Utilisation de la profondeur statistique pour la détection de courbes atypiques

Henri Klajnmic

UTILISATION DE LA PROFONDEUR STATISTIQUE POUR LA DETECTION DE COURBES ATYPIQUES

Henri Klajnmic

EDF R & D, Département ICAME, Groupe Statistique et Outils d'Aide à la Décision
1, avenue du Général de Gaulle
92141 Clamart Cedex
henri.klajnmic@edf.fr

RESUME
Pour détecter des valeurs atypiques (outliers), on peut se servir de la profondeur statistique, due à J. W. Tukey (1975, 1977). Pour des variables à une dimension, la médiane de l'échantillon est la plus profonde (ou centrale), et les outliers sont à une « faible profondeur ». Il est possible de généraliser cette notion à des données multidimensionnelles. Des travaux récents proposent des méthodes basées sur la profondeur pour les données fonctionnelles (des courbes). Nous les avons utilisées sur des courbes de consommation d'électricité et en montrons l'intérêt mais aussi les limites de ces techniques qui sont globales (la courbe entière est ou non atypique) plutôt que locales (des valeurs atypiques dans la courbe).

ABSTRACT
To detect outliers, we can use the notion of statistical depth, due to J. W. Tukey (1975, 1977). For univariate data, the median is the most deepest and outliers have a 'small depth'. We can generalise it to multivariate data. Recently, the notion of depth has been extended to functional data. We use such methods to load curves (electricity consomption) and we show the interest but also the limitations of these techniques which are global (the whole curve is considered as an outlier) but not necessarily local (some points of the curve are outliers).

MOTS-CLES
profondeur statistique ; outliers ; données fonctionnelles ; projections aléatoires ; bootstrap

INTRODUCTION

Dans une étude statistique, la recherche et l'identification de valeurs « atypiques » est importante car les méthodes sont sensibles soit aux données aberrantes (une erreur importante a été commise dans le recueil ou la transmission), soit à des valeurs « éloignées » de la majorité des données (« outlier ») : modèles et conclusions peuvent en être affectés. Lorsque les données sont des courbes (données fonctionnelles), les difficultés de telles détections augmentent.
Une courbe de charge consiste à enregistrer à intervalles de temps fixés les puissances électriques appelées par un client. Nous avons disposé de courbes enregistrées pendant un an au pas demi-horaire (48 points par jour) au niveau d'un client.
Nous avons utilisé des techniques basées sur la profondeur statistique permettant de détecter **globalement** des courbes de charges journalières atypiques. Selon l'activité ou le comportement du client, il est conseillé de traiter séparément différents types de jours comme jours ouvrables, jours fériés, vacances. L'examen visuel montre que les courbes peuvent être très différentes au niveau des horaires et des puissances appelées. Il serait plus rigoureux d'utiliser une « analyse de la variance fonctionnelle » (M. Febrero, P. Galeano, W. González-Manteiga (2006) et A. Cuevas, M. Febrero, R. Fraiman (2004)) pour tester si les courbes sont différentes selon le type de jour, mais cette méthode est compliquée à mettre en oeuvre.
Les méthodes utilisées déclarent atypiques une courbe journalière, ce qui signifie qu'elle est

« différente » des autres, mais pas forcément que ponctuellement des valeurs sont aberrantes. C'est un des écueils rencontrés. Nous travaillons au niveau d'un seul client. Tous les travaux ont été effectués **en R 2.8.0** (http://cran.r-project.org/) avec des fonctions écrites par M. Febrero, P. Galeano et W. González-Manteiga (2006, 2008) pour le traitement de données de pollution par les oxydes d'azote.
L'auteur tient à remercier M. Febrero (Universidad de Santiago de Compostella) pour son aide précieuse dans cette communication.

PROFONDEURS STATISTIQUES

Il existe plusieurs définitions de la « profondeur statistique » pour des observations unidimensionnelles de fonction de répartition $F(x) = Pr\{X \leq x\}$. On se reportera à la bibliographie.

Tukey

La profondeur de Tukey (1975, 1977), dite de demi-espace, s'écrit, à une dimension : $D(x) = min\,(F(x), 1 - F(x^-))$ avec $x^- \approx x - \varepsilon$. La valeur la plus profonde est la médiane.
En effet, si l'on trie les observations par ordre croissant, il y a autant de points avant et après (la moitié des points de l'échantillon) et $D(x) = 1/2$. Sinon, $D(x)$ est plus petite que 0.5, par exemple pour le troisième quartile, $D(x) = min\,(3/4, 1/4)$ et $D(x) = 1/4$.
La généralisation à p dimensions conduit à : $D(x) = inf\{P(H) : H \quad demi-espace\ fermé \quad x \in H\}$.

Simpliciale

La profondeur simpliciale de R. Y Liu (1984) : à une dimension, s'écrit $D(x) = P\left(x \in X_1^- X_2\right)$, $x \in IR$, $X_1^- X_2$ étant le segment reliant deux observations indépendantes X_1 et X_2 de fonction de répartition $F(x)$, $D(x) = 2F(x) \times \left(1 - F\left(x^-\right)\right)$. On vérifie encore que la médiane a la plus grande profondeur : 1/2.
A p dimensions, si $X_1, \ldots, X_{p+1}$ sont des observations iid de loi F et $S\left\lfloor X_1, \ldots, X_{p+1}\right\rfloor$ le simplexe de sommets $X_1, \ldots, X_{p+1}$, la définition est $D(x) = P\left(x \in S\left\lfloor X_1, \ldots, X_{p+1}\right\rfloor\right)$. Il est alors possible de déterminer un estimateur empirique de la profondeur.

Fraiman-Muñiz

La profondeur proposée par R. Fraiman et G. Muñiz (2001) : si l'échantillon n'a pas d'ex-aequos : à une dimension $D(x) = 1 - \left|1/2 - F(x)\right|$. Là encore, la médiane est la valeur la plus « profonde » et les extrêmes (minimum ou maximum) les moins profonds : $1/2 \leq D(x) \leq 1$.

Et à plusieurs dimensions ?

La difficulté concerne les données à plusieurs dimensions : comment les « ordonner » comme on le fait pour des valeurs à une dimension (des nombres) ? Les algorithmes pour calculer ces profondeurs (multidimensionnelles) sont toujours l'objet de recherches (G. Aloupis(2006)).

LES DIFFERENTES PROFONDEURS PROPOSEES POUR DES COURBES

Introduction

Dans tous les cas, on se ramène à des profondeurs à une dimension définies explicitement à partir de la fonction de répartition empirique.

Méthode des projections aléatoires

Cette méthode (fonction R utilisée par M. Febrero et al. (2006) [4]) consiste à projeter l'échantillon des 365 courbes sur des directions aléatoires (i.e. des vecteurs aléatoires normés de même dimension que les courbes, ici 48) et de calculer la profondeur simpliciale à une dimension des projections obtenues (365 valeurs unidimensionnelles). On en prend ensuite la moyenne sur le nombre de projections. La médiane est la courbe de l'échantillon qui a la plus grande profondeur. Pour trouver les outliers, on recherche un seuil en dessous duquel la courbe est atypique. Celui-ci est fixé au quantile 0.01 de la distribution des profondeurs et on utilise une méthode de bootstrap lissé pour le trouver en prenant la médiane de tous les quantiles à 1% des échantillons ainsi générés.

Profondeur de Fraiman-Muñiz

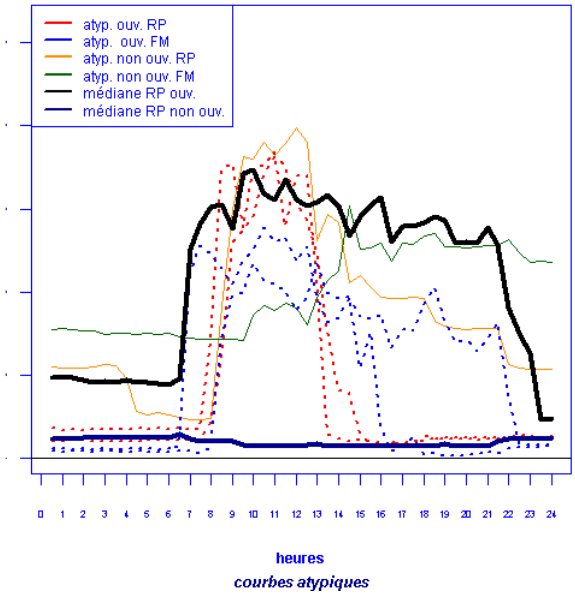
R. Fraiman et G. Muñiz (2001) proposent d'intégrer la « courbe des profondeurs statistiques » à une dimension, c'est-à-dire de calculer : $FD(x(t)) = \int_0^1 D(x(t))dt$, en supposant que la courbe $x(t)$ est définie sur $[0,1]$. Pratiquement, en chaque instant (il y en a ici 48) d'une courbe donnée, on calcule les profondeurs de l'échantillon des 365 valeurs et on effectue la moyenne des 48 profondeurs de la courbe ainsi déterminées. La profondeur de Fraiman-Muñiz est utilisée ($D(u) = 1 - [1/2 - F(u)]$) en chacun des 48 points d'une courbe journalière. Le seuil pour déclarer une courbe comme outlier est obtenu de la même façon que pour la méthode des projections aléatoires du paragraphe précédent.

MISE EN OEUVRE ET EXEMPLES

Nous présentons deux exemples d'applications.

Cas A

L'examen des courbes journalières selon les mois et les jours de la semaine montre qu'il n'y pas d'arrêt pendant le mois d'août et que l'entreprise ne fonctionne généralement pas les samedis, dimanches et jours fériés. Nous avons remarqué que le vendredi les horaires de fonctionnement sont souvent différents de ceux des autres jours ouvrables. Les vendredis non fériés ont été éliminés du traitement. Nous n'avons pas remarqué de valeurs atypiques et le maximum de la courbe annuelle est 2.8 en données centrées-réduites.
Nous avons recherché séparément les courbes atypiques sur les jours ouvrables (sans les vendredis) et les jours non ouvrables (samedis, dimanches et jours fériés) en utilisant la méthode des projections aléatoires avec la profondeur simpliciale et la profondeur de Fraiman-Muñiz. Les fonctions fournissent un seuil en-dessous duquel la courbe, dont la profondeur est alors jugée faible, est considérée comme atypique.



courbes atypiques

En gras, les courbes médianes (i.e. les plus centrales dans l'échantillon) : on voit les horaires d'activité et le fait qu'il y a peu de consommation les samedis, dimanches et jours fériés. En trait fin, on voit deux courbes atypiques le 1er et le 8 mai où l'entreprise a travaillé (le 1er mai est détecté par les deux méthodes mais pas le 8 mai). En pointillé, les courbes atypiques pour les jours ouvrables : le 14 août est détecté par les deux méthodes (projections aléatoires et profondeur de Fraiman-Muñiz), mais on détecte avec les projections aléatoires le 24 et le 31 décembre où il n'y a pas eu de consommation l'après-midi et avec Fraiman-Muñiz le 16 septembre.

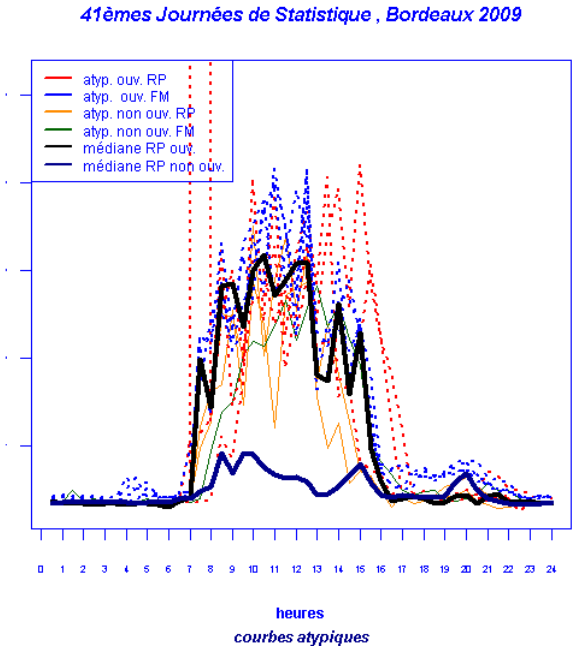
Si on fait varier le germe du générateur de nombre aléatoires, les résultats sont un peu différents, mais les méthodes détectent en général le 1er mai, les 24 et 31 décembre, le 14 août et le 16 septembre.

La profondeur de Fraiman-Muñiz est reliée au niveau moyen de la consommation journalière, ce qui fait que les jours atypiques ainsi détectés ont un niveau global de consommation différent des autres. Par contre, la méthode des projections aléatoires permet de détecter des plages de consommations qui diffèrent du comportement habituel. L'intérêt de telles techniques est leur automaticité et qu'il n'y a pas besoin de vérifier manuellement et visuellement toutes les courbes journalières.

Cas B

Ce client est présenté car il y a réellement une erreur de saisie ou de mesure qui fournit une valeur grossièrement aberrante. Ce maximum représente 124 en données centrées-réduites.
En examinant les courbes selon les mois et les jours de la semaine, nous avons décidé de séparer jours ouvrables et jours fériés. La valeur erronée est celle du vendredi 19 avril. Les méthodes des projections aléatoires et de Fraiman-Muñiz ne donnent pas les mêmes outliers.
Pour les jours fériés, on trouve samedi 18 mai et samedi 2 novembre (il y a une activité ce samedi contrairement aux autres) avec la méthode des projections aléatoires. L'autre méthode (Fraiman-Muñiz) donne le samedi 3 mars (précédant Pâques).
Pour les jours ouvrables, la méthode des projections aléatoires donne lundi 25 mars, vendredi 19 avril (avec la grosse erreur) et vendredi 10 mai (après le 8 mai et le 9 mai (Ascension)). Par contre la méthode Fraiman-Muñiz donne des jours complètement différents : vendredi 3, mardi 8, mercredi 23 et jeudi 24 janvier.

En comparant les courbes atypiques à celles de la même semaine, on peut voir des pics de consommation à des instants plutôt différents.
En ce qui concerne l'erreur de mesure, on peut faire les commentaires suivants : la méthode Fraiman-Muñiz ne le détecte pas. L'autre méthode le détecte. Mais le fait de passer par l'intermédiaire de la fonction de répartition empirique ne permet pas de bien refléter l'amplitude de l'erreur concernant ce maximum. Au-dessus du maximum, la fonction de répartition empirique ne changera pas.
Sur le dessin ci-dessous : en gras les courbes médianes pour les jours ouvrables et les jours fériés, en pointillés les courbes atypiques pour les jours ouvrables et en traits fins les courbes atypiques des jours fériés dont les niveaux de consommations sont inférieurs à ceux des jours ouvrables. On peut remarquer que selon les méthodes, les pics de consommation des outliers ne se produisent pas aux mêmes instants.



Nous avons commencé à modifier artificiellement des courbes, en créant des plages de zéros, soit des pics. Nous avons introduit aléatoirement dans la courbe médiane (initiale) trois pics valant 1.7 fois le maximum de cette courbe. La courbe médiane ainsi modifiée est devenue atypique avec la méthode des projections aléatoires. Par contre, l'introduction d'une plage de zéros n'a pas été détectée. Nous poursuivons nos analyses dans cette direction et en essayant de traiter plus de clients.

CONCLUSIONS ET PERSPECTIVES

Nous disposons d'un moyen de détecter assez facilement des comportement atypiques dans une famille de courbes. Cependant, selon les définitions des profondeurs, les germes des générateurs de nombres aléatoires ou encore le nombre de projections, les résultats peuvent être différents et il n'est donc pas simple d'être sûr qu'une courbe est atypique. Utiliser des profondeurs différentes et différentes méthodes de détermination du seuil en gardant les outliers communs est une piste intéressante proposée par M. Febrero, P. Galeano, W. González-Manteiga (2006). Il semble plus facile de détecter des courbes globalement différentes que des valeurs aberrantes plus ou moins isolées (erreur de mesure, plage de zéros). Un diagnostic d'expert serait utile et permettrait de mieux apprécier les performances. La méthode des projections aléatoires semble cependant un peu plus attractive.
Ce qui est intéressant dans ces techniques est que l'on détecte comme atypiques des jours fériés,

veilles ou lendemains de jours fériés : les horaires de forte ou faible consommation sont différents de ceux des périodes habituelles qui figurent dans la courbe médiane ou dans la courbe moyenne robuste. Par contre, ce n'est pas toujours stable et des anomalies volontairement créées ne sont pas retrouvées. Nous continuons à travailler sur la détermination du seuil ainsi que la part apportée à la profondeur d'une courbe par une portion de courbe pour améliorer le diagnostic d'atypicité.
D'autres données sont fournies lors de la construction de ces profondeurs : la courbe médiane (profondeur maximum), une moyenne robuste (α-trimmed) et une courbe « modale». Ces courbes représentent le comportement le « plus habituel » ou moyen et cela peut être intéressant à étudier pour définir une sorte de profil.
Enfin, une future direction d'investigation sera d'utiliser les techniques proposées dans l'étude des séries temporelles pour les outliers additifs (décalage de niveau, changement de variance) comme le propose par exemple R.S. Tsay (1988).

Bibliographie

[1] Aloupis G., Geometric Measures of Data Depth *DIMACS Series in Discrete Mathematics and Theoretical Computer Science (vol.72 Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications)*, R.Liu, R.Serfling, D.Souvaine eds, American Mathematical Society, (2006), 147-158.
[2] Cuevas A., Febrero M., Fraiman R., (2004) An anova test for functional data *Computational Statistics & Data Analysis* 47, 1063-1074
[3] Cuevas A., Febrero M., Fraiman R., (2006) On the use of the bootstrap for estimating functions with functional data *Computational Statistics & Data Analysis* 51, 1063-1074
[4] Febrero M., Galeano P., González-Manteiga W.,(2006) A functional analysis of NOx levels: location and scale estimation and outlier detection Universidade de Santiago de Compostella Report 06-03
[5] Febrero M., Galeano P., González-Manteiga W., (2006) Outliers detection for functional data by depth measures *International Workshop on spatio-temporal modelling, Pamplona (Spain)*
[6] Febrero M., Galeano P., González-Manteiga W., (2007) Outliers detection for functional data, *5èmes Journées de Statistique Fonctionnelle et Opératorielle, Lille*
[7] Febrero M., Galeano P., González-Manteiga W., (2008) Outlier detection in functional data by depth measures with application to identify abnormal NOx levels *Environmetrics*, Vol. 19, 331-345
[8] Fraiman R., Muñiz G. (2001) Trimmed Means for Functional Data, *Test*, Vol.10, No. 2, 419-440
[9] Liu R. Y. (1988) On a notion of simplicial depth *Proceedings of the National Academy of Sciences of the USA*, vol. 85, 1732-1734
[10] Liu R. Y, Parelius J. M., Singh K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference *Annals of Statistics*, Vol. 27, No. 3, 783-858
[11] Ramsay J. O., Silverman B. W. (2005) *Functional Data Analysis, second edition*, Springer
[12] Tsay R. S. (1988) Outliers, Level Shits and Variance Changes in Time Series, *Journal of Forecasting*, Vol. 7, 1-20
[13] Tukey, J. W. (1975) Mathematics and the picturing of Data *Proceedings of the International Congress of Mathematicians 1974, Vancouver*, volume 2, 523-531
[14] Tukey, J. W. (1977) *Exploratory Data Analysis* Addison-Wesley