Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n2p385

Associated kernel discriminant analysis for multivariate mixed data
By Somé, Kokonendji, Ibrahim

Published: 14 October 2016

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 02, October 2016, 385-399
DOI: 10.1285/i20705948v9n2p385

Associated kernel discriminant analysis for multivariate mixed data

Sobom M. Somé*[a], Célestin C. Kokonendji[a], and Mona Ibrahim[b]

[a] *Bourgogne Franche-Comté University, LMB Besançon, 16 route de Gray, 25030 Besançon cedex, France*
[b] *FCLAB/Femto-ST Belfort, rue Thierry Mieg, F-90010 Belfort, France*

Published: 14 October 2016

Associated kernels have been introduced to improve the classical (symmetric) continuous kernels for smoothing any functional on several kinds of supports such as bounded continuous and discrete sets. In this paper, an associated kernel for discriminant analysis with multivariate mixed variables is proposed. These variables are of three types: continuous, categorical and count. The method consists of using a product of adapted univariate associated kernels and an estimate of the misclassification rate. A new profile version cross-validation procedure of bandwidth matrices selection is introduced for multivariate mixed data, while a classical cross-validation is used for homogeneous data sets having the same reference measures. Simulations and validation results show the relevance of the proposed method. The method has been validated on real coronary heart disease data in comparison to the classical kernel discriminant analysis.

keywords: Bandwidth matrix, non-classical kernel, profile cross-validation.

1 Introduction

Discriminant analysis is a classical method in many scientific domains, such as bankruptcy prediction, face recognition, marketing and medicine. The dataset can be a mix of discrete (categorical, count) and continuous (rates, positive) data, such as the biomedical

*Corresponding author: sobom.some@univ-fcomte.fr

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

study in Rousseauw et al. (1983). Henceforth, the set $\mathbb{T}_d$ ($\subseteq \mathbb{R}^d$) denotes the support of the d-variate mixed data

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

and $\boldsymbol{\nu} = \nu_1 \times \ldots \times \nu_L$ represents the reference measure on $\mathbb{T}_d$, where ν_ℓ $(1 \leq \ell \leq L)$ is the measure related to the corresponding support $\mathbb{T}_{k_\ell}^{[\ell]}$ with fixed k_ℓ-dimension (that has a given correlation structure).

In a discriminant analysis problem, a decision rule is used to classify a d-dimensionnal observation $\mathbf{x}$ belonging to $\mathbb{T}_d$ in one of the J classes. Specifically, the optimal Bayes rule assigns an observation to the class with the largest posterior probability. It can be described as follows

$$\text{Allocate } \mathbf{x} \text{ to group } j_0 \text{ where } j_0 = \arg \max_{j \in \{1,\ldots,J\}} \pi_j f_j(\mathbf{x}), \tag{2}$$

where π_j is the prior probability and $f_j(\mathbf{x})$ is the probability density function (pdf) of the jth class. These pdfs are usually unknown in practice, and can be estimated from the *training data set* using either parametric or nonparametric approaches. In parametric approaches, the underlying populations distributions are assumed to be known except for some unknown parameters such as mean vector and dispersion matrix; see, e.g., Nath et al. (1992) and Simonoff (1996). The linear and quadratic discriminant techniques are the most used classic parametric approaches. However, they suffer from the restrictive assumption of normality. In nonparametric approaches, this assumption is relaxed and therefore more complex cases can be investigated. Kernel density estimation is a well-known method for constructing nonparametric estimations of population densities, namely in discriminant analysis; see for instance Duong (2007) and Ghosh and Chaudhury (2004). Other methods have also been performed for nonparametric density estimation such as splines in Wahba (1990), Gu (1993) and Koo et al. (2009) and wavelets in Antoniadis (1997) and Shi et al. (2006).

More precisely, let $\mathcal{X}_j = \{\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}\}$ be a d-dimensional observations drawn from an unknown density function f_j, for $j = 1, \ldots, J$, where the sample sizes n_j are known and non-random. Moreover, the observations $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}$ are independent and identically distributed (iid) and are they belong to the jth population on the support $\mathbb{T}_d (\subseteq \mathbb{R}^d)$. The classical kernel estimator $\widehat{f_j}$ of f_j in (2) which uses continuous symmetric kernels is of the form:

$$\widehat{f_j}(\mathbf{x}; K, \mathbf{H}_j) = \frac{1}{n_j \det \mathbf{H}_j} \sum_{i=1}^{n_j} K \left\{ \mathbf{H}_j^{-1} (\mathbf{x} - \mathbf{X}_{ji}) \right\} = \widehat{f_j}(\mathbf{x}; \mathbf{H}_j), \qquad \forall \mathbf{x} \in \mathbb{T}_d := \mathbf{R}^d, \tag{3}$$

where $\mathbf{H}_j$ is the jth symmetric and positive definite bandwidth matrix of dimension $d \times d$ and the function $K(\cdot)$ is the multivariate kernel assumed to be a spherically symmetric pdf. Since the choice of the kernel K does not change the result in classical case, a common notation will be used $\widehat{f_j}(\mathbf{x}; \mathbf{H}_j)$ for the jth density estimation with classical kernel. Thus, the Kernel Discriminant Rule (KDR) is obtained from the Bayes rule, in equation (2), by replacing f_j by $\widehat{f}_j$ and π_j is usually replaced by the sample proportion $\widehat{\pi}_j = n_j/n$ with $\sum_{j=1}^{J} n_j = n$; that is

$$\text{KDR: Allocate } \mathbf{x} \text{ to group } \widehat{j}_0 \text{ where } \widehat{j}_0 = \arg \max_{j \in \{1,\ldots,J\}} \widehat{\pi}_j \widehat{f}_j(\mathbf{x}; \mathbf{H}_j). \tag{4}$$

This now raises the question of the choice of the most appropriate multivariate kernel according to the mix of discrete and continuous variables. In fact, because of symmetry, the multivariate classical kernel (e.g. Gaussian) suits only for estimating the densities f_j with unbounded supports (i.e. $\mathbb{R}^d$); see Scott (1992), Duong (2004) and Zougab et al. (2014). Moreover, the multivariate Gaussian kernel is well-known in discriminant analysis. In order to estimate different functionals, Racine and Li (2007) proposed multiple kernels composed of univariate Gaussian kernels for continuous variables and Aitchison and Aitken (1976) kernels for categorical variables. Some implementations of the multiple kernels using the R software (R Development Core Team, 2015) have been performed in Hayfield and Racine (2007). Besides, it should be noted that the application of Gaussian kernels (i.e. symmetric) gives weights outside variables with bounded or discrete supports. Also, the resulting bias increases with the dimension d and will affect the accuracy of the KDR in (4), especially in the purpose of classifying points at the boundary of the classes. In the univariate continuous case, Chen (1999, 2000) is one of the pioneers who has proposed asymmetric kernels, as beta and gamma, whose supports coincide with those of the unknown functions to be estimated. Zhang (2010) and Zhang and Karunamuni (2010) studied the performance of these beta and gamma kernel estimators at the boundaries in comparison with those of the classical kernels; see also Malec and Schienle (2014) and Igarashi and Kakizawa (2015). Libengué (2013) investigated several families of these univariate continuous kernels that he called univariate associated kernels; see also Kokonendji et al. (2007), Kokonendji and Senga Kiessé (2011), Zougab et al. (2012) for univariate discrete situations. As for the multivariate case, Bouerzmarni and Rombouts (2009) proposed a product of univariate gamma and beta kernels. Another multivariate version with correlation structure of these associated kernels has been studied in Kokonendji and Somé (2015) for continuous density functions and in Somé and Kokonendji (2016) for multiple regression; see Sarmanov (1966) for construction of multivariate densities with some correlation structure from independent components.

In this work, a new application of multivariate associated kernels for the discriminant analysis is proposed. The associated kernels are appropriated for both mixed training data $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}$ for $j = 1, \ldots, J$ and test data $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ drawn from $f = \sum_{j=1}^{J} \pi_j f_j$. In order to estimate the densities f_j in (2), we propose multiple (or product of) associated kernels composed by univariate discrete (e.g. binomial) and continuous (e.g. beta, gamma) associated kernels. The bandwidth matrices selection remains crucial to minimize the misclassification rate. A new profile cross-validation will be introduced for mixed variables, whereas a classic cross-validation will be used for homogeneous data sets. It should be noted that these associated kernels are adapted for this situation of mixing axis, since they fully respect the support of each explanatory variable. Some appropriated type of associated kernels, denoted by $\boldsymbol{\kappa}$, will be used for discriminant analysis by mean of simulations and validations.

The rest of the paper is organized as follows. Section 2 represents a general definition of multivariate associated kernels including both of the continuous classical symmetric and the multiple composed by univariate discrete and continuous. Then, the corresponding KDR appropriated for both continuous and discrete explanatory variables is given. Also, we present the profile cross-validation suitable for bandwidth matrices selection in the case of mixed variables. In Section 3, we investigate only the appropriated associate kernels according to the support of the variables through simulations studies and real data analysis. Finally, concluding remarks are drawn in Section 4.

2 Associated kernels for discriminant analysis

With the assumptions of equation (1), the associated kernel $K_{\mathbf{x},\mathbf{H}}(\cdot)$ which replaces the classical kernel $K(\cdot)$ of (3) is a pdf according to some measure $\boldsymbol{\nu}$. This kernel $K_{\mathbf{x},\mathbf{H}}(\cdot)$ can be defined as follows:

Definition 2.1 *Let $\mathbb{T}_d \left(\subseteq \mathbb{R}^d\right)$ be the support of the densities f_j, to be estimated, $\mathbf{x} \in \mathbb{T}_d$ a target vector and $\mathbf{H}$ a bandwidth matrix. A parametrized pdf $K_{\mathbf{x},\mathbf{H}}(\cdot)$ of support $\mathbb{S}_{\mathbf{x},\mathbf{H}} \left(\subseteq \mathbb{R}^d\right)$ is called "multivariate (or general) associated kernel" if the following conditions are satisfied:*

$$\mathbf{x} \in \mathbb{S}_{\mathbf{x},\mathbf{H}}, \tag{5}$$

$$\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x} + \mathbf{a}(\mathbf{x},\mathbf{H}), \tag{6}$$

$$\operatorname{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{B}(\mathbf{x},\mathbf{H}), \tag{7}$$

where $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$ denotes the random vector with pdf $K_{\mathbf{x},\mathbf{H}}$ and both

$$\mathbf{a}(\mathbf{x},\mathbf{H}) = \left(a_1(\mathbf{x},\mathbf{H}),\ldots,a_d(\mathbf{x},\mathbf{H})\right)^{\top}$$

and

$$\mathbf{B}(\mathbf{x},\mathbf{H}) = \left(b_{ij}(\mathbf{x},\mathbf{H})\right)_{i,j=1,\ldots,d}$$

tend, respectively, to the null vector $\mathbf{0}$ and the null matrix $\mathbf{0}_d$ as $\mathbf{H}$ goes to $\mathbf{0}_d$.

From this definition and in comparison with (3), the jth associated kernel estimator $\widetilde{f}_j$ of f_j is

$$\widetilde{f}_j(\mathbf{x}) = \frac{1}{n_j}\sum_{i=1}^{n_j} K_{\mathbf{x},\mathbf{H}_j}\left(\mathbf{X}_{ji}\right) = \widetilde{f}_j(\mathbf{x};\boldsymbol{\kappa},\mathbf{H}_j), \tag{8}$$

where $\mathbf{H}_j \equiv \mathbf{H}_{jn_j}$ is the bandwidth matrix such that $\mathbf{H}_{jn_j} \to \mathbf{0}$ as $n_j \to \infty$, and $\boldsymbol{\kappa}$ represents the type of the associated kernel $K_{\mathbf{x},\mathbf{H}_j}$, parametrized by $\mathbf{x}$ and $\mathbf{H}_j$. Without loss of generality, and hereafter $\widetilde{f}_j(\mathbf{x};\boldsymbol{\kappa},\mathbf{H}_j) \equiv \widetilde{f}_j(\mathbf{x};\boldsymbol{\kappa})$ will be used in order to point out the effect of $\boldsymbol{\kappa}$, since the bandwidth matrix is here investigated only by cross-validation. Therefore, the classical KDR of (4) becomes the associated KDR (AKDR) using (8)

$$\text{AKDR: Allocate } \mathbf{x} \text{ to group } \widetilde{j}_0 \text{ where } \widetilde{j}_0 = \arg \max_{j\in\{1,\ldots,J\}} \widehat{\pi}_j \widetilde{f}_j(\mathbf{x};\boldsymbol{\kappa}). \tag{9}$$

The following two examples provide the well-known multivariate associated kernel estimators and they represent particular cases that can be used in (1). The first can be seen as an interpretation of classical associated kernels through continuous symmetric kernels. The second deals on non-classical associated kernels without correlation structure.

Given a target vector $\mathbf{x} \in \mathbb{R}^d =: \mathbb{T}_d$ and a bandwidth matrix $\mathbf{H}$, it follows that the classical kernel in (3) with null mean vector and covariance matrix $\boldsymbol{\Sigma}$ induces the so-called (multivariate) classical associated kernel:

$$(i) \;\; K_{\mathbf{x},\mathbf{H}}(\cdot) = \frac{1}{\det \mathbf{H}} K\left\{\mathbf{H}^{-1}(\mathbf{x}-\cdot)\right\} \tag{10}$$

on $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \mathbf{x} - \mathbf{H}\mathbb{S}_d$ with $\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x}$ (i.e. $\mathbf{a}(\mathbf{x},\mathbf{H}) = \mathbf{0}$) and $\operatorname{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}$;

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

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

A second particular case of Definition 2.1, appropriated for a mix of both continuous and count explanatory variables without correlation structure, is presented as follows.

Let $\mathbf{x} = (x_1,\ldots,x_d)^\top \in \times_{\ell=1}^{d}\mathbb{T}_1^{[\ell]} =: \mathbb{T}_d$ with $k_\ell = 1$ of (1) and $\mathbf{H} = \mathbf{Diag}(h_{11},\ldots,h_{dd})$ with $h_{\ell\ell} > 0$. Let $K^{[\ell]}_{x_\ell,h_{\ell\ell}}$ be a (discrete or continuous) univariate associated kernel (see Definition 2.1 for $d = 1$) with its corresponding random variable $\mathcal{Z}^{[\ell]}_{x_\ell,h_{\ell\ell}}$ on $\mathbb{S}_{x_\ell,h_{\ell\ell}}(\subseteq \mathbb{R})$ for all $\ell = 1,\ldots,d$. Then, the multiple associated kernel is also a multivariate associated kernel:

$$K_{\mathbf{x},\mathbf{H}}(\cdot) = \prod_{\ell=1}^{d} K^{[\ell]}_{x_\ell,h_{\ell\ell}}(\cdot) \tag{11}$$

on $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \times_{\ell=1}^{d}\mathbb{S}_{x_\ell,h_{\ell\ell}}$ with $\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \left(x_1 + a_1(x_1,h_{11}),\ldots,x_d + a_d(x_d,h_{dd})\right)^\top$ and $\operatorname{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{Diag}\left(b_{\ell\ell}(x_\ell,h_{\ell\ell})\right)_{\ell=1,\ldots,d}$. In other words, the random variables $\mathcal{Z}^{[\ell]}_{x_\ell,h_{\ell\ell}}$ are independent components of the random vector $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$.

In the following two subsections, some examples of associated kernels are illustrated and then criteria of discriminant analysis are presented.

2.1 Some associated kernels

In order to point out the importance of associated kernel in discriminant analysis, below some kernels that will be used in numerical studies are motivated. These concern four basic univariate associated kernels for which two of them are discrete and the two last are continuous.

- The binomial kernel is defined on the support $\mathbb{S}_x = \{0, 1, \ldots, x+1\}$ with $x \in \mathbb{T}_1 := \mathbb{N} = \{0, 1, \ldots\}$ and then $h \in (0, 1]$:

$$B_{x,h}(u) = \frac{(x+1)!}{u!(x+1-u)!} \left(\frac{x+h}{x+1}\right)^u \left(\frac{1-h}{x+1}\right)^{x+1-u} \mathbb{1}\{u \in \mathbb{S}_x\},$$

  where $\mathbb{1}\{A\}$ denotes the indicator function of any given event A. Note that $B_{x,h}$ is the probability mass function (pmf) of the binomial distribution $\mathcal{B}(x+1; (x+h)/(x+1))$ with its number of trials $x+1$ and its success probability in each trial $(x+h)/(x+1)$. It is appropriated for count data with small or moderate sample sizes and, also, it does not satisfy equation (7); see Kokonendji and Senga Kiessé (2011) and also Zougab et al. (2012) for a bandwidth selection by Bayesian method.

- From Aitchison and Aitken (1976), Kokonendji and Senga Kiessé (2011) deduced the following discrete kernel which is here labelled DiracDU as "Dirac Discrete Uniform". For fixed $c \in \{2, 3, \ldots\}$ the number of categories, one has $\mathbb{S}_c = \{0, 1, \ldots, c-1\}$ and

$$DU_{x,h;c}(u) = (1-h)^{1-\mathbb{1}\{u \in \mathbb{S}_c \setminus \{x\}\}} \left(\frac{h}{c-1}\right)^{\mathbb{1}\{u \in \mathbb{S}_c \setminus \{x\}\}},$$

  where $h \in (0, 1]$ and $x \in \mathbb{T}_1$. This DiracDU kernel is symmetric around the target, satisfying Definition 2.1 and appropriated for categorical set $\mathbb{T}_1$.

- The gamma kernel is defined on $\mathbb{S}_{x,h} = [0, \infty) = \mathbb{T}_1$ with $x \in \mathbb{T}_1$ and $h > 0$:

$$GA_{x,h}(u) = \frac{u^{x/h}}{\Gamma(1 + x/h)\, h^{1+x/h}} \exp\left(-\frac{u}{h}\right) \mathbb{1}\{u \in [0, \infty)\},$$

  where $\Gamma(\cdot)$ is the classical gamma function. It is the pdf of the gamma distribution $\mathcal{G}a(1 + x/h, h)$ with scale parameter $1 + x/h$ and shape parameter h. It satisfies Definition 2.1 and it is appropriated for non-negative real set $\mathbb{T}_1$; see Chen (2000).

- Finally, the beta kernel is defined on $\mathbb{S}_{x,h} = [0, 1] = \mathbb{T}_1$ with $x \in \mathbb{T}_1$ and $h > 0$:

$$BE_{x,h}(u) = \frac{u^{x/h}(1-u)^{(1-x)/h}}{\mathscr{B}(1 + x/h, 1 + (1-x)/h)} \mathbb{1}\{u \in [0, 1]\},$$

  where $\mathscr{B}(r, s) = \int_0^1 t^{r-1}(1-t)^{s-1} dt$ is the usual beta function with $r > 0$ and $s > 0$. It is the pdf of the beta distribution $\mathcal{B}e(1 + x/h, (1-x)/h)$ with shape parameters $1 + x/h$ and $(1-x)/h$. This pdf satisfies Definition 2.1 and is appropriated for rates, proportions and percentages dataset $\mathbb{T}_1$; see Chen (1999).

Figure 1 shows some forms of the above-mentioned univariate associated kernels. These kernels are completed by the Gaussian in its associated form constructed with

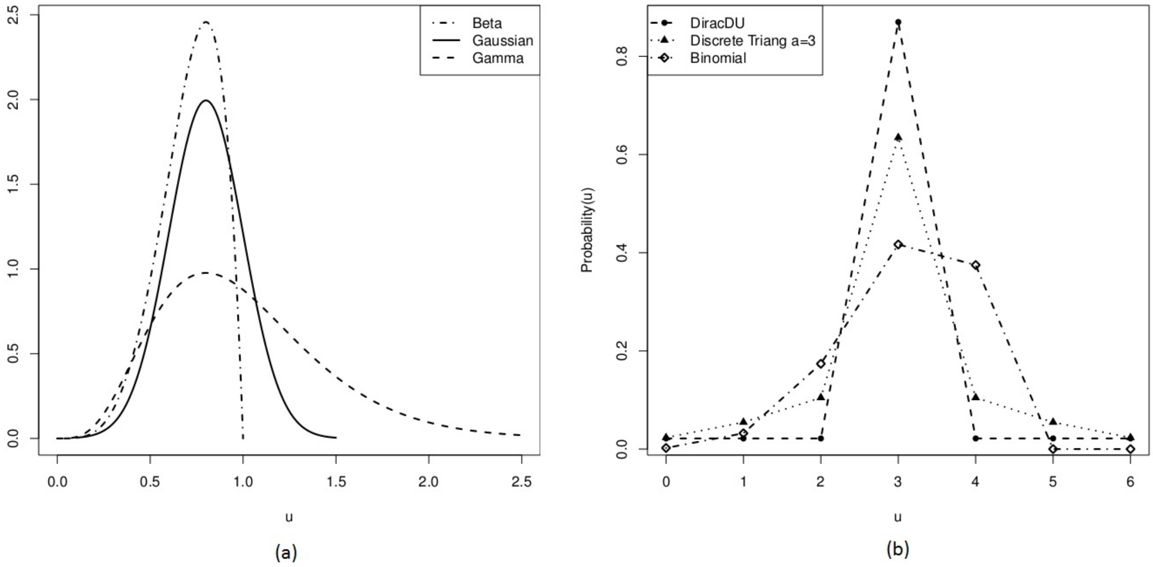

Figure 1: Shapes of univariate (discrete and continuous) associated kernels: (a) DiracDU, discrete triangular $a = 3$ and binomial with same target $x = 3$ and bandwidth $h = 0.13$; (b) Beta, Gaussian and gamma with same $x = 0.8$ and $h = 0.2$.

equation (10), and the discrete triangular of Kokonendji et al. (2007); see also Kokonendji and Zocchi (2010) for an asymmetric version. The plots highlight the importance of the target point and around it in discrete (a) and continuous (b) cases. Furthermore, for a fixed bandwidth h, the Gaussian keeps its same shape along the support; however, they change according to the target for other non-classical associated kernels. This explains the inappropriateness of the Gaussian kernel for density estimation in any bounded interval; see Part (b) of Figure 1. This argument remains valid for discrete triangular kernel for any bounded count support; see Part (a) of Figure 1.

2.2 Misclassification rate and bandwidth matrix selection

The performance of the AKDR method is investigated by the misclassification rate denoted by MR. This error rate is the proportion of the points that are assigned to an incorrect group based on the discriminant rule in (9). Then, we have

$$\begin{aligned} 1 - \mathrm{MR} &= \mathbb{P}(\mathbf{Y} \text{ is correctly classified}) \\ &= \mathbb{E}_{\mathbf{Y}}(\mathbb{1}\{\mathbf{Y} \text{ is correctly classified}\}), \end{aligned}$$

where $\mathbb{E}_{\mathbf{Y}}$ is the expectation with respect to $\mathbf{Y}$ or $\sum_{j=1}^{J} \pi_j f_j$. See also Hall and Wand (1988) who proposed to find optimal bandwidth that directly optimises this MR for a two-class problem. In practice, an estimate of MR is used, denoted $\widehat{\mathrm{MR}}$, for test data

$\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ and also for training data:

$$\widehat{\mathrm{MR}} = 1 - m^{-1} \sum_{k=1}^{m} \mathbb{1}\{\mathbf{Y}_k \text{ is correctly classified using AKDR}\}. \tag{12}$$

The optimal bandwidth matrices that give the best $\widehat{\mathrm{MR}}$ are chosen by the least squared cross-validation (LSCV) method. The multivariate cross-validation is a straightforward generalization of the one dimensional case. For the jth group, the LSCV estimator is

$$\mathrm{LSCV}(\mathbf{H}_j) = \int_{\mathbb{T}_d} \left\{ \widetilde{f}_j(\mathbf{x}) \right\}^2 \boldsymbol{\nu}(d\mathbf{x}) - \frac{2}{n_j} \sum_{i=1}^{n_j} \widetilde{f}_{j,-i}(\mathbf{X}_{ji}), \tag{13}$$

where $\widetilde{f}_{j,-i}\left(\mathbf{X}_{ji}\right) = (n_j - 1)^{-1} \sum_{k \neq i}^{n_j} K_{\mathbf{X}_{ji},\mathbf{H}_j}(\mathbf{X}_{jk})$ is being computed as $\widetilde{f}_j(\mathbf{X}_{ji})$ excluding the observation $\mathbf{X}_{ji}$. Practically, for homogeneous (continuous and discrete) data the first term of (13) is calculated by successive sums or integrals according to the appropriated measure $\boldsymbol{\nu}$ (Lebesgue or count). In that case, the optimal bandwidth matrix $\widetilde{\mathbf{H}}_j$ obtained by LSCV rule in (13) with multiple associated kernels in equation (11) is defined as follows:

$$\widetilde{\mathbf{H}}_j = \arg \min_{\mathbf{H}_j \in \mathcal{D}} \mathrm{LSCV}(\mathbf{H}_j), \tag{14}$$

where $\mathcal{D}$ is the set of all positive definite diagonal bandwidth matrices. Its algorithm is described below and used for numerical studies in the following section.

A1. *Algorithms of LSCV method in (14) for some type of associated kernels and their corresponding bandwidth matrices*

Multiple associated kernels (i.e. diagonal bandwidth matrices) for $d \geq 2$.

1. Choose d intervals H_{11}, …, H_{dd} related to h_{11}, …, h_{dd}, respectively.
2. For $\delta_1 = 1, \ldots, \ell(H_{11})$, …, $\delta_d = 1, \ldots, \ell(H_{dd})$,
   Compose the diagonal bandwidth matrix $\mathbf{H}_j(\delta_1, \ldots, \delta_d) := \mathbf{Diag}\left(H_{11}(\delta_1), \ldots, H_{dd}(\delta_d)\right)$.
3. Apply LSCV method on the set $\mathcal{D}$ of all diagonal bandwidth matrices $\mathbf{H}_j(\delta_1, \ldots, \delta_d)$.

However, for mixed case where the Fubini theorem is not applicable in (13), one uses the so called profile cross-validation since there is no convergence of the classical cross-validation algorithm below. The method is presented in bivariate case for any group ($j = 1, \ldots, J$). For instance, for $\mathbb{T}_2 = [0, 1] \times \mathbb{N}$ and a multiple associated kernel (11) beta×binomial, the smoothing parameter $h_{11(j)}$ of the beta kernel is fixed, followed by a minimisation of the cross-validation function on $h_{22(j)} \in (0, 1]$ of the binomial kernel:

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

with $\mathrm{LSCV}_{h_{11(j)}}(h_{22(j)}) = \mathrm{LSCV}(h_{11(j)}, h_{22(j)}) := \mathrm{LSCV}(\mathbf{H}_j)$. Furthermore, $h_{11(j)}$ is unknown, hence for each $h_{11(j)}$ we can evaluate $\widetilde{h}_{22(j)[h_{11(j)}]}$ and estimate the corresponding misclassification rate $\widehat{\mathrm{MR}}(h_{11(j)}, \widetilde{h}_{22(j)[h_{11(j)}]})$. The jth optimal bandwidth matrix using the profile cross-validation and denoted $\widehat{\mathbf{H}}_{jp}$ is

$$\widehat{\mathbf{H}}_{jp} = \arg \min_{h_{11(j)}, \widetilde{h}_{22[h_{11(j)}]}} \widehat{\mathrm{MR}}(h_{11(j)}, \widetilde{h}_{22[h_{11(j)}]}). \tag{15}$$

This example of profile cross-validation in (15) and the classical one in (13) are implemented below. It should be noted that, other bandwidth matrix selection by Bayesian methods is possible; see, e.g., Ziane et al. (2015) for the adaptive case using asymmetric kernel.

3 Numerical studies

Before presenting some simulations results and real data analysis, let us start by the algorithm of the associated kernel discriminant analysis.

A2. *Algorithm for associated kernel discriminant analysis*

1. For each training sample $\mathcal{X}_j = \{\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}\}$, $j = 1, 2, \ldots, J$, compute a kernel density estimate with (8) using the optimal bandwidth matrix obtained by cross-validation in (14).
2. Use the prior probabilities once they are known. Otherwise, estimate them using the training sample proportions $\widehat{\pi}_j = n_j/n$.
3. (a) Allocate test data points $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ according to AKDR of (9).

   (b) Allocate all points $\mathbf{x}$ from the sample space according to AKDR of (9).
4. (a) If we have test data then the estimate of the misclassification rate is $\widehat{\mathrm{MR}}$ of (12).

   (b) If the test data are not available, the cross-validation estimate of misclassification rate is

   $$\widehat{\mathrm{MR}}_{\mathrm{cv}} = 1 - n^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{1}\{\mathbf{X}_{ji} \text{ is correctly classified using } \mathrm{AKDR}_{ji}\} \tag{16}$$

   where AKDR_{ji} is similar to AKDR except that $\widehat{\pi}_j$ and $\widetilde{f}_j(\mathbf{x}; \boldsymbol{\kappa})$ are replaced by their leave-one-out estimates by removing $\mathbf{X}_{ji}$ i.e. $\widehat{\pi}_{j,-i} = (n_j - 1)/n$ and

   $$\widetilde{f}_{j,-i}(\mathbf{x}; \boldsymbol{\kappa}) = \frac{1}{n_j - 1} \sum_{r \neq i}^{n_j} K_{\mathbf{x}, \mathbf{H}_{j,-i}}(\mathbf{X}_{jr}).$$

   That is, we repeat step 3 to classify all $\mathbf{X}_{ji}$ using AKDR_{ji}.

3.1 Simulation studies

In this section, the results of a simulation study that was conducted for evaluating the performance of the algorithm of AKDR is presented. Computations have been performed on the supercomputer facilities of the "*Mésocentre de calcul de Franche-Comté*" using the `R` software; see `R` Development Core Team (2015). This simulation study has two objectives. First, we investigate the ability of multiple associated kernels in (13) which can scrupulously respect any bounded, count and mixed support $\mathbb{T}_d$ of dataset, and more give good estimate $\widehat{\mathrm{MR}}$ of (12). We, therefore, use only appropriate multiple associated kernels for these simulations studies. Second, we evaluate the sensitivity of the proposed method in relation to the sample size n.

Three scenarios denoted A, B and C are considered in dimension $d = 2$. With Scenario A data are generated using a mixture of two bivariates Dirichlet density

$$\begin{aligned} f_A(x_1,x_2) &= \frac{3\Gamma(\alpha_1+\alpha_2+\alpha_3)}{7\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1-1}x_2^{\alpha_2-1}(1-x_1-x_2)^{\alpha_3-1}\mathbb{1}_{\{x_1,\ x_2\geq 0,\ x_1+x_2\leq 1\}}(x_1,x_2) \\ &\quad + \frac{4\Gamma(\beta_1+\beta_2+\beta_3)}{7\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_3)} x_1^{\beta_1-1}x_2^{\beta_2-1}(1-x_1-x_2)^{\beta_3-1}\mathbb{1}_{\{x_1,\ x_2\geq 0,\ x_1+x_2\leq 1\}}(x_1,x_2), \end{aligned}$$

where $\Gamma(\cdot)$ is the classical gamma function, with parameter values $\alpha_1 = \alpha_2 = 5$, $\alpha_3 = 6$, $\beta_1 = \beta_2 = 2$ and $\beta_3 = 10$. With Scenario B data are generated using a mixture of two bivariates Poisson with a correlation structure

$$\begin{aligned} f_B(x_1,x_2) &= \frac{2e^{-(\theta_1+\theta_2+\theta_{12})}}{5} \sum_{i=0}^{min(x_1,x_2)} \frac{\theta_1^{x_1+i}\theta_2^{x_2+i}\theta_{12}^{i}}{(x_1+i)!(x_2+i)!i!}\mathbb{1}_{\mathbb{N}\times\mathbb{N}}(x_1,x_2) \\ &\quad + \frac{3e^{-(\theta_a+\theta_b+\theta_{ab})}}{5} \sum_{i=0}^{min(x_1,x_2)} \frac{\theta_a^{x_1+i}\theta_b^{x_2+i}\theta_{ab}^{i}}{(x_1+i)!(x_2+i)!i!}\mathbb{1}_{\mathbb{N}\times\mathbb{N}}(x_1,x_2), \end{aligned}$$

with parameter values $\theta_1 = 2$, $\theta_2 = 3$, $\theta_{12} = 4$, $\theta_a = 3$, $\theta_b = 4$ and $\theta_{ab} = 5$. Finally, with Scenario C is a bivariate beta without correlation

$$f_C(x_1,x_2) = \left(\frac{3e^{-2}2^{x_1}}{7x_1!} + \frac{4e^{-3}3^{x_2}}{7x_2!}\right) \times \frac{x_2^{p_1-1}(1-x_2)^{q_1-1}}{\mathscr{B}(p_1,q_1)}\ \mathbb{1}_{\mathbb{N}}(x_1)\mathbb{1}_{[0,1]}(x_2),$$

with $(p_1, q_1) = (2, 7)$. The use of these bivariate distributions is motivated by the aim of investigating unbounded continuous, count and mixed situations and, if possible, correlation structure in the data. For each scenario we generate $N_{sim} = 100$ dataset of different sizes.

The choice of the multiple beta kernel is motivated by the cumbersome procedures of the bivariate beta kernel with correlation structure; see Kokonendji and Somé (2015) for further details. Also, the authors show that the so-called beta-Sarmanov kernel with correlation structure gives similar result to the multiple beta without correlation structure. Thus, we here focus on the bivariate case. The test data size is $m = 200$ for each replication. We consider sample size $n = 250$ and 500 in continuous case and the multiple beta kernel which is the more suitable for bivariate rates data. Table 2 reports the average $\widehat{\mathrm{MR}}$ which we denote $\overline{\widehat{\mathrm{MR}}}$. We can observe low values that prove the appropriateness of the discriminant analysis algorithm mentioned above in this section. Furthermore, the errors become smaller when the sample size increases.

Table 1: Some expected values of $\overline{\widehat{\mathrm{MR}}}$ and their standard errors in parentheses with $N_{sim} = 100$ of the multiple beta kernel with test data size $m = 200$.

	m	n	Beta×Beta
A	200	250	0.0421(0.0146)
		500	0.0329(0.0097)

Since the binomial kernel is the most interesting of discrete (count) associated kernels for small sample sizes, the samples sizes $n = 50, 100$ and 200 are considered; see Somé and Kokonendji (2016) for comparisons with the discrete (symmetric) triangular kernel. In fact, unlike the binomial kernel, the discrete triangular kernel induces a bias on any partially or totally bounded count support. Of course, this boundary bias problem might have an effect on discriminant analysis at the boundaries of two classes.

In each replication, the test data size is $m = 100$. Table 2 reports the values of $\overline{\widehat{\mathrm{MR}}}$ for this appropriate multiple binomial kernel. Once again, we observe low values of $\overline{\widehat{\mathrm{MR}}}$ which has a tendency to become better when the sample size increases. For scenario C of mixed case, we have to take into account the effects of the continuous and discrete sample size. Thus, we consider sample sizes $n = 80, 250$ and 500 and test data size $m = 150$. The values of $\overline{\widehat{\mathrm{MR}}}$ in Table 3 show the effectiveness of the profile cross-validation method in equation (15). Also, the errors rate is getting improved when the sample size increases.

Table 2: Some expected values of $\overline{\widehat{\mathrm{MR}}}$ and their standard errors in parentheses with $N_{sim} = 100$ of the multiple binomial kernel with test data size $m = 100$.

	m	n	Binomial×Binomial
		50	0.0747(0.0334)
B	100	100	0.0677(0.0255)
		200	0.0581(0.0221)

For now, it is not possible to make simulations studies with more than two variables mainly due to the time consuming of the cross-validation methods.

Table 3: Some expected values of $\widehat{\mathrm{MR}}$ and their standard errors in parentheses with $N_{sim} = 100$ of the beta×binomial kernel with test data size $m = 150$.

	m	n	Beta×Binomial
C	150	80	0.0182(0.0116)
		250	0.0171(0.0080)
		500	0.0152(0.0042)

3.2 Real data analysis

The algorithm A2 of AKDR in Section 3 is applied using multiple associated kernels in equation (11). The dataset represents a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. These data are taken from a larger dataset, described in Rousseauw et al. (1983) and available in Hastie et al. (2009); see also the `R` package `ElemStatLearn` of Halvorsen (2015). It has 462 observations on 10 variables composed of five continuous (positive) variables, three count variables, one categorical variable and the classification variable Coronary Heart Disease (CHD). This CHD variable has two groups: the "group 1" of patients with CHD and the "group 2" of those without CHD. The used multiple associated kernel is composed by gamma kernels for continuous (positive) variables, binomial kernels for count variables and DiracDU kernel for the categorical one. We must here use the cross-validation estimate MR_{cv} of (16) since we do not have test data. Also, the profile cross-validation is appropriate for this mixed dataset. This method is already computationally intense for two variables, and thus is not recommended for this dataset of ten mixed variables. Two diagonal bandwidth matrices $\mathbf{H}_{\text{group1}} = \mathbf{Diag}\,(0.01, 0.02, 0.03, 0.015, 0.004, 0.02, 0.03, 0.02, 0.02)$ and $\mathbf{H}_{\text{group2}} = \mathbf{Diag}\,(0.87, 0.3, 0.2, 0.15, 0.4, 0.2, 0.4, 0.2, 0.56, 0.6)$ are chosen for both groups of CHD and the cross-validation estimate of the misclassification rate MR_{cv} is 30.090%. Also, we use the LSCV selector with full bandwidth matrices and multivariate Gaussian kernel of Duong (2007), and the misclassification rate is equal to 30.952%. Thus, the associated kernel discriminant analysis with chosen bandwidth matrices is slightly better than the classical one of Duong (2007).

4 Concluding remarks

We have presented associated kernels for discriminant analysis and in presence of a mixture of discrete and continuous explanatory variables. Two particular cases including the continuous classical and the multiple (or product of) associated kernels are highlighted. Also, four univariate associated kernels are presented. These kernels, namely binomial, DiracDU, beta and gamma are used for simulations studies and real data analysis. The bandwidth selection is obtained by classical cross-validation for homogeneous (continuous or discrete) data while a profile version is provided for mixed data.

Simulation experiments and analysis of a real dataset provide insight into the appropriateness of associated kernel for small and moderate sample sizes and also for homogeneous or mixed data. Table 1 and equation (2) show the efficiency of the discriminant analysis with appropriate kernels while Table 3 gives an effective bandwidth selection using profile cross-validation. The method needs some improvements, particularly in terms of computation time, to be applicable to real mixed datasets with more than two variables. Further research can be done on bandwidth matrix selection, especially to speed it up for example by parallelism processing or also by Bayesian methods (Zougab et al., 2014). Also, regarding to symmetric kernels such as Gaussian or discrete triangular, comparisons can be done to measure the potential impacts of boundary bias on the accuracy of discriminant analysis.

Acknowledgements

We sincerely thank the Co-Editors and an anonymous referee for their valuable comments.

References

Aitchison, J. and Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion), *Journal of the Italian Statistical Society SeriesB*, 6(2):97–144.

Bouezmarni, T. and Rombouts, J.V.K. (2010). Nonparametric density estimation for multivariate bounded data, *Journal of Statistical Planning and Inference*, 140(1):139–152.

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

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

Duong, T. (2004). *Bandwidth Selectors for Multivariate Kernel Density Estimation.* Ph.D. Thesis Manuscript to University of Western Australia, Perth, Australia, October 2004.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R, *Journal of Statistical Software*, 21(7):1–16.

Gosh, A.K. and Chaudhury, P. (2004). Optimal smoothing in kernel analysis discriminant, *Statistica Sinica*, 14(2):457–483.

Gosh, A.K. and Hall, P. (2008). On error-rate estimation in nonparametric classification, *Statistica Sinica*, 18:1081–1100.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association*, 88(422):495–504.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York.

Hall, P. and Wand, M.P. (1988). On nonparametric discrimination using density differences. *Biometrika*, 75(3):541–547.

Halvorsen, K. (2015). `ElemStatLearn`: Data sets, functions and examples from the book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman, URL `http://cran.r-project.org/web/packages/ElemStatLearn/index.html`.

Hayfield, T. and Racine, J.S. (2007). Nonparametric econometrics: the np package, *Journal of Statistical Software*, 27(5):1–32.

Igarashi, G. and Kakizawa, Y. (2015). Bias correction for some asymmetric kernel estimators, *Journal of Statistical Planning and Inference*, 159:37–63.

Kokonendji, C.C. and Senga Kiessé, T. (2011). Discrete associated kernels method and extensions, *Statistical Methodology*, 8(6):497–516.

Kokonendji, C.C., Senga Kiessé, T. and Zocchi, S.S. (2007). Discrete triangular distributions and nonparametric estimation for probability mass function, *Journal of Nonparametric Statistics*, 19(6-8):241–254.

Kokonendji, C.C. and Somé, S.M. (2015). On multivariate associated kernels for smoothing some density function, `arXiv:1502.01173`.

Kokonendji, C.C., Zocchi, S.S. (2010). Extensions of discrete triangular distribution and boundary bias in kernel estimation for discrete functions, *Statistics and Probability Letters*, 80(21):1655–1662.

Koo, J.-Y, Park, C. and Jun, M. (2009). A classification spline machine for building a credit scorecard, *Journal of Statistical Computation and Simulation*, 79(5):681–689.

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, June 2013, LMB no. 14334, Besançon , URL `http://hal.in2p3.fr/tel-01124288/document`.

Malec, P. and Schienle, M. (2014). Nonparametric kernel density estimation near the boundary, *Computational Statistics and Data Analysis*, 72:57–76.

Nath, R., Jackson, W.M. and Jones, T.W. (1992). A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis, *Journal of Statistical Computation and Simulation*, 41(1-2):73–93.

`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/`.

Racine, J. and Li, Q. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. and Ferreira, J. (1983).

Coronary risk factor screening in three rural communities, *South African Medical Journal*, 64(12):430–436.

Sarmanov, O.V. (1966). Generalized normal correlation and two-dimensionnal Frechet classes, *Doklady (Soviet Mathematics)*, 25:1207–1222.

Scott, W.D. (1992). *Multivariate Density Estimation*, John Wiley and Sons, New York.

Shi, B., Moloney, K.P., Pan, Y., Leonard, V.K., Vidakovic, B., Jacko, J.A. and Sainfort, F. (2006). Wavelet classification of high frequency pupillary responses, *Journal of Statistical Computation and Simulation*, 76(5):431–445.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

Somé, S.M. and Kokonendji, C.C. (2016). Effects of associated kernels in nonparametric multiple regressions, *Journal of Statistical Theory and Practice*, 10(2):456–471.

Wahba, G. (1990). *Splines Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.

Ziane, Y., Adjabi, S. and Zougab, N. (2015). Adaptive Bayesian bandwidth selection in asymmetric kernel density estimation for nonnegative heavy-tailed data, *Journal of Applied Statistics*, 42(8):1645–1658.

Zhang, S. (2010). A note on the performance of the gamma kernel estimators at the boundary, *Statistics and Probability Letters*, 80(7):548–557.

Zhang, S. and Karunamuni, R.J. (2010). Boundary performance of the beta kernel estimators, *Journal of Nonparametric Statistics*, 22(1):81–104.

Zougab, N., Adjabi, S. and Kokonendji, C.C. (2012). Binomial kernel and Bayes local bandwidth in discrete functions estimation, *Journal of Nonparametrics Statistics*, 24(3):783–795.

Zougab, N., Adjabi, S. and Kokonendji, C.C. (2014). Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation, *Computational Statistics and Data Analysis*, 75:28–38.