Analyse Discriminante Décisionnelle

Résumé

Une variable qualitative Y à m modalités est modélisée par p variables quantitatives $X^j, j = 1, \ldots, p$. L'objectif est la prévision de la classe d'un ou de nouveaux individus sur lesquels les variables $X^j, j = 1, \ldots, p$ sont également observées. Différents modèles d'analyse discriminante décisionnelle sont considérés : règle linéaire et quadratique de décision dans le cas gaussien, règle non paramétrique par k plus proches voisins.

Retour au plan du cours

1 Introduction

Il s'agit de la modélisation d'une variable qualitative Y à m modalités par p variables quantitatives $X^j, j = 1, \ldots, p$ observées sur un même échantillon Ω de taille n. L'objectif de l'analyse discriminante décisionnelle déborde le simple cadre descriptif de l'analyse factorielle discriminante (AFD). Disposant d'individus sur lesquels les X^j sont observées mais pas Y, il s'agit de *décider* de la modalité $\mathcal{T}_\ell$ de Y (ou de la classe correspondante) de ces individus. L'ADD s'applique donc également à la situation précédente de la régression logistique ($m = 2$) mais aussi lorsque le nombre de classes est plus grand que 2. Les variables explicatives devant être quantitatives, celles qualitatives sont remplacées par des indicatrices.

L'objectif est de définir des *règles de décision* (ou d'affectation); $\mathbf{x} = (x^1, \ldots, x^p)$ désigne les observations des variables explicatives sur un individu, $\{\mathbf{g}_\ell; \ell = 1, \ldots, m\}$ les barycentres des classes calculés sur l'échantillon et $\overline{\mathbf{x}}$ le barycentre global.

La matrice de covariance empirique se décompose en

$$\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r.$$

où $\mathbf{S}_r$ est appelée variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}'\mathbf{D}\overline{\mathbf{X}}_r = \sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell} w_i(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)',$$

et $\mathbf{S}_e$ la variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}'\mathbf{D}\overline{\mathbf{G}} = \overline{X}_e'\mathbf{D}\overline{X}_e = \sum_{\ell=1}^{m}\overline{w_\ell}(\mathbf{g}_\ell - \overline{\mathbf{x}})(\mathbf{g}_\ell - \overline{\mathbf{x}})'.$$

2 Règle de décision issue de l'AFD

2.1 Cas général : *m* quelconque

DÉFINITION 1. — *L'individu x est affectée à la modalité de Y minimisant :*

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell), \ell = 1, \ldots, m.$$

Cette distance se décompose en

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell) = \|\mathbf{x} - \mathbf{g}_\ell\|^2_{\mathbf{S}_r^{-1}} = (\mathbf{x} - \mathbf{g}_\ell)'\mathbf{S}_r^{-1}(\mathbf{x} - \mathbf{g}_\ell)$$

et le problème revient donc à maximiser

$$\mathbf{g}_\ell'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_\ell'\mathbf{S}_r^{-1}\mathbf{g}_\ell.$$

Il s'agit bien d'une règle linéaire en $\mathbf{x}$ car elle peut s'écrire : $\mathbf{A}_\ell\mathbf{x} + \mathbf{b}_\ell$.

2.2 Cas particulier : *m*=2

Dans ce cas, la dimension r de l'AFD vaut 1. Il n'y a qu'une seule valeur propre non nulle λ_1, un seul vecteur discriminant v^1 et un seul axe discriminant Δ_1. Les 2 barycentres $\mathbf{g}_1$ et $\mathbf{g}_2$ sont sur Δ_1, de sorte que v^1 est colinéaire à $\mathbf{g}_1 - \mathbf{g}_2$.

L'application de la règle de décision permet d'affecter $\mathbf{x}$ à $\mathcal{T}_1$ si :

$$\mathbf{g}_1'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_1'\mathbf{S}_r^{-1}\mathbf{g}_1 > \mathbf{g}_2'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_2'\mathbf{S}_r^{-1}\mathbf{g}_2$$

c'est-à-dire encore si

$$(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{S}_r^{-1}\mathbf{x} > (\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{S}_r^{-1}\frac{\mathbf{g}_1 + \mathbf{g}_2}{2}.$$

Remarque

La règle de décision liée à l'AFD est simple mais elle est limitée et insuffisante notamment si les variances des classes ne sont pas identiques. De plus, elle ne tient pas compte de l'échantillonnage pour $\mathbf{x}$: tous les groupes n'ont pas nécessairement la même probabilité d'occurrence.

3 Règle de décision bayésienne

3.1 Introduction

La variable Y, qui indique le groupe d'appartenance d'un individu, prend ses valeurs dans $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ et est munie d'une loi de probabilité $\pi_1, \ldots, \pi_m$. Les probabilités $\pi_\ell = P[\mathcal{T}_\ell]$ représentent les probabilités *a priori* des classes ou groupes ω_ℓ. Les vecteurs $\mathbf{x}$ des observations des variables explicatives sont supposés suivent, connaissant leur classe, une loi de densité

$$h_\ell(\mathbf{x}) = P[\mathbf{x} \mid \mathcal{T}_\ell]$$

par rapport à une mesure de référence [1].

3.2 Définition

Une règle de décision est une application δ de Ω dans $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ qui, à tout individu, lui affecte une classe connaissant $\mathbf{x}$. Sa définition dépend du contexte de l'étude et prend en compte la

- connaissance ou non de coûts de mauvais classement,
- connaissance ou non des lois *a priori* sur les classes,
- nature aléatoire ou non de l'échantillon.

Soit $c_{\ell \mid k}$ le coût du classement dans $\mathcal{T}_\ell$ d'un individu de $\mathcal{T}_k$. Le *risque de Bayes* d'une règle de décision δ exprime alors le coût moyen :

$$R_\delta = \sum_{k=1}^{m} \pi_k \sum_{\ell=1}^{m} c_{\ell \mid k} \int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} h_k(\mathbf{x}) d\mathbf{x}$$

où $\int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} h_k(\mathbf{x}) d\mathbf{x}$ représente la probabilité d'affecté $\mathbf{x}$ à $\mathcal{T}_\ell$ alors qu'il est dans $\mathcal{T}_k$.

1. La mesure de Lebesgues pour des variables réelles, celle de comptage pour des variables qualitatives

3.3 Coûts inconnus

L'estimation des coûts n'est pas du ressort de la Statistique et, s'ils ne sont pas connus, on suppose simplement qu'ils sont tous égaux. La minimisation du risque ou règle de Bayes revient alors à affecter tout $\mathbf{x}$ à la classe la plus probable c'est-à-dire à celle qui maximise la probabilité conditionnelle *a posteriori* : $P[\mathcal{T}_\ell \mid \mathbf{x}]$. Par le théorème de Bayes :

$$P[\mathcal{T}_\ell \mid \mathbf{x}] = \frac{P[\mathcal{T}_\ell \text{ et } \mathbf{x}]}{P[\mathbf{x}]} = \frac{P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]}{P[\mathbf{x}]}$$

avec le principe des probabilités totales : $P[\mathbf{x}] = \sum_{\ell=1}^{m} P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]$.

Comme $P[\mathbf{x}]$ ne dépend pas de ℓ, la règle consiste à choisir $\mathcal{T}_\ell$ maximisant

$$P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell] = \pi_\ell.P[\mathbf{x} \mid \mathcal{T}_\ell];$$

$P[\mathbf{x} \mid \mathcal{T}_\ell]$ est la probabilité d'observer $\mathbf{x}$ au sein de la classe $\mathcal{T}_\ell$. Pour une loi discrète, il s'agit d'une probabilité du type $P[\mathbf{x} = \mathbf{x}_k^l \mid \mathcal{T}_\ell]$ et d'une densité $h(\mathbf{x} \mid \mathcal{T}_\ell)$ pour une loi continue amis dans tous les cas la notation $h_\ell(\mathbf{x})$ est utilisée.

La règle de décision s'écrit finalement sous la forme :

$$\delta(\mathbf{x}) = \arg \max_{\ell=1,\ldots,m} \pi_\ell h_\ell(\mathbf{x}).$$

3.4 Détermination des *a priori*

Les probabilités *a priori* π_ℓ peuvent effectivement être connues : proportions de divers groupes dans une population, de diverses maladies… ; sinon elles sont estimées sur l'échantillon d'apprentissage :

$$\widehat{\pi}_\ell = w_\ell = \frac{n_\ell}{n} \quad \text{(si tous les individus ont le même poids)}$$

à condition qu'il soit bien un échantillon aléatoire susceptible de fournir des estimations correctes des fréquences. Dans le cas contraire il reste à considérer tous les π_ℓ égaux.

3.5 Cas particuliers

- Dans le cas où les probabilités *a priori* sont égales, c'est par exemple le cas du choix de probabilités non informatives, la règle de décision

bayésienne revient alors à maximiser $h_\ell(\mathbf{x})$ qui est la vraisemblance, au sein de $\mathcal{T}_\ell$, de l'observation $\mathbf{x}$. La règle consiste alors à choisir la classe pour laquelle cette vraisemblance est maximum.

- Dans le cas où $m = 2$, $\mathbf{x}$ est affectée à $\mathcal{T}_1$ si :

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1},$$

faisant ainsi apparaître un rapport de vraisemblance. D'autre part, l'introduction de coûts de mauvais classement différents selon les classes amène à modifier la valeur limite π_2/π_1.

Finalement, il reste à estimer les densités conditionnelles $h_\ell(\mathbf{x})$. Les différentes méthodes d'estimation considérées conduisent aux méthodes classiques de discrimination bayésienne objets des sections suivantes.

4 Règle bayésienne avec modèle gaussien

Dans cette section, conditionnellement à $\mathcal{T}_\ell$, $\mathbf{x} = (x_1, \ldots, x_p)$ est supposée être l'observation d'un vecteur aléatoire gaussien $\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$; $\boldsymbol{\mu}_\ell$ est un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\Sigma}_\ell$ une matrice $(p \times p)$ symétrique et définie-positive. La densité de la loi, au sein de la classe $\mathcal{T}_\ell$, s'écrit donc :

$$h_\ell(\mathbf{x}) = \frac{1}{\sqrt{2\pi}(\det(\boldsymbol{\Sigma}_\ell))^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell)\right].$$

L'affectation de $\mathbf{x}$ à une classe se fait en maximisant $\pi_\ell.h_\ell(\mathbf{x})$ par rapport à l soit encore la quantité :

$$\ln(\pi_\ell) - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_\ell)) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell).$$

4.1 Hétéroscédasticité

Dans le cas général, il n'y a pas d'hypothèse supplémentaire sur la loi de $\mathbf{x}$ et donc les matrices $\boldsymbol{\Sigma}_\ell$ sont fonction de ℓ. Le critère d'affectation est alors *quadratique* en $\mathbf{x}$. Les probabilités π_ℓ sont supposées connues mais il est nécessaire d'estimer les moyennes $\boldsymbol{\mu}_\ell$ ainsi que les covariances $\boldsymbol{\Sigma}_\ell$ en maximisant, compte tenu de l'hypothèse de normalité, la vraisemblance. Ceci conduit à estimer la moyenne

$$\widehat{\boldsymbol{\mu}_\ell} = \mathbf{g}_\ell$$

par la moyenne empirique de $\mathbf{x}$ dans la classe l pour l'échantillon d'apprentissage et $\boldsymbol{\Sigma}_\ell$ par la matrice de covariance empirique $\mathbf{S}^*_{Rl}$:

$$\mathbf{S}^*_{Rl} = \frac{1}{n_\ell - 1}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

pour ce même échantillon.

4.2 Homoscédasticité

Les lois de chaque classe sont supposées partager la même structure de covariance $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$. En supprimant les termes indépendants de l, le critère à maximiser devient

$$\ln(\pi_\ell) - \frac{1}{2}\boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell + \boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}_\ell^{-1}\mathbf{x}$$

qui est cette fois *linéaire* en $\mathbf{x}$. Les moyennes $\boldsymbol{\mu}_\ell$ sont estimées comme précédemment tandis que $\boldsymbol{\Sigma}$ est estimée par la matrice de covariance intraclasse empirique :

$$\mathbf{S}^*_R = \frac{1}{n-m}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'.$$

Si, de plus, les probabilités π_ℓ sont égales, après estimation le critère s'écrit :

$$\overline{\mathbf{x}_\ell}'\mathbf{S}_R^{*-1}\mathbf{x} - \frac{1}{2}\overline{\mathbf{x}_\ell}'\mathbf{S}_R^{*-1}\overline{\mathbf{x}_\ell}.$$

qui coïncide avec le critère initial issu de l'AFD..

4.3 Commentaire

Les hypothèses : normalité, éventuellement l'homoscédasticité, doivent être vérifiées par la connaissance *a priori* du phénomène ou par une étude préalable de l'échantillon d'apprentissage. L'hypothèse d'homoscédasticité, lorqu'elle est vérifiée, permet de réduire très sensiblement le nombre de paramètres à estimer et d'aboutir à des estimateurs plus fiables car de variance moins élevée. Dans le cas contraire, l'échantillon d'apprentissage doit être de taille importante.

5 Règle bayésienne avec estimation non paramétrique

5.1 Introduction

En Statistique, l'estimation est non paramétrique ou fonctionnelle lorsque le nombre de paramètres à estimer est infini. L'objet statistique à estimer est alors une fonction par exemple de régression $y = f(x)$ ou encore une densité de probabilité h. Dans ce cas, au lieu de supposer que la densité est de type connu (gaussien) dont les paramètres sont estimés. c'est la fonction de densité h qui est directement estimée. Pour tout x de $\mathbb{R}$, $h(x)$ est donc estimée par $\widehat{h}(x)$.

Cette approche très souple a l'avantage de ne pas nécessiter d'hypothèse particulière sur la loi (seulement la régularité de h pour de bonnes propriétés de convergence), en revanche elle n'est applicable qu'avec des échantillons de grande taille d'autant plus que le nombre de dimensions p est grand (*curse of dimensionality*).

Dans le cadre de l'analyse discriminante, ces méthodes permettent d'estimer directement les densités $h_\ell(\mathbf{x})$. On considère ici deux approches : la méthode du noyau et celle des k plus proches voisins.

5.2 Méthode du noyau

Estimation de densité

Soit $y_1, \ldots, y_n$ n observations équipondérées d'une v.a.r. continue Y de densité h inconnue. Soit $K(y)$ (le *noyau*) une densité de probabilité unidimensionnelle (sans rapport avec h) et λ un réel strictement positif. On appelle estimation de h par la méthode du noyau la fonction

$$\widehat{h}(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{y - y_i}{\lambda}\right).$$

Il est immédiat de vérifier que

$$\forall y \in \mathbb{R}, \widehat{h}(y) \geq 0 \quad \text{et} \quad \int_{-\infty}^{+\infty} \widehat{h}(y)dy = 1;$$

λ est appelé *largeur de fenêtre* ou paramètre de *lissage* ; plus λ est grand, plus l'estimation $\widehat{h}$ de h est régulière. Le noyau K est choisi centré en 0, unimodal et symétrique. Les cas les plus usuels sont la densité gaussienne, celle uniforme sur $[-1, 1]$ ou triangulaire : $K(x) = [1 - |x|]\mathbf{1}_{[-1,1]}(x)$. La forme du noyau n'est pas très déterminante sur la qualité de l'estimation contrairement à la valeur de λ.

Application à l'analyse discriminante

La méthode du noyau est utilisée pour calculer une estimation non paramétrique de chaque densité $h_\ell(\mathbf{x})$ qui sont alors des fonctions définies dans $\mathbb{R}^p$. Le noyau K^* doit donc être choisi multidimensionnel et

$$\widehat{h_\ell}(\mathbf{x}) = \frac{1}{n_\ell \lambda^p} \sum_{i \in \Omega_\ell} K^*\left(\frac{\mathbf{x} - \mathbf{x}_i}{\lambda}\right).$$

Un noyau multidimensionnel peut être défini à partir de la densité usuelle de lois : multinormale $\mathcal{N}_p(0, \mathbf{\Sigma}_p)$ ou uniforme sur la sphère unité ou encore par produit de noyaux unidimensionnels :

$$K^*(\mathbf{x}) = \prod_{j=1}^{p} K(x^j).$$

5.3 *k* plus proches voisins

Cette méthode d'affectation d'un vecteur $\mathbf{x}$ consiste à enchaîner les étapes décrites dans l'algorithme ci-dessous.

Algorithme des k plus proches voisins (k-nn)

1. Choix d'un entier k : $1 \geq k \geq n$.
2. Calculer les distances $d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}_i)\,, \quad i = 1, \ldots, n$ où $\mathbf{M}$ est la métrique de Mahalanobis c'est-à-dire la matrice inverse de la matrice de variance (ou de variance intraclasse).
3. Retenir les k observations $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$ pour lesquelles ces distances sont les plus petites.
4. Compter les nombres de fois $k_1, \ldots, k_m$ que ces k observations appraissent dans chacune des classes.

TABLE 1 – Cancer : estimations des taux d'erreurs de prévision obtenus par différents types d'analyse discriminante

Méthode	apprentissage	validations croisée	test
linéaire	1,8	3,8	3,6
kNN	2,5	2,7	2,9

5. Estimer localement les densités conditionnelles par

$$\widehat{h_\ell}(\mathbf{x}) = \frac{k_\ell}{kV_k(\mathbf{x})};$$

où $V_k(\mathbf{x})$ est le volume de l'ellipsoïde $\{\mathbf{z}|(\mathbf{z}-\mathbf{x})'\mathbf{M}(\mathbf{z}-\mathbf{x}) = d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}_{(k)})\}$.

Pour $k = 1$, $\mathbf{x}$ est affecté à la classe du plus proche élément.

Comme toute technique, celles présentées ci-dessus nécessitent le réglage d'un paramètre (largeur de fenêtre ou nombre de voisins considérés). Ce choix s'apparente à un choix de modèle et nécessite le même type d'approche à savoir l'optimisation d'un critère (erreur de classement, validation croisée.

6 Exemples

6.1 Cancer du sein

Par principe, l'analyse discriminante s'applique à des variables explicatives quantitatives. Ce n'est pas le cas des données qui sont au mieux ordinales. Il est clair que construire une fonction de discrimination comme combinaison de ces variables n'a guère de sens. Néanmoins, en s'attachant uniquement à la qualité de prévision sans essayer de construire une interprétation du plan ou de la surface de discrimination, il est d'usage d'utiliser l'analyse discriminante de façon "sauvage". Les résultats obtenus sont résumés dans le tableau 1. L'analyse discriminante quadratique, avec matrice de variance estimée pour chaque classe n'a pas pu être calculée. Une des matrices n'est pas inversible.

TABLE 2 – Ozone : estimations des taux d'erreurs de prévision obtenus par différents types d'analyse discriminante

Méthode	apprentissage	validations croisée	test
linéaire	11,9	12,5	12,0
quadratique	12,7	14,8	12,5

TABLE 3 – Banque : estimations des taux d'erreurs de prévision obtenus par différents types d'analyse discriminante

Méthode	apprentissage	validations croisée	test
linéaire	16,5	18,3	18
quadratique	17,8	22,0	30
kNN	23,5	29,8	29

6.2 Concentration d'ozone

Dans cet exemple aussi, deux variables sont qualitatives : le type de jour à 2 modalités ne pose pas de problème mais remplacer la station par un entier est plutôt abusif. D'ailleurs, la méthode des plus proches voisins ne l'acceptent pas, une transformation des données serait nécessaire.

6.3 Carte visa

Comme pour les données sur le cancer, les données bancaires posent un problème car elles associent différents types de variables. Il est possible de le contourner, pour celles binaires, en considérant quantitative, l'indicatrice de la modalité (0 ou 1). Pour les autres, certaines procédures (DISQUAL pour discrimination sur variables qualitatives) proposent de passer par une analyse factorielle multiple des correspondances pour rendre tout quantitatif mais ceci n'est pas implémenté de façon standard dans les logiciels d'origine américaine.

Pour l'analyse discriminante, R ne propose pas de sélection automatique de variable mais inclut une estimation de l'erreur par validation croisée. Les résultats trouvés sont résumés dans le tableau 3. Seule une discrimination linéaire

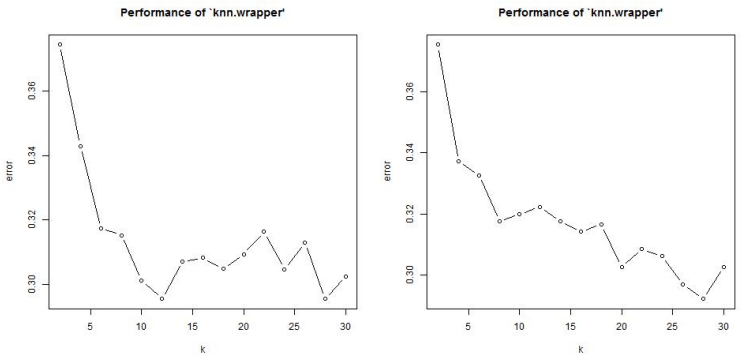

FIGURE 1 – *Banque : Deux exécutions de l'optimisation du choix de k par validation croisée.*

semble fournir des résultats raisonnables, la recherche d'une discrimination quadratique n'apporte rien pour ces données. De son côté, SAS propose une sélection automatique (procédure stepdisc) mais les résultats obtenus ne sont pas sensiblement meilleurs après sélection.

Le choix de k dans la méthode des k plus proches voisins est souvent délicat ; chaque exécution de l'estimation de l'erreur par validation croisée conduit à des résultats aléatoires et très différents et k optimal oscille entre 10 et 30 (fig. 1) !