Analyse discriminante par noyaux associés pour données mixtes

Sobom M. Somé & Célestin C. Kokonendji

Université de Franche-Comté
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS-UFC
16 route de Gray, 25030 Besançon cedex, France.
prenom.nom@univ-fcomte.fr

Résumé. L'objet de ce travail est de proposer une méthode non-paramétrique d'analyse discriminante pour des variables mixtes : continues, catégorielles et comptages. Après la présentation du modèle à l'aide des noyaux associés multiples composés, nous proposons deux types de validation croisées pour la sélection appropriée des matrices des fenêtres liée à chaque famille de données. En particulier, la classique est utilisée pour les données homogènes ayant la même mesure de référence ; tandis qu'une version profilée de validation croisée est introduite pour les données mixtes. Des études de simulations ainsi qu'une application aux données réelles seront présentées.

Mots-clés. Matrice des fenêtres, noyau non-classique, validation croisée profilée.

Abstract. The purpose of this communication is to propose a nonparametric method for discriminant analysis for mixed variables: continuous, categorical and count. After the presentation of the model with multiple of associated kernels, we propose two types of cross-validation for bandwidth matrices selection appropriated to each data family. In particular, the classical is used for homogeneous data with the same measure; while a profile version of cross-validation is introduced for mixed data. Simulations studies and an application to real data are provided.

Keywords. Bandwidth matrix, non-classical kernel, profile cross-validation.

1 Introduction

Nous nous intéressons à l'analyse discriminante non-paramétrique sur des données multivariés composées de variables continues (bornées ou non) et de variables discrètes (catégorielles ou dénombrement). Dans la suite, on notera par $\mathbb{T}_d$ ($\subseteq \mathbb{R}^d$) le support de ces données mixtes

$$\mathbb{T}_d = \mathbb{T}_{k_1}^{[1]} \times \cdots \times \mathbb{T}_{k_L}^{[L]} \text{ avec } \sum_{\ell=1}^{L} k_\ell = d \tag{1}$$

et par $\boldsymbol{\nu} = \nu_1 \times \cdots \times \nu_L$ la mesure de référence sur $\mathbb{T}_d$, où ν_ℓ est la mesure liée au support correspondant à $\mathbb{T}_{k_\ell}^{[\ell]}$ de (1). La règle de décision en analyse discriminante est de classer

n'importe quel observation $\mathbf{x}$ de $\mathbb{T}_d$ dans un des J *groupes* prédéfinies (J entier naturel non nul et fini). Précisement, la règle de Bayes assigne une observation $\mathbf{x}$ au *groupe* ayant la plus grande probabilité a priori. Elle est donnée par

$$\text{Allouer } \mathbf{x} \text{ au } \textit{groupe } j_0, \text{ où } j_0 = \arg \max_{j\in\{1,\ldots,J\}} \pi_j f_j(\mathbf{x}), \tag{2}$$

où les π_j sont les probabilités a priori connues et les $f_j(\mathbf{x})$ sont des fonctions de densité de probabilités (f.d.p.) inconnues par rapport à la mesure $\boldsymbol{\nu}$ des J *groupes* respectifs. Les noyaux classiques multivariés (e.g. gaussiens) ne sont adaptés que pour l'estimation des densités f_j de données non bornées (i.e. $\mathbb{R}^d$) ; voir Scott (1992) et Zougab et al. (2014). Pour l'utilisation du noyau gaussien multivarié en analyse discriminante, on peut se référer par exemple à Duong (2007), Gosh et Chaudhury (2004), Liu et al. (2009) et He et al. (2012). Racine et Li (2007) ont proposé, pour l'estimation de différentes fonctionnelles, des noyaux multiples composés de noyaux univariés gaussiens pour les variables continues et de noyaux d'Aitchison et Aitken (1976) pour les variables catégorielles ; voir aussi Hayfield et Racine (2007) pour des implémentations et utilisations de ces noyaux multiples sous le logiciel **R** (2015). Signalons que l'utilisation des noyaux gaussiens (i.e. symétriques) produisent des poids en dehors des variables à support bornés ou discrets. Dans le cas univarié continu, Chen (1999, 2000) est l'un des premiers à avoir proposé des noyaux asymétriques (e.g. bêta, gamma) dont les supports coïncident avec celles des densités à estimer ; voir aussi Malec et Schienle (2014) et Igarashi et Kakizawa (2015). De même, Libengué (2013) a étudié plusieurs familles de ces noyaux univariés qu'il a appelé noyaux associes univariés ; voir aussi Kokonendji et al. (2007), Kokonendji et Senga Kiéssé (2011), Zougab et al. (2012) pour les cas univariés discrets. Une version multivariée continue des noyaux associés a été étudié par Kokonendji et Somé (2015) pour des fonctions de densités continues et Somé et Kokonendji (2015) pour la régression multiple.

Pour l'estimation des densités f_j de (2), nous proposons des noyaux associés multiples composés de noyaux associés discrets univariés (e.g. binomial) et continues (e.g. bêta, gamma). Ces noyaux associés sont adaptés à cette situation de mélange d'axes car ils respectent le support de chaque variable. C'est pourquoi, dans ce qui suit, nous rappelons une définition des noyaux associés multivariés (discret, continu ou mixte) et présentons trois cas particuliers dont associé classique et associé multiple. Nous présentons alors la méthode d'analyse discriminante par noyaux associés en explicitant la procédure de validation croisée profilée. Des simulations et applications étudieront l'effet du type de noyau associé, noté $\boldsymbol{\kappa}$, en classification.

2 Analyse discriminante par noyaux associés

Considérons une suite d'échantillons ou données d'apprentissage $\mathcal{X}_j = \{\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}\}$, pour $j = 1, \ldots, J$, de densités inconnues f_j sur $\mathbb{T}_d$ et où les tailles d'échantillons n_j sont connues et non aléatoires. Un estimateur à noyau associé multivarié $\widehat{f_j}$ de f_j dans (2) est alors

simplement défini par

$$\widehat{f_j}(\mathbf{x}) = \frac{1}{n_j}\sum_{i=1}^{n_j} K_{\mathbf{x},\mathbf{H}_j}(\mathbf{X}_{ji}) = \widehat{f_j}(\mathbf{x};\boldsymbol{\kappa},\mathbf{H}_j), \quad \forall \mathbf{x} \in \mathbb{T}_d \subseteq \mathbb{R}^d, \tag{3}$$

où $\mathbf{H}_j$ est une matrice des fenêtres d'ordre $d \times d$ (i.e. symétrique et définie positive) telle que $\mathbf{H}_j \equiv \mathbf{H}_{n_j} \to \mathbf{0}$ quand $n_j \to +\infty$, et $\boldsymbol{\kappa}$ est le type de noyau associé $K_{\mathbf{x},\mathbf{H}_j}(\cdot)$, paramétré par $\mathbf{x}$ et $\mathbf{H}_j$. Sans perte de généralité, nous utiliserons dans la suite $\widehat{f_j}(\mathbf{x};\boldsymbol{\kappa},\mathbf{H}_j) \equiv \widehat{f_j}(\mathbf{x};\boldsymbol{\kappa})$ puisque la matrice de lissage se fait ici par validation croisée (classique et profilée). Pour toute matrice de lissage $\mathbf{H}$, ce noyau associé $K_{\mathbf{x},\mathbf{H}}(\cdot)$ est une f.d.p. par rapport à la mesure $\boldsymbol{\nu}$, et est défini comme suit.

Définition 2.1 *(Somé et Kokonendji, 2015) Soit $\mathbb{T}_d$ le support de la densité à estimer avec $\mathbb{T}_d \subseteq \mathbb{R}^d$ et $\mathbf{H}$ une matrice des fenêtres. Pour un vecteur cible $\mathbf{x} \in \mathbb{T}_d$, on considère un vecteur aléatoire $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$ de f.d.p. paramétrée $K_{\mathbf{x},\mathbf{H}}(\cdot)$ et de support $\mathbb{S}_{\mathbf{x},\mathbf{H}}\left(\subseteq \mathbb{R}^d\right)$. La fonction $K_{\mathbf{x},\mathbf{H}}(\cdot)$ est appelée "noyau associé multivarié (ou général)" si les conditions suivantes sont satisfaites :*

$$\mathbf{x} \in \mathbb{S}_{\mathbf{x},\mathbf{H}}, \ \ \mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x} + \mathbf{a}(\mathbf{x},\mathbf{H}) \ \ \textit{et} \ \ \mathrm{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{B}(\mathbf{x},\mathbf{H}),$$

où $\mathbf{a}(\mathbf{x},\mathbf{H}) = (a_1(\mathbf{x},\mathbf{H}),\ldots,a_d(\mathbf{x},\mathbf{H}))^\top$ et $\mathbf{B}(\mathbf{x},\mathbf{H}) = (b_{k\ell}(\mathbf{x},\mathbf{H}))_{k,\ell=1,\ldots,d}$ tendent respectivement vers le vecteur nul et la matrice nulle quand $\mathbf{H} \to \mathbf{0}$.

Ainsi, la *règle de discrimination par noyaux associés* (en anglais "Associated Kernel Discriminant Rule (AKDR)") s'obtient en remplaçant dans (2) f_j par $\widehat{f_j}$ de (3) et π_j par la taille d'échantillon empirique n_j/n avec $\sum_{j=1}^{J} n_j = n$:

$$\text{AKDR : Allouer } \mathbf{x} \text{ au } \textit{groupe } \widehat{j_0}, \text{ où } \widehat{j_0} = \arg\max_{j\in\{1,\ldots,J\}} \widehat{\pi}_j \widehat{f_j}(\mathbf{x};\boldsymbol{\kappa}). \tag{4}$$

De la Définition 2.1 on déduit trois cas particuliers de noyaux associés multivariés. Le premier est une interprétation des noyaux classiques (symétriques) à travers les noyaux symétriques continus. Pour une cible $\mathbf{x} \in \mathbb{R}^d =: \mathbb{T}_d$ et une matrice des fenêtres $\mathbf{H}$, le noyau classique K de moyenne nulle et de matrice de variance-covariance $\boldsymbol{\Sigma}$ est un noyau associé classique (multivarié) :

$$(i) \ \ K_{\mathbf{x},\mathbf{H}}(\cdot) = \frac{1}{\det \mathbf{H}} K\left\{\mathbf{H}^{-1}(\mathbf{x} - \cdot)\right\}$$

sur $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \mathbf{x} - \mathbf{H}\mathbb{S}_d$ avec $\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x}$ (i.e. $\mathbf{a}(\mathbf{x},\mathbf{H}) = \mathbf{0}$) et $\mathrm{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}$;

$$(ii) \ \ K_{\mathbf{x},\mathbf{H}}(\cdot) = \frac{1}{(\det \mathbf{H})^{1/2}} K\left\{\mathbf{H}^{-1/2}(\mathbf{x} - \cdot)\right\}$$

sur $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \mathbf{x} - \mathbf{H}^{1/2}\mathbb{S}_d$ avec $\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x}$ (i.e. $\mathbf{a}(\mathbf{x},\mathbf{H}) = \mathbf{0}$) et $\mathrm{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{H}^{1/2}\boldsymbol{\Sigma}\mathbf{H}^{1/2}$.

Un autre cas particulier de Définition 2.1, appropriée pour un mélange de variables continues et discrètes est présenté comme suit. Soit $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \times_{\ell=1}^{d} \mathbb{T}_1^{[\ell]} =: \mathbb{T}_d$ avec $k_\ell = 1$ de (1) et $\mathbf{H} = \mathbf{Diag}(h_{11}, \ldots, h_{dd})$ avec $h_{\ell\ell} > 0$. Soit $K_{x_\ell, h_{\ell\ell}}^{[\ell]}$ un noyau associé univarié (continu ou discret) (voir Définition 2.1 pour $d = 1$) de variable aléatoire correspondante $\mathcal{Z}_{x_\ell, h_{\ell\ell}}^{[\ell]}$ sur $\mathbb{S}_{x_\ell, h_{\ell\ell}} (\subseteq \mathbb{R})$ pour tout $\ell = 1, \ldots, d$. Alors, le noyau associé multiple est aussi un noyau associé multivarié :

$$K_{\mathbf{x},\mathbf{H}}(\cdot) = \prod_{\ell=1}^{d} K_{x_\ell, h_{\ell\ell}}^{[\ell]}(\cdot) \tag{5}$$

sur $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \times_{\ell=1}^{d} \mathbb{S}_{x_\ell, h_{\ell\ell}}$ avec $\mathbb{E}(\mathcal{Z}_{\mathbf{x},\mathbf{H}}) = (x_1 + a_1(x_1, h_{11}), \ldots, x_d + a_d(x_d, h_{dd}))^\top$ et $\mathrm{Cov}(\mathcal{Z}_{\mathbf{x},\mathbf{H}}) = \mathbf{Diag}\left(b_{\ell\ell}(x_\ell, h_{\ell\ell})\right)_{\ell=1,\ldots,d}$. En d'autres termes, les variables aléatoires $\mathcal{Z}_{x_\ell, h_{\ell\ell}}^{[\ell]}$ sont les composantes indépendantes du vecteur aléatoire $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$.

Le troisième cas particulier de noyaux associés multivariés est construit à partir d'une f.d.p. constituée de produit de f.d.p. univariés et d'une structure de corrélation utilisant la technique de Sarmanov (1966). De tels noyaux associés permettent d'atteindre certains endroits du lissage multidimensionnel. Kokonendji et Somé (2015) ont illustré l'effet de cette technique pour le noyau bêta bivarié.

L'algorithme d'analyse discriminante (4) est alors présenté en utilisant des noyaux associés multiples (5) et qui respectent le support de chaque variable d'intérêt. Le choix des J matrices de lissage optimale donnant le meilleur *taux d'erreur de classification* (en anglais "Misclassification Rate (MR)") se fait par minimisation de la fonction de validation croisée :

$$\mathrm{LSCV}(\mathbf{H}_j) = \int_{\mathbb{T}_d} \left\{\widehat{f_j}(\mathbf{x}; \boldsymbol{\kappa})\right\}^2 \boldsymbol{\nu}(d\mathbf{x}) - \frac{2}{n_j} \sum_{i=1}^{n_j} \widehat{f_{j,-i}}(\mathbf{X}_{ji}; \boldsymbol{\kappa}), \quad j = 1, \ldots, J, \tag{6}$$

où $\widehat{f_{j,-i}}\left(\mathbf{X}_{ji}; \boldsymbol{\kappa}\right) = (n_j - 1)^{-1} \sum_{k \neq i}^{n_j} K_{\mathbf{X}_{ji}, \mathbf{H}_j}(\mathbf{X}_{jk})$ est calculé à partir de $\widehat{f_j}(\mathbf{X}_{ji}; \boldsymbol{\kappa})$ en excluant l'observation $\mathbf{X}_{ji}$. On rappelle que le MR est la proportion des points $\mathbf{x} \in \mathbb{T}_d$ mal classés selon (4):

$$1 - \mathrm{MR} = \mathbb{E}_{\mathbf{x}}(\mathbb{1}\{\mathbf{x} \text{ est bien classé}\}), \quad \mathbf{x} \in \mathbb{T}_d,$$

où $\mathbb{E}_{\mathbf{x}}$ est l'espérance conditionnelle suivant $\mathbf{x}$. En pratique, pour des données homogènes (continues ou discrètes) le premier terme de (6) se calcule par intégrales ou sommes successives par rapport à la mesure $\boldsymbol{\nu}$ appropriée (Lebesgue ou comptage). Par contre, dans le cas mixte où le théorème de Fubini n'est pas applicable dans (6), on utilise la validation croisée profilée puisqu'il n'ya pas convergence de l'algorithme de (6). Par exemple, pour $\mathbb{T}_2 = [0, 1] \times \mathbb{N}$ et un noyau associé multiple (5) bêta×binomial, on fixe le paramètre de lissage h_{11} du noyau bêta suivi d'une minimisation de la fonction (6) sur $h_{22} \in]0, 1]$ du noyau binomial :

$$\widetilde{h}_{22[h_{11(j)}]} = \arg \min_{h_{22(j)} \in (0,1]} \mathrm{LSCV}_{h_{11(j)}}(h_{22(j)}),$$

avec $\mathrm{LSCV}_{h_{11(j)}}(h_{22(j)}) = \mathrm{LSCV}(h_{11(j)}, h_{22(j)}) := \mathrm{LSCV}(\mathbf{H}_j)$. En réalité $h_{11(j)}$ est inconnu, et donc pour chaque $h_{11(j)}$ on évalue $\widetilde{h}_{22[h_{11(j)}]}$ et on estime le MR correspondant $\widehat{\mathrm{MR}}(h_{11(j)}, \widetilde{h}_{22[h_{11(j)}]})$. Finalement, la meilleure estimation $\widehat{\mathrm{MR}}$ obtenue par validation croisée profilée et noté $\widehat{\mathrm{MR}}_p$ est

$$\widehat{\mathrm{MR}}_p = \min_{h_{11(j)}, \widetilde{h}_{22[h_{11(j)}]}} \widehat{\mathrm{MR}}(h_{11(j)}, \widetilde{h}_{22[h_{11(j)}]}).$$

Suivant Duong (2007), le $\widehat{\mathrm{MR}}$ de la règle de discrimination par noyaux associés (4) dépend de la disponibilité ou non de données test. Pour des données test $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ appartenant à $\mathbb{R}^d$, on a alors

$$\widehat{\mathrm{MR}} = 1 - m^{-1} \sum_{k=1}^{m} \mathbb{1}\{\mathbf{Y}_k \text{ est bien classé suivant AKDR}\},$$

où $\mathbb{1}\{A\}$ désigne la fonction indicatrice d'un événement quelconque A donné. Sinon, il est plus approprié d'utiliser une estimation par validation croisée de MR :

$$\widehat{\mathrm{MR}}_{\mathrm{cv}} = 1 - n^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{1}\{\mathbf{X}_{ji} \text{ est bien classé suivant AKDR}_{-ji}\},$$

où AKDR_{-ji} est similaire à AKDR sauf que $\widehat{\pi}_j$ et $\widehat{f_j}(\mathbf{x}; \boldsymbol{\kappa})$ sont remplacés par $\widehat{\pi}_{j,-i} = (n_j - 1)/n$ et $\widehat{f_{j,-i}}(\mathbf{x}; \boldsymbol{\kappa}) = (n_j - 1)^{-1} \sum_{r \neq i}^{n_j} K_{\mathbf{x}, \mathbf{H}_{j,-i}}\left(\mathbf{X}_{jr}\right)$ obtenues sans l'observation $\mathbf{X}_{ji}$. Des simulations pour les cas continu, discret et mixte avec respectivement les noyaux associés multiples (5) bêta×bêta, binomial×binomial et bêta×binomial montrent le caractère approprié et efficace de cette méthode. Une application sur un jeu de données provenant d'une région à haut risque de crise cardiaque de l'Afrique du Sud est présentée ; voir Rousseauw et al. (1983). Ce jeu de données comprend cinq variables continues, trois de comptages, une catégorielle et la variable de classification. On utilise alors (4) avec un noyau associé multiple approprié et composé de cinq noyaux gamma, trois binomial et d'un catégoriel de Aitchison et Aitken (1976).

Bibliographie

[1] Aitchison, J. et Aitken, C.G.G. (1976), Multivariate binary discrimination by the kernel method, *Biometrika* **63**, 413–420.

[2] Chen, S.X. (1999), A beta kernel estimation for density functions, *Computational Statistics and Data Analysis* **31**, 131–145.

[3] Chen, S.X. (2000), Probability density function estimation using gamma kernels, *Annals of the Institute of Statistical Mathematics* **52**, 471–480.

[4] Duong, T. (2007), `ks`: Kernel density estimation and kernel discriminant analysis for multivariate data in `R`, *Journal of Statistical Software* **21**, 1–16.

[5] Gosh, A.K. et Chaudhury, P. (2004), Optimal smoothing in kernel analysis discriminant, *Statistica Sinica* **14**, 457–483.

[6] Hayfield, T. et Racine, J.S. (2007), Nonparametric econometrics: the np package, *Journal of Statistical Software* **27**, 1–32.
[7] He, J., Yang, G., Rao, H., Li, Z., Ding, X. et Chen, Y. (2012), Prediction of human major histocompatibility complex class II binding peptides by continuous kernel discrimination method, *Artificial Intelligence in Medecine* **55**, 107–115.
[8] Igarashi, G. et Kakizawa, Y. (2015), Bias correction for some asymmetric kernel estimators, *Journal of Statistical Planning and Inference* **159**, 37–63.
[9] Kokonendji, C.C. et Senga Kiéssé, T. (2011), Discrete associated kernels method and extensions, *Statistical Methodology* **8**, 497–516.
[10] Kokonendji, C.C., Senga Kiéssé, T. et Zocchi, S.S. (2007), Discrete triangular distributions and nonparametric estimation for probability mass function, *Journal of Nonparametric Statistics* **19**, 241–254.
[11] Kokonendji, C.C. et Somé, S.M. (2015), On multivariate associated kernels for smoothing some density function, `arXiv:1502.01173`.
[12] Libengué, F.G. (2013), *Méthode Non-Paramétrique par Noyaux Associés Mixtes et Applications*. Ph.D. Thesis Manuscript (in French) to Université de Franche-Comté, Besançon, France & Université de Ouagadougou, Burkina Faso, Juin 2013, **LMB no. 14334**, Besançon.
[13] Liu, B., Yang, Y., Webb, I. et Boughton, J. (2009), A comparative study of bandwidth choice in kernel density estimation for naive Bayesian classification, *Lecture Notes in Computer Science* **5476**, 302–313.
[14] Malec, P. et Schienle, M. (2014), Nonparametric kernel density estimation near the boundary, *Computational Statistics and Data Analysis* **72**, 57–76.
[15] R Development Core Team (2015), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL `http://cran.r-project.org/`.
[16] Racine, J. et Li, Q. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.
[17] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. et Ferreira, J. (1983), Coronary risk factor screening in three rural communities, *South African Medical Journal* **64**, 430–436.
[18] Sarmanov, O.V. (1966), Generalized normal correlation and two-dimensionnal Frechet classes, *Doklady (Soviet Mathematics)* **168**, 596–599.
[19] Scott, W.D. (1992), *Multivariate Density Estimation*, John Wiley and Sons, New York.
[20] Somé, S.M. et Kokonendji, C.C. (2015), Effects of associated kernels in nonparametric multiple regressions, `arXiv:1502.01488`.
[21] Zougab, N., Adjabi, S. et Kokonendji, C.C. (2012), Binomial kernel and Bayes local bandwidth in discrete functions estimation, *Journal of Nonparametrics Statistics* **24**, 783–795.
[22] Zougab, N., Adjabi, S. et Kokonendji, C.C. (2014), Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation, *Computational Statistics and Data Analysis* **75**, 28–38.