement classifiées dans la classe cible ne sont pas compensées par les sorties de chiffre d'affaires des unités qui ont été à tort non classifiées dans la classe cible. Dans le

contexte des tableaux de contingence, Schwartz (1985) a souligné l'importance des classifications erronées dans l'exactitude des sorties en qualifiant cette situation de « problème négligé ».

Les expressions analytiques du biais et de la variance de moyennes ou de totaux des sous-populations touchés par des classifications erronées ont été publiées par Selén (1986) et Van Delden et coll. (2016). En outre, Kooiman, Willenborg et Gouweleeuw (1997) ont traité du biais et de la variance des totaux de sous-populations dans le contexte de la confidentialité des données. Ces expressions analytiques présentent deux inconvénients. Le premier est qu'elles reposent sur l'hypothèse selon laquelle les probabilités de tous les types de classification erronée sont connues ou qu'elles ont été estimées avec exactitude au moyen d'un échantillon. Dans le contexte de la confidentialité des données, ces probabilités sont connues, étant donné qu'on applique intentionnellement des classifications erronées pour éviter toute divulgation (Kooiman et coll., 1997). Toutefois, dans la plupart des applications, les probabilités de classification erronée sont inconnues. Les probabilités d'erreur de classification estimées sont souvent obtenues à partir de connaissances antérieures ou d'un ensemble comparable de données (Edwards, Bakoyannis, Yiannoutsos, Mburu et Cole, 2019; Edwards, Cole et Fox, 2020), ou encore à partir d'un échantillon de vérification (Gravel et Platt, 2018). L'estimation exacte de ces probabilités au moyen d'échantillons suffisamment grands exige une quantité considérable de travail manuel. Le deuxième inconvénient est que les estimations de biais et de variance fondées sur ces expressions analytiques sont elles-mêmes biaisées. Kooiman et coll. l'ont constaté (1997), et cet inconvénient est décrit plus en détail par Van Delden et coll. (2016). La principale raison est que les expressions du biais et de la variance dépendent des vrais totaux de la population qui sont inconnus. Lors de l'estimation du biais et de la variance, ces vrais totaux de la population sont remplacés par leurs estimateurs biaisés des totaux de la population, ce qui mène à des estimations biaisées du biais et (dans une moindre mesure) de la variance des totaux. Dans le cas particulier où l'on souhaite estimer seulement une proportion sans variable numérique, différentes méthodes de correction du biais peuvent être trouvées dans Kloos, Meertens, Scholtus et Karch (2021).

Des méthodes numériques ont été élaborées comme solution de rechange à l'utilisation d'expressions analytiques. La méthode bootstrap peut servir à estimer l'exactitude des données de sortie en cas de classifications erronées. Zhang (2011) a utilisé la méthode bootstrap pour mesurer la variance des totaux observés de sous-populations causée par la classification erronée des ménages. Van Delden et coll. (2016) ont proposé une méthode bootstrap pour estimer le biais et la variance de moyennes ou de totaux de sous-populations. La méthode bootstrap est très souple et adaptable à de nombreux estimateurs. Malheureusement, elle présente les mêmes inconvénients que l'approche analytique, c'est-à-dire qu'elle exige que les probabilités de classification erronée soient connues avec exactitude et, qu'en général, elle donne des estimations biaisées du biais et de la variance. De fait, pour les paramètres de domaine simples comme les totaux et les proportions, la méthode bootstrap et les expressions analytiques mènent à des résultats presque identiques; voir par exemple Van Delden et coll. (2016). Pour les paramètres non linéaires de domaine comme les ratios, les résultats bootstrap sont généralement plus exacts dans le cas de distributions asymétriques en raison de la possibilité de violation de certaines hypothèses sous-jacentes de l'approximation

utilisée dans le biais analytique et les expressions de variance; voir Van Delden, Scholtus, Burger et Meertens (2023).

Une autre méthode numérique a été proposée pour les classifications erronées : la méthode SIMEX (« simulation et extrapolation »), introduite pour la première fois par Cook et Stefanski (1994). Elle a été élaborée pour estimer l'effet des erreurs de mesure et a été adaptée dans le champ de la classification erronée par Küchenhoff, Mwalili et Lesaffre (2006) et Hopkins et King (2010). Comme pour la méthode bootstrap, la méthode SIMEX commence par appliquer des classifications erronées aux données observées, après quoi les estimateurs (totaux, moyennes) sont recalculés. De plus, la méthode SIMEX applique un processus de simulation qui introduit de multiples ensembles d'erreurs supplémentaires pour estimer l'effet des classifications erronées. Enfin, on extrapole les estimations à la condition sans erreur. Nous avons appliqué ici les idées qui sous-tendent la procédure SIMEX pour vérifier si elle peut servir à éliminer le biais dans les estimations du biais et de la variance de moyennes et de totaux en présence de classifications erronées.

Dans la présente étude, nous présentons une nouvelle méthode qui s'appuie sur un modèle de mélange pour estimer le biais et la variance de moyennes et de totaux en présence de classifications erronées. Un algorithme espérance-maximisation (EM) est utilisé pour estimer le modèle de mélange. Étant donné que les vraies classes d'unités sont inconnues en cas de classifications erronées, ces vraies classes peuvent être considérées comme une variable non observée (« latente ») dans le modèle de mélange. Nous modélisons la variable cible numérique dans chaque classe comme un mélange de différentes lois normales (un mélange gaussien), qui peut également tenir compte des variables cibles qui n'ont pas une loi normale; voir McLachlan et Peel (2000). La méthode est appelée « méthode bootstrap EM ». Dans la présente étude, nous nous limitons à une variable de classification binaire. Notre méthode peut être étendue à une situation comportant des classes multiples, qui sera traitée dans la suite de l'exposé.

La nouvelle méthode donne de meilleurs résultats que la méthode bootstrap et la méthode SIMEX eu égard aux deux inconvénients mentionnés. Premièrement, il n'est pas nécessaire d'estimer les probabilités de classification erronée au préalable, bien que dans les cas plus complexes, il soit efficace pour ce qui est du calcul d'avoir une estimation approximative des taux de classification erronée, qui peut servir de point de départ informé pour l'algorithme EM. Deuxièmement, nous montrerons que – du moins dans le cas binaire – la nouvelle méthode donne des estimations plus précises du biais et de la variance des moyennes et des totaux touchés par les classifications erronées que la méthode bootstrap et la méthode SIMEX.

Dans le présent article, nous évaluons l'exactitude des estimations de biais et de variance en cas de classification erronée, en comparant la méthode bootstrap EM proposée avec la méthode bootstrap et la méthode SIMEX adaptée (appelée « méthode bootstrap SIMEX ») dans une étude par simulations et dans une étude de cas. Dans les deux études, les proportions de deux classes sont variées, tout comme les taux de classification erronée. Dans l'étude par simulations, la distribution des données dans chaque classe est un mélange gaussien. Dans l'étude de cas, nous utilisons des distributions empiriques provenant d'un ensemble de données avec le logarithme du chiffre d'affaires par entreprise aux Pays-Bas. Pour les données empiriques, nous utilisons un modèle de mélange gaussien dans lequel le nombre optimal de composantes doit

être estimé. Dans les deux études, nous comparons les vraies valeurs de biais et de variance avec leurs estimations.

L'article est structuré comme suit. La section 2 présente en détail les trois méthodes. L'évaluation des méthodes est présentée à la section 3 pour l'étude par simulations et à la section 4, pour l'étude de cas. La section 5 expose les résultats de notre étude et propose des pistes de recherche futures. En outre, les annexes A, B et C fournissent des éléments complémentaires. L'annexe A présente trois fonctions d'extrapolation différentes dans la méthode SIMEX. L'annexe B porte sur la différence entre l'utilisation de BIC et de sBIC comme critère de sélection du nombre optimal de composantes dans l'étude de cas. L'annexe C comprend certaines propriétés théoriques de la méthode bootstrap et de la méthode bootstrap EM. Le code R qui met en œuvre les méthodes utilisées dans la présente étude se trouve à l'adresse https://github.com/Yanzhee/EM-bootstrapping.

2. Méthodologie

2.1 Paramètres généraux

Nous considérons une situation où une population de N unités est classée dans deux classes. La variable indicatrice pour la vraie appartenance à l'une de ces classes est notée $z \in \{0,1\}$. Par souci de commodité, nous désignerons dorénavant les classes d'intérêt par classe 1 et classe 0, et nous utiliserons l'indicateur z de façon interchangeable avec la classification elle-même. Les valeurs de z sont censées ne contenir aucune erreur de classification et sont considérées comme fixes pour la population d'intérêt. En pratique, on ne connaît pas les vraies classes de toutes les unités. Nous pouvons seulement les établir par inférence au moyen, par exemple, d'une vérification manuelle par des spécialistes, d'une classification automatique à partir de multiples algorithmes d'apprentissage automatique, etc. La vraie proportion de classe 1 dans la population est désignée par α_1; la vraie proportion de la classe 0 est $1-\alpha_1$.

La variable de classification observée est désignée par $\hat{z} \in \{0,1\}$. En pratique, les classes observées d'unités contiennent des erreurs de classification. La source des erreurs de classification peut être une mauvaise compréhension de la définition de la classe, une mauvaise communication ou simplement des fautes de frappe.

Dans la pratique, on s'intéresse souvent à l'estimation des paramètres de population d'une variable continue y dans chaque classe, appelée « statistiques de domaine ». Par exemple, quand y représente le chiffre d'affaires d'entreprises dans plusieurs industries, le chiffre d'affaires total de chaque industrie sera un indicateur intéressant. Nous utilisons ζ pour désigner un paramètre de population vrai, basé sur y et la vraie variable de classification z. Dans ce qui suit, nous supposons que la variable continue y est sans erreur.

Les exemples de paramètres de domaine communs ζ compris dans notre étude sont la somme totale de y pour la classe 1 (T_1), la proportion de la classe 1 (α_1), et l'écart-type de y pour la classe 1 (σ_1). Le tableau 2.1 énumère les formules pour ζ étant donné les variables y et z. Étant donné que notre étude est

dans une configuration de classificateur binaire, l'exactitude des statistiques de domaine dans la classe 0 est directement liée à l'exactitude des estimations correspondantes dans la classe 1.

Tableau 2.1
Liste de formules pour des exemples de paramètres de domaine ζ.

Statistiques d'intérêt	Notation	Formule étant donné y et z
Somme totale de y pour la classe 1	T_1	$\sum_i z_i y_i$
Proportion de la classe 1	α_1	$\sum_i z_i / N$
Écart-type de y pour la classe 1	σ_1	$\sqrt{\frac{1}{\sum_i z_i} \sum_i z_i (y_i - \mu_1)^2}$

Notes : 1. i désigne une unité de la population.
2. $\mu_1 = \sum_i z_i y_i \big/ \sum_i z_i = T_1/(N\alpha_1)$ est la moyenne de y pour la classe 1.
3. En remplaçant z_i par $\hat{z}_i$ dans les formules, on peut calculer $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$.

Le paramètre ζ requiert les vraies valeurs de z et ne peut donc pas être calculé en pratique. Nous obtenons un estimateur évident en remplaçant la valeur vraie inconnue z par la valeur observée $\hat{z}$ dans les expressions pour ζ; ceci donne la statistique de domaine $\hat{\zeta}$. Par exemple, en remplaçant z_i par $\hat{z}_i$ dans le tableau 2.1, nous obtenons les formules pour les statistiques de domaine $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$. Notons que l'accent circonflexe dans $\hat{\zeta}$ signifie un estimateur, alors que dans $\hat{z}_i$, il n'a aucune autre signification que de permettre de distinguer l'indicateur observé du vrai indicateur z_i.

Notre étude s'appuie sur le biais et la variance pour mesurer les effets que les erreurs de classification ont sur $\hat{\zeta}$. Le biais est défini comme la différence entre les valeurs attendues du résultat estimé et la vraie valeur des paramètres de domaine. La variance mesure l'ampleur attendue du changement des statistiques de domaine estimées si différentes variables de classification ayant la même distribution d'erreurs sont utilisées. Mathématiquement, nous définissons :

$$\begin{aligned} \textbf{Biais} \quad &= \mathrm{E}(\hat{\zeta} - \zeta), \\ \textbf{Variance} \quad &= \mathrm{Var}(\hat{\zeta}) = \mathrm{E}\left(\left(\hat{\zeta} - \mathrm{E}(\hat{\zeta})\right)^2\right). \end{aligned} \tag{2.1}$$

Par souci de simplicité, nous supposons qu'à part les erreurs de classification, aucune autre erreur ne se produit. Conformément à un type d'application courant dans les statistiques officielles, nous nous intéressons ici au biais et à la variance attribuables aux erreurs de classification pour une population finie fixe. Autrement dit, dans (2.1), nous imposons implicitement des conditions aux valeurs réalisées de $z_1, \ldots, z_N$ et $y_1, \ldots, y_N$.

2.2 Modèle de mélange

2.2.1 Paramétrage du modèle

La nouvelle méthode que nous proposons pour estimer le biais et la variance de $\hat{\zeta}$ nécessite un modèle pour les valeurs observées $y_1, \ldots, y_N$ et $\hat{z}_1, \ldots, \hat{z}_N$. Dans le cas présent, nous proposons d'utiliser un modèle

de mélange, où les distributions dans la classe $z=1$ et la classe $z=0$ peuvent être différentes. Par souci de simplicité, nous supposons que les valeurs des unités $i=1,\ldots,N$ sont établies indépendamment les unes des autres. De plus, nous supposons que les erreurs de classification dans $\hat{z}_i$ se produisent avec les mêmes probabilités pour toutes les unités, ce qui signifie également qu'elles sont indépendantes de la variable continue y_i. Cette dernière hypothèse nous permet de modéliser les valeurs observées y_i et $\hat{z}_i$ séparément, car elle sous-entend que dans chaque classe vraie, la densité conjointe de y_i et $\hat{z}_i$ est factorisée comme suit :

$$f(y_i, \hat{z}_i = b \mid z_i = a) = f(y_i \mid z_i = a) \cdot P(\hat{z}_i = b \mid z_i = a) \tag{2.2}$$

pour tous les $i=1,\ldots,N,$ avec $a,b \in \{0,1\}$. Ici, $P(\hat{z}=b \mid z=a)$ désigne une probabilité d'erreur de classification et $f(y \mid z=a)$ désigne la densité de la variable continue dans la classe a. Décrivons maintenant plus en détail ces deux parties du modèle.

Matrice de probabilités. Les probabilités d'erreurs de classification sont modélisées par une matrice de transition 2×2 $\mathbf{P}$ (formule 2.3). Celle-ci décrit la relation entre les vraies classes z (lignes) et les classes observées $\hat{z}$ (colonnes).

$$\mathbf{P} = \begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{00} & p_{00} \end{pmatrix}. \tag{2.3}$$

La valeur p_{ab} indique la probabilité d'observer une unité dans la classe b quand sa vraie classe est a. Ainsi, pour une unité i, si sa vraie classe est 1, la probabilité de l'observer dans la classe 1 est $P(\hat{z}_i=1 \mid z_i=1) = p_{11},$ et la probabilité de l'observer dans la classe 0 est $P(\hat{z}_i=0 \mid z_i=1) = 1-p_{11};$ de même, pour une unité dans la vraie classe 0, $P(\hat{z}_i=0 \mid z_i=0) = p_{00}$ et $P(\hat{z}_i=1 \mid z_i=0) = 1-p_{00}.$ Pour les classificateurs raisonnables, les valeurs de p_{11} et p_{00} doivent être supérieures à 0,5.

Modèle de mélange gaussien. Nous supposons que la distribution de y pour chaque classe z suit un modèle de mélange gaussien. Le nombre de composantes gaussiennes dans la classe 1 est $q_1,$ et leur nombre dans la classe 0 est q_0. Notons que cela signifie que le modèle global pour y peut être considéré comme un « mélange de mélanges », le premier niveau de mélange étant donné par la vraie classe z et le deuxième niveau de mélange par les composantes du mélange gaussien dans une vraie classe.

Pour mieux l'expliquer, disons qu'une variable de composante m est définie pour déterminer à quelle composante du modèle de mélange gaussien chaque unité appartient. La variable m n'est pas observée. La distribution de y_i dépend de la classe à laquelle il appartient (la valeur de z_i) et également de la composante de cette classe à laquelle il appartient (la valeur de m_i); nous posons l'hypothèse classique que chaque unité i appartient à une paire classe-composante unique $(z_i, m_i).$ Selon la loi de la probabilité totale, la densité de y_i conditionnelle à z_i est

$$f\left(y_i \mid z_i = a\right) = \begin{cases} \displaystyle\sum_{j=1}^{q_1} P\left(m_i = j \mid z_i = 1\right) \cdot f\left(y_i \mid z_i = 1, m_i = j\right), & \text{si } a = 1, \\ \displaystyle\sum_{k=1}^{q_0} P\left(m_i = k \mid z_i = 0\right) \cdot f\left(y_i \mid z_i = 0, m_i = k\right), & \text{si } a = 0, \end{cases}$$

où $P(m=j \mid z=1)$ est le poids du mélange d'une composante de la classe 1, désigné par ξ_{1j} ($j \in \{1,\ldots,q_1\}$); $P(m=k \mid z=0)$ est le poids du mélange d'une composante de la classe 0, désigné par ξ_{0k} ($k \in \{1,\ldots,q_0\}$). Ces poids de mélange satisfont $\sum_{j=1}^{q_1} \xi_{1j} = 1$ et $\sum_{k=1}^{q_0} \xi_{0k} = 1$.

Dans un mélange gaussien, nous supposons que chaque composante suit une loi normale. Ainsi, nous obtenons dans ce cas :

$$f(y_i \mid z_i = a) = \begin{cases} \sum_{j=1}^{q_1} \xi_{1j} \cdot \varphi\left(y_i;\ \mu_{1j}, \sigma_{1j}^2\right), & \text{si } a = 1, \\ \sum_{k=1}^{q_0} \xi_{0k} \cdot \varphi\left(y_i;\ \mu_{0k}, \sigma_{0k}^2\right), & \text{si } a = 0, \end{cases} \tag{2.4}$$

où μ_{1j} est la moyenne de la composante j dans la classe 1 et σ_{1j} est son écart-type; μ_{0k} est la moyenne de la composante k dans la classe 0 et σ_{0k} est son écart-type; $\varphi(y;\mu,\sigma^2) = \sigma^{-1}(2\pi)^{-1/2}\exp\{-(y-\mu)^2/(2\sigma^2)\}$ désigne la densité d'une loi normale avec des paramètres μ et σ^2.

Soit $\boldsymbol{\theta} = (\alpha_1, p_{11}, p_{00}, \xi_{11}, \mu_{11}, \sigma_{11}, \ldots, \xi_{1q_1}, \mu_{1q_1}, \sigma_{1q_1}, \xi_{01}, \mu_{01}, \sigma_{01}, \ldots, \xi_{0q_0}, \mu_{0q_0}, \sigma_{0q_0})'$ le vecteur de paramètres inconnus du modèle de mélange gaussien. L'identification de tous les paramètres dans $\boldsymbol{\theta}$ exige que l'ordre des composantes dans chaque classe soit fixe. Par souci de simplicité, nous supposons ici que $\mu_{11} < \ldots < \mu_{1j} < \ldots < \mu_{1q_1}$ et $\mu_{01} < \ldots < \mu_{0k} < \ldots < \mu_{0q_0}$.

En pratique, le nombre approprié de composantes dans les modèles de mélange gaussien pour les classes 1 et 0, q_1 et q_0, n'est pas connu et doit être déterminé à partir des données observées. Dans le cas présent, le principal objectif du modèle de mélange est de fournir une façon souple de modéliser la distribution de y dans chaque classe. Pour ce type d'application, une approche courante consiste à adapter plusieurs modèles de mélange aux données avec différents nombres de composantes et à utiliser le critère d'information bayésien (BIC pour *bayesian information criterion*) afin de sélectionner le nombre optimal de composantes (McLachlan et Peel, 2000, pages 175 et 209-210). Cependant, des études plus récentes laissent entendre que quand les composantes d'un modèle de mélange ont une moyenne et une variance très semblables, la matrice d'information de Fisher peut devenir singulière et le BIC n'est plus justifié comme critère (Drton et Plummer, 2017). Pour ces situations, Drton et Plummer (2017) ont proposé un critère BIC modifié, appelé sBIC. Dans l'étude de cas que nous traitons dans la section 4, nous comparons le nombre optimal de composantes selon les critères BIC et sBIC.

2.2.2 Algorithme espérance-maximisation

L'estimation par le maximum de vraisemblance du modèle de mélange gaussien ci-dessus peut être obtenue au moyen d'un algorithme EM (Dempster, Laird et Rubin, 1977; Little et Rubin, 2002). Ce type d'algorithme est souvent appliqué en présence de variables non observées dans les modèles statistiques. Comme son nom l'illustre, il comporte deux étapes : une étape E et une étape M. L'étape E construit la valeur espérée du logarithme de la fonction de vraisemblance des données complètes, conditionnellement aux données observées. L'étape M estime ensuite les paramètres du modèle, qui, à leur tour, fournissent des données pour l'étape E suivante. L'algorithme procède de manière itérative jusqu'à la convergence.

Notons que l'algorithme EM peut estimer tous les paramètres du modèle de mélange (y compris la matrice $\mathbf{P}$) à partir d'un ensemble de données d'observations $(\hat{z}_i, y_i)$; il n'est donc pas nécessaire d'avoir observé la vraie classe z_i d'une unité. En gros, cela est possible parce que, d'une part, la probabilité qu'une observation $(\hat{z}_i, y_i)$ appartienne à la classe 1 ou à la classe 0 peut être prédite à partir des différences dans la distribution de y pour la classe 1 et la classe 0 (ce qui est fait pendant l'étape E) et, d'autre part, les distributions de y et $\hat{z}$ dans chaque vraie classe peuvent être estimées à partir de ces probabilités prédites (ce qui est fait pendant l'étape M). Techniquement, l'estimation est possible parce qu'il existe (dans des circonstances normales) un ensemble unique de valeurs de paramètres $\boldsymbol{\theta}$ pour lequel le logarithme de la fonction de vraisemblance des données complètes atteint son maximum global. Autrement dit, le modèle est *identifié*; voir McLachlan et Peel (2000) pour en savoir plus.

Pour calculer le logarithme de la fonction de vraisemblance des données complètes, nous notons à partir des expressions (2.2), (2.3) et (2.4) que

$$f(z_i, m_i, \hat{z}_i, y_i; \boldsymbol{\theta}) = \prod_{j=1}^{q_1} \omega_{1ji}^{z_i \mathbf{1}_{(m_i=j)}} \prod_{k=1}^{q_0} \omega_{0ki}^{(1-z_i)\mathbf{1}_{(m_i=k)}},$$

où

$$\omega_{1ji} \triangleq \alpha_1 p_{11}^{\hat{z}_i} (1-p_{11})^{1-\hat{z}_i} \frac{\xi_{1j}}{\sigma_{1j}\sqrt{2\pi}} \exp\left\{-\frac{(y_i-\mu_{1j})^2}{2\sigma_{1j}^2}\right\},$$

$$\omega_{0ki} \triangleq (1-\alpha_1)(1-p_{00})^{\hat{z}_i} p_{00}^{1-\hat{z}_i} \frac{\xi_{0k}}{\sigma_{0k}\sqrt{2\pi}} \exp\left\{-\frac{(y_i-\mu_{0k})^2}{2\sigma_{0k}^2}\right\},$$

et les fonctions indicatrices pour m_i sont définies comme suit :

$$\mathbf{1}_{(m_i=j)} = \begin{cases} 1 & m_i = j \\ 0 & m_i \neq j \end{cases}; \qquad \mathbf{1}_{(m_i=k)} = \begin{cases} 1 & m_i = k, \\ 0 & m_i \neq k \end{cases}.$$

Notons que $z_i \sum_{j=1}^{q_1} \mathbf{1}_{(m_i=j)} = z_i$ et $(1-z_i)\sum_{k=1}^{q_0} \mathbf{1}_{(m_i=k)} = 1-z_i$ pour toutes les unités i.

Il s'ensuit que le logarithme de la fonction de vraisemblance (LL pour log-likelihood en anglais) pour les données complètes du modèle de mélange gaussien peut s'écrire ainsi :

$$\mathrm{LL}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f(z_i, m_i, \hat{z}_i, y_i; \boldsymbol{\theta}) = \sum_{i=1}^{N} \left\{ z_i \sum_{j=1}^{q_1} \mathbf{1}_{(m_i=j)} \log \omega_{1ji} + (1-z_i) \sum_{k=1}^{q_0} \mathbf{1}_{(m_i=k)} \log \omega_{0ki} \right\}, \qquad (2.5)$$

où nous avons utilisé l'hypothèse selon laquelle les unités sont tirées indépendamment les unes des autres.

Étape E. Dans l'étape E de l'algorithme, les quantités non observées $z_i \mathbf{1}_{(m_i=j)}$ et $(1-z_i)\mathbf{1}_{(m_i=k)}$ dans (2.5) sont remplacées par leurs espérances conditionnelles, compte tenu des valeurs observées $\hat{z}_i$ et y_i :

$$\mathrm{E}\left(z_i \mathbf{1}_{(m_i=j)} \mid \hat{z} = \hat{z}_i, y = y_i\right) = P\left(z_i = 1, m_i = j \mid \hat{z} = \hat{z}_i, y = y_i\right) = \frac{\omega_{1ji}}{\sum_{j=1}^{q_1} \omega_{1ji} + \sum_{k=1}^{q_0} \omega_{0ki}} \triangleq A_{1ji};$$

$$\mathrm{E}\left((1-z_i)\,\mathbf{1}_{(m_i=k)} \mid \hat{z}=\hat{z}_i,\, y=y_i\right) = P\left(z_i=0, m_i=k \mid \hat{z}=\hat{z}_i,\, y=y_i\right) = \frac{\omega_{0ki}}{\sum_{j=1}^{q_1}\omega_{1ji}+\sum_{k=1}^{q_0}\omega_{0ki}} \triangleq A_{0ki}.$$

Pendant l'itération t de l'algorithme, ces expressions sont évaluées au moyen des estimations des paramètres actuels $\hat{\boldsymbol{\theta}}^{(t-1)}$, ce qui produit les valeurs $A_{1ji}^{(t)}$ et $A_{0ki}^{(t)}$.

Étape M. Dans l'étape M de l'algorithme, le logarithme de la fonction de vraisemblance (2.5) est maximisée par rapport aux paramètres du modèle, les quantités non observées étant remplacées par $A_{1ji}^{(t)}$ et $A_{0ki}^{(t)}$ à partir de l'étape E la plus récente. En paramétrant à zéro les dérivées partielles de premier ordre de cette fonction de vraisemblance espérée, on obtient les formules suivantes pour mettre à jour les paramètres du modèle :

$$\hat{\alpha}_1^{(t+1)} = \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_1}A_{1ji}^{(t)}\right)}{N};$$

$$\hat{p}_{11}^{(t+1)} = \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_1}A_{1ji}^{(t)}\right)\hat{z}_i}{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_1}A_{1ji}^{(t)}\right)}; \qquad \hat{p}_{00}^{(t+1)} = \frac{\sum_{i=1}^{N}\left(\sum_{k=1}^{q_0}A_{0ki}^{(t)}\right)(1-\hat{z}_i)}{\sum_{i=1}^{N}\left(\sum_{k=1}^{q_0}A_{0ki}^{(t)}\right)};$$

$$\hat{\xi}_{1j}^{(t+1)} = \frac{\sum_{i=1}^{N}A_{1ji}^{(t)}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{q_1}A_{1ji}^{(t)}\right)}; \qquad \hat{\xi}_{0k}^{(t+1)} = \frac{\sum_{i=1}^{N}A_{0ki}^{(t)}}{\sum_{i=1}^{N}\left(\sum_{k=1}^{q_0}A_{0ki}^{(t)}\right)};$$

$$\hat{\mu}_{1j}^{(t+1)} = \frac{\sum_{i=1}^{N}A_{1ji}^{(t)}y_i}{\sum_{i=1}^{N}A_{1ji}^{(t)}}; \qquad \hat{\mu}_{0k}^{(t+1)} = \frac{\sum_{i=1}^{N}A_{0ki}^{(t)}y_i}{\sum_{i=1}^{N}A_{0ki}^{(t)}};$$

$$\hat{\sigma}_{1j}^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{N}A_{1ji}^{(t)}\left(y_i-\hat{\mu}_{1j}^{(t+1)}\right)^2}{\sum_{i=1}^{N}A_{1ji}^{(t)}}}; \qquad \hat{\sigma}_{0k}^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{N}A_{0ki}^{(t)}\left(y_i-\hat{\mu}_{0k}^{(t+1)}\right)^2}{\sum_{i=1}^{N}A_{0ki}^{(t)}}}.$$

Comme nous l'avons indiqué précédemment, l'algorithme EM peut servir à estimer le modèle de mélange même quand la vraie classe z_i n'est jamais observée dans les données disponibles. Toutefois, en général, l'algorithme EM convergera simplement vers un maximum local du logarithme de la fonction de vraisemblance. Pour que le maximum global soit trouvé, il faut parfois exécuter l'algorithme plusieurs fois en utilisant des valeurs de départ différentes (choisies au hasard) et retenir la meilleure solution. Pour des propositions générales concernant la façon de choisir des valeurs de départ aléatoires appropriées pour les modèles de mélange, voir McLachlan et Peel (2000, section 2.12).

Par ailleurs, des observations $(z_i, \hat{z}_i, y_i)$ contenant la vraie classe peuvent avoir été obtenues pour un sous-échantillon aléatoire des données (un *échantillon de vérification*). Si un échantillon de vérification est disponible, il peut servir à améliorer la convergence de l'algorithme EM vers le maximum global de la

fonction de vraisemblance, ce qui réduit la nécessité de l'exécuter de nouveau avec des valeurs de départ différentes. Premièrement, les paramètres α_1, p_{11} et p_{00} peuvent être estimés directement à partir de l'échantillon de vérification pour fournir des valeurs de départ raisonnables à l'algorithme EM. De plus, on pourrait obtenir de meilleures valeurs de départ pour les autres paramètres (liés aux composantes du mélange gaussien à l'intérieur de chaque classe) en appliquant un algorithme de mise en grappe de k moyennes à chaque classe vraie de l'échantillon de vérification (Li, 2020b). Enfin, pour examiner les observations de l'échantillon de vérification, A_{1ji} et A_{0ki} peuvent être remplacés pendant l'étape E par des valeurs espérées définies plus étroitement :

$$\mathrm{E}\left(z_i \mathbf{1}_{(m_i=j)} \mid z_i=1, \hat{z}=\hat{z}_i, y=y_i\right)=P\left(m_i=j \mid z_i=1, \hat{z}=\hat{z}_i, y=y_i\right)=\frac{\omega_{1ji}}{\sum_{j=1}^{q_1}\omega_{1ji}} \triangleq Q_{1ji};$$

$$\mathrm{E}\left((1-z_i)\mathbf{1}_{(m_i=k)} \mid z_i=0, \hat{z}=\hat{z}_i, y=y_i\right)=P\left(m_i=k \mid z_i=0, \hat{z}=\hat{z}_i, y=y_i\right)=\frac{\omega_{0ki}}{\sum_{k=1}^{q_0}\omega_{0ki}} \triangleq Q_{0ki};$$

tandis que $\mathrm{E}\left(z_i \mathbf{1}_{(m_i=j)} \mid z_i=0, \hat{z}=\hat{z}_i, y=y_i\right)=\mathrm{E}\left((1-z_i)\mathbf{1}_{(m_i=k)} \mid z_i=1, \hat{z}=\hat{z}_i, y=y_i\right)=0.$

2.3 Méthodes

Dans la présente étude, nous comparerons trois méthodes par lesquelles nous tentons d'estimer le biais et la variance d'une statistique de domaine $\hat{\zeta}$ selon la définition de (2.1) : la méthode bootstrap, la méthode bootstrap EM et la méthode bootstrap SIMEX. Nous n'avons pas comparé les résultats avec les expressions analytiques, puisque nous savons déjà que pour les statistiques de domaine, ces expressions donnent des résultats qui sont soit semblables à ceux de la méthode bootstrap, soit moins précis (voir l'introduction).

Comme nous l'avons indiqué à la fin de la section 2.1, nous nous intéressons ici au biais et à la variance dus aux erreurs de classification pour une population finie, conditionnellement aux valeurs $z_1, \ldots, z_N$ et $y_1, \ldots, y_N$. En raison de l'hypothèse (2.2), cela signifie que le biais et la variance d'intérêt sont entièrement déterminés par le processus aléatoire décrit par la matrice $\mathbf{P}$, appliqué aux valeurs fixes $z_1, \ldots, z_N$.

Une considération pratique importante est que les méthodes bootstrap et bootstrap SIMEX supposent que (une estimation de) la matrice $\mathbf{P}$ est disponible au préalable, tandis qu'une estimation de $\mathbf{P}$ est obtenue dans le cadre de la méthode bootstrap EM. Pour les autres méthodes, $\mathbf{P}$ pourrait être estimé en pratique à partir d'un échantillon de vérification ou par l'exécution séparée de l'algorithme EM. Dans notre étude abordée dans les sections 3 et 4, nous avons utilisé l'estimation de $\mathbf{P}$ tirée de l'algorithme EM comme donnée d'entrée pour les méthodes bootstrap et bootstrap SIMEX.

2.3.1 Méthode bootstrap

La méthode bootstrap appliquée ici provient de Van Delden et coll. (2016). Elle est résumée dans l'algorithme 1. On applique la matrice de classification $\mathbf{P}$ à la classe observée $\hat{z}$ et on obtient les classes

découlant du bootstrap z^*. La probabilité que la classe bootstrap soit 1 étant donné la classe observée 1 est $P(z_i^* = 1 \mid \hat{z}_i = 1) = p_{11}$ et elle est $P(z_i^* = 1 \mid \hat{z}_i = 0) = 1 - p_{00}$ quand la classe observée est 0. Au moyen de la méthode bootstrap, des erreurs de classification aléatoires sont introduites dans les classes observées. Par conséquent, le biais estimé est calculé en comparant les statistiques bootstrap ζ^* à la statistique observée $\hat{\zeta}$, et la variance des statistiques bootstrap ζ^* est une estimation de la variance de la statistique de domaine observée $\hat{\zeta}$. En pratique, la variance et le biais bootstrap théoriques sont habituellement calculés approximativement au moyen d'un nombre fini (S) d'échantillons bootstrap. Pour certaines statistiques simples comme α_1 et T_1, il est également possible de calculer une formule exacte pour la variance et le biais bootstrap théoriques (Van Delden et coll., 2016).

Algorithme 1 La méthode bootstrap

Entrée : Observations $(y_i, \hat{z}_i)$ pour $i = 1, \ldots, N$, la matrice $\mathbf{P}$ et S.

1 : **for** $s = 1 \ldots S$ **do**

2 : Générer z_i^* par $\mathbf{P}$, conditionnellement à $\hat{z}_i$, pour chaque unité i dans l'ensemble de données.

3 : Calculer la valeur correspondante ζ^* basée sur (y_i, z_i^*) au lieu de $(y_i, \hat{z}_i)$.

4 : **end for**

5 : Calculer $\mathbf{Biais}_{\mathbf{boot}} = \mathrm{E}(\zeta^* \mid \hat{z}) - \hat{\zeta}$, $\mathbf{Var}_{\mathbf{boot}} = \mathrm{Var}(\zeta^* \mid \hat{z})$ à partir de S simulations

Sortie : $\mathbf{Biais}_{\mathbf{boot}}$ et $\mathbf{Var}_{\mathbf{boot}}$

On sait qu'en général, les estimations du biais et de la variance de l'algorithme 1 sont biaisées, parce que les classes observées $\hat{z}_i$ sont utilisées comme point de départ du bootstrap. Ce problème est illustré dans l'annexe C.1 au moyen du paramètre $\zeta = T_1$, pour lequel une analyse exacte est possible. Les deux méthodes suivantes tentent de corriger ce biais.

2.3.2 Méthode bootstrap espérance-maximisation

Dans la méthode bootstrap EM, nous supposons que les données observées $(y, \hat{z})$ suivent le modèle de mélange gaussien de la section 2.2. L'algorithme EM permet d'estimer les paramètres $\boldsymbol{\theta}$ du modèle, qui comprennent les probabilités de classification p_{11} et p_{00}. Nous appliquons ensuite un processus de simulation emboîté; voir l'algorithme 2. L'objectif du premier niveau de simulation est de rétablir (en espérance) un état sans erreur. Cela génère les classes $\tilde{z}$ à partir de $P(z \mid y, \hat{z}; \hat{\boldsymbol{\theta}})$, ce qui mène aux statistiques sans biais $\tilde{\zeta}$. (Notons que les probabilités requises sont disponibles directement à partir de l'algorithme EM, puisque $P(z_i = 1 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{q_1} A_{1ji}$ et $P(z_i = 0 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{q_0} A_{0ki}$.) Par la suite, les classes $\tilde{z}^*$ sont générées par un processus bootstrap avec une matrice $\mathbf{P}$, ce qui mène à une estimation du biais et de la variance semblable à celle de l'algorithme 1. Enfin, la moyenne des résultats de cette simulation bootstrap interne est calculée au premier niveau de la simulation, afin que soit réduit l'effet du bruit attribuable au fait de tirer $\tilde{z}$ d'une loi de probabilité.

Algorithme 2 La méthode bootstrap EM

Entrée : Observations $(y_i, \hat{z}_i)$ pour $i = 1, \ldots, N$ et S_1, S_2.

1 : Estimer les paramètres du modèle $\boldsymbol{\theta}$, y compris p_{11} et p_{00}, à partir de l'algorithme EM, conditionnellement à $\hat{z}$ et y

2 : Calculer $P(z_i \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}})$ pour chaque unité i dans l'ensemble de données

3 : **for** $s_1 = 1, \ldots, S_1$ **do**

4 : Générer $\tilde{z}_i$ par $P(z_i \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}})$ pour chaque unité i dans l'ensemble de données

5 : Calculer la valeur correspondante $\tilde{\zeta}$ basée sur $(y_i, \tilde{z}_i)$ au lieu de $(y_i, \hat{z}_i)$

6 : **for** $s_2 = 1, \ldots, S_2$ **do**

7 : Générer $\tilde{z}_i^*$ par $\mathbf{P}$, conditionnellement à $\tilde{z}_i$, pour chaque unité i dans l'ensemble de données

8 : Calculer la valeur correspondante $\tilde{\zeta}^*$ basée sur $(y_i, \tilde{z}_i^*)$ au lieu de $(y_i, \hat{z}_i)$

9 : **end for**

10 : **end for**

11 : Calculer $\mathbf{Biais}_{\mathbf{comb}} = \mathrm{E}_{\tilde{z}}(\mathrm{E}(\tilde{\zeta}^* \mid \hat{z}, \tilde{z}) - \tilde{\zeta} \mid \hat{z})$ et $\mathbf{Var}_{\mathbf{comb}} = \mathrm{E}_{\tilde{z}}(\mathrm{Var}(\tilde{\zeta}^* \mid \hat{z}, \tilde{z}) \mid \hat{z})$ à partir de S_1 et S_2 simulations

Sortie : $\mathbf{Biais}_{\mathbf{comb}}$, $\mathbf{Var}_{\mathbf{comb}}$ et la matrice estimée $\mathbf{P}$

L'annexe C.2 montre que, pour $S_1, S_2 \to \infty$, l'algorithme 2 donne effectivement des estimateurs du biais et de la variance approximativement sans biais dans les cas particuliers $\zeta = T_1$ et $\zeta = \alpha_1$, à condition que les données observées suivent le modèle de mélange supposé. Pour d'autres paramètres non linéaires tels que $\zeta = \sigma_1$, aucune preuve exacte n'est disponible, mais nous étudierons le comportement de l'algorithme 2 dans une étude par simulations.

2.3.3 Méthode bootstrap SIMEX

La méthode SIMEX a été introduite par Cook et Stefanski (1994) pour les variables numériques. Elle s'appuie sur une méthode bootstrap pour ajouter diverses erreurs supplémentaires, au moyen desquelles on obtient une séquence d'estimations dans différentes conditions d'inclusion d'erreurs. Une fonction est ensuite appliquée pour extrapoler la séquence à l'estimation sans erreur. Ici, nous utilisons une extension de la méthode SIMEX aux variables catégoriques introduite par Küchenhoff et coll. (2006).

De façon classique, nous appliquerions la méthode SIMEX pour obtenir une estimation avec correction du biais du paramètre cible lui-même (ζ dans notre notation). Dans le cas présent, nous l'utilisons pour obtenir des estimations bootstrap du biais et de la variance de $\hat{\zeta}$ avec correction du biais. Nous appelons cette approche la méthode bootstrap SIMEX.

La méthode bootstrap SIMEX simule plusieurs conditions dans lesquelles des erreurs de classification sont ajoutées aux données selon la matrice $\mathbf{P}^{\lambda}$ ($\mathbf{P}$ étant donné par (2.3)), pour différentes valeurs de $\lambda \geq 0$. Pour chaque valeur de λ, la méthode bootstrap de l'algorithme 1 est appliquée aux données ajustées pour obtenir des estimations du biais et de la variance. Enfin, nous obtenons les estimations du biais et de la variance de la méthode SIMEX en extrapolant la séquence des estimations du biais et de la variance de λ à la valeur $\lambda = -1$. Cela peut se comprendre comme suit. Les données observées disponibles peuvent être

considérées comme une réalisation de l'application de la matrice $\mathbf{P}$ aux vraies données. Par conséquent, à partir des données observées à $\lambda = 0$, nous extrapolons la séquence des estimations du biais et de la variance à ce qui aurait été trouvé au point non observé de zéro classification erronée, ce qui correspond à $\lambda = -1$. La méthode bootstrap SIMEX est résumée dans l'algorithme 3.

Algorithme 3 La méthode bootstrap SIMEX

Entrée : Observations $(y_i, \hat{z}_i)$ pour $i = 1, \ldots, N$, la matrice $\mathbf{P}$ et S_1, S_2.

1 : **for** λ une plage de 0 à 5 avec incréments de 0,5 **do**
2 : Calculer $\mathbf{P}^{\lambda}$
3 : **for** $s_1 = 1, \ldots, S_1$ **do**
4 : Générer $z_i^{(\lambda)}$ par $\mathbf{P}^{\lambda}$, conditionnellement à $\hat{z}_i$, pour chaque unité i dans l'ensemble de données
5 : Calculer la valeur correspondante $\zeta^{(\lambda)}$ basée sur $(y_i, z_i^{(\lambda)})$ au lieu de $(y_i, \hat{z}_i)$
6 : **for** $s_2 = 1, \ldots, S_2$ **do**
7 : Générer $\hat{z}_i^{(\lambda)}$ par $\mathbf{P}$, conditionnellement à $z_i^{(\lambda)}$, pour chaque unité i dans l'ensemble de données
8 : Calculer la valeur correspondante $\hat{\zeta}^{(\lambda)}$ basée sur $(y_i, \hat{z}_i^{(\lambda)})$ au lieu de $(y_i, \hat{z}_i)$
9 : **end for**
10 : **end for**
11 : Calculer $\mathbf{Biais}_{\mathbf{simex}}^{(\boldsymbol{\lambda})} = \mathrm{E}_{z^{(\lambda)}}(\mathrm{E}(\hat{\zeta}^{(\lambda)} \mid \hat{z}, z^{(\lambda)}) - \zeta^{(\lambda)} \mid \hat{z})$ et $\mathbf{Var}_{\mathbf{simex}}^{(\boldsymbol{\lambda})} = \mathrm{E}_{z^{(\lambda)}}(\mathrm{Var}(\hat{\zeta}^{(\lambda)} \mid \hat{z}, z^{(\lambda)}) \mid \hat{z})$ à partir de S_1 et S_2 simulations
12 : **end for**
13 : Extrapoler $\mathbf{Biais}_{\mathbf{simex}}^{(\boldsymbol{\lambda})}$ et $\mathbf{Var}_{\mathbf{simex}}^{(\boldsymbol{\lambda})}$ à $\lambda = -1$ pour obtenir $\mathbf{Biais}_{\mathbf{simex}}$ et $\mathbf{Var}_{\mathbf{simex}}$

Sortie : $\mathbf{Biais}_{\mathbf{simex}}$, $\mathbf{Var}_{\mathbf{simex}}$

Calculer $\mathbf{P}^{\boldsymbol{\lambda}}$. Quand $\lambda = 0$, $\mathbf{P}^0$ est une matrice d'identité. Dans ce cas particulier, aucune autre erreur de classification n'est introduite à la ligne 4 de l'algorithme bootstrap SIMEX. Pour toute valeur réelle de $\lambda > 0$, $\mathbf{P}^{\lambda}$ peut être calculé au moyen de la décomposition de la valeur propre de $\mathbf{P}$. Grâce à elle, nous obtenons $\mathbf{P} = \mathbf{QAQ}^{-1}$, où $\mathbf{A}$ est une matrice diagonale de valeurs propres et $\mathbf{Q}$ est une matrice de vecteurs propres de $\mathbf{P}$. Ensuite, $\mathbf{P}^{\lambda}$ est calculé par $\mathbf{P}^{\lambda} = \mathbf{QA}^{\lambda}\mathbf{Q}^{-1}$. Nous obtenons les probabilités correspondantes $p_{11}^{(\lambda)}$ et $p_{00}^{(\lambda)}$:

$$\mathbf{P}^{\lambda} = \begin{pmatrix} p_{11}^{(\lambda)} & 1 - p_{11}^{(\lambda)} \\ 1 - p_{00}^{(\lambda)} & p_{00}^{(\lambda)} \end{pmatrix},$$

où $p_{11}^{(\lambda)}$ est la probabilité de $z^{(\lambda)} = 1$ quand $\hat{z} = 1$, et $p_{00}^{(\lambda)}$ est la probabilité de $z^{(\lambda)} = 0$ quand $\hat{z} = 0$. Ces probabilités servent à tirer $z_i^{(\lambda)}$, étant donné $\hat{z}_i$, à la ligne 4 de l'algorithme. Notons que cette procédure fonctionne seulement si $\mathbf{P}^{\lambda}$ est une vraie matrice de probabilités dans le sens que $0 \leq p_{11}^{(\lambda)}, p_{00}^{(\lambda)} \leq 1$. Pour $\lambda \geq 0$, cela est garanti à condition que $p_{11} + p_{00} > 1$ (Küchenhoff et coll., 2006). Pour $\lambda < 0$, cette propriété ne se vérifie pas. Par conséquent, l'extrapolation est une étape nécessaire de la méthode SIMEX.

Fonctions d'extrapolation. Une méthode SIMEX donne un estimateur convergent d'un paramètre d'intérêt (dans notre cas : le biais et la variance de $\hat{\zeta}$) quand la fonction d'extrapolation, laquelle décrit dans quelle

mesure l'estimateur non corrigé varie en tant que fonction de λ, est correctement spécifiée; voir Küchenhoff et coll. (2006) pour en savoir plus. La littérature sur la méthode SIMEX propose plusieurs fonctions d'extrapolation : une régression linéaire locale (LOESS) sur λ (Hopkins et King, 2010) et une régression standard sur un polynôme quadratique ou cubique de λ (Küchenhoff et coll., 2006). Nous avons comparé les trois approches dans l'étude par simulations de la section 3.

3. Étude par simulations

3.1 Paramétrage

Nous avons simulé une population de taille $N = 2\,000$ en utilisant deux classes. Pour la variable cible y, nous avons utilisé la distribution de mélange gaussien suivante : la classe 0 a une composante avec $\mu_0 = 15$, $\sigma_0 = 3$; la classe 1 a deux composantes avec $(\xi_{11}, \xi_{12}) = (0{,}5;\ 0{,}5)$, $(\mu_{11}, \mu_{12}) = (2, 4)$ et $(\sigma_{11}, \sigma_{12}) = (1, 2)$.

En ce qui concerne la variable de classification vraie z, nous avons mis à l'essai deux proportions différentes de la classe 1 : α_1 est 0,3 ou 0,5. La variable de classification observée $\hat{z}$ a été générée à partir de z au moyen de la matrice de transition $\mathbf{P}$ donnée dans l'équation (2.3). Dans l'étude par simulations, les valeurs de p_{11} et p_{00} ont été définies à 0,6, 0,75 ou 0,9. Pour chaque paramètre de α_1, p_{11} et p_{00}, nous avons utilisé $S_0 = 100$, ce qui signifie que 100 ensembles de $\hat{z}$ ont été générés à partir de z. Dans chaque ensemble s_0, 5 % des unités de la population (donc 100 au total) ont été sélectionnées au hasard comme échantillon de vérification, lequel a servi à obtenir les valeurs de départ pour l'algorithme EM de $\hat{p}_{11}$, $\hat{p}_{00}$ et les autres paramètres au niveau de la classe. Les valeurs de départ des paramètres au niveau des composantes ont été obtenues par des k-moyennes; voir Li (2020b) pour en savoir plus. Dans une étude préliminaire, nous avions également mis à l'essai l'algorithme EM sans échantillon de vérification (voir la section 3.3).

Pour chaque ensemble s_0, nous avons appliqué la méthode bootstrap, la méthode bootstrap SIMEX et la méthode bootstrap EM pour estimer le biais et la variance correspondants des paramètres de domaine estimés $\hat{\zeta}$ qui sont donnés dans le tableau 2.1 : la somme totale pour la classe 1 (T_1), la proportion de la classe 1 (α_1) et l'écart-type de la classe 1 (σ_1). Les estimations du biais et de la variance sont présentées ici comme la moyenne sur les ensembles S_0.

Pour ce qui est de la méthode bootstrap SIMEX, nous avons mis à l'essai le paramétrage ci-dessus de l'étude par simulations de trois fonctions d'extrapolation différentes, comme nous l'avons indiqué dans la section 2.3.3; voir aussi l'annexe A. Les estimations du biais et de l'erreur-type de la fonction LOESS et du polynôme de troisième degré étaient plus proches des vraies valeurs que celles du polynôme du second degré. Étant donné que la fonction LOESS a été utilisée dans une étude précédente sur les classifications erronées (voir Hopkins et King (2010)), nous avons employé cette fonction d'extrapolation dans le reste du présent article.

L'algorithme EM était interrompu soit quand aucune des estimations de paramètre ne variait de plus de 0,001 entre deux itérations, soit après 5 000 itérations. En pratique, tant dans la présente étude par

simulations que dans l'étude de cas de la section 4, ce nombre maximal d'itérations n'a jamais été atteint. Pour 1 % des ensembles de données simulés environ, l'algorithme EM n'a pas convergé correctement en raison de problèmes numériques. Ceux-ci ont été causés par un choix malheureux de valeurs de départ estimées à partir de l'échantillon de vérification, par exemple une valeur de départ de p_{11} exactement égale à 1. En principe, ce problème serait facilement évité par un léger changement des valeurs de départ. Cependant, comme ce problème ne touchait qu'un petit nombre de cas, nous n'avons pas tenu compte de ces cas par souci de commodité dans les résultats ci-dessous.

Les nombres d'itérations dans les méthodes S, S_1 et S_2 ont été fixés à 100. Pour obtenir des données repères pour le biais et la variance vrais des paramètres de domaine estimés $\hat{\zeta}$, 1 000 ensembles de $\hat{z}$ ont été simulés.

3.2 Résultats

Estimation du biais. La figure 3.1 montre le biais de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ pour $\alpha_1 = 0{,}3$ (ligne 1, 3 et 5 respectivement) et pour $\alpha_1 = 0{,}5$ (ligne 2, 4 et 6 respectivement). Chaque colonne affiche les résultats sous les trois paramètres p_{11}. Dans chaque sous-graphique, l'axe horizontal indique les valeurs de p_{00}, l'axe vertical définit les estimations dans différentes conditions et méthodes. De plus, les différentes méthodes sont représentées par différents symboles, soit les vraies valeurs (•), le bootstrap (Δ), le bootstrap SIMEX (∇) et la méthode bootstrap EM ($\times$). La figure 3.2 montre l'erreur-type (racine carrée de la variance) de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ pour $\alpha_1 = 0{,}3$ et $\alpha_1 = 0{,}5$ pour les mêmes paramètres.

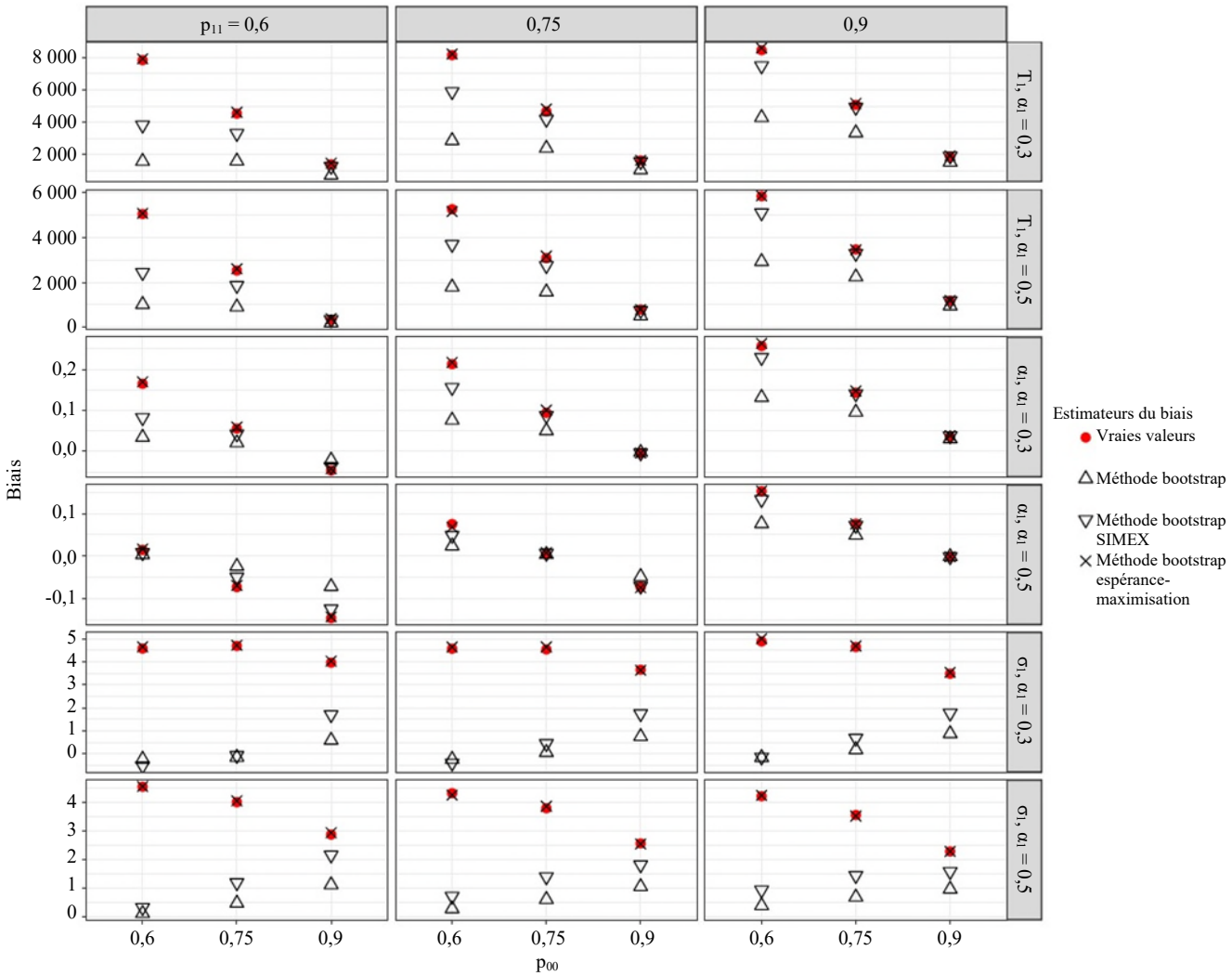
Dans l'ensemble, pour une valeur donnée de p_{11}, le biais de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ diminuaient quand la valeur de p_{00} était plus grande. Pour une valeur donnée de p_{00}, le biais augmentait quand la valeur de p_{11} était plus grande. Ce résultat peut se comprendre comme suit. Dans le cas du total, une expression analytique du biais de $\hat{T}_1$ est donnée par (C.1). Normalement, on ne peut pas calculer cette expression directement puisque T_0 et T_1 sont inconnus dans des situations réelles, mais dans les simulations, nous connaissons leurs valeurs. Dans notre exemple, à $\alpha_1 = 0{,}3$; nous constatons que $T_0 = 15 \times (1 - 0{,}3) \times 2\,000 = 21\,000$ et $T_1 = (0{,}5 \times 2 + 0{,}5 \times 4) \times (0{,}3) \times 2\,000 = 1\,800$. Une augmentation de p_{00} de 0,6 à 0,9 pour une valeur donnée de p_{11} réduit le biais puisque la contribution $(1 - p_{00})\,T_0$ passe de 8 400 à 2 100. Il s'agit des unités de la vraie classe 0 qui sont observées par erreur comme étant de la classe 1 (surestimation). À l'inverse, une augmentation de p_{11} passant de 0,6 à 0,9 pour une valeur donnée de p_{00} entraîne une légère augmentation du biais en raison de l'augmentation de la contribution $(p_{11} - 1)\,T_1$ de -720 à -180. Cette contribution fait référence aux unités de la vraie classe 1 qui sont observées par erreur comme étant de la classe 0 (sous-estimation). En général, pour une valeur donnée de p_{11}, le biais des statistiques d'intérêt ($\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$) diminue quand la valeur de p_{00} est plus grande parce que leur surestimation diminue. À l'inverse, pour une valeur donnée de p_{00}, le biais des statistiques augmente quand la valeur de p_{11} est plus grande parce que sa sous-estimation diminue.

Pour ce qui est des trois méthodes d'estimation (voir la figure 3.1), nous avons constaté que les estimations du biais pour les statistiques d'intérêt de la méthode bootstrap EM étaient les plus proches des vraies valeurs. Les estimations du biais de la méthode SIMEX étaient plus proches des vraies valeurs que

les estimations de la méthode bootstrap, mais elles s'écartaient tout de même considérablement du vrai biais. De plus, nous avons constaté que les estimations du biais des trois méthodes étaient plus proches du biais pour T_1 et σ_1 quand les probabilités de classification erronée étaient réduites (valeurs de p_{11} ou p_{00} plus proches de 1). Quand p_{11} et p_{00} étaient égaux à 0,9, les estimations du biais de la méthode bootstrap et de la méthode bootstrap SIMEX se chevauchaient même avec le vrai biais correspondant. Enfin, il convient de mentionner que quand $\alpha_1 = 0{,}5$ et $p_{00} = p_{11}$, le vrai biais de $\hat{\alpha}_1$ est égal à 0 et que les trois méthodes estimaient ce biais correctement.

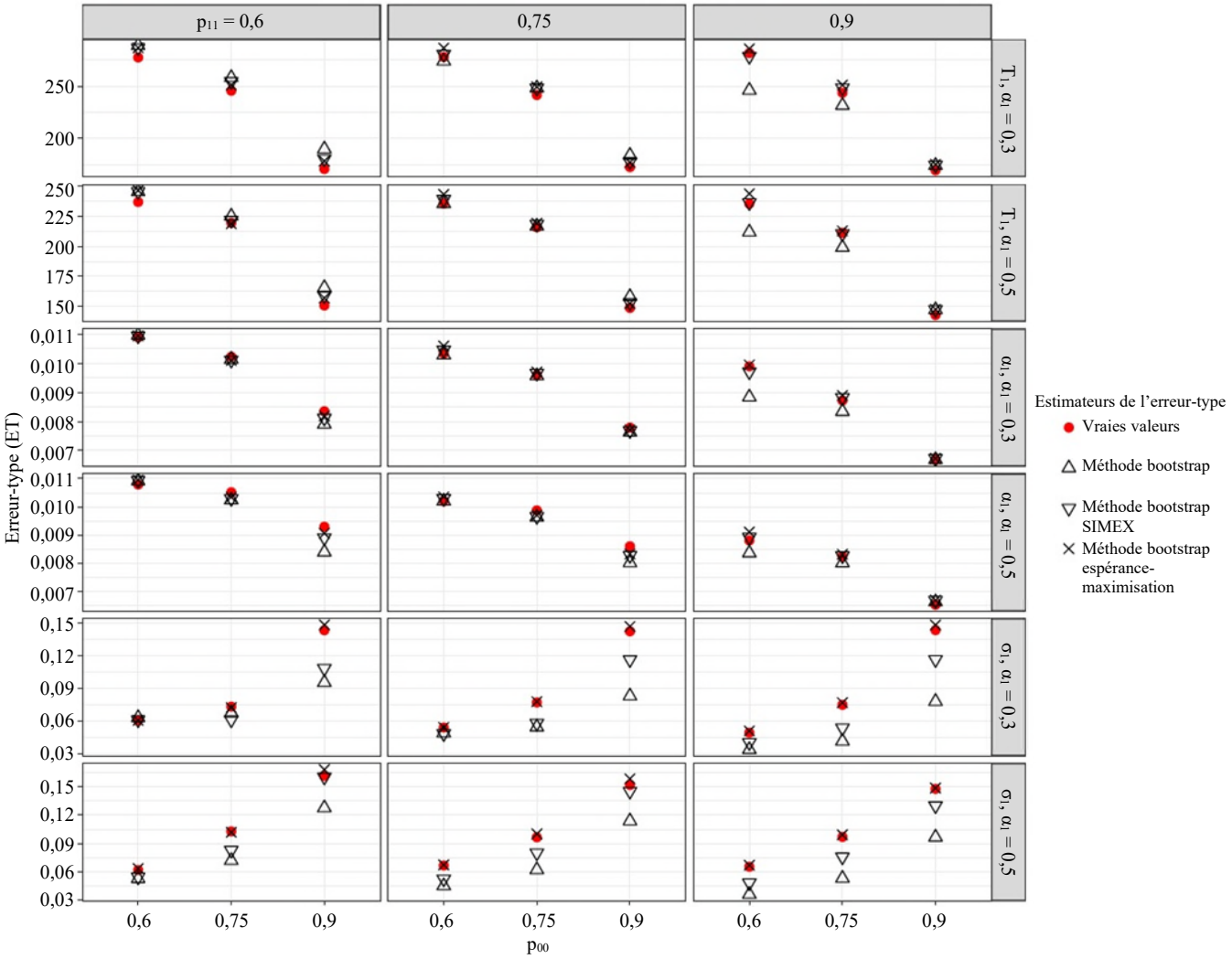
Estimation de la variance. La figure 3.2 montre l'erreur-type vraie et estimée de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ pour $\alpha_1 = 0{,}3$ (ligne 1, 3 et 5 respectivement) et pour $\alpha_1 = 0{,}5$ (ligne 2, 4 et 6 respectivement). Les vraies erreurs-types de $\hat{T}_1$ et de $\hat{\alpha}_1$ ont été réduites avec moins de classifications erronées avec p_{00} et alors que p_{11} passe de 0,6 à 0,9. Ce résultat découle de l'expression analytique (C.2) pour la variance de $\hat{T}_1$.

Figure 3.1 Estimation du biais dans l'étude par simulations.



Notes : Les lignes représentent différents paramètres de domaine (T_1, α_1, σ_1), pour deux configurations de α_1 : $\alpha_1 = 0{,}3$ et $\alpha_1 = 0{,}5$. Les colonnes représentent différentes valeurs de p_{11}; différentes valeurs de p_{00} sont données dans chaque colonne.

Figure 3.2 Estimation de l'erreur-type dans l'étude par simulations.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure 3.1.

Étonnamment, l'erreur-type de $\hat{\sigma}_1$ augmentait clairement pour une valeur donnée de p_{11} quand p_{00} augmente en passant de 0,6 à 0,9. De plus, il y avait une très faible réduction de cette erreur-type pour une valeur donnée de p_{00} quand p_{11} passait de 0,6 à 0,9. Ce résultat peut s'expliquer comme suit. Le niveau de chiffre d'affaires moyen est beaucoup plus élevé dans la classe 0 que dans la classe 1. Quand la plupart des unités observées dans la classe 1 proviennent vraiment de la classe 1 et que relativement peu d'unités proviennent de la classe 0 ($p_{00} = 0{,}9$), l'erreur-type de $\hat{\sigma}_1$ est grande puisqu'il y a une variation considérable selon les répliques dans lesquelles les valeurs de chiffre d'affaires de la classe 0 seront observées comme étant de la classe 1. Dans certaines des répliques, cela concerne les valeurs aberrantes comparativement à la vraie distribution dans la classe 1. Quand le nombre d'unités de la classe 0 augmente ($p_{00} = 0{,}75$), cette variation des valeurs de chiffre d'affaires des répliques diminue et donc l'erreur-type de $\hat{\sigma}_1$ diminue. Puisque les valeurs de chiffre d'affaires de la classe 0 sont plus grandes que celles de la classe 1, l'effet des valeurs variables de p_{00} est plus grand que pour p_{11}.

Pour ce qui est des trois méthodes d'estimation (voir la figure 3.2), nous avons constaté, comme pour le biais, que les estimations de l'erreur-type pour les statistiques d'intérêt de la méthode bootstrap EM étaient

les plus proches des vraies valeurs. Pour $\hat{T}_1$ et $\hat{\alpha}_1$, les estimations de l'erreur-type des trois méthodes étaient presque tout aussi bonnes dans la plupart des conditions. Dans les autres conditions, les estimations de la méthode bootstrap étaient les moins exactes, suivies de celles de la méthode bootstrap SIMEX, qui étaient proches des vraies valeurs. Les erreurs-types estimées de $\hat{\sigma}_1$ étaient proches de leurs vraies valeurs dans le cas de la méthode bootstrap EM, beaucoup plus proches que celles des deux autres méthodes. Des valeurs plus grandes de l'erreur-type de $\hat{\sigma}_1$ entraînent des différences d'estimation plus grandes entre les trois méthodes.

3.3 Autres études par simulations

Nous résumons ci-dessous les résultats de trois autres études par simulations.

Échantillon de vérification. Dans une étude préliminaire, nous avons comparé l'estimation du biais et de l'erreur-type avec et sans échantillon de vérification (Li, 2020a). Nous avons mis à l'essai des valeurs de p_{00} et p_{11} de 0,6; 0,75 et 0,9; $N = 2\,000$, α_1 égal à 0,1; 0,3; 0,5; 0,7 et 0,9; une seule composante gaussienne dans chaque classe, les valeurs de μ_1 de 2, 10, 12 et 15, μ_0 étant fixées à 15, $\sigma_1 = 1$ et $\sigma_0 = 2$. Dans les situations sans échantillon de vérification, différentes valeurs de départ ont été mises à l'essai pour l'algorithme EM. Pour p_{00} et p_{11}, des valeurs de départ de 0,6; 0,75 et 0,9 ont été utilisées, ce qui donnait neuf combinaisons pour la paire (p_{00}, p_{11}). En choisissant différentes valeurs pour p_{00} et p_{11}, les points de départ peuvent être considérés comme représentatifs dans l'espace des paramètres. La valeur de départ de α_1 a été établie selon $\alpha_1 = \sum \hat{z}_i / N \times (3 - p_{11} - p_{00}) - (1 - p_{00})$ qui est une estimation sans biais de α_1 (Kloos et coll., 2021). Les moyennes et les variances ont commencé par des statistiques robustes obtenues à partir de classes observées, où μ_g pour les deux classes a été initialisé à la médiane de la variable cible correspondante, et σ_g commençait par $k \times$ MAD où $k = 1/\Phi^{-1}(3/4) \approx 1{,}48$ et MAD est l'écart absolu médian; voir Rousseeuw et Croux (1993).

Nous avons constaté que le biais estimé des statistiques d'intérêt au moyen de la méthode bootstrap EM avec et sans échantillon de vérification produisait presque les mêmes résultats, sauf dans des conditions d'estimation difficiles. Ces conditions difficiles se présentaient quand $\mu_0 = \mu_1$ était combiné à des valeurs inférieures pour α_1 et à de plus petites valeurs de p_{00} et p_{11} (non illustré). Dans ces conditions, les estimations du biais étaient moins exactes, mais le vrai biais (relatif) était petit. Les estimations du biais étaient proches de zéro, tandis que le vrai biais relatif atteignait 0,03. Pour l'erreur-type des statistiques d'intérêt, la méthode bootstrap EM avec et sans échantillon de vérification produisait presque les mêmes résultats dans toutes les conditions testées.

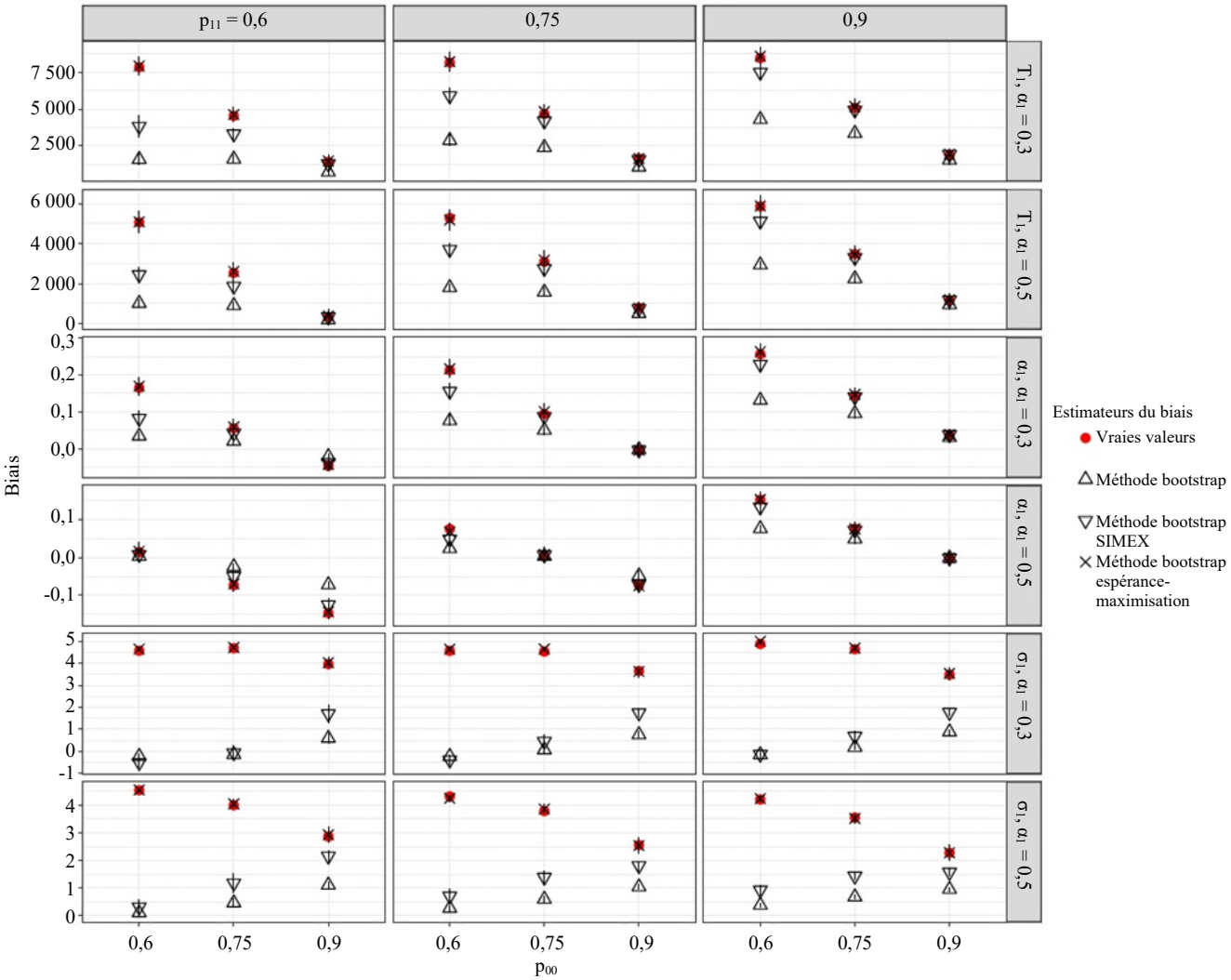
Résultats pour μ_1. En plus d'avoir estimé le biais et l'erreur-type de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ pour $\alpha_1 = 0{,}3$; nous les avons également calculés pour $\hat{\mu}_1$. Notons que $\hat{\mu}_1 = \hat{T}_1 / (\hat{\alpha}_1 \times N)$ et que donc les résultats pour $\hat{\mu}_1$ découlent de ceux de $\hat{T}_1$ et $\hat{\alpha}_1$. En ce qui concerne la comparaison des trois méthodes, nous avons conclu que le vrai biais et l'erreur-type étaient estimés le plus précisément par la méthode bootstrap EM. Les résultats détaillés se trouvent dans Li (2020a).

Intervalles de confiance. Jusqu'à présent, nous avons présenté les résultats correspondant à la moyenne sur $S_0 = 100$ ensembles de $\hat{z}$ qui ont été générés à partir de z. En pratique, on aurait un seul ensemble de valeurs $\hat{z}$. Cela nous amène à nous demander si la méthode bootstrap EM donne aussi des estimations du

biais plus exactes que les deux autres méthodes dans le cas d'un seul échantillon ou seulement en moyenne. Afin de répondre à cette question, nous avons estimé un intervalle de confiance de 95 % pour les estimations du biais des statistiques d'intérêt pour les trois méthodes au moyen de $S_0 = 100$ répliques. Cet intervalle a été estimé à 1,96 fois l'erreur-type du biais estimé qui a été calculé à son tour à partir des $S_0 = 100$ estimations du biais.

À partir de la section 3.2, nous avons conclu que dans certaines configurations, les trois méthodes produisaient des estimations exactes du biais pour $\hat{T}_1$ et $\hat{\alpha}_1$, mais que sinon, l'estimation du biais par la méthode bootstrap EM était visiblement plus proche de la vraie valeur que celle de la méthode bootstrap SIMEX et de la méthode bootstrap, en moyenne sur S_0 répliques. Étant donné que les intervalles de confiance de 95 % étaient très petits pour les trois méthodes (voir la figure 3.3), dans la plupart des configurations aussi pour un seul ensemble de valeurs $\hat{z}$, l'estimation bootstrap EM du biais était plus proche du vrai biais que les estimations de la méthode bootstrap et de la méthode bootstrap SIMEX.

Figure 3.3 Estimation du biais dans l'étude par simulations avec intervalle de confiance de 95 %.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure 3.1.

4. Étude de cas

4.1 Données

Nous avons réalisé une étude de cas afin d'évaluer le rendement de nos méthodes dans des applications réelles. Dans l'étude de cas, le biais et la variance des statistiques de domaine estimées $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ ont été estimés au moyen des mêmes méthodes que celles de l'étude par simulations.

Pour l'étude de cas, nous avons commencé par un ensemble de données qui contient le logarithme du chiffre d'affaires annuel pour une population d'entreprises dont nous connaissons l'adresse de site Web; voir Oosterveen (2020). Les entreprises sont classées par code d'activité économique, selon la classification européenne NACE Rév. 2 (Nomenclature générale des activités économiques dans la Communauté européenne) (Eurostat, 2008). Les entreprises, leurs codes de la NACE et les adresses de site Web ont été obtenus à partir du registre statistique des entreprises (RSE). Concernant certaines entreprises, nous avons obtenu l'adresse de site Web à partir d'un ensemble de données contenant des adresses URL que nous avons récupérées d'une entreprise externe DataProvider.

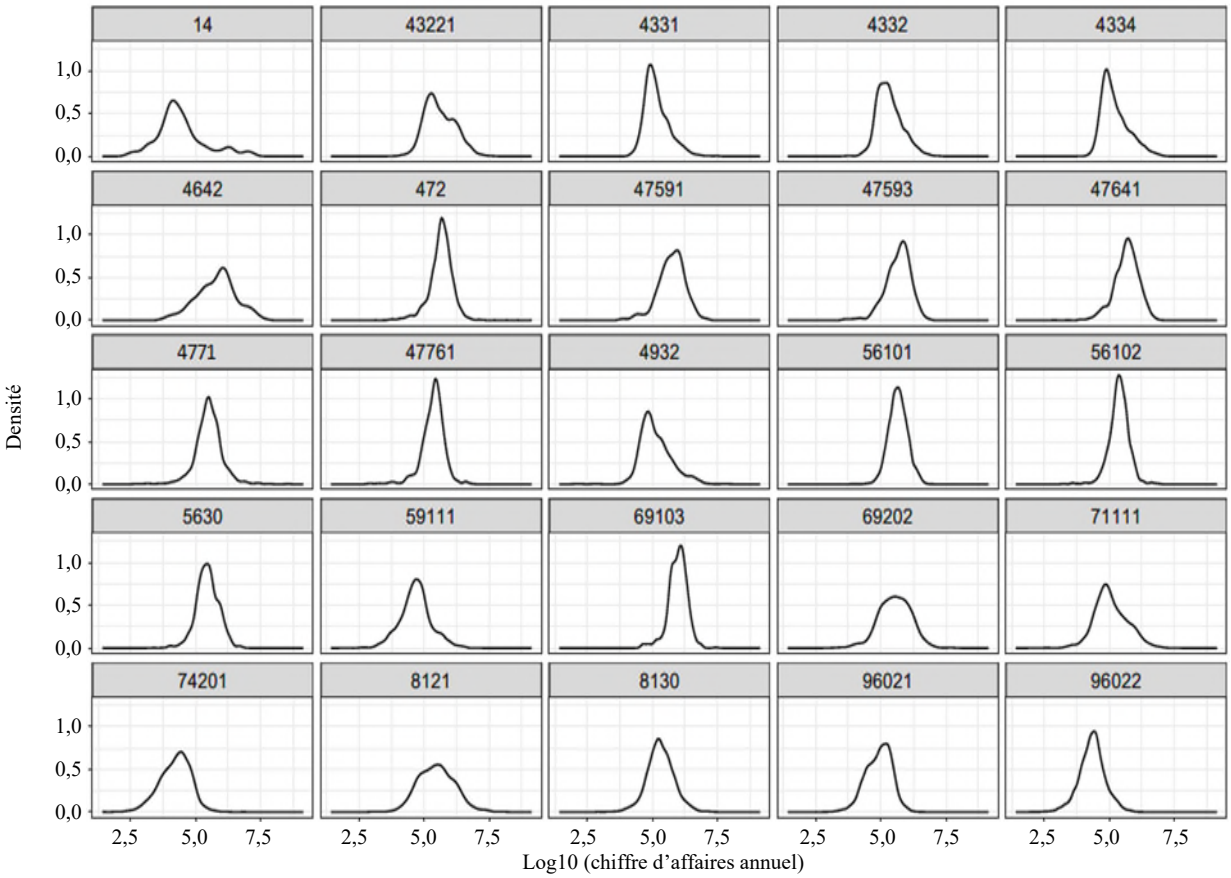
Ces valeurs de code de la NACE peuvent présenter des erreurs de classification. Les codes de la NACE des entreprises plus grandes et plus complexes sont vérifiés manuellement et corrigés au besoin. C'est la raison pour laquelle notre étude de cas se limite aux entreprises les plus simples, composées au maximum de trois unités juridiques, car ce sont celles dont les codes de la NACE sont les plus sujets aux classifications erronées en pratique. De plus, nous avons commencé par une liste restreinte de 25 activités économiques; les codes de la NACE sont donnés dans la figure 4.1. Cette liste restreinte contient quelques groupes de codes de la NACE ayant des classes semblables au sein d'un groupe, comme le commerce de gros de vêtements et le commerce de détail de vêtements, et des classes différentes des autres, comme le transports de voyageurs par taxis.

Comme dans l'étude par simulations, nous voulions que l'étude de cas commence par un ensemble de données exempt de classifications erronées, pour ensuite introduire des classifications erronées délibérément afin de mettre à l'essai le rendement des trois méthodes. Cependant, les données que nous avons extraites du RSE concernaient des codes de la NACE qui pouvaient déjà contenir des classifications erronées. Nous n'avons donc pas utilisé toutes les petites entreprises ayant un site Web (76 270 entreprises), mais plutôt une sélection d'entreprises qui avaient une probabilité relativement élevée d'avoir le bon code de la NACE. Nous avons effectué cette sélection en prédisant le code de la NACE des entreprises à partir du texte de la page principale du site Web de l'entreprise. Ces textes ont été moissonnés et prétraités; voir Oosterveen (2020) sur la façon dont cela a été fait. Nous avons entraîné trois algorithmes d'apprentissage automatique différents (bayésien naïf, machine à vecteurs de support et forêt aléatoire) à prédire le code de la NACE, au moyen d'une procédure de validation croisée à dix blocs. Dans chaque bloc, le modèle a été entraîné sur 90 % des données et le modèle ajusté a servi à prédire les 10 % restants des données. Un code de la NACE observé était considéré comme correct quand les modèles ajustés des trois algorithmes prédisaient le même code ou quand il était prédit par deux algorithmes et que la confiance de prédiction des modèles était

relativement élevée; voir la description plus détaillée de la sélection dans Oosterveen (2020). Cette sélection a donné 45 965 entreprises.

Dans le présent article, nous nous limitons aux classifications erronées binaires. Nous avons donc effectué une autre sélection de paires de deux codes de la NACE à partir de la liste restreinte de 25 codes de la NACE. Nous avons sélectionné différentes paires, chaque paire sélectionnée étant un « cas ». Les paires différaient dans l'étendue du chevauchement entre les distributions de logarithme du chiffre d'affaires et dans les formes des distributions. De plus, nous avons veillé à ce que les deux groupes ne soient pas trop petits et pas trop déséquilibrés, ce qui signifie que ce α_1 n'était près ni de 0 ni de 1. L'ensemble complet des essais que nous avons effectués se trouve dans Li (2020b). Nous présentons ici seulement les résultats des deux paires les plus intéressantes, donnés dans le tableau 4.1.

Figure 4.1 Distribution de la densité du logarithme du chiffre d'affaires pour tous les groupes.



Notes : Les étiquettes font référence aux codes de la NACE (Nomenclature générale des activités économiques dans la Communauté européenne).

Le tableau 4.2 présente quelques statistiques de base pour chacune des quatre sous-populations définies par les codes de la NACE sélectionnés dans le tableau 4.1, à savoir la taille, le total, la moyenne et l'écart-type du logarithme du chiffre d'affaires annuel par entreprise. Le cas 1 comporte deux distributions bien séparées et un grand nombre d'entreprises par classe. En revanche, dans le cas 2, les deux distributions sont

moins bien séparées et ont un plus petit nombre d'entreprises par classe. Le tableau 4.2 présente les vraies statistiques de domaine ζ de chaque classe.

Nous avons supprimé les valeurs aberrantes avant d'appliquer nos méthodes. Nous avons retiré les entreprises dont la valeur du chiffre d'affaires était inférieure à $Q_1 - 1{,}5 \times \text{IQR}$ ou supérieure à $Q_3 + 1{,}5 \times \text{IQR}$, où Q_1 est le premier quartile des entreprises de la vraie classe, Q_3 est le troisième quartile et $\text{IQR} = Q_3 - Q_1$ correspond à l'écart interquartile. La suppression des valeurs aberrantes était nécessaire seulement dans le cas 2. Nous avons supprimé 17 valeurs aberrantes de la NACE 4932 et 6, de la NACE 8121.

Tableau 4.1
Groupes sélectionnés et leurs allocations dans l'étude de cas.

Code de la NACE	Description de l'activité économique	Cas	Classe
56101	Restauration	Cas 1	Classe 1
96022	Soins de beauté, pédicures et manucures, maquillage et conseil en image	Cas 1	Classe 0
4932	Transports de voyageurs par taxi	Cas 2	Classe 1
8121	Nettoyage courant des bâtiments	Cas 2	Classe 0

Note : NACE = Nomenclature générale des activités économiques dans les Communautés européennes.

4.2 Paramétrage

Contrairement à l'étude par simulations, le nombre de composantes du modèle de mélange gaussien n'est pas connu ici et il doit être déterminé. À cette fin, nous avons ajusté un modèle standard de mélange gaussien (c'est-à-dire sans composante de classification erronée) pour chaque classe par cas séparément. Nous avons ensuite déterminé les critères BIC et sBIC de ces modèles ajustés, puisque ces mesures pouvaient servir à sélectionner le nombre de composantes; voir la section 2.2. Par souci de commodité, nous avons ajusté ce modèle de mélange gaussien aux données réelles plutôt qu'aux données générées par des classifications erronées, ce qui nous a évité d'avoir à l'exécuter pour 100 ensembles (S_0) multipliés par neuf conditions de classification erronée (voir ci-dessous). En pratique, il faut ajuster le modèle standard de mélange gaussien aux données observées contenant des classifications erronées, ce qui devrait donner un peu plus de composantes que quand le modèle est exécuté sur des données vraies. Dans l'annexe B, le nombre optimal estimé de composantes pour les cas 1 et 2 est indiqué selon les critères BIC et sBIC. Le nombre optimal de composantes dans la classe 0 du cas 2 était deux quand nous utilisions sBIC et un quand nous utilisions BIC. Pour les trois autres classes, nous n'a constaté aucune différence dans le nombre optimal de composantes. Pour le reste, nous avons utilisé le nombre de composantes selon le critère sBIC, comme l'illustre le tableau 4.2.

Comme dans le paramétrage de l'étude par simulations (section 3.1), la variable de classification observée $\hat{z}$ a été générée à partir de z au moyen de $\mathbf{P}$ (équation 2.3) avec des valeurs de p_{11} et p_{00} de 0,6, 0,75 ou 0,9. Pour chaque paramètre de p_{11} et p_{00}, nous avons généré $S_0 = 100$ ensembles de $\hat{z}$ à partir de z. Dans chaque ensemble S_0, 5 % des unités de la population ont été sélectionnées aléatoirement comme échantillon de vérification, lequel a servi à estimer $\hat{p}_{11}$ et $\hat{p}_{00}$. Ces estimations ont été utilisées comme

valeurs de départ de l'algorithme EM. Les estimations finales de $\hat{p}_{11}$ et $\hat{p}_{00}$ par l'algorithme EM ont été entrées pour les méthodes bootstrap et la méthode bootstrap SIMEX. Le nombre d'itérations S, S_1 et S_2 était de 100. Les estimations globales du biais et de la variance des trois méthodes ont été calculées comme la moyenne sur S_0 estimations du biais et de la variance. Comme auparavant, le vrai biais et la vraie variance estimés des paramètres de domaine estimés étaient fondés sur 1 000 ensembles de $\hat{z}$.

Tableau 4.2
Statistiques de domaine pour chaque classe de l'étude de cas.

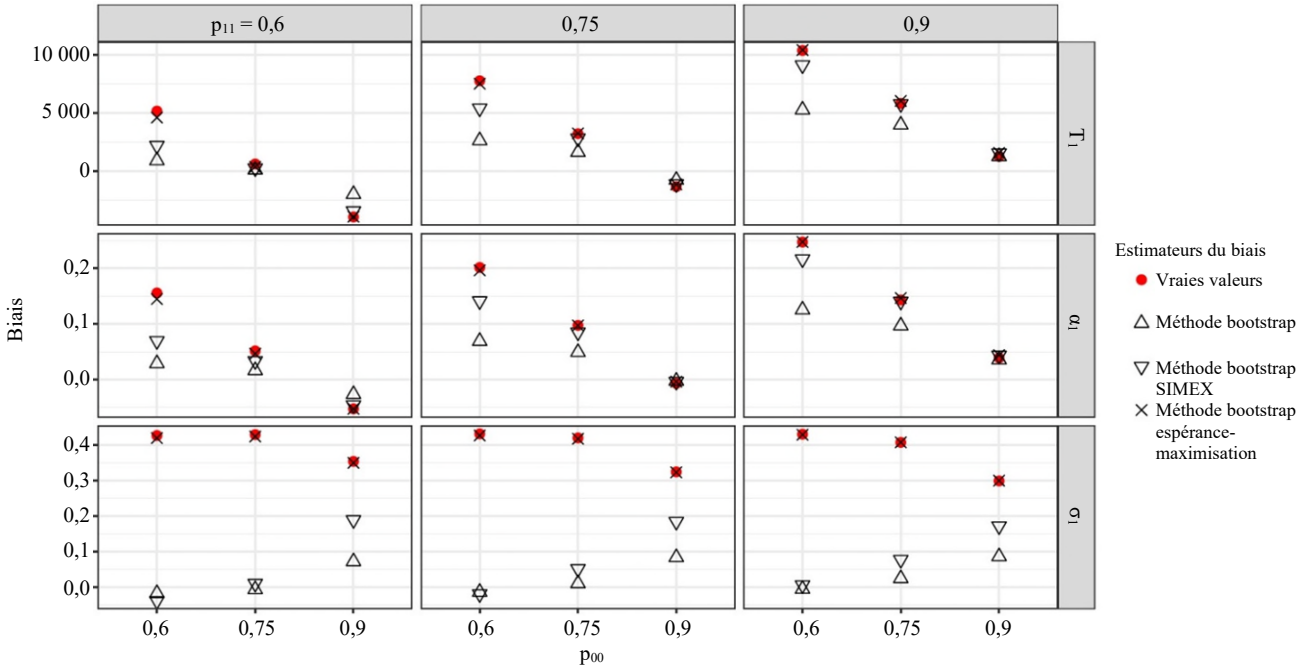
Cas	Classe	Nombre de composantes	Taille	Total	Moyenne	Écart-type
Cas 1	Classe 1	1	3 076	17 415	5,66	0,358
	Classe 0	2	6 993	30 377	4,34	0,501
Cas 2	Classe 1	2	642	3 294	5,13	0,597
	Classe 0	2	1 067	5 847	5,48	0,687

4.3 Résultats

La tendance du biais (voir les figures 4.2 et 4.3) et de l'erreur-type (figures 4.4 et 4.5) en tant que fonction de p_{00} et de p_{11} est la même que celle trouvée précédemment dans l'étude par simulations. Dans la plupart des configurations mises à l'essai, nous obtenions l'estimation du biais de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ la plus exacte par la méthode bootstrap EM, suivie de la méthode bootstrap SIMEX, tandis que la méthode bootstrap donnait les résultats les moins exacts. Une exception s'est produite dans le cas 2, $p_{00} = 0{,}9$ pour $\hat{\sigma}_1$ où la méthode bootstrap SIMEX était la plus exacte, suivie de la méthode bootstrap EM. Dans certaines des configurations avec $p_{00} = 0{,}75$, la méthode bootstrap EM et la méthode bootstrap SIMEX ont produit des résultats presque identiques pour le biais.

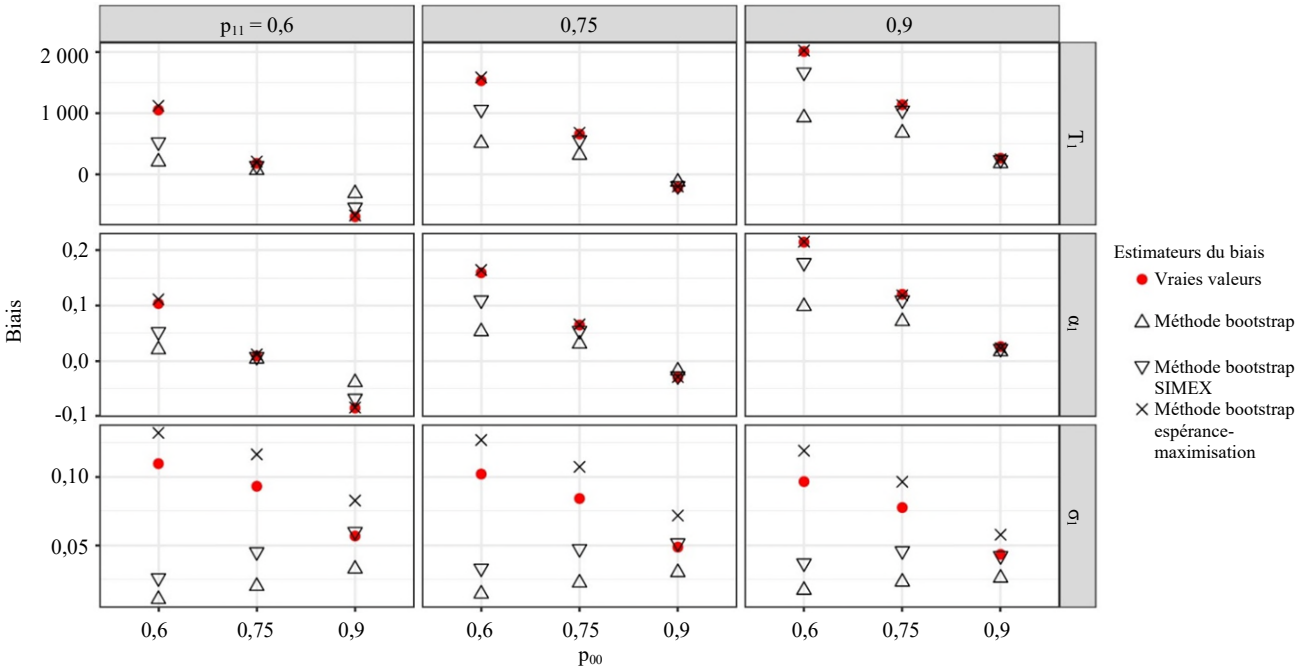
Les estimations du biais par la méthode bootstrap EM se chevauchaient presque avec les vraies valeurs correspondantes. Dans le cas 2 seulement, les estimations du biais de $\hat{\sigma}_1$ présentaient une certaine distance par rapport aux vraies valeurs, tandis que le biais pour $\hat{T}_1$ et $\hat{\alpha}_1$ demeurait exact. La figure 4.1 montre que, dans ce cas, la vraie distribution de y dans la classe 1 (code de la NACE 4932) est plus asymétrique à droite que les autres distributions examinées dans l'étude. Cela peut toucher particulièrement la statistique $\hat{\sigma}_1$, qui est relativement sensible aux valeurs de la queue droite de la distribution. Une analyse de suivi a montré que la méthode bootstrap EM donnait de meilleurs résultats dans cet exemple quand le nombre de composantes par classe passait de deux à trois (voir les figures B.1 et B.2 de l'annexe B). Cela indique qu'il pourrait être avantageux de choisir un nombre relativement élevé de composantes du mélange si l'on sait que la distribution est asymétrique et si des paramètres non linéaires et non robustes comme σ_1 sont d'intérêt. Il convient toutefois de mentionner que la statistique $\hat{\sigma}_1$ n'est pas publiée directement en tant que sortie dans les statistiques officielles. Dans l'ensemble, les résultats indiquent que la méthode bootstrap EM a habituellement des résultats exacts, même dans des situations difficiles, quand les deux classes contiennent moins d'unités et que leurs distributions se chevauchent.

Figure 4.2 Estimation du biais dans le cas 1.



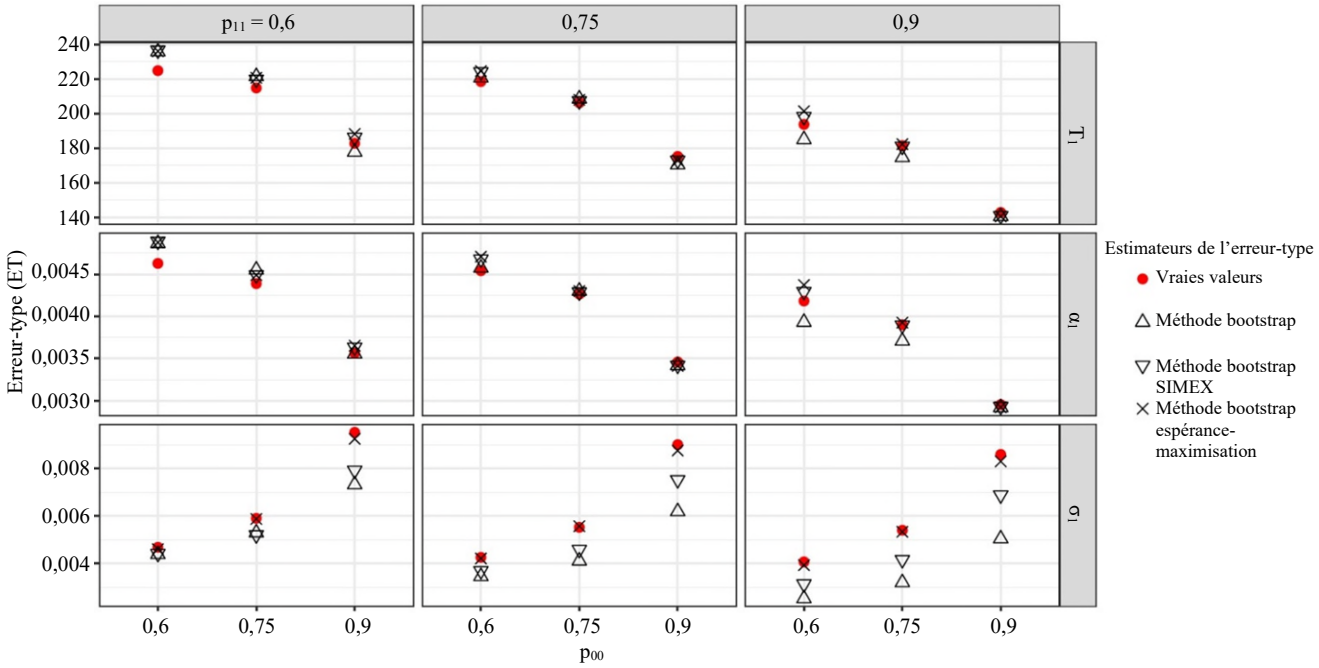
Notes : Les lignes représentent différents paramètres de domaine (T_1, α_1, σ_1). Les colonnes représentent différentes valeurs de p_{11}; différentes valeurs de p_{00} sont données dans chaque colonne.

Figure 4.3 Estimation du biais dans le cas 2.



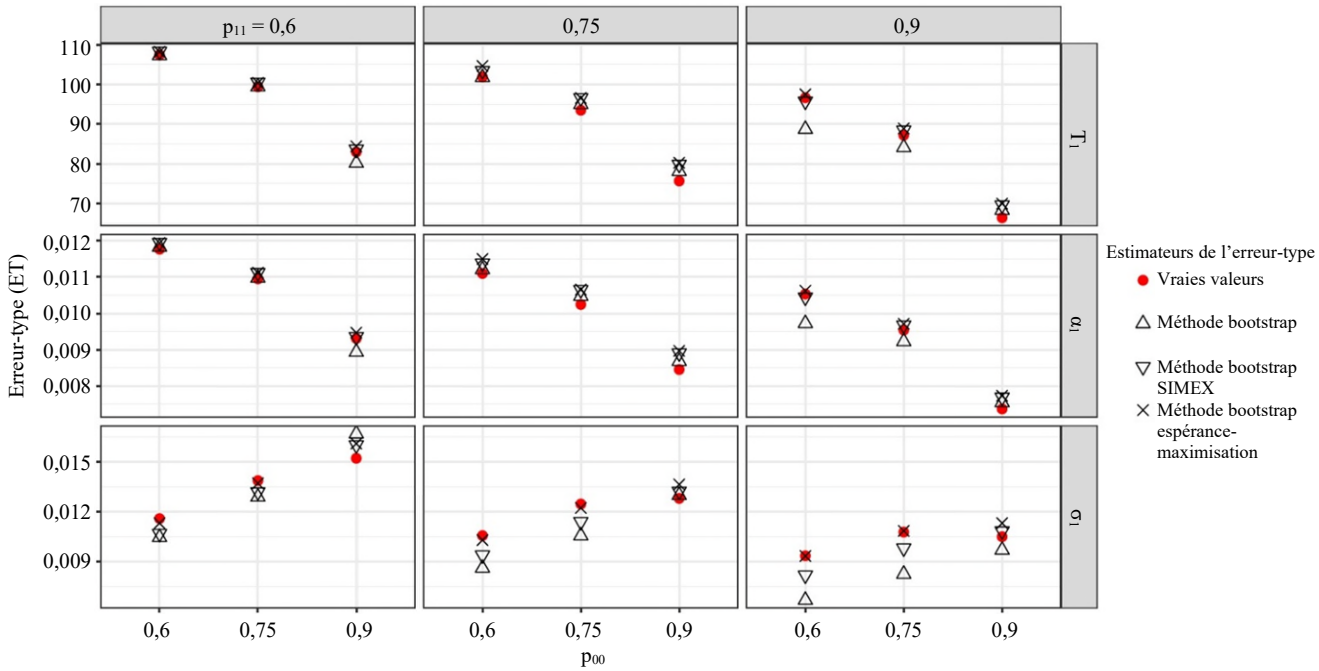
Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure 4.2.

Figure 4.4 Estimation de l'erreur-type dans le cas 1.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure 4.2.

Figure 4.5 Estimation de l'erreur-type dans le cas 2.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure 4.2.

Les différences entre les trois méthodes concernant l'exactitude de l'erreur-type estimée de $\hat{T}_1$, $\hat{\alpha}_1$ et $\hat{\sigma}_1$ étaient plus petites que celles du biais estimé. Dans les configurations où les estimations des trois méthodes étaient nettement différentes, la plupart du temps, l'estimation de l'erreur-type par la méthode bootstrap EM se rapprochait le plus de sa vraie valeur, suivie de la méthode bootstrap SIMEX, tandis que la méthode bootstrap était la moins exacte. Parfois, l'erreur-type estimée de la méthode bootstrap SIMEX se rapprochait le plus de la vraie erreur-type, par exemple pour le cas 1, $p_{00} = 0{,}6$; $p_{11} = 0{,}9$ et la statistique cible $\hat{T}_1$, $\hat{\alpha}_1$ mais les différences avec la méthode bootstrap EM étaient petites.

5. Discussion

Dans le présent article, nous avons proposé une méthode bootstrap EM pour estimer l'exactitude des statistiques de domaine, pour ce qui est du biais et de la variance, en présence de classifications erronées. L'utilisation d'un algorithme EM n'est pas nouvelle dans les études de situations dans lesquelles des erreurs de classification se produisent. À titre d'exemple, Sinclair et Hooker (2017) et Kosinski et Flanders (1999) ont cherché à estimer les paramètres d'un modèle en présence d'erreurs de classification dans une ou plusieurs de ses variables. Notre méthode bootstrap EM suppose un modèle de mélange gaussien. Dans notre étude, nous n'utilisons pas le modèle de mélange pour améliorer directement l'exactitude de nos estimations (totales ou moyennes), mais plutôt pour estimer l'exactitude d'un estimateur donné. La principale raison est que les organismes nationaux de statistique ont tendance à éviter d'utiliser des estimateurs fondés sur un modèle directement à des fins de production, particulièrement si les hypothèses du modèle ne sont pas vérifiables (Van den Brakel et Bethlehem, 2008). En revanche, l'utilisation de modèles pour estimer la qualité de la production est acceptée. Une autre raison pour laquelle nous avons utilisé la méthode pour estimer l'exactitude plutôt que pour l'améliorer (directement) est que le rendement nécessaire dans le premier cas est inférieur à celui exigé dans le second cas. Une fois que le biais et la variance des estimations (données) sont obtenus, on peut soit conclure que ces estimations sont suffisamment exactes pour être publiées, soit viser d'abord à améliorer l'exactitude des estimations. Dans ce dernier cas, on pourrait d'abord appliquer la vérification des données pour réduire le taux de classifications erronées. Par ailleurs, le modèle de mélange pourrait servir à construire une estimation améliorée, si le modèle est considéré comme suffisamment fiable.

Nous avons comparé la mesure dans laquelle il est possible d'estimer le biais et la variance de statistiques en présence d'une classification erronée au moyen de la méthode bootstrap EM, de la méthode bootstrap et de la méthode bootstrap SIMEX en utilisant des ensembles de données simulées et des applications réelles. Les données simulées concernaient des (mélanges de) lois normales, alors que les données réelles concernaient des distributions empiriques. Pour la plupart des conditions mises à l'essai, nous avons constaté que la méthode bootstrap EM dépassait la méthode bootstrap et la méthode bootstrap SIMEX. Le biais

estimé des statistiques de la méthode bootstrap EM était plus proche du vrai biais que celui de la méthode bootstrap et de la méthode bootstrap SIMEX. La variance estimée des statistiques fondée sur la méthode bootstrap EM était également plus exacte que celle des autres méthodes, mais dans ce cas, les différences relatives entre les valeurs estimées des trois méthodes étaient plus petites que pour le biais. C'est seulement dans les situations où les moyennes des deux distributions étaient très rapprochées et où les populations étaient petites que la statistique $\hat{\sigma}_1$ était mieux estimée par la méthode SIMEX. L'estimation du biais du total demeurait toutefois exacte.

D'après les résultats obtenus, nous nous attendons à ce que la méthode bootstrap EM offre un meilleur rendement que la méthode bootstrap pour une variable binaire, à condition que les paramètres du modèle soient bien estimés et que la distribution de y soit bien représentée. Nous avons constaté que l'algorithme EM avait de la difficulté à bien estimer les paramètres du modèle seulement quand les deux distributions avaient des moyennes rapprochées et que la taille de la population était petite. Néanmoins, malgré ces conditions difficiles, le biais du total était bien estimé. Nous nous attendons à ce que la combinaison de moyennes de classes similaires et de grands écarts-types pose également plus de problèmes. La méthode SIMEX fonctionne bien seulement quand le paramètre d'intérêt est une fonction lisse de λ qui peut être extrapolée correctement. Comme notre méthode d'estimation ne dépend pas de ces conditions, nous nous attendons à ce qu'elle donne de meilleurs résultats que la méthode SIMEX pour les variables de classification binaire. Quand la distribution de y est très asymétrique et contient un certain nombre de valeurs aberrantes, l'algorithme SIMEX peut alors donner de meilleures résultats que l'algorithme EM.

Les résultats ci-dessus ont été obtenus en faisant la moyenne de plus de 100 exécutions de simulation. En comparant les intervalles de confiance de 95 % du biais estimé dans l'étude par simulations, nous avons montré que même dans une situation pratique où il n'y a qu'*un seul ensemble* de $\hat{z}$, la méthode bootstrap EM mène à de meilleures estimations du biais que la méthode bootstrap et la méthode bootstrap SIMEX. Nous avons obtenu ces résultats en utilisant les probabilités d'erreur de classification estimées par l'algorithme EM comme données d'entrée pour la méthode bootstrap et la méthode bootstrap SIMEX. Nous avons ainsi donné à la méthode bootstrap et à la méthode bootstrap SIMEX la meilleure position de départ possible pour faire concurrence à la méthode bootstrap EM. Comme solution de rechange, nous avons aussi utilisé les probabilités estimées d'un échantillon de vérification comme données d'entrée pour les méthodes bootstrap et bootstrap SIMEX, ce qui a donné des résultats moins bons qu'auparavant (non indiqués ici).

En plus de donner des estimations plus exactes du biais et de la variance, la méthode bootstrap EM présente deux autres avantages par rapport aux méthodes bootstrap et bootstrap SIMEX. Le premier avantage est que le bootstrap EM n'exige pas que les probabilités d'erreur de classification p_{11} et p_{00} soient estimées avec exactitude au préalable. De fait, l'algorithme EM fournit des estimations de ces probabilités dans ses données de sortie. Dans des études antérieures, ces probabilités de classification erronée ne pouvaient être obtenues qu'à partir d'échantillons de vérification (Van Delden et coll., 2016). Les résultats

de Li (2020a) et de Li (2020b) indiquent que le bootstrap EM fonctionne aussi bien sans échantillon de vérification, sauf dans certains cas extrêmes (où deux classes ont la même moyenne). Quand aucun échantillon de vérification n'est disponible, il faut utiliser plusieurs valeurs de départ pour éviter de trouver un maximum local. Si un échantillon de vérification est disponible, il est utile de l'intégrer à l'algorithme EM (Li, 2020a). Dans notre étude, les valeurs de départ obtenues à partir de l'échantillon de vérification étaient suffisantes pour obtenir des estimations exactes du biais et de la variance. Le deuxième avantage de la méthode bootstrap EM est que sa sortie comprend les probabilités propres à l'unité $P(z_i = 1 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{q_1} A_{1ji}$ et $P(z_i = 0 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{q_0} A_{0ki}$, qui peuvent servir à prédire la probabilité d'une erreur de classification dans $\hat{z}_i$ pour chaque unité de l'ensemble de données. Ce résultat peut ensuite servir à vérifier et corriger manuellement les unités susceptibles d'avoir un code incorrect de façon efficace. Cela permettrait d'économiser beaucoup de temps et d'efforts comparativement au simple fait de vérifier toutes les unités dans une (sous) population à partir de p_{11} et p_{00}.

Deux éléments sont à prendre en considération dans l'utilisation pratique de la méthode bootstrap EM. Premièrement, la méthode bootstrap EM suppose que la distribution réelle peut être décrite au moyen d'un modèle de mélange gaussien. Selon McLachlan et Peel (2000), un modèle de mélange gaussien peut prendre en charge diverses distributions pour la variable continue. Dans les études de cas, nous avons constaté qu'il était possible de calculer approximativement les distributions empiriques avec deux ou trois composantes. Nous ne pensons pas que l'hypothèse du mélange gaussien présente de grandes limites à son application dans la pratique, bien qu'il faille préciser que nous n'avons pas testé notre méthode dans des situations qui nécessitent plus de trois composantes gaussiennes par classe. De plus, dans nos études de cas, nous avons déterminé le nombre de composantes par classe en utilisant le critère sBIC sur les vraies données, alors qu'en pratique, il faudrait le déterminer en appliquant la méthode sur des données comportant des classifications erronées. Les résultats de la section 4.3 laissent entendre qu'en cas de données asymétriques, il pourrait être avantageux en pratique de privilégier l'inclusion d'un trop grand nombre de composantes plutôt que d'un trop petit nombre. Ces questions pourraient faire l'objet d'études plus approfondies.

Deuxièmement, il faudra soigneusement déterminer le nombre d'itérations des deux boucles, car il influera sur l'exactitude des estimations du biais et de la variance de la méthode bootstrap EM. (Cela vaut également pour les deux autres méthodes.) Dans notre étude, nous avons utilisé un nombre fixe d'itérations dans les expériences. À la lumière des résultats, nous avons jugé que nous avions effectué assez d'itérations pour tirer des conclusions valides. En pratique cependant, le nombre d'itérations doit être ajusté selon les propriétés des ensembles de données, comme la taille, la distribution de chaque classe, la distance entre les deux classes, etc. Efron et Tibshirani (1993) présentent quelques éléments dont il faut tenir compte quand on choisit le nombre d'itérations dans un algorithme bootstrap; par exemple, il faut généralement plus d'itérations pour une plus petite taille de population (*op. cit.*, section 6.4). De plus, il a été démontré que pour des algorithmes emboîtés similaires, le nombre d'itérations dans la boucle « for » externe influe le plus

sur la convergence (Chang et Hall, 2015), ce qui indique que l'augmentation de S_1 dans l'algorithme 2 est plus bénéfique que l'augmentation de S_2. Toutefois, comme chaque application est différente, il est recommandé d'examiner la convergence des estimations bootstrap, par exemple en représentant graphiquement les résultats intermédiaires par rapport au nombre d'itérations pour vérifier le moment auquel ils deviennent suffisamment stables. Enfin, si l'on veut seulement estimer le biais, et non la variance, la boucle bootstrap secondaire de la méthode bootstrap EM n'est pas nécessaire, ce qui permet de gagner du temps de calcul; voir la « méthode EM » dans Li (2020b).

Il serait utile d'étudier plusieurs extensions de notre approche dans de futures travaux de recherche. Une première extension importante consisterait à généraliser l'estimation du biais et de la variance à une situation au moyen d'une variable de classification ayant $D \geq 2$ classes. En présence de D classes, les probabilités de classification erronée seront données par une matrice $D \times D$ $\mathbf{P}$ ayant $D(D-1)$ probabilités à estimer. Pour chaque domaine supplémentaire $d \in \{1, \ldots, D\}$, le nombre d'autres paramètres dans le modèle de mélange augmente linéairement selon $(q_d - 1) + 2q_d + 1 = 3q_d$ où q_d est le nombre de composantes par domaine supplémentaire d. Puisque le nombre de paramètres augmente avec le nombre de domaines, l'échantillon de vérification gagne en importance pour donner au modèle des valeurs de départ raisonnables. Il est également important de vérifier par un essai si le bootstrap EM de D classes converge bien. De plus, il faut vérifier par un essai si le modèle de mélange donne de meilleurs résultats que les méthodes bootstrap et SIMEX existantes pour $D > 2$.

Une deuxième extension consisterait à tenir compte de données d'échantillon plutôt que de données de recensement. Dans ce cas, la qualité de la sortie est influencée par l'erreur de classification et l'erreur d'échantillonnage. Dans le cas d'un échantillonnage aléatoire simple, le biais estimé n'est pas touché par l'erreur d'échantillonnage. Concernant la variance, on pourrait utiliser deux approches. La première est que la procédure d'échantillonnage est également par bootstrap parce qu'elle est incluse dans la procédure bootstrap EM. Une autre méthode serait une approche hybride dans laquelle l'estimation bootstrap EM est utilisée pour les erreurs de classification, tandis qu'une expression analytique est utilisée pour l'erreur d'échantillonnage.

Une troisième extension consisterait à assouplir les hypothèses pour les probabilités d'erreurs de classification. Dans notre étude, les probabilités de faire des erreurs de classification, p_{11} et p_{00}, sont supposées indépendantes de la variable continue. Cependant, dans des situations réelles, cette hypothèse ne tient pas toujours, du moins pas sans qu'elle soit conditionnée à d'autres covariables. Il serait donc intéressant de réaliser une extension dans laquelle les probabilités de classification erronée dépendent de covariables.

Une quatrième extension consisterait à utiliser plusieurs variables numériques cibles, comme la taille et le poids de patients dans des dossiers médicaux. Il y aurait ensuite plus d'une variable cible dans le modèle général. Un modèle multivarié de mélange gaussien peut être un modèle approprié dans ce cas (McLachlan

et Peel, 2000). Une cinquième extension possible consisterait à prendre en compte les valeurs manquantes dans la ou les variables cibles.

Enfin, nous remarquons que l'algorithme 1 de notre étude est une méthode bootstrap unique standard. Dans la littérature, des bootstraps doubles et à plusieurs degrés ont également été proposés comme moyen de corriger le biais dans les estimateurs du biais et de la variance à partir d'un seul bootstrap (Chang et Hall, 2015; Hall et Martin, 1988). Ces méthodes ne nécessitent pas de modèle explicite pour les données, mais comme le bootstrap unique, elles nécessitent (une estimation) de la matrice $\mathbf{P}$ comme entrée. Il serait intéressant de comparer le rendement de la méthode bootstrap EM et celui d'un bootstrap double dans une future étude, en particulier pour le problème étendu ayant $D \gg 2$ classes ou plusieurs variables cibles numériques, pour lesquelles l'estimation d'un modèle de mélange gaussien peut devenir difficile.

Remerciements

Les opinions exprimées dans l'article sont celles des auteurs et ne correspondent pas nécessairement aux politiques du Bureau central de la statistique des Pays-Bas. Les auteurs aimeraient remercier le rédacteur associé et les deux examinateurs anonymes pour leurs commentaires constructifs et détaillés concernant une version préliminaire du présent article.

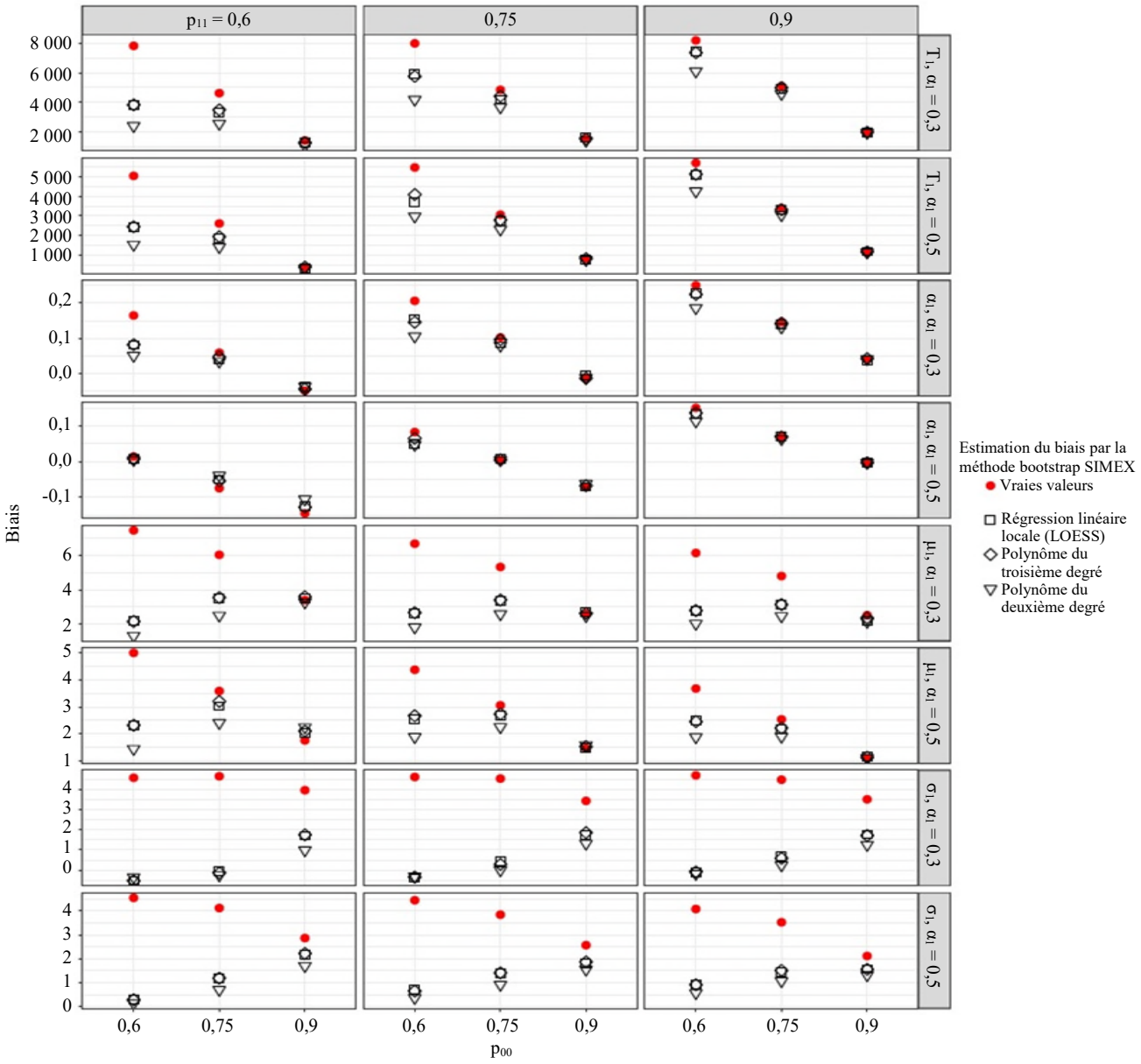
Annexe

A. Fonction d'extrapolation dans la méthode bootstrap SIMEX

Dans la présente annexe, nous comparons trois fonctions d'extrapolation utilisées dans la méthode bootstrap SIMEX, soit une régression par polynômes locaux (LOESS), une régression par polynômes du second degré (poly2) et une régression par polynômes du troisième degré (poly3). Pour la fonction LOESS, nous avons utilisé les paramètres par défaut de la fonction LOESS dans R (span = 0,75 et nls.control (nombre maximal d'itérations dans l'algorithme = 1000)).

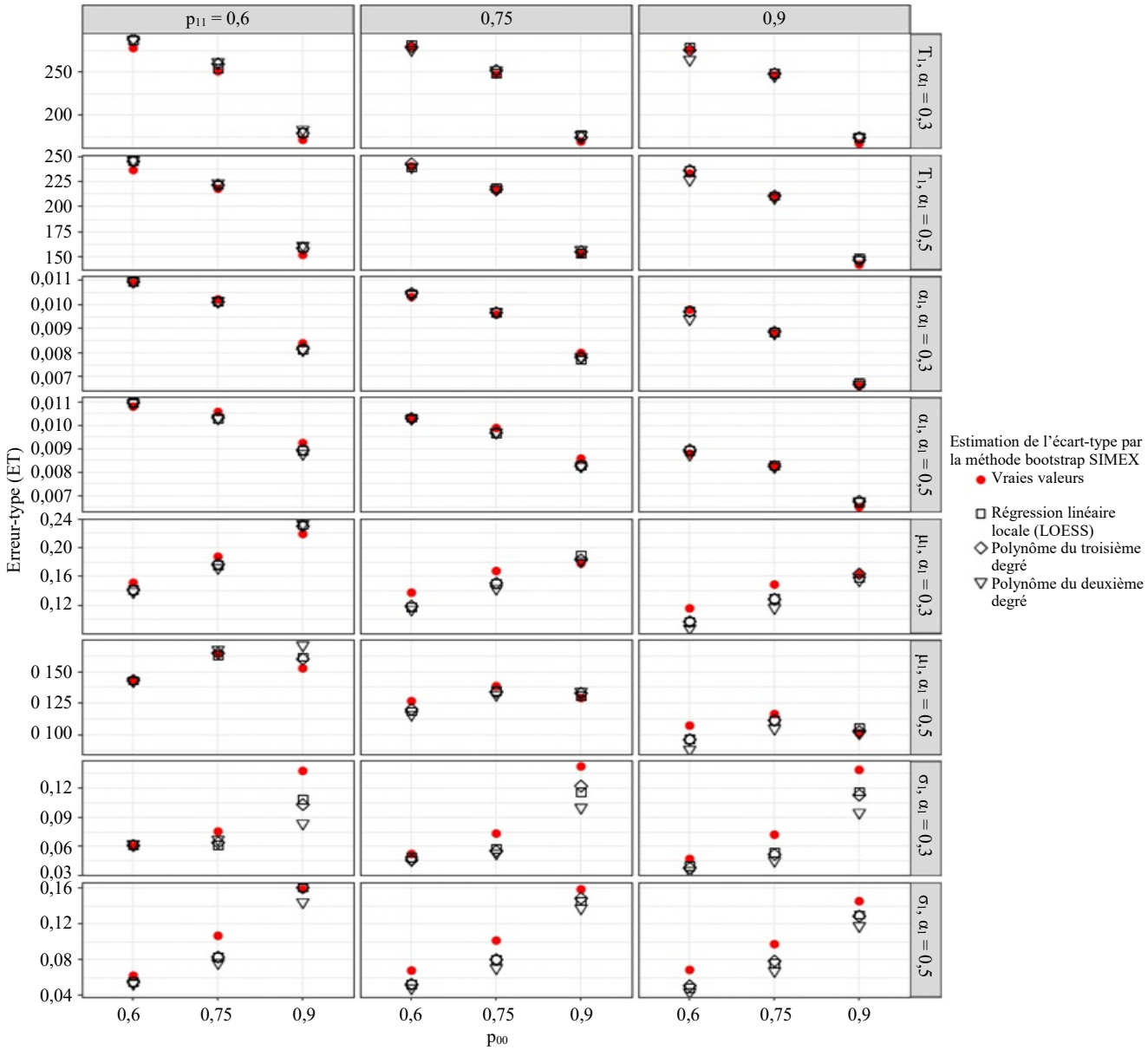
Les estimations du biais et de l'erreur-type de la fonction LOESS et de la régression par polynômes du troisième degré étaient semblables; voir les figures A.1 et A.2. Les estimations des deux modèles étaient plus proches des vraies valeurs que celles de la régression par polynômes du second degré. Étant donné que la fonction LOESS avait été utilisée dans une étude précédente sur les classifications erronées (Hopkins et King, 2010), nous avons décidé d'appliquer la fonction LOESS dans le présent article.

Figure A.1 Comparaison du rendement de l'estimation du biais des trois fonctions d'extrapolation dans la méthode bootstrap SIMEX.



Notes : Les lignes représentent différents paramètres de domaine (T_1, α_1, μ_1, σ_1), pour deux configurations de α_1 : $\alpha_1 = 0{,}3$ et $\alpha_1 = 0{,}5$. Les colonnes représentent différentes valeurs de p_{11}; différentes valeurs de p_{00} sont données dans chaque colonne.

Figure A.2 Comparaison du rendement de l'estimation de l'erreur-type des trois fonctions d'extrapolation dans la méthode bootstrap SIMEX.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure A.1.

B. Critère d'information bayésien BIC comparativement au critère BIC modifié sBIC

Avant d'ajuster le modèle de mélange gaussien, il faut sélectionner son nombre de composantes. Dans la présente annexe, nous comparons les résultats de l'estimation du biais et de l'erreur-type quand nous utilisons le critère BIC par rapport au critère sBIC comme critère de sélection du nombre de composantes. Le critère sBIC est plus justifié quand les composantes du modèle de mélange ont des moyennes et une variance très semblables; voir la section 2.2.1.

Le tableau B.1 présente le nombre optimal de composantes qui a été sélectionné selon les critères BIC et sBIC. Les deux critères donnent le même nombre de composantes pour les deux classes du cas 1 et pour la classe 1 du cas 2. Pour la classe 0 du cas 2, le nombre optimal de composantes qui a été sélectionné selon le critère sBIC était deux et celui selon le critère BIC était un. Par conséquent, pour le cas 2 seulement, nous avons comparé les estimations du biais et de l'erreur-type, en utilisant deux composantes (sBIC) ou une composante (BIC) pour la classe 0.

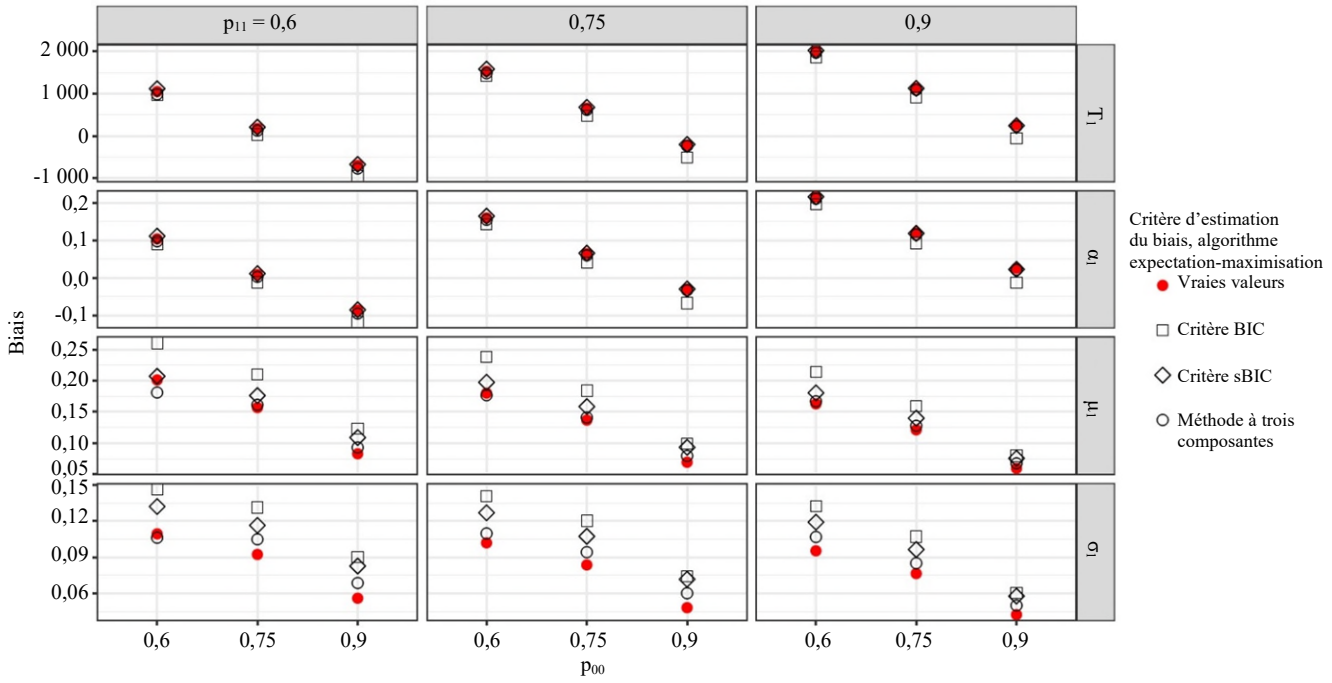
Tableau B.1
Nombre optimal de composantes qui a été sélectionné selon le critère d'information bayésien BIC et le critère BIC modifié sBIC.

Nº du cas	Classe	Nombre optimal	
		Critère BIC	Critère sBIC
Cas 1	Classe 1	1	1
	Classe 0	2	2
Cas 2	Classe 1	2	2
	Classe 0	1	2

L'utilisation de deux composantes pour la classe 0 a permis d'obtenir des estimations du biais plus exactes que l'utilisation d'une seule composante; voir la figure B.1. Pour certains paramétrages, les estimations des erreurs-types fondées sur deux composantes étaient également plus exactes que celles fondées sur une composante (voir la figure B.2), même si les différences d'exactitude étaient plus petites pour l'erreur-type que pour le biais. Nous avons donc décidé d'utiliser le critère sBIC dans notre étude de cas pour estimer le nombre optimal de composantes.

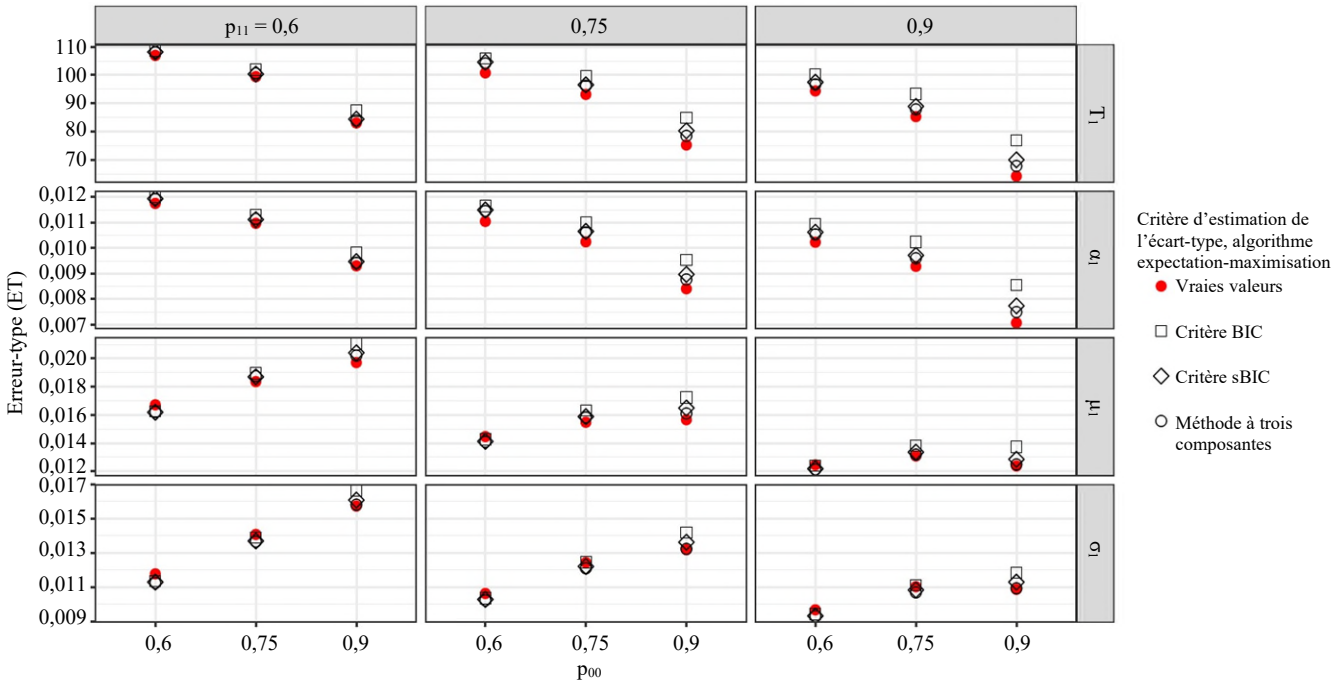
Pour le cas 2, nous avons également essayé un modèle comportant trois composantes au lieu de deux dans les deux classes. Les estimations du biais et les erreurs-types qui en résultent sont présentées dans les figures B.1 et B.2 (« Méthode à trois composantes »). Comme cela a été mentionné dans la section 4.3, nous constatons que l'augmentation du nombre de composantes à trois permet d'obtenir des résultats plus exacts.

Figure B.1 Estimation du biais au moyen du critère d'information bayésien BIC par rapport à celle au moyen du critère BIC modifié sBIC dans le cas 2.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure A.1.

Figure B.2 Estimation de l'erreur-type au moyen du critère d'information bayésien BIC par rapport à celle au moyen du critère BIC modifié sBIC dans le cas 2.



Notes : Pour obtenir la configuration des lignes et des colonnes, voir la figure A.1.

C. Propriétés théoriques des algorithmes 1 et 2

C.1 Algorithme 1 : la méthode bootstrap

En général, l'algorithme 1 produit des estimations biaisées du biais et de la variance de $\hat{\zeta}$. Nous illustrerons cela au moyen du total estimé du domaine $\hat{T}_1$. Notons que les résultats ci-dessous s'appliquent aussi à $\hat{\alpha}_1$, car il est obtenu comme un cas particulier de $\hat{T}_1$ avec $y_i \equiv 1/N$.

Pour $\hat{T}_1$, on peut calculer que son vrai biais est égal à

$$\text{Biais}(\hat{T}_1) = (1-p_{00})\,T_0 + (p_{11}-1)\,T_1, \tag{C.1}$$

tandis que, pour $S \to \infty$, l'estimateur du biais bootstrap de l'algorithme 1 converge vers $(1-p_{00})\,\hat{T}_0 + (p_{11}-1)\,\hat{T}_1$; voir, par exemple, Burger, van Delden et Scholtus (2015). Ainsi, en général, l'estimateur du biais bootstrap est biaisé à moins que $\hat{T}_1$ lui-même ne soit un estimateur sans biais de T_1. Cette dernière situation ne se produit que pour des combinaisons particulières de (p_{11}, p_{00}) (Kloos et coll., 2021).

De même, on peut calculer que la vraie variance de $\hat{T}_1$ est

$$\text{Var}(\hat{T}_1) \;=\; p_{00}(1-p_{00})\,K_0 + p_{11}(1-p_{11})\,K_1, \tag{C.2}$$

avec $K_1 = \sum_{i=1}^{N} z_i y_i^2$ et $K_0 = \sum_{i=1}^{N}(1-z_i)\,y_i^2$, tandis que l'estimateur bootstrap de la variance converge vers la même expression, K_1 et K_0 étant remplacés respectivement par $\hat{K}_1 = \sum_{i=1}^{N} \hat{z}_i y_i^2$ et $\hat{K}_0 = \sum_{i=1}^{N}(1-\hat{z}_i)\,y_i^2$ (Burger et coll., 2015). Notons qu'en général, quand $T_1 \neq \alpha_1$, les conditions dans lesquelles l'estimateur du biais bootstrap et l'estimateur de la variance bootstrap sont sans biais ne sont pas équivalentes.

C.2 Algorithme 2 : la méthode bootstrap espérance-maximisation

Comme l'indique la section 2.3.2, le but de générer des valeurs 0-1 $\tilde{z}_i$ dans la boucle « for » externe de l'algorithme 2, avec $P(\tilde{z}_i = 1 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}}) = P(z_i = 1 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{q_1} A_{1ji}$, est de corriger les estimateurs bootstrap du biais et de la variance pour le biais qui se produit dans l'algorithme 1. Pour illustrer l'idée sous-jacente, nous montrerons ici que dans le cas du total de domaine estimé $\hat{T}_1$, pour lequel des expressions analytiques exactes sont disponibles, la correction de biais est effectivement réalisée par $\mathbf{Biais_{comb}}$ et $\mathbf{Var_{comb}}$ à condition que le modèle de mélange supposé se vérifie. Une version simplifiée de la dérivation ci-dessous peut être donnée pour la proportion estimée $\hat{\alpha}_1$.

Notons $\tilde{T}_1 = \sum_{i=1}^{N} \tilde{z}_i y_i$ et $\tilde{T}_0 = \sum_{i=1}^{N}(1-\tilde{z}_i)\,y_i$; notons aussi $\tilde{T}_1^* = \sum_{i=1}^{N} \tilde{z}_i^* y_i$ et $\tilde{T}_0^* = \sum_{i=1}^{N}(1-\tilde{z}_i^*)\,y_i$. Nous pouvons calculer d'une manière analogue à (C.1) que $\text{E}(\tilde{T}_1^* - \tilde{T}_1 \mid \hat{z}, \tilde{z}) = (1-p_{00})\,\tilde{T}_0 + (p_{11}-1)\,\tilde{T}_1$. Par conséquent, pour $S_1, S_2 \to \infty$, la valeur espérée de l'estimateur du biais dans la méthode bootstrap EM est

$$\text{E}\left\{\mathbf{Biais_{comb}}(\hat{T}_1)\right\} \;=\; (1-p_{00})\,\text{E}\left\{\text{E}_{\tilde{z}}(\tilde{T}_0 \mid \hat{z})\right\} + (p_{11}-1)\,\text{E}\left\{\text{E}_{\tilde{z}}(\tilde{T}_1 \mid \hat{z})\right\}. \tag{C.3}$$

De plus, nous constatons que

$$\text{E}_{\tilde{z}}(\tilde{T}_1 \mid \hat{z}) \;=\; \sum_{i=1}^{N} \text{E}(\tilde{z}_i \mid \hat{z}_i)\,y_i \;=\; \sum_{i=1}^{N} P(z_i = 1 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}})\,y_i \;=\; \sum_{i=1}^{N}\sum_{j=1}^{q_1} A_{1ji} y_i = N\hat{\alpha}_1 \sum_{j=1}^{q_1} \hat{\xi}_{1j}\hat{\mu}_{1j},$$

où la dernière expression découle des formules appliquées à l'étape M de l'algorithme EM (voir la section 2.2.2). En supposant que le modèle de mélange se vérifie, il s'ensuit que

$$\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}(\tilde{T}_1 \mid \hat{z})\right\} \;=\; \mathrm{E}\left(N\hat{\alpha}_1 \sum_{j=1}^{q_1} \hat{\xi}_{1j}\hat{\mu}_{1j} \right) \;\approx\; N\alpha_1 \sum_{j=1}^{q_1} \xi_{1j}\mu_{1j} \;=\; N\alpha_1\mu_1 \;=\; T_1;$$

Voir la note 2 du tableau 2.1 pour les deux dernières égalités. De même, on peut démontrer que $\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}(\tilde{T}_0 \mid \hat{z})\right\} \approx T_0$ si le modèle se vérifie. En substituant les deux résultats dans (C.3) et en se rappelant (C.1), nous concluons que $\mathrm{E}\left\{\mathbf{Biais}_{\mathbf{comb}}(\hat{T}_1)\right\} \approx \mathrm{Biais}(\hat{T}_1)$.

Pour l'estimateur de la variance, nous pouvons procéder de la même façon. Notons $\tilde{K}_1 = \sum_{i=1}^{N} \tilde{z}_i y_i^2$ et $\tilde{K}_0 = \sum_{i=1}^{N} (1-\tilde{z}_i)\, y_i^2$. De manière analogue à (C.2), nous pouvons démontrer que $\mathrm{Var}(\tilde{T}_1^* \mid \hat{z}, \tilde{z}) = p_{00}(1-p_{00})\,\tilde{K}_0 + p_{11}(1-p_{11})\,\tilde{K}_1$. Par conséquent, pour $S_1, S_2 \to \infty$,

$$\mathrm{E}\left\{\mathbf{Var}_{\mathbf{comb}}(\hat{T}_1)\right\} = p_{00}(1-p_{00})\,\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}(\tilde{K}_0 \mid \hat{z})\right\} + p_{11}(1-p_{11})\,\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}(\tilde{K}_1 \mid \hat{z})\right\}. \tag{C.4}$$

À partir des formules appliquées dans l'algorithme EM, il s'ensuit que

$$\mathrm{E}_{\tilde{z}}(\tilde{K}_1 \mid \hat{z}) \;=\; \sum_{i=1}^{N} \mathrm{E}(\tilde{z}_i \mid \hat{z}_i)\, y_i^2 \;=\; \sum_{i=1}^{N} P(z_i = 1 \mid y_i, \hat{z}_i; \hat{\boldsymbol{\theta}})\, y_i^2 \;=\; \sum_{i=1}^{N}\sum_{j=1}^{q_1} A_{1ji}\, y_i^2 \;=\; N\hat{\alpha}_1 \sum_{j=1}^{q_1} \hat{\xi}_{1j}(\hat{\sigma}_{1j}^2 + \hat{\mu}_{1j}^2).$$

Par conséquent, en supposant que le modèle de mélange se vérifie, nous obtenons ce qui suit :

$$\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}(\tilde{K}_1 \mid \hat{z})\right\} \;\approx\; N\alpha_1 \sum_{j=1}^{q_1} \xi_{1j}(\sigma_{1j}^2 + \mu_{1j}^2) \;=\; N\alpha_1(\sigma_1^2 + \mu_1^2) \;=\; K_1.$$

De la même façon, nous pouvons en déduire que $\mathrm{E}\left\{\mathrm{E}_{\tilde{z}}(\tilde{K}_0 \mid \hat{z})\right\} \approx K_0$. Par conséquent, nous pouvons voir en utilisant (C.2) que $\mathrm{E}\left\{\mathbf{Var}_{\mathbf{comb}}(\hat{T}_1)\right\} \approx \mathrm{Var}(\hat{T}_1)$ si le modèle de mélange se vérifie.

Bibliographie

Bakker, B.F.M., Van Rooijen, J. et Van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, 30, 411-424.

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics*, 10, 478-486.

Buonaccorsi, J.P. (2010). *Measurement Error: Models, Methods and Applications*. Chapman and Hall/CRC Press.

Burger, J., van Delden, A. et Scholtus, S. (2015). Sensitivity of mixed-source statistics to classification errors. *Journal of Official Statistics*, 31(3), 489-506.

Chang, J., et Hall, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika*, 102, 203-214.

Cook, J.R., et Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428), 1314-1328.

Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22.

Drton, M., et Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society*, 79(2), 323-380.

Edwards, J.K., Bakoyannis, G., Yiannoutsos, C.T., Mburu, M.W. et Cole, S.R. (2019). Non-parametric estimation of the cumulative incidence function under outcome misclassification using external validation data. *Statistics in Medicine*, 38(29), 5512-5527.

Edwards, J.K., Cole, S.R. et Fox, M.P. (2020). Flexibly accounting for exposure misclassification with external validation data. *American Journal of Epidemiology*, 189(8), 850-860. Récupéré de https://doi.org/10.1093/aje/kwaa011 doi: 10.1093/aje/kwaa011.

Efron, B., et Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Londres: Chapman and Hall/CRC.

Eurostat (2008). *NACE Rev.2 Statistical Classification of Economic Activities in the European Community*. Récupéré de https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2.

Eurostat (2009). *ESS Handbook for Quality Reports* (Rapport technique). Office for Official Publications of the European Communities, Methodologies and Working papers. Récupéré de https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-08-016.

Eurostat (2022). *Table: In-Work at-Risk-of-Poverty Rate by Educational Attainment Level - EU-SILC survey*. Récupéré de https://ec.europa.eu/eurostat/databrowser/view/ilciw04/default/table?lang=en.

Gravel, C.A., et Platt, R.W. (2018). Weighted estimation for confounded binary outcomes subject to misclassification. *Statistics in Medicine*, 37(3), 425-436.

Hall, P., et Martin, M.A. (1988). On bootstrap resampling and iteration. *Biometrika*, 75, 661-671.

Hopkins, D.J., et King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1), 229-247.

Keogh, R.G., Shaw, P.A., Gustafson, P., Carroll, R.J., Deffner, V., Dodd, K.W., Küchenhoff, H., Tooze, J.A., Wallace, M.P., Kipnis, V. et Freedman, L.S. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1-Basic theory and simple methods of adjustments. *Statistics in Medicine*, 39(16), 2197-2231. doi: 10.1002/sim.8532.

Kloos, K., Meertens, Q., Scholtus, S. et Karch, J. (2021). Comparing correction methods to reduce misclassification bias. Dans *Proceedings of BNAIC/BeneLearn*, (Éds., L. Cao, W.A. Kosters et J. Lijffijt), 103-129. Springer, Leiden. Récupéré de https://bnaic.liacs.leidenuniv.nl/wordpress/wp-content/uploads/papers/BNAICBENELEARN_2020_Final_paper_64.pdf.

Kooiman, P., Willenborg, L. et Gouweleeuw, J. (1997). *PRAM: A Method for Disclosure Limitation of Microdata*. Document de recherche nº 9705. Voorburg/Heerlen: Statistics Netherlands.

Kosinski, A.S., et Flanders, W.D. (1999). Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: A regression approach. *Statistics in Medicine*, 18(20), 2795-2808.

Küchenhoff, H., Mwalili, S.M. et Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62(1), 85-96.

Li, Y. (2020a). *Bias Correction for Classification Errors*. Récupéré de https://www.researchgate.net/publication/362862838_Bias_Correction_for_Classification_Errors (Internship Report, Leiden University).

Li, Y. (2020b). *Estimating the Effect of Classification Errors on Domain Statistics* (Thèse de doctorat, Leiden University). Récupéré de https://hdl.handle.net/1887/3280845.

Little, R.J., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2e éd., Vol. 793). New York: John Wiley & Sons, Inc.

Magnusson, P., Palm, A., Branden, E. et Mörner, S. (2017). Misclassification of hypertrophic cardiomyopathy: Validation of diagnostic codes. *Clinical Epidemiology*, 9, 403.

McLachlan, G.J., et Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

Meertens, Q., Diks, C., Van den Herik, H. et Takes, F. (2020). A data-driven supply-side approach for estimating cross-border internet purchases within the European Union. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1), 61-90.

Nations Unies (2015). *Guidelines on Statistical Business Registers*. New York et Genève: Nations Unies.

Oosterveen, V. (2020). *Notice the Noise: Detecting Misclassifications in Register Data* (Thèse de doctorat, Utrecht University, les Pays-Bas). Récupéré de https://www.researchgate.net/publication/354533688_Notice_the_noise_detecting_misclassifications_in_register_data.

Rousseeuw, P.J., et Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424), 1273-1283.

Schwartz, J.E. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods and Research*, 13(4), 435-466.

Selén, J. (1986). Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association*, 81, 75-81.

Shaw, P.A., Gustafson, P., Carroll, R.J., Deffner, V., Dodd, K.W., Keogh, R.H., Kipnis, V., Tooze, J.A., Wallace, M.P., Küchenhoff, H. et Freedman, L.S. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2-More complex methods of adjustment and advanced topics. *Statistics in Medicine*, 39(16), 2232-2263. doi: 10.1002/sim.8531.

Sinclair, D.G., et Hooker, G. (2017). An Expectation Maximization algorithm for high-dimensional model selection for the Ising model with misclassified states. *arXiv preprint arXiv:1704.05995*. Récupéré de https://arxiv.org/abs/1704.05995.

Van Delden, A., Scholtus, S. et Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, 32(3), 619-642. Récupéré de https://content.sciendo.com/view/journals/jos/32/3/article-p619.xml doi: https://doi.org/10.1515/jos-2016-0032.

Van Delden, A., Scholtus, S., Burger, J. et Meertens, Q.A. (2023). Accuracy of estimated ratios as affected by dynamic classification errors. *Journal of Survey Statistics and Methodology*, 11(4), 942-966. doi: https://doi.org/10.1093/jssam/smac015.

Van den Brakel, J., et Bethlehem, J. (2008). *Model-Based Estimation for Official Statistics* (Rapport technique). Récupéré de https://www.cbs.nl/nl-nl/achtergrond/2008/10/model-based-estimation-for-official-statistics.

Van den Hout, A., et Van der Heijden, P.G.M. (2002). Randomised response, statistical disclosure control and misclassification: A review. *Revue Internationale de Statistique*, 70(2), 269-288.

Zhang, L.C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, 27(3), 415-432.