JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. CAZÈS
F. GOUPIL
J. GOUPIL
C. PARDOUX

Traitement statistique du signal de contrôle en ligne par spectrométrie proche infra-rouge

Journal de la société statistique de Paris, tome 138, n° 3 (1997), p. 83-95

<http://www.numdam.org/item?id=JSFS_1997__138_3_83_0>

TRAITEMENT STATISTIQUE DU SIGNAL DE CONTROLE EN LIGNE PAR SPECTROMETRIE PROCHE INFRA-ROUGE[1]

P. CAZES – F. GOUPIL – J. GOUPIL – C. PARDOUX
Université de Paris IX-Dauphine, LISE-CEREMADE [2]

M. LAMBERT
Société Weber & Broutin France [3]

Résumé

Une méthodologie relative à l'automatisation du contrôle de qualité multivarié de la composition d'un produit a été développée et mise en œuvre. On fait appel à des lois statistiques multivariées comme le T^2 de HOTELLING, et les valeurs aberrantes sont éliminées si le carré de leur distance de MAHALANOBIS à la composition moyenne, calculée au cours de la phase de calibration, dépasse un seuil fixé. Le contrôle en ligne est réalisé à l'aide d'une carte de contrôle établie en fonction de ce seuil.

Mots-clés : Calibration, Carte de contrôle, Contrôle de qualité, Contrôle en ligne, D^2 de MAHALANOBIS, Points aberrants, T^2 de HOTELLING.

Summary

Methodology for automating multivariate quality control of a product was both developed and implemented. For this, we first used multivariate statistical distributions such as HOTELLING's T^2 ; the outliers were eliminated if the square of their MAHALANOBIS distance to the standard composition (calibration phase) was over a fixed level. Then on-line control was determined with a control chart based on that level.

Keywords : Calibration, Control Chart, Quality Control, On-line Control, MAHALANOBIS D^2, Outliers, HOTELLING's T^2.

---

1. Cet article a donné lieu à une communication présentée lors des troisièmes journées MODULAD "*Applications industrielles de l'analyse des données*" qui se sont tenues à Trégastel les 9 et 10 octobre 1997.
2. Place du Maréchal-de-Lattre-de-Tassigny, 75775 Paris cedex 16
3. BP 84, rue de Brie Servon, 77253 - Brie Comte Robert cedex

I. Introduction

On veut automatiser le contrôle de la composition d'un produit en cours de fabrication à partir de son spectre proche Infra-Rouge. A intervalle de temps régulier (toutes les trois minutes, par exemple), on prélève automatiquement une certaine quantité de produit qui passe dans une chambre, et un spectromètre fournit les absorbances de ce produit sur un nombre p de longueurs d'onde, p étant de l'ordre de 20 pour les spectromètres utilisés actuellement, mais pouvant dépasser 100 avec les spectromètres les plus récents. Au cours de la même journée, des produits différents peuvent être fabriqués sur la même chaîne de fabrication, ces produits différant par suite de dosages non identiques des matières premières qui rentrent dans leur composition, et de réglages différents de l'appareillage. Pour effectuer le contrôle, on se servira du D^2 de MAHALANOBIS comme mesure de l'écart entre le spectre du produit analysé et sa composition moyenne établie lors de la calibration.

L'article comprend trois parties. Dans la première partie, on définit les notations, puis les caractéristiques (phase de calibration) qui permettront d'effectuer le contrôle en ligne qui sera étudié dans la deuxième partie. Deux méthodologies (suivant le choix de la métrique) ayant été définies, on indiquera dans la dernière partie la méthodologie effectivement adoptée, ainsi que les problèmes pratiques rencontrés.

Cette étude a donné lieu à la réalisation d'un logiciel opérationnel fonctionnant en continu en usine.

II. Notations • Phase de calibration

Dans ce paragraphe, on suppose qu'on dispose d'un fichier de n spectres (chaque spectre étant caractérisé par p valeurs numériques associées à p longueurs d'onde) correspondant à r produits, et après avoir éliminé les spectres considérés comme aberrants, on définira les caractéristiques qui permettront d'établir la carte de contrôle.

II.1 Notations

On désigne par P_k $(1 \leq k \leq r)$ le $k^{\text{ème}}$ produit et par n_k le nombre de spectres relevés sur ce produit $(n = \sum\{n_k \mid k = 1, \cdots, r\})$.

Pour chaque produit P_k, on a donc n_k spectres,
soit n_k vecteurs $\underline{y}_{k_1}, \cdots, \underline{y}_{k_i}, \cdots, \underline{y}_{k_{nk}}$ à p composantes.

On désigne par $\underline{g}_k$ le centre de gravité des $\underline{y}_{k_i} (1 \leq i \leq n_k)$, par V_k la matrice variance associée, et par W la matrice variance intraclasses :

$$\underline{g}_k = \frac{1}{n_k} \sum \{\underline{y}_{ki} \mid i = 1, \cdots, n_k\}$$

$$\underline{V}_k = \frac{1}{n_k} \sum \{(\underline{y}_{ki} - \underline{g}_k)(\underline{y}_{ki} - \underline{g}_k)' \mid i = 1, \cdots, n_k\}$$

$$W = \frac{1}{n} \sum \{n_k V_k \mid k = 1, \cdots, r\}$$

Nous supposerons dans la suite que les $\underline{y}_{k_i} (1 \leq i \leq n_k, 1 \leq k \leq r)$ suivent des lois gaussiennes indépendantes de moyenne $\underline{m}_k$ et de matrice variance Σ_k.

Nous supposerons également que les V_k ne diffèrent pas significativement, ce qui revient à faire l'hypothèse que les matrices variances théoriques Σ_k associées à chaque produit sont égales ($\Sigma_k = \Sigma$). Nous indiquerons comment se modifient les formules si on ne peut pas faire l'hypothèse précédente.

Supposant la matrice W inversible, on adoptera dans l'espace R^p la métrique de MAHALANOBIS $M = W^{-1}$. Le carré de la distance entre deux vecteurs $\underline{x}$ et $\underline{y}$ de R^p muni de la métrique précédente sera appelé D^2 de MAHALANOBIS entre $\underline{x}$ et $\underline{y}$ et notée $D^2(\underline{x}, \underline{y})$[4].

II.2 Elimination des points aberrants • Obtention de l'échantillon de calibration

Pour voir si le $i^{\text{ème}}$ spectre $\underline{y}_{k_i}$ du produit P_k est aberrant, on calcule le carré de la distance de MAHALANOBIS entre $\underline{y}_{k_i}$ et le centre de gravité $\underline{g}_k$, et on élimine $\underline{y}_{k_i}$ si le carré de cette distance est trop important (i.e. diffère significativement de 0). De façon précise, et en se fixant un risque de première espèce égal à α de considérer à tort que $\underline{y}_{k_i} - \underline{g}_k$ diffère significativement de $\underline{0}$, on élimine, avec les hypothèses faites à la fin du § 2.1, $\underline{y}_{k_i}$ si (cf. annexe B) :

$$\frac{n-r-p}{np} \cdot \frac{n_k}{n_k - 1} \cdot \frac{D^2(\underline{y}_{k_i}, \underline{g}_k)}{1 - \frac{n_k}{n_k - 1} \cdot \frac{D^2(\underline{y}_{k_i}, \underline{g}_k)}{n}} > FS_{1-\alpha}(p, n-r-p) \quad (1)$$

$FS_{1-\alpha}(p, n-r-p)$ étant le quantile d'ordre $(1-\alpha)$ d'une loi de FISHER-SNEDECOR à p et $(n-r-p)$ degrés de liberté.

Après avoir retiré les spectres considérés comme aberrants, on recalcule les caractéristiques de chaque produit (centres de gravité, matrices variances) ainsi que la matrice variance intraclasse W, et on peut le cas échéant retester sur l'échantillon ainsi apuré s'il reste des prélèvements aberrants, les éliminer, recalculer les caractéristiques des produits et itérer le processus jusqu'à ce qu'il n'y ait plus de spectres aberrants.

Après cette phase d'élimination des prélèvements aberrants, on obtient l'échantillon de calibration dont les caractéristiques des produits vont permettre de déterminer les caractéristiques du contrôle en ligne (cf. § 3).

---

4. Le D^2 de MAHALANOBIS usuel est calculé à partir de la métrique $\left(\frac{n}{n-r} \cdot W\right)^{-1}, \frac{n}{n-r}$. W étant l'estimateur sans biais de la matrice variance Σ supposée commune des produits, le calcul étant en général effectué dans le cas de deux groupes ($r = 2$) entre les centres de gravité de ces deux groupes.

Remarques

i) En général, le nombre total de prélèvements (ou de spectres) n est très élevé et la quantité $\frac{n_k}{n_k - 1} \cdot \frac{D^2\left(\underline{y}_{k_i}, \underline{g}_k\right)}{n}$ est négligeable devant 1.

La formule (1) se simplifie et s'écrit alors :

$$\frac{n-r-p}{np} \cdot \frac{n_k}{n_k - 1} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) > FS_{1-\alpha}(p, n-r-p) \qquad (1')$$

Si $(n - r - p)$ est supérieur à 200, la formule précédente se simplifie encore dans la mesure où la loi de FISHER-SNEDECOR à p et $(n - r - p)$ degrés de liberté peut être considérée, au facteur $1/p$ près, comme une loi du Chi-Deux à p degrés de liberté. La formule (1') s'écrit alors :

$$\frac{n-r-p}{n} \cdot \frac{n_k}{n_k - 1} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) > \chi^2_{1-\alpha}(p) \qquad (1'')$$

$\chi^2_{1-\alpha}(p)$ étant le quantile d'ordre $(1 - \alpha)$ d'une loi du Chi-Deux à p degrés de liberté.

ii) Si on ne peut plus faire l'hypothèse d'égalité des matrices variances théoriques Σ_k associées aux spectres des différents produits pour voir si $\underline{y}_{k_i}$ diffère significativement du centre de gravité $\underline{g}_k$ du produit P_k, il faut munir l'espace R^p de la métrique V_k^{-1}. La formule (1) demeure valable à condition :

- de calculer le D^2 entre $\underline{y}_{k_i}$ et $\underline{g}_k$ avec la métrique V_k^{-1} et non plus W^{-1},
- de remplacer n par n_k, et r par 1, ce qui revient à se restreindre aux prélèvements du produit k ;

posant $D'^2 = \left(\underline{y}_{k_i} - \underline{g}_k\right)' V_k^{-1} \left(\underline{y}_{k_i} - \underline{g}_k\right)$ la formule (1) s'écrit alors :

$$\frac{n_k - p - 1}{p} \cdot \frac{D'^2}{n_k - 1 - D'^2} > FS_{1-\alpha}(p, n_k - p - 1) \qquad (2)$$

On peut noter que quand $\underline{y}_{k_i} - \underline{g}_k$ ne diffère pas significativement de $\underline{0}$, $D'^2/(n_k - 1)$ suit une loi beta de type I de paramètres $p/2$ et $(n_k - p - 1)/2$ [cf. fin de l'annexe B et TRACY et al., 1997].

III. Contrôle en ligne

Pour effectuer le contrôle en ligne du processus de fabrication, on se sert des caractéristiques $\underline{g}_k (1 \leq k \leq r)$ et W [et le cas échéant, $V_k (1 \leq k \leq r)$] calculées sur l'échantillon de calibration après élimination des spectres aberrants.

Soit $\underline{y}_k$ le spectre issu d'un prélèvement du produit k en cours de fabrication. Pour tester si ce prélèvement est conforme, on va encore regarder si $\underline{y}_{k_i} - \underline{g}_k$ diffère significativement de $\underline{0}$. Désignant toujours par $D^2\left(\underline{y}_{k_i}, \underline{g}_k\right)$ le carré de la distance entre $\underline{y}_{k_i}$ et $\underline{g}_k$ avec la métrique W^{-1}, on déclarera le prélèvement

non conforme, et on arrêtera la production avec un risque α de fausse alarme si (cf. annexe A) :

$$\frac{n-r-p+1}{np} \cdot \frac{n_k}{n_k+1} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) > FS_{1-\alpha}(p, n-r-p+1) \qquad (3)$$

$FS_{1-\alpha}(p, n-r-p+1)$ étant le quantile d'ordre $(1-\alpha)$ d'une loi de FISHER-SNEDECOR à p et $(n-r-p+1)$ degrés de liberté. En général, on prend $\alpha = 0,1\%$.

Dans la mesure où sur la carte de contrôle (cf. figure 1), on reporte les D^2, la limite de contrôle pour un spectre issu du produit k est donnée d'après la formule (3) par :

$$\frac{np \cdot (n_k+1)}{(n-r-p+1) \cdot n_k} \cdot FS_{99,9\%}(p, n-r-p+1) \qquad (4)$$

et à cette limite de contrôle, on associe une limite de surveillance correspondant classiquement à un degré de confiance égal à 95 %.

Deux exemples de cartes de contrôle sont montrés dans la figure 1. Outre les limites de contrôle et de surveillance, on reporte sur chaque carte la valeur moyenne du D^2 de MAHALANOBIS calculé pour le produit considéré sur l'échantillon de calibration. On indique également les caractéristiques du dernier spectre analysé, à savoir son D^2, le nom des produits associés aux deux spectres les plus proches, et on précise si le produit est bon (en dessous de la limite de surveillance), bon à surveiller (entre les limites de surveillance et de contrôle), mauvais (au dessus de la limite de contrôle).

Remarque

Si on ne suppose plus l'égalité des matrices variances théoriques Σ_k des produits, on peut faire la même remarque que celle effectuée à la fin du § 2.2. La formule (3) s'applique à condition de remplacer n par n_k, r par 1 et la métrique W^{-1} par V_k^{-1}. Posant $D'^2 = \left(\underline{y}_{k_i} - \underline{g}_k\right) V_k^{-1} \left(\underline{y}_{k_i} - \underline{g}_k\right)$, la formule (3) s'écrit alors :

$$\frac{n_k - p}{p \cdot (n_k+1)} \cdot D'^2 > FS_{1-\alpha}(p, n_k - p) \qquad (5)$$

Les limites de contrôle et de surveillance pour D'^2 se déduisent alors immédiatement de (5).

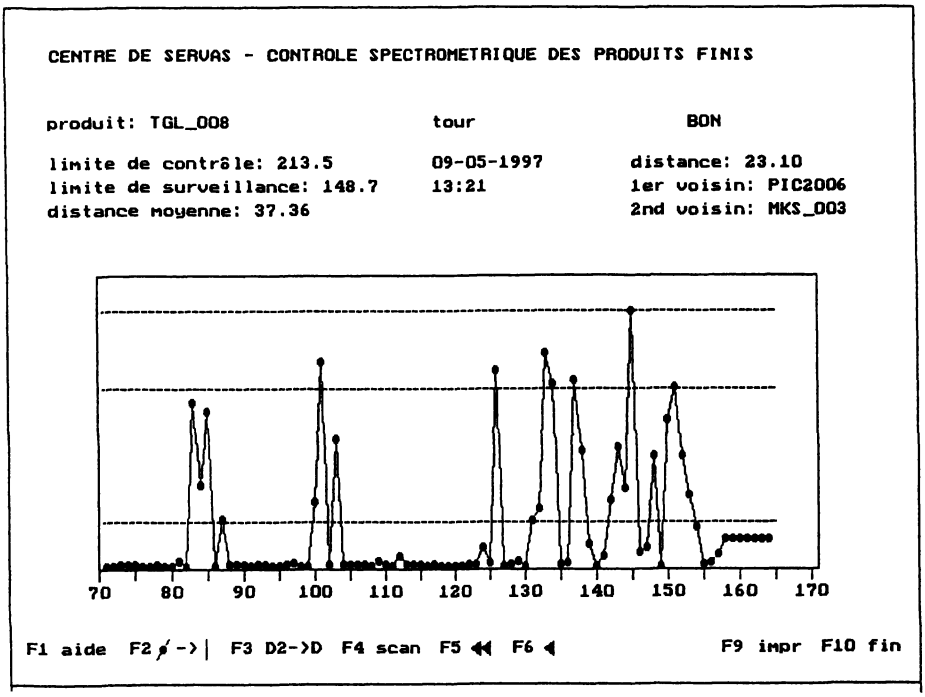

FIGURE 1 – Exemple de cartes de contrôle

IV. Mise en œuvre • Recalibration

IV.1 La métrique

Dans la mesure où certains produits ne sont fabriqués qu'en faible quantité, il n'est pas possible de prendre une métrique différente pour chaque produit. On adoptera donc, comme cela avait été fait au § 3 la métrique W^{-1} dans R^p pour calculer les D^2 de MAHALANOBIS. Par ailleurs, les absorbances étant très corrélées, la matrice W n'est pas inversible, ce qui pose problème pour calculer les D^2. On remplace alors W^{-1} par l'Inverse Généralisée de MOORE-PENROSE W^+ de W, ce qui revient à se restreindre à l'espace engendré par les $\underline{y}_{k_i} - \underline{g}_k (1 \leq i \leq n_k, 1 \leq k \leq r)$ avec les notations du § 2. De façon précise, soit :

$$W = \sum \left\{ \lambda_\alpha \cdot \underline{u}_\alpha \, \underline{u}'_\alpha \mid \alpha = 1, \cdots, p \right\}$$

la décomposition spectrale de W en fonction de ses valeurs propres λ_α et des vecteurs propres orthonormés (pour la métrique usuelle) $\underline{u}_\alpha$.

Si les valeurs propres λ_α sont rangées par valeurs décroissantes, et si W est de rang $t (\lambda_{t+1} = \lambda_{t+2} = \cdots = \lambda_p = 0)$, on a :

$$W = \sum \left\{ \lambda_\alpha \cdot \underline{u}_\alpha \, \underline{u}'_\alpha \mid \alpha = 1, \cdots, t \right\}$$

et

$$W^+ = \sum \left\{ \frac{1}{\lambda_\alpha} \cdot \underline{u}_\alpha \underline{u}'_\alpha \mid \alpha = 1, \cdots, t \right\}$$

Le carré de la distance de MAHALANOBIS entre un vecteur $\underline{y}$ dans R^p (par exemple $\underline{y} = \underline{y}_{k_i} - \underline{g}_k$) et l'origine s'écrit si $F_\alpha(\underline{y}) = \underline{u}'_\alpha \underline{y}$ est la coordonnée de la projection (avec la métrique usuelle) de $\underline{y}$ sur $\underline{u}_\alpha$:

$$D^2(\underline{y}, \underline{0}) = \underline{y}' W^+ \underline{y} = \sum \left\{ \frac{F_\alpha^2(\underline{y})}{\lambda_\alpha} \mid \alpha = 1, \cdots, t \right\}$$

formule très simple et qu'on utilisera effectivement. Il faudra alors dans les formules des § 2 et 3 remplacer p par t.

D'un point de vue pratique, dans les formules précédentes, on élimine les valeurs propres λ_α associées à un pourcentage d'inertie faible, par exemple inférieur à 1 %. On peut aussi conserver les valeurs propres associées à un pourcentage d'inertie cumulé supérieur à une valeur donnée, 95 % par exemple. Dans le logiciel réalisé c'est la première solution qui a été retenue.

Remarque

Des Analyses en Composantes Principales préliminaires effectuées sur les spectres d'un même produit ont montré, ce qui est classique avec de telles données, un effet «taille» souvent important, le pourcentage d'inertie du premier axe variant entre 60 % et 99 % suivant le produit étudié et étant dans la plupart des cas supérieur à 90 %.

IV.2 Recalibration

Si au cours du contrôle en ligne, un nouveau produit P_s est fabriqué, dès que le nombre de prélèvements de P_s est supérieur à une valeur n_0 ($n_0 = 10$ par exemple), on calcule le centre de gravité $\underline{g}_s$ des spectres associés, et on peut ainsi pour les prélèvements ultérieurs de P_s tester (sans changer la métrique déterminée lors de la phase de calibration) si ces prélèvements sont conformes (i.e. ne s'écartent pas significativement de $\underline{g}_s$).

Par ailleurs, toujours dans le contrôle en cours de fabrication, dès que le nombre de prélèvements d'un produit P_s atteint (en comptant les spectres de ce produit ayant servi à la calibration) certaines valeurs (20, 30, 50, 100, 200, etc.), on recalcule le centre de gravité de ce produit (sans changer la métrique W).

Enfin, périodiquement, ou dès que le nombre total de prélèvements dépasse une certaine valeur, on refait la calibration en gardant les derniers prélèvements issus de la production.

Annexe A

Dans l'espace R^p, on considère r nuages de points $N_1, N_2, \cdots, N_k, \cdots, N_r$, le nuage N_k étant constitué des n_k vecteurs $\underline{y}_{k1}, \cdots, \underline{y}_{ki}, \cdots, \underline{y}_{kn_k}$. On désigne par $\underline{g}_k$ et V_k le centre de gravité et la matrice variance du nuage N_k (tous les points $\underline{y}_{ki}$ ayant la même masse) et par W la matrice variance intraclasse, la définition de tous ces éléments ayant été rappelée au § 2.1.

Nous supposerons que les $\underline{y}_{ki} (1 \leq i \leq n_k, 1 \leq k \leq r)$ suivent des lois gaussiennes indépendantes de moyenne $\underline{m}_k$ et de matrice variance Σ_k, ce que nous noterons de façon classique :

$$\underline{y}_{ki} \in \mathcal{N}_p(\underline{m}_k; \Sigma_k)$$

et nous supposerons par la suite que les Σ_k sont égaux, et on désignera par Σ cette matrice variance commune.

Désignons par $W_p(d, \Sigma)$ la loi de WISHART à p dimensions, d degrés de liberté (d.d.l.), et de matrice variance Σ. Compte tenu de ce que $n_k V_k \in W_p(n_k - 1, \Sigma)$ et de ce que $nW = \sum\{n_k V_k \mid k = 1, \cdots, r\}$ avec $n = \sum\{n_k \mid k = 1, \cdots, r\}$,

$$nW \in W_p(n - r, \Sigma)$$

puisque les V_k sont indépendants.

Soit $\underline{y}_{ki} \in \mathcal{N}_p(\underline{m}_k; \Sigma)$ un vecteur gaussien indépendant des $\{\underline{y}_{mj} \mid 1 \leq j \leq n_k; 1 \leq m \leq r\}$, alors :

$$\underline{y}_{ki} - \underline{g}_k \in \mathcal{N}_p\left(\underline{0}; \Sigma\left(1 + \frac{1}{n_k}\right)\right)$$

et donc :

$$\underline{u} = \sqrt{\frac{n_k}{n_k+1}} \cdot \left(\underline{y}_{k_i} - \underline{g}_k\right) \in \mathcal{N}_p(\underline{0}; \Sigma)$$

Compte tenu de ce que $\underline{g}_k$ est indépendant de V_k, et donc de W, $\underline{u}$ est indépendant de W, et donc [Saporta, 1990] :

$$\underline{u}' \left(\frac{nW}{n-r}\right)^{-1} \underline{u} \in T_p^2(n-r)$$

$T_p^2(n-r)$ désignant la loi du T^2 de HOTELLING à p dimensions et $(n-r)$ d.d.l., d'où l'on déduit [Saporta, 1990] que :

$$F = \frac{n-r-p+1}{p \cdot (n-r)} \cdot \underline{u}' \left(\frac{nW}{n-r}\right)^{-1} \underline{u} \in FS(p, n-r-p+1)$$

$FS(p, n-r-p+1)$ désignant la loi de FISHER-SNEDECOR à p et $(n-r-p+1)$ degrés de liberté.

Ce qui s'écrit encore, en posant :

$$D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) = \left(\underline{y}_{k_i} - \underline{g}_k\right)' W^{-1} \left(\underline{y}_{k_i} - \underline{g}_k\right)$$

et tenant compte de la définition de $\underline{u}$:

$$F = \frac{n-r-p+1}{np} \cdot \frac{n_k}{n_k+1} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) \in FS(p, n-r-p+1) \qquad (A1)$$

ce qui permet de tester

l'hypothèse H_0 : $E\left(\underline{y}_{k_i}\right) = \underline{m}_k$

contre l'hypothèse alternative H_1 : $E\left(\underline{y}_{k_i}\right) \neq \underline{m}_k$

Sous H_0, F suivant la loi de FISHER-SNEDECOR à p et $(n-r-p+1)$ d.d.l., on rejettera H_0 avec un risque de première espèce égal à α si :

$$F > FS_{1-\alpha}(p, n-r-p+1)$$

$FS_{1-\alpha}(p, n-r-p+1)$ désignant le quantile d'ordre $(1-\alpha)$ d'une loi de FISHER-SNEDECOR à p et $(n-r-p+1)$ degrés de liberté.

Annexe B

B1. Introduction

Supposons maintenant que $\underline{y}_{k_i}$ soit un des éléments du nuage N_k : alors le test précédent n'est plus valable dans la mesure où $\underline{y}_{k_i}$ n'est plus indépendant de $\underline{g}_k$ et de V_k, et en particulier $\underline{y}_{k_i} - \underline{g}_k$ n'est plus indépendant de V_k et donc de W. Pour pouvoir faire le test, il suffit d'enlever le point $\underline{y}_{k_i}$ et d'appliquer le test précédent à partir des caractéristiques des $(n-1)$ points restants.

Désignant respectivement par $\underline{g}_{k-i}, V_{k-i}, W_{-i}$ le centre de gravité, la matrice variance du nuage N_k privé du point $\underline{y}_{k_i}$, et la matrice variance intraclasse associée, la formule (A1), où n est remplacé par $(n-1)$ et n_k par $n_k - 1$, s'écrit :

$$F = \frac{n-r-p}{(n-1)\cdot p} \cdot \frac{n_k - 1}{n_k} \cdot \left(\underline{y}_{k_i} - \underline{g}_{k-i}\right)' W_{-i}^{-1} \left(\underline{y}_{k_i} - \underline{g}_{k-i}\right) \in FS(p, n-r-p) \quad (B1)$$

Calculons maintenant les différents termes intervenant dans (B1).

B2. Calcul de $\underline{g}_{k-i}$

On a :

$$\underline{g}_{k-i} = \sum \left\{ \underline{y}_{k_j} \mid j = 1, \cdots, n_k; j \neq i \right\} \Big/ (n_k - 1) = \left(n_k \underline{g}_k - \underline{y}_{k_i}\right) \Big/ (n_k - 1)$$

d'où on déduit que :

$$\underline{y}_{k_i} - \underline{g}_{k-i} = n_k \left(\underline{y}_{k_i} - \underline{g}_k\right) \Big/ (n_k - 1)$$

$$\underline{g}_k - \underline{g}_{k-i} >= \left(\underline{y}_{k_i} - \underline{g}_k\right) \Big/ (n_k - 1)$$

On considère alors le vecteur $\underline{u}_i$ suivant, qui nous servira dans le calcul de W_{-i} :

$$\underline{u}_i = \sqrt{\frac{n_k - 1}{n_k}} \cdot (\underline{y}_{k_i} - \underline{g}_{k-i}) = \sqrt{\frac{n_k}{n_k - 1}} \cdot (\underline{y}_{k_i} - \underline{g}_k) = \sqrt{n_k \cdot (n_k - 1)} \cdot (\underline{g}_k - \underline{g}_{k-i}) \quad (B2)$$

On peut noter que compte tenu de l'indépendance entre $\underline{y}_{k_i}$ et $\underline{g}_{k-i}$,

$$\underline{u}_i \in \mathcal{N}(\underline{0}\,;\, \Sigma).$$

B3. Calcul de V_{k-i} et W_{-i}

On a en appliquant le théorème de KOENIG et en tenant compte de (B2) :

$$\begin{aligned}(n_k-1)\cdot V_{k-i} &= \sum\left\{\left(\underline{y}_{kj}-\underline{g}_{k-i}\right)\left(\underline{y}_{kj}-\underline{g}_{k-i}\right)' \mid j=1,\cdots,n_k; j\neq i\right\}\\ &= \sum\left\{\left(\underline{y}_{kj}-\underline{g}_k\right)\left(\underline{y}_{kj}-\underline{g}_k\right)' \mid j=1,\cdots,n_k\right\}\\ &\quad -\left(\underline{y}_{ki}-\underline{g}_k\right)\left(\underline{y}_{ki}-\underline{g}_k\right)'-(n_k-1)\cdot\left(\underline{g}_k-\underline{g}_{k-i}\right)\left(\underline{g}_k-\underline{g}_{k-i}\right)'\\ &= n_k\cdot V_k-\underline{u}_i\cdot\underline{u}'_i\end{aligned}$$

On a alors :

$$(n-1)\cdot W_{-i}=\sum\{n_m V_m \mid m=1,\cdots,r; m\neq k\}+(n_k-1)V_{k-i}=nW-\underline{u}_i\cdot\underline{u}'_i$$

soit :

$$W_i=\frac{n}{n-1}\cdot W-\frac{\underline{u}_i\cdot\underline{u}'_i}{n-1}=\frac{n}{n-1}\cdot\left(W-\frac{\underline{u}_i\cdot\underline{u}'_i}{n}\right)\qquad (B3)$$

B4. Calcul de la quantité $S=\left(\underline{y}_{ki}-\underline{g}_{k-i}\right)'W_i^{-1}\left(\underline{y}_{ki}-\underline{g}_{k-i}\right)$ intervenant dans (B1), puis calcul de F

Compte tenu de (B2) et (B3), on a :

$$S=\frac{n-1}{n}\cdot\frac{n_k}{n_k-1}\cdot\underline{u}'_i\cdot\left(W-\frac{\underline{u}_i\cdot\underline{u}'_i}{n}\right)^{-1}\cdot\underline{u}_i$$

Or, il est immédiat de vérifier, en multipliant par $\left(W-\frac{\underline{u}_i\cdot\underline{u}'_i}{n}\right)$ que :

$$\left(W-\frac{\underline{u}_i\cdot\underline{u}'_i}{n}\right)^{-1}\cdot\underline{u}_i=\frac{W^{-1}\underline{u}_i}{1-\underline{u}'_i\cdot W^{-1}\cdot\underline{u}_i/n}$$

d'où :

$$S=\frac{n-1}{n}\cdot\frac{n_k}{n_k-1}\cdot\frac{\underline{u}'_i\cdot W^{-1}\underline{u}_i}{1-\underline{u}'_i\cdot W^{-1}\cdot\underline{u}_i/n}$$

Posant :

$$D^2\left(\underline{y}_{ki},\underline{g}_k\right)=\left(\underline{y}_{ki}-\underline{g}_k\right)'W^{-1}\left(\underline{y}_{ki}-\underline{g}_k\right)$$

on a, compte tenu de (B2) :

$$\underline{u}'_i\cdot W^{-1}\cdot\underline{u}_i=\frac{n_k}{n_k-1}\cdot D^2\left(\underline{y}_{ki},\underline{g}_k\right)$$

d'où :

$$S = \frac{n-1}{n} \cdot \left(\frac{n_k}{n_k - 1}\right)^2 \cdot \frac{D^2\left(\underline{y}_{k_i}, \underline{g}_k\right)}{1 - \frac{n_k}{n_k - 1} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) \Big/ n}$$

On déduit alors de (B1) que :

$$F = \frac{n-r-p}{n \cdot p} \cdot \frac{n_k}{n_k - 1} \cdot \frac{D^2\left(\underline{y}_{k_i}, \underline{g}_k\right)}{1 - \frac{n_k}{n_k - 1} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right) \Big/ n} \qquad (B4)$$

Remarques

i) L'expression (B4) correspond à la formule (6.2.4.) de [McLachlan, 1992] reprenant la formule (2) de [Hawkins, 1981], ces formules étant données sans démonstration détaillée.

ii) Comme d'après (B1), F suit une loi de Fisher-Snedecor à p et $(n-r-p)$ degrés de liberté, $\frac{p}{n-r-p} \cdot F$, qui s'écrit sous la forme $\frac{y}{1-y}$ avec $y = \frac{n_k}{(n_k-1) \cdot n} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right)$, suit une loi beta de type II de paramètres $p/2$ et $(n-r-p)/2$, d'où l'on déduit que $y = \frac{n_k}{(n_k-1) \cdot n} \cdot D^2\left(\underline{y}_{k_i}, \underline{g}_k\right)$, suit une loi beta de type I de paramètres $p/2$ et $(n-r-p)/2$.

Dans le cas particulier où l'on a un seul groupe ($r=1$, $n_k = n$ $\underline{y}_{k_i} = \underline{y}_i$, $\underline{g}_k = g$), on retrouve un résultat classique donné par [Tracy et al., 1997], à savoir que $D^2(\underline{y}_i, \underline{g})/(n-1)$ suit une loi beta de type I de paramètres $p/2$ et $(n-p-1)/2$.

iii) On peut noter que la quantité y définie ci-dessus n'est autre que $1 - \Lambda$ où Λ est le lambda de Wilks défini par :

$$\Lambda = \det\left((n-1) \cdot W_{-i}\right) \Big/ \det(nW)$$

En effet, compte tenu de (B3), puis de (B2), on a :

$$\Lambda = \frac{\det\left(nW - \underline{u}_i \cdot \underline{u}_i'\right)}{\det(nW)} = \det\left((nW - \underline{u}_i \cdot \underline{u}_i')(nW)^{-1}\right)$$

$$= \det\left(I_n - \frac{\underline{u}_i \cdot \underline{u}_i' \cdot W^{-1}}{n}\right) = 1 - \frac{\underline{u}_i' \cdot W^{-1} \cdot \underline{u}_i}{n} = 1 - y$$

puisque le déterminant de $I_n - \dfrac{\underline{u}_i \cdot \underline{u}_i' \cdot W^{-1}}{n}$ est égal au produit des valeurs propres qui sont :

- $1 - \dfrac{\underline{u}_i' \cdot W^{-1} \cdot \underline{u}_i}{n}$, associé au vecteur propre $\underline{u}_i$,
- 1 avec la multiplicité $n-1$, associé au sous-espace propre orthogonal (pour la métrique W^{-1}) à $\underline{u}_i$.

Le test basé sur F est donc équivalent au test basé sur le lambda de WILKS, résultat classique avec un seul groupe ($r = 1$) qu'on peut, par exemple, trouver dans [HAWKINS, 1980].

Bibliographie

CAZES, P., GOUPIL, F., GOUPIL, J., LAMBERT, M., PARDOUX, C. (1997) *Contrôle de qualité multivarié à l'aide du D^2 de* MAHALANOBIS, XXIX^e Journées de l'ASU, Carcassonne.

HAWKINS, D.M. (1980) *Identification of outliers*, 198p., Chapman and Hall.

HAWKINS, D.M. (1981) "A Nnew Test for Multivariate Normality and Homoscedasticity", *Technometrics, vol. 23, n°1, 105-110.*

MCLACHLAN, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*, 526 p., John Wiley & Sons.

SAPORTA, G. (1990) *Probabilités, analyse des données et statistique*, 530p., Technip.

TRACY, N.D., YOUNG, J.C., MASON, R.L. (1997) "Some Aspects of Hotelling's T^2 Statistic for Multivariate Quality Control", *in Statistics of Quality*, édité par S.GHOSH, W.R.SCHUCANY, W.B.SMITH, Dekker, 77

REMERCIEMENT

Les auteurs remercient Jean-Jacques DAUDIN pour avoir relu l'article et suggéré des références bibliographiques complémentaires.