Algorithmes séquentiels pour l'analyse de données par méthodes à noyau

Sequential algorithms for data analysis with kernel-based methods

C. Richard, F. Abdallah, R. Lengellé

Équipe M2S, Institut des Sciences et Technologies de l'Information de Troyes (ISTIT - FRE CNRS 2732)
Université de Technologie de Troyes (UTT), 12 rue Marie Curie, B.P. 2060, 10010 Troyes cedex, France
tél. : +33.3.25.71.58.47 fax. : +33.3.25.71.56.99, email : prenom.nom@utt.fr

Manuscrit reçu le 10 février 2004

Résumé et mots clés

Au cours de la dernière décennie, de multiples méthodes pour l'analyse et la classification de données fondées sur la théorie des espaces de Hilbert à noyau reproduisant ont été développées. Elles reposent sur le principe fondamental du kernel trick, initialement exploité par Vapnik et col. dans le cadre des Support Vector Machines. Celui-ci permet d'étendre au cas non-linéaire des traitements initialement linéaires en utilisant la notion de noyau. La méthode KFD, pour *Kernel Fisher Discriminant*, constitue ainsi une généralisation non-linéaire de l'analyse discriminante de Fisher. Bien que son efficacité soit indiscutable, on déplore le fait que sa mise en œuvre nécessite le stockage et la manipulation de matrices de dimension égale au nombre de données traitées, point critique lorsque l'ensemble d'apprentissage est de grande taille. Cet article présente un algorithme séquentiel palliant cette difficulté puisqu'il ne nécessite, ni la manipulation, ni même le stockage de telles matrices. Un parallèle est également proposé entre KFD et KPCA, acronyme de Kernel *Principal Component Analysis*, cette dernière méthode constituant une extension au cas non-linéaire de l'analyse en composantes principales. Cet article s'achève par la présentation d'un algorithme séquentiel à noyau pour la méthode GDA, *Generalized Discriminant Analysis*, qui étend l'analyse discriminante de Fisher au cas multi-classe.

Algorithmes séquentiels, ACP à noyau, AFD à noyau, noyaux de Mercer, SVM.

Abstract and key words

In recent years, many methods of analysis and classification of data based on reproducing kernel Hilbert spaces have been developed. Most of these methods incorporate the fundamental principle dictated by Vapnik et al. in Support Vector Machines, which consists in extending linear algorithms to the non-linear case by using kernels. Kernel Fisher Discriminant (KFD) is one of these nonlinear methods which provides interesting results in many practical cases. However, the use of KFD requires storage and processing of matrices whose size equals the number of available data. This may be critical when the training set is large. This paper presents a sequential KFD algorithm which does not require the manipulation of large matrices. Sequential algorithms that fulfil the same requirements as KFD are also presented to perform Kernel Principal Component Analysis (KPCA) and Kernel Generalized Discriminant Analysis (KGDA).

Sequential algorithms, KPCA, KFD, Mercer kernels, SVM.

1. Introduction

Le domaine de la Reconnaissance des Formes connaît une révolution depuis le milieu des années 90 avec l'avènement des noyaux reproduisants pour la résolution de problèmes de détection/classification et régression [16], [27]. Ceux-ci permettent en effet de conférer un caractère non-linéaire à nombre de traitements linéaires, sans qu'il soit nécessaire de recourir à d'importants développements théoriques. Aussi l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle Discriminante (AFD), outils standards en analyse de données, ont-elles été rapidement reformulées afin d'intégrer des caractéristiques non-linéaires. Aujourd'hui, elles sont communément désignées par KPCA [20] et KFD [9], acronymes respectifs de *Kernel Principal Component Analysis* et *Kernel Fisher Discriminant*. Si l'efficacité de ces deux méthodes est indiscutable, elles s'avèrent toutefois délicates à mettre en œuvre lorsque la taille de l'ensemble d'apprentissage est importante. Appliquées à une base regroupant n individus, toutes deux requièrent en effet le stockage et la manipulation de matrices de taille $(n \times n)$. Un algorithme de type EM a été récemment développé afin de remédier à cette situation dans le cadre de la méthode KPCA [8], [11], évolution directe d'une technique séquentielle initialement proposée pour l'ACP [19]. Plus complexe parce qu'elle ne repose pas directement sur la diagonalisation d'une matrice de covariance, la méthode KFD n'a pas encore suscitée autant d'intérêt du point de vue de sa mise en œuvre, exception faite à l'algorithme d'optimisation avec contraintes exposé dans [10].

L'un des objectifs de cet article est de remédier à ce manque en proposant un algorithme séquentiel pour la méthode KFD qui ne nécessite pas la manipulation de matrices de grande taille. Celui-ci est comparé à une approche séquentielle à noyau dédiée à la minimisation de l'erreur quadratique. On rappelle en effet que pour des sorties désirées convenablement choisies, la solution minimisant cette fonction coût est également optimum au sens du critère de Fisher [2], [4]. Un parallèle avec la méthode de KPCA est alors proposé, montrant au lecteur qu'un algorithme séquentiel de même type peut également être utilisé dans ce dernier cas. Enfin, l'article s'achève par la présentation d'un algorithme séquentiel à noyau pour la méthode GDA, *Generalized discriminant analysis* [1], [18], qui constitue une extension de l'analyse de Fisher au cas multi-classe. Auparavant, il convient de décrire le cadre algébrique offert par les espaces de Hilbert à noyau reproduisant, permettant par là-même de rappeler les propriétés clé ayant contribué au succès des noyaux de Mercer.

2. Espaces à noyau reproduisant et condition de Mercer

Soit $\mathcal{H}$ un espace fonctionnel hilbertien réel de produit scalaire $< \cdot ; \cdot >_{\mathcal{H}}$, composé de fonctions ψ continues sur un ensemble $\mathcal{X}$. D'après le théorème de représentation de Riesz, il existe une fonction unique $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ de la variable $\boldsymbol{x}_i$, étant donné $\boldsymbol{x}_j$, telle que

$$\psi(\boldsymbol{x}_j) = < \psi ; \kappa(\cdot,\boldsymbol{x}_j) >_{\mathcal{H}} , \qquad \forall \psi \in \mathcal{H}. \tag{1}$$

Dans cette expression, $\kappa(\cdot,\boldsymbol{x}_j)$ désigne une fonction définie sur $\mathcal{X}$, obtenue en fixant le second argument de κ à $\boldsymbol{x}_j$. Il en résulte que l'ensemble $\{\kappa(\cdot,\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ engendre $\mathcal{H}$, et que le produit scalaire $< \cdot ; \cdot >_{\mathcal{H}}$ ne nécessite d'être défini que sur cet ensemble de générateurs. Au vu de cette propriété, κ est appelé *noyau reproduisant* de $\mathcal{H}$. En notant $\phi(\boldsymbol{x})$ la fonction $\kappa(\cdot,\boldsymbol{x})$, l'équation (1) implique

$$< \phi(\boldsymbol{x}_i) ; \phi(\boldsymbol{x}_j) >_{\mathcal{H}} = \kappa(\boldsymbol{x}_j,\boldsymbol{x}_i), \tag{2}$$

pour tout $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. Ce résultat signifie que $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ fournit le produit scalaire des images dans $\mathcal{H}$ de toute paire d'éléments de l'ensemble $\mathcal{X}$. En autorisant l'exploitation de ce concept sans nécessairement connaître explicitement $\mathcal{H}$ et ϕ, la condition de Mercer a contribué aux plus récents développements des structures à noyau [12], [27]. On présente ci-dessous trois noyaux usuels qui vérifient cette condition, ce qui signifie qu'ils fournissent à moindre coût de calcul le produit scalaire des images de deux observations $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ par une application ϕ. Une liste plus complète de noyaux de Mercer peut être consultée dans [27].

2.1. Noyaux polynômiaux

Afin d'élaborer une règle de décision basée sur une statistique polynômiale de degré q, on utilise le noyau reproduisant suivant

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = (1 + < \boldsymbol{x}_i ; \boldsymbol{x}_j >)^q. \tag{3}$$

On peut en effet montrer que les composantes de l'application $\phi(\boldsymbol{x})$ associée sont alors les monômes de degrés inférieurs à q constitués des composantes de $\boldsymbol{x}$. Parce qu'ils sont fonction du produit scalaire des observations, de tels noyaux sont dits projectifs.

2.2. Noyaux exponentiels radiaux

Les noyaux de type *radial* dépendent de la distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ entre les observations. Ils ont fait l'objet d'une attention particulière dans la littérature en raison du rôle central qu'ils jouent

dans les méthodes d'estimation et de classification à base de noyaux ou de potentiels [3]. On compte parmi eux le noyau gaussien, défini par

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\beta_0), \tag{4}$$

où β_0 est appelé *largeur de bande*. Ce noyau est caractérisé par un continuum de valeurs propres, ce qui signifie que les composantes de $\boldsymbol{\phi}$ ne sont pas en nombre fini comme dans l'exemple (3). Enfin, le noyau exponentiel (5) offre souvent des solutions intéressantes en fournissant une surface de décision linéaire par morceaux dans l'espace des observations.

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\beta_0). \tag{5}$$

2.3. Noyaux sigmoïdaux

On peut élaborer un réseau de neurones à une couche cachée en choisissant le noyau sigmoïdal

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \tanh(\alpha_0 <\boldsymbol{x}_i ; \boldsymbol{x}_j> +\beta_0). \tag{6}$$

La qualité de noyau reproduisant de κ dépend des paramètres α_0 et β_0 sélectionnés, contrairement aux noyaux polynômiaux et radiaux présentés ci-dessus. S'ils ne sont pas convenablement choisis, il en résulte la perte du cadre rigoureux offert par les espaces de Hilbert à noyau reproduisant. Toutefois, les erreements de la pratique font que cette contrainte se trouve parfois relaxée. Des exemples de mise en œuvre de ces noyaux sont présentés plus loin.

3. Méthode KFD et algorithme de calcul séquentiel

3.1. Présentation générale

Supposons qu'on dispose d'un ensemble d'apprentissage constitué de n individus $\boldsymbol{x}_k$, se répartissant en n_1 et n_2 représentants des deux classes en compétition $\mathcal{C}_1$ et $\mathcal{C}_2$. La méthode KFD suppose que l'on applique préalablement une transformation non-linéaire $\boldsymbol{\phi}(\cdot)$ aux données, les représentant ainsi dans un espace image noté $\mathcal{F}$, puis que l'on recherche un vecteur $\boldsymbol{w}$ de sorte que la statistique $\lambda(\boldsymbol{x}) = \boldsymbol{w}^t\,\boldsymbol{\phi}(\boldsymbol{x})$ maximise le critère de Fisher défini ainsi [4] :

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^t\boldsymbol{B}\boldsymbol{w}}{\boldsymbol{w}^t\boldsymbol{V}\boldsymbol{w}}, \tag{7}$$

où

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1,2} n_i(\boldsymbol{m}_i^{\phi} - \boldsymbol{m}^{\phi})(\boldsymbol{m}_i^{\phi} - \boldsymbol{m}^{\phi})^t \tag{8}$$

$$\boldsymbol{V} = \frac{1}{n}\sum_{i=1,2}\sum_{\boldsymbol{x}\in\mathcal{C}_i}(\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{m}_i^{\phi})(\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{m}_i^{\phi})^t. \tag{9}$$

Dans les expressions ci-dessus, $\boldsymbol{m}_i^{\phi}$ et $\boldsymbol{m}^{\phi}$ désignent les moyennes dans $\mathcal{F}$ des données de la base d'apprentissage, respectivement avec et sans distinction de leur classe d'appartenance, soit :

$$\boldsymbol{m}^{\phi} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{\phi}(\boldsymbol{x}_k) \qquad \boldsymbol{m}_i^{\phi} = \frac{1}{n_i}\sum_{\boldsymbol{x}\in\mathcal{C}_i}\boldsymbol{\phi}(\boldsymbol{x}). \tag{10}$$

Notons au passage la *relation de Huygens* stipulant que $\boldsymbol{S} = \boldsymbol{V} + \boldsymbol{B}$, où $\boldsymbol{S}$ désigne la matrice de covariance estimée à partir des observations sans distinction de leur classe d'appartenance :

$$\boldsymbol{S} = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{m}^{\phi})(\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{m}^{\phi})^t. \tag{11}$$

La maximisation de $J(\boldsymbol{w})$ revient à trouver une direction $\boldsymbol{w}$ maximisant conjointement la dispersion inter-classe et minimisant la dispersion intra-classe, respectivement estimées à partir de $\boldsymbol{B}$ et $\boldsymbol{V}$. On montre que l'expression (7), connue sous le nom de *Quotient de Rayleigh*, est maximale lorsque $\boldsymbol{w}$ est un vecteur propre vérifiant l'équation

$$\boldsymbol{B}\boldsymbol{w} = \gamma\boldsymbol{V}\boldsymbol{w}, \tag{12}$$

où γ désigne l'unique valeur propre non nulle associée. On a alors $J(\boldsymbol{w}) = \gamma$. En combinant l'expression ci-dessus et la relation de Huygens, on aboutit au problème équivalent suivant

$$\boldsymbol{B}\boldsymbol{w} = \gamma(\boldsymbol{S} - \boldsymbol{B})\boldsymbol{w} = \gamma'\boldsymbol{S}\boldsymbol{w}, \tag{13}$$

avec $\gamma' = \dfrac{\gamma}{1+\gamma}$. Si l'on souhaite s'affranchir de la résolution des problèmes (12) ou (13), une alternative consiste à remarquer directement que $\boldsymbol{B}$ peut être reformulée selon

$$\boldsymbol{B} = \frac{n_1\,n_2}{n^2}\,(\boldsymbol{m}_2^{\phi} - \boldsymbol{m}_1^{\phi})(\boldsymbol{m}_2^{\phi} - \boldsymbol{m}_1^{\phi})^t, \tag{14}$$

et que $\boldsymbol{B}\boldsymbol{w}$ est en conséquence colinéaire à $(\boldsymbol{m}_2^{\phi} - \boldsymbol{m}_1^{\phi})$. À partir de la relation (12), on aboutit alors au système linéaire suivant

$$\boldsymbol{V}\boldsymbol{w} = (\boldsymbol{m}_2^{\phi} - \boldsymbol{m}_1^{\phi}), \tag{15}$$

la constante γ ayant été ignorée parce qu'elle n'influe pas sur $\boldsymbol{w}$. On note que la dimension de l'espace image $\mathcal{F}$ de $\boldsymbol{\phi}(\cdot)$ limite l'intérêt pratique de cette approche qui, en théorie, permet de donner un caractère non-linéaire à l'analyse discriminante de

Fisher grâce à une transformation préalable des données. Il est possible de contourner cet obstacle en remarquant que le domaine de recherche de $\boldsymbol{w}$ peut être limité à l'espace linéaire induit par $\boldsymbol{\phi}(\boldsymbol{x})$, toute composante $\boldsymbol{w}^{\perp}$ extraite de l'espace complémentaire étant sans effet sur la statistique $\lambda(\boldsymbol{x}) = \boldsymbol{w}^t\, \boldsymbol{\phi}(\boldsymbol{x})$. Plus précisément encore et conformément au *Representer Theorem* [21], l'information disponible sur l'observation $\boldsymbol{x}$ se bornant aux données $\boldsymbol{x}_k$ de l'ensemble d'apprentissage, le domaine de recherche de $\boldsymbol{w}$ peut être restreint à l'espace linéaire induit par les éléments $\boldsymbol{\phi}(\boldsymbol{x}_k)$, c'est-à-dire :

$$\boldsymbol{w} = \sum_{k=1}^{n} \alpha_k\, \boldsymbol{\phi}(\boldsymbol{x}_k). \tag{16}$$

Ceci mène directement à l'expression duale $\lambda(\boldsymbol{x}) = \boldsymbol{\alpha}^t\, \boldsymbol{\kappa}(\boldsymbol{x})$ de la statistique, avec $\boldsymbol{\alpha}$ le vecteur de composantes α_k et $\boldsymbol{\kappa}(\boldsymbol{x}) = [\kappa(\boldsymbol{x},\boldsymbol{x}_1)\ \ldots\ \kappa(\boldsymbol{x},\boldsymbol{x}_n)]^t$. En substituant $\boldsymbol{S}$ à $\boldsymbol{V}$ dans la définition (7) compte tenu de l'équivalence observée entre les problèmes (12) et (13), puis en combinant celle-ci avec l'expression (16), on aboutit finalement au critère suivant :

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^t \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}. \tag{17}$$

Dans cette expression, $\boldsymbol{N} = \boldsymbol{K}\boldsymbol{K}^t - n\boldsymbol{\mu}\boldsymbol{\mu}^t$ et $\boldsymbol{M} = \boldsymbol{\delta}\,\boldsymbol{\delta}^t$ avec $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, où $\boldsymbol{\mu}$ et $\boldsymbol{\mu}_i$ désignent les moyennes des vecteurs $\boldsymbol{\kappa}(\boldsymbol{x})$ de la base d'apprentissage, respectivement sans et avec distinction de leur classe $\mathcal{C}_i$ d'appartenance. De plus, $\boldsymbol{K}$ désigne la matrice de Gram de terme général $K_{ij} = \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$. Par analogie avec le critère (7) et le résultat (15) de sa maximisation, on en déduit que $\boldsymbol{\alpha}$ peut être obtenu en résolvant le système linéaire suivant :

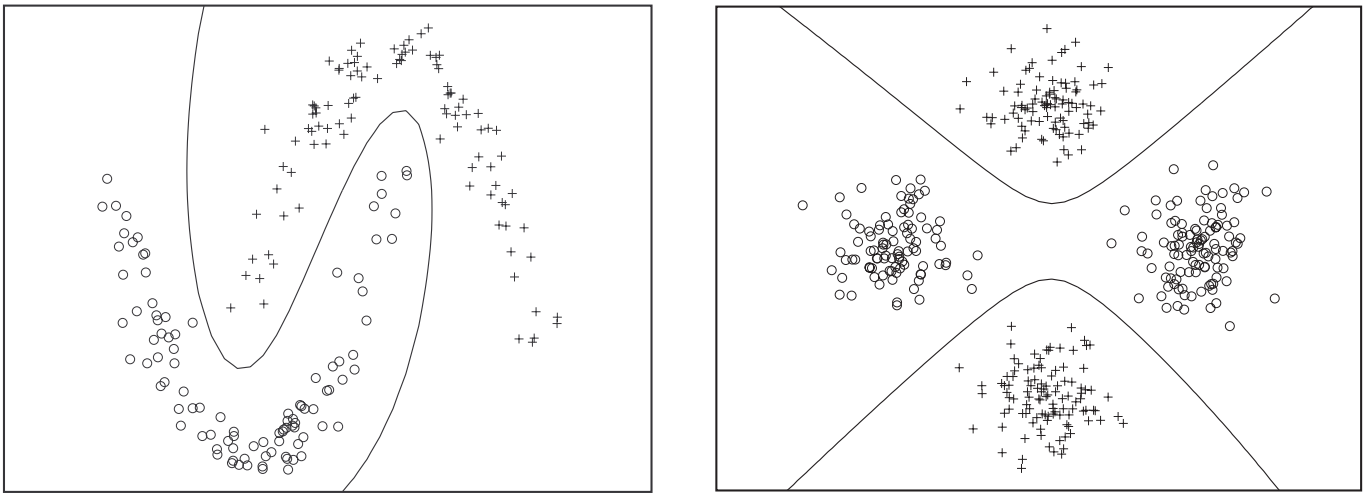
$$\boldsymbol{N}\boldsymbol{\alpha} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{18}$$

On note que la taille de ce problème est indépendante de la dimension de l'espace image $\mathcal{F}$, ce qui permet d'accéder à une très grande diversité d'espaces d'analyse au prix d'une charge calculatoire fixe. La figure 1 illustre les possibilités offertes par cette méthode sur des problèmes issus de [6], où ils servent à illustrer les limites d'une approche linéaire. En particulier, la figure 1.(a) correspond au choix d'un noyau polynômial de degré 3, tandis que la figure 1.(b) résulte d'un noyau gaussien de largeur de bande unité. Si les possibilités offertes par une telle approche semble illimitées, il convient toutefois de noter que la taille du système (18) est égale à celle de la base d'apprentissage. Parce que cette singularité peut mener à une situation de blocage lorsque les données disponibles abondent, on présente un algorithme séquentiel ne nécessitant pas la manipulation de matrices de grandes tailles.

3.2. Algorithme séquentiel

Afin de maximiser le critère de Fisher sans avoir à résoudre le système (18), on peut envisager de recourir à un algorithme de *descente du gradient*. Celui-ci débute avec une estimation de la solution $\boldsymbol{\alpha}$ qu'il met à jour itérativement en suivant la ligne de plus grande pente de la fonction coût, indiquée par le gradient de celle-ci.
Pour des raisons qui apparaîtront clairement au cours des calculs à suivre, on s'intéresse à l'inverse du critère de Fisher qu'il s'agit donc de minimiser :

$$J_{\mathrm{KFD}}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}{(\boldsymbol{\alpha}^t \boldsymbol{\delta})^2}. \tag{19}$$



(a) **(b)**

Figure 1. Application de la méthode KFD à des données synthétiques en utilisant un (a) noyau polynômial de degré 3 et un (b) noyau gaussien de largeur de bande 1

L'expression du gradient $\nabla J_{\mathrm{KFD}}(\boldsymbol{\alpha})$ est donnée par :

$$\nabla J_{\mathrm{KFD}}(\boldsymbol{\alpha}) = \frac{2}{(\boldsymbol{\alpha}^t \boldsymbol{\delta})^2} \left[\boldsymbol{N}\boldsymbol{\alpha} - \frac{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{\delta}} \boldsymbol{\delta} \right]. \quad (20)$$

Clairement, la norme de $\boldsymbol{\alpha}$ n'influe aucunement sur la valeur du critère (19). Aussi $\boldsymbol{\alpha}$ peut-il être normalisé de sorte que l'on ait $\boldsymbol{\alpha}^t \boldsymbol{\delta} = 1$. Dans ces conditions et après quelques calculs, l'expression (20) mène directement au terme de mise-à-jour suivant

$$\Delta \boldsymbol{\alpha} = \sum_{k=1}^{n} y_k \left[\boldsymbol{\kappa}(\boldsymbol{x}_k) - y_k \, \boldsymbol{\delta} \right] + n \, (\boldsymbol{\alpha}^t \boldsymbol{\mu}) \left[(\boldsymbol{\alpha}^t \boldsymbol{\mu}) \, \boldsymbol{\delta} - \boldsymbol{\mu} \right], \quad (21)$$

où l'on a posé $y_k = \boldsymbol{\alpha}^t \, \boldsymbol{\kappa}(\boldsymbol{x}_k)$. Il est à noter que l'expression (21) est plus simple que celle figurant dans [17]. Ceci résulte de la substitution de $\boldsymbol{S}$ à $\boldsymbol{V}$ dans la définition du critère (7). On aboutit finalement à l'algorithme séquentiel qui suit.

1. Initialiser aléatoirement $\boldsymbol{\alpha}$
2. Calculer $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ et $\boldsymbol{\delta}$
3. Normaliser $\boldsymbol{\alpha}$ selon $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}/(\boldsymbol{\alpha}^t \boldsymbol{\delta})$, puis calculer $\Delta \boldsymbol{\alpha}$
4. Rafraîchir $\boldsymbol{\alpha}$ selon $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \, \Delta \boldsymbol{\alpha}$, avec $\eta > 0$
5. Retourner en 3. jusqu'à convergence de l'algorithme.

On constate que cette nouvelle approche ne nécessite, ni le stockage et la manipulation d'une matrice de taille $(n \times n)$ [9], ni l'optimisation d'une fonction quadratique avec $(n+2)$ contraintes d'égalité [10]. Cette méthode séquentielle s'avère ainsi être une alternative intéressante à l'algorithme KFD original lorsque la base d'apprentissage est de taille importante. Il convient de noter que cet algorithme s'inspire d'une technique séquentielle proposée dans le cadre de l'analyse discriminante linéaire de Fisher [15]. Elle est elle-même une réminiscence d'une méthode d'apprentissage appelée Règle d'Oja [13].

3.3. Expérimentations et comparaison

La méthode proposée a été expérimentée sur un ensemble de données synthétiques distribuées dans le plan selon deux hyperboloïdes, comme l'indique la figure 2. Ces dernières représentent les classes $\mathcal{C}_1$ et $\mathcal{C}_2$ en compétition, et regroupent chacune 200 individus. La figure considérée montre la frontière de décision obtenue avec un noyau exponentiel de largeur de bande 0.2. Afin d'apprécier les qualités de convergence de l'algorithme proposé, celui-ci a été comparé à l'algorithme séquentiel proposé pour la méthode KMSE, acronyme de *Kernel Mean Square Error* [2]. Cette dernière a pour vocation d'élaborer des détecteurs à noyau qui minimisent l'erreur quadratique entre sorties obtenues et sorties désirés :

$$J_{\mathrm{KMSE}}(\boldsymbol{\alpha}) = \|\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}_d\|^2, \quad (22)$$

où $\boldsymbol{y}_d$ est un vecteur regroupant les sorties désirées, tandis que $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{\alpha}$ représente les sorties obtenues. On rappelle qu'un choix approprié des sorties désirées conduit à une solution optimum au sens du critère de Fisher. En l'occurrence, il s'agit de poser $y_d[k] = n/n_1$ si $\boldsymbol{x}_k$ est issu de $\mathcal{C}_1$, et $y_d[k] = -n/n_2$ si il provient de $\mathcal{C}_2$. Le calcul du gradient de $J_{\mathrm{KMSE}}(\boldsymbol{\alpha})$ donne

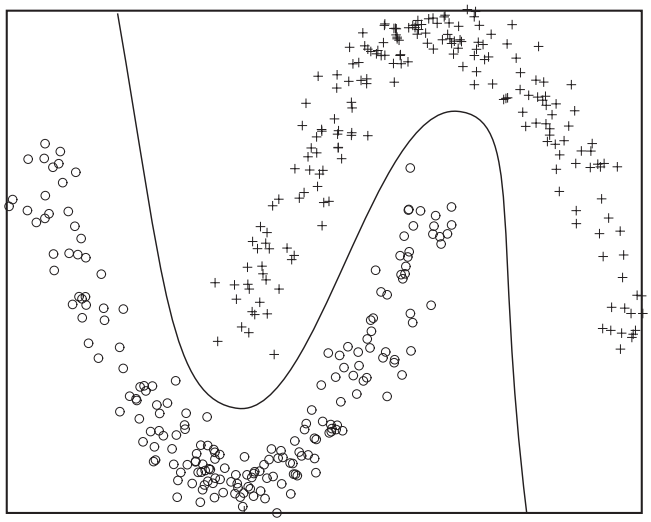


Figure 2. Frontière de décision obtenue grâce à la méthode KFD séquentielle. Un noyau exponentiel de largeur de bande 0.2 a été utilisé

$$\nabla J_{\mathrm{KMSE}}(\boldsymbol{\alpha}) = \boldsymbol{K}^t \boldsymbol{K} \boldsymbol{\alpha} - \boldsymbol{K}^t \boldsymbol{y}_d. \quad (23)$$

Finalement, le terme de mise-à-jour recherché est fournit par

$$\Delta \boldsymbol{\alpha} = \boldsymbol{K}^t \boldsymbol{y} - \boldsymbol{K}^t \boldsymbol{y}_d, \quad (24)$$

que l'on peut écrire pour chaque composante de $\boldsymbol{\alpha}$ ainsi : $\Delta \alpha_k = \boldsymbol{\kappa}^t(\boldsymbol{x}_k)(\boldsymbol{y} - \boldsymbol{y}_d)$. L'algorithme séquentiel proposé pour la méthode KMSE prend finalement la forme qui suit.

1. Initialiser aléatoirement $\boldsymbol{\alpha}$
2. Calculer les composantes $\Delta \alpha_k$
3. Rafraîchir α_k selon $\alpha_k \leftarrow \alpha_k - \eta \, \Delta \alpha_k$, avec $\eta > 0$
4. Retourner en 2. jusqu'à convergence de l'algorithme

La figure 3 compare les vitesses de convergence des algorithmes séquentiels KFD et KMSE. Ces expérimentations ont été conduites sur le problème représenté en figure 2, en fixant arbitrairement le pas η à 0.05. Si les critères optimisés (19) et (22) mènent théoriquement à des solutions équivalentes, ils diffèrent toutefois par leur nature. Ceci se traduit par des surfaces d'erreur de types différents, comme cela est montré par [15] dans le cas linéaire. Il en résulte au final des vitesses de convergence différentes.

3.4. Connections avec la méthode KPCA

La méthode KFD a pour but de concentrer le caractère discriminant des données tandis que la méthode KPCA a trait à leur représentation fidèle. Il est toutefois intéressant de constater

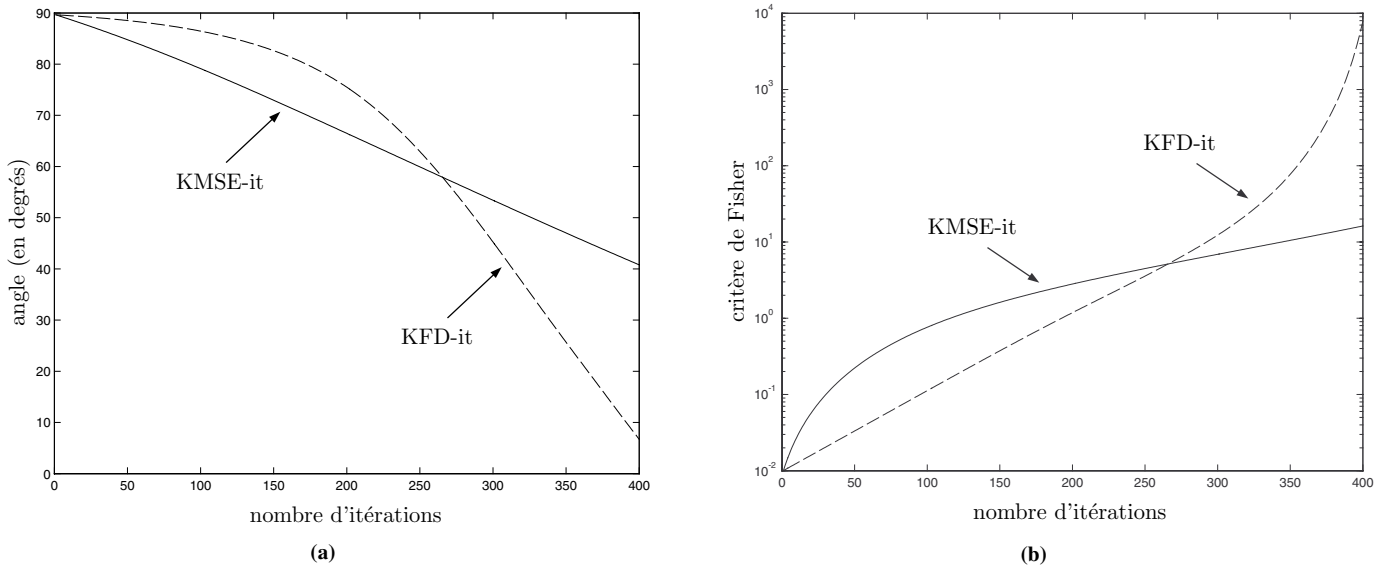


Figure 3. (a) Angle entre les vecteurs α obtenus à l'aide des méthodes KFD et KMSE séquentielles et le vecteur optimal au sens du critère de Fisher obtenu par un calcul direct [9], en fonction du nombre d'itérations. (b) Valeur du critère de Fisher en fonction du nombre d'itérations

que, malgré cette divergence d'objectifs, des techniques comparables peuvent être employées pour résoudre ces deux problèmes. On rappelle que l'extraction d'un axe principal d'inertie par la méthode KPCA consiste en la maximisation du critère suivant [20] :

$$J_{\mathrm{KPCA}}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{K} \boldsymbol{\alpha}}. \tag{25}$$

Dans cette expression, $\boldsymbol{N} = \boldsymbol{K}\boldsymbol{K}^t - n\,\boldsymbol{\mu}\,\boldsymbol{\mu}^t$ où $\boldsymbol{\mu}$ désigne la moyenne des $\boldsymbol{\kappa}(\boldsymbol{x})$ de la base d'apprentissage. Le calcul du gradient de $J_{\mathrm{KPCA}}(\boldsymbol{\alpha})$ mène à l'expression

$$\nabla J_{\mathrm{KPCA}}(\boldsymbol{\alpha}) = \frac{2}{\boldsymbol{\alpha}^t \boldsymbol{K} \boldsymbol{\alpha}} \left[\boldsymbol{N}\boldsymbol{\alpha} - \frac{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{K} \boldsymbol{\alpha}} \boldsymbol{K}\boldsymbol{\alpha} \right]. \tag{26}$$

Le vecteur $\boldsymbol{K}\boldsymbol{\alpha}$ a pour composantes $y_k = \boldsymbol{\alpha}^t \boldsymbol{\kappa}(\boldsymbol{x}_k)$. Il est en conséquence noté $\boldsymbol{y}$. Comme dans la section précédente, seule la direction de $\boldsymbol{\alpha}$ importe. Ce vecteur peut donc être normalisé de sorte que $\boldsymbol{\alpha}^t \boldsymbol{K} \boldsymbol{\alpha} = 1$, soit $\boldsymbol{\alpha}^t \boldsymbol{y} = 1$. Après quelques calculs destinés à éliminer $\boldsymbol{N}$ dans l'expression (26), on aboutit au terme de mise-à-jour suivant

$$\Delta\boldsymbol{\alpha} = \sum_{k=1}^{n} y_k \left[\boldsymbol{\kappa}(\boldsymbol{x}_k) - y_k\, \boldsymbol{y} \right] + n\, (\boldsymbol{\alpha}^t \boldsymbol{\mu}) \left[(\boldsymbol{\alpha}^t \boldsymbol{\mu})\, \boldsymbol{y} - \boldsymbol{\mu} \right]. \tag{27}$$

On retrouve ainsi l'expression (21) dans laquelle $\boldsymbol{y}$ est venu se substituer à $\boldsymbol{\delta}$. Il en résulte des algorithmes semblables, mise à part la différence mineure qui vient d'être soulignée, ainsi que le choix de la normalisation de $\boldsymbol{\alpha}$.

4. Analyse discriminante multi-classe

4.1 Présentation générale

L'analyse discriminante multi-classe repose sur la projection des n individus $\boldsymbol{x}_k$ d'une base d'apprentissage, regroupés selon c classes $\mathcal{C}_i$ de cardinal n_i, sur un espace de dimension $(c-1)$ de sorte à conjointement maximiser la dispersion inter-classe et minimiser la dispersion intra-classe [4], [6]. Ces dernières sont respectivement estimées à partir des matrices $\boldsymbol{B}$ et $\boldsymbol{V}$ définies ci-dessous, généralisant les définitions (8) et (9) établies dans le cas de deux classes. Si l'on suppose comme dans les sections précédentes que les données ont été préalablement transformées au moyen d'une application non-linéaire $\boldsymbol{\phi}(\cdot)$, ces matrices s'expriment respectivement ainsi

$$\boldsymbol{B} = \frac{1}{n} \sum_{i=1}^{c} n_i (\boldsymbol{m}_i^{\boldsymbol{\phi}} - \boldsymbol{m}^{\boldsymbol{\phi}})(\boldsymbol{m}_i^{\boldsymbol{\phi}} - \boldsymbol{m}^{\boldsymbol{\phi}})^t, \tag{28}$$

$$\boldsymbol{V} = \frac{1}{n} \sum_{i=1}^{c} \sum_{\boldsymbol{x} \in \mathcal{C}_i} (\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{m}_i^{\boldsymbol{\phi}})(\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{m}_i^{\boldsymbol{\phi}})^t. \tag{29}$$

La projection d'une observation $\boldsymbol{\phi}(\boldsymbol{x})$ sur l'espace de dimension $(c-1)$ recherché est pratiquée avec $(c-1)$ fonctions discriminantes $z_i = \boldsymbol{w}_i^t\, \boldsymbol{\phi}(\boldsymbol{x})$ qu'il s'agit de déterminer, opération que traduit la formulation matricielle suivante

$$\boldsymbol{z} = \boldsymbol{W}^t \boldsymbol{\phi}(\boldsymbol{x}), \tag{30}$$

où $\boldsymbol{z}$ est un vecteur de composantes z_i et $\boldsymbol{W}$ une matrice comportant $(c-1)$ colonnes $\boldsymbol{w}_i$. La méthode *Generalized Discriminant Analysis* (GDA) consiste à rechercher $\boldsymbol{W}$ de sorte à maximiser la fonction coût suivante [1], [4], [6]

$$J(\boldsymbol{W}) = \frac{|\boldsymbol{W}^t \boldsymbol{B} \boldsymbol{W}|}{|\boldsymbol{W}^t \boldsymbol{V} \boldsymbol{W}|} \tag{31}$$

où la notation $|\cdot|$ désigne le déterminant d'une matrice. On montre que les colonnes $\boldsymbol{w}_i$ de la solution sont des vecteurs propres vérifiant l'équation[1]

$$\boldsymbol{B}\boldsymbol{w}_i = \gamma_i \boldsymbol{S}\boldsymbol{w}_i \tag{32}$$

associés aux plus grandes valeurs propres γ_i non nulles. Ainsi note-t-on que le vecteur propre $\boldsymbol{w}$ associé à la plus grande valeur propre γ maximise le quotient de Rayleigh suivant :

$$\gamma = \frac{\boldsymbol{w}^t \boldsymbol{B} \boldsymbol{w}}{\boldsymbol{w}^t \boldsymbol{S} \boldsymbol{w}}. \tag{33}$$

En pratique, la résolution de ce problème peut s'avérer délicate puisqu'il est de la dimension de l'espace image $\mathcal{F}$ de $\boldsymbol{\phi}(\cdot)$. Comme dans les sections précédentes, il est toutefois possible de contourner cet obstacle est constatant que le domaine de recherche de chaque colonne $\boldsymbol{w}_i$ de la matrice $\boldsymbol{W}$ peut être restreint à l'espace linéaire induit par les éléments $\boldsymbol{\phi}(\boldsymbol{x}_k)$ constituant l'ensemble d'apprentissage, soit :

$$\boldsymbol{w}_i = \sum_{k=1}^{n} \alpha_{ik}\, \boldsymbol{\phi}(\boldsymbol{x}_k). \tag{34}$$

Il en résulte immédiatement que $z_i = \boldsymbol{w}_i^t\, \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\alpha}_i^t\, \boldsymbol{\kappa}(\boldsymbol{x})$ où $\boldsymbol{\alpha}_i$ est le vecteur dual associé à $\boldsymbol{w}_i$ et $\boldsymbol{\kappa}(\boldsymbol{x}) = [\kappa(\boldsymbol{x},\boldsymbol{x}_1)\ \ldots\ \kappa(\boldsymbol{x},\boldsymbol{x}_n)]^t$. Dans ces conditions, l'expression (33) que l'on cherche à maximiser se réécrit ainsi

$$\gamma = \frac{\boldsymbol{\alpha}^t \boldsymbol{P} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}. \tag{35}$$

Ci-dessus, $\boldsymbol{P} = \sum_{i=1}^{c} n_i\, \boldsymbol{\delta}_i\, \boldsymbol{\delta}_i^t$ avec $\boldsymbol{\delta}_i = \boldsymbol{\mu} - \boldsymbol{\mu}_i$, où $\boldsymbol{\mu}$ et $\boldsymbol{\mu}_i$ désignent toujours les moyennes des vecteurs $\boldsymbol{\kappa}(\boldsymbol{x})$ de la base d'apprentissage, respectivement sans et avec distinction de leur classe $\mathcal{C}_i$ d'appartenance. La définition de $\boldsymbol{N}$ demeure quant à elle évidemment inchangée. On note que la taille du problème est à présent indépendante de la dimension de l'espace image $\mathcal{F}$, augurant une grande souplesse de la méthode. Elle est en revanche égale à celle de l'ensemble d'apprentissage, source de difficultés lorsque les données disponibles abondent. La section suivante est consacrée à la présentation d'un algorithme séquentiel à même de pallier ce problème puisqu'il ne nécessite pas la manipulation de matrices de grandes tailles.

[1] Compte tenu de la relation de Huygens, on peut considérer indifféremment les équations $\boldsymbol{B}\boldsymbol{w}_i = \gamma_i \boldsymbol{V}\boldsymbol{w}_i$ et $\boldsymbol{B}\boldsymbol{w}_i = \gamma_i \boldsymbol{S}\boldsymbol{w}_i$. La seconde se prête toutefois mieux aux développements à venir, ce qui justifie le choix effectué.

4.2 Algorithme séquentiel

Dans le but de mettre en œuvre une stratégie de descente du gradient pour extraire le premier axe discriminant, on considère la fonction coût suivante qu'il convient de minimiser :

$$J_{\mathrm{KGDA}}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{P} \boldsymbol{\alpha}}. \tag{36}$$

Le gradient de celle-ci est donnée par l'expression

$$\nabla J_{\mathrm{KGDA}}(\boldsymbol{\alpha}) = \frac{2}{\boldsymbol{\alpha}^t \boldsymbol{P} \boldsymbol{\alpha}} \left[\boldsymbol{N}\boldsymbol{\alpha} - \frac{\boldsymbol{\alpha}^t \boldsymbol{N} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^t \boldsymbol{P} \boldsymbol{\alpha}} \boldsymbol{P}\boldsymbol{\alpha} \right]. \tag{37}$$

Seule la direction du vecteur $\boldsymbol{\alpha}$ comptant pour la résolution du problème, celui-ci peut être normalisé de sorte que $\boldsymbol{\alpha}^t \boldsymbol{P} \boldsymbol{\alpha} = 1$. On peut effectuer cette opération en divisant $\boldsymbol{\alpha}$ par $\|\boldsymbol{\alpha}^t \boldsymbol{M} \boldsymbol{D}^{\frac{1}{2}}\|$, où $\boldsymbol{D}$ résulte de la diagonalisation de $\boldsymbol{P}$ par la matrice de passage $\boldsymbol{M}$. Afin d'éviter toute manipulation de $\boldsymbol{P}$ puisque celle-ci est de la taille de la base d'apprentissage, on constate préalablement qu'elle se réécrit sous la forme $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^t$ avec $\boldsymbol{A} = [\sqrt{n_1}\boldsymbol{\delta}_1\ \ldots\ \sqrt{n_c}\boldsymbol{\delta}_c]$ et qu'elle est de rang maximal c. Il s'en suit que $\boldsymbol{A}^t\boldsymbol{A}$ et $\boldsymbol{P}$ ont mêmes valeurs propres non-nulles, et que les vecteurs propres $\boldsymbol{v}_i$ de la première fournissent ceux de la seconde puisqu'ils sont donnés par $\boldsymbol{A}\boldsymbol{v}_i$. La combinaison de cette normalisation, peu exigeante en calculs, avec la relation (37) mène au terme de mise-à-jour suivant

$$\Delta\boldsymbol{\alpha} = \sum_{k=1}^{n} y_k\, [\boldsymbol{\kappa}(\boldsymbol{x}_k) - y_k\, \boldsymbol{\xi}] + n\, (\boldsymbol{\alpha}^t \boldsymbol{\mu}) \left[(\boldsymbol{\alpha}^t \boldsymbol{\mu})\, \boldsymbol{\xi} - \boldsymbol{\mu} \right], \tag{38}$$

où l'on a introduit la notation $\boldsymbol{\xi} = \sum_{i=1}^{c} n_i\, (\boldsymbol{\alpha}^t \boldsymbol{\delta}_i)\, \boldsymbol{\delta}_i$ afin de mettre en évidence l'homogénéité des expressions (21), (27) et (38). L'algorithme séquentiel d'extraction du premier axe discriminant prend finalement la forme qui suit.

1. Initialiser aléatoirement $\boldsymbol{\alpha}$
2. Calculer $\boldsymbol{\mu}$, $\boldsymbol{\mu}_i$ et $\boldsymbol{\delta}_i$
3. Diagonaliser $\boldsymbol{A}^t\boldsymbol{A}$ pour déterminer $\boldsymbol{M}$ et $\boldsymbol{D}$
4. Normaliser $\boldsymbol{\alpha}$ selon $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} / \|\boldsymbol{\alpha}^t \boldsymbol{M} \boldsymbol{D}^{\frac{1}{2}}\|$ puis calculer $\Delta\boldsymbol{\alpha}$
5. Rafraîchir $\boldsymbol{\alpha}$ selon $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta\, \Delta\boldsymbol{\alpha}$, avec $\eta > 0$
6. Retourner en 4. jusqu'à convergence de l'algorithme.

Afin de déterminer le deuxième axe discriminant $\boldsymbol{\alpha}_2$, on suggère d'adopter une procédure de déflation. Celle-ci vise à éliminer la contribution du premier axe $\boldsymbol{\alpha}_1$ calculé ci-dessus, afin de pouvoir réappliquer un algorithme similaire. On exploite pour ce faire le fait que les axes $\boldsymbol{\alpha}_i$ sont orthogonaux au sens de la métrique P. Cette propriété résulte de l'expression (32) reconsidérée au jour de la formulation (35) du critère optimisé. On a en effet

$$\boldsymbol{\alpha}_j^t (\boldsymbol{P}\boldsymbol{\alpha}_i) = \boldsymbol{\alpha}_j^t (\gamma_i\, \boldsymbol{N}\boldsymbol{\alpha}_i) = \frac{\gamma_i}{\gamma_j}\, \boldsymbol{\alpha}_j^t \boldsymbol{P} \boldsymbol{\alpha}_i \tag{39}$$

quel que soit le problème traité. Il en résulte directement le résultat escompté, c'est-à-dire

$$\boldsymbol{\alpha}_i^t P \boldsymbol{\alpha}_j = 0, \quad \forall i \neq j. \tag{40}$$

Afin de déterminer l'axe discriminant $\boldsymbol{\alpha}_2$ à la suite de $\boldsymbol{\alpha}_1$, on suggère d'adopter l'algorithme présenté ci-dessus dans lequel l'opération complémentaire suivante fait suite à l'étape 5.

$$\boldsymbol{\alpha}_2 \leftarrow \boldsymbol{\alpha}_2 - \frac{\boldsymbol{\alpha}_2^t \boldsymbol{P} \boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^t \boldsymbol{P} \boldsymbol{\alpha}_1} \boldsymbol{\alpha}_1, \tag{41}$$

garantissant ainsi que $\boldsymbol{\alpha}_2^t \boldsymbol{P} \boldsymbol{\alpha}_1 = 0$. On note que l'on a $\boldsymbol{\alpha}_1^t \boldsymbol{P} \boldsymbol{\alpha}_1 = 1$ si l'on a pris soin de normaliser préalablement $\boldsymbol{\alpha}_1$ selon l'étape 4. avant de procéder à la recherche de $\boldsymbol{\alpha}_2$. Sous cette condition et après quelques calculs destinés à éliminer $\boldsymbol{P}$ dans l'expression (41), cette dernière se réécrit ainsi :

$$\boldsymbol{\alpha}_2 \leftarrow \boldsymbol{\alpha}_2 - \sum_{i=1}^{c} n_i \, (\boldsymbol{\alpha}_1^t \boldsymbol{\delta}_i) \, (\boldsymbol{\alpha}_2^t \boldsymbol{\delta}_i) \, \boldsymbol{\alpha}_1. \tag{42}$$

La généralisation de cet algorithme à l'extraction de $(c-1)$ axes ne présente aucune difficulté. Elle est laissée au soin du lecteur.

4.3 Experimentations

Les données Iris consistent en 150 individus de dimension 4 répartis en 3 classes [1]. Celles-ci sont représentées en figure 4.(a) dans le plan défini par les deux premiers axes factoriels obtenus à l'issu d'une ACP classique, donnant un aperçu de la configuration des classes. La figure 4.(b) représente les données dans le plan défini par les axes discriminants $\boldsymbol{\alpha}_1$ et $\boldsymbol{\alpha}_2$, obtenus à l'issu de l'algorithme décrit au cours de la section précédente. Afin d'obtenir ce résultat, un noyau gaussien de largeur de bande unité a été utilisé et le pas η de descente a été arbitrairement fixé à 0,02. La figure 4.(c) présente l'évolution du critère (36) lors de la détermination de l'axe $\boldsymbol{\alpha}_2$ en fonction du nombre d'itérations, laissant apprécier les qualités intéressantes de convergence de l'algorithme. Les trois représentations composant la figure 5 reprennent les mêmes informations pour un jeu de données différent. Celles-ci se répartissent ici selon 3 classes décrivant des cercles concentriques.

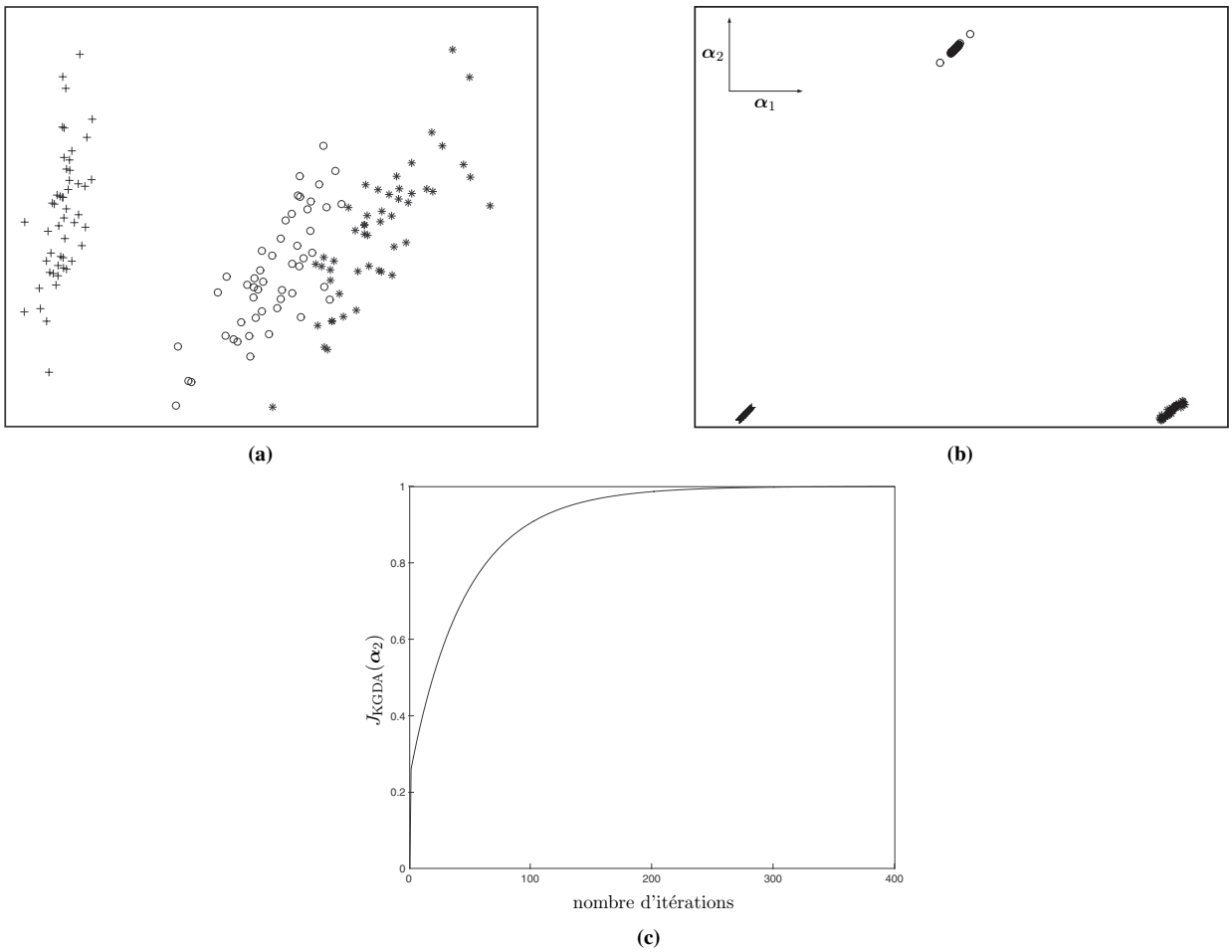


Figure 4. Représentations des données Iris obtenues au terme (a) d'une ACP conventionnelle pour illustrer la configuration des classes, (b) dans le plan défini par les axes discriminants $\boldsymbol{\alpha}_1$ et $\boldsymbol{\alpha}_2$ fournis par l'algorithme séquentiel proposé. La convergence de celui-ci est illustrée en (c), où l'on représente l'évolution du critère $J_{KGDA}(\boldsymbol{\alpha}_2)$ en fonction du nombre d'itérations.

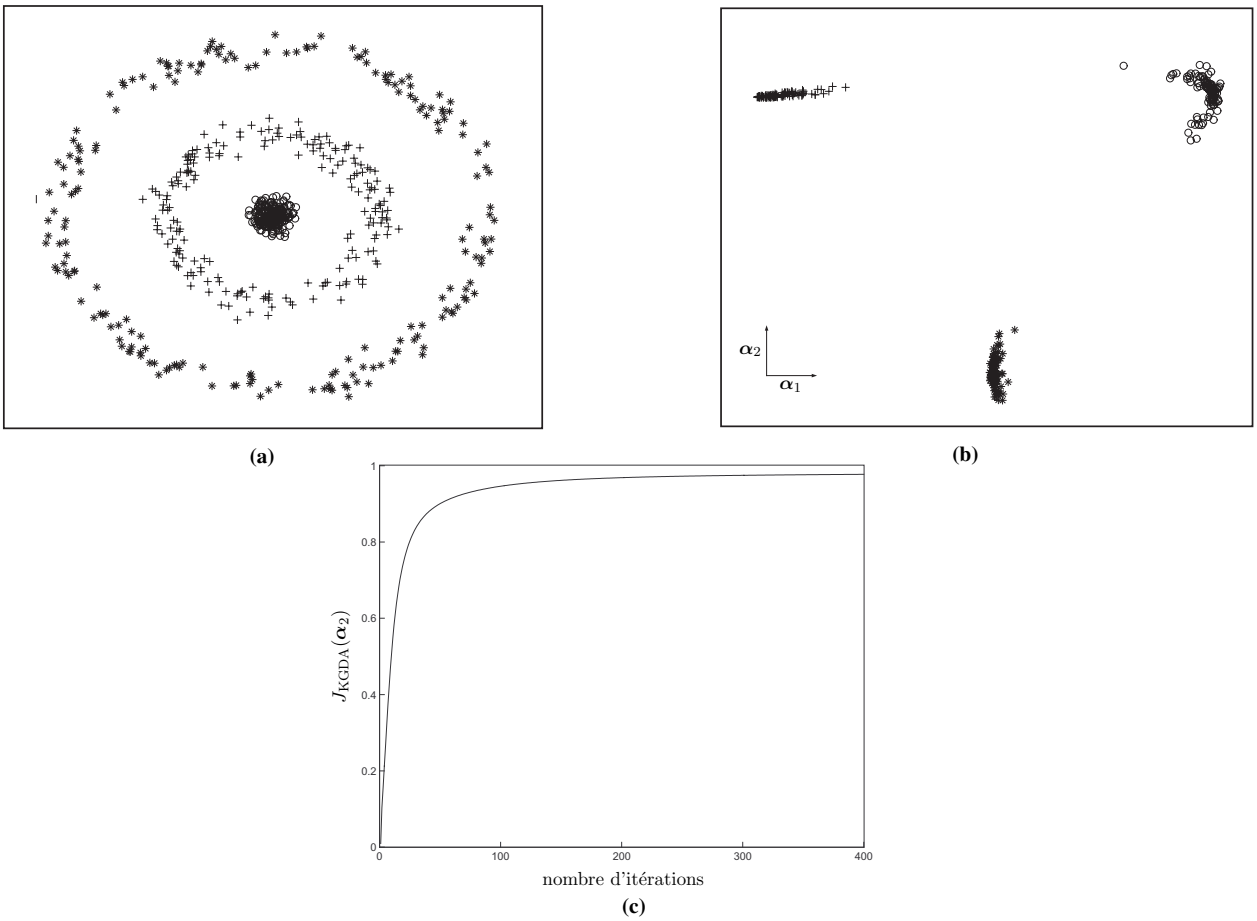


Figure 5. Voir légende de la Fig. 4, ici avec un autre jeu de données.

5. Conclusion

En raison de leurs performances et de leur simplicité de mise en œuvre exceptionnelles, les méthodes à noyau suscitent depuis près de dix ans un intérêt sans cesse grandissant de l'ensemble des communautés ayant trait au traitement automatique de l'information. Parmi les techniques largement bénéficiaires de cette révolution méthodologique, on compte les méthodes classiques d'analyse de données telles que l'analyse en composantes principales et l'analyse factorielle discriminante. Celles-ci constituent le sujet central du présent article, où elles font l'objet d'une description succincte accompagnée de points d'entrée bibliographiques. Les aspects algorithmiques sont ensuite privilégiés, partant de la constatation que ces méthodes nécessitent le stockage et la manipulation de matrices dont la taille est égale au nombre de données traitées. Afin de s'affranchir de cette source de difficultés lorsque les données disponibles sont nombreuses, on propose une famille d'algorithmes séquentiels reposant sur une descente de gradient. La présentation adoptée dans le présent article met l'accent sur le fait qu'une même structure peut être utilisée selon les besoins pour effectuer une analyse en composantes principales ou une analyse factorielle discriminante à noyau, à deux classes ou plus. Si les méthodes proposées remplissent parfaitement leur rôle comme l'illustrent les simulations, elles sont toutefois susceptibles de bénéficier encore des différentes stratégies d'optimisation existantes et dédiées au type de problème traité, répertoriées dans [5].

Références

[1] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Computation*, vol. 12, n° 10, pp. 2385-2404, 2000.

[2] S. A. Billings, K. L. Lee, Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm, *Neural Networks*, vol. 15, pp. 263-270, 2002.

[3] L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, New York : Springer-Verlag, 1996.

[4] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, New York : Wiley and Sons, 2001.

[5] S. C. Douglas, S.-I. Amari, S.-Y. Kung, Gradient adaptation with unit-norm constraints, *Technical Report*, n° EE-99-003, Southern Methodist University, Dallas, 1999.

[6] K. Fukunaga, *Statistical Pattern Recognition*, San Diego : Academic Press, 1990.

[7] R. Lengellé, C. Richard, Apprentissage de règles de décision à structure imposée et contrôle de la complexité (33 p.), In R. Lengellé, (éd.), *Reconnaissance des Formes et Décision en Signal*, Paris : Hermès Sciences, Traité IC2, 2002.

[8] S. Mika, *Kernalgorithmen zur nichtlinearen Signalverarbeitung in Mermalsrôumen*, Master's thesis, Technische Universität Berlin, 1998.

[9] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K. R. Müller, Fisher discriminant analysis with kernels, In Y. H. Hu, J. Larsen, E. Wilson, S. Douglas, (éds), *Proc. Advances in Neural Information Processing Systems*, San Mateo : Morgan Kaufmann, pp. 41-48, 1999.

[10] S. Mika, G. Rätsch, K. R. Müller, A mathematical programming approach to the kernel Fisher algorithm, In T. K. Leen, T. G. Dietterich, V. Tresp, (éds), *Proc. Advances in Neural Information Processing Systems*, Cambridge : MIT Press, pp. 591-597, 2001.

[11] P. Moerland, *Mixture models for unsupervised and supervised learning*, Ph. D. thesis, École Polytechnique Fédérale de Lausanne, 2000.

[12] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schökopf, An introduction to kernel-based learning algorithms, *IEEE Transactions on Neural Networks*, vol. 12, n°. 2, pp. 181-201, 2001.

[13] E. Oja, A simplified neuron model as a principal component analyzer, *Journal of Mathematical Biology*, vol. 15, pp. 267-277.

[14] H. V. Poor, *An Introduction to Signal Detection and Estimation*, New York : Springer-Verlag, 1994.

[15] J. Principe, D. Xu, C. Wang, Generalized Oja's rule for linear discriminant analysis with Fisher criterion, *Proc. ICASSP'97*, vol. 4, pp. 3401-3404, Seattle, WA, 1997.

[16] C. Richard, *Méthodes à noyau et critères de contraste pour la détection à structure imposée*, Habilitation à diriger des recherches, Université de Technologie de Compiègne, 2002.

[17] C. Richard, F. Abdallah, Algorithme d'apprentissage séquentiel pour la méthode KFD. Relations avec la méthode KPCA, *Proc. Colloque GRETSI*, Paris, 2003.

[18] V. Roth and V. Steinhage, Nonlinear discriminant analysis using kernel functions, *Advances in Neural Information Processing Systems*, S.A. Solla, T.K. Leen, and K.-R. Müller, editors, vol. 12, pp. 568-574, MIT Press, 2000.

[19] S. Roweis, EM algorithm for PCA and SPCA, In M. I. Jordan, M. J. Kearns, S. A. Solla, (éds), *Proc. Advances in Neural Information Processing Systems*, Cambridge, MA : MIT Press, vol. 10, pp. 626-632, 1998.

[20] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, vol. 10, n° 5, pp. 1299-1319, 1998.

[21] B. Schölkopf, R. Herbrich, A. Smola, A generalized representer theorem, In *Proceedings of the Annual Conference on Computational Learning Theory*, pp. 416-426, 2001.

[22] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vol. 1, New York : John Wiley & Sons, 1968.

[23] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and its Applications*, vol. 16, pp. 264-280, 1971.

[24] V. Vapnik, A. Chervonenkis, *Theory of Pattern Recognition*, Moscou : Nauka, 1974.

[25] V. Vapnik, *Estimation of Dependencies based on Empirical Data*, New York : Springer-Verlag, 1982.

[26] V. Vapnik, A. Chervonenkis, The necessary and sufficient conditions for consistency of the method of empirical risk minimization (in Russian), *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification and Forecasting*, vol. 2, pp. 217-249, 1989.

[27] V. Vapnik, *The Nature of Statistical Learning Theory*, New York : Springer-Verlag, 1995.

Cédric **Richard**

Cédric Richard est né à Sarrebourg, France, en 1970. Il a obtenu un diplôme d'ingénieur en 1994, un doctorat en 1998 et une habilitation à diriger des recherches en 2002, dans la spécialité contrôle des systèmes de l'Université de Technologie de Compiègne. Il a été maître de conférence de 1999 à 2003 à l'Université de Technologie de Troyes, où il a été nommé professeur des universités depuis. Cédric Richard effectue ses recherches à l'ISTIT (FRE CNRS 2732), au sein de l'équipe M2S. Celles-ci concernent l'analyse temps-fréquence, les théories statistiques de l'estimation et de la décision, et la reconnaissance des formes.

Régis **Lengellé**

Régis Lengellé a obtenu le diplôme de docteur ingénieur de l'Université de Technologie de Compiègne, spécialité automatique et traitement du signal, en 1983 et l'habilitation à diriger des recherches de l'Université Henri Poincaré, Nancy I en 1994. Professeur à l'Université de Technologie de Troyes depuis 1994, ses activités de recherche sont relatives à la détection d'événements dans les signaux et systèmes, avec applications à la sûreté de fonctionnement.

Fahed **Abdallah**

Fahed Abdallah est né au Liban en 1976. Il a obtenu le diplôme d'ingénieur et le diplôme d'études approfondies de la faculté de génie de l'université libanaise en 1999 et 2000, et a soutenu sa thèse de doctorat en optimisation et sûreté des systèmes de l'Université de Technologie de Troyes en 2004. Il est actuellement attaché temporaire d'enseignement et de recherche dans cet établissement. Les travaux de recherche de Fahed Abdallah concernent la théorie de l'apprentissage et les méthodes à noyau.