VNIVERSITAT DE VALÈNCIA

FACULTAT DE CIÈNCIES MATEMÀTIQUES

Departament d'Estadística i Investigació Operativa

Análisis Discriminante Discreto

Mediante

Suavización de las Correspondencias múltiples

Tesis doctoral en C. Matemáticas

José Vicente Pruñonosa

(Texto parcial en castellano. Para completarlo y consultar fórmulas, demostraciones, gráficos y referencias hay que revisar el original en catalán en http://perso.wanadoo.es/jvicent/TEXTOS/Tesi.pdf . Esta traducción está especialmente dedicada a los compañeros de la Facultad de Ciencias de la UNAN-León de Nicaragua)

Índice completo (*Se señalan con * los apartados traducidos*)

4 Análisis Discriminante Discreto por el método ADDSUC (* parcial)

4.1 El análisis discriminante como correlación canónica
4.1.1 Expresión del análisis discriminante lineal (LDA) como correlación canónica simple
4.1.2 El triplete de la LDA con ponderación de individuos
4.1.3 Correlación canónica simple versus Correlación canónica generalizada
4.2 Las propuestas previas para el análisis discriminante de correspondencias múltiples
4.2.1 Las correspondencias múltiples no simétricas de Benzècri-Palumbo
4.2.2 El análisis discriminante de correspondencias de Chessel-Thioulouse
4.2.3 El análisis discriminante sobre variables cualitativas de Saporta
4.3 La propuesta ADDSUC
4.3.1 Resumen de conceptos previos
4.3.2 El Planteamiento de la propuesta (*)
4.3.3 La fundamentación matemática: la generalización del teorema de Lancaster (* parcial)
4.3.4 El algoritmo ADDSUC (*)
4.3.5 La convergencia del algoritmo ADDSUC

5 Resultados numéricos (*)

5.1 El organigrama del ADDSUC
5.2 Comparación con los métodos de estructura pareciendo
5.2.1 Selección de los conjuntos de datos por hacer las simulaciones de prueba
5.2.2 Selección de los métodos de estructura parecida para comparar
5.2.3 Resultados comparativos de las simulaciones
5.3 Comparación con la logística-redes neuronales
5.4 Comparación con datos reales
5.4.1 Los datos del estudio de mercadotecnia
5.4.2 Los datos del proyecto AFIPE
5.5 Comentarios de los resultados
5.6 Aspectos computacionales

Conclusiones y líneas de investigación (*)

A Conclusiones
B Sugerencias y posibilidades de mejora

Apéndices (*)

A Descripción de las categorías de los datos de mercadotecnia
B Descripción de las categorías de los datos de AFIPE

Bibliografía

Introducción

La motivación para realizar el presente estudio proviene del análisis epidemiológico de los factores influyentes en el patrón de evolución de las enfermedades, en el cual se pretende determinar que variables y en que grado influyen en los cambios, tanto favorables como desfavorables, que puede tener una persona en los niveles de salud cuando recibe un tratamiento determinado.

El hecho de que estos factores son, en gran parte, variables categóricas dificulta considerablemente la aplicación de las técnicas estadísticas específicas incluidas dentro el ámbito del conocido como análisis discriminante.

Como es sabido, este análisis, en este contexto, nos debe permitir asignar a una persona el patrón de evolución más probable de su enfermedad en función de los datos sociosanitarios disponibles, tomando como referencia un conjunto de personas de evolución conocida (datos de aprendizaje).

La dificultad matemática proviene de que la simplificación que introduce la suposición de continuidad, muy estudiada y con resultados que pueden considerarse satisfactorios, no es aplicable a la mayoría de las variables disponibles, y se hace necesario adaptar el método sin forzar la naturaleza de éstas.

Hay que ahondar, por lo tanto, a pesar de que pueda parecer una reflexión demasiado filosófica, aunque sea brevemente, sobre los conceptos de categórico y continuo para orientar adecuadamente esta adaptación.

Si consideramos el nivel perceptivo como la base de la aproximación continua, podemos enfocar esta como correspondiendo a un pequeño, pero muy significativo, intervalo sensorial, de manera que por debajo de él la realidad la podemos imaginar discreta y por encima la volvemos a percibir categórica como identificación de objetos diferenciados.

Desde este punto de vista podemos considerar que muchos fenómenos discretos, especialmente los de natura biológica, son el resultado de un proceso de acumulación-umbralización a partir de variables subyacentes continuas.

En el contexto epidemiológico mencionado se puede suponer que una combinación de factores continuos subyacentes determina la aparición de un determinado patrón de evolución, y que a medida que nos alejamos de esta combinación, la probabilidad de que se presente este patrón disminuye de manera que al acercarse a la combinación que determina otro patrón, la probabilidad de este último llega a ser la dominante.

La traducción matemática de esta idea consiste en suponer que las variables categóricas proceden de la discretización de subyacentes continuas, que siguen un modelo probabilístico conocido como mixtura de normales.Este es el punto de partida del método que se presenta en este trabajo, ya que de esta manera podemos considerar, como es habitual en la literatura, que los factores significativos en la determinación del patrón afectan a la media de las variables subyacentes mientras que los no significativos determinan una dispersión gaussiana alrededor de los valores centrales.

El esfuerzo se centrará, como consecuencia, en reencontrar lo más cuidadosamente posible la distribución probabilística continua subyacente, y posteriormente aplicar una metodología de discriminación con variables continuas.

Para lograr este objetivo "reconstructor", tendremos, a su vez, dos fases: En la primera, y mediante una análisis de correspondencias múltiples convenientemente adaptado al objetivo discriminante, buscaremos cuantificaciones que aproximen las medias de las celdas resultantes de la discretización. En la segunda, empleando un procedimiento de suavización, completaremos la reproducción de la distribución subyacente aplicando una dispersión alrededor de estas medias.

El capítulo 1 analizará las definiciones básicas del análisis discriminante y hará una revisión de los métodos existentes con el objetivo mencionado.

El segundo y el tercer capítulos se centrarán en hacer lo equivalente con el análisis de correspondencias y los métodos de suavización (fundamentalmente Kernel y EM) como elementos básicos a combinar, para lograr la mencionada reconstrucción.

En el capítulo 4 se hará la propuesta metodológica y se demostrará el resultado que le da fundamento matemático. En el último capítulo, el 5, se discutirán los resultados con datos simulados y reales, comparando con otros métodos de frecuente utilización.

Finalmente, con posterioridad a las conclusiones, se harán sugerencias, tanto para la posible continuación de la búsqueda teórica como para su aplicación práctica.

4.3.2 El Planteamiento de la propuesta

Las dificultades que los métodos comentados en la sección anterior presentan provienen, en esencia, de la necesidad de equilibrar dos objetivos que han estado tratados por separado: la discriminación propiamente dicha y la reconstrucción de las variables subyacentes contínuas mediante correspondencias.

Nuestra propuesta consiste en hacer un análisis de correspondencias múltiples "ponderado - iterado" en el que las variables serán pesadas por su valor discriminante (máxima separación de los centroides) y, posteriormente, recalcular los centroides teniendo en cuenta lo resultados, iterando hasta que obtengamos la convergencia.

Partimos de unos centroides obtenidos simplemente de los originales valores categóricos.
La matriz a diagonalizar cada vez es $D^{-1}X'XA$ donde A representa los pesos atribuidos a cada variable colocados forma diagonal y con el mismo coeficiente para todas las categorías de la misma variable. Esta matriz corresponde, introduciendo una ponderación, al enfoque del punto C-vi del apartado anterior.

Por otra parte, si utilizamos el esquema de cuantificación recíproca, $\psi = XA\zeta$ nos da las cuantificaciones de los individuos como suma ponderada de las que le otorgan cada una de las variables, y $\zeta = D^{-1}X'\psi$ representa las cuantificaciones variable a variable obtenidas por proyección de las de los individuos sobre los espacios determinados por cada una.

La matriz diagonal A que, como se ha dicho, recoge los pesos atribuidos a cada variable se obtendrá a partir de los coeficientes de la combinación lineal de estas que maximice la varianza de los centroides (B) con la varianza total (T) igual a la unidad, es decir mediando un LDA sobre las variables cuantificadas por el paso anterior.

Naturalmente, en cada paso al variar A variará ζ y el proceso será iterativo. Habrá que, por lo tanto asegurarse la convergencia, lo que haremos en el apartado 4.3.5 .La idea básica es que, mediando este proceso, lograremos equilibrar el efecto reconstructor de la Normal evidenciado por Lancaster para el correspondencias simples con la finalidad de la discriminación en un contexto multidimensional.

En definitiva, al proyectar sobre un espacio donde las variables son pesadas por su poder discriminante es como realmente logramos la normalización al reconstruir el eje principal de las normales originales o, mejor dicho, al reconstruir lo más cuidadosamente posible la disposición relativa de los centroides. Ambos objetivos van por tanto unidos, como se evidenciará en uso del resultado demostrado a continuación, donde la maximización de la correlación que Lancaster empleó con las correlaciones canónicas simples se transfiere a una maximización de la varianza sobre el eje principal de la matriz de correlaciones, R del caso multidimensional.

4.3.3 La fundamentación matemática: la generalización del teorema de Lancaster

Como mencionábamos en la sección 4.1 el objetivo reconstructor de una multinormal tendría que partir de una generalización multivariable del teorema de Lancaster, pero ésta no es fácil debido a que el que era un único valor de correlación ρ es substituido, ahora, por una matriz R y por lo tanto debemos seleccionar cuál aspecto de ésta se debe maximizar.

La respuesta viene dada por el mismo objetivo discriminante que perseguimos: lo haremos en la dirección principal discriminante, es decir en aquella que maximice la varianza entre los centroides de clase.

El fundamento de este proceso se establece mediante el siguiente teorema:

Teorema 4.1 Generalización multidimensional del teorema de Lancaster

Sean Zi, i = 1, . . . , p variables aleatorias normales tipificadas con R como matriz de correlaciones y sean Xi, i = 1, . . . , p transformadas de las Zi respectivamente y también tipificadas.

Si v es el vector propio correspondiente al mayor valor propio de R entonces:

$$Var(w'X) \leq Var(v'Z) \quad \forall w \in Rp \ \ con\ w'w = 1$$

El significado de este teorema en nuestro caso es claro: cualquier combinación lineal de las variables normales discretizadas que maximice la varianza en la dirección del primer vector propio de R, va en la dirección de la reconstrucción de el eje principal de la normal subyacente, dado que esta distribución es la que la tiene máxima entre todas las derivadas de una transformación suya variable a variable.

4.3.4 El algoritmo ADDSUC

Si ahora nos planteamos buscar una cuantificación que maximice la varianza en la dirección del primer vector propio *a* de la matriz de covarianzas de los centroides de clase, lograremos el equilibrio perseguido, ya que el componente de la varianza debido al efecto de clase (la proveniente de B) quedará maximizada, y la debida a la parte común ($\sum$) será reconstruida como residual por el análisis de correspondencias múltiple en la dirección de la normalidad.

Resulta claro, del Teorema 4.1, que esta reconstrucción será más "eficiente" en la medida en que *a* se acerque a la dirección del primer vector propio de la matriz de correlaciones R (normalización de $\sum$).

Por otra parte, al desconocer la situación continua subyacente no podemos calcular con exactitud *a*, por lo que nos introduciremos en un proceso iterativo que, partiendo de las cuantificaciones habituales (ζ_0) nos calcule una aproximación de *a*, la cual iremos mejorando posteriormente de manera iterativa. Si ζ_j ,R_j , a_j son, respectivamente, las cuantificaciones, la matriz de correlaciones y la aproximación al vector *a*, logrados en una determinada iteración j, el proceso iterativo sería:

$\zeta_0 \rightarrow R_0 \rightarrow a_0 \rightarrow \zeta_1 \rightarrow R_1 \rightarrow a_1 \cdots\cdots$

La convergencia de este procedimiento se garantiza en 4.3.5 y nos asegura encontrar finalmente unas cuantificaciones ζ y un vector *a*, de manera que éste sea el primer vector propio de los centroides determinados por aquellas, habiendo logrado reconstruir, con dimensión 1, lo más cuidadosamente posible, la situación subyacente de partida.

El resto de las dimensiones las obtendremos por el habitual proceso canónico que se detendrá cuando no se obtenga ninguna mejora significativa del error real final.

Finalmente, si hacemos intervenir la suavización propuesta en la sección 3.6 el algoritmo del método propuesto se puede esquematizar como:

1ª fase: Realizar un análisis de correspondencias múltiples ponderado-iterado que, de forma canónica, obtendrá cuantificaciones con ponderaciones para las variables que corresponderán al eje principal de la descomposición de la varianza entre los centroides de clase (B). Este análisis será descrito en detalle al apartado siguiente.

2ª fase: Aplicar una suavización EM (sección 3.6) sobre las cuantificaciones aportadas por la fase 1 para reconstruir lo más cuidadosamente las normales subyacentes a cada clase y, posteriormente, proceder a una discriminación LDA-canónica.

Capítulo 5: Resultados numéricos

En este capítulo probaremos el método ADDSUC comparándolo primero, por simulación, con los de estructura parecida (sección 5.2) y después con el método más utilizado actualmente para hacer análisis discriminante discreto: el logístico-redes neuronales (sección 5.3) y finalmente con los dos tipos al mismo tiempo empleando datos reales (sección 5.4)

5.1 El organigrama del ADDSUC

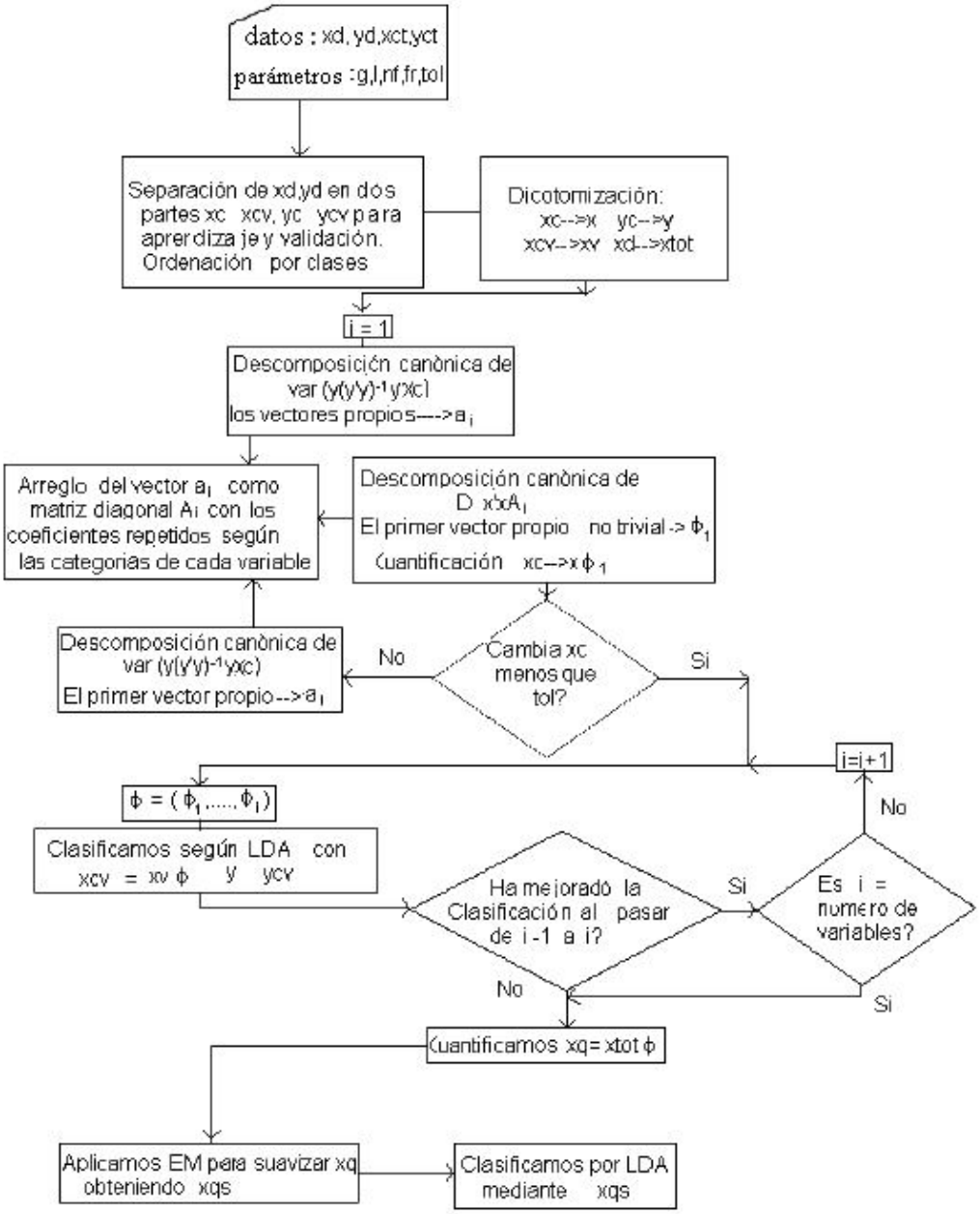
Presentaremos, en primer lugar en la figura de la página siguiente, el organigrama del método ADDSUC que hemos utilizado como base para su programación.

En cuanto a la interpretación de los símbolos que en él aparecen, comentaremos que la entrada, *xd*, debe ser una matriz con tantas filas como individuos y columnas correspondientes a las variables clasificadoras; coincide exactamente con la matriz X definida en la sección 1.1. El vector *yd* contendrá la clase correspondiente a cada individuo y se corresponde exactamente con la Y de la misma sección. Ambas se refieren a las datos de aprendizaje mientras que *xct* y *yct* son las equivalentes para las muestras de verificación.

En cuanto a los parámetros: *g* representa el número de clases, *l* es un vector con el número de categorías de cada variable y *nf* es el número de esos a considerar si es que este se quiere fijar. En caso de que sea *0* el programa lo estimará por validación cruzada mínimo-cuadrática.

Por otra parte, *fr* es la fracción de los datos de aprendizaje que se empleará para esta validación cruzada(redondeando la frecuencia resultante al número entero más próximo). Si *nf* = 0 no se utilizará validación cruzada y se hará *fr* = 1, tomando todos los datos de aprendizaje para la estimación.

Finalmente, *tol* representa el umbral de tolerancia para la convergencia del algoritmo y se utiliza de la manera que queda reflejada en el diagrama. Por defecto se toma *tol* = 0.0001. Los parámetros *g* y *l* son también calculados por el programa, pero se considera más conveniente que sean dados por el usuario con el fin de detectar posibles errores informando mediante un mensaje de esta circunstancia para dar la oportunidad de corregirla.


datos : xd, yd,xct,yct
parámetros :g,l,nf,fr,tol
Separación de xd,yd en dos partes xc xcv, yc ycv para aprendizaje y validación. Ordenación por clases
Dicotomización:
xc-->x yc-->y
xcv-->xv xd-->xtot
i = 1
Descomposición canònica de var (y(y'y)-1y'xc)
los vectores propios--->ai
Arreglo del vector ai como matriz diagonal Ai con los coeficientes repetidos según las categorias de cada variable
Descomposición canònica de D x'xAi
El primer vector propio no trivial -> ϕ1
Cuantificación xc-->x ϕ1
Descomposición canònica de var (y(y'y)-1y'xc)
El primer vector propio-->a1
No
Cambia xc menos que tol?
Si
i=i+1
No
ϕ = (ϕ1,....., ϕi)
Clasificamos según LDA con xcv = xv ϕ y ycv
Ha mejorado la Clasificación al pasar de i -1 a i?
Si
Es i = numero de variables?
No
Si
Cuantificamos xq= xtot ϕ
Aplicamos EM para suavizar xq obteniendo xqs
Clasificamos por LDA mediante xqs

5.2 Comparación con los métodos de estructura parecida

Comenzaremos las pruebas de ADDSUC con esta sección, la cual está dedicada ala comparación con los métodos que, como él, utilizan el análisis de correspondencias y/o la suavización mediante el algoritmo EM.

5.2.1 Selección de los conjuntos de datos para hacer las simulaciones de prueba del método.

Para elegir un conjunto de datos simulados que, sin pretensiones de exhaustividad, pueda ser representante de una gama bastante amplia de situaciones, debemos seleccionar unos criterios que nos permitan valorar la "peculiaridad" clasificatoria de cada uno de ellos.

Después de hacer una revisión de la literatura sobre medidas previas de separabilidad de clases, donde destacan las propuestas recogidas en Hand (1981) [107], que tienen el problema de limitarse a la separabilidad entre dos clases (y que se reducirán a la distancia de Mahalanobis en caso de que nos ocupa), llegamos a la conclusión de que los parámetros más convenientes son:

1.- Número de variables y de categorías como medida de la complejidad inicial tratada a la sección 1.2.3

2.- El error óptimo continuo ec, descrito en el apartado 1.2.2.1, que nos refleja el grado de solapamiento entre clases del que partimos.

3.- El porcentaje de varianza entre clases absorbido por el primer eje, el cual nos refleja el grado aproximado de la dimensionalidad de los centroides (cerca del 100% indica un grado aproximado de 1, mientras que por debajo del 90% podemos considerar este grado mayor de 1). Otras medidas de dimensionalidad pueden aplicarse pero elegimos ésta por su relación con el planteamiento canónico que se sigue a lo largo de este estudio.

Teniendo en cuenta todo eso hemos seleccionado los siguientes conjuntos de datos todos con 3 clases de pesos 0.2, 0.3, 0.5:

Conjunto 1 Número de Variables: 3, Número de Categorías: 9

Medias M1	Varianzas V1	Cortes para la discretización T1	ec	Peso 1r eje
$\begin{bmatrix} (-0.5,-0.3,0.1) \\ (0.1,0.2,0.2) \\ (0.7,0.6,0.7) \end{bmatrix}$	$\begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & 0.5 \\ 0.3 & 0.5 & 1 \end{pmatrix}$	$\begin{bmatrix} (-0.3) \\ (-0.6,0.4) \\ (-0.3,0.4,0.8) \end{bmatrix}$	0.48	99%

Conjunto 2 Número de Variables: 3, Número de Categorías: 8

Medias M2	Varianzas V2	Cortes para la discretización T2	ec	Peso 1r eje
$\begin{bmatrix} (-0.5,-0.5,-0.25) \\ (0.5,0.5,0) \\ (1,-0.5,0.25) \end{bmatrix}$	V_1	$\begin{bmatrix} (0,0.75) \\ (0) \\ (-0.125,0.125) \end{bmatrix}$	0.27	75%

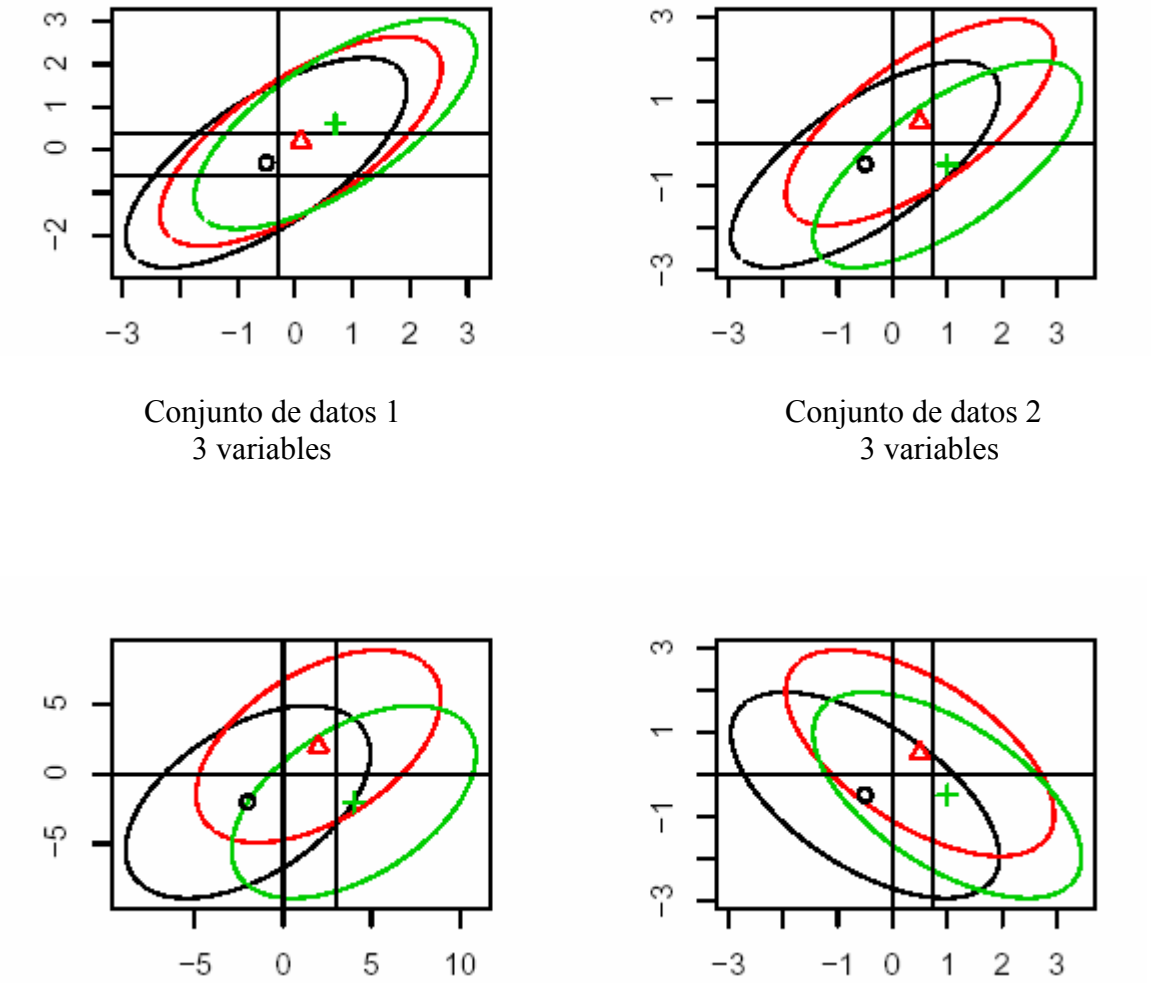
Conjunto 3 Número de Variables: 2, Número de Categorías: 5

Medias M3	Varianzas V3	Cortes para la discretización T3	ec	Peso 1r eje
$\begin{bmatrix} (-2,2) \\ (2,2) \\ (4,2) \end{bmatrix}$	$\begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix}$	$\begin{bmatrix} (0,3) \\ (0) \end{bmatrix}$	0.23	80%

Conjunto 4 Número de Variables: 3, Número de Categorías: 8

Medias M4	Varianzas V4	Cortes para la discretización T4	ec	Peso 1r eje
M_2	$\begin{pmatrix} 1 & -0.6 & 0.3 \\ -0.6 & 1 & 0.5 \\ 0.3 & 0.5 & 1 \end{pmatrix}$	T_2	0.16	68%

Estos conjuntos de datos pueden representarse gráficamente utilizando los elipsoides del 95% correspondientes a la distancia de Mahalanobis. En la figura 5.2 podemos observar las representaciones cartesianas de sus dos primeras variables, con los correspondientes centroides y rectas horizontales y verticales de corte. Conviene aclarar que los conjuntos de datos 2 y 3, que en el gráfico parecen similares, no lo son, dado que el primero dispone de tres variables y el segundo sólo de dos.



Conjunto de datos 1
3 variables

Conjunto de datos 2
3 variables

Conjunto de datos 3
2 variables

Conjunto de datos 4
3 variables

Figura 5.2: Variables 1 y 2 de los conjuntos de datos simulados

5.2.2 Selección de los métodos de estructura parecida para comparar

Una vez seleccionados los conjuntos de datos, hay que decidir cuáles serán los métodos de estructura parecida a comparar. Teniendo en cuenta la revisión del capítulo 1, queda claro que, en nuestro caso, los métodos de referencia deben ser el LDA-Canónico y el MDA.

Ya dentro del ámbito de las correspondencias utilizaremos las variantes mencionadas en la sección 4.2 (pàg. 78), excepto la de Chessel-Thioulouse dado que, en el caso multivariable, nos conduce, como hemos probado, al LDA-Canónico.

Añadiremos cuatro posibilidades más (aparte de ADDSUC), que describiremos brevemente a continuación, con otras exploraciones que hicimos por utilizar el análisis de correspondencias con objetivo discriminante. Por lo tanto el cuadro de métodos a comparar será:

1. LDA-Canónico.

2. MDA.

3. Basados en correspondencias:

   (a) Benzècri-Palumbo: Consiste en la descomposición de la matriz $(PY\ X)'PY\ XD^{-1}$. Es el también llamado análisis interclases. Complementariamente se puede hacer el intraclases que descompondría canónicamente la matriz $(X - PY\ X)'(X - PY\ X)D^{-1}$

   (b) Saporta-Volle: Un análisis de correspondencias completo con la matriz $X'XD^{-1}$ seguido de una correlación canónica simple con Y .

   (c) Método ADDSUC.

   (d) Otras Exploraciones propias (OEP)basadas en las correspondencias:
      i. Descomposición de $(X'X - (PY\ X)'PY\ X)D^{-1}$.
      ii. Correspondencias incluyendo Y que después se suprime para el cálculo de las cuantificaciones.
      iii. Descomposición de $x'iXiD^{-1}$ $i = 1, \ldots, g$, seguida de una ponderación por las frecuencias relativas a cada categoría dentro de cada clase.
      iv. Descomposición de $(X - PY\ X)'(X - PY\ X)\Omega^{-1}$.

La pauta de comparación se establecerá en función del error e_{fr} dado que, como se explicó en la sección 1.2.2.4 representa un error que, a diferencia del aparente, nos permite valorar si se ha alcanzado un adecuado equilibrio con la complejidad .Utilizaremos muestras de 500 datos divididos aleatoriamente en dos mitades, una para estimar y la otra para ajustar empleando la validación cruzada. También utilizaremos muestras de verificación de la misma medida.

Necesitaremos, también, dos medidas de las repeticiones necesarias para estimar el error: una de repeticiones de la extracción originaria de la muestra de aprendizaje y otra de repeticiones de las muestras de verificación una vez fija la muestra de aprendizaje (ver sección 1.2.2.3). Para ambos casos utilizaremos el valor 50.

5.2.3 Resultados comparativos de las simulaciones

Los resultados se pueden resumir con el siguiente cuadro de errores finales reales, e_{fr} :

Conjunto de datos	LDA canónico	MDA	ADDSUC	Benzècri	Saporta	Mejor OEP
1	0.492	0.464	**0.461**	0.480	0.479	0.467(iv)
2	0.473	0.430	**0.369**	0.478	0.406	0.443(iv)
3	0.449	0.421	**0.312**	0.503	0.337	0.428(iii)
4	0.456	0.483	**0.387**	0.450	0.402	**0.387(ii)**

dónde la columna de Mejor OEP contiene los errores mínimos para cada conjunto de datos de entre todos los métodos reseñados al ítem 3.(d) del apartado anterior como otras exploraciones propias de discriminación mediante análisis de correspondencias. Entre paréntesis figura la identificación del método que ha obtenido este mínimo.

Hay que señalar que ADDSUC supera a todos los métodos en todos los conjuntos de datos, obteniendo sobre el siguiente método (Saporta) una ventaja relativa del 7% y siendo igualado sólo por otro método de exploración propia (ii) en el cuarto conjunto de datos. Dado que este último obtiene frente a ADDSUC unos errores superiores al 15% en media no parece conveniente retenerlo como método comparable. En cambio el tercero (después de ADDSUC y Saporta) el MDA si que lo tendremos en consideración para las comparaciones con datos reales, dado que representa la aplicación directa del algoritmo EM (la segunda parte de ADDSUC) y nos da una idea muy clara de como influye la primera parte (correspondencias) sobre el resultado final. Es muy importante indicar que para evitar que el efecto del orden implícito restase generalidad a los resultados se han permutado las categorías 1 y 2 de cada variable. Es decir si los puntos de corte son , por ejemplo, -0.6 y 0.4 (T1 , 2ª variable) el valor 1 corresponderá al intervalo (-0.6, 0.4], el valor 2 al intervalo (-1,-0.6] y el valor 3 al intervalo (0.4,1).

En estas condiciones más difíciles donde se ha deshecho el orden subyacente (observemos como aumenta significativamente e_{fr} con relación a ec en los conjuntos de datos 2, 3 y 4), es donde los métodos basados en las correspondencias tienen la oportunidad de demostrar sus propiedades reconstructoras.

No hemos añadido los resultados cuando no hay permutación ya que en este caso MDA, ADDSUC y Saporta (los principales métodos a comparar) dan errores casi equivalentes; tampoco hemos reseñado los de la combinación del MDA con las otras posibilidades de aplicación de las correspondencias, ya que no aportan ninguna mejora.

5.3 La comparación con el método híbrido logística-redes neuronales

Una vez asegurado que el método ADDSUC supera con claridad tanto a los métodos que nada más utilizan Correspondencias (del que destacamos el propuesto por Saporta), como al que nada más utiliza el EM (el MDA) y a cualquier combinación entre ellos, debemos pasar a comparar con los métodos que actualmente destacan como los más eficientes para el análisis discriminante discreto.

Entre ellos sobresale por sus buenos resultados la adaptación de la logística empleando la idea de las redes neuronales (LRN) (sección 1.3.3).Se trata, también, de un método mixto que utiliza la filosofía del de redes neuronales de la manera que ADDSUC utiliza la suavización mediante EM empleando para la cuantificación previa una logística en lugar de un correspondencias. Eso le da una gran potencia y versatilidad y lo hace el recomendado en estos momentos por la mayoría de los autores que no forman parte de una escuela específica.

Comenzaremos por hacer la comparación con los conjuntos de datos referidos a la sección anterior obteniendo una sorprendente igualdad (con solo diferencias a nivel de la cuarta cifra decimal lo que nos la hace significativas) en todos los casos.

Esta situación que podríamos calificar "de empate" aunque bastante estimulante, dado que ADDSUC como veremos al apartado de sugerencias, es aún un método acabado de nacer y con muchas posibilidades de mejora y ajuste, nos dejaba en el punto de intentar averiguar donde podría haber diferencias que nos guiasen en las posteriores búsquedas.

Frente a la alternativa de "perturbar" los conjuntos de datos buscando pequeñas diferencias, las cuales, tratándose de simulaciones, no serían demasiado relevantes y se podrían atribuir a peculiaridades específicas, optamos por emplear para la comparación, por una parte los datos reales que analizaremos a la siguiente sección y, por otra, los datos de referencia que son la pauta de comparación para todos los nuevos métodos de discriminación: los aportados por Fisher bajo el nombre de IRIS.

Este conjunto de datos, muy estudiado, se compone de 150 individuos divididos en tres clases (de 50 miembros cada una) y cuatro variables continuas: longitud y anchura de los sépalos y longitud y anchura de los pétalos. Para aplicar un análisis discriminante discreto debemos proceder a la discretización de estas variables. Con este objetivo haremos primero los correspondientes histogramas.

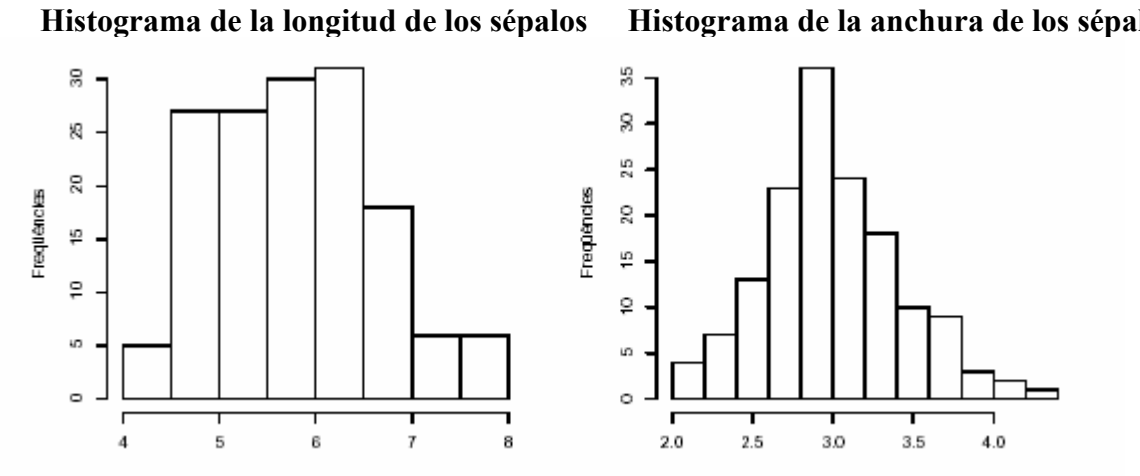


Figura 5.3: Histogramas de los datos de IRIS (sépalos)

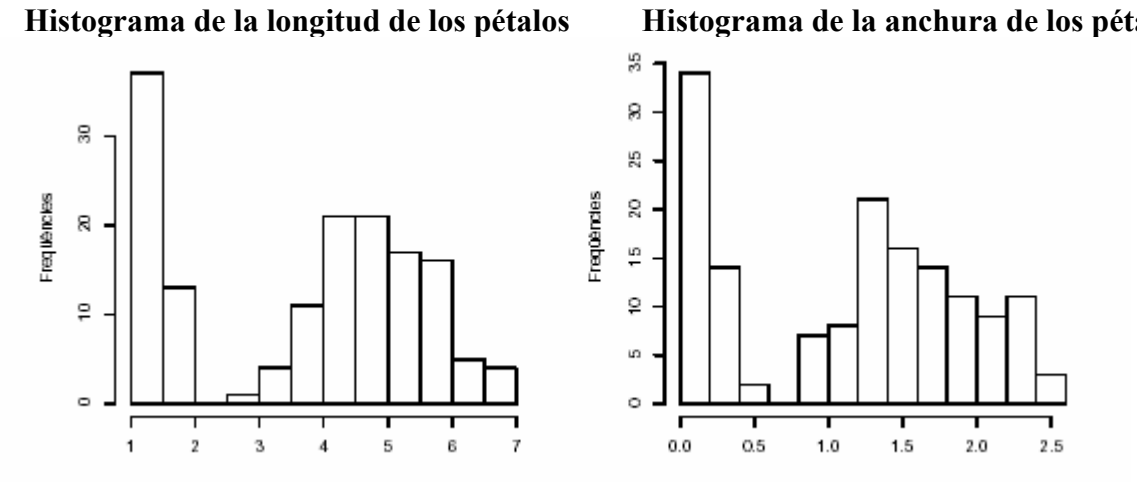


Figura 5.4: Histogramas de los datos de IRIS (pétalos)

Observando los histogramas de las figuras 5.3 y 5.4, resulta claro que las dos variables relativas a los pétalos tienen puntos de corte evidentes a los centros de los correspondientes valles: 2.3 para la longitud y 0.7 para la anchura. En cambio, las variables correspondientes a los sépalos no presentan ningún corte claro dado que debemos considerar las variables en sus marginales sin tener en cuenta por nada las especies (variable clasificadora).

Por eso si queremos considerarlas como variables discretas, hecho que implica tener por lo menos dos categorías, lo más natural parece cortar por la mediana en ambos casos. Procediendo, pues, de esta manera y tomando 75 datos al azar con 25 de cada especie tal y como hizo Ripley, 2002 [218] como muestra de aprendizaje y dejando el resto como datos de verificación, hicimos 1000 repeticiones y comparamos los resultados.

En 53 de las repeticiones fue mejor el ADDSUC y en el resto se produjo un empate exacto (mismo número de bien clasificados) con un error final medio para el ADDSUC de 0.217 y para la Logística-redes neuronales de 0.223 por el que podemos decir que, en estas condiciones, el ADDSUC supera "uniformemente" a la Logística-Redes neuronales.

5.4 Comparación con datos reales

Finalmente haremos una comparación de los métodos que mejor resultado han dado en las simulaciones: MDA, ADDSUC, Saporta y logística-redes neuronales (LRN) con dos conjuntos de datos reales.

5.4.1 Los datos del estudio de mercadotecnia

Este conjunto de datos lo obtuvimos mediante una búsqueda por Internet de valores empleados como prueba de los métodos de análisis discriminante. Se trata de datos provenientes de un estudio de mercadotecnia (ver [126]) sobre 9409 residentes en San Francisco (California) y nos sirve por estudiar el método en situaciones reales de una gran cantidad de variables e individuos mucho mayor que la probada en las simulaciones. Las variables y sus categorías son descritas en el apéndice A.

Después de la supresión de los casos con algún dato faltante se realizó el análisis discriminante con 2000 datos de aprendizaje y 4000 de verificación obteniendo los siguientes errores finales reales:

$$\textbf{\textit{MDA}} = 0.392 \quad \textbf{\textit{ADDSUC}} = 0.363 \quad \textbf{\textit{Saporta}} = 0.404 \quad \textbf{\textit{LRN}} = 0.385$$

En términos absolutos eso significa que ADDSUC clasifica bien 2548 casos de los 4000 mientras que el método que le sigue (logística-redes neuronales) lo hace con sólo 2460.

5.4.2 Los datos del proyecto AFIPE

El segundo conjunto de datos corresponde al tipo de los que han servido por motivar este trabajo. Se trata de un pilotaje del proyecto AFIPE (Análisis de los Factores Influyentes en el Patrón de Evolución de las enfermedades) que formó parte del SISNICA (Sistema de Información Sanitaria de Nicaragua) desarrollado durante el período 1990-1994 (ver [175]).

Se trata de 1144 personas de las que las variables y sus categorías están descritas en el apéndice B. Aplicando el análisis discriminante con 550 datos de aprendizaje y 594 como datos de verificación se obtuvieron los siguientes errores finales reales:

$$\textbf{\textit{MDA}} = 0.387 \quad \textbf{\textit{ADDSUC}} = 0.318 \quad \textbf{\textit{Saporta}} = 0.461 \quad \textbf{\textit{LRN}} = 0.388$$

Lo que, en términos absolutos, significa que ADDSUC clasifica bien 405 de los 594 mientras que el método que le sigue (el MDA) lo hace solo en 364.

5.5 Comentarios de los resultados

- En primer lugar debemos destacar que el método propuesto: el ADDSUC (Análisis discriminante discreto mediante suavización de las correspondencias múltiples) parece representar una ventaja significativa sobre cualquiera método que se base en suavizaciones mediante EM, correspondencias o una combinación de los dos procedimientos, cuando los datos provienen de una multinormal discretitzada.

- Esta ventaja se hace especialmente relevante cuando se ha aplicado una permutación al orden natural de las discretizaciones, situación que puede considerarse bastante frecuente en la práctica cuando las variables llegan al investigador desprovistas de cualquiera indicación ordinal, lo que ocurre en la gran mayoría de los casos de las búsquedas sanitarias a las que se hacía referencia en la introducción como motivación de este trabajo.

- Naturalmente, si los datos no pueden considerarse provenientes de un modelo como el que ahí se ha supuesto (sección 1.1) no podemos asegurar la permanencia de esta ventaja, pero las pruebas con datos reales parecen confirmar que las suposiciones son de un alcance bastante amplio en la práctica.

- Por otra parte si comparamos el ADDSUC con el método considerado más avanzado actualmente para realizar el análisis discriminante discreto: el perfeccionamiento de la logística basado en las redes neuronales, observamos una ligera ventaja del ADDSUC si las discretizaciones se han realizado en los puntos de cortes naturales: los valles de las distribuciones marginales (sección 5.3).

- Finalmente, las pruebas realizadas con datos reales de una cierta complejidad, una procedente de datos analizados por J.Friedman que está al alcance de los investigadores mediante Internet [126] y otra procedente de la experiencia propia con datos epidemiológicos, invitan a la continuación de la búsqueda en la línea iniciada, dado que el método propuesto logra los mejores resultados con una diferencia significativa.

5.6 Aspectos computacionales

Los programas tanto por el uso convencional del ADDSUC como para su prueba empleando simulaciones, han sido realizados mediando R versión 1.7.1, ya que este lenguaje se ha convertido en el medio habitual de programación GNU en estadística.

En R disponíamos del paquete *mda* desarrollado por Hastie (2002) [119], de la librería MASS diseñada por Ripley (2002) [218] con la subrutina *mvrnorm* que hemos empleado para la simulación de multinormals y del paquete *nnet*, que dispone de las rutinas asociadas a la metodología de redes neuronales, de donde hemos extraído la renombrada *multinom*, que realiza la logística-redes neuronales.

La prueba de los otros métodos inspirados en correspondencias, tanto los de Bènzecri y Saporta como los que hemos llamado de exploración propia, han sido programados directamente en R, ya que este lenguaje nos provee de una potencia de programación y de una simplicidad de uso considerable.

Se debe comentar, también, que el tiempo de procesamiento no es ningún inconveniente, ya que en todas las pruebas realizadas, la convergencia del algoritmo ADDSUC no ha requerido más de 10 iteraciones.

Todos los programas y datos utilizados en este capítulo se encuentran, comprimidos, en el enlace Programes R de la página web http://perso.wanadoo.es/jvicent/Cient%EDficCatal%E0.htm

Conclusiones y líneas de investigación

Resumiremos aquí, brevemente, las conclusiones del estudio y las sugerencias para ampliar la investigación.

A Conclusiones

En este trabajo hemos procedido a realizar una revisión sintetizadora y unificadora de la teoría y de los procedimientos tanto del análisis discriminante como de los métodos de correspondencias y de suavización.

Posteriormente, se ha desarrollado y fundamentado una nueva metodología para realizar el análisis discriminante discreto estructurado en dos fases: en la primera se procede a cuantificar empleando un análisis de correspondencias múltiples ponderado-iterado y en la segunda se lleva a cabo una suavización mediante el algoritmo EM.

La prueba del método con datos simulados utilizando un modelo de Normales subyacentes con medias diferente por clase y varianza común, puede considerarse positiva, ya que sus resultados superan los de los otros procedimientos con los que se ha comparado (secciones 5.2 y 5.3).

En nuestra opinión estos esperanzadores resultados se deben a la solidez del resultado matemático probado a la sección 4.3.3, el cual nos garantiza que la reconstrucción de los datos subyacentes continuos se realiza en la dirección correcta.

Si a eso se añade que la suposición de una multinormal subyacente puede considerarse el final de un amplio abanico de procesos investigadores cuando, finalmente, se logra separar la parte relevante de la que no lo es (en términos probabilísticos), no nos debe sorprender que un método, basado en estas premisas, obtenga buenos resultados prácticos, tal y como sucede a los dos ejemplos reales analizados.

Se debe tener en cuenta, también, que la cuantificación propuesta puede utilizarse no solo con objetivos clasificatorios sino también con intenciones descriptivas y comparativas.

Por todas estas razones, consideramos que la metodología desarrollada, la cual, programada en lenguaje R, se pone a la disposición de los investigadores a la página web anteriormente citada, representa una aportación a tener en cuenta dentro del campo del análisis discriminante discreto.

B Sugerencias para la mejora del método

Una posibilidad que ha estado explorada en la realización de este estudio, pero que necesita más trabajo, tanto teórico como práctico, consiste en hacer posteriormente al análisis de correspondencias un análisis canónico generalizado, agrupando todas las cuantificaciones (por ejes) de una misma variable dentro del mismo bloque. De esta manera encontraríamos, para cada variable, la combinación lineal de las aproximaciones de sus polinomios de l'Hermite que mejor se proyecten sobre la combinación global.

Este proceso sería similar a un FDA (consultar sección 1.3.4), dado que se trata de una expansión polinómica con selección posterior, y hemos probado que, en algunos casos, mejora lo resultados. Así cubriríamos también la posibilidad que la matriz de covarianzas fuera diferente por clase, ya que el QDA correspondiente sería incluido dentro la mencionada expansión polinomial.

Otra ampliación perfectamente factible del método se comentaba cuando precisábamos la situación en estudio en la sección 1.1 y consistiría en ampliar la consideración de que las distribuciones por clase son Normales incorporando la posibilidad de que puedan ser mixturas de Normales, lo que es perfectamente compatible con la utilización del EM en la segunda fase del proceso, siguiendo un esquema similar al del MDA.

También se puede tener al alcance la posibilidad de utilizar suavización con el Kernel adaptable, explicado en la sección 3.5 , lo que nos permitiría dar al método una mayor flexibilidad, al poderse emplear con un conjunto de funciones de densidad por clase más amplia que no se limitara a Normales o mixtura de Normales.

En cuanto al caso del análisis mixto (X contiene variables categóricas y continuas) se propone investigar la posibilidad de incluir al proceso iterativo las variables continuas o bien emplear estas como a covariables.

Finalmente, hay que asegurar, también, un tratamiento adecuado de los datos incompletos, adaptando los procedimientos desarrollados con este objetivo, y probar la sensibilidad del método a las cuantificaciones de partida, buscando un procedimiento rápido (permutando, por ejemplo) que nos diese la que tuviera menor error aproximado inicial.

Estos son los aspectos por donde se sugiere que debería continuar la búsqueda con perspectivas que nos parecen positivas.

ANEXOS

A) Descripción de las categorías de los datos de mercadotecnia

Las variables son:

Y = Nivel de Ingresos anuales familiares, con categorías:

1.- Menos de $20,000
2.- De $20,000 a $40,000
3.- Más de $40,000

X1 = Género, con categorías:

1.- Hombre
2.- Mujer

X2= Estado civil, con categorías:

1.- Casado
2.- Unión estable de hecho
3.- Divorciado o separado
4.- Viudo
5.- Soltero

X3 = Edad, con categorías:

1.- 14 a 17
2.- 18 a 24
3.- 25 a 34
4.- 35 a 44
5.- 45 a 54
6.- 55 a 64
7.- 65 y más

X4 = Nivel educativo, con categorías:

1.- Hasta octavo grado de primaria
2.- Grados 9 a 11 de primaria
3.- Graduado del Instituto (High school)
4.- 1 a 3 años de Universidad
5.- Graduado universitario
6.- Con estudios de postgrado

X5 = Ocupación, con categorías:

1.- Profesional/Gerente
2.- Vendedor
3.- Obrero/Conductor
4.- Clero/Trabajadores de Servicios
5.- Amo/a de casa
6.- Estudiante
7.- Militar
8.- Retirado
9.- Parado

X6 = Años de residencia en la zona, con categorías:

1.- Menos de un año
2.- De uno a tres años
3.- De cuatro a seis años
4.- De siete a diez años
5.- Más de diez años

X7 = Hay dos o más ingresos en la familia?, con categorías:

1.- No casado
2.- Si
3.- No

X8 = Número de personas en la familia, con categorías:

1.- Una
2.- Dos
3.- Tres
4.- Cuatro
5.- Cinco
6.- Seis
7.- Siete
8.- Ocho
9.- Nueve o más

X9 = Número de personas de menos de 18 años en la familia, con categorías:

1.- Ninguna
2.- Una
3.- Dos
4.- Tres
5.- Cuatro
6.- Cinco
7.- Seis
8.- Siete
9.- Ocho
10.- Nueve o más

X10 = Propiedad de la casa, con categorías:

1.- Propiedad
2.- Alquiler
3.- Con los padres o familiares

X11 = Tipo de casa, con categorías:

1.- Casa
2.- Condominio
3.- Apartamento
4.- Casa Móvil
5.- Otro

X12 = Clasificación étnica, con categorías:

1.- Indio americano
2.- Asiático
3.- Negro
4.- Indio del Este
5.- Hispánico
6.- Islas del Pacífico
7.- Blanco
8.- Otro

X13 = Lengua empleada más frecuentemente en el hogar, con categorías:

1.- Inglés
2.- Español
3.- Otro

B Descripción de las categorías de los datos de AFIPE

Y = Patrón de evolución de ERA ("Enfermedad Respiratoria aguda"), con categorías:

1.- Sanos
2.- Episodios aislados leves
3.- Episodios repetitivos y crónicos leves
4.- Episodios aislados graves
5.- Episodios repetitivos de gravedad decreciente
6.- Episodios repetitivos de gravedad creciente
7.- Episodios repetitivos graves
8.- Crisis larga y crónica grave

X1 = Frecuencia de la atención recibida, con las categorías:

1.- Nunca
2.- Aisladamente (1 o 2 veces no consecutivas en los seis períodos semanales del estudio)
3.- Repetitivamente

X2 = Tipo de tratamiento aportado por el sistema sanitario, con las categorías:

1.- Receta
2.- Tratamiento
3.- Referencia

X3 = Profesional que atiende, con las categorías:

1.- Solo el auxiliar de enfermería
2.- Médico/Médica

X4 = Perfil de incremento o disminución de la atención recibida, con las categorías:

1.- Creciente (de auxiliar a médico o de receta a tratamiento)
2.- Igual (se incluyen los casos de menos de 2 atenciones)
3.- Decreciente

X5 = Municipio, con las categorías:

1.- León: Cabecera regional
2.- El Sauce: Municipio muy extenso y disperso
3.- El Jicaral: Municipio periférico

X6 = Tipo de comunidad, con las categorías:

1.- Urbana populosa
2.- Urbana periférica
3.- Rural concentrada
4.- Rural dispersa