Record Number: 16440
Author, Monographic: Ondo, J. C.//Ouarda, T. B. M. J.//Bobée, B.
Author Role:
Title, Monographic: Revue de littérature des procédures bayésiennes pour la détection d'observations singulières
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 1999
Original Publication Date: Février 1999
Volume Identification:
Extent of Work: x, 128
Packaging Method: pages inlcuant 2 annexes
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, rapport de recherche
Series Volume ID: 540
Location/URL:
ISBN: 2-89146-315-3
Notes: Rapport annuel 1998-1999
Abstract: Chaire en hydrologie statistique Hydro-Québec / CRSNG
Call Number: R000540
Keywords: rapport/ ok/ dl

REVUE DE LITTÉRATURE DES PROCÉDURES BAYÉSIENNES POUR LA DÉTECTION D'OBSERVATIONS SINGULIÈRES

REVUE DE LITTÉRATURE DES PROCÉDURES BAYÉSIENNES POUR LA DÉTECTION D'OBSERVATIONS SINGULIÈRES

par

Jean-Cléophas Ondo
Taha B.M.J. Ouarda
Bernard Bobée

Chaire industrielle Hydro-Québec / CRSNG en Hydrologie statistique
Institut national de la recherche scientifique, INRS-Eau
2800 rue Einstein, Case postale 7500, SAINTE-FOY (Québec) G1V 4C7

Rapport de recherche No R-540

Février 1999

ISBN 2-89146-315-3

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

1. INTRODUCTION

Le quotidien d'un statisticien, analyste de données est caractérisé par la confrontation régulière avec des données que l'on peut qualifier d'imparfaites. Il s'agit soit de matrices de données manquantes soit de matrices avec des points douteux voire aberrants ou encore des matrices dont les observations ne respectent pas les hypothèses prérequises à l'emploi d'une méthode (exemple, ne suivent pas une loi de probabilité donnée.)

En analyse de données comme d'ailleurs en statistique, il est essentiel de travailler sur des données expérimentales fiables. Pour évaluer la qualité d'un échantillon le statisticien doit disposer de techniques permettant de déceler des données douteuses ou franchement erronées. Ces types de données ont parfois un intérêt en elles-mêmes, par ailleurs elles peuvent être la source de contamination dans des analyses futures, telles que dans l'ajustement d'un modèle, l'estimation de paramètres ou dans les tests d'hypothèses pour ne citer que ceux-là. Il est donc nécessaire de s'en préoccuper au début de toute étude.

Dans ce travail, nous allons traiter le point de vue analyse de données singulières, appelées encore données douteuses ou étrangères («outliers»). Le problème des données singulières a retenu l'attention de nombreux auteurs ces dernières années et la littérature est plus qu'abondante sur ce sujet traité tantôt de façon générale, tantôt dans un contexte particulier comme les séries chronologiques, les modèles linéaires et les plans d'expérience. Les méthodes statistiques qui se rattachent à leur développement sont du type *classique* ou *bayésienne*. Pour une vision plus détaillée de ce sujet, on peut se référer à la synthèse complète de *Barnett et Lewis (1994)*.

L'objectif de ce travail consiste donc à présenter une revue de littérature des procédures bayésiennes de détection d'observations singulières. Nous allons premièrement rapprocher le domaine des données singulières dans trois contextes à savoir les concepts généraux - les différents tests de détection dans le cas univarié pour un échantillon aléatoire simple et pour un modèle de régression linéaire - puis nous tenterons d'élargir certains de ces différents tests dans le contexte multivarié. Nous donnerons alors les tests de détection d'observations

singulières dans un échantillon multivarié et dans un modèle de régression linéaire multivarié. Nous esquisserons au besoin, sur la base d'une compréhension facile, à partir des exemples d'application existant dans la littérature, des règles pratiques.

2. CONCEPTS GÉNÉRAUX

2.1. Définition d'une donnée singulière

Aucune observation ne peut garantir d'être totalement dépendante de la manifestation du phénomène physique auquel elle provient. Intuitivement, la probable fiabilité d'une observation est reflétée par sa relation avec d'autres observations obtenues dans les mêmes conditions. Bien qu'il existe plusieurs techniques de détection d'observations singulières dans un échantillon de données, une définition rigoureuse à leur sujet semble aussi vague aujourd'hui qu'il y a cent ans. 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 frequecy from other observations with which they are combined».*

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 membres of the sample in which it occurs».*

C'est donc une question de *jugement subjectif* de la part du décideur (analyste de données) de décider si oui ou non une observation (ou un ensemble d'observations) doit être soumise à examen minutieux.

D'une façon générale, on peut donner une définition «simple» d'une observation singulière dans un cadre unidimensionnel. En effet, il s'agit d'un point situé loin dans les queues de la distribution mais, il devient pratiquement impossible de donner une définition plus générale dans une situation multidimensionnelle. Dans l'effort d'éviter certaines ambiguïtés, nous adoptons les définitions suivantes dans le reste de ce rapport :

- **observation discordante** : *toute observation qui apparaît surprenante ou hors de propos du point de vue du décideur (analyste de données ou 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 ou une observation discordante.*

De ce point de vue, si l'on démontre, dans une étape de validation de données, que l'observation singulière est fausse, c'est-à-dire qu'elle est survenue dans l'échantillon à la suite d'une condition irrégulière (erreur de mesure etc.), on dira alors qu'elle est *aberrante* (exemple la donnée de la crue du Saguenay en 1996 est singulière mais non aberrante.)

2.2 Procédures statistiques appropriées à la manipulation d'observations singulières

Bâtir un corps de méthodes statistiques pour examiner les observations singulières consiste, en termes généraux, à fournir un moyen d'évaluer si notre déclaration subjective sur la présence éventuelle d'observations douteuses dans un échantillon de données a des implications objectives pour une analyse ultérieure de ces données. La présence d'observations singulières dans un échantillon peut résulter de deux raisons. L'une étant de nature purement déterministe et l'autre de nature aléatoire ou non explicable. Dans la première raison, les observations singulières pourraient simplement être enlevées dans l'échantillon ou être remplacées par des données corrigées. La deuxième raison requiert quant à elle beaucoup plus de précautions sur le traitement de ce type d'observations et c'est elle que nous analyserons dans la suite.

L'approche statistique servant à manipuler les observations singulières est basée sur deux méthodes distinctes :

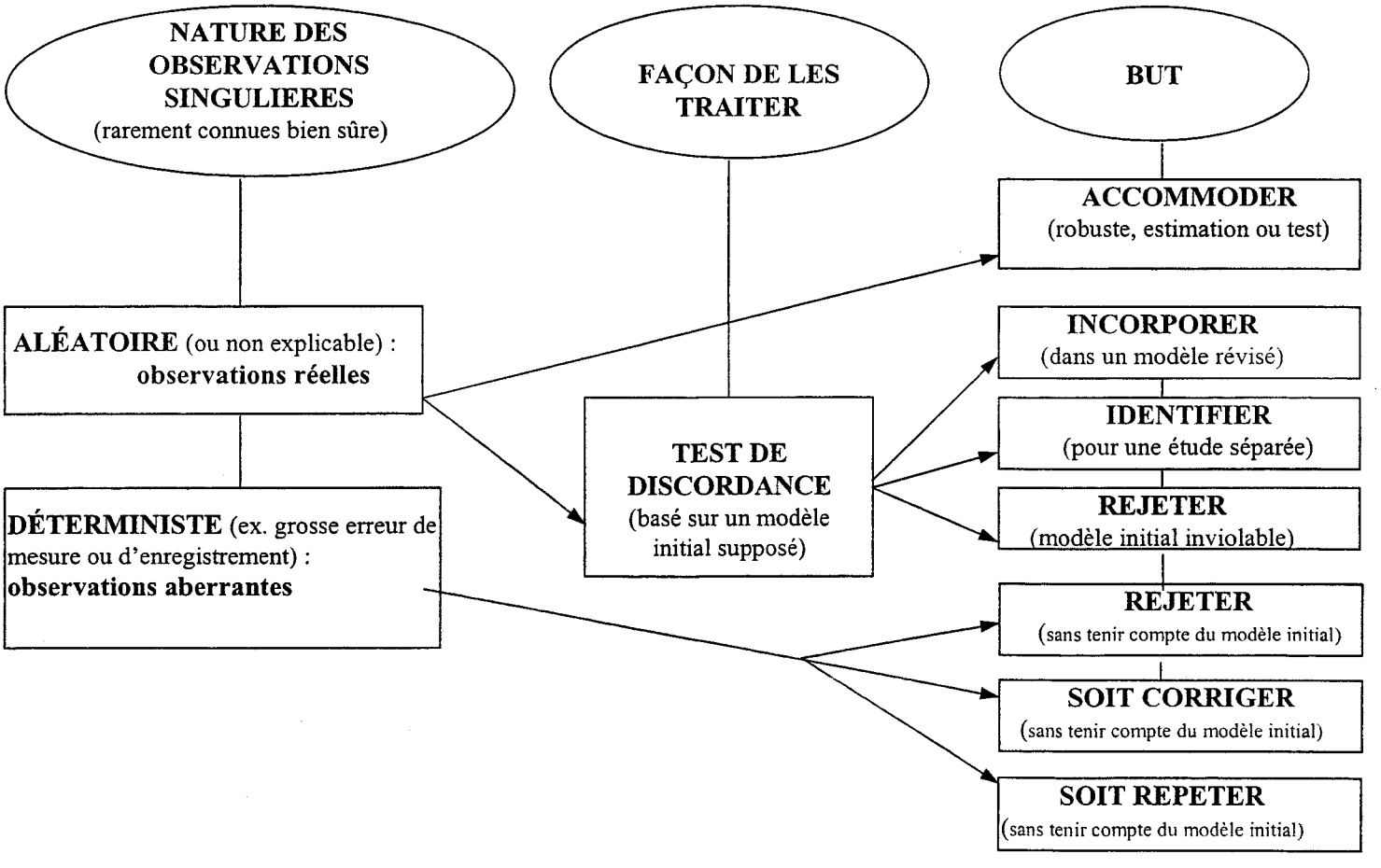
1. *l'identification* : elle sert à identifier les observations singulières pour une étude ultérieure ;
2. *l'accommodation* : elle sert à accommoder la possibilité d'observations singulières par des modifications confortables du modèle et /ou méthodes d'analyse.

Les méthodes d'identification sont caractérisées, en principe, par des tests statistiques que l'on appelle *tests de discordance.* Ces derniers consistent à tester une observation singulière avec la perspective de la rejeter de l'échantillon de données auquel elle a été issue. Par contre, les méthodes d'accommodation sont des méthodes statistiques conçues pour tirer des inférences valides sur la population de laquelle un échantillon a été obtenu, et qui ne sera pas sérieusement déformée par la présence d'observations singulières. Ces méthodes

accommodent donc les observations singulières à aucun inconvénient sérieux ou sont encore robustes à leur présence. Par exemple les modèles mixtes peuvent être utilisés pour accommoder certains types de contaminants (observations singulières) et les *M-estimateurs* sont souvent utilisés pour fournir certaines protections contre les observations quand le modèle mixte est symétrique.

De ces deux méthodes, nous jugeons la méthode d'identification comme étant la plus pertinente aux objectifs de ce rapport. Les méthodes d'accommodation d'observations singulières tendent à demander beaucoup d'informations sur le processus générateur de ces observations ou sont conçues d'être immunisées à leur présence. Toutefois, bien que la philosophie sous-jacente aux notions d'identification et d'accommodation semble distincte, elles sont souvent confondues puisqu'une méthode d'accommodation pourrait produire une méthode d'identification comme un dérivé, ou vice versa. En somme, le traitement d'observations singulières peut facilement être mieux assimilé à l'aide de la figure 2.1 empruntée à *Barnet et Lewis (1994).*

Figure 2.1 : traitement des observations singulières



2.3. Construction des tests de discordance

Tout test statistique doit inévitablement examiner deux types d'hypothèses : une hypothèse nulle (H_0), que l'on appelle encore hypothèse de travail, et une hypothèse alternative (H_1). Pour un test de discordance d'observations singulières, l'hypothèse nulle, (H_0), exprimera un certain modèle de probabilité de base qui caractérisera la génération de toutes les données sans tenir compte d'une présence éventuelle de données singulières ; l'hypothèse alternative, (H_1), exprimera quant à elle une façon dans laquelle le modèle de probabilité de base pourrait être modifié pour incorporer ou expliquer les observations singulières comme résultats de contamination. Pour une bonne compréhension de ces deux types d'hypothèses dans le contexte de traitement d'observations singulières, il est judicieux d'apporter quelques précisions sur leurs formes.

2.3.1. Forme de l'hypothèse nulle

De par la nature suspecte des observations singulières, l'hypothèse nulle, (H_0), sera principalement l'énoncé d'un modèle de probabilité de base. Il sera donc suffisant de déclarer, à cet effet, que les données $x_i \quad (i = 1,\ldots,n)$ sont des observations indépendantes provenant d'une certaine distribution de probabilité $\Im$, ce qui peut encore s'écrire mathématiquement :

$$H_0 : x_i \in \Im \quad (i = 1,2,\ldots,n)$$

2.3.2. Formes de l'hypothèse alternative

Sur la base d'un test statistique de discordance, lorsque nous rejetons une observation singulière (ou un ensemble d'observations singulières) d'être discordante, nous rejetons implicitement l'hypothèse nulle, (H_0), en faveur d'une certaine hypothèse alternative, H_1. Il nous faut donc connaître l'alternative appropriée à la dite hypothèse nulle. De plus, toute évaluation de la puissance du test de discordance va dépendre de la forme de l'hypothèse alternative adoptée.

Ainsi, parmi les formes possibles d'hypothèses alternatives que l'on retrouve dans la littérature, nous avons retenu celles qui suivent :

(a) ***Alternative déterministe***

Cette alternative reflète la situation où nous sommes quasiment certains que la présence d'observations singulières dans l'échantillon est due à de grosses erreurs de mesure ou d'enregistrements qui sont identifiables dans le processus de cueillette de données. Le test de discordance revient alors à tester les deux hypothèses suivantes :

$$\begin{cases} H_0: & \text{tout } x_k \ (k \neq i) \quad \text{provient de la loi } \Im \\ H_1: & x_k \text{ ne provient pas de la loi } \Im \end{cases}$$

Dans cette situation, l'observation x_k identifiée comme étant singulière est aberrante. Elle est donc soit corrigée ou simplement rejetée de l'échantillon de données.

(b) ***Alternative inhérente***

Cette forme d'hypothèse alternative correspond à la situation où l'observation singulière peut refléter une forme différente de variabilité inhérente à un certain modèle de probabilité de base $\Im$. Le test de discordance revient alors à tester les hypothèses suivantes :

$$\begin{cases} H_0: & x_i \in \Im \quad (i = 1,2,\ldots,n) \\ & (\textit{toutes les observations proviennent d'une loi de probabilité } \Im) \\ \\ H_1: & x_i \in G \qquad (i = 1,2,\ldots,n) \\ & (\textit{toutes les observations proviennent d'une loi de probabilité } G) \end{cases}$$

(c) ***Alternative mixte***

Pour cette hypothèse alternative, au lieu de supposer, comme précédemment, que l'observation singulière reflète un degré de variabilité auquel on ne s'attend pas, nous

admettons plutôt, la possibilité que quelques observations de l'échantillon, autres que celles qui proviennent de la distribution de probabilité de base $\Im$, sont des contaminants (observations singulières provenant de la distribution de probabilité G). Le test de discordance revient alors à tester les hypothèses suivantes :

$$\begin{cases} H_0: \ x_i \in \Im \qquad (i = 1,2,\ldots,n) \\ H_1: \ x_i \in (1-\lambda)\Im - \lambda G \quad (i = 1,2,\ldots,n) \\ \qquad (\lambda \text{ est la probabilité que l'observation } x_i \\ \qquad\quad \text{provient de la distribution de probabilité } G) \end{cases}$$

(d) ***Alternative de décalage*** **(«slippage alternative»)**

Dans sa forme la plus usuelle, cette hypothèse alternative stipule que toutes les observations, excepté un petit nombre k (1, 2 ou plus) proviennent indépendamment d'une distribution de probabilité de base $\Im$ indexée par les paramètres de position et d'échelle, μ et σ^2, tandis que les k autres observations restantes (observations singulières) proviennent indépendamment d'une version modifiée de $\Im$, où ses paramètres de position et d'échelle ont changé, μ dans l'une ou l'autre direction, σ^2 typiquement augmenté. Le test de discordance revient alors à tester les hypothèses suivantes :

$$\begin{cases} H_0: \ x_i \in \Im(\mu,\ \sigma^2) \quad (i = 1,2,\ldots,n) \\ H_1: \ x_i \in \Im(\mu + a,\ \sigma^2) \ \text{ou} \ x_i \in \Im(\mu,\ b\sigma^2) \\ \qquad\qquad (b > 1) \quad (i = 1,2,\ldots,n) \end{cases}$$

(e) ***Alternative d'échange*** **(«exchangeable alternative»)**

Cette hypothèse alternative est une extension de la formulation de l'alternative de décalage précédente. Dans sa formulation la plus générale, le modèle le plus simple pour un contaminant suppose que $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ sont des observations indépendantes

provenant de la distribution de probabilité de base $\Im$. On suppose qu'il est également probable que l'indice i du contaminant soit n'importe lequel des n indices. Les variables aléatoires $X_1, X_2, \ldots, X_N$ sont ainsi, dans le modèle, non pas indépendantes mais échangeable.

2.4 Approche bayésienne dans la manipulation d'observations singulières

Dans les sections précédentes, nous avons fait remarqué que des facteurs subjectifs intervenaient dans le jugement du décideur (analyste de données) sur la présence éventuelle d'une observation singulière (ou un ensemble d'observations singulières) dans un échantillon de données à analyser. Ainsi, la caractéristique d'une observation singulière est donc le degré de surprise qu'il engendre dans un échantillon de données. L'approche de la statistique classique que nous avons présentée précédemment se résumait à l'utilisation de deux méthodes distinctes pour manipuler de telles observations : une méthode d'identification et une méthode d'accommodation. Dans cette section, nous examinons l'utilité de l'approche bayésienne dans la manipulation d'observations singulières.

2.4.1. Introduction aux méthodes bayésiennes

Notre objectif dans cette sous-section, n'est pas de présenter une revue de littérature exhaustive sur les méthodes bayésiennes, mais de résumer la méthodologie afin de permettre au lecteur d'apprécier l'utilité de telles méthodes dans ce survol. Toutefois, il existe une vaste littérature sur les fondements philosophiques de l'approche bayésienne à l'inférence statistique ; on peut consulter à ce titre les ouvrages de *Zellner (1971)*, *Poirier (1988)*, *Leamer (1978)*, *Berger (1985)*, *Robert (1992)* ou *Bernardo et Smith (1994)* pour une présentation détaillée.

Les méthodes statistiques dites *classiques* ont pour objectif principal de faire, au vu d'observations des résultats expérimentaux, une *inférence* au sujet de la loi générant ces

observations, afin, soit d'analyser un phénomène passé, soit de prévoir un événement futur, en nous attachant tout particulièrement aux aspects décisionnels de cette inférence. Une telle définition laisse évidemment de côté certaines branches de la statistique comme par exemple la *collecte des données* (les sondages ou les plans d'expérience), mais elle correspond à ses buts variés parmi lesquels les plus couramment rencontrés sont :

- *l'estimation* (soit ponctuelle ou par intervalle de confiance) ;
- *les tests d'hypothèses.*

On suppose donc que les observations sur lesquelles l'analyse statistique classique (en ne considérant que le modèle paramétrique) est fondée, $x_1, x_2, \ldots, x_{n-1}, x_n$, proviennent de modèle de probabilité paramétrés, c'est-à-dire x_i suit une loi de densité $f(x_i/\theta)$ sur IR^p, dont le paramètre θ, scalaire ou vectoriel, est inconnu, au contraire de *f*. Ce modèle caractérise le comportement des observations *x*, *conditionnellement* au paramètre θ ; il fournit des probabilités d'échantillonnage $p(x/\theta)$. Étant donné l'observation *x*, $p(x/\theta)$ peut être regardé comme une fonction, non pas de *x*, mais de θ. Considérée ainsi, cette probabilité est appelée la *fonction de vraisemblance* de θ pour la valeur de *x* donnée, et s'écrit :

$$L(\theta/x) = \prod_{i=1}^{n} f(x_i/\theta)$$

Formellement, elle permet de réécrire le modèle probabiliste dans le «bon ordre», dans la mesure où l'objectif est de faire une inférence sur θ :

$$L(\theta/x) = p(x/\theta)$$

L'inférence de la statistique classique est donc fondamentalement une démarche d'inversion puisqu'elle vise à remonter des effets aux causes, c'est-à-dire des observations aux paramètres. La théorie bayésienne réalise cette inversion de façon légitime et cohérente. En

effet, la théorie bayésienne associe aux résultats expérimentaux des distributions de probabilités *a priori* sur les valeurs des paramètres. L'information «objective» apportée par l'expérimentateur (analyste des données) intervient pour modifier les distributions a priori, et les remplacer par des distributions *a posteriori*, qui concernent des probabilités conditionnelles (et plus précisément conditionnées par l'expérimentation). Ainsi, du point de vue de l'expérimentateur, les données x sont observées, mais le vecteur des paramètres θ est inconnu ; la distribution d'intérêt est donc la densité conditionnelle $p(\theta/x)$. Cette densité est dérivée à l'aide du théorème de Bayes (*Berger (1985)*) :

$$p(\theta/x) = \frac{p(x/\theta)p(\theta)}{p(x)} \equiv \frac{p(x/\theta)p(\theta)}{\int p(x/\theta)p(\theta)d\theta}$$

où $p(\theta)$ est la distribution marginale de θ. Ainsi, partant d'un état d'incertitude initial sur le paramètre, le théorème de Bayes permet d'exprimer la nouvelle incertitude sur les valeurs possibles du paramètre, une fois les données collectées. Comme la distribution $p(\theta)$ représente l'incertitude sur la valeur de θ avant d'avoir observé les données, elle est donc la distribution a priori de θ. La distribution conditionnelle $p(\theta/x)$ est alors la distribution a posteriori de θ, car elle décrit l'incertitude de l'expérimentateur concernant θ après avoir observé les données x. Le rôle fondamental de la vraisemblance dans le raisonnement statistique est ainsi mis en avant par l'analyse bayésienne qui établit que la probabilité a posteriori $p(\theta/x)$ est proportionnelle au produit de la vraisemblance de θ (pour x donné) par la probabilité a priori $p(\theta)$, ce qui s'écrit :

$$p(\theta/x) \propto f(x/\theta)p(\theta)$$

Ainsi, l'outil de base de la statistique bayésienne est une formule particulièrement simple à utiliser, une fois admis et compris les deux concepts fondamentaux que sont la vraisemblance et la probabilité a priori.

Le propre de la méthode bayésienne est précisément d'utiliser dans un contexte décisionnel une distribution a priori choisie indépendamment de l'échantillon observé. On définit alors *le risque de Bayes* associé à la densité a priori $p(\theta)$ de la règle de décision δ par :

$$R(\delta) = \int W(\delta,\theta)p(\theta)d\theta,$$

où $W(\delta,\theta)$ est la *fonction de perte* associée à la décision δ. Comme il est d'usage pour toute règle de décision, δ sera appelée estimateur de θ et sera notée $\hat{\theta}$. Pour une fonction de perte quadratique, l'estimateur de Bayes de θ, associée à la loi a priori $p(\theta)$, est la moyenne de la distribution a posteriori. Dans le contexte des tests bayésiens, classiquement on notera :

Θ_0 : sous-ensemble des paramètres θ correspondant à l'hypothèse nulle ;

Θ_1 : sous-ensemble des paramètres θ correspondant à l'hypothèse alternative ;

$\Theta = \Theta_0 \cup \Theta_1$ *et* $\Theta_0 \cap \Theta_1 = \varnothing$;

$\delta = (\delta_0, \delta_1)$ avec δ_0 : acceptation de l'hypothèse nulle ;

δ_1 : acceptation de l'hypothèse alternative ;

Comme il est naturel de retenir une fonction de perte s'annulant uniquement pour la fonction bonne décision, c'est-à-dire telle que :

$$\begin{cases} W(\delta_1,\theta) > 0 \text{ et } W(\delta_0,0) = 0, & \forall \theta \in \Theta_0 \\ W(\delta_0,\theta) > 0 \text{ et } W(\delta_1,0) = 0, & \forall \theta \in \Theta_1 \end{cases}$$

pour un test φ, on peut définir le risque bayésien :

$$\int_{\Theta} R(\varphi,\theta)p(\theta)d\theta,$$

et comparer les tests à l'aide de ce nombre. Il est donc clair que le test optimal φ_0 est défini par :

$$\varphi_0(x)=\begin{cases}0, & si \quad \int_{\Theta_0} W(\delta_1,\theta)p(\theta/x)d\theta \geq \int_{\Theta_1} W(\delta_0,\theta)p(\theta/x)d\theta \\ 1, & si \quad \int_{\Theta_0} W(\delta_1,\theta)p(\theta/x)d\theta < \int_{\Theta_1} W(\delta_0,\theta)p(\theta/x)d\theta\end{cases}$$

Autrement dit, x est dans la région critique si la perte moyenne a posteriori encourue en choisissant l'hypothèse nulle H_0 est plus grande que celle en courue en choisissant l'hypothèse alternative H_1. Pour une fonction de perte constante ayant la forme :

$$W(\delta_0,\theta)=\begin{cases}0 \ si \quad \theta\in\Theta_0 \\ c_0 \quad si \quad \theta\in\Theta_1\end{cases}$$

$$W(\delta_1,\theta)=\begin{cases}0 \ si \quad \theta\in\Theta_0 \\ c_1 \quad si \quad \theta\in\Theta_1\end{cases}$$

on a :

$$\varphi_0(x)=\begin{cases}0, & si \quad c_0 p_{\Theta_0}(\theta/x) > c_1 p_{\Theta_1}(\theta/x) \\ 1, & si \quad c_0 p_{\Theta_0}(\theta/x) > c_1 p_{\Theta_1}(\theta/x)\end{cases}$$

Ceci revient à comparer le rapport des probabilités a posteriori de Θ_0 et de Θ_1 et le rapport des pertes. En particulier si le fait de penser à tort que l'hypothèse nulle est vraie entraîne une perte infinie $(c_1 \to +\infty)$ et si l'autre perte est bornée, le test optimal conduit à accepter H_1 dès que $p_{\Theta_1}(\theta/x)$ est strictement positif.

En conclusion, un principe fondamental de l'approche bayésienne est que toute inférence statistique devrait se fonder sur la détermination rigoureuse de trois facteurs :

1. la famille des lois des observations, $f(x/\theta)$,
2. la distribution a priori des paramètres, $p(\theta)$,
3. la fonction de perte, $W(\delta,\theta)$.

De ces facteurs, nous pouvons dégager les points suivants :

- Il est conceptuellement immédiat de passer de la vraisemblance à la distribution a posteriori. En outre le problème crucial reste alors le choix de la distribution a priori, qui a été souvent la pierre d'achoppement de l'inférence bayésienne. En effet, il est souvent difficile d'approcher les distributions a priori, et on se contente d'une approche intuitive «à vue de nez». Suivant la conception bayésienne la plus large, les distributions initiales permettent d'incorporer toutes les connaissances et opinions a priori sur les paramètres disponibles avant le recueil des données - notamment des résultats antérieurs - ou encore par exemple l'opinion d'experts du phénomène naturel sous étude. Si l'analyste de données veut utiliser une loi *a priori non informative* pour refléter l'ignorance totale, il peut adopter une *loi diffuse* avec une variance très élevée ou même une *loi impropre* (c'est-à-dire, une densité qui ne s'intègre pas à 1 comme la loi $p(\theta) = k$, une constante). Alors la loi a posteriori aura la même forme que la fonction de vraisemblance. Il y a d'ailleurs une importante littérature sur la meilleure façon de caractériser l'ignorance *Bernardo et Smith (1994)* par exemple.

- La détermination de la fonction de perte, $W(\delta, \theta)$, est fortement liée à une connaissance plus poussée du problème, donc à des informations a priori sur le modèle qu'une analyse bayésienne pourrait utiliser plus efficacement.

Clairement, si les méthodes bayésiennes apportent une plus grande souplesse dans la méthodologie statistique de l'analyse de données. elles ne valent en fait que ce que vaut l'information a priori. Si celle-ci est bonne l'application est possible, et conduit effectivement à des décisions plus économiques. Si elle est mauvaise, ou vague, la prudence est de règle.

2.4.2. Utilité des méthodes bayésiennes dans la manipulation d'observations singulières

L'aspect le plus important dans l'approche bayésienne est sa formalisation des impressions subjectives comme élément de l'analyse statistique. *de Finetti (1961)* fut le premier à étudier de manière approfondie le problème de la manipulation d'observations singulières par les

méthodes bayésiennes. Son implication sur ce sujet a été surtout orientée sur les principes de base de l'approche bayésienne plutôt que sur les techniques de développement à proprement dites sur la manipulation de telles observations. Il a proposé que toute approche bayésienne sur la manipulation d'observations singulières soit posée en ces termes : une certaine quantité X ayant une distribution a priori qui peut être modifiée à la lumière d'un échantillon de données observées, $x_1, x_2, \ldots, x_{n-1}, x_n$, pour produire une distribution finale (a posteriori) de X. Vu sous cet angle, l'approche bayésienne s'avère donc comme étant un outil indispensable dans la manipulation d'observations singulières. En effet, contrairement à l'approche classique de l'inférence statistique présentée au début de ce rapport, la distribution finale ou distribution a posteriori (celle servant à l'inférence totale) dépend maintenant sur toutes les données de l'échantillon. Ainsi, toute observation dans l'échantillon est un candidat potentiel pour être considéré comme observation singulière, selon que son influence sur la distribution finale est faible ou pratiquement négligeable. Cela semble donc constituer une base pour, soit identifier, ou accommoder les observations singulières, si leur présence a une influence négligeable sur la portée inférentielle des données.

3. DÉTECTION D'OBSERVATIONS SINGULIÈRES DANS UN ÉCHANTILLON ALÉATOIRE UNIVARIÉ

Dans ce chapitre, nous présentons divers tests bayésiens de détection d'observations singulières dans un échantillon univarié. L'hypothèse de base stipule que toutes les observations de l'échantillon, $x_1, x_2, \ldots, x_{n-1}, x_n$, ont été générées par une distribution de base mais, nous craignons que certaines de ces observations soient des contaminants, c'est-à-dire qu'elles aient été générées par une autre distribution.

3.1. Tests de détection suivant le modèle *de Guttman (1973) et Guttman et Khatri (1975)*

Dans ces procédures, on suppose que l'on dispose d'un échantillon de n observations indépendantes $x_1, x_2, \ldots, x_n$, provenant d'une loi normale $N(\mu, \sigma^2)$ mais, on craint toutefois qu'une (ou deux) de ces observations pourrait avoir été générée par une loi normale $N(\mu + a, \sigma^2)$, $N(\mu + a, \sigma^2/\delta)$ ou $N\left(\mu + a\frac{\sigma}{\delta}, \frac{\sigma^2}{\delta^2}\right)$. Le paramètre δ (respectivement δ_1 et δ_2 pour deux contaminants) est supposé connu. L'inférence repose donc ici sur le paramètre de décalage à la moyenne, a (respectivement a_1 et a_2 dans le cas de deux contaminants). En supposant, qu'avant d'avoir collecté les données, une vague information est disponible sur les paramètres μ, σ^2, et probablement sur les valeurs de a (respectivement a_1 et a_2 dans le cas de deux contaminants), on utilise alors une distribution a priori non informative pour μ, σ^2, a (respectivement μ, σ^2, a_1 et a_2) qui est :

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

(respectivement $p(\mu, \sigma^2, a_1, a_2) \propto (\sigma^2)^{-1}$)

La distribution a posteriori prend l'une ou l'autre des formes suivantes :

$$p\left(a \mid \delta^2; x_1,\ldots x_n\right) = \sum_{i=1}^{n} c_i h\left(a \mid n_i; B; k\right) \qquad \text{(pour un contaminant)}$$

$$p\left(a_1, a_2 \mid \delta_1^2, \delta_2^2 ; \quad x_1,\ldots,x_n\right) = \sum\sum_{i\neq j} d_{ij} g\left(a_1, a_2 \mid n_{ij}; B_{ij}; l\right) \text{ (pour deux contaminants)}$$

Il devient alors important d'observer que :

- La distribution a posteriori de a (respectivement de a_1 et a_2) est une pondération de la distribution h (respectivement de celle de g). Une inspection des poids est donc utile pour nous aider à faire une inférence sur a (respectivement sur a_1 et a_2). L'uniformité des poids peut en conséquence constituer un bon moyen pour soutenir l'énoncé que « $a = 0$» (respectivement « $a_1 = 0$», « $a_2 = 0$»). Ainsi, pour $n \geq 3$, on montre, d'après *Guttman (1973)*, que les poids sont uniformes s'ils sont tous compris dans l'intervalle $\left[0, \frac{1}{n} + \frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$.
- Si toutes les observations avaient été générées par une même source, les intervalles a posteriori des valeurs de a (respectivement celles de a_1 *et* a_2) doivent nous permettre de vérifier, par exemple que :

$$p\left(a > 0 \mid \delta^2; x_1,\ldots,x_n\right) \approx p\left(a < 0 \mid \delta^2; x_1,\ldots,x_n\right)$$

(respectivement, $p\left(a_l > 0 \mid \delta_1^2, \delta_2^2; x_1,\ldots,x_n\right) \approx p\left(a_l < 0 \mid \delta_1^2, \delta_2^2; x_1,\ldots,x_n\right), l = 1 \textit{ ou } 2$)

Le test bayésien de détection de contaminants est alors basé sur le rapport des probabilités a posteriori suivant :

$$\gamma = \frac{p\left(a > 0 \mid \delta^2; x_1,\ldots,x_n\right)}{p\left(a < 0 \mid \delta^2; x_1,\ldots,x_n\right)}$$

(respectivement $\gamma_s = \dfrac{p\left(a_s > 0 \mid \delta_1^2, \delta_2^2; x_1,\ldots,x_n\right)}{p\left(a_s < 0 \mid \delta_1^2, \delta_2^2; x_1,\ldots,x_n\right)}$, $s = 1 \textit{ ou } 2$)

La détermination du rapport γ (respectivement γ_s) avec les poids associés à chaque observation, constitue un moyen efficace pour le décideur de vérifier si une observation quelconque de l'échantillon a été générée par une source contaminante. Ainsi, si γ (respectivement γ_s) est proche de *1* et si tous les poids sont à l'intérieur de l'intervalle, $\left[0, \frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, nous devrions soutenir l'énoncé que toutes les observations ont été générées par une même source. Toutefois, si γ (respectivement γ_s) a une grande valeur, par exemple plus grande ou égale à *5* ou inférieure ou égale à *1/5*, et que certaines observations ont des poids à l'extérieur de l'intervalle ci-haut mentionné, alors la décision qu'un contaminant existe dans l'échantillon de données devrait être maintenue, et l'observation dont le poids est à l'extérieur de cet intervalle doit être considérée comme étant un contaminant. Ces remarques inspirent les tests qui vont suivre.

a) Test de *Guttman (1973)* : *cas où on craint qu'une observation ait été générée par la source contaminante* $N(\mu+a, \sigma^2)$

On suppose que $x_1, x_2, \ldots, x_n$ sont *n* observations indépendantes qui ont été générées par la loi $N(\mu, \sigma^2)$ mais, on craint toutefois que l'observation x_i ait été générée par la source contaminante $N(\mu+a, \sigma^2)$. Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0: \; x_j \in N(\mu, \sigma^2) & (j=1,\ldots,n) \\ H_1: \; x_i \in N(\mu+a, \sigma^2) & (a>1) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). On démontre (*Guttman (1973)*) que la densité a posteriori de a est de la forme :

$$p(a \mid x_1, x_2, \ldots, x_n) \propto \sum_{i=1}^{n} c_i h(a \mid \eta_i ; B^{(i)} ; n-2),$$

où

- la fonction h est la densité généralisée d'une distribution de *Student*, de degré de liberté $(n-2)$, de moyenne η_i et de constante $B^{(i)} > 0$. Cette densité est donnée, pour une observation par

$$h\left(y \mid \eta ; \beta ; \upsilon\right) = \frac{(B)^{1/2}\,\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)(\pi\upsilon)^{1/2}}\left[1+\frac{B(y-\eta)^2}{\upsilon}\right]^{-(\upsilon+1)/2}$$

où,

- $$\eta_j = \frac{n}{n-1}\left(x_j - \bar{x}\right), \quad \text{avec} \quad \bar{x} = \frac{\sum_{l=1}^{n} x_l}{n}$$

- $$B^{(j)} = \frac{1}{\frac{nA^{(j)}}{(n-1)(n-2)}}, \text{ avec } A^{(j)} = \sum_{l\neq j}\left(x_l - \bar{x}^{(j)}\right)^2 \text{ et } \bar{x}^{(j)} = \frac{\sum_{l\neq j} x_l}{n-1}$$

- les poids c_j sont donnés par la formule : $c_j = \dfrac{\left(A^{(jj)}\right)^{-(n-2)/2}}{\sum_{l=1}^{n}\left(A^{(l)}\right)^{-(n-2)/2}}, \qquad j = 1,\ldots,n$;

D'après *Guttman (1973)*, on montre aussi que

$$\eta = E\left(a \mid x_1, x_2, \ldots, x_n\right) = \sum_{l=1}^{n} c_l \eta_l$$

$$Var\left(a \mid x_1, x_2, \ldots, x_n\right) = \sum_{l=1}^{n} c_l \eta_l^2 + \frac{n}{(n-1)(n-4)}\sum_{l=1}^{n} c_l A^{(l)} - \eta^2$$

et que le rapport des probabilités a posteriori est donné par :

$$\gamma = \frac{p\left(a > 0 \mid x_1, \ldots, x_n\right)}{p\left(a < 0 \mid x_1, \ldots, x_n\right)} = \frac{\sum_{l=1}^{n} c_l G_{n-2}\left(\sqrt{B^{(l)}}\eta_l\right)}{1 - \sum_{l=1}^{n} c_l G_{n-2}\left(\sqrt{B^{(l)}}\eta_l\right)}$$

où G_{n-2} est la fonction de répartition d'une distribution de Student avec $(n-2)$ degrés de liberté, c'est-à-dire :

$$G_{n-2}(y)=\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi(n-2)}}\int_{-\infty}^{y}\left(1+\frac{t^2}{n-2}\right)^{-\frac{n-1}{2}}dt$$

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale* $N(\mu,\sigma^2)$.

Étape 2 *: on craint qu'une donnée,* x_i*, de l'échantillon ait été générée par la loi normale* $N(\mu+a,\sigma^2)$.

Étape 3 *: calculer les poids* c_i *en remplissant le tableau suivant :*

i	x_i	$\bar{x}^{(i)}$	$A^{(i)}$	$\left(A^{(i)}\right)^{-\frac{n-2}{2}}$	c_i
1	x_1				
2	x_2				
⋮	⋮				
n	x_n				
Somme					

Étape 4 : *calculer le rapport* γ *en remplissant le tableau suivant :*

i	x_i	η_i	$B^{(i)}$	$c_i G_{n-2}\left(\eta_i\sqrt{B^{(i)}}\right)$	$1-c_i G_{n-2}\left(\eta_i\sqrt{B^{(i)}}\right)$
1	x_1				
2	x_2				
$\vdots$	$\vdots$				
n	x_n				
somme					

$$\gamma = \frac{p\left(a>0 \mid \delta^2; x_1,\ldots,x_n\right)}{p\left(a<0 \mid \delta^2; x_1,\ldots,x_n\right)} = \frac{\sum_{l=1}^{n} c_l G_{n-2}\left(\sqrt{B^{(l)}\eta_l}\right)}{1-\sum_{l=1}^{n} c_l G_{n-2}\left(\sqrt{B^{(l)}\eta_l}\right)}$$

Étape 5 : si $\left(\gamma \geq 5\right)$ *ou* $\left(\gamma \leq \frac{1}{5}\right)$ *et si* $c_i \notin \left[0, \frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que l'observation* x_i *est singulière.*

b) Test de *Guttman et Khatri (1975)* : *cas où on craint qu'une observation ait été générée par la source contaminante* $N\left(\mu + a\frac{\sigma}{\delta}, \frac{\sigma^2}{\delta^2}\right)$

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi $N\left(\mu, \sigma^2\right)$ mais, on craint toutefois que l'observation x_i ait été générée par la source

contaminante $N\left(\mu + a\frac{\sigma}{\delta}, \frac{\sigma^2}{\delta^2}\right)$ où, contrairement au test précédent, le paramètre de position est proportionnel à l'écart-type de la loi contaminante. De plus, le paramètre δ est supposé connu de sorte que l'inférence se fera sur a pour un δ^2 fixé. Les hypothèses de ce test bayésien sont alors :

$$\begin{cases} H_0 : \; x_j \in N(\mu, \sigma^2) & (j = 1, \ldots, n) \\ H_1 : \; x_i \in N\left(\mu + a\frac{\sigma}{\delta}, \frac{\sigma^2}{\delta^2}\right) & (a > 1) \end{cases}$$

Nous sommes aussi en présence d'une alternative de décalage («slippage»). D'après *Guttman et Khatri*, la densité a posteriori de a, pour δ^2 fixé, est de la forme :

$$p\left(a \mid \delta^2 ; x_1, x_2, \ldots, x_n\right) = \sum_{l=1}^{n} c_l f_l(a_1) \quad , \qquad a_1 = a\sqrt{\frac{n-1}{n-1+\delta}},$$

où,

- la fonction f_l est définie par :

$$f_l(a_1) = \sum_{t=0}^{\infty}\left(1-\rho_l^2\right)^{(n-1)/2} \rho_l^t \frac{\Gamma\left(\frac{n+t-1}{2}\right)}{\Gamma\left(\frac{t}{2}+1\right)\Gamma\left(\frac{n-1}{2}\right)} h_t(a_1) \quad \text{avec}$$

$$\overline{x} = \frac{\sum_{l=1}^{n} x_l}{n}$$

$$h_t(a_1) = \frac{\Gamma\left(\frac{t}{2}+1\right)2^{t/2}}{\sqrt{2\pi}\,\Gamma(t+1)}\sqrt{\frac{n-1}{n-1+\delta^2}}\, a_1^t \exp\left(-a_1^2/2\right)$$

- les poids c_i sont donnés par la formule : $c_j = \dfrac{\left(A^{(ji)}\right)^{-(n-2)/2}}{\sum_{l=1}^{n}\left(A^{(l)}\right)^{-(n-2)/2}}$, $\quad j = 1,\ldots,n$ avec

$$A^{(j)} = \sum_{l\neq j}\left(x_l - \bar{x}^{(j)}\right)^2 \quad \text{et} \quad \bar{x}^{(j)} = \frac{\sum_{l\neq j} x_l}{n-1}$$

Guttman et Khatri montrent aussi que

$$E\left(a \mid \delta^2 ; x_1, x_2, \ldots, x_n\right) = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\delta \sum_{l=1}^{n} c_l \frac{\left(x_l - \bar{x}^{(l)}\right)}{\sqrt{A^{(l)}}}$$

$$Var\left(a \mid \delta^2 ; x_1, x_2, \ldots, x_n\right) = \frac{n-1+\delta^2}{n-1} + (n-1)\delta^2 \sum_{l=1}^{n} \frac{c_l\left(x_l - \bar{x}^{(l)}\right)^2}{\left(A^{(l)}\right)} - \left\{E\left(a \mid \delta^2 ; x_1, x_2, \ldots, x_n\right)\right\}^2$$

et que le rapport des probabilités a posteriori est donné par la formule :

$$\gamma = \frac{p\left(a > 0 \mid \delta^2; x_1, \ldots, x_n\right)}{p\left(a < 0 \mid \delta^2; x_1, \ldots, x_n\right)} = \frac{p\left(a > 0 \mid \delta^2; x_1, \ldots, x_n\right)}{1 - p\left(a > 0 \mid \delta^2; x_1, \ldots, x_n\right)}, \text{ avec}$$

- $p\left(a > 0 \mid \delta^2 ; x_1, \ldots, x_n\right) = \dfrac{1}{2}\sum_{l=1}^{n}\sum_{t=0}^{\infty} c_l\left(1-\rho_l^2\right)^{(n-1/2)}\rho_l^t \dfrac{\Gamma\left(\frac{n+t-1}{2}\right)}{\Gamma\left(\frac{t}{2}+1\right)\Gamma\left(\frac{n-1}{2}\right)}$

Une application pratique de ce test consistera à suivre les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu, \sigma^2\right)$.

Étape 2 *: on craint qu'une donnée,* x_i*, de l'échantillon ait été générée par la loi normale*

$$N\left(\mu + a\frac{\sigma}{\delta}, \frac{\sigma^2}{\delta^2}\right).$$

Étape 3 *: calculer les poids* c_i *en remplissant le tableau suivant :*

i	x_i	$\overline{x}^{(i)}$	$A^{(i)}$	$\left(A^{(i)}\right)^{-\frac{n-2}{2}}$	c_i
1	x_1				
2	x_2				
$\vdots$	$\vdots$				
n	x_n				
Somme					

Étape 4 *: calculer le rapport* γ *en remplissant le tableau suivant :*

i	x_i	ρ_i	$\left(1-\rho_i^2\right)^{(n-1)/2}$
1	x_1		
2	x_2		
$\vdots$	$\vdots$		
n	x_n		
somme			

$$\gamma = \frac{p\left(a>0 \mid \delta^2; x_1,\ldots,x_n\right)}{p\left(a<0 \mid \delta^2; x_1,\ldots,x_n\right)} = \frac{p\left(a>0 \mid \delta^2; x_1,\ldots,x_n\right)}{1-p\left(a>0 \mid \delta^2; x_1,\ldots,x_n\right)}, \text{ avec}$$

$$p\left(a>0 \mid \delta^2 ; x_1,\ldots,x_n\right) = \frac{1}{2}\sum_{l=1}^{n}\sum_{t=0}^{\infty} c_l\left(1-\rho_l^2\right)^{(n-1/2)}\rho_l^t \frac{\Gamma\left(\frac{n+t-1}{2}\right)}{\Gamma\left(\frac{t}{2}+1\right)\Gamma\left(\frac{n-1}{2}\right)}$$

Étape 5 : *si* $\left(\gamma \geq 5\right)$ *ou* $\left(\gamma \leq \frac{1}{5}\right)$ *et si* $c_i \notin \left[0\,,\frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que l'observation* x_i *est singulière.*

c) Test de *Guttman et Khatri (1975)* : *cas où on craint que deux observations x_i et x_k aient été générées simultanément par les sources contaminantes $N\left(\mu+a_i\,,\frac{\sigma^2}{\delta_i^2}\right)$ et $N\left(\mu+a_k\,,\frac{\sigma^2}{\delta_k^2}\right)$, avec a_i et a_k non liés par une relation*

On suppose que $x_1, x_2, \ldots, x_n$ sont *n* observations indépendantes qui ont été générées par la loi $N(\mu\,,\sigma^2)$ mais, on craint toutefois que l'observation x_i ait été générée par la source contaminante $N\left(\mu+a_i\,,\frac{\sigma^2}{\delta_i^2}\right)$ et que l'observation x_k ait été générée par la source contaminante $N\left(\mu+a_k\,,\frac{\sigma^2}{\delta_k^2}\right)$. De plus, les paramètres δ_i et δ_k sont supposés connus de sorte que l'inférence se fera sur le couple $(a_i\,,a_k)$ pour δ_i^2 et δ_k^2 fixés. Les hypothèses de ce test bayésien sont de la forme :

$$\begin{cases} H_0: \quad x_j \in N(\mu\,,\sigma^2) \qquad (j=1,\ldots,n) \\ H_1: \quad x_i \in N\left(\mu+a_i\,,\frac{\sigma^2}{\delta_i^2}\right) \quad , \quad x_k \in N\left(\mu+a_k\,,\frac{\sigma^2}{\delta_k^2}\right) \end{cases}$$

Nous sommes aussi en présence d'une alternative de décalage («slippage»). *Guttman et Khatri* montrent que la densité a posteriori du couple $(a_i\,,a_k)$, pour δ_i^2 et δ_k^2 fixés, est de la forme :

$$p\left(a_i\,,a_k \mid \delta_i^2\,,\delta_k^2\,;x_1,x_2,\ldots,x_n\right)=\sum\sum_{j\neq l} d_{jl}^{(1)}\,h_2\left(a_i\,,a_k \mid \vec{\eta}_{jl}\,;S_{jl}\,;n-3\right)$$

où,

- la fonction h_2 est la densité généralisée d'une distribution de *Student* bivariée, de degré de liberté $(n-3)$, de moyenne $\vec{\eta}_{jl}$ et de constante S_{jl}.

- les poids $d_{jl}^{(1)}$ sont donnés par la formule : $d_{jl}^{(1)} = \dfrac{\left(A^{(jl)}\right)^{-(n-3)/2}}{\sum\sum_{j\neq l}^{n}\left(A^{(jl)}\right)^{-(n-3)/2}}$, avec,

$$A^{(jl)} = \sum_{c\neq j,l}\left(x_c - \bar{x}^{(jl)}\right)^2 \text{ et } \bar{x}^{(jl)} = \frac{\sum_{c\neq j,l} x_c}{n-2} ;$$

Ils montrent aussi que

$$E\left(a_i \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \sum\sum_{j\neq l} d_{jl}^{(1)}\left(x_i - \bar{x}^{(jl)}\right) ;$$

$$Var\left(a_i \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \sum\sum_{j\neq l} d_{jl}^{(1)}\left[\left(x_j - \bar{x}^{(jl)}\right)^2 + \frac{A^{(jl)}}{n-5}\frac{n-2+\delta_i^2}{(n-2)\delta_i^2}\right] - \left[\sum\sum_{j\neq l} d_{jl}^{(1)}\left(x_j - \bar{x}^{(j)}\right)\right]^2$$

Des formules similaires sont aussi obtenues pour a_k. Le rapport des probabilités a posteriori est alors donné par :

$$\gamma_s = \frac{p\left(a_s > 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)}{p\left(a_s < 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)} = \frac{p\left(a_s > 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)}{1 - p\left(a_s > 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)}, \qquad s = i \quad ou \quad s = k,$$

avec, $p\left(a_s > 0 \mid \delta_i^2\ \delta_k^2\ ; x_1, \ldots, x_n\right) = \sum\sum_{j\neq l} d_{jl}^{(1)}\, G_{n-3}\left(\dfrac{(n-3)(n-2)\delta_s^2}{\left[A^{(jl)}\left(n-2+\delta_s^2\right)\right]}\left(x_j - \bar{x}^{(jl)}\right)\right)$

où G_{n-3} est la fonction de répartition d'une distribution de *Student* avec $(n-3)$ degrés de liberté

Une application pratique de ce test, consistera à suivre les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu, \sigma^2\right)$.

Étape 2 : *on craint que deux observations* x_i *et* x_k *aient été générées simultanément par les lois* $N\left(\mu + a_i\ ,\ \frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu + a_k\ ,\ \frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *non liés par une relation*

Étape 3 : *calculer les poids* $d_{jl}^{(1)}$ *en remplissant le tableau suivant :*

x_i	j / i	1 2 ... n	j / i	1 2 ... n	j / i	1 2 ... n	j / i	1 2 ... n
x_1 x_2 ⋮ x_n	1 2 ⋮ n	$\bar{x}^{(ij)}$	1 2 ⋮ n	$A^{(ij)}$	1 2 ⋮ n	$\left[A^{(ij)}\right]^{-\frac{n-3}{2}}$	1 2 ⋮ n	$d_{ij}^{(1)}$
					somme			

Étape 4 : *calculer le rapport* γ_s :

$$\gamma_s = \frac{p\left(a_s > 0 \mid \delta_i^2\ ,\delta_k^2; x_1,\ldots,x_n\right)}{p\left(a_s < 0 \mid \delta_i^2\ ,\delta_k^2; x_1,\ldots,x_n\right)} = \frac{p\left(a_s > 0 \mid \delta_i^2\ ,\delta_k^2; x_1,\ldots,x_n\right)}{1 - p\left(a_s > 0 \mid \delta_i^2\ ,\delta_k^2; x_1,\ldots,x_n\right)}, \qquad s = i \quad ou \quad s = k,$$

$$avec,\ p\left(a_s > 0 \mid \delta_i^2\ \delta_k^2\ ; x_1,\ldots,x_n\right) = \sum\sum_{j \neq l} d_{jl}^{(1)}\ G_{n-3}\left|\frac{(n-3)(n-2)\delta_s^2}{\left[A^{(jl)}\left(n-2+\delta_s^2\right)\right]}\left(x_j - \bar{x}^{(jl)}\right)\right|$$

Étape 5 : si $\left(\gamma_s \geq 5\right)$ *ou* $\left(\gamma_s \leq \frac{1}{5}\right)$, $s = i \quad ou \quad s = k$ *et si* $d_{ik}^{(1)} \notin \left[0\ ,\ \frac{1}{n} + \frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que les observations* x_i *et* x_k *sont singulières.*

d) Test de *Guttman et Khatri (1975)* : *cas où on craint que deux observations x_i et x_k aient été générées simultanément par les sources contaminantes* $N\left(\mu+a_i\,,\frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu+a_k\,,\frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *liés par la relation* $a_i=-a_k=a$

Les hypothèses de ce test sont les mêmes que précédemment avec une hypothèse additionnelle qui stipule que $a_i=-a_k=a$. *Guttman et Khatri* montrent que la densité a posteriori de a, pour δ_i^2 et δ_k^2 fixés, est de la forme :

$$p\left(a\mid \delta_i^2\,,\delta_k^2\,;x_1,x_2,\ldots,x_n\right)=\sum\sum_{j\neq l}t_{jl}\,h_1\left(a\mid \eta(jl)\,;f_{jl}\,;n-2\right),$$

où,

- la fonction h_1 est la densité généralisée d'une distribution de *Student*, de degré de liberté $(n-2)$, de moyenne $\eta(jl)$ et de constante f_{jl}.
- les poids t_{jl} sont donnés par la formule : $t_{jl}=\dfrac{\left(u_{jl}\right)^{-(n-2)/2}}{\sum\sum_{j\neq l}\left(u_{jl}\right)^{-(n-2)/2}}$, avec,

$$u_{jl}=A^{(jl)}+\frac{\delta_i^2\delta_k^2(n-2)}{(n-2)\left(\delta_i^2+\delta_k^2\right)+4\delta_i^2\delta_k^2}\left[\left(x_j-\bar{x}^{(jl)}\right)+\left(x_j-\bar{x}^{(jl)}\right)\right]^2$$

$$A^{(jl)}=\sum_{e\neq j,l}\left(x_e-\bar{x}^{(jl)}\right)^2 \quad\text{et}\quad \bar{x}^{(jl)}=\frac{\sum_{e\neq j,l}x_e}{n-2}\;;$$

Ils montrent aussi que

$$E\left(a\mid \delta_i^2\;\delta_k^2\;;x_1,x_2,\ldots,x_n\right)=\sum\sum_{j\neq l}t_{jl}\,\eta(jl)\;;$$

$$Var\left(a\mid \delta_i^2\;\delta_k^2\;;x_1,x_2,\ldots,x_n\right)=\sum\sum_{j\neq l}t_{jl}\left[\eta^2(jl)+\frac{n-2}{n-4}f_{jl}^{-1}\right]-\left[E\left(a\mid \delta_i^2\;\delta_k^2\;;x_1,x_2,\ldots,x_n\right)\right]^2$$

avec,

$$f_{jl} = \frac{(n-2)\left[(n-2)\left(\delta_i^2 + \delta_k^2\right) + 4\delta_i^2\delta_k^2\right]}{u_{jl}\left(n-2+\delta_i^2+\delta_k^2\right)}$$

Le rapport des probabilités a posteriori est alors donné par :

$$\gamma = \frac{p\left(a>0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)}{p\left(a<0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)} = \frac{p\left(a>0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)}{1-p\left(a>0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)},$$

avec

$$p\left(a>0 \mid \delta_i^2\ \delta_k^2\ ; x_1,\ldots,x_n\right) = \sum\sum_{j\neq l} t_{jl}\, G_{n-2}\left(\sqrt{f_{jl}}\ \eta(jl)\right)$$

où G_{n-2} est la fonction de répartition d'une distribution de Student avec $(n-2)$ degrés de liberté.

Les différentes étapes conduisant à une application pratique de ce test sont les suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu, \sigma^2\right)$.

Étape 2 : *on craint que deux observations* x_i *et* x_k *aient été générées simultanément par les lois* $N\left(\mu + a_i, \frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu + a_k, \frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *liés par la relation :*

$a_i = -a_k = a$.

Étape 3 : *calculer les poids* $d_{jl}^{(1)}$ *en remplissant le tableau suivant :*

x_i	i / j	1 2 ... n	j / i	1 2 ... n	j / i	1 2 ... n	j / i	1 2 ... n
x_1 x_2 ⋮ x_n	1 2 ⋮ n	$\overline{x}^{(ij)}$	1 2 ⋮ n	$A^{(ij)}$	1 2 ⋮ n	u_{ij}	1 2 ⋮ n	t_{ij}
					somme			

Étape 4 : *calculer le rapport* γ :

$$\gamma = \frac{p\left(a>0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)}{p\left(a<0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)} = \frac{p\left(a>0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)}{1-p\left(a>0 \mid \delta_i^2, \delta_k^2; x_1,\ldots,x_n\right)},$$

avec,

$$p\left(a>0 \mid \delta_i^2\ \delta_k^2; x_1,\ldots,x_n\right) = \sum\sum_{j\neq l} t_{jl}\, G_{n-2}\left(\sqrt{f_{jl}}\,\eta(jl)\right)$$

Étape 5 : *si* $\left(\gamma \geq 5\right)$ *ou* $\left(\gamma \leq \frac{1}{5}\right)$, *et si* $t_{jk} \notin \left[0, \frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que les observations* x_i *et* x_k *sont singulières.*

e) Test de *Guttman et Khatri (1975)* : *cas où on craint que deux observations* x_i *et* x_k *aient été générées simultanément par les sources contaminantes* $N\left(\mu + a_i\frac{\sigma}{\delta_i}, \frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu + a_k\frac{\sigma}{\delta_k}, \frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *non liés par une relation*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi $N\left(\mu, \sigma^2\right)$ mais, on craint toutefois que l'observation x_i ait été générée par la source

contaminante $N\left(\mu + a_i \frac{\sigma}{\delta_i}, \frac{\sigma^2}{\delta_i^2}\right)$ et que l'observation x_k ait été générée par la source contaminante $N\left(\mu + a_k \frac{\sigma}{\delta_k}, \frac{\sigma^2}{\delta_k^2}\right)$. Contrairement au test précédent, le paramètre de position est proportionnel à l'écart-type de la loi contaminante. De plus, les paramètres δ_i et δ_k sont supposés connus de sorte que l'inférence se fera sur le couple (a_i, a_k) pour δ_i^2 et δ_k^2 fixés. Les hypothèses de ce test bayésien sont de la forme :

$$\begin{cases} H_0: \quad x_j \in N(\mu, \sigma^2) \qquad (j = 1,\ldots,n) \\ H_1: \quad x_i \in N\left(\mu + a_i \frac{\sigma}{\delta_i}, \frac{\sigma^2}{\delta_i^2}\right) \quad , \quad x_k \in N\left(\mu + a_k \frac{\sigma}{\delta_k}, \frac{\sigma^2}{\delta_k^2}\right) \end{cases}$$

Nous sommes aussi en présence d'une alternative de décalage («slippage»). *Guttman et Khatri* montrent que la densité a posteriori du couple (a_i, a_k), pour δ_i^2 et δ_k^2 fixés, est de la forme :

$$p(a_i, a_k \mid \delta_i^2, \delta_k^2; x_1, x_2, \ldots, x_n) = \sum\sum_{j \neq l} d_{jl}^{(2)} h_{jl}(a_i, a_k)$$

où,

- la fonction h_{jl} est donnée par la formule :

$$h_{jl}(a_i, a_k) = \sum_{t=0}^{\infty} \frac{2^{t/2}\Gamma\left(\frac{n-1+t}{2}\right)|M|^{1/2}}{t!\Gamma\left(\frac{n-1}{2}\right)2\pi} \left[\rho_{i(jl)}a_i + \rho_{k(jl)}a_k\right]^t \left(1-\varepsilon_{jl}\right)^{(n-1)/2} \exp(-Q/2)$$

  les différents paramètres apparaissant dans cette formule sont donnés par *Guttman et Khatri (1975)*

- les poids $d_{jl}^{(2)}$ sont donnés par la formule : $d_{jl}^{(2)} = \dfrac{\left(A^{(jl)}\right)^{-(n-1)/2}}{\sum\sum_{j\neq l}^{n}\left(A^{(jl)}\right)^{-(n-1)/2}}$, avec,

$$A^{(jl)} = \sum_{e\neq j,l}\left(x_e - \bar{x}^{(jl)}\right)^2 \text{ et } \bar{x}^{(jl)} = \frac{\sum_{e\neq j,l} x_e}{n-2}$$

Ils montrent aussi que

$$E\left(a_s \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sum\sum_{j\neq l} \frac{d_{jl}^{(2)}}{\sqrt{\left(1-\varepsilon_{jl}\right)}} z_{s(jl)}$$

$$Var\left(a_s \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \left(1 + \frac{\delta_s^2}{n-2}\right) + \frac{(n-1)\sum\sum_{j\neq l} d_{js}^{(2)}\, z_{s(jl)}^2}{\left(1-\varepsilon_{jl}\right)} - \left[E\left(a_s \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right)\right]^2$$

$$\text{avec} \quad z_{s(jl)} = \left[\rho_{s(jl)} + \frac{\delta_s}{n-2}\left(\rho_{i(jl)}\delta_i + \rho_{k(jl)}\delta_k\right)\right]$$

Le rapport des probabilités a posteriori est alors donné par la formule :

$$\gamma_s = \frac{p\left(a_s > 0 \mid \delta_i^2\ , \delta_k^2; x_1, \ldots, x_n\right)}{p\left(a_s < 0 \mid \delta_i^2\ , \delta_k^2; x_1, \ldots, x_n\right)} = \frac{p\left(a_s > 0 \mid \delta_i^2\ , \delta_k^2; x_1, \ldots, x_n\right)}{1 - p\left(a_s > 0 \mid \delta_i^2\ , \delta_k^2; x_1, \ldots, x_n\right)}, \qquad s = i \quad ou \quad s = k,$$

- avec,

$$p\left(a_s > 0 \mid \delta_i^2\ \delta_k^2\ ; x_1, \ldots, x_n\right) = \frac{1}{2} + \sum\sum_{j\neq l} d_{jl}^{(2)} \sum_{m=0}^{\infty} \frac{2^{m+1/2}\,\Gamma\left(\frac{n}{2}+m\right)}{(2m+1)!\,\Gamma\left(\frac{n-1}{2}\right)} C_{2m+1}^{(s)}\left(j, l\right)\left[1-\varepsilon_{jl}\right]^{(n-1)/2},$$

et,

$$C_{2m+1}^{(s)}(j,l)=\frac{2^{m+1/2}}{\pi}\Gamma\left(m+\frac{3}{2}\right)\int_0^{v_{s(jl)}}\left(\varepsilon_{jl}-t^2\right)^m dt, \qquad v_{s(jl)}=\frac{z_{s(jl)}}{\sqrt{1+\frac{\delta_s^2}{(n-2)}}}$$

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi normale* $N(\mu,\sigma^2)$.

Étape 2 : *on craint que deux observations* x_i *et* x_k *aient été générées simultanément par les sources contaminantes* $N\left(\mu+a_i\frac{\sigma}{\delta_i},\frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu+a_k\frac{\sigma}{\delta_k},\frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *non liés par une relation.*

Étape 3 : *calculer les poids* $d_{jl}^{(1)}$ *en remplissant le tableau suivant :*

x_i	j / i	1 2 ... n	j / i	1 2 ... n	j / i	1 2 ... n	j / i	1 2 ... n
x_1 x_2 ⋮ x_n	1 2 ⋮ n	$\bar{x}^{(ij)}$	1 2 ⋮ n	$A^{(ij)}$	1 2 ⋮ n	$\left[A^{(ij)}\right]^{-\frac{n-1}{2}}$	1 2 ⋮ n	$d_{ij}^{(2)}$
					somme			

Étape 4 : *calculer le rapport* γ_s :

$$\gamma_s=\frac{p\left(a_s>0\mid \delta_i^2,\delta_k^2;x_1,\ldots,x_n\right)}{p\left(a_s<0\mid \delta_i^2,\delta_k^2;x_1,\ldots,x_n\right)}=\frac{p\left(a_s>0\mid \delta_i^2,\delta_k^2;x_1,\ldots,x_n\right)}{1-p\left(a_s>0\mid \delta_i^2,\delta_k^2;x_1,\ldots,x_n\right)}, \qquad s=i \quad ou \quad s=k,$$

avec,

- $p\left(a_s > 0 \mid \delta_i^2\ \delta_k^2\ ; x_1,\ldots,x_n\right) = \frac{1}{2} + \sum\sum_{j\neq l} d_{jl}^{(2)} \sum_{m=0}^{\infty} \frac{2^{m+1/2}\,\Gamma\left(\frac{n}{2}+m\right)}{(2m+1)!\,\Gamma\left(\frac{n-1}{2}\right)} C_{2m+1}^{(s)}\left(j\ ,\ l\right)\left[1-\varepsilon_{jl}\right]^{(n-1)/2}$,

et,

$$C_{2m+1}^{(s)}\left(j\ ,\ l\right) = \frac{2^{m+1/2}}{\pi}\Gamma\left(m+\frac{3}{2}\right)\int_0^{v_{s(jl)}} \left(\varepsilon_{jl}-t^2\right)^m dt, \qquad v_{s(jl)} = \frac{z_{s(jl)}}{\sqrt{1+\frac{\delta_s^2}{(n-2)}}}$$

Étape 5 : *si* $\left(\gamma_s \geq 5\right)$ *ou* $\left(\gamma_s \leq \frac{1}{5}\right)$, $s = i$ *ou* $s = k$ *et si* $d_{ik}^{(2)} \notin \left[0\ ,\ \frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que les observations* x_i *et* x_k *sont singulières.*

f) Test de *Guttman et Khatri (1975)* : *cas où on craint que deux observations* x_i *et* x_k *aient été générées simultanément par les sources contaminantes* $N\left(\mu + a_i\frac{\sigma}{\delta_i}, \frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu + a_k\frac{\sigma}{\delta_k}, \frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *liés par la relation :* $a_i = -a_k = a$

Les hypothèses de ce test sont les mêmes que précédemment avec une hypothèse additionnelle qui stipule que $a_i = -a_k = a$. *Guttman et Khatri* montrent que la densité a posteriori de a, pour δ_i^2 et δ_k^2 fixés, est de la forme :

$$p\left(a \mid \delta_i^2\ ,\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \sum\sum_{j\neq l} w_{jl}\, h_{jl}\left(a\right),$$

où,

- la fonction h_{jl} est donnée par la formule :

$$h_{jl}(a) = \sum_{t=0}^{\infty} \frac{2^{t/2}\,\Gamma\left(\frac{n-1+t}{2}\right)\left(1-\varepsilon_{i(jl)}^2\right)^{(n-1)/2}}{t!\;\Gamma\left(\frac{n-1}{2}\right)\sqrt{2\pi}} \left(\varepsilon_{i(jl)}a\right)^t c^{(t+1)/2} \exp\left(-ca^2/2\right),$$

les différents paramètres apparaissant dans cette formule sont donnés par *Guttman et Khatri (1975)*

- les poids w_{jl} sont donnés par la formule :

- $$w_{jl} = \frac{\left(u_{i(jl)}\right)^{-(n-1)/2}}{\sum\sum_{j\neq l}\left(u_{i(jl)}\right)^{-(n-1)/2}}, \qquad \text{avec,}$$

$$u_{i(jl)} = A^{(jl)} + B_{jl} - cp_{jl}^2$$

Ils montrent aussi que

$$E\left(a \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \frac{\frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sum\sum_{j\neq l} w_{jl}\ \varepsilon_{i(jl)}}{\sqrt{c\left(1-\varepsilon_{i(jl)}^2\right)}}\ ;$$

$$Var\left(a \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right) = \frac{1}{c}\left\{1 + (n-1)\sum\sum_{j\neq l}\left(\frac{\varepsilon_{i(jl)}^2}{1-\varepsilon_{i(jl)}^2}\right)\right\} - \left[E\left(a \mid \delta_i^2\ \delta_k^2\ ; x_1, x_2, \ldots, x_n\right)\right]^2,$$

avec,

$$c = \frac{2n - 4 + \left(\delta_i + \delta_k\right)^2}{n - 2 + \delta_i^2 + \delta_k^2}$$

Le rapport des probabilités a posteriori est alors donné par la formule :

$$\gamma = \frac{p\left(a > 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)}{p\left(a < 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)} = \frac{p\left(a > 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)}{1 - p\left(a > 0 \mid \delta_i^2, \delta_k^2; x_1, \ldots, x_n\right)},$$

avec,

- $$p\left(a > 0 \mid \delta_i^2\ \delta_k^2\ ; x_1, \ldots, x_n\right) = \frac{1}{2}\sum\sum_{j \neq l} w_{jl} \left\{ \sum_{m=0}^{\infty} \frac{\Gamma\left(\frac{n}{2} + m\right) 2^{2m} \left(1 - \varepsilon_{i(jl)}^2\right)^{(n-1)/2} \varepsilon_{i(jl)}^{2m+1} \Gamma(m+1)}{\Gamma(2m+2)\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}} \right\}$$
-

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu, \sigma^2\right)$.

Étape 2 : *on craint que deux observations* x_i *et* x_k *aient été générées simultanément par les sources contaminantes* $N\left(\mu + a_i, \frac{\sigma^2}{\delta_i^2}\right)$ *et* $N\left(\mu + a_k, \frac{\sigma^2}{\delta_k^2}\right)$, *avec* a_i *et* a_k *liés par la relation :* $a_i = -a_k = a$.

Étape 3 : *calculer les poids* w_{jl} *en remplissant le tableau suivant :*

x_i	i , j	1 2 ... n	j / i	1 2 ... n	j / i	1 2 . . n
x_1 x_2 ⋮ x_n	1 2 ⋮ n	$\bar{x}^{(ij)}$	1 2 ⋮ n	$A^{(ij)}$	1 2 ⋮ n	w_{ij}
					somme	

Étape 4 *: calculer le rapport* γ *:*

$$\gamma = \frac{p\left(a>0 \mid \delta_i^2 , \delta_k^2; x_1,\ldots,x_n\right)}{p\left(a<0 \mid \delta_i^2 , \delta_k^2; x_1,\ldots,x_n\right)} = \frac{p\left(a>0 \mid \delta_i^2 , \delta_k^2; x_1,\ldots,x_n\right)}{1-p\left(a>0 \mid \delta_i^2 , \delta_k^2; x_1,\ldots,x_n\right)},$$

avec,

- $$p\left(a>0 \mid \delta_i^2\ \delta_k^2\ ; x_1,\ldots,x_n\right) = \frac{1}{2}\sum\sum_{j\neq l} w_{jl}\left\{\sum_{m=0}^{\infty}\frac{\Gamma\left(\frac{n}{2}+m\right)2^{2m}\left(1-\varepsilon_{i(jl)}^2\right)^{(n-1)/2}\varepsilon_{i(jl)}^{2m+1}\Gamma(m+1)}{\Gamma(2m+2)\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}}\right\}$$

Étape 5 : *si* $\left(\gamma \geq 5\right)$ *ou* $\left(\gamma \leq \frac{1}{5}\right)$, *et si* $t_{jk} \notin \left[0\ ,\ \frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que les observations* x_i *et* x_k *sont singulières.*

En conclusion, le modèle de détection de *Guttman (1973)* et *Guttman et Khatri (1975)* suppose que l'on dispose d'un échantillon de n observations indépendantes $x_1, x_2, \ldots, x_n$, provenant d'une loi normale $N\left(\mu, \sigma^2\right)$ mais, on craint toutefois qu'une (ou deux) de ces

observations pourrait avoir été générée par une loi normale $N(\mu+a\,,\sigma^2)$, $N(\mu+a\,,\sigma^2/\delta)$ ou $N\left(\mu+a\frac{\sigma}{\delta}\,,\frac{\sigma^2}{\delta^2}\right)$. Les tests résultants sont donnés dans le tableau 3.1.1. De plus, une application du test de *Guttman (1973)* est donnée à l'appendice A.

Numéro du test	**Source contaminante** (modèle sous H_1)	**Nombre d'observations douteuses**
a)	$N(\mu+a\,,\sigma^2)$	1
b)	$N\left(\mu+a\frac{\sigma}{\delta}\,,\frac{\sigma^2}{\delta^2}\right)$	1
c)	$N\left(\mu+a_i\,,\frac{\sigma^2}{\delta_i^2}\right)$ et $N\left(\mu+a_k\,,\frac{\sigma^2}{\delta_k^2}\right)$	2 (a_i et a_k non liés par une relation)
e)	$N\left(\mu+a_i\frac{\sigma}{\delta_i}\,,\frac{\sigma^2}{\delta_i^2}\right)$ et $N\left(\mu+a_k\frac{\sigma}{\delta_k}\,,\frac{\sigma^2}{\delta_k^2}\right)$	2 (a_i et a_k non liés par une relation et, contrairement à c, le paramètre de position est proportionnel à la variance de la source contaminante)
d)	$N\left(\mu+a_i\,,\frac{\sigma^2}{\delta_i^2}\right)$ et $N\left(\mu+a_k\,,\frac{\sigma^2}{\delta_k^2}\right)$	2 (a_i et a_k liés par la relation : $a_i=-a_k=a$)
f)	$N\left(\mu+a_i\frac{\sigma}{\delta_i}\,,\frac{\sigma^2}{\delta_i^2}\right)$ et $N\left(\mu+a_k\frac{\sigma}{\delta_k}\,,\frac{\sigma^2}{\delta_k^2}\right)$	2 (a_i et a_k liés par la relation : $a_i=-a_k=a$ et, contrairement à c, le paramètre de position est proportionnel à la variance de la source contaminante.)

Tableau 3.1.1 : résumé des tests de détection de *Guttman (1973)* et *Guttman et Khatri (1975)*

3.2. Tests de détection suivant le modèle de *de Alba et Van Ryzin (1979)*

Le modèle de *de Alba et Van Ryzin (1979)* peut être formulé de deux façons : un modèle *de changement de moyenne* (appelé encore Modèle A) et un modèle de *changement de variance* (appelé encore Modèle B). Ces deux types de modèles ont les formulations suivantes :

Modèle A

Dans ce modèle, on suppose que dans l'échantillon de taille n, $(n-k)$ observations ont été générées par la loi normale $N(\mu, \sigma^2)$ tandis que les k observations restantes ont été générées par la source contaminante $N(\mu+\delta, \sigma^2)$, $-\infty \leq \delta \leq \infty$. Le paramètre δ est supposé connu. Un tel modèle peut encore s'écrire mathématiquement :

$$\begin{cases} x_{i_1},\ldots,x_{i_{(n-k)}} \sim N(\mu, \sigma^2) \\ x_{i_j} \sim N(\mu+\delta, \sigma^2) \qquad j = n-k+1,\ldots,n \end{cases}$$

où $(i_1,\ldots,i_n)$ est une permutation des indices $(1,\ldots,n)$.

Modèle B

Dans ce modèle, on suppose que dans l'échantillon de taille n, $(n-k)$ observations ont été générées par la loi normale $N(\mu, \sigma^2)$ alors que les k observations restantes ont été générées par la source contaminante $N(\mu, \sigma^2\lambda)$, $(\lambda > 1)$. Ce modèle peut encore s'écrire mathématiquement :

$$\begin{cases} x_{i_1},\ldots,x_{i_{(n-k)}} \sim N(\mu,\sigma^2) \\ x_{i_j} \sim N(\mu,\sigma^2\lambda) \qquad j=n-k+1,\ldots,n \end{cases}$$

Les tests de détection d'observations singulières issus de ces modèles sont basés sur une *approche bayésienne empirique non standard*. Cette approche peut être résumée de la façon suivante :

On considère $(X_1,\Lambda_1),\ldots,(X_n,\Lambda_n)$, n paires mutuellement indépendantes de variables aléatoires, où $X_r\ (r=1,\ldots,n)$ est défini sur l'espace échantillonnal χ et $\Lambda_r\ (r=1,\ldots,n)$ sur l'espace paramétré Θ. Les $\Lambda_r\ (r=1,\ldots,n)$, sont supposés avoir une distribution commune a priori G sur Θ et la densité conditionnelle de X_r étant donne Λ_r est $f_{\Lambda_r}(x_r)$, $r=1,\ldots,n$.

La règle de Bayes empirique pour la r *ième* observations, $r=1,\ldots,n$, est notée $t_n^{(r)}(x_r)$. Si $t_n^{(r)}(x_r)$ est obtenu pour chaque $r=1,\ldots,n$ et que son risque relatif à G est noté par $r^*(t_n^{(r)},G)$, alors le vecteur, de dimension n, des fonctions de décision, $\vec{t}_n=\{t_n^{(r)}(X_r):\ r=1,\ldots,n\}$ est appelé une procédure empirique de décision de Bayes. Une telle procédure nous permet alors de décider si chaque observation $x_r\ (r=1,\ldots,n)$ est singulière ou non. La règle de Bayes empirique, ou encore le test de Bayes empirique, pour rejeter une observation singulière est :

$$t_n^{(r)}(x)=\begin{cases} 0 \quad si \quad \Delta_n^{(r)}(x)\geq 0 \quad (ne\ pas\ rejeter\ l'observation\ x) \\ 1 \quad si \quad \Delta_n^{(r)}<0 \quad (rejeter\ l'observation\ x) \end{cases}$$

Cette procédure conduit alors aux tests suivants :

a) Test de *de Alba et Van Ryzin (1979)* : *cas d'un modèle de changement de moyenne (Modèle A)*

On suppose que $x_1,\ldots,x_n$ sont n observations indépendantes générées par la loi normale $N(\mu,\sigma^2)$ mais, on craint qu'un certain nombre d'observations ait été généré par la source contaminante $N(\mu+\delta,\sigma^2)$, $-\infty \le \delta \le \infty$. Les hypothèses du test de Bayes empirique sont les suivantes :

$$\begin{cases} H_0 : x_r \sim N(\mu,\sigma^2) \\ H_1 : x_r \sim N(\mu \pm \delta,\sigma^2), \quad r=1,\ldots,n \end{cases} \Leftrightarrow \begin{cases} H_0 : \Lambda_r = 0 \\ H_1 : \Lambda_r = \delta \quad ou \quad \Lambda_r = -\delta \end{cases}$$

Nous sommes en présence d'une alternative de décalage («slippage»). La fonction de perte est de la forme

$$\begin{cases} L_0(\delta) = L_0(-\delta) = L_1(0) = constante \\ L_0(0) = L_1(\delta) = 0 \end{cases}$$

On suppose que la distribution a priori de Λ_r est de la forme

$$G = \left\{ \frac{(1-q)}{2}, q, \frac{(1-q)}{2} \right\},$$

où,

- $$\begin{cases} p(\Lambda_r = 0) = q \\ p(\Lambda_r = \delta) = p(\Lambda_r = -\delta) = q\dfrac{(1-q)}{2} \end{cases}$$

La règle de décision utilisée ou encore le test de Bayes empirique est alors de la forme :

- $t_n(x_r) = \begin{cases} 0 \quad si \quad \Delta_n(x_r) \geq 0 \quad (ne\ pas\ rejeter\ l'observation\ x_r) \\ \\ 1 \quad ailleurs \quad (rejeter\ l'observation\ x_r) \end{cases}, \qquad r = 1,\ldots,n,$

où,

$$\Delta_n(x) = (1-q)L_0(\delta)\tilde{f}_1(x) - qL_1(0)\tilde{f}_0(x),$$

- $\tilde{f}_{\delta^*}(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)\exp\left\{-\left(x-\tilde{\mu}-\delta^*\right)^2/2\tilde{\sigma}^2\right\}, \qquad \delta^* = -\delta, 0, \delta$

- $\tilde{f}_1(x) = \frac{1}{2}\tilde{f}_{-\delta}(x) + \frac{1}{2}\tilde{f}_{\delta}(x)$

- $\tilde{\sigma}^2 = \begin{cases} \sigma^{*2} \quad si \quad \sigma^{*2} > 0 \\ \\ \frac{1}{n} \quad ailleurs, \quad avec \quad \sigma^{*2} = s_2 - \frac{\delta^2}{6} + \frac{\sqrt{\max\left\{0, \left(\sigma^4 - 12s_4 + 36s_2^2\right)\right\}}}{6} \end{cases}$

- $\tilde{q} = \min\left\{1, q^*\right\}, \quad q^* = \max\left\{0, \frac{\left(1-s_2-\tilde{\sigma}^2\right)}{\delta^2}\right\}$

- $\tilde{\mu} = \bar{x}$

- $s_j = \frac{1}{n-1}\sum_{r=1}^{n}\left(x_r - \bar{x}\right)^j$

Une application pratique de ce test, conduira aux étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi normale* $N(\mu, \sigma^2)$.

Étape 2 : *on craint que la source* $N(\mu \pm \delta, \sigma^2)$ *(avec* δ *connu) a généré quelques contaminants dans l'échantillon de données.*

Étape 3 : *spécifier la fonction de perte* L. *On peut toutefois utiliser simplement, dans nos calculs, la relation* $L_0(\delta) = L_1(0)$.

Étape 4 : *calculer* $t_n(x_r)$.

Étape 5 : pour $r = 1,\ldots,n$, *si* $t_n(x_r) = 0$ *conclure que l'observation* x_r *n'est pas singulière. Si par contre* $t_n(x_r) = 1$, *conclure que l'observation* x_r *est singulière.*

b) Test de *de Alba et Van Ryzin (1979)* : *cas d'un modèle de changement de variance (Modèle B)*

on suppose que $x_1,\ldots,x_n$ sont n observations indépendantes générées par la loi normale $N(\mu, \sigma^2)$. On craint toutefois que quelques observations aient été générées par la source contaminante $N(\mu, \lambda\sigma^2)$, $\lambda > 1$. Les hypothèses du test de Bayes empirique sont les suivantes :

$$\begin{cases} H_0 : x_r \sim N(\mu, \sigma^2) \\ H_1 : x_r \sim N(\mu, \lambda\sigma^2), \quad r = 1,\ldots,n \end{cases} \Leftrightarrow \begin{cases} H_0 : \Lambda_r = 1 \\ H_1 : \Lambda_r = \lambda > 1 \end{cases}$$

Nous sommes aussi en présence d'une alternative de décalage («slippage»). La fonction de perte est de la forme

$$\begin{cases} L_0(\lambda) = L_1(1) = constante \\ L_1(\lambda) = L_0(1) = 0 \end{cases}$$

On suppose aussi que la distribution a priori de Λ_r est de la forme

$$G = \left\{ \frac{(1-q)}{2}, q, \frac{(1-q)}{2} \right\},$$

où

- $$\begin{cases} p(\Lambda_r = 1) = q \\ p(\Lambda_r = \lambda) = q\dfrac{(1-q)}{2} \end{cases}$$

Le test de Bayes empirique prend la forme suivante :

$$t_n(x_r) = \begin{cases} 0 \quad si \quad \Delta_n(x_r) \geq 0 \quad (ne\ pas\ rejeter\ l'observation\ x_r) \\ \\ 1 \quad ailleurs \quad (rejeter\ l'observation\ x_r) \end{cases}, \qquad r = 1,\ldots,n,$$

où,

$$\Delta_n(x) = (1-\tilde{q})L_0(\lambda)\tilde{f}_\lambda(x) - \tilde{q}L_1(1)\tilde{f}_1(x),$$

- $$\tilde{f}_{\lambda^*}(x) = \left(\frac{1}{\sigma\sqrt{2\pi\lambda^*}}\right)\exp\left\{-(x-\tilde{\mu})^2/2\tilde{\sigma}^2\lambda^*\right\}, \qquad \lambda^* = 1, \lambda$$

- $$\tilde{\sigma}^2 = \frac{3(m_2 - \tilde{\mu}^2)(1+\lambda) + \sqrt{\max\left\{0, \left[(m_2 - \tilde{\mu}^2)^2\ 9(1+\lambda)^2 - 12\lambda(m_4 - 6\tilde{\mu}^2 m_2 + 5\tilde{\mu}^4)\right]\right\}}}{6\lambda}$$

- $\widetilde{\mu} = \overline{x} \quad si \quad \mu \neq 0$

- $m_j = \frac{1}{n}\sum_{r=1}^{n} x_r^j$

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi normale* $N(\mu, \sigma^2)$.

Étape 2 : *on craint que la source* $N(\mu, \lambda\sigma^2)$ *(avec* $\lambda > 1$ *connu) ait généré quelques contaminants dans l'échantillon de données.*

Étape 3 : *spécifier la fonction de perte* L. *On peut toutefois utiliser, dans nos calculs, la relation* $L_0(\lambda) = L_1(1)$.

Étape 4 : *calculer* $t_n(x_r)$.

Étape 5 : *pour* $r = 1,\ldots,n$, *si* $t_n(x_r) = 0$ *conclure que l'observation* x_r *n'est pas singulière. Si par contre* $t_n(x_r) = 1$, *conclure que l'observation* x_r *est singulière.*

3.3. Tests de détection suivant le modèle de *Pettit et Smith (1983,1985)*

Dans les modèles précédents, on supposait toujours que les observations étaient générées soit par une loi normale $N(\theta, \sigma^2)$ ou par une source contaminante de même loi mais, avec un changement de moyenne ($N(\theta+\delta, \sigma^2)$), ou avec une variance modifiée ($N(\theta, \delta\sigma^2)$). De plus, on supposait toujours que le paramètre δ était connu, ce qui a beaucoup facilité le développement des procédures bayésiennes de détection d'observations singulières.

En pratique, la situation la plus commune est celle où nous sommes conscients de la possibilité d'observations singulières dans l'échantillon de données que nous avons à analyser mais où nous sommes par contre incapables de spécifier des a priori appropriés sur le paramètre δ parce que nous ne sommes pas sûres de son mécanisme générateur. *Freeman (1980)* avait déjà montré que si nous fixons des a priori inappropriés sur ce paramètre, cela aura comme conséquence que la procédure de Bayes utilisée va détecter plus d'observations singulières qu'il n'y en a en réalité. Dans le but d'éviter cette problématique, le présent modèle nous permet de rechercher les observations singulières parmi les observations les plus grandes (ou les moins grandes) de l'échantillon de données à analyser. Le *facteur de Bayes* utilisé, dans ce modèle, comme critère de détection d'observations singulières est exprimé de la façon suivante.

Supposons que nous avons deux modèles, M_0 et M_1 de la forme :

$$M_0 : x_1,\ldots,x_n \sim \xi(x \mid \mu) \qquad (\textit{une densité unimodale})$$

$$M_1 : x_1,\ldots,x_{n-1} \sim \xi(x \mid \mu),\ x_n \sim \xi(x \mid \mu+\delta)$$

(on suppose que l'observation x_n est le contaminant qui a été générée par la source de changement de moyenne de paramètre δ). Le *facteur de Bayes* est alors

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

où, ϑ est fonction des données de l'échantillon et c_0/c_1 un rapport de constantes inconnus dont la valeur est déterminée à l'aide de la technique des observations imaginaires proposée par *Spiegelhalter et Smith (1982)* . La détection d'une observation singulière à l'aide du *facteur de Bayes* est faite à partir de l'observation suivante :

Supposons que nous avons n observations et que nous souhaitons tester si l'une d'entre elles est singulière. Si de plus nous supposons, par hypothèse, qu'une observation singulière survient dans l'échantillon avec la probabilité α, alors la probabilité a priori qu'exactement

une observation est singulière est égale à $n\alpha(1-\alpha)^{n-1}$ et la probabilité qu'aucune observation est singulière est égale à $(1-\alpha)^n$. Donc, le rapport de probabilités a priori est donné par

$$\frac{P(\textit{aucune observation n'est singuliere})}{P(\textit{une observation est singuliere})} = \frac{1-\alpha}{\alpha}$$

Ainsi, pour qu'une observation singulière existe dans l'échantillon, il faut que le *facteur de Bayes* soit plus petit que le rapport $\alpha/(1-\alpha)$. En pratique, *Pettit (1992)* suggère que pour qu'une observation singulière soit survenu dans l'échantillon de données à analyser, il faut que le *facteur de Bayes* soit compris entre 0.005 et 0.015, c'est-à-dire :

$$\textit{facteur de Bayes} \in [0.005\,,\,0.015].$$

a) Test de *Pettit et Smith (1983, 1985)* : *cas où on craint qu'une observation ait été générée par une source normale*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi $N(\mu, \sigma^2)$ mais, on craint toutefois que l'observation x_n ait été générée par la source contaminante $N(\mu+\delta, \sigma^2)$. On pose alors :

$$M_0 : x_1, \ldots, x_n \sim N(\mu, \sigma^2)$$

$$M_1 : x_1, \ldots, x_{n-1} \sim N(\mu, \sigma^2),\quad x_n \sim N(\mu+\delta, \sigma^2)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0: \quad x_j \in N(\mu, \sigma^2) \qquad (j = 1,\ldots,n) \\ H_1: \quad x_n \in N(\mu + \delta, \sigma^2) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). *Pettit (1992)* démontre que le facteur de Bayes est de la forme

$$B_{01} = \frac{c_0}{c_1}\sqrt{\frac{n-1}{n}\left\{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n-1}(x_i - x^*)^2}\right\}^{-n}}$$

où,

- $x^* = \dfrac{\sum_{i=1}^{n-1} x_i}{n-1}$;
- $\dfrac{c_0}{c_1} = \sqrt{\dfrac{3}{2}}$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi normale* $N(\mu, \sigma^2)$.

Étape 2 : *on craint que la source* $N(\mu + \delta, \sigma^2)$ *a généré une observation dans l'échantillon de données et on ne possède pas d'a priori sur le paramètre* δ.

Étape 3 : *calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	Facteur de Bayes
x_1	B_{01} (x_1 *est considérée comme douteuse*)
x_2	B_{01} (x_2 *est considérée comme douteuse*)
$\vdots$	$\vdots$
x_n	B_{01} (x_n *est considérée comme douteuse*)

Étape 4: *pour* $i = 1,\ldots,n$, *si l'observation* x_i *est considérée comme étant douteuse et que le facteur de Bayes associé,* B_{01}, *est tel que :* $B_{01} \in [0.005\ ,\ 0.015]$, *conclure que l'observation* x_i *est singulière.*

b) Test de *Pettit et Smith (1983, 1985)* : *cas où on craint que deux observations aient été générées par deux sources normales différentes*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi $N(\mu\ ,\ \sigma^2)$ mais, on craint toutefois que les observations x_n et x_{n-1} aient été générées par les sources contaminantes $N(\mu+\delta_1\ ,\ \sigma^2)$ et $N(\mu+\delta_2\ ,\ \sigma^2)$. On pose alors :

$$M_0:\ x_1,\ldots,x_n \ \sim\ N(\mu\ ,\ \sigma^2)$$

$$M_2: x_1,\ldots,x_{n-2} \ \sim\ N(\mu\ ,\ \sigma^2),\ \ x_{n-1} \ \sim\ N(\mu+\delta_1\ ,\ \sigma^2), et\ \ x_n \ \sim\ N(\mu+\delta_2\ ,\ \sigma^2)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0:\ \ x_j \in N(\mu\ ,\ \sigma^2) \qquad (j = 1,\ldots,n) \\ H_1:\ \ x_{n-1} \in N(\mu+\delta_1\ ,\ \sigma^2) \ \ et \ \ x_n \in N(\mu+\delta_2\ ,\ \sigma^2) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). *Pettit (1992)* montre que le facteur de Bayes est de la forme

$$B_{02} = \frac{c_0}{c_2}\sqrt{\frac{n-2}{n}\left\{\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n-2}\left(x_i-x^{**}\right)^2}\right\}^{-n}},$$

où,

- $x^{**} = \dfrac{\sum_{i=1}^{n-2} x_i}{n-2}$;
- $\dfrac{c_0}{c_2} = \sqrt{2}$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu, \sigma^2\right)$.

Étape 2 : *on craint que les sources* $N\left(\mu+\delta_1, \sigma^2\right)$ *et* $N\left(\mu+\delta_2, \sigma^2\right)$ *ont généré chacune une observation dans l'échantillon de données et on ne possède pas d'a priori sur les paramètres* δ_1 *et* δ_2.

Étape 3 *: calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	**Facteur de Bayes**
x_1	B_{02} (*le couple* $\left(x_1, x_2\right)$ *est considéré comme douteux*)
x_2	B_{02} (*le couple* $\left(x_1, x_3\right)$ *est considéré comme douteux*)
$\vdots$	$\vdots$
x_n	B_{02} (*le couple* $\left(x_n, x_1\right)$ *est considéré comme douteux*)

Étape 4: *pour* $i = 1,\ldots,n$, *si le couple d'observations* $\left(x_i\ ,\ x_j\right)$, $i \neq j$ *est considéré comme douteux et que le facteur de Bayes associé,* B_{02}, *est tel que :* $B_{02} \in \left[0.005\ ,\ 0.015\right]$, *conclure que les observations* x_i *et* x_j *sont singulières.*

c) Test de *Pettit (1988)* : *cas où on craint qu'une observation singulière ait été générée par une source exponentielle*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi exponentielle $Exp(\theta)$ mais, on craint toutefois que l'observation x_n ait été générée par la source contaminante $Exp(\theta\delta)$. On pose alors :

$$M_0 :\ x_1, \ldots, x_n \ \sim\ Exp(\theta)$$

$$M_1 : x_1, \ldots, x_{n-1} \ \sim\ Exp(\theta),\ \ x_n \ \sim\ Exp(\theta\delta)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0 :\ \ x_j \in Exp(\theta) \qquad \left(j = 1,\ldots,n\right) \\ H_1 :\ \ x_n \in Exp(\theta\delta) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). *Pettit (1988)* a dérivé un test en supposant des a priori sur le paramètre δ. En supposant que

$$p\left(\theta \mid M_0\right) = c_0\theta^{-1},$$
$$p\left(\theta\ ,\ \delta \mid M_1\right) = c_1(\theta\delta)^{-1}$$

Pettit (1992) démontre que le facteur de Bayes est de la forme

$$B_{01} = \frac{c_0}{c_1} \frac{(n-1)\, x_n \left(\sum_{i=1}^{n-1} x_i \right)^{n-1}}{(n\bar{x})^n}$$

où,

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- $\dfrac{c_0}{c_1} = 4$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi exponentielle* $Exp(\theta)$.

Étape 2 : on *craint que la source* $Exp(\theta\delta)$ *a généré une observation dans l'échantillon de données et on ne possède pas d'a priori sur le paramètre* δ.

Étape 3 : *calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	**Facteur de Bayes**
x_1	B_{01} (x_1 *est considérée comme douteuse*)
x_2	B_{01} (x_2 *est considérée comme douteuse*)
⋮	⋮
x_n	B_{01} (x_n *est considérée comme douteuse*)

Étape 4: pour $i = 1,\ldots,n$, *si l'observation* x_i *est considérée comme étant douteuse et que le facteur de Bayes associé,* B_{01}, *est tel que :* $B_{01} \in [0.005\,,\, 0.015]$, *conclure que l'observation* x_i *est singulière.*

d) Test de *Pettit (1988)* *: cas où on craint que deux observations singulières aient été générées par deux sources exponentielles différentes*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi $Exp(\theta)$ mais, on craint toutefois que les observations x_n et x_{n-1} aient été générées par les sources contaminantes $Exp(\theta\delta_1)$ et $Exp(\theta\delta_2)$ On pose alors :

$$M_0 : x_1, \ldots, x_n \sim Exp(\theta)$$

$$M_2 : x_1, \ldots, x_{n-2} \sim Exp(\theta),\ x_{n-1} \sim Exp(\theta\delta_1),\ et\ x_n \sim Exp(\theta\delta_2)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0 : x_j \in Exp(\theta) \quad (j = 1, \ldots, n) \\ H_1 : x_{n-1} \in Exp(\theta\delta_1) \quad et \quad x_n \in Exp(\theta\delta_2) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). En supposant que

$$p(\theta, \delta_1, \delta_2) = c_2 (\theta\delta_1\delta_2)^{-1}$$

Pettit (1992) démontre que le facteur de Bayes est alors de la forme

$$B_{02} = \frac{c_0}{c_2} \frac{(n-1)(n-2)\, x_{n-1}\, x_n \left(\sum_{i=1}^{n-2} x_i \right)^{n-2}}{(n\bar{x})^n}$$

où,

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

- $\dfrac{c_0}{c_2} = \dfrac{27}{2}$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi exponentielle* $Exp(\theta)$.

Étape 2 : *on craint que les sources* $Exp(\theta\delta_1)$ *et* $Exp(\theta\delta_2)$ *ont généré chacune une observation dans l'échantillon de données et on ne possède pas d'a priori sur les paramètres* δ_1 *et* δ_2.

Étape 3 *: calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	**Facteur de Bayes**
x_1	B_{02} (*le couple* (x_1, x_2) *est considéré comme douteux*)
x_2	B_{02} (*le couple* (x_1, x_3) *est considéré comme douteux*)
$\vdots$	$\vdots$
x_n	B_{02} (*le couple* (x_n, x_1) *est considéré comme douteux*)

Étape 4: *pour* $i = 1,\ldots,n$, *si le couple d'observations* (x_i, x_j), $i \neq j$ *est considéré comme douteux et que le facteur de Bayes associé,* B_{02}, *est tel que :*

$B_{02} \in [0.005, 0.015]$, *conclure que les observations* x_i *et* x_j *sont singulières.*

e) **Test de *Pettit (1988)*** : *cas où on craint que deux observations aient été générées par deux sources exponentielles identiques*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi $Exp(\theta)$ mais, on craint toutefois que les observations x_n et x_{n-1} aient été générées par la source contaminante $Exp(\theta\delta)$. On pose alors :

$$M_0 : x_1, \ldots, x_n \sim Exp(\theta)$$

$$M_2 : x_1, \ldots, x_{n-2} \sim Exp(\theta),\quad x_{n-1} \sim Exp(\theta\delta),\ et\ \ x_n \sim Exp(\theta\delta)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0 : & x_j \in Exp(\theta) \quad (j = 1, \ldots, n) \\ H_1 : & x_{n-1} \in Exp(\theta\delta) \quad et \quad x_n \in Exp(\theta\delta) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). En supposant que

$$p(\theta, \delta) = c_2 (\theta\delta)^{-1},$$

Pettit (1992) démontre que le facteur de Bayes est alors de la forme

$$B_{02} = \frac{c_0}{c_2} \frac{(n-1)(n-2)(x_{n-1} + x_n)\left(\sum_{i=1}^{n-2} x_i\right)^{n-1}}{(n\bar{x})^n}$$

où,

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

- $\dfrac{c_0}{c_2} = \dfrac{27}{4}$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi exponentielle* $Exp(\theta)$.

Étape 2 : on *craint que la source* $Exp(\theta\delta)$ *a généré deux observations dans l'échantillon de données et on ne possède pas d'a priori sur le paramètre* δ.

Étape 3 : *calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	**Facteur de Bayes**
x_1	B_{02} (*le couple* (x_1, x_2) *est considéré comme douteux*)
x_2	B_{02} (*le couple* (x_1, x_3) *est considéré comme douteux*)
$\vdots$	$\vdots$
x_n	B_{02} (*le couple* (x_n, x_1) *est considéré comme douteux*)

Étape 4: pour $i = 1,\ldots,n$, *si le couple d'observations* (x_i, x_j), $i \neq j$ *est considéré comme douteux et que le facteur de Bayes associé,* B_{02}, *est tel que :* $B_{02} \in [0.005, 0.015]$, *conclure que les observations* x_i *et* x_j *sont singulières.*

f) Test de *Pettit (1994)* : *cas où on craint que k observations aient été générées par une source de Poisson $P(\theta\delta)$, avec δ connu*

On suppose que $x_1, x_2, \ldots, x_n$ sont n observations indépendantes qui ont été générées par la loi de Poisson $P(\theta)$ mais, on craint toutefois que k observations, $x_{n-k+1}, \ldots, x_n$ aient été générées par la source contaminante $P(\theta\delta)$, $\delta > 1$. On pose alors :

$$M_0: x_1, \ldots, x_n \sim P(\theta)$$

$$M_k: x_1, \ldots, x_{n-k} \sim P(\theta), \quad x_{n-k+1}, \ldots, x_n \sim P(\theta\delta)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0: & x_j \in P(\theta) \qquad (j = 1, \ldots, n) \\ H_1: & x_{n-k+1}, \ldots, x_n \in P(\theta\delta) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). Si l'on assigne une densité gamma a priori pour le paramètre θ, $G(a, b)$, il vient que le facteur de Bayes est de la forme

$$B_{0k} = \frac{c_0}{c_k} \left\{ \frac{b + n - k + k\delta}{b + n} \right\}^{a + n\bar{x}} \frac{1}{\delta^{\left(n\bar{x} - \sum_{i=1}^{n-k} x_i \right)}}$$

où,

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- $\dfrac{c_0}{c_k} = 1$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi de Poisson* $P(\theta)$

Étape 2 : *on craint que la source* $P(\theta\delta)$ *a généré k observations dans l'échantillon de données et on connaît la valeur du paramètre* δ.

Étape 3 *: calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	**Facteur de Bayes**	
x_1	B_{0k}	(x_1 *est considérée comme douteuse*)
x_2	B_{0k}	(x_2 *est considérée comme douteuse*)
$\vdots$		$\vdots$
x_n	B_{0k}	(x_n *est considérée comme douteuse*)

Étape 4: *pour* $i = 1,\ldots,n$, *si l'observation* x_i *est considérée comme étant singulière et que le facteur de Bayes associé,* B_{01}, *est tel que :* $B_{01} \in [0.005, 0.015]$, *conclure que l'observation* x_i *est singulière.*

g) Test de *Pettit (1994)* : *cas où on craint qu'une observation ait été générée par une source de Poisson* $P(\theta\delta)$, *avec* δ *inconnu et* θ *connu*

On suppose que $x_1, x_2, \ldots, x_n$ sont *n* observations indépendantes qui ont été générées par la loi de Poisson $P(\theta)$ mais, on craint toutefois que l'observation x_n ait été générée par la source contaminante $P(\theta\delta)$, $\delta > 1$. On pose alors :

$$M_0 : x_1, \ldots, x_n \sim P(\theta)$$

$$M_1 : x_1, \ldots, x_{n-k} \sim P(\theta), \quad x_n \sim P(\theta\delta)$$

Les hypothèses de ce test bayésien sont données par :

$$\begin{cases} H_0: & x_j \in P(\theta) \qquad (j = 1,\ldots,n) \\ H_1: & x_n \in P(\theta\delta) \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). Si l'on assigne une densité gamma a priori pour le paramètre θ, $G(a, b)$ et si l'on pose comme densité a priori pour le paramètre δ, $p(\delta) = c_1\delta^{-1}$, il est possible d'exprimer le facteur de Bayes sous la forme

$$B_{01} = \frac{c_0}{c_1} \frac{\Gamma(a + n\bar{x})}{\Gamma(x_n)\Gamma\left(a + \sum_{i=1}^{n-1} x_i\right)} \frac{(b+n-1)^{a+\sum_{i=1}^{n-1} x_i}}{(b+n)^{a+n\bar{x}}}$$

où,

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- $\dfrac{c_0}{c_1} = \dfrac{1}{\dfrac{\Gamma(a+2m)(b+1)^{a+m}}{\Gamma(m)\Gamma(a+m)(b+2)^{a+2m}}}$, on peut prendre $m = \dfrac{a}{b}$;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi de Poisson* $P(\theta)$.

Étape 2 : *on craint que la source* $P(\theta\delta)$ *a généré une observation dans l'échantillon de données et on connaît la valeur du paramètre* θ *et non celle de* δ.

Étape 3 *: calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	Facteur de Bayes
x_1	B_{01} (x_1 *est considérée comme douteuse*)
x_2	B_{01} (x_2 *est considérée comme douteuse*)
$\vdots$	$\vdots$
x_n	B_{01} (x_n *est considérée comme douteuse*)

Étape 4: *pour* $i = 1,\ldots,n$, *si l'observation* x_i *est considérée comme singulière et que le facteur de Bayes associé,* B_{01}, *est tel que :* $B_{01} \in [0.005\,,\, 0.015]$, *conclure que l'observation* x_i *est singulière.*

h) Test de *Pettit (1994)* : *cas où une observation singulière ait été générée par une source de Poisson* $P(\theta\delta)$, *avec* δ *inconnu et* θ *inconnu*

On suppose que $x_1, x_2, \ldots, x_n$ sont *n* observations indépendantes qui ont été générées par la loi de Poisson $P(\theta)$ mais, on craint toutefois que l'observation x_n ait été générée par la source contaminante. $P(\theta\delta)$, $\delta > 1$ On pose alors :

$$M_0 : x_1, \ldots, x_n \sim P(\theta)$$

$$M_1 : x_1, \ldots, x_{n-k} \sim P(\theta),\quad x_n \sim P(\theta\delta)$$

Les hypothèses de ce test bayésien sont alors données par :

$$\begin{cases} H_0 : x_j \in P(\theta) & (j = 1,\ldots,n) \\ H_1 : x_n \in P(\theta\delta) \end{cases}$$

Nous sommes aussi en présence d'une alternative de décalage («slippage»). Si on assigne une densité a priori pour le paramètre θ, $p(\theta) = c_0\theta^{-1}$ et si on pose que la densité a priori des paramètres δ et θ est : $p(\theta, \delta) = c_1\theta^{-1}\delta^{-1}$, il vient que le facteur de Bayes est de la forme

$$B_{01} = \frac{c_0}{c_1} \frac{\Gamma(n\bar{x})}{\Gamma(x_n)\Gamma\left(\sum_{i=1}^{n-1} x_i\right)} \frac{(n-1)^{\sum_{i=1}^{n-1} x_i}}{n^{n\bar{x}}}$$

où,

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

- $\dfrac{c_0}{c_1} = \dfrac{\Gamma(2x)}{\Gamma(x)^2 2^{2x}}$

  on peut prendre $x = 1$, *moyenne échantillonnale, ou médiane des observations* ;

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 : *vérifier que l'échantillon de données ait été généré par la loi de Poisson* $P(\theta)$

Étape 2 : on *craint que la source* $P(\theta\delta)$ *a généré une observation dans l'échantillon de données et on ne connaît ni la valeur du paramètre* θ *ni celle de* δ.

Étape 3 : *calculer le facteur de Bayes en remplissant le tableau suivant :*

Observations	Facteur de Bayes
x_1	B_{01} (x_1 *est considérée comme douteuse*)
x_2	B_{01} (x_2 *est considérée comme douteuse*)
$\vdots$	$\vdots$
x_n	B_{01} (x_n *est considérée comme douteuse*)

Étape 4: pour $i = 1,\ldots,n$, *si l'observation* x_i *est considérée comme étant singulière et que le facteur de Bayes associé,* B_{01}, *est tel que :* $B_{01} \in [0.005\,,\,0.015]$, *conclure que l'observation* x_i *est singulière.*

En conclusion, les tests de détection suivant le modèle de *Pettit et Smith (1983, 1985)* peuvent être résumés dans le tableau 3.3.1.

Test	Modèle sous H_0	Modèle sous H_1	Nombre de données douteuses
Pettit et Smith (1983, 1985)	$N(\mu,\sigma^2)$	$N(\mu+\delta,\sigma^2)$	1
Pettit et Smith (1983, 1985)	$N(\mu,\sigma^2)$	$N(\mu+\delta_i,\sigma^2)$	2
Pettit (1988)	$E(\theta)$	$E(\theta\delta)$	1
Pettit (1988)	$E(\theta)$	$E(\theta\delta_i)$	2
Pettit (1994)	$P(\theta)$	$P(\theta\delta)$	1
Pettit (1994)	$P(\theta)$	$P(\theta\delta_i)$	k

Tableau 3.3.1 : résumé des tests de détection suivant le modèle de *Pettit et Smith (1983, 1985)*

3.4. Tests de détection suivant le modèle *de Genshiro Kitagawa (1984)*

Dans ce modèle, la *vraisemblance prédictive* d'un modèle de Bayes est utilisée pour construire une procédure de détection d'observations singulières dans un échantillon. On suppose que celles-ci sont générées par des classes de modèles distincts (ou encore de distributions). La problématique associée consiste alors à dériver la *vraisemblance prédictive* d'*Akaike* («Akaike's prédictive likelihood») pour ensuite obtenir une *probabilité a posteriori quasi-bayésienne* qui permettra de discerner entre une observation singulière et une observation provenant du modèle de base.

a) «Test» de *Genshiro (1984)* : *cas où l'on soupçonne m observations singulières*

On considère un échantillon , $x_1,\ldots,x_n$, de n observations indépendantes générées par une distribution de base $N(\mu_0, \sigma^2)$. On craint toutefois que m de ces observations aient été générées par l'une ou l'autre des distributions alternatives $N(\mu_i, \sigma^2)$ $(i = 1,\ldots,k)$ qui sont des sources potentielles de contamination. Les μ_i désignent les changements de la moyenne de la population de base ($\mu_0 \pm$ *constante*). k, m *et* σ^2 représentent les paramètres inconnus. On suppose que $J = (j_1,\ldots,j_n)$ est un vecteur de n variables aléatoires indiquant la provenance de chacune des n observations. Si $j_i = 0$, on admettra que l'observation x_i a été générée par la source $N(\mu_0, \sigma^2)$, tandis que si $j_i \in \{1,\ldots,k\}$, x_i est une observation singulière qui a été générée par la source $N(\mu_{j_i}, \sigma^2)$. Les hypothèses de ce «test» bayésien sont données par :

$$\begin{cases} H_0: \; x_l \in N(\mu, \sigma^2) & (l = 1,\ldots,n) \\ H_1: \; x_l \in N(\mu_{j_i}, \sigma^2) & i \in \{1,\ldots,k\} \end{cases}$$

Nous sommes donc en présence d'une alternative de décalage («slippage»). Pour obtenir la *vraisemblance prédictive*, nous rappelons que dans la procédure *quasi-bayésienne* présentée par *Akaike (1980)* («Akaike's prédictive likelihood»), on part du principe que le meilleur ajustement d'un modèle statistique sur un ensemble de données est évalué à l'aide de la *log vraisemblance espérée* du modèle ajusté, $E_y \log\{f(y \mid \theta)\}$ qui est donnée par :

$$E_y \log\{f(y \mid \theta)\} = \int \log\{f(y \mid \theta)\}\, g(y)\, dy$$

où,

- $g(y)$ est la vraie densité et $f(y \mid \theta)$ est celle du modèle ;
- dans notre modèle, $\theta = (\mu_0, \mu_1, \ldots, \mu_k, \sigma^2)$

C'est une mesure de similarité entre la vraie distribution et celle associée avec le modèle. Plus cette mesure est petite et meilleur est l'ajustement du modèle estimé. Comme dans notre problématique nous avons à évaluer l'ajustement de plusieurs modèles sur un échantillon de données, la procédure du minimum du critère *AIC d'Akaike (1973)* permet d'obtenir une estimation non biaisée de la *log vraisemblance espérée* du modèle ajusté. Ainsi, en supposant qu'avant d'avoir collecté les données nous n'avons aucune information sur le paramètre θ, nous utilisons une distribution a priori non informative pour θ qui est :

$$\begin{aligned} p(\theta) &= p(\mu_0)\, p(\mu_1) \ldots p(\mu_k)\, p(\sigma^2) \\ &\propto \sigma^{-2} \end{aligned}$$

La distribution a posteriori de θ est alors

$$p(\theta \mid J ; x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid J ; \theta)\, p(\theta \mid J)}{\int f(x_1, \ldots, x_n \mid J ; \theta)\, p(\theta \mid J)\, d\theta}$$

où,

- $f\left(x_1,\ldots,x_n \mid J\right)=\pi^{-\frac{n-k-1}{2}}\left(n_0 n_1 \ldots n_k\right)^{-\frac{1}{2}}\Gamma\left(\frac{n-k-1}{2}\right)\left\{\sum_{l=0}^{k}\sum_{j_i=l}\left(x_i-\bar{x}_l\right)^2\right\}^{-\frac{n-k-1}{2}}, \qquad \bar{x}_l=\frac{\sum_{j_i=l} x_i}{n_l}$

Pour le $J^{ième}$ modèle, la probabilité a posteriori est de la forme :

$$p\left(J \mid x_1,\ldots,x_n\right)=\frac{p\left(x_1,\ldots,x_n \mid J\right)p(J)}{\sum p\left(x_1,\ldots,x_n \mid J\right)p(J)}$$

Or, pour que la procédure *quasi-bayésienne* soit significative, on doit remplacer $p\left(J \mid x_1,\ldots,x_n\right)$ par la *distribution prédictive* $p\left(y \mid J; x_1,\ldots,x_n\right)$:

$$f\left(y \mid J; x_1,\ldots,x_n\right)=2^{-\frac{k+1}{2}}\pi^{-\frac{2n-k-1}{2}}\left(n_0 n_1 \ldots n_k\right)^{\frac{1}{2}}\Gamma\left(\frac{2n-k-1}{2}\right)\left(d^2\right)^{-\frac{n-k-1}{2}}\frac{1}{f\left(x_1,\ldots,x_n \mid J\right)}, \quad \textit{avec}$$

$$d^2=\sum_{l=0}^{k}\sum_{j_i=l}\left\{\left(x_i-\bar{z}_l\right)^2+\left(y_i-\bar{z}_l\right)^2\right\}, \qquad \bar{z}_l=\frac{\sum_{j_i=l}\left(x_i+y_i\right)}{2n_l},$$

De façon similaire,

$$f\left(y \mid J; x_1,\ldots,x_n\right)=2^{-\frac{k+1}{2}}\pi^{-\frac{2n-k-1}{2}}\left(n_0 n_1 \ldots n_k\right)^{-\frac{1}{2}}\Gamma\left(\frac{2n-k-1}{2}\right)r\frac{1}{f\left(x_1,\ldots x_n \mid J\right)}, \quad \textit{avec}$$

$$r=\left\{\sum_{l=0}^{k}\sum_{j_i=l}\left\{\left(x_i-\bar{z}_l\right)^2\right\}\right\}^{-\frac{2n-k-1}{2}}$$

L'estimateur non biaisé de la log vraisemblance espérée est alors :

$$I\left(x_1,\ldots,x_n \mid J\right) = \log\left\{p\left(x \mid J\,;x\right)\right\} + E_x\left\{E_y \log\left\{p\left(y \mid J\,,x\right)\right\} - \log\left\{p\left(x \mid J\,;x\right)\right\}\right\}$$

$$= -\frac{n}{2}\log\left\{2\pi \sum_{l=0}^{k}\sum_{j_i=l}\left(x_i - \bar{x}_l\right)^2\right\} + \frac{n-k-1}{2}\log\{2\}$$

$$+ \log\left\{\Gamma\left(\frac{2n-k-1}{2}\right)\right\} - \log\left\{\Gamma\left(\frac{n-k-1}{2}\right)\right\}$$

$$-\frac{2n-k-1}{2}\left\{\psi\left(\frac{2n-k-1}{2}\right) - \psi\left(\frac{n-k-1}{2}\right)\right\},$$

où,

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

Il s'ensuit que la *vraisemblance prédictive* du modèle est donnée par :

$$L\left(x_1,\ldots,x_n \mid J\right) = \exp\left\{I\left(x_1,\ldots,x_n \mid J\right)\right\}$$

La procédure de détection, ou encore le «test» de détection d'observations singulières est définie à partir des observations suivantes :
La probabilité du modèle spécifié par J est de la forme :

$$p(J) = \frac{(m+1)!\,(n-m)!}{3^m\; n!}\sum_{k=h}^{n}\frac{\Gamma\left(k+\frac{1}{2}\right)}{k^m\;(k-h)!}$$

où,

- h dénote le nombre de sources de contamination apparaissant dans $J = \left(j_1,\ldots,j_n\right)$.

La probabilité a posteriori spécifiée par J est alors

$$p\left(J \mid x_1,\ldots,x_n\right) = \frac{L\left(x_1,\ldots,x_n \mid J\right)p(J)}{\sum_J L\left(x_1,\ldots,x_n \mid J\right)p(J)}$$

où,

- $L\left(x_1,\ldots,x_n \mid J\right)$ est la *vraisemblance prédictive* du modèle ;
- la somme est effectuée sur tous les J possibles, $j_i \neq 0$.

Cette probabilité a posteriori est exprimée en fonction de la *vraisemblance prédictive* Si elle est élevée, on admettra que les m observations spécifiées par le modèle J sont singulières.

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu_0 , \sigma^2\right)$

Étape 2 *: on craint que* m *observations aient été générées par l'une ou l'autre des* k *distributions alternatives,* $N\left(\mu_{j_i} , \sigma^2\right) \quad i \in \left\{1,\ldots,k\right\}$.

Étape 3 *: afin d'éviter un examen laborieux de plusieurs modèles* J *(c'est-à-dire examiner* k^m *modèles), on pourra supposer que les observations soupçonnées sont situées aux deux extrémités de l'échantillon de données, c'est-à-dire examiner les observations les plus élevées et les moins élevées. On calculera ensuite la probabilité a posteriori en remplissant le tableau suivant :*

J (les l modèles proposés)	Les m_i observations soupçonnées	Probabilité a posteriori
1 $\{j_{11},\ldots,j_{1n}\}$	$x_1,\ldots,x_{m_1}$	
2 $\{j_{21},\ldots,j_{2n}\}$	$x_1,\ldots,x_{m2}$	$p(J \mid x_1,\ldots,x_n)=\frac{L(x_1,\ldots,x_n \mid J)p(J)}{\sum_J L(x_1,\ldots,x_n \mid J)p(J)}$
$\vdots$ $\vdots$.	$\vdots$	
l $\{j_{l1},\ldots,j_{ln}\}$	$x_1,\ldots,x_{m_l}$	

Étape 4 : *Pour le modèle J spécifié, si sa probabilité a posteriori est élevée, conclure que les m observations soupçonnées sont singulières.*

b) «Test» de *Genshiro (1984)* : *cas où l'on soupçonne une observation singulière*

On considère toujours un échantillon , $x_1,\ldots,x_n$, de n observations indépendantes générées par une distribution de base $N(\mu_0,\sigma^2)$ mais on craint que l'observation x_j ait été générée par l'une ou l'autre des distributions alternatives $N(\mu_i,\sigma^2)$ $(i=1,\ldots,k)$ qui sont des sources potentielles de contamination. Les μ_i designent les changements de la moyenne de la population de base ($\mu_0 \pm$ *constante*). k, m *et* σ^2 sont des paramètres inconnus. On pose toujours $J=(j_1,\ldots,j_n)$ un vecteur de n variables aléatoires indiquant la provenance de chacune des n observations. Les hypothèses de ce «test» bayésien sont données par :

$$\begin{cases} H_0: & x_l \in N(\mu,\sigma^2) \qquad (l=1,\ldots,n) \\ H_1: & x_j \in N(\mu_{j_i},\sigma^2) \qquad i \in \{1,\ldots,k\} \end{cases}$$

Nous sommes aussi en présence d'une alternative de décalage («slippage»). Les résultats obtenus précédemment sont toujours valides dans ce cas-ci. En reprenant l'expression de la probabilité a posteriori spécifiée par J ,

$$p\left(J \mid x_1,\ldots,x_n\right) = \frac{L\left(x_1,\ldots,x_n \mid J\right)p(J)}{\sum_J L\left(x_1,\ldots,x_n \mid J\right)p(J)}$$

où,

- $L\left(x_1,\ldots,x_n \mid J\right)$ est la vraisemblance prédictive du modèle ;
- la somme est effectuée sur tous les J possibles, $j_i \neq 0$,

Une façon simple de détecter une observation singulière avec un tel modèle consiste à évaluer la probabilité marginale a posteriori qu'une observation spécifique, x_i, soit singulière, c'est-à-dire calculer la probabilité $p\left(x_i \mid x_1,\ldots,x_n\right)$ comme :

$$p\left(x_i \mid x_1,\ldots,x_n\right) = \sum_J p\left(J \mid x_1,\ldots,x_n\right)$$

Ainsi, si cette probabilité est élevée, on conclura que l'observation x_i du modèle J spécifié est singulière.

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale* $N\left(\mu_0 , \sigma^2\right)$.

Étape 2 *: on craint qu'une observation ait été générée par l'une ou l'autre des* k *distributions alternatives,* $N\left(\mu_{j_i} , \sigma^2\right) \quad i \in \left\{1,\ldots,k\right\}$.

Étape 3 *: afin d'éviter un examen laborieux de plusieurs modèles J (c'est-à-dire examiner k^m modèles), on commencera par ordonner l'échantillon de données de la façon suivante :* $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

Étape 4 *: pour un certain L fixé, calculer la probabilité marginale a posteriori que* $x_{(i)}$ *est une observation singulière,* $p\left(x_i \mid x_1,\ldots,x_n\right) = \sum_J p\left(J \mid x_1,\ldots,x_n\right)$, *avec* $L < i \leq n - L$, *en remplissant le tableau suivant :*

J **(les** l **modèles proposés)**	**L'observation** x_i **soupçonnée**	**Probabilité marginale a posteriori que** x_i **est singulière**
1 $\{j_{11},\ldots,j_{1n}\}$	x_1	
2 $\{j_{21},\ldots,j_{2n}\}$	x_2	$p\left(x_i \mid x_1,\ldots,x_n\right) = \sum_J p\left(J \mid x_1,\ldots,x_n\right)$
$\vdots$ $\vdots$	$\vdots$	
l $\{j_{l1},\ldots,j_{ln}\}$	x_l	

Étape 5 : p*our le modèle J spécifié, si la probabilité marginale a posteriori que l'observation* $x_{(i)}$ *est élevée, conclure que* x_i *est une observation singulière.*

3.5. Conclusion

Dans ce chapitre, nous avons fourni une revue bibliographique non exhaustive des tests bayésiens de détection d'observations singulières dans un échantillon aléatoire univarié. Ces différents tests sont les plus utilisés en pratique et /ou proposés récemment dans la littérature.

L'approche de détection d'observations singulières dans un échantillon de données comporte essentiellement trois grandes phases : l'ajustement d'une loi de probabilité à l'échantillon de données, puis l'identification d'observations apparaissant surprenantes dans l'ensemble de l'échantillon et, enfin, à partir du tableau 3.5.1, choisir le test de détection s'approchant le mieux à la problématique de vérification de la qualité de l'échantillon.

La première phase est surtout descriptive ; elle a pour but de définir la loi de probabilité ou encore la population statistique à partir de laquelle l'échantillon de données a été tiré. La seconde phase par contre permet d'identifier les observations qui, à notre avis, ne semblent pas avoir été générées par la population statistique retenue. Enfin vient la troisième et dernière phase, celle du choix du test de détection approprié, qui consiste à tirer des conclusions du travail d'analyse, phase décisive car le test retenu va déterminer l'action d'admettre ou non que les observations retenues à la deuxième phase sont singulières.

À la lumière de toutes ces informations, nous pouvons dériver les recommandations suivantes pour l'expérimentateur, analyste de données qui souhaite utiliser ces tests pour vérifier la qualité d'un échantillon :

a) en accord avec les données à étudier, trouver le modèle statistique s'ajustant à vos données ;

b) relever le nombre de données douteuses de l'échantillon ;

c) à partir du tableau 3.5.1, choisir le test de détection s'approchant le mieux à la problématique de vérification de la qualité des données de l'échantillon.

Test	Modèle sous H_0	Modèle sous H_1	Nombre de données douteuses	Jugement sur l'application du test
Guttman (1973)	$N(\mu,\sigma^2)$	$N(\mu+a,\sigma^2)$	1	calculs faciles
Khatri (1975)	$N(\mu,\sigma^2)$	$N\left(\mu+a\frac{\sigma}{\delta},\frac{\sigma^2}{\delta}\right)$	1	calculs longs
Guttman et Khatri (1975)	$N(\mu,\sigma^2)$	$N\left(\mu+a_i,\frac{\sigma^2}{\delta^2}\right)$	2	calculs longs
Guttman et Khatri (1975)	$N(\mu,\sigma^2)$	$N\left(\mu+a_i\frac{\sigma}{\delta_i},\frac{\sigma^2}{\delta_i^2}\right)$	2	calculs longs
de Alba et Vanryzin (1979)	$N(\mu,\sigma^2)$	$N(\mu\pm\delta,\sigma^2)$	1	calculs très faciles
de Alba et Vanryzin (1979)	$N(\mu,\sigma^2)$	$N(\mu,\lambda\sigma^2)$	1	calculs très faciles
Pettit et Smith (1983, 1985)	$N(\mu,\sigma^2)$	$N(\mu+\delta,\sigma^2)$	1	calculs très faciles
Pettit et Smith (1983, 1985)	$N(\mu,\sigma^2)$	$N(\mu+\delta_i,\sigma^2)$	2	calculs très faciles
Genshiro (1984)	$N(\mu,\sigma^2)$	$N(\mu_i,\sigma^2)$	1	calculs longs
Genshiro (1984)	$N(\mu,\sigma^2)$	$N(\mu_{ij},\sigma^2)$	m	calculs longs
Pettit (1988)	$E(\theta)$	$E(\theta\delta)$	1	calculs très faciles
Pettit (1988)	$E(\theta)$	$E(\theta\delta_i)$	2	calculs très faciles
Pettit (1994)	$P(\theta)$	$P(\theta\delta)$	1	calculs très faciles
Pettit (1994)	$P(\theta)$	$P(\theta\delta_i)$	k	calculs faciles

Tableau 3.5.1 : résumé des tests de détection dans un échantillon univarié

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

Dans le cas univarié, nous avons identifié subjectivement une observation singulière comme une donnée qui engendre la «surprise» dans sa valeur extrême relative à celle d'autres membres de l'échantillon. De ce fait, il suffit d'examiner les points dans les queues de la distribution échantillonnale pour tester la présence d'observations singulières.

En situation multivarié, cette détection requiert beaucoup plus de précaution. En effet, nous aimerions garder le stimuli de donnée singulière comme celle qui engendre la «surprise» dans un échantillon de IR^n $(n > 1)$. Mais, qu'est-ce qu'une donnée singulière dans un tel échantillon ? Une définition exempte de toute critique est impossible à donner dans l'absolu, car si une observation est singulière pour un modèle elle ne l'est pas forcément pour un autre et de nombreux types de singularités peuvent survenir en statistique multivariée comme le souligne *Gnanadesikan (1977)*. Le concept de donnée qui engendre la «surprise» devient donc beaucoup plus nébuleux.

Dans un modèle de régression linéaire simple où nous avons une variable dépendante et une variable explicative, si nous suivons la définition très générale de *Grubbs (1969)* donnée à la section 2.1, les données douteuses se retrouveront à la périphérie du nuage de points formé par l'échantillon. Cependant, il est clair que dans un modèle de régression linéaire multiple, la détection de données douteuses apparaîtra beaucoup moins évidente. De plus, l'utilisation de techniques univariées appliquées aux projections sur chaque axe ne conduit pas nécessairement à un bon résultat comme le montre la figure 4.1

Dans ce chapitre, nous examinons le problème de la détection d'observations singulières dans un modèle de régression linéaire.

Le modèle classique de régression linéaire univarié a la forme suivante :

$$Y = X\beta + \varepsilon \qquad (4.1)$$

où,

- Y est un vecteur $(n \times 1)$ de variables aléatoires normales appelées variables réponse ;
- X est une matrice $(n \times p)$ connu de plein rang $p < n$ appelée matrice des variables

explicatives ;

- β est un vecteur $(n \times 1)$ des paramètres ;
- ε est un vecteur $(n \times 1)$ des erreurs indépendantes et identiquement distribuées selon une loi normale $N(o, \sigma^2 I_n)$.

En ajustant une droite de régression à ce modèle, nous sommes confrontés à la présence d'observations x_{ij}, $i = 1,\ldots,n \quad j = 1,\ldots,p$ aussi bien que y_i qui peuvent être singulières. Ainsi, une observation de l'échantillon $(x_{i1},\ldots,x_{ip},y_i)$, $i = 1,\ldots,n$ ayant une valeur extrême sur la variable réponse, sera appelé une *observation singulière dans le sens vertical*, de même qu'une observation ayant une valeur extrême pour une variable explicative sera appelée *une observation singulière à effet de levier*, c'est-à-dire une observation qui exerce une influence sur l'ajustement de la droite de régression linéaire.

Deux principales procédures de détection d'observations singulières ont émergées dans l'usage de l'approche bayésienne. La première se borne à postuler, sous l'hypothèse nulle H_0, que les données ont été générées suivant le modèle (4.1) et aucune hypothèse alternative n'est cependant spécifiée. La deuxième procédure prend en compte un modèle alternatif pour la génération d'un sous ensemble de l'échantillon de données, caractérisant ainsi l'hypothèse alternative H_1. Les modèles alternatifs les plus couramment utilisés sont des modèles de changement de moyenne et de changement de variance.

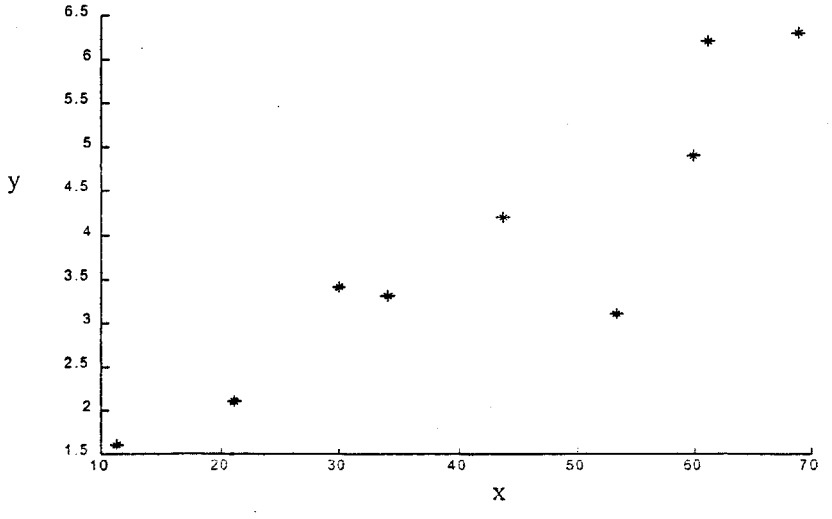


Figure 4.1 exemple d'un nuage de points illustrant la relation linéaire entre les variables x et y

4.1. Procédures bayésiennes de détection utilisant essentiellement le modèle sous l'hypothèse nulle

Les procédures de détection retenues dans cette partie soutiennent que sous l'hypothèse nulle, H_0, l'échantillon de données $\left(x_{i1},\ldots,x_{ip},y_i\right)$, $i=1,\ldots,n$ a été généré conformément avec le modèle de régression linéaire spécifié à l'équation (4.1). L'estimation des paramètres $\left(\beta,\sigma^2\right)$ est faite en supposant une distribution a priori non informative. Ainsi,

$$p\left(\beta,\sigma^2\right)\propto\left(\sigma^2\right)^{-1},$$

et dans ce cas la distribution a posteriori pour les paramètres d'un tel modèle est conditionnée sur σ^2, nous avons alors :

$$\beta\sim N\left(\hat{\beta},\sigma^2(X'X)^{-1}\right),\ \text{avec}\ \hat{\beta}=(X'X)^{-1}X'Y$$

et $p\left(\sigma^2/y,x\right)$ est la densité d'une variable aléatoire $\chi^2_{n-p}/\left\{(n-p)s^2\right\}$,

où,

$$(n-p)s^2=\hat{\varepsilon}'\hat{\varepsilon}=Y'\left(I_n-H\right)Y$$

$H=X(X'X)^{-1}X'$ est la matrice chapeau.

Ces remarques inspirent les procédures de détection présentées dans cette section.

a) Procédure de détection de *Chaloner et Brant (1988)*

Dans cette procédure une observation singulière est définie comme étant celle qui a un résidu, $\hat{\varepsilon}$, élevé dans le modèle de régression linéaire spécifié sous l'hypothèse nulle, ou encore sous le modèle spécifié à l'équation (4.1). Les observations singulières sont alors identifiées en

regardant les probabilités a posteriori des résidus. Ainsi, dans l'échantillon de données $(x_{i1},\ldots,x_{ip},y_i)$, $i=1,\ldots,n$, on dira que la i ème observation est singulière si $|\varepsilon_i| > k\sigma$ pour un certain choix de k. La probabilité a posteriori, p_i, que $|\varepsilon_i| > k\sigma$ est donnée par :

$$p_i = P(|\varepsilon_i| > k\sigma / y,x) = \int_0^{\infty} \{1-\Phi(z_1)+\Phi(z_2)\} p(\tau/y,x) d(\tau)$$

où,

- $\tau = \sigma^{-2}$;

- $p(\tau/y,x)$ est la densité d'une distribution gamma, $\Gamma\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right)$;

- $z_1 = \frac{k-\hat{\varepsilon}_i\sqrt{\tau}}{\sqrt{h_i}}$, $\quad z_2 = \frac{k+\hat{\varepsilon}_i\sqrt{\tau}}{\sqrt{h_i}}$

avec, h_i le i ème élément de la diagonale principale de la matrice H et

$$\hat{\varepsilon} = Y - X\hat{\beta}$$

- $\Phi(z)$ est la fonction de répartition d'une loi normale centrée réduite.

La valeur de k peut être choisie de telle sorte que la probabilité a priori qu'il n'existe aucune observation singulière soit grande, c'est-à-dire :

$$k = \Phi^{-1}\left\{0.5 + \frac{(0.95)^{1/n}}{2}\right\}$$

La i ème observation sera alors suspectée d'être singulière si $p_i > 2\Phi(-k)$.

Pour une application pratique de cette procédure de détection, on suivra les étapes suivantes :

Étape 1 *: s'assurer que les n observations aient été tirées conformément au modèle* (4.1)

Étape 2 *: on craint que la i ème observation soit singulière.*

Étape 3 *: calculer*

$$k = \Phi^{-1}\left\{0.5 + \frac{(0.95)^{1/n}}{2}\right\}$$

$$p_i = P\left(|\varepsilon_i| > k\sigma / y, x\right) = \int_0^{\infty} \left\{1 - \Phi(z_1) + \Phi(z_2)\right\} p(\tau / y, x) d(\tau)$$

Étape 4 *: si pour la i ème observation, nous avons* $p_i > 2\Phi(-k)$, *décider qu'elle est singulière.*

b) Procédure de détection de *Pena et Guttman (1993)*

Cette procédure améliore l'approche de *Chaloner et Brant* dans laquelle la probabilité p_i, ne prend pas en compte l'effet de levier pour des échantillons relativement grands. Toutefois, le concept de base reste le même que celui présenté précédemment.

Dans l'approche de *Chaloner et Brant*, nous avions :

$$p_i = P\left(|\varepsilon_i| > k\sigma / y, x\right) = \int_0^{\infty} \left\{1 - \Phi(z_1) + \Phi(z_2)\right\} p(\tau / y, x) d(\tau)$$

ainsi, en écrivant,

$$p\left(\varepsilon_i, \sigma^2 / y, x\right) = p\left(\varepsilon_i / \sigma^2, y, x\right) p\left(\sigma^2 / y, x\right)$$

La probabilité a posteriori p_i peut encore s'écrire :

$$p_i = p_{i1} + p_{i2},$$

où,

$$p_{i1} = p(\varepsilon_i > k\sigma / y, x) \text{ et } p_{i2} = p(\varepsilon_i < -k\sigma / y, x)$$

Johnson et Welch (1940) ont montré qu'une approximation de cette expression pour n grand est donné par :

$$p_i \approx 1 - \Phi(u_1) + \Phi(u_2)$$

où,

$$u_j = \frac{r_i/\sqrt{l_i} - (-1)^j k/\sqrt{h_i}}{\sqrt{1 + \frac{1}{2}(n-p)^{-1} r_i^2 / l_i}} \quad (j = 1,2)$$

où,

- $r_i = \dfrac{\hat{\varepsilon}_i}{\sqrt{s^2(1-h_i)}}$, est le résidu studentisé ;

- $l_i = \dfrac{h_i}{1-h_i}$, est la mesure de l'effet de levier.

En observant que : $\lim_{h_i} p_i = 2\Phi(-k)$, nous voyons que la probabilité a posteriori qu'une observation ayant un grand effet de levier $(h_i \approx 1)$ est singulière au sens de *Chaloner et Brant* pour des échantillons relativement grands.

En pratique, nous suivrons les différentes étapes suivantes pour appliquer ce test :

Étape 1 *: s'assurer que les n observations aient été tirées conformément au modèle* (4.1)

Étape 2 *: on craint que la i ème observation soit singulière.*

Étape 3 *: calculer*

$$k = \Phi^{-1}\left\{0.5 + \frac{(0.95)^{1/n}}{2}\right\} \text{ et } p_i \approx 1 - \Phi(u_1) + \Phi(u_2)$$

Étape 4 *: si pour la i ème observation, nous avons* $p_i \approx 2\Phi(-k)$, pour $(h_i \approx 1)$, décider qu'elle est singulière.

4.2. Procédures bayésiennes de détection utilisant un modèle alternatif à l'hypothèse nulle

Dans ces procédures, bien que l'équation (4.1) soit le modèle souhaité pour la génération de l'échantillon de données $\left(x_{i1},\ldots,x_{ip},y_i\right)$, $i=1,\ldots,n$, l'expérimentateur craint (à cause de son expérience) que certaines observations, c'est-à-dire $\left(x_{i_t1},\ldots,x_{i_tp},y_{i_t}\right)$, $t=1,\ldots,k$, avec k fixé et tel que $k << n/2$, aient été générées par une source contaminante se manifestant par un changement de moyenne ou de variance. Afin de présenter ces procédures avec plus de clarté, nous adoptons la notation suivante :

- $I=\left\{i_1,\ldots,i_k\right\}$ est un ensemble de k entiers distincts choisi parmi l'ensemble $\left\{1,\ldots,n\right\}$, dans l'échantillon de données, les observations ayant un indice appartenant à l'ensemble I sont considérées comme avoir été générées par la source contaminante envisagée ;
- Y peut être décomposé comme $Y'=\left(Y_I',Y_{(I)}'\right)$, où (I) signifie «supprimer le groupe d'observations ayant un indice appartenant à I» ;
- X peut être partitionné comme $X'=\left(X_I',X_{(I)}'\right)$;
- $\hat{\beta}_{(I)}$ et $s_{(I)}^2$ sont les estimateurs de β et σ^2 basés sur $\left(X_{(I)},Y_{(I)}\right)$, etc...

Par opposition avec la procédure précédente basée essentiellement sur l'hypothèse nulle, la présente approche utilise deux modèles alternatifs. Le premier est de type changement de moyenne qui prend la forme suivante, pour la génération des observations :

$$Y=\begin{pmatrix}y_{(I)}\\ \cdots\\ y_I\end{pmatrix}=\begin{pmatrix}X_{(I)}\\ \cdots\\ X_I\end{pmatrix}\beta+\begin{pmatrix}0\\ \cdots\\ a\end{pmatrix}+\begin{pmatrix}\varepsilon_{(I)}\\ \cdots\\ \varepsilon_I\end{pmatrix}$$

où, a est un vecteur $(k \times 1)$ d'un changement constant de la moyenne, et $\varepsilon_I \sim N(0, \sigma^2 I_k)$ indépendant de $\varepsilon_{(I)} \sim N(0, \sigma^2 I_{n-k})$.

Le second modèle sera impliqué avec un changement de la variance qui prend la forme, pour la génération des observations :

$$Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix}$$

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 $\delta^2 > 1$.

c) Procédure de détection de *Guttman et al. (1978)*

Dans cette procédure, bien qu'il soit souhaitable que les n observations de l'échantillon aient été générées conformément au modèle (4.1), on craint toutefois que k d'entre elles aient été générées par une source contaminante associée à un changement de la moyenne. Les hypothèses reliées à cette procédure sont alors les suivantes :

$$\begin{cases} H_0 : Y = X\beta + \varepsilon \\ H_1 : Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} 0 \\ \cdots \\ a \end{pmatrix} + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix} \end{cases}$$

En supposant que l'on ne possède qu'une information vague sur les paramètres de changement $a = (a_1, \ldots, a_k)$, la distribution a priori non informative de (a, β, σ^2) est :

$$p(a, \beta, \sigma^2) \propto (\sigma^2)^{-1}$$

La distribution a posteriori de β est alors :

$$p(\beta/Y,X)=\sum c_I h\left(\beta\Big/\hat{\beta}_{(I)}\,;\frac{n-k-p}{S_{(I)}}X'_{(I)}X_{(I)}\,;n-k-p\,;p\right)$$

où,

- $c_I=\dfrac{\sqrt{\left\{S_{(I)}^{-(n-k)}\left[S_{(I)}^{p}\left|\left(X'_{(I)}X_{(I)}\right)^{-1}\right|\right]\right\}}}{\sum\sqrt{\left\{S_{(I)}^{-(n-k)}\left[S_{(I)}^{p}\left|\left(X'_{(I)}X_{(I)}\right)^{-1}\right|\right]\right\}}}$

  les $\binom{n}{k}$ poids sont les probabilités a posteriori que les observations $y_{i_1},\ldots,y_{i_k}$ sont singulières ;

- $\sum$ représente la somme à travers tous les $\binom{n}{k}$ ensembles possibles de I ;

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

De plus, on montre que :

$$E(\beta/Y,X)=\sum c_I\hat{\beta}_{(I)}.$$

et que $V(\beta/Y,X)=\sum c_I\left[\dfrac{S_{(I)}}{n-k-p-2}\left(X'_{(I)}X_{(I)}\right)^{-1}+\hat{\beta}'_{(I)}\hat{\beta}_{(I)}\right]-\left(E(\beta/Y,X)\right)'\left(E(\beta/Y,X)\right)$

Ainsi, les k observations $\left(x_{i_t1},\ldots,x_{i_tp},y_{i_t}\right)$, $t=1,\ldots,k$ seront singulières si leur poids c_I est le plus important parmi tous les $\binom{n}{k}$ poids. Par ailleurs, le choix de la valeur k à retenir pourra être celle qui minimise la trace de la matrice de variance-covariance a posteriori du paramètre β qui est une mesure de la précision de l'estimation de ce paramètre, c'est-à-dire :

$$TrV(\beta/Y,X)=\sum_{i=1}^{p}V_{ii}(\beta/k)$$

Pour une application pratique de cette procédure de détection, nous suivrons les différentes étapes suivantes :

Étape 1 *: s'assurer que les n observations aient été tirées conformément au modèle* (4.1)

Étape 2 *: on craint que parmi les n observations, k d'entre elles aient été générées par une source contaminante associée à un changement dans la moyenne de Y.*

Étape 3 *: choisir dans un premier temps différentes valeurs de k et déterminer la meilleure valeur de k qui minimise* $TrV(\beta/Y,X)=\sum_{i=1}^{p}V_{ii}(\beta/k)$.

Étape 4 *: avec la meilleure valeur de k retenue à l'étape précédente, calculer les* $\binom{n}{k}$ *poids* c_l.

Étape 5 *:décider que les k observations* $\left(x_{i_t1},\ldots,x_{i_tp},y_{i_t}\right)$, $t=1;\ldots,k$ *sont singulières si leur poids* c_l *est le plus important parmi tous les* $\binom{n}{k}$.

Un exemple d'application de cette procédure est donné à l'appendice B.

d) Procédure de détection de *Pena et Tiao (1992)*

Contrairement à la procédure précédente, bien qu'il soit désirable que les n observations de l'échantillon aient été générées conformément au modèle décrit par l'équation (4.1), on craint maintenant que k d'entre elles aient été générées par une source contaminante associée à un changement dans la variance. Les hypothèses de cette procédure sont alors les suivantes :

$$\begin{cases} H_0 : Y = X\beta + \varepsilon \\ H_1 : Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix} \end{cases}$$

En supposant que l'on ne possède qu'une information vague sur les paramètres $(\delta^2, \beta, \sigma^2)$ on pose que la distribution a priori non informative de $(\delta^2, \beta, \sigma^2)$ est :

$$p(\delta^2, \beta, \sigma^2) \propto (\sigma^2)^{-1}$$

La probabilité a posteriori que le groupe d'observations $y_{i_1}, \ldots, y_{i_k}$ ait été généré par le modèle de changement dans la variance est donnée par :

$$P(I) = K_0 \sqrt{\left\{ \frac{|X'X|}{|X'X - \phi X_I' X_I|} \right\} \left\{ \frac{s^2}{\hat{s}^2_{(I)}} \right\}^{(n-p)}}$$

avec,

$$\phi = 1 - \delta^{-2}, \qquad (n-p)s^2 = S = Y'(I - H)Y$$

Après certaines considérations, *Pena et Tiao (1992)* ont montré que si δ^2 est grand, et que $\phi \approx 1$, cette probabilité devient :

$$P(I) = K_1 S_{(I)}^{-\frac{n-p}{2}} \left| X'_{(I)} X_{(I)} \right|^{-1/2}$$

qui, pour des échantillons relativement grands, est essentiellement égale à c_I (c'est-à-dire la probabilité a posteriori que les observations $y_{i_1}, \ldots, y_{i_k}$ soient singulières) trouvé dans la procédure précédente.

Une application pratique de cette procédure de détection consistera à suivre les différentes étapes qui suivent :

Étape 1 *: s'assurer que les n observations aient été tirées conformément au modèle* (4.1).

Étape 2 *: on craint que parmi les n observations, k d'entre elles aient été générées par une source contaminante associée à un changement dans la variance de Y.*

Étape 3 *: s'assurer que la taille de l'échantillon est suffisamment grande.*

Étape 4 *: calculer les* $\binom{n}{k}$ *poids* c_l.

Étape 5 *:décider que les k observations* $\left(x_{i_t 1},\ldots,x_{i_t p},y_{i_t}\right)$, $t=1,\ldots,k$ *sont singulières si leur poids* c_l *est le plus important parmi tous les* $\binom{n}{k}$.

4.3. Conclusion

Dans les modèles de régression linéaires univariés, deux principales procédures pour la détection d'observations singulières ont été présentées. La première postule un modèle sous l'hypothèse nulle que les données ont été générées conformément au modèle classique de régression, la deuxième prend en compte un modèle alternatif pour la génération de sous ensembles d'observations singulières. Les procédures de détection présentées dans l'un ou l'autre cas sont les plus couramment utilisés et /ou proposés récemment dans la littérature.

Pour vérifier la qualité d'un échantillon, le statisticien, analyste de données, devra décider de la meilleure procédure à appliquer. Dans la situation où il ne possède aucun modèle sous-jacent à la génération de données douteuses, il pourra utiliser les procédures de détections présentées au tableau 4.3.1, sinon, les procédures de détections présentées au tableau 4.3.2 seront les plus appropriées à sa problématique.

Procédures de détection	**Modèle sous H_0**	**Modèle sous H_1**	**Nombre de données douteuses**	**Jugement sur l'application du test**
Chaloner et Brant	$Y = X\beta + \varepsilon$	aucun	k	calculs longs
Pena et Guttman (1993)	$Y = X\beta + \varepsilon$	aucun	k	calculs faciles

Tableau 4.3.1 : procédures de détection dans un modèle de régression linéaire univarié basées essentiellement sur l'hypothèse nulle

Procédures de détection	**Modèle sous H_0**	**Modèle sous H_1**	**Nombre de données douteuses**	**Jugement sur l'application du test**
Guttman et al. (1978)	$Y = X\beta + \varepsilon$	$Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} 0 \\ \cdots \\ a \end{pmatrix} + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix}$	k	calculs longs
Pena et Tia (1992)	$Y = X\beta + \varepsilon$	$Y = \begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{(I)} \\ \cdots \\ \varepsilon_I \end{pmatrix}$	k	calculs longs

Tableau 4.3.2 : procédures de détection dans un modèle de régression linéaire univarié utilisant un modèle alternatif

5.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 mais elle l'est également pour des échantillons multivariés. 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 ce chapitre, nous présentons les procédures bayésiennes de détection d'observations singulières dans le cas multivarié pour des échantillons aléatoires simples et des modèles linéaires de régression.

5.1. Procédures bayésiennes de détection dans un échantillon aléatoire multivarié

Les procédures présentées ici considèrent que les observations de l'échantillon, $\{Y_1,\ldots,Y_n\}$, ont été tirées indépendamment dans une population normale multivariée de dimension p, $N(\mu,\Sigma)$, où, chaque observation, Y_i est un vecteur de dimension $p\times 1$, μ est un vecteur $(p\times 1)$ représentant la moyenne et Σ est la matrice $(p\times p)$ de variance-covariance, elle est symétrique positive définie.

a) Test de *Guttman (1973)*

Ce test est une généralisation de celui de *Guttman (1973)* présenté au chapitre 3. Plus spécifiquement, on souhaite que chaque observation Y_i, $i=1,\ldots,n$ ait été tiré dans une population normale multivariée $N(\mu,\Sigma)$. Toutefois, nous avons une certaine crainte que l'une d'entre elle puisse avoir été générée par la loi normale multivariée $N(\mu+a,\Sigma)$, où a est un vecteur de dimension p. Les hypothèses de ce test bayésien sont alors données par :

$$\begin{cases} H_0: & Y_j \in N(\mu, \Sigma) \qquad (j = 1,\ldots,n) \\ H_1: & Y_i \in N(\mu + a, \Sigma) \qquad (a > 1) \end{cases}$$

Nous sommes en présence d'une alternative de décalage («slippage»). En utilisant des a priori non informatifs, la densité a posteriori de a est de la forme :

$$p(a \mid Y_1, Y_2, \ldots, Y_n) \propto \sum_{i=1}^{n} c_i h_p\left(a \mid \eta_i ; B_p^{(i)} ; n-p-1\right),$$

où la fonction h_p est la densité généralisée d'une distribution de *Student* de dimension p, de degré de liberté $(n-p-1)$, de moyenne η_i et de constante $B_p^{(i)} > 0$ et les poids c_i sont maintenant donnés par la formule :

$$c_j = \frac{\left|A_p^{(j)}\right|^{-(n-2)/2}}{\sum_{l=1}^{n}\left|A_p^{(l)}\right|^{-(n-2)/2}}, \qquad j = 1,\ldots,n$$

avec,

$$A_p^{(i)} = \sum_{i \neq l}\left(Y_l - \overline{Y}^{(i)}\right)\left(Y_l - \overline{Y}^{(i)}\right)'$$

où,

- $\overline{Y} = \dfrac{\sum_{l=1}^{n} Y_l}{n}$; $\qquad \overline{Y}^{(j)} = \dfrac{\sum_{l \neq j} Y_l}{n-1}$

Le rapport des probabilités a posteriori est donné par :

$$\gamma_i = \frac{p\left(a>0 \mid Y_1,\ldots,Y_n\right)}{p\left(a<0 \mid Y_1,\ldots,Y_n\right)} = \frac{\sum_{j=1}^{n} c_j G_{n-p-1}\left(\sqrt{d_{ii}^{(i)}}\,\eta_i\right)}{\sum_{j=1}^{n} c_j G_{n-p-1}\left(-\sqrt{d_{ii}^{(i)}}\,\eta_i\right)}$$

où, G_{n-p-1} est la fonction de répartition d'une distribution généralisée de *Student* avec $(n-p-1)$ degrés de liberté, $d_{ii}^{(i)} = \left\{n\,A_{p_{ii}}^{(i)}/(n-1)(n-p-1)\right\}^{-1}$ et η_i est donné par la formule :

$$\eta_i = \frac{n}{n-1}\left(Y_i - \overline{Y}\right)$$

Pour une application pratique de ce test, on suivra les étapes suivantes :

Étape 1 *: vérifier que l'échantillon de données ait été généré par la loi normale multivariée* $N(\mu, \Sigma)$.

Étape 2 *: on craint qu'une donnée,* x_i*, de l'échantillon ait été générée par la loi normale multivariée* $N(\mu + a, \Sigma)$.

Étape 3 *: calculer les poids* c_i *en remplissant le tableau suivant :*

i	Y_i	$\overline{Y}^{(i)}$	$A_p^{(i)}$	$\left\|A_p^{(i)}\right\|^{-\frac{n-p-1}{2}}$	c_i
1	Y_1				
2	Y_2				
⋮	⋮				
n	Y_n				
Somme					

Étape 4 : *calculer le rapport* γ *en remplissant le tableau suivant :*

i	Y_i	$d_{ii}^{(i)}$	η_i	$c_i G_{n-p-1}\left(\sqrt{\eta_i d_{ii}^{(i)}}\right)$	$c_i G_{n-p-1}\left(\sqrt{\eta_i d_{ii}^{(i)}}\right)$
1	Y_1				
2	Y_2				
⋮	⋮				
n	Y_n				
somme					

$$\gamma_i = \frac{p\left(a>0 \mid Y_1,\ldots,Y_n\right)}{p\left(a<0 \mid Y_1,\ldots,Y_n\right)} = \frac{\sum_{j=1}^{n} c_j G_{n-p-1}\left(\sqrt{d_{ii}^{(i)}\eta_i}\right)}{\sum_{j=1}^{n} c_j G_{n-p-1}\left(-\sqrt{d_{ii}^{(i)}\eta_i}\right)}$$

Étape 5 : si $\left(\gamma_i \geq 5\right)$ *ou* $\left(\gamma_i \leq \frac{1}{5}\right)$ *et si* $c_i \notin \left[0, \frac{1}{n}+\frac{2}{n}\sqrt{\frac{n-1}{n+1}}\right]$, *conclure que l'observation* Y_i *est singulière.*

b) Procédure de détection de *Varbanov (1998)*

Varbanov(1998) utilise une approche semblable aux méthodes fréquentistes de détection d'observations singulières dans un échantillon multivarié qui sont basées essentiellement sur la statistique suivante :

$$R(Y_i,\mu,\Sigma)=(Y_i-\mu)^{'}\,\Sigma^{-1}(Y_i-\mu)$$

En posant :

$$\delta_i = R(Y_i,\mu,\Sigma)$$

on déclare que la i ème observation de l'échantillon est singulière si la probabilité a posteriori que $\delta_i > k$ est plus grande qu'une certaine valeur pour un choix approprié de k. La valeur de k peut être posée comme suit :

$$k = F_p^{-1}(0.95^{1/n})$$

où, $F_p(\ )$ est la fonction de répartition d'une chi-deux centrée avec p degrés de libertés (χ_p^2). Pour la détection d'observations singulières, le test bayésien pour chacune des n observations utilise alors les hypothèses suivantes :

$$\begin{cases} H_{0i}: \ \delta_i > k \\ H_{1i}: \ \delta_i \le k, \qquad i=1,\ldots,n \end{cases}$$

Le rapport de probabilité a posteriori associé à ce test est alors :

$$B_i = \frac{P(\delta_i > k/Y)F_p(k)}{(1-P(\delta_i > k/Y))(1-F_p(k))}$$

Kass et Raftery (1995) suggèrent d'accepter l'hypothèse nulle H_{0i} si B_i est plus grand que 10. On aura plus de confiance cependant si B_i est plus grand que 100. En utilisant des a priori non informatifs pour les paramètres Σ et μ, il vient que :

$$p_i = P(\delta_i > k/Y) = E_{\Sigma^{-1}/Y}\left\{P(W_i > nk/Y,\Sigma)\right\}$$

où, Σ^{-1}/Y suit une distribution de *Wishart*, $W(S^{-1},p,n-p)$, avec $S = \sum_{i=1}^{n}(Y_i - \overline{Y})(Y_i - \overline{Y})'$.

L'application pratique de ce test nécessitera les différentes étapes suivantes :

Étape 1 *: s'assurer que les n observations aient été générées par une loi normale multivariée.*

Étape 2 *: on craint la présence d'observations singulières dans l'échantillon.*

Étape 3 *: calculer* p_i *et* B_i

Étape 4 *: si* B_i *est plus grand que 10, conclure que la i ème observation de l'échantillon est singulière.*

5.2. Procédures bayésiennes de détection dans un modèle de régression linéaire multivarié

Dans cette section l'approche bayésienne de détection d'observations singulières dans un modèle de régression linéaire multivarié est appliquée. Les procédures de détection appropriées considèrent le modèle classique suivant :

$$Y = X\Theta + E \qquad (5.1)$$

où,

- $Y = (Y_1,\ldots,Y_n)'$ est une matrice $(n \times p)$ où les p colonnes sont représentées par les

variables réponses de dimension $p \times 1$;

- X est une matrice $(n \times q)$ connu de plein rang $q < n$ appelée matrice des variables explicatives ;

- $\Theta = (\theta_1, \ldots \theta_q)$ est un vecteur $(q \times 1)$ des paramètres ;

- E est une matrice $(n \times p)$ contenant les erreurs sur les vecteurs Y_i, distribuées indépendamment suivant une même loi normale multivariée de dimension p, $N(0, \Sigma)$.

c) Procédure de détection de *Dutter et Guttman (1979)*

Cette procédure est une généralisation de la procédure de *Guttman et al. (1978)* présentée au chapitre 4. Nous rappelons que, bien qu'il soit enviable que les n observations de l'échantillon aient été générées conformément au modèle (5.1), on craint toutefois que k d'entre elles aient été générées par une source contaminante associé à un changement de la moyenne. Les hypothèses de cette procédure sont alors les suivantes :

$$\begin{cases} H_0 : Y = X\Theta + E \\ H_1 : Y = \begin{pmatrix} y_{(l)} \\ \cdots \\ y_l \end{pmatrix} = \begin{pmatrix} X_{(l)} \\ \cdots \\ X_l \end{pmatrix} \Theta + \begin{pmatrix} 0 \\ \cdots \\ a \end{pmatrix} + \begin{pmatrix} E_{(l)} \\ \cdots \\ E_l \end{pmatrix} \end{cases}$$

On suppose par ailleurs que la matrice X ne contient qu'une variable explicative $(q = 1)$. En supposant que l'on ne possède qu'une information vague sur les paramètres de changement $a = (a_1, \ldots, a_k)$, on pose que la distribution a priori non informative de (a, Θ, Σ) est :

$$p(a, \Theta, \Sigma) \propto |\Sigma|^{-(p+1)/2}$$

La distribution a posteriori de Θ est alors donnée par

$$p(\Theta/Y,X)=\sum c_I^{(p)} t_p\left(\Theta/\hat{\Theta}_{(I)}, S_I^{(p)}/(n-k)(n-p--1);n-p-1\right)$$

où,

- t_p est une distribution généralisée de *student* de dimension p.

- $c_I^{(p)}=\dfrac{\sqrt{\left|S_I^{(p)}\right|^{2-n}}}{\sum\sqrt{\left|S_I^{(p)}\right|^{2-n}}}$, les $\binom{n}{k}$ poids qui sont les probabilités a posteriori que les observations $Y_{i_1},\ldots,Y_{i_k}$ sont singulières ;

- $\sum$ représente la somme à travers tous les $\binom{n}{k}$ ensembles possibles de I ;

- $S_I^{(p)}=\left(Y-1\hat{\mu}_I'\right)' I_I\left(Y-1\hat{\mu}_I'\right)$ avec, $\hat{\mu}_I'=\dfrac{1}{n-k}\sum_{j=k+1}^{n} Y_{i_j}$

Ainsi, les k observations $\left(X_{i_t},Y_{i_t}\right)$, $t=1,\ldots,k$ seront singulières si leur poids c_I est le plus important parmi tous les $\binom{n}{k}$ poids.

Une application pratique de cette procédure de détection, conduira aux différentes étapes suivantes :

Étape 1 *: s'assurer que les n observations aient été générées conformément au modèle spécifié à l'équation* (5.1) *avec, dans la matrice X, q = 1.*

Étape 2 *: on craint que parmi les n observations, k d'entre elles aient été générées par une source contaminante associée à un changement dans la moyenne de Y.*

Étape 3 *: calculer les* $\binom{n}{k}$ *poids* c_I.

Étape 4 *: décider que les k observations* $\left(X_{i_t},Y_{i_t}\right)$, $t=1,\ldots,k$ *sont singulières si leur poids*

c_l est le plus important parmi tous les $\binom{n}{k}$.

d) Procédure de détection de *Varbanov (1998)*

La présente procédure de détection est une généralisation de la procédure de *Chaloner et Brant (1988)*, présentée au chapitre 4, pour des modèles de régression linéaires multivariés. On utilise donc l'idée qu'une observation donnée est singulière si son résidu associé au modèle décrit par l'équation (5.1) est plus grand qu'une certaine valeur. Dans ce même modèle, nous pouvons réécrire la matrice E sous la forme suivante :

$$E = \begin{pmatrix} \varepsilon_1' \Sigma^{1/2} \\ \varepsilon_2' \Sigma^{1/2} \\ \vdots \\ \varepsilon_n' \Sigma^{1/2} \end{pmatrix}$$

où les ε_i, $i = 1,\ldots n$ sont des vecteurs $(p \times 1)$ indépendants et identiquement distribués suivant la loi normale multivariée, $N(0, I)$.

Le modèle (5.1) peut encore s'écrire :

$$Y_i = \Theta' X_i + \Sigma^{1/2} \varepsilon_i, \qquad i = 1,\ldots,n$$

Ainsi, en posant $\delta_i = \varepsilon_i' \varepsilon_i = R(Y_i, \mu, \Sigma)$, on déclare que la i ème observation de l'échantillon est singulière si la probabilité a posteriori que $\delta_i > k$ est plus grande qu'une certaine valeur pour un choix approprié de k. La valeur de k peut encore être donnée par :

$$k = F_p^{-1}\left(0.95^{1/n}\right)$$

où, $F_p(\ )$ est la fonction de répartition d'une chi-deux centrée avec p degrés de libertés $\left(\chi_p^2\right)$. Le test bayésien associé, pour chacune des n observations, à la détection des singularités considère alors les hypothèses suivantes :

$$\begin{cases} H_{0i}: \ \delta_i \geq k \\ H_{1i}: \ \delta_i \leq k, \qquad i = 1,\ldots,n \end{cases}$$

Le rapport de probabilité a posteriori associé à ce test est alors :

$$B_i = \frac{P\left(\delta_i > k/Y\right)F_p(k)}{\left(1-P\left(\delta_i > k/Y\right)\right)\left(1-F_p(k)\right)}$$

D'après *Kass et Raftery (1995)* on accepte l'hypothèse nulle H_{0i} si B_i est plus grand que 10. On aura plus de confiance cependant si B_i est plus grand que 100. En utilisant des a priori non informatifs pour les paramètres Σ et μ, il vient que :

$$p_i = P\left(\delta_i > k/Y\right) = E_{\Sigma^{-1}/Y}\left\{P\left(W_i > \frac{k}{\sigma_{(i)}}\Big/ Y, \Sigma^{-1}\right)\right\}$$

où, Σ^{-1}/Y suit une distribution de *Wishart*, $W\left(S^{-1}, p, n-q-p+1\right)$, avec

$S = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(Y_i - \overline{Y}\right)'$.

L'application pratique de ce test nécessitera les différentes étapes suivantes :

Étape 1 *: s'assurer que les n observations aient été générées conformément au modèle décrit par l'équation* (5.1).

Étape 2 *: on craint que parmi les n observations, k d'entre elles aient été générées par une source contaminante associée à un changement dans la moyenne de Y.*

Étape 3 *: calculer* p_i *et* B_i

Étape 4 *: si* B_i *est plus grand que 10, conclure que la i ème observation de l'échantillon est singulière.*

5.3. Conclusion

Dans ce chapitre, nous avons présenté une généralisation du concept de détection d'observations singulières dans un échantillon aléatoire multivarié et dans un modèle de régression linéaire multivarié. La détection d'observations singulières est importante non seulement pour des échantillons univariés mais aussi pour des échantillons multivariés. Nous n'avons pas la prétention d'avoir présenté tous les tests existant dans la littérature. Nous avons voulu montré l'importance de la détection d'observations singulières dans le cas multivarié et montrer que les principales procédures de détection sont une généralisation de celles présentées aux chapitres 3 et 4. Dans les échantillons aléatoires multivariés la grande majorité des procédures postulent que les données ont été tirées dans une population normale multivariée. L'hypothèse de normalité est donc au centre de toutes ces procédures de détection. Concernant les deux situations retenues, une procédure récente de détection due à *Varbanov (1998)* a été présentée.

Enfin, Pour vérifier la qualité d'un échantillon multivarié, le statisticien, analyste de données, utilisera les procédures de détection présentées au tableau 5.3.1

Procédure de détection	**Type**	**jugement sur l'application du test**
Guttman (1973)	Échantillon multivarié	calculs faciles
Varbanov (1998)	Échantillon multivarié	calculs relativement faciles
Dutter et Guttman (1979)	régression multivarié	calculs faciles
Varbanov (1998)	régression multivarié	calculs longs

Tableau 5.3.1 :procédures de détection dans le cas multivarié.

6. CONCLUSION GÉNÉRALE

L'aspect le plus important dans l'approche bayésienne est de permettre la formalisation des impressions subjectives comme élément de l'analyse statistique. Toute approche bayésienne sur la manipulation d'observations singulières doit être posée en ces termes : une certaine quantité X ayant une distribution a priori peut être modifiée à la lumière d'un échantillon de données observées, $x_1, x_2, \ldots, x_{n-1}, x_n$, pour produire une distribution finale (a posteriori) de X. Vu sous cet angle, l'approche bayésienne s'avère donc comme étant un outil indispensable dans la manipulation d'observations singulières. En effet, contrairement à l'approche classique de l'inférence statistique qui est fondamentalement une démarche d'inversion puisqu'elle vise à remonter des effets aux causes, c'est-à-dire des observations aux paramètres, la distribution finale ou distribution a posteriori (celle servant à l'inférence totale) dépend maintenant de toutes les données de l'échantillon. Toute observation dans l'échantillon est donc un candidat potentiel pour être soupçonnée d'être singulière. Ainsi, seule la théorie bayésienne réalise cette inversion de façon légitime et cohérente.

Dans ce travail, différents tests et procédures bayésiens ont été présentées pour détecter la présence d'observations singulières dans un échantillon. Des conclusions spécifiques ont été apportées dans chaque chapitre quant à la façon d'appliquer ces procédures de détection afin d'optimiser leur usage. De ces conclusions, nous pouvons retenir les recommandations qui suivent :

D'abord, dans l'optique bayésienne, la comparaison des tests se fait théoriquement à l'aide du risque bayésien présenté à la section 2.4. Cependant, comme la plupart des tests et procédures n'ont pas été construits autour de cette notion, d'autres critères sont alors à prendre en considération pour un meilleur classement de ces tests.

Deux dangers importants peuvent sérieusement affecter l'efficacité des tests et procédures de détection présentés dans ce rapport. Le premier est le phénomène de «*Masking*», il concerne les tests ou procédures de détection d'une observation singulière. Le second appelé effet de «*Swamping*» touche essentiellement les tests et procédures de détection de deux ou plusieurs observations singulières. Pour fixer les idées, supposons que nous voulons vérifier que les observations A et B (voir la figure 6.1 emprunté à *McCulloch et Meeter (1983)*) sont discordantes

(singulières). Une possibilité est de procéder consécutivement en testant d'abord A avec le reste de l'échantillon et en suite appliquer le même test 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 le test consécutif ou encore que l'observation B a masqué l'observation A. C'est le phénomène de «Masking». L'autre approche est d'utiliser un test ou une procédure de détection de deux ou trois observations discordantes (singulières). Si nous choisissons par exemple de tester la paire d'observations B et C avec le reste de l'échantillon, ce test déclarera qu'elle est discordante (singulière) ; en réalité, l'observation discordante B a emporté l'observation C avec elle, ce qui a eu pour résultat de faire un faux jugement, car l'observation C appartient bien au nuage central et elle n'est pas discordante. Ce danger affectant les tests et procédures de détection de deux ou plusieurs observations singulières s'appelle l'effet de «Swamping».

Se prémunir de ces deux dangers constituent donc, à notre avis, un critère valable pour classer les différents tests et procédures de détection présentés suivant leur ordre d'efficacité. En effet, le phénomène de «Masking» peut contribuer à diminuer la puissance d'un test ou d'une procédure de détection d'une observation singulière, alors que l'effet de «Swamping» peut concourir à identifier plusieurs observations singulières dans l'échantillon qu'il n'y en a en réalité. Ainsi, une méthode qui évite ces deux dangers est donc souhaitable.

À la lumière de toutes ces considérations, une bonne démarche à suivre pour vérifier la qualité d'un échantillon est la suivante :

a) en accord avec les données à étudier, trouver le modèle statistique s'ajustant aux données de l'échantillon ;
b) relever le nombre de données douteuses de l'échantillon ;
c) à partir des tableaux 6.1 (pour un échantillon univarié), 6.2 (pour un modèle de régression linéaire univarié) et 6.3 (pour le cas multivarié), choisir le test de détection s'approchant le mieux à la problématique de vérification de la qualité des données de l'échantillon.

D'autre part, comme les procédures et tests de détection que nous avons présentés utilisaient principalement des distributions a priori non informatives, elles ne valent en fait que ce que vaut

l'information a priori. Si celle-ci est bonne, l'application est possible, et conduit effectivement à des décisions plus économiques. Si elle est mauvaise, ou vague, la prudence est de règle. Cela semble donc présenter une limite dans l'approche bayésienne de la détection d'observations singulières. Une extension intéressante pour ces procédures et tests serait de choisir la loi a priori caractérisant véritablement les paramètres qui entrent en jeu dans leur édification.

Comme nous l'avons déjà mentionné au début de ce rapport, il est très difficile de définir ce qu'est une observation singulière. Dans le cas multivarié par exemple, une donnée peut être identifiée comme une singularité par une méthode et ne pas l'être par une autre. Dès lors, nous pensons qu'en toute circonstance le statisticien, analyste de données, qui trouve une ou des données douteuses doit discuter avec l'expérimentateur et qu'ensemble ils répondent à la question : est-on en présence de réelles aberrances et si oui qu'en fait-on ? à ce sujet, nous recommandons au lecteur la discussion intéressante que font *McCulloch et Meeter (1983)*. Si toutefois aucune décision ne peut être prise, il est certainement préférable de faire l'inférence statistique en utilisant les méthodes d'accommodation.

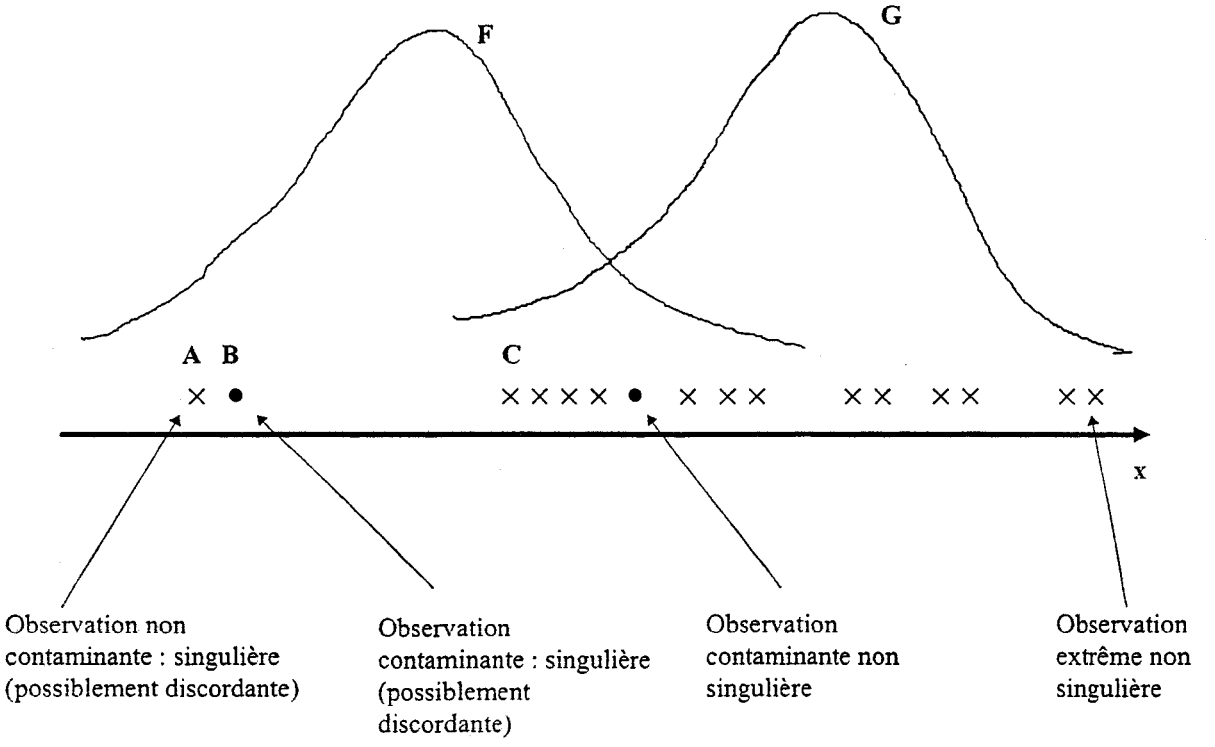


Figure 6.1 : graphique illustratif du phénomène de «Masking » et de l'effet de «Swamping»

Test	Modèle sous H_0	Modèle sous H_1
Pettit et Smith (1983, 1985)	$N(\mu,\sigma^2)$	$N(\mu+\delta_i,\sigma^2)$
Pettit (1988)	$E(\theta)$	$E(\theta\delta)$
Pettit (1988)	$E(\theta)$	$E(\theta\delta_i)$
Pettit (1994)	$P(\theta)$	$P(\theta\delta)$
Pettit (1994)	$P(\theta)$	$P(\theta\delta_i)$

Tableau 6.1 : résumé des tests de détection dans un échantillon univarié qui ne sont pas affectés par le phénomène de «Masking» et de l'effet de «Swamping»

Procédures de détection	Nombre de données douteuses	Jugement sur l'application du test
Guttman et al. (1978)	k	calculs longs
Pena et Tia (1992)	k	calculs longs

Tableau 6.2 : procédures de détection dans un modèle de régression linéaire univarié qui sont protégées contre le phénomène de «Masking» et de l'effet de «Swamping»

Procédure de détection	Type
Varbanov (1998)	Échantillon multivarié
Varbanov (1998)	régression multivarié

Tableau 6.3 : procédures de détections dans le cas multivariés qui sont immunisées contre le phénomène de «*Masking*» et de l'effet de «*Swamping*».

7. REVUE BIBLIOGRAPHIQUE

Akaike, H. (1973). Information Theory and an Extention of the Maximum Likelihood Principle. *In 2nd International Symposium on Information Theory* (B. N. Petrov and F. Csaki, Eds.). Budapest, Akademial Kiado, p267-281.

Akaike, H. (1980). On the Use of Predictive Likelihood of a Gaussian Model. *Ann. Inst. Statist. Math.*, 32, pp. 311-324.

Barnet V. et Lewis T. (1994). Outliers in Statistical Data, Wiley, 584 pages.

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. New York : Springer Verlag, 617 pages.

Bernardo, J. M. et A.F.M. Smith (1994). Bayesian Theory, Chichester, UK : John Wiley and sons.

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.

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*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Valencia, Spain : Valencia University Press, pp. 349-365.

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*, 11, pp. 1-21

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

Guttman, Irwin, and C. G. Khatri (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., 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

Johnson, N. L. and Welch, B. L. (1940). Applications of the Noncentral *t* Distribution. *Biometrika* 31, pp 362-89.

Kass, R., Raftery, A. (1995). Bayes Factors. *JASA*, 90, pp.773-95

Leamer, E.E. (1978). Specification Searches. John Wiley and Sons, 370 pages.

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

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

Pena, 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. 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.

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, *Commun. Statist. Theory Meth.*, 23(6), pp. 1785-1795.

Pettit, L. I. and Smith A. F. M. (1983). Bayesian modele comparaison in presence of outliers. *Bull. Int. Statist. Inst.*, 50, pp. 292-309.

Poirier, D.J. (1988). Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics, *journal of Economic Perspectives 2* : 121-170.
Robert Christian (1992). L'Analyse Statistique Bayésienne. Paris : Economica.

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 Alexandre (1998). Bayesian Approach to Outliers Detection in Multivariate Normal Samples and Linear Models. *Commun. Statist.-Theory Meth.*, 27(3), pp.—547-557.

Zellner, Arnold (1971). An Introduction to Bayesian Inference in Econometrics, John Wiley and Sons, 431 pages.

APPENDICE A

Exemple d'application dans le cas d'un échantillon aléatoire univarié

(Test de détection de *Guttman (1973)*)

8.1 Description de l'exemple

Nous avons un échantillon de cinq observations tel que les quatre premières aient été générées par la loi normale $N(13,1)$ et la dernière par la loi normale $N(17,1)$. Les données obtenues sont montrées dans le tableau A1.

i	y_i
1	12.9624
2	13.3273
3	14.1892
4	14.1909
5	17.1746

Tableau A1 : échantillon de données appliqué à l'exemple d'un échantillon aléatoire univarié

8.2 Programme utilisé dans Matlab

```
function y =exoech(serie)
% La fonction EXOECH est un exemple d'application
% du test de détection de Guttman (1973) dans un
% échantillon univarié. SERIE désigne la série de
% données.


% Déclaration des constantes

n = length(serie);


% Échantillon de données

xy=serie;
x=sort(xy);


% Calcul de xbar(j)
```

```
x3=x;

for i=1:n
    x(i,:)=[];
    xbar(i)=mean(x);
    x=x3;
end
xbar


% Calcul de A(j)

x4=x3;
x=x3;
xbar1=xbar';
for i=1:n
   x(i,:)=[];
   x4=x;
   x4(:)=xbar1(i);
   xx=(x-x4).*(x-x4);
   a(i)=sum(xx);
   x=x3;
end
a


% Calcul des poids ci

for i=1:n
   a1=(a(i))^(-(n-2)/2);
    for j=1:n
     a2(j)=(a(j))^(-(n-2)/2);
    end
   ci(i)=a1/sum(a2);
end

ci
t1=(1/n)+(2/n)*sqrt((n-1)/(n+1));
(['Intervalle de non singularité
=','[',int2str(0),',',num2str(t1),']' ])


% Calcul du rapport de probabilité

x=x3;
for i=1:n
   moy(i)=(n/(n-1))*(x(i)-mean(x));
   b(i)=(n*a(i)/((n-1)*(n-2)))^(-1);
   num(i)=ci(i)*tcdf(sqrt(b(i))*moy(i),n-2);
```

```
   den(i)=ci(i)*tcdf(-sqrt(b(i))*moy(i),n-2);
end
proba=sum(num)/sum(den)
% Calcul de la moyenne et de la variance

moyenne=sum(ci.*moy)
variance=sum(ci.*(moy.*moy))+(n/(n-1)*(n-4))*sum(ci.*a)-
moyenne.*moyenne


% Graphique de la densité a posteriori des observations

l1=linspace(-variance,variance,50);
bb=l1;
moy1=l1;
for j=1:n
  for i=1:50
   bb(:)=b(j);
   moy1(:)=moy(j);
   yy(j,i)=ci(j)*tpdf(sqrt(bb(i))*(l1(i)-moy1(i)),n-2);
  end
end

plot(l1,sum(yy(1:5,:)),'-.');
hold on

for j=1:n
   plot(l1,5*yy(j,:))
   xx44=[l1;5*yy(j,:)];
   [i11,j11]=find(xx44'==max(5*yy(j,:)));
   text(l1(i11),5*yy(j,i11),(['c',int2str(j)]))
end

xlabel('valeurs de a')
ylabel('ci*h(a/ (),(),n-2)')
hold off
```

8.3 Résultats obtenus

En suivant la procédure du test de détection de *Guttman (1973)*, l'observation y_5 est celle qui apparaît surprenante dans l'échantillon de données présenté au tableau A1. Les poids de chaque observation sont donnés dans le tableau A2 :

i	y_i	$\bar{y}^{(i)}$	$A^{(i)}$	c_i
1	12.9624	14.7205	8.5266	0.0432
2	13.3273	14.6293	9.6433	0.0359
3	14.1892	14.4138	10.9591	0.0296
4	14.1909	14.4134	10.9599	0.0296
5	17.1746	13.6674	1.1591	0.8616

Tableau A2 : calcul des poids (Exemple d'application dans le cas univarié)

À la lumière de ce tableau, nous observons que les observations y_1, y_2, y_3 et y_4 ont des poids relativement faibles qui appartiennent à l'intervalle [0,0.5266]. Par contre, l'observation y_5 a un poids élevé qui n'appartient pas à cet intervalle. Selon toute apparence, l'observation y_5 semble influencer considérablement la densité a posteriori. Ce fait tend à se confirmer, car en observant la figure A1, nous remarquons que la densité a posteriori de a est dominée par l'observation y_5. Par ailleurs, le rapport de probabilité est donné par $\gamma = 9.1727$. Ainsi, comme ce rapport est plus grand que 5 et que $c_5 \notin [0, 0.5266]$, nous concluons que l'observation y_5 est singulière. Le test de détection de *Guttman (1973)* a donc été concluant pour les données choisies dans cet exemple.

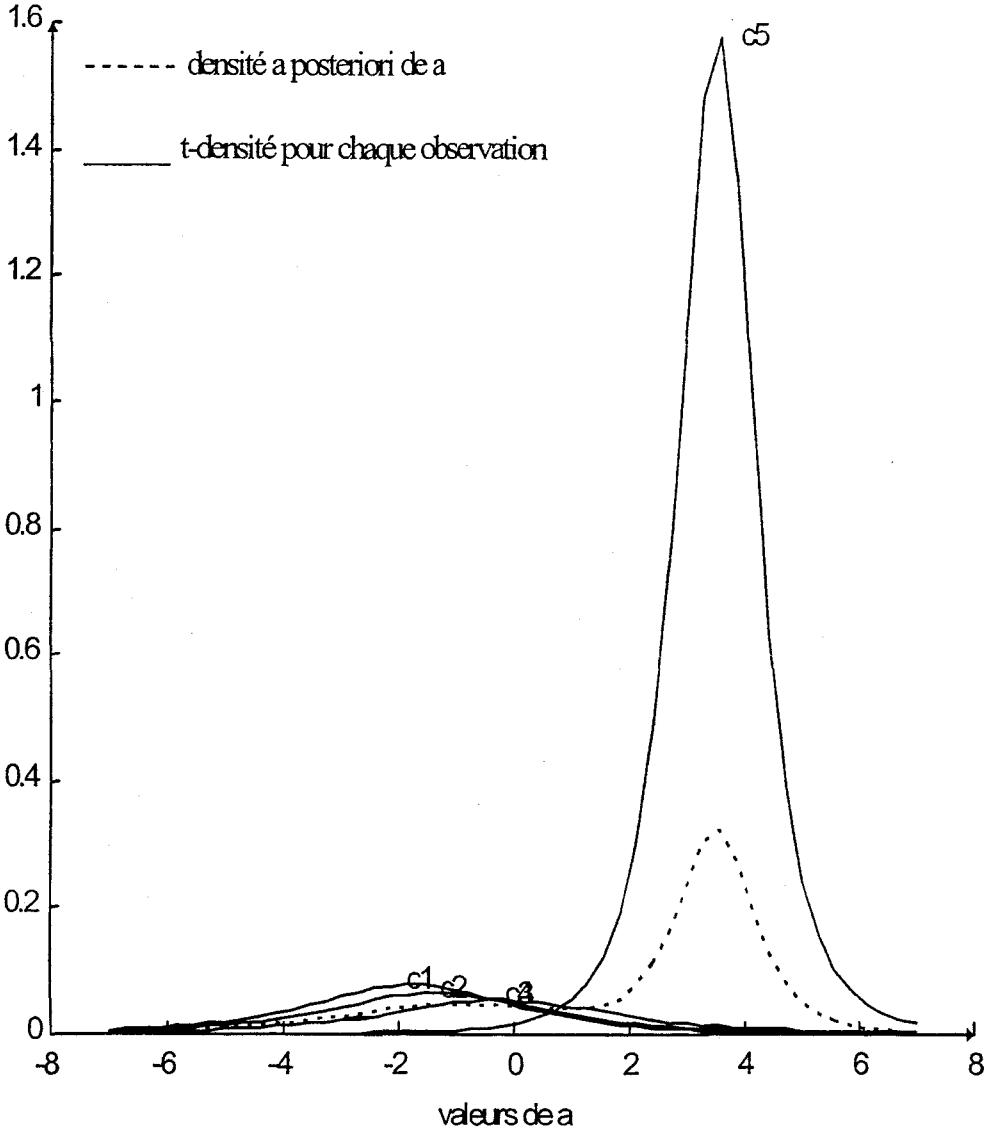


Figure A1 : densités de l'exemple d'application dans le cas univarié

APPENDICE B

Exemple d'application dans le cas d'un modèle de régression linéaire univarié

(Procédure de détection de *Guttman et al. (1978)*)

9.1 Description de l'exemple

Dans cet exemple, nous avons un échantillon de 10 observations tel que les 8 premières ont été générées par la loi normale $N(0,1)$ et les deux dernières par la loi normale $N(2,0)$. Le paramètre de décalage à la moyenne est donc $a = (2,2)$. Les données obtenues sont consignées dans le tableau B1.

i	y_i
1	-0.5883
2	-0.1867
3	-0.1364
4	0.0593
5	0.1139
6	0.7258
7	1.0668
8	1.1677
9	1.9044
10	2.1832

Tableau B1 : échantillon de données appliqué à l'exemple d'un modèle de régression linéaire univarié

Nous avons limité le modèle de régression en prenant $X = \mathbf{1}$, c'est-à-dire $p = 1$.

9.2 Programme utilisé dans Matlab

```
function y =exoreg(y,x,k)
% La fonction EXOREG est un exemple d'application
% du test de détection de Guttman et al. (1978) dans un
% modèle de régression linéaire univarié. Y est la
% variable réponse, X est la variable explicative et
```

```
% k (≤3) désigne les k observations soupçonnées d'être singulières

% Déclaration des constantes

p=1;
n = length(y);
k=k;

% Calcul du paramètre béta avec les k dernières
% observations soupçonnées.

y_i=y(n-k+1:n);
x_i=x(n-k+1:n);
beta_i=inv(x_i'*x_i)*x_i'*y_i;


% Calcul de béta et SI après omission des k dernières
% observations soupçonnées.

yi=y(1:n-k);
xi=x(1:n-k);
betai=inv(xi'*xi)*xi'*yi;
si=(yi-xi*betai)'*(yi-xi*betai);

% Calcul des poids CI

yy=y;
xx=x;
w=1;
e=1;

for f=1:k
   c(k)=0;
end

 for i=1:n
   if i<=(n-k+1)
      nombr=(gamma(n-i+1))/((gamma(k)*gamma(n-i-k+2)));
   else
      nombr=0;
   end

   e4=i-1;
   e1=1;

   for j=1:n-1
      e=1;
      while ((e+j+k-2+e4)<=n) & (e1<=nombr)
         x([i:j:(j+k-2+e4) (j+k-2+e4+e)])=[];
         y([i:j:(j+k-2+e4) (j+k-2+e4+e)])=[];
```

```
        c=[i:j:(j+k-2+e4) (j+k-2+e4+e)];
        (['c(',int2str(w),')=']), c
        const(w)=x'*x;
        beta1i =inv(x'*x)*x'*y;
        s1i =(y-x*beta1i)'*(y-x*beta1i);
        poids1(w) =(s1i^(-(n-p-1)/2))/(sqrt(abs(x'*x)));
        beta11(w)=beta1i;
        s11(w)=s1i;
        w=w+1;
        y=yy;
        x=xx;
        e=e+1;
        e1=e1+1;
    end
end

    e4=e4+1;

end

    for l=1:w-1
       ci(l)=(poids1(l))/(sum(poids1));
    end

    ([num2str(ci)]); %Le vecteur de poids



% GRAPHIQUES

    % Calcul de la moyenne a posteriori du paramètre beta

       for i=1:w-1
          moy(i)=beta11(i)*ci(i);
       end

        moyenne=sum(moy)


    % Calcul de la variance a posteriori du paramètre beta

      beta111=beta11';

      for j=1:w-1
       var1(j)=ci(j)*(((s11(j)/(n-k-p-2))*
        ((const(j))^(-1)))+((beta11(j)*beta111(j))));
      end

      variance=sum(var1)-moyenne.*moyenne'
```

```
%Graphique de la densité a posteriori des observations

   l1=linspace(-1,1,50);
   s11i=l1;
   beta11i=l1;
   moyy=l1;



   for j=1:w-1
    for i=1:50
     s11i(:)=s11(j);
     beta11i(:)=beta11(j);
     coef_b = ((n-k-p)/(s11(j)))*const(j);
     coef_m = n-k-p;
     yy(j,i)=ci(j)*tpdf((sqrt(coef_b)*(l1(i)-
     beta11i(i))),coef_m);
    end
   end

   plot(l1,sum(yy(1:w-1,:)),'-.');
   hold on
   legend('Graphique a posteriori de beta',2)

  for j=1:w-1
   plot(l1,5*yy(j,:))
   xx44=[l1;5*yy(j,:)];
   [i11,j11]=find(xx44'==max(5*yy(j,:)));
   text(l1(i11),5*yy(j,i11),(['c',int2str(j)]))
  end

   title('LES DENSITES DES OBSERVATIONS')
   xlabel('valeurs de beta')
   ylabel('ci*h(beta/beta_hat,(),n-k-p)')
   hold off
```

9.3 Résultats obtenus

En suivant la procédure de détection de Guttman et al (1978), les observations y_9 et y_{10} sont celles qui apparaissent surprenantes dans l'échantillon de données présenté au tableau B1. Les poids de chaque observation sont donnés dans le tableau suivant :

	i_1	i_2	c_i
c1	1	2	0.024666
c2	1	3	0.022763
c3	1	4	0.017560
c4	1	5	0.016557
c5	1	6	0.012020
c6	1	7	0.012825
c7	1	8	0.013516
c8	1	9	0.035018
c9	1	10	0.073360
c10	2	3	0.011404
c11	2	4	0.009265
c12	2	5	0.008841
c13	2	6	0.006948
c14	2	7	0.007497
c15	2	8	0.007895
c16	2	9	0.018885
c17	2	10	0.036128
c18	3	4	0.008742
c19	3	5	0.008350
c20	3	6	0.006613
c21	3	7	0.007146
c22	3	8	0.007526
c23	3	9	0.017913
c24	3	10	0.034053
c25	4	5	0.006946
c26	4	6	0.005646
c27	4	7	0.006138
c28	4	8	0.006470
c29	4	9	0.015229
c30	4	10	0.028471
c31	5	6	0.005453
c32	5	7	0.005938
c33	5	8	0.006262
c34	5	9	0.014724
c35	5	10	0.027453
c36	6	7	0.005271
c37	6	8	0.005594
c38	6	9	0.013804
c39	6	10	0.026326
c40	7	8	0.006404
c41	7	9	0.017098
c42	7	10	0.034575
c43	8	9	0.018927
c44	8	10	0.039280
c45	9	10	0.278490

Tableau B2 : Calcul des poids (Exemple d'application dans le modèle de régression linéaire univarié)

Le tableau B2, nous montre que la paire d'observations $\{y_9, y_{10}\}$ réalise le plus grand poids parmi les 45 poids possibles. De plus, en observant la figure B.1, nous remarquons que la densité a posteriori de θ est dominée par la paire d'observations $\{y_9, y_{10}\}$. Nous concluons donc que les observations y_9 *et* y_{10} sont singulières.

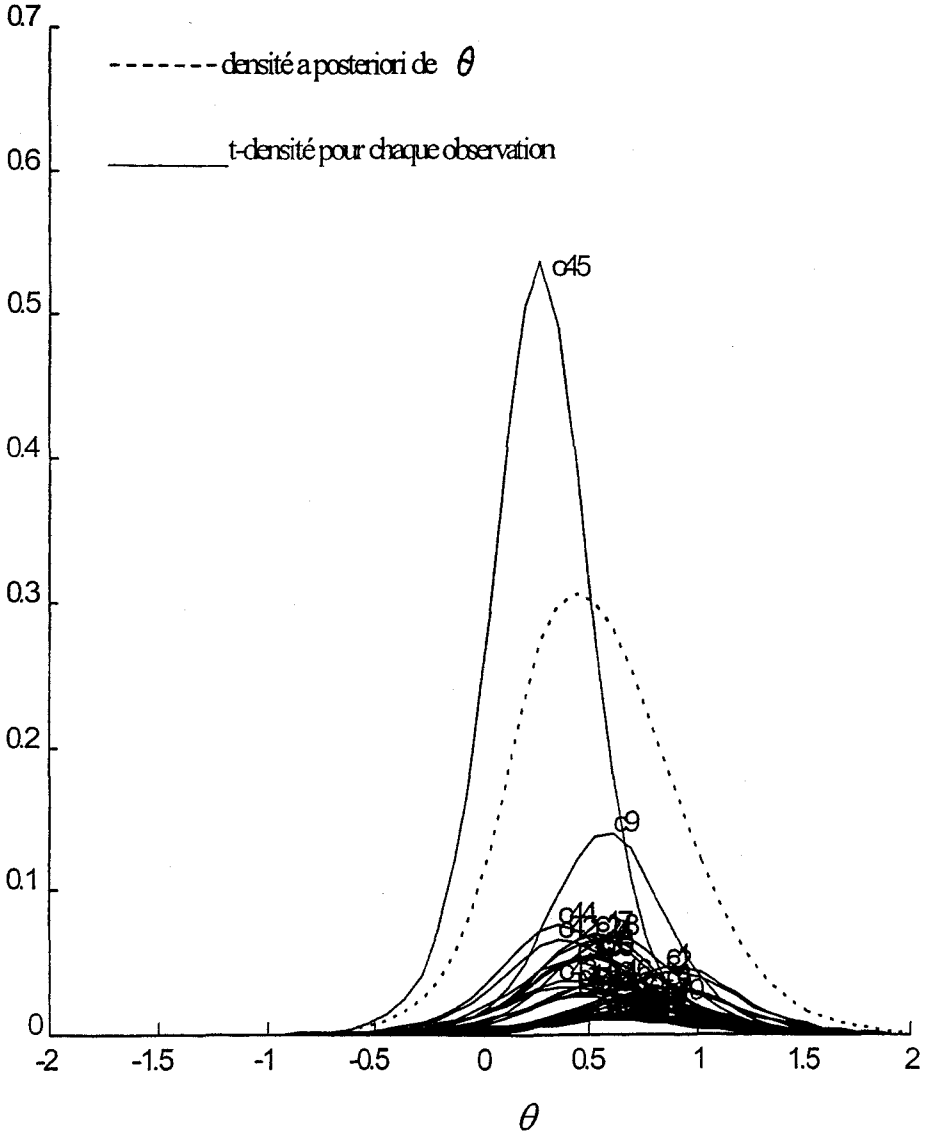


Figure B1 : densités de l'exemple d'application dans le cas du modèle de régression linéaire univarié