La classification de données quantitatives avec SPAD

SPAD effectue toujours une ACP de la matrice des données quantitatives $X_{n\times p}$ avant de faire la classification des *n* individus. Les méthodes de classification s'appliquent alors non pas à la matrice initiale $X_{n\times p}$, mais à la matrice $\Psi_{n\times q}$ des $q \leq p$ premières composantes principales.

Remarques :

- Dans le cas de données qualitatives, SPAD effectue une AFCM (Analyse Factorielle des Correspondances multiples) et applique la classification à la matrice des *q* premières composantes principales. Si elles sont mixtes pour pouvoir faire l'AFCM, il faut au préalable recoder les variables quantitatives en variables qualitatives.
- Il est équivalent d'appliquer la CAH de Ward ou l'algorithme des centres mobiles :
  - aux individus décrits par X (resp. la matrice centrée-réduite Z)
  - aux individus décrits par Ψ où Ψ est la matrice de toutes les composantes principales ($q=p$) issues de l'ACP non normée (resp. l'ACP non normée).

En général, avec SPAD on choisit d'appliquer la méthode de classification à Ψ avec $q < p$. Le nombre de composantes conservées peut-être choisi de nombreuses manières. Souvent, il est choisi aux vues de l'éboulis (parfois appelée histogramme) des valeurs propres, du pourcentage d'inertie expliquée par les *q* composantes, ou encore en fonction du nombre de valeurs propres plus grandes que 1.

1 Classification hiérarchique de Ward et classification mixte

Dans SPAD la méthode de classification est toujours appliquée à la matrice $\Psi_{n\times q}$ des $q \leq p$ premières composantes principales issue d'une ACP normée ou d'une ACP non normée.

Deux méthodes de classification sont proposées :

- La classification hiérarchique de Ward (RECIP)
- La classification mixte (SEMIS)

La classification mixte utilise conjointement la CAH de Ward et l'algorithme des centres mobiles afin de chercher à réunir leurs avantages et palier à leurs inconvénients :

- La CAH de Ward fournit en général une partition moins bonne (au sens du critère d'inertie intra-classe) que les centres mobiles. De plus, si une mauvaise agrégation a été effectuée à une étape, cela se répercute sur toutes les partitions suivantes.
- Les inconvénients de la méthode des centres mobiles sont :
  - le nombre de classes fixées au départ
  - la partition finale dépend des choix initiaux

En revanche

- La méthode des centres mobiles a l'avantage d'être rapide et de permettre de traiter de grands jeux de données.
- La CAH fournit une suite de partitions emboîtées et donne une indication du nombre de classes.

La classification mixte dans SPAD procède donc de la manière suivante :

- Première étape : On définit d'abord une partition de l'ensemble des individus, en un nombre K de classes plus grand que le nombre de classes que l'on cherche. Pour trouver cette partition, SPAD propose deux possibilités :
    - appliquer les centres mobiles pour trouver cette partition en K classes. Par exemple, si l'on a n=10000 individus, on peut prendre K=100 par exemple.
    - trouver L partitions en K classes par la méthode des centres mobiles, en changeant à chaque fois les individus tirés au hasard dans l'étape d'initialisation. Définir ensuite les *groupements stables* c'est à dire les sous-ensembles d'individus qui ont toujours été dans la même classe dans les L partitions. Si L=2, on considère deux partitions P^1, P^2 en K classes et on obtient la partition-produit ayant $K \times K$ classes. La classe $C_{kk'}$ de cette partition produit contient les individus appartenant à la classe k de P^1, et à la classe k' de P^2. Les classes de la partition « produit » contenant au moins un individus constituent les groupements stables. En pratique, L=2 deux partitions suffisent pour déterminer les groupements stables. Dans l'exemple ci-dessous, 5 individus ont été classés dans la classe 1 de la partition 1 et dans la classe 1 de la partition 2,…. . et il y a 9 groupements stables

	Partition 1			
Partition 2		38	35	40
	30	5	25	0
	43	30	8	5
	40	3	2	25

    Si l'on a n=10000 individus par exemple que l'on veut un partition ayant au plus 100 groupement stable, on applique deux fois les centres mobiles pour trouver deux partitions en K=10 classes.
- Deuxième étape : On applique ensuite la CAH sur les classes de la partition obtenue à l'étape précédente. Pour cela, un nouveau tableau de données est construit ou chaque ligne est le centre de gravité des individus d'une classe (calculés sur la matrice $\Psi_{n \times q}$). Il y a donc autant de lignes que de classes. Chaque ligne (représentant une classe) est pondérée par la somme des poids des individus de la classe.

Les singletons de la hiérarchie trouvée par classification mixte ne sont donc pas les n individus de départ mais les centres de gravités classes obtenues à la première étape. Cette première étape est une étape de réduction du nombre d'individus. Cette étape est particulièrement utile lorsque le nombre d'individus au départ est grand. En effet, le dendrogramme d'une hiérarchie est alors vite illisible.

2 Coupure du dendrogramme

Si on veut maintenant sélectionner des partitions issues de cette hiérarchie, on applique la méthode PARTI-DECLA.

Cette méthode permet :

- soit de choisir les coupures qui nous intéressent : il faut spécifier le nombre de classes des partitions que l'on souhaite étudier.
- soit de laisser le logiciel définir automatiquement ces coupures. Il faut juste lui dire combien on veut de partitions et il définira alors automatiquement leurs nombres de classes.

En pratique, il vaut souvent mieux regardez successivement les partition en 2, 3,… classes et s'arrêter lorsqu'on ne trouve plus de bonne interprétation.

Ensuite, pour chaque partition ainsi retenue, la méthode PARTI-DECLA va consolider la partition. Pour cela, l'algorithme des centres mobiles est appliqué en prenant comme centres initiaux les centres de gravité des classes de cette partition. Une nouvelle partition dont le pourcentage d'inertie expliquée est nécessairement supérieur ou égale à celui de la partition de Ward, est ainsi obtenue.
Pour chaque partition après consolidation, la méthode peut fournir :

- pour chaque classe, les *parengons* des classes (dont le nombre est à fixer par l'utilisateur). Les *parengons* d'une classe sont les individus les plus proches de son centre de gravité. En ce sens, ce sont des individus représentatifs de la classe,
- l'interprétation de ses classes en fonctions des variables (voir section suivante).

3 Interprétation des classes d'une partition

On peut interpréter une classe d'une partition :

- à partir du centre de gravité de la classe ou des parangons,
- à partir des variables.

Si on veut interpréter les classes d'une partition en terme de variable, on peut fournir une interprétation

- unidimensionnelle des classes à partir des variables de l'analyse. C'est ce qui est fait avec la méthode PARTI-DECLA de SPAD
- multidimensionnelle des classes à partir des variables. Dans ce cas, la partition définit une nouvelle variable dite variable de classe, qui joue le rôle de la variable à expliquer dans une analyse discriminante. On peut entre autre effectuer :
  - Une méthode de segmentation comme CART (approche non paramétrique)
  - Une analyse discriminante (linéaire ou quadratique) à partir des composantes factorielles, et revenir ensuite aux variables initiales

Ici, nous nous intéressons uniquement à l'interprétation unidimensionnelle, les méthodes multidimensionnelles de segmentation ou d'analyse discriminante faisant l'objet d'un autre cours.

L'objectif de la description unidimensionnelle des classes d'une partition dans SPAD consiste à définir les variables et les modalités qui caractérisent une classe. On peut caractériser les classes par les variables illustrative **ou** par les variables actives.

La méthode PARTI-DECLA permet donc de caractériser les classes par les modalités des variables qualitatives ou par les variables quantitatives. Dans les deux cas (quantitatif ou qualitatif), on compare la moyenne (ou la fréquence) d'une variable (ou d'une modalité) sur la classe et dans l'échantillon global.

3.1 *Caractérisation d'une classe par les modalités des variables qualitatives*

3.1.1 Définition d'une valeur test

On définit la variable N_{ks}=nombre d'individus de la classe C_k, ayant la modalité s. Sous l'hypothèse H_0 que les individus qui constituent la classe soient tirés au hasard et sans remise

dans l'échantillon global, cette variable suit une distribution hypergéométrique d'espérance et de variance :

- $E(N_{ks}) = n_k \frac{n_s}{n}$
- $\sigma^2(N_{ks}) = n_k \frac{n - n_s}{n - 1} \frac{n_s}{n} \left(1 - \frac{n_s}{n}\right)$

où n est le nombre d'individus, n_k est l'effectif de la classe C_k et n_s est le nombre d'individus possédant la modalité s.

Cette distribution peut être approximée par une distribution normale si les effectifs des classes sont assez élevés. On a alors la statistique :

$t(N_{ks}) = \frac{N_{ks} - E(N_{ks})}{\sigma(N_{ks})}$ qui suit une loi normale centrée réduite

On calcule alors la p-valeur :

$p(s) = P(|t(N_{ks})| > t(n_{ks}))$

où n_{ks} est le nombre d'individus de la classe C_k possédant la modalité s.

La valeur $t(n_{ks})$ est appelée la *valeur test* de la modalité s dans la classe C_k.

Plus $t(n_{ks})$ est grand en valeur absolue, plus cette probabilité est faible et plus l'hypothèse H_0 d'un tirage aléatoire est rejeté et donc plus la modalité s est caractéristique de la classe C_k. Enfin, plus $|t(n_{ks})|$ est grand $t(n_{ks})$ est positif, plus $N_{ks} > E(N_{ks})$ et plus on dira que la modalité s est sur-représentée dans la classe C_k. A l'inverse, plus $|t(n_{ks})|$ est grand $t(n_{ks})$ est négatif, plus $N_{ks} < E(N_{ks})$ et plus on dira que la modalité s est sous-représentée dans la classe C_k.

3.1.2 Lecture des résultats

Le listing de la méthode PARTI-DECLA donne pour chaque classe des partitions la liste des modalité (ordonnée par valeurs de $t(n_{ks})$ décroissant) pour lesquelles $|t(n_{ks})|$ est supérieur à un certain seuil que l'on peut modifier dans les paramètres de la méthode. On a ainsi une sélection des modalités qui caractérisent bien la classe.

Pour chacune de ces modalités, le listing donne comme résultat :

- "MOD/CLA" = proportion d'individus de la classe qui possèdent cette modalité
  $= \frac{n_{ks}}{n_k}$
- "CLA/MOD" = proportion d'individus qui possèdent cette modalité qui se trouvent dans cette classe
  $= \frac{n_{ks}}{n_s}$

Une modalité caractérise d'autant mieux une classe que ces deux indicateurs sont grands, simultanément. En effet, si 100% des individus de la classe possèdent cette modalité ("MOD/CLA"=100%), ont peut retrouver cette modalité dans d'autres classes. Pour mesurer cela, on regarde quel est le pourcentage d'individus possédant cette modalité qui se trouvent dans la classe. Ainsi, si "CLA/MOD"=100%, la classe contient tous les individus ayant la modalité.

3.2 Caractérisation d'une classe par une variable quantitative

On calcule là aussi pour chaque variable quantitative une valeur test définie de la manière suivante :

On appelle Y la variable quantitative. Sous l'hypothèse H_0 que les individus qui constituent la classe soient tirés au hasard et sans remise dans l'échantillon global (i.e. la moyenne dans la classe est « égale » à la moyenne globale), la variable $\overline{Y}_k$ = moyenne de Y dans la classe C_k, a une espérance et variance égale à :

- $E(\overline{Y}_k) = \overline{Y}$
- $\sigma_k^2(Y) = \dfrac{n - n_k}{n - 1} \dfrac{\sigma^2(Y)}{n_k}$

où $\sigma^2(Y)$ est la variance empirique de Y.

D'après le théorème central limite, la statistique :

$t(\overline{Y}_k) = \dfrac{\overline{Y}_k - \overline{Y}}{\sigma_k(Y)}$ suit une normale centrée réduite

On calcule alors la p-valeur :

$$p(\overline{y}_k) = P(\left|t(\overline{Y}_k)\right| > t(\overline{y}_k))$$

où $\overline{y}_k$ est la moyenne de la variable Y dans la classe C_k.

La valeur $t(\overline{y}_k)$ est appelée la *valeur test* de la variable Y dans la classe C_k.

On raisonne ainsi exactement comme dans le cas de la caractérisation d'une classe par une modalité. Plus la valeur test $t(\overline{y}_k)$ est grande (en valeur absolue) et plus la probabilité est petite, plus la variable continue caractérise la classe. Une valeur test négative indique que la moyenne de cette variable dans la classe est plus faible que la moyenne dans tout l'échantillon. Et vice et versa.

Le listing de la méthode PARTI-DECLA donne pour chaque classe des partitions la liste (ordonnée par valeurs de $t(\overline{y}_k)$ décroissant) des variables pour lesquelles $\left|t(\overline{y}_k)\right|$ est supérieur à un certain seuil que l'on peut modifier dans les paramètres de la méthode. Pour chacune de ces variables sélectionnées, on lit également les valeurs suivantes : sa moyenne dans la classe et dans tout l'échantillon, idem pour l'écart-type.

Remarque 1 : Lorsque les variables sont actives, on ne peut pas donner d'interprétation statistique aux valeurs test calculées puisque ces variables ont participé à la construction de ces classes. Les valeurs test permettent néanmoins d'opérer un tri sur les variables actives et s'interprètent comme des écarts entre les valeurs prises par cette variable dans la classe, et dans tout l'échantillon.

Remarque 2 : La méthode PARTI-DECLA permet d'autres types de caractérisations comme par exemple, la caractérisation des classes par les axes factoriels. Cela permet de déterminer les directions dans lesquels les classes sont bien représentées.

Référence :

J.P. Nakacha, J. Confais, "Approche pragmatique de la classification", Dunod.