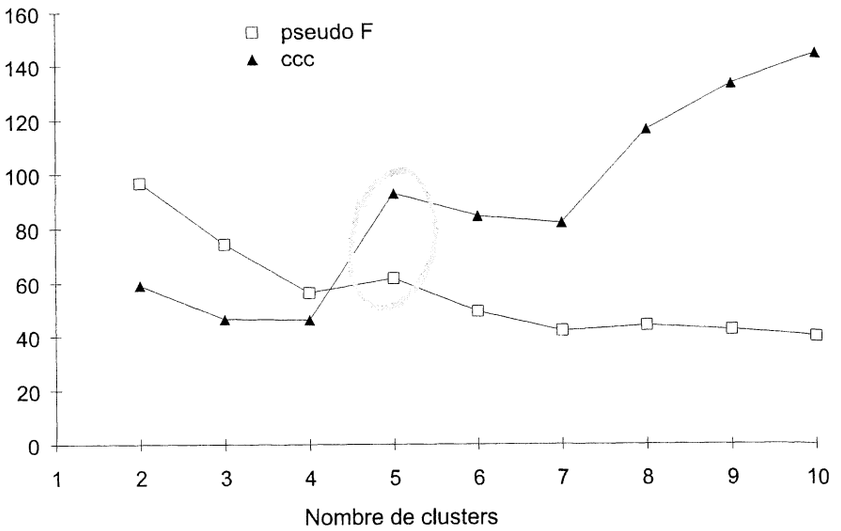

Fig. 3.7 — *Indices permettant de déterminer un nombre optimal de clusters (ce nombre correspond à un maximum local des courbes).*

Les deux courbes présentent chacune un seul maximum local, ce dernier est obtenu dans les deux cas pour n = 5.

4.
Méthodes de segmentation, CART

Les méthodes de segmentation, dont la méthode « CART » (*classification and regression tree*) est la plus utilisée, peuvent être considérées comme un compromis entre les méthodes de classification hiérarchiques et les modèles de régression.

Nous retrouvons en effet la dichotomie variable à expliquer / variables explicatives, mais, contrairement à un modèle de régression, l'objectif est ici de segmenter l'échantillon au moyen des variables explicatives, de façon que les segments obtenus soient le plus homogènes possible relativement à la variable à expliquer.

En médecine, un premier exemple d'application de ces méthodes pourrait être la constitution de catégories pronostiques en cancérologie, comme les classifications TNM ([1]). Ainsi, le choix optimal des seuils choisis pour les variables T, N et M pourrait-il reposer, au moins pour partie, sur une méthode de segmentation.

Un deuxième exemple porterait sur la constitution de groupes homogènes de patients en termes de coût de prise en charge. Une telle catégorisation pourrait ensuite être utilisée par les économistes de la santé pour proposer une rémunération des soins par types de patients traités et non par actes réalisés.

En quelques mots

La méthode CART procède par itérations successives. Sur un principe voisin de celui rencontré dans les méthodes de classification hiérarchique (voir p. 275), l'échantillon étudié est découpé dans un premier temps en deux sous-groupes homogènes, puis chaque sous-groupe ainsi obtenu est à son tour segmenté en deux parties, etc.

Sur un plan technique, il faut cependant définir formellement ce qu'est un « sous-groupe homogène » de sujets.

Prenons un exemple comprenant une variable à expliquer Y et une variable explicative X, toutes deux quantitatives (fig. 4.1). Si l'on examine les valeurs x_i prises par X, on ne constate aucun sous-groupe homogène de sujets. Il en est d'ailleurs de même pour les y_i, valeurs prises par Y.

[1] Si l'on prend l'exemple du cancer de la prostate, cette classification caractérise le stade de la maladie suivant que la tumeur (T) est cliniquement inapparente, limitée au tissu prostatique, envahissant la capsule ou fixée aux tissus avoisinants ; suivant qu'il y ait des ganglions régionaux envahis ou pas (N) ; et, enfin, suivant le niveau d'envahissement métastatique (M). En fonction du stade TNM le pronostic, et éventuellement la prise en charge thérapeutique, seront différents.

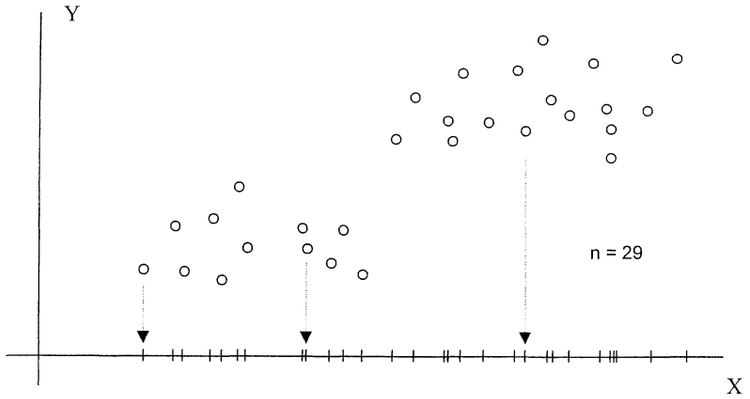

Fig. 4.1 — *Que ce soit selon X ou selon Y, il n'y a pas de regroupement évident des sujets en plusieurs « clusters ».*

On constate cependant qu'il existe une valeur x_{AB} de X déterminant deux groupes A et B de sujets (pour lesquels $X < x_{AB}$ et $X > x_{AB}$) plus homogènes relativement à la variable Y.

Formellement, x_{AB} est la valeur de X qui minimise :

$$[n_A Var(Y_A) + n_B Var(Y_B)] / n$$

(où A est l'ensemble des $X < x_{AB}$ et B l'ensemble des $X > x_{AB}$)

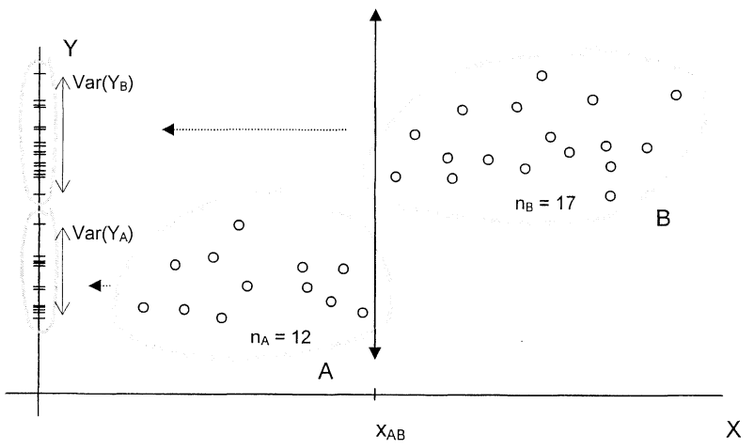

Fig. 4.2 — *Il existe une segmentation possible de l'échantillon selon une valeur x_{AB} de X qui conduit à deux sous-groupes de sujets plus homogènes pour la variable Y.*

Si l'on envisage maintenant le cas plus général d'une variable à expliquer quantitative Y et d'une liste de variables explicatives X_1, ..., X_p (quantitatives ou pas), l'algorithme CART fonctionne de la façon suivante.

– Pour chaque variable explicative X_i, détermination d'une valeur seuil optimale (optimale au sens d'une meilleure homogénéité de la variable Y, c'est-à-dire qui minimise la variance intra-sous-groupe : $[n_A Var(Y_A) + n_B Var(Y_B)] / n$). Pour les variables explicatives quantitatives ou qualitatives ordonnées, l'échantillon est découpé en deux suivant que X_i est inférieur ou pas à la valeur seuil. Pour les variables qualitatives non ordonnées, toutes les partitions en deux de l'échantillon de sujets sont étudiées.

– La variable ayant le seuil optimal qui conduit au découpage le plus homogène possible pour Y est retenue pour décider de la première segmentation.

– La même technique est ensuite utilisée pour chacun des deux sous-groupes obtenus lors de la première segmentation. Puis chacun des quatre sous-groupes sera de nouveau segmenté suivant un processus similaire, etc.

– On notera qu'à chaque segmentation, toutes les variables explicatives sont passées en revue. Une variable explicative peut donc être utilisée à plusieurs reprises, sur la base de seuils différents.

– Si l'on conserve toutes les segmentations proposées par l'algorithme le schéma peut être difficilement lisible. Il est donc généralement conseillé de terminer l'analyse par un « élagage » de l'arbre de segmentation. Il existe de nombreuses techniques d'élagage, chaque logiciel ayant généralement sa propre approche.

En pratique

Dans le chapitre « Données de survie » nous avons étudié les facteurs pronostiques de 65 patients atteints de myélome. La fonction « rpart » ([2]), de la librarie « rpart » de R, permet de générer des arbres de segmentation pour des variables à expliquer quantitatives, qualitatives ou censurées.

Nous obtenons ici ([3]) :

[2] Le document « Atkinson, A.J., Therneau, T.M. An Introduction to Recursive Partitioning Using the RPART Routines, Mayo Foundation, February 11, 2000 » est très utile pour l'utilisation de la fonction rpart. Il est disponible à l'adresse : http://www.mayo.edu/ hsr/ techrpt/rpartmini.pdf

[3] Les données et la syntaxe R de cet exemple sont disponibles sur le site Internet du livre.

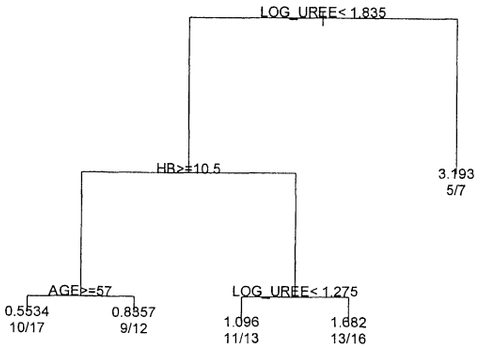

Fig. 4.3 — *Arbre de segmentation de facteurs pronostiques dans le myélome.*

Les cinq catégories proposées sont donc, par ordre croissant de bon pronostic :

- Log_urée ≥ 1,835 ;
- Log_urée < 1,835 et Hb < 10,5 et Log_urée ≥ 1,275 ;
- Log_urée < 1,835 et Hb < 10,5 et Log_urée < 1,275 (la première condition est donc redondante) ;
- Log_urée < 1,835 et Hb ≥ 10,5 et Age < 57 ;
- Log_urée < 1,835 et Hb ≥ 10,5 et Age ≥ 57.

Le logiciel est paramétré par défaut pour proposer un arbre élagué. Rien ne prouve cependant que cet élagage soit le plus pertinent. Pour vérifier ce point nous utilisons la fonction « printcp » ; nous obtenons :

```
Survival regression tree:
rpart(formula = Surv(T, DECES) ~ BENCE_J + LOG_UREE + CALCIUM +
    HB + AGE, method = "exp")

Variables actually used in tree construction:
[1] AGE      HB       LOG_UREE

Root node error: 75.03/65 = 1.1543
n= 65

        ❷       ❶                    ❸
        CP nsplit rel error xerror    xstd
1 0.108298      0   1.00000 1.0249 0.13479
2 0.086282      1   0.89170 1.2191 0.17824
3 0.017454      2   0.80542 1.1618 0.16903
4 0.011694      3   0.78797 1.1932 0.15378
5 0.010000      4   0.77627 1.1932 0.15378
```


En fonction de chaque segmentation ❶ nous avons l'évolution d'un indice d'adéquation, le « CP » ❷. En ❸ nous avons « xerror », un nombre moyen d'erreurs de classement calculé par une méthode de validation croisée ([4]). On constate que « xerror » est toujours supérieur à 1 alors que par construction il est standardisé à 1 avant toute segmentation… Les erreurs de classement ne sont donc pas réduites par la segmentation, c'est un mauvais signe pour sa pertinence.

Prenons pour l'exemple une segmentation en trois classes. Nous obtenons alors :

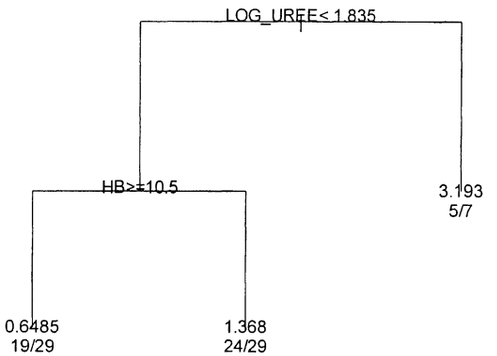

Fig. 4.4 — *Arbre de segmentation comptant trois classes.*

Il est possible en outre de tracer les courbes de survie pour chacune de ces trois classes de patients :

[4] 90 % des patients de l'échantillon sont prélevés, un arbre de segmentation est construit sur la base de ce sous-échantillon et les 10 % de patients restants sont utilisés pour évaluer les performances de la segmentation. Ce procédé est utilisé une dizaine de fois, « xerror » (❸) est alors le nombre moyen d'erreurs de classement constatées.

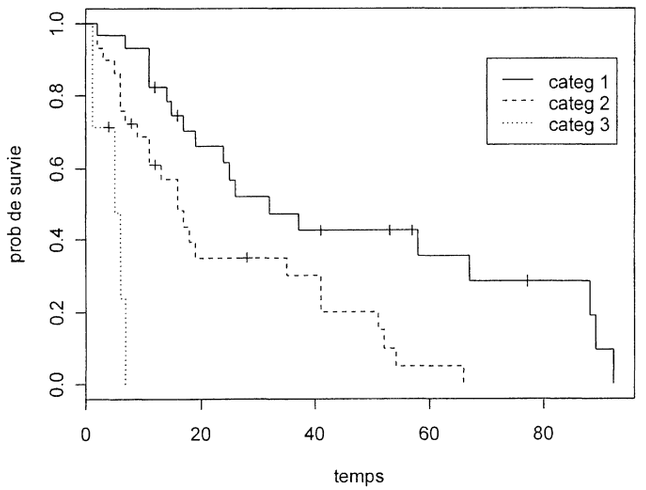

Fig. 4.5 — *Courbes de survie des trois classes de patients déterminées par l'algorithme de segmentation.*

5. Analyse en composantes principales

L'analyse en composantes principales (ACP) est une technique qui permet de faire la synthèse de l'information contenue dans un grand nombre de variables.

Les « composantes principales » sont de nouvelles variables, indépendantes, combinaisons linéaires des variables initiales, possédant une variance maximale. Ces nouvelles variables permettent parfois d'éclairer les mécanismes intimes mis en œuvre dans la genèse des données. Elles permettent aussi d'utiliser dans de meilleures conditions des techniques multivariées classiques comme la régression linéaire.

Les composantes principales autorisent en outre la représentation graphique de grands tableaux de données trop complexes à décrire par les méthodes graphiques habituelles. C'est incontestablement cette dernière propriété qui est à l'origine de leur large utilisation.

Pour des raisons essentiellement historiques, l'analyse en composantes principales est une méthode à l'origine de *quiproquos* encore tenaces dont le plus important est sûrement le lien équivoque qui la relie à l'analyse factorielle. Les relations entre ces deux méthodes ainsi que la notion de rotation ne seront abordées que dans le chapitre 9 (« Analyse factorielle »).

> En toute rigueur, une analyse en composantes principales ne nécessite aucune condition de validité. L'usage de variables qualitatives binaires ou ordonnées est cependant théoriquement à éviter, des techniques alternatives étant plus performantes (analyse des correspondances, *multidimensional scaling*).

Une réduction de dimensions

A la base de tout problème multivarié, se trouve un tableau de données, correspondant par exemple à n sujets chez lesquels sont mesurées p variables.

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$