Détection de valeurs aberrantes

Dans les données incomplètes multivariées d'enquêtes par échantillonnage

Cédric Béguin · Beat Hulliger

e-mail: cedric.beguin@bfs.admin.ch, beat.hulliger@bfs.admin.ch

Résumé Dans le cadre du projet européen EUREDIT (5ème programme cadre de recherche de la CE), l'Office fédéral de la statistique suisse a développé des méthodes de détection de valeurs aberrantes dans des données multivariées obtenues lors d'enquêtes par échantillonnage. Ces méthodes doivent pouvoir tenir compte des poids d'échantillonnage et fonctionner en présence de nombreuses valeurs manquantes. Trois nouvelles méthodes sont présentées ici, les deux premières utilisant une distance de Mahalanobis basée sur des estimateurs robustes et la troisième basée sur une notion d'épidémie se propageant dans l'ensemble des données.

---

1. Introduction

Le projet EUREDIT ou "Développement et évaluation de nouvelles méthodes de contrôle, correction et imputation de données" implique 12 partenaires provenant de 7 pays européens et se déroule sur une période de 3 ans, de mars 2000 à février 2003. La participation de la Suisse à ce projet, financée par l'Office fédéral de l'éducation et de la science, est assurée par le service de méthodes statistiques de l'Office fédéral de la statistique (OFS).

Outre les méthodes standards de contrôle et de correction des données, un des intérêts d'EUREDIT est de développer et de comparer des méthodes robustes de détection de valeurs aberrantes dans des données d'enquêtes par échantillonnage. Cette partie du projet, à laquelle participent également l'Université de Southampton, l'Office national hollandais de statistique (CBS) et la firme allemande Quantaris, est conduite par l'OFS. La phase de développement de ces nouvelles méthodes est suivie d'une phase d'évaluation basée sur un certain nombre de jeux de données choisis au début du projet.

La présence de valeurs aberrantes est un phénomène relativement commun dans les données d'enquêtes et en particulier dans les enquêtes auprès des entreprises. Ce sont des observations dont le comportement est extrêmement différent de celui de la majorité des données et dont la prise en compte ou non lors de la phase d'estimation peut conduire à des résultats totalement différents et imprécis. La détection et la correction est par conséquent une étape importante du processus d'enquête. C'est la première phase - celle de détection - du traitement des valeurs aberrantes qui est l'objet de cet exposé. La phase de correction idéale consiste bien évidement en un retour des questionnaires suspects auprès des répondants afin de distinguer les valeurs valides - on parle alors de valeurs aberrantes représentatives [3] - des erreurs - les valeurs aberrantes non représentatives. Un exemple classique de valeurs aberrantes non représentatives lors d'enquêtes auprès des entreprises survient lorsque le questionnaire demande au répondant de donner l'information dans une certaine unité (par exemple en milliers d'euros) et que le répondant en utilise une autre (par exemple en euros). Laisser inchangées de telles valeurs peut perturber sensiblement les estimations. Un traitement automatique de telles valeurs comme leur remplacement par des valeurs plus fréquentes ou leur troncature est possible mais il aboutit en général à une estimation biaisée.

Si la détection de valeurs aberrantes reste relativement aisée en une dimension, elle se complique notablement en dimensions supérieures. Pour des jeux de données multivariés complets, différentes méthodes existent et certaines sont même devenues standards comme MCD (Minimum Covariance Determinant estimator, [11] et [12]). La plupart d'entre elles utilisent une distance de Mahalanobis basée sur des estimations robustes du centre et de la dispersion des données, c'est le cas de MCD, mais aussi entre autres

---

Cédric Béguin: Office fédéral de la statistique, Espace de l'Europe 10, 2010 Neuchâtel, Suisse;
Beat Hulliger: Office fédéral de la statistique, Espace de l'Europe 10, 2010 Neuchâtel, Suisse.

Auteur à contacter: Cédric Béguin

Recherche effectuée pour le projet EUREDIT du 5ème programme cadre de recherche de la CE. La participation financière de la Suisse a été assurée par l'Office fédéral de l'éducation et de la science.

de SD (Stahel-Donoho estimator, [14], [2] et [10]) et de BACON (Blocked Adaptative Computationally-efficient Outlier Nominator, [1]). Parmi ces méthodes, celles préservant l'équivariance affine semblent être les plus coûteuses en terme de ressources informatiques. Une autre famille de méthodes se base sur des notions de profondeur des données (data-depth, [9]). Ce second groupe de méthodes a l'avantage de ne rien supposer sur la distribution des données mais en contrepartie est souvent très coûteux voir impossible en terme de calcul.

Si les estimateurs robustes univariés, tels que des médianes ou des moyennes tronquées, sont maintenant fréquemment utilisés dans les enquêtes à grande échelle ([6], [7],...), les méthodes multivariées en sont encore a leurs premiers balbutiements et ne sont que très rarement appliquées dans les différents offices nationaux à l'exception de Statistics Canada [4]. Parmi les principaux obstacles s'opposant à l'application des méthodes robustes multivariées on peut citer:

a. les méthodes existantes fonctionnent relativement bien avec des jeux de données petits, mais deviennent extrêmement coûteuses en temps de calcul déjà avec des jeux de taille modérée tels que 5000 observations et 10 dimensions;
b. les méthodes existantes ne tiennent en général pas compte des poids d'échantillonnage;
c. les méthodes existantes ne fonctionnent en général pas en présence de valeurs manquantes.

Dans le cadre d'EUREDIT l'OFS a développé trois nouvelles méthodes qui tentent de surmonter ces trois difficultés simultanément. Les deux premières utilisent une distance de Mahalanobis basée sur des estimateurs robustes, alors que la troisième utilise une épidémie virtuelle pour définir une notion proche de celle de profondeur. La deuxième section présentera une méthode simple estimant de façon robuste la matrice de covariance à l'aide de transformations appliquées aux corrélations des rangs (algorithme TRC ou "Corrélations de Rangs Transformées"). La troisième section adaptera la méthode BACON aux jeux de données réelles en s'appuyant sur l'algorithme EM (algorithme BEM ou "BACON-EM"); l'estimation robuste est ici obtenue en partant d'un petit sous-ensemble d'observations correctes et en le faisant croître. Enfin la dernière section présentera la méthode dynamique basée sur la distance euclidienne (algorithme EA ou "Algorithme d'Epidémie").

2. Algorithme TRC ou "Corrélations de Rangs Transformées"

La matrice de covariance usuelle peut être calculée composante par composante à l'aide des covariances entre deux variables. Cette idée simple est utilisée par Gnanadesikan et Kettenring [5] pour obtenir une première estimation robuste de la matrice de covariance en remplaçant la covariance de deux variables x et y par $\mathrm{cov}(x,y) = (\sigma^2(x+y) - \sigma^2(x-y))/4$ et en utilisant un estimateur robuste de la variance σ^2. La matrice obtenue n'est en général pas définie positive et une transformation supplémentaire est nécessaire pour rétablir cette propriété. Cette première approximation de la matrice est obtenue ici de façon différente en utilisant les corrélations de rangs de Spearman. La covariance de x et y est estimée de façon robuste par $mad(x) \cdot 2\sin(\pi R(x,y)/6) \cdot \mathrm{mad}(y)$, où $R(x,y)$ est la corrélation de rangs de Spearman et $\mathrm{mad}(x) = \mathrm{med}(|x - \mathrm{med}(x)|)$. Pour assurer la positivité de la matrice S_1 ainsi obtenue les transformations existantes avaient le défaut d'être purement algébriques [13] et de ne pas avoir d'interprétation statistique basée sur les données. La solution adoptée ici est, elle, basée sur les données. La matrice obtenue S_1, si elle n'est pas définie positive, est néanmoins symétrique, donc orthogonalement diagonalisable. Ainsi il est possible d'écrire $S_1 = QDQ^{-1}$ où Q est une matrice orthogonale dont les colonnes définissent une base de "composantes principales robustes" et D est diagonale avec des valeurs propres non nécessairement strictement positives. L'idée est alors d'exprimer les données dans la nouvelle base et de remplacer les valeurs propres de D par une estimation robuste des variances des données (le mad est ici à nouveau choisi) sur chacune des nouvelles variables pour obtenir une matrice diagonale $\tilde{D}$ cette fois-ci définie positive. L'estimateur de la matrice de covariance est alors obtenu en se ramenant à la base initiale $S_{TRC} = Q\tilde{D}Q^{-1}$. De façon similaire, un estimateur robuste du centre $\tilde{M}$ (les médianes) dans les nouvelles variables est ramené dans la base initiale pour obtenir l'estimateur robuste du centre $M_{TRC} = Q\tilde{M}$. Ces deux estimateurs sont alors utilisés pour calculer les distances de Mahalanobis des observations et pour déterminer quelles observations sont aberrantes.

L'introduction des poids d'échantillonnage dans la méthode est purement technique et concerne uniquement les estimations des rangs, de la médiane et du mad. En exprimant les rangs comme une fonctionnelle de la distribution de la population, on obtient facilement l'estimateur naturel pour le rang de

l'observation x_i: $\widehat{R}(x_i) = \sum_{x_k<x_i} w_k + \frac{1}{2}\sum_{x_k=x_i} w_k + \frac{1}{2}$, où les w_k sont les poids d'échantillonnage. La médiane pondérée et donc l'estimation du mad sont obtenus similairement.

La présence de valeurs manquantes ne perturbe que très peu les statistiques univariées (rang, médiane et mad) et bivariées (corrélation) qui sont calculées uniquement sur les valeurs présentes, sélection qui est nettement moins restrictive qu'en dimensions supérieures. Le principal problème créé par les valeurs manquantes survient lors du changement de base des données. Il est, en effet, impossible de projeter sur un axe une observation ayant une ou plusieurs composantes manquantes. La solution adoptée reste dans l'esprit bivarié de la construction de TRC en utilisant l'information récoltée par les corrélations. Soit une observation $x = (x_1,...,x_m,x_{m+1},...,x_p)$, supposons sans nuire à la généralité que les m premières composantes sont observées et que les $p-m$ dernières composantes sont manquantes. Pour une composante manquante x_j, $j \in \{m+1,...,p\}$, on déterminera la variable X_i, $i \in \{1,...,m\}$, ayant la plus haute corrélation $R(X_i,X_j)$ avec la variable X_j et on utilisera une régression robuste de X_j sur X_i et la valeur observée x_i pour imputer une valeur x_j. Ces valeurs imputées seront alors utilisées pour la projection de x dans la nouvelle base. La qualité de ces imputations est contrôlée par divers paramètres comme le nombre minimum d'observations simultanées de X_i et X_j ou la valeur minimum de $R(X_i,X_j)$). Remarquons que ces valeurs imputées ne sont utilisées que pour le changement de base. Les distances de Mahalanobis finales sont, elles, calculées uniquement sur les composantes disponibles et amplifiées d'un facteur inversement proportionnel à la proportion de valeurs observées.

3. Algorithme BEM ou "BACON-EM"

L'algorithme BACON décrit dans [1] fait partie de la classe des méthodes de "recherche en avant" (forward search methods). Ces méthodes commencent par estimer le centre et la dispersion des données en se basant sur un petit sous-ensemble d'observations. Les distances de Mahalanobis de toutes les observations sont alors calculées en fonction de ces estimations et le sous-ensemble est agrandi en utilisant l'ordre donné par les distances. La vitesse et la fin de l'accroissement de la taille du sous-ensemble initial dépendent de la méthode. Par rapport aux premières méthodes de recherche en avant, BACON présente l'avantage de ne pas exiger un ensemble initial dépourvu de toutes valeurs aberrantes; la proportion de celles-ci ne doit cependant pas être trop importante. BACON propose deux sélections possibles de l'ensemble de départ, soit en se basant sur la distance de Mahalanobis standard basée sur l'ensemble des données, soit en utilisant la distance euclidienne par rapport à la médiane. La seconde version est plus robuste que la première mais perd la propriété d'équivariance affine. La croissance est contrôlée par un test de chi carré sur la distribution des distances de Mahalanobis. L'algorithme s'arrête soit lorsque l'ensemble des données est atteint soit lorsque le sous-ensemble se stabilise, les données exclues étant alors considérées comme aberrantes.

L'insertion des poids d'échantillonnage est directe, la moyenne et la matrice de covariance étant simplement estimées par des estimateurs de Horvitz-Thompson ou de Hájek si la taille de la population n'est pas connue. Dans le cas du départ robuste une médiane pondérée est utilisée.

Les valeurs manquantes constituent un problème plus délicat puisque toutes les estimations peuvent potentiellement être perturbées. Les comparaisons entre méthodes robustes sur des jeux de données complets ayant démontré l'extrême efficacité de BACON pour des données dont la distribution est proche d'une multivariée normale, la solution choisie est d'utiliser l'algorithme EM basé sur une telle distribution. Une adaptation de cet algorithme aux données d'enquête par échantillonnage est donc proposée. Pour chaque sous-ensemble intermédiaire, les estimations de moyenne et de matrice de covariance sont alors obtenues par cet EM modifié. La structure croissante de BACON permet une insertion décomposée de l'algorithme EM évitant des calculs superflus.

Les distances de Mahalanobis finales sont, comme pour TRC, calculées uniquement sur les composantes disponibles et amplifiées d'un facteur inversement proportionnel à la proportion de valeurs observées.

4. Algorithme EA ou "Algorithme d'Epidémie"

L'algorithme d'épidémie [8] est une nouvelle méthode non paramétrique de détection de valeurs aberrantes. Il se base sur les distances euclidiennes entre les observations calculées après standardisation des variables (basée sur les médianes et mad). Il simule une épidémie qui débute à la médiane spatiale de l'échantillon (observation minimisant la somme des distances à toutes les autres). La probabilité

d'infection d'un point sain diminue avec la distance aux autres points déjà infectés. Les infections sont indépendantes les unes des autres. L'épidémie se déroule selon un temps discret, un certain nombre d'observations étant infecté à chaque étape. L'épidémie s'arrête lorsque toutes les observations sont infectées ou lorsque plus aucun point n'est atteint pendant un temps donné. A la différence des méthodes de plus proche voisin standards, EA est plus ou moins sensible à la densité locale des données selon la fonction d'infection choisie. Les valeurs aberrantes, normalement isolées des bonnes observations, sont en général infectées très tardivement ou ne sont jamais atteintes. La détection se fait donc selon les temps d'infection.

Les poids d'échantillonnage interviennent partout dans l'algorithme: lors de la standardisation, par l'utilisation de médianes et de mad pondérés, lors de la détermination de la médiane spatiale, par la minimisation de la somme pondérée des distances et lors de chaque estimation de la probabilité d'infection d'un point. Cette dernière estimation a une interprétation relativement intuitive en terme d'épidémie: un point infecté x_i de l'échantillon représente un potentiel ou un poids d'infection au niveau de la population qui est égal à son poids d'échantillonnage.

Les valeurs manquantes perturbent l'algorithme uniquement au niveau du calcul initial des distances. Les standardisations des variables sont appliquées sur les valeurs observées. Le calcul de la distance entre deux observations se fait sur les valeurs observées en commun, résultat amplifié par un facteur inversement proportionnel à la proportion de celles-ci. Pour éviter de baser la probabilité d'infection sur une information trop restreinte un paramètre permet d'annuler celle-ci si la distance se base sur un nombre trop faible de composantes communes.

Références

1. Billor, N., Hadi, A. S. and Vellemann, P. F. (2000). "BACON: Blocked Adaptative Computationally-efficient Outlier Nominators", *Computational Statistics and Data Analysis*, **34**, pp. 279-298.
2. Donoho, D.L. (1982). "Breakdown Properties of Multivariate Location Estimators", Ph.D. Qualifying Paper, *Department of Statistics, Harvard University.*
3. Chambers, R.L. (1986). "Outlier Robust Finite Population Estimation", *Journal of the American Statistical Association*, **81**, pp. 1063-1069.
4. Franklin, S., Thomas, S. and Brodeur, M. (2000). "Robust Multivariate Outlier Detection Using Mahalanobis' Distance and a modified Stahel-Donoho Estimator", Techn. Rep., *Statistics Canada.*
5. Gnanadesikan, R. and Kettenring, J. R. (1972). "Robust estimates, residuals, and outlier detection with multiresponse data", *Biometrics*, **28**, pp. 81-124.
6. Hulliger, B. (1995). "Outlier Robust Horvitz-Thompson Estimators", *Survey Methodology*, **21**, pp. 79-87.
7. Hulliger, B. (1999). "Simple and Robust Estimators for Sampling", in *Proceedings of the Section on Survey Research Methods*,edited by the American Statistical Association, pp. 54-63.
8. Hulliger, B. and Béguin, C. (2001). "Detection of Multivariate Outliers by a Simulated Epidemic", in *Proceedings of the ETK/NTTS 2001 Conference*,edited by Eurostat, pp. 667-676.
9. Liu, R. Y., Parelius, J. M. and Singh, K. (1999). "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference", *The Annals of Statistics*, **27**, pp. 783-858.
10. Patak, Z. (1990). "Robust principal component analysis via projection pursuit", Master Thesis, *University of British Columbia, Canada.*
11. Rousseeuw, P.J. (1985). "Multivariate Estimation with High Breakdown Point", in *Mathematical Statistics and Applications*, edited by Elsevier, pp. 283–297.
12. Rousseeuw, P.J. and van Driessen, K. (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, **41**, pp. 212-223.
13. Rousseeuw, P.J. and Molenberghs, G. (1993). "Transformation of non positive semidefinite correlation matrices", *Commun. Statist.-Theory Meth.*, **22**, pp. 965-984.
14. Stahel, W.A. (1981). "Robuste Schätzungen: infinitesimale optimalität und Schätzungen von Kovarianzmatrizen", Ph.D. Thesis, *Swiss Federal Institute of Technology.*