UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR
SAID LAHLOU MIMI

PROFONDEUR DE TUKEY ET SON APPLICATION EN CONTRÔLE DE QUALITÉ MULTIVARIÉ

DÉCEMBRE 2000

À MES PARENTS
À TOUTE MA FAMILLE

Résumé

Le contrôle statistique de la qualité est un outil pour le suivi et la mise à jour de la qualité des produits. Les techniques univariées classiques ont certaines limitations dans leur applicabilité. Parmi ces limitations, est qu'elles ne peuvent typiquement surveiller qu'une seule caractéristique de qualité. Cependant, c'est totalement insatisfaisant pour la plupart des procédés de fabrication modernes. Le contrôle statistique multivarié de la qualité s'adresse à ce genre de situations. Il considère toutes les données simultanément et extrait l'information sur la directionnalité des variations du procédé, *i.e.* le comportement d'une observation relativement aux autres.

Le but de ce travail est d'employer la profondeur de Tukey pour introduire des nouvelles cartes de contrôle de qualité pour les observations multivariées. La philosophie derrière notre approche est de réduire la dimensionnalité du problème en formant un nouvel ensemble de variables latentes. Et ce, pour obtenir une meilleure compréhension du comportement du procédé de fabrication. Ainsi, pour n'importe quelle dimension des observations, ces cartes se présentent sous forme de graphiques bidimensionnels. Elles peuvent cependant être visualisées et interprétées facilement comme c'est le cas pour les cartes univariées $X, \bar{X}, CUSUM$. En plus, elles ont plusieurs avantages significatifs. D'abord, elles peuvent détecter simultanément le décalage d'emplacement et mesurer la croissance d'échelle de fabrication, à la différence des méthodes existantes qui ne peuvent détecter que le décalage d'emplacement. En second lieu, leur construction est complètement non-paramétrique ; en particulier, elle n'exige pas l'hypothèse de normalité pour la distribution de qualité, qui est nécessaire dans des approches standards telles que les cartes χ^2 et Hotelling T^2. Ainsi, ces nouvelles cartes généralisent le principe des cartes de contrôle au cas multivarié et s'appliquent à une classe beaucoup plus large des distributions de qualité.

Abstract

Statistical quality control is a tool for achieving and maintaining product quality. Classical univariate statistical techniques have certain limitations in their applicability. One such limitation is that they typically monitor only one quality characteristic. However this is totally inadequate for most modern process industries. Multivariate statistical process control addresses several limitations of univariate monitoring. It considers all data simultaneously and extracts information on the directionality of the process variations, *i.e.* the behaviour of one observation relative to the others.

The purpose of this work is to use Tukey depth to introduce several new control charts for monitoring processes of multivariate quality measurements. The philosophy behind our approach is to reduce the dimensionality of the problem by forming a new set of latent variables to obtain an enhanced understanding of the process behaviour. Thus for any dimension of the measurements, these charts are in the form of two-dimensional graphs that can be visualized and interpreted just as easily as the well-known univariate $X, \bar{X}$, and $CUSUM$ charts. Moreover, they have several significant advantages. First, they can detect simultaneously the location shift and scale increase of the process, unlike the existing methods, which can detect only the location shift. Second, their construction is completely nonparametric; in particular, it does not require the assumption of normality for the quality distribution, which is needed in standard approaches such as the χ^2 and Hotelling's T^2 charts. Thus these new charts generalize the principle of control charts to multivariate settings and apply to a much broader class of quality distributions.

Remerciements

Je désire avant tout remercier sincèrement mon directeur de recherche, le professeur Belkacem Abdous, d'avoir accepté de diriger ce travail, de m'avoir soutenu et de m'avoir orienté dans mes moments les plus critiques. Je lui suis également très reconnaissant pour sa disponibilité et pour m'avoir laissé tant de liberté. L'aboutissement de ce travail doit beaucoup à ses conseils comme à ses critiques.

Mes plus vifs remerciements vont également au professeur Kilani Ghoudi, mon co-directeur, dont la grande expertise en contrôle statistique de qualité m'a été d'un grand secours. Je tiens aussi à le remercier pour la confiance qu'il m'a accordée dès le début de ce travail.

Je désire également remercier les membres du jury d'avoir accepté d'évaluer ce travail et d'y avoir consacré autant de temps et d'attention.

Je remercie tous mes collègues et amis particulièrement, Julie Pelletier, Adil Lahyane, Arturo Figueroa, ainsi que tous les autres qui ont participé à un moment ou un autre, à enrichir l'environnement stimulant du laboratoire de la maîtrise de mathématiques et informatique appliquées.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Avant-propos

Au cours de ces dernières années, l'industrie a été confrontée à une concurrence de plus en plus féroce. L'internationalisation de la compétition, et la course au développement, ont poussé les entreprises à rechercher des atouts leur permettant de gagner la partie. La recherche de la qualité est alors devenue un point clé de la compétitivité des entreprises. Cette recherche ne date pas d'aujourd'hui. En effet, la qualité a toujours été un objectif important depuis que l'homme fabrique des objets. Cependant, le nouveau contexte de concurrence mondiale ranime cette quête de la qualité en demandant plus de formalisme dans son approche.

Ainsi la qualité a toujours été un critère qualifiant. Et pour mieux l'apprécier, il faut savoir le mesurer tout en incluant notre capacité de traitement des données disponibles. Pour cela, il faut faire appel à des outils et des méthodes qui permettent la réalisation d'une telle tâche.

Le contrôle statistique de la qualité, ou plus spécialement le contrôle statistique des procédés, se révèle être un moyen privilégié qui permet cette démarche indispensable à l'obtention d'une meilleure qualité. Développé aux États-Unis dans les années 20, le contrôle statistique des procédé n'a pas connu l'emploi généralisé qu'en ont fait les Japonais et qu'en font aujourd'hui un nombre de plus en plus croissant d'entreprises occidentales. Grâce aux techniques liées à ce dernier, les problèmes peuvent être évalués objectivement en se basant sur des faits plutôt que sur de l'intuition et des opinions souvent subjectives. À l'aide de ces techniques, nous pouvons rechercher les causes de ces problèmes et leur trouver des solutions adéquates.

Le contrôle statistique des procédés consiste à appliquer des techniques statistiques pour suivre et corriger un procédé avant même que la qualité ne se détériore. Pour cette fin, il utilise les cartes de contrôle qui à leur tour se basent sur le fait que tout procédé produit des variations dont les causes peuvent être identifiables ou non identifiables. L'objectif d'une carte de contrôle est de donner une image du déroulement du procédé de fabrication. Cette image doit permettre de discerner si, à un moment donné, il y a présence d'une cause spéciale ou si les variations observées ne sont dues qu'à des causes communes. Ainsi, les causes identifiables peuvent être éliminées ou réduites, si cela est nécessaire. De plus, nous pouvons savoir à l'avance si un procédé sera capable de respecter des spécifications données en évaluant sa capacité opérationnelle. Si un procédé n'a pas la capacité de respecter certaines spécifications, nous pouvons calculer la proportion probable d'unités défectueuses qu'il produira si ces spécifications étaient effectivement exigées.

La première carte de contrôle a été introduite par Walter A. Shewhart, de la société de téléphone bell, aux États-Unis en 1924. Ce dernier a ouvert la voie devant de nombreuses équipes de recherche qui ont travaillé sur le sujet. Ce qui a donné naissance à d'autres cartes de contrôle dont les plus utilisées sont : la carte $\mathbf{X}$, la carte $\bar{\mathbf{X}}$ (connue aussi sous le nom de carte Shewhart) et la carte **CUSUM**. Ces dernières sont utilisables seulement pour contrôler des observations univariées, et leur validité repose sur l'hypothèse de normalité des observations, ce qui n'est pas toujours le cas.

En réalité, la qualité des produits est souvent déterminée par plusieurs caractéristiques. Par exemple, la qualité de certains types de plaques métalliques peut être déterminée par le poids, le degré de solidité, l'épaisseur, la largeur et la longueur. Ces caractéristiques sont sans doute corrélées, et les cartes de contrôle univariées ne peuvent être adéquates pour détecter les variations du procédé de fabrication. Cependant, il est désirable d'avoir des cartes qui peuvent contrôler les mesures multivariées directement. Il existe dans la littérature quelques méthodes pour la construction de telles cartes de contrôle. Mais ces méthodes sont généralement restreintes au cas de la normalité des observations d'une part et d'autre part elles sont difficiles à visualiser et à interpréter.

Ce travail de recherche consiste à développer des cartes de contrôle multivariées en se basant sur la notion de profondeur de Tukey des données multivariées. L'idée derrière les cartes que nous proposons est de réduire un problème multidimensionnel en un problème univarié en utilisant la profondeur de Tukey des observations. Par la suite nous développons des cartes de

contrôle, basées sur ces mesures de profondeur, suivant les mêmes principes que les cartes **X**, $\bar{\mathbf{X}}$ et **CUSUM**. Les résultats de simulation que nous présentons au chapitre 4 montrent que nos cartes performent bien sans aucune hypothèse sur la distribution des données du procédé.

1.2 Définitions

Avant d'entamer les détails statistiques, il est nécessaire de spécifier quelques définitions.

1.2.1 Procédé "sous-contrôle"

Un procédé de fabrication est dit "sous-contrôle"(ou maîtrisé) lorsqu'aucune cause précise ne vient modifier les caractéristiques du produit. Ces caractéristiques varient légèrement dans le temps, mais ces variations restent à l'intérieur d'une fourchette de variation naturelle. Elles sont dues à de multiples causes aléatoires, dites "causes communes", de faible importance et que nous ne pouvons pas supprimer.

1.2.2 Procédé "hors-contrôle"

Un procédé de fabrication est dit "hors-contrôle" (ou non maîtrisé) quand une cause particulière et importante, dite "cause spéciale", vient modifier son fonctionnement. Par exemple l'usure d'une pièce, la modification d'une matière première ou d'un additif, une modification des conditions de production (humidité, température...), un déréglage dû à des vibrations, etc.

1.2.3 Les 5 "M" d'un procédé

Tous les procédés, quels qu'ils soient, sont incapables de produire le même produit sans aucune variation. Quelle que soit la machine étudiée et la caractéristique observée, nous notons toujours une dispersion dans la répartition de la caractéristique. Ces variations proviennent de l'ensemble du procédé de fabrication. L'analyse des procédés de fabrications permet de dissocier 5 éléments élémentaires qui contribuent à créer cette dispersion. Nous désignons généralement par les 5 M ces 5 causes fondamentales responsables de dispersion, et donc de non-qualité.

- **M**achine
- **M**ain d'oeuvre

- Matière
- Méthodes
- Milieu

Le contrôle statistique des procédés a pour objectif la maîtrise des procédés de fabrication en partant de l'analyse de ces 5 M responsables de la non-qualité. Il n'apporte pas une grande révolution dans la façon de faire, mais une plus grande rigueur et des outils méthodologiques qui vont aider les opérateurs dans leur tâche d'amélioration de la qualité.

1.3 Plan et contenu

Dans ce travail nous abordons directement les techniques multivariées pour la construction des cartes de contrôle. Au besoin, nous donnons des rappels au fur et à mesure sur les techniques univariées.

Chapitre 2 : Les cartes de contrôle multivariées

Ce chapitre présente une revue de littérature portant sur les méthodes les plus répandues pour la construction et l'interprétation des cartes de contrôle multivariées. Le lecteur remarquera que tous les tests statistiques associés à ces méthodes reposent sur l'hypothèse de normalité des observations. Ce qui n'est pas toujours le cas en pratique. Ce point situe d'emblée l'orientation de notre recherche.

Chapitre 3 : La profondeur de Tukey

Ce chapitre est consacré à la profondeur de Tukey. Nous y rassemblons les propriétés théoriques nécessaires pour la réalisation de notre projet. Ensuite, nous traitons l'aspect algorithmique de la profondeur de Tukey.

Chapitre 4 : Contribution de la profondeur de Tukey au contrôle de qualité multivarié

Ce chapitre est consacré à notre application. Elle consiste à utiliser la profondeur de Tukey pour la construction des cartes de contrôle multivariées. Nous décrivons dans un premier temps les statistiques derrières l'approche que nous proposons. Ensuite, nous exposons les résultats obtenus par l'application de notre approche ainsi que les résultats de comparaison avec une carte multivariée conventionnelle.

Chapitre 5 : Conclusion et perspectives

Dans ce dernier chapitre, nous résumons les principaux résultats présentés dans ce mémoire ainsi que les extensions et les suites possibles de ce travail.

Chapitre 2

Les cartes de contrôle multivariées

2.1 Introduction

En général, les cartes de contrôle permettent de surveiller un procédé de fabrication en s'assurant que les pièces fabriquées demeurent stables ou conformes aux spécifications (compte tenu d'une certaine variabilité inévitable). Elles consistent en un tracé d'une mesure statistique d'une caractéristique quantitative ou qualitative en fonction du temps. Cette mesure est généralement évaluée à partir d'un ensemble d'échantillons ou de sous-groupes. Le tracé d'une carte de contrôle s'effectue en indiquant en ordonnée la mesure statistique que l'on veut maîtriser et en abscisse, le numéro d'ordre chronologique (ou l'instant de prélèvement) de l'échantillon à partir duquel a été calculée cette statistique.

Dans le cas où il s'agit de contrôler une seule caractéristique de qualité (cartes de contrôle univariées), le problème du choix de la mesure statistique à tracer sur la carte est bien résolu. On trouve dans la littérature une grande variété de ce type de cartes. Elles sont d'ailleurs intégrées dans les manufactures qui adoptent les techniques de contrôle statistique de la qualité. En pratique, la qualité d'un produit est souvent déterminée par plusieurs caractéristiques. Prenons, par exemple, le cas des films plastiques. L'utilité de ces derniers dépend de leur transparence (X_1) ainsi que de leur résistance (X_2). Ces deux caractéristiques sont sans doute corrélées. Les cartes de contrôle univariées ne pourront pas détecter adéquatement les variations du procédé de fabrication. Il est cependant désirable de disposer de cartes de contrôle

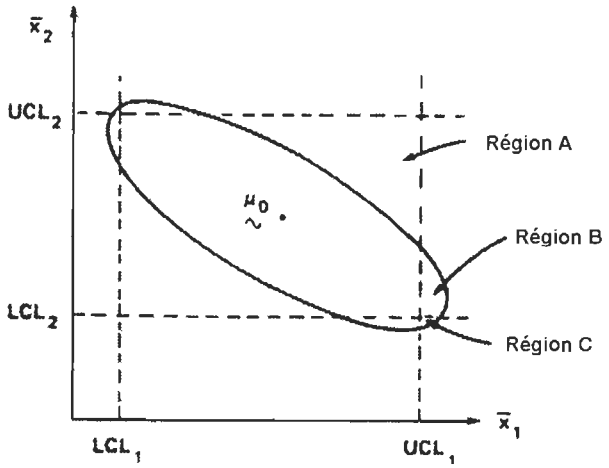

FIG. 2.1 – Superposition des régions rectangulaire et elliptique de contrôle

qui peuvent répondre à ce type de situations.

Une approche possible, est d'ignorer la corrélation entre les deux caractéristiques (X_1) et (X_2) et de les contrôler individuellement. Pour chaque échantillon de taille déterminée, on obtient une estimation de μ_{01} (resp. de μ_{02}), qu'on note par $\bar{x}_1$ (resp. $\bar{x}_2$). On la dessine en fonction du temps sur une carte $\bar{x}$ par exemple. Si les deux moyennes restent dans leurs régions de contrôle respectives, alors le procédé est déclaré en état de contrôle statistique.

On peut remarquer facilement que cette approche est équivalente à tracer la paire $(\bar{x}_1, \bar{x}_2)$ sur une seule carte en superposant les deux cartes $\bar{x}$ précédentes, tel qu'illustré par la figure 2.1. Si la paire des moyennes de l'échantillon se trouve dans le rectangle de contrôle, le procédé est déclaré en état de contrôle statistique.

L'utilisation séparée des cartes de contrôle univariées peut être très trompeuse. En effet, on peut montrer que dans le cas de deux caractéristiques de qualité corrélées que la vraie région de contrôle est une ellipse plutôt qu'un rectangle. Ainsi, le procédé est déclaré hors contrôle si et seulement si la paire $(\bar{x}_1, \bar{x}_2)$ se trouve à l'extérieur de cette ellipse. Cependant, si on utilise l'approche précédente, il est possible de conclure à tort que le procédé est sous contrôle (Région A), une caractéristique seulement est sous contrôle (Région

B), ou les deux caractéristiques sont hors contrôle (Région C). Le degré de corrélation entre ces deux caractéristiques, affecte donc la taille des régions de contrôle ainsi que leurs erreurs respectives. En outre, la probabilité que la paire des moyennes d'un échantillon se trouve à l'intérieur de l'ellipse de contrôle est $1 - \alpha$, tandis qu'avec la région de contrôle rectangulaire, cette même probabilité est au moins égale à $1 - 2\alpha$.

Ce chapitre se veut une revue de littérature portant sur les méthodes les plus répandues pour la construction et l'interprétation des cartes de contrôles multivariées en se basant sur l'hypothèse de normalité des observations.

2.2 La carte Shewhart multivariée

Comme dans le cas univarié, la construction d'une carte de contrôle multivariée passe par deux phases essentielles. La première (phase I) consiste en une analyse d'un large ensemble de données préliminaires, supposées être en état de contrôle statistique. Cette analyse conduit à l'estimation des paramètres qui vont être utilisés par la suite durant le monitorage du procédé (phase II).

Dans cette section, on s'intéresse à l'aspect multivarié de la carte de contrôle Shewhart, (connue aussi sous le nom de carte de contrôle pour la moyenne ou encore la carte $\bar{X}$). On va ainsi détailler les deux phases de construction de cette dernière. Nous commençons par la présentation de la phase II.

2.2.1 Phase II

Dans cette partie, on suppose que l'estimation des paramètres a été déjà réalisée durant la Phase I.

Quand il s'agit de contrôler une seule caractéristique de qualité qu'on suppose être distribuée suivant une loi normale de moyenne μ et d'écart type σ, la probabilité que la moyenne ($\bar{X}$) d'un échantillon du procédé soit entre $\mu_0 \pm z_{\alpha/2}(\sigma_0/\sqrt{n})$ est $(1 - \alpha)$, où $z_{\alpha/2}$ est le percentile standard de la loi normale tel que $P(Z > z_{\alpha/2}) = \alpha/2$. Donc, si $\bar{X}$ se trouve à l'extérieur de cet intervalle, une inspection du procédé s'impose.

La philosophie derrière la carte Shewhart consiste donc en une succession de tests statistiques de la forme :

$$H_0 : \mu = \mu_0 \ \ vs. \ H_1 : \mu \neq \mu_0$$

À ce stade, deux approches sont possibles. On peut utiliser la carte Shewhart avec les deux limites de contrôle (supérieure UCL[1] et inférieure LCL[2]), ou bien on peut l'utiliser juste avec la limite supérieure, où les valeurs à tracer sont $[\sqrt{n}(\bar{X}-\mu_0)/\sigma_0]^2$. Dans ce dernier cas, $UCL = \chi^2_{1,\alpha}$, où $\chi^2_{1,\alpha}$ représente le percentile d'ordre $1-\alpha$ d'une χ^2_1, *i.e.* $P(\chi^2_1 > \chi^2_{1,\alpha}) = \alpha$.

Ainsi l'hypothèse du test est rejetée si :

$$[\sqrt{n}(\bar{X}-\mu_0)/\sigma_0]^2 = n(\bar{X}-\mu_0)(\sigma_0^2)^{-1}(\bar{X}-\mu_0) > \chi^2_{1,\alpha} \tag{2.1}$$

L'extension de la carte Shewhart au cas multivarié est basée sur cette même idée, on rejette l'hypothèse H_0 si :

$$\chi^2_0 = n(\bar{X}-\mu_0)'\Sigma_0^{-1}(\bar{X}-\mu_0) > \chi^2_{p,\alpha} \tag{2.2}$$

où $\bar{X}$ est le vecteur $(p \times 1)$ moyen de l'échantillon et Σ_0^{-1} est l'inverse de la matrice variance-covariance $(p \times p)$.

Dans le cas de deux caractéristiques de qualité, la formule (2.2) devient :

$$\begin{aligned} \chi^2_0 \ = \ & n(1-\rho_0^2)^{-1}[(\bar{X}_1-\mu_{01})^2\sigma_{01}^{-2} + (\bar{X}_2-\mu_{02})^2\sigma_{02}^{-2} \\ & -2\rho_0\sigma_{01}^{-1}\sigma_{02}^{-1}(\bar{X}_1-\mu_{01})(\bar{X}_2-\mu_{02})] > \chi^2_{2,\alpha}. \end{aligned} \tag{2.3}$$

Le terme de gauche représente l'équation d'une ellipse centrée au point (μ_{01}, μ_{02}). Ainsi, dans le cas de deux caractéristiques de qualité, la région de contrôle est représentée par l'intérieur et la frontière d'une ellipse. Si le vecteur moyen, d'un échantillon bivarié, se trouve à l'extérieur de cette région, le procédé est déclaré hors contrôle et l'inspection visuelle peut révéler la caractéristique qui a causé le problème. (voir figure 2.1).

---

[1] "UCL" correspond à "Upper Control Limit"

[2] "LCL" correspond à "Lower Control Limit"

D'une manière générale, quand il s'agit de contrôler plus de deux caractéristiques de qualité à l'aide de la carte Shewhart multivariée, cette dernière prend comme limite de contrôle $UCL = \chi^2_{p,\alpha}$. Si $\chi^2_0 > UCL$, le procédé est déclaré hors contrôle statistique. La caractéristique derrière la détérioration du produit doit être déterminée. Cette dernière étape constitue l'handicap majeur pour la carte Shewhart multivariée. Une approche possible, est d'utiliser des cartes Shewhart univariées supplémentaires. La probabilité d'erreur de type I, pour chacune de ces cartes doit être égale à α/p.

2.2.2 Estimation des paramètres (phase I)

Durant la phase II, la carte de contrôle Shewhart est utilisée pour contrôler les moyennes des échantillons du procédé. Généralement dans la Phase I, les paramètres μ_0 et Σ_0 sont inconnus, et doivent être estimés à partir d'échantillons préléminaires. Dans la suite, on suppose qu'on dispose de m échantillons de même taille n et que ces échantillons ont été recueillis lorsque le procédé est sous contrôle statistique.

Dans le cas univarié, la procédure ordinaire, utilisée pour déterminer les limites de contrôle consiste à remplacer les valeurs théoriques μ_0 et σ_0 dans la phase II par leurs estimés non biaisés, obtenus à partir des m échantillons. Par exemple, μ_0 doit être remplacé par la moyenne des moyennes de chaque sous-groupe, et toutes les mesures de variabilité doivent être utilisées pour remplacer σ_0. À cet égard, Hillier [34] et Yang and Hillier [81], ont développé une procédure, basée sur deux étapes. Son objectif est de vérifier si les données des m sous-groupes proviennent réellement du procédé en état de contrôle statistique (étape I), et si les futurs sous-groupes de données représentent le contrôle statistique (étape II). Cette procédure a été généralisée au cas mutivarié par Alt and al. [1].

Étape I

Si les m échantillons préliminaires ont été prélevés d'un procédé en état de contrôle statistique, alors des estimations non biaisées du vecteur moyen et la matrice variance-covariance du procédé, sont données par :

$$\bar{\bar{X}} = (1/m)\sum_{i=1}^{m} \bar{X}_i \quad \text{et} \quad \bar{S} = (1/m)\sum_{i=1}^{m} S_i,$$

où $\bar{X}_i$ et S_i désignent la moyenne et la matrice variance-covariance de l'échan-

tillon i.

Dans le cas où les valeurs standards μ et Σ sont disponibles, le test statistique est celui donné par l'équation (2). Dans le cas contraire, le test sera basé sur la statistique

$$T_{0,1}^2 = n(\bar{X}_i - \bar{\bar{X}})'\bar{S}^{-1}(\bar{X}_i - \bar{\bar{X}}). \tag{2.4}$$

$i = 1, 2, ..., m$. Dans le cas de deux caractéristiques de qualité,

$$\begin{aligned} T_{0,1}^2 &= (n/det(\bar{S}))[(\bar{X}_{1,i} - \bar{\bar{X}}_1)^2\bar{s}_2^2 + (\bar{X}_{2,i} - \bar{\bar{X}}_2)^2\bar{s}_1^2 \\ &\quad -2(\bar{X}_{1,i} - \bar{\bar{X}}_1)(\bar{X}_{2,i} - \bar{\bar{X}}_2)\bar{s}_{12}] \end{aligned} \tag{2.5}$$

où $det(\bar{S}) = \bar{s}_1^2\bar{s}_2^2 - \bar{s}_{12}^2$, $\bar{s}_1^2 = (1/m)\sum_{i=1}^m s_{1,i}^2$, $\bar{s}_2^2 = (1/m)\sum_{i=1}^m s_{2,i}^2$, et $\bar{s}_{12}^2 = (1/m)\sum_{i=1}^m s_{12,i}^2$.

Alt and al. [5] ont démontré que $T_{0,1}^2$ suit la même loi que la variable aléatoire $c_1(m,n,p)F_{p,mn-m-p+1}$, où $F_{p,mn-m-p+1}$ est la distribution F de premier paramètre (degré de liberté numérateur) p et de deuxième paramètre (degré de liberté dénominateur) $mn - m - p + 1$, et,

$$c_1(m,n,p) = p(m-1)(n-1)(mn - m - p + 1). \tag{2.6}$$

Pour vérifier si le procédé était sous contrôle lors de la collecte des m premiers sous-groupes de données, on trace les m valeurs de $T_{0,1}^2$ sur la carte avec comme limites de contrôle $UCL = c_1(m,n,p)F_{p,mn-m-p+1}$ et $LCL = 0$. Si une valeurs $T_{0,1}^2$ pour les m premiers sous-groupes se trouvent hors contrôle, les sous-groupes correspondants sont éliminés et on répète l'étape I.

Étape II

Durant l'étape I, la limite supérieure du contrôle a été révisée et la statistique du test pour les sous-groupes restants ne dépasse pas cette limite. Durant l'étape II, la carte de contrôle commence à fonctionner avec les futurs sous-groupes. Soit $\bar{X}_f$ le $(p \times 1)$ vecteur des moyennes de l'échantillon pour les futurs sous-groupes. En remplaçant $\bar{X}_i$ par $\bar{X}_f$ dans l'équation (2.4), on obtient la statistique du test pour l'étape II :

$$T_{0,2}^2 = n(\bar{x}_f - \bar{\bar{x}})'\bar{S}^{-1}(\bar{x}_f - \bar{\bar{x}}) \tag{2.7}$$

où $\bar{\bar{x}}$ et $\bar{S}$ sont obtenues durant l'étape I. Alt and al. (voir [6]) ont démontré que $T^2_{0,2}$ est distribuée suivant : $c_2(m,n,p)F_{p,mn-m-p+1}$

où

$$c_2(m,n,p) = p(n-1)(m+1)(mn-m-p+1). \tag{2.8}$$

Pour déterminer si les moyennes restent sous contrôle durant cette étape, on trace les valeurs $T^2_{0,2}$ sur la carte de contrôle dont les limites de contrôle sont $UCL = c_2(m,n,p)F_{p,mn-m-p+1,\alpha}$ et $LCL = 0$. Si $T^2_{0,2}$ dépasse UCL, on dit qu'il y a présence d'une cause de variation. Généralement, pour détecter la variable qui a causé cet effet, on introduit des cartes univariées supplémentaires pour chaque caractéristique de qualité.

2.3 Les cartes de contrôle multivariées pour la dispersion

Bien que les cartes de contrôle pour la moyenne nous donnent une information précieuse sur les caractéristiques de qualité durant le processus de fabrication d'un procédé, elles ne permettent pas de quantifier l'éparpillement des valeurs de ces dernières et de préciser ainsi l'ampleur avec laquelle les valeurs observées s'écartent les unes des autres ou s'écartent de leur valeur centrale. Ainsi, pour compléter la caractérisation de qualité d'un procédé, on fait appel également aux cartes de contrôle pour la dispersion (ou variabilité).

Une carte de contrôle pour la dispersion consiste à tracer les valeurs d'une certaine mesure de dispersion de chaque échantillon en fonction du temps. Les principales mesures de dispersion sont l'étendue et l'écart type.

Dans ce paragraphe, on va présenter quelques généralités sur les cartes de contrôle de dispersion. En particulier, on s'attardera sur la carte $|S|$ qui est la plus recommandée dans la littérature.

2.3.1 Généralités

Considérons un procédé dont la qualité du produit final est caractérisée par p variables $X_1, X_2, ..., X_p$ qu'on suppose être distribuées suivant une loi normale multivariée.

Comme pour les cartes de contrôle basées sur la moyenne, on peut toujours penser à placer une carte de contrôle de dispersion univariée (s^2 par exemple) pour chacune des p caractéristiques. Ce qui nous amène à appliquer le test statistique de la forme :

$$H_0 : \sigma_i^2 = \sigma_{0i}^2 \quad vs. \quad H_1 : \sigma_i^2 \neq \sigma_{i0}^2 \quad i = 1, 2, ..., p \tag{2.9}$$

où $\sigma_0 = (\sigma_{10}, \sigma_{20}, ..., \sigma_{p0})$ est le vecteur écart type des p caractéristiques.

Mais, en procédant ainsi, on ne prend pas en considération la structure de dépendance éventuelle entre les p caractéristiques. Ceci peut engendrer des erreurs désastreuses comme nous l'avons montré dans la section précédente.

Pour éviter ce problème, une approche possible est basée sur la même philosophie que celle de la carte s^2 univariée. Elle consiste à effectuer le test multivarié suivant :

$$H_0 : \Sigma = \Sigma_0 \quad vs. \quad H_1 : \Sigma \neq \Sigma_0$$

À ce stade, en utilisant le résultat du test du rapport de vraisemblance asymptotique, on peut penser à calculer la statistique W suivante pour chaque échantillon :

$$W = -pn + pn \ \ln\left(|A|/|\Sigma_0|\right) + tr(\Sigma_0^{-1} A) \tag{2.10}$$

où $A = (n-1)S$, S étant la matrice $(p \times p)$ variance-covariance, tr est l'opérateur trace d'une matrice, et $|\cdot|$ désigne une norme matricielle.

Sous l'hypothèse H_0, la statistique W suit approximativement la loi du *Chi-carrée* (voir [10]). Donc, il est possible de calculer les limites de contrôle appropriées donnant une probabilité d'erreur de Type I, mais la distribution exacte de W est inconnue.

Une autre approche possible, est d'utiliser la variance généralisée (qu'on note par $|S|$). On peut montrer, dans le cas de deux caractéristiques de qualité, que :

$$\frac{2(n-1)|S|^{1/2}}{|\Sigma_0|^{1/2}} \equiv \chi^2_{2n-4}. \tag{2.11}$$

donc, il est possible de spécifier un test statistique pour une probabilité d'erreur de Type I donnée.

2.3.2 La distribution de $|S|$

On considère un échantillon de taille n pris quand le procédé est supposé être en état de contrôle statistique. Anderson [7] a donné une expression de la distribution de $|S|$ quand il s'agit de contrôler p caractéristiques de qualité.

$$\frac{|S|(n-1)^p}{|\Sigma_0|} \equiv \prod_{i=1}^{p} \chi^2_{n-i} \tag{2.12}$$

où $|\Sigma_0|$ est le déterminant de la matrice variance-covariance et les variables *Chi-carrée* sont indépendantes. Donc, si on prévoit la construction d'une carte $|S|$ pour tester s'il y a des changements dans la dispersion du procédé, on doit spécifier une ou deux limites de contrôle pour cette carte.

2.3.3 Les limites de contrôle pour la carte $|S|$

Bien que la distribution de $|S|$ puisse s'exprimer en fonction de lois χ^2, elle n'est pas une loi standard et il n'existe pas de table pour cette loi. Cet état de fait présente un léger inconvénient pour la mise en œuvre pratique de la carte $|S|$.

Par ailleurs, Aparisi, Jabaloyes, and Carrión (voir [10]), ont déterminé la densité de probabilité de la statistique $J_{n,p} \equiv \frac{|S|(n-1)^p}{|\Sigma_0|}$.

Pour construire une limite supérieure UCL pour la carte $|S|$, il suffit de fixer une erreur de Type I, α et de poser

$$\begin{aligned} \alpha &= P(|S| > UCL) \\ &= P\left[\frac{|S|(n-1)^p}{|\Sigma_0|} > \frac{UCL(n-1)^p}{|\Sigma_0|}\right] \\ &= P\left[J_{n,p} > \frac{UCL(n-1)^p}{|\Sigma_0|}\right] \end{aligned}$$

Désignons par $J^{\alpha}_{n,p}$ le quantile d'ordre $(1-\alpha)$ de la distribution de $J_{n,p}$. La

limite de contrôle pour cette carte est donc donnée par :

$$UCL = \frac{J_{n,p}^{\alpha}|\Sigma_0|}{(n-1)^p} \tag{2.13}$$

Dans le cas où on désire utiliser la carte $|S|$ avec une limite inférieure LCL et une limite supérieure UCL, il suffit de fixer une erreur de type I, α, et de procéder de la même manière que celle de la carte précédente, *i.e.*

$$P(LCL < |S| < UCL) = 1 - \alpha$$

$$\begin{aligned} 1-\alpha &= P(LCL < |S| < UCL) \\ &= P\left[\frac{LCL(n-1)^p}{|\Sigma_0|} < \frac{|S|(n-1)^p}{|\Sigma_0|} < \frac{UCL(n-1)^p}{|\Sigma_0|}\right] \\ &= P\left[\frac{LCL(n-1)^p}{|\Sigma_0|} < J_{n,p} < \frac{UCL(n-1)^p}{|\Sigma_0|}\right] \end{aligned}$$

Soient $J_{n,p}^{1-\alpha/2}$ et $J_{n,p}^{\alpha/2}$ les percentiles de $J_{n,p}$ d'ordres respectifs $\alpha/2$ et $(1-\alpha/2)$. Ceci permet d'exprimer les deux limites de contrôles comme suit :

$$LCL = \frac{J_{n,p}^{1-\alpha/2}|\Sigma_0|}{(n-1)^p}$$

et

$$UCL = \frac{J_{n,p}^{\alpha/2}|\Sigma_0|}{(n-1)^p}$$

Remarque

Le plus souvent, les cartes de contrôle de dispersion ne comporte qu'une limite de contrôle supérieure, puisqu'on cherche à se prémunir contre les

augmentations de variabilité du procédé. Parfois on indique également une limite inférieure pour repérer ainsi les cas où la variabilité est plus faible que la normale, ce qui peut provenir d'une avarie de l'appareil de mesure. D'autre part, une diminution accidentelle de la variabilité est intéressante car elle donne une piste pour rechercher des conditions de fonctionnement du procédé qui favorise cette diminution qui est toujours profitable à la qualité.

2.4 Carte de contrôle multivariée pour l'information cumulée : Carte MCUSUM [3]

Jusqu'à présent, les cartes de contrôle qu'on a traitées (dites cartes Shewhart) consistaient à reporter en ordonnée une mesure statistique des caractéristiques de qualité ($\bar{X}$, $|S|$) en fonction de l'instant de l'échantillonnage du procédé. Toutefois, chaque point d'une carte de contrôle conventionnelle est basé sur l'information obtenue d'un sous-groupe ou d'un certain nombre de sous-groupes si on applique les tests présentés dans les sections précédents. Elles permettent de détecter rapidement des écarts importants des caractéristiques de qualité. Elles sont néanmoins moins performantes pour détecter de faibles variations de la moyenne du procédé par rapport à une cible donnée.

Les cartes de contrôle pour l'information cumulée des caractéristiques de qualité (**MCUSUM**), qu'on va présenter dans cette section, ont été proposées pour obtenir un outil d'aide à la décision plus sensible aux faibles variations que pourraient subir les caractéristiques de qualité. Dans un premier temps, on va parler de la carte $CUSUM$ univariée, ensuite de sa généralisation dans le cas multivarié.

2.4.1 La carte de contrôle $CUSUM$

La carte de contrôle $CUSUM$, due à Page [61], comporte en ordonnée la somme cumulative des écarts d'une mesure statistique (valeur individuelle, moyenne, écart type, nombre de non-conformités,...) par rapport à une valeur cible en fonction des instants d'échantillonnage du procédé.

En effet, à partir des moyennes $\bar{X}_i$ des échantillons prélevés, on calcule :

[3] "**MCUSUM**" correspond à "Multivariate Cumulative-Sum Control Chart"

$$\begin{aligned} z_i &= \sqrt{n}\frac{\bar{X}_i - \mu_0}{\sigma} \\ S_i^+ &= \max\{0, S_{i-1}^+ + (z_i - k)\} \\ S_i^- &= \min\{0, S_{i-1}^- + (z_i + k)\} \end{aligned}$$

La statistique S_i^+ permet de détecter des augmentations de moyenne, tandis que S_i^- permet de détecter ses diminutions. k est un coefficient de sensibilité qu'il est conseillé de prendre égal à $\delta/2$, où $\delta = \sqrt{n}\frac{|\mu-\mu_0|}{\sigma}$ est le déréglage qu'on souhaite mettre en évidence. μ étant l'espérance théorique de l'échantillon soumis au contrôle et μ_0 représente la valeur cible de cette dernière.

On construit ainsi une carte de contrôle où l'on note les valeurs de S_i^+ et de S_i^- en fonction de i. Comme la carte $\bar{X}$, cette carte comporte des limites de contrôle supérieure et inférieure situées à h et $-h$. On ne fait rien tant que S_i^+ et S_i^- sont situées entre ces deux limites, et on décide que le procédé est hors contrôle dès que l'une des deux valeurs sort des limites. S'il y a un décalage positif de la moyenne de z supérieur à k, les écarts $(z - k)$ seront cumulés dans S_i^+ jusqu'à ce que S_i^+ atteigne h ; on décidera alors de faire un réglage du procédé. Il en est de même du coté négatif pour S_i^-.

La valeur recommandée pour k, pour des valeurs de δ entre 0.5 et 2.0 est $k = \frac{|\delta|}{2}$. Ainsi k correspond à la variation permise pour la somme cumulative pour une variation unitaire sur l'axe horizontal de la carte (instants de prélèvement).

2.4.2 La carte MCUSUM

Comme toutes les autres cartes multivariées qu'on a vues jusqu'à présent, une carte "**MCUSUM**" peut être obtenue de sa version univariée (carte "*CUSUM*"). Pour cette fin, il y a deux différentes approches. La première consiste à placer une carte *CUSUM* pour chaque caractéristique de qualité et puis analyser l'ensemble des ces dernières simultanément. Le procédé est ainsi déclaré hors-contrôle si l'une des p caractéristiques est déclarée hors-contrôle. La deuxième approche consiste à modifier la carte *CUSUM* en se basant sur l'une des deux stratégies suivantes. La première consiste à transformer chaque observation multivariée en un scalaire et former par la suite une carte *CUSUM* univariée à partir des scalaires obtenus. La deuxième méthode consiste à accumuler le vecteur d'observations (X) avant de les réduire

en scalaire. Ceci permet de construire la carte **MCUSUM** directement des observations. Crosier [18] a montré que la carte **MCUSUM** dépend uniquement du paramètre de non-centralité.

Une carte **MCUSUM** avec une meilleure performance d'ARL[4] (pour plus d'information sur l'ARL voir [29]), pour détecter les changements anticipés et imprévus, Z, est développée par Pignatiello and Ranger [62] dont les équations sont données par :

$$\begin{cases} Z_t = \max(\|C_t\| - kn_t, 0) \\ \|C_t\| = \sqrt{C_t' \Sigma^{-1} C_t} \\ C_t = \sum_{i=t-n_t+1}^{t} (X_i - \mu_0) \\ n_t = \begin{cases} n_{t-1} & \text{si } Z_{t-1} > 0 \\ 1 & \text{ailleurs} \end{cases} \end{cases}$$

où Z est la statistique à tracer, $\|C_t\|$ est la norme de C_t, n_t est le nombre de sous-groupes et k est un coefficient de sensibilité qui dépend uniquement de la distance entre μ_0 et μ_1. Le coefficient k et les limites de contrôle sont choisis de façon à obtenir une efficacité donnée à priori.

À la différence de la carte Shewhart multivariée, il n'est pas possible de donner une probabilité de déceler un déréglage pour un échantillon, car elle dépend de son numéro de contrôle ; avec la carte **MCUSUM**, il est probable de déceler un déréglage au bout de quelques contrôles que lors du premier d'entre eux. L'efficacité de la carte **MCUSUM** se juge d'après la période opérationnelle moyenne. Si le procédé est sous contrôle, on souhaite que les fausses alarmes (un point situé par hasard hors des limites de contrôle) soient peu nombreuses. Par contre s'il y a un déréglage, on souhaite qu'il soit décelé le plus vite possible.

La carte **MCUSUM** a une très bonne efficacité comparée à la carte Shewhart multivariée. Il faut noter, en particulier le gain d'efficacité sur cette dernière pour des déréglages faibles ou moyens, ceci étant d'autant plus marqué que le coefficient k est faible. Par contre la carte Shewhart multivariée est plus efficace pour des déréglages élevés.

[4] "*ARL*" correspond à (Average Run Lengths) ou (La période opérationnelle moyenne)

2.5 Conclusion

Dans ce chapitre, on a détaillé les principales cartes de contrôle multivariées qu'on peut trouver dans la littérature, ainsi que leurs conditions de mise en œuvre dans les procédés de fabrication. Il existe d'autres cartes de contrôle multivariées qu'on n'a pas présentées. On cite à titre d'exemple la carte $MEWMA$[5] (voir [46] pour plus d'information) qui est utilisée dans les mêmes conditions que la carte **MCUSUM** et les cartes basées sur l'analyse en composantes principales, etc.

Cette revue de littérature nous a également permis de constater aussi que la majorité des procédures de construction des cartes de contrôle multivariées, sont basées sur l'hypothèse de normalité des observations. Malheureusement, cette hypothèse n'est pas toujours rencontrée en pratique.

L'approche qu'on va présenter dans le chapitre 4, n'impose aucune hypothèse sur la distribution des données du procédé. Elle se base sur la notion de profondeur de Tukey qu'on va détailler au chapitre suivant.

[5] "$MEWMA$" correspond à Multivariate Exponentially Weighted Moving Average Control Chart

Chapitre 3

La Profondeur de Tukey

3.1 Introduction

Étant donné un nuage de points régis par une distribution de probabilité multivariée, et un point x fixé, la profondeur de x relativement à ce nuage est une notion qui mesure le degré de centralité de x par rapport à l'ensemble du nuage. Il existe plusieurs définitions de la profondeur d'un point (voir [42]). Dans ce travail nous nous intéressons exclusivement à la profondeur de John Tukey (voir Tukey 1975 [77]). Elle se définit comme suit :

Soit $X = \{X_1, X_2, ..., X_n\}$ un ensemble de données univariées et x un point donné. La profondeur de Tukey $\mathcal{D}(x, X)$ (quand il n'y a pas de confusion, on peut également utiliser la notation $\mathcal{D}_1(x, X)$, l'indice 1 correspond à la dimension) du point x relativement à X est égale au minimum entre le nombre de points de l'ensemble X, situés à gauche et ceux à droite de x. *i.e.*

$$\mathcal{D}_1(x, X) \; = \; n^{-1} \min(\#\{i : X_i \leq x\}, \#\{i : X_i \geq x\}).$$

Par exemple, pour $n = 19$ et $x_1 < x_2 < ... < x_{19}$, on obtient pour $x_3 \leq \theta \leq x_4$, $\mathcal{D}_1(\theta, X) = 3/19$ et pour $x_{15} \leq \xi \leq x_{16}$, $\mathcal{D}_1(\xi, X) = 4/19$. La médiane dans ce cas est $\tau = x_{10}$. Sa profondeur est maximale et vaut $\mathcal{D}_1(\tau, X) = 10/19$.

Dans le cas multivarié, la profondeur de Tukey d'un point $\theta \in \mathbb{R}^d$ par rapport à un ensemble X de données de dimension d est définie comme la plus petite profondeur de θ dans toutes les projections univariées. Plus précisément,

Définition 1 *Soit $u \in \mathbb{R}^d$ tel que $\|u\| = 1$, alors l'ensemble $\{u^\top X_i\}$ est une projection unidimensionnelle de X sur la direction u et on définit :*

$$\begin{aligned} \mathcal{D}_d(\theta; X) &= min_{\|u\|=1} \mathcal{D}_1(u^\top \theta; \{u^\top X_i\}) \\ &= n^{-1} min_{\|u\|=1} \#\{i : u^\top X_i \geq u^\top \theta\}. \end{aligned} \tag{3.1}$$

La figure 3.1 illustre la profondeur de Tukey d'un point arbitraire $\theta \in \mathbb{R}^2$ par rapport à un ensemble bivarié X (de taille $n = 9$). Quelques demi-plans dont la frontière passe par θ sont indiqués. On peut remarquer facilement que tels demi-plans contiennent au moins deux points de X. Donc, d'après la définition 1, $\mathcal{D}_2(\theta; X) = 2/9$.

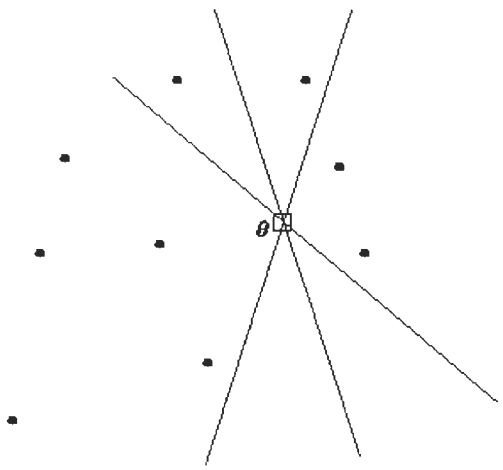

FIG. 3.1 – Illustration de la profondeur de Tukey dans le cas bivarié.

Tukey a utilisé la notion de profondeur comme moyen pour indiquer la forme des données bivariées. Il a suggéré par la suite que cette notion peut nous permettre de définir un ordre statistique raisonnable dans le cas multivarié. Il a ainsi introduit la définition de la médiane d'un ensemble de données multivariées comme le point le plus profond par rapport à ce dernier. D'autres réalisations statistiques ont vu le jour grâce à cette nouvelle profondeur, notamment dans le domaine des statistiques multivariées.

Ce chapitre est consacré à la profondeur de Tukey. Nous commencerons par un résumé de quelques propriétés théoriques nécessaires pour la réalisation de notre projet. Ensuite, nous traitons l'aspect algorithmique de la

profondeur de Tukey, pour ce, on se base sur les travaux de Rousseeuw (voir [66, 67]).

3.2 Définitions et propriétés

La définition de la profondeur de Tukey a été introduite dans le cas discret seulement. D'une manière générale, on peut la définir de la façon suivante.

Définition 2 *Soit F une distribution de probabilité sur $\mathbb{R}^d$. Notons par $\mathcal{H}$ la classe des semi-espaces fermés H dans $\mathbb{R}^d$. La profondeur de Tukey d'un point $x \in \mathbb{R}^d$ par rapport à F est définit par :*

$$\mathcal{D}(x) = \inf\{F(H) : H \in \mathcal{H}, x \in H\}. \tag{3.2}$$

Il est clair que les points ayant une grande profondeur sont concentrés au centre de la distribution F, tandis que ceux avec une petite profondeur appartiennent aux queues de F.

Si la distribution F d'un nuage de points X_i, $i = 1, ..., n$ est inconnue, on estime F à l'aide de la distribution empirique définie par :

$$F_n(S) = n^{-1}\#\{i, X_i \in S\}$$

où S est un ensemble mesurable. Ainsi, pour obtenir la version empirique de la profondeur de Tukey (qu'on note par $\mathcal{D}_n(x)$), il suffit de remplacer F par sa version empirique F_n dans 3.2.

D'après la définition 2, il est facile de remarquer que $\mathcal{D}(x)$ est déterminée par la classe des semi-espaces fermés H tels que $x \in \partial H$, la frontière topologique de H. Soit $U = \{u \in \mathbb{R}^d : \|u\| = 1\}$ et $H[x, u] = \{v \in \mathbb{R}^d : u^\top v \geq u^\top x\}$ pour $u \in U$, on a alors les deux expressions suivantes :

$$\mathcal{D}(x) = \inf_{u \in U} FH[x, u] \tag{3.3}$$

$$\mathcal{D}_n(x) = \inf_{u \in U} F_n H[x, u] \tag{3.4}$$

dans certains cas, on utilise la notation $\mathcal{D}_F(.)$ (resp. $\mathcal{D}_{F_n}(.)$) pour indiquer la dépendance de F (resp. de F_n quand F est inconnue).

En général, pour comparer la version théorique et empirique d'une distribution, on utilise la métrique μ_H suivante :

$$\mu_H(F_n, F) = \sup_{u,x} |F_n H[x, u] - FH[x, u]|. \tag{3.5}$$

C'est la plus grande discrépance entre F_n et F sur tous les sous-espaces. On remarque que μ_H vérifie la propriété de Glivenko-Cantelli ; *i.e.* si les X_i sont *i.i.d.* de loi F, alors

$$\mu_H(F_n, F) \to 0 \ p.s.^1 \text{ quand } n \to \infty. \tag{3.6}$$

Pour plus d'information sur la métrique μ_H, voir [72, 64].

Proposition 1 *La profondeur de Tukey vérifie la propriété de l'invariance affine :*

$$\mathcal{D}(Ax + b; \{AX_i + b\}) = \mathcal{D}(x, X) \tag{3.7}$$

pour tout vecteur b et toute transformation affine non singulière A. Dans le cas continu l'équation (3.7) devient :

$$\mathcal{D}(F_{A,b}; \{A \cdot + b\}) = \mathcal{D}(F, \cdot) \tag{3.8}$$

où $F_{A,b}$ est la distribution de $AX + b$ et $X \sim F$.

Démonstration:
Nous avons

$$X_i \in H[x, u] \Longleftrightarrow AX_i + b \in AH[x, u] + b$$

pour tout b et toute A non singulière. Par conséquent,

$$min_{||u||=1} \#\{i, X_i \in H[x, u]\} = min_{||u||=1} \#\{i, AX_i + b \in AH[x, u] + b\}.$$

et le résultat découle d'après la formule (3.1) de la définition 1.

Proposition 2 *Comme fonction de x, $\mathcal{D}$ est semi-continue supérieurement. Si de plus F est absolument continue, $\mathcal{D}$ est continue.*

Démonstration:
Notons initialement que pour chaque demi-espace fermé H, la fonctionnelle linéaire $F \to F(H)$ est semi-continue supérieurement (**s.c.s**) pour la convergence faible. On a aussi, pour $\nu \to 0$ la mesure $F(. - \nu)$ converge faiblement vers F. Comme $H[x, u] = H[w, u] + (x - w)$, $f_u(x) = F(H[x, u])$ est donc semi-continue supérieurement en x. Or, d'après l'expression (3.3), on a :

[1] *p.s.* correspond à la convergence presque sûre.

$$\mathcal{D}(x) = \inf_{u \in U} F(H[x,u]) = \inf_{u \in U} f_u(x).$$

ainsi, $\mathcal{D}$ est l'infimum d'une collection de fonctions **s.c.s**, donc elle est **s.c.s**.

Maintenant, on va montrer que si F est absolument continue, $\mathcal{D}$ est aussi semi-continue inférieurement (**s.c.i**) et par conséquent continue. En effet, soit $x_n \to x_0$, et soit u_n une suite de directions vérifiant :

$$F(H[x_0,u_n]) \leq \mathcal{D}(x_n) + 1/n, \forall n.$$

comme toutes les u_n appartiennent à la sphère unitaire de $\mathbb{R}^d$, alors la suite u_n possède un point d'accumulation, on peut donc supposer que u_n converge vers u (dans le cas contraire on en extrait une sous-suite convergente). On a alors :

$$F(H[x_0,u]) - F(H[x_n,u_n]) = \int (\mathbf{I}_{H[x_0,u]} - \mathbf{I}_{H[x_n,u_n]})\, dF,$$

où $\mathbf{I}_S$ est la fonction indicatrice de l'ensemble S. La différence des fonctions indicatrices est bornée en valeur absolue par la constante 1. Comme $u_n \to u$ et $x_n \to x_0$, la différence tend vers zéro presque partout (plus précisément $(\mathbf{I}_{H[x_0,u]} - \mathbf{I}_{H[x_n,u_n]}) \to 0$ sauf peut-être sur l'hyperplan $\{y; y^\top u = u^\top x_0\}$). Une application du théorème de la convergence dominée, nous donne $F(H[x_0,u]) - F(H[x_n,u_n]) \to 0$ si $u_n \to u$, $x_n \to x_0$. Ainsi,

$$\liminf_{n\to\infty} \mathcal{D}(x_n) = \liminf_{n\to\infty} F(H[x_n,u_n]) = F(H[x_0,u]) \geq \mathcal{D}(x_0).$$

d'où $\mathcal{D}$ est semi-continue inférieurement.

Définition 3 *On dit qu'une distribution de probabilité F est centro-symétrique autour de x_0 si $F(x_0 + S) = F(x_0 - S)$ pour tout ensemble S mesurable.*

Proposition 3 *Si F est centro-symétrique par rapport à x_0, alors $\mathcal{D}(x_0) \geq 1/2$. Si, de plus F est absolument continue, alors $\mathcal{D}(x_0) = 1/2$.*

Démonstration:
Nous avons $H[x_0, u] = x_0 + H[0, u]$ et d'après la centro-symétrie de F par rapport à x_0

$$F(H[x_0, u]) = F(x_0 + H[0, u]) = F(x_0 - H[0, u]) = F(H[x_0, -u]).$$

Comme $H[x_0, u] \cup H[x_0, -u] = \mathbb{R}^d$ et $H[x_0, u] \cap H[x_0, -u]$ est un hyperplan, on a

$$2F(H[x_0, u]) = F(H[x_0, u]) + F(H[x_0, -u]) \geq 1$$

d'où $F(H[x_0, u]) \geq 1/2$.
Si F est absolument continue, alors $F(\partial H[x_0, u]) = 0$ pour tous les semi-espaces. Donc $F(H[x_0, u]) + F(H[x_0, -u]) = 1$, et par la suite $F(H[x_0, u]) = 1/2$.

Les deux propositions suivantes montrent que la profondeur de Tukey atteint toujours son maximum.

Proposition 4

$$\sup_{x \in \mathbb{R}^d} \mathcal{D}(x) \geq 1/(d+1).$$

Démonstration:
La démonstration de ce résultat est assez technique. Le lecteur peut se référer à [25] pour plus d'informations.

Proposition 5 *$\mathcal{D}$ atteint sa borne supérieure.*

Démonstration:
Soit B_R la boule fermée de centre 0 et de rayon R. On choisit R assez large de telle sorte que $F(B_R) > 1 - 1/(d+2)$, alors on peut montrer facilement que

$$\sup_{x \in \mathbb{R}^d \setminus B_R} \mathcal{D}(x) < 1/(d+2).$$

En effet, pour chaque $x \in B_R^c$, il existe une direction $u \in U$ telle que $H[x, u] \cap B_R = \emptyset$, d'où $\mathcal{D}(x) \leq F(H[x, u]) \leq F(B_R^c) < 1/(d+2)$. En appliquant les propositions 1 et 2, on déduit que $A = \{x : \mathcal{D}(x) \geq 1/(d+2)\}$ est un sous-ensemble fermé non vide de B_R, et par conséquent un ensemble compact. Or, $\mathcal{D}$ est semi-continue supérieurement, elle atteint donc sa borne supérieure sur A. Ce qui achève la démonstration.

Proposition 6

$$\lim_{R\to\infty} \sup_{\|x\|>R} \mathcal{D}(x) = 0$$

Démonstration:
La preuve de ce résultat est similaire à celle de la proposition 3.2. Il suffit de remplacer $1/(d+2)$ par un $\epsilon > 0$ arbitraire.

Définition 4 *Une fonction f définie sur $\mathbb{R}^d$ et à valeurs réelles est dite quasi-concave si $\{x, f(x) \geq c\}$ est un ensemble convexe $\forall x \in \mathbb{R}$.*

Proposition 7 *Comme fonction de x, $\mathcal{D}$ est quasi-concave.*

Démonstration:
Pour α fixe, soient $x_1 \neq x_2$ tels que $\mathcal{D}(x_i) \geq \alpha$, $i = 1, 2$. On pose $y = \lambda x_1 + (1-\lambda)x_2$ pour un certain $\lambda \in (0,1)$, et on suppose par absurde que $\mathcal{D}(y) < \alpha$. Alors, puisque $\min\{F(H[x_1,u]), F(H[x_2,u])\} \leq F(H[y,u]) < \alpha$ pour un certain $u \in U$, on obtient une contradiction.

Définition 5 *On dit que la distribution de probabilité F vérifie la condition de régularité **(R)**, si :*

(R) *$F(\partial H) = 0$ pour tout $H \in \mathcal{H}$.*

Proposition 8 *Supposons que F satisfait la condition **(R)** alors :*

(a) *la fonction $(x,u) \mapsto F(H[x,u])$ est continue sur $\mathbb{R}^d \times U$.*

(b) *la fonction $x \mapsto \mathcal{D}(x)$ est continue.*

Une question qui se pose à ce stade est à propos de la profondeur maximale pour un nuage de points X donné. Dans le cas univarié, il est facile de voir que la médiane possède la profondeur maximale (qui vaut 1/2). Pour une dimension > 1, la profondeur maximale peut être inférieure à 1/2. Ceci dépend de la forme du nuage de points X. Les propositions suivantes répondent à cette question.

Pour commencer, on introduit les notations suivantes :

$$k^*(X) = \max_x \mathcal{D}(x, X)$$
$$k^+(X) = \max_i \mathcal{D}(X_i, X)$$

qui représentent la profondeur maximale pour $x \in \mathbb{R}^d$ et pour $X_i \in X$ respectivement.

Définition 6 *On dit qu'un nuage de points X de dimension d est en position générale si tout sous-espace affine de dimension $(d-1)$ contient au plus d points de X.*

En particulier, pour $d = 2$, un nuage de points est dit en position générale si chaque droite du plan contient au plus 2 points. Pour $d = 3$, chaque plan contient au plus 3 points, et ainsi de suite.

Proposition 9 *On suppose que l'ensemble X est en position générale. Alors la profondeur maximale $k^*(X)$ est comprise entre $1/(d+1)$ et $1/2$.*

Démonstration:
Pour X en position générale, il existe une direction $v \in U$ telle que les points de l'ensemble $\{v^T X_i\}$ ne soient pas liés. Dans cette direction, la profondeur maximale est $1/2$. D'autre part, on a :

$$\mathcal{D}_d(x, X) = \min_{|u|=1} \mathcal{D}_1(u^\top x, \{u^\top x_i\}) \leq 1/2.$$

Donc $k^*(X) \leq 1/2$. L'autre inégalité découle de la proposition 4.

Concernant $k^+(X)$, on peut seulement dire que $1/n \leq k^+(X) \leq k^*(X)$. Si en plus l'ensemble X est presque symétrique, la profondeur maximale devient strictement supérieure à $1/(d+1)$.

Proposition 10 *Soit $X^{(n)} = \{X_1, ..., X_n\}$ un nuage de points suivant une distribution de probabilité F absolument continue et centro-symétrique. Alors $k^*(X^{(n)})$ converge en probabilité et presque sûrement vers $1/2$ quand $n \to +\infty$. Si de plus, F a une densité positive en x_0, alors $k^+(X^{(n)})$ converge en probabilité et presque sûrement vers $1/2$.*

Démonstration:
Comme F est centro-symétrique et absolument continue, la proposition 2 implique que $\mathcal{D}(x_0) = 1/2$. D'autre part, on a :

$$\mathcal{D}(x, X^{(n)}) = \mathcal{D}_n(x_0) \to \mathcal{D}(x_0) \quad p.s.$$

et puisque F est absolument continue, le nuage $X^{(n)}$ est alors en position générale. Ainsi, par la proposition 9,

$$1/2 \geq k^*(X^{(n)}) \geq \mathcal{D}(x_0, X^{(n)}).$$

En combinant les deux assertions précédentes, on a $k^*(X^{(n)}) \to 1/2 \quad p.s.$
On considère maintenant $k^+(X^{(n)})$. Soit X_{i_n} le plus proche voisin de x_0 dans $X_1, ..., X_n$. En appliquant la positivité de la densité de F au point x_0 et le lemme de Borel-Cantelli, $\{X_{i_n}\}_{n=1}^{\infty}$ converge vers x_0 presque sûrement. D'où

$$k^+(X^{(n)}) \geq \mathcal{D}(X_{i_n}, X^{(n)}) = \mathcal{D}_n(X_{i_n}) \geq \mathcal{D}(X_{i_n}) - \mu_H(F, F_n).$$

Du fait que F est absolument continue, on peut appliquer la proposition 3 pour déduire que

$$\mathcal{D}(X_{i_n}) \to \mathcal{D}(x_0) \quad p.s.$$

L'assertion $k^+ \to 1/2$ découle de la propriété de Glivenko-Cantelli (3.6) et du fait que $k^+ \leq k^*$.

3.3 Aspect algorithmique

La section précédente, traite l'aspect théorique de la profondeur de Tukey. On en a ainsi rassemblé les principales propriétés établies par Donoho and Gasko [25] et Massé, J.-C. [55]. Le lecteur peut sans doute remarquer la richesse de cette nouvelle mesure de profondeur et de ce qu'elle peut apporter au domaine des statistiques multivariées. A présent, on s'intéresse à l'aspect algorithmique de la profondeur de Tukey. Rousseeuw, P. J. and Ruts, I.(voir [66]) ont développé un algorithme robuste pour le calcul de la profondeur de Tukey d'un point arbitraire de $\mathbb{R}^2$ par rapport à un nuage de points bi-dimensionnels, en se basant sur quelques propriétés géométriques combinées avec des mécanismes de tris et de mise à jour. Ce dernier a une complexité temporelle d'ordre $O(n \log n)$. Suivant la même philosophie, Rousseeuw, P.J.

and Struyf, A. (voir [67]) ont introduit un algorithme exact dans le cas de la dimension 3, il nécessite une complexité temporelle d'ordre $O(n^2 \log n)$. Ainsi, pour une dimension $p > 3$, ils ont remarqué qu'un algorithme similaire nécessitera une complexité temporelle d'ordre $O(n^{p-1} \log n)$, ce qui devient trop long pour un p ou n large. Cependant, ces auteurs ont développé un algorithme approximatif très rapide, pour une dimension p quelconque. Ce dernier donne des résultats très compétitifs avec ceux de l'algorithme exact.

Le but de cette section est d'expliciter ces trois algorithmes. Dans un premier temps on s'attardera sur le cas bidimensionnel. Par la suite on parlera de son extension au cas de la dimension 3. Enfin, on décrira l'algorithme approximatif dans le cas d'une dimension quelconque.

3.3.1 Cas bivarié

Soient $X = \{x_1, ..., x_n\}$ un nuage de points bivariées et θ un point arbitraire de $\mathbb{R}^2$. On se propose de calculer la profondeur de Tukey du point θ par rapport à l'ensemble X.

La première étape consiste à calculer l'angle α_i pour chaque vecteur $u_i = (x_i - \theta)/\|x_i - \theta\|$, $i = 1, .., n$, comme suit :

$$\begin{aligned} &\text{Si } u_{i_1} > u_{i_2} \geq 0 \text{ alors } \alpha_i = \arcsin(u_{i_2}). \\ &\text{Si } u_{i_2} > u_{i_1} \geq 0 \text{ alors } \alpha_i = \arccos(u_{i_1}). \end{aligned}$$

où u_{i_1} et u_{i_1} sont les coordonnées de u_i.

Les α_i ainsi calculés, sont stockés dans un tableau α. Par la suite, les coordonnées des points x_i sont ignorées et l'algorithme se base sur le tableau α.

Dans la deuxième étape, on ordonne le tableau α et on cherche par la suite la plus grande différence entre deux composantes consécutives de α. Si cette différence est supérieure à π, on dit que θ se trouve à l'extérieur du nuage X et par conséquent $\mathcal{D}(\theta, X) = 0$. Sinon (la différence est inférieure à π), on soustrait α_1 (le plus petit α_i) de chaque α_i, ce qui correspond à une rotation des données initiales d'angle θ. Ceci n'a pas d'influence sur la profondeur de Tukey à cause de l'invariance affine. Ainsi, par construction, on a :

$$0 = \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_n < 2\pi.$$

Par la suite, on cherche le plus grand indice (qu'on note par n_u) tel que $\alpha_{n_u} < \pi$. Il est facile de voir que n_u correspond au nombre des α_i dans le demi-cercle supérieur, incluant l'extrémité 0 et excluant π.

La dernière étape consiste à calculer le tableau F définit par :

$$F(i) = \#\{j, 0 \leq \alpha_j < \alpha_i + \pi\} \tag{3.9}$$

où $\alpha_i + \pi$ peut s'envelopper autour de 2π au besoin. Donc $F(i) \leq 2n$.

Une façon élégante pour calculer le tableau F est de considérer l'angle antipodal β_i pour chaque α_i, $i = 1, ..., n$,

$$\begin{aligned} \beta_i &= \alpha_i + \pi \text{ si } 0 \leq \alpha_i \leq \pi \\ &= \alpha_i - \pi \text{ si } \pi \leq \alpha_i \leq 2\pi \end{aligned}$$

on range ensuite les α_i et les β_i dans un même tableau γ de longueur $2n$ qu'on va trier par la suite. La figure 3.2 illustre cette opération pour $n = 5$.

$$\gamma = (\alpha_1, ..., \alpha_n, \beta_1, ..., \beta_n).$$

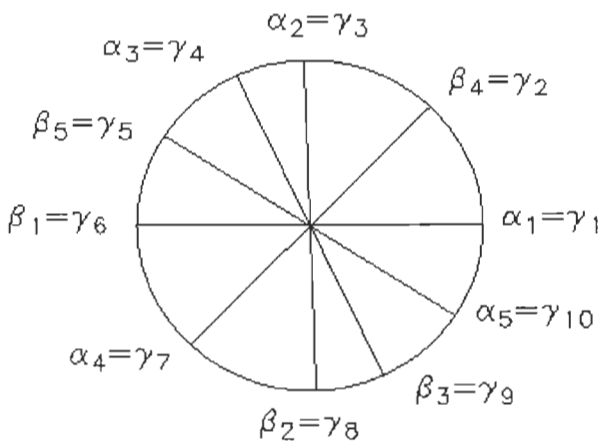

FIG. 3.2 – Illustration du calcul de $F(i)$

Initialement ($i = 1$), on sait que $F(1) = n_u$ ($= 3$ dans l'exemple de la figure 3.2) qui est le nombre des α_i qui viennent avant β_1 ($= \gamma_6$ pour $n = 5$). Pour calculer les autres éléments du tableau F, on vérifie les angles qui restent ($\gamma_7, ..., \gamma_{10}, \gamma_1, ..., \gamma_5$ pour $n = 5$). Pour chaque α on introduit un compteur NF qu'on initialise à n_u, pour chaque β on incrémente i et par la suite on affecte NF à $F(i)$. Dans l'exemple de la figure 3.2, γ_7 est un α, alors $NF \leftarrow NF + 1 = 4$, et parce que γ_8 est un β on incrémente i (*i.e.* $i \leftarrow i + 1 = 2$), par conséquent $F(2) = 4$, et ainsi de suite.

En utilisant la définition 1, on peut vérifier facilement que la profondeur de Tukey d'un point $\theta \in \mathbb{R}^2$ par rapport à un nuage de points bivariés X peut s'exprimer de la façon suivante :

$$\mathcal{D}(\theta, X) = \frac{1}{n} \min_i \min(k_i, n - k_i) \tag{3.10}$$

où

$$k_i = F(i) - G(i) \tag{3.11}$$

avec $G(i) = \#\{j; 0 \leq \alpha_j < \alpha_i\}$. Ce qui achève l'algorithme.

3.3.2 Cas de la dimension 3

Pour la dimension 3, la profondeur de Tukey d'un point $\theta \in \mathbb{R}^3$ peut être calculée en comptant le nombre des observations des deux côtés de chaque plan qui passe par θ. Naturellement, cette définition ne peut être implémentée comme telle. Il est suffisant par contre de considérer un nombre fini de plans qui passent par θ.

L'idée derrière l'algorithme qu'on va présenter est la suivante. Notons par L la droite liant θ avec un point x_i. Le plan η contenant L est tourné autour de L dans des étapes discrètes, comme le montre la figure 3.3.a. Une fois il dépasse le point x_j qui n'appartient pas à l'axe de rotation L, on compte les points de ses deux côtés, comme l'indique la définition de la profondeur de Tukey. Pour s'ajuster aux observations qui se trouve sur L, on considère pour chaque plan η deux inclinaisons (notées η_1 et η_2 dans la figure 3.3.b). Ainsi, toutes les observations qui se trouvent sur L du même côté que x_i sont situées à gauche de η_1, tandis que celles dans le côté opposé par rapport à θ sont situées à droite de η_1. Pour η_2 on interchange le rôle des côtés droit et gauche. Les observations qui n'appartiennent pas à L gardent la même position pour les deux plans η_1 et η_2. En faisant varier i entre 1 et n, on obtient toutes les positions possibles des plans passant par θ dans le nuage

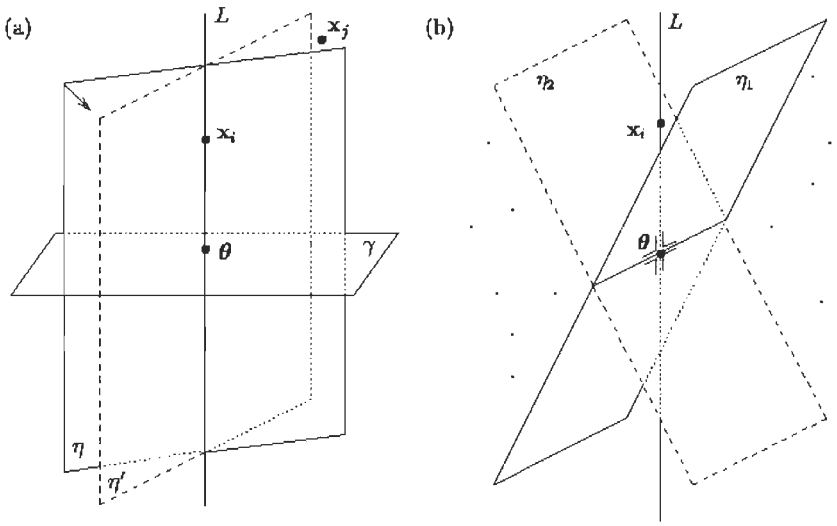

FIG. 3.3 – La profondeur de Tukey du point θ peut être calculée en considérant seulement un nombre fini de plans : **(a)** pour chaque point x_i, le plan η est tourné autour de L. **(b)** pour chaque plan η, on considère deux inclinaisons.

de points $X = \{x_1, ..., x_n\}$.

Ainsi, au lieu de compter le nombre des observations à droite et à gauche de chaque plan η pour une ligne particulière L (ceci demande un nombre d'opérations d'ordre $O(n^2)$), on projette toutes les observations sur le plan γ qui passe par θ et perpendiculaire à L (voir figure 2a). Ensuite, on fait appel à l'algorithme de calcul de la profondeur de Tukey dans le cas bivarié qu'on a explicité dans le section 3.3.1, pour compter le nombre de points dans les deux côtés d'un ensemble fini de lignes qui passent par θ.

Le pseudo-code suivant résume les étapes de calcul de la profondeur de Tukey d'un point $\theta \in \mathbb{R}^3$ par rapport à un nuage de points $X = \{x_1, ..., x_n\}$ de dimension 3.

1. Centrer toutes les données autour de θ , *i.e.* poser $x_i \leftarrow x_i - \theta$ ensuite $\theta \leftarrow 0$.
2. Poser $\mathcal{D}(\theta, X) \leftarrow n$.
3. Pour chaque $x_i \neq \theta$, notons par L la droite qui passe par θ et x_i et :
   - **a.** Projeter tous les points sur le plan η qui passe par θ et perpendiculaire à L.
   - **b.** Classifier les points dont la projection coïncide avec θ en tenant compte de sa position par rapport au plan η.
   - **c.** Utiliser l'algorithme bivarié pour trouver le plus petit nombre k de points dans un seul côté de chaque plan qui passe par θ, en prenant en considération quelques inclinaisons.
   - **d.** Mettre $\mathcal{D}(\theta, X) \leftarrow \frac{1}{n} \min(\mathcal{D}(\theta, X), k)$.

Cet algorithme a une complexité temporelle d'ordre $O(n^2 \log n)$. Il performe même si le nuage de points X est en position générale (voir définition 6).

Théorème 1 *L'algorithme ci-dessus est exact.*

Démonstration:
On doit montrer que chaque plan η qui passe par θ est soit considéré par l'algorithme soit il n'atteint pas le minimum dans (2). Pour cette fin on passe par deux étapes :

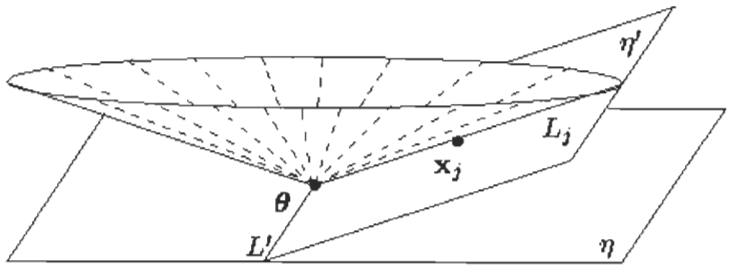

FIG. 3.4 – Visualisation des plans passant par θ et considérés pour la démonstration du théorème 1.

Etape I : Notons qu'on ne doit pas considérer les plans qui contiennent des points de X différents de θ. En effet, si un plan contient un point x_k, alors on considère la droite $\tilde{L} \subset \eta$ tel que θ appartient à $\tilde{L}$ et $\tilde{L} \perp [x_k, \theta]$. Ensuite, on tourne le plan η autour de $\tilde{L}$ d'un angle très petit de telle sorte que le plan résultant η' ne contient aucun point de l'ensemble X (si par hasard un point se trouve sur $\tilde{L}$, on modifie L légèrement). Ainsi, on remarque que le plan η' passe lui aussi par θ, mais il donne une valeur inférieure dans (2) que le plan η. Ce qui prouve qu'on peut supposer que le plan η ne contient aucun point de X différent de θ.

Etape II : On note par L_j la droite qui passe par θ et un point $x_j \in X$ pour chaque j, on cherche ensuite le point x_j dont lequel la valeur absolue de l'angle entre L_j et le plan η est minimale. Soit la droite $L' \subset \eta$ la perpendiculaire à L_j. On tourne η autour de L' de telle sorte que le plan résultant η' passe par le point x_j. Ainsi, η' est le plan tangent du cône de tous les points x tels que la droite $[x, \theta]$ forme le même angle avec η que la droite L_j (comme le montre la figure 3.4). Cependant, par construction, il n'y a pas de points entre les plans η et η'. En plus, les seuls points de X qui appartiennent à η' sont ceux appartenant à L_j. Ainsi, on sait que les plans qui ont la même position entre les points x_i (comme η) sont parcourus par l'algorithme, comme une inclinaison de η' durant la rotation autour de L_j

3.3.3 Approximation dans le cas d'une dimension quelconque

L'algorithme précédent permet de calculer la profondeur de Tukey exacte d'un point $\theta \in \mathbb{R}^3$. Ce dernier peut être généralisé dans le cas d'une dimension $p > 3$. La complexité temporelle va donc être de l'ordre $O(n^{p-1} \log n)$,

ce qui devient trop long pour un grand n ou p. Cependant, Rousseeuw, P.J. and Struyf, A. [67] ont développé un algorithme pour approximer la profondeur de Tukey dans de telles situations. L'idée de base est de prendre un nombre fini m de directions u (voir Définition 1). La question qui se pose est comment choisir ces m directions ? Rousseeuw, P.J. and Struyf, A. ont essayé plusieurs possibilités. Ils les choisissent aléatoirement parmi :

1. Toutes les directions qui lient les point θ et x_i pour tous i.
2. Toutes les directions qui lient deux points x_i et x_j.
3. Toutes les directions perpendiculaires aux hyperplans qui contiennent θ et $p-1$ points de X.
4. Toutes les directions perpendiculaires aux hyperplans contenant p points de X.

En se basant sur l'étude de plusieurs exemples (voir [67]), Rousseeuw, P.J. and Struyf, A. ont montré que la quatrième méthode donne les meilleures résultats. En plus, si l'ensemble X est en position générale, la profondeur de Tukey exacte peut être calculée en considérant toutes les C_n^p directions. Le pseudo-code suivant décrit l'algorithme proposé. Son implémentation nécessite une complexité temporelle d'ordre $O(mp^3 + mpn)$.

1. Mettre $\mathcal{D}(\theta, X) \leftarrow 1$.
2. Itérer m fois :
   a. Tirer un échantillon aléatoire de taille p de l'ensemble X.
   b. Déterminer la direction u perpendiculaire à cet échantillon.
   c. Projeter tous les points de X sur la droite L de direction u et qui passe par θ.
   d. Calculer la profondeur univariée k de θ sur L.
   e. Mettre $\mathcal{D}(\theta, X) \leftarrow \min(\mathcal{D}(\theta, X), k)$.

3.4 Conclusion

Dans ce chapitre, on a parlé de la profondeur de Tukey qui constitue la principale composante des cartes de contrôles qu'on va présenter au chapitre suivant. Dans un premier temps, on a décrit les principales propriétés étudiées par Donoho and Gasko (voir [24, 25]) et J-C. Massé (voir [54, 55]).

Dans un second temps, on a traité l'aspect algorithmique de cette profondeur. Pour ce, on s'est basé sur les travaux de Rousseeuw (voir [66, 67]). Le code en Fortran des trois algorithmes qu'on a présenté est disponible au site suivant : *http : //win-www.uia.ac.be/u/statis/index.html.* Pour notre application, on a implémenté ces trois algorithmes en **C++**.

Il existe toutefois d'autres mesures de profondeur qui sont également riches en propriétés. On cite à titre d'exemple, la profondeur du simplexe, la profondeur de Mahalanobis et celle de la majorité. Regina Y. Liu (voir [44]) a utilisé les profondeurs de simplexe et de Mahalanobis pour construire des cartes de contrôle multivariées. Indépendamment de la dimension des observations à contrôler, ces dernières prennent la forme d'un graphique bivarié facile à visualiser et interpréter. Elles possèdent la robustesse de détecter simultanément les déréglages ainsi que la croissance d'échelle du procédé qu'on veut contrôler. En plus, contrairement aux approches standards (comme la carte χ^2 et la carte Hotelling's T^2) leurs constructions est complètement non-paramétrique. Elles ne reposent pas sur l'hypothèse de normalité des observations. La difficulté majeure derrière ces cartes est d'un caractère algorithmique. Pour une dimension > 3 ou une taille d'échantillon élevée, le calcul des deux profondeurs utilisées par Liu devient très coûteux en matière de complexité temporelle. Dans le prochain chapitre, nous utilisons la profondeurs de Tukey pour la construction de ces cartes. Nous profiterons ainsi de la robustesse des algorithmes présentés dans ce chapitre pour le calcul de la profondeur de Tukey.

Chapitre 4

Contribution au contrôle de qualité multivarié (CQM[1])

4.1 Introduction

Le chapitre 2 nous a permis un survol des principales cartes de contrôle multivariées qu'on peut trouver dans la littérature. Toutefois, on a remarqué que ces dernières se basent sur l'hypothèse de normalité des observations. Ce qui n'est pas toujours le cas dans la vie réelle. L'idée derrière les cartes, que nous allons proposer dans ce chapitre, est de réduire chaque observation multivariée à un indice univarié. Pour ce, on utilise la profondeur de Tukey. Ainsi, en représentant les observations originales par leur profondeur de Tukey, on est capable de développer des cartes de contrôle suivant les mêmes principes que les cartes de contrôle univariées ($X, \bar{X}, CUSUM, ...$). Comme la profondeur de Tukey est basée sur la fonction de répartition empirique, notre approche est complètement non-paramétrique.

Ce chapitre est organisé de la façon suivante. La deuxième section consiste en une brève description des statistiques dérivées de la notion de profondeur. Ces dernières ont été introduites par Liu and Singh [43]. En 1995, Liu [44] les a employées pour construire des cartes de contrôle multivariées. Il a utilisé pour cette fin, la profondeur du simplexe et la profondeur de Mahalanobis. Il a toutefois montré que son approche fonctionne avec toute autre profondeur qui posséde la propriété de l'invariance affine. D'où l'idée d'utiliser la profondeur de Tukey dans notre approche. En plus de sa richesse en matière de propriétés mathématiques et statistiques, nous profitons aussi de la robus-

[1] **CQM** correspond à "Contrôle de qualité multivarié"

tesse des algorithmes développés par Rousseeuw [66, 67] et que nous avons décrit au chapitre 3 pour le calcul de cette dernière. Ce qui donne à notre méthode une élégance computationnelle remarquable. Dans la troisième section, nous proposons et nous justifions trois types de cartes de contrôle ; r, Q et S. Nous utilisons des données qui suivent une loi normale bivariée pour illustrer les cartes que nous proposons. La quatrième section donne une discussion détaillée des simulations proposées. La cinquième section consiste en une étude comparative de la carte Q proposée avec la carte Hotelling T^2. Enfin, la sixième et dernière section présente quelques remarques et conclusions sur le contenu du chapitre.

4.2 Quelques statistiques dérivées de la notion de profondeur

Comme nous l'avons indiqué, cette section consiste à décrire des statistiques basées sur la notion de profondeur. Ces dernières vont servir pour la construction des cartes de contrôle que nous allons détailler dans la suite de ce chapitre.

Soient F et G les distributions de deux populations indépendantes données dans $\mathbb{R}^p, p \geq 1$, et soient X et Y tels que, $X \sim F$ et $Y \sim G$. Ici la notation $X \sim F$ indique que la variable aléatoire X suit la distribution de probabilité F. Ainsi, On définit les deux statistiques R et Q suivantes :

$$R(G; y) = P\{\mathcal{D}_G(Y) \leq \mathcal{D}_G(y) \mid Y \sim G\} \tag{4.1}$$

et

$$\begin{aligned} Q(G, F) &= P\{\mathcal{D}_G(Y) \leq \mathcal{D}_G(X) \mid Y \sim G, X \sim F\} \\ &(= E_F[R(G; X)]). \end{aligned} \tag{4.2}$$

où y est un point arbitraire de $\mathbb{R}^p$. En terme de contrôle de qualité, G désigne la population sous contrôle (*i.e.* la population qui rencontre les spécifications de qualité considérées), et y représente une observation de la future population F.

Soient $Y_1, ..., Y_m$ et $X_1, X_2, ...$ des observations aléatoires issues de G et de F respectivement. Dans toute la suite de ce travail, les m observations $Y_1, ..., Y_m$ vont représenter l'échantillon de référence du procédé, et les

$X_1, X_2, ...$ les observations qu'on veut contrôler. Notons par $F_n(.)$ la distribution empirique de l'échantillon $\{X_1, ..., X_n\}$ et par $G_m(.)$ celle de $\{Y_1, ..., Y_m\}$. Dans ce cas, les équations 4.1 et 4.2 deviennent :

$$R(G_m; y) = \#\{Y_j \mid \mathcal{D}_{G_m}(Y_j) \leq \mathcal{D}_{G_m}(y), j = 1, ..., m\}/m. \tag{4.3}$$

$$Q(G, F_n) = \frac{1}{n} \sum_{i=1}^{n} R(G, X_i), \tag{4.4}$$

et

$$Q(G_m, F_n) = \frac{1}{n} \sum_{i=1}^{n} R(G_m, X_i). \tag{4.5}$$

Notons que l'intervalle des valeurs prises par Q est $[0, 1]$. Dans la littérature, la statistique Q porte souvent le nom d'indice de qualité. Ceci provient du fait qu'elle permet de comparer les dispersions de deux populations F et G. Sous l'hypothèse nulle $H_0 : F = G$, les valeurs $Q = 1/2$ et $Q < 1/2$, indiquent la présence d'un décalage d'emplacement ou une augmentation d'échelle de F par rapport à G. Si, par contre, $Q > 1/2$, alors G possède une dispersion minimale et peut-être le même (un décalage mineur dans le cas échéant) emplacement que F. En termes de contrôle de qualité, supposons que G représente la population des observations sous-contrôle d'un procédé manufacturier et F la population des futures observations du même procédé. La statistique Q semble être un outil attrayant pour voir si la population F répond aux mêmes spécifications que la population G.

En plus de son interprétation dans le contexte de contrôle de qualité, Q possède des propriétés mathématiques très intéressantes.

Proposition 11 *Si la profondeur $\mathcal{D}(G; \cdot)$ est à invariance affine, alors il en est de même pour $R(G; Y)$ et $Q(G; F)$; c'est à dire,*

$$R(G; Y) = R(G_{\mathbf{A},b}; \mathbf{A}Y + b)$$

et

$$Q(G; F) = Q(G_{\mathbf{A},b}; F_{\mathbf{A},b})$$

avec $F_{\mathbf{A},b}$ est la distribution de $\mathbf{A}X + b$, $G_{\mathbf{A},b}$ celle de $\mathbf{A}Y + b$, $\mathbf{A}$ est une matrice non-singulière $p \times p$ et b est un vecteur de $\mathbb{R}^p$.

Démonstration:
En effet, dire que la profondeur $\mathcal{D}_F(\cdot)$ est à invariance affine est équivalent à dire que pour toute matrice $(p \times p)$ non-singulière $\mathbf{A}$ et tout vecteur b de $\mathbb{R}^p$, on a :

$$\mathcal{D}_{G_{\mathbf{A},b}}(\mathbf{A}x + b) = \mathcal{D}_G(x) \quad \forall x \in \mathbb{R}^p$$

Ainsi, en remplaçant $\mathcal{D}_G(x)$ par $\mathcal{D}_{G_{\mathbf{A},b}}(\mathbf{A}x + b)$ dans les équations 4.1 et 4.2 on obtient :

$$\begin{aligned} R(G; y) &= P\{\mathcal{D}_G(Y) \leq \mathcal{D}_G(y) \mid Y \sim G\} \\ &= P\{\mathcal{D}_{G_{\mathbf{A},b}}(\mathbf{A}Y + b) \leq \mathcal{D}_{G_{\mathbf{A},b}}(\mathbf{A}y + b) \mid Y \sim G\} \\ &= R(G_{\mathbf{A},b}; \mathbf{A}Y + b). \end{aligned}$$

et

$$\begin{aligned} Q(G, F) &= P\{\mathcal{D}_G(Y) \leq \mathcal{D}_G(X) \mid Y \sim G, X \sim F\} \\ &= P\{\mathcal{D}_{G_{\mathbf{A},b}}(\mathbf{A}Y + b) \leq \mathcal{D}_{G_{\mathbf{A},b}}(\mathbf{A}X + b) \mid Y \sim G, X \sim F\} \\ &= Q(G_{\mathbf{A},b}; F_{\mathbf{A},b}). \end{aligned}$$

ce qui achève la démonstration.

Cette proposition montre que les valeurs de la statistique Q ne dépendent pas des échelles de mesures des populations F et G.

Proposition 12 *Supposons que* $\lim_{m\to\infty} \sup_{x\in\mathbb{R}^p} |\mathcal{D}_{G_m}(x) - \mathcal{D}_G(x)| = 0$, *presque sûrement et que la distribution de* $\mathcal{D}_G(Y)$ *est continue. Alors,* $Q(G_m, F_n) \to Q(G, F)$ *presque sûrement pour* $\min(m, n) \to \infty$.

Démonstration:
On définit,

$$R_\epsilon^+(G; x) = P_G\{X : \ \mathcal{D}_G(X) \leq \mathcal{D}_G(x) + \epsilon\}$$

et

$$R_\epsilon^-(G; x) = P_G\{X : \ \mathcal{D}_G(X) \leq \mathcal{D}_G(x) - \epsilon\}.$$

Notons que pour tout $\epsilon > 0$, et pour tous les grands m et n, on a,

$$\frac{1}{n}\sum_{i=1}^{n} R_\epsilon^-(G; X_i) \leq Q(G_m, F_n) \leq \frac{1}{n}\sum_{i=1}^{n} R_\epsilon^+(G; X_i)$$

presque sûrement. Ainsi, le résultat peut être obtenu en utilisant la loi forte des grands nombres et suivant les assertions suivantes. Pour $\epsilon \to 0$, on a,

$$E_F R_\epsilon^+(G;X) \to E_F R(G;X) = Q(G,F)$$

et

$$E_F R_\epsilon^-(G;X) \to E_F R(G;X) = Q(G,F),$$

les deux dernières assertions proviennent du théorème de la convergence monotone et de la condition que la distribution de $\mathcal{D}_G(X)$ est continue.

4.3 Les cartes de contrôle basées sur la mesure de profondeur

Dans cette section nous allons introduire quatre cartes de contrôle, la carte r, la carte Q, la carte S et la carte S^*. Les définitions de ces dernières sont basées sur les statistiques R et Q définies dans la section précédente. Pour justifier la performance de notre approche, nous présentons au fur et à mesure les graphiques obtenus par l'application des quatres cartes de contrôle sur un échantillon aléatoire.

4.3.1 La carte r

La carte r que nous allons présenter dans ce section est similaire à la carte X univariée. Elle est basée sur la statistique $R(\cdot;\cdot)$ définie dans les équations (4.1) et (4.3). Pour commencer, nous allons parler de la carte X univariée.

Supposons que les observations $Y_1, ..., Y_m$ et $X_1, ..., X_n$ sont univariées et que nous nous proposons de contrôler les X_i, $i = 1, .., n$. Il arrive que pour certains procédés industriels, on ne peut former de sous-groupe et que l'on doit se contenter d'une seule donnée. La carte X répond à ce genre de situations. Sa mise en œuvre repose sur l'hypothèse de normalité de la caractéristique soumise au contrôle. La figure 4.1 suivante montre un schéma typique de la carte X,

Dans cet exemple, $UCL = CL + z_{\alpha/2}\sigma$, $LCL = CL - z_{\alpha/2}\sigma$, et $CL = \mu$ si μ est connue et $= \bar{Y}$ sinon. z_α indique le percentile standard de la loi normale tel que $\alpha = P(Z > z_\alpha)$, où $Z \sim \mathcal{N}(0,1)$. μ représente la moyenne de la loi de distribution de l'échantillon $Y_1, ..., Y_m$.

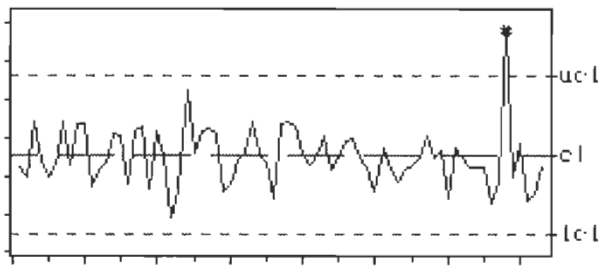

FIG. 4.1 – Exemple de la carte Shewhart univariée

La carte X est simple mais pertinente pour le contrôle des procédés univariés. Cependant elle ne se généralise pas facilement au cas multivarié. Dans le cas bivarié, Alt and smith [6] ont introduit la carte X bivariée avec un contour elliptique comme limites de contrôle (également appelé ellipse de contrôle). En plus de la restriction de normalité des observations, il est également difficile de visualiser et de détecter les déréglages du procédé qu'on veut contrôler. Ceci est une conséquence du fait qu'on ne peut conserver l'ordre chronologique des observations sur le schéma. En outre, quand la dimension p des observations dépasse 3, il ne semble pas possible de suivre la même idée de construire des cartes faciles à visualiser et à interpréter.

La carte r que nous proposons est construite comme suit. Nous calculons les valeurs $\{R(G; X_1), R(G; X_2), ...\}$ (ou $\{R(G_m; X_1), R(G_m; X_2), ...\}$ si on ne connaît pas la distribution G), Suivant la formule (4.1) (ou (4.3)). Ainsi, la carte r est le tracé des $R(G; X_i)$ (ou $R(G_m; X_i)$) en fonction du temps i, avec la ligne centrale $CL = .5$ et la limite de contrôle inférieure $LCL = \alpha$. Ici, rappelons que α est le taux de fausses alarmes. Il est souvent choisi proche de 0. Nous remarquons que la carte r ne contient qu'une seule limite de contrôle $LCL = \alpha$. La motivation et la justification de ce choix sont données dans la suite.

L'expression (4.3) montre que les $\{R(G_m; X_i)\}$ indiquent comment les X sont positionnés par rapport aux Y. Une très petite valeur de $R(G_m; X_i)$ montre que juste une petite proportion des Y sont positionnés dans la même périphérie que les X. Ceci se traduit par le fait que les X ne sont pas conformes à l'ensemble de données Y. Supposons que $X \sim F$. Une petite valeur de $R(G_m; \cdot)$ s'interprète par une possible déviation de G par rapport

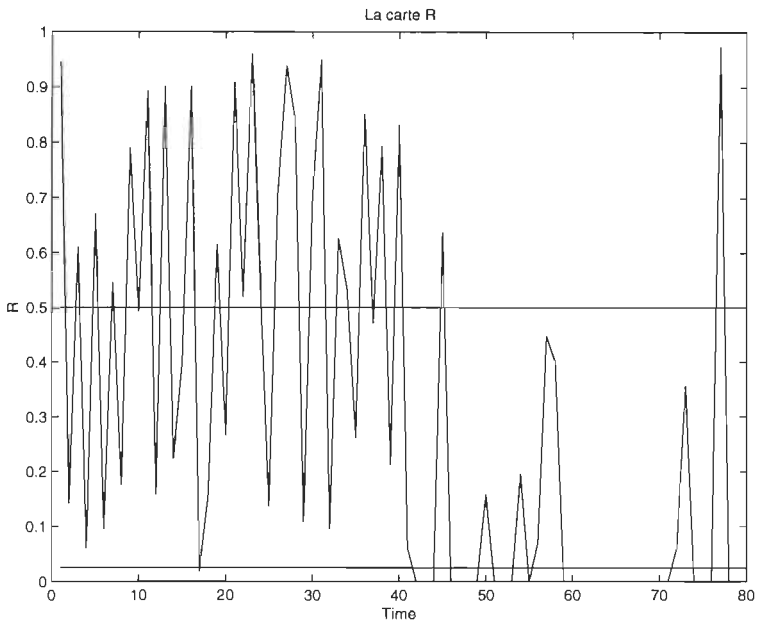

FIG. 4.2 – La carte r

à F. Comme la définition de $R(G_m; \cdot)$ est basée sur la notion de profondeur, une telle déviation peut être le résultat d'un hors-contrôle et/ou une croissance d'échelle. La justification détaillée de cette interprétation peut être dérivée du travail de Liu and Singh [43], sec. 3). D'après la proposition 13, la carte r avec $LCL = \alpha$ correspond au test statistique d'ordre α avec les hypothèses suivantes :

$$\begin{cases} H_0 : & F = G \\ & vs. \\ H_1 : & \text{Présence d'un changement ou d'une} \\ & \text{croissance d'échelle de } G \text{ par rapport à } F. \end{cases} \tag{4.6}$$

Nous remarquons que l'hypothèse alternative (H_1) présente une perte de précision. Ceci la rend particulièrement appropriée pour détecter la détérioration de la qualité durant le processus de contrôle. Ceci justifie également la déclaration du procédé hors-contrôle quand l'hypothèse H_0 est rejetée ou, d'une manière équivalente, quand une observation dépasse la limite α sur la carte r.

Maintenant, pour expliquer le choix des limites de contrôle de la carte r ($CL = .5$ et $LCL = \alpha$), on considère les propriétés de $R(G;X)$ et $R(G_m;X)$ établies par Liu and Singh [43] et que nous citons dans la proposition 13 suivante,

Proposition 13 *Supposons que $F = G$ et $X \sim F$. Notons par $U[0,1]$ la distribution uniforme sur l'intervalle $[0,1]$ et par $\to^{\mathcal{L}}$ la convergence en loi. Si $\mathcal{D}_G(X)$ possède une distribution continue, alors*

a. *$R(G;X) \sim U[0,1]$, et*

b. *pour $m \to \infty$, $R(G_m;X) \to^{\mathcal{L}} U[0,1]$ conditionnellement sur Y à condition que $\mathcal{D}_{G_m}(\cdot)$ vérifie :*

$$\lim_{m\to\infty} \sup_{x\in\mathbb{R}^p} |\mathcal{D}_{G_m}(x) - \mathcal{D}_G(x)| = 0 \tag{4.7}$$

sur presque toute la suite $\{Y_1, ..., Y_m\}$.

Démonstration:

a. Ce résultat découle facilement de la transformation des distributions de probabilité. En effet, posons

$$\mathcal{H}(x) = P(\mathcal{D}_G(X) \leq x)$$

alors

$$\begin{aligned} P(\mathcal{D}_G(X) \leq \mathcal{D}_G(x)) &= \mathcal{H}(\mathcal{D}_G(x)) \\ &= R(G,x) \end{aligned}$$

d'autre part,

$$\begin{aligned} P(R(G,X) \leq x) &= P(\mathcal{H}(\mathcal{D}_G(X) \leq x) \\ &= x \quad \text{car } \mathcal{H} \text{ est continue.} \end{aligned}$$

b. En effet, pour montrer ce résultat, il suffit de montrer que $R(G_m;x)$ converge vers $R(G;x)$ pour tout x fixé (en respectant G) sur presque toute la suite $\{Y_1, ..., Y_m\}$. Soit Y une suite qui vérifie la condition (4.7). Alors, pour un $\epsilon > 0$ donné, il existe un m_0 tel que, $\sup_{x\in\mathbb{R}^p} |\mathcal{D}_{G_m}(x) - \mathcal{D}_G(x)| < \epsilon/2$ pour tout $m > m_0$. Donc, pour tout $m > m_0$ on a,

$$\begin{aligned} \{Y : \mathcal{D}_G(Y) \geq \mathcal{D}_G(x) + \epsilon\} &\subseteq \{Y : \mathcal{D}_{G_m}(Y) \geq \mathcal{D}_{G_m}(x)\} \\ &\subseteq \{Y : \mathcal{D}_G(Y) \geq \mathcal{D}_G(x) - \epsilon\}. \end{aligned}$$

et le résultat découle en faisant tendre ϵ vers 0.

Sous l'hypothèse H_0, la proposition 13 implique que l'espérance de $R(G;X)$ est égale à .5 et celle de $R(G_m;X)$ est égale à .5 presque sûrement pour toute la suite $\{Y_1,...,Y_m\}$ pour $m \to \infty$. Ceci justifie le choix de CL pour la carte r. Quand $R(G;X)$ (ou $R(G_m;X)$) est largement plus petite que .5, on dit qu'il y a des doutes pour l'hypothèse H_0 et une évidence pour supporter H_1. Ceci signale la possibilité de détérioration de la qualité. Quand $R(G;X)$ (ou $R(G_m;X)$) est plus grande que .5, ceci se traduit par une décroissance d'échelle avec une possibilité de présence d'un décalage négligeable. Ceci est vu comme une amélioration de la qualité, (ou encore un gain dans la précision) et le procédé ne devrait pas être déclaré hors-contrôle. C'est la raison pour laquelle on trouve seulement une limite de contrôle inférieure (LCL) dans la carte r. La distribution uniforme de $R(G;X)$ (ou $R(G_m;X)$) implique clairement que LCL devrait être égale à α.

Quoique la carte r n'ait pas de limite supérieure (UCL) pour faire de CL la ligne centrale de la région de contrôle, le CL ici sert comme référence pour nous permettre d'observer si une configuration ou une tendance se développe dans une suite d'échantillons.

4.3.2 La carte Q

L'idée derrière la carte Q est similaire à celle derrière la carte $\bar{X}$ univariée. Supposons que $X_1, X_2, ...$ sont univariées et suivent une loi normale G. La carte $\bar{X}$ est un graphique sur lequel sont pointées les valeurs des moyennes des échantillons successifs prélevés du procédé de fabrication et reliées entre elles en préservant l'ordre chronologique de prélèvement. La carte $\bar{X}$ peut empêcher une fausse alarme quand le procédé est réellement en contrôle mais quelques observations individuelles se trouvent en dehors des limites de contrôle simplement à cause des fluctuations aléatoires. Ceci est un avantage par rapport à la carte X standard.

Dans le cas multivarié, nous traçons les moyennes des sous-groupes des $R(G;X_i)$ (ou $R(G_m;X_i)$). Supposons qu'on choisit des sous-groupes de taille n. D'après les formules (4.4) et (4.5), les moyennes des $R(G;X_i)$ et $R(G_m;X_i)$ sont données par $Q(G,F_n^j)$ et $Q(G_m,F_n^j)$ respectivement, avec F_n^j étant la distribution empirique des X_i du $j^{\text{ième}}$ sous-groupe, $j = 1,2,....$ Ainsi, sur la carte Q on trace,

$$\{Q(G,F_n^1), Q(G,F_n^2), ...\}$$

ou

$$\{Q(G_m, F_n^1), Q(G_m, F_n^2), ...\}$$

si on ne connaît pas la distribution des $Y_1, ..., Y_m$.

À cette étape, la question est de trouver les valeurs adéquates pour la ligne centrale CL et la limite de contrôle inférieure LCL pour la carte Q. Ceci dépend du choix de n. En effet, quand n est grand, la proposition 14 montre que CL devrait être égale à .5, tandis que LCL devrait être égale à $(.5 - z_\alpha(12n)^{-1/2})$ pour tracer les $\{Q(G, F_n^j)\}$ et $.5 - z_\alpha\sqrt{\frac{1}{12}[(1/m) + (1/n)]}$ pour tracer les $\{Q(G_m, F_n^j)\}$. Cette approximation semble être tout à fait raisonnable même lorsque n est plus petit ($n = 3$ ou 4). Dans ce cas, nous pouvons utiliser les distributions pour $Q(G, F_n)$ données dans la proposition 15. Il découle de cette proposition que pour une petite valeur de α, la carte Q doit avoir $CL = .5$ et $LCL = (n!\alpha)^{1/n}/n$.

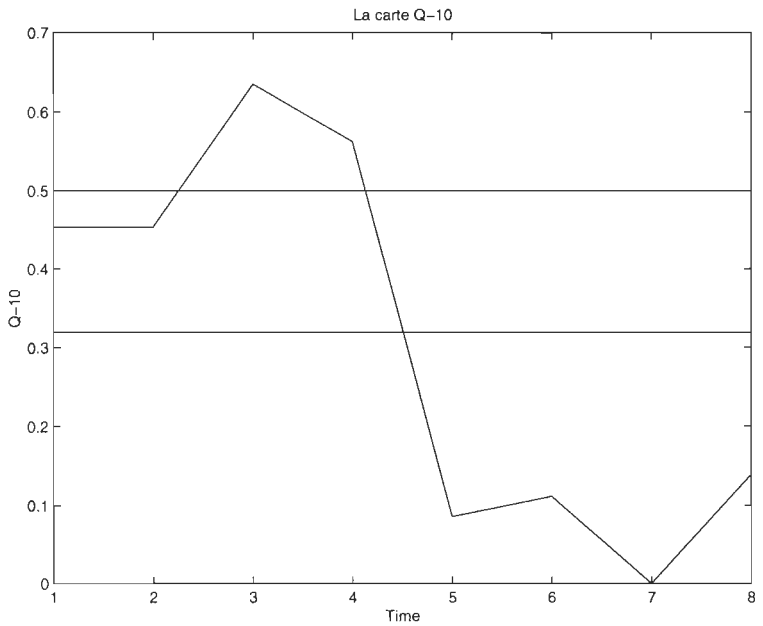

FIG. 4.3 – La carte Q avec $n = 10$

Dans un premier temps, nous allons décrire le cas d'un grand n. La carte Q correspond à un test statistique de niveau α basé sur les $Q(G, F_n)$ (ou

$Q(G_m, F_n)$) pour tester le même ensemble d'hypothèses dans (4.6). Il existe actuellement deux tests multivariés étudiés par Liu [42] et Liu and Singh [44]. La proposition 14 suivante résume les propriétés asymptotiques principales de ces derniers.

Proposition 14 *Supposons que les conditions de la proposition 13 sont vérifiées. Alors*

a. *pour* $n \to \infty$, $\sqrt{n}[Q(G, F_n) - \frac{1}{2}] \to^{\mathcal{L}} \mathcal{N}(0, 1/12)$ *; et*

b. *pour* $\min(m, n) \to \infty$, $[(\frac{1}{m} + \frac{1}{n})\frac{1}{12}]^{-1/2}[Q(G_m, F_n) - \frac{1}{2}] \to^{\mathcal{L}} \mathcal{N}(0, 1)$.

La première partie (a) de la proposition découle directement du théorème central limite. Ceci provient du fait que $Q(G, F_n)$ n'est rien d'autre que la moyenne de n variables aléatoires d'une loi uniforme *i.i.d.* La partie (b) a été établie par Liu and Singh [44] pour toutes les profondeurs qui vérifient, sous certaines conditions, la consistance uniforme (équation 4.7). Dans le cas de la profondeur de Tukey, on a la consistance uniforme si la distribution G de la fonction de profondeur est absolument continue. Ainsi, le choix des limites de contrôles CL et LCL dans le cas d'un grand n est clair d'après la proposition 14.

Quand il s'agit d'un n petit, les résultats antérieurs peuvent ne pas être applicables. Puisque LCL dans ce cas est le quantile d'ordre α de la distribution de $Q(G, F_n)$, nous avons besoin de la distribution de la moyenne des variables aléatoires uniformes (proposition 13). Ceci découle directement de la formule pour la distribution de la somme de variables aléatoires uniformes fournies dans la proposition 15.

Proposition 15 *Soit* $\{U_1, ..., U_n\}$ *un échantillon i.i.d provenant d'une loi uniforme* $U[0, 1]$, *et notons par* $H_n(t)$ *la distribution de* $\sum_{i=1}^{n-1} U_i$, *c'est à dire* $H_n(t) = P\{\sum_{i=1}^{n} U_i \leq t\}$. *Alors pour chaque* $n = 1, 2, ...$, $H_n(t) = 0$ *pour* $t \leq 0$ *et*

$$H_n(t) = \frac{1}{n!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (t - k)_+^n, \tag{4.8}$$

ou

$$\begin{aligned} (x)_+^n &= 0, \ si\ x \leq 0; \\ &= x^n \ si\ x > 0. \end{aligned}$$

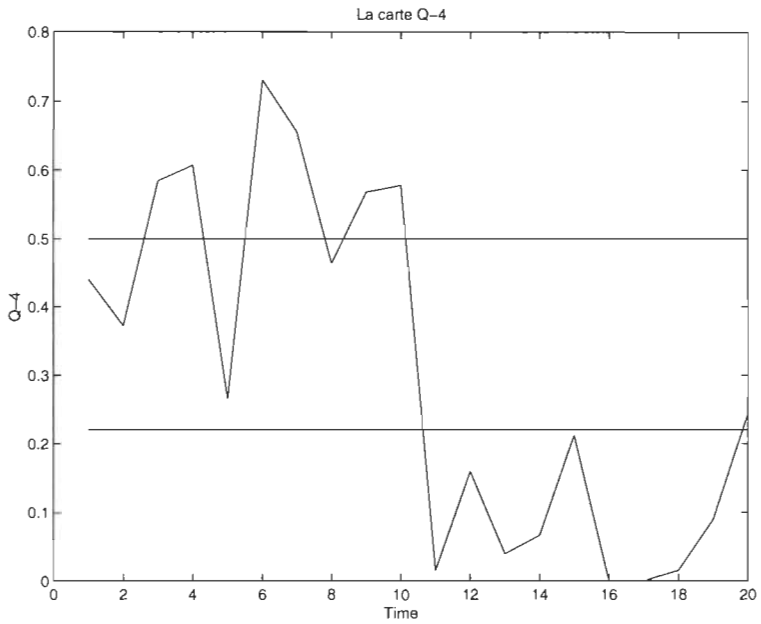

FIG. 4.4 – La carte Q avec $n = 4$

Cette dernière formule a été obtenue par Feller [28]. L'expression (4.8) montre que $H_n(\cdot)$ est un polynôme par morceaux. Pour notre propos, les parties les plus appropriées du polynôme sont,

$$\begin{aligned} H_n(t) &= \frac{1}{n!}t^n, \text{ si } 0 \leq t < 1; \\ &= \frac{1}{n!}(t^n - n(t-1)^n), \text{ si } 1 \leq t < 2; \\ &= \frac{1}{n!}(t^n - n(t-1)^n + \frac{n(n-1)}{2}(t-2)^n), \text{ si } 2 \leq t < 3. \quad (4.9) \end{aligned}$$

Pour déterminer LCL de la carte Q dans le cas d'un petit n, nous avons besoin de trouver la valeur w_α telle que $P(1/n\sum_{i=1}^{n} U_i \leq w_\alpha) \equiv \alpha$, ou d'une manière équivalente, $H_n(nw_\alpha) = \alpha$. La formule (4.9) implique que pour $\alpha \leq 1/n!$, $(n!w_\alpha)^n/n! = \alpha$. Par conséquent, $w_\alpha = (n!\alpha)^{1/n}/n$. Ce qui justifie le choix de LCL pour la carte Q. Par exemple pour $n = 4$ et $\alpha = .025$, $W_{.025} = [24(.025)]^{1/4}/4 = .220$. Cette dernière valeur est utilisée comme LCL pour la carte Q présentée dans la figure 4.4, où les X_i sont

partagées en groupes de 4. Il est également clair que CL ici devrait être .5, parce que c'est l'espérance des moyennes de n variables aléatoires uniformes $U[0,1]$ *i.i.d.* Notons que dans des situations pratiques dans le contrôle de qualité, α est habituellement choisi inférieur ou égal à .0027. Ceci est une convention établie par les spécialistes du domaine afin d'avoir un nombre convenable de fausses alarmes. Ainsi, pour un n inférieur ou égal à 4, LCL est donnée par $(n!\alpha)^{1/n}/n$ comme nous l'avons montré plus haut. Cependant, si pour une raison ou une autre, α est choisi supérieur à $1/n!$, alors la formule par morceaux appropriée dans (4.9) devrait être utilisée pour déterminer la valeur de w_α. Par exemple, prenons $n = 4$ et $\alpha = 0.1$. Nous devrions résoudre l'équation $1/4!((4w_{.1})^4 - 4((4w_{.1}) - 1)^4) = .1$. La solution est unique parce que $H_n(\cdot)$ est une fonction strictement croissante. En général, il n'y a aucune méthode explicite pour la résolution des équations polynômiales de degré élevé. Cependant les solutions peuvent être facilement obtenues en utilisant la méthode de Newton.

4.3.3 La carte S

La motivation de la carte S est similaire à celle de la carte $CUSUM$ univariée. Quand les X_i sont univariées, la carte $CUSUM$ la plus simple qu'on peut trouver est le tracé de $\sum_{i=1}^{n}(X_i - \mu)$. Cette dernière reflète la configuration de toutes les déviations des observations par rapport à la valeur moyenne. Ce schéma est plus pertinent que la carte X et la carte $\bar{X}$ pour détecter les petits changements dans un procédé et est peut-être le plus utilisé en pratique.

Dans le cas multivarié, l'idée de la carte $CUSUM$ consiste simplement à tracer les valeur $S_n(G)$ et $S_n(G_m)$ définies par :

$$\begin{aligned} S_n(G) &= \sum_{i=1}^{n}\left[R(G;X_i) - \frac{1}{2}\right] \\ &= n[Q(G,F_n) - \frac{1}{2}] \end{aligned} \tag{4.10}$$

et

$$\begin{aligned} S_n(G_m) &= \sum_{i=1}^{n}\left[R(G_m;X_i) - \frac{1}{2}\right] \\ &= n[Q(G_m,F_n) - \frac{1}{2}] \end{aligned} \tag{4.11}$$

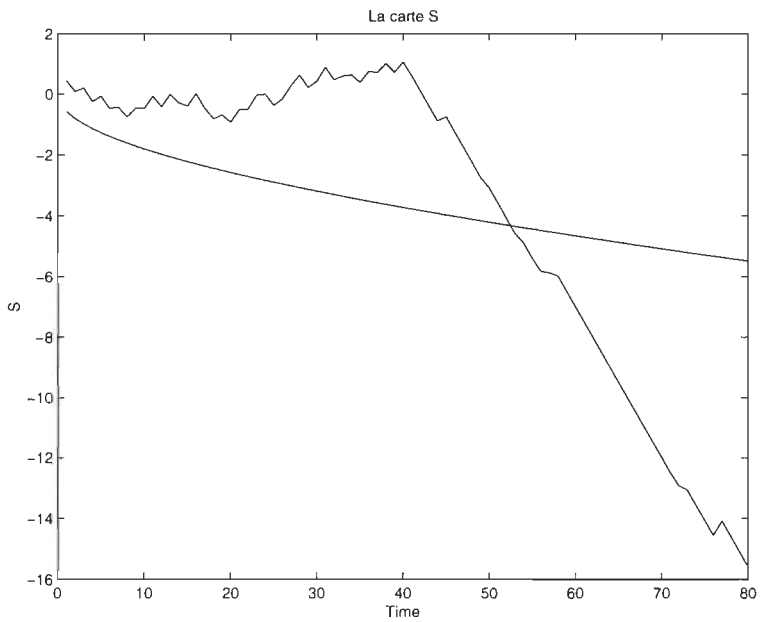

FIG. 4.5 – La carte S

Proposition 16 *Sous les conditions de la proposition 14, nous avons*

a. *pour* $n \to \infty$, $(n/12)^{-1/2} S_n(G) \to^{\mathcal{L}} \mathcal{N}(0,1)$ *; et*

b. *pour* $\min(m,n) \to \infty$, $[(\frac{1}{m} + \frac{1}{n})\frac{n^2}{12}]^{-1/2} S_n(G_m) \to^{\mathcal{L}} \mathcal{N}(0,1)$.

La proposition 16 implique que LCL pour la carte S basée sur les $S_n(G)$ est égale à $-(z_\alpha(n/12)^{1/2})$ et la LCL pour la carte S basée sur les $S_n(G_m)$ est égale à $-\{z_\alpha\sqrt{n^2[(1/m)+(1/n)]/12}\}$. Nous remarquons que la limite de contrôle dans ce cas est une courbe au lieu d'une ligne comme le montre la figure 4.5. En fait, la courbe limite de contrôle tend inférieurement vers $\sqrt{n}$. Quand n est grand, la carte S peut facilement excéder le format standard du papier, ce qui n'est pas pratique. Cependant, il est préférable de normaliser toutes les sommes cumulatives ($CUSUM$) pour avoir une ligne droite comme limite de contrôle (voir figure 4.6). Ceci revient à tracer $S_n^*(G) = S_n(G)/\sqrt{n/12}$ ou $S_n^*(G_m) = S_n(G_m)/\sqrt{n^2[(1/m)+(1/n)]/12}$ pour $n = 1, 2, ...$ La carte S^* a une $CL = 0$ et $LCL = -z_\alpha$.

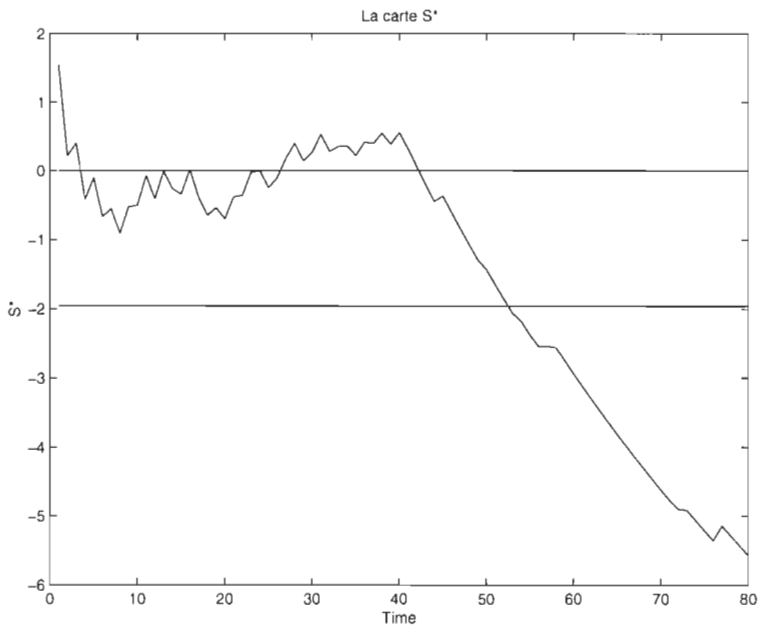

FIG. 4.6 – La carte S^*

4.4 Résultats de simulations

Dans cette section, nous mettons en œuvre les quatres cartes de contrôle présentées dans la section précédente. Pour cette fin, nous utilisons une base de 580 observations bivariées obtenues de la façon suivante :

Soit G telle que, $G \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$. Nous générons un échantillon de 540 points suivant G. Nous notons les 500 premières observations par $Y_1, ..., Y_{500}$ (c'est les observations qui vont servir comme modèle sous-contrôle) et les 40 dernières par $X_1, ..., X_{40}$. Nous générons également un échantillon de 40 points suivant la distribution $F \sim \mathcal{N}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$ et nous les notons par $X_{41}, ..., X_{80}$. ($X_1, X_2, ..., X_{80}$ représentent l'échantillon qui va être soumis au contrôle)

Ici, nous avons choisi des distributions normales dans le but de faciliter l'évaluation des résultats. La normalité n'est pas exigée pour l'applicabilité de nos cartes. Notez qu'il y a un net décalage dans la moyenne et une croissance

i	$\mathcal{D}(\mathbf{X_i})$	$\mathbf{R_m(X_i)}$	i	$\mathcal{D}(\mathbf{X_i})$	$\mathbf{R_m(X_i)}$
1	.35	.95	41	.01	.06
2	.02	.14	42	0	0
3	.16	.61	43	0	0
4	.01	.06	44	0	0
5	.19	.67	45	.17	.64
6	.01	.10	46	0	0
7	.13	.55	47	0	0
8	.03	.18	48	0	0
9	.24	.79	49	0	0
10	.12	.49	50	.02	.16
11	.31	.89	51	0	0
12	.02	.16	52	0	0
13	.32	.90	53	0	0
14	.04	.22	54	.03	.20
15	.09	.40	55	0	0
16	.32	.90	56	.01	.07
17	0	.02	57	.10	.45
18	.02	.16	58	.09	.40
19	.17	.62	59	0	0
20	.05	.27	60	0	0
21	.33	.91	61	0	0
22	.13	.52	62	0	0
23	.37	.96	63	0	0
24	.13	.53	64	0	0
25	.02	.14	65	0	0
26	.2	.71	66	0	0
27	.35	.94	67	0	0
28	.28	.84	68	0	0
29	.02	.11	69	0	0
30	.20	.70	70	0	0
31	.35	.95	71	0	0
32	.01	.10	72	.01	.06
33	.17	.63	73	.07	.36
34	.13	.53	74	0	0
35	.05	.26	75	0	0
36	.28	.85	76	0	0
37	.11	.47	77	.38	.97
38	.24	.79	78	0	0
39	.04	.21	79	0	0
40	.27	.83	80	0	0

TAB. 4.1 – Les valeurs de la profondeur de Tukey et de la statistique R_m

d'échelle dans la distribution pour les 40 derniers X_i. En principe, nous devrons nous attendre à ce que toutes nos cartes détectent ce changement. Ce qui est confirmé par les figures 4.2 à 4.6.

Pour chaque X_i, on calcule sa profondeur de Tukey, en utilisant l'algorithme développé par Rousseeuw et Ruts(1992) et que nous avons implémenté en **VC++ 6.0**. Les valeurs de la profondeur de Tukey des X_i sont stockées dans la deuxième colonne de la table 4.1. En se basant sur ces valeurs, nous pouvons calculer toutes les $R(G_m; X_i)$ en se basant sur la formule 4.3. Les $R(G_m; X_i)$ sont ainsi calculées, nous les stockons dans la troisième colonne de la table 4.1. La figure 4.2 donne le graphique des $R(G_m; X_i)$ avec $CL = .5$ et $LCL = .025$. La valeur .025 correspond à la valeur α que nous avons choisie pour les 5 cartes. La carte R ainsi tracée, elle montre clairement que le procédé est hors contrôle pour la deuxième moitié des X_i. Une partie majeure des $R_{G_m}(X_i)$ tombe au-dessous de LCL. Les quelques fausses alarmes dans la première moitié des X_i sont attribuées aux fluctuations aléatoires. Elle peuvent être caractérisées de la même manière que celle dans la carte X univarié.

i	$\mathbf{Q(G, F^i_{10})}$
1	.32
2	.45
3	.45
4	.64
5	.56
6	.09
7	.11
8	.13

TAB. 4.2 – Les valeurs de la statistique $Q(G_m, F^i_{10})$

Les figures 4.4 et 4.3 représentent la carte Q avec des sous-groupes de tailles $n = 4$ et $n = 10$ respectivement. Les valeurs $\{Q(G_m, F^j_n), j = 1, 2, ...\}$ sont calculées suivant la définition 4.5 et sont stockées dans les tables 4.3 et 4.2 respectivement. Pour la figure 4.4, la limite CL est égale à .5 et la limite $LCL = .22$ d'après la proposition 15. Dans la figure 4.3, le résultat de la proposition 14 prouve le choix de $CL = .5$ et $LCL = \{\frac{1}{2} - z_\alpha\sqrt{1/12[(1/m) + (1/n)]}\}$, qui est égale à .32 quand $\alpha = .025$. Les deux graphiques (figures 4.4 et 4.3) montrent que le procédé est hors-contrôle dans la

deuxième moitié. Nous observons également que le calcul des moyennes des valeurs de la statistique R dans Q, a éliminé les fluctuations aléatoires qui apparaissent dans la première moitié de la carte r dans la figure 4.2.

i	$\mathbf{Q(G, F_4^i)}$
1	.44
2	.37
3	.58
4	.60
5	.26
6	.73
7	.65
8	.46
9	.57
10	.58
11	.02
12	.16
13	.04
14	.07
15	.21
16	0
17	0
18	.01
19	.09
20	.24

TAB. 4.3 – Les valeurs de la statistique $Q(G_m, F_4^i)$

La figure 4.5 illustre la carte S pour les valeurs $S_n(G_m)$ de la table 4.4. Puisque les valeurs de S ne sont pas normalisées, la limite de contrôle inférieure LCL est $-z_\alpha\sqrt{(n^2/12)[(1/m)+(1/n)]}$, qui est une courbe. Ainsi, pour garder le graphique dans le format standard du papier, nous devons adopter une échelle beaucoup plus petite pour l'axe des valeurs de S. En revanche, sur la figure 4.6, les valeurs de S ont été normalisées, et par conséquent aucun changement d'échelle n'est nécessaire. Les valeurs normalisées de S (notées par S^*) sont stockées dans la table 4.5. La limite de contrôle inférieure LCL est ainsi égale à $-z_\alpha$ et qui vaut -1.96 dans ce cas. Pour les deux figures, la ligne centrale CL est égale à zéro.

i	S(i)	LCL(i)	i	S(i)	LCL(i)
1	0.45	-0.57	41	0.62	-3.78
2	0.09	-0.80	42	0.12	-3.83
3	0.2	-0.98	43	-0.38	-3.88
4	-0.24	-1.14	44	-0.88	-3.93
5	-0.07	-1.27	45	-0.75	-3.98
6	-0.47	-1.40	46	-1.25	-4.03
7	-0.43	-1.51	47	-1.75	-4.07
8	-0.75	-1.61	48	-2.25	-4.12
9	-0.46	-1.71	49	-2.75	-4.17
10	-0.47	-1.81	50	-3.09	-4.21
11	-0.07	-1.90	51	-3.59	-4.26
12	-0.41	-1.99	52	-4.09	-4.30
13	-0.01	-2.07	53	-4.59	-4.35
14	-0.29	-2.15	54	-4.89	-4.40
15	-0.39	-2.23	55	-5.39	-4.44
16	0.02	-2.30	56	-5.82	-4.48
17	-0.47	-2.38	57	-5.87	-4.53
18	-0.81	-2.45	58	-5.97	-4.57
19	-0.69	-2.52	59	-6.47	-4.62
20	-0.92	-2.59	60	-6.97	-4.66
21	-0.52	-2.65	61	-7.47	-4.70
22	-0.50	-2.72	62	-7.97	-4.75
23	-0.03	-2.78	63	-8.47	-4.79
24	-0.002	-2.84	64	-8.97	-4.83
25	-0.37	-2.91	65	-9.47	-4.87
26	-0.16	-2.97	66	-9.97	-4.92
27	0.28	-3.03	67	-10.47	-4.96
28	0.62	-3.08	68	-10.97	-5.0
29	0.23	-3.14	69	-11.47	-5.04
30	0.43	-3.20	70	-11.97	-5.08
31	0.88	-3.25	71	-12.47	-5.12
32	0.47	-3.31	72	-12.91	-5.16
33	0.6	-3.37	73	-13.06	-5.20
34	0.63	-3.42	74	-13.56	-5.24
35	0.39	-3.47	75	-14.06	-5.28
36	0.75	-3.53	76	-14.56	-5.32
37	0.72	-3.58	77	-14.08	-5.36
38	1.01	-3.63	78	-14.58	-5.40
39	0.72	-3.68	79	-15.08	-5.44
40	1.06	-3.73	80	-15.58	-5.48

TAB. 4.4 – Les valeurs de la statistique S et de la courbe limite de contrôle

i	$S^*(i)$	i	$S^*(i)$
1	0.46	41	-0.52
2	-0.57	42	-2.65
3	0.089	43	-0.50
4	-0.80	44	-2.72
5	0.2	45	-0.03
6	-0.98	46	-2.78
7	-0.24	47	-0.002
8	-1.14	48	-2.84
9	-0.07	49	-0.37
10	-1.27	50	-2.91
11	-0.47	51	-0.16
12	-1.40	52	-2.97
13	-0.43	53	0.28
14	-1.51	54	-3.03
15	-0.75	55	0.62
16	-1.61	56	-3.08
17	-0.46	57	0.23
18	-1.71	58	-3.14
19	-0.47	59	0.43
20	-1.81	60	-3.20
21	-0.07	61	0.88
22	-1.90	62	-3.25
23	-0.41	63	0.47
24	-1.99	64	-3.31
25	-0.01	65	0.6
26	-2.07	66	-3.37
27	-0.29	67	0.63
28	-2.15	68	-3.42
29	-0.39	69	0.39
30	-2.22	70	-3.47
31	0.02	71	0.75
32	-2.30	72	-3.53
33	-0.47	73	0.72
34	-2.38	74	-3.58
35	-0.81	75	1.01
36	-2.45	76	-3.63
37	-0.69	77	0.72
38	-2.52	78	-3.68
39	-0.92	79	1.06
40	-2.59	80	-3.73

TAB. 4.5 – Les valeurs de la statistique S^*

Dans cette simulation, nous avons choisi $m = 500$. Clairement, de plus grandes valeurs de m donnent de meilleures approximations aux distributions limites indiquées dans les propositions 13, 14 et 16, et au LCL pour les cartes r, Q, et S. Notre expérience montre que, pour le cas bivarié, les résultats d'approximations sont raisonnables pour le m que nous avons choisi. Toutefois, nous recommandons de plus grandes valeurs pour des observations de grandes dimensions.

4.5 Étude comparative

Pour mieux apprécier la performance de notre modèle de construction des cartes de contrôle multivariées, nous nous proposons de faire une étude comparative avec une carte conventionnelle. Pour cette fin, nous avons choisi la carte Hotelling T^2. Pour commencer, rappelons les équations sous-jacentes à cette carte.

$$\begin{cases} T^2 = n(\bar{x} - \bar{\bar{x}})^\top S^{-1}(\bar{x} - \bar{\bar{x}}) \\ UCL = \frac{p(m-1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \\ LCL = 0 \end{cases}$$

où $\bar{\bar{x}}$ est la valeur en contrôle du vecteur moyen, m est le nombre de sous-groupes dans l'ensemble de données, n est la taille des sous-groupes, et S est l'estimée de la matrice covariance. La limite UCL précédente est utilisée pour la première phase d'analyse (voir chapitre 2). Pour la phase II, la limite de contrôle UCL est estimée par :

$$UCL = \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1}$$

Nous avons expérimenté ce schéma (avec $n = 4$ et $n = 10$) sur une base de 10000 observations. Nous avons également appliqué la carte Q à ces mêmes observations. Nous résumons les résultats obtenus dans les tables 4.6 et 4.7 suivantes. Nous appelons :

- **P.F.A**, le pourcentage des fausses alarmes
- **P.H.C.D**, le pourcentage des observations hors-contrôles détectées
- **P.H.C.N.D**, le pourcentage des observations hors-contrôles non-détectées

Nous estimons ainsi la qualité de la méthode selon sa sensibilité (pourcentage des observations hors-contrôles détectées divisé par le pourcentage des observations hors-contrôles : $\frac{\text{P.H.C.D}}{\text{P.H.C.D+P.H.C.N.D}}$) et selon le nombre des fausses alarmes. Bien évidemment, il s'agit d'atteindre une haute sensibilité sans que le nombre de fausses alarmes soit trop important.

	P.F.A	**P.H.C.D**	**P.H.C.N.D**	**Sensibilité**
Carte Q $(n = 4)$	7.73 %	91.8 %	8.2 %	0.91
Carte Q $(n = 10)$	7.83 %	99 %	1 %	0.99
Carte T^2 $(n = 4)$	53.13 %	100 %	0 %	1
Carte T^2 $(n = 10)$	9.66 %	100 %	0 %	1

TAB. 4.6 – Résultats de comparaison de la carte Q avec la carte Hotelling T^2 avec des données qui suivent une loi normale multivariée

Dans cette première comparaison (Table 4.6), nous supposons que les observations du procédé suivent une loi normale multivariée. Nous utilisons ainsi un échantillons dont les 6000 premières observations suivent une $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$, et les 4000 dernières suivent une $\mathcal{N}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$.

	P.F.A	**P.H.C.D**	**P.H.C.N.D**	**Sensibilité**
Carte Q $(n = 4)$	8.86 %	84 %	16 %	0.84
Carte Q $(n = 10)$	9 %	98.5 %	1.5 %	0.99
Carte T^2 $(n = 4)$	44.2 %	82.9 %	17.1 %	0.82
Carte T^2 $(n = 10)$	10.1 %	89.75 %	10.25 %	0.90

TAB. 4.7 – Résultats de comparaison de la carte Q avec la carte Hotelling T^2 avec des données qui ne suivent pas une loi normale multivariée

Dans la deuxième comparaison (Table 4.7), nous ne posons aucune hypothèse sur les observations du procédé. Nous utilisons donc un échantillon qui

ne provient pas d'une loi normale. Pour cette fin, on utilise un générateur standard de nombres aléatoires multivariés.

Ainsi, l'analyse des résultats présentés dans les tables 4.6 et 4.7 montre que la carte Q que nous proposons performe mieux même quand le procédé est supposé suivre une loi normale. Cette conclusion est basée sur la comparaison des deux méthodes en tenant compte de leur sensibilités ainsi que de leurs pourcentages de fausses alarmes comme nous l'avons déjà indiqué précédemment. Notons qu'on ne doit accorder à ces résultats qu'une importance relative. En effet, pour valider de telles cartes de contrôle, il est nécessaire d'utiliser une base d'observations provenant d'un vrai procédé de fabrication. Ce qui n'est pas le cas ici. Néanmoins, ces résultats sont très encourageants et prouvent clairement la performance des cartes de contrôle proposées.

4.6 Conclusion

Dans ce chapitre, nous avons proposé une méthode générale pour la mise au point des cartes de contrôle multivariées. Elle se base sur la notion de profondeur des données multivariées. Dans un premier temps, nous avons décrit les statistiques (dérivées de la notion de profondeur) derrières la méthode proposée. Ces dernières ont été établies par Liu, R. Y. and Singh, K. [43] en 1993. Dans un second temps, nous avons utilisé la profondeur de Tukey pour introduire et expérimenter quatre cartes de contrôle multivariées (la carte r,la carte Q,la carte S, et la carte S^*). Les résultats obtenus par l'application de ces dernières montrent que notre modèle est très prometteur. D'une part, du fait que les cartes de contrôle qui en résultent sont valides sans aucune considération paramétrique sur le procédé. D'autre part, les résultats de comparaison de la carte Q que nous avons proposée avec la carte Hotelling T^2 montrent que notre méthode est meilleure même avec la restriction de normalité sur le procédé.

Chapitre 5

Conclusion et perspectives

Lorsqu'on cherche à piloter un procédé de fabrication, le contrôle statistique de la qualité se présente comme une boîte à outils bien fournie et qui ne cesse de s'enrichir au fur et à mesure que ses utilisateurs explorent des domaines nouveaux. Parmi ces outils, la méthode la plus répandue, dans le milieu industriel, est celle des cartes de contrôle. Cette technique graphique permet d'observer les variations du procédé dans le temps et de juger statistiquement si un déréglage ou une variation inhabituelle se sont produits. Le présent travail porte sur les cartes de contrôle multivariées. Notre premier but était de bien comprendre les techniques de construction de ces dernières. Pour ce faire, nous avons décortiqué les méthodes qui existent dans la littérature. Cette étape nous a permis de constater que la plupart de ces méthodes sont restreintes au cas de la normalité des observations. Ce qui n'est pas toujours le cas en pratique. Nous avons donc cherché tout au long de ce travail, à édifier une approche pour la construction des cartes de contrôle multivariées faciles à interpréter et à visualiser d'une part, et qui ne posent aucune condition paramétrique sur le procédé d'autre part.

Au terme de cette présentation, il semble utile de situer l'apport de notre travail à l'intérieur de la grande boîte du contrôle statistique de la qualité.

5.1 Apport de ce mémoire

Nous avons basé notre approche sur la profondeur de Tukey qui constitue la colonne vertébrale de notre travail. Dans un premier temps, nous avons rassemblé ses propriétés théoriques nécessaires pour la mise au point de notre approche. Ensuite, nous avons décrit les statistiques basées sur la notion de

profondeur en général. Ces dernières ont été établies par Liu and Singh [43]. Liu [44] les a utilisées pour introduire trois cartes de contrôle en se basant sur la profondeur du simplexe.

Les cartes que nous avons proposées dans ce travail se distinguent par rapport aux cartes classiques (Hotelling T^2, Shewhart,...) par le fait qu'elles n'imposent pas l'hypothèse de normalité des observations. En plus, les résultats de comparaison, réalisée au chapitre précédent, montrent que notre modèle performe mieux même sous l'hypothèse de normalité des observations. Enfin, on ne doit pas oublier la rapidité des algorithmes de calcul de la profondeur de Tukey. Par exemple, pour un échantillon bidimentionnel de taille n, le calcul de la profondeur de Tukey d'un point par rapport à cet échantillon nécessite une complexité temporelle d'ordre $O(n \log n)$ plutôt que $O(n^3)$ pour la profondeur du simplexe que Liu [44] a utilisée pour la mise au point de ces cartes.

5.2 Extensions et suites possibles de ce travail

Comme l'objectif du contrôle statistique multivarié de la qualité est de surveiller les procédés de fabrication dans le temps, afin de détecter les événements inhabituels causant la détérioration de la qualité. Il est essentiel de pouvoir dépister la cause d'un signal hors-contrôle. Cependant, contrairement aux cartes de contrôle univariées, la complexité des cartes multivariées et la corrélation entre les variables, rendent l'analyse et l'interprétation d'un signal hors-contrôle difficiles. Ceci se traduit par le fait qu'un signal hors-contrôle n'est pas habituellement provoqué par une seule variable, mais plutôt par plusieurs variables qui sont en général corrélées.

Un autre but est d'étudier la période opérationnelle moyenne **ARL**[1] des cartes que nous avons proposées. Ceci pourra nous permettre une comparaison facile et rigoureuse de nos cartes avec les autres cartes de la littérature.

Bien entendu cette liste est non exhaustive et bon nombre d'extensions supplémentaires peuvent certainement être envisagées...

[1] **ARL** correspond à "Average Run Lengths"

Bibliographie

[1] Alt. F. B., Goode, J. J., and Wadsworth, H. M. (1976). *Ann. Tech. Conf. Trans.* ASCQ, pp. 170-176.

[2] Alt. F. B., Walker, J. W., and Goode, J. J.(1980). *Ann. Tech. Conf. Trans.* ASCQ, pp. 754-759.

[3] Alt, F. (1982). *Multivariate Quality Control : State of Art.* ASQC Annual Quality Congress Transactions, pp. 886-893.

[4] Alt, F. (1984). *Multivariate Quality Control.* The Encyclopidia of statistical Sciences, pp. 110-122.

[5] Alt. F. B. (1985). *Multivariate Control Charts*, Encyclopidia of statistical Sciences, 6. (S. Kotz and N. L. Johson, Eds. Wiley, New York), pp. 110-122.

[6] Alt, F., and Smith, N. (1988). *Multivariate Quality Control.* in Handbook of Statistics, 7, eds. P. R. Krishnaiah and C. R. Rao, Amsterdam : Elsevier, pp. 333-351.

[7] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York.

[8] Aparisi, F. (1996). *Hotelling's T^2 control chart with adaptative sample sizes*, International Journal of Production Research, 34(10), pp. 2853-2862.

[9] Aparisi, F. (1997). *Sampling Plans for the Multivariate T^2 Control Chart*, Quality Engineering, 10(1), pp. 141-147.

[10] Aparisi, F., Jabaloyes, J., and Carrión, A. (1999). *Statistical Properties of the $|S|$ Multivariate Control Chart.* Commun. Statist.–Theory Meth., 28(11), pp. 2671-2686.

[11] Bélisle, C. and Massé, J.-C. (1994). *Using the Tukey depth as a trimming tool.* Unpublished manuscript.

[12] Brown, B. M. and Hettmansperger, T. P. (1989). *An affine invariant bivariate version of the sign test.* J. R. Statist. Soc. B, 51, pp. 117-125.

[13] Chan, L. K., and Li, G. C. (1995). *Bivariate Control Charts for Testing $\bar{X}$ Data Patterns*, Canadian Journal of Statistics, 23, pp. 85-99.

[14] Chan, L. K., and Zhang, J. (1996a). *Some Issues on the Design of CUSUM and EWMA Charts*, Research Report, Departement of Management Sciences, City University of Hong Kong.

[15] Chan, L. K., and Zhang, J. (1996b). *Cumulative Sum Control Charts for Covariance Matrix*, Research Report, Departement of Management Sciences, City University of Hong Kong.

[16] Chua, M. K., and D. C. Montgomery (1992). *Investigation and Characterization of a Control Scheme for Multivariate Quality Control*, Quality and Reliability Engineering International, 8, pp. 37-44.

[17] Crosier, R. B. (1986). *A New Two-Sided Cumulative Sum Quality Control Scheme*, Technometrics, 28, pp. 187-194.

[18] Crosier, R. B. (1988). *Multivariate Generalization of Cumulative Sum Quality Control Schemes*, Technometrics, 30, pp. 291-303.

[19] Dempster, A. P. and Gasko, M. G. (1981). *New tools for residual analysis.* Ann. Statist. 9, pp. 945-959.

[20] Dharmadikhari, S. and Joag-Dev K. (1988). *Unimodality, Convexity, and Applications*. Acad. Press, New York.

[21] Doganaksoy, N., F. W. Faltin and W.T. Tucker (1991). *Identification of Out of Control Quality Characteristics in a Multivariate Manufacturing Environment*, Communications in Statistics Theory and Methods, 20, pp. 2775-2790.

[22] Donoho, D. L. (1982). *Breakdown propreties of multivariate location estimators*. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

[23] Donoho, D., Jhonstone, I., Rousseeuw, P. and Stahel, W. (1985). *Comment on "Projection pursuit" by P. J. Huber*. Ann. Statist. 13, pp. 496-500.

[24] Donoho, D. L. and Liu, R. C. (1988). *The "automatic" robustness of minimum distance estimators*. Ann. Statist. 16. pp. 552-586.

[25] Donoho, D. L. and Gasko, M. (1992). *Breakdown propreties of location estimates based on halfspace depth and projected outlyingness*. Ann. Statist. 20, pp. 1803-1827.

[26] Dudley, R. M. (1989). *Real Analysis and Probability*. Wadsworth & Books/Cole, Pacific Grove, Ca.

[27] Dümbgen, L. (1992). *Limit theorems for the simplicial depth*. Statist. Probab. Letters, 14, pp. 119-128.

[28] Feller, W. (1971). *Introduction to Probability Theory and Its Applications (2nd ed.)*, New York : John Wiley.

[29] Fellner, W. H. (1990). *Average Run Lengths for cumulative Sum Schemes. Algorithm AS258*, Applied Statistics, 39, 3, pp. 402-412.

[30] Hawkins, D. M. (1991). *Multivariate quality control based on regression-adjusted variables*, Technometrics, 33, pp. 61-75.

[31] Hawkins, D. M. (1993). *Regression adjustement for variables in multivariate quality control*, Journal of Quality Technology, 26, 3, pp. 170-182.

[32] Hayter, A. J. and K. Tsui (1994). *Identification and Quantification in Multivariate Quality Control Problems* , Journal of Quality Technology, 26, 3, pp. 197-208.

[33] Healy, J. D. (1987). *A note on multivariate* **CUSUM** *procedures*, Technometrics, 29, pp. 409-412.

[34] Hillier, F. S. (1969). $\bar{X}-$ *and* $R-$*Chart Control Limits Based on A Small Number of Subgroups*, J. Qual. Tech. 1, pp. 17-26

[35] Hotelling, H. (1947). *Multivariate Quality Control Illustrated by the Air Testing of Sample Bombsights*, Techniques of Statistical Analysis, (Eds. C. Eisenhart, M. Hastay and W. A. Wallis, McGraw-Hill), 111-184.

[36] Jackson, J. E. (1957). *An Application of Multivariate Quality Control to Photographic Processing*, Journal of the American Statistical Association, pp. 186-199.

[37] Jackson, J. E. (1959). *Quality Control Methods for Several Related Variables*, Technometrics, 1(4), 359-377.

[38] Jackson, J. E. (1985). *Multivariate Quality Control*, Communications in Statistics, 14(11), pp. 2657-2688.

[39] Kourti, T. and J. F. MacGregor (1996). *Multivariate SPC Methods for Process and Product Monitoring*, Journal of Quality Technology, 28, 4, pp. 409-428.

[40] Kresta, J. V., J. F. MacGregor and T. E. Marlin (1991). *Multivariate Statistical Monitoring of Process Operating Performance*, Canadian Journal of Chemical Engineering, 69, pp. 35-47.

[41] Liu, R. Y. (1990). *On a notion of data depth based on random simplices.* Ann. Statist. 18, pp. 405-414.

[42] Liu, R. Y. (1992). *Data Depth and Multivariate Rank Tests.* In L_1-Statistical Analysis and Related Methods (Y. Dodge, ed.), pp. 279-294. North-Holland.

[43] Liu, R. Y. and Singh, K. (1993). *A quality index based on data depth and multivariate rank tests.* Journal of the American Statistical Association, 88, pp. 252-260.

[44] Liu, R. Y. (1995). *Control charts for Multivariate Process.* Journal of the American Statistical Association, 90, pp. 1380-1387.

[45] Liu, R. Y. and Singh, K. (1997). *Notions of limiting P-values based on data depth and bootstrap.* Journal of the American Statistical Association, 92, pp. 266-277.

[46] Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992). *A Multivariate Exponentially Weighted Moving Average Control Chart*, Technometrics, 34, pp. 46-53.

[47] Lowry, C. A. and D. C., Montgomery (1995). *A review of multivariate control charts*, IIE Transactions, 27, pp. 800-810.

[48] Lucas, J. M. (1973). *A Modified V-Mask Control Scheme*, Technometrics, 15, pp. 833-847.

[49] Lucas, J. M., Saccucci, M. S. (1990). *Exponentially Weighted Moving Average Control Schemes : Properties and Enhancements*, Technometrics, 32(1), pp. 1-30.

[50] Mason, R. L., Tracy, N. D. and Young, J. C. (1995). *Decomposition of T^2 for Multivariate Control Chart Interpretation*, Journal of Quality Technology, 27, pp. 99-108.

[51] Mason, R. L., Tracy, N. D. and Young, J. C. (1996). *Monitoring a Multivariate Step Process*, Journal of Quality Technology, 28(1), pp. 39-50.

[52] Mason, R. L., Champ, C.W., Tracy, N. D., Wierda, S. J. and Young, J. C. (1997). *Assessment of Multivariate Process Control Techniques*, Journal of Quality Technology, 29(2), pp. 140-143.

[53] Mason, R. L., Tracy, N. D. and Young, J. C. (1997). *A Practical Approch for Interpreting Multivariate T^2 Control Chart Signals*, Journal of Quality Technology, 29(4), pp. 396-406.

[54] Massé, J.-C. and Theodorescu, R. (1994). *Halfplane trimming for bivariate distributions*, Journal of Multivariate Analysis, 48, pp. 188-202.

[55] Massé, J.-C. (1998). *Asymptotics for the Tukey Depth*, Technical Report, dépt. de Mathématiques et de statistique, Univ. Laval, Canada.

[56] Morud, T. E. (1996). *Multivariate Statistical Process Control; Example from the Chemical Process Industry*, Journal of Chemometrics, 10, pp. 669-675.

[57] Murphy, B. J. (1987). *Selecting Out of Control Variables with the T^2 Multivariate Quality Control Procedure*, The Statistician, 36, pp. 571-583.

[58] Nelson, L. S. (1985). *Interpreting Shewhart $\bar{X}$ Control Charts*, Journal of Quality Technology, 17, pp. 114-116.

[59] Nomikos, P. and J., MacGregor (1995). *Multivariate SPC Charts for Monitoring Batch Processes*, Technometrics, 37(1), pp. 41-59.

[60] Oja, H. (1983). *Descriptive Statistics for multivariate distributions.* Statist. Probab. Lett., 1, pp. 327-332.

[61] Page, E. S. (1954). *Continuous inspection schemes*, Biometrika, 41, pp. 100-115.

[62] Pignatello, J., Jr. and Runger, G. C. (1990). *Comparisons of multivariate CUSUM charts*, Journal of Quality Technology, 22, pp. 173-186.

[63] Prabhu, S. S. and Runger, G. C. (1997). *Designing a Multivariate EWMA Control Chart*, Journal of Quality Technology, 29(1), pp. 8-15.

[64] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

[65] Rius, A., M. P. Callao, and F. X. Rius (1997). *Multivariate Statistical Process Control Applied to Sulfate Determination By Sequential Injection Analysis.* Analyst, 122, pp. 737-741.

[66] Rousseeuw, P. J. and Ruts, I. (1996). *Algorithm AS 307 : Bivariate location depth.* Applied Statistics (JRSS-C), 45, pp. 512-526.

[67] Rousseeuw, P.J. and Struyf, A. (1998). *Computing location depth and regression depth in higher dimensions.* Statistics and Computing, 8, pp. 193-203.

[68] Ruiz, A. (1992). *Estimation robuste d'une matrice de dispersion et projections révélatrices.* Ph.D. Thesis, University of Toulouse, France.

[69] Runger, G. C., F. B., Alt, and D. C., Montgomery (1996). *Contributions to a Multivariate Statistical Process Control Chart Sgnal*, Communications in Statistics - Theory and Methods , 25(10), pp. 2203-2213.

[70] Runger, G. C. and S. S., Prabhu (1996). *A Markov Chain Model for the Multivariate Exponentially Weighted Moving Averages Control Chart*, Journal of the American Statistical Association, 91(436), pp. 1701-1706.

[71] Small, C. G. (1987). *Measures of centrality for multivariate and directional distributions*, Canad. J. Statist. 15. pp. 31-39.

[72] Steele, J. M. (1978). *Empirical discrepancies and subadditive processes*, Ann. Probab., 6, pp. 118-127.

[73] Sultan, T. I. (1986). *An Acceptance Chart for Raw Material of Two Correlated Properties*, Quality Assurance, 12, pp. 70-72.

[74] Timm, N. H. (1996). *Multivariate Quality Control Using Finite Intersection Tests*, Journal of Quality Technology, 28(2), pp. 233-243.

[75] Tukey, J. W. (1974.a). *T6 : Order Statistics.* In mimeographed notes for Statistics 411, Princeton Univ.

[76] Tukey, J. W. (1974.b). *Adress to International Congress of Mathematicians*, Vancouver.

[77] Tukey, J. W. (1975). *Mathematics and Picturing Data.* Proceedings of International Congress of Mathematicians, Vancouver, 2, pp. 523-531.

[78] Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.

[79] Woodall, W. H., and Ncube, M. M. (1985), *Multivariate CUSUM Quality-Control procedures*, Technometrics, 27(3), pp. 285-292.

[80] Yashchin, E. (1985). *On the Analysis and Design of CUSUM-Shewhart Control Schemes*, IBM Journal of Research and Development, 29, pp. 377-391.

[81] Yang, C. H. and Hillier, F. S. (1970). *J. Qual. Tech.* 2, 9-16.

[82] Yeh, A. B. and Singh, K. (1992). *Bootstrap confidence regions based on data depths.* Presented at the Annual Meeting of the IMS, Boston, August 9-13.

[83] Yeh, A. B. and Singh, K. (1997). *Balanced confidence regions based on Tukey's depth and the bootstrap.* Journal of the Royal Statistical Society Series B, 59, pp. 639-652.

[84] Zhang, W. (1996). *Multivariate Control Charts Based on a Small Number of Sub-groups*, Chinese Journal of Applied Probabilities and Statistics, 12, pp. 88-94.