

DONNÉES ABERRANTES OU DONNÉES ENTACHÉES D'ERREURS ? D'UN EXEMPLE «JOUET» VERS UNE MÉTHODOLOGIE UNIFIÉE !

Abdellatif Demnati ¹

RÉSUMÉ

Le problème de données aberrantes comporte deux étapes: la détection et le traitement. La détection consiste à classifier les unités selon deux groupes dont un groupe contiendrait les unités ayant été entachées par des valeurs aberrantes ou grossières. La variable de classification est bien entendu non observée, autrement l'étape de détection serait résolue. Une fois les unités aberrantes détectées, un traitement déterministe ou stochastique s'ensuit. La détection et le traitement des valeurs aberrantes fait partie d'un problème plus vaste qui consiste à reconstruire des variables non observées à partir de variables observées.

Comme le problème des données aberrantes est complexe autant du point de vue pratique que théorique, il est important d'évaluer constamment de nouvelles approches. Ce document présente une approche envisagée par la Division des données régionales et administratives de Statistique Canada. Une variable multivariée est considérée pour des raisons évidentes : nous voulons traiter des données longitudinales et nous voulons profiter de l'interrelation entre les variables pour améliorer les estimations. La méthodologie tient alors compte des données manquantes inhérentes aux données multivariées. En s'appuyant sur un exemple «jouet», ce document passe en revue la méthodologie basée sur une approche d'un modèle de mélange de distributions pour la détection des valeurs aberrantes. Le document porte une attention particulière à l'étape d'optimisation. Le maniement de données multivariées ajoute une autre complexité : il s'agit non seulement d'identifier les unités possédant au moins une valeur aberrante mais d'identifier les variables contaminées. Une approche basée sur un modèle de données entachées d'erreurs est envisagée pour l'identification des variables contaminées ainsi que pour le traitement des données aberrantes.

MOTS CLÉS : Variable manquante; problème inverse; mélange de distributions; algorithme EM; algorithme de recuit simulé; variable multivariée; données manquantes.

ABSTRACT

Outlier problems must be dealt with in two steps: detection and treatment. Detection is done by classifying units according to two groups, one of which would contain units affected by outliers. The classification variable is obviously not observed, otherwise the detection step would be solved. Once outliers have been detected, a deterministic or stochastic treatment follows. The detection and treatment of outliers is part of a larger problem which consists of reconstructing non observed variables from observed variables.

Since the problem of outliers is complex from both a practical and theoretical viewpoint, it is important to constantly evaluate new approaches. This document presents an approach considered by the Small Area and Administrative Data Division of Statistics Canada. A multivariate variable is considered for obvious reasons: we want to process longitudinal variables and we want to take advantage of the interrelation between variables to improve estimations. The methodology then takes into account missing variables inherent in multivariate data. Supported by a "toy" example, this document presents the methodology based on a model of mixture of distributions to detect outliers. The document gives special attention to the optimizing step. Handling multivariate data adds another complexity, it's not just about identifying units with at least one outlier values but also it's about identifying contaminated variables. An approach based on errors in variables is proposed for the identification of contaminated variables as well as for the treatment of outliers.

KEY WORDS: Missing variable; Inverse problem; Mixture of distributions; EM algorithm; Simulated Annealing algorithm; Multivariate variable; Missing data.

¹ Division des méthodes d'enquêtes sociales, Statistique Canada, immeuble R.-H.-Coats, 15-G, Ottawa (Ontario), K1A 0T6, demnabd@statcan.ca.

1. INTRODUCTION

« Est-ce-que tous les gens gentils sont gentils ? Est-ce-que tous les gens méchants sont méchants? » me demanda ma fille âgée de cinq ans. «.. Dieu n'a pas mis assez de ressources pour avoir un monde complètement gentil. Mais comment je peux reconnaître les personnes gentilles? et comment je peux reconnaître les personnes méchantes? », continua-t-elle à s'interroger et à m'interroger.

Ces interrogations concernant les gens méchants sont aussi valables pour les données aberrantes.

« Pourquoi n'a-t-on pas mis assez de ressources pour n'avoir que des données gentilles?
Comment peut-on identifier les données gentilles?
Comment peut-on traiter les données méchantes? »

Le problème de données aberrantes comporte deux étapes: la détection et le traitement. La détection consiste à classer les unités selon deux groupes dont un groupe contiendrait les unités ayant été entachées par des valeurs aberrantes ou grossières. La variable de classification est bien entendu non observée, autrement l'étape de détection serait résolue. Une fois les unités aberrantes détectées, un traitement déterministe ou stochastique s'ensuit. La Division des données régionales et administratives de Statistique Canada utilise une méthodologie similaire à la méthode de Tukey (1977). La méthodologie consiste à :

- traiter une seule variable à la fois;
- utiliser des statistiques d'ordre pour la détection;
- répéter la processus dans certains cas afin d'identifier un nombre restreint pour le traitement manuel;
- traiter manuellement les enregistrement identifiés.

Il est souhaitable d'apporter certaines améliorations à la méthodologie existante dont:

- Tenir compte des inter-relations entre les variables. La conformité aux relations entre variables est une indication de la véracité des valeurs observées. Il faut donc traiter des variable multivariées;
- Contrôler le taux de contamination des unités. Il est impensable de tolérer plus qu'un certain nombre d'enregistrements pour des vérifications manuelles;
- Produire une estimation direct ou par intervalle de la vraie valeur non observée;
- Tenir compte des données manquantes inhérentes aux variables multivariées;
- Tenir compte du plan d'échantillonnage.

Chaque approche débute par une définition de la méthodologie et se termine par l'implantation sur des données réelles. Ce document présente l'approche envisagée et discute de l'estimation des éléments inconnus. Certains changements pourront survenir au cours du processus avant l'implantation finale. D'autres méthodes font aussi l'objet d'un examen à Statistique Canada. Nous nous limitons dans cet exposé à une seule approche.

Le problème des valeurs aberrantes fait partie d'un problème plus vaste qui consiste à estimer ou à imputer des quantités non observées ou observées indirectement à partir de variables observées. Ce type de problème est connu aussi sous le nom de problème inverse². La série de données observées est dénotée par Y et la série de données non observées est dénotée par Z .

On suppose que les deux séries de données sont liées par une distribution de la forme :

$$f(Y, Z; \varphi)$$

où $f()$ est une fonction connue et φ représente l'ensemble des paramètres à estimer.

Le problème se pose en deux étapes :

Étape 1 : modélisation de $f(Y, Z; \varphi)$

Étape 2 : estimation de φ et de z

L'étape 1 est constituée d'hypothèses assez fortes. Ces hypothèses relient d'une part les données observées à quelques paramètres inconnus en utilisant un modèle de distribution:

Sous-étape 1.1 : choix de $f(Y/Z; \varphi)$

² Un ensemble assez large de problèmes pratiques implique la reconstruction de données non observées.

Le traitement du signal, comme le traitement de l'image, est présent en communication comme en médecine. Il consiste à reconstituer ou à restaurer des signaux ou des images émis à partir de données obtenues à partir des satellites, d'ultrasons ou d'autres moyens. On émet un signal d'intensité S_z lorsqu'on veut communiquer une valeur z . Le signal émis est une variable aléatoire qui peut être non observée. La variable Y représente les signaux reçus. Cependant la transmission est affectée par des perturbations, dites bruit. Le lecteur est invité à consulter, par exemple, l'article de Valdi et al (1985) qui donne une introduction à l'aspect statistique de la tomographie par des émissions de positrons «PET» (Positron Emission Tomography) qui est une technique médicale de diagnostic. Elle consiste à reconstruire des intensités d'émissions radioactives d'une partie du corps humain à partir d'observations externes obtenus par des émissions de positrons.

et modélisent d'autre part la distribution non observée:

Sous-étape 1.2 : choix de $f(Z;\varphi)$

Ces deux sous-étapes constituent une première façon de décomposer la distribution conjointe:

$$f(Y, Z; \varphi) = f(Y/z; \varphi) f(Z; \varphi)$$

La deuxième façon de décomposer la fonction conjointe est:

$$f(Y, Z; \varphi) = f(Z/y; \varphi) f(Y; \varphi)$$

Dans cette dernière décomposition, la distribution de la variable d'intérêt Y est spécifiée pour la population entière $f(Y; \varphi)$, tandis qu'à la première décomposition, la distribution de la variable d'intérêt est spécifiée pour chaque valeur de la variable non observée $f(Y/z; \varphi)$. De plus, cette modélisation est généralement suivie d'hypothèses d'indépendance³ entre observations,

$$f(Y, Z; \varphi) = \prod_i f(Y_i, Z_i; \varphi)$$

Ce document n'aborde pas cette étape en profondeur. On se contente plutôt d'un exemple «jouet» pour illustrer la méthodologie et montrer d'autres problématiques. L'exemple est présenté à la section 2.

Nous nous intéressons particulièrement à l'étape d'estimation (section 3). Cette étape porte aussi le nom d'étape d'identification ou de reconstruction. Plusieurs approches peuvent être envisagées pour l'estimation des paramètres dont l'approche Bayésien. L'approche Bayésien nécessite une distribution à priori pour (Z, φ) et cherche alors à estimer la distribution a posteriori. Binder (1978) décrit une classe assez générale de mélanges de distributions normales et discute des outils Bayésiens pour la classification. Cependant l'analyse Bayésien conduit à des calculs intraitables à l'exception de quelques cas bien choisis. Dans ce document, nous privilégierons une approche basée sur la notion du maximum de vraisemblance.

L'étape d'estimation englobe l'estimation des paramètres des modèles et des données manquantes. Le problème d'estimation se pose comme suit:

$$\max_{(z, \varphi)} f(z, \varphi; y)$$

³ En traitement de l'image, les pixels voisins sont souvent supposés similaires

$$f(z) = c^{-1} \exp(\beta \sum_{\{s-t\}} \delta(z_s, z_t))$$

où $\sum_{\{s-t\}}$ représente la somme sur les paires de pixels voisins selon une structure de voisinage qu'on définit, $\delta(z_s, z_t) = 1$; si $z_s = z_t$, β représente le degré de dépendance locale et c est une constante de normalisation.

Il arrive que l'estimation de φ et de z se fasse simultanément. Cependant, on procède généralement en deux étapes:

Sous-étape 2.1 : on estime les paramètres du vecteur φ ;

Sous-étape 2.2 : une fois le vecteur φ estimé, on impute des valeurs à Z .

Cette façon est la plus utilisée à cause de la quantité de données à traiter et de la quantité assez substantielle d'opérations pour achever l'optimisation nécessaire. En général, les deux difficultés majeures de la sous-étape 2.1 sont l'identifiabilité des distributions et l'opération d'intégration, tandis que les deux difficultés majeures de la sous-étape 2.2 viennent du fait que le domaine Z est large et que la fonction à optimiser est rarement convexe.

Il n'en reste pas moins que c'est une approximation de l'algorithme itératif suivant :

- étant donné $(Z^{(c)}, \varphi^{(c)})$, on maximise $f(z^{(c)}, \varphi; y)$ par rapport à φ pour obtenir $\varphi^{(c+1)}$
- étant donné $(Z^{(c)}, \varphi^{(c+1)})$, on maximise $f(z, \varphi^{(c+1)}; y)$ par rapport à z pour obtenir $z^{(c+1)}$

Cet algorithme a, par construction, la propriété de converger vers un maximum,

$$f(z^{(c+1)}, \varphi^{(c+1)}; y) \geq f(z^{(c)}, \varphi^{(c)}; y)$$

Quelque soit la méthodologie utilisée, il est souhaitable qu'elle puisse être appliquée aux données multivariées comme des données longitudinales. L'inter-relation entre variables ne peut qu'améliorer la détection et le traitement des valeurs aberrantes ainsi que l'estimation des données manquantes. Il est aussi souhaitable que la méthodologie tienne compte des données manquantes inhérents aux données multivariées. Une fois les unités possédant au moins une valeur aberrante détectées, il reste deux complexités : il s'agit d'abord d'identifier les variables contaminées et ensuite d'estimer la vraie valeur avant contamination de chaque variable. La section 4 présente la détection et le traitement des variables contaminées. Finalement, à la section 5, des commentaires sur la méthodologie sont présentés.

2. UN EXEMPLE «JOUET»

On suppose que les Y_i les variables observées, $i=1, \dots, n$, sont issues de deux⁴ populations $U^{(0)}$ et $U^{(1)}$.

⁴La variable Y peut être considérée issue d'un mélange de $(G+1)$ distributions.

L'ensemble $U^{(0)}$ est l'ensemble des unités ayant des valeurs aberrantes. Soit le vecteur de variables de classification non observé $I=(I^{(0)},I^{(1)})$ qui prend ses valeurs dans $\{0,1\}$ et telle que

$$\begin{aligned} I_i^{(1)} &= 1 && \text{si l'unité } i \text{ ne contient aucune donnée} \\ &&& \text{aberrante et} \\ I_i^{(0)} &= 1 && \text{si l'unité } i \text{ contient au moins une} \\ &&& \text{donnée aberrante} \\ \text{avec } & I_i^{(0)} + I_i^{(1)} = 1. \end{aligned}$$

Modèle pour la variable de classification

Nous supposons le cas le plus simple pour la distribution de I , où I ne dépend pas des unités,

$$Pr(i \in U^{(0)}) = Pr(I_i^{(0)} = 1) = \theta$$

où θ représente le taux de contamination des individus. Cette supposition peut cependant s'avérer non réaliste dans plusieurs cas. Certains modèles permettent de tenir compte de la variabilité entre unités. Le modèle fixe ou mixte tient compte de la différence observée entre individus.

Modèle pour la variable d'intérêt

Si le vecteur des variables n'est pas entaché d'erreurs, alors on suppose que Y est une variable⁵ aléatoire normale de moyenne μ et de variance Σ :

$$f(Y/(I^{(1)}=1)) = (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-(y-\mu)'\Sigma^{-1}(y-\mu)}$$

Si le vecteur des variables d'intérêts est entaché d'erreurs, alors $Y/(I^{(0)}=1)$ est supposée être distribuée uniformément⁶ sur un compact de R^p (Lavielle et Moulines, 1997), et donc il existe une constante c telle que :

$$f(Y/(I^{(0)}=1)) = (2\pi c^2)^{-p/2}$$

Données manquantes

Si certaines valeurs des variables d'intérêts sont manquantes en raison par exemple de la non réponse, il est alors préférable de décomposer le vecteur des observations en deux parties $Y_i' = (Y_{io}', Y_{im}')$, où Y_{io} représente le vecteur des variables observées de

⁵ La variable Y peut être considérée issue d'un mélange de G distributions :

$$Y/(I^{(g)}=1) \sim N(\mu_g, \Sigma_g), \quad g=1, \dots, G$$

⁶ Little et Rubin (1987) utilisent la distribution $Y/(I^{(0)}=1) \sim N(\mu, \Sigma/\lambda)$, avec $\lambda < 1$

l'individu i et Y_{im} représente le vecteur des variables manquantes de l'individu i . Dans le cas d'une réponse complète on a $Y_i = Y_{io}$.

3. ESTIMATION VIA LE MAXIMUM DE VRAISEMBLANCE

L'objectif de cette section est de fournir une méthode pour estimer les paramètres $\phi' = (\mu, \Sigma, \theta)$ ainsi que les valeurs des variables manquantes $Z' = (I, Y_m)$ (n valeurs de la variable de classification I et un certain nombre de valeurs des variables d'intérêts Y_m correspondant aux valeurs manquantes).

3.1 Approche via la distribution conditionnelle

Une première approche est de maximiser la distribution conditionnelle $f(Z/y_o)$. Puisque la distribution conditionnelle est proportionnelle à la distribution conjointe

$$f(Z/y_o) \propto f(Z, Y_o)$$

cela revient à maximiser la distribution conjointe. La fonction conjointe des données complètes est

$$\begin{aligned} L(Y_o, Z) &= \prod_i \prod_g \left(\theta^{(g)} f(Y_i/g) \right)^{I_i^{(g)}} \\ &= \prod_i \prod_g \left(\theta^{(g)} f(Y_{im}/y_{io}, g) f(y_{io}/g) \right)^{I_i^{(g)}} \end{aligned}$$

pour $g=0,1$, et le logarithme de la fonction de vraisemblance est

$$\begin{aligned} l(Y_o, Z) &= \sum_i \sum_g I_i^{(g)} \log \theta^{(g)} \\ &= \sum_i \sum_g I_i^{(g)} \log f(Y_{im}/y_{io}, g) \\ &= \sum_i \sum_g I_i^{(g)} \log f(Y_{io}/g) \end{aligned}$$

Dans le cas de notre exemple, le logarithme de la fonction conjointe est

$$\begin{aligned} l(Y_o, Z) &= n^{(0)}(\log \theta - p \log c) + n^{(1)} \log(1 - \theta) \\ &\quad - \frac{1}{2} \sum_{i=1}^n I_i^{(1)} \left((y_{im} - \mu_{m/o}) \Sigma_{m/o}^{-1} (y_{im} - \mu_{m/o})' + \log |\Sigma_{m/o}| \right) \\ &\quad - \frac{1}{2} \sum_{i=1}^n I_i^{(1)} \left((y_{io} - \mu_o) \Sigma_o^{-1} (y_{io} - \mu_o)' + \log |\Sigma_o| \right) \\ &\quad - \frac{1}{2} n p \log(2\pi) \end{aligned}$$

où $n^{(g)} = \sum I_i^{(g)}$, $\mu_{m/o} = \mu_m + \sum_{mo} \Sigma_{oo}^{-1} (y_o - \mu_o)$ and

$$\Sigma_{n/o} = \Sigma_{nm} - \sum_{mo} \Sigma_{oo}^{-1} \Sigma_{om}$$

Chaque terme de la fonction conjointe est relié à un ajustement. Le premier et le deuxième terme contrôlent le nombre de données aberrantes. Le troisième terme ajuste les données manquantes au modèle et le quatrième terme ajuste les données observées au modèle.

Pour Y_m et I données, les paramètres à estimer sont $\varphi=(\mu,\Sigma,\theta)$ et peuvent être estimés par des estimateurs empiriques. Pour Y_m et φ données, les valeurs de I peuvent être estimées par la méthode de recuit simulé. D'une façon similaire, pour φ et I données, les valeurs de Y_m peuvent être estimées par la méthode de recuit simulé.

Algorithme de Recuit Simulé

Les fonctions à optimiser peuvent comporter des complexités; dans ce cas, les méthodes usuelles échouent dans l'application à trouver des solutions. La première complexité est que la fonction à optimiser $f(x)$ est rarement convexe et la solution est un point stationnaire pour laquelle la dérivée est nulle. Ce point peut être un maximum local, un point-selle ou un minimum suivant les valeurs initiales choisies.

La deuxième complexité est que la fonction à optimiser est définie sur un ensemble fini de très grande dimension. Dans ce cas, on est confronté à un problème combinatoire énorme. L'ensemble des valeurs de $I^{(g)}$ est l'ensemble de tous les choix possibles (donc 2^n). Un choix possible est un point de 2^n , $(I_1^{(g)}, \dots, I_n^{(g)})$ où $I_i^{(g)} \in \{0,1\}$.

Dans de pareilles circonstances, Geman et Geman (1984) a utilisé l'algorithme de recuit simulé (Kirkpatrick et al, 1983) pour s'attaquer au problème. L'idée de l'algorithme de recuit simulé consiste à définir une suite de lois: loi de Gibbs. La loi de Gibbs associée à la fonction $f(x)$ et à la température τ est la loi définie par

$$g_\tau(x) = c_\tau e^{-\frac{f(x)}{\tau}}$$

où c_τ est une constante de normalisation et τ est un paramètre connu et appelé, en terme physique, température du système. Cette loi se base sur un principe statistique de la physique (Boltzmann 1877) décrit plus loin et disponible dans les livres de base de physique.

On a

- quand $\tau \rightarrow \infty$ la fonction g_τ tend vers une distribution uniforme sur le domaine de X ,
- $\tau = 1$ correspond à la distribution initiale,

- quand $\tau \rightarrow 0$ la fonction g_τ tend vers une distribution uniforme sur l'ensemble des maxima globaux de $f(x)$. Ce fait est essentiel pour la recherche de maxima de $f(x)$.

Pour fabriquer un acier de bonne qualité, on fusionne les composantes de l'acier à haute température et on réduit la température graduellement. Réduire la température, $\tau: \infty \rightarrow 0$, graduellement par étape permet au système d'atteindre son équilibre à chaque étape. À température élevée, les atomes ont une grande mobilité. Au fur et à mesure que la température baisse, les atomes perdent de cette liberté et tendent à se cristalliser afin d'atteindre un état rigide. Cette état rigide correspond, dans un cas idéal, à l'état avec énergie minimale.

Le premier système à reprendre cette idée est la dynamique de Metropolis (Metropolis et al, 1953). L'algorithme de Metropolis, génère une nouvelle estimation ou observation $X^{(c+1)}$ à partir de l'estimation courante $X^{(c)}$ en générant en premier lieu une estimation candidate X en utilisant une distribution conductrice $g(X|x^{(c)})$ (conductrice au sens de piloter la série d'estimation). Par la suite, une décision probabiliste est prise concernant le rejet ou l'acceptation de la candidate en fonction de son poids par rapport à la précédente en utilisant la fonction $f(x)$. Si $f(x) > f(x^{(c)})$, on sait que x est bon (on maximise), on prend $X^{(c+1)} = x$. Dans le cas contraire, on veut éviter de rester piégé en un éventuel maximum local, on prend $X^{(c+1)} = x$ si $f(x^{(c)}) - f(x)$ est inférieur à la variable aléatoire $-\log(u)$ et $X^{(c+1)} = x^{(c)}$ dans le cas contraire, où u est une variable uniforme sur $(0,1)$. Ainsi $f(x^{(c)}) - f(x)$ est comparé à une variable exponentielle⁷ de moyenne τ . On répète ce procédé jusqu'à convergence. La dynamique de recuit simulé réduit la température et recommence la dynamique de Metropolis. La température est alors réduite jusqu'à convergence. Si tous les points communiquent entre eux, alors la chaîne est récurrente et sa loi stationnaire est $g_\tau(x)$. Le choix d'une candidate est basé sur la relation de voisinage. Le choix d'un voisin peut, par exemple, consister à tirer au hasard une unité, puis ayant tiré l'unité i , à modifier la composante X_i selon la transition $f(X/\cdot)$. Cependant l'algorithme de recuit simulé nécessite un temps d'ordinateur élevé.

⁷ En effet une variable exponentielle de moyenne τ a comme fonction de densité $f(x) = e^{-x/\tau}/\tau$ et comme fonction de répartition $F(x) = 1 - e^{-x/\tau}$. On posant $u = 1 - e^{-x/\tau}$, on obtient $x = -\log(1-u)\tau$. Puisque u a la même distribution que $1-u$, certains auteurs recommandent d'utiliser la transformation plus simple $x = -\log(u)\tau$.

3.2 Approche via l'algorithme EM

Si l'intérêt est d'estimer uniquement les paramètres des distributions, on pourrait songer à utiliser la distribution marginale des variables observées,

$$f(Y_o) = \int f(Y_o, Z) dz$$

L'évaluation de la distribution marginale nécessite une sorte d'intégration. Cette intégration peut être une lourde tâche dans certains cas, si ce n'est une tâche impossible à effectuer. La phase de maximisation pose aussi certains problèmes de calcul. La vraisemblance peut être presque aplatie dans certaines directions, ce qui rend la matrice des dérivées secondes presque singulière. Dans ce cas, les méthodes, comme celle de Newton-Raphson, deviennent inefficaces (Demnati et Beaumont, 1998). Cependant, il est beaucoup plus facile d'obtenir les estimations du maximum de vraisemblance en traitant le problème comme un cas de données manquantes et d'appliquer l'algorithme itérative EM (Hartley 1958, Dempster et al 1977). Les estimateurs de l'algorithme EM peuvent varier légèrement d'une situation à une autre mais ils ont la forme générale suivante:

$$\hat{\theta}^{(g)} = \sum_{i=1}^n w_i^{(g)} / n$$

$$\hat{\mu} = \sum_{i=1}^n w_i^{(1)} \hat{y}_i / \sum_{i=1}^n w_i^{(1)}$$

$$\hat{\sigma}_{jk} = \frac{\sum_i w_i^{(1)} (\hat{y}_{ij} - \mu_j)(\hat{y}_{ik} - \mu_k) + c_{ijklo}}{\sum_i w_i^{(1)}}$$

Les notations suivantes 1 à 3 définissent les paramètres de la forme générale des estimateurs:

Notation 1

$$w_i^{(g)} = \begin{cases} I_i^{(g)} & \text{si } I_i^{(g)} \text{ est observée} \\ E(I_i^{(g)} / y_o, \varphi) & \text{si } I_i^{(g)} \text{ n'est pas observée} \end{cases}$$

Il suffit de remplacer la variable de classification par son espérance conditionnelle, étant donné Y_o et φ . Par exemple,

$$\begin{aligned} w_i^{(0)} &= E(I_i^{(0)} / y_o; \varphi) = Pr(I_i^{(0)} = 1 / y_o; \varphi) \\ &= \frac{\theta^{(0)} f(y_o / I_i^{(0)} = 1; \varphi)}{\theta^{(0)} f(y_o / I_i^{(0)} = 1; \varphi) + (1 - \theta^{(0)}) f(y_o / I_i^{(0)} = 0; \varphi)} \end{aligned}$$

Notation 2

$$\hat{Y}' = (Y_o', \hat{Y}_m') \quad \text{où} \quad \hat{Y}_m = \mu_m + \Sigma_{mo} \Sigma_{oo}^{-1} (y_o - \mu_o)$$

$$\text{car si} \quad Y = \begin{bmatrix} Y_o \\ Y_m \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_o \\ \mu_m \end{bmatrix}, \begin{bmatrix} \Sigma_{oo} & \Sigma_{om} \\ \Sigma_{mo} & \Sigma_{mm} \end{bmatrix} \right)$$

alors

$$Y_m / y_o \sim N \left(\mu_m + \Sigma_{mo} \Sigma_{oo}^{-1} (y_o - \mu_o), \Sigma_{mm} - \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} \right)$$

et donc on estime la vraie valeur manquante par la meilleure estimation, dans le sens d'un carré moyen de l'erreur et d'après l'observation de y_o ,

$$\hat{Y}_m = E(Y_m / y_o) = \mu_m + \Sigma_{mo} \Sigma_{oo}^{-1} (y_o - \mu_o)$$

$$\text{On a } E(\hat{Y}_m) = \mu_m$$

$$\begin{aligned} \text{Cov}(\hat{Y}_m) &= E(\hat{Y}_m - \mu_m)(\hat{Y}_m - \mu_m)' \\ &= \Sigma_{mo} \Sigma_{oo}^{-1} E(y_o - \mu_o)(y_o - \mu_o)' \Sigma_{oo}^{-1} \Sigma_{om} \\ &= \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} \end{aligned}$$

$$\begin{aligned} \text{Cov}(\hat{Y}_m, Y_o) &= E(\hat{Y}_m - \mu_m)(Y_o - \mu_o)' \\ &= \Sigma_{mo} \Sigma_{oo}^{-1} E(y_o - \mu_o)(y_o - \mu_o)' \\ &= \Sigma_{mo} \end{aligned}$$

et

$$\begin{aligned} \text{Cov}(\hat{Y}_m, Y_m) &= E(\hat{Y}_m - \mu_m)(Y_m - \mu_m)' \\ &= \Sigma_{mo} \Sigma_{oo}^{-1} E(y_o - \mu_o)(y_m - \mu_m)' \\ &= \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} \end{aligned}$$

Notation 3

$$c_{ijklo} = \begin{cases} \text{Cov}(y_j, y_k) / y_o, \varphi & \text{si } Y_{ij} \text{ et } Y_{ik} \text{ sont manquantes} \\ 0 & \text{sinon} \end{cases}$$

Le logarithme de la fonction de densité d'une loi normale multivariée est :

$$\begin{aligned} \log f(Y / \mu, \Sigma) &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^p (y_{ij} - \mu_j) \delta_{jk} (y_{ik} - \mu_k) \\ &\quad - \frac{1}{2} n \log |\Sigma| - \frac{1}{2} n p \log(2\pi) \end{aligned}$$

où δ_{jk} dénote l'élément jk de la matrice inverse Σ^{-1} .

L'espérance conditionnelle (Beale et Little, 1975) est

$$E(\log f(Y/\mu, \Sigma)/y_{o, \varphi}) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^p (\hat{y}_{ij} - \mu_j)(\hat{y}_{ik} - \mu_k) + c_{ijk|o} \delta_{jk} - \frac{1}{2} n \log |\Sigma| - \frac{1}{2} n p \log(2\pi)$$

L'algorithme EM est sensible aux valeurs initiales et la convergence peut être lente. L'algorithme doit aussi être répété avec différentes valeurs initiales pour s'assurer que le maximum global est atteint. La partition des unités en deux groupes peut se faire en affectant l'unité i au groupe pour lequel l'unité a la plus haute probabilité d'appartenance. Par exemple, au groupe $U^{(0)}$ si $w_i^{(0)} > w_i^{(1)}$. Une variante de l'algorithme EM, SEM (Celleux et Diebolt 1986), permet la classification automatique des unités. Elle repose sur un principe de tirage aléatoire des Z_i au cours, d'une étape S insérée entre les étapes E et M. Cette version du EM permet d'éviter les convergences du EM vers des points selle de la vraisemblance.

4. TRAITEMENT DES DONNÉES

Une fois que les unités ayant au moins une valeur aberrante ont été détectées, il reste deux complexités : il s'agit d'abord d'identifier les variables contaminées et ensuite d'estimer la vraie valeur avant contamination. Soit le vecteur de variables de classification non observées $J_i = (J_{i1}, \dots, J_{ip})$ qui prend ses valeurs dans $\{0, 1\}$, telle que

$J_{ij} = 1$ si la variable j de l'unité i est contaminée
et
 $J_{ij} = 0$ sinon

Le domaine de J_i a $(2^p - 1)$ éléments soient : $\{(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (1, 1, \dots, 1)\}$

Le taux de contamination de chaque variable est dénoté par $E(J_{ij}) = \lambda_{ij}$. Il est difficile de connaître la distribution conjointe des J_i . Le cas le plus simple est de supposer l'indépendance du processus de contamination,

$$Pr(J_{ij} = 1 / J_{i(-j)}) = Pr(J_{ij} = 1)$$

où $J_{i(-j)} = (J_{i1}, \dots, J_{i(j-1)}, J_{i(j+1)}, \dots, J_{ip})$. Puisque les indicateurs $J_i = J_{i1}, \dots, J_{ip}$ sont inconnus on les remplace par leur espérance conditionnelles, étant donné Y_o et φ

$$E(J_{ij} / y_{oj}; \varphi) = Pr(J_{ij} = 1 / y_{oj}; \varphi) = \frac{\lambda_{ij} f(y_{oj} / J_{ij} = 1; \varphi)}{\lambda_{ij} f(y_{oj} / J_{ij} = 1; \varphi) + (1 - \lambda_{ij}) f(y_{oj} / J_{ij} = 0; \varphi)}$$

Étant donné que certaines valeurs observées des variables d'intérêts sont entachées d'erreurs, il est pratique de décomposer le vecteur observé en deux parties: $Y'_{io} = (Y'_{ior}, Y'_{ioc})$, où Y'_{ior} représente le vecteur des variables observées non contaminées de l'individu i et Y'_{ioc} représente le vecteur des variables contaminées de l'individu i .

Le modèle complet est donc $Y_i = X_i + I_i^{(0)} \text{Diag}(J_i) \varepsilon_i$, où X est la vraie valeur, avant contamination.

On estime les vraies valeurs non observées par leurs espérances conditionnelles.

Si

$$\begin{bmatrix} Y_{oc} \\ X_{oc} \end{bmatrix} = \begin{bmatrix} X_{oc} + \varepsilon_{oc} \\ X_{oc} \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_c + m_c \\ \mu_c \end{bmatrix}, \begin{bmatrix} \Sigma_{cc} + Q_{cc} & \Sigma_{cc} \\ \Sigma_{cc} & \Sigma_{cc} \end{bmatrix} \right)$$

alors

$$\hat{X}_{oc} = E(X_{oc} / y_{oc}) = \mu_c + \Sigma_{cc} (\Sigma_{cc} + Q_{cc})^{-1} (y_c - \mu_c - m_c)$$

L'erreur ou la contamination pour l'individu i est estimée par $e_{io} = Y_{io} - \hat{X}_{io}$, où $\hat{X}'_{io} = (X'_{ior}, \hat{X}'_{ioc})$. Il est possible d'introduire une certaine similitude entre les variables indicatrices quant à la contamination, la probabilité conditionnelle dans ce cas dépendra de l'état des autres variables,

$$Pr(J_{ij} = 1 / J_{i(-j)})$$

Le processus d'estimation devient alors plus coûteux puisque on doit avoir recours à un algorithme de type Metropolis, par exemple l'échantillonneur Gibbs (Geman et Geman, 1984) pour générer des observations de cette distribution,

$$Pr(J_{ij} = 1 / y_{oj}, J_{i(-j)}) \propto f(y_{oj} / J_{ij}) Pr(J_{ij} = 1 / J_{i(-j)})$$

L'échantillonneur Gibbs permet de générer des observations de distributions conjointes ou marginales à partir de distributions conditionnelles.

5. COMMENTAIRES

Nous avons fait le survol d'une méthodologie permettant la détection et le traitement des valeurs aberrantes. La modélisation de variables est certes un élément important à l'application de toute méthodologie aux données réelles. D'autres méthodologies font aussi partie d'un examen à Statistique Canada. Certaines méthodes utilisent des statistiques «robustes» pour l'estimation des paramètres. Cette approche utilise aussi des statistiques «robustes» dans le sens où elle pondère les observations quant à leur appartenance au groupe. La

méthode de recuit simulé est une méthode à considérer si ce n'est que pour échantillonner la fonction à optimiser afin d'avoir une idée sur sa concavité.

En cas de données répétées, il est souhaitable qu'il y ait un processus d'apprentissage. L'apprentis-sage peut s'effectuer (Lavine et West, 1992) en remplaçant la distribution inconditionnelle $f(Z/\varphi)$ de la variable manquante par la distribution conditionnelle $f(Z/y,\varphi)$ obtenue, qui résume l'information obtenue jusqu'à maintenant.

Afin de tenir compte de la différence entre répondant et non-répondant, Little et Wang (1996) suggèrent de stratifier les données selon les profils de la non-réponse et de spécifier un modèle pour chaque strate. Des hypothèses sur le mécanisme générant la non-réponse doivent alors être spécifiées. Supposons que le vecteur R_i représente le schéma de réponse de l'unité i , où $R_{ij}=1$ si la variable y_{ij} est observée, sinon $R_{ij}=0$. La fonction de vraisemblance maintenant peut être écrite comme un mélange de distributions, $f(Y_o,Z,R) = f(Y_o,Z/r)f(R=r)$, où $f(Y_o,Z/r)$ représente maintenant la distribution conjointe pour chaque profil de réponse. Finalement, signalons que les interrogations de ma fille ont été légèrement perturbées afin ...

REMERCIEMENTS

Ce travail est rendu possible grâce à la Division des données régionales et administratives de Statistique Canada. L'auteur tient à remercier Linda Standish et Richard Burgess pour leurs conseils précieux. Il tient aussi à remercier Réjeanne Loranger pour son aide à la révision de ce document.

BIBLIOGRAPHIE

Beale, E. M. et Little, R. J. A., (1975), Missing Values in Multivariate Analysis, Journal of the Royal Statistics Society, B37, pp.129-145.

Binder, D. A. (1978), Bayesian cluster analysis, Biometrika, 65,1,pp. 31-38.

Celleux et Diebolt (1986), Un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densité, Revue de Statistiques appliquées, Vol.XXXIV, 2, 35-52.

Demnati, A. et Beaumont, J.-F. (1998), Données administratives et Hétérogénéité non observée: Un mariage sans divorce pour une analyse longitudinale efficace !, Recueil de la section des méthodes d'enquêtes, Société Statistique du Canada.

Dempster, A. P., Laird, N. M. et Rubin D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm(with discussion). J. R. Statist. Soc. B 39, 1-38.

German, S. et German, D. (1984) Stochastic relaxation, Gipps distributions and the Bayesian restoration of images. I.E.E.E. Trans.Pattern. Anal. Machine Intell.,6, 721-741.

Hartley, H. O. (1958), Maximum likelihood procedures for incomplete data, Biometrics, 14, 174-194.

Kirkpatrick, S., Gelatt, C. D. and Vecchi. M. P. (1983) Optimization by simulated annealing, Science, 220,671-680.

Lavielle, M. et Moulines, E. (1997), Quelques exemples de problèmes inverses en statistique et en traitement du signal, Rev. Statistique Appliquée, XLV(4),5-38.

Lavine, M. et West, M. (1992), A Bayesian method for classification and discrimination, The Canadian Journal of Statistics, Vol.20, No.4, p.451-561.

Little, R. J. A. et Rubin, D. B. (1986), Statistical Analysis with Missing Data, Wiley.

Little, R. J. A. et Wang, Y. (1996), Pattern-Mixture Models for Multivariate Incomplete Data with Covariates, Biometrics, 52, 98-111.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller A. H. et Teller, E. (1953), Equations of state calculations by fast computing machines, J. Chem. Phys., 21, p.1087-1091.

Tukey, J. W. (1977), Exploratory data analysis, Addison-Wesley.

Vardi, Y. Shepp, A. et Kaufman L (1985), A Statistical Model for Positron Emission Tomography, Journal of the American Statistical Association, Vol.80, No. 389, pp. 8-37.