DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

ANÀLISI DISCIMINANT DISCRETA MITJANÇANT SUAVITZACIÓ DE LES CORRESPONDÈNCIES MÚLTIPLES

JOSÉ VICENTE PRUÑONOSA REVERTER

UNIVERSITAT DE VALENCIA
Servei de Publicacions
2003

Aquesta Tesi Doctoral va ser presentada a Valencia el día 04 de Desembre de 2003 davant un tribunal format per:

- D. Francisco Montes Suay
- D. Emilio Carbonell Guevara
- D. Jorge Mateu Mahiques
- D. Mario Plaza Delgado
- D. José Domingo Bermúdez Edo

Va ser dirigida per:
D. Mario Sendra Pina

Depòsit legal:
I.S.B.N.:84-370-5880-5
Edita: Universitat de València
Servei de Publicacions
C/ Artes Gráficas, 13 bajo
46010 València
Spain
Telèfon: 963864115

VNIVERSITAT DE VALÈNCIA

Facultat de Ciències Matemàtiques

Departament d'Estadística i Investigació Operativa

Anàlisi Discriminant Discreta Mitjançant Suavització de les Correspondències Múltiples

Memòria presentada per
J. Vicent Pruñonosa Reverter
per optar al grau de Doctor en
Ciències Matemàtiques

Dirigida pel
Dr. Mario Sendra Pina

Don Mario Sendra Pina, Professor Titular d'Estadística i Investigació Operativa del Departament d'Estadística i Investigació Operativa de la Universitat de València

CERTIFICA que la present memòria d'investigació:

"Anàlisi Discriminant Discreta Mitjançant Suavització de les Correspondències Múltiples"

ha estat realitzada sota la seva direcció al Departament d'Estadística i Investigació Operativa per J. Vicent Pruñonosa Reverter, i constitueix la seva tesi per optar al grau de Doctor en Ciències Matemàtiques.

I perquè així conste, en compliment amb la normativa vigent, autoritza la seva presentació a la Facultat de Matemàtiques de la Universitat de València per a que pugui ser tramitada la seva lectura i defensa pública.

Burjassot, 2 de setembre de 2003

EL DIRECTOR

Mario Sendra Pina

Índex general

Índex de figures

Agraïments

Agraeixo al Departament d'Estadística i Investigació Operativa de la Universitat de València el seu suport i, molt especialment, al meu director, Mario Sendra, sense el qual aquest treball hagués estat del tot impossible de realitzar.

Vull agrair, també de manera especial, als brigadistes de salut de les comunitats de El Jicaral, El Sauce i León de Nicaragua per l'esforç amb l'aportació de les dades que han estat el punt de partida d'aquest estudi.

J. Vicent Pruñonosa Reverter

Introducció

La motivació per a realitzar el present estudi prové de l'anàlisi epidemiològica dels factors influents en el patró d'evolució de les malalties, on es pretén determinar quines variables i en quin grau influeixen en els canvis, tant favorables com desfavorables, que pugui tenir una persona en els nivells de salut quan rep un tractament determinat.

El fet que aquests factors són, en gran part, variables categòriques dificulta considerablement l'aplicació de les tècniques estadístiques específiques incloses dins l'àmbit de l'anomenada anàlisi discriminant.

Com és sabut, aquesta anàlisi, en aquest context, ens ha de permetre assignar a una persona el patró d'evolució més probable de la seva malaltia en funció de les dades sociosanitàries disponibles, prenent com a referència un conjunt de persones d'evolució coneguda (dades d'aprenentatge).

La dificultat matemàtica prové del fet que la simplificació que introdueix la suposició de continuïtat, molt estudiada i amb resultats que poden considerar-se satisfactoris, no és aplicable a la majoria de les variables disponibles, i es fa necessari adaptar el mètode sense forçar la natura d'aquestes.

Cal, per tant, malgrat que pugui semblar una reflexió massa filosòfica, aprofundir, encara que sigui breument, sobre els conceptes de categòric i continu per tal d'orientar adequadament aquesta adaptació.

Si considerem el nivell perceptiu com la base de l'aproximació contínua, podem enfocar aquesta com corresponent a un petit, però molt significatiu, interval

sensorial, de manera que per sota d'ell la realitat la podem imaginar discreta i per amunt la tornem a percebre categòrica com a identificació d'objectes diferenciats.

Des d'aquest punt de vista podem considerar que molts fenòmens discrets, especialment els de natura biològica, són el resultat d'un procés d'acumulació-umbralització a partir de variables subjacents contínues.

En el context epidemiològic esmentat es pot suposar que una combinació de factors continus subjacents determina l'aparició d'un determinat patró d'evolució, i que a mesura que ens allunyen d'aquesta combinació, la probabilitat de que es presenti aquest patró disminueix de manera que a l'apropar-se a la combinació que determina un altre patró, la probabilitat d'aquest últim arriba a ser la dominant.

La traducció matemàtica d'aquesta idea consisteix a suposar que les variables categòriques procedeixen de la discretització de subjacents contínues, que segueixen un model probabilístic conegut com a mixtura de normals.

Aquest és el punt de partida del mètode que es presenta en aquest treball, ja que d'aquesta manera podem considerar, com es habitual a la literatura, que els factors significatius en la determinació del patró afecten a la mitjana de les variables subjacents mentre que els no significatius determinen una dispersió gaussiana arreu dels valors centrals.

L'esforç es centrarà, com a conseqüència, en retrobar el més acuradament possible la distribució probabilística contínua subjacent, i posteriorment aplicar una metodologia de discriminació amb variables contínues.

Per tal d'aconseguir aquest objectiu "reconstructor", en tindrem, al seu torn, dues fases: A la primera, i mitjançant una anàlisi de correspondències múltiples convenientment adaptada a l'objectiu discriminant, cercarem quantificacions que aproximen les mitjanes de les cel·les resultants de la discretització. A la segona, emprant un procediment de suavització, completarem la reproducció de la distribució subjacent aplicant una dispersió al voltant d'aquestes mitjanes.

El capítol 1 analitzarà les definicions bàsiques de l'anàlisi discriminant i farà una revisió dels mètodes existents amb l'objectiu esmentat.

El segon i el tercer capítols es centraran a fer l'equivalent amb l'anàlisi de correspondències i els mètodes de suavització (fonamentalment *Kernel* i *EM*) com a elements bàsics a combinar, per tal d'aconseguir l'esmentada reconstrucció.

Al capítol 4 es farà la proposta metodològica i es demostrarà el resultat que li dóna fonament matemàtic.

Al darrer capítol, el 5, es discutiran els resultats amb dades simulades i reals, comparant amb altres mètodes de freqüent utilització.

Finalment, amb posterioritat a les conclusions, es faran suggeriments, tant per a la possible continuació de la recerca teòrica com per a la seva aplicació pràctica.

Capítol 1

L'anàlisi discriminant

Aquest capítol conté una revisió dels conceptes i mètodes de l'anàlisi discriminant, emfatitzant els que tenen una aplicació més directa a la situació en estudi. Començarem per precisar matemàticament aquesta (1.1) passant, a continuació, a la presentació dels conceptes bàsics hi involucrats (1.2), especialment els relatius a l'error de classificació (1.2.1), eina imprescindible per a la selecció de variables (1.2.2) i per tenir una valoració orientadora dins del conjunt de mètodes i models, els quals són revisats a l'apartat (1.3).

1.1 Precisió de la situació en estudi

Anomenarem y la variable que recull les diferents classes a la que pertanyen els individus de la població en estudi. La y serà, per tant, una variable categòrica amb valors $1, \ldots, g$.

Per altra banda identificarem com a x el conjunt de p variables, que mesurades als individus de la mateixa població, ens han de servir per aproximar el valor de y quan aquesta sigui desconeguda. La x representarà, per tant, un vector de p components x_j que al llarg de l'estudi es consideraran categòriques amb valors $1, \ldots, k_j$. A la secció de suggeriments, pàgina 116, es farà un comentari sobre com es podrien incorporar noves variables de natura continua.

Suposarem que les x_j procedeixen de discretitzacions unidimensionals de la variable contínua, $p-$dimensional subjacent u, la qual, al seu torn, serà una mixtura de densitat $\sum_{i=1}^{g} p(i)\phi(u;\mu_i,\Sigma)$ on ϕ representa la densitat Normal i $p(i)$ la probabilitat de cada classe i dins de la població general.

També se suggerirà a la pàgina 116 quina pot ser l'adaptació del mètode proposat en aquest treball si considerem que en lloc d'una Σ comuna en tenim Σ_i diferents per classe o en lloc d'una mixtura de g normals en tenim una d'un nombre $g' > g$.

Finalment, considerarem la mostra de dades d'aprenentatge formada per Y, vector de mida n, realització d'y i X, matriu de mida $(n \times p)$, realització, al seu torn d'x.

1.2 Els conceptes bàsics de l'anàlisi discriminant

Donat que, per tractar una situació com la que acabem de descriure, la metodologia que s'utilitza es coneix sota el nom d'anàlisi discriminant, revisarem en aquesta secció els conceptes bàsics d'aquest tipus d'anàlisi.

1.2.1 Notacions bàsiques

Si prenem $\mathbb{R}_p$ com a espai subjacent i $f(u/i)$ com a les densitats de probabilitat atribuïdes a les p variables continues subjacents per a cada classe $i = 1, \ldots, g$ (Normals al model proposat) serà $f_i(u) = f(u/i) \cdot p(i)$ la component i-sima de la mixtura corresponent a la classe i i, per tant, $Z_i = \{u / f_i(u) \geq f_{i'}(u)\ \forall i'\}$ la zona on domina aquesta component.

Definirem la matriu quadrada M amb:

$$M_{ij} = \int_{Z_j} f_i(u)du \qquad i = 1, \ldots, g \quad j = 1, \ldots, g \tag{1.1}$$

de manera que a la diagonal tindrem les probabilitats "d'encerts" a la classificació emprant directament les densitats contínues subjacents i considerant aquestes conegudes.

Si procedim ara a discretitzar mitjançant p particions de manera que per a la variable j tinguem $\mathbb{R} = \bigcup_{r=1}^{k_j} S_{jr}$ on k_j representa el nombre de categories resultants per a aquesta variable, la partició global serà $\mathbb{R}_p = \bigotimes_{j=1}^{p} \bigcup_{r=1}^{k_j} S_{jr} = \bigcup_{r'=1}^{k'} {S'}_{r'}$ amb $k' = \prod_{j=1}^{p} k_j$.

S'ha de fer constar que, encara que aquesta discretització és general, les particions que utilitzarem en aquest treball seran les més freqüents, basades en intervals, tal i com es representa a la figura 1.1.

L'efecte de la discretització sobre l'error discriminant es pot valorar tenint en compte que aquest procés ens porta a les probabilitats $p_i(r') = \int_{S'_{r'}} f_i(u)du$ i a la corresponent matriu M^d amb:

$$M^d_{ij} = p_i(Z'_j) \qquad \text{on} \qquad Z'_j = \{r'/p_j(r') \geq p_{j'}(r') \quad \forall j'\} \tag{1.2}$$

Finalment, si anomenem $p^n_j(r')$ i ${Z'_j}^n = \{r'/p^n_j(r') \geq p^n_{j'}(r') \quad \forall j'\}$ a les probabilitats i zones de domini resultants quan substituïm la distribució original per la resultant d'una mostra de mida n, podem definir les matrius M^{da} i M^{dr} com:

$$M^{da}_{ij} = p^n_i({Z'_j}^n) \qquad M^{dr}_{ij} = p_i({Z'_j}^n) \tag{1.3}$$

on, en el primer cas s'ha aplicat la substitució per l'equivalent mostral tant a la probabilitat com a les zones de domini, mentre que, en el segon, només s'ha fet la substitució en aquestes darreres conservant la probabilitat poblacional. Les matrius equivalents per al cas continu serien M^{ca} i M^{cr} amb:

$$M^{ca}_{ij} = \int_{Z^n_j} f^n_i(u)du \qquad M^{cr}_{ij} = \int_{Z^n_j} f_i(u)du \tag{1.4}$$

on $Z^n_j = \{u/f^n_j(u) \geq f^n_{j'}(u) \quad \forall j'\}$.

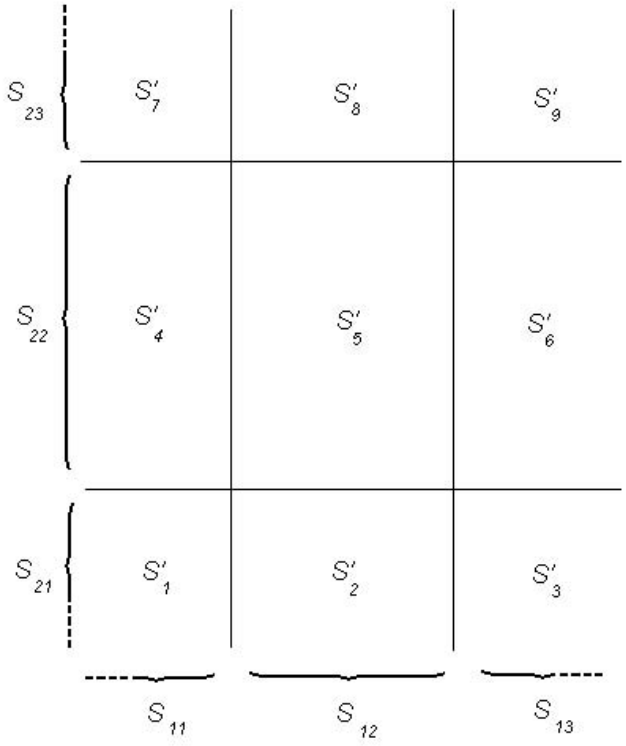

Figura 1.1: *Esquema de la discretització*

1.2.2 Els diferents tipus d'errors a considerar

La mesura més directa per valorar la qualitat d'un procés d'anàlisi discriminant és l'error de classificació que proporciona. Aquest serà, per tant, el primer concepte que revisarem i formalitzarem per tal que ens serveixi de guia al llarg de tot l'estudi.

Hem de començar per verificar que la definició d'aquest error no és senzilla, donat que només s'apropem al tema veiem de seguida que, amb les suposicions departida, no podem considerar un únic error, sinó més ben bé una cadena que caldrà seguir per veure com es va transmetent per totes les etapes del procediment de discriminació.

1.2.2.1 Error òptim continu

El primer error a considerar és el que anomenarem e_c o error òptim continu, que vindrà determinat pel grau de solapament entre les densitats contínues subjacents:

$$e_c = 1 - traça(M) \tag{1.5}$$

Si considerem $f(u/i) \sim N(\mu_i, \Sigma)$ aquest és l'error òptim per aplicació del teorema de Neyman-Pearson (a la fase final una anàlisi discriminant sempre es pot veure com un contrast d'hipòtesi simple o múltiple) i és el que Fisher [60], va demostrar equivalent a l'obtingut classificant mitjançant les distàncies de Mahalanobis (mètode *LDA*: *Linear Discriminant Analysis*).

També és conegut que si cerquem direccions ortogonals de manera que e_c sigui mínim en cada etapa, cal trobar els vectors propis de la matriu $B\Sigma^{-1}$ on B és la matriu de covariància dels centroïds μ_i (LDA-canònic).

1.2.2.2 Error òptim de la discretització

El segon error a considerar o, millor dit, la segona etapa en la consideració de l'error, ens porta a l'e_d o error òptim resultant després de la discretització. Es tracta, ara, d'adjudicar a cada cel·la resultant de la discretització la classe on resulti la probabilitat més alta.

$$e_d = 1 - traça(M^d) \tag{1.6}$$

Aquest error tendirà a e_c quan la mida dels S_{jr} mencionats al punt (1.2.1) tendeixi a 0 (com a conseqüència els $k_j \to \infty$) i serà menor en la mesura que la partició S' s'aproximi a un recobriment de la Z. També sabem que en la mesura que la direcció de discretització s'apropi a la principal dels centroïds, l'error disminueix, com pot observar-se gràficament a la figura 1.2.

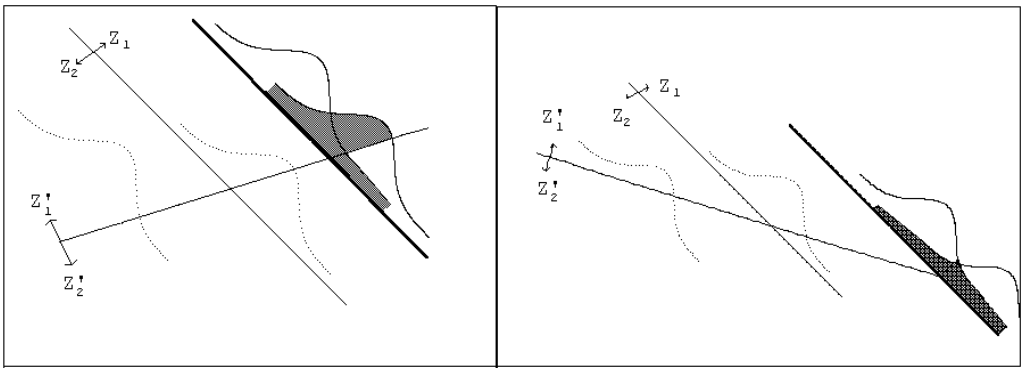

Figura 1.2: *Error òptim segons l'angle de la discretització.*

1.2.2.3 Error mostral

La tercera etapa fa intervenir l'error mostral, i depenent de si la substitució de les p_i poblacional per les p_i^n es fa una vegada o dues a la fórmula, tenim l'error real (estimat com a mitjana) e_{dr} o l'aparent e_{da}:

$$e_{dr} = 1 - traça\big(\mathrm{E}(M^{dr})\big) \qquad e_{da} = 1 - traça\big(\mathrm{E}(M^{da})\big) \tag{1.7}$$

L'equivalent abans de discretitzar seria:

$$e_{cr} = 1 - \mathit{traça}\big(\mathrm{E}(M^{cr})\big) \qquad e_{ca} = 1 - \mathit{traça}\big(\mathrm{E}(M^{ca})\big) \tag{1.8}$$

Un resultat bàsic de l'anàlisi discriminant és que, el que hem anomenat error aparent (o l'esperança de l'error aparent com es coneix a la literatura), té un biaix negatiu, mentre que l'error real el té de positiu, és a dir que:

$$e_{da} \leq e_d \leq e_{dr} \qquad \text{i, equivalentment,} \qquad e_{ca} \leq e_c \leq e_{cr}$$

S'ha de tenir en compte, de totes maneres, que l'operador $1-\mathit{traça}(M)$ que s'ha utilitzat en totes aquestes expressions correspon a l'anomenada pèrdua 0–1 (ben o mal classificat) i que existeixen altres possibilitats com a funcions de pèrdua, entre les que podríem destacar la provinent de la teoria de la informació, que empraria l'operador $-2\mathit{traça}(\log(M))$, conegut com a *entropia creuada ("deviance")*, amb el 2 com a factor que el fa igual a la pèrdua mínimo-quadràtica per al cas normal (Hastie et all, 2001, [119], pàg.195)

També s'ha de tenir en compte, que la suposició implícita que tots els M_{ij} amb $i \neq j$ tinguin el mateix pes pel càlcul de l'error, pot ser variada per consideracions pràctiques associades al context del problema, i en aquest cas, s'hauria d'anar a un operador de tipus suma ponderada.

Per altra banda, hem de considerar que els biaixos $|e_{ca} - e_c|$ i $|e_{cr} - e_c|$ tenen expressions conegudes per al cas Normal i que són el límit a què tendirien $|e_{da} - e_d|$ i $|e_{dr} - e_d|$ amb els assenyalaments expressats a l'apartat anterior.

Glick (73)[81] va establir unes cotes per a aquests biaixos en el cas de dues classes:

$$|e_{da} - e_d| \leq (0.5m/\sqrt{n})\alpha^n \qquad \text{i} \qquad |e_{dr} - e_d| \leq (0.5 - e_d)\alpha^n$$

on m és el nombre de cel·les (de les k' totals) amb valors de p_i diferents:

$$m = Card\,\{r'/p_1(r') \neq p_2(r')\}$$

i α, valor real positiu menor de la unitat, és una mesura de "separabilitat" entre classes donada per:

$$\alpha = 1 - \left(\inf_{r'} \left| \sqrt{p_1(r')} - \sqrt{p_2(r')} \right| \right)^2$$

d'on es conclou que el biaix de l'error aparent convergeix més ràpidament a 0 que el de l'error real .

Goldstein i Dillon (1978) [84] analitzen amb detall la situació estudiada per Glick partint de la base que les seves cotes son massa àmplies, i fan una descomposició del biaix per categories, assenyalant, entre d'altres aspectes que, respecte a la mida relativa per classe dels elements d'una determinada categoria el biaix mínim s'obté quan aquesta s'iguala.

També fan referència a una situació que pot resultar sorprenent: les cotes de Glick, en determinats casos, poden ser inferiors als valors dels biaixos estimats per al cas normal subjacent, el que faria considerar en aquestes circumstàncies una possible discretització com a mesura de reducció de biaix (veure conjunt de dades 1 comparant els errors de les seccions 5.2.1 (pàg. 101) i 5.2.3).

Per altra banda, és important, des del punt de vista pràctic, abordar en aquest context com es farà l'estimació de l'error real donat que, per estimar les p_i que utilitzem a la matriu M^{dr} hem d'emprar una mostra independent de la d'aprenentatge, anomenada habitualment de *test* i, per tant, l'esperança que apareix a la formula del e_d vindrà determinada a partir d'un doble procés d'estimació amb mitjanes mostrals: de la mostra d'aprenentatge i de la mostra de test.

En cas que no es disposi de la possibilitat d'obtenir mostres de test es, poden utilitzar mètodes com la validació creuada i el *bootstrap* per tal de corregir el biaix de l'error aparent. Com aquests mètodes són també molt utilitzats per a l'estimació de paràmetres d'un model, els analitzarem breument a l'apartat següent.

Finalment, cal assenyalar que l'anàlisi discriminant també pot enfocar-se com una regressió (simple o múltiple depenent del nombre de classes) si es prenen $Y_i = p_i(x)$ i, en aquest context, es pot definir l'error in-mostra, el qual, per simplicitat, considerarem per al cas de dues classes amb una única $Y = t(x) + \varepsilon$ i amb funció de pèrdua quadràtica, com :

$$\sigma_\varepsilon^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(t(x_i) - E\big(\hat{t}(x_i)\big)\Big)^2 + \frac{p}{n}\sigma_\varepsilon^2$$

el qual resulta de calcular l'error quadràtic mitjà deixant fixa la mostra de les X i variant la Y. Aquest error, d'utilitat més teòrica que pràctica, ens serveix per separar del biaix de l'error real la part corresponent a la "extrapolació", entenent per aquesta, el fet d'utilitzar uns valors X diferents dels de les dades d'aprenentatge. Analitzant la fórmula observem que el primer component és, per dir-ho així, l'error irreductible que prové de la variabilitat essencial de la Y i els altres dos components representen els clàssics del biaix i la variància d'una estimació mínimo-quadràtica. Podem retenir la proporció p/n com a indicador de la "complexitat" del model amb afectació proporcional sobre el tercer component.

1.2.2.4 Error final

A la darrera etapa en aquest procés d'aproximació successiva a l'error de classificació, hem de fer intervenir el fet que les probabilitats i les zones de domini discretes mostrals per a cada classe p_i^n, Z'^n_i no seran conegudes amb exactitud sinó estimades finalment a partir d'un mètode, obtenint p_i^{fn}, Z'^{fn}_i.

Substituint a la definició dels errors de l'apartat anterior les matriu corresponents per les equivalents amb els valors estimats, obtindrem els errors finals:

$$e_{fr} = 1 - traça\big(\mathrm{E}(M^{fr})\big) \qquad e_{fa} = 1 - traça\big(\mathrm{E}(M^{fa})\big) \tag{1.9}$$

amb $M_{ij}^{fa} = p_i^{fn}(Z'^{fn}_j)$ i $M_{ij}^{fr} = p_i^{fr}(Z'^{fn}_j)$.

Naturalment, el mètode més simple consisteix a prendre com a p_i^{fn} les freqüències relatives, però aquest procediment presenta, entre altres possibles problemes, el

de no poder classificar si no s'ha trobat a la mostra d'aprenentatge cap element d'una determinada categoria.

Aquesta dificultat il·lustra, amb prou claredat, la necessitar de trobar models que copsen la natura bàsica de les p_i i, mitjançant un procediment que les aproximi, aconseguir una classificació més complerta i exacta.

En terminologia paramètrica en tindrem $p_i(\theta)$ i dividirem la mostra d'aprenentatge en dues parts, de manera que una d'elles ens serveixi per l'ajust de θ i l'altra per a la seva validació.

El determinar quina fracció ha de correspondre a cada part, i quin serà el procediment de validació, entra de ple al tema ja esmentat de la validació creuada i el *bootstrap*.

En relació a la validació creuada hem de tenir en compte que aquest procediment, àmpliament documentat a la literatura (veure, per exemple, Bowman, 1984 [14], Davison, 1992 [45] i Sain, 1994 [185]) té moltes variants depenent de la funció de pèrdua que s'utilitzi. Les de millors propietats reconegudes són la de Kullback-Leibler ($\hat{p}\log p$) i la mínimo-quadràtica $(p-\hat{p})^2$. La decisió dependrà d'una sèrie de consideracions ad-hoc que s'han de discutir per a cada cas concret com, per exemple, el rang de valors del paràmetre en el que ens movem, com evidencia el gràfic de la figura 1.3, on es representen les funcions bàsiques hi involucrades $(x-1)^2$ i $x\log(x)$ amb $x=\frac{\hat{p}}{p}$. A aquest gràfic observem l'avantatge d'emprar la pèrdua de Kullback-Leibler si infraestimem p ($\frac{\hat{p}}{p}<1$) i , al contrari, utilitzar la pèrdua mínimo-quadràtica si el sobreestimem. En cas de dubte sembla més adient utilitzar la mínimo-quadràtica per la seva simetria.

En quant al *bootstrap* s'ha de tenir en compte, de forma general, que, en relació a la validació creuada aquest aconsegueix una reducció del biaix a canvi d'un augment de la variància, efecte que es pot compensar si s'agafa una fracció d'ajustament petita o s'utilitza la correcció 0.632 (Efron, 1982 [58] i Fitzmaurice et all, 1991 [62]). En molts casos, encara que no necessàriament en tots, serà

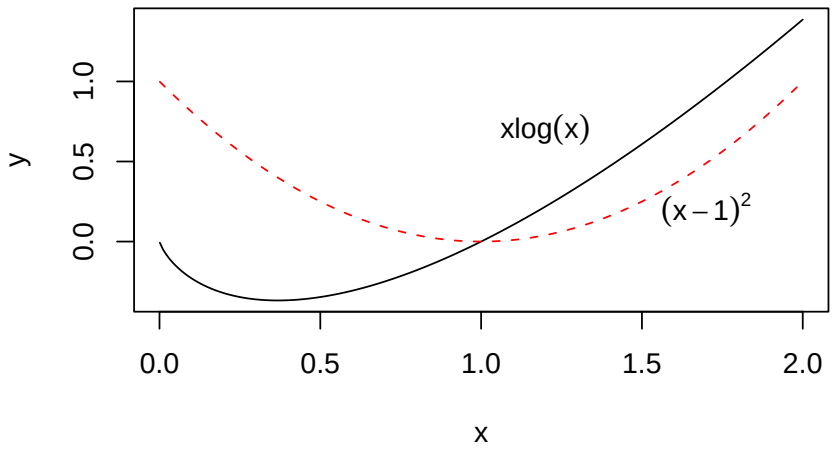

Figura 1.3: *Comparació de les funcions de pèrdua minímo-quadràtica i de Kullback-Leibler.*

l'error quadràtic mitjà $ECMI = biaix^2 + variància$ qui ha d'ajudar a prendre la decisió final.

Cal, finalment, considerar que no és imprescindible estimar les p_i per obtenir l'estimació de l'error real, ja que existeixen procediments per a estimar-lo directament, ben mitjançant quocients de versemblances, ben emprant distribucions en lloc de densitats al cas continu, ben per qualsevol altre procediment dels que han estat estudiats en aquest objectiu, entre els que cal destacar l'ajustament de l'error in-mostra (mitjana d'errors de submostres de la mostra original) partint de l'error aparent tal i com es resumirà a continuació.

Per fer aquest ajustament utilitzarem l'interessant resultat que, tant per a la pèrdua quadràtica que l'ha servit d'inspiració, com per a la més general que hem utilitzat en aquesta presentació, l'anomenada 0-1, el biaix de l'error aparent en

relació a l'in-mostra és (Hastie et all, 2001, [119] pàg.202):

$$\frac{2}{n}\mathrm{Cov}(\hat{Y}, Y) \tag{1.10}$$

Aquesta estimació del biaix ens permet, quan $\hat{Y}$ és obtingut mitjançant un model lineal, calcular l'error in-mostra, simplement sumant a l'error aparent la quantitat $\frac{2p}{n}\sigma_\varepsilon^2$. A més, aquesta fórmula ens reflecteix d'una manera molt clara i sintètica que la millora de la qualitat de l'estimació (alta $Cov(\hat{Y}, Y)$) ens farà augmentar l'error aparent (reduint el seu biaix negatiu), al separar-se de la mostra en el seu camí d'adaptació a la població general .

1.2.3 Selecció de variables

Encara que conservem el títol clàssic d'aquest apartat hem d'assenyalar que aquest concepte pot considerar-se tal i com fan Hastie et all, 2001 [119] inclòs dins d'un de més ampli identificat sota el nom de complexitat d'un model. Efectivament, podem entendre que reduir variables mitjançant la selecció de la seva eficiència discriminant forma part de tractar d'identificar la complexitat real que ha de tenir un model per reflectir una situació concreta. Conseqüentment, l'abordatge més senzill d'aquest concepte seria la seva definició a partir del nombre de variables seleccionades pel model discriminant del conjunt de les disponibles, tal i com ja es va fer a l'apartat 1.2.2.3 (pàg. 10) al tractar de l'error in-mostra (p/n). És àmpliament conegut que, d'aquesta manera, l'error aparent e_{fa} disminueix amb la complexitat mentre que el real e_{fr} presenta un mínim tal com poder sintetitzar amb la figura 1.4:

Cal ressenyar, com a conseqüència, que, a part naturalment dels avantatges de simplificació i interpretabilitat de la reducció del nombre de variables, hem de tenir en compte que, en discriminació, aquesta reducció pot disminuir també l'error real de classificació, per més que no ho faci sobre l'error aparent. Aquesta n'és, per tant, una altra evidencia que l'error aparent no ens serveix com a mesura de la bondat de classificació, donat que sempre millorarà amb la complexitat (i per

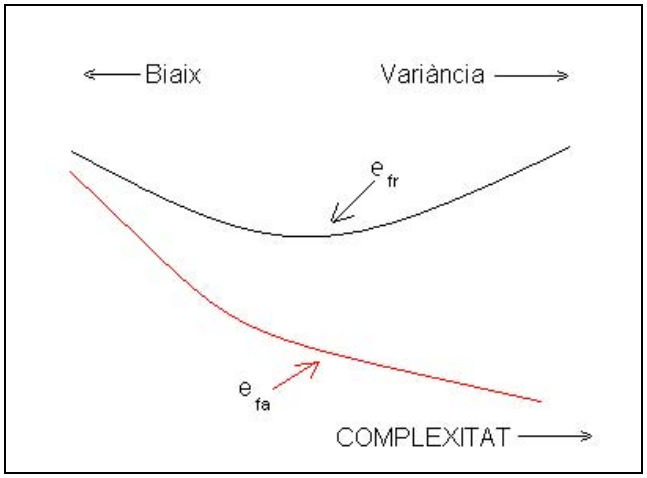

Figura 1.4: *Evolució dels errors finals real i aparent segons la complexitat del model.*

tant amb el nombre de variables per molt que les que afegim siguin completament espúries).

Ara bé, el càlcul de l'error real per a cada subconjunt de variables candidat és molt tediós, pel que Hand (1982) [107] proposa mesures ràpides de separabilitat que, en definitiva, són extensions de la distància de Mahalanobis.

Altra proposta molt utilitzada empra mètodes seqüencials i especialment els de dos cap endavant i un cap endarrere suggerits per Kittler (1978) [134]. L'inconvenient, comú a tots les del mateix tipus és el possible problema de coherència que es pot presentar depenent del punt de partida.

Rao (1995) [177] va proposar un contrast F per regular l'entrada i sortida de variables però va alertar en relació a l'augment de l'α global.

Per evitar aquest problema McKay i Campell (1982) [152] suggereixen procediments amb α global fixe per seleccionar un subconjunt inicial, seguits per l'aplicació de mètodes seqüencials amb regles d'aturada per criteris probabilístics com els que va proposar McLachan (1976) [148].

Existeix també la possibilitat de seleccionar categories més que variables, com ja van proposar Goldstein i Dillon, 1978 (Cap.4) [84] utilitzant la χ^2 sobre una distància de Kullback-Leibler tal i com es fa, més modernament, als mètodes d'arbres que explicarem posteriorment a l'apartat 1.3.6.1 (pàg. 25).

Finalment, podem ressenyar els mètodes canònics que, front a l'avantatge d'emprar una operativa estàndard i ben coneguda, tenen l'inconvenient de transformar les variables originals dificultant la interpretabilitat dels resultats.

1.3 Revisió de mètodes discriminants

En aquesta secció revisarem els principals mètodes d'anàlisi discriminant sota la perspectiva de la seva aplicació a la situació en estudi descrita a la secció 1.1.

1.3.1 El difícil equilibri local-global

A l'apartat anterior avançàvem que per definir la complexitat d'un model calia alguna cosa més que el nombre de variables o de categories. Com hi comentàvem l'excessiva adaptabilitat d'un mètode a les dades mostrals pot resultar contraproduent de cara a complir el seu paper a la població sencera. Aquestes observacions continuen sent vàlides, en forma general, si com a mesura de complexitat fem intervenir l'adaptabilitat local. És a dir si, per exemple, l'augmentem substituint un model lineal global per un lineal a trossos.

Donat això, l'objectiu serà trobar el nivell de complexitat adient el qual, si considerem ja realitzada la selecció de variables esmentada a l'apartat anterior, es tradueix en un equilibri local-global que, al seu torn, ens portarà a un equilibri biaix-variància que hem de procurar que estigui a l'entorn de l'error mínim (ver figura 1.4).

Si aquest equilibri s'esbiaixa cap a la banda local podem "pegar" massa la regla discriminant a les nostres dades d'aprenentatge i obtenir resultats decebedors amb les dades de test (alta variància) i si s'esbiaixa cap a un model global

massa genèric trobarem també resultats dolents (alt biaix) per molt que l'estimació dels paràmetres ens amagui part del problema quan fem servir les dades d'aprenentatge.

Tenint això present farem un recorregut pels diferents mètodes situant-nos inicialment als dos extrems: El model lineal com la globalitat màxima i els dels 1-veïns (comparació directa de freqüències relatives) com el de la major localització.

Com el model lineal va ser el primer mètode discriminant i continua tenint avui una innegable importància, li dedicarem un apartat específic analitzant aquí els dels k-veïns com a exemple de la manera de tractar la localització i com a indicador dels problemes pràctics que es poden donar si aquesta és excessiva.

Al mètode dels k-veïns estimarem les p_i^n com:

$$p_i^n(x) = \frac{1}{k} Card\{x_j \in V_k(x) \ / y_j = i\}$$

on $V_k(x)$ representa l'entorn d'x que conté k punts en una mètrica prèviament definida. Per tant els 1-veïns es correspon amb la comparació de les freqüències relatives ja esmentada a l'1.2.2.3 i és, sense cap dubte, el mètode més local que podem considerar. A més en tenim amb k un paràmetre molt intuïtiu per graduar la localització, el qual ens servirà per veure el que pot succeir si el nombre de variables p és gran.

Per evidenciar-ho prendrem $p = 10$ i els punts x repartits uniformement per l'hypercub unitari amb $k = 0.01n$ resultant que com $0.01^{1/10} = 0.63$ hem d'allargar-nos fins el 63% de l'interval total de cada variable si volem copsar un entorn de només l'1% dels punts totals. Evidentment, la localització queda seriosament en entredit quan el nombre de variables augmenta, i és fàcil veure que el problema s'agreuja per als valors situats a la frontera. A l'apartat 1.3.6.5 (pàg. 29) es comentarà un procediment dissenyat modernament, de forma específica, per intentar resoldre aquesta dificultat.

1.3.2 Els models basats en la Normal: la robustesa de l'*LDA*

Passem ara a revisar el mètode més global de tots: l'*LDA* (*Linear Discriminant Analysis*) el qual, desenvolupat inicialment per Fisher [60] , es basa en una anàlisi canònica de la matriu $B\Sigma^{-1}$ esmentada a l'apartat 1.2.2.1 (pàg. 9) per tal de trobar les transformacions lineals ortogonals de màxima discriminació. Com es deia a aquell apartat, aquest procediment condueix a l'error òptim quan les X segueixen distribucions Normals de variància comuna. Pel cas de variàncies diferents per classe una modificació directa del mètode ens porta al *QDA* (*Quadratic Discriminant Analysis*) on les funcions discriminants són polinomis de grau 2.

Les principals característiques del mètode per al nostre objectiu són:

1. Si les X són categòriques i tenim només dues classes amb $p(1) = p(2)$, l'*LDA* coincideix amb una regressió lineal sobre les variables indicadores (dicotomització de les X). Si $p(1) \neq p(2)$ la direcció discriminant és la mateixa, però cal modificar el punt de tall. En general, l'*LDA* pot veure's com una regressió sobre indicadores, seguida d'una descomposició canònica d' $\hat{Y}'Y$, el que evita el problema que les classes intermèdies queden anul·lades, indesitjable fenomen que, per la pròpia natura de la regressió, ocorreria si s'apliqués aquesta directament a la discriminació.

2. *QDA* és pràcticament equivalent a un *LDA* on s'han incorporat a les X els quadrats de les variables originals.

3. *LDA* funciona acceptablement encara que no es compleixi la suposició de normalitat perquè la seva estabilitat sol compensar el seu biaix. Aquesta propietat que explica la continuïtat del seu ús és la que hem anomenat robustesa del *LDA*.

1.3.3 El models basats en la multinomial: La versatilitat de la logística

Una vegada descrits els dos extrems de l'interval de complexitat local-global, farem un recorregut pels mètodes intermedis respecte a aquest criteri.

Començarem amb els que es basen en el model probabilístic més general per a una situació discreta: el multinomial.

D'entre ells el més important és el logístic, sorgit per evitar que l'aplicació directa d'una regressió de les variables indicadores sobre les $p_i(x)$ ens menés a la possibilitat desagradable de tenir estimadors de probabilitat fora de l'interval $[0, 1]$.

En aquest cas se suposa que la distribució base és multinomial, i prenent una classe com a referència (per exemple, la primera), ajustem els paràmetres del següent esquema:

$$\log \frac{p_i(x)}{p_1(x)} = \beta_i(x) \qquad i = 2, \ldots, g$$

Aquest model presenta la interessant propietat de ser aplicable tant per a distribucions normals com per a multinomials (passant per la dicotomització), el que facilita la seva aplicació a situacions mixtes. També es pot convertir l'ajustament dels β en un procés de mínims quadrats iteratiu, emprant una estimació χ^2 de les diferències entre Y i $\hat{Y}$ (Hastie et all, 2001 [119] pàg.103)

A més, la logística permet incloure al model els productes corresponents a les interaccions, el quals ens donen la flexibilitat necessària per poder gestionar les desviacions de la independència, d'una forma semblant a la que la majoria dels estadístics estem acostumats a utilitzar en els models lineals habituals, i que és molt més clara que la dels primers intents que, com el de Bahadur, es van fer amb aquest objectiu partint directament de la multinomial (Goldstein i Dillon, 1978 [84]).

Un altre enfocament que parteix de la multinomial és el de Cuadras, 1990, [41] el qual relaciona les distàncies entre els membres d'una classe amb la distància

del candidat a classificar a la classe en conjunt, el que també és aplicable al cas mixt donat que per a distribucions Normals es poden utilitzar les distàncies de Mahalanobis. Aquest mètode tindrà aplicabilitat quan el context del problema ens suggereixi una mètrica adient, el que no és pas el cas que nosaltres hem considerat.

S'ha de tenir en compte, però, (tornant a la logística, encara que el resultat que comentarem pot extendre's, en l'essencial, a tots els mètodes basats en la multinomial) que si les distribucions són Normals, la logística pot incrementar l'error real fins a un 30% sobre l'*LDA* (Hastie et all, 2001 [119] pàg.105), el que la fa desaconsellable, en principi, pel cas que ens ocupa en el que partim, precisament, de Normals subjacents. A més, una gran quantitat d'exemples proven que la logística és molt útil als casos mixtos quan el nombre de variables contínues supera àmpliament al de categòriques, situació que és exactament la contrària de la del nostre punt de partida.

Modernament, Venables i Ripley, 2002 [218] han proposat una modificació de la logística fent servir el procediment de les xarxes neurals que es comentarà a 1.3.6.3 (pàg. 27), la qual s'ha revelat d'una gran qualitat discriminant fins al punt de ser la preferida en situacions genèriques discretes.

1.3.4 L'expansió en funcions base: el Discriminant Flexible (*FDA*)

Després de la breu incursió del paràgraf precedent pels models basats en la multinomial, hem arribat al convenciment que cal aprofitar la informació sobre les Normals subjacents si volem obtenir una classificació més acurada.

Tornem, per tant, als mètodes basats en la Normal, els quals, partint de l'*LDA*, han cercat rebaixar la seva globalitat per donar-li flexibilitat localment adaptable.

El primer que revisarem té el seu origen en els treballs d'Andreas Buja sobre la correlació canònica no lineal (Buja, 1990 [22]) els quals, conjuntament amb

els estudis fets per ajustar polinomis locals com les *natural cubic splines* i les *wavelets* (Donoho et all, 1995 [56]), han quallat de la ma d'Hastie i Tibshirani en l'anomenat *FDA* (*Flexible Discriminant Analysis*). Aquest mètode, de manera sintètica, es composa de les següentes passes:

1. Regressió lineal múltiple adaptable no paramètrica $\hat{Y}$ d'Y sobre X.
2. Anàlisi canònica d'$\hat{Y}'Y$.
3. Tornar al punt 1 amb els resultats de 2.

Es tracta d'una idea senzilla i poderosa: per alliberar l'*LDA* de les seves restriccions massa rígides, conservant al màxim les seves propietats derivades de ser un model lineal, el que es fa és expandir polinòmicament les variables originals (com es feia al *QDA* per al grau 2), permetent que una adaptació local seguida d'una anàlisi canònica global faci el corresponent treball de selecció entre las variables expandides.

L'*FDA* connecta amb els mètodes previs basats en una expansió en sèries de Fourier, penalitzant, a l'estil de la *Ridge Regression*, les freqüències altes. Al seu fonament matemàtic utilitza la important *Kernel property*, basada en funcions *Kernel* com les que descriurem a la secció 3.3 (pàg. 56), la qual pot resumir-se com:

> *Si es fa una estimació lineal emprant un Kernel K i s'utilitza una funció de pèrdua quadràtica amb el mateix tipus de Kernel, l'error mínim s'obté, simplement, aplicant K als punts d'aprenentatge* (Girosi et all, 1995 [78]).

El plantejament de l'*FDA* d'expansió polinomial i selecció canònica es retrobarà a les conclusions quan se suggereixen possibles ampliacions de la proposta que fonamenta aquest treball (pàgina 116).

1.3.5 La relaxació de la hipòtesi unimodal: l'*MDA* (*Mixture Discriminant Analysis*)

Una altra manera de, partint de l'*LDA*, relaxar la seva suposició que les X_j es distribueixen amb $\sum_{i=1}^{g} p(i)\phi(u;\mu_i,\Sigma)$ és la d'eliminar la unimodalitat per classe, el que significa que la distribució serà:

$$\sum_{i=1}^{g}\sum_{s=1}^{c_i} p_i(s)\phi(u;\mu_{is},\Sigma) \quad \text{amb} \quad \sum_{i=1}^{g}\sum_{s=1}^{c_i} p_i(s) = 1$$

on c_i és el nombre de subclasses de la classe i.

L'algorisme que es farà servir per l'ajustament dels paràmetres és de tipus *EM* amb les següents etapes:

1. Partim d'uns valors inicials dels paramètres $p(s),\ \mu_{is}\ ,\Sigma$, obtinguts per un procediment de conglomerats dins de cada classe com, per exemple, el *k-means*.

2. *Etapa E*: Per a cada observació t estimem les aportacions a una subclasse s dins d'una classe i com:

$$W_{ts}(i) = \frac{p_i(s)\phi(u_t;\mu_{is},\Sigma)}{\sum_{s=1}^{c_i} p_i(s)\phi(u_t;\mu_{is},\Sigma)}$$

   on ϕ representa, com és habitual, la funció de densitat normal.

3. *Etapa M*: Utilitzant els valors W fem una anàlisi canònica d'$\hat{Y}'Y$ i estimem per màxima versemblança els paràmetres $p_i(s),\ \mu_{is},\ \Sigma$.

4. Tornem a 2 iterant fins que es produeixi una diferència menor que un valor fixat com a "tolerància" amb una determinada mètrica definida sobre l'espai dels paràmetres.

La gran aportació de l'*MDA* és que al relaxar la hipòtesi d'unimodalitat, amplia considerablement el camp d'aplicació del model sense perdre els trets fonamentals

de l'*LDA*. De fet, tornarem a aquest model quan discutim, a la pàgina 116, les possibilitats d'ampliar el camp d'actuació de la nostra proposta, i emprarem la seva estratègia *EM* com el referent bàsic de la suavització proposada al capítol 4.

1.3.6 Altres mètodes d'anàlisi discriminant

Encara que ja han estat comentats als apartats anteriors els mètodes més directament implicats amb l'objectiu definit d'aquest estudi, acabarem aquest capítol amb la revisió d'altres que formen part de l' ampli ventall desplegat als darrers anys per tal de cercar possibles idees complementàries. La conclusió, que ja avancem, es que no hi ha cap mètode que tingui un avantatge definitiu sobre la resta, però tots poden aportar elements a tenir en compte per fer un refinament que perfeccioni els resultats per a unes dades concretes.

És a dir, com el mètode que proposarem al capítol 4 pressuposa un model que, com qualsevol altre, només pot complir-se aproximadament en una situació pràctica real, cal tenir pressent el recull que ara descriurem breument per provar d'encetar una segona fase que utilitzant alguna o varies d'aquestes propostes pugui millorar els resultats.

1.3.6.1 La discriminació taxonòmica: els arbres

Una idea molt natural per aplicar a la classificació és la que va fer servir el naturalista Linneo: un conjunt de regles de decisió encadenades les quals poden estructurar-se formant un arbre, de manera que, a cada passa avancem des d'una branca més grossa a una de les més petites que sorgeixen d'ella fins a arribar a un dels extrems, el qual ja ens dóna la classe corresponent.

Aquesta metodologia àmpliament desenvolupada (Ripley, 1996 [180]) per la seva interpretabilitat, té els inconvenients d'una gran inestabilitat, falta de suavitat i dificultat per detectar la possible estructura additiva de les variables, si aquesta existeix.

Per combatre la inestabilitat s'apliquen les tècniques anomenades de *bagging* inspirades en el *bootstrap*, però la disminució de la variància porta a la pèrdua de la interpretabilitat.

En quant a la falta de suavitat, s'ha posat a punt el mètode conegut com a *Mixtura d'experts*, el qual substitueix la decisió esquerpa dels arbres per una de format probabilístic provinent d'una regressió amb estructura de mixtura (mitjançant l'*EM*). Aquest mètode, però, només té aplicabilitat en situacions precises (Jordan i Jacobs, 1994 [131]).

Respecte a l'estructura aditiva, el mètode *MARS* (*Multiple Adaptive Regression Splines*) utilitza un model lineal de selecció entre *funcions base* construïdes com a frontisses als punts d'aprenentatge. Es tracta, per tant, d'un parent de l'*FDA*. Al mateix temps, pot considerar-se una variant dels arbres, ja que aquestos surten com a casos particulars quan es prenen variables indicadores com a frontisses. L'inconvenient és que el problema es resol a canvi d'un augment considerable de la complexitat.

1.3.6.2 Un anàlisi discriminant que aprèn dels seus errors: el *boosting*

La idea del *boosting* original circula des de 1996, any en què Freund i Schapire [63] varen proposar el *Adaboost* com a mètode de classificació. La idea consisteix a millorar successivament un classificador feble però flexible combinant el resultat de totes les passes, però ponderant més alt en cada etapa les observacions mal classificades de l'anterior.

El resultat matemàtic més interessant és que *Adaboost* és equivalent, quan s'utilitzen dues classes retolades com $-1, 1$, a utilitzar una pèrdua $\exp(-\hat{Y}'Y)$ la qual, al seu torn, equival poblacionalment a la de l'entropia creuada (Kullback-Leibler).

El classificador feble que se sol utilitzar és el d'arbres, donant lloc al mètode conegut com *MART* (*Multiple Additive Regression Trees*), el qual utilitza com a

paràmetres de "sintonització" el nombre d'interaccions i el nombre de branques. Respecte al primer s'empra un factor de regularització on s'ha substituït la penalització mínimo-quadràtica de la *ridge* per una de valor absolut, que es coneix com a *lasso*, que te l'avantatge d'enviar cap a zero els coeficients no significatius, fent una mena de selecció de variables molt més interpretable. En quant al nombre de branques, una sèrie de consideracions heurístiques, aconsellen establir-la en l'interval 4-8, pel que es pren 6 com a punt de partida.

L'inconvenient del *boosting* és que si existeix un alt nivell de *soroll*, és a dir, si les variables X no aconsegueixen un elevat nivell de predicció de les Y els resultats es poden degradar ràpidament evidenciant la falta de robustesa del mètode.

1.3.6.3 La sinapsis com a inspiradora: les xarxes neurals

Si, inspirant-se en la sinapsis entre neurones, considerem que entre les X i la Y existeixen unes variables intermèdies que es connecten amb les X com a funcions sigmoidees de combinacions lineals (intentant reproduir l'activació per umbralització de la connexió neural) i amb les Y de nou en base a combinacions lineals, tindrem un procés lineal-no lineal-lineal que encaixa bé amb els mètodes coneguts com de *projection pursuit* (Friedman, 1987 [64]).

Tenint en compte que les funcions sigmoidees es poden aproximar per rectes a la zona d'activació prenent un adequat punt d'inici, les xarxes neurals poden considerar-se un model de regressió que evolucionarà "convolucionant" amb les dades fins que un determinat paràmetre de regularització l'indique que ha d'aturar-se.

Aquesta evolució, i d'aquí ve l'impuls que han tingut recentment aquests models, es pot realitzar mitjançant un ajust anomenat propagació cap enrere, que consisteix a prendre els errors finals i transmetre'ls cap a les variables intermèdies a la fi d'obtenir els gradients necessaris per al mètode d'optimització.

L'inconvenient principal de les xarxes neurals és la seva dificultat de "sintonització" donat que cal ajustar els valors inicials (els mínims són locals), la regularització, l'escala de les X i el nombre de variables intermèdies.

Ara bé, combinant aquest mètode amb la logística per al cas de variables discretes, s'han obtingut bons resultats, tal i com s'esmentava a la secció 1.3.3.

1.3.6.4 Els hiperplans separadors: *SVM* (*Support Vector Machines*)

Malgrat el que hem anomenat robustesa de l'*LDA* per indicar la seva estabilitat front a alteracions de la suposició bàsica de normalitat, és a dir el que en termes de la Teoria de la robustesa s'anomena robustesa qualitativa, l'*LDA* té dificultats a l'anomenada robustesa quantitativa o més precisament a la capacitat de no ser alterat greument per *outliers*. Per evitar això, es van desenvolupar mètodes que cerquen directament els hiperplans separadors de classes, entre els que cal destacar el *perceptron* de Rosenblatt (1958) [182], avantpassat de les xarxes neurals que acabem de comentar.

Existeix, però, una manera diferent d'abordatge del problema: Si tenim dues classes codificades com $-1, 1$ per a un punt x_s determinat, el producte $\hat{Y}_s' Y_s$ serà positiu si el punt està ben classificat i negatiu en cas contrari. A més es pot considerar que el valor absolut serà més gran quan més allunyat estigui de la frontera i que, per tant, el millor hiperplà separador és el que aconsegueix maximitzar $\sum_{s=1}^{n} \hat{Y}_s' Y_s$.

Tanmateix si les classes no són completament separables podem establir per cada punt un marge de desplaçament cap a l'altre costat, de forma que el total d'aquestos percentatges de desplaçament estigui acotat.

D'aquesta manera, com l'objectiu continua sent maximitzar la separabilitat calculada com suma de les distàncies a la frontera, afegint a cadascuna un marge de tolerància globalment controlat, els punts que per estar lluny de la frontera no necessiten d'aquesta tolerància, influeixen menys en la determinació de l'hiperplà.

Aquesta és la base del mètode discriminant conegut com *SVM* (*Support Vector Machines* o Discriminació amb punts de suport), el qual pot ser considerat com una aplicació de la penalització directa de l'error de classificació (mitjançant la suma de les toleràncies esmentades), la més lògica i directa que podem tenir en compte. A més, aquest mètode utilitza també la *Kernel property* esmentada a la secció 1.3.4 (pàg. 22), el que el connecta amb tot el conjunt dels que utilitzen suavitzacions tipus *Kernel* (que serà extensament descrita al capítol 3) obtenint un sòlid suport matemàtic.

L'inconvenient del *SVM* és que, malgrat el que podria pensar-se, no és un mètode "resistent" a l'augment de la dimensionalitat i a més pot distorsionar-se considerablement si s'introdueixen variables no significatives per a la classificació.

1.3.6.5 Els veïns millorats: *DANN* (*Discriminant Adaptive Nearest Neighboors*)

Finalment, acabarem la nostra revisió tornant al començament: Hastie i Tibshirani (1996) [117] van fer una millora del mètode dels *k-veïns* que va en la línia de superar el problema esmentat a la secció 1.3.1 (pàg. 18), deixant que la definició de l'entorn variï localment.

Es tracta de seleccionar en cada punt les direccions més discriminants amb una anàlisi canònica, on la matriu Σ està sotmesa a una penalització tipus *ridge* (una ponderació entre ella i la identitat). Prèviament es proposa fer una determinació global d'eixos significatius per eliminar variables no significatives o de *soroll*. La dificultat torna a ser la falta d'interpretabilitat que complica la selecció adequada del nivell correcte de localització.

Capítol 2

Anàlisi de Correspondències

El recorregut pels diversos mètodes d'anàlisi discriminant que es podrien aplicar a la situació en estudi (1.1), ens ha portat a la conclusió que és necessari dissenyar una metodologia específica que s'adapti a les seves característiques de discretització d'una multinormal subjacent, donat que els mètodes que parteixen de la normal: *LDA*, *QDA*, *FDA*, *MDA* no tenen en compte la discretització posterior, i els que s'adaptarien a la situació discreta com la logística , els arbres o els *k-veïns* (*DANN*) ignoren la Normal subjacent. Finalment, els que són més versàtils tenen o una complexitat alta o una interpretabilitat baixa o ambdues coses al mateix temps.

La idea que desenvoluparem en aquest treball consisteix a explorar les possibilitats de generalització multidimensional del resultat que va demostrar Lancaster per a l'anàlisi de correspondències simples entès com una correlació canònica (Lancaster, 1957 [139]), consistent en que aquest podria interpretar-se com l'aproximació a una binormal subjacent.

D'aconseguir això, tindríem un camí per retrobar la multinormal subjacent en el nostre cas i aplicar posteriorment els mètodes que, inspirats amb l'*LDA*, treballen amb prou eficiència en aquestes situacions.

Necessitem, per tant, revisar àmpliament l'anàlisi de correspondències per tal de fonamentar els resultats que s'utilitzaran al capítol 4 on presentarem el mètode proposat. A aquest objectiu està consagrat el present capítol.

2.1 La dualitat individu-variable

Començarem per aprofundir en un concepte que està per sota de tot el plantejament de l'anàlisi de correspondències: la dualitat individu-variable.

Una situació real plasmada en una matriu d'individus (fileres) per variables (columnes) permet una doble relació, de manera que les variables (característiques) expliquen als individus (tota dualitat té un costat més obvi) però també els individus expliquen les variables.

Paga la pena analitzar una mica aquesta segona part, perquè significa la substitució del caràcter "essencialista" de les variables i dels individus per a donar pas a una concepció molt més relativista: en el nostre context, una variable pot veure's com un conjunt de dades preses (se suposa que de manera semblant) als individus de la nostra mostra i pot considerar-se, per tant, una mena de **meta-individu** que té per elements els valors que pren en cadascun dels individus. De manera simètrica els individus poden veure's com a portadors de les variables (**meta-variables**).

2.1.1 El producte escalar d'individus i variables

Repassem ara, la formulació matemàtica d'aquest plantejament:

L'aplicació matemàtica més coneguda d'una idea tan poderosa com la de dualitat, és la dels espais duals de l'àlgebra lineal, on les formes lineals (funcions amb valors a $\mathbb{R}$) de l'espai $p-$dimensional, $\mathbb{R}_p$, conformen un espai $\mathbb{R}_p^*$ de la mateixa dimensió que aquest. Si e representa una base de $\mathbb{R}_p$ i e^* una altra de $\mathbb{R}_p^*$ les podem configurar com a duals mitjançant la senzilla definició $e_i^*(e_j) = \delta_{ij}$.

Això permet definir el conegut producte escalar entre $v^* \in \mathbb{R}_p^*$ i $w \in \mathbb{R}_p$ com a:

$$< v^*, w > = < \sum_i v_i e_i^*, \sum_j w_j e_j > = v^*(w) = \sum_i \sum_j v_i w_j e_i^*(e_j) \; =$$

$$= \sum_i \sum_j v_i w_j \delta_{ij} \; = \; \sum_i v_i w_i \; = \; v'w$$

donant a aquest producte escalar una altra visió de l'originària de la física (aplicació d'una força en una direcció obliqua) al enfocar-lo com a combinació lineal simètrica entre membres de dos espais duals (la simetria és essencial a la dualitat).

Al seu torn, aquest plantejament il·lumina la interpretació original, al fer-nos reflexionar sobre la simetria inherent al fet que l'aprofitament d'una força que te una direcció diferent de la restringida pel moviment, seria la mateixa si ambdues direccions s'intercanviessin.

De manera semblant, la dualitat ens aportarà una simetria a la relació individus-variables, que no era obvia quan pensàvem a aquestes només com característiques d'aquells.

És a dir, una variable pot veure's, també, com una forma lineal (un membre de $\mathbb{R}_p^*$) de forma que el producte escalar (entre individus i variables, ja que en aquest context tenen la mateixa dimensió) representaria o bé una combinació lineal de variables aplicada sobre un individu o bé una variable aplicada sobre una combinació lineal d'individus (això té sentit després de l'explicat anteriorment).

Aquest producte escalar, de gran importància ja que representa la relació entre un espai i el seu dual, pot enunciar-se com una combinació lineal de variables base aplicada a una combinació lineal d'individus base (una variable genèrica sobre un individu genèric expressats sobre bases duals).

2.1.2 Les transferències entre espais segons l'esquema dual

De l'apartat anterior podem extractar que una matriu de dades X, determina la consideració de quatre espais vectorials duals dos a dos:

$\mathbb{R}_p$ = Individus meta-variables $\quad \mathbb{R}_n^*$ = Individus essencials
$\mathbb{R}_p^*$ = Variables essencials $\quad \mathbb{R}_n$ = Variables meta-individus

Analitzarem ara les relacions tant horitzontals com verticals entre tots quatres espais.

2.1.2.1 La transferència horitzontal mitjançant X

Començarem per recordar que, en el nostre context, emmarcat per X, una variable pot abordar-se des de dues perspectives: com a generada a partir de p variables essencials ($\mathbb{R}_p^*$) o com a definida per la descripció dels seus valors en els n individus bàsics (meta-individus, $\mathbb{R}_n$). Naturalment ambdues concepcions estan estrictament relacionades mitjançant la matriu X, ja que si anomenem c_j^* a la seva columna j entesa com a *variable essencial* i f_i^* al seu individu i entès, amb forma simètrica, com a *individu essencial* tenim

$$c_j^*(f_i^*) \;=\; x_{ij}$$

i prenent c^*, f^* com a base de variables i individus essencials:

$c_j^*(f^*) \;=\;$ columna j d'X = variable j entesa com a meta-individu.
$f_i^*(c^*) \;=\;$ filera i d'X = individu i entès com a meta-variable.

Com a conseqüència:

Propietat 2.1 (Transferència horitzontal) *Les* variables essencials *prenent forma de* meta-individus *mitjançant les columnes de X, ja que si una variable té λ per vector de coordenades dins del'espai de variables essencials ($\mathbb{R}_p$), la seva realització en els individus de la nostra mostra i per tant la seva expressió coma meta-individu serà $X\lambda$ i, per altra banda, els* individus essencials *prendran forma de* meta-variables *a través de les fileres de X (o les columnes de X'), donat*

que si μ és el seu vector de coordenades a l'espai $\mathbb{R}_n$, la seva expressió com a meta-variable *serà $X'\mu$.*

Això ho podem reflectir mitjançant l'esquema de la figura (2.1).

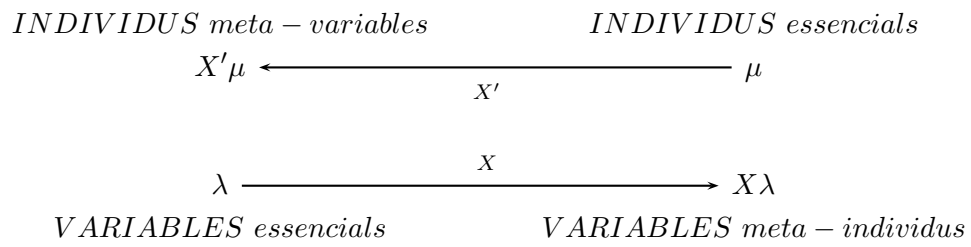

Figura 2.1: *Esquema de la transferència horitzontal*

2.1.2.2 La transferència vertical mitjançant la inversa

Per relacionar verticalment els espais considerats, podem observar que la matriu del producte exterior $M_{ij} =< e_i^*, e_j >= e_i^*(e_j)$ reflecteix la relació entre les bases dels dos espais duals, de forma que si aquestes són duals, M és converteix en la identitat, el que pot interpretar-se com una mena d'ortonormalitat conjugada (si una de les dues bases ho és, ho serà la dual). Tanmateix, si M és diferent de la identitat (és a dir les bases ja no són duals), poden retrobar aquesta conjugació mitjançant un canvi de variable en un dels dos espais, ja que si:

$$< v^*, w >_M = v'Mw$$

la veritable base dual d'e^* serà $\hat{e}_j = M^{-1}e_j$ donat que

$$< e_i^*, \hat{e}_j >_M \; = \; e_i{}'M\hat{e}_j = e_i{}'MM^{-1}e_j =< e_i^*, e_j >= \; \delta_{ij}$$

Per tant, si tenim una mètrica M a un espai de la nostra consideració i aquesta la prenem també com a producte escalar amb el seu dual, poden trobar una base

dual mitjançant el canvi de variable abans mencionat de matriu M^{-1}, on si a representa les coordenades a la base e_j i $\hat{a}$ les coordenades a la base $\hat{e}_j$ tenim que:

$$\hat{e}_j = M^{-1}e_j \Rightarrow \hat{a}'\hat{e}_j = \hat{a}'M^{-1}e_j = a'e_j \Rightarrow a = M^{-1}\hat{a} \Rightarrow \hat{a} = Ma$$

i la mètrica conjugada que obtindrem al dual serà:

$$\|a\|_M = a'Ma = a'\hat{a} = \hat{a}M^{-1}\hat{a} = \|\hat{a}\|_{M^{-1}}$$

D'aquesta manera, arribem al resultat que cercàvem, el qual utilitzarem conjuntament amb les transferències mitjançant X, amb tots els esquemes de l'anàlisi de correspondències que es desenvoluparan en aquest capítol:

Propietat 2.2 (Transferència vertical) *Una mètrica amb matriu M a un espai indueix al seu dual la mètrica de matriu M^{-1} sobre la base dual (definida utilitzant M com a matriu del producte escalar).*

El que pot esquematitzar-se tal i com es veu a la figura (2.2).

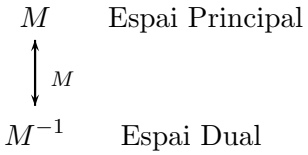

Figura 2.2: *Esquema de la transferència vertical*

2.2 El triplet bàsic de l'anàlisi de components principals

L'enfocament anterior ens permet expressar l'anàlisi de correspondències mitjançant un esquema aclaridor de les relacions entre els quatre espais involucrats: el de les variables essencials, el de les variables enteses com a meta-individus, el dels individus essencials i el dels individus entesos com a meta-variables.

Començarem per fer l'esquema d'un *ACP* (*Anàlisi de Components Principals*), donat que com després veurem, l'anàlisi de correspondències no és més que un conjunt d'*ACP* relacionats.

L'article clau en aquest plantejament va ser el de Tenenhaus i Young (1985) [207], pel que relacionarem la nomenclatura que els adopten amb els conceptes tal i com s'han definit aquí.

Els esmentats autors prenen com a nomenclatura principal la que hem anomenat *meta* és a dir el que ells anomenen variables són aquestes enteses com a meta-individus (realitzacions concretes plasmades a les columnes de X, també anomenats *eixos*) i el que anomenen individus són els entesos com a meta-variables (*components*).

Al que hem anomenat essencial al paràgraf anterior l'anomenen *coeficients* (de variables o d'individus) o bé *factors* per les variables essencials i *cofactors* pels individus essencials.

També és de notar que, als seus esquemes, no utilitzen l'asterisc per identificar els duals però aquí els inclourem per evitar confusions.

Amb aquests aclariments, el seu esquema de l'ACP col·loca a cada vèrtex d'un quadrat un dels quatre espais esmentats amb la seva identificació i la mètrica principal. També relaciona horitzontalment mitjançant X, X' (propietat 2.1) i verticalment segons la inversa (propietat 2.2).

L'enfocament general amb aquest esquema serà que a cada vèrtex tenim la mètrica original i la induïda pel recorregut arreu els quatre costats del quadre (figura 2.3).

Ara bé, com sempre que tenim dues mètriques a un mateix espai, el nostre interès serà el de "harmonitzar-les" (com s'explicarà a la secció 2.3.2, pàg. 45) mitjançant una *ortogonalització conjugada progressiva*, trobant les direccions (vectors propis) que siguin invariables per a les dues mètriques i que, considerant com a producte escalar el que té per matriu la d'una de les dues mètriques,

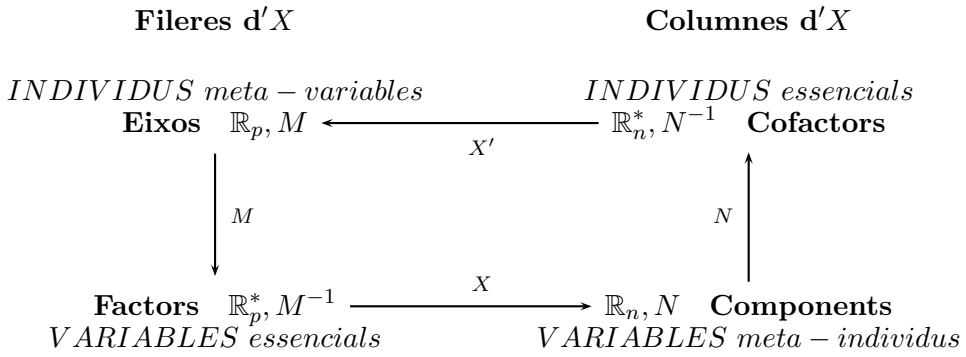

Figura 2.3: *Esquema de l'ACP segons Tenenhaus i Young*

es maximitze successivament la norma respecte a l'altra, sota la ben coneguda restricció d'ortogonalitat entre vectors propis.

Detallem una mica aquest procés:

Siguin T, S les matrius definides positives corresponents a les dues mètriques que pretenem harmonitzar. El nostre objectiu serà trobar uns vectors propis que, estructurats com a columnes de la matriu Q, compleixen:

$$Q'TQ = \mathbf{I} \quad \text{i} \quad Q'SQ = D$$

Amb $\mathbf{I}$ matriu identitat i D matriu diagonal.

Es ben sabut que la solució consisteix a trobar la matriu V dels vectors propis de $T^{-1/2}ST^{-1/2}$ i fer $Q = UT^{-1/2}$ i que si ordenem D (matriu de valors propis) de major a menor, Q_1 serà el vector de màxima $S-norma$ (amb 1 com a $T-norma$), Q_2 farà el mateix però sobre l'espai ortogonal a Q_1 i així successivament seguint a Gram-Schmidt.

Aquest procés l'anomenarem sintèticament diagonalitzar $S \circledast T^{-1}$ i el podem formalitzar de la següent manera:

Definició 2.1 *Siguin T, S matrius quadrades definides positives que representen l'expressió de dues mètriques en una base d'un espai vectorial, diagonalitzar*

$S \circledast T^{-1}$ consisteix a trobar la matriu Q tal que $Q'SQ$ sigui diagonal i $Q'TQ$ la matriu identitat de manera que :

$$Q_i'SQ_i = \max\left\{v'Sv \ \ \forall v / v'SQ_j = 0 \qquad j = 1 \ldots (i-1)\right\}$$

on, Q_i és la columna i-sima de Q i, per simplicitat notacional, considerarem $Q_0 = \mathbf{0}$

Amb aquesta definició es compleix la següent proposició:

Proposició 2.1 *Diagonalitzar $S \circledast T^{-1}$ és equivalent a fer la reducció canònica de la matriu simètrica $T^{-1/2}ST^{-1/2}$ i prendre $Q = UT^{-1/2}$ sent U la matriu de vectors propis obtinguda.*

No hem d'oblidar que això està ben definit pel fet que, al ser T una matriu que expressa una mètrica és definida positiva i, per tant, podem assegurar l'existència de $T^{-1/2}$.

S'ha de tenir en compte, també, que al llarg de la resta de l'estudi considerarem que, una vegada fixades les bases canòniques als espais essencials, quan es referim a una determinada mètrica ho fem, equivalentment, a la matriu que la representa a la base fixada a l'espai corresponent.

Finalment, tot el procés seguit en aquesta secció es pot esquematitzar mitjançant la taula de la figura 2.4.

En forma resumida l'anàlisi de components principals se simbolitzarà amb el $triplet\big(X, M, N\big)$ on:
X **= matriu de dades**
M **= mètrica per als individus(meta-variables)**
N **= mètrica per a les variables (meta-individus)**

	Matriu T de la mètrica principal transferida del dual	Matriu S de la mètrica obtinguda pel recorregut pels altres vèrtexs	Matriu $S \circledast T^{-1}$ a diagonalitzar
Fileres	N^{-1}	XMX'	$XMX' \circledast N$
Columnes	M^{-1}	$X'NX$	$X'NX \circledast M$

Figura 2.4: *Esquema de les diagonalitzacions de l'anàlisi de components principals*

2.3 Els triplets equivalents de l'anàlisi de correspondències simples

A continuació veurem que l'anàlisi de correspondències simples, i totes les seves bones propietats de simetria fileres-columnes, pot enfocar-se com un conjunt de quatre ACP equivalents, representats cadascun mitjançant el corresponent triplet, tal i com acabem d'analitzar a l'apartat anterior. En definitiva, es tracta de situar-nos en cadascun dels quatre vèrtex de l'esquema que acabem de presentar, adaptant la nomenclatura a l'específica de la situació discreta que tracta l'anàlisi de correspondències simples.

Considerarem com a dades la matriu F de freqüències absolutes i les seves diagonals marginals de fileres D_f i de columnes D_c

El triplet inicial que anomenarem **doble-inercial** del correspondències simple és:

$$\left(D_f^{-1} F D_c^{-1} - \mathbf{1}_{f,c},\ D_c,\ D_f\right) \tag{2.1}$$

on veiem la X de l'ACP obtinguda a partir de "normalitzar" F dividint tant pel total de fileres com el de columnes, i després restant una matriu d'uns per tenir les desviacions χ^2.

Les mètriques naturals son: la de les ponderacions per sumes de columnes per a les fileres (D_c) i la de la ponderació per la suma de fileres per a les columnes (D_f). Aquest plantejament resulta completament simètric i ens porta al conegut doble anàlisi d'inèrcia, el qual segons l'enfocament de la taula de la figura 2.4 ens portarà a diagonalitzar:

$$(D_f^{-1}FD_c^{-1} - \mathbf{1}_{f,c})D_c(D_f^{-1}FD_c^{-1} - \mathbf{1}_{f,c})' \circledast D_f$$

o, equivalentment:

$$(D_f^{-1}FD_c^{-1} - \mathbf{1}_{f,c})'D_f(D_f^{-1}FD_c^{-1} - \mathbf{1}_{f,c}) \circledast D_c$$

Donada la simetria i centrament d'aquesta anàlisi es pot interpretar com la cerca dels eixos que acumulen ortogonalment més inèrcia global, el que justifica el nom de doble inercial que hem donat a aquest triplet.

També podem arribar a la mateixa matriu a diagonalitzar (i per tant es tracta d'un procés equivalent) si invertim les dues mètriques passant als duals (fileres i columnes *essencials*) i analitzem la taula d'observats menys esperats:

$$\left(F - D_f\mathbf{1}_{f,c}D_c \ , \ D_c^{-1} \ , \ D_f^{-1}\right) \tag{2.2}$$

Aquest triplet el podem anomenar **doble-discriminant**, donat que si considerem que fileres i columnes d'un correspondències procedeixen d'agrupacions dels individus sota classes determinades per les categories de cada una de les dues variables que es creuen, podem interpretar que un correspondències simple utilitza les files per discriminar sobre columnes i viceversa, el que encaixa perfectament amb la selecció dels espais essencials (mètriques inverses) que es fa en aquest plantejament.

Ara bé, si entenem que no és molt lògica la doble discriminació en un context real, sinó que és més raonable d'afavorir la interpretació d'una de les dues variables en funció de l'altra, trencarem la simetria que inspira la χ^2 (mantenint la equivalència de la matriu a diagonalitzar) fent dos nous enfocaments:

- El primer seria una mena de discriminant de les columnes sobre les fileres, triplet que anomenarem de **perfil de fileres** :

$$\left(D_f^{-1}F - \mathbf{1}_{f,c}Dc\ ,\ D_c^{-1}\ ,\ D_f\right) \tag{2.3}$$

  on la matriu que ara analitzem és la de les desviacions de la distribució condicional de les fileres a la marginal, i la mètrica que ara apliquem a les fileres (en principi enteses com a *meta-columnes*) és la inversa, que correspon a l'espai dual (*columnes essencials*), ja que ara no cerquem una anàlisi simètrica, sinó trobar els eixos principals que ens expliquem les direccions fonamentals que prenen els perfils de les fileres. Per això, quan comparem fileres hem de fer-ho procurant que la "mida" de les columnes no ens afecte, pel que hem de ponderar aquestes per l'invers de la seva mida.

- I el segon (discriminant de les fileres sobre les columnes) seria, simètricament, el triplet de **perfil de columnes**:

$$\left(FD_c^{-1} - D_f\mathbf{1}_{f,c}\,,\, D_c\,,\, D_f^{-1}\right) \tag{2.4}$$

2.3.1 L'aproximació dels polinomis de l'Hermite

Per poder completar l'estudi de l'anàlisi de correspondències simples, ens cal presentar l'esmentat teorema de Lancaster (1957) [139] el qual desvetlla el factor normal subjacent als mètodes canònics.

Efectivament, tots el mètodes canònics poden veure's com la maximització d'alguna correlació, però hem de tenir en compte que maximitzar una correlació és aproximar una binormal en el sentit següent:

Teorema 2.1 (Lancaster) *Suposem que X_1 resulte d'una transformació d'una variable Normal tipificada Z_1 i X_2 de forma semblant de Z_2, aleshores:*

$$corr(X_1, X_2) < |corr(Z_1, Z_2)| = |\rho|$$

La importància d'aquest teorema és que es pot interpretar en el sentit que qualsevol parella de transformacions Y_1, Y_2 de les X_1, X_2, que maximitze la correlació, dins d'un determinat conjunt de transformacions possibles (T), la podem considerar un intent d'aproximar les Normals Z_1, Z_2 dins de T, en el benentès que si T fóra el conjunt de totes les transformacions possibles, resultaria: $Y_i = Z_i + constant$.

Demostració Qualsevol transformació tipificada $x_1 = \frac{X_1 - \mathrm{E}(X_1)}{\sqrt{\mathrm{Var}(X_1)}}$ pot veure's com $x_1 = \sum a_\alpha Z_1^\alpha$ amb $\sum a_\alpha^2 = 1$ on per simplicitat hem anomenat Z^α al polinomi ortogonal de l'Hermite estandarditzat de grau α (coeficient de t en $\exp(tz - t^2/2)$, Kendall i Stuart (1977) [133] (vol.II, pàg. 600).

De la mateixa forma $x_2 = \sum b_\alpha Z_2^\alpha$ amb $\sum b_\alpha^2 = 1$ resultant que si tenim en compte que:

$$\mathrm{E}\big(\exp(tz) - \frac{t^2}{2} + uz_2 - \frac{u^2}{2}\big) = \exp(\rho t u)$$

s'obté, per desenvolupament en sèrie en les variables t, u :

$$\mathrm{E}(Z_1^\alpha Z_2^\beta) = \rho^\alpha \delta_{\alpha\beta}$$

i ,per tant:

$$\mathrm{corr}(X_1, X_2) = \mathrm{corr}(x_1, x_2) = \sum\nolimits_\alpha a_\alpha b_\alpha \rho^\alpha < |\rho| = |\mathrm{corr}(Z_1, Z_2)|$$

∎

Una segona parella de transformacions ortogonals a la primera farà la seva correlació inferior a ρ^2 i successivament tindrem una cota de $|\rho|^3$,$\rho^4 \ldots$.

De fet quan fem una correlació canònica (tipus correspondències simples) podem considerar que si $\psi_{\alpha i}$ són les quantificacions de les fileres i $\zeta_{\alpha j}$ les de les columnes per al valor propi α, aquestes poden ser interpretades com les aproximacions tipificades de:

$$\varphi_{\alpha i} = \frac{\int_{S_i} x^\alpha \phi(x) \mathrm{d}x}{\int_{S_i} \phi(x) \mathrm{d}x} \tag{2.5}$$

$$\theta_{\alpha j} = \frac{\int_{S_j} y^\alpha \phi(y)\mathrm{d}y}{\int_{S_j} \phi(y)\mathrm{d}y}$$

on amb ϕ representarem, per simplicitat, la densitat Normal tipificada (tant uni com bivariable), amb x^α, y^α els polinomis de l'Hermite tipificats de grau α de x, y respectivament, i S_i, S_j recobriments d'intervals tals que:

$$\int_{S_i}\int_{S_j} \phi(x,y)\mathrm{d}x\mathrm{d}y = f_{ij} \quad \Rightarrow \quad \int_{S_i} \phi(x)\mathrm{d}x = f_{i\cdot} \quad , \quad \int_{S_j} \phi(y)\mathrm{d}y = f_{\cdot j}$$

A més sabem que:

$$\phi(x,y) = \phi(x)\phi(y)\big(1 + \sum_{\alpha=1}^{\infty} \rho^\alpha x^\alpha y^\alpha\big)$$

i per tant si la subjacent fos binormal:

$$f_{ij} = f_{i\cdot} f_{\cdot j}\big(1 + \sum_{\alpha=1}^{\infty} \rho^\alpha \varphi_{\alpha i}\theta_{\alpha j}\big)$$

Com per altra banda del resultat de correspondències tenim :

$$f_{ij} = f_{i\cdot} f_{\cdot j}\big(1 + \sum_{\alpha=1}^{r} \lambda_\alpha \psi_{\alpha i}\zeta_{\alpha j}\big)$$

resulta que λ_α pot considerar-se també una aproximació de ρ^α, de la mateixa manera que $\psi_{\alpha i}$ ho és de $\varphi_{\alpha i}$ i $\zeta_{\alpha j}$ de $\theta_{\alpha j}$.

El fet que estem aproximant una sèrie infinita, tipus Taylorià, per una que no ho és, ens indica fins quin punt la discretització que disposem ens limita el grau del polinomi ortogonal al que podem arribar.

Aquest resultat ens permet obtenir una visió de l'anàlisi de correspondències com un desenvolupament en sèrie finita aproximant l'infinit que correspondria a una binormal, el que té un gran interès per als nostres objectius reconstructors esmentats al començament del capítol, sempre que siguem capaços de generalitzarlos al cas multidimensional, el que farem al capítol 4.

2.3.2 Interpretació geomètrica del teorema de Lancaster

Per poder aprofundir en el teorema de Lancaster, que està en la base de la idea que es proposa com a nou mètode d'anàlisi discriminant discreta al capítol 4, procedirem ara, detalladament, a la seva interpretació geomètrica amb la consideració que les intuïcions d'aquest ordre, són en moltes ocasions, la base per a plantejaments on la seva expressió analítica pot amagar el sentit (Carbonell i altres, 1983, cap.7 [26]).

Començarem per analitzar el significat probabilístic de l'ortogonalitat, donat que la inspiració purament física de la representació cartesiana a la que es feia referència a la secció 2.1.1, quan comentàvem la reinterpretació del producte escalar en el context dual, ha de reconsiderar-se en un context aleatori.

Per veure-ho amb més claredat, suposem que a una variable aleatòria x amb distribució Normal estàndard se li afegeix una nova font de variació independent que anomenarem ε amb variància $\gamma < 1$. Imaginem que no ens és possible accedir directament a ε i que ho fem mitjançant la variable $y = rx + \varepsilon$.

Per visualitzar la transformació de l'ortogonalitat resultant de la falta d'observabilitat d'ε, farem el corresponent anàlisi canònic de la parella (x, y), el resultat del qual s'il·lustra a la figura (2.5).

A la figura (2.5) s'ha tingut en compte que:

$$\sigma_x^2 = 1 \qquad \sigma_y^2 = r^2 + \gamma$$

pel que, definint:

$$\Delta = \sqrt{(1 - r^2 - \gamma)^2 + 4r^2} = \sqrt{(1 + r^2 + \gamma)^2 - 4\gamma^2}$$

ens resulta:

$$\lambda_1 = \frac{\sigma_y^2 + \sigma_x^2 + \Delta}{2} = \frac{1 + r^2 + \gamma + \sqrt{(1 + r^2 + \gamma)^2 - 4\gamma^2}}{2}$$

$$\lambda_2 = \frac{\sigma_y^2 + \sigma_x^2 - \Delta}{2} = \frac{1 + r^2 + \gamma - \sqrt{(1 + r^2 + \gamma)^2 - 4\gamma^2}}{2}$$

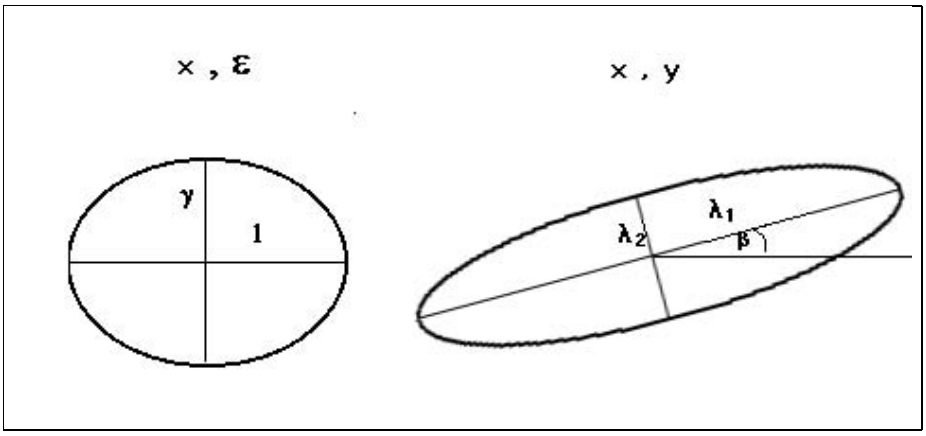

Figura 2.5: *Deformació introduïda per la falta d'observabilitat d'una font de variació.*

i

$$\cos(2\beta) = \frac{\sigma_y^2 - \sigma_x^2}{\Delta} = \frac{1 - r^2 - \gamma}{\sqrt{(1 - r^2 - \gamma)^2 + 4r^2}}$$

Això ens col·loca front a la reinterpretabilitat de l'ortogonalitat en aquest context, donat que si la considerem, com és habitual, associada a la independència en sentit probabilístic, no té gaire sentit procedir a una representació cartesiana de (x, y), perquè aquestes variables no són independents. Malgrat tot, i a falta d'altra idea millor, emprem, per comoditat, una representació on apareixen com a ortogonals.

Precisament, la importància de l'anàlisi canònica (anàlisi de components principals en aquest context) és que ens permet reconciliar els dos aspectes de l'ortogonalitat, el físic i el probabilístic, al trobar els eixos principals que són independents amb els dos sentits i que, per tant, poden representar-se ortogonalment sense cap mena de dubte. El resultat és una deformació de l'el·lipse de probabilitat tal i com hem vist a la figura (2.5).

Per tant el "gir" i "l'aplanament" d'aquesta el·lipse no és cap cosa intrínseca de la distribució binormal que analitzem, sinó que depèn de la correlació i de l'escala (en definitiva, de la matriu de covariàncies) de les variables observades dins del pla sobre el que es troben definides.

Com a il·lustració la figura (2.6) ens representa el sentit d'aquest "gir" i d'aquest "aplanament" en funció de $\rho = \text{corr}(x, y)$, prenent com a referència θ, l'angle a què tendeix β quan $\rho \to 1$ (per simplicitat suposem $\rho > 0$).

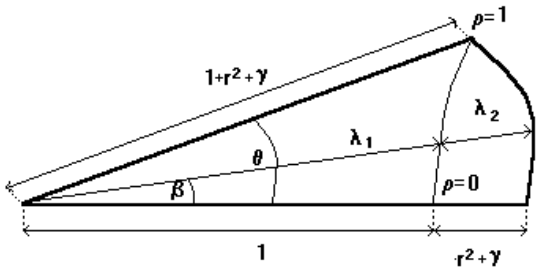

Figura 2.6: *Variació dels valors propis i de l'angle β segons ρ*

Considerem ara com a mesura de la deformació la χ^2 definida com:

$$\chi^2 = \int\int D^2\Big(\phi(x,y)\ ,\ \phi(x)\phi(y)\Big)\ \phi(x)\phi(y)dxdy$$

on s'ha considerat que la funció distància D^2 està definida per:

$$D^2\Big(o, e\Big) = \frac{(o-e)^2}{e}$$

L'avantatge d'aquesta formulació és que ens fa palesa la similitud de l'estructura de la χ^2 amb la de la variància, on el paper de la mitjana ho fa la distribució d'independència, i on els quadrats de les desviacions són expressats en unitats determinades per aquest model de referència.

Per tant, a l'anàlisi de correspondències, que utilitza la χ^2 com a base, les distancies a l'origen representen les contribucions de les categories a la dispersió

respecte a la independència, de la mateixa forma que qualsevol element d'una binormal ho fa respecte a la seva mitjana.

Amb això, i suprimint l'eix trivial resultant del centrament, podem interpretar el teorema de Lancaster en el sentit següent:

> *Cada cel·la de la taula de contingència pot considerar-se com una zona rectangular, estesa arreu de la seva quantificació, de forma que: la probabilitat sobre la binormal subjacent sigui la freqüència relativa corresponent, estimació de la corresponent probabilitat d'acord amb les fórmules 2.5 (pàg. 43).*

Un exemple gràfic el podem veure a la figura 2.7.

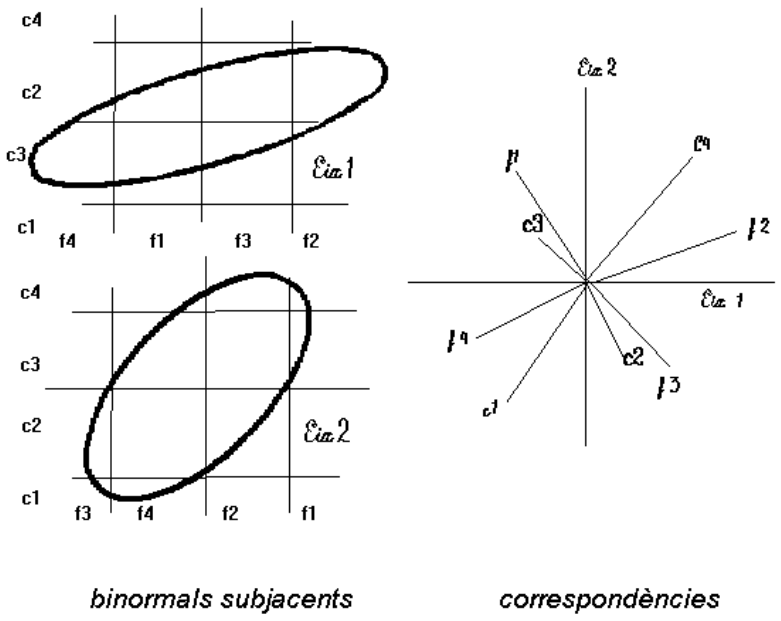

Figura 2.7: *Interpretació geomètrica del teorema de Lancaster*

A més, tenint en compte que els eixos canònics corresponen als polinomis ortogonals de l'Hermite, tal i com ens demostra Lancaster, cada un d'ells ens aporta un grau en l'aproximació successiva de la reconstrucció de la deformació de l'el·lipse introduïda per la dependència, partint dels valors mitjans d'aquests polinomis a les zones corresponents a cada cel·la.

El primer eix, per tant, es revela en aquesta anàlisi com el fonamental al ser el que a l'utilitzar el grau 1 ens proporciona la millor aproximació directa de la binormal, actuant els altres com a informació complementària per realitzar un millor ajustament progressiu.

És important ressaltar aquest paper del primer eix, donat que resultarà bàsic a l'hora d'interpretar la generalització del teorema de Lancaster (capítol 4), i hem de tenir present que al situar dins d'ell les coordenades de tots els centroïds de les cel·les, fem una mena de "col·lapsament" de les dimensions originals (veure figura 2.8) projectant alhora sobre la direcció que proporciona la màxima variància (entesa com a χ^2).

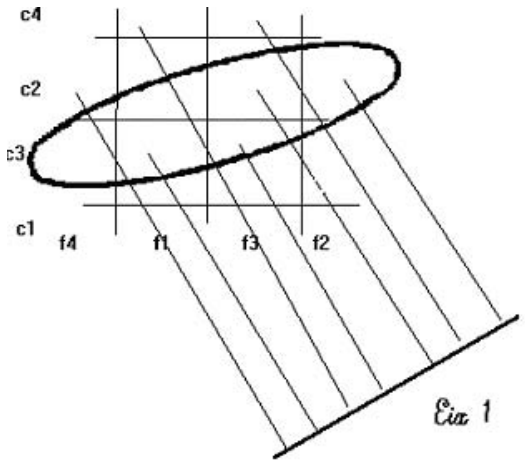

Figura 2.8: *Efecte de col·lapsament sobre l'eix principal*

2.4 Els triplets conjugats de l'anàlisi de correspondències múltiples

Naturalment, l'anàlisi de correspondències simples no ens serveix per a l'objectiu d'aproximar una mixtura de multinormals discretitzades, però el teorema de Lancaster ens estimula a la recerca d'algun resultat semblant per al cas multidimensional.

Hem de traslladar, per tant, l'anàlisi mitjançant triplets al cas múltiple, per veure si d'aquesta manera podem posar ordre dins de les múltiples variants de generalització del correspondències simples, amb el propòsit d'esbrinar quina és la més adient per a la nostra situació.

Seguint de nou a Tenenhaus i Young estudiarem els dos triplets corresponents als dos ACP conjugats següents:

- Triplet de l'ACP de **fileres**: $\left(X/p,\ npD^{-1},\ \mathbf{I}/n\right)$
- Triplet de l'ACP de **columnes**: $\left(D^{-1}X,\ n\mathbf{I},\ D/np\right)$

on D és la matriu diagonal amb les freqüències de totes les categories.

Observem que fent $F = X/np$, $D_c = D/np$ i $D_f = \mathrm{I}/n$ estem exactament (centrament a banda) amb els que hem anomenat triplets de *perfil de fileres* i de *perfil de columnes* a la secció 2.3 (pàg. 40) per l'anàlisi de correspondències simples.

La diferència essencial consisteix a que ara es consideren tant les fileres, com les columnes, com a variables estructurades en columnes (i així D_c passa a ser D/np), i es deixa per als individus la mètrica que abans tenien les fileres (i així D_f passa a ser $\mathbf{I}/n$) per indicar que tots els individus pesen igual.

La matriu bàsica (es repeteix dues vegades) a diagonalitzar serà, segons l'esquema de la figura (2.4), $XD^{-1}X'$ matriu $(n\times n)$ amb $D^{-1}\circledast X'X$ i $X'X\circledast D^{-1}$, matrius $(k\times k)$, com a relacionades.

Aquest plantejament permet emmarcar els diferents abordatges de l'anàlisi de correspondències múltiples segons quina de les tres matrius possibles es proposen diagonalitzar:

Nom del mètode	**Objectiu principal**	**Matrius a diagonalitzar**
Quantificació recíproca	Quantificació d'individus i de categories que respecte les relacions naturals entre les dues.	$XD^{-1}X'$ i $D^{-1} \circledast X'X$
Component Principals	Màxima variància dels individus essencials (ACP fileres)	$XD^{-1}X'$
Canònica Generalitzada	Màxima suma de correlacions al quadrat de les quantificacions dels individus amb les de les variables (ACP *columnes*)	$D^{-1} \circledast X'X$
Inèrcia	Màxima variància dels individus (meta-variables)	$X'X \circledast D^{-1}$

Figura 2.9: *Esquema de les diagonalitzacions de l'anàlisi de correspondències múltiples*

Capítol 3

Mètodes de suavització

Revisats els conceptes de l'anàlisi de correspondències i plantejada la dificultat fonamental a resoldre en la reconstrucció de la distribució continua subjacent, la solució de la qual resta aplaçada fins al capítol 4, analitzarem aquí l 'altre element que hi es combinarà per tal de completar l'esmentada reconstrucció: la suavització de les quantificacions obtingudes mitjançant l'anàlisi de correspondències.

3.1 La Suavització com a operació pseudoinversa de la discretització

Formalitzarem ara el concepte de discretització ja esbossat a la secció 1.2.1 (pàg. 6).

Sigui $\mathbb{R}_p$ l'espai subjacent que considerem i sigui $\mathcal{P}$ el conjunt de distribucions de probabilitat definida en ell.

Una discretització d serà una funció de $\mathbb{R}_p$ amb valors a $\mathbb{N}_p$ de forma que, en cada component, la inversa defineixi una partició dins l'espai original. Es a dir:

$$d_j^{-1}(r) = S_{jr} \quad \text{amb} \quad \mathbb{R}_p = \otimes_{j=1}^{p} \cup_{r=1}^{k_j} S_{jr}$$

Per tant per a cada $p \in \mathcal{P}$ tindrem p_d com a la distribució induida per p mitjançant d dins de $\mathbb{N}_p$ i al conjunt ho anomenarem $\mathcal{P}^d$.

Una suavització s_d serà una funció de l'espai $\mathcal{P}^d$ amb valors a $\mathcal{P}$.

Naturalment l'ideal seria que $s_d(p_d) = p \quad \forall p \in \mathcal{P}$, o al menys que:

$$[s_d(p_d)]_d = p_d \quad \forall p \in \mathcal{P} \tag{3.1}$$

però això és pràcticament impossible, donat que desconeixem p i, només en casos molt especials, podem disposar per mig de p_d de prou informació com per obtenir aquests tipus de resultats.

Per tant, rebaixem el nostre objectiu i considerem que disposem de una mostra de p_d que anomenarem p_d^n, la qual pot ser considerada com a distribució i, per tant, suavitzada i parametritzada. Imposarem la condició:

$$\lim_{n\to\infty} \theta(s_d(p_d^n)) = \theta(p) \quad \forall p_d^n \to p_d \tag{3.2}$$

que és el mateix que dir que, en definitiva, alló que ens interessa de p (el paràmetre θ) és aproximat asintóticament per la suavització proposada.

En aquest cas, anomenarem a s_d *θ-consistent* i la funció de l'espai de les discretitzacions en el de les suavitzacions (que dóna com imatge de d, s_d) rebrà el mateix qualificatiu, deixant que sigui el context el que diferenciï entre les dues, per evitar nomenclatures complexes. Finalment, si considerem una successió de discretitzacions d_m tal que:

$$m \leq m' \Rightarrow S_{jr}^m \supseteq S_{jr}^{m'}$$

i si $\mathcal{L}$ és la mesura de Lebesgue en $\mathbb{R}^p$ amb

$$\lim_{m\to\infty} \mathcal{L}(S_{jr}^m) = 0 \qquad \text{i} \qquad \lim_{m\to\infty} k_j^m = \infty$$

anomenarem a s_d *θ-dèbil consistent* sii:

$$\lim_{n\to\infty} \lim_{m\to\infty} \theta(s_{d^m}(p_d^n)) = \theta(p) \qquad \forall p_{d^m}^n \to p_{d^m}$$

Es clar que, en el nostre context discriminant qui fa el paper d'θ és el que hem anomenat error óptim continu e_c (fórmula 1.5), i que el requisit mínim a garantir per a qualsevol mètode de suavització en condicions de discriminació serà la e_c-dèbil consistència.

Hi ha dos mètodes de suavització del quals està demostrada aquesta condició sota hipotèsis molt més generals que les que hem fixat en aquest estudi: la estimació de densitats mitjançant *Kernel* per a un espai $\mathcal{P}$ genèric (Titterington, 1980 [208] i Hall, 1989 [97]) i el mètode EM basat en la màxima versemblança per al espai format per les mixtures de Normals (Tanner, 1989 [206]), raó per la qual seràn els analitzats al llarg de la resta del capítol.

Tanmateix no és la condició de la e_c-dèbil consistència la única questió a tenir en compte quan procedint a una suavització, donat que necessitem alguna mesura de suavitat que ens permeta controlar el procés, o dit més formalment cal asegurar-se també la ν-dèbil consistència sent ν la mesura esmentada.

3.2 Mesures de suavitat

La primera mesura de suavitat d'una funció que és natural plantejar està asociada directament a la segona derivada mitjançant la fórmula sintètica:

$$\nu_1(f) = \int (f'')^2$$

Per observar de forma ràpida com treballa ν_1, prenguem la família de densitats paràboliques:

$$f_a \propto \begin{cases} a^2 - x^2, & \text{si } -a \leq x \leq a \\ 0, & \text{a la resta} \end{cases} \quad \Rightarrow \quad \nu_1(f_a) = \frac{9}{2a^5}$$

es a dir que si a creix fent la paràbola més "aplanada" aleshores ν_1 disminueix.

També cal assenyalar, com a referència, que si f és Normal $\nu_1(f) = 0.53$.

Per altra banda, entre l'ampli ventall de possibles mesures, poden considerar una que té per als nostres objectius l'avantatge de que recull directament un patró de suavitat inspirat en la Normal:

$$\nu_2(f) = \int ((\log f)'')^2$$

la qual té la propietat de que val 0 sii la distribució és gaussiana.

Cal observar que ambdues mesures ho són en realitat de la falta de suavitat, ja que tendeixen des de valors possitius a 0 quan aquesta augmenta.

Els mètodes *Kernel* de nucli gaussià i *EM* de mixtures d'ajustament Normal (West, 1991 [224] i Dempster et all, 1977 [49]) que es desenvoluparan en aquest capítol són ν_1-consistents i ν_2-consistents.

3.3 La suavització *Kernel* i les seves propietats globals

Per començar l'estudi del estimadors *Kernel* definirem una amplia classe d'estimadors no paramètrics d'una funció de densitat probabilística mitjançant:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} w(X_i, x) \quad \text{amb} \quad w \geq 0 \quad \text{i} \quad \int w = 1$$

on X_i són els punts de la mostra d'aprenentatge i w una funció general d'estimació. El *Kernel* de finestra fixa h será un cas particular de w, amb:

$$w(X_i, x) = \frac{1}{h}K(\frac{x - X_i}{h}) \quad \text{amb} \quad K \geq 0 \quad \text{i} \quad \int K = 1$$

Altres exemples de w són sèries de Fourier com la esmentada al realitzar el desenvolupament en polinomis de l'Hermite (secciò 2.3.1, pàg. 42).

La primera consideració per a aquest conjunt d'estimadors prové del teorema de Rosenblatt (1956) [183], que demostra que els estimadors no paramètrics de la densitat són sempre esbiaixats.

Aquest resultat, aparentment advers, que pot explicar-se perquè les exigències de regularitat de la funció (Borel-mesurable) no la deixen "pegar-se" completament a la mostra, ens obliga a introduir les consideracions clàssiques de l'estudi de l'error quadràtic mitjà *ECMI* ($biaix^2 + variància$) per a l'anàlisi de la qualitat de l'estimació.

Afortunadament quan:

$$\lim_{n\to\infty} \int wf = f$$

podem (Prakasa-Rao, 1983 [171]) garantir la convergència en probabilitat $\hat{f} \to f$, els que en dona per als *Kernels* les propietats de consistència esmentades anteriorment.

3.4 La selecció de la funció nucli i l'ajustament de la finestra fixa del *Kernel* unidimensional

Una vegada assegurada la consistència, ens interessa analitzar l'*ECMI* i la seva velocitat de convergència, per tal de decidir-nos sobre la funció K i el paramètre h conegut com a finestra del *Kernel*.

En primer lloc hem de tenir pressent que Boyd y Steele (78) [16] van demostrar que no existeixen estimadors no paramètrics que superin el "llistó" d'una velocitat de convergència més ràpida que $o(n^{-1})$.

Per altra banda i donada la incompatibilitat entre la consistència i la robustesa qualitativa, sabem que hem d'aplicar un filtre previ que detecti i elimini els *outliers*, interpretant la suavització variant h com un procés intermedi entre estimar amb la distribució mostral (massa inestable) i suavitzar amb finestra fixa (massa rígid).

En aquest sentit se sap que si $h \to 0$ i $nh \to \infty$, tenim garantida la convergència de l'*ECMI* sota condicions de regularitat de K, les quals compleixen totes les funcions que es proposen habitualment(Devroye, 1988 [52]). Afinant més, podem dir amb Nadaraya (1965) [161] que la condició necessària i suficient d'aquesta convergència és que:

$$\sum_{n=1}^{\infty} \exp(-\gamma n h^2) \text{ sigui convergent } \forall \gamma > 0$$

obtenint-se condicions de convergència uniforme si $nh/\log(n) \to \infty$.

El resultat fonamental és que la finestra òptima (*ECMI* mínim) és:

$$h_{opt} = k_2^{-\frac{2}{5}} K_2^{\frac{1}{5}} \nu_1(f)^{-\frac{1}{5}} n^{-\frac{1}{5}}$$

on $K_2 = \int K(t)^2 \mathrm{d}t$ i k_2 és el moment d'ordre 2 de K.

L'*ECMI* resultant és:

$$ECMI_{opt} = \frac{5}{4} k_2^{-\frac{2}{5}} K_2^{\frac{1}{5}} \nu_1(f)^{-\frac{1}{5}} n^{-\frac{4}{5}}$$

Per a interpretar correctament aquests resultats cal tenir present que el *Kernel* funciona com una convolució de la distribució mostral, amb la K regulada per h, de manera que: a menor h menys biaix però més variància.

Retindrem de les fórmules que l'equilibri *ECMI*, el qual és de l'ordre de $n^{-\frac{4}{5}}$, s'obté per a valors de h proporcionals a $n^{-\frac{1}{5}}$ amb coeficients que depenen tant de la pròpia K com de la desconeguda f.

Respecte a la forma de K i després de molta discussió arran de si és preferible utilitzar nuclis de suport compacte, com el d'Epanechnikov (tipus parabòlic) o gaussians, s'ha vist que la influència d'aquesta selecció és mínima respecte a la de la finestra (Hall i Marron, 1987 [101]). En qualsevol cas si, com al nostre cas, la densitat que volem estimar prové d'una mixtura de normals la decisió pels nuclis gaussians és evident.

Fixat el tipus de K podem tornar a la determinació de la finestra h amb més concreció: en casos de K Normal el procediment utilitzat és partir del valor $n^{-\frac{1}{5}}$ que és prop del òptim ($k_2 = 1,\ K_2 = 0.71,\ \nu_1 = 0.53$) i fer petites variacions cap amunt i cap avall cercant el millor valor per un procediment de validació creuada.

També hi hagut molta controvèrsia (Marron, 1987 [151]) en relació a si és preferible utilitzar en la determinació de la finestra una validació creuada mínimo-quadràtica (funció de pèrdua tipus L_2) o de Kullback-Leibler (funció de pèrdua tipus L_1), però la conclusió final és que la diferencia pràctica és mínima, tal i com ho és també si en lloc de la validació creuada emprem un procediment tipus *bootstrap*.

Cal ressenyar, més que res per la seva potència explicativa, la possibilitat d'utilitzar una metodologia bayesiana ja que, de forma semblant al que succeeix en el cas de la ridge regression [174], ens aporta la interpretació de que si prenem la funció nucli K com a densitat a priori, la finestra h resulta el quocient entre les variàncies a priori i mostral.

Finalment, s'han de considerar també els treballs de Schucany (1989) [189], que mitjançant desenvolupaments en sèrie realitza una aproximació més fina de l'estimació d'h que li porta a proposar la substitució de l'exponent 1/5 per 1/9.

3.5 La suavització mitjançant *Kernel* adaptable multidimensional

Pel cas multidimensional la finestra óptima resulta per p variables (Silverman(1986) [198] pàg.85):

$$\left(pk_2^{-2} K_2 \left(n \int (\nabla^2 f)^2 \right)^{-1} \right)^{\frac{1}{p+4}}$$

on s'observa la substitució del coeficient $n^{-\frac{1}{5}}$ per $n^{-\frac{1}{p+4}}$

Tanmateix l'*ECMI* pot incrementar-se significativament si la f té les cues "pesades", donat que al tenir que fixar la mateixa h per a totes les zones de la distribució, si es suavitza "massa" la zona central, es deixen "fluctuants" les cues. En conseqüència, Silverman proposa un estimador *Kernel* adaptable que ajusta una finestra diferent segon la zona de l'espai de que es tracti. L'algorisme consisteix a:

1. Partir d'un estimador pilot f^*, normalment un *Kernel* de finestra fixa.

2. Definir factors d'amplada de banda $\lambda_i = \left(\frac{f^*(X_i)}{g} \right)^{-\alpha}$ on g és la mitjana geomètrica dels $f^*(X_i)$ i α un factor de sensibilitat entre 0 i 1. Breiman i Meisel (1977) [20] proposen $\alpha = 1/p$ però Silverman suggereix $\alpha = 1/2$

per reduir el biaix fins al grau 4. Incementar α significa aproximar-se a l'estimador pilot.

3. La estimació ve definida per:

$$\hat{f}(t) = n^{-1} \sum_i (h\lambda_i)^{-p} K\left((h\lambda_i)^{-p}(t - X_i)\right)$$

L'interessant d'aquesta proposta és que és equivalent a un estimador de màxima versemblança amb penalització quadràtica a partir d'un polinomi en les derivades de f (recordem la *Kernel property*, pàgina 23), pel que es convertirà en una referència que hi ha que tenir en compte per al cas multidimensional.

3.6 Combinació *Kernel*–Correspondències

En aquesta secció mostrarem les dificultats observades al aplicar la suavització *Kernel* a les quantificacions resultants d'una anàlisi de correspondències simples.

Com a introducció, l'apartat 3.6.1 ens indicarà el problemes a l'hora de realitzar la pseudoinversa de la discretització d'una distribució Normal. Al 3.6.2 mostrarem, amb un exemple, el que pot ocórrer al combinar *Kernel* i correspondències simples.

3.6.1 La deformació introduïda per *Kernel* quan s'aplica a la discretització d'una Normal

És fonamental tenir en compte, abans de qualsevol altra consideració, que el *Kernel*, al ser un mètode d'estimació de densitats, requereix una mostra que tingui "llibertat" per moure's per tot l'espai, lo que no s'avé massa bé en la quantificació provinent de les correspondències amb l'interpretació provinent del teorema de Lancaster (secció 2.3.1, pàg. 42), on una part de la mostra s'ha "col·lapsat" al seu valor central de zona, el que produeix deformacions en l'estimació de la densitat resultant (recordar figura 2.8, pàg. 49).

Per veure aquest efecte gràficament, utilitzarem un exemple amb una estimació basada en una mostra de 10000 punts d'una Normal tipificada i una discretització amb punts de tall -0.3, 0.2, 0.6. Farem les suavitzacions a partir de les mitjanes de cada interval amb dues finestres fixes de partida: 0.4 i 0.8.

Compararem les següents densitats:

1. Normal tipificada.
2. Suavització amb *Kernel* de finestres fixes.
3. Suavització amb *Kernel* de finestra variable ajustant localment amb les variàncies obtingudes a cada interval.
4. Mixtura amb les dades de la discretització de cada interval.

Les densitats 1 i 4 actuen com a referència donat que la 1 és la densitat de partida i la 4 el resultat "cru" de la discretització amb una mixtura de normals per cada interval amb la seva corresponent mitjana, variància i freqüència.

Per la seva part les densitats 2 i 3 ens donaran el resultat d'una suavització amb *Kernel* de finestra fixa i variable respectivament. El resultat el podem apreciar a la figura 3.1.

Observem que quan s'aconsegueix la suavització requerida se'ns desajusta la kurtosis (totes les variables han estat tipificades) amb un excés de 2.48 pel *Kernel* de finestra fixa i de 4.60 pel de finestra variable (veure gràfica Finestra 0.8 a la figura 3.1). Per altra banda és clar que, amb un *Kernel* de finestra variable s'obté una suavització més ràpida i acurada, tal i com caldria esperar de l'explicat a la secció 3.4.

3.6.2 *Kernel* i correspondències simples

Per explorar les possibilitats reconstructores de la binormal, combinant *Kernel* i l'anàlisi de les correspondències simples, vam utilitzar com a punt de partida la taula que empra Kendall (vol.II, pàg 595 [133]) per a il·lustrar el teorema de Lancaster.

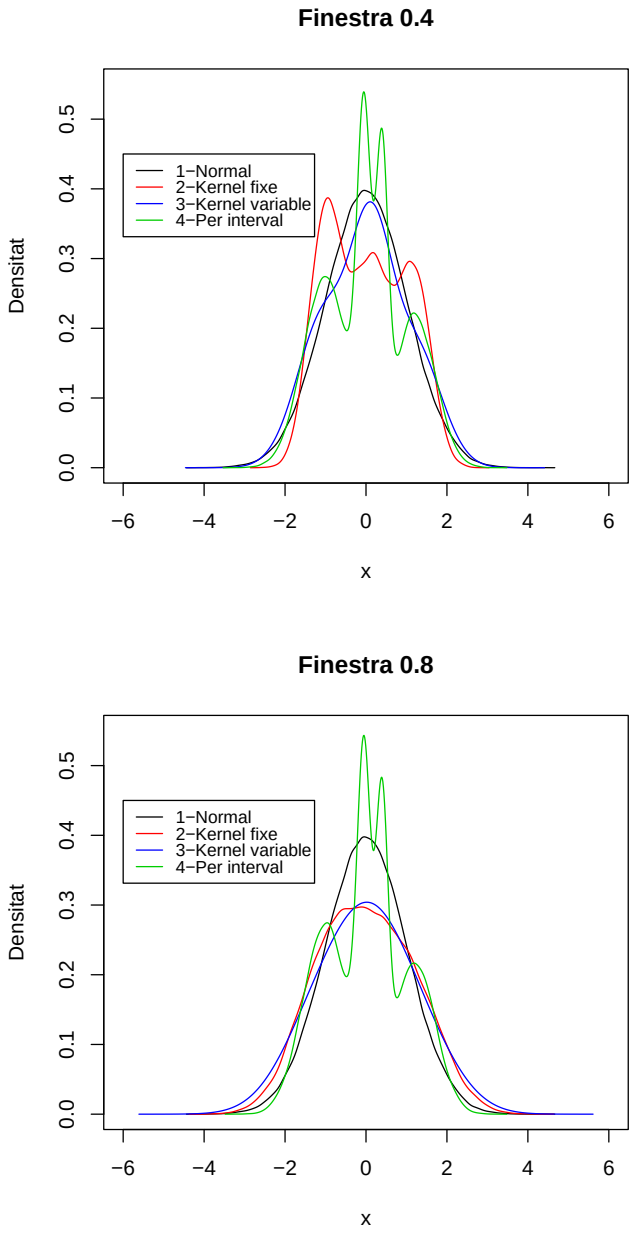

Figura 3.1: *Kernel sobre la discretització d'una distribució Normal*

Amb aquest objectiu prendrem de partida un coeficient de correlació de 0.69 (que és el que estima Kendall per a la distribució subjacent) i calcularem els punts de tall de manera que es corresponguin a les freqüències marginals obtenint-se per a la primera variable: 0.328, 0.568, 0.841 i per a la segona: 0.324, 0.565, 0.842.

Introduirem ara la variable que determina la classe, la qual no era present a l'exemple de Kendall ja que la taula esmentada no es presenta en un context d'anàlisi discriminant.

Ho farem considerant que a cada cel·la (i, j), determinada per la discretització, existeix una probabilitat p_{ij} de pertànyer a la classe 1, i una probabilitat $1 - p_{ij}$ de fer-ho a la classe 2.

La matriu dels corresponents pesos, p_{ij}, que vam aplicar va ser:

$$\begin{pmatrix} 0.8 & 0.6 & 0.4 & 0.2 \\ 0.6 & 0.8 & 0.6 & 0.4 \\ 0.4 & 0.6 & 0.8 & 0.6 \\ 0.2 & 0.4 & 0.6 & 0.8 \end{pmatrix} \tag{3.3}$$

per reflectir un domini de la classe 1 sobre la direcció marcada per la diagonal principal. També suposarem que la probabilitat de pertànyer a la classe 1 ès $p(1) = \frac{2}{3}$.

El resultat, si anomenem $Y12[, 1]$ a la primera variable i $Y12[, 2]$ a la segona sent Yc la corresponent a la classe, pot veure's a la figura 3.2, tal i com resulta de la utilització de la rutina *density* del llenguatge **R** (veure secció 5.6, pàg. 114).

Ara considerarem com a punts de tall "naturals" els valls de la distribució i discretitzarem d'acord amb aquestos, realitzant posteriorment l'anàlisi de correspondències sobre la taula resultant sense fer intervenir la classe (el que explica l'ús de la notació $Y12$ per indicar que les dues classes estàn incloses).

Al representar conjuntament les quantificacions proporcionades per l'anàlisi de correspondències (rodones) i els centroïds de zona inicials(triangles) obtenim

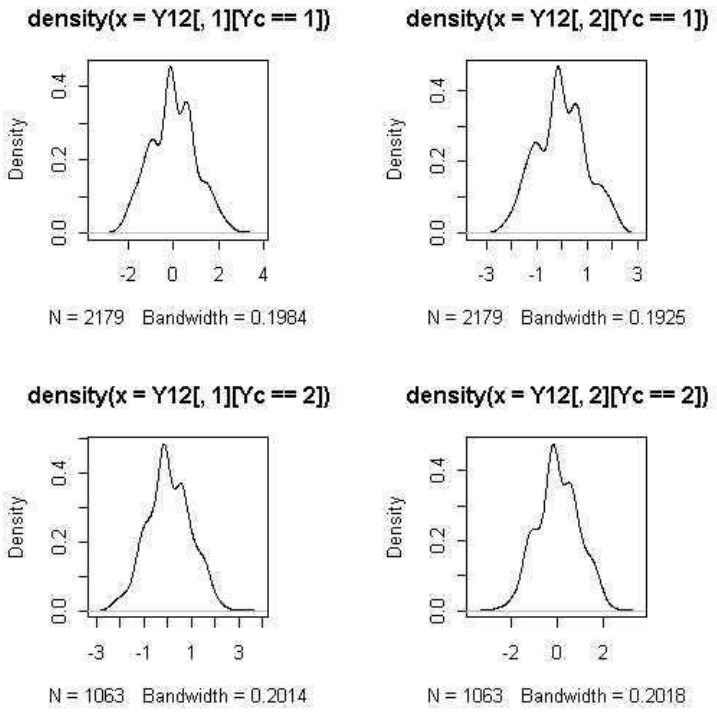

Figura 3.2: *Densitat original (Y*12*) per variable i classe*

el gràfic de la figura 3.3, on observem un efecte que podem anomenar de "rectangularització", degut al fet de que l'anàlisi de correspondències ha de donar una quantificació fixa per a cada filera o columna.

Si procedim a suavitzar mitjançant *Kernel* a partir dels centres de zona comuns a les dues classes, però atorgant a cadascuna el pes que el correspon segons la matriu (3.3), i deixant que la finestra s'ajusti per validació creuada sobre l'error final discriminant, podrem tornar a fer, a efectes comparatius, la mateixa gràfica que a la figura 3.2 substituint la nomenclatura $Y12$ per $Z12$. El resultat pot observar-se a la figura 3.4.

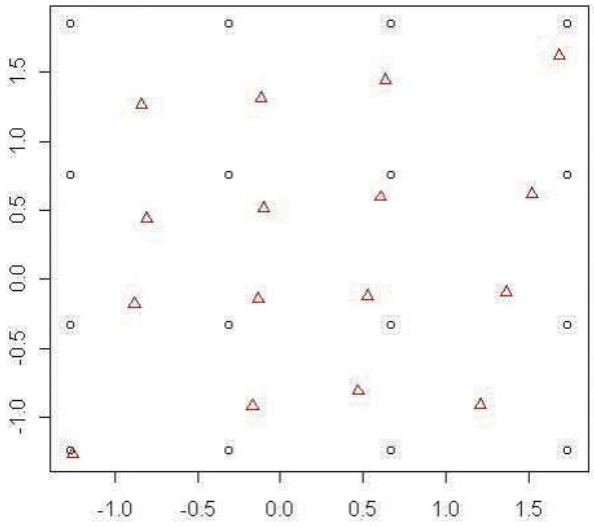

Figura 3.3: *Reconstrucció dels centroïds de zona amb les quantificacions de les correspondències*

L'efecte d'una menor suavitat que la inicial és prou evident si comparem aquestes figures, però encara pot apreciar-se millor si comparem els gràfics de les realitzacions de la densitat inicial (ja separada per classes, negre = classe 1, roig = classe 2) i la seva reconstrucció mitjançant *Kernel*-correspondències (figura 3.5).

Com a conclusió d'aquestes exploracions observem les limitacions de l'aplicació d'aquest procediment quan la suavitat de la que procedim és, com en aquest cas, molt alta. Això es degut al fet que si es reconstrueix amb una suavitat propera a la inicial, és a costa de la kurtosis (com vam veure a la figura 3.1) i això ens dificulta la tasca discriminant. Per contra si tractem de millorar l'error final discriminant s'ha de fer a costa de la suavitat en la reconstrucció (figura 3.5) i el resultat no aconsegueix un equilibri prou satisfactori.

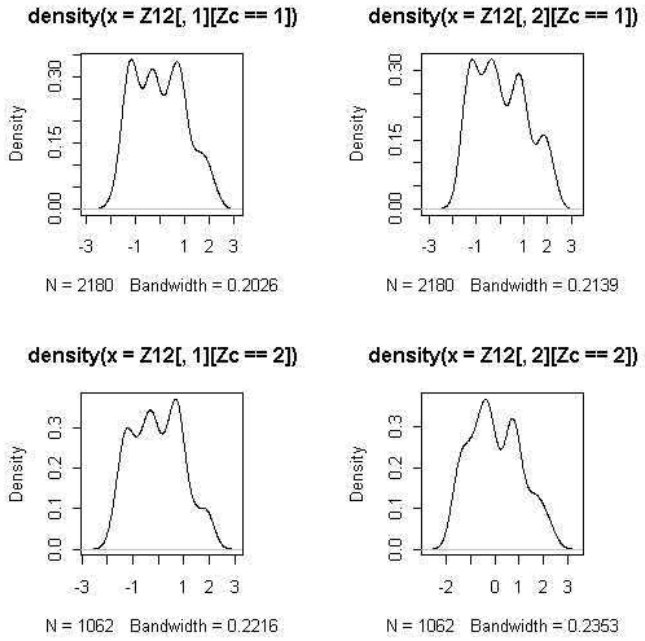

Figura 3.4: *Densitat reconstruïda (Z*12*) per variable i classe*

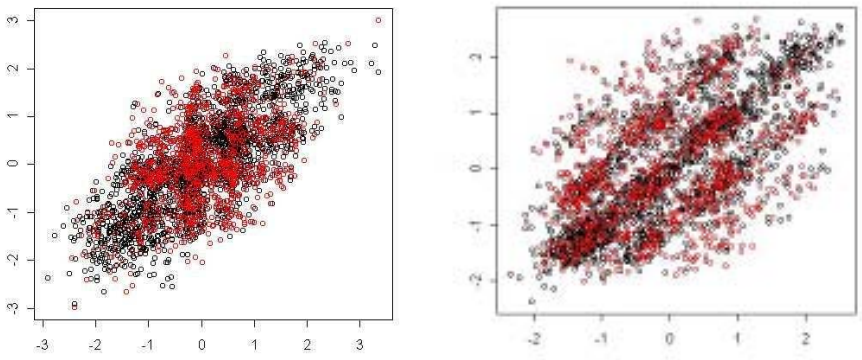

Figura 3.5: *Comparació dels núvols corresponents a Y*12 *i a la seva reconstrucció Z*12

Tanmateix, cal esperar que si la densitat contínua subjacent no és Normal o mixtura de Normals i el nombre de categories és alt, amb discretitzacions realitzades als punts de vall de les marginals (tal i com resulta natural), el *Kernel* adaptable ens proveirà d'un mètode d'una gran flexibilitat.

Per aquesta raó s'han detallat les seves propietats bàsiques en aquest capítol i s'inclourà en els suggeriments per a ampliar la recerca (pàg. 116) quan les suposicions de normalitat subjacent detallades a la secció 1.1 (pàg. 5) no es compleixin.

3.7 El procediment *EM*

Donades les dificutats esmentades a la secció anterior per suavitzar la sortida d'un anàlisi de correspondències (quan les distribucions subjacents per classe se suposen normals) amb el mètode del *Kernel*, analitzarem ara una alternativa de suavització especiament dissenyada per a situacions de mixtura de normals. Es tracta d'un procediment que té el seu origen en els sistemes de *Data Augmentation* Tanner, 1991 [206].

La filosofia es pot resumir amb "complicar un per a simplificar dos". Efectivament, si volem estimar θ de la versemblança $L(\theta; x)$, i aquesta té una expressió difícil de manegar, incorporarem unes noves variables "latents" z de manera que $L(\theta; x, z)$ sigui més simple. Necessitem, això si, conèixer $p(z/\theta, x)$ donat que el procediment serà:

- Partim d'un valor inicial θ_0.
- **Etapa E (Esperança)**: Calculem $Z = E(z/\theta_0, x)$.
- **Etapa M (Màxima versemblança)**: Estimem θ_1 per màxima versemblança a partir de $L(\theta; x, Z)$.
- Tornem a l'*Etapa E* per iterar mentre $|\theta_0 - \theta_1| > \varepsilon$

Aquesta és la forma més estàndard e intuïtiva de presentar el mètode però es pot demostrar (Neal i Hinton, 1998, [162]) que les dues passes es resumeixen a:

$$\theta_1 = \text{argmax}_\theta \{\text{E}(L(\theta, z/x, \theta_0)\}$$

L' estructura d'aquest procediment el fa molt adient a situacions de mixtura de normals, donat que si fem intervenir z com a la variable que determina el component de la mixtura al que pertany x, les $L(\theta; x, z)$ seran versemblances gaussianes.

Aquesta és la raó per la qual el mètode *MDA* la utilitzarà per suavitzar les dades d'un problema discriminant continu, suposat de distribució base mixtura de Normals, emprant l'algorisme explicat a la secció 1.3.5 (pàg. 24). Per la nostra part serà el mètode de suavització que triarem com a primera alternativa a la proposta del capítol 4 mantenint el *Kernel* adaptable multidimensional esmentat al paràgraf anterior com a segona alternativa, ja que, com hi es deia, pot servir per millorar la cobertura del mètode quan les suposicions bàsiques (secció 1.1, pàg. 5) no es compleixin.

Capítol 4

Anàlisi Discriminant Discreta pel mètode *ADDSUC*

Després de la incursió del capítol precedent pels mètodes de suavització, disposem ja de les eines teòriques necessàries per passar a desenvolupar, en aquest capítol, la nostra proposta.

En primer lloc reprendrem l'anàlisi de correspondències al punt que ho havíem deixat al final del capítol 2: la conveniència de trobar una generalització multidimensional de les propietats de reconstrucció de la normal bivariable, demostrades per Lancaster per a la correlació canònica simple, de manera que tingui utilitat per l'objectiu discriminant que perseguim.

La secció 4.1 és fonamentalment notacional, mentre que a la secció 4.2 es revisen les principals propostes que, a hores d'ara, han utilitzat les correspondències amb objectius discriminants. Finalment, a la secció 4.3 presentarem la nostra proposta fonamentant-la matemàticament a l'apartat 4.3.3 (pàg. 88) i provant la convergencia de l'algorisme al 4.3.5 (pàg. 94).

4.1 L'anàlisi discriminant com a correlació canònica

En aquesta secció començarem per enfocar l'anàlisi discriminant "clàssica" com una correlació canònica simple entre variables indicadores, amb l'objectiu de proveïrnos d'una adequada notació que ens permeti la posterior aplicació de l'anàlisi de correspondències en aquest context.

4.1.1 Expressió d'una anàlisi discriminant lineal (*LDA*) com a correlació canònica simple

Com s'esmentava a la secció 1.3.2 (pàg. 20), l'*LDA-canònic* ens proporciona la solució lineal discriminant óptima de dimensió r, mitjançant la selecció dels r primers vectors propis de la matriu $B\Sigma^{-1}$ on Σ representa la matriu de covariàncies comuna i B la dels centroïds de classe. Ara bé, és prou conegut i fàcilment demostrable, que aquest procés dona el mateix resultat si substituïm Σ per la anomenada matriu de covariàncies totals $T = B + \Sigma$.

El nostre objectiu serà trobar l'expressió de la matriu a diagonalitzar $B \circledast T^{-1}$, en termes de les matrius de dades, per tal de poder analitzar les possibilitats d'adaptació al cas discret múltiple. D'aquesta manera superarem els problemes que la aplicació directa sobre matrius indicadores produeix, però sense perdre els avantatges ja esmentats a 1.3.2 (pàg. 20).

Començarem per denotar com a Y la matriu indicadora de classes, de dimensions $(n \times g)$, i com a X, la matriu amb les variables discriminadores (dicotomitzades) com a columnes, de dimensions $(n \times k)$ amb $k = \sum_{j=1}^{p} k_j$.

Hem d'aclarir que, per simplicitat notacional, a la resta d'aquest capítol considerarem Y, X aquestes matrius i no les definides a la secció 1. Es a dir que si, per exemple, el vector de les classes per individus correspon al membre esquerre de l'expressió (4.1) en lloc de considerar-ho com a Y, reservarem aquesta notació

per a l'equivalent dicotomitzat (membre dret).

$$\begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.1}$$

El mateix farem per a cada columna de la matriu inicial de variables classificadores unint-les com es reflecteix al petit exemple de l'expressió (4.2).

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \\ 2 & 1 \\ 3 & 2 \\ 3 & 2 \\ 1 & 1 \\ 2 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{4.2}$$

Si ara, projectem X sobre el subespai generat per les variables indicadores Y, obtenim: $P_Y X = Y(Y'Y)^{-1}Y'X$. Aquesta matriu té per fileres les mitjanes de les classes G_i, $i, i = 1, \ldots, g$ amb $G_i = ({Y_{(i)}}'Y_{(i)})^{-1}{Y_{(i)}}'X$, on $Y_{(i)}$ representa la columna $i-$sima de la Y.

Per altra banda, centrar una matriu X és fer $X_c = (\mathbf{I} - \frac{1}{n}\mathbf{1}'_n\mathbf{1}_n)X$ i per tant ${X_c}'X_c = X'(\mathbf{I} - \frac{1}{n}\mathbf{1}'_n\mathbf{1}_n)X$ pel que si anomenem $H=$ matriu centradora $= (\mathbf{I} - \frac{1}{n}\mathbf{1}'_n\mathbf{1}_n)$, tenim $X_c = HX$ i per tant ${X_c}'X_c = X'HX$ (per idempotència d'H). A partir d'aquí suposarem per raons de simplicitat notacional X centrada, és a dir notarem X com si fos X_c.

Amb aquestes definicions la matriu de covariàncies total serà:

$$T = X'X$$

i la de les variàncies entre classes:

$$B = (P_Y X)' P_Y X = X'Y(Y'Y)^{-1}Y'X$$

resultant la matriu a diagonalitzar:

$$B \circledast T^{-1} = X'Y(Y'Y)^{-1}Y'X \circledast (X'X)^{-1} \tag{4.3}$$

El que correspon exactament a una correlació canònica simple entre Y i X.

4.1.2 El triplet de l'*LDA* amb ponderació d'individus

Una vegada repassada la notació i establerta la matriu a diagonalitzar a l'anàlisi discriminant clàssica, en termes de les matrius indicadores, veurem la forma de representar aquest procés mitjançant la terminologia de triplets (secció 2.2, pàg. 36).

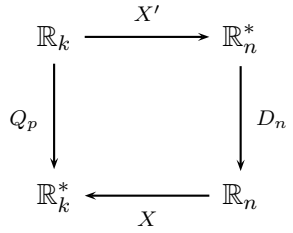

Figura 4.1: *Esquema de l'ACP de X*

Utilitzarem, en primer lloc, el diagrama de la figura 4.1, on Q_p representa la mètrica dels individus entesos com a metavariables (el que hem anomenat $M = (X'X)^{-1}$ a l'ACP i npD^{-1} a l'ACM), i D_n la ponderació als individus (el que hem anomenat $N = \mathbf{I}$ al'ACP i $\mathbf{I}/n$ a l'ACM).

A una anàlisi discriminant tindrem aquest esquema per a les X i un altre de semblant per a les Y, compartint la D_n (ponderació dels individus), segons

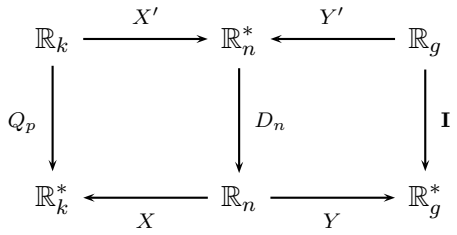

Figura 4.2: *Esquema dels ACP de* X i Y combinats

s'observa a la figura 4.2, on amb la simplificació $D_n = \mathbf{I}/n$ (tots els individus pesen igual) tindríem la situació expressada al discriminant clàssic.

Ara bè, si volem obtenir la matriu a diagonalitzar de la fórmula (4.3) hem de creuar X i Y mitjançant el triplet:

$$\Big(Y'D_nX,\ (X'D_nX)^{-1},\ (Y'D_nY)^{-1}\Big)$$

On podem dir que retrobem sobre $\mathbb{R}_k$ la mètrica $(X'D_nX)^{-1}$ que ha introduït el pes dels individus al càlcul de les covariancies i sobre $\mathbb{R}_g$ la mètrica $(Y'D_nY)^{-1}$, la qual, donat que Y representa a les variables indicadores de les classes, correspon a la divisió per el total de pesos de cada classe (veure figura 4.3).

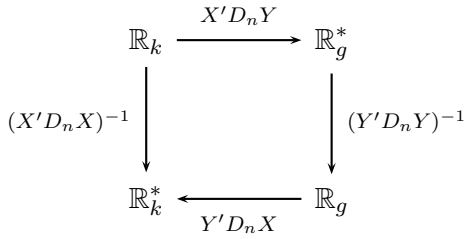

Figura 4.3: *Esquema de la Correlació Canònica Simple*

La matriu de treball serà ara $Y'D_nX$ on les files representen les classes amb valors que són les sumes ponderades de cada variable dins de la classe corresponent. Això és equivalent a treballar amb $(Y'D_nY)^{-1}Y'D_nX$, on tindríem mitjanes ponderades per classe, és a dir els individus entesos com a meta-variables han estat substituits pels individus genèrics de cada classe, obtinguts per mitjana

ponderada i on la mètrica $(Y'D_nY)^{-1}$ ha estat invertida. Es a dir, substituïm un triplet doble-discriminant per un de perfil de fileres. El resultat serà:

$$\Big((Y'D_nY)^{-1}Y'D_nX,\ (X'D_nX)^{-1},\ Y'D_nY\Big)$$

Seguint l'esquema, en els dos casos la matriu a diagonalitzar serà la mateixa:

$$X'D_nY(Y'D_nY)^{-1}Y'D_nX \circledast (XD_n{}'X)^{-1} \tag{4.4}$$

Que en el cas clàssic (no ponderant els individus, $D_n = \mathbf{I}/n$) porta al conegut discriminant vist com correlació canònica (fórmula 4.3):

$$X'Y(Y'Y)^{-1}Y'X \circledast (X'X)^{-1} \tag{4.5}$$

corresponent a la $X'NX \circledast M$ de l'ACP de fileres de l'esquema de Tenenhaus (Individus com a meta-variables, pàgina 40).

És a dir, l'anàlisi discriminant pot veure's com un ACP interclasses on tots els individus meta-variables entren per calcular la mètrica entre variables essencials $(X'X)^{-1}$, però després a l'hora de considerar-los com a individus essencials només es tenen en compte les mitjanes de cada classe.

Tenim també una forma alternativa d'arribar a aquesta diagonalització que és la següent: sigui $G(Y)$ el subespai engendrat a $\mathbb{R}_n$ per les indicadores de classes les quals conformen amb columnes la matriu Y i sigui $P_Y = Y(Y'D_nY)^{-1}Y'D_n$ el corresponent projector sobre $G(Y)$. Llavors, l'anàlisi discriminant quedarà reflectida pel triplet:

$$\Big(P_YX,\ (X'D_nX)^{-1},\ D_n\Big) \tag{4.6}$$

que es pot interpretar com trobar vectors (columnes "essencials") normalitzats respecte a $X'D_nX$ i que siguin ortogonals respecte a $(P_YX)'D_nP_YX$ (mètrica transferida dels individus amb restricció a les classes).

Finalment, mirant l'esquema de la figura (4.2) com un acoplament de les anàlisis de la X i de la Y i, per tant, de les taules de dades corresponents a cadascuna, poden tenir en compte el que diuen Chessel i Thioulouse, 1997

(Mètodes K-tableaux, pàg.51) [28]: *Acoplar taules per les fileres significa o bé considerar-les diferents conjunts de variables sobre els mateixos individus, com fa l'anàlisi canònica, o bé com les mateixes variables sobre individus diferents, com fa l'anàlisi de co-inèrcia; però no les dues coses alhora, com en el cas d'una única taula.*

Hem citat aquesta frase perquè ens sembla que és una manera interessant d'expressar també el problema de la generalització del correspondències simples al cas múltiple. És a dir, al replicar les taules hem de decidir-nos quin costat del rectangle de la figura 2.3 (pàg. 38), amb el que hem representat les *ACP* (individus metavariables i variables essencials o variables metaindividus i individus essencials) hem d'utilitzar per anar "pegant" els diferents anàlisis. En el cas del discriminant l'enllaç s'ha representat a la figura 4.2 i al fer-se sobre $\mathbb{R}_n^* - \mathbb{R}_n$ vol dir que "juntem", com és lògic, per la segona d'aquestes opcions.

4.1.3 Correlació canònica simple versus Correlació canònica generalitzada

L'anàlisi discriminant clàssica és, com acabem de veure, essencialment simètrica, ja que es tracta, bàsicament, d'una correlaciò canònica simple amb Y com a matriu d'indicadores. Ho podem enfocar, tant com un triplet doble discriminant (amb consonància amb la seva estructura simètrica, veure fórmula 2.2, pàg. 41):

$$\left(Y'X,\ (X'X)^{-1},\ (Y'Y)^{-1}\right)$$

o com un triplet del perfil de fileres (veure fórmula 2.3 (pàg. 42)):

$$\left((Y'Y)^{-1}Y'X,\ (X'X)^{-1},\ Y'Y\right)$$

que reflecteix la intenció de discriminació.

Pel contrari l'*ACM* (*Anàlisi de Correspondències Múltiples*) és essencialment una canònica generalitzada (recordem la taula de la pàgina 51), la qual expressada en la forma habitual, es centra en trobar un vector genèric unitari ψ ($\|\psi\|$

$= \psi'\psi = 1$), de forma que la suma de les normes al quadrat de les seves projeccions ortogonals al subespai de cada una de les variables (agrupant categories) $\sum_i \|\mathrm{P}_{X_i}\psi\|^2 = \lambda$, sigui màxima.

Com a conseqüència tenim que $\sum_i \mathrm{corr}^2(\mathrm{P}_{X_i}\psi,\ \psi) = \lambda$, que és l'altra manera de veure la maximització que es realitza. La matriu a diagonalitzar és $\sum_i \mathrm{P}_{X_i}$, però si definim Ω com una matriu amb caixes diagonals $X_i'X_i$ el resultat és equivalent a diagonalitzar $\Omega^{-1} \circledast X'X$ i fer la transformació $\psi = X\Omega^{-1/2}v/\sqrt{\lambda}$ on v és el corresponent vector propi.

Sabem, això si, que la canònica generalitzada es redueix a la canònica simple quan fem $Y = X_1$, $X = X_2$ resultant ψ el vector unitari a la direcció bisectriu de les projeccions (amb la mateixa norma al quadrat $= \lambda/2$) i la correlació $\lambda - 1$ (Volle, 1981 [221]), el que està en consonància amb l'explicat al començament de la secció 2.4 quan es tractava del pas de les correspondències simples a les múltiples.

Amb terminologia de triplets, aquesta canònica simple pot veure's com:

$$\Big(X_1'X_2,\ (X_2'X_2)^{-1},\ (X_1'X_1)^{-1}\Big)$$

mentre que la canònica generalitzada ens portaria, en aquest cas de dues variables, a:

$$\Big((X_1, X_2),\ \Omega^{-1},\ \mathbf{I}\Big)$$

on observem que la estructura de creuament s'ha desplaçat de la matriu de dades a la mètrica a $\mathbb{R}_k$ (recordem que Ω seria la matriu $X'X$ deixant iguals les caixes diagonals $X_1'X_1$ i $X_2'X_2$ i fent les corresponents a $X_1'X_2 = \mathbf{0}$).

En resum:

Anàlisi	**Mètrica fileres**	**Matriu de dades**	**Mètrica de columnes**
Canònica Simple	$(X_1'X_1)^{-1}$	$X_1'X_2$	$(X_2'X_2)^{-1}$
Canònica Generalitzada	I	(X_1, X_2)	$\begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix}$

L'interessant és que el ψ de la canònica generalitzada que ens dóna les quantificacions variable a variable per projecció:

$$P_{X_1}\psi = X_1\zeta_1 \qquad P_{X_2}\psi = X_2\zeta_2$$

i la maximització de:

$$|X_1\zeta_1|^2 + |X_2\zeta_2|^2 = \lambda = \text{corr}^2(X_1\zeta_1,\ X_1\zeta_1 + X_2\zeta_2) + \text{corr}^2(X_2\zeta_2,\ X_1\zeta_1 + X_2\zeta_2)$$

resulta, en aquest cas de només dues variables:

$$|X_1\zeta_1|^2 = |X_2\zeta_2|^2 = \lambda/2 = \text{corr}^2(X_1\zeta_1,\ X_1\zeta_1 + X_2\zeta_2) = \text{corr}^2(X_2\zeta_2,\ X_1\zeta_1 + X_2\zeta_2)$$

obtenint també la maxitmització de l'equivalent del coeficient de correlació al que es referia Lancaster:

$$\zeta_1' X_1' X_2 \zeta_2 = \lambda(\lambda - 1)/2$$

Però aquest resultat, que pot obtenir-se per una certa simplificació i simetria del cas de dues variables (maximitzar la suma de les correlacions al quadrat entre el vector global i les projeccions és equivalent a maximitxar la correlació entre aquestes dues), no pot, malauradament, generalitzar-se al cas d'un nombre superior de variables.

En definitiva, l'*LDA* és una canònica simple entre Y i X, entenent aquesta X sense l'estructura multidimensional, i l'*ACM* és una canònica generalitzada que té en compte aquesta disposició, però no la relaciona amb la Y.

Es tracta, doncs, d'aconseguir una adequada combinació entre les dues, de manera que, puguem utilitzar l'estructura multidimensional de les X per discriminar sobre la Y.

4.2 Les propostes prèvies per a l'anàlisi discriminant de correspondències múltiples

En aquesta secció revisarem els tres intents previs més importants d'utilització de l'anàlisi de correspondències amb intencions discriminants: el de Benzècri-Palumbo (apartat 4.2.1), el de Chessel i Thiolouse (apartat 4.2.2) i el de Saporta (apartat 4.2.3).

4.2.1 Les correspondències múltiples no simètriques de Benzècri-Palumbo

Palumbo reprèn al 1998 [168] l'anomenat *Baricentric Discriminant Analysis* de Benzécri, (1982) [8], el qual és, en essència, una anàlisi de correspondències (múltiple) de la Y amb les X (juxtaposades) on la matriu a diagonalitzar és:

$$X'Y(Y'Y)^{-1}Y'X \circledast D^{-1}$$

Si la comparem amb la que hem anomenat del discriminant clàssic amb ponderació d'individus (fórmula 4.4 (pàg. 74)):

$$X'D_nY(Y'D_nY)^{-1}Y'D_nX \circledast (X'D_nX)^{-1}$$

observem la suposició $D_n = \mathbf{I}$ (tots els individus pesen igual) i la substitució de la mètrica $(X'X)^{-1} = (X'D_nX)^{-1}$ per D^{-1} (on D és la matriu diagonal amb les freqüències de les k categories) que és el que justifica l'apelatiu de no-simètrica a aquesta anàlisi.

Hi ha que resaltar que no podem considerar aquesta anàlisi de Palumbo-Benzècri com a correlació canònica sobre indicadores ja que la D està formada

per les caixes diagonals d'$X'X$ (matriu de Burt), però a diferència d'aquesta matriu la resta és 0. Es tracta més ben bé d'una canònica generalitzada tipus ACM sobre les projeccions per classe $\left(P_Y X,\ D^{-1},\ \mathbf{I}\right)$ a l'estil de la fórmula 4.6 (pàg. 74).

L'inconvenient d'aquesta proposta és que el canvi de mètrica no és prou per recollir l'estructura multivariant i d'ahí que els resultats pràctics no siguin gaire satisfactoris com veurem al capítol 5.

4.2.2 L'anàlisi discriminant de correspondències de Chessel-Thioulouse

Chessel i Thioulouse, (1997) [28] van fer una proposta diferent que van anomenar ADC (*Analyse Discriminante de Correspondences*) pero limitada al cas simple.

Veurem en aquest apartat, que la generalització de la seva proposta al cas múltiple, ens condueix directament a l'LDA-canònic, i per tant no aporta cap solució nova al problema que estem estudiant.

Els autors començen indicant que per aplicar l'esquema de triplets de l'anàlisi discriminant al cas de X discreta bivariable, hem d'aplicar els triplets especificats a la secció 2.3 (pàg. 40) a una matriu de dades adient.

Aquesta ha de resultar per la lògica de discriminació ja comentada a l'apartat (4.1), de la aplicació del projector P_Y a X però fent el necessari centrament $P_{Yo} = P_Y - \mathbf{1}_n\mathbf{1}_n' D_n$ on hem modificat el subíndex f de D_f per l'n de D_n, a fi i efecte d'evidenciar que les fileres actuen ara com a individus. Això és el que Chessel i Thioulouse anomenen un anàlisi amb variables instrumentals per remarcar que el projector transforma les variables originals amb uns nous instruments que, en aquest cas, són la projecció de les variables originals al subespai generat per les indicadores centrades (equivalent a treballar sobre mitjanes de classe).

Si partim del triplet de fileres (suposant que és aquí on ens interessa la discriminació) i eliminem l'altre centrament al que es feia referència a (2.3) per

comoditat de notació (i perquè només afegeix el valor trivial 1) obtindrem:

$$\left(P_{Y_o} Dn^{-1} F,\ D_c^{-1},\ D_n\right)$$

Fins aquí només hem formulat el triplet corresponent a una anàlisi de correspondències simples. L'observació dels autors és que aquest procés no ha respectat les mètriques que caldrien per a que fos un veritable anàlisi discriminant i proposen, conseqüentment, substituir la mètrica D_c^{-1} per $(F' D_n^{-1} F)^{-1}$ és a dir agafar com a mètrica de les columnes "essencials" la transferida dels individus (fileres) en lloc de la de dividir pels pesos total de les columnes, ja que d'aquesta manera si es fan les oportunes traduccions notacionals ($X = D_n^{-1} F$) arribem a $(F' D_n^{-1} F)^{-1} = (X' D_n X)^{-1}$ que és, en definitiva, la matriu que s'aplica al discriminant vist com a correlació canònica amb individus ponderats (veure apartat 4.1.2).

Aquesta substitució de mètriques ens porta a les següents observacions:

- El canvi és de la mateixa natura que el que es fa a l'anàlisi de components principals al substituir com mètrica de variables "essencials" la matriu bàsica **I**, que considera ortonormals i d'igual pes totes les variables contínues, per $X' D_n X$ que fa intervenir la matriu de dades per a induir una mètrica que té en compte, com a estructura de covariàncies la que aporten les realitzacions concretes d'aquestes variables ponderant els individus amb D_n. És a dir, es canvia una matriu diagonal amb totals de columnes (mètrica bàsica a les discretes) per la transferida per dualització mitjancant X (o F), $X' D_n X = F' D_n^{-1} F$.

- La quantificació que s'obté fa que les columnes, enteses com metafileres, tinguin variància 1 (la total) i maximitza la variància entre centroïds B, tal i com corresponen a una anàlisi discriminant, en lloc de fer 1 la variància de la quantificació de les columnes "essencials" i després maximitzar B que seria al que ens portaria directament l'aplicació de la lògica del correspondències a la situació discriminant (veure apartat 4.1.1). Recordem que un canvi

de mètrica és equivalent a un canvi de normalització, pel que aquest és un enfocament complementari de l'anterior. Chessel i Thiolouse en trien aquest punt de vista i ho expresen diguent que el seu plantejament canvia la lògica d'una anàlisi canònica (la doble discriminació simètrica de fileres i columnes que fa un correspondències) per la d'una anàlisi de variables intrumentals (substituïm les columnes originals per unes on la normalització tingui més sentit en el nostre context).

Per analitzar la possible extensió multivariable de la proposta de Chessel i Thioulouse, hem de traduir primer a notació matricial més precisa el seu plantejament. Hem de tenir en compte que els autors consideren tres variables: les dos que determinen les fileres i columnes de l'anàlisi de correspondències i la que descriu els agrupaments de fileres que es pretenen discriminar.

Com observació que doni sentit pràctic a aquest procés, és convenient saber que els autors treballen en Ecologia de rius, i que les fileres representen estacions situades al llarg del riu en estudi, mentre que les columnes són les diferents espècies, els efectius dels quals es compten a cada estació. L'agrupament de les estacions ve determinat per zones del riu, on es preveu un comportament ecològic semblant (per eixemple des del naixement fins a un primer embassament situat a pocs Km.). L'objectiu és determinar com les diferents espècies intervenen en la caracterització de cada zona.

Si reservem Y com a la variable que conté les classes a discriminar, hem de cercar una altra lletra, posem L per a la matriu resultat de la dicotomització de les fileres, deixant X per a la de les columnes com abans.

D'aquesta manera fent les oportunes traduccions tindrem:

$$D_c = X'X \qquad D_n = L'L \qquad F = L'X$$

I per tant la matriu a analitzar serà:

$$D_n^{-1}F = (L'L)^{-1}L'X$$

La mètrica de fileres per a l'anàlisi de Chessel i Thioulouse sería:

$$(F'D_n^{-1}F)^{-1} = (X'L(L'L)^{-1}L'X)^{-1}$$

Mentre que per al de Benzècri-Palumbo tindríem:

$$D_c^{-1} = (X'X)^{-1}$$

La diferència, per tant, de la proposta de Chessel amb la de Benzècri-Palumbo pot veure's no només com un canvi de mètrica, sinò que, si considerem que els dos treballen amb mètriques euclídees del tipus $(X'DX)^{-1}$, la diferència està en que Benzècri en el seu discriminant baricèntric utilitza directament en aquesta mètrica la X fent $D = \mathbf{I}$, mentre que la de Chessel-Thioulouse és una de base contínua aplicable als perfils fileres $D_n^{-1}F$ i utilitza aquesta, que en expressió d'indicadores seria $(L'L)^{-1}L'X$, en el lloc de la X, fent també $D = D_n$.

Per tant, Benzècri-Palumbo amb el seu discriminant baricèntric mantenen l'estructura d'un correspondències amb la presentació de perfil de fileres, aplicant una correlació canònica sobre indicadors, però on les fileres han estat projectades per la classe, mentre Chessel i Thioulouse recullen aquests perfils de fileres per fer la correlació canònica amb les indicadores de columnes (naturalment incorporant també la projecció) recollint l'esperit del discriminant clàssic $(X'X)^{-1}(P_YX)'P_YX$, i trencant la simetria baricèntrica del correspondències.

Com a conseqüència, aquesta anàlisi no aporta res de nou al cas múltiple ja que la seva generalització ens porta directament de nou al discriminant clàssic.

4.2.3 L'anàlisi discriminant sobre variables qualitatives de Saporta

Finalment, la proposta de millors resultats entre les tres conegudes que enfoquen l'anàlisi discriminant discreta emprant el correspondències com a punt de partida és, com es veurà al capítol 5, la que va fer inicialment Saporta amb la seva tesi i que després ha anat perfeccionant, Gautier i Saporta, (1983) [70].

Es tracta d'un mètode consistent en:

1. Una anàlisi de correspondències múltiples (ACM) sobre X.
2. Una anàlisi discriminant d'Y sobre el resultat del pas anterior.

Encara que la proposta té una lògica molt convincent i justifica un procediment habitual entre els investigadors aplicats de l'escola "francesa", el problema està en que al no considerar l'estructura de la Y condicionant a la X al primer pas els resultats ja queden distorsionats i el segon pas només pot corregir-ho parcialment. *És a dir, en terminologia reconstructora, aplicada al nostre cas, s'està intentant reconstruir una normal on el que hi ha com a subjacent és una mixtura.*

4.3 La proposta *ADDSUC*

En aquesta secció, donat que els mètodes revisats a l'anterior no són satisfactoris, presentarem la nostra proposta en els següents apartats: Al 4.3.1 farem una revisió prèvia dels conceptes, al 4.3.2 expressarem el plantejament guia, el qual fonamentarem matemàticament al 4.3.3. Aquest prendrà expressió pràctica mitjançant l'algorisme explicat al 4.3.4, del qual demostrarem la convergència a 4.3.5.

4.3.1 Resum de conceptes previs

Abans de presentar la nostra proposta començarem per fer un resum dels conceptes hi involucrats que hem desenvolupat fins ara:

C-i) Tota matriu X de dades que pot abordar-se des de les fileres com a individus a l'espai de les variables ($\mathbb{R}_p$) i des de les columnes com a variables a l'espai dels individus ($\mathbb{R}_n$) té també dos **espais duals** complementaris inclosos implícitament amb aquestos enfocaments, el de les variables

essencials (coeficients de les variables, $\mathbb{R}_p^*$) i el dels individus essencials (coeficients dels individus, $\mathbb{R}_n^*$).

C-ii) Cada espai principal ha d'estar dotat d'una matriu definida positiva simètrica com a mètrica (M^{-1} a $\mathbb{R}_p$, N^{-1} a $\mathbb{R}_n$). També es pot **transferir la mètrica** de $\mathbb{R}_n$ a $\mathbb{R}_p$ mitjançant $X'NX$ i de $\mathbb{R}_p$ a $\mathbb{R}_n$ mitjançant XMX'. Per tant a cada espai dual tenim dues mètriques: la "essencial" (inversa del seu corresponent espai principal) i la transferida de l'altra parella dual mitjançant X. A $\mathbb{R}_p$ seran M^{-1} i $X'NX$; a $\mathbb{R}_n$ tindrem N^{-1} i XMX' (esquema de la figura 2.4, pàg. 40).

C-iii) **Una anàlisi de components principals (*ACP*) genèrica** serà bàsicament un procés d'harmonització d'aquestes dues mètriques (que per simplificació anomenarem aquí S i T) prenent l'essencial (T) com a referència i fent una ortogonalització conjugada progressiva de l'altra (S) per reducció canònica de la matriu $S \circledast T^{-1}$. Això ens portarà a una entre dos possibles reduccions canòniques segons quin sigui l'espai principal triat: $X'NX \circledast M$ si triem $\mathbb{R}_p^*$ (variables "essencials") i $XMX' \circledast N$ si triem $\mathbb{R}_n^*$ (individus "essencials"). Aquesta estructura la representarem amb el triplet (X, M, N) i equival a la coneguda com anàlisi factorial de $X'NX$ sota normalització en mètrica M.

C-iv) **En el cas de l'*ACP* clàssica**, si considerem que $M = I$ i $N = H$ matriu de centrament de les columnes, situant-nos a $\mathbb{R}_p$ (variables "essencials"), el procés es centra en reduir canònicament la matriu de les covariancies mostrals $X'HX$ (també existeix la possibilitat d'aplicar una ponderació D_n als individus i fer $N = HD_nH$ reduïnt $X'NX$). És a dir, es tracta de trobar noves variables unitàries (direccions z_i amb $z_i'z_i = 1$) que maximitzen en forma d'ortogonalització ($z_i \perp z_j$) conjugada progressiva la norma transferida dels individus *(inèrcia=variància)* $= z_i'X'NXz_i$.

C-v) **A l'anàlisi de correspondències simples** tota aquesta estructura ens porta a quatre triplets (partim d'una F doblement centrada en les file-

res i en les columnes): dos que recullen la simetria de la χ^2: el doble inercial $\left(D_p^{-1}FD_n^{-1},\ D_p,\ D_n\right)$ i el doble discriminant $\left(F,\ D_p^{-1},\ D_n^{-1}\right)$, i dos que privilegien la relació d'una variable com explicadora de l'altra: el del perfil de fileres $\left(D_n^{-1}F,\ D_p^{-1},\ D_n\right)$ i el del perfil de columnes $\left(FD_p^{-1},\ D_p,\ D_n^{-1}\right)$. Veure secció 2.3 (pàg. 40).

C-vi) **A l'anàlisi de correspondències múltiples (*ACM*)** se'n agafen aquestos dos últims: perfils de fileres i de columnes, anomenant-los respectivament *ACP de fileres* i *ACP de columnes* mitjançant la definició $F = HX\frac{1}{np}$, $D_p = D\frac{1}{np}$ (D = matriu diagonal de totals de categories) i $D_n = \mathbf{I}\frac{1}{n}$. Enfocaments com l'anomenat de Components principals faran l'*ACP de fileres* i d'altres, com el de Quantificació recíproca, cercaran un equilibri entre els dos (tipus doble inercial o doble discriminant). Recordem també, i de manera especial, que l'*ACM* pot enfocar-se com una anàlisi Canònica generalitzada i que, en aquesta, l'objectiu és trobar ψ vector global "auxiliar" (quantificació dels individus) amb $\psi'\psi = 1$ tal que $\sum_l \psi P_l \psi$ sigui màxima sent P_l el projector sobre el subespai X_l definit per la variable l, el que ens porta a diagonalitzar $D^{-1} \circledast X'X$ (taula 2.9, pàg. 51).

C-vii) **A l'anàlisi discriminant clàssica (*LDA*)**, al seu torn, pretenem maximitzar $u'Bu$ amb $u'Tu = 1$, on si suposem X centrades i Y la matriu d'indicadores de classe: $T = X'X$ i $B = (P_Y X)'P_Y X = XY(Y'Y)^{-1}Y'X$. Amb això el que fem (geomètricament) és discriminar pel subespai de màxima variància (estimada com a mitjana de les de cada classe ja que la suposem comuna, $\Sigma = T - B$), tallant amb ell les rectes que uneixen els centroïds. Substituint X per $\Sigma^{-1/2}X$ "esferem" les dades (en teminologia de Volle, 1981 [221]) i podem tallar pel punt mig i amb direcció perpendicular ja que ara la mètrica de referència té per matriu $\mathbf{I}$.

C-viii) **En terminología factorial l'*LDA* és equivalent** a fer l'anàlisi de $X'P_Y X$ sota mètrica T o bé l'anàlisi de $X'T^{-1}P_Y T^{-1}X$ sota mètrica

T^{-1} i en terminologia de triplets tindrem $\Big(P_Y X,\ (X'D_nX)^{-1},\ D_n\Big)$ si acceptem una ponderació D_n pels individus o bé l'equivalent

$$\Big((Y'D_nY)^{-1}Y'D_nX,\ (X'D_nX)^{-1},\ Y'D_nY\Big)$$

que procedeix de l'anàlisi canònic de la parella Y, X consistent a maximitzar $u'X'Yw$ sota $u'X'Xu = w'Y'Yw = 1$ (el que és equivalent a una canònica generalitzada amb $[X, Y] = [X_1, X_2]$).

C-ix) **L' anomenat anàlisi discriminant baricèntric** per a X dicotòmiques (proposat per Benzècri al 1982 i defensat per Palumbo al 1998) és un anàlisis de correspondències de perfil de fileres sobre les mitjanes de classes, i que, per tant, utilitza el triplet $\Big(P_Y X,\ D_p^{-1},\ D_n\Big)$.

C-x) **L'anàlisi discriminant de correspondències** (també per a X dicotòmiques, però amb $p = 2$) proposat per Chessel i Thioulouse al 1997 consisteix, simplement, a tornar al triplet del *LDA* (centrant el projector, això si) per al que s'ha de canviar la mètrica de $\mathbb{R}_p$ (fileres) de D_p^{-1} =dividir pels totals de categoria a $(X'D_nX)^{-1}$ = normalitzar per la covariància mostral. Els autors anomenen a això canviar la lògica canònica per una d'instrumental.

C-xi) Finalment, i per completar aquest resum conceptual previ al plantejament de la nostra proposta, referirem **l'anàlisi de Saporta** consistent en realitzar una anàlisi de correspondències múltiples (ACM) sobre X com pas ortogonalitzador previ a l' anàlisi discriminant sobre Y.

4.3.2 El Plantejament de la proposta

Les dificultats que els mètodes comentats a la secció anterior presenten provenen, en essència, de la necessitat d'equilibrar dos objectius que han estat tractats per separat: la discriminació pròpiament dita i la reconstrucció de les variables subjacents contínues mitjançant correspondències.

La nostra proposta consisteix a fer un anàlisi de correspondències múltiples "ponderada - iterada" en el que les variables seran pesades pel seu valor discriminant (màxima separació dels centroïds) i, posteriorment, recalcular els centroïds tenint en compte el resultats, iterant fins que arribem a la convergència.

Partim d'uns centroïds obtinguts simplement dels originals valors categòrics. La matriu a diagonalitzar cada vegada és $D^{-1} \circledast X'XA$ on A representa els pesos atribuïts a cada variable col·locats en forma diagonal i amb el mateix coeficient per a totes les categories de la mateixa variable. Aquesta matriu correspon, introduint una ponderació, a l'enfocament del punt **C**-vi de l'apartat anterior. Per altra banda, si utilitzem l'esquema de quantificació recíproca, $\psi = XA\zeta$ ens dóna les quantificacions dels individus com a suma ponderada de les que li atorguen cadascuna de les variables, i $\zeta = D^{-1}X'\psi$ representa les quantificacions variable a variable obtingudes per projecció de las dels individus sobre els espais determinats per cadascuna.

La matriu diagonal A que, com s'ha dit, recull els pesos atribuïts a cada variable s'obtindrà a partir dels coeficients de la combinació lineal d'aquestes que maximitze la variància dels centroïds (B) amb la variància total (T) igual a la unitat, és a dir mitjançant un *LDA* sobre les variables quantificades pel pas anterior.

Naturalment, en cada pas al variar A variarà ζ i el procès serà iteratiu. Caldrà, per tant asegurar-ne la convergència, el que farem a l'apartat 4.3.5 (pàg. 94).

La idea bàsica és que, mitjançant aquest procès, aconseguirem equilibrar l'efecte reconstructor de la Normal evidenciat per Lancaster pel correspondències simples amb la finalitat de la discriminació en un context multidimensional. En definitiva, al projectar sobre un espai on les variables són pesades pel seu poder discriminant és com realment aconseguim la normalització al reconstruir l'eix principal de les normals originals o, millor dit, al reconstruir el més acuradament possible la disposició relativa dels centroïds. Ambdós objectius van per tant lligats, com s'evidenciarà fent ús del resultat demostrat a continuació, on

la maximització de la correlació que Lancaster va emprar amb les correlacions canòniques simples es transfereix a una maximització de la variància sobre l'eix principal de la matriu de correlacions, R del cas multidimensional.

4.3.3 La fonamentació matemàtica: la generalització del teorema de Lancaster

Com esmentàvem a la secció 4.1 l'objectiu reconstructor d'una multinormal hauria de partir d'una generalització multivariable del teorema de Lancaster, però aquesta no és fàcil degut a que el que era un únic valor ρ és substituit per una matriu R i per tant hem de seleccionar quin aspecte d'aquesta s'ha de maximitzar. La resposta ve donada pel mateix objectiu discriminant que perseguim: ho farem en la direcció principal discriminant, és a dir en aquella que maximitzi la variància entre els centroïds de classe.

El fonament d'aquest procés s'estableix al següent teorema:

Teorema 4.1
Generalització multidimensional del teorema de Lancaster

Siguin $Z_i, i = 1, \ldots, p$ variables aleatòries normals tipificades amb R com a matriu de correlacions i siguin $X_i, i = 1, \ldots, p$ transformades de les Zi respectivament i també tipificades.

Si v és el vector propi corresponent al major valor propi de R aleshores:

$$\mathrm{Var}(w'X) \leq \mathrm{Var}(v'Z) \qquad \forall w \in \mathbb{R}_p \quad \text{amb} \quad w'w = 1$$

El significat d'aquest teorema en el nostre cas és clar: qualsevol combinació lineal de les variables normals discretitzades que maximitze la variància en la direcció del primer vector propi de R, va en la direcció de la reconstrucció de l'eix principal de la normal subjacent, donat que aquesta distribució és la que la té màxima entre totes les derivades d'una transformació seva variable a variable.

Per fer la demostració començarem per provar els següents lemes:

Lema 4.1 *Sigui R una matriu de correlació (de mida $(p\times p)$ amb element genèric r_{ij} i primer vector propi $v = (v_1, v_2, \cdots, v_p)$ i sigui R_n la matriu de dimensió $(np \times np)$ formada per caixes de mida $(n \times n)$ de manera que l'element r_{ij} ha estat substituït per:*

$$\begin{pmatrix} r_{ij} & & & 0 \\ & r_{ij}^2 & & \\ & & \ddots & \\ 0 & & & r_{ij}^n \end{pmatrix}$$

Aleshores: el primer vector propi de R_n és $(v_1, 0, \ldots, 0, v_2, 0, \ldots, 0, \ldots, v_p, 0, \ldots, 0)$.

Demostració *Lema* 4.1 Suposem que $w = (w_{11}, w_{12}, \ldots, w_{1n}, w_{21}, w_{22}, \ldots, w_{2n}, \ldots)$ fóra el primer valor propi de R_n.

Si anomenen $w_1 = (w_{11}, w_{21}, \ldots, w_{p1})$, $w2 = (w_{12}, w_{22}, \ldots, w_{p2})$, $\ldots$,$w_n = (w_{1n}, w_{2n}, \ldots, w_{pn})$ tindrem:

$$w' R_n\, w = w_1' R w_1 + w_2' R^2 w_2 + ... + w_n' R^n w_n$$

Ara bé donat que, per ser R una matriu de correlació, $R^i \leq R \quad \forall i > 1$ i que $\|w\| = 1$ és clar que maximitzar $w' R_n w$ és equivalent a maximitzar $w_1' R w_1$ amb $\|w_1\| = 1$, $w_2 = 0 \ldots w_n = 0$ i per tant $w_1 = v$ i

$$w = (v_1, 0, \ldots, 0, v_2, 0, \ldots, 0, \ldots, v_p, 0, \ldots, 0)$$

∎

Corol·lari 4.1 *Sigui R_n^{or} la matriu de correlació resultant de R_n per una reordenació de fileres i columnes resultant de substituir l'ordre*

$$(w_{11}, w_{12}, \ldots, w_{1n}, w_{21}, w_{22}, \ldots, w_{2n}, \ldots, w_{p1}, w_{p2}, \ldots, w_{pn})$$

per

$$(w_{11}, w_{21}, \ldots, w_{p1}, w_{12}, w_{22}, \ldots, w_{p2}, \ldots, w_{1n}, w_{2n}, \ldots, w_{pn})$$

El primer vector propi de R_n^{or} *serà* $(v, 0, \ldots, 0)$ *on* v *és el primer vector propi de* R*.*

A partir d'ara considerarem realitzada aquesta ordenació i per tant prendrem, per simplicitat notacional, $R_n = R_n^{or}$

Lema 4.2 *Sigui* R *una matriu de correlació (de mida* $p \times p$*) amb primer vector propi* v *i sigui* $E_{p\infty}$ *l'espai format per successions de vectors de mida* p *tals que:*

$$\|w\| = \sum_{i=1}^{\infty} w_i' w_i = 1 \quad \text{on} \quad w_i = (w_{i1}, \ldots, w_{ip})$$

Es compleix:

1. *La norma* $\|w\|_\infty = \sum_{i=1}^{\infty} w_i' R^i w_i$ *està ben definida a* $E_{p\infty}$*.*
2. *Si* v^* *és el primer vector propi de la mètrica definida per la norma* $\| \ \|_\infty$ *respecte a la mètrica definida per* $\| \ \|$*, és a dir si:*

   $$\|v^*\| = 1 \text{ i } \|v^*\|_\infty = \max \|w\|_\infty, \ w \in E_{p\infty}$$

   resulta $v^* = (v, 0, 0, \ldots)$

Demostració *Lema* 4.2

1. Sabem que $\lim_{i\to\infty} R^i = \mathbf{I}$, i per tant, $\lim_{i\to\infty} (R^{i+1} - R^i) = 0$ d'on com $w_i' w_i \leq 1 \quad \forall i$ aleshores, $\lim_{i\to\infty} (w_{i+1}' R^{i+1} w_{i+1} - w_i' R^i w_i) = 0$ el que ens garanteix la convergència de la sèrie i, com a conseqüència, $\| \ \|_\infty$ està ben definida a $E_{p\infty}$ ∎
2. Si definim $\|w\|_n = w' R_n w = \sum_{i=1}^{n} w_i' R^i w_i$ tenim que $\| \ \|_n \to \| \ \|_\infty$ i , per tant, si v_n^* és el valor propi de $\| \ \|_n$ $v_n^* \to v^*$ i , pel corol.lari 4.1 $v^* = (v, 0, 0, ...)$ ∎

Demostració *Teorema* 4.1

Generalització multidimensional del teorema de Lancaster

Siguin Z_{ik} els polinomis de l'Hermite de grau k de Z_i i a_{ik} els coeficients de X_i de la corresponent descomposició de manera que:

$$X_i = \sum_{k=1}^{\infty} a_{ik} Z_i^k \quad \text{amb} \quad \sum_{k=1}^{\infty} a_{ik}^2 = 1 \quad \forall i = 1, \ldots, n$$

Tenim que:

$$\begin{aligned} \mathrm{Var}(w'X) = \mathrm{E}((w'X)^2) &= \sum_{i,j} w_i w_j \mathrm{E}(X_i X_j) \\ &= \sum_{i,j} \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} w_i w_j \mathrm{E}(Z_i^{k_1} Z_j^{k_2}) \\ &= \sum_{i,j} \sum_{k=1}^{\infty} w_i w_j a_{ik} a_{jk} r_{ij}^k \end{aligned} \tag{4.7}$$

donat que $\mathrm{E}(Z_i^{k_1} Z_j^{k_2}) = r_{ij}^{k_1} \delta_{k_1,k_2}$

Si definim

$$M_{ij} = (r_{ij}, r_{ij}^2, r_{ij}^3, \ldots)$$

a l'espai d'Hilbert de les successions amb valors a $[-1, 1]$ i convergents a zero, i construïm amb elements d'aquest espai la matriu $(p \times p)$ M amb els corresponents M_{ij}, podem, per analogia amb les expressions d'un espai finit, definir:

$$s'Ms = \sum_{i,j} \sum_{k=1}^{\infty} s_{ik} s_{jk} r_{ij}^k$$

i el vector aw:

$$aw = \left(\{w_1 a_{1k}\}_{k=1}^{\infty},\ \{w_2 a_{2k}\}_{k=1}^{\infty}, \ldots, \{w_p a_{pk}\}_{k=1}^{\infty} \right)$$

Amb aquesta notació podem escriure l'expressió 4.7 com:

$$\mathrm{Var}(w'X) = (aw)' M aw \tag{4.8}$$

resultant que, si apliquem la reordenació expressada al corol.lari 4.1 sense canviar la notació, per evitar una complicació innecessària, $v^* = (v, 0, 0, \ldots)$ serà el primer vector propi de M corresponent al mateix primer valor propi de R, tal i com ens assegura el lema 4.2.

Per tant, per qualsevol altre vector i en particular per aw:

$$(aw)'M(aw) \leq v^{*\prime}Mv^*$$

donat que l'expressió del vector propi en un espai d'Hilbert conserva les propietats definides per a espais finits (Koster, 1989 [135]). Com a conseqüència, per (4.8):

$$Var(w'X) \leq v^{*\prime}Mv^*$$

I com:

$$\mathrm{Var}(v'Z) = \mathrm{E}((v'Z)^2) = \sum_{i,j} v_i v_j \mathrm{E}(Z_i Z_j) = \sum_{i,j} v_i v_j r_{ij} = v'Rv$$

per altra banda, per construcció de M i pel fet que $v^* = (v, 0, 0, ...)$:

$$v'Rv = v^{*\prime}Mv^*$$

tenim:

$$\mathrm{Var}(w'X) \leq \mathrm{Var}(v'Z)$$

∎

Completarem aquesta fonamentació matemàtica amb l'enunciat i demostració del següent:

Corol·lari 4.2 *Amb les mateixes condicions del Teorema 4.1*

$$\mathrm{Var}(w'X) \leq \mathrm{Var}(w'Z) \qquad \forall w \in \mathbb{R}_p \qquad \text{amb } w'w = 1$$

Demostració *Corol·lari* 4.2

Mitjançant l'habitual procés de projecció ortogonal sobre el subespai generat pels vectors propis previs s'aconsegueix estendre el resultat del Teorema 4.1 per a tots els vectors propis ortogonals (independents) $v_i,\ i = 1, \cdots, p$ de R . Aleshores qualsevol vector $w \in \mathbb{R}_p$ pot considerar-se com $w = \sum_{i=1}^{p} \tau_i v_i$ i, per tant, pel Teorema 4.1:

$$\mathrm{Var}(w'X) = \sum_{i=1}^{p} \tau_i^2 \mathrm{Var}(v_i'X) \leq \sum_{i=1}^{p} \tau_i^2 \mathrm{Var}(v_i'Z) = \mathrm{Var}(w'Z)$$

∎

4.3.4 L'algorisme *ADDSUC*

Si ara ens plantegem cercar una quantificació que maximitzi la variància en la direcció del primer vector propi a de la matriu de covariancies dels centroïds de classe, aconseguirem l'equilibri perseguit, ja que el component de la variància degut a l'efecte de classe (la provinent de B) quedarà maximitzada, i la deguda a la part comuna (Σ) serà reconstruïda com a residual per l'anàlisi de correspondències múltiple en la direcció de la normalitat.

És clar, del Teorema 4.1, que aquesta reconstrucció serà més "eficient" en la mesura que a s'apropi a la direcció del primer vector propi de la matriu de correlacions R (normalització de Σ).

Per altra banda, al desconèixer la situació contínua subjacent no podem calcular amb exactitud a, pel que ens introduirem en un procés iteratiu que, partint de les quantificacions habituals $1, \cdots, k_i\ i = 1, \cdots, p$, ens calculi una aproximació d'a, la qual anirem millorant posteriorment de manera iterativa. Si Q_j, R_j, a_j són, respectivament, les quantificacions, la matriu de correlacions i l'aproximació al vector a, aconseguits a una determinada iteració j, el procés iteratiu serà:

$$Q_0 \to R_0 \to a_0 \to Q_1 \to R_1 \to a_1 \cdots\cdots$$

La convergència d'aquest procediment es garanteix a 4.3.5 i ens assegura trobar finalment unes quantificacions Q i un vector a, de manera que aquest sigui el primer vector propi dels centroïds determinats per aquelles, havent aconseguit reconstruir, amb dimensió 1, el més acuradament possible, la situació subjacent de partida.

La resta de les dimensions les obtindrem per l'habitual procés canònic que es detindrà quan no s'obtingui cap millora significativa a l'error real final.

Finalment, si fem intervenir la suavització proposada a la secció 3.6 (pàg. 60) l'algorisme del mètode proposat es pot esquematitzar com:

1a fase: Realitzar una anàlisi de correspondències múltiple ponderada-iterada que de forma canònica obtindrà quantificacions amb ponderacions per a les variables que correspondran a l'eix principal de la descomposició de la variància entre els centroïds de classe (B). Aquesta anàlisi serà descrita en detall a l'apartat següent.

2a fase: Aplicar una suavització *EM* (secció 3.6, pàg. 60) sobre les quantificacions aportades per la fase 1 per a reconstruir el més acuradament les normals subjacents a cada classe i, finalment, procedir a una discriminació *LDA-canònica*

4.3.5 La convergència de l'algorisme *ADDSUC*

Naturalment, una vegada fonamentat matemàticament el mètode mitjançant el teorema 4.1 i el corol·lari 4.2, hem de garantir l'existència de les solucions que proposa, el que, tractant-se a la primera fase d'un algorisme iteratiu, significa provar la seva convergència.

Per a aquest objectiu emprarem com a base el resultat donat per Gifi (1990) [72], capítol 3.5, en el que s'utilitza un mètode de *Alternating Least Squares* (ALS) convergent per a l'anàlisi de correspondències múltiple o canònica generalitzada.

La diferència que presenta *ADDSUC* està en la ponderació establerta i en el fet de que aquesta es fa, al seu torn, depenent d'un altre procés canònic.

Es a dir, mitjançant Gifi sabem que si les quantificacions provinguesin de la equació:

$$X'D^{-1}X\psi = \mu\psi \quad \text{amb} \quad \zeta_i = P_{X_i}\psi$$

aleshores l'**algorisme 1**:

1. Seleccionem un valor inicial ψ_0 , un valor de tolerància ε i fem $r = 0$.
2. $\zeta_r = D^{-1}X'\psi_r$
3. $\psi_{r+1} = X\zeta_r$, amb normalització per a que $\psi'_{r+1}\psi_{r+1} = 1$
4. Si $\|\psi_{r+1} - \psi_r\| > \varepsilon$ fem $(r+1) \to r$ i tornem a 2.

ens dona la successió convergent $\psi_r \to \psi$ (si $\varepsilon \to 0$).

Per altra banda i basant-se en un principi simètric (comencem per les quantificacions de les variables en lloc de les dels individus), l'**algorisme 2**:

1. Seleccionem un valor inicial a_0, un valor de tolerància ε i fem $s = 0$
2. $\tau_s = Za_s$
3. $a_{s+1} = Z'P_Y\tau_s$ amb normalització per a que $a'_{s+1}a_{s+1} = 1$
4. Si $\|a_{s+1} - a_s\| > \varepsilon$ fem $(s+1) \to s$ i tornem a 2.

ens dona la successió convergent $a_s \to a$ on a és la solució principal de $ZP_YZ'a = \lambda a$.

En el nostre cas, es tracta de provar la convergència de l'**algorisme *ADDSUC***:

1. Seleccionem un valor inicial de quantificacions de variables $\zeta_0 = (\zeta_{01}, \zeta_{02}, \ldots, \zeta_{0p})$, un valor de tolerància ε i fem $n = 0$ i $Z_0 = (X_1\zeta_{01}, X_2\zeta_{02}, \ldots, X_p\zeta_{0p})$.

2. Trobem a_n tal que $Z_n' P_Y Z_n a_n = \lambda a_n$ amb $a_n' a_n = 1$

3. Construim A_n com una matriu diagonal obtinguda per expansió d'a en el sentit següent: per cada categoria j de la variable i fem $A_{j(i),j(i)} = a_i$, és a dir li atribuïm la ponderació corresponent al valor adjudicat a la seva variable a l'apartat anterior.

4. Trobem ψ_{n+1} tal que $X'D^{-1}A_nX\psi_{n+1} = \mu\psi_{n+1}$ amb $\psi_{n+1}'\psi_{n+1} = 1$

5. Si $\|\psi_{r+1} - \psi_r\| > \varepsilon$, calculem les quantificacions de les variables per projecció de les dels individus $\zeta_{n+1} = D^{-1}X'\psi_{n+1}$, fem $(n+1) \to n$ calculem $Z_n = (X_1\zeta_{n1}, X_2\zeta_{n2}, \ldots, X_p\zeta_{np})$ i tornem a 2.

En primer lloc hem de garantir que el resultat de la convergència de l'**algorisme 1** per a $X'D^{-1}X\psi = \mu\psi$ pot estendre's al cas $X'D^{-1}AX\psi = \mu\psi$ amb A matriu diagonal. Això és evident quan A és definida positiva substituint X per $A^{1/2}X$ i per al cas que no ho sigui, aplicarem el conegut resultat segons el qual $M'M$ té els mateixos vectors propis "per la dreta" que M i valors propis que són el quadrat dels d'aquesta matriu. Aleshores, procedirem amb l'equació $X'D^{-1}A^2D^{-1}X\psi = \mu^2\psi$ substituint X per $AD^{-1/2}X$.

Ara bé, Gifi fa recaure la convergència dels seus algorismes amb el fet de que la funció de pèrdua definida per la suma de quadrats de les diferències entre la quantificació directa dels individus i l'obtinguda mitjançant les variables, està acotada per zero i disminueix a cada passa de l'algorisme.

Aquesta funció de pèrdua seria per l'**algorisme 1**:

$$\ell_1(\psi, \zeta) = \sum_{j=1}^{p} \sum_{i=1}^{n} (\psi_i - X_{ij}\zeta_j)^2$$

i per l'**algorisme 2**:

$$\ell_2(\tau, a) = \sum_{j=1}^{p} \sum_{i=1}^{n} (\tau_i - Z_{ij}a_j)^2$$

i si definim per a *ADDSUC* (fent la descomposició del pas 4 segons l'algorisme 1 amb l'observació relativa a les ponderacions ja comentada i la del pas 2 segons l'algorisme 2):

$$\ell(\psi,\zeta,\tau,a)=\sum_{j=1}^{p}\sum_{i=1}^{n}(\psi_i-X_{ij}\zeta_j)^2+\sum_{j=1}^{p}\sum_{i=1}^{n}(\tau_i-Z_{ij}a_j)^2$$

Es clar que ℓ està acotada inferiorment per zero i que a l'augmentar n disminueix, donat que al pas 2 ho fa la segona part del segon membre (ℓ_2) mantenent-se fixa la primera part (ℓ_1) i al pas 4 ocorre a l'inrevés.

Això ens garanteix la convergència de l'algorisme *ADDSUC*. ∎

Capítol 5

Resultats numèrics

En aquest capítol provarem el mètode *ADDSUC* comparant-lo primer, per simulació, amb els d'estructura semblant (secció 5.2) i després amb el mètode més utilitzat actualment per fer anàlisi discriminant discreta: la logística-xarxes neurals (secció 5.3) i finalment amb els dos tipus emprant dades reals (secció 5.4)

5.1 El fluxegrama de l'*ADDSUC*

Presentarem, en primer lloc a la figura 5.1, el fluxegrama del mètode ADDSUC que hem utilitzat com a base per a la seva programació.

En quant a la interpretació dels símbols que hi apareixen, comentarem que a l'entrada, xd, ha de ser una matriu amb tantes fileres com a individus i columnes corresponents a les variables classificadores; coincideix exactament a la matriu X de la pàgina 5. El vector yd contindrà la classe corresponent a cada individu i es correspon exactament a la Y de la mateixa secció. Ambdues es refereixen a les dades d'aprenentatge mentre que xct i yct són les equivalents per a les mostres de test. En quant als paramètres: g representa el nombre de classes, l és un vector amb el nombre de categories de cada variable i nf és el nombre d'eixos a considerar si és que aquest es vol fixar. En cas de que sigui 0 el programa l'estimarà per validació creuada mínimo-quadràtica. Per altra banda, fr és la

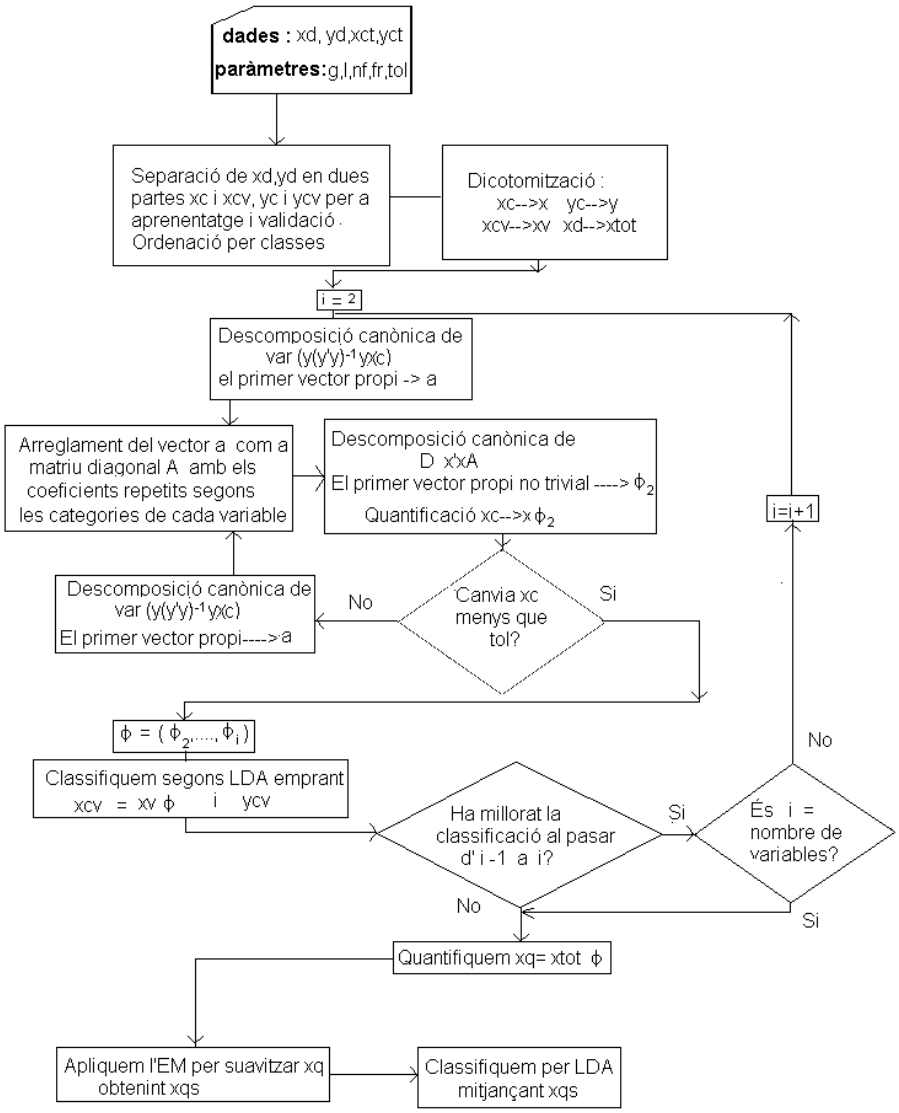

Figura 5.1: *Fluxegrama del mètode ADDSUC*

fracció de les dades d'aprenentatge que s'emprarà per a aquesta validació creuada (arrodonint la freqüència resultant al nombre enter més pròxim). Si $nf = 0$ no s'utilitzarà validació creuada i es farà $fr = 1$, prenent totes les dades d'aprenentage per a l'estimació. Finalment, *tol* representa l'umbral de tolerància per a la convergència de l'algorisme i s'utiliza de la manera que queda reflexada al diagrama. Per defecte es pren $tol = 0.0001$. Els paràmetres g i l són també calculats pel programa, però es considera més convenient que siguin donats per l'usuari a fi i efecte de detectar possibles errors informant mitjançant un missatge d'aquesta circumstància per a donar l'oportunitat de corregir-la.

5.2 Comparació amb els mètodes d'estructura semblant

Començarem les proves de l'*ADDSUC* amb aquesta secció, la qual està dedicada a la comparació amb els mètodes que, com ell, utilitzen l'anàlisi de correspondències i/o la suavització mitjançant l'algorisme *EM*.

5.2.1 Selecció dels conjunts de dades per fer les simulacions de prova del mètode

Per triar un conjunt de dades simulades que, sense pretensions d'exhaustivitat, pugui ser representat d'una gamma prou àmplia de situacions, hem de seleccionar uns criteris que ens permetin valorar la "peculiaritat" classificatòria de cadascun d'ells.

Després de fer una revisió a la literatura sobre mesures prèvies de separabilitat de classes, on destaquen les propostes recollides a Hand (1981) [107], que tenen el problema de limitar-se a la separabilitat entre dues classes (i que es reduiran a la distància de Mahalanobis en el cas que ens ocupa), arribem a la conclussió de que les paràmetres més adients són:

1. Nombre de variables i de categories com a mesura de la complexitat inicial tractada a la secció 1.2.3, pàg. 16.

2. L'error òptim continu *ec*, descrit a l'apartat 1.2.2.1, pàg. 9 que ens reflecteix el grau de solapament entre classes del que partim.

3. El percentatge de variància entre classes absorbit pel primer eix, que ens reflecteix el grau aproximat de la dimensionalitat dels centroïds (prop del 100% indica un grau aproximat d'1, mentre que per sota del 90% podem considerar aquest grau major d'1). Altres mesures de dimensionalitat poden aplicar-se però triem aquesta per la seva relació en el plantejament canònic que se segueix al llarg d'aquest estudi.

Tenint en compte tot això hem seleccionat els següents conjunt de dades tots amb 3 classes de pesos $0.2, 0.3, 0.5$:

Conjunt 1 *Nombre de Variables:* 3, *Nombre de Categories:* 9

Mitjanes M_1	**Variàncies** V_1	**Talls per a la discretització** T_1	$\mathbf{e_c}$	**Pes 1r eix**
$\begin{bmatrix} (-0.5,-0.3,0.1) \\ (0.1,0.2,0.2) \\ (0.7,0.6,0.7) \end{bmatrix}$	$\begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & 0.5 \\ 0.3 & 0.5 & 1 \end{pmatrix}$	$\begin{bmatrix} (-0.3) \\ (-0.6,0.4) \\ (-0.3,0.4,.8) \end{bmatrix}$	0.48	99%

Conjunt 2 *Nombre de Variables:* 3, *Nombre de Categories:* 8

Mitjanes M_2	**Variàncies** V_2	**Talls per a la discretització** T_2	$\mathbf{e_c}$	**Pes 1r eix**
$\begin{bmatrix} (-0.5,-0.5,-0.25) \\ (0.5,0.5,0) \\ (1,-0.5,0.25) \end{bmatrix}$	V_1	$\begin{bmatrix} (0,0.75) \\ (0) \\ (-0.125,0.125) \end{bmatrix}$	0.27	75%

Conjunt 3 *Nombre de Variables:* 2, *Nombre de Categories:* 5

Mitjanes M_3	**Variàncies** V_3	**Talls per a la discretització** T_3	$\mathbf{e_c}$	**Pes 1r eix**
$\begin{bmatrix} (-2,2) \\ (2,2) \\ (4,2) \end{bmatrix}$	$\begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix}$	$\begin{bmatrix} (0,3) \\ (0) \end{bmatrix}$	0.23	80%

Conjunt 4 *Nombre de Variables:* 3, *Nombre de Categories:* 8

Mitjanes M_4	**Variàncies** V_4	**Talls per a la discretització** T_4	$\mathbf{e_c}$	**Pes 1r eix**
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%

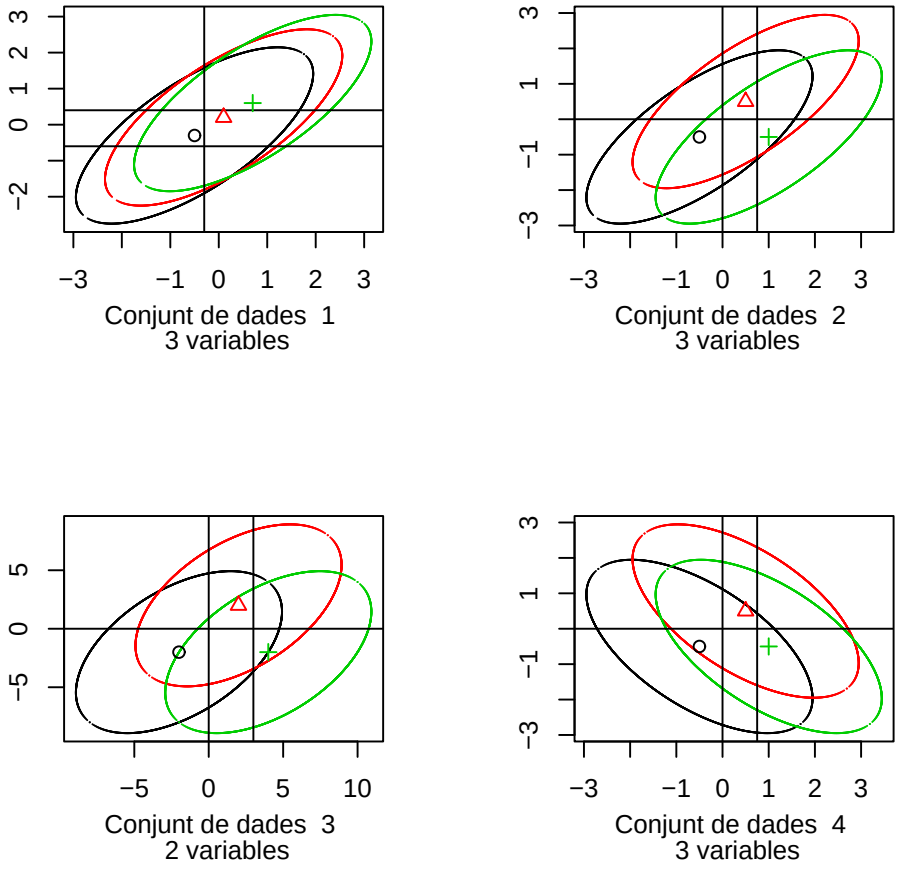

Figura 5.2: *Variables 1 i 2 dels conjunts de dades simulades*

Aquests conjunts de dades poden representar-se gràficament utilitzant els el·lipsoïds del 95% corresponents a la distància de Mahalanobis. A la figura 5.2 podem observar les representacions cartesianes de les seves dues primeres variables, amb els corresponents centroïds i rectes horitzontals i verticals de tall.

Convé aclarir que els conjunts de dades 2 i 3, que al gràfic semblen similars, no ho són, donat que el primer disposa de tres variables i el segon només de dues.

5.2.2 Selecció dels mètodes d'estructura semblant a comparar

Una vegada seleccionat els conjunt de dades, cal decidir quins seràn els mètodes d'estructura semblant a comparar. Tenint en compte la revisió del capítol 1, és clar, que en el nostre cas, els mètodes de referència han de ser l'*LDA-Canònic* i l'*MDA*. Ja dins del àmbit de les correspondències utilitzarem les variants esmentades a la secció 4.2 (pàg. 78), llevat de la de Chessel-Thioulouse donat que, en el cas multivariable, ens condueix, com hem provat, a l'*LDA-Canònic*. Afegirem quatre possibilitats més (a banda d'*ADDSUC*), que descriurem breument a continuació, amb altres exploracions que vam fer per utilitzar l'anàlisi de correspondències amb objectiu discriminant. Per tant el quadre de mètodes a comparar serà:

1. *LDA-Canònic*.
2. *MDA*.
3. Bassats en correspondències:
   (a) Benzècri-Palumbo: Consisteix en la descomposició canònica de la matriu $(P_Y X)' P_Y X \circledast D^{-1}$. És el també anomenat anàlisi inter-clases. Complementàriament es pot fer l'intra-classes que descomposaria la matriu $(X - P_Y X)'(X - P_Y X) \circledast D^{-1}$
   (b) Saporta-Volle: Un anàlisi de correspondències complert amb la matriu $X'X \circledast D^{-1}$ seguit d'una correlació canònica simple amb Y.
   (c) Mètode *ADDSUC*.
   (d) Altres Exploracions pròpies basades en les correspondències:
      i. Descomposició de $(X'X - (P_Y X)' P_Y X) \circledast D^{-1}$.

ii. Correspondències incluint Y que després se suprimeix per al càlcul de les quantificacions.

iii. Descomposició de $X_i'X_i \circledast D_i^{-1}, \quad i = 1, \ldots, g$, seguida d'una ponderació per les freqüències relatives a cada categoria dins de cada classe.

iv. Descomposició de $(X - P_Y X)'(X - P_Y X) \circledast \Omega^{-1}$.

La pauta de comparació s'establirà en funció de l'error e_{fr} donat que com es va explicar a la secció 1.2.2.4 (pàg. 13) representa un error que, a diferència de l'aparent, ens permeteix valorar si s'alcançat un adequat equilibri amb la complexitat (figura 1.4, pàg. 17). Utilitzarem mostres de 500 dades dividides aleatòriament amb dues meitats, una per estimar i l'altra per ajustar emprant la validació creuada. També utilitzarem mostres de test de la mateixa mida.

Necessitarem, també, dues mides de les repeticions neccessaries per estimar l'error: Una de repeticions de l'extracció originària de la mostra d'aprenentatge i un altra de repeticions de les mostres de test una vegada fixa la mostra d'aprenentatge (veure secció 1.2.2.3, pàg. 10). Per ambdós casos utilitzarem el valor 50.

5.2.3 Resultats comparatius de les simulacions

Els resultats es poden resumir amb el següent quadre d'errors finals reals, e_{fr}:

Conjunt de dades	LDA-canònic	MDA	ADDSUC	Benzècri	Saporta	Millor AEP
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)**

on la columna de *Millor AEP* conté els errors mínims per a cada conjunt de dades de entre tots els mètodes resenyats a l'item 3.(d) de l'apartat anterior com altres exploracions pròpies de discriminació mitjançant anàlisi de correspondències. Entre parèntesi figura la identificació del mètode que ha obtingut aquest mínim.

Cal assenyalar que *ADDSUC* supera a tots els mètodes en tots els conjunts de dades, obtenint sobre el següent mètode (Saporta) un avantatge relatiu del 7% i sent igualat només per altre mètode d'exploració pròpia (ii) en el quart conjunt de dades. Donat que aquest darrer obté front a *ADDSUC* uns errors superiors al 15% en mitjana no sembla convenient retenir-ho com a mètode comparable. En canvi el tercer (després de *ADDSUC* i Saporta) l'*MDA* si que el tindrem en consideració per a les comparacions amb dades reals, donat que representa l'aplicació directa de l'algorisme *EM* (la segona part d'*ADDSUC*) i ens dóna una idea molt clara de com influeix la primera part (correspondències) sobre el resultat final.

És molt important indicar que per evitar que l'efecte de l'ordre implicit restés generalitat als resultats s'han permutat les categories 1 i 2 de cada variable. Es a dir si els punts de tall són , per exemple, -0.6 i 0.4 (T_1 ,2a variable) el valor 1 correspondrà al interval $(-0.6, 0.4]$, el valor 2 a l'interval $(-\infty, -0.6]$ i el valor 3 a l'interval $(0.4, \infty)$.

En aquestes condicions més difícils (observem com augmenta significativament e_{fr} amb relació a e_c als conjunts de dades 2, 3 i 4) on s'ha desfet l'ordinalitat subjacent, és on els mètodes basats en les correspondències tenen la oportunitat de demostrar les seves propietats reconstructores.

No hem afegit els resultats quan no hi ha permutació ja que en aquest cas *MDA*, *ADDSUC* i Saporta (els principals mètodes a comparar) donen errors gairebé equivalents; ni hem ressenyat els de la combinació de l'*MDA* amb les altres posibilitats d'aplicació de correspondències, ja que no aporten cap millora.

5.3 La comparació amb el mètode híbrid logística-xarxes neurals

Una vegada assegurat que el mètode *ADDSUC* supera amb claredat tant als mètodes que només utilitzen Correspondències (del que destaquem el proposat per Saporta), com al que només utilitza l'*EM* (l'*MDA*) i a qualsevol combinació entre ells, hem de passar a comparar amb els mètodes que actualment destaquen com els més eficients per a l'anàlisi discriminant discreta.

Entre ells sobresurt pels seus bons resultats l'adaptació de la logística emprant la idea de les xarxes neurals (*LXN*) (secció 1.3.3, pàg. 21).

Es tracta, també, d'un mètode mixt que utilitza la filosofia del de xarxes neurals de la manera que ADDSUC utilitza la suavització mitjançant *EM* emprant per a la quantificació prèvia una logística en lloc d'un correspondències. Això li dona una gran potència i versatilitat i el fa el recomanat a hores d'ara per la majoria dels autors que no formen part d'una escola específica.

Vam començar per fer la comparació amb els conjunts de dades referits a la secció anterior obtenint una sorprenent igualtat (amb només diferències a nivell de la quarta xifra decimal el que nos la fa significatives) amb tots els casos.

Aquesta situació que podríem qualificar "d'empat" encara que prou estimulant , donat que *ADDSUC* com veurem a l'apartat de suggeriments (pàg. 116), és encara un mètode acabat de nàixer i amb moltes possibilitats de millora i ajustament, ens deixava en el punt d'intentar esbrinar on podríem haver-hi diferències que ens guiessin en les posteriors recerques.

Front a la alternativa de "pertorbar" els conjunt de dades cercant petites diferències les quals, tractant-se de simulacions, no serien massa rellevants i es podrien atribuir a peculiaritats específiques, vam optar per emprar per a la comparació , per una banda les dades reals que analitzarem a la següent secció i, per altra, les dades de referència que són la pauta de comparació per a tots els nous mètodes de discriminació: les aportades per Fisher sota el nom d'*IRIS*.

Es tracta d'un conjunt molt estudiat de 150 individus dividits en tres classes (de 50 membres cadascuna) i quatre variables contínues: longitud i amplada dels sèpals i longitud i amplada dels pètals.

Per aplicar-hi una anàlisi discriminant discreta hem de procedir a la discretització d'aquestes variables. Amb aquest objectiu farem primer els corresponents histogrames:

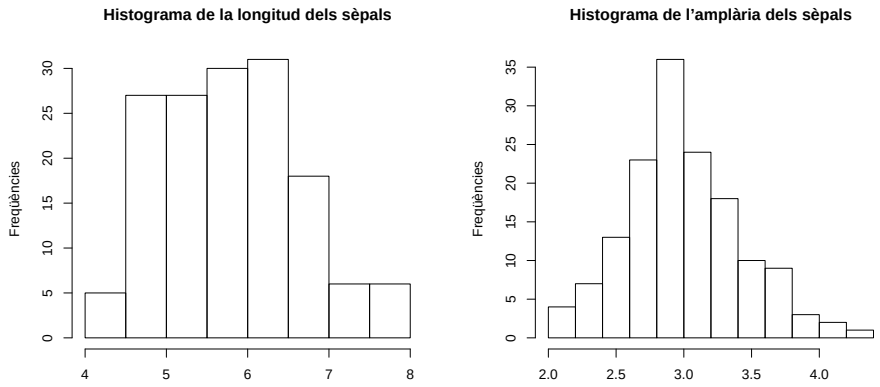

Figura 5.3: *Histogrames de les dades d'IRIS (sèpals)*

Observant els histogrames de les figures 5.3 i 5.4, resulta clar que les dues variables relatives als pètals tenen punts de tall evidents als centres dels corresponents valls: 2.3 per a la longitud i 0.7 per a l'amplària. En canvi, les variables corresponents als sèpals no presenten cap tall clar donat que hem de considerar les variables a les seves marginals sense tenir en compte per res les espècies (variable classificadora). Per això si volen considerar-les com a variables discretes, fet que implica tenir almenys dues categories, el més natural sembla tallar per la mediana en ambdós casos.

Procedint, doncs, d'aquesta manera i prenent 75 dades a l'atzar amb 25 de cada espècie tal i com va fer Ripley, 2002 [218] com a mostra d'aprenentatge

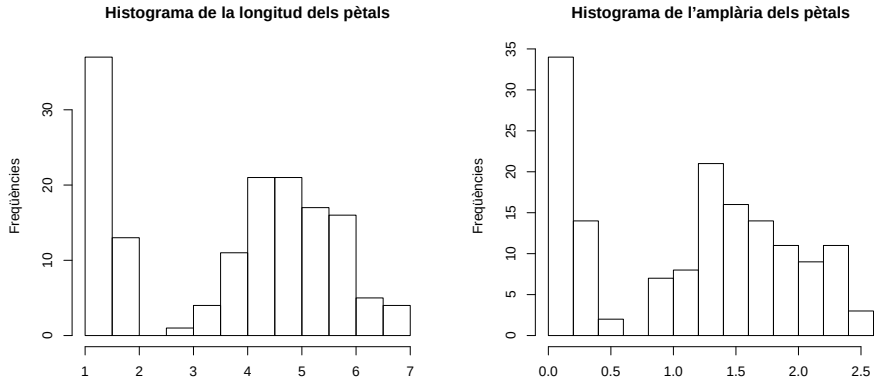

Figura 5.4: *Histogrames de les dades d'IRIS (pètals)*

deixant la resta com a dades de test, vam fer 1000 repeticions i vam comparar els resultats.

En 53 de les repeticions va ser millor l'*ADDSUC* i en la resta es va produir un empat exacte (mateix nombre de ben classificats) amb un error final mitjà per l'*ADDSUC* d'0.217 i per la Logística-xarxes neurals de 0.223 pel que podem dir que, en aquestes condicions, l'*ADDSUC* supera "uniformement" a la Logística-Xarxes neurals.

5.4 Comparació amb dades reals

Finalment farem una comparació dels mètodes que millor resultat han donat a les simulacions: *MDA*, *ADDSUC*, Saporta i logística-xarxes neurals (*LXN*) amb dos conjunts de dades reals.

5.4.1 Les dades de l'estudi de màrqueting

Aquest conjunt de dades ho vam extreure mitjançant una cerca per Internet de dades emprades com a prova dels mètodes d'anàlisi dicriminant. Es tracta de dades provinents d'un estudi de màrqueting (veure [126]) sobre 9409 residents a San Francisco (California) i ens serveix per estudiar el mètode en situacions reals d'una gran quantitat de variables i dades molt més extensa que la provada a les simulacions.

Les variables i les seves categories són descrites a l'apèndix A.

Després de la supressió dels casos amb alguna dada faltant es va realitzar l'anàlisi discriminant amb 2000 dades d'aprenentatge i 4000 de test obtenint els següents errors finals reals:

$MDA = 0.392$ $ADDSUC = \mathbf{0.363}$ $Saporta = 0.404$ $LXN = 0.385$

En termes absoluts això significa que $ADDSUC$ classifica bé 2548 casos dels 4000 mentre que el mètode que el segueix (logística-xarxes neurals) ho fa amb només 2460.

5.4.2 Les dades del projecte *AFIPE*

El segon conjunt de dades correspon al tipus dels que han servit per motivar aquest treball. Es tracta d'un pilotatge del projecte *AFIPE* (Anàlisi dels Factors Influents en el Patró d'Evolució de les malalties) que forma part del SISNICA (Sistema de Informació Sanitària de Nicaragua) desenvolupat durant el període 1990-1994 (veure [175]).

Es tracta de 1144 persones de les que les variables i les seves categories són descrites a l'apèndix B.

Aplicant l'anàlisi discriminant amb 550 dades d'aprenentatge i 594 com a dades de test s'obtenen els següents errors finals reals:

$MDA = 0.387$ $ADDSUC = \mathbf{0.318}$ $Saporta = 0.461$ $LXN = 0.388$

El que, en termes absoluts, significa que *ADDSUC* classifica bé 405 dels 594 mentre que el mètode que el segueix (l'*MDA*) ho fa només en 364.

5.5 Comentaris dels resultats

◇ En primer lloc hem de destacar que el mètode proposat: l'*ADDSUC* (Anàlisi discriminant discreta mitjançant suavització de les correspondències múltiples) sembla presentar un avantatge significatiu sobre qualsevol mètode que es basi amb suavitzacions mitjançant *EM*, correspondències o una combinació dels dos procediments, quan les dades provenen d'una multinormal discretitzada.

◇ Aquest avantatge es fa especialment rellevant quan s'ha aplicat una permutació a l'ordre natural de les discretitzacions, situació que pot considerar-se prou freqüent a la pràctica quan les variables arriben a l'investigador desprovistes de qualsevol indicació ordinal, el que ocorreix en la gran majoria dels casos de les recerques sanitàries a les que es feia referència a la introducció com a motivació d'aquest treball.

◇ Naturalment, si les dades no poden considerar-se provinents d'un model com el que aquí s'ha suposat (secció 1.1, pàg. 5) no podem assegurar la permanència d'aquest avantatge, però les proves amb dades reals semblen confirmar que les suposicions són d'un abast prou ampli a la pràctica.

◇ Per altra banda si comparem l'*ADDSUC* amb el mètode considerat més avançat actualment per realitzar l'anàlisi discriminant discreta: el perfeccionament de la logística basat en les xarxes neurals, observem un lleuger avantatge de l'*ADDSUC* si les discretitzacions s'han realitzat als punts de talls naturals: als valls de les distribucions marginals (secció 5.3).

◇ Finalment, les proves realitzades amb dades reals d'una certa complexitat, una procedent de dades analitzades per J.Friedman que és a l'abast dels investigadors mitjançant *Internet* [126] i una altra procedent de l'experiència pròpia amb dades epidemiològiques, inviten a la continuació de la recerca en la línia iniciada, donat que el mètode proposat aconsegueix els millors resultats amb una diferència significativa.

5.6 Aspectes computacionals

Els programes tant per l'ús convencional de l'*ADDSUC* com per a la seva prova emprant simulacions, han estat realitzats mitjançant **R** versió 1.7.1, ja que s'ha convertit en el mitjà habitual de programació *GNU* en estadística.

En **R** disposàvem del paquet *mda* desenvolupat per Hastie (2002) [119], de la llibreria *MASS* dissenyada per Ripley (2002) [218] amb la subrutina *mvrnorm* que hem emprat para la simulació de multinormals i del paquet *nnet*, que disposa de les rutines asociades a la metodologia de xarxes neurals, d'on hem extret l'anomenada *multinom*, que realitza la logística-xarxes neurals.

La prova dels altres mètodes inspirats en correspondències, tant els de Bènzecri i Saporta com els que hem anomenat d'exploració pròpia, han estat programades directament en **R**, ja que aquest llenguatge ens proveix d'una potència de programació i d'una simplicitat d'ús considerable.

S'ha de comentar, també, que el temps de processament no és cap entrebanc, ja que en totes les proves realitzades, la convergència de l'algorisme *ADDSUC* no ha requerit més de 10 iteracions.

Tots el programes i dades utilitzats en aquest capítol es troben, comprimits, al enllaç **Programes R. ADDSUC** de la plana web `www.uv.es\~msen`.

Conclusions i línies de recerca

Resumirem aquí, breument, les conclusions de l'estudi i els suggeriments per ampliar la recerca.

A Conclusions

En aquest treball hem procedit a realitzar una revisió sintetitzadora i unificadora de la teoria i dels procediments tant de l'anàlisi discriminant com dels mètodes de correspondències i de suavització.

Posteriorment, s'ha desenvolupat i fonamentat una nova metodologia per realitzar l'anàlisi discriminant discreta estructurada en dues fases: a la primera es procedeix a quantificar emprant una anàlisi de correspondències múltiples ponderada-iterada i a la segona es porta a terme una suavització mitjançant l'algorisme *EM*.

La prova del mètode amb dades simulades utilitzant un model de Normals subjacents amb mitjana diferent per classe i variància comuna, pot considerar-se positiva, ja que els seus resultats superen els altres procediments amb què s'ha comparat (seccions 5.2 i 5.3).

En la nostra opinió aquests esperançadors resultats es deuen a la solidesa del resultat matemàtic provat a la secció 4.3.3 (pàg. 88), el qual ens garanteix que la reconstrucció de les dades subjacents contínues es realitza en la direcció correcta.

Si a això s'afegeix que la suposició d'una multinormal subjacent pot considerar-se el final d'un ampli ventall de processos investigadors quan, finalment, s'aconsegueix destriar la part rellevant de la que no ho és (en termes probabilístics), no ens ha de sorprendre que un mètode, basat en aquestes premisses, obtingui bons resultats pràctics, tal i com succeeix als dos exemples reals analitzats.

S'ha de tenir en compte, també, que la quantificació proposada pot utilitzar-se no només amb objectius classificatoris sinò amb intencions descriptives i comparatives.

Per totes aquestes raons, considerem que la metodologia desenvolupada, la qual, programada en llenguatge **R**, es posa a la disposició dels investigadors a la pàgina web `www.uv.es\~msen`, representa una aportació a tenir en compte dins del camp de l'anàlisi discriminant discreta.

B Suggeriments i possibilitats de millora

Una possibilitat que ha estat explorada en la realització d'aquest estudi, però que necessita més treball, tant teòric com pràctic, consisteix a fer posteriorment a l'anàlisi de correspondències una anàlisi canònica generalitzada, agrupant totes les quantificacions (per eixos) d'una mateixa variable dins del mateix bloc. D'aquesta manera trobaríem, per a cada variable, la combinació lineal de les aproximacions dels seus polinomis de l'Hermite que millor es projectés sobre la combinació global. Això seria semblant a un *FDA* (consultar secció 1.3.4, pàg. 22), donat que fem una expansió polinòmica amb selecció posterior, i hem provat que, en alguns casos, millora el resultats. Així cobriríem també la possibilitat que la matriu de covariàncies Σ fóra diferent per classe, ja que el *QDA* corresponent seria inclòs dins l'esmentada expansió polinomial.

Una altra ampliació perfectament factible del mètode s'esmentava quan precisàvem la situació en estudi secció 1.1 (pàg. 5), i consistiria a ampliar la consideració que les distribucions per classe són Normals incorporant la possibilitat

que puguin ser mixtures de Normals, el que és perfectament compatible amb la utilització del *EM* en la segona fase del procés, seguint un esquema similar al de l'*MDA*.

També es pot tenir a l'abast la possibilitat d'utilitzar suavització amb el *Kernel* adaptable, explicat a la secció 3.5 (pàg. 59), el que ens permetria donar al mètode una major flexibilitat, al poder-se emprar amb un conjunt de funcions de densitat per classe més àmplia que no es limiti a Normals o mixtura de Normals.

En quant al cas de l'anàlisi mixt (X conté variables categòriques i contínues) es proposa investigar la possibilitat d'incloure al procés iteratiu les variables contínues o bé emprar aquestes com a covariables.

Finalment, cal també assegurar un tractament acurat de les dades incomplertes, adaptant els procediments desenvolupats amb aquest objectiu, i provar la sensibilitat del mètode a les quantificacions de partida, cercant un procediment ràpid (permutant, per exemple) que ens donés la que té menor error aproximat inicial.

Aquests són els aspectes per on se suggereix que hauria de continuar la recerca amb perspectives que ens semblen prou positives.

Apèndixs

A Descripció de les categories de les dades de màrqueting

Les variables són:

Y = Nivell d'Ingressos anuals familiars amb categories:

1.- Menys de \$20,000
2.- De \$20,000 a \$40,000
3.- Més de \$40,000

X_1 = Gènere amb categories:

1.- Home
2.- Dona

X_2= Estat civil amb categories:

1.- Casat
2.- Unió estable de fet
3.- Divorciat o separat
4.- Vidu
5.- Solter

X_3 = Edat amb categories:

1.- 14 fins 17
2.- 18 fins 24
3.- 25 fins 34
4.- 35 fins 44
5.- 45 fins 54
6.- 55 fins 64
7.- 65 i més

X_4 = Nivell educatiu amb categories:

1.- Grau 8 or menys
2.- Graus 9 a 11
3.- Graduat de l'Institut (High school)
4.- 1 a 3 anys d'Universitat
5.- Graduat universitari
6.- Amb estudis de postgrau

X_5 = Occupació amb categories:

1.- Profesional/Gerent
2.- Vendedor
3.- Obrer/Conductor
4.- Clergat/Treballadors de Serveis
5.- Ama de casa
6.- Estudiant
7.- Militar
8.- Retirat
9.- Aturat

X_6 = Anys de residència a la zona amb categories:

1.- Menys d'un any
2.- D'un a tres anys
3.- De quatre a sis anys
4.- De set a deu anys
5.- Més de deu anys

X_7 = Hi ha dos o més ingressos a la família? amb categories:

1.- No casat
2.- Si
3.- No

X_8 = Nombre de persones en la família amb categories:

1.- Una
2.- Dues
3.- Tres
4.- Quatre
5.- Cinc
6.- Sis
7.- Set
8.- Vuit
9.- Nou o més

X_9 = Nombre de persones de menys de 18 anys en la família amb categories:

1.- Cap
2.- Una
3.- Dues
4.- Tres
5.- Quatre
6.- Cinc
7.- Sis
8.- Set
9.- Vuit
10.- Nou o més

X_{10} = Propietat de la casa amb categories:

1.- Propietat
2.- LLoger
3.- Amb els pares o familiars

X_{11} = Tipus de casa amb categories:

1.- Casa
2.- Condominio
3.- Apartament
4.- Casa Mòbil
5.- Altres

X_{12} = Clasificació étnica amb categories:

1.- Indi americà
2.- Asiàtic
3.- Negre
4.- Indi de l'Est
5.- Hispànic
6.- Illes del Pacífic
7.- Blanc
8.- Altres

X_{13} = Llengua emprada mès freqüentment a l'hogar amb categories:

1.- Anglès
2.- Espanyol
3.- Altres

B Descripció de les categories de les dades d'AFIPE

Y = Patrò d'evolució de *ERA* ("Enfemedad Respiratoria aguda") amb categories:

1.- Sans
2.- Episodis aïllats lleus
3.- Episodis repetitius i crònics lleus
4.- Episodis aïllats greus
5.- Episodis repetitius de gravetat decreixent
6.- Episodis repetitius de gravetat creixent
7.- Episodis repetitius greus
8.- Crisis llargues i crònics greus

X_1 = Freqüència de l'atenció rebuda amb les categories:

1.- Mai
2.- Aïlladament (1 o 2 vegades no consecutives en els sis períodes setmanals de l'estudi)
3.- Repetitivament

X_2 = Tipus de tractament aportat pel sistema sanitari:

1.- Recepta
2.- Tractament
3.- Referència

X_3 = Professional que atén:

1.- Només l'auxiliar d'infermeria
2.- Metge/Metgessa

X_4 = Perfil d'increment o decrement de l'atenció rebuda:

1.- Creixent (d'auxiliar a metge o de recepta a tractament)
2.- Igual (s'inclouen els casos de menys de 2 atencions)
3.- Decreixent

X_5 = Municipi amb les categories:

1.- León: Capçalera regional

2.- El Sauce: Municipi molt extens i dispers

3.- El Jicaral: Municipi perifèric

X_6 = Tipus de comunitat amb les categories:

1.- Urbana populosa

2.- Urbana perifèrica

3.- Rural concentrada

4.- Rural dispersa

Bibliografia

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. John Wiley and Sons.

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

[3] Aitchison, J., Habbema, J. D. F., and Kay, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Applied Statistics*, **26**:15–25.

[4] Aluja-Banet, T. and Nonell-Torrent, R. (1993). Multiple correspondence analysis on panel data. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[5] Anderson, J. A. and Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, **21**:71–78.

[6] Baccini, A., Caussinus, H., and Falguerolles, A. (1993). Analysing dependence in large contingency tables: Dimensionality and patterns in scatter-plots. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[7] Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, **12**:2385–2404.

[8] Benzècri, J. P. (1982). *L'Analyse des Données*. Dunod.

[9] Benzècri, J. P. (1992). *Correspondence Analysis Handbook*. Marcel Dekker.

[10] Bermúdez, J. (1984). *Modelos de Clasificación Regulares*. Ph.D. thesis, Universitat de València.

[11] Bhattacharya, P. K. and Mack, P. (1990). Multivariate data-driven k-nn function estimation. *Journal of Multivariate Analysis*, **35**:1–11.

[12] Boos, D. D. (1985). A converse to Scheffe's theorem. *Annals of Statistics*, **13**:423–427.

[13] Bowman, A. W. (1984). An althernative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**:353–360.

[14] Bowman, A. W., Hall, P., and Titterington, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, **71**:341–351.

[15] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of Royal Statistical Society B*, **26**:211–253.

[16] Boyd, D. and Steele, J. M. (1978). Lower bounds for non-parametric density estimation. *Annals of Statistics*, **6**:932–934.

[17] Boys, R. J. (1992). On a kernel approach to a screening problem. *Journal of The Royal Statistical Society B*, **54**:157–169.

[18] Breiman, L. (1991). The p method for estimating multivariate functions from noisy data. *Technometrics*, **33**:125–143.

[19] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of The American Statistical Association*, **80**:580–619.

[20] Breiman, L. and Meisel, W. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, **19**:135–144.

[21] Broniatowsky, M., Deheuvels, P., and Devroye, L. (1989). On the relationship between stability of extreme order statistics and convergence of the maximum likekihood kernel density estimate. *Annals of Statistics*, **17**:1070–1086.

[22] Buja, A. (1990). Remarks on functional canonical variates alternating least squares methods and ace. *Annals of Statistics*, **18**:1032–1069.

[23] Buja, A., Hastie, T., and Tibshirani, R. (1989). Lineal smoothers and additive models. *Annals of Statistics*, **17**:453–555.

[24] Bull, S. B. and Donner, A. (1987). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of The American Statistical Association*, **82**:1118–1122.

[25] Burman, P. and Nolan, D. (1992). Location-adaptive density estimation and nearest-neighbor distance. *Journal of Multivariate Analysis*, **40**:132–157.

[26] Carbonell, E., Denis, J. B., Calvo, R., González, F., and Pruñonosa, J. V. (1983). *Análisis de Regresión: Un Enfoque Conceptual y Práctico para Investigadores en Ciencias de la Vida*. Instituto Nacional de Investigaciones Agrarias/Institut National de Recherches Agraires, Madrid/Paris.

[27] Chanda, K. C. and Ruymgaart, F. H. (1988). Asymptotic estimate of probability of misclassification for discriminant rules based on density estimates. *Statistics and Probability Letters*, **8**:81–88.

[28] Chessel, D. and Thioulouse, J. (1997). Fiches thematiques ade-4.analyse discriminante des correspondances. Technical report, Université de Lyon.

[29] Chiu, S. T. (1990). On the asymptotics distributions of bandwith estimates. *Annals of Statistics*, **18**:1696–1711.

[30] Chiu, S. T. (1990). Why bandwith selectors tend to choose smaller bandwiths and a remedy. *Biometrika*, **77**:222–226.

[31] Chiu, S. T. (1991). The effect of discretization error on bandwith selection for kernel density estimation. *Biometrika*, **78**:436–441.

[32] Chow, Y. S., Geman, S., and Wu, L. D. (1983). Consistent cross-validated density estimation. *Annals of Statistics*, **11**:25–38.

[33] Cleveland, W. S. and Loader, C. R. (1996). Smoothing by local regression:principles and methods. Technical report, AT&T Bell Laboratories.

[34] Cline, D. B. H. (1988). Admissible kernel estimators of a multivariate density. *Annals of Statistics*, **16**:1421–1427.

[35] Cohen, A. and Sacrowitz, H. B. (1991). Test for independence in contingency tables with ordered categories. *Journal of Multivariate Analysis*, **36**:56–67.

[36] Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis. *Australian and New Zealand Journal of Statistics*, **43**:147–199.

[37] Coomans, D., Broeckaert, I., Jonckheer, M., and Massart, D. L. (1983). Comparison of multivariate discrimination techniques for clinical data-application to the thyroid functional state. *Meth.Inform.Med.*, **22**:93–101.

[38] Cox, T. F. and Ferry, G. (1991). Robust logistic discrimination. *Biometrika*, **78**:841–849.

[39] Cuadras, C. M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In *Recent Developments in Statistical Data Analysis and Inference* (Dodge, Y., editor), pages 459–474. Elsevier Science.

[40] Cuadras, C. M. (1992). Probability distributions with given multivariate marginals and given dependence estructure. *Journal of Multivariate Analysis*, **42**:51–66.

[41] Cuadras, C. M. and Arenas, C. (1990). A distance based regression model for prediction with mixed data. *Communications in Statistics A. Theory and Methods*, **19**:2261–2279.

[42] Cuevas, A. (1981). *Robustez en inferencia Bayesiana: Un Estudio Cualitativo*. Ph.D. thesis, Universidad Complutense de Madrid.

[43] Cuevas, A. (1989). Una revisión de resultados recientes en estimación de densidades. *Estadística Española*, **31**:'7–62.

[44] Cwik, J. and Mielniczuk, J. (1989). Estimating density ratio with application to discriminant analysis. *Communications in Statistics A. Theory and Methods*, **18**:3057–3069.

[45] Davison, A. C. and Hall, P. (1992). On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika*, **79**:279–284.

[46] De Leeuw, J. (1984). Statistical properties of multiple correspondence analysis. In *Joint Summer Research Conference Series in the Mathematical Sciences*. Bowdoin College, Brunswick, Maine.

[47] De Leeuw, J. (1993). Some generalizations of correspondence analysis. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[48] De Leeuw, J. and Groenen, P. J. F. (1995). Inverse multidimensional scaling. URL = http://gifi.stat.ucla.edu.

[49] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of The Royal Statistical Society B*, **39**:1–38.

[50] Devroye, L. (1983). The equivalence of weak , strong and complete convergence in l1 for kernel density estimates. *Annals of Statistics*, **11**:896–904.

[51] Devroye, L. (1985). A note on the l consistency of variable kernel estimates. *Annals of Statistics*, **13**:1041–1049.

[52] Devroye, L. (1988). Asymptotic perfomance bounds for the kernel estimate. *Annals of Statistics*, **16**:1162–1179.

[53] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L1 View*. John Wiley and Sons.

[54] Doksum, K. A. and Lo, A. Y. (1990). Consistent and robust bayes procedures for location based on partial information. *Annals of Statistics*, **18**:443–453.

[55] Donoho, D. L. and Johnstone, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Annals of Statistics*, **17**:58–106.

[56] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage. *Journal of The Royal Statistical Society B*, **57**:301–337.

[57] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of The American Statistical Association*, **70**:892–898.

[58] Efron, B. (1982). *The Jacknife, the bootstrap and other resamling plans*. Society for Industrial and Applied Mathematics.

[59] Efron, B. (1992). Jackniffe-after-bootstrap standard errors and influence functions. *Journal of The Royal Statistical Society B*, **54**:83–127.

[60] Fisher, R. A. (1971). *Collected Papers of R.A. Fisher*. University of Adelaide, South Australia.

[61] Fitzmaurice, G. M. and Hand, D. J. (1987). A comparison of two average conditional error rate estimators. *Pattern Recognition Letters*, **6**:221–224.

[62] Fitzmaurice, G. M., Krzanowsky, W. J., and Hand, D. J. (1991). A Monte Carlo study of the 632 bootstrap estimator of error rate. *Journal of Classification*, **8**:239–250.

[63] Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceddings of the Thirteen International Conference*. Morgan, Kaufman,S.Francisco.

[64] Friedman, J. (1987). Exploratory projection pursuit. *Journal of The American Statistical Association*, **82**:249–266.

[65] Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of The American Statistical Association*, **84**:165–175.

[66] Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, **19**:1–141.

[67] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, **31**:3–21.

[68] Frühwirth-Schnatter, S. (1995). Bayesian model discrimination and bayes factors for linear gaussian state space models. *Journal of The Royal Statistical Society B*, **57**:237–246.

[69] Gabriel, K. R. (1971). The biplot graphic display of matrices with applications to principal component analysis. *Biometrika*, **58**:453–466.

[70] Gautier, J. M. and Saporta, G. (1983). Methodes non parametriques en analyse discriminante: Quelques propositions nouvelles. In *Troisièmes Journées d'Analyse Des Données et Informatique*. INRIA, Versailles.

[71] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of The American Statistical Association*, **85**:398–409.

[72] Gifi, A. (1990). *Nonlinear Multivariate Analysis*. John Wiley and Sons.

[73] Gilula, Z. (1984). On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika*, **71**:523–529.

[74] Gilula, Z. and Haberman, S. J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of The American Statistical Association*, **81**:780–788.

[75] Gilula, Z. and Haberman, S. J. (1995). Dispersion of categorical variables and penalty functions: Derivation, estimation and comparability. *Journal of The American Statistical Association*, **90**:1447–1452.

[76] Gilula, Z. and Krieger, A. M. (1989). Collapsed two-way contingency tables and the chi-square reduction principle. *Journal of The Royal Statistical Society B*, **51**:425–433.

[77] Gilula, Z. and Ritov, Y. (1990). Inferential ordinal correspondence analysis: Motivation,derivation and limitations. *International Statistical Review*, **58**:99–108.

[78] Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural network architectures. *Neural Computation*, **7**:219–269.

[79] Gittins, R. (1980). *Canonical Analysis*. Springer Verlag.

[80] Glick, N. (1972). Sample-based classification procedures derived from density estimators. *Journal of The American Statistical Association*, **67**:116–122.

[81] Glick, N. (1973). Sample-based multinomial classification. *Biometrics*, **29**:241–256.

[82] Glonek, G. F. V. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, **83**:15–28.

[83] Gnadesikan, R. (1989). Discriminant analysis and clustering. Panel on discriminant analysis, classification and clustering. *Statistical Science*, **4**:34–69.

[84] Goldstein, M. and Dillon, W. R. (1978). *Discrete Discriminant Analysis*. John Wiley and Sons.

[85] Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of The American Statistical Association*, **75**:42–73.

[86] Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, **54**:243–309.

[87] Goodman, L. A. (1991). Measures models and graphical displays in the analysis of cross-classified data. *Journal of The American Statistical Association*, **86**:1085–1138.

[88] Goodman, L. A. (1993). Correspondence analysis, association analysis and generalized nonindependence analysis of contingency tables: Saturated and unsaturated models, and appropiate graphical displays. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[89] Gordon, A. D. (1990). Constructing dissimilarity measures. *Journal of Classification*, **90**:257–269.

[90] Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.

[91] Greenacre, M. J. (1987). The geometric interpretation of correspondence analysis. *Journal of The American Statistical Association*, **82**:437–447.

[92] Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, **75**:457–467.

[93] Greenacre, M. J. (1993). Multivariate generalisations of correspondence analysis. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[94] Habbema, J. D. F., Hermans, J., and Remme, J. (1978). Variable kernel density estimation in discriminant analysis. In *COMPSTAT*. Physica-Verlag.

[95] Habbema, J. D. F., Hermans, J., and Van der Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In *COMPSTAT*. Physica-Verlag.

[96] Hall, P. (1987). On the use of compactly supported density estimates in problems of discrimination. *Journal of Multivariate Analysis*, **23**:131–158.

[97] Hall, P. (1989). On convergence rates in nonparametric problems. *International Statistical Review*, **57**:45–58.

[98] Hall, P. (1990). On the bias of variable bandwith curve estimators. *Biometrika*, **77**:529–535.

[99] Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smooting parameter in nonparametrics problems. *Journal of Multivariate Analysis*, **32**:177–203.

[100] Hall, P., DiCiccio, T., and Romano, J. (1989). On smoothing and the bootstrap. *Annals of Statistics*, **17**:692–704.

[101] Hall, P. and Marron, J. S. (1987). Choice of kernel order in density estimation. *Annals of Statistics*, **15**:161–173.

[102] Hall, P. and Marron, J. S. (1987). On the amount of noise inherent in bandwith selection for a kernel density estimator. *Annals of Statistics*, **15**:163–181.

[103] Hall, P. and Marron, J. S. (1988). Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields*, **75**:37–49.

[104] Hall, P., Sheather, S. J., and Jones, M. C. (1991). On optimal data-based bandwith selection in kernel density estimation. *Biometrika*, **78**:263–269.

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

[106] Halperin, M., Blackwelder, C., and Verter, J. I. (1971). Estimation of the multivariate logistic risk function: A comparison of the discriminant function and maximum likelihood approaches. *Journal of Chronical Diseases*, **24**:125–158.

[107] Hand, D. J. (1981). *Discrimination and Classification*. John Wiley and Sons.

[108] Hand, D. J. (1982). *Kernel Discriminant Analysis*. John Wiley and Sons.

[109] Hand, D. J. (1986). An optimal error rate estimator based on average conditional error rate: Asymptotic results. *Pattern Recognition Letters*, **4**:347–350.

[110] Hand, D. J. (1987). Screening vs prevalence estimation. *Journal of The Royal Statistical Society C*, **36**:1–7.

[111] Hand, D. J. (1987). A shrunken leaving-one-out of error rate. *Computational Mathematics Applications*, **14**:161–167.

[112] Hand, D. J. (1992). Statistical methods in diagnosis. *Statistical Methods in Medical Research*, **1**:49–67.

[113] Hand, D. J. (1994). Assessing classification rules. *Journal of Applied Statistics*, **21**:3–16.

[114] Härdle, W., J., H., Marron, J. S., and Tsybakov, A. B. (1992). Bandwith choice for average derivative estimation. *Journal of The American Statistical Association*, **87**:218–226.

[115] Hart, J. and Vieu, P. (1990). Data-driven bandwith choice for density estimation based on dependent data. *Annals of Statistics*, **18**:873–890.

[116] Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, **23**:73–102.

[117] Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of The Royal Statistical Society B*, **58**:155–176.

[118] Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of The American Statistical Association*, **89**:1255–1270.

[119] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Verlag.

[120] Hermans, J. and Habbema, J. D. F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Medizin and Biologie*, **1/**:14–19.

[121] Herrmann, E. and Wand, M. P. (1995). A bandwith selector for bivariate kernel regression. *Journal of The Royal Statistical Society B*, **57**:171–180.

[122] Hills, M. (1966). Allocation rules and their error rates. *Journal of The Royal Statistical Society B*, **28**:1–21.

[123] Hirst, D. J., Ford, I., and Crichtley, F. (1990). An empirical investigation of methods for interval estimation of the log odds ratio in discriminant analysis. *Biometrika*, **73**:609–615.

[124] Hoel, P. G. and Peterson, R. P. (1949). A solution to the problem of optimum classification. *Annals of Mathemathical Statistics*, **20**:433–438.

[125] Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. John Wiley and Sons.

[126] Impact Resources, I. (1987). Income data (San Francisco Bay Area). Technical report, Columbus, Ohio. URL = http://www.stat-stanford.edu/~jhf/ftp/trebst.ps.

[127] Isogai, E. (1987). On the asymptotic normality for nonparametric sequential density estimation. *Biometrical Journal*, **87**:215–224.

[128] Israëls, A. (1987). *Eigenvalue Techniques for Qualitative data*. DSWO Press, Leiden.

[129] John, S. (1961). Errors in discrimination. *Annals of Mathemathical Statistics*, **32**:1125–1144.

[130] Jones, M. C. (1989). Discretized and interpolated kernel density estimates. *Journal of The American Statistical Association*, **84**:733–741.

[131] Jordan, M. and R., J. (1994). Hierarchical mixtures of experts and the em alghorithm. *Neural Computation*, **6**:181–214.

[132] Kato, T. (1980). *Perturbation Theory for Linear Operators*. Springer Verlag.

[133] Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics (Three-Volume Edition)*. Charles Griffin & Company Limited.

[134] Kittler, J. (1978). Feature set search algorithms. *Pattern Recognition and Signal Processing*, pages 41–60.

[135] Koster, J. T. A. (1989). *Mathematical Aspects of Multiple Correspondence Analysis for Ordinal Variables*. DSWO Press,University of Leiden.

[136] Krusinska, E. (1989). New procedure for selection of variables in location model for mixed variable discrimination. *Biometrical Journal*, **31**:511–523.

[137] Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of Royal Statistical Society B*, **27**:251–263.

[138] Lancaster, H. O. (1957). The estructure of bivariate distribution. *Annals of Mathemathical Statistics*, **28**:719–735.

[139] Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, **44**:289–292.

[140] Lancaster, H. O. (1963). Canonical correlations and partitions of χ^2. *Quart.J.Math.Oxford*, **14**:220–224.

[141] Lancaster, H. O. (1975). Joint probabilty distributions in the meixner classes. *Journal of Royal Statistical Society B*, **37**:434–441.

[142] Lebart, L. and Mirkin, B. G. (1993). Correspondence analysis in classification. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[143] Lebart, L. and Morineau, A. i Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley and Sons.

[144] Leonard, T. (1973). A bayesian method for histograms. *Biometrika*, **60**:297–308.

[145] Lerman, I. C. (1981). *Classification et Analyse Ordinales des Données*. Dunod.

[146] Lerman, I. C. (1982). Correlation partielle dans le cas "qualitatif". In *Actes des Journées de Classification*.

[147] Loh, W. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of The American Statistical Association*, **83**:715–728.

[148] MacLahan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons.

[149] Marron, J. S. (1985). An asymptotically eficient solution to the bandwith problem of kernel density estimation. *Annals of Statistics*, **13**:1011–1023.

[150] Marron, J. S. (1986). Will the art of smoothing ever become a science? *Contemporary Mathematics*, **59**:169–177.

[151] Marron, J. S. (1987). Comparison of cros-validation techniques in density estimation. *Annals of Statistics*, **15**:152–162.

[152] McKay, R. J. and Campbell, N. A. (1982). Variable selection techniques in discriminant analysis: I.description ii.allocation. *British Journal of Mathematical and Statistical Psychology*, **35**:1–29.

[153] Meulepas, E. (1990). On a criterium for omitting variables in discriminant analysis. *Biometrics*, **46**:1181–1183.

[154] Michailidis, G. and De Leeuw, J. (1997). The Gifi system of nonlinear multivariate analysis. URL = http://gifi.stat.ucla.edu.

[155] Michalek, J. E. and Tripathi, R. C. (1980). The effect of errors in diagnosis and measurament on the estimation of the probability of an event. *Journal of The American Statistical Association*, **75**:713–721.

[156] Miller, M. E. and Landis, J. R. (1991). Generalized variance component models for clustered categorical response variables. *Biometrics*, **47**:33–44.

[157] Moore, D. H. (1973). Evaluation of five discrimination procedures for binary variables. *Journal of The American Statistical Association*, **68**:399–404.

[158] Muirhead, R. J. (1980). Asymptotic distribution in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika*, **67**:31–43.

[159] Müller, H. G. and Stadtmüller, U. (1987). Variable bandwith kernel estimators of regression curves. *Annals of Statistics*, **15**:182–201.

[160] Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**:67–79.

[161] Nadaraya, E. A. (1965). On nonparametric estimation of density functions and regression curves. *Theory of Probability and its Applications*, **10**:186–190.

[162] Neal, R. and Hinton, G. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models, M. Jordan (ed.)*. Dordrecht: Kluwer Academic Publishers, Boston.

[163] Nelder, J. A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: Some comparisons. *Journal of The Royal Statistical Society B*, **54**:273–284.

[164] Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. B. (1992). The effects of mixture distribution missespecification when fitting mixed-effects logistic models. *Biometrika*, **79**:755–762.

[165] Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Annals of Mathemathical Statistics*, **34**:1287–1301.

[166] O'Neill, M. E. (1978). Distributional expansions for canonical correlations from contingency tables. *Journal of The Royal Statistical Society B*, **40**:303–312.

[167] O'Neill, T. J. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *Journal of The American Statistical Association*, **75**:154–160.

[168] Palumbo, F. and Siciliano, R. (1998). Factorial discriminant analysis and probabilistic models. *Metron*, **56**:186–198.

[169] Pederson, S. P. and Johnson, M. E. (1990). Estimating model discrepancy. *Technometrics*, **32**:305–314.

[170] Perrière, G. and Thioulouse, J. (2003). Use of correspondence discriminant analysis to predict the subcellular location of bacterian proteins. *Computer Methods and Programs in Biomedicine*, **70**:99–105.

[171] Prakasa Rao, D. L. S. (1983). *Nonparametric Functional Estimation*. Academic Press.

[172] Press, S. J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of The American Statistical Association*, **73**:699–705.

[173] Priebe, C. E. (1994). Adaptive mixtures. *Journal of The American Statistical Association*, **89**:796–806.

[174] Pruñonosa, J. V. (1980). *Estimación Sesgada en el Modelo Lineal: Un Enfoque Bayesiano*. Master's thesis, Universidad Complutense de Madrid.

[175] Pruñonosa, J. V. (1994). Sisnica: Un proyecto de cooperación para la mejora del sistema de información en salud de Nicaragua. Technical report, Agencia Española de Cooperación Internacional, Managua.

[176] Raatgever, J. W. and Duin, R. P. W. (1978). On the variable kernel model for multivariate nonparametric density estimation. In *COMPSTAT*. Physica-Verlag.

[177] Rao, C. and Toutenburg, H. (1995). *Linear Models, Least Squares and Alternatives*. Springer Verlag.

[178] Rao, C. R. and Caligiuri, P. (1993). On scaling of ordinal categorical data. In *Seventh International Conference on Multivariate Analysis*. Elsevier Science, Barcelona.

[179] Rijckevorsel, J. L. A. and De Leeuw, J. (1988). *Component and Correspondence Analysis*. John Wiley and Sons.

[180] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

[181] Rogers, G. S. (1980). *Matrix Derivatives*. Marcel Dekker.

[182] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**:386–408.

[183] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. In *Proceedings of the National Academy of Science U.S.A.*, pages 43–47.

[184] Ruiz-Velasco, S. (1991). Asymptotic efficiency of logistic regression relative to linear discriminant analysis. *Biometrika*, **78**:235–243.

[185] Sain, R. S., Baggerly, K. A., and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of The American Statistical Association*, **89**:807–817.

[186] Samiuddin, M. and El-Sayyad, G. M. (1990). On nonparametric kernel density estimates. *Biometrika*, **77**:865–874.

[187] Schott, J. R. (1990). Canonical mean projection and confidence regions in canonical variate analysis. *Biometrika*, **76**:587–596.

[188] Schriever, B. F. (1983). Scaling of order dependent categorical variables with correspondence analysis. *International Statistical Review*, **51**:225–238.

[189] Schucany, W. R. (1989). Locally optimal window widths for kernel density estimation with large samples. *Statistics and Probability Letters*, **7**:401–405.

[190] Schucany, W. R. and Sommers, J. P. (1977). Improvement of kernel type density estimators. *Journal of American Statistical Association*, **72**:420–423.

[191] Scott, D. A. (1992). *Multivariate Density Estimation*. John Wiley and Sons.

[192] Scott, D. M. and Wand, M. P. (1991). Feasibility of multivariate density estimates. *c*, **78**:197–205.

[193] Seber, G. A. F. (1984). *Multivariate Observations*. John Wiley and Sons.

[194] Sendra, M. (2002). Análisis estadístico de datos. Technical report, Universitat de València. URL = http://www.uv.es/~msen.

[195] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density. *Journal of The Royal Statistical Society B*, **53**:683–690.

[196] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of The Royal Statistical Society B*, **43**:97–99.

[197] Silverman, B. W. (1984). Spline smoothing:the equivalent variable method. *Annals of Statistics*, **12**:898–916.

[198] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

[199] Silverman, B. W. and Young, G. A. (1987). The bootstrap: To smooth or not smooth? *Biometrika*, **74**:469–479.

[200] Silverman, S. W. and Jones, M. C. (1989). E. Fix and J.L. Hodges(1951): An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, **57**:233–247.

[201] Snapinn, S. W. and Knoke, J. D. (1989). Estimation of error rates in discriminant analysis with selection of variables. *Biometrics*, **45**:289–299.

[202] Sorum, M. J. (1971). Estimating the conditional probability of missclassification. *Technometrics*, **13**:333–342.

[203] Staniwaslis, J. G. (1989). The kernel estimate of a regression function in likekihood based models. *Journal of The American Statistical Association*, **84**:276–288.

[204] Stefanski, L. A. and Bay, J. M. (1996). Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. *Biometrika*, **83**:407–417.

[205] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**:1285–1297.

[206] Tanner, M. (1991). *Tools for Statistical inference: Observed Data and Data Augmentation Methods*. Springer.

[207] Tenenhaus, M. and Young, F. W. (1985). An analysis and sinthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**:91–119.

[208] Titterington, D. M. (1980). A comparative study of kernel based density estimates for categorical data. *Technometrics*, **22**:259–268.

[209] Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review*, **53**:141–170.

[210] Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F., and Geltke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of The Royal Statistical Society A*, **14**:145–171.

[211] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons.

[212] Tran, L. T. (1990). Kernel density estimation on random fields. *Journal of Multivariate Analysis*, **34**:37–53.

[213] Tsujitani, M. (1992). A note on the additive and multiplicative models in two-way contingency tables. *Biometrics*, **48**:267–269.

[214] Tutz, G. (1986). An alternative choice of smoothing for kernel-based density estimates in discrete discriminant analysis. *Biometrika*, **73**:405–411.

[215] Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Annals of Statistics*, **9**:725–736.

[216] Van der Burg, E. (1988). *Nonlinear Canonical Correlation and Some Related Techniques*. DSWO Press,University of Leiden.

[217] Van der Heijden, P. G. H., Falguerolles, A., and De (1989). A combined approach to contingency tables analysis using correspondence analysis and loglinear. *Journal of The Royal Statistical Society C*, **38**:249–292.

[218] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics*. Springer.

[219] Vlachonikolis, I. G. (1990). Predictive discrimination and classification with mixed binary and continuous variables. *Biometrika*, **77**:656–662.

[220] Vlachonikolis, I. G. and Marriott, F. H. C. (1982). Discrimination with mixed binary and continous data. *Applied Statistics*, **31**:23–31.

[221] Volle, M. (1981). *Analyse des Données*. Economica,Paris.

[222] Wand, M. P., Marron, J., and Ruppert, D. (1991). Transformations in density estimation. *Journal of The American Statistical Association*, **86**:343–361.

[223] Wernecke, K. D., Haerting, J., Kalb, G., and Stuerze (1989). On model-choice in discrimination with categorical variables. *Biometrical Journal*, **31**:289–296.

[224] West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika*, **78**:421–425.

[225] Wilkinson, M. A. (1992). *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

[226] Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, **46**:358–388.

[227] Young, G. A. (1990). Alternative smoothed bootstraps. *Journal of The Royal Statistical Society B*, **52**:477–484.