Journal de la Société Française de Statistique
Vol. 159 No. 3 (2018)

Détection non-supervisée d'observations atypiques en contrôle de qualité : un survol

Title: Unsupervised outlier detection in quality control: an overview

Aurore Archimbaud[1]

Résumé : La détection d'observations atypiques ou d'anomalies est un challenge dans de nombreux domaines. Dans cet article, une revue de la littérature des méthodes non-supervisées est dressée et l'accent est principalement mis sur le contrôle de qualité. Tout d'abord il est important de noter que la notion d'anormalité retenue suit celle donnée par Hawkins (1980), à savoir qu'une observation est atypique si elle est générée par un mécanisme différent de celui de la majorité des données. Une première section se focalise sur le contexte du contrôle de qualité dans l'industrie des composants électroniques destinés aux applications automobiles, afin d'établir un inventaire des différentes méthodes utilisées en pratique. Il apparaît que ce sont principalement des méthodes univariées qui sont intégrées aux différents processus de détection de défauts. Seules quelques méthodes multivariées de type distance de Mahalanobis ou Analyse en Composantes Principales semblent connues de quelques industriels. Les sections suivantes essaient de résumer l'ensemble de la palette de possibilités destinées à la détection d'observations atypiques de manière non-supervisée ainsi que leur mise en œuvre sous le logiciel R (R Core Team, 2017). Une distinction est faite entre les méthodes ne traitant que des données en dimension standard, i.e avec plus d'observations que de variables, et celles acceptant des données en grande dimension et avec une faible taille d'échantillon.

Abstract: The outlier or anomaly detection is quite a challenge in many areas. In this article, we mainly focus on quality control and we do a review of the literature of unsupervised methods. All along this work, the notion of outlyingness follows the definition given by Hawkins (1980), namely that an observation is outlying if it is generated by a different mechanism than the one of the bulk of the data. A first section focuses on the context of quality control for the electronic components for automotive applications. It reviews all the common methods used in practice. It appears that mainly univariate methods are integrated into the fault detection processes. Only a few multivariate methods like the Mahalanobis distance or the Principal Components Analysis are used by some manufacturers. The next sections attempt to summarize all the unsupervised methods for outlier detection as well as their implementation in the R software (R Core Team, 2017). A distinction is made between methods designed for standard data, i.e. with more observations than variables, and those adapted to high dimensional data with a small sampling size.

Mots-clés : détection d'anomalies, analyse multivariée, faible taille d'échantillon, haute fiabilité
Keywords: anomaly detection, multivariate analysis, low sample size, high reliability
Classification AMS 2000 : 62-02, 62H99, 62P30

1. Introduction

La détection d'observations présentant un comportement atypique est un sujet de préoccupation depuis le XIXème siècle au moins. Les chercheurs écartaient les valeurs de l'échantillon qu'ils considéraient comme n'étant pas en accord avec la majorité de la population. Cette analyse n'étant basée que sur l'expertise de chacun à définir ce qu'est une anormalité, Peirce (1852) fut le premier à proposer un critère objectif. Il s'ensuivit de nombreuses recherches dont les

[1] TSE-R, Université Toulouse 1 Capitole
E-mail : aurore.archimbaud@ut-capitole.fr

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

avancées les plus pertinentes effectuées jusqu'en 1931 ont été consignées par Rider (1933). Toutefois, c'est véritablement l'article de Pearson et Sekar (1936) qui marqua un tournant dans la discipline. En effet, ils mirent en évidence le problème désormais très connu d'effet de masque (*masking effect* en anglais), détaillé en Section 2.3, qui peut se manifester dès que plus d'une observation se comporte de manière anormale. Or, jusqu'à cette publication, les différents critères utilisés n'étaient adaptés qu'à l'identification d'une seule observation atypique. Notons aussi que des solutions basées sur un processus itératif de détection se sont avérées être une stratégie infructueuse (Barnett et Lewis, 1994).

Dans le même temps que la recherche de critères pouvant définir et identifier des observations anormales se poursuivait, la problématique de savoir comment gérer ces valeurs s'est imposée. Dès 1977, Daniel Bernoulli remis en question la stratégie d'écarter ces mesures de toutes les analyses. Il est vrai que plusieurs cas de figure peuvent se présenter : il est possible que ces valeurs proviennent d'erreurs humaines de copie, ou véritablement qu'elles traduisent une réalité physique, physionomique, chimique ou autre. Comme l'analyste ne peut être certain que les mesures de l'échantillon ne contiennent aucune erreur, les recherches se sont portées sur une manière de pouvoir gérer la présence de telles valeurs en évitant qu'elles n'impactent trop les résultats des analyses statistiques. Le concept de Robustesse se développa alors rapidement et de nouvelles notions ont été introduites pour appréhender quantitativement la robustesse des méthodes statistiques (voir Maronna *et al.* (2006), Droesbeke *et al.* (2015) pour une présentation détaillée). Pour pallier au problème du manque de robustesse des analyses classiques, de nombreuses méthodes sont proposées, notamment pour l'estimation de paramètres de position et d'échelle en univarié et multivarié.

En dépit de l'importance des challenges à relever et de la recherche effectuée, il fallut attendre Hawkins (1980) pour un premier ouvrage dédié aux méthodes de détection ainsi qu'une définition formelle du concept d'observation atypique :
« An outlier is an observation which deviates so much from the other observations as to arouse suspicious that is was generated by a different mechanism. »
Cette intuition, qu'une observation est atypique par rapport au reste de la population parce qu'elle a été générée par un mécanisme différent, va s'instaurer comme un fondement du domaine. Le second ouvrage incontournable est la contribution de Barnett et Lewis (1994) qui présentent des outils pour pouvoir identifier ces observations anormales ou réussir à s'en accommoder sans trop impacter les analyses. Enfin, il faut attendre Aggarwal (2013) pour avoir de nouveau une revue de la littérature qui prend en compte les avancées des dernières décades. L'originalité de son ouvrage tient à ce qu'il regroupe les méthodes émergentes dans la communauté statistique aussi bien que celles de la communauté informatique.

Paradoxalement au (trop) petit nombre de livres publiés sur le sujet, à l'heure actuelle, tous les domaines sont concernés par la gestion des observations atypiques, seul leur but final diffère. Dans certains domaines, il est seulement nécessaire d'identifier et de supprimer ces individus ; dans d'autres, il faut que les analyses ne soient pas trop impactées par la présence potentielle de ces observations, et enfin pour les derniers, l'objectif même de leurs études est de les détecter. Pour exemple, on peut vouloir mettre en place un système de détection d'intrusion en cybersécurité, de détection de fraudes aux cartes de crédit, traquer les variations de capteurs de tous genres pour anticiper des pannes, diagnostiquer des maladies ou bien renforcer la fiabilité des composants électroniques dans le cadre du contrôle de qualité.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

C'est ce dernier objectif qui nous intéresse particulièrement dans cet état de l'art. En effet, afin d'assurer un niveau de qualité élevé des composants électroniques dans les domaines de l'automobile ou du spatial, la détection des composants potentiellement défectueux est un enjeu crucial qui s'inscrit dans le contrôle de qualité. D'un point de vue statistique, plusieurs challenges se posent. Tout d'abord, en fonction de la complexité des composants électroniques, le nombre de mesures à prendre en compte peut être très important. Ensuite, principalement dans le contexte spatial, le nombre de mesures excède généralement le nombre de composants testés. Or les méthodes statistiques classiques multivariées de détection d'anomalies ne peuvent s'appliquer que dans les cas où la taille d'échantillon n est plus grande que la dimension p ($n > p$). La gestion de données de type HDLSS (*High Dimension - Low Sample Size*), i.e. en grande dimension et avec une faible taille d'échantillon ($n < p$), est donc également un enjeu critique.

Cet article se focalise sur des méthodes de détection d'observations anormales dans le cas non-supervisé, i.e. sans avoir d'information à priori, afin de se placer dans le contexte industriel de détection d'une infime proportion de défauts de fabrication. La première section présente les objectifs et les méthodes utilisées dans le cadre du contrôle de qualité principalement des semi-conducteurs. La deuxième partie synthétise la revue de la littérature d'Aggarwal (2013) en présentant les principales caractéristiques des méthodes ainsi que leur possible mise en œuvre sous le logiciel R (R Core Team, 2017). Finalement, la troisième section est dédiée aux approches s'appliquant aux échantillons de type HDLSS, qu'elles proviennent des communautés informatique ou statistique.

2. Contrôle de qualité dans le domaine des semi-conducteurs

Dans les industries automobile, aéronautique ou spatiale, la recherche du zéro-défaut est la finalité à atteindre. L'émergence des marchés à bas prix et le nombre toujours croissant de composants électroniques qui équipent les différents systèmes, comme par exemple la voiture autonome, maintiennent la pression sur les entreprises pour qu'elles livrent des produits de qualité toujours supérieure. Pour parvenir à cet objectif, un long processus de qualification des produits, décrit ci-dessous, est mis en place pendant toutes les étapes de fabrication.

Dans cette section, on présente brièvement le contexte des semi-conducteurs. Ensuite, on se focalise sur les méthodes statistiques non-supervisées univariées et multivariées utilisées couramment dans ce domaine pour valider la fiabilité des puces fabriquées. La section suivante propose des critères d'évaluation de la performance des méthodes employées dans ce contexte particulier et introduit le rôle important des contributeurs. Enfin, un exemple réel de l'industrie est présenté afin de pouvoir comparer certaines méthodes multivariées.

2.1. Contexte des semi-conducteurs

Dans le domaine des semi-conducteurs, Moreno-Lizaranzu et Cuesta (2013) décrivent en détail les quatre phases importantes du processus de création : la fabrication du wafer, le probe, l'assemblage et le test final. Précisions qu'un wafer est une plaque (généralement en silicium) sur laquelle les circuits intégrés sont fabriqués couche par couche. L'étape du probe consiste à effectuer des mesures électriques sur chaque circuit pour s'assurer qu'ils ne sont pas défectueux et pouvoir choisir lesquels envoyer en assemblage. Les circuits intégrés assemblés en paquets sont

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

de nouveau soumis à une série de mesures afin de filtrer toutes les pièces qui présenteraient des défauts. En dépit de toutes ces étapes de vérification, certains appareils contiennent des pièces présentant des défauts latents qui se manifesteront plus tard et qui pourront être à l'origine d'un incident de qualité chez le client (CQI). Afin de réduire ces incidents au taux le plus bas, cinq niveaux de vérification sont appliqués dans le cadre du contrôle global de la qualité :

— FDC (*Fault Detection and Classification*) : tout d'abord, les paramètres machines sont testés pour détecter des défauts et les classer. Cette première étape permet de s'assurer que les paramètres machine ne dérivent pas et permet d'intervenir si nécessaire. D'un point de vue statistique, chacun des paramètres machines correspond à une variable et la problématique est donc une analyse multivariée. Entre autres, Lee *et al.* (2004b) proposent d'utiliser la méthode d'analyse en composantes indépendantes (ICA) pour identifier ces déviations. Cette méthode est originellement connue comme méthode de séparation aveugle de source et est très utilisée dans le traitement du signal. Quant à Lee *et al.* (2004a) et Taouali *et al.* (2016) ils suggèrent d'exploiter une variante de l'analyse en composantes principales (ACP) à noyau (KPCA).
— SPC (*Statistical Process Control*) ou MSP (Maîtrise Statistique des Procédés) en français : cette vérification a lieu au cours de la fabrication des wafers et repose sur des méthodes statistiques et graphiques comme les cartes de contrôle. Une littérature abondante existe sur le sujet, voir entre autres Mercier et Bergeret (2011); Mnassri *et al.* (2008); Jensen *et al.* (2007); Harkat *et al.* (2002); Cinar et Undey (1999); Rocke (1992) pour une synthèse des principales approches utilisées.
— PT (Test Paramétrique) : de nombreux tests électriques sont mesurés sur les wafers en fin de fabrication, afin d'écarter les plaques qui présenteraient un comportement électrique anormal. À titre d'exemple, les travaux de thèse d'Hassan (2014) visaient la mise en place d'un système de détection en temps réel pour l'entreprise STMicroelectronics.
— Probe : à ce stade ce sont les puces de chaque wafer qui sont testées pour détecter de potentiels défauts fonctionnels. Elles sont donc caractérisées en deux groupes : les « pass » ou les « fail » . Seules les premières, qui sont jugées comme « bonnes » sont mises en assemblage. En fonction du type de produit, les mesures, généralement de tension, de courant, de temps de réponse, etc, peuvent varier en intensité et en nombre.
— FT (Test Final) : une fois les puces assemblées, elles sont testées dans différentes conditions pour s'assurer de leur performance. Ces insertions regroupent des températures froide, ambiante et chaude (*cold, room, hot* en anglais) souvent suivies par une phase de *burn-in* où les puces sont placées dans des étuves. Cette dernière étape est réservée aux produits finis les plus complexes qui nécessitent une rigueur supplémentaire dans la détection des possibles pièces défaillantes.

Idéalement, il est pertinent de détecter les éventuels défauts dans les premières étapes du processus de validation. Toutefois rejeter un wafer au niveau du Test Paramétrique (PT) est déjà très coûteux car on écarte de ce fait toutes les puces créées sur cette plaque. Dans cette section, nous nous limitons à l'analyse non-supervisée du contrôle de qualité des puces, ce qui correspond aux niveaux du Probe et du Test Final (FT).

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

2.2. *Les standards en univarié pour les produits finis*

Le comité du JEDEC (2009) (Joint Electron Device Engineering Council) développe et publie les standards dans l'industrie du semi-conducteur. Dans l'industrie automobile, les entreprises Chrysler, Ford et General Motors ont créé un autre comité, Automotive Electronic Council (2011), qui donne également des lignes directrices pour assurer une production fiable et de haute qualité. Moreno-Lizaranzu et Cuesta (2013) proposent une synthèse des méthodes recommandées et mises en œuvre dans l'entreprise de semi-conducteurs Freescale (maintenant NXP). Nous présentons ici brièvement les approches non-supervisées.

2.2.1. *LSL et USL : les limites de spécification*

Avant l'arrivée de la statistique dans le processus de qualification, les ingénieurs de tests vérifiaient uniquement si les valeurs de chaque mesure étaient comprises entre les limites de spécifications. Ces limites, notées LSL (*Lower Specification Limit*) et USL (*Upper Specification Limit*), sont déterminées théoriquement ou par l'expertise des clients. Par définition, une pièce est considérée comme « bonne » (*pass*, en anglais) si elle est comprise entre ces limites LSL et USL. Même si cette méthode s'avère incontournable pour répondre au cahier des charges défini pour chaque produit, les limites sont généralement très larges et ne permettent pas de détecter toutes les pièces au comportement anormal.

2.2.2. *PAT : Part Average Testing*

Afin de renforcer le filtrage des pièces potentiellement atypiques, un second niveau de détection est mis en place : le PAT. Il permet d'identifier et de rejeter les pièces qui se comportent de manière anormale sur le plan statistique. Sous l'hypothèse de normalité de l'échantillon, cela revient à identifier les valeurs qui ne semblent pas avoir été générées par une distribution normale. Les valeurs qui se trouvent à plus de k écarts types σ par rapport à la moyenne μ mais toujours dans les limites de spécifications LSL et USL, sont de potentielles pièces atypiques, comme illustré sur la Figure 1. Formellement les limites basses et hautes, PAT_L et PAT_U, sont déterminées en fonction de k par :

$$PAT_L = \mu - k\sigma \quad \& \quad PAT_U = \mu + k\sigma \tag{1}$$

Généralement, la valeur de $k = 6$ est préconisée.

Moreno-Lizaranzu et Cuesta (2013) présentent dans leur article différentes variantes du PAT. Dans la pratique, les limites peuvent être calculées de manière statique ou dynamique. Dans le premier cas, les paramètres statistiques σ et μ sont calculés sur des pièces de référence et ensuite appliqués aux nouvelles pièces testées. Dans le deuxième cas, les paramètres σ et μ sont estimés depuis l'échantillon par la moyenne empirique μ_n et l'écart type empirique σ_n. Enfin ces limites peuvent aussi être déterminées de manière robuste, c'est-à-dire en choisissant des estimateurs de position et de dispersion qui ne sont pas impactés par la présence de valeurs atypiques, comme la médiane et l'écart inter-quartiles (IQR, *Inter-quartile Range*) pondéré par exemple. Toutefois l'utilisation des estimateurs robustes peut induire le rejet de groupes entiers de pièces et non pas

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

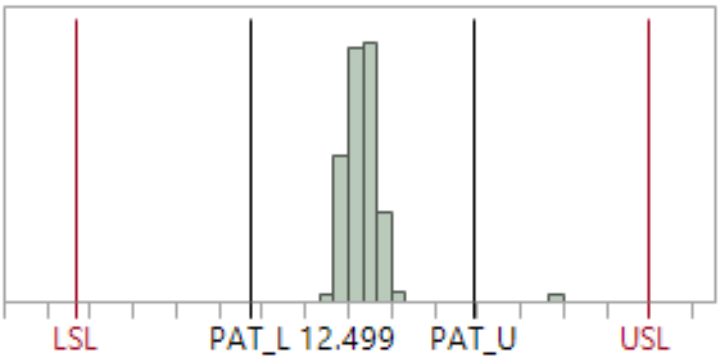

FIGURE 1. *PAT_L et PAT_U sont les limites de PAT calculées avec* $k = 6$. *LSL et USL sont les limites de spécifications.*

des pièces réellement atypiques. En effet, l'avantage d'utiliser des estimateurs robustes consiste à ne se focaliser que sur la partie centrale des données. Or par exemple, si on prend le cas de pièces provenant de deux lots différents, l'analyse des données va permettre d'identifier ces deux lots mais les pièces d'un des lots vont être identifiées comme différentes de la majorité des pièces et donc atypiques alors qu'en réalité ce ne sont pas des anomalies mais juste des pièces qui proviennent d'un autre lot.

2.2.3. GPAT ou GDBC, Geographic Part Average Testing ou Good Die in Bad Cluster

Le GPAT ou GDBC est une variante du PAT qui prend en compte la dimension spatiale des pièces sur le wafer, comme rappelés par Moreno-Lizaranzu et Cuesta (2013). Lors de la phase de Probe, toutes les pièces sont testées et si elles respectent les limites de spécification sur tous les tests, alors elles sont identifiées comme bonnes (« pass »), et sont caractérisées par une couleur blanche sur la cartographie du wafer représenté sur la Figure 2. Au contraire, elles sont considérées comme défectueuses (« fail »), et représentées par des couleurs si au moins un test est en dehors de ces limites. En fonction du type de tests hors limites, la couleur de la puce est différente. L'idée de la méthode est alors de rejeter les pièces « pass » situées dans un cluster de pièces « fail », comme par exemple la pièce blanche entourée par un carré noir en haut à gauche du wafer, qui est entourée de pièces qui vont être rejetées.

2.2.4. NNR : Nearest Neighbor Residual

Cette approche prend en compte la notion de voisinage géographique par rapport à la position des pièces sur le wafer. L'idée est de rejeter la pièce si sa valeur mesurée est statistiquement différente de la valeur attendue, calculée comme une moyenne des pièces voisines, pour un test donné (voir Moreno-Lizaranzu et Cuesta (2013) pour une description complète de la méthode). La Figure 3 illustre l'idée de cette approche.

Toutes ces méthodes permettent une amélioration de la fiabilité, mais leur coût par rapport à la détection devient rapidement prohibitif. En effet, les méthodes sont appliquées sur chaque

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

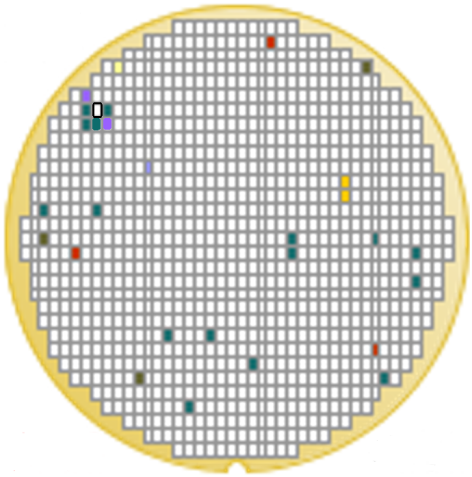

FIGURE 2. *Cartographie du wafer représentant la position géographique des puces électroniques. Les couleurs caractérisent la qualification de la puce : blanche pour « pass » et en couleur pour « fail ».*

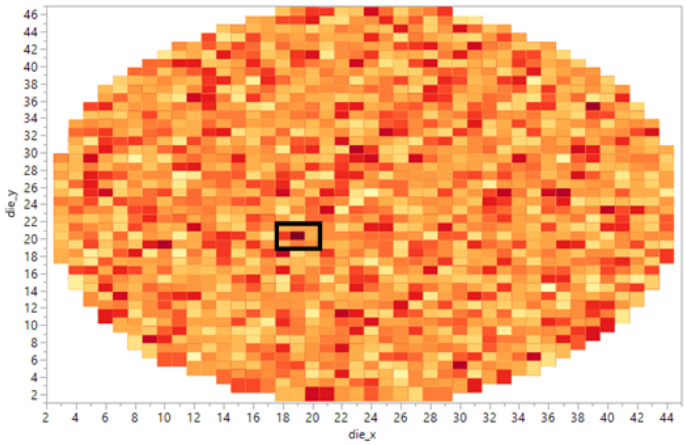

FIGURE 3. *Wafer dont les couleurs des puces représentent les différences entre les valeurs mesurées et attendues de chaque pièce sur un test électrique, par la méthode NNR. La puce en rouge foncée au centre du rectangle noir est différente de ses voisines sur ce test.*

mesure de manière indépendante et donc le taux de fausses alarmes augmente de façon spectaculaire avec le nombre de mesures. Par exemple, comme expliqué par Mercier et Bergeret (2011), si nous considérons p tests électriques au niveau de risque de $\alpha\%$, nous multiplions le taux de fausses alarmes par un facteur voisin de p :

$$\begin{aligned} &P[\text{Au moins une mesure est hors contrôle}] \\ &= 1 - P[\text{Toutes les mesures sont sous contrôle}] = 1 - (1-\alpha)^p \approx p\alpha \end{aligned} \tag{2}$$

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

Dans ce cas-là il est important de considérer une procédure de tests multiples afin de contrôler ce taux de fasses alarmes. Une solution, appelée correction de Bonferroni (Holm, 1979), consiste à ajuster le niveau de risque α en le divisant par le nombre total p de tests effectués. Néanmoins, comme le nouveau seuil α/p devient de plus en plus petit avec le nombre de tests p, il arrive fréquemment que cette solution soit trop stringente car plus aucune pièce n'est déclarée atypique avec ce nouveau niveau de risque.

De plus, comme les tests de détection sont réalisés de manière indépendante, les corrélations entre les mesures ne sont pas prises en compte. Ces méthodes ne permettent que de détecter des anormalités univariées, ce qui est un véritable inconvénient pour l'analyse de produits complexes nécessitant une fiabilité extrême. Enfin, les méthodes GPAT et NNR, qui se basent sur la localisation géographique de la puce sur le wafer, ne peuvent être utilisées que lors de la phase de probe car cette information n'est plus disponible lorsque les puces sont découpées et assemblées pour la phase de test final.

2.3. Les méthodes multivariées usuelles non-supervisées

2.3.1. La statistique du T^2 de Hotelling ou la distance de Mahalanobis (MD)

Une méthode parfois utilisée dans la maîtrise statistique des procédés (MSP ou SPC en anglais) est le calcul de la statistique du T^2 de Hotteling (1947), comme expliqué dans Jensen *et al.* (2007); Lafaye de Micheaux (2000); Mnassri *et al.* (2008). Cette statistique, équivalente à la distance de Mahalanobis (MD) au carré, peut également être appliquée pour renforcer le contrôle de qualité des puces (Aggarwal, 2013). Elle est souvent décrite comme une carte de contrôle multivariée.

Formellement, soient $\mathbf{x}_1, \ldots, \mathbf{x}_n$, n observations caractérisées par p variables quantitatives. Concrètement dans notre contexte de semi-conducteurs, $\mathbf{x}_i$ représente l'ensemble des p mesures effectuées sur la puce i. Lors du probe (resp. du test final), toutes les n puces du wafer (resp. du lot) sont testées avec le même nombre de p mesures. Chaque $\mathbf{x}_i$ est supposé être un vecteur aléatoire réel p-multivarié généré par une distribution normale multivariée avec un paramètre de localisation μ et une matrice de variance-covariance Σ : $\mathbf{x}_i \sim N(\mu, \Sigma)$. On définit la distance de chaque puce $\mathbf{x}_i$ au centre de la distribution μ par :

$$\mathrm{T}^2_{\mu,\Sigma}(\mathbf{x}_i) = \mathrm{MD}^2_{\mu,\Sigma}(\mathbf{x}_i) = (\mathbf{x}_i - \mu)'\Sigma^{-1}(\mathbf{x}_i - \mu) \tag{3}$$

Généralement les paramètres statistiques de position et de dispersion sont inconnus et doivent être estimés, le plus souvent à l'aide de la moyenne empirique

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

et de la matrice de variance-covariance empirique

$$\Sigma_n = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \mu_n)(\mathbf{x}_i - \mu_n)'.$$

En fonction des estimateurs utilisés et sous l'hypothèse de normalité, la distribution des distances peut être déduite. Les distances peuvent donc être testées pour savoir si elles dépassent le

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

quantile théorique et donc si les observations associées sont de potentielles anomalies ou non. Mardia *et al.* (1979) notamment rappellent que lorsque les paramètres μ et Σ sont connus la distribution de $\mathrm{MD}^2_{\mu,\Sigma}$ est une loi du χ^2 à p degrés de liberté. Wilks (1962) et Gnanadesikan et Kettenring (1972) démontrent que lorsque les paramètres sont estimés empiriquement par μ_n et Σ_n, la distribution exacte est une loi Beta. Plus précisément :

$$\frac{n}{(n-1)^2}\,\mathrm{MD}^2_{\mu_n,\Sigma_n}(\mathbf{y}_i) \sim \mathrm{Beta}\left(\frac{p}{2}, \frac{n-p-1}{2}\right),$$

qui peut être approximée par une loi de χ^2_p pour n grand. C'est cette dernière approximation qui est le plus largement utilisée en pratique, même si son exactitude dépend de la dimension considérée. Une observation est donc identifiée comme une anomalie au niveau $\alpha\%$ si :

$$\mathrm{T}^2_{\mu_n,\Sigma_n}(\mathbf{x}_i) = \mathrm{MD}^2_{\mu_n,\Sigma_n}(\mathbf{x}_i) > c_{p,1-\alpha} \tag{4}$$

avec $c_{p,\alpha}$ le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à p degrés de liberté.

Cette méthode est l'une des plus répandues pour la détection d'anomalies, principalement parce qu'elle est "affine invariante" tant que les estimateurs de position et de dispersion choisis sont affine équivariants, i.e. que par une transformation affine des données, ils vérifient : $\mu_n(\mathbf{A}\mathbf{x}_i+\mathbf{b}) = \mathbf{A}\mu_n(\mathbf{x}_i)+\mathbf{b}$ et $\Sigma_n(\mathbf{A}\mathbf{x}_i+\mathbf{b}) = \mathbf{A}\Sigma_n(\mathbf{x}_i)\mathbf{A}'$ avec $\mathbf{A}$ une $p\times p$ matrice non singulière et $\mathbf{b}$ un p-vecteur.

La méthode précédente soulève également certaines critiques. Tout d'abord, comme discuté dans Cerioli (2010), le choix du niveau du test est encore un point délicat car il influence le type de taux d'erreurs qui va être contrôlé : le PCER (*per-comparison error rate*), basé sur l'espérance du nombre de faux-positifs, ou le FWER (*family-wise error rate*), basé sur la probabilité d'avoir au moins un faux-positif. L'approche la plus courante consiste à contrôler le PCER qui garantit d'identifier $\alpha\%$ d'observations atypiques même s'il n'y a pas de valeur aberrante dans les données considérées. Becker et Gather (1999) et Cerioli *et al.* (2009) conseillent donc de commencer par tester de manière simultanée l'absence de valeurs aberrantes en contrôlant le niveau du test par des ajustements de type Bonferroni (Holm, 1979).

Ensuite, l'utilisation de paramètres empiriques tels que μ_n et Σ_n, sensibles à la présence de valeurs atypiques, est critiquée. En effet, ce choix de paramètres peut entraîner les effets bien connus de masque ou de débordement (*swamping*, en anglais), comme illustré sur la Figure 4 (Filzmoser et Todorov, 2011). Ce diagramme de dispersion représente les deux dimensions d'un jeu de données contenant 15 observations atypiques, représentées par les symboles « x » et « * ». Une observation est atypique si elle présente un comportement différent de la majorité des données, mais ne prend pas forcément des valeurs extrêmes sur chaque dimension. Les observations identifiées par des « + » ne sont pas considérées comme atypiques car bien qu'elles prennent des valeurs basses sur les deux dimensions, leur comportement est similaire à la majorité des données. De plus, deux ellipses de tolérance à 97,5% sont tracées, une en pointillé, calculée avec les estimateurs non robustes usuels et une en trait plein, obtenue avec les estimateurs robustes du $\mathrm{MCD}_{0.8}$. Les estimateurs du déterminant minimum MCD_α sont les estimateurs repondérés du déterminant minimum qui correspondent aux estimateurs empiriques non robustes usuels du sous échantillon contenant αn observations qui minimise le déterminant de la matrice de variance-covariance (Rousseeuw, 1986). Ces ellipses sont censées contenir 95% de la population la plus « centrale » d'une distribution normale et donc révéler les observations atypiques

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

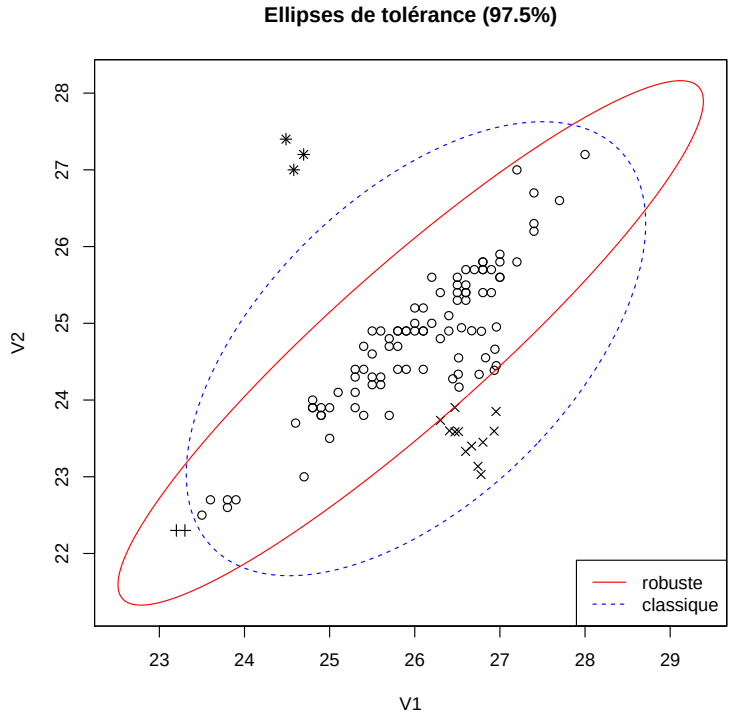

FIGURE 4. *Jeu de données bi-dimensionnel contenant plusieurs observations anormales. Deux ellipses de tolérance à 97.5% sont représentées pour essayer de révéler ces anomalies.*

comme des points extérieurs. En deux dimensions, elles illustrent parfaitement le fonctionnement de la distance de Mahalanobis calculée avec des estimateurs robustes ou non. Il apparaît ici que l'ellipse robuste est beaucoup plus plate que l'ellipse classique et qu'elles n'identifient donc pas les mêmes observations comme anormales. Dans le cas robuste, les 15 anomalies sont bien identifiées comme telles, alors que dans le cas classique, seules les trois « * » le sont et deux autres observations (les « + ») sont identifiées à tort comme atypiques. Ce phénomène est dû aux effets de masque (pour « x ») et de débordement (pour « + »).

L'effet de masque se produit lorsqu'un groupe de points distants de la majorité des observations attire les estimations de moyenne et de covariance, comme le font les observations « x ». La distance résultante des points périphériques par rapport à la moyenne devient alors petite et les observations ne peuvent pas être détectées comme atypiques, ce qui entraîne des faux-négatifs (atypiques détectés comme non-atypiques). En fait, les anomalies « x » masquent leur atypicité en influençant les valeurs des estimateurs de position et de dispersion.

L'effet de débordement, survient lorsqu'un groupe de points périphériques fléchit les estimations de moyenne et de covariance dans leur direction et les éloigne des autres points. La distance résultante entre les points initiaux et la moyenne est grande. Dans ce cas, le nombre de faux-positifs augmente, i.e. des observations non atypiques sont détectées comme atypiques comme l'illustrent les observations « + ».

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

Utiliser des estimateurs de position et de dispersion robustes permet de prévenir ces effets. Pour obtenir une distance de Mahalanobis robuste, il suffit de considérer des estimateurs de position et de dispersion robustes μ_R et Σ_R, comme illustré précédemment :

$$\mathrm{T}^2_{\mu_R,\Sigma_R}(\mathbf{x}_i) = \mathrm{RD}^2_{\mu_R,\Sigma_R}(\mathbf{x}_i) = (\mathbf{x}_i - \mu_R)'\Sigma_R^{-1}(\mathbf{x}_i - \mu_R) \tag{5}$$

Comme pour leur version non robuste, ces distances restent affines invariantes tant que les estimateurs considérés sont affines équivariants. Toutefois, il est important de noter qu'elles ne suivent pas exactement la même distribution.

Par exemple, il est possible d'utiliser des estimateurs du déterminant minimum (MCD), comme proposés par Rousseeuw et Van Zomeren (1990). Cette approche est encore actuelle et mise en pratique par certains dans le SPC (Jensen *et al.*, 2007). Toutefois, avec des estimateurs MCD, les distributions des distances présentées précédemment ne sont plus valides. Sous l'hypothèse de normalité, Hardin et Rocke (2005) donnent des arguments pour approximer la queue de distribution des distances de Mahalanobis calculées avec des estimateurs MCD par une loi de Fisher. Cerioli *et al.* (2009), Cerioli (2010) et Green et Martin (2017b) proposent d'ajuster le nombre de degrés de liberté de cette distribution de Fisher, dans le cas où le point de rupture [1] n'est plus de 50% ou bien lorsque le MCD est repondéré.

Enfin, même si la distance de Mahalanobis, dans sa version robuste ou non, a fait ses preuves au cours des années, elle a l'inconvénient de prendre en considération toutes les mesures. Or, d'après notre expérience dans le domaine industriel, il est possible que les pièces défectueuses ne se comportent anormalement que sur un sous-ensemble de mesures, ce qui implique de rechercher le sous-espace dans lequel elles se révèlent être atypiques. Une méthode très utilisée pour réduire l'espace initial est l'ACP.

2.3.2. ACP : Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) est un procédé bien connu pour réduire la dimension d'un ensemble de données corrélées. Elle est couramment appliquée en MSP multivariée comme l'expliquent Lafaye de Micheaux et Vieux (2005); Lafaye de Micheaux *et al.* (2007). L'idée clé est de transformer les variables initiales en des composantes principales (PC) non corrélées et ordonnées de sorte que les premières k composantes principales expliquent la plus grande partie de la variation des données initiales. Ce nouvel hyperplan k-dimensionnel est obtenu en minimisant l'erreur de projection au carré (SPE). Formellement, comme expliqué dans l'ouvrage de Jolliffe (2002), on diagonalise la matrice de variance-covariance Σ de dimension $p \times p$:

$$\Sigma = \mathbf{PDP}' \tag{6}$$

où $\mathbf{D}$ est une matrice diagonale contenant les valeurs propres $\gamma_1 \geq \cdots \geq \gamma_p$ de Σ, tandis que les colonnes de la matrice (orthonormée) $\mathbf{P}$ contiennent les vecteurs propres correspondants. Ces vecteurs contiennent en fait les coefficients de la combinaison linéaire des variables initiales permettant de construire les composantes. Les nouvelles coordonnées des observations sont donc

[1] Le point de rupture mesure la proportion de contamination qu'un estimateur peut tolérer sans se rompre, i.e. pour des estimateurs de position et de dispersion sans exploser ni imploser. Voir Droesbeke *et al.* (2015) pour une présentation plus détaillée.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

obtenues en projetant le tableau de données initiales $\mathbf{X}$ dans le sous-espace formé par les k premiers axes :

$$\mathbf{C} = \mathbf{X}\mathbf{P}_k \tag{7}$$

Nonobstant la simplicité apparente de la méthode, plusieurs difficultés se présentent pour identifier formellement les possibles observations atypiques.

Tout d'abord, le nombre k de composantes à retenir n'est pas un choix aisé. L'idée est de trouver le sous-espace qui réduit significativement le nombre d'axes à prendre en compte tout en expliquant une part importante de la variation des données initiales. Pour sélectionner au mieux ce nombre de composantes, plusieurs critères existent. Le premier consiste à retenir le nombre d'axes nécessaires pour expliquer au moins une fraction ϕ de l'inertie totale. Formellement les valeurs propres, qui correspondent aux variances des composantes, permettent de déterminer la part d'inertie ou de variance expliquée. Avec k composantes, le pourcentage cumulé de variance est donc :

$$PCV(k) = \frac{\sum_{l=1}^{k} \gamma_l}{\sum_{l=1}^{p} \gamma_l} \tag{8}$$

Il suffit donc de chercher le plus petit nombre k qui permette d'expliquer la fraction ϕ de l'inertie totale souhaitée :

$$k = \arg\min_{l} \{PCV(l) \geq \phi\} \tag{9}$$

Un autre critère consiste à analyser le graphique des valeurs propres, appelé « éboulis de valeurs propres », qui représente les valeurs propres γ_l par ordre décroissant. L'idée est de rechercher un « coude » dans cet éboulis de valeurs, qui serait suivi par une décroissance régulière. Cette décroissance est le signe que les axes associés à ces faibles valeurs ne sont pas pertinents pour l'analyse et que les axes associés ne permettent d'expliquer qu'une très faible part de la variation initiale.

Ensuite, utiliser une ACP pour détecter des anomalies n'est pas une tâche facile car les valeurs aberrantes peuvent être révélées sur les premiers et/ou les derniers axes. Ainsi, Hubert *et al.* (2005) ont introduit l'idée d'analyser un diagramme de dispersion qui représente deux scores : SD (*Score Distance*) et OD (*Orthogonal Distance*). Pour chaque observation $\mathbf{x}_i$, la distance SD est calculée dans l'espace formé par les k premières composantes sélectionnées et la distance OD dans l'espace orthogonal :

$$\begin{aligned} \mathrm{SD}^2_{\mu_n,\Sigma_n}(\mathbf{x}_i,k) &= ||\mathrm{diag}(\frac{1}{\sqrt{\gamma_1}},\ldots,\frac{1}{\sqrt{\gamma_k}})\mathbf{P}'_k(\mathbf{x}_i-\mu_n)||^2 \\ \mathrm{OD}^2_{\mu_n,\Sigma_n}(\mathbf{x}_i,k) &= ||(\mathbf{I}_d-\mathbf{P}_k\mathbf{P}'_k)(\mathbf{x}_i-\mu_n)||^2 \end{aligned} \tag{10}$$

Remarque 1. Si $k = p$ alors la distance $\mathrm{SD}^2_{\Sigma_n}(\mathbf{x}_i, k = p)$ est équivalente à la distance de Mahalanobis $\mathrm{MD}^2_{\mu_n,\Sigma_n}(\mathbf{x}_i)$.

Puis, une observation $\mathbf{x}_i$ est identifiée comme atypique si sa distance SD et/ou OD dépasse l'un des quantiles théoriques dérivés par Hubert *et al.* (2005) :

$$\mathrm{SD}(\mathbf{x}_i) > c_{p,1-\alpha} \text{ et/ou } \mathrm{OD}(\mathbf{x}_i) > (\mathrm{med}_{\mathrm{OD}} + \mathrm{MAD}_{\mathrm{OD}}\, z_{1-\alpha})^{3/2} \tag{11}$$

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

où $c_{p,1-\alpha}$ est le quantile d'une distribution du χ^2 à p degrés de liberté d'ordre $1-\alpha$, $z_{1-\alpha}$ est le quantile d'une distribution gaussienne d'ordre $1-\alpha$, med_{OD} représente la médiane des distances OD et MAD_{OD} la médiane des valeurs absolues des écarts à la médiane des distances OD (*Median Absolute Deviation*, en anglais) ajusté par un facteur (1,4826) pour rendre l'estimateur asymptotiquement normal : $\text{MAD}_{\text{OD}} = 1.4826\ \text{med}(|x - \text{med}_{\text{OD}}|)$.

Remarque 2. Avec cette méthode, il est nécessaire d'analyser les valeurs des distances SD et/ou OD d'une observation $\mathbf{x}_i$ car il n'est pas possible de déterminer en amont si l'observation va être atypique seulement dans l'espace des premières composantes principales (SD), dans l'espace orthogonal (OD) ou bien dans les deux.

Remarque 3. Cette méthode présente l'avantage de prendre en compte l'espace orthogonal aux premières composantes principales. Toutefois, contrairement à la distance de Mahalanobis, les distances SD et OD sont uniquement orthogonales invariantes, c'est-à-dire que les résultats diffèrent selon que les données sont préalablement standardisées par un changement d'échelle (division par l'écart-type notamment) ou pas.

Remarque 4. La distance SD est également référencée comme étant le calcul de la statistique du T^2 de Hotelling sur les k premiers axes principaux (cf Harkat *et al.* (2002); Hassan (2014)). La distance OD correspond quant à elle à la SPE, soit l'erreur de projection au carré.

Remarque 5. Comme pour la distance de Mahalanobis, il est possible de rendre robuste la méthode en diagonalisant un estimateur de dispersion robuste comme le MCD par exemple, au lieu de la matrice de la matrice de variance-covariance.

2.4. Évaluation des méthodes et contributeurs

2.4.1. Évaluation des méthodes

Afin de pouvoir comparer les différentes approches, il est important de définir certains critères de performance qui mesurent l'efficacité des méthodes. L'objectif est d'identifier les « vraies » valeurs aberrantes, c'est-à-dire les incidents de qualité du client (CQI) avérés dans le domaine industriel, tout en minimisant les fausses détections. Les résultats d'identification peuvent être de quatre types, résumés dans le Tableau 1, avec :

- TP : le nombre de vrais positifs, c'est-à-dire le nombre de CQI détectés comme valeurs aberrantes.
- FN : le nombre de faux négatifs, c'est-à-dire le nombre de CQI détectés comme valeurs non aberrantes.
- FP : le nombre de faux positifs, c'est-à-dire le nombre d'observations non CQI, détectées comme valeurs aberrantes.
- TN : le nombre de vrais négatifs, c'est-à-dire le nombre d'observations non CQI, détectées comme valeurs non aberrantes.

Généralement les critères suivants, qui sont utilisés dans la Section 2.5.2, sont également

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

TABLE 1. *Classification des résultats des méthodes de détection de valeurs aberrantes.*

Réalité \ Identification	Valeurs aberrantes	Valeurs non aberrantes
CQI	TP	FN
Non CQI	FP	TN

calculés afin de synthétiser les résultats de la classification :

$$\begin{aligned} \text{Sensibilité} &= \frac{TP}{TP+FN} \\ \text{Spécificité} &= \frac{TN}{FP+TN} \\ \text{1-Spécificité} &= 1-\frac{TN}{FP+TN}=\frac{FP}{FP+TN} \end{aligned} \tag{12}$$

D'un point de vue industriel, la sensibilité est considérée comme le taux de bonne détection (DR) et 1−Spécificité comme le taux de fausses alarmes (FAR). De façon optimale, le DR devrait être de 100 % et le FAR de 0 %. Dans la pratique, en fonction des industries, le taux acceptable de fausses alarmes est variable. Il est compris entre 1 à 2% dans l'automobile et il peut être un peu plus élevé dans l'industrie spatiale s'il permet d'augmenter significativement le taux de bonne détection.

2.4.2. Contributeurs ou signature des défauts

Un autre point très important dans l'industrie est la nécessité de pouvoir remonter à la cause du défaut. Les ingénieurs de tests ont besoin de savoir sur quelles mesures la puce s'est comportée de manière anormale. Dans le contexte présenté ici, seules des méthodes de type non-supervisé sont introduites, ce qui signifie qu'il n'est pas possible d'apprendre des données. Ces méthodes vont donc identifier sans distinction des observations prenant des valeurs extrêmes sur certains tests bruités et de « vraies » anomalies. L'ingénieur de test, en s'appuyant sur les mesures qui expliquent l'anormalité des observations, est quant à lui capable de faire la différence entre ces deux types d'atypiques. La combinaison de la puissance de l'analyse exploratoire des méthodes statistiques et de l'expertise des ingénieurs permet donc de maximiser le taux de détection tout en minimisant le taux de fausses alarmes.

À titre d'illustration, pour l'ACP, Mnassri *et al.* (2008) ont cherché à déterminer la contribution des mesures initiales aux distances SD et OD résultantes de l'analyse de données réelles issues du processus de fabrication de l'entreprise STMicroelectronics. Pour ce faire il ont cherché à déterminer dans quelle mesure chaque test influence les distances SD et OD (voir aussi Harkat *et al.* (2002) pour une approche alternative).

Dans la communauté informatique, la phase d'interprétabilité des défauts est connue sous le nom *Intensional Knowledge* (en anglais), comme expliqué dans Aggarwal (2013, 2017). Un des travaux les plus notables dans le domaine est celui de Knorr et Ng (1999). Toutefois, la communauté statistique s'intéresse également à cette problématique comme le montrent Debruyne *et al.* (2017) dans leurs recherches.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

2.5. Exemple réel de l'industrie des semi-conducteurs et mise en œuvre en R

Dans cette section nous souhaitons comparer les différentes méthodes de détection d'observations atypiques sur un exemple réel de l'industrie des semi-conducteurs à l'aide du logiciel R (R Core Team, 2017). Pour information, le domaine des semi-conducteurs fabrique des produits critiques, i.e. requérant une fiabilité maximale, puisque les conséquences d'un mauvais fonctionnement d'un composant électronique, par exemple dans la fabrication d'une automobile, peuvent être désastreuses et peuvent porter atteinte à la sécurité des personnes. Les résultats des tests effectués pour assurer la fiabilité de ces produits sont donc soumis à une sévère politique de confidentialité et réussir à avoir des données en accès libre est une véritable opportunité.

2.5.1. Mise en œuvre en R

De nombreux logiciels propriétaires mettent en œuvre les méthodes présentées précédemment. Les entreprises créent généralement leur propre logiciel afin de faciliter le traitement et la traçabilité des données des puces testées. Toutefois ces méthodes sont également disponibles sous le logiciel R. La fonction de base *mahalanobis* permet de calculer la distance de Mahalanobis dans ses versions robustes ou non, en précisant les estimateurs de position et de dispersion que l'on souhaite utiliser. Les quantiles des différentes distributions sont également disponibles afin de pouvoir déterminer les valeurs seuils permettant l'identification des observations anormales. Des ajustements de ces limites sont disponibles dans le package mvoutlier pour identifier des valeurs aberrantes seulement dans les queues de distribution. Enfin le package CerioliOutlierDetection permet de tester simultanément si des observations sont atypiques en se basant sur les distances de Mahalanobis calculées avec les estimateurs MCD. En ce qui concerne l'ACP, nous avons choisi de travailler avec le package rrcov qui calcule les distances SD et OD, ainsi que le package robustbase pour le calcul des estimateurs robustes.

2.5.2. Exemple réel : HTP

L'exemple choisi est un jeu de données de l'industrie des semi-conducteurs, nommé HTP, qui est disponible dans le package ICSOutlier. Il contient $n = 902$ pièces high-tech conçues pour des produits de consommation et $p = 88$ variables quantitatives. Ces variables correspondent à 88 mesures, principalement de température, de tension, de courant et de temps de réponse, qui sont effectuées pour assurer une haute qualité de production. Seules les pièces considérées comme fonctionnelles et qui ont été vendues sont présentes ici. Deux pièces se sont avérées être des incidents de qualité client (CQIs). Par conséquent, ces deux éléments peuvent être considérés comme des anomalies. Le but de l'analyse est de pouvoir déterminer si une méthode statistique aurait pu les identifier comme telles avant la vente.

Dans cette étude, on compare les résultats obtenus avec la distance de Mahalanobis et l'ACP. Les variantes robustes de ces méthodes multivariées sont également analysées en considérant les estimateurs MCD repondérés avec un point de rupture de 25%. L'ACP n'étant pas affine invariante, la question de la standardisation des données est pertinente. Toutefois, ici les 88 tests relèvent du même type de mesures et sont donc dans des unités similaires, ce qui ne rend pas la standardisation nécessaire. En ce qui concerne le choix du nombre de composantes à retenir,

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

les fonctions du package rrcov proposent une décision automatique basée sur deux critères : ne garder que des valeurs propres assez grandes par rapport à la première, $\gamma_l/\gamma_1 \geq 10^{-3}$ puis parmi celles-ci expliquer 80% de la variance totale soit $k = \arg\min_l\{PCV(l) \geq 0.8\}$. L'identification des observations atypiques est réalisée au niveau $\alpha = 2\%$ pour pouvoir déduire le taux de fausses alarmes (FAR). Le seuil des distances de Mahalanobis robustes a été ajusté par la technique de Green et Martin (2017b). Pour l'ACP, la détection basée sur les SD et les OD est effectuée au niveau $\alpha/2$ pour chacune des deux distances.

Les résultats sont illustrés sur la Figure 5. La première ligne fait référence aux méthodes non robustes et la deuxième ligne concerne les méthodes robustes. Les observations représentées en noir sont celles qui sont identifiées comme atypiques par la méthode considérée. Les symboles de type triangle permettent de repérer les deux observations de type CQI qu'il faut détecter.

Tout d'abord on s'aperçoit que toutes les méthodes permettent de détecter les deux CQIs avec un test à 2%. Toutefois le taux de fausses alarmes n'est clairement pas comparable entre les méthodes. Si on analyse plus en détail le graphique en haut à gauche représentant la distance de Mahalanobis de chaque observation, le taux de fausses alarmes est de 13.22%. Le CQI n°2 est clairement identifié comme observation anormale avec la distance la plus éloignée, alors que CQI n°1 est mélangé dans le reste des observations détectées comme atypiques. On aurait pu penser qu'utiliser des estimateurs robustes du MCD permettrait d'améliorer la détection puisqu'ils prémunissent contre les effets de masque ou de débordement, mais sur notre exemple, on constate l'effet inverse. Désormais le CQI n°2 n'est plus l'observation avec la distance la plus élevée et n'est donc plus aussi clairement identifié comme atypique. De plus, bien que le seuil utilisé pour l'identification des observations anormales soit ajusté, le taux de fausses alarmes a presque doublé par rapport à la version non robuste. Cet exemple illustre les limites de l'utilisation de la distance de Mahalanobis ou du T^2 de Hotelling et leurs versions robustes dans le domaine des semi-conducteurs. Ce phénomène peut s'expliquer d'une part par la nature même des observations atypiques, qui ne se comportent anormalement que dans un sous-espace et non pas forcément sur toutes les mesures. D'autre part, les analyses basées sur des estimateurs robustes de position et de dispersion se focalisent sur la partie la plus centrale des données (composée ici d'environ 75% des observations). Or dans cet exemple, la proportion d'anomalies est très faible (bien inférieure à 2%), il est donc peu probable que des effets de masque ou de débordement apparaissent. Les résultats de l'ACP en haut à droite semblent confirmer cette hypothèse. Avec seulement 3 composantes parmi les 88, les deux CQIs apparaissent comme atypiques à la fois sur les distances OD et SD, tout en gardant un taux de fausses alarmes relativement faible, inférieur à 3%. Toutefois, ici aussi l'utilisation de la version robuste ne permet plus d'identifier aussi aisément les deux CQIs et amènent à considérer presque le triple de fausses alarmes.

Avec ce petit exemple illustratif des phénomènes propres aux semi-conducteurs, on s'aperçoit que les méthodes les plus robustes ne sont pas forcément celles les plus performantes dans ce domaine. En effet, dans ce secteur, seul un faible pourcentage d'anomalies est présent et il n'est donc pas nécessaire de se prémunir des effets de masque ou de débordement qui sont peu probables. Toutefois, il est difficile de pouvoir généraliser ces résultats car très peu de données sont publiées pour des raisons de confidentialité ou de disponibilité de l'information des incidents de qualité client. En ce qui concerne les méthodes univariées, Moreno-Lizaranzu et Cuesta (2013) ont réussi à mener une des premières analyses globales. En se basant sur des données de production de l'entreprise Freescale (maintenant NXP), ils ont testé différentes méthodes sur plus

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

de 205 671 puces dont 26 CQIs. Ils montrent que, quand le nombre de mesures effectuées est de l'ordre de 1500, utiliser la méthode PAT mène à détecter 34.6% des CQIs mais en éliminant 40% de pièces, ce qui est contreproductif et moins efficace qu'une identification aléatoire des mauvaises pièces. De manière générale, les experts estiment qu'une des méthodes les plus performantes dans le domaine des semi-conducteurs est le PAT dynamique (voir Section 2.2.2) avec une détection maximale de 5% des CQIs pour un taux de fausses alarmes de moins de 1%. Ces

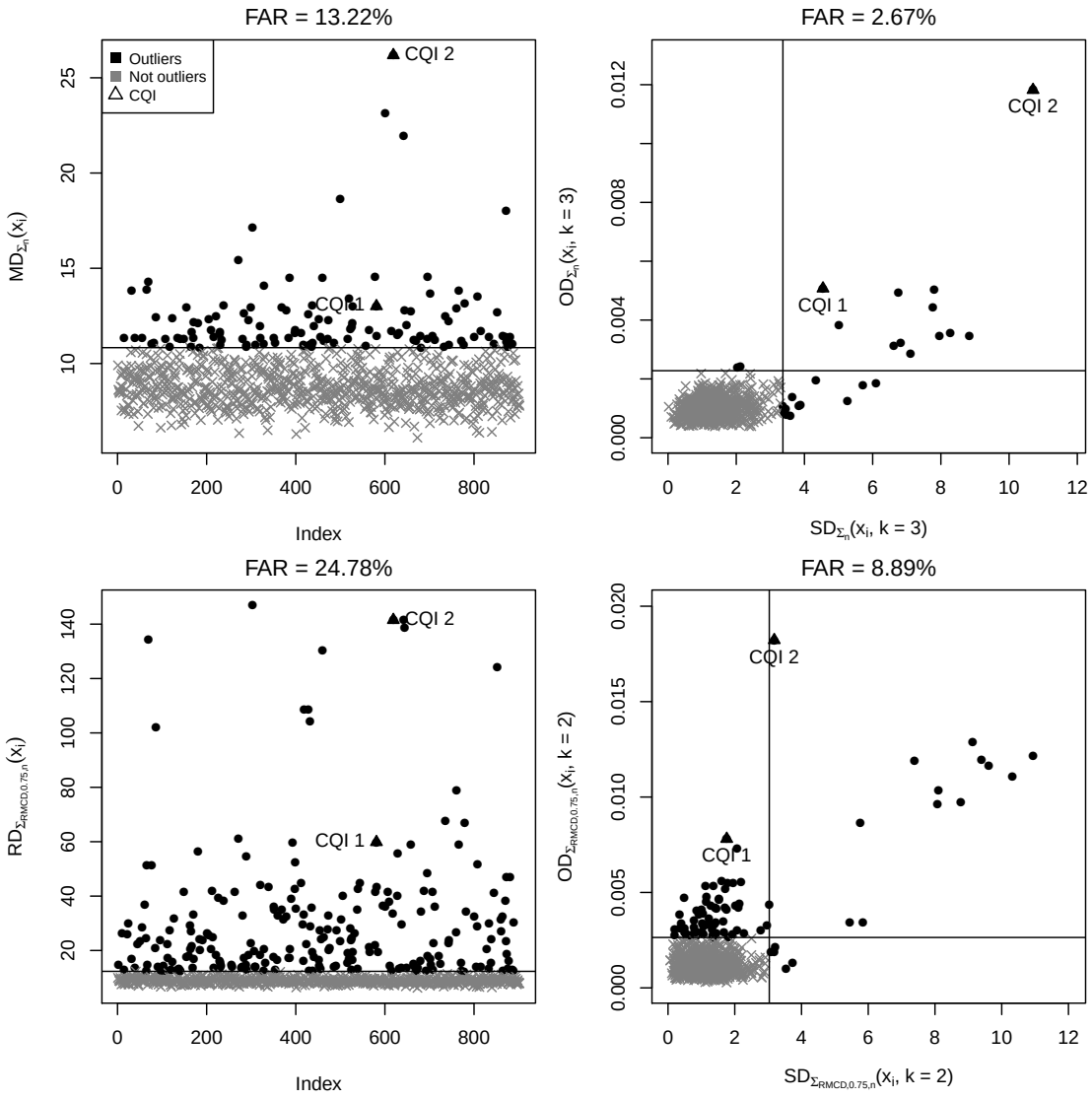

FIGURE 5. *Comparaison des résultats de détection des CQIs des données HTP. La première colonne représente les distances de Mahalanobis calculées, sur la première ligne de manière non robuste et, sur la seconde avec les estimateurs du MCD. La deuxième colonne donne les distances SD et/ou OD (cf Remarque 2) obtenues après une ACP non robuste sur la première ligne et robuste sur la seconde.*

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

résultats laissent donc une nette marge d'amélioration dans l'identification des anomalies.
Il est donc intéressant de dresser un état de l'art des méthodes de détection existantes de manière générale, sans se restreindre seulement à celles utilisées en pratique dans le domaine du contrôle de qualité.

3. Approches en dimension standard : $n > p$

De très nombreuses méthodes de détection d'atypiques existent, même en se limitant aux approches non supervisées seulement dédiées aux variables quantitatives. De manière générale, et non plus en se focalisant sur le contrôle de qualité, nous nous concentrons sur les méthodes de détection usuelles applicables lorsque la taille de l'échantillon est supérieure à sa dimension, soit $n > p$.

Dans la première section, nous rappelons les difficultés pour définir une observation atypique ainsi que sur les caractéristiques attendues d'une méthode de détection. Ensuite, en se basant sur l'ouvrage très complet d'Aggarwal (2017), nous synthétisons les différentes approches existantes dans la littérature statistique et informatique. Nous rajoutons également les caractéristiques importantes de chaque méthode ainsi que les packages R qui les mettent en œuvre.

3.1. Généralités

Tout d'abord, une observation atypique peut être définie comme étant une anormalité, une discordance, une déviation ou une anomalie. Toutes ces appellations ont en commun de renvoyer à un comportement qui semble être incompatible avec le reste de l'ensemble de données, comme défini dans Barnett et Lewis (1994). La subtilité cachée dans l'emploi de ces termes réside dans la caractérisation de l'observation atypique. Par exemple, Aggarwal (2017) explique que la définition d'atypicité est subjective et dépend du degré à partir duquel on considère qu'une déviation est assez importante par exemple. Généralement le terme d'observation « atypique » est utilisé pour définir n'importe quel comportement différent de l'ensemble du reste de la population, alors que le terme « anomalie » fait référence à quelque chose qui intéresse l'analyste, comme un incident de qualité dans le domaine des semi-conducteurs par exemple. Toutefois, dans le cadre des approches non-supervisées, la notion de déviation significativement intéressante n'est pas clairement définie. Les méthodes identifient donc seulement des observations atypiques sans faire la distinction entre celles qui sont de vraies anomalies ou celles produites à cause de mesures bruités.

Ensuite, pour caractériser les individus, les méthodes de détection retournent communément un score d'atypicité pour chaque observation (par exemple la distance de Mahalanobis ou les distances SD et OD pour l'ACP précédemment présentées en Section 2.3) et/ou une identification binaire des observations comme normale ou non. Ce score permet de pouvoir ordonner les individus en fonction de leur atypicité. L'identification formelle est généralement basée sur la détermination statistique de la valeur seuil du score à partir duquel les observations sont considérées comme atypiques. D'un point de vue pratique, cette étape est cruciale pour la prise de décision. Toutefois, le seuil statistique utilisé ne permet pas d'assurer que les observations signalées soient significatives d'une réelle anormalité.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

Enfin, nous pouvons dresser une liste de caractéristiques souhaitables pour les méthodes de détection. Tout d'abord, nous notons les propositions de Serfling et Mazumder (2013), à savoir : (i) la robustesse de la méthode en présence de valeurs aberrantes, (ii) l'invariance affine faible (i.e. qu'une transformation affine des coordonnées, si elle modifie les valeurs d'atypicité des observations, ne devrait pas affecter le classement relatif des scores d'atypicité), (iii) l'efficacité computationnelle dans n'importe quelle dimension pratique, et (iv) la non-imposition de l'hypothèse de distributions elliptiques. De plus, Cerioli (2010) fait également remarquer qu'il est pertinent de (v) tester l'absence d'observation atypique. Enfin, Aggarwal (2017) et Markou et Singh (2003) mettent en garde contre le nombre de paramètres à ajuster pour utiliser certaines méthodes et il peut donc être préférable de (vi) privilégier des méthodes avec un nombre de paramètres faible. En effet, chaque choix va avoir des conséquences sur la performance de la détection et, dans le cas d'approches non-supervisées, il est souvent très difficile de connaître le paramétrage optimal de la méthode.

En pratique, aucune méthode ne remplit toutes ces caractéristiques et il est donc difficile de pouvoir déterminer celle qui est la meilleure. Les comparaisons des différentes approches sont donc relatives à l'importance donnée à chacun des six critères décrits ci-dessus. Dans notre contexte industriel, nous considérons qu'il est important que la méthode optimale soit affine invariante pour se prémunir des changements d'échelle et idéalement qu'elle soit sans paramétrage et rapide d'exécution pour pouvoir être utilisée par des ingénieurs de tests non statisticiens. Idéalement, la méthode doit être en mesure de tester l'absence d'observation atypique. Par contre, on ne se focalise pas sur l'hypothèse de la distribution des données car il est courant que les propriétés théoriques requièrent une distribution elliptique alors qu'en pratique les méthodes tendent à fonctionner correctement même lorsque les hypothèses ne sont pas vérifiées (voir Hastie *et al.* (2001) Section 4.3 pour l'analyse discriminante par exemple). Enfin, la faible proportion d'anomalies, typique des données de semi-conducteurs, ne rend pas nécessaire l'utilisation de méthodes de détection très robustes, puisqu'il est peu probable que des effets de masque ou de débordement apparaissent.

3.2. Présentation succincte des méthodes

Dans cette section, nous proposons une synthèse des méthodes de détection d'observations atypiques. En plus des ouvrages mentionnés dans l'introduction (Hawkins, 1980; Barnett et Lewis, 1994; Aggarwal, 2013, 2017), de nombreuses revues de la littérature existent déjà dont notamment celles de Markou et Singh (2003); Venkatasubramanian *et al.* (2003); Hodge et Austin (2004); Rousseeuw et Leroy (2005); Agyemang *et al.* (2006); Chandola *et al.* (2007); Cateni *et al.* (2008); Chandola *et al.* (2009); Kriegel *et al.* (2010); Singh et Upadhyaya (2012); Zimek *et al.* (2014a); Pimentel *et al.* (2014). Toutes ces revues regroupent les différentes méthodes en plusieurs classes. Toutefois, cette classification n'est pas homogène en fonction des chercheurs et de leur communauté (statistique ou informatique). Nous avons choisi ici de suivre celle proposée récemment par Aggarwal (2017), qui distingue les approches en trois groupes : celles basées (i) sur un modèle probabiliste, (ii) sur la détermination d'un sous-espace et (iii) sur la notion de proximité.

L'idée n'est pas de refaire le travail exhaustif réalisé par Aggarwal (2017) dans son ouvrage, mais de s'appuyer sur celui-ci pour proposer un tableau synthétique des méthodes non-

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

supervisées existantes et leurs principales caractéristiques. Des références provenant principalement de la littérature statistique (liste non exhaustive) sont également ajoutées ainsi que les packages R mettant en œuvre ces méthodes. Il faut bien entendu noter que cette synthèse n'est qu'une image à un instant donné, car le développement de nouvelles méthodes et des packages R associés est particulièrement rapide. De plus, les méthodes considérées comme trop complexes sur le plan calculatoire ne sont pas présentées plus avant.

3.2.1. Méthodes basées sur un modèle probabiliste

Les approches de détection basées sur les modèles probabilistes peuvent être regroupées en trois grandes classes : les valeurs extrêmes univariées, multivariées et la modélisation probabiliste de mélanges.

Si la distribution des données est connue, il est possible d'analyser les valeurs extrêmes par des tests statistiques et d'identifier les observations qui se comportent véritablement de manière anormale. Cette approche, appelée théorie des valeurs extrêmes, est principalement univariée mais adaptable au cas multivarié. Par exemple, la distance de Mahalanobis, présentée en Section 2.3, qui calcule l'éloignement des observations par rapport au centre de la distribution sous l'hypothèse de loi elliptique, peut être considérée comme une méthode linéaire de détection par valeurs extrêmes multivariées. Cette distance peut également être appréhendée comme un cas particulier de mesure de profondeur. En effet, elle ordonne les observations de manière multivariée par rapport à un concept de centralité, ce qui est la définition même de la notion de profondeur. Une autre approche multivariée consiste à considérer une méthode hybride, nommée ABOD (*Angle-based Outlier Detection*), qui se base sur des angles et des distances. La méthode est considérée comme hybride car les angles formés par les observations sont inversement pondérés par la distance entre les points, comme expliqué plus en détail en Section 4.2.1.

Comme le type de distribution est généralement inconnu, on peut réaliser des tests d'adéquation afin de déduire la loi potentielle sous-jacente ainsi que d'estimer ses paramètres. Dans le cas multivarié, une procédure souvent utilisée est la modélisation par un mélange de gaussiennes à l'aide d'algorithme de type EM (pour *Expectation-Maximisation*), suivie du calcul de la probabilité des points d'appartenir à ce mélange. Toutefois, comme le nombre de paramètres augmente avec la complexité de la distribution sous-jacente des données, des problèmes de sur-ajustement peuvent se produire. De plus, les paramètres de ces modèles sont souvent difficiles à interpréter pour l'analyste. Or les experts du domaine ont besoin de déterminer si les observations détectées comme atypiques sont des anomalies liées à la fiabilité.

Le Tableau 2 permet de synthétiser les méthodes appartenant aux approches de détection présentées succinctement précédemment ainsi que de préciser leurs principaux avantages et inconvénients par rapport à nos attentes décrites en Section 3.1. Les packages R mettant en œuvre les différentes méthodes listées sont également mentionnés.

3.2.2. Méthodes basées sur la détermination d'un sous-espace

La deuxième approche telle que définie par Aggarwal (2017) concerne les méthodes basées sur les modèles linéaires de type régression ou détermination d'un sous-espace. Aggarwal (2017) interprète les modèles linéaires non pas avec une définition statistique classique mais en termes

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

TABLE 2. *Synthèse des méthodes de détection d'atypiques basées sur les modèles probabilistes et leurs propriétés.*

Caractéristiques des méthodes	Références
VALEURS EXTRÊMES UNIVARIÉES	
. Tests statistiques de discordance. . Test de Grubb's. . Tests en fonction de la distribution. . Règles du boxplot. . Cartes de contrôles. R : alphaOutlier, outliers, extremevalues	Barnett et Lewis (1994); Beckman et Cook (1983); Gao et Tan (2006); Grubbs (1950, 1969); Laurikkala *et al.* (2000); Rocke (1989, 1992); Tatum (1997); Vargas N. (2003)
VALEURS EXTRÊMES MULTIVARIÉES	
- **Distance de Mahalanobis (MD),** . Hypothèse de normalité requise. . Affine invariante. . Scores d'atypicité et test formel d'identification. . Pas de paramètres à ajuster. . Calcul peu complexe en $O(p^2)$. R : mvoutlier; rrcovHD	Becker et Gather (1999); Cerioli *et al.* (2009); Cerioli (2010); Geun Kim (2000); Gnanadesikan et Kettenring (1972); Laurikkala *et al.* (2000); Mardia *et al.* (1979); Rocke et Woodruff (1996); Wilks (1962)
T^2 de Hotelling, . Cartes de contrôles multivariées. . Equivalente à la distance de Mahalanobis.	Hotelling (1931); Hotteling (1947); Jensen *et al.* (2007); Mnassri *et al.* (2008); Sullivan et Woodall (1996); Vargas N. (2003)
Distance de Mahalanobis robuste . Test d'absence d'atypiques possible. . Cas particulier de profondeur. R : mvoutlier; CerioliOutlierDetection; faoutlier; robustX; rrcovHD	Billor *et al.* (2000); Campbell (1980); Cerioli *et al.* (2009); Cerioli (2010); Filzmoser *et al.* (2005); Hadi *et al.* (2009); Hardin et Rocke (2005); Jobe et Pokojovy (2015); Mardia *et al.* (1979); Maronna et Zamar (2002); Rousseeuw et Van Zomeren (1990); Serfling (1980)
- **Profondeur** . Calcul potentiellement très complexe. R : depth	Aggarwal (2013, 2017); Johnson et Wichern (1998); Kriegel *et al.* (2010); Ruts et Rousseeuw (1996)
- **Angles (ABOD)** . Approche hybride basée sur des distances et des angles. . Seulement des scores d'atypicité, pas d'identification formelle. . Calcul complexe en $O(n^3)$, mais des optimisations existent. . Adaptée si $n < p$. R : abodOutlier; HighDimOut	Kriegel *et al.* (2008); Campos *et al.* (2015); Kriegel *et al.* (2010); Pham et Pagh (2012); Radovanović *et al.* (2010); Laurikkala *et al.* (2000); Shyu *et al.* (2003); Hadi *et al.* (2009)
MODÉLISATION PROBABILISTE DE MÉLANGES	
. Nombreux paramètres à ajuster. . Calcul complexe.	Dempster *et al.* (1977); Gao et Tan (2006); Kriegel *et al.* (2011)

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

de variables latentes. Toutefois, comme nous nous intéressons exclusivement aux méthodes non-supervisées, nous ne présentons pas les méthodes de régression. Nous nous focalisons sur l'approche qui consiste à déterminer un sous-espace dans lequel le comportement atypique des observations est plus aisément identifiable.

Une des méthodes les plus connues est l'Analyse en Composantes Principales, présentée en Section 2.3.2, qui cherche à résumer l'information concernant la structure de variance-covariance des p variables dans les premières $k < p$ composantes principales. Une fois que les observations sont projetées dans ce sous-espace, celles qui se comportent différemment sont considérées comme atypiques. Cette approche est donc particulièrement efficiente dès lors que les données sont très corrélées. Bien que cette méthode ne soit pas affine invariante mais orthogonale invariante, elle est une référence très utilisée en pratique.

L'ACP peut également être considérée comme un cas particulier de projections révélatrices (ou *Projection Pursuit* en anglais). Ces méthodes ont été introduites par Friedman et Tukey (1974) et comme le présentent entres autres Hadi *et al.* (2009), l'idée est de trouver un sous-espace de projection de dimension inférieure, dans lequel la structure intéressante apparaît. Dans le contexte de la détection d'atypiques, les méthodes optimisent un indice caractéristique de la présence de valeurs aberrantes afin de trouver la projection qui met le plus en évidence ces observations. Dans le cas de l'ACP, cet indice est la maximisation de la variance de chaque composante. D'après Huber (1985), les méthodes de ce type ont l'avantage de ne se focaliser que sur les représentations contenant la structure des données et de faire abstraction du bruit. Toutefois, les algorithmes pour ces méthodes sont très coûteux en temps de calcul et nous ne les utiliserons pas. Une méthode de ce type sera toutefois citée lorsque nous présenterons la méthode ROBPCA en Section 4.2.4.

Aggarwal (2017), dans la Section 3.6 de son ouvrage, appréhende un perceptron, la forme la plus simple d'un réseau de neurones, comme une ACP. Il explique en détail comment définir ce réseau de neurones dans un cadre non-supervisée, nommé auto-encodeur, puisque la phase d'apprentissage est réalisée de manière non supervisée. Finalement, pour une certaine fonction d'activation, il conclut à l'équivalence entre ce réseau de neurones et l'ACP si $p-1$ composantes sont sélectionnées.

D'autres méthodes de factorisation de matrices que l'ACP existent et permettent de déterminer un sous-espace qui peut s'avérer intéressant pour mettre en évidence le comportement atypique de certaines observations. La méthode ICS (Tyler *et al.*, 2009) est une méthode de factorisation de matrices particulière, largement étudiée et adaptée pour la détection d'observations atypiques par Archimbaud *et al.* (2018); Archimbaud *et al.* (2018) ainsi que dans le travail de thèse de Archimbaud (2018) et mis en œuvre dans les packages R ICSOutlier et ICSShiny. Plus spécifiquement, la méthode ICS (*Invariant Coordinate Selection*) consiste à diagonaliser conjointement deux estimateurs de matrices de variances-covariances, ce qui permet de révéler la structure des données s'il y en a une. En présence d'observations atypiques, si on compare un estimateur robuste à un estimateur non-robuste, ceux-ci vont se différencier et faire émerger la ou les directions où se situent les observations atypiques. Ces directions correspondent aux premières et/ou aux dernières composantes (voir Archimbaud *et al.* (2018) pour plus de détails). Un score d'atypicité peut être calculé en mesurant la distance à l'origine des individus à partir de ces composantes. Enfin, contrairement à l'ACP, la méthode est affine invariante au sens que les composantes obtenues restent identiques lors d'une transformation affine des données.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

Quelques références sont données dans le Tableau 3 suivant. Ce tableau synthétise l'ensemble des méthodes, mentionnées ci-dessus, qui se basent sur la détermination d'un sous-espace.

3.2.3. Méthodes basées sur la notion de proximité

Contrairement à la détermination de sous-espaces contenant de l'information sur la structure des données, l'approche basée sur la proximité cherche des régions de l'espace dans lesquelles certaines observations vont être isolées. Ces individus sont ceux identifiés comme atypiques. Les principales méthodes présentées dans le Tableau 4 s'appuient sur la classification non supervisée (*clustering* en anglais), la densité ou les plus proches voisins.

La première méthode de classification non supervisée n'est pas dédiée à la détection d'observations atypiques, puisqu'elle permet de déterminer des groupes d'observations. Toutefois, elle peut s'avérer intéressante si on considère que les groupes contenant peu d'observations peuvent correspondre aux observations potentiellement atypiques. Par contre, seule une identification des observations atypiques est possible et aucune mesure d'atypicité n'est disponible.

À l'inverse, les méthodes basées sur la densité calculent un indice d'atypicité pour chaque observation mais ne permettent pas de dériver un seuil de manière théorique pour identifier celles qui sont atypiques. Une des méthodes les plus connues est le LOF (*Local Outlier Factor* en anglais) qui calcule le degré d'éloignement d'une observation à ses plus proches voisins en termes de densités locales. Toutefois, les calculs peuvent être complexes et longs, c'est pourquoi nous ne détaillons pas plus avant cette méthode.

Les méthodes basées sur les k plus proches voisins utilisent la distance de chaque observation à ses k plus proches voisins comme score d'atypicité. Plus la distance est importante, plus l'observation est isolée. Ces approches en terme de distances sont sans doute les plus utilisées pour leur facilité de mise en œuvre et leur interprétabilité en terme de variables initiales. Toutefois, les méthodes nécessitent généralement des temps de calcul assez importants.

Pour conclure, il apparaît qu'aucune méthode ne répond aux six caractéristiques décrites en introduction de la Section 3.2. En fait, chacune présente ses propres avantages et faiblesses dont il convient de tirer profit en fonction des applications. À partir de ce constat, un nouveau concept s'est développé : combiner les résultats de plusieurs méthodes afin d'améliorer la détection finale. Un vaste domaine de recherche est consacré à cet objectif, appelé *outlier ensembles* (Aggarwal et Sathe, 2017) ou méthodes d'ensemble en français. L'idée générale est d'appliquer différentes procédures de détection d'atypiques au jeu de données considéré, et de combiner les scores d'atypicité. Avant de pouvoir les agréger d'une quelconque manière, il convient de les normaliser car ceux-ci ne sont généralement pas comparables. En termes de fonction d'agrégation, la moyenne ou la maximisation sont les deux méthodes les plus couramment employées. Entre autres, Aggarwal (2017); Aggarwal et Sathe (2017); Dang *et al.* (2014); Gao et Tan (2006); Keller *et al.* (2012); Kriegel *et al.* (2011); Lazarevic et Kumar (2005); Müller *et al.* (2010b,a, 2011); Nguyen *et al.* (2010); Schubert *et al.* (2012); Zimek *et al.* (2012, 2013, 2014b) décrivent les principales méthodes utilisées dans la littérature. Ces approches sont relativement récentes car elles peuvent également résoudre les problèmes liés à la grande dimension, comme expliqué dans la section suivante.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

TABLE 3. *Synthèse des méthodes de détection d'atypiques basées sur la détermination d'un sous-espace et leurs propriétés.*

Caractéristiques des méthodes	Références
ANALYSE EN COMPOSANTES PRINCIPALES	
- **ACP** . Hypothèse : données dans un sous-espace de dimension $k < p$. . Seulement orthogonale invariante. . Choix du paramètre k. . Interprétation assez difficile. . Calcul peu complexe. R : rrcov	Barnett et Lewis (1994); Campbell (1980); Dutta *et al.* (2007); Filzmoser et Todorov (2013); Fujimaki *et al.* (2005); Gnanadesikan et Kettenring (1972); Jolliffe (2002); Mnassri *et al.* (2008); Parra *et al.* (1996); Rousseeuw et Leroy (2005)
- **ACP robuste** . Mêmes caractéristiques que l'ACP. R : rrcov; rrcovHD	Candès *et al.* (2011); Cardot et Godichon (2015); Devlin *et al.* (1981); Hubert *et al.* (2002, 2005); Kwitt et Hofmann (2006); Locantore *et al.* (1999); She *et al.* (2016); Shyu *et al.* (2003)
- **ACP non linéaire** (ex : ACP à noyau) . Adaptation possible de l'ACP pour des dépendances non linéaires.	Barnett et Lewis (1994); Rousseeuw et Leroy (2005)
PROJECTIONS RÉVÉLATRICES	
. Hypothèse : données dans un sous-espace de dimension $k < p$. . Choix de l'indice à optimiser. . Calcul complexe. R : rrcov; REPPlab	Maronna et Yohai (1995); Peña et Prieto (2001a,b); Croux et Ruiz-Gazen (2005); Ruiz-Gazen *et al.* (2010)
RÉSEAUX DE NEURONES : PERCEPTRONS ET AUTO-ENCODEURS	
. Cas particulier de l'ACP. . Calcul complexe.	An et Cho (2015); Aggarwal (2017)
FACTORISATION DE MATRICES	
- **ICS** . ACP Généralisée. . Affine invariante. . Calcul peu complexe. . Robustification possible. R : ICS; ICSOutlier; ICSShiny	Caussinus et Ruiz-Gazen (1990); Caussinus *et al.* (2003); Penny et Jolliffe (1999); Tyler *et al.* (2009); Bookstein et Mitteroecker (2014); Alashwali et Kent (2016); Archimbaud *et al.* (2018); Fischer *et al.* (2017); Archimbaud *et al.* (2018); Archimbaud (2018)
- **Autres méthodes** . Généralisation de méthodes comme l'ACP à d'autres fonctions objectifs. . Robustification possible. R : denoiseR	Farcomeni et Greco (2016); Josse et Sardy (2016); Josse *et al.* (2016a); Xiong *et al.* (2011)

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

TABLE 4. *Synthèse des méthodes de détection d'atypiques basées sur la notion de proximité et leurs propriétés.*

Caractéristiques des méthodes	Références
CLASSIFICATION NON SUPERVISÉE	
. Nombreux paramètres : choix du nombre de clusters, du modèle et de la méthode d'initialisation. . Calcul peu complexe. . Pas de scores d'atypicité. R : CrossClustering; kmodR	Duda *et al.* (2012); Jain et Dubes (1988); Rousseeuw et Kaufman (1990); Smith *et al.* (2002)
DENSITÉ	
- **LOF**	
. Hypothèse : la densité autour d'un atypique est différente de celle autour de ses voisins. . Détection d'atypiques locaux et globaux. . Choix du nombre k à ajuster. . Seulement une mesure d'atypicité, pas d'identification des atypiques. . Calcul souvent complexe. . Nombreuses variantes. R : DMwR2; Rlof	Breunig *et al.* (1999, 2000); Hadi *et al.* (2009); Tang *et al.* (2002); Zimek *et al.* (2012)
DISTANCES, KNN	
. Hypothèse : les atypiques ont un voisinage moins dense (ils sont éloignés de leurs voisins). . Facilement généralisable à différents types de données. . Calcul complexe en $O(n^2)$.	Bay et Schwabacher (2003); Campos *et al.* (2015); Ghoting *et al.* (2006); Knorr et Ng (1998, 1999); Pimentel *et al.* (2014); Ramaswamy *et al.* (2000); Rohlf (1975); Tao *et al.* (2006); Wu et Jermaine (2006)

4. Approches en grande dimension - faible taille d'échantillon (HDLSS) : $n < p$

Avec l'émergence des nouvelles technologies, mesurer des caractéristiques et les stocker est devenu de plus en plus facile, à tel point qu'aujourd'hui il est courant de devoir traiter des jeux de données avec plus de variables que d'observations. Ce contexte de « grande dimension - faible taille d'échantillon », plus connu sous son acronyme anglais HDLSS (*High dimension-low sample size*) intéresse fortement les communautés informatique et statistique. En effet, la majorité des méthodes de détection d'observations atypiques présentées dans la section précédente ne peuvent plus être appliquées principalement à cause du « fléau de la dimension ». Plus concrètement, ce phénomène, connu en anglais sous le nom de *curse of dimensionality*, regroupe différents challenges : l'effet de la concentration des distances, l'ajout d'attributs non pertinents pour la détection d'observations atypiques ou simplement les problèmes d'efficacité de certains algorithmes.

La première section caractérise plus précisément les différents challenges que l'on vient d'évoquer. Les sections suivantes se consacrent aux grandes approches utilisées dans ce contexte : l'analyse en dimension globale ou l'analyse en sous-espaces de l'espace originel. Dans ce travail, nous utilisons indistinctement les termes "HDLSS" et "grande dimension" pour désigner ce contexte particulier.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

4.1. Le fléau de la dimension

Tout d'abord, Beyer *et al.* (1999) ont mis en évidence un problème de « concentration des distances » ou de « concentration des mesures d'atypicité » avec l'augmentation de la dimensionnalité. En considérant des distances calculées avec une norme L^k et $k \geq 1$, ils montrent que toutes les paires de points de données deviennent presque équidistantes les unes des autres, et cela pour un grand nombre de distributions. En conséquence, en HDLSS, la notion de proximité ne s'applique plus car il n'est plus possible de discriminer des voisins comme proches ou lointains. De plus, sous l'hypothèse d'uniformité des données, Aggarwal *et al.* (2001) vont plus loin et démontrent que le problème de la pertinence de notions comme la proximité, la distance ou les plus proches voisins, est en fait sensible à la valeur de k lors de calculs de distances avec la norme L^k. Plus spécifiquement, ils stipulent que seules les normes L^1 et L^2, ne donnent pas des indices d'atypicité égaux pour toutes les observations. À partir de ce constat, ils suggèrent même d'utiliser une norme fractionnelle avec $0 < k < 1$.

Ce problème de concentration des distances n'apparaît que lorsque les variables ajoutées n'apportent pas d'information pertinente concernant l'atypicité des observations. Concrètement, si un individu se comporte de manière anormale sur toutes les mesures considérés alors on parle plutôt d'une bénédiction (*self-similarity blessing*) car l'information à retrouver est très présente dans les données initiales. Ce problème se confirme également si les attributs sont fortement corrélés entre eux. Toutefois, dans le contexte industriel, on constate généralement la situation inverse avec la présence d'une grande proportion d'attributs non pertinents pour détecter les anormalités. En conséquence, les données deviennent de plus en plus éparses (*sparse data*) car le nombre d'observations est faible comparé au nombre de dimensions, les mesures d'atypicités basées sur des distances sont donc faussées puisque pratiquement identiques. Néanmoins, Zimek *et al.* (2012) notent que le classement est toujours raisonnable même s'il devient presque impossible de choisir un seuil basé sur ces distances pour identifier les observations atypiques. En grande dimension, le principal challenge est donc de trouver le sous-espace d'attributs pertinents qui met en évidence le comportement anormal de certains individus. Or, avec l'accroissement de la dimensionnalité, le nombre de sous-espaces à considérer augmente exponentiellement, ce qui rend cette recherche très complexe.

Enfin, concernant la grande dimension, une croyance assez répandue est que chaque point dans un espace en grande dimension est un atypique. En fait, Zimek *et al.* (2012) expliquent que ce n'est pas exactement le cas. Ils précisent que pour chaque observation, il est toujours possible de trouver un sous-espace dans lequel celle-ci apparaît comme anormale. Il faut donc être vigilant à ce problème de biais de sollicitation de données (*data-snooping bias*) qui peut être associé à du sur-ajustement (*overfitting*) et doit être correctement traité, principalement dans le cas de procédure d'apprentissage.

À la vue de ces challenges, Aggarwal et Yu (2001) considèrent que pour traiter correctement des jeux de données en grande dimension, les méthodes de détection doivent avoir les caractéristiques suivantes :

— Gestion efficace des problèmes liés à la présence de données clairsemées.
— Interprétabilité de l'anomalie, i.e. la raison pour laquelle on peut dire que cette observation s'est comportée différemment.
— Comparabilité de la mesure d'atypicité. Une distance calculée dans un sous-espace de

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

dimension k n'est pas directement comparable à celle calculée dans un sous-espace de dimension $k+1$, par exemple.
— Calcul peu complexe, et cela même pour des problèmes de très grande dimension. Par exemple, les algorithmes basés sur une exploration combinatoire de l'espace ne sont pas efficaces.
— Prise en compte du comportement local des données pour déterminer si une observation est atypique ou pas.

À ces propriétés proposées par des chercheurs de la communauté informatique, on peut vouloir rajouter le test de l'absence d'observation atypique, comme dans le cas des données en dimension standard. Par contre, la propriété d'affine invariance des scores est difficile à obtenir lorsque l'on analyse des données parcimonieuses. Dans ce contexte, nous pouvons être amenés à relâcher notre exigence et à considérer des méthodes qui sont invariantes par transformation orthogonale. On peut aussi s'intéresser, comme le proposent Serfling et Mazumder (2013), à une affine invariance « faible » qui conserve seulement l'ordre des observations par rapport à une mesure d'atypicité.

Les revues de la littérature de Kriegel *et al.* (2010) et Zimek *et al.* (2012) suggèrent de classer les méthodes en deux grandes approches : les analyses basées sur l'ensemble des dimensions et celles basées sur la projection dans des sous-espaces de dimension inférieure. En suivant cette classification, nous présentons succinctement les méthodes qui semblent les plus pertinentes par rapport à notre contexte industriel et peu complexes en calculs.

4.2. Les analyses en dimension globale

Cette première approche, dite en dimension globale, rassemble des méthodes qui analysent l'espace dans sa dimension globale, i.e. sans avoir à traiter séparément des sous-espaces. Ce concept est le même que celui de la plupart des méthodes usuelles en dimension standard, à savoir celles basées sur la proximité, sur de la classification non supervisée, sur des distances ou sur l'ACP. On s'intéresse particulièrement aux adaptations des méthodes de type distance de Mahalanobis et ACP, qui semblent répondre le mieux à nos attentes dans le contexte industriel. On investigue également une autre démarche très utilisée en pratique qui consiste à pré-traiter les données en les projetant dans un sous-espace de dimension égale au rang des données, puis d'appliquer les méthodes usuelles de détection en dimension standard. Les autres méthodes sont quant à elles trop complexes en calculs pour notre application. Toutefois, on n'écarte pas l'approche basée sur les angles qui est capable d'analyser directement des données de type HDLSS.

4.2.1. Méthode basée sur les angles : ABOD

La méthode hybride ABOD (*Angle-based Outlier Detection*), brièvement introduite dans la Section 3.2.1 est une des seules méthodes qui ne nécessite aucune adaptation pour pouvoir être utilisée sur des données en grande dimension. La mesure d'atypicité est un facteur nommé ABOF (*Angle-Based Outlier Factor*) qui mesure la variabilité du spectre des angles formés à partir du point $\mathbf{x}$ avec l'ensemble des autres observations de l'espace. Ces angles sont inversement pondérés par la distance entre les points, ce qui rend la méthode hybride. Cette approche est souvent

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

considérée comme plus adaptée au cas de données en grande dimension que les méthodes calculant des distances. Toutefois, Radovanović *et al.* (2010) ont montré que les mesures basées sur les angles ne sont pas immunisées contre le fléau de la dimension en raison des effets de concentration dans la mesure du cosinus, ce qui impacte le spectre des angles. De plus, la distribution de l'indice d'atypicité est inconnue et il n'existe donc pas de règle d'identification des observations anormales. Enfin, la complexité des calculs en $O(n^3)$ est importante mais peut être réduite à l'ordre de $O(n^2)$ (Kriegel *et al.*, 2008) ou $O(n \log n)$ (Pham et Pagh, 2012).

4.2.2. Réduction de la dimension comme prétraitement

Une autre approche consiste à prétraiter les données en réduisant leur dimensionnalité, afin d'être ensuite en mesure d'appliquer les méthodes standards de détection d'observations atypiques. La décomposition en valeurs singulières, connue sous son acronyme anglais SVD (*Singular Value Decomposition*) est l'une des méthodes les plus populaires pour effectuer cette réduction (voir Schott (2005), Chapitre 4 pour une présentation détaillée). En quelques mots, cette procédure consiste à appliquer une transformation affine sur les données afin de les projeter dans un sous-espace de dimension inférieure ou égale au rang r des données. S'assurer de garder les r premières composantes de la décomposition garantit de ne pas perdre d'information, contrairement à l'ACP qui ne garde que les $k < r$ premières composantes. Il est alors possible d'appliquer les méthodes classiques de détection d'atypiques comme la distance de Mahalanobis ou l'ACP par exemple.

Parmi d'autres, Filzmoser *et al.* (2008) proposent différents algorithmes basés sur des composantes obtenues après une réduction de dimension. Nous présentons ici seulement l'algorithme Sign1 qu'ils proposent et qui est par exemple testé sur des données réelles par Archimbaud *et al.* (2018). Cette méthode se base sur l'ACP sphérique proposée par Locantore *et al.* (1999) qui consiste à normaliser les données de manière robuste et à les projeter sur une sphère avant de calculer les composantes principales robustes qui en découlent. Elle garde $r = \min(n-1, p-1)$ composantes et calcule l'inverse de la matrice de variance-covariance empirique afin de pouvoir déduire les distances de Mahalanobis de chaque observation dans le cas de la grande dimension.

Néanmoins, Branco et Pires (2015) remarquent que si le rang des données r est égal à $n-1$ et qu'on projette les données sur les $n-1$ premières composantes, alors les distances de Mahalanobis calculées à partir de ces nouvelles observations $\mathbf{x}_i^*$ sont constantes :

$$\mathrm{MD}^2_{\mu,\mathrm{COV}}(\mathbf{x}_i^*) = (\mathbf{x}_i^* - \mu)' \mathrm{COV}^{-1} (\mathbf{x}_i^* - \mu) = \frac{(n-1)^2}{n} \tag{13}$$

Tyler (2010) note également que dans le cas où les données se trouvent en position générale, alors tous les estimateurs de dispersion affine équivariants sont proportionnels à la matrice de variance covariance. Une solution envisageable est de ne projeter les données que dans un sous-espace de dimension inférieur au rang r, mais cela à condition d'accepter de perdre éventuellement de l'information sur la structure des données. À ce sujet, She *et al.* (2016) mettent en garde contre cette pratique dans le cas où les observations atypiques sont présentes dans le sous-espace orthogonal complémentaire car les anomalies peuvent ne pas être détectées.

Archimbaud (2018) discute également plus en détail des problèmes pouvant survenir lors de ce pré-traitement des données dans le Chapitre 4 de ses travaux de thèse. Plus spécifiquement,

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

l'auteure utilise ce pré-traitement avant d'appliquer la méthode ICS présentée en Section 3.2.2 et démontre que, dans ce cas, les scores obtenus ne sont plus affines invariants. Afin de préserver cette propriété intéressante, l'auteure propose également d'adapter directement la méthode ICS en décomposant en valeur singulière généralisée (GSVD) les deux estimateurs. Toutefois, cette approche n'est valide que dans le cas où les estimateurs sont semi-définis positifs affines équivariants et suppose que les données ne sont pas en position générale. Or, dans le contexte de la grande dimension ou en présence de variables colinéaires, les estimateurs, à l'exception de la matrice de variance-covariance, ne sont généralement qu'au mieux orthogonalement équivariants.

4.2.3. Adaptation de la distance de Mahalanobis

La distance de Mahalanobis est une méthode très utilisée en pratique au vu de ses nombreux avantages. Elle ne peut malheureusement pas être calculée dans le cas de données en grande dimension. En effet, l'estimateur de la matrice de variance-covariance est nécessairement singulier à partir du moment où le nombre de dimensions dépasse le nombre d'observations et ne peut plus être inversé. De nombreux chercheurs ont donc proposé des solutions pour adapter cette distance au cas HDLSS.

Une des solutions les plus simples consiste à déterminer l'inverse généralisée de l'estimateur de covariance ou réussir à définir un estimateur de la matrice de variance-covariance qui soit inversible, comme le proposent Ledoit et Wolf (2004, 2012). Sinon, il est également possible de régulariser d'autres estimateurs plus robustes de la variance-covariance de manière à obtenir une estimation de la dispersion qui soit inversible (voir entre autres Ollila et Tyler (2014); Verbanck *et al.* (2015); Ro *et al.* (2015)).

4.2.4. Adaptation de l'ACP

Certaines variantes de l'ACP présentent l'avantage d'être directement applicables sur des données HDLSS. Ces variantes regroupent les ACP basées sur des projections révélatrices (Filzmoser *et al.*, 2008; Croux *et al.*, 2007) et les ACP creuses (Bernard et Saporta, 2013; Croux *et al.*, 2013; Hubert *et al.*, 2016; Reynkens *et al.*, 2015; Shen et Huang, 2008; Zou *et al.*, 2006). Pour adapter la version classique de l'ACP, Ledoit et Wolf (2004, 2012); Ollila et Tyler (2014); Verbanck *et al.* (2015) proposent de considérer des estimateurs de dispersion régularisés, comme dans le cas de la distance de Mahalanobis.

Une autre approche, nommée ROBPCA, a été mise au point par Hubert *et al.* (2005) spécialement pour des données de type HDLSS. Cette méthode combine des idées de projections révélatrices et d'estimation robuste de la matrice de variance-covariance. Plus spécifiquement, les données sont projetées dans un sous-espace de dimension inférieure ou égale à $n-1$ grâce à une décomposition en valeurs singulières. Après cette réduction de dimension, l'idée est d'identifier un sous-ensemble de h observations les moins susceptibles d'être anormales selon l'estimateur de Stahel-Donoho, en se basant sur des projections révélatrices. La décomposition spectrale de la matrice de variance-covariance calculée à partir de ces h observations va permettre de décider du nombre de composantes à retenir dans la suite de l'analyse. Les données sont également projetées sur le sous-espace formé par les premiers vecteurs propres de la décomposition. Enfin, un

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

estimateur robuste de type MCD repondéré peut être utilisé pour calculer les composantes d'une ACP robuste. Les observations sont ensuite identifiées comme anormales à l'aide d'un diagnostic graphique qui prend en compte les distances SD et OD. Cette méthode, qui est devenue une référence dans le cas de données en grande dimension, est mise en œuvre dans le package rrcov. Quelques améliorations de l'algorithme ont été proposées dans Engelen *et al.* (2005). Les propriétés de robustesse et le comportement asymptotique de ROBPCA ont été étudiés par Debruyne et Hubert (2009).

4.3. Les analyses de sous-espaces de l'espace originel

À l'inverse des méthodes présentées précédemment, une nouvelle approche remet en question l'analyse des données dans l'espace global. En effet, les données de type HDLSS sont souvent contaminées par un nombre non négligeable de variables non pertinentes pour l'identification d'anormalités. On peut donc légitimement supposer que les atypiques sont davantage visibles dans un sous-espace local de dimension inférieure, constitué des attributs pertinents. Une analyse en pleine dimension va masquer ce comportement déviant en dimension inférieure. Les méthodes basées sur la proximité qui prennent en compte toutes les dimensions ne sont donc plus adaptées. L'idée principale est de découvrir le sous-espace qui permet d'identifier les observations anormales. Toutefois, ce n'est pas une tâche aisée.

Tout d'abord, la recherche de sous-espaces dans l'espace total est très complexe en calculs surtout lorsqu'une approche combinatoire est utilisée. Ensuite, il est souvent difficile de réussir à sélectionner le « bon » sous-espace. D'autant plus qu'Aggarwal (2017) constate que l'omission de certaines variables pertinentes a des effets plus graves sur l'efficience de la détection que l'inclusion de variables non pertinentes dans l'analyse. Pourtant, réussir à identifier les sous-espaces appropriés rend plus aisée l'interprétabilité des anormalités, principalement quand ceux-ci sont décrits en fonction des attributs originels.

Aggarwal (2017) consacre le Chapitre 5 de son ouvrage à passer en revue les différentes méthodes permettant de détecter les observations atypiques en se basant sur des sous-espaces de l'espace originel. Nous évoquons ici seulement les idées directrices de ces approches ainsi que celles présentées par Zimek *et al.* (2012) dans leur revue de la littérature. Les méthodes ne sont pas présentées plus avant au vu de leur complexité mais les détails techniques peuvent être trouvés dans les articles mentionnés.

4.3.1. Feature Bagging et méthodes d'ensembles

Une des approches les plus simples consiste à constituer des sous-espaces de $r < n < p$ attributs, à analyser ces sous-espaces avec les méthodes classiques de détection d'atypiques et à combiner les résultats avec des méthodes d'ensembles. Toutefois, plusieurs difficultés apparaissent. Elles concernent principalement trois aspects : (i) le choix du nombre de sous-espaces à considérer, (ii) la création des sous-espaces disjoints ou pas et (iii) l'agrégation et la normalisation des résultats. Le dernier point est particulièrement important si les scores d'atypicités sont calculés dans des sous-espaces de différentes dimensions. À titre d'exemple, le package HighDimOut met en œuvre la méthode présentée par Lazarevic et Kumar (2005) qui consiste à tirer à chaque itération

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

un échantillon aléatoire de $p/2 < r < p$ variables et à calculer des scores pour chaque observation à l'aide de la méthode LOF. La mesure d'anormalité finale est la somme cumulative de tous les scores obtenus à chaque itération.

4.3.2. Autres algorithmes

De très nombreux algorithmes ont été développés principalement par la communauté informatique. Ils se basent sur les notions classiques présentées à la Section 3.2, à savoir la distance entre observations, la densité, la classification non supervisée, les probabilités ou les projections révélatrices. Entre autres, on peut mentionner les méthodes HOS (*High-dimensional Outlying Subspaces*) de Zhang *et al.* (2004) ; SOD (*Subspace Outlier Degree*) de Kriegel *et al.* (2009) mise en œuvre dans le package HighDimOut ; HiCS (*High-Contrast Subspaces*) de Keller *et al.* (2012) ; OutRank (*Projected Clustering Ensembles*) de Muller *et al.* (2008, 2012) ; OUTRES (*Local Selection of Subspace Projections*) de Müller *et al.* (2010b, 2011) et COP (*Correlation Outlier Probability*) de Kriegel *et al.* (2012).

Par contre, comme en dimension standard, aucune de ces méthodes ne remplit toutes les conditions décrites en Section 4.1. Ces algorithmes souffrent en effet d'au moins un des trois inconvénients majeurs suivants : (i) la non normalisation des scores d'atypicité avant agrégation, (ii) une recherche pas assez extensive des sous-espaces, ce qui amène à ne pas pouvoir identifier les observations atypiques ou (iii) les calculs engendrés sont trop complexes.

En conclusion, il apparaît clairement que la détection d'observations atypiques dans le cas de la grande dimension est un sujet d'actualité particulièrement investigué par la communauté statistique comme informatique et présentant de nombreux challenges. Toutefois, jusqu'à présent, aucune méthode proposée ne permet de répondre à toutes les propriétés souhaitables mentionnées dans la Section 4.1, à savoir l'invariance par transformation affine ou orthogonale, le test d'absence d'observation atypique, l'interprétabilité des mesures d'atypicité ou encore la faible complexité de la méthode. Bien qu'il semble illusoire d'arriver à développer une méthode qui remplisse tous ces objectifs, Archimbaud (2018) consacre les Chapitres 4 et 5 de ses travaux de thèse à proposer différentes approches, se basant sur méthode ICS, qui permettent de satisfaire le plus grand nombre de ces critères dans le contexte de grande dimension ou en présence de données colinéaires.

5. Conclusion et perspectives

La première partie de cet article s'est focalisée sur le contrôle de la qualité dans le contexte industriel. Cette partie a mis en évidence que les standards du domaine automobile publiés par le comité du JEDEC (2009) (Joint Electron Device Engineering Council) ainsi que par NXP (Moreno-Lizaranzu et Cuesta, 2013) préconisent l'utilisation de méthodes univariées. Toutefois, comme ces approches ne sont pas satisfaisantes car elles engendrent trop de rejets pour un niveau de détection acceptable, certaines entreprises ont recours à des méthodes de détection multivariées de type distance de Mahalanobis ou ACP, mais qui ne remplissent pas non plus toutes leurs attentes. Après une étude extensive de la littérature des différentes approches non-supervisées de détection d'atypiques, nous concluons qu'aucune méthode ne remplit tous les critères attendus, à savoir être robuste en présence de valeurs aberrantes, invariante par transformation affine, non

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

basée sur l'hypothèse de distributions elliptiques, capable d'interpréter les anomalies ainsi que d'identifier un cas d'absence d'observation atypique, computationnellement efficace, requérant un nombre de paramètres faible et pouvant analyser des données de type HDLSS. Il est donc difficile de pouvoir mener une comparaison "juste" entre les différentes méthodes puisque les résultats vont dépendre du type de données. Le domaine d'application s'impose donc comme critère premier pour choisir parmi les différentes méthodes et conditionne les propriétés à privilégier.

En se focalisant sur le cas de la détection de défauts industriels telle que présentée dans cet article, la principale caractéristique du contexte est le faible pourcentage d'anomalies à identifier (généralement $< 2\%$) dont l'atypicité est contenue dans un sous-espace de petite dimension en comparaison du nombre total de variables. Comme les risques d'effets de masque ou de débordement sont peu probables, la robustesse de la méthode n'est donc pas une priorité. Par contre, l'affine invariance de la méthode est très importante car les données réelles analysées peuvent être dans des unités très différentes et l'échelle des valeurs ne doit pas affecter les résultats. De plus, comme le nombre de mesures sur les composants électroniques est en constante augmentation, il est nécessaire que la méthode soit peu complexe en terme calculatoire et qu'elle ne dépende pas d'un grand nombre de paramètres. Enfin, il faut également qu'elle soit en mesure d'analyser des jeux de données contenant des variables colinéaires ou en plus grand nombre que les observations. Pour répondre au mieux à ces spécificités, Archimbaud *et al.* (2018); Archimbaud *et al.* (2018) proposent, en dimension standard, une méthodologie principalement basée sur la méthode ICS qui est mise en œuvre en R à travers deux packages ICSOutlier; ICSShiny. Ils comparent les résultats obtenus avec certaines des approches présentées dans cet article et la méthode ICS sur des données réelles provenant de l'industrie. Dans ses travaux de thèse, Archimbaud (2018) adapte également la méthode à un contexte de grande dimension, dans le cas où des variables sont colinéaires ou en plus grand nombre que les observations, et l'applique à des données réelles issues de l'industrie spatiale.

Références

AGGARWAL, C. C. (2013). *Outlier Analysis*. Springer Publishing Company, Incorporated.

AGGARWAL, C. C. (2017). *Outlier Analysis, 2nd edition*. Springer Publishing Company, Incorporated.

AGGARWAL, C. C., HINNEBURG, A. et KEIM, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *In International Conference on Database Theory*, pages 420–434. Springer.

AGGARWAL, C. C. et SATHE, S. (2017). *Outlier Ensembles : An Introduction*. Springer.

AGGARWAL, C. C. et YU, P. S. (2001). Outlier detection for high dimensional data. *In ACM Sigmod Record*, pages 37–46. ACM.

AGYEMANG, M., BARKER, K. et ALHAJJ, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10(6):521–538.

ALASHWALI, F. et KENT, J. (2016). The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis*, 152:145–161.

AN, J. et CHO, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center - Special Lecture on IE*.

ARCHIMBAUD, A. (2018). *Statistical methods for outlier detection for high-dimensional data*. Thèse de doctorat, Université Toulouse 1 Capitole.

ARCHIMBAUD, A., MAY, J., NORDHAUSEN, K. et RUIZ-GAZEN, A. (2017). *ICSShiny : Invariant Coordinate Selection With a Shiny App*. R package version 0.5.

ARCHIMBAUD, A., NORDHAUSEN, K. et RUIZ-GAZEN, A. (2016). *ICSOutlier : Outlier Detection Using Invariant Coordinate Selection*. R package version 0.3-0.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

ARCHIMBAUD, A., NORDHAUSEN, K. et RUIZ-GAZEN, A. (2018). ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, 128:184 – 199.

ARCHIMBAUD, A., NORDHAUSEN, K. et RUIZ-GAZEN, A. (2018). ICSOutlier : Unsupervised Outlier Detection for Low- Dimensional Contamination Structure. *The R Journal*, 10(1):234–250.

AUTOMOTIVE ELECTRONIC COUNCIL (2011). Guidelines for part average testing. *AEC-Q001, rev-D*.

BARNETT, V. et LEWIS, T. (1994). *Outliers in Statistical Data*. Wiley.

BAY, S. D. et SCHWABACHER, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *In Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38. ACM.

BECKER, C. et GATHER, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94(447):947–955.

BECKMAN, R. J. et COOK, R. D. (1983). Outliers. *Technometrics*, 25(2):119–149.

BERNARD, A. et SAPORTA, G. (2013). Analyse en composantes principales sparse pour données multiblocs et extensiona l'analyse des correspondances multiples sparse. *In 45emes Journées de Statistique*.

BEYER, K., GOLDSTEIN, J., RAMAKRISHNAN, R. et SHAFT, U. (1999). When is "nearest neighbor" meaningful ? *In International Conference on Database Theory*, pages 217–235. Springer.

BILLOR, N., HADI, A. S. et VELLEMAN, P. F. (2000). BACON : blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34(3):279–298.

BOOKSTEIN, F. L. et MITTEROECKER, P. (2014). Comparing covariance matrices by relative eigenanalysis, with applications to organismal biology. *Evolutionary Biology*, 41(2):336–350.

BRANCO, J. A. et PIRES, A. M. (2015). High dimensionality : the trouble with Mahalanobis distance. WOMAT : Workshop On Multivariate Analysis Today.

BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T. et SANDER, J. (1999). Optics-of : Identifying local outliers. *In European Conference on Principles of Data Mining and Knowledge Discovery*, pages 262–270. Springer.

BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T. et SANDER, J. (2000). LOF : identifying density-based local outliers. *In ACM Sigmod Record*, pages 93–104. ACM.

CAMPBELL, N. A. (1980). Robust procedures in multivariate analysis I : Robust covariance estimation. *Applied Statistics*, 29(3):231–237.

CAMPOS, G. O., ZIMEK, A., SANDER, J., CAMPELLO, R. J., MICENKOVÁ, B., SCHUBERT, E., ASSENT, I. et HOULE, M. E. (2015). On the evaluation of unsupervised outlier detection : measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, pages 1–37.

CANDÈS, E. J., LI, X., MA, Y. et WRIGHT, J. (2011). Robust principal component analysis ? *Journal of the ACM (JACM)*, 58(3):11.

CARDOT, H. et GODICHON, A. (2015). Robust principal components analysis based on the median covariation matrix. *arXiv preprint arXiv :1504.02852*.

CATENI, S., VANNUCCI, M. et COLLA, V. (2008). *Outlier detection methods for industrial applications*. INTECH Open Access Publisher.

CAUSSINUS, H., FEKRI, M., HAKAM, S. et RUIZ-GAZEN, A. (2003). A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44(1):237–252.

CAUSSINUS, H. et RUIZ-GAZEN, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analyses. *In Proceedings of COMPSTAT'1990*, pages 121–126. Springer.

CERIOLI, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, 105(489):147–156.

CERIOLI, A., RIANI, M. et ATKINSON, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, 19(3):341–353.

CHALMERS, R. P. et FLORA, D. B. (2015). faoutlier : An R package for detecting influential cases in exploratory and confirmatory factor analysis. *Applied Psychological Measurement*, 39(7):573–574.

CHANDOLA, V., BANERJEE, A. et KUMAR, V. (2007). Outlier detection : A survey. Rapport technique, University of Minnesota.

CHANDOLA, V., BANERJEE, A. et KUMAR, V. (2009). Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

CINAR, A. et UNDEY, C. (1999). Statistical process and controller performance monitoring. a tutorial on current methods and future directions. *In Proceedings of the American Control Conference*, volume 4, pages 2625–2639. IEEE.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

CROUX, C., FILZMOSER, P. et FRITZ, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2):202–214.

CROUX, C., FILZMOSER, P. et OLIVEIRA, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225.

CROUX, C. et RUIZ-GAZEN, A. (2005). High breakdown estimators for principal components : the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

DANG, X. H., ASSENT, I., NG, R. T., ZIMEK, A. et SCHUBERT, E. (2014). Discriminative features for identifying and interpreting outliers. *In IEEE 30th International Conference on Data Engineering (ICDE)*, pages 88–99. IEEE.

DEBRUYNE, M., HÖPPNER, S., SERNEELS, S. et VERDONCK, T. (2017). Outlyingness : why do outliers lie out ? *arXiv preprint arXiv :1708.03761v1*.

DEBRUYNE, M. et HUBERT, M. (2009). The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness. *Statistics & Probability Letters*, 79(3):275–282.

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)*, 2:1–38.

DEVLIN, S. J., GNANADESIKAN, R. et KETTENRING, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362.

DROESBEKE, J.-J., SAPORTA, G. et THOMAS-AGNAN, C. (2015). *Méthodes robustes en statistique*.

DUDA, R. O., HART, P. E. et STORK, D. G. (2012). *Pattern classification*. John Wiley & Sons.

DUTTA, H., GIANNELLA, C., BORNE, K. et KARGUPTA, H. (2007). Distributed top-k outlier detection from astronomy catalogs using the demac system. *In Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 473–478. SIAM.

ENGELEN, S., HUBERT, M. et BRANDEN, K. V. (2005). A comparison of three procedures for robust PCA in high dimensions. *Austrian Journal of Statistics*, 34(2):117–126.

FAN, C. (2015). *HighDimOut : Outlier Detection Algorithms for High-Dimensional Data*. R package version 1.0.0.

FARCOMENI, A. et GRECO, L. (2016). *Robust methods for data reduction*. CRC press.

FILZMOSER, P., GARRETT, R. G. et REIMANN, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31(5):579–587.

FILZMOSER, P. et GSCHWANDTNER, M. (2015). *mvoutlier : Multivariate outlier detection based on robust methods*. R package version 2.0.6.

FILZMOSER, P., MARONNA, R. et WERNER, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

FILZMOSER, P. et TODOROV, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta*, 705(1):2–14.

FILZMOSER, P. et TODOROV, V. (2013). Robust tools for the imperfect world. *Information Sciences*, 245:4–20.

FISCHER, D., BERRO, A., NORDHAUSEN, K. et RUIZ-GAZEN, A. (2016). *REPPlab : R Interface to 'EPP-Lab', a Java Program for Exploratory Projection Pursuit*. R package version 0.9.4.

FISCHER, D., HONKATUKIA, M., TUISKULA-HAAVISTO, M., NORDHAUSEN, K., CAVERO, D., PREISINGER, R. et VILKKI, J. (2017). Subgroup detection in genotype data using invariant coordinate selection. *BMC Bioinformatics*, 18(1):173.

FRIEDMAN, J. H. et TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 100(9):881–890.

FUJIMAKI, R., YAIRI, T. et MACHIDA, K. (2005). An approach to spacecraft anomaly detection problem using kernel feature space. *In Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 401–410. ACM.

GAO, J. et TAN, P.-N. (2006). Converting output scores from outlier detection algorithms into probability estimates. *In ICDM'06 - Sixth International Conference on Data Mining*, pages 212–221. IEEE.

GENEST, M., MASSE, J.-C. et PLANTE, J.-F. (2012). *depth : Depth functions tools for multivariate analysis*. R package version 2.0-0.

GEUN KIM, M. (2000). Multivariate outliers and decompositions of Mahalanobis distance. *Communications in Statistics-Theory and Methods*, 29(7):1511–1526.

GHOTING, A., PARTHASARATHY, S. et OTEY, M. E. (2006). Fast mining of distance-based outliers in high-dimensional datasets. *In Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 609–613. SIAM.

GNANADESIKAN, R. et KETTENRING, J. R. (1972). Robust estimates, residuals, and outlier detection with multires-

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

ponse data. *Biometrics*, 28(1):81–124.
GREEN, C. G. et MARTIN, D. (2017a). *CerioliOutlierDetection : Outlier Detection Using the Iterated RMCD Method of Cerioli (2010)*. R package version 1.1.9.
GREEN, C. G. et MARTIN, R. D. (2017b). An extension of a method of Hardin and Rocke, with an application to multivariate outlier detection via the IRMCD method of Cerioli. Rapport technique, Working Paper, 2017. Available from http ://christopherggreen.github.io/papers/hr05_extension.pdf.
GRUBBS, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1):27–58.
GRUBBS, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.
HADI, A. S., IMON, A. et WERNER, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews : Computational Statistics*, 1(1):57–70.
HARDIN, J. et ROCKE, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946.
HARKAT, M.-F., MOUROT, G. et RAGOT, J. (2002). Différentes méthodes de localisation de défauts basées sur les dernières composantes principales. *In Conférence Internationale Francophone d'Automatique (CIFA)*.
HASSAN, A. H. (2014). *Détection multidimensionnelle au test paramétrique avec recherche automatique des causes*. Thèse de doctorat, Université Grenoble.
HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
HAWKINS, D. M. (1980). *Identification of outliers*, volume 11. Springer.
HODGE, V. J. et AUSTIN, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.
HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.
HOTELLING, H. (1931). The generalization of Student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.
HOTTELING, H. (1947). Multivariate quality control illustrated by the air testing of sample bombsites. *Selected Techniques of Statistical Analysis*, page 111.
HOWE, D. C. (2015). *kmodR : K-Means with Simultaneous Outlier Detection*. R package version 0.1.0.
HU, Y., MURRAY, W., SHAN, Y. et AUSTRALIA (2015). *Rlof : R Parallel Implementation of Local Outlier Factor (LOF)*. R package version 1.1.1.
HUBER, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2):435–475.
HUBERT, M., REYNKENS, T., SCHMITT, E. et VERDONCK, T. (2016). Sparse PCA for high-dimensional data with outliers. *Technometrics*, 58(4):424–434.
HUBERT, M., ROUSSEEUW, P. J. et VANDEN BRANDEN, K. (2005). ROBPCA : a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.
HUBERT, M., ROUSSEEUW, P. J. et VERBOVEN, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1):101–111.
JAIN, A. K. et DUBES, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.
JEDEC (2009). Outlier identification and management system for electronic components. Rapport technique, JEDEC, Replaced by JESD50.
JENSEN, W. A., BIRCH, J. B. et WOODALL, W. H. (2007). High breakdown estimation methods for phase I multivariate control charts. *Quality and Reliability Engineering International*, 23(5):615–629.
JIMENEZ, J. (2015). *abodOutlier : Angle-Based Outlier Detection*. R package version 0.1.
JOBE, J. M. et POKOJOVY, M. (2015). A cluster-based outlier detection scheme for multivariate data. *Journal of the American Statistical Association*, 110(512):1543–1551.
JOHNSON, R. A. et WICHERN, D. W. (1998). *Applied Multivariate Statistical Analysis (6th Edition)*. Prentice Hall.
JOLLIFFE, I. (2002). *Principal component analysis*. Wiley Online Library.
JOSSE, J. et SARDY, S. (2016). Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724.
JOSSE, J., SARDY, S. et WAGER, S. (2016a). denoiseR : A package for low rank matrix estimation. *arXiv preprint arXiv :1602.01206*.
JOSSE, J., SARDY, S. et WAGER, S. (2016b). *denoiseR : Regularized Low Rank Matrix Estimation*. R package version 1.0.
KELLER, F., MULLER, E. et BOHM, K. (2012). HiCS : High contrast subspaces for density-based outlier ranking. *In IEEE 28th International Conference on Data Engineering (ICDE)*, pages 1037–1048. IEEE.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

KNORR, E. M. et NG, R. T. (1998). Algorithms for mining distancebased outliers in large datasets. *In Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer.

KNORR, E. M. et NG, R. T. (1999). Finding intensional knowledge of distance-based outliers. *In VLDB*, volume 99, pages 211–222.

KOMSTA, L. (2011). *outliers : Tests for outliers*. R package version 0.14.

KRIEGEL, H.-P., KRÖGER, P., SCHUBERT, E. et ZIMEK, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 831–838. Springer.

KRIEGEL, H.-P., KROGER, P., SCHUBERT, E. et ZIMEK, A. (2011). Interpreting and unifying outlier scores. *In Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 13–24. SIAM.

KRIEGEL, H.-P., KROGER, P., SCHUBERT, E. et ZIMEK, A. (2012). Outlier detection in arbitrarily oriented subspaces. *In IEEE 12th International Conference on Data Mining (ICDM)*, pages 379–388. IEEE.

KRIEGEL, H.-P., KRÖGER, P. et ZIMEK, A. (2010). Outlier detection techniques. *Tutorial at KDD*, 10.

KRIEGEL, H.-P., ZIMEK, A. *et al.* (2008). Angle-based outlier detection in high-dimensional data. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 444–452. ACM.

KWITT, R. et HOFMANN, U. (2006). Robust methods for unsupervised PCA-based anomaly detection. *Proc. of IEEE/IST WorNshop on Monitoring, AttacN Detection and Mitigation*, pages 1–3.

Lafaye de MICHEAUX, D. (Juillet 2000). Prolonger la MSP par la "maîtrise globale du processus". *Qualité références*, pages 47–55.

Lafaye de MICHEAUX, D., CEMBRYNSKI, T., DALANCON, T. et DEMONSANT, J. (2007). Réduction de la dispersion des caracteristiques produit, méthodologie GPC et application en carrosserie automobile. *In 7ième édition du Congrès International Pluridisciplinaire Qualita 2007, Tanger (Maroc)*.

Lafaye de MICHEAUX, D. et VIEUX, D. (Janvier 2005). MSP multidimensionnelle, détecter et identifier "l'invisible". *Qualité références*, pages 79–82.

LAURIKKALA, J., JUHOLA, M., KENTALA, E., LAVRAC, N., MIKSCH, S. et KAVSEK, B. (2000). Informal identification of outliers in medical data. *In Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, volume 1, pages 20–24.

LAZAREVIC, A. et KUMAR, V. (2005). Feature bagging for outlier detection. *In Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 157–166. ACM.

LEDOIT, O. et WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

LEDOIT, O. et WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2).

LEE, J.-M., YOO, C., CHOI, S. W., VANROLLEGHEM, P. A. et LEE, I.-B. (2004a). Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 59(1):223–234.

LEE, J.-M., YOO, C. et LEE, I.-B. (2004b). Statistical process monitoring with independent component analysis. *Journal of Process Control*, 14(5):467–485.

LOCANTORE, N., MARRON, J., SIMPSON, D., TRIPOLI, N., ZHANG, J., COHEN, K., BOENTE, G., FRAIMAN, R., BRUMBACK, B., CROUX, C. *et al.* (1999). Robust principal component analysis for functional data. *Test*, 8(1):1–73.

MARDIA, K. V., KENT, J. T. et BIBBY, J. M. (1979). *Multivariate analysis*. Academic press.

MARKOU, M. et SINGH, S. (2003). Novelty detection : a review part 1 : statistical approaches. *Signal Processing*, 83(12):2481–2497.

MARONNA, R., MARTIN, R. D. et YOHAI, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

MARONNA, R. A. et YOHAI, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

MARONNA, R. A. et ZAMAR, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4).

MERCIER, S. et BERGERET, F. (2011). *Maîtrise Statistique des procédés - Principes et cas industriels*. Dunod/Usine Nouvelle.

MNASSRI, B., ANANOU, B., OULADSINE, M., GASNIER, F. *et al.* (2008). Détection et localisation de défauts des wafers par des approches statistiques multivariees et calcul des contributions. *In CIFA 2008, Conférence Internationale Francophone d'Automatique*.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

MORENO-LIZARANZU, M. J. et CUESTA, F. (2013). Improving electronic sensor reliability by robust outlier screening. *Sensors*, 13(10):13521–13542.

MULLER, E., ASSENT, I., IGLESIAS, P., MULLE, Y. et BOHM, K. (2012). Outlier ranking via subspace analysis in multiple views of the data. *In IEEE 12th International Conference on Data Mining (ICDM)*, pages 529–538. IEEE.

MULLER, E., ASSENT, I., STEINHAUSEN, U. et SEIDL, T. (2008). OutRank : ranking outliers in high dimensional data. *In ICDEW 2008, IEEE 24th International Conference on Data Engineering Workshop*, pages 600–603. IEEE.

MÜLLER, E., SCHIFFER, M., GERWERT, P., HANNEN, M., JANSEN, T. et SEIDL, T. (2010a). SOREX : Subspace outlier ranking exploration toolkit. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 607–610. Springer.

MÜLLER, E., SCHIFFER, M. et SEIDL, T. (2010b). Adaptive outlierness for subspace outlier ranking. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1629–1632. ACM.

MÜLLER, E., SCHIFFER, M. et SEIDL, T. (2011). Statistical selection of relevant subspace projections for outlier ranking. *In IEEE 27th International Conference on Data Engineering (ICDE)*, pages 434–445. IEEE.

NGUYEN, H. V., ANG, H. H. et GOPALKRISHNAN, V. (2010). Mining outliers with ensemble of heterogeneous detectors on random subspaces. *In International Conference on Database Systems for Advanced Applications*, pages 368–383. Springer.

NORDHAUSEN, K., OJA, H. et TYLER, D. E. (2008). Tools for exploring multivariate data : The package ICS. *Journal of Statistical Software*, 28(6):1–31.

OLLILA, E. et TYLER, D. E. (2014). Regularized-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070.

PARRA, L., DECO, G. et MIESBACH, S. (1996). Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation*, 8(2):260–269.

PEARSON, E. S. et SEKAR, C. C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28(3/4):308–320.

PEIRCE, B. (1852). Criterion for the rejection of doubtful observations. *The Astronomical Journal*, 2:161–163.

PEÑA, D. et PRIETO, F. J. (2001a). Cluster identification using projections. *Journal of the American Statistical Association*, 96(456).

PEÑA, D. et PRIETO, F. J. (2001b). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3).

PENNY, K. I. et JOLLIFFE, I. T. (1999). Multivariate outlier detection applied to multiply imputed laboratory data. *Statistics in Medicine*, 18(14):1879–1895.

PHAM, N. et PAGH, R. (2012). A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. *In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 877–885. ACM.

PIMENTEL, M. A., CLIFTON, D. A., CLIFTON, L. et TARASSENKO, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

R CORE TEAM (2017). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RADOVANOVIĆ, M., NANOPOULOS, A. et IVANOVIĆ, M. (2010). On the existence of obstinate results in vector space models. *In Proceedings of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 186–193. ACM.

RAMASWAMY, S., RASTOGI, R. et SHIM, K. (2000). Efficient algorithms for mining outliers from large data sets. *In ACM Sigmod Record*, pages 427–438. ACM.

REHAGE, A. et KUHNT, S. (2016). *alphaOutlier : Obtain Alpha-Outlier Regions for Well-Known Probability Distributions*. R package version 1.2.0.

REYNKENS, T., HUBERT, M., SCHMITT, E. et VERDONCK, T. (2015). Sparse PCA for high-dimensional data with outliers. *Technometrics*.

RIDER, P. R. (1933). *Criteria for rejection of observations*. Washington University Studies, St. Louis.

RO, K., ZOU, C., WANG, Z., YIN, G. *et al.* (2015). Outlier detection for high-dimensional data. *Biometrika*, 102(3): 589–599.

ROCKE, D. M. (1989). Robust control charts. *Technometrics*, 31(2):173–184.

ROCKE, D. M. (1992). X_Q and R_Q charts : Robust control charts. *The Statistician*, 41(1):97–104.

ROCKE, D. M. et WOODRUFF, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

ROHLF, F. J. (1975). Generalization of the gap test for the detection of multivariate outliers. *Biometrics*, 31(1):93–101.

ROUSSEEUW, P., CROUX, C., TODOROV, V., RUCKSTUHL, A., SALIBIAN-BARRERA, M., VERBEKE, T., KOLLER, M. et MAECHLER, M. (2016). *robustbase : Basic Robust Statistics*. R package version 0.92-6.

ROUSSEEUW, P. J. (1986). Multivariate estimation with high breakdown point. *In* GROSSMAN, W., PFLUG, G., VINCZE, I. et WERTZ, W., éditeurs : *Mathematical Statistics and Applications*, pages 283–297. Reidel, Dordrecht.

ROUSSEEUW, P. J. et KAUFMAN, L. (1990). *Finding Groups in Data*. Wiley Online Library.

ROUSSEEUW, P. J. et LEROY, A. M. (2005). *Robust regression and outlier detection*, volume 589. John wiley & sons.

ROUSSEEUW, P. J. et VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

RUIZ-GAZEN, A., MARIE-SAINTE, S. L. et BERRO, A. (2010). Detecting multivariate outliers using projection pursuit with particle swarm optimization. *In Proceedings of COMPSTAT'2010*, pages 89–98. Springer.

RUTS, I. et ROUSSEEUW, P. J. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis*, 23(1):153–168.

SCHOTT, J. R. (2005). *Matrix analysis for statistics*. Wiley.

SCHUBERT, E., WOJDANOWSKI, R., ZIMEK, A. et KRIEGEL, H.-P. (2012). On evaluation of outlier rankings and outlier scores. *In Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 1047–1058. SIAM.

SERFLING, R. et MAZUMDER, S. (2013). Computationally easy outlier detection via projection pursuit with finitely many directions. *Journal of Nonparametric Statistics*, 25(2):447–461.

SERFLING, R. J. (1980). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

SHE, Y., LI, S. et WU, D. (2016). Robust orthogonal complement principal component analysis. *Journal of the American Statistical Association*, 111(514):763–771.

SHEN, H. et HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034.

SHYU, M.-L., CHEN, S.-C., SARINNAPAKORN, K. et CHANG, L. (2003). A novel anomaly detection scheme based on principal component classifier. Rapport technique, DTIC Document.

SINGH, K. et UPADHYAYA, S. (2012). Outlier detection : applications and techniques. *International Journal of Computer Science Issues*, 9(1):307–323.

SMITH, R., BIVENS, A., EMBRECHTS, M., PALAGIRI, C. et SZYMANSKI, B. (2002). Clustering approaches for anomaly based intrusion detection. *Proceedings of Intelligent Engineering Systems Through Artificial Neural Networks*, pages 579–584.

STAHEL, W., MAECHLER, M. et potentially OTHERS (2013). *robustX : eXperimental Functionality for Robust Statistics*. R package version 1.1-4.

SULLIVAN, J. H. et WOODALL, W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28(4):398–408.

TANG, J., CHEN, Z., FU, A. W.-C. et CHEUNG, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 535–548. Springer.

TAO, Y., XIAO, X. et ZHOU, S. (2006). Mining distance-based outliers from large databases in any metric space. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 394–403. ACM.

TAOUALI, O., JAFFEL, I., LAHDHIRI, H., HARKAT, M. F. et MESSAOUD, H. (2016). New fault detection method based on reduced kernel principal component analysis (RKPCA). *The International Journal of Advanced Manufacturing Technology*, 85(5-8):1547–1552.

TATUM, L. G. (1997). Robust estimation of the process standard deviation for control charts. *Technometrics*, 39(2): 127–141.

TELLAROLI, P. et DONATO, M. (2016). *A Partial Clustering Algorithm with Automatic Estimation of the Number of Clusters and Identification of Outliers*. R package version 3.0.

TODOROV, V. (2016). *rrcovHD : Robust Multivariate Methods for High Dimensional Data*. R package version 0.2-4.

TODOROV, V. et FILZMOSER, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.

TORGO, L. (2016). *Data Mining with R, learning with case studies, 2nd edition*. Chapman and Hall/CRC.

TYLER, D. E. (2010). A note on multivariate location and scatter statistics for sparse data sets. *Statistics & Probability Letters*, 80(17):1409–1413.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238

TYLER, D. E., CRITCHLEY, F., DÃ¼MBGEN, L. et OJA, H. (2009). Invariant coordinate selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(3):549–592.

VAN DER LOO, M. (2010). *extremevalues, an R package for outlier detection in univariate data*. R package version 2.3.

VARGAS N., J. A. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35(4):367–376.

VENKATASUBRAMANIAN, V., RENGASWAMY, R., YIN, K. et KAVURI, S. N. (2003). A review of process fault detection and diagnosis : Part I : Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3): 293–311.

VERBANCK, M., JOSSE, J. et HUSSON, F. (2015). Regularised PCA to denoise and visualise data. *Statistics and Computing*, 25(2):471–486.

WILKS, S. (1962). *Mathematical Statistics*. John Wiley & Sons.

WU, M. et JERMAINE, C. (2006). Outlier detection by sampling with accuracy guarantees. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–772. ACM.

XIONG, L., CHEN, X. et SCHNEIDER, J. (2011). Direct robust matrix factorizatoin for anomaly detection. *In IEEE 11th International Conference on Data Mining (ICDM)*, pages 844–853. IEEE.

ZHANG, J., LOU, M., LING, T. W. et WANG, H. (2004). Hos-miner : a system for detecting outlyting subspaces of high-dimensional data. *In Proceedings of the Thirtieth International Conference on Very Large Data Bases*, volume 30, pages 1265–1268. VLDB Endowment.

ZIMEK, A., CAMPELLO, R. J. et SANDER, J. (2014a). Data perturbation for outlier detection ensembles. *In Proceedings of the 26th International Conference on Scientific and Statistical Database Management*, page 13. ACM.

ZIMEK, A., CAMPELLO, R. J. et SANDER, J. (2014b). Ensembles for unsupervised outlier detection : challenges and research questions a position paper. volume 15, pages 11–22. ACM.

ZIMEK, A., GAUDET, M., CAMPELLO, R. J. et SANDER, J. (2013). Subsampling for efficient and effective unsupervised outlier detection ensembles. *In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 428–436. ACM.

ZIMEK, A., SCHUBERT, E. et KRIEGEL, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, 5(5):363–387.

ZOU, H., HASTIE, T. et TIBSHIRANI, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

© Société Française de Statistique et Société Mathématique de France (2018) ISSN: 2102-6238