JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN-CLÉOPHAS ONDO
TAHA B. M. J. OUARDA
VINCENT FORTIN
BERNARD BOBÉE

# Procédures bayésiennes pour la détection d'observations singulières : synthèse bibliographique

*Journal de la société française de statistique*, tome 142, n° 2 (2001), p. 41-64

<http://www.numdam.org/item?id=JSFS_2001__142_2_41_0>

# PROCÉDURES BAYÉSIENNES POUR LA DÉTECTION D'OBSERVATIONS SINGULIÈRES : SYNTHÈSE BIBLIOGRAPHIQUE

Jean-Cléophas ONDO [1] *, Taha B.M.J. OUARDA [1],
Vincent FORTIN [2] et Bernard BOBÉE [1]

RÉSUMÉ

Le problème de la détection des données dites singulières, aberrantes, douteuses, erronées ou hors « norme » ('outliers') est posé par les collecteurs de données depuis de nombreuses années. Étant donné l'abondance de littérature et l'importance de ce thème, le présent article se propose de passer en revue les travaux à son sujet relevant de la statistique bayésienne. L'approche proposée est fondamentalement pratique. On montre l'optimalité des procédures bayésiennes de détection pour une large classe de lois *a priori* sur les paramètres. Par ailleurs, on élargit le cadre des travaux les plus usuels dans deux directions : d'une part on propose des tableaux résumant les conditions d'application des différentes procédures bayésiennes de détection en vue de faciliter leur usage, d'autre part on propose une classification de ces procédures de façon à permettre à l'utilisateur de se prémunir de l'effet de masque ('masking effect'), pour les méthodes de détection d'une observation singulière et de l'effet d'entraînement ('swamping effect'), pour les méthodes de détection de plusieurs observations singulières.

*Mots clés* — détection d'observations singulières, données singulières multivariées, données singulières univariées, effet d'entraînement ('swamping effect'), effet de masque ('masking effect'), observation influente, procédures bayésiennes, régression.

ABSTRACT

Increasing attention is being devoted to the problem of outlier detection. This paper presents a comprehensive review of the most prominent bayesian procedures which have been proposed in the literature for the detection of outliers. A practice oriented approach is adopted in order to enhance the use of bayesian techniques for outlier detection in practice. The paper intends to illustrate the optimality of bayesian procedures for the detection of outliers for a large class of prior distributions for the

*. Auteur correspondant

1. Chaire en Hydrologie Statistique, Institut National de la Recherche Scientifique, INRS-Eau, 2800 rue Einstein, C P. 7500, Sainte-Foy (Québec) G1V 4C7, E-mail : chaire_hydro@inrs-eau.uquebec.ca

2. Institut de Recherche d'Hydro Québec, 1800 boul Lionel Boulet, Varennes (Québec), J3X 1S1

parameters. The present paper proposes a set of summary tables underlining the hypothesis and the conditions of applicability of the various procedures identified. The paper presents also a classification of these procedures, allowing the user to get around the phenomenon of "masking" in the case of detection of a single outlier, and to avoid the effect of "swamping" in the case of detection of a group of two or several outliers.

*Key words* — detection of outliers, multivariate outliers, univariate outliers, swamping, influential observations, masking, bayesian procedures, regression.

# 1. INTRODUCTION

Recueillir de manière optimale et analyser adéquatement les données sont les deux objectifs fondamentaux de la statistique. Pris dans leur ensemble, l'une de ses tâches les plus importantes est l'épuration de données : vérification, contrôle et traitement des données singulières. Il est en effet essentiel de travailler sur des données fiables. Le statisticien doit disposer d'outils et de pistes permettant de valider la qualité des prélèvements. Il s'agit de vérifier, par des essais statistiques, que les données ne sont pas contaminées, c'est-à-dire que les échantillons ne comportent pas de données douteuses ou erronées, que le système dans lequel on a fait les relevés est stable ou encore que la distribution dans laquelle on a prélevé l'échantillon correspond à une distribution attendue (exemple la loi Gaussienne). La collecte des données est une étape importante de la démarche expérimentale et on doit y apporter le plus grand soin. Dans une série de mesures, il est cependant possible que l'on voie apparaître des valeurs s'écartant notablement des autres : ce sont les valeurs singulières. Mais on peut également trouver d'autres valeurs apparues dans l'échantillon à la suite d'une condition irrégulière (erreur de saisie ou d'expérimentation) : ce sont les valeurs erronées qui, elles ne font pas partie de la population : on les appelle *valeurs aberrantes.*

Le problème de la détection de données singulières a donné lieu à divers travaux dans la littérature statistique. Barnett et Lewis (1984) recensaient déjà près de 700 références bibliographiques sur le sujet. Un des problèmes posés par ces travaux est que les auteurs ne s'accordent sur aucune définition rigoureuse d'une donnée singulière. Par exemple, Edgeworth (1887) a écrit : « *Discordant observations may be defined as those which present the appearance of differing in respect to their law of frequency from other observations with which they are combined*fg. Quatre-vingt deux ans plus tard, Grubbs (1969) a écrit : « *An outlying observation, or outlier, is one that appears to deviate markely from the other members of the sample in which it occurs* ». Intuitivement, en statistique univariée, une observation singulière est un point situé loin dans les queues de la distribution. De nombreux types de singularités peuvent survenir en statistique multivariée comme le souligne Gnanadesikan (1977). Toutefois, si l'on s'en tient à la définition de Grubbs (1969) on peut dire qu'en situation multivariée, les données singulières se trouveront à la périphérie du nuage de points formé par l'échantillon. Dans l'effort d'éviter

certaines ambiguïtés, on adopte les définitions suivantes dans le reste de cet article :

- *observation discordante :* toute observation qui apparaît surprenante ou hors de propos du point de vue du décideur (souvent l'analyste de données ou l'expérimentateur) ;
- *contaminant :* toute observation qui n'est pas une réalisation de la distribution de probabilité de la population de base ;
- *observation singulière :* une observation qui est soit un contaminant, soit une observation discordante ;
- *observation aberrante :* une observation qui est apparue dans l'échantillon à la suite d'une condition irrégulière (erreurs de mesures, erreurs de transmission,...).

De ce point de vue, une observation discordante et une observation aberrante se situeront très éloignées du centre formé par le nuage central de l'échantillon de données. En revanche, un contaminant peut se situer dans le centre de l'échantillon mais on ne le verra jamais. De fait, un contaminant n'est pas nécessairement un 'outlier' au sens de Grubbs (1969) puisqu'il se définit sans faire référence aux autres observations de l'échantillon.

L'identification, et, si justifié, l'élimination rationnelle de ces valeurs est importante puisque ce type de valeurs peut avoir une très grande influence sur la construction d'un modèle statistique. Par exemple, les observations singulières exercent sur $\overline{X}$, la moyenne de l'échantillon, un effet de levier trop important et font de $\overline{X}$ un estimateur non fiable du centre de la population. Il existe plusieurs tests statistiques permettant d'identifier de telles valeurs. Ces tests vont permettre de s'assurer si une valeur suspecte a ou n'a pas une faible probabilité d'apparition, mais ils ne peuvent distinguer les valeurs singulières des valeurs aberrantes. Le présent travail traite la détection des données singulières. La caractéristique d'une donnée singulière est le degré de surprise qu'elle engendre dans un échantillon. Ce type de données peut renseigner sur le premier degré d'adéquation d'un modèle et ainsi contribuer à initier des travaux visant à l'amélioration ou à l'abandonner. C'est pourquoi il importe de s'intéresser aux procédures de détection de valeurs singulières et d'être très prudent dans l'utilisation systématique de telles techniques.

Pour sa part, la statistique inférentielle classique procède en deux temps. D'abord on repère les observations atypiques de l'échantillon. Puis, dans un second temps, on effectue un test d'hypothèse nulle $H_0$ (les observations atypiques repérées sont singulières) au niveau de signification $\alpha$ contre l'hypothèse alternative $H_1$ qui prétend le contraire. De fait, le statisticien peut justifier que les observations atypiques suspectes sont singulières avec une probabilité $\alpha$ de se tromper, choisie à l'avance. Lorsqu'on adopte ce point de vue, cher à la pensée et à la pratique d'un grand nombre de statisticiens en faveur de l'approche dite classique, on raisonne conditionnellement aux observations atypiques préalablement suspectées. En d'autres termes, on ne juge qu'un groupe d'observations donné. On ne regarde pas cependant si d'autres observations de l'échantillon sont susceptibles d'être singulières.

L'approche bayésienne est en grande partie redevable aux travaux pionniers de De Finetti (1961). De Finetti s'est surtout préoccupé de lancer les bases exploratrices de la question traitant de la détection des observations singulières plutôt que de développer une technique proprement dite. Son approche s'insère dans les fondements même des méthodes bayésiennes. Ces méthodes associent aux résultats expérimentaux des distributions de probabilité *a priori* sur les valeurs des paramètres. L'information « objective » apportée par l'expérimentation intervient pour modifier les distributions *a priori*, et les remplacer par des distributions *a posteriori*, qui concernent des probabilités conditionnelles (plus précisément, conditionnées par l'expérimentation). De Finetti (1961) argumente que, comme la distribution *a posteriori* dépend de toutes les données de l'échantillon, on peut donc étayer la détection d'observations singulières en exploitant toute l'information disponible sur cette distribution. De fait, toute observation de l'échantillon fait office de candidat potentiel pouvant être considéré comme observation singulière, selon que son influence sur la distribution *a posteriori* est faible ou pratiquement négligeable. Cela revient donc à considérer qu'une observation est singulière si sa présence dans l'échantillon a une influence sur la portée inférentielle des données. C'est le principe de base de détection énoncé par De Finetti (1961). Les travaux de De Finetti (1961) ont beaucoup influencé la littérature bayésienne de détection de données singulières. Le champ de recherche s'est élargi principalement avec les travaux de Box et Tiao (1968) et Freeman (1980) qui ont inspiré l'essentiel des procédures de détection présentées dans ce survol.

En somme, l'approche bayésienne de détection d'observations singulières est un peu plus flexible que sa compère classique. Elle consiste pour l'essentiel à considérer un modèle statistique de base comme un résumé du contenu informatif d'un ensemble de données; puis, à employer une variable aléatoire pour décrire l'incertitude relative aux valeurs du paramètre de ce modèle statistique et à en conditionner la distribution aux résultats de l'échantillonnage. La distribution *a posteriori* résultante intègre ainsi aussi bien les informations issues de l'échantillonnage que celles provenant de connaissances subjectives. Ce faisant, toute observation de l'échantillon peut être considérée comme étant une observation singulière, selon que son influence sur *cette distribution a posteriori* est faible ou pratiquement négligeable. Cette dernière approche semble donc plus naturelle et à ce titre plus facile à expliquer. En effet, elle renverse le point de vue classique en adoptant une approche directe conditionnelle aux données plutôt qu'un raisonnement indirect vis-à-vis d'une hypothèse nulle. Cette approche statistique non classique met donc en évidence une démarche qui peut éviter de désorienter certains praticiens, peu rompus aux raisonnements par l'absurde (on repère d'abord les observations atypiques, puis on vérifie à l'aide d'un test statistique si elles sont singulières) de l'approche classique et favoriser ainsi une plus large utilisation des techniques statistiques de détection d'observations singulières. D'où l'intérêt de s'en préoccuper dans ce travail.

On se propose dans le présent article de passer en revue les différentes procédures bayésiennes développées pour la détection de données singulières. L'objectif n'est certainement pas de présenter une revue bibliographique ex-

haustive sur ces procédures, mais de résumer la méthodologie afin de permettre au lecteur d'apprécier l'utilité de telles méthodes dans ce survol. Ce travail complète dans un certain sens, pour les procédures bayésiennes, la revue de littérature de Barnett et Lewis (1994) : il aborde plus en détail les procédures bayésiennes de détection d'observations singulières, et il présente un survol de développements récents sur le sujet. La suite de l'article distingue cinq sections. La section 2 présente une revue de littérature des différentes procédures bayésiennes de détection d'observations singulières dans un échantillon univarié. La section 3 fait le même développement pour le modèle de régression linéaire univarié tandis que la section 4 est une généralisation des sections 2 et 3 au cas multivarié. Dans chacune de ces sections des tableaux résumés sont présentés afin d'illustrer les techniques bayésiennes retenues. Quelques discussions font l'objet de la section 5, concernant en particulier les effets de masque et d'entraînement. Enfin, une conclusion est présentée dans la section 6.

## 2. DÉTECTION D'OBSERVATIONS SINGULIÈRES DANS UN ÉCHANTILLON ALÉATOIRE UNIVARIÉ

Soit $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$ un échantillon aléatoire univarié d'une loi $f(x/\theta_1; \theta_2)$ à valeurs dans $\mathbb{R}$, de moyenne (ou autre paramètre de localisation) $\theta_1$ inconnue et de variance (ou autre paramètre de dispersion) $\theta_2$ inconnue. L'étude des procédures bayésiennes de détection dans le cas univarié se situe dans le cadre de trois modèles spécifiques. Le premier suppose qu'une (ou plusieurs) valeur singulière provient d'un glissement de la moyenne, soit d'une loi de distribution $f(x/\theta_1 + \Delta; \theta_2)$, $\Delta \in \mathbb{R}$, $\Delta \neq 0$. Le deuxième modèle prétend qu'une (ou plusieurs) valeur singulière provient d'une dilatation de la variance, soit d'une loi $f(x/\theta_1; \lambda\theta_2)$ avec $\lambda \in \mathbb{R}$, $\lambda > 1$. Enfin, le troisième modèle suppose qu'une (ou plusieurs) valeur singulière provient à la fois d'un glissement de la moyenne et d'une dilatation de la variance.

### 2.1. Procédures bayésiennes de détection suivant le modèle de Guttman (1973) et Guttman et Khatri (1975)

On considère un échantillon gaussien unidimensionnel d'une loi normale $N(\mu, \sigma^2)$ à valeurs dans $\mathbb{R}$ de moyenne $\mu$ inconnue et de variance $\sigma^2$ supposée inconnue. L'étude de Guttman (1973) et Guttman et Khatri (1975) se situe dans le cadre des trois modèles énoncés plus haut. En d'autres termes, on suppose que s contaminants proviennent d'une loi normale gaussienne $N(\mu + a_s, \sigma^2)$, $N(\mu + a_s, \sigma^2/\delta_s)$ ou $N(\mu + a_s, \sigma^2/\delta_s)$. Le paramètre $\delta_s$ est supposé connu. L'inférence repose donc ici sur le paramètre de changement de moyenne, $a_s$. Les difficultés de la formulation d'une connaissance *a priori* sur les paramètres $\mu$, $\sigma^2$ et $a_s$ ont conduit à rechercher des distributions *a priori* « *non informatives* » ou « *peu informatives* ». Guttman (1973) et Guttman et Khatri (1975) ont considéré la distribution *a priori* non informative suivante :

$$p(\mu, \sigma^2, a_s) = p(\mu)p(a_s)p(\sigma^2) \propto (\sigma^2)^{-1} \qquad (1)$$

La distribution *a posteriori* du paramètre $a_s$ prend alors la forme :

$$p(a_s|\delta_s^2; x_1, \ldots, x_n) = \begin{cases} \sum_{i=1}^{n} c_i h(a_s|\eta_i; B^{(i)}; n-2) & \text{pour } s = 1 \\ \sum_{l=1}^{n}\sum_{j=1}^{n} c_{jl} h_2(a_s|\overrightarrow{\eta}_{jl}; S_{jl}; n-3) & \text{pour } s = 1, 2 \end{cases} \tag{2}$$

où,

- $c_i$ représente le poids de l'observation $x_i$ et $c_{jl}$ celui du couple $(x_l, x_j)$ ;
- $h$ est la densité d'une distribution de Student généralisée à $n-2$ degrés de liberté, de moyenne $\eta_i$ et de constante $B^{(i)}$ et $h_2$ est la densité d'une Student bivariée à $n-3$ degrés de liberté, de moyenne $\overline{\eta}_{jl}$ et de constante $S_{jl}$ ;
- pour $s = 1$, on a les formules :

  $\eta_i = n(n_i - \overline{x})/x - 1$ ; $B^{(i)} = \left[nA^{(i)}/(n-1)(n-2)\right]^{-1}$ ;

  avec : $A^{(j)} = \sum_{l \neq j}(x_l - \overline{x}^{(j)})^2$ ; $\overline{x} = \sum_{l=1}^{n} x_l/n$ ; et $\overline{x}^{(j)} = \sum_{l \neq j} x_l/n - 1$.

La distribution *a posteriori* de $a_s$ est une combinaison pondérée de densités de Student. Elle correspond précisément à l'incertitude sur l'amplitude du glissement de la moyenne $a_s$ lorsque la position du groupe de $s$ observations suspectes est connue dans l'échantillon. Ce faisant, $c_i$ le poids (respectivement $c_{jl}$) correspond à la probabilité que l'observation $x_i$ (respectivement le couple d'observations $(x_l, x_j)$) de l'échantillon soit un contaminant. Ainsi, on admettra que l'observation $x_i$ (respectivement le couple d'observations $(x_l, x_j)$) est un contaminant si son poids $c_i$ (respectivement $c_{jl}$) n'appartient pas à l'intervalle $\left[0, \frac{1}{n} + \frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$ (Guttman, 1973). Par ailleurs, on peut aussi juger tout l'échantillon en admettant la présence de contaminants dans l'échantillon si le rapport de probabilités *a posteriori* $\gamma_s$,

$$\gamma_s = \frac{p(a_s > 0|\delta_s^2; x_1, \ldots, x_n)}{1 - p(a_s > 0|\delta_s^2; x_1, \ldots, x_n)} \tag{3}$$

est supérieur ou égal à 5 ou inférieur ou égal à 1/5. En somme, la procédure de prise de décision est la suivante :

1. à l'aide du tableau 1, si on suspecte un contaminant (respectivement un couple de contaminants) dans l'échantillon, calculer tous les poids $c_i$ (respectivement $c_{jl}$) des $n$ observations (respectivement des $n(n-1)/2$ couples) de l'échantillon ainsi que $\gamma_s$ ;

2. déclarer que l'observation $x_i$ (respectivement le couple d'observations $(x_l, x_j)$) est un contaminant si $\gamma_s \notin \left]\frac{1}{5}, 5\right[$ et si $c_i$ (respectivement $c_{jl}$) $\notin \left[0, \frac{1}{n} + \frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$.

TABLEAU 1. — Résumé des procédures de Guttman (1973) et Guttman et Khatri (1975).

| Procédure | Observations suspectes | Poids | Valeur de $p(a_s > 0 \mid \delta_s^2; x_1, \ldots, x_n)$ |
|---|---|---|---|
| Guttman (1973) | $x_j \sim N(\mu + a, \sigma^2)$ $s=1$ | $c_j = \frac{(A^{(j)})^{(n-2)/2}}{\sum_{l=1}^{n}(A^{(l)})^{(n-2)/2}}$ | $\sum_{l=1}^{n} c_l G_{n-2}\left(\sqrt{B^{(l)}}\eta_l\right)$ ($G_{n-2}$ est la fonction de distribution de $t_{(n-2)}$) |
| Guttman et Khatri (1975) | $x_j \sim N\left(\mu + a\frac{\sigma}{\delta}, \sigma^2\right)$ $s=1$ $\delta^2$ fixé et $a>1$ | $c_j = \frac{(A^{(j)})^{(n-2)/2}}{\sum_{l=1}^{n}(A^{(l)})^{(n-2)/2}}$ | $\frac{1}{2}\sum_{l=1}^{n}\sum_{t=0}^{\infty} c_l(1-\rho_l^2)^{(n-1)/2}\rho_l^t \frac{\Gamma\left(\frac{n+t-1}{2}\right)}{\Gamma\left(\frac{t}{2}+1\right)\Gamma\left(\frac{n-1}{2}\right)}$ |
| Guttman et Khatri (1975) | $x_s \sim N\left(\mu + a_s, \frac{\sigma^2}{\delta_s^2}\right)$, $s=1,2$ ($a_1$ et $a_2$ non liés et $\delta_1$, $\delta_2$ fixés) | $c_{jl} = \frac{(A^{(jl)})^{(n-3)/2}}{\sum_{j=1}^{n}\sum_{l\neq j}(A^{(jl)})^{(n-3)/2}}$ | $\sum_{j=1}^{n}\sum_{l\neq j} c_{jl} G_{n-3}\left(\frac{(n-3)(n-2)\delta_s^2}{[A^{(jl)}(n-2+\delta_s^2)]}\left(x_s - \bar{x}^{(jl)}\right)\right)$ ($G_{n-3}$ est la fonction de distribution de $t_{(n-3)}$) |
| Guttman et Khatri (1975) | $x_s \sim N\left(\mu + a_s, \frac{\sigma^2}{\delta_s^2}\right)$, $s=1,2$ ($a_1 = -a_2 = a$ et $\delta_1$, $\delta_2$ fixés) | $c_{jl} = \frac{(u_{jl})^{-(n-2)/2}}{\sum_{j=1}^{n}\sum_{l\neq j}(u_{jl})^{-(n-2)/2}}$ | $\sum_{j=1}^{n}\sum_{l\neq j} c_{jl} G_{n-2}\left(\sqrt{f_{jl}}\eta(jl)\right)$ ($G_{n-2}$ est la fonction de distribution de $t_{(n-2)}$) |
| Guttman et Khatri (1975) | $x_s \sim N\left(\mu + a_s\frac{\sigma}{\delta_s}, \frac{\sigma^2}{\delta_s^2}\right)$, $s=1,2$ ($a_1$ et $a_2$ non liés et $\delta_1$, $\delta_2$ fixés) | $c_{jl} = \frac{(A^{(jl)})^{-(n-1)/2}}{\sum_{j=1}^{n}\sum_{l\neq j}(A^{(jl)})^{-(n-1)/2}}$ | $\frac{1}{2} + \sum_{j=1}^{n}\sum_{l\neq j} c_{jl}\sum_{m=0}^{\infty}\frac{2^{m-1/2}\Gamma\left(\frac{n}{2}+m\right)}{(2m+1)!\Gamma\left(\frac{n-1}{2}\right)} C_{2m+1}^{(n)}(j,l)[1-\varepsilon_{jl}]^{m-1/2}$ |
| Guttman et Khatri (1975) | $x_s \sim N\left(\mu + a_s\frac{\sigma}{\delta_s}, \frac{\sigma^2}{\delta_s^2}\right)$, $s=1,2$ ($a_1 = -a_2 = a$ et $\delta_1$, $\delta_2$ fixés) | $c_{jl} = \frac{(u_{s(jl)})^{-(n-1)/2}}{\sum_{j=1}^{n}\sum_{l\neq j}(u_{s(jl)})^{-(n-1)/2}}$ | $\frac{1}{2}\sum_{j=1}^{n}\sum_{l\neq j} c_{jl}\left\{\sum_{m=0}^{\infty}\frac{\Gamma\left(\frac{n}{2}+m\right)2^{2m}(1-\varepsilon_{s(jl)}^2)^{(n-1)/2}\varepsilon_{s(jl)}^{2m+1}\Gamma(m+1)}{\Gamma(2m+2)\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}}\right\}$ |

$$A^{(j)} = \sum_{l\neq j}(x_l - \bar{x}^{(j)})^2 \;;\; \bar{x}^{(j)} = \frac{\sum_{l\neq j} x_l}{n-1} \;;\; A^{(jl)} = \sum_{s\neq j,l}(x_s - \bar{x}^{(jl)})^2 \text{ et } \bar{x}^{(jl)} = \frac{\sum_{s\neq j,l} x_s}{n-2} \;;$$

$$u_{jl} = A^{(jl)} + \frac{\delta_j^2\delta_k^2(n-2)}{(n-2)(\delta_j^2+\delta_k^2)+4\delta_j^2\delta_k^2}\left[(x_j - \bar{x}^{(jl)}) + (x_k - \bar{x}^{(jl)})\right]^2 \text{ et } f_{jl} = \frac{(n-2)[(n-2)(\delta_j^2+\delta_k^2)+4\delta_j^2\delta_k^2]}{u_{jl}(n-2+\delta_j^2+\delta_k^2)} \;;$$

$$C_{2m+1}^{(n)}(j,l) = \frac{2^{m-1/2}}{\pi}\Gamma\left(m+\frac{3}{2}\right)\int_0^{+\infty}(\varepsilon_{jl} - t^2)^m\,dt, \nu_{s(jl)} = \frac{z_{s(jl)}}{\sqrt{1+\frac{\delta_s^2}{(n-2)}}} \quad \text{et} \quad u_{s(jl)} = A^{(jl)} + B_{jl} - cp_{jl}^2$$

## 2.2. Procédures bayésiennes de détection suivant le modèle de De Alba et Van Ryzin (1979)

On suppose que $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$ est un échantillon aléatoire univarié d'une loi normale $N(\mu, \sigma^2)$ à valeurs dans $\mathbb{R}$ de moyenne $\mu$ inconnue et de variance $\sigma^2$ supposée inconnue. L'approche de De Alba et Van Ryzin (1979) de détection de contaminants concerne deux modèles spécifiques. Le premier suppose que, parmi les $n$ observations de l'échantillon, $k$ d'entre elles proviennent d'un glissement de la moyenne $\mu$, soit d'une loi $N(\mu \pm \delta, \sigma^2)$ avec $\delta$ $(-\infty < \delta < +\infty)$ connu. Le deuxième modèle prétend que, parmi les $n$ observations de l'échantillon, $k$ d'entre elles proviennent d'une dilatation de la variance $\sigma^2$, soit d'une loi $N(\mu, \lambda\sigma^2)$ avec $\lambda \in \mathbb{R}$, $\lambda > 1$ inconnu. Dans un cas comme dans l'autre, De Alba et Van Ryzin (1979) traitent le problème de la détection des $k$ contaminants par la méthode *bayésienne empirique* : on considère $(X, \Lambda_1), \ldots, (X_n, \Lambda_n)$, $n$ paires mutuellement indépendantes de variables aléatoires, où $X_r$ $(r = 1, \ldots, n)$ est définie sur l'espace échantillonnal $\chi$ et $\Lambda_s$ $(r = 1, \ldots, n)$ sur l'espace paramétré $\Theta$ et on vérifie l'hypothèse $H_0 : \Lambda_r = 0$ (absence de contaminant) contre l'alternative $H_1 : \Lambda_r = \delta$ ou $\Lambda_r = -\delta$, tandis que dans le cas d'une dilatation de variance, on vérifie l'hypothèse $H_0 : \Lambda_r = 1$ (absence de contaminant) contre l'alternative $H_1 : \Lambda_r = \lambda > 1$. Dans tous les cas les $\Lambda_r$ $(r = 1, \ldots, n)$ sont supposés avoir une distribution commune *a priori* $G$ sur $\Theta$ et la densité conditionnelle de $X_r$ étant donné $\Lambda_r$ est $f_{\Lambda_r(x_r)}$, $r = 1, \ldots, n$. La règle de décision empirique se résume alors à calculer, pour chaque observation de l'échantillon, la probabilité $t_n^{(r)}(x_r)$ de ne pas la rejeter comme contaminant, conditionnellement aux données :

$$t_n^{(r)}(x_r) = \begin{cases} 0 & \text{si } \Delta_n^{(r)}(x_r) \geqslant 0 \text{ (l'observation } x_r \text{ n'est pas un contaminant)} \\ 1 & \text{si } \Delta_n^{(r)}(x_r) < 0 \text{ (l'observation } x_r \text{ est un contaminant)} \end{cases} \tag{4}$$

où,

- $\Delta_n^{(r)}(x) = \begin{cases} (1-p)L_0(\delta)\widetilde{f}_\delta(x) - pL_1(0)\widetilde{f}_0(x) & \text{pour un glissement de moyenne} \\ (1-q)L_0(\lambda)\widetilde{f}_\lambda(x) - qL_1(1)\widetilde{f}_1(x) & \text{pour une dilatation de variance} \end{cases}$
- $L_i(\lambda)$ est la perte encourue en prenant l'action $i$ alors que la vraie valeur du paramètre est $\lambda$ (même interprétation pour $L_i(\delta)$).

La procédure résultante de prise de décision est alors la suivante :

1. à l'aide du tableau 2, calculer, pour chaque observation $x_r$, $r = 1, \ldots, n$ de l'échantillon la quantité $\Delta_n^r(x_r)$ et déduire sa probabilité $t_n^{(r)}(x_r)$ ;
2. décider que l'observation $x_r$, $r = 1, \ldots, n$ est singulière si $t_n^{(r)}(x_r) = 1$.

TABLEAU 2. — Résumé des procédures de De Alba et Van Ryzin (1979).

| Procédure | Observations suspectes | Aide au calcul de $\Delta_n^{(r)}(x_r)$ |
|---|---|---|
| De Alba et Van Ryzin (1979) | $x_{i_j} \sim N(\mu \pm \delta, \sigma^2)$ $j = n-k+1, \ldots, n$ (c'est à dire $\Lambda_r = \delta$ *ou* $\Lambda_r = -\delta$) • $(i_1, \ldots, i_n)$ est une permutation des indices $(1, \ldots, n)$. | $\begin{cases} L_0(\delta) = L_0(-\delta) = L_1(0) = \textit{constante} \\ L_0(0) = L_1(\delta) = 0 \end{cases}$ • $\tilde{f}_s(x)$ est la densité d'une loi normale $N(\bar{x} + \delta^*, \tilde{\sigma}^2)$, $\delta^* = \delta, 0, -\delta$ • $\tilde{f}_1(x) = \frac{1}{2}\tilde{f}_{-\delta}(x) + \frac{1}{2}\tilde{f}_{\delta}(x)$ et, $\tilde{\sigma}^2 = \begin{cases} \sigma^{*2} \ \textit{si} \ \sigma^{*2} > 0 \\ \qquad \textit{avec} \quad \sigma^{*2} = s_2 - \frac{\delta^2}{6} + \frac{\sqrt{\max\{0, (\sigma^4 - 12s_4 + 36s_2^2)\}}}{6} \\ \frac{1}{n} \quad \textit{ailleurs} \end{cases}$ $p = \min\{1, p^*\}$, $p^* = \max\left\{0, \frac{(1 - s_2 - \tilde{\sigma}^2)}{\delta^2}\right\}$ $s_j = (1/n-1)\sum_{r=1}^{n}(x_r - \bar{x})^j$ |
| De Alba et Van Ryzin (1979) | $x_{i_j} \sim N(\mu, \lambda\sigma^2)$ $j = n-k+1, \ldots, n$ (c'est à dire : $\Lambda_r = \lambda > 1$) $(i_1, \ldots, i_n)$ est une permutation des indices $(1, \ldots, n)$. | $\begin{cases} L_0(\lambda) = L_1(1) = \textit{constante} \\ L_1(\lambda) = L_0(1) = 0 \end{cases}$ $\tilde{f}_x(x)$ est la densité d'une loi normale $N(\mu, \lambda^* \tilde{\sigma}^2)$, $\lambda^* = 1, \lambda$ et, $\tilde{\sigma}^2 = \frac{3(m_2 - \mu^2)(1+\lambda) + \sqrt{\max\{0, [(m_2 - \mu^2)^2 9(1+\lambda)^2 - 12\lambda(m_4 - 6\mu^2 m_2 + 5\mu^4)]\}}}{6\lambda}$ $q = \min\{1, q^*\}$, $q^* = \max\{0, [(m_2 - \bar{x} - \lambda\tilde{\sigma}^2)/(1-\lambda)\tilde{\sigma}^2]\}$ $\tilde{\mu} = \bar{x}$ *si* $\mu \neq 0$ $m_j = \frac{1}{n}\sum_{r=1}^{n} x_r^j$ |

## 2.3. Procédures bayésiennes de détection suivant le modèle de Pettit et Smith (1983, 1985)

Soit $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$ un échantillon aléatoire univarié d'une loi $f(x/\theta_1; \theta_2)$ à valeurs dans $\mathbb{R}$, de moyenne $\theta_1$ inconnue et de variance $\theta_2$ inconnue. Les procédures bayésiennes de Pettit et Smith (1983, 1985) supposent qu'une (ou plusieurs) valeur singulière provient soit d'un glissement de la moyenne, c'est-à-dire d'une loi de distribution $f(x/\theta_1 + \Delta; \theta_2)$, soit d'une dilatation de la variance, c'est-à-dire d'une loi $f(x/\theta_1; \lambda\theta_2)$. Le problème crucial dans ces deux types de modèles reste le choix de la distribution *a priori* pour les paramètres $\Delta$ et $\lambda$. En effet, si l'on fixe des *a priori* inadéquats sur ces paramètres, cela aura comme conséquence que la procédure de Bayes utilisée pourra conduire à des détections fallacieuses d'observations singulières (Freeman, 1980). C'est

TABLEAU 3. — Résumé des procédures suivant le modèle de Pettit et Smith (1983, 1985).

| Procédure | Observations suspectes | Facteur de Bayes $B_{0i}$ |
|---|---|---|
| Pettit et Smith (1983, 1985) | $x_i \sim N(\mu+\delta,\sigma^2)$ | $B_{01}=\frac{c_0}{c_1}\sqrt{\frac{n-1}{n}\left\{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i\neq i}(x_i-\bar{x}^{(i)})^2}\right\}^n}$ ; avec $\frac{c_0}{c_1}=\sqrt{\frac{3}{2}}$ |
| Pettit et Smith (1983, 1985) | $x_i \sim N(\mu+\delta_1,\sigma^2)$<br>$x_j \sim N(\mu+\delta_2,\sigma^2)$ | $B_{02}=\frac{c_0}{c_2}\sqrt{\frac{n-2}{n}\left\{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i\neq j}(x_i-\bar{x}^{(ij)})^2}\right\}^n}$ ; avec $\frac{c_0}{c_2}=\sqrt{2}$ |
| Pettit (1988) | $x_i \sim Exp(\theta\delta)$ | $B_{01}=\frac{c_0}{c_1}\frac{(n-1)x_i\left(\sum_{i=1}^{n}x_i\right)^{n-1}}{(n\bar{x})^n}$ ; avec $\frac{c_0}{c_1}=4$ |
| Pettit (1988) | $x_i \sim Exp(\theta\delta_1)$<br>$x_j \sim Exp(\theta\delta_2)$ | $B_{02}=\frac{c_0}{c_2}\frac{(n-1)(n-2)x_j x_i\left(\sum_{i\neq j}x_i\right)^{n-2}}{(n\bar{x})^n}$ ; avec $\frac{c_0}{c_2}=\frac{27}{2}$ |
| Pettit (1988) | $x_i \sim Exp(\theta\delta)$<br>$x_j \sim Exp(\theta\delta)$ | $B_{02}=\frac{c_0}{c_2}\frac{(n-1)(n-2)(x_j+x_i)\left(\sum_{i\neq j}x_i\right)^{n-1}}{(n\bar{x})^n}$ ; avec $\frac{c_0}{c_2}=\frac{27}{4}$ |
| Pettit (1994) | $x_{n-k+1},\cdots,x_n \sim P(\theta\delta)$<br>$(\delta>1)$<br>( $P$ : loi de poisson ) | $B_{0k}=\frac{c_0}{c_k}\left\{\frac{b+n-k+k\delta}{b+n}\right\}^{a+n\bar{x}}\frac{1}{\delta^{\left(a+\sum_{n-k+1}^{n}x_i\right)}}$ ; avec $\frac{c_0}{c_k}=1$;<br>$\theta \sim$ Gamma($a$, $b$) |
| Pettit (1994) | $x_i \sim P(\theta\delta)$<br>$(\delta>1)$<br>( $P$ : loi de poisson ) | $B_{01}=\frac{c_0}{c_1}\frac{\Gamma(a+n\bar{x})}{\Gamma(x_n!)\Gamma\left(a+\sum_{i=1}^{n-1}x_i\right)}\frac{(b+n-1)^{a+\sum x_i}}{(b+n)^{a+n\bar{x}}}$ ; avec<br>$\frac{c_0}{c_1}=\frac{1}{\frac{\Gamma(a+2m)(b+1)^{a+m}}{\Gamma(m)\Gamma(a+m)(b+2)^{a+2m}}}$ et $\theta \sim$ Gamma($a$ , $b$) ;<br>$m=\frac{a}{b}$ |

$\bar{x}^{(i)}=\frac{\sum x_i}{n-1}$ ; et $\bar{x}^{(ij)}=\frac{\sum x_i}{n-2}$ ;

ainsi que, dans les procédures bayésiennes précédentes, ces *a priori* étaient supposés connus, ce qui a beaucoup facilité leur développement.

Pour contourner cette difficulté, Pettit et Smith (1983, 1985) proposent de rechercher les valeurs singulières de l'échantillon à l'aide d'un critère bayésien : le *facteur de Bayes*. Formellement, si l'on suppose qu'on possède deux modèles,

$M_0$ et $M_1$ de la forme suivante :

$$M_0 : x_1, \ldots, x_n \sim \xi(x/\mu) \quad \text{(une densité unimodale)}$$
$$M_1 : x_1, \ldots, x_{n-1} \sim \xi(x/\mu), \quad x_n \sim \xi(x/\mu + \delta)$$

(on suppose que l'observation suspecte $x_n$ est apparue dans l'échantillon à la suite d'un glissement de la moyenne), le facteur de Bayes correspondant s'exprime alors comme suit :

$$B_{01} = \frac{p(M_0|x_1, \ldots, x_n)}{p(M_1|x_1, \ldots, x_n)} = \frac{c_0}{c_1}\, \vartheta(x) \tag{5}$$

où $\vartheta$ est une fonction des données de l'échantillon et $c_0/c_1$ un rapport de constantes inconnues dont la valeur est déterminée à l'aide de la technique des observations imaginaires proposée par Spiegelhalter et Smith (1982). En pratique, Pettit (1992) suggère qu'une observation singulière (ou un groupe) existe dans l'échantillon si le facteur de Bayes associé est compris entre les valeurs 0.005 et 0.015. Ce faisant, la prise de décision se fait comme suit : pour $i = 1, \ldots, n$, on admettra que l'observation $x_i$ (ou un groupe d'observations) est singulière si le facteur de Bayes correspondant, calculé à l'aide du tableau 3, est tel que $B_{0i} \in [0.005, 0.015]$.

## 2.4. Procédures bayésiennes de détection suivant le modèle de Kitagawa (1984)

Soit $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$ un échantillon aléatoire d'une loi normale $N(\mu_0, \sigma^2)$ à valeurs dans $\mathbb{R}$, de moyenne $\mu_0$ inconnue et de variance $\sigma^2$ inconnue. Le modèle de Kitagawa (1984) suppose que $m$ valeurs singulières sont générées par des classes de modèles distinctes caractérisées par des glissements de la moyenne, c'est-à-dire des distributions normales $N(\mu_i, \sigma^2)$ avec $\mu_i = \mu_0 + \Delta_i$ et $i = 1, \ldots, k$. Les paramètres $k$, $m$ et $\sigma^2$ sont supposés inconnus. Par ailleurs, on suppose que $J = (j_1, \ldots, j_n)$ est un vecteur de $n$ variables aléatoires indiquant la provenance de chacune des $n$ observations de l'échantillon : $j_i = 0$ indique que l'observation $x_i$ a été générée par le modèle de base $N(\mu_0, \sigma^2)$, alors que $j_i \in \{1, \ldots, k\}$ indique que l'observation $x_i$ est un contaminant qui a été généré par la loi normale $N(\mu_j, \sigma^2)$. L'approche proposée par Kitagawa (1984) est principalement basée sur la *vraisemblace prédictive* d'un modèle de Bayes. Le résultat est assez intuitif, puisque celle-ci peut apparaître à l'utilisateur facile à appréhender, car elle est relative aux quantités observables. En posant $\theta = (\mu_0, \mu_1, \ldots, \mu_k, \sigma^2)$, la méthode proposée est la suivante : d'abord on utilise une distribution *a priori* non informative pour $\theta$ :

$$p(\theta) = p(\mu_0)p(\mu_1)\ldots p(\mu_k)p(\sigma^2) \propto \sigma^{-2} \tag{6}$$

ensuite on calcule la distribution *a posteriori* correspondant au $J^{ième}$ vecteur :

$$p(J|x_1, \ldots, x_n) = \frac{L(x_1, \ldots, x_n|J)p(J)}{\sum_J L(x_1, \ldots, x_n|J)p(J)} \tag{7}$$

où,

- $L(x_1, \ldots, x_n|J) = \exp\{l(x_1, \ldots, x_n|J)\}$ est la vraisemblance prédictive du modèle ;
- $l(x_1, \ldots, x_n|J)$ est l'estimation du logarithme de la vraisemblance espérée du modèle.

Le critère de décision peut alors porter sur cette probabilité. D'autre part, on peut éviter un examen laborieux de plusieurs modèles $J$ (c'est-à-dire examiner $k^m$ modèles), en supposant que les observations suspectes sont situées aux deux extrémités de l'échantillon de données, c'est-à-dire examiner les observations les plus élevées et les moins élevées. Dans ce cas, la décision peut être prise à partir de l'évaluation de la probabilité marginale *a posteriori* qu'une observation spécifique, $x_i$, soit singulière : $p(x_i|x_1, \ldots, x_n) = \sum_J p(J|x_1, \ldots, x_n)$.

En somme, le critère de décision résultant est le suivant :

1. à l'aide du tableau 4, pour le modèle $J$ spécifié, calculer sa probabilité *a posteriori* $p(J|x_1, \ldots, x_n)$ (respectivement, pour un $L$ $(L \leqslant i \leqslant n - L)$ fixé calculer la probabilité marginale *a posteriori* $p(x_i|x_1, \ldots, x_n)$ que $x_{(i)}$ est une observation singulière) ;
2. si la probabilité calculée est élevée, conclure que les observations suspectes sont singulières (respectivement que l'observation $x_{(i)}$ est singulière).

## 3. DÉTECTION D'OBSERVATIONS SINGULIÈRES DANS UN MODÈLE DE RÉGRESSION LINÉAIRE UNIVARIÉ

Le modèle de régression linéaire univarié (ou étude de l'évolution d'une variable dépendante) est sans doute l'outil statistique le plus souvent mis en œuvre. La régression linéaire multiple constitue la généralisation naturelle de la régression simple. Le modèle de régression s'écrit matriciellement :

$$Y = X\beta + \varepsilon \tag{8}$$

où, $Y$ est un vecteur $(n \times 1)$ de variables aléatoires normales appelé variable réponse (ou dépendante), $X$ est une matrice $(n \times p)$ connue de plein rang $p < n$ appelée matrice des variables explicatives (ou indépendantes), $\beta$ est le vecteur $(p \times 1)$ des paramètres et $\varepsilon$ est le vecteur $(n \times 1)$ des erreurs indépendantes et identiquement distribuées selon une loi normale centrée. Ainsi définies, les données sont supposées provenir de l'observation d'un échantillon statistique de taille $n$ de $\mathbb{R}^{(p+1)}$ : $(x_{i1}, \ldots, x_{ip}, y_i)$, $i = 1, \ldots, n$. Le critère des moindres carrés qui sert dans la majorité des cas à ajuster un hyperplan de régression à ce modèle est très sensible aux observations *singulières* et aux observations *influentes*. Les observations singulières sont caractérisées par leur éloignement du barycentre du nuage central (*observations singulières à effet de levier*) ou

TABLEAU 4. — Résumé des procédures suivant le modèle de Kitagawa (1984).

| Procédure | Observations suspectes | Calcul de la probabilité *a posteriori* $p(J\|x_1,\ldots,x_n)$ |
|---|---|---|
| Kitagawa (1984) | $x_i \sim N(\mu_0+\delta,\sigma^2)$ (1 contaminant) | Pour un $L$ $(L<i\leq n-L)$ fixé calculer la probabilité marginale *a posteriori* que $x_{(i)}$ est une observation singulière : $p(x_i\|x_1,\ldots,x_n)=\sum_J p(J\|x_1,\ldots,x_n)$ Où, $L(x_1,\ldots,x_n\|J)=\exp\{l(x_1,\ldots,x_n\|J)\}$ $p(J)=\frac{(m+1)!\,(n-m)!}{3^n n!}\sum_{k=h}^{n}\frac{\Gamma\left(k+\frac{1}{2}\right)}{k^n(k-h)!}$ avec, • $h$ dénote le nombre de sources de contamination apparaissant dans $J=(j_1,\ldots,j_n)$ et $m$ est le nombre d'observations suspectes ; la somme est effectuée sur tous les $J$ possibles, $j_i\neq 0$. |
| Kitagawa (1984) | $x_j \sim N(\mu_0+\delta,\sigma^2)$ $j=n-m+1,\ldots,n$ ($m$ contaminants) | Pour le modèle $J$ spécifié, calculer sa probabilité *a posteriori* : $p(J\|x_1,\ldots,x_n)$ Où, $L(x_1,\ldots,x_n\|J)=\exp\{l(x_1,\ldots,x_n\|J)\}$ ; $p(J)=\frac{(m+1)!\,(n-m)!}{3^n n!}\sum_{k=h}^{n}\frac{\Gamma\left(k+\frac{1}{2}\right)}{k^n(k-h)!}$ avec, • $h$ dénote le nombre de sources de contamination apparaissant dans $J=(j_1,\ldots,j_n)$ et $m$ est le nombre d'observations suspectes ; la somme est effectuée sur tous les $J$ possibles, $j_i\neq 0$. |

$$l(x_1,\ldots,x_n\|J)=-\frac{n}{2}\log\left\{2\pi\sum_{i=0}^{k}\sum_{j_i=i}(x_i-\bar{x}_i)^2\right\}+\frac{n-k-1}{2}\log\{2\}+\log\left\{\Gamma\left(\frac{2n-k-1}{2}\right)\right\}-\log\left\{\Gamma\left(\frac{n-k-1}{2}\right)\right\}$$
$$-\frac{2n-k-1}{2}\left\{\psi\left(\frac{2n-k-1}{2}\right)-\psi\left(\frac{n-k-1}{2}\right)\right\},$$

où, $\psi$ est la fonction digamma (digamma function) définie par $\psi(t)=\frac{d}{dt}\log\{\Gamma(t)\}$

par la taille des résidus (*observations singulières dans le sens vertical*) tandis que les observations influentes sont celles dont une faible variation du couple $(x_{i1},\ldots,x_{ip},y_i)$ induisent une modification importante des estimations des caractéristiques du modèle. En fait, les observations singulières à effet de levier exercent une influence sur l'ajustement de l'hyperplan de régression. Par contre, les observations singulières dans le sens vertical (ayant de grands résidus) signalent plutôt des valeurs atypiques de la variable à expliquer.

Un diagnostic doit donc être établi dans le cadre spécifique du modèle de régression recherché afin d'identifier de telles observations.

On distingue deux approches bayésiennes de détection d'observations singulières dans les modèles de régression. La première se restreint à postuler un modèle principal de génération de données et cherche alors des méthodes d'identification des observations singulières sans toutefois considérer de modèle alternatif. En revanche, la deuxième approche prend en compte un modèle alternatif pour caractériser la génération d'un sous-ensemble de contaminants dans l'échantillon de données. Les modèles alternatifs les plus couramment utilisés concernent les glissements de la moyenne et la dilatation de la variance.

### 3.1. Procédures bayésiennes de détection utilisant uniquement un modèle de génération

Les principales méthodes d'identification d'observations singulières que l'on trouve dans cette catégorie utilisent une distribution prédictive pour la détection, ou encore des probabilités *a posteriori* de diverses perturbations non observées. Pour la première méthode, on dénote par $I$ un ensemble de $k$ entiers distincts choisis dans l'ensemble $(1, \ldots, n)$ de sorte que le vecteur $Y$ de l'équation 8 peut être décomposé en $Y = (Y_1', Y_{(I)}')$, où $(I)$ signifie « enlève l'ensemble $I$ ». L'idée principale est de calculer la densité prédictive $p(Y_I/Y_{(I)})$ pour détecter les observations singulières. La procédure de décision est alors la suivante : pour chaque groupe de $k$ observations calculer $p(Y_I/Y_{(I)})$ à l'aide du tableau 5; déclarer le groupe pour lequel $p(Y_I/Y_{(I)})$ est la plus petite comme étant composé d'observations singulières à effet de levier (Geisser, 1985). La deuxième méthode concerne la détection d'observations singulières à effet de levier et dans le sens vertical (avec de grands résidus). La technique de détection consiste à calculer, pour chaque résidu non observé $\varepsilon_i = y_i - x_i'\beta$ de l'équation (8), la probabilité *a posteriori* que $\varepsilon_i$ soit grand pour une constante $k$ fixée :

$$p_i = P(|\varepsilon_i| > k\sigma/Y, X) \tag{9}$$

La constante $k$ peut être choisie telle que le nombre d'observations singulières attendues *a priori* soit égal à un petit nombre $\alpha$. Par exemple, pour $\alpha = 0.05$, on choisit la constante $k$ comme suit : $k = \Phi^{-1}\{0.5 + (0.95)^{1/n}/2\}$, où $\Phi(z)$ est la fonction de répartition d'une loi normale centrée réduite et n est la taille de l'échantillon (Chaloner et Brant, 1988). La règle de décision consiste alors à calculer, à l'aide du tableau 5, la probabilité *a posteriori* $p_i$ pour chaque résidu $\widehat{\varepsilon_i}$ et à décider que l'observation $(x_{i1}, \ldots, x_{ip}, y_i)$ est singulière dans le sens vertical si $p_i > 2\Phi(-k)$. Par ailleurs, on peut aussi décider que l'observation $(x_{i1}, \ldots, x_{ip}, y_i)$ est singulière à effet de levier si $h_i$ est proche de 1 (voir tableau 5).

TABLEAU 5. — Résumé des procédures bayésiennes de détection utilisant les résidus d'une régression.

| Procédure | Type de singularité | Calcul des probabilités |
|---|---|---|
| Geisser (1985) | Observations singulières à effet de levier | $p(y_I/y_{(I)}) = K_1(s^2_{(I)})^{-n/2}\lvert I-H_I\rvert^{\frac{1}{2}}(1-Q_I)^{(n-p-k)/2}$<br>où,<br>• $H = X(X'X)^{-1}X'$ et $H_I$ est une matrice de $k\times k$ formée par les $k$ lignes et colonnes de H indexées par $I$.<br>• $Q_I = \dfrac{\hat{\varepsilon}'_I(I-H_I)^{-1}\hat{\varepsilon}_I}{(n-p-k)s^2_{(I)}}$ ; $K_1 = \left\{\dfrac{(n-p)s^2}{n-p-k}\right\}^{-n/2} K$<br>• $K = \dfrac{\Gamma((n-p)/2)}{\Gamma\left(\frac{1}{2}\right)^k \Gamma((n-p-k)/2)(n-p-k)^{k/2}}$<br>• $s^2$ est l'estimation de $\sigma^2$ dans le modèle de régression de l'équation 8 et $s^2_{(I)}$ est l'estimation de $\sigma^2_{(I)}$.<br>• $\hat{\varepsilon}'$ représente le vecteur des résidus du modèle de régression de l'équation 8 |
| Peña et Guttman (1993) | Observations singulières dans le sens vertical (avec de grands résidus) & Observations singulières à effet de levier | $p_i \approx 1-\Phi(u_1)+\Phi(u_2)$<br>où,<br>• $u_j = \dfrac{r_i/\sqrt{l_i} - (-1)^j k/\sqrt{h_i}}{\sqrt{1+\frac{1}{2}(n-p)^{-1}r_i^2/l_i}}$ $(j=1,2)$<br>• $r_i = \dfrac{\hat{\varepsilon}_i}{\sqrt{s^2(1-h_i)}}$ est le résidu studentisé et $l_i = \dfrac{h_i}{1-h_i}$ est l'effet de levier<br>• $h_i$ est le $i^{ème}$ élément de la diagonale de la matrice $H = X(X'X)^{-1}X'$ |

## 3.2. Procédures bayésiennes de détection utilisant un modèle alternatif de contamination

Dans ces procédures, bien que l'équation (8) soit le modèle supposé pour la génération de l'échantillon de données $(x_{i1},\ldots,x_{ip},y_i)$, $i=1,\ldots,n$ l'expérimentateur craint (à cause de son expérience) que certaines observations, c'est-à-dire $(x_{i_t,1},\ldots,x_{i_t,p},y_{i_t})$, $t=1,\ldots,k$ avec $k$ fixé et tel que $k \ll n/2$, aient été générées par une source contaminante se manifestant par un glissement de la moyenne ou une dilatation de la variance. La procédure bayésienne de détection de contaminants est alors basée sur une inspection des poids. Ainsi, les $k$ observations $(x_{i_t1},\ldots,x_{i_tp},y_{i_t})$, $t=1,\ldots,k$ seront déclarées singulières si leur poids $c_I$ est le plus important parmi tous les $n!/(n-k)!k!$ poids. Par ailleurs, le choix de la valeur $k$ à retenir pourra être celui qui minimise la trace de la matrice de variance-covariance *a posteriori* du paramètre $\beta$. Cette

grandeur est une mesure de la précision de l'estimation de ce paramètre. La prise de décision est

1. à l'aide du tableau 6 calculer les poids $c_I$ de tous les $n!/(n-k)!k!$ groupes de $k$ observations;
2. décider que le groupe ayant le plus grand poids est constitué d'observations singulières.

TABLEAU 6. — Résumé des procédures bayésiennes de détection utilisant un modèle de régression alternatif.

| Procédure | Modèle alternatif pour les $k$ observations suspectes | Poids |
|---|---|---|
| Guttman et al. (1978) | $Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} 0 \\ \cdots \\ a \end{pmatrix} + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix}$ où, $\varepsilon_I \sim N(0, \sigma^2 I_k)$ indépendant de $\varepsilon_{(I)} \sim N(0, \sigma^2 I_{n-k})$ | $c_I = \dfrac{\sqrt{\left\{S_{(I)}^{(n-k)} \left\|S'_{(I)}\right\| \left\|(X'_{(I)} X_{(I)})^{-1}\right\|\right\}}}{\sum \sqrt{\left\{S_{(I)}^{(n-k)} \left\|S'_{(I)}\right\| \left\|(X'_{(I)} X_{(I)})^{-1}\right\|\right\}}}$ |
| Peña et Tiao (1992) | $Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdot \\ X_I \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix}$ où, $\varepsilon_I \sim N(0, \delta^2 \sigma^2 I_k)$ indépendant de $\varepsilon_{(I)} \sim N(0, \sigma^2 I_{n-k})$ et où $\delta^2 > 1$. | $C_I = K_1 S_{(I)}^{-\frac{n-p}{2}} \left\|X'_{(I)} X_{(I)}\right\|^{1/2}$ où, • $K_1 = \left\{\dfrac{(n-p)s^2}{n-p-k}\right\}^{k/2} K$ • $K = \dfrac{\Gamma((n-p)/2)}{\Gamma\left(\frac{1}{2}\right)^k \Gamma((n-p-k)/2)(n-p-k)^{k/2}}$ • $s^2$ est l'estimation de $\sigma^2$ dans le modèle de régression de l'équation 8 |

$S_{(I)} = \left(Y_{(I)} - X_{(I)}(X'_{(I)}X_{(I)})^{-1}X'_{(I)}Y_{(I)}\right)'\left(Y_{(I)} - X_{(I)}(X'_{(I)}X_{(I)})^{-1}X'_{(I)}Y_{(I)}\right)$

*(I)* signifie «supprimer le groupe d'observations ayant un indice appartenant à *I*»

# 4. PROCÉDURES BAYÉSIENNES DE DÉTECTION DANS UN ÉCHANTILLON MULTIVARIÉ ET DANS UN MODÈLE DE RÉGRESSION LINÉAIRE MULTIVARIÉ

La détection d'observations singulières est importante non seulement pour des données univariées (données issues d'une variable aléatoire) mais également pour des données multivariées (données issues d'un vecteur aléatoire). Les observations douteuses sont dans ce cas-ci celles qui sont discordantes avec la majorité des données et présentent des écarts relatifs à un certain modèle de base. Dans cette section, on présente les procédures bayésiennes de détection d'observations singulières dans le cas multivarié pour des échantillons ainsi que pour des modèles linéaires de régression.

## 4.1. Procédures bayésiennes de détection dans un échantillon aléatoire multivarié

Les procédures présentées ici considèrent que les observations de l'échantillon, $\{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$, ont été tirées indépendamment dans une population normale multivariée de dimension $p$, $\boldsymbol{N}(\mu, \boldsymbol{\Sigma})$, où chaque observation $\boldsymbol{X_i}$ est un vecteur de dimension $p \times 1$, $\mu$ est un vecteur $(p \times 1)$ représentant la moyenne et $\Sigma$ est la matrice $(p \times p)$ de variance-covariance, elle est symétrique positive définie.

Les procédures de détection présentées dans ce cas-ci sont essentiellement des généralisations des procédures examinées en situation univariée. Guttman (1973) a proposé une généralisation de la procédure vue plus haut dans le cas de l'échantillon univarié. Plus spécifiquement, on suppose que l'on dispose d'un échantillon multivarié de $n$ observations indépendantes $\{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$, provenant d'une loi normale multivariée $\boldsymbol{N}(\mu, \boldsymbol{\Sigma})$, on suppose toutefois qu'une observation suspecte provient d'une loi normale multivariée $\boldsymbol{N}(\mu + \mathbf{a}, \boldsymbol{\Sigma})$. Ici, $\mathbf{a}$ est un vecteur de dimension $p$. Comme en situation univariée, la prise de décision se fait à l'aide du poids $c_i$ de l'observation $\boldsymbol{X_i}$ ainsi que de la probabilité *a posteriori* :

$$\gamma = \frac{p(\mathbf{a} > 0 | \boldsymbol{X_1}, \ldots, \boldsymbol{X_n})}{p(\mathbf{a} < 0 | \boldsymbol{X_1}, \ldots, \boldsymbol{X_n})} \tag{10}$$

La procédure de prise de décision est alors

1. à l'aide du tableau 7, calculer tous les poids $c_i$ des $n$ observations de l'échantillon ainsi que $\gamma$;
2. déclarer que l'observation $\boldsymbol{X_i}$ est un contaminant si $\gamma \notin \left]\frac{1}{5}, 5\right[$ et si $c_i \notin \left[0, \frac{1}{n} + \frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$.

Plus récemment, Varbanov (1998) a proposé une procédure bayésienne de détection dans un échantillon multivarié. Son approche considère la statistique suivante :

$$\boldsymbol{R}(\boldsymbol{X_i}, \mu, \boldsymbol{\Sigma}) = (\boldsymbol{X_i} - \mu)' \boldsymbol{\Sigma^{-1}} (\boldsymbol{X_i} - \mu) \tag{11}$$

de sorte qu'en posant $\delta_i = \boldsymbol{R}(\boldsymbol{X_i}, \mu, \boldsymbol{\Sigma})$, on admet que la $i^{ème}$ observation de l'échantillon est singulière si la probabilité *a posteriori* $p_i = P(\delta_i > k / \boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ est plus grande qu'une certaine valeur pour un choix approprié de $k$. En posant $k = F_p^{-1}(0.95^{1/n})$ (où $F_p(\ \ )$ est la fonction de répartition d'une loi de chi-deux avec $p$ degrés de libertés $(\chi_p^2)$), le critère de décision est alors basé sur le rapport de probabilité *a posteriori* :

$$B_i = \frac{p_i F_p(k)}{p_i(1 - F_p(k))} \tag{12}$$

Ainsi, si $B_i$ est plus grand que 10, Kass et Raftery (1995) suggèrent de conclure que la $i^{ème}$ observation est singulière. Notons cependant que les tables de ces auteurs ne font pas l'unanimité dans la littérature. Quoi qu'il en soit la procédure de prise de décision est

TABLEAU 7. — Résumé des procédures bayésiennes de détection dans un échantillon multivarié.

| Procédure | Aide à la prise de décision |
|---|---|
| Guttman . (1973) | $c_j = \dfrac{\left\|A_p^{(j)}\right\|^{(n-2)/2}}{\sum_{i=1}^{n}\left\|A_p^{(i)}\right\|^{(n-2)/2}}, \quad j = 1,\ldots,n$<br>$p(a > 0 \mid \delta^2; X_1,\ldots,X_n) = \sum_{j=1}^{n} c_j G_{n-p-1}\left(\sqrt{d_n^{(i)}\eta_i}\right)$ , avec;<br>$d_n^{(i)} = \{nA_{pn}^{(i)}/(n-1)(n-p-1)\}^{-1}$ et $\eta_i = \dfrac{n}{n-1}(X_i - \overline{X})$<br>(voir le tableau 1 pour d'autres détails de calcul) |
| Varbanov (1998) | $p_i = P(\delta_i > k/X_1,\ldots,X_n)$<br>$= E_{\Sigma^{-1}/X}\{P(W_i > nk/X_1,\ldots,X_n,\Sigma)\}$ où,<br>$\Sigma^{-1}/X$ suit une distribution de *Wishart*, $W(S^{-1}, p, n-p)$. |

1. à l'aide du tableau 7, calculer pour chaque observation $X_i$ la probabilité *a posteriori* $p_i$ et déduire sa valeur $B_i$ ;
2. déclarer que l'observation $X_i$ est singulière si $B_i$ est plus grand que 10.

### 4.2. Procédures bayésiennes de détection dans un modèle de régression linéaire multivarié

Dans cette section on considère l'approche bayésienne de détection d'observations singulières dans un modèle de régression linéaire multivarié (c'est-à-dire le cas où la variable dépendante est multidimensionnelle). Les procédures de détection appropriées considèrent le modèle classique suivant :

$$\boldsymbol{Y} = \boldsymbol{X\Theta} + \boldsymbol{E} \tag{13}$$

où $\boldsymbol{Y} = (\boldsymbol{Y_1},\ldots,\boldsymbol{Y_n})'$ est une matrice $(n \times p)$ où les $p$ colonnes sont représentées par une variable réponse de dimension $p \times 1$, $\boldsymbol{X}$ est une matrice $(n \times q)$ connue de plein rang $q < n$ appelée matrice des variables explicatives, $\boldsymbol{\Theta} = (\theta_1,\ldots,\theta_q)$ est une matrice $(q \times p)$ des paramètres et $\boldsymbol{E}$ est une matrice $(n \times p)$ contenant les erreurs sur les vecteurs $\boldsymbol{Y}_i$, distribuées indépendamment suivant une même loi normale multivariée de dimension $p$, $\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Les procédures bayésiennes de détection présentées ici sont les adaptations de Guttman *et al.* (1978) et de Varbanov (1998) en régression linéaire multivariée. Dans la cas de Guttman *et al.* (1978), Dutter et Guttman (1979) ont traité le cas où la source contaminante est caractérisée par un glissement de la moyenne. La prise de décision est résumée dans le tableau 8.

TABLEAU 8. — Résumé des procédures bayésiennes de détection en régression multivariée.

| Procédure | Aide à la prise de décision | Critère de décision |
|---|---|---|
| Dutter et Guttman (1979) | $c_i^{(\nu)} = \frac{\sqrt{\lvert S_i^{(\nu)}\rvert^2}^n}{\sum \sqrt{\lvert S_i^{(\nu)}\rvert^2}^n}$ avec $S_i^{(\nu)} = (Y - \hat{\mu}_i')' I_i (Y - \hat{\mu}_i')$ et $\hat{\mu}_i' = \frac{1}{n-k} \sum_{j=k+1}^{n} Y_{i_j}$ | $K$ observations $(x_{i_1,1},\ldots,x_{i_p,p},y_{i_n})$, $i = 1,\ldots,k$ seront singulières si leur poids $c_i^{(\nu)}$ est le plus important parmi tous les $\binom{n}{k}$ poids. |
| Varbanov (1998) | $p_i = P(\delta_i > k/Y) = E_{\Sigma^{-1}/Y}\left\{P\left(W_i > \frac{k}{\sigma_{(i)}}/Y, \Sigma^{-1}\right)\right\}$ où, $\Sigma^{-1}/X \sim Wishart, W(S^{-1}, p, n-p+1)$. | Si $B_i$ (voir équation 12) est plus grand que 10, conclure que la $i^{ème}$ observation de l'échantillon est singulière. |

# 5. DISCUSSION

La majorité des procédures bayésiennes présentées dans ce survol procèdent avec le schéma général suivant. On définit une observation (un groupe d'observations) singulière comme étant une observation (un groupe d'observations) qui n'a pas été générée par le mécanisme générateur de la majorité des observations de l'échantillon de données; ce mécanisme générateur est spécifié par un modèle principal de génération. Le principe de détection peut soit se baser sur un modèle alternatif de génération d'observations singulières, soit ne pas exiger un tel modèle. Les procédures bayésiennes qui traitent de l'échantillon univarié se retrouvent dans le premier cas. Il appert que les procédures bayésiennes de détection suivant le modèle de Pettit et Smith (1983, 1985) sont les plus intéressantes dans le cas univarié. En effet, elles offrent plusieurs choix de modèles alternatifs de génération d'observations singulières.

Les difficultés auxquelles la littérature bayésienne fait face sont posées par les calculs complexes, parfois intraitables, des probabilités *a posteriori* indispensables à la prise de décision. Pour surmonter cette difficulté, plusieurs auteurs ont proposé de calculer ces probabilités *a posteriori* à l'aide des techniques d'échantillonnage de type *Monte Carlo Markov Chain* (MCMC). La principale contribution des techniques MCMC est de faciliter les applications empiriques de la méthodologie bayésienne. Ainsi, l'échantillonneur de Gibbs (Gelfand et Smith, 1991) qui est la technique MCMC la plus simple à utiliser, a connu jusqu'à très récemment un essort considérable dans la littérature bayésienne de détection des observations singulières (Verdinelli et Wasserman, 1992; Merwe et Botha, 1993; Justel et Peña, 1996a; Bayarri et Morales, 2000).

Les procédures (bayésiennes ou non) de détection d'observations singulières peuvent aussi faire l'objet d'une critique fondamentale. En effet, si l'ensemble de données qu'on analyse contient plus qu'une observation singulière, le problème de la détection devient plus difficile à cause de l'*effet de masque* ('*masking effect*') et de l'*effet d'entraînement* ('*swamping effect*'). L'effet de masque se produit lorsqu'un sous-ensemble d'observations singulières n'est

pas détecté à cause de sa proximité avec une autre observation singulière. L'effet d'entraînement se produit par contre lorsque de « bonnes » observations sont incorrectement identifiées comme singulières à cause de la présence d'un autre sous-ensemble d'observations singulières, habituellement éloigné de ces dernières. Pour fixer les idées, supposons qu'on veut vérifier que les observations A et B (voir la figure 1 tirée de McCulloch et Meeter (1983)) sont discordantes.

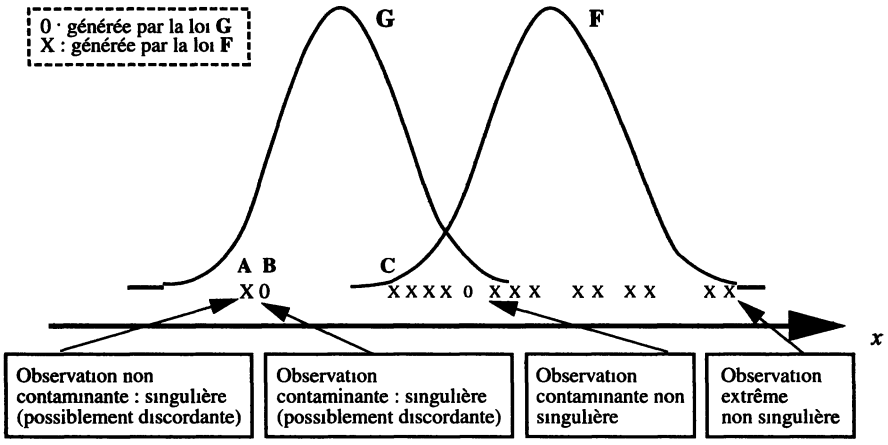


FIGURE 1. — Illustration de l'effet de masque et de l'effet d'entraînement (McCulloch et Meeter, 1983).

Une possibilité est de procéder consécutivement en vérifiant d'abord A avec le reste de l'échantillon et ensuite appliquer la même procédure à B avec le reste de l'échantillon. En procédant ainsi, l'observation A ne sera pas jugée discordante à cause de sa proximité avec l'observation B. On dit alors que l'observation B a eu un effet de masque sur l'identification de l'observation A dans la procédure consécutive ou encore que l'observation B a masqué l'observation A. C'est l'effet de masque. L'autre approche est d'utiliser une procédure de détection de deux ou trois observations discordantes. Ainsi par exemple si on choisit de vérifier la paire d'observations B et C avec le reste de l'échantillon, la procédure de vérification déclarera qu'elle est discordante; en réalité, l'observation discordante B a entraîné l'observation C avec elle, ce qui a eu pour résultat de conduire à un faux jugement, car l'observation C appartient bien au nuage central et elle n'est pas discordante. Ce danger affectant les procédures de détection de deux ou plusieurs observations singulières s'appelle l'effet d'entraînement. La capacité à se prémunir de ces deux dangers constitue donc, à notre avis, un bon critère pour classer les différentes procédures de détection présentées suivant leur ordre d'efficacité. En effet, l'effet de masque peut contribuer à diminuer l'efficacité d'une procédure de détection d'une observation singulière, alors que l'effet d'entraînement peut concourir à identifier plus d'observations singulières dans l'échantillon qu'il

n'y en a en réalité. Ainsi, une méthode qui évite ces deux écueils est donc souhaitable. C'est pourquoi on recommande, parmi les procédures bayésiennes de détection présentées dans ce survol, celles référées aux tableaux 9 (pour un échantillon univarié), 10 (pour un modèle de régression linéaire univarié) et 11 (pour le cas multivarié). En effet, elles sont immunisées contre l'effet de masque et l'effet d'entraînement. D'autre part, Justel et Peña (1996b) proposent une procédure bayésienne, basée sur l'echantillonneur de Gibbs, pour vaincre le problème posé par l'effet de masque dans la détection des observations singulières pour les modèles de régression linéaires.

TABLEAU 9. — Procédures de détection dans un échantillon univarié qui sont immunisées contre l'effet de masque ('masking effect') et l'effet d'entraînement ('swamping effect').

| Procédure | Modèle de contamination | Nombre de contaminants |
|---|---|---|
| Pettit et Smith (1983, 1985) | $N(\mu+\delta,\sigma^2)$ | 1 |
| Pettit et Smith (1983, 1985) | $N(\mu+\delta_i,\sigma^2)$ | 2 |
| Pettit (1988) | $E(\theta\delta)$ | 1 |
| Pettit (1988) | $E(\theta\delta_i)$ | 2 |
| Pettit (1994) | $P(\theta\delta)$ | 1 |
| Pettit (1994) | $P(\theta\delta_i)$ | $k$ |

TABLEAU 10. — Procédures de détection dans un modèle de régression univarié qui sont immunisées contre l'effet de masque et l'effet d'entraînement.

| Procédure | Modèle alternatif pour les $k$ contaminants | Nombre d'observations suspectes |
|---|---|---|
| Guttman et al. (1978) | $Y=\begin{pmatrix}y_{(i)}\\ \cdots\\ y_i\end{pmatrix}=\begin{pmatrix}X_{(i)}\\ \cdots\\ X_i\end{pmatrix}\beta+\begin{pmatrix}0\\ \cdots\\ a\end{pmatrix}+\begin{pmatrix}\varepsilon_{(i)}\\ \cdots\\ \varepsilon_i\end{pmatrix}$ où, $\varepsilon_i \sim N(0,\sigma^2 I_k)$ indépendant de $\varepsilon_{(i)} \sim N(0,\sigma^2 I_{n-k})$ | $k$ |
| Peña et Tiao (1992) | $Y=\begin{pmatrix}y_{(i)}\\ \cdots\\ y_i\end{pmatrix}=\begin{pmatrix}X_{(i)}\\ \cdots\\ X_i\end{pmatrix}\beta+\begin{pmatrix}\varepsilon_{(i)}\\ \cdots\\ \varepsilon_i\end{pmatrix}$ où, $\varepsilon_i \sim N(0,\delta^2\sigma^2 I_k)$ indépendant de $\varepsilon_{(i)} \sim N(0,\sigma^2 I_{n-k})$ et où $\delta^2>1$. | $k$ |

TABLEAU 11. — Procédures de détection dans le cas multivarié qui sont immunisées contre l'effet de masque et l'effet d'entraînement.

| Procédure de détection | Type de problème |
|---|---|
| Varbanov (1998) | Échantillon multivarié |
| Varbanov (1998) | Régression multivariée |

## 6. CONCLUSION

Dans cet article, on a passé en revue les différentes procédures bayésiennes pour la détection d'observations singulières. Le but visé dans ce travail était d'éclairer l'utilisateur pour qu'il puisse mieux percevoir la philosophie et les enjeux entourant ces techniques de manière à lui permettre une lecture plus aisée de ces méthodes. Des tableaux résumés d'aide à l'application des différentes méthodes ont été fournis dans chaque cas.

Quatre idées peuvent motiver le lecteur à la découverte des techniques bayésiennes de détection : la probabilité traduit un «degré» de croyance; il existe des distributions *a priori* qui incorporent toutes connaissances et opinions *a priori* sur les paramètres disponibles avant le recueil des données expérimentales; le théorème de Bayes fournit des distributions *a posteriori* qui probabilisent les valeurs possibles des paramètres; ces distributions *a posteriori* sont utilisées pour l'inférence finale. Le problème crucial qui a souvent été la pierre d'achoppement de l'inférence bayésienne reste le choix de la distribution *a priori*, car il faut veiller à ce que celle-ci soit jugée acceptable, compte tenu à la fois des connaissances antérieures, même peu précises, que l'on possède, et des considérations générales. Toutefois, les difficultés de formulation d'une connaissance *a priori* ont conduit à rechercher des distributions *a priori* non informatives. Dans les deux cas, si l'étude est fondée sur un faible effectif d'échantillon, la distribution *a priori* retenue aura une incidence très forte sur la qualité des résultats, en ce sens qu'elle imprimera sa marque sur la sensibilité de l'inférence. Quoi qu'il en soit, on pense que le statisticien, analyste de données, qui trouve une ou plusieurs données singulières doit discuter avec l'expérimentateur (ou le collecteur de données) et qu'ensemble ils répondent à la question : est-on en présence de réelles singularités et si oui qu'en fait-on ? (McCulloch et Meeter, 1983). Si par contre aucune décision ne peut être prise, il est certainement préférable de faire l'inférence statistique en utilisant les méthodes bayésiennes d'accommodation (Barnett et Lewis, 1994).

## REMERCIEMENTS

Ce travail a été rendu possible grâce au soutien financier d'Hydro-Québec et du Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG). Les auteurs remercient un rapporteur pour ses judicieuses remarques ayant permis d'améliorer la rédaction initiale.

## RÉFÉRENCES BIBLIOGRAPHIQUES

BARNETT V. and LEWIS T. (1984). Outliers in Statistical Data, Wiley, Second edition, 463 pages.

BARNETT V. and LEWIS T. (1994). Outliers in Statistical Data, Wiley, Third edition, 584 pages.

BAYARRI M.J. and MORALES J. (2000). Bayesian Measures of Surprise for Outlier Detection. Journal Statistical Planning and Inference, to appear.

BOX G. E. P. and TIAO G. C. (1968). A Bayesian Approach to some Outlier Problems. *Biometrika*, 55, pp. 119-129.

CHALONER K. and BRANT R. (1988). A Bayesian Approach to Outlier Detection and Residual Analysis. *Biometrika*, 75, pp. 651-659.

DE ALBA E. and VAN RYZIN J. (1979). An Empirical Bayes Test for Multiple Outliers, University Statistics Center, *Technical Report No. 35*, New Mexico State University.

DE FINETTI B. (1961). The Bayesian Approach to the Rejection of Outliers, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 199-210.

DUTTER R. and GUTTMAN I. (1979). On Estimation in the Linear Model when Spurious Observations are Present – a Bayesian Approach. *Communication in Statistics : Theory and Methods.*, 8, pp. 611 -635.

EDGEWORTH F. Y. (1887). On Discordant Observations. *Philosophical Magazine*, 23, Ser. 5, pp. 364-375.

FREEMAN P. R. (1980). On the Number of Outliers in Data From a Linear Model (with discussion), in *Bayesian Statistics*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F.M. Smith, Valencia, Spain : Valencia University Press, pp. 349-365.

GEISSER S. (1985). On the Predicting of Observables : a Selective Update. In *Bayesian Statistics 2* Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A.F M. Smith, pp. 203-30. Amsterdam : North Holland.

GELFAND A. E. and SMITH A. F. M. (1991). Gibbs Sampling for Marginal Posterior Expectations. *Communications in Statistics : Theory Methods*, 20 (5 et 6), pp. 1747-1766.

GNANADESIKAN R. (1977). Methods for Statistical Data Analysis of Multivariate Observations, New York, John Wiley and Sons, 311 pages.

GRUBBS F. E. (1969). Procedures for Detecting Outlying Observations in Samples, *Technometrics*, Vol. 11, pp. 1-21.

GUTTMAN I (1973). Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity-A Bayesian Approach, *Technometrics*, Vol.15, pp. 723-738.

GUTTMAN I, and KHATRI C. G. (1975). A Bayesian Approach to some Problems Involving the Detection of Spuriosity, *Applied Statistics*, North-Holland, Amsterdam, pp.111-145 of Gupta, R. P. (Ed.).

GUTTMAN I., FREEMAN P. R. and DUTTER R.(1978). Care and Handling of Univariate Outliers in the General Linear Model to Detect Spuriosity A Bayesian Approach, *Technometrics*, Vol. 20, No. 2, pp. 187-193.

JUSTEL A. and PEÑA D. (1996a). Gibbs Sampling will Fail in Outlier Problem with Strong Masking. Journal of Computational and Graphical Statistics, 5, pp. 176-189.

JUSTEL A. and PEÑA D. (1996b). Bayesian Unmasking in Linear Models. Core Discussion Paper 9619, Université Catholique de Louvain.

KASS R. and RAFTERY A. (1995). Bayes Factors. Journal of the American Statistical Association, 90, pp.773-95.

KITAGAWA G. (1984). Bayesian Analysis of Outliers via Akaike's Predictive Likelihood of a Model. Communications in Statistics : Simulation and Computation., 13, pp. 107-126.

MCCULLOCH C. E. and MEETER D. (1983). Discussion of Outliers by Beckman, R. J. and Cook, R. D., *Technometrics*, 25, pp.152-155.

MERWE A. J. and BOTHA T. J. (1993). A Bayesian Approach to Outlier Detection in Regression Analysis Using Gibbs Sampling. *South African Statistics Journal*, 27, pp.181-202.

PEÑA D. and GUTTMAN I. (1993). Comparing Probabilistic Methods for Outlier Detection in Linear Models, *Biometrika*, 80, 3, pp 603-10.

PEÑA D. and TIAO G. C. (1992). Bayesian Robustness Functions for Linear Models. *Bayesian Statistics*, 4 (Edited by M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 365-388. Oxford University Press.

PETTIT L. I. (1988). Bayes Methods for Outliers in Exponential Samples, *Journal of the Royal. Statistical Society*, Ser. B, 50, pp. 371-380.

PETTIT L. I. (1992). Bayes Factors for Outlier Models Using the Device of Imaginary Observations, *Journal of the American Statistical Association*, Vol. 87, No. 418, pp. 541-545.

PETTIT L. I. (1994). Bayesian Approaches to the Detection of Outliers in Poisson Samples, *Communication in Statistics : Theory and Methods*, 23(6), pp. 1785-1795.

PETTIT L. I. and SMITH A. F. M. (1983). Bayesian Model Comparaison in Presence of Outliers. *Bulletin International Statistics Institute.*, 50, pp. 292-309.

PETTIT L. I. and SMITH A. F. M. (1985). Outliers and Influential Observations in Linear Models. In *Bayesian Statististics 2*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 473-494. Amsterdam : North-Holland.

SPIEGELHALTER D. J. and SMITH A. F. M. (1982). Bayes Factor for Linear and Log-Linear Models With Vague Prior Information, *Journal of the Royal Statistical Society*, Ser. B, 44, pp. 377-387.

VARBANOV A. (1998). Bayesian Approach to Outliers Detection in Multivariate Normal Samples and Linear Models. *Communication in Statistics : Theory and Methods*, 27(3), pp.-547-557.

VERDINELLI I and WASSERMAN L. (1992). Bayesian Analysis of Outlier Problem Using the Gibbs Sampler. Statistics and Computing, 1, pp.105-117.