UNIVERSITAT DE BARCELONA

Departament de Física Aplicada i Òptica

Programa de Micro i Optoelectrònica Física

Bienni 1994-96

DISSENY D'UN PROTOCOL NUMÈRIC PER A LA CLASSIFICACIÓ INVARIANT D'IMATGES APLICANT TÈCNIQUES MULTIVARIANTS

Memòria presentada per optar al títol de doctor en Ciències Físiques

Directors:
Dr. Arturo Carnicer González
Dr. Ignacio Juvells Prades

Jordi-Roger Riba Ruíz

Barcelona, maig de 2000

7. Resultats experimentals. Signatures

En aquest capítol es fa un reconeixement invariant de les signatures de sis persones diferents. Per tal d'efectuar el reconeixement es posaran en pràctica quasi totes les tècniques explicades en els capítols anteriors. Es disposa de cent signatures diferents per persona, cinquanta de les quals formaran part del conjunt de calibratge; les cinquanta restants seran el conjunt de predicció o test. Les signatures han estat digitalitzades una a una amb el mateix digitalitzador, a una resolució de 256 x 256 píxels i amb només dos nivells de gris (blanc o negre). Les signatures són un clar exemple d'imatges binàries. Per a informació complementària sobre aquest problema, consulteu: [Pla89], [Amm90], [Qi94], [Yan95], [Lee96], [Hua97] i [Wu98].

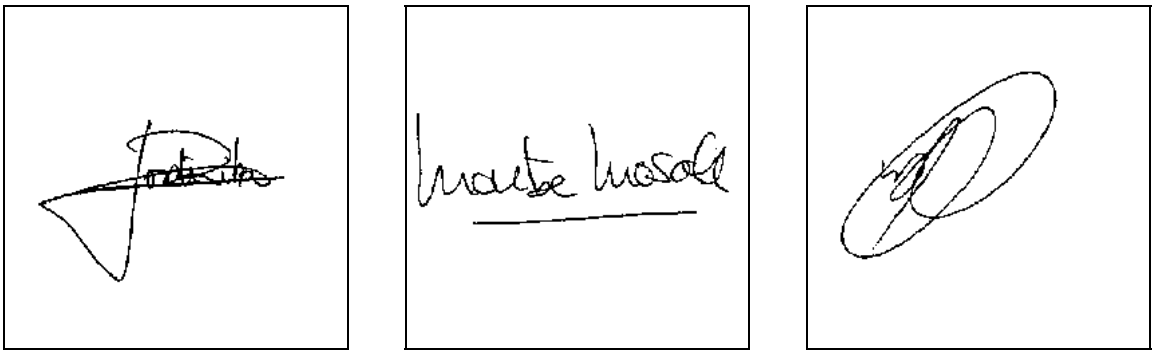
El procediment que s'ha de realitzar és el següent:

1. Càlcul de quaranta-vuit característiques discriminants per objecte. Això genera una matriu de calibratge de cinquanta files i quaranta-vuit columnes per classe. Aquest punt es desenvolupa en l'apartat 7.1.
2. Estudi i selecció de la millor tècnica de reducció de dimensions. S'apliquen la CVA, la DPCA, l'OCVA i la PCA. Aquest punt es desenvolupa en l'apartat 7.2.
3. Selecció del nombre òptim de variables que s'han de retenir. Aquest punt es desenvolupa en l'apartat 7.3.

4. Determinació del nombre mínim d'objectes de calibratge. Aquest punt es desenvolupa en l'apartat 7.4.
5. Estudi comparatiu dels diferents mètodes de predicció i de classificació. Aquest punt es desenvolupa en l'apartat 7.5.

Les figures següents mostren els sis tipus diferents de signatures tractades:

Firmes 1 Firmes 2 Firmes 3



Firmes 4 Firmes 5 Firmes 6

Figura 7.1.

Quan una persona fa més d'una signatura, és normal que la posició, la mida i l'orientació de les diferents signatures siguin lleugerament diferents. Això explica la importància d'utilitzar característiques invariants a totes les transformacions no deformatives.

La figura 7.2. mostra diferents posicions, orientacions i mides d'una mateixa signatura:

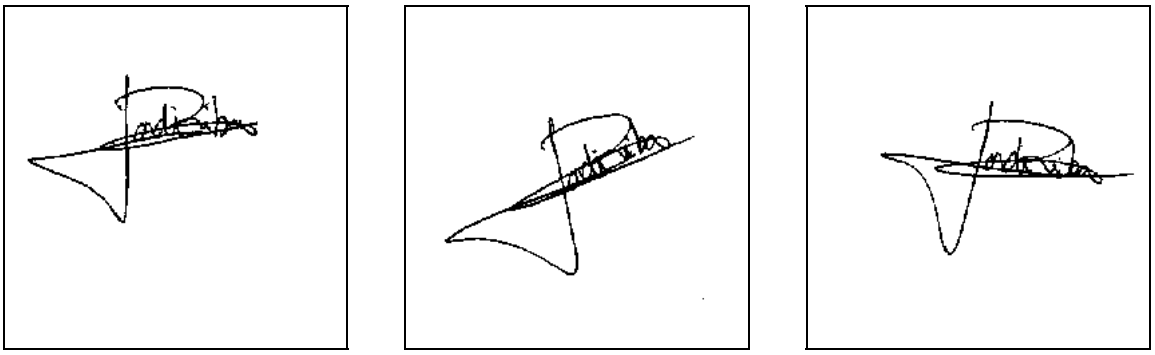


Figura 7.2.

7.1. Càlcul de les característiques

De totes les característiques invariants explicades en el capítol 2, en el cas de les signatures només es poden aplicar les que són vàlides per a imatges binàries.

Per cada signatura s'han calculat les 48 característiques següents:

- Els primers cinc moments invariants de Hu:

  $\varphi[1]$, $\varphi[2]$, $\varphi[3]$, $\varphi[4]$ i $\varphi[5]$.

- Els tres primers moments de Reiss:

  Reiss[1], Reiss[2] i Reiss[3].

- Els següents set moments calculats a partir del contorn:

  $\overline{m}_{c,1}, \overline{m}_{c,2}, \overline{m}_{c,3}, \overline{M}_{c,1}, F_{c,1}, F_{c,2}$ i $F_{c,3}$.

- Els següents set moments calculats a partir de l'esquelet:

  $\overline{m}_{e,1}, \overline{m}_{e,2}, \overline{m}_{e,3}, \overline{M}_{e,1}, F_{e,1}, F_{e,2}$ i $F_{e,3}$.

- Les cinc característiques *%AO* calculades a partir d'un cercle situat en el CDM de l'objecte, prenent com a factors: 0,7, 0,8, 0,9, 1,0 i 1,1.

- El quocient $C = P^2/(4.\pi.A)$

- Els catorze moments de les projeccions següents:

  $\eta_1^H, \eta_1^V, \eta_2^H, \eta_2^V, \eta_3^H, \eta_3^V, \eta_4^H, \eta_4^V, \eta_5^H, \eta_5^V, K^H, K^V, S^H$ i S^V.

- Les sis característiques calculades a partir de l'EPI i l'EM següents:

  D, a, b, $E = b/a$, $\varepsilon = c/a$ i A_I

No s'han calculat els descriptors de Fourier pel mateix motiu exposat en el capítol anterior: a més de requerir un càlcul molt intensiu, en molts casos no presenten un nivell de discriminació massa elevat.

Amb un ordinador *Pentium II* a 233 MHz i amb programari (*software*) compilat a 32 bits, el temps de càlcul mitjà de les quaranta-vuit característiques és inferior a tres dècimes de segon per signatura, però amb un PC d'última generació aquest temps es pot reduir substancialment. S'ha de tenir en compte que el temps de càlcul depèn molt del nombre de píxels (àrea) de cada signatura. Per tant, aquest temps és orientatiu; hi haurà signatures les característiques de les quals requeriran una mica més de temps de càlcul i altres les característiques de les quals seran més ràpides de calcular.

7.2. Selecció de la millor tècnica de reducció de dimensions

En aquest treball tractem diferents mètodes de reducció de dimensions, que són la CVA, la DPCA, l'OCVA i la PCA. S'empra l'índex I_M definit en l'apartat 6.2. per determinar quin dels mètodes de reducció de dimensions ens proporciona màxima separació entre classes. Ens quedarem amb el mètode de reducció de dimensions que, amb un nombre reduït de variables, maximitzi el valor de l'índex I_M donat per l'expressió:

$$I_M = traça(W^{-1}.B) = \sum_{i=1}^{m} \lambda_i \,. \qquad (7.1)$$

La taula 7.1. mostra els valors proporcionats per l'índex I_M per al problema de les 300 signatures del conjunt de calibratge. S'han considerat les quatre tècniques de reducció de variables i tres tipus de pretractament per a cadascuna d'aquestes: dades sense cap mena de pretractament, dades centrades i dades autoescalades. D'aquesta taula (7.1.) es dedueix que la separació entre classes augmenta en fer-ho el nombre de variables que es retenen. Aquest resultat és bastant lògic, ja que en tenir en compte més variables, aquestes aporten informació –per poca que sigui–, que ajuda sempre a la resolució del problema. Per tant, en treballar amb variables (variables canòniques o components principals) amb valors propis petits estem augmentant tant la separació entre classes com el cost computacional, fent-se necessari arribar a un compromís entre aquests dos paràmetres.

D'altra banda, en el cas de les firmes, la taula 7.1. mostra que la tècnica que amb un nombre reduït de variables ens proporciona major separació de les diferents classes és la

CVA (sense fer cap pretractament de les dades), seguida a força distància de l'OCVA (amb les dades sense pretractament o centrades).

Índex I_M	Nombre de variables retingudes	**CVA**	**DPCA**	**OCVA**	**PCA**
Dades sense pretractament	1	0,85	0,15	0,48	0,09
	2	1,20	0,21	0,65	0,11
	3	1,38	0,27	0,70	0,20
	4	1,54	0,31	0,75	0,23
	5	**1,66**	0,36	**0,82**	0,26
Dades centrades	1	0,50	0,15	0,32	0,02
	2	0,90	0,21	0,64	0,13
	3	1,09	0,27	0,74	0,19
	4	1,27	0,31	0,77	0,22
	5	1,38	0,36	**0,82**	0,25
Dades autoescalades	1	0,05	0,16	0,44	0,01
	2	0,07	0,25	0,46	0,15
	3	0,09	0,34	0,53	0,25
	4	0,12	0,45	0,57	0,30
	5	0,18	**0,52**	0,58	**0,32**

Taula 7.1.

En les figures següents (7.3. a 7.6.) es mostren els 300 objectes del conjunt de calibratge representats en l'espai de les variables. Aquestes figures permeten visualitzar el grau de separació entre les diferents classes.

La figura 7.3. mostra els resultats proporcionats per la CVA sense haver fet cap mena de pretractament previ de les dades:

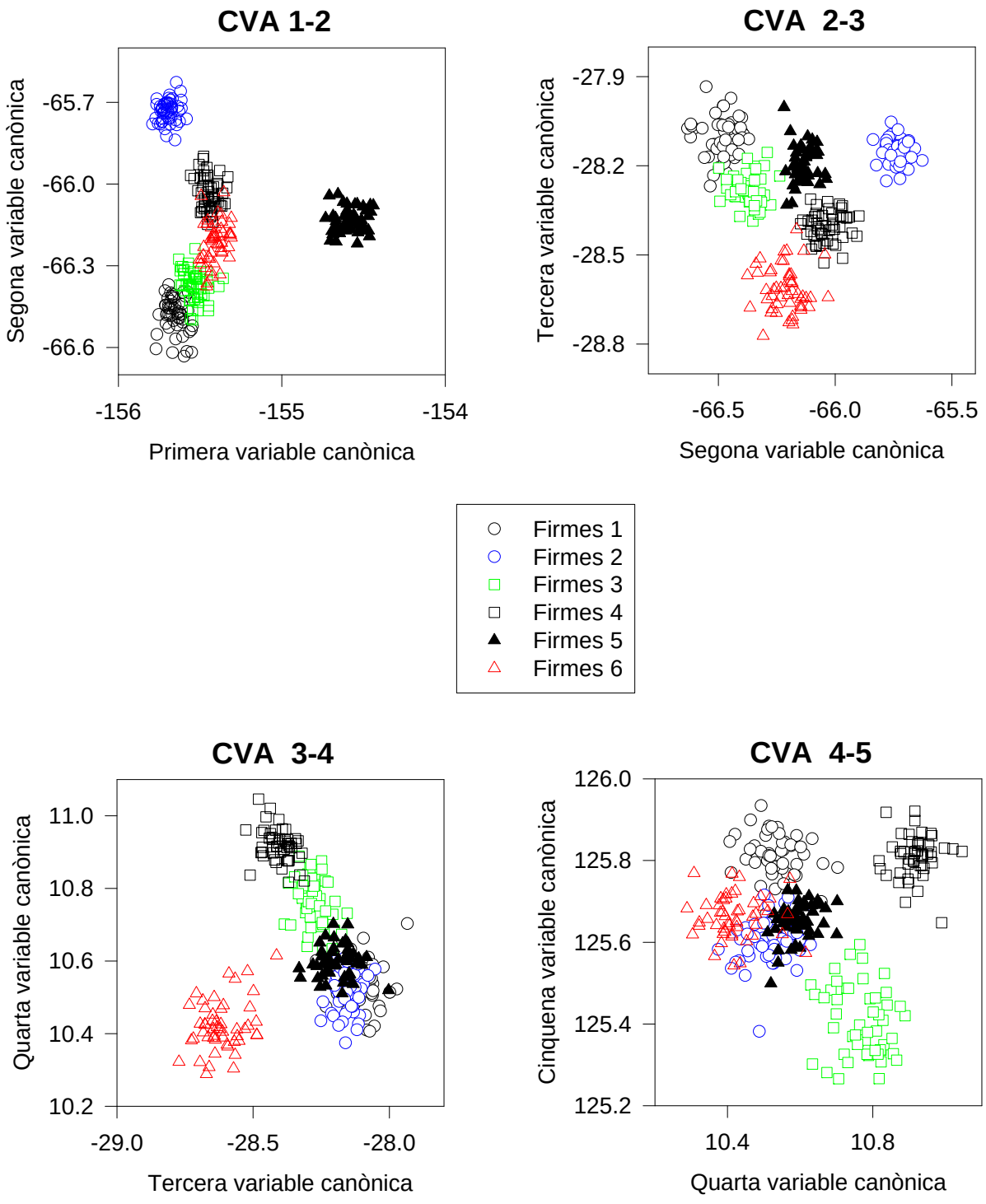


Figura 7.3.

La figura 7.4. mostra els resultats proporcionats per la DPCA havent autoescalat prèviament les dades:

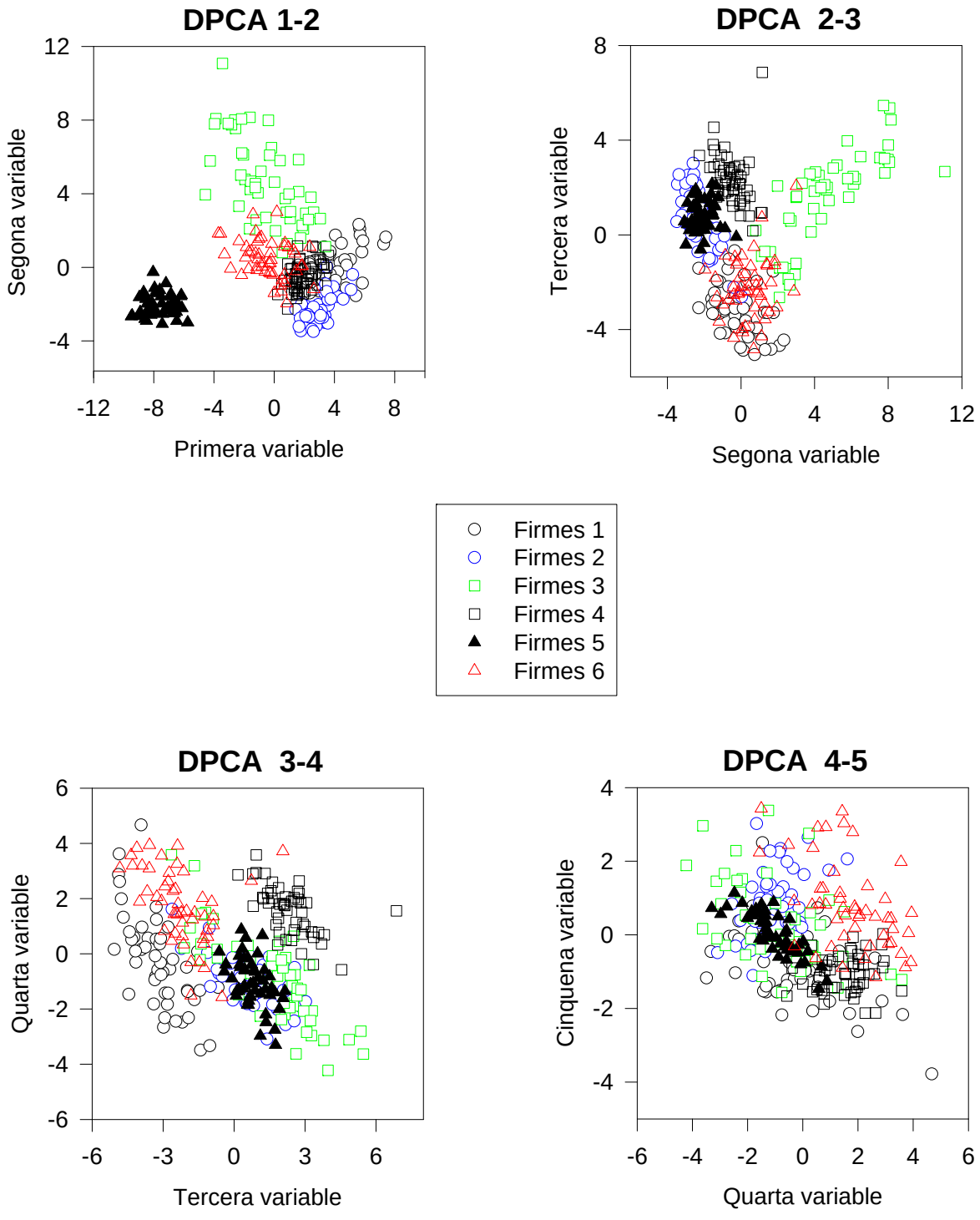


Figura 7.4.

La figura 7.5. mostra els resultats proporcionats per l'OCVA havent centrat prèviament les dades:

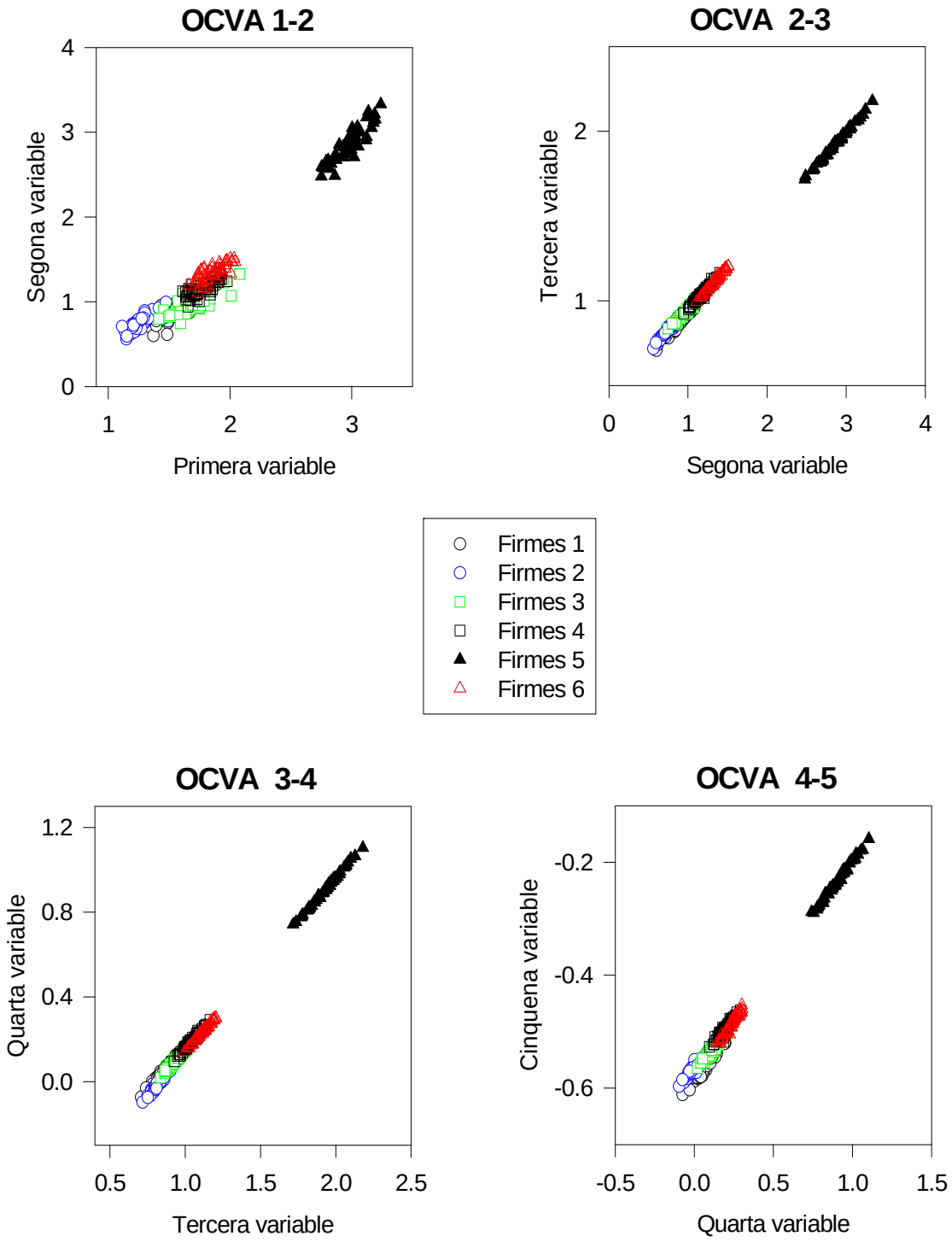


Figura 7.5.

La figura 7.6. mostra els resultats proporcionats per la PCA sense haver fet cap mena de pretractament previ de les dades:

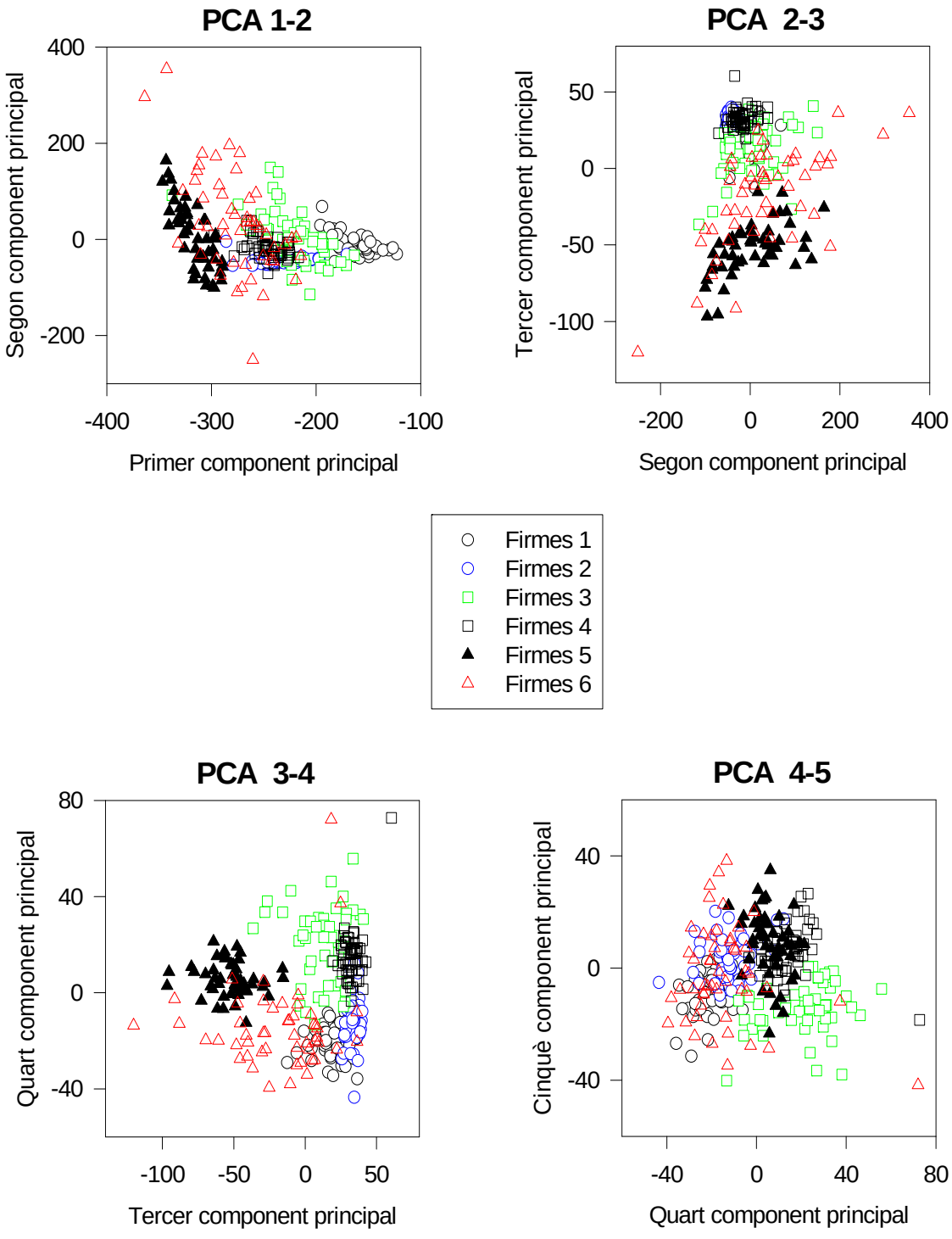


Figura 7.6.

Les figures 7.3. a 7.6. confirmen que l'algorisme que amb les primeres variables proporciona millors resultats és la CVA.

7.3. Selecció del nombre òptim de variables que s'han de retenir

En l'apartat anterior s'ha vist que, en el cas de les signatures, la tècnica de reducció de dimensions que millors resultats proporciona és la CVA. Per tant, una vegada s'ha aplicat la CVA, es fa necessari decidir quin és el nombre òptim de variables que s'han de retenir. En l'apartat 3.6. hi ha exposades diferents tècniques útils per seleccionar el nombre de variables que s'han de retenir. No totes , però, són aplicables al cas de la CVA.

Els valors propis associats a totes les cinc variables canòniques de la CVA són els següents:

Valors propis	Valor numèric
λ_1	0,8308
λ_2	0,3506
λ_3	0,1805
λ_4	0,1689
λ_5	0,1144

Taula 7.2

Ara aplicarem els diferents criteris de selecció del nombre òptim de variables:

- **Criteri del 95 % de la variància explicada**

D'acord amb aquest criteri, s'ha de calcular la variància acumulada per les primeres variables i retenir les primeres *i* variables, que acumulen el 95 % de la variància (informació) total.

Valors propis acumulats	Variància acumulada
λ_1	50,50 %
$\lambda_1+\lambda_2$	73,02 %
$\lambda_1+\lambda_2+\lambda_3$	82,78 %
$\lambda_1+\lambda_2+\lambda_3+\lambda_4$	93,05 %
$\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5$	100%

Taula 7.3.

La taula 7.3. mostra el tant per cent de la variància total acumulada per les primeres variables. D'acord amb aquest criteri, tenint en compte els resultats de la taula 7.3., cal retenir les cinc primeres variables canòniques.

- **Criteri del diagrama de caigudes**

Aquest criteri indica que el nombre de variables que cal retenir ve donat per l'últim canvi abrupte del pendent del diagrama de caigudes (*Scree Diagram*: valor numèric dels valors propis, enfront del nombre de variables retingudes). La figura 7.7. mostra el diagrama de caigudes resultant per al cas de les signatures:

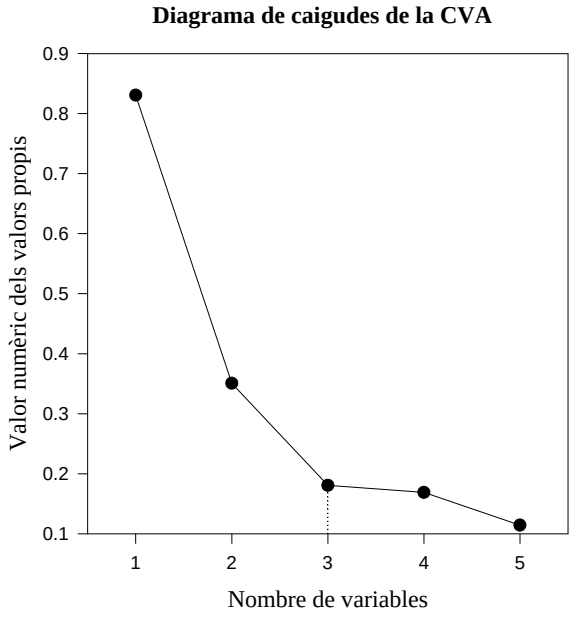


Figura 7.7.

Segons el criteri del diagrama de caigudes, com que l'últim canvi abrupte del pendent de la corba correspon a la tercera variable canònica, cal retenir les tres primeres variables canòniques.

- **Criteri dels valors propis no degenerats**

Segons aquest criteri, cal retenir només les variables corresponents a valors propis no degenerats. Recordem que dos valors propis són degenerats quan es compleix:

$$\lambda_i < \lambda_{i+1} \cdot \frac{1+\alpha}{1-\alpha}, \quad \text{on} \quad \alpha \approx \sqrt{\frac{2}{n}},$$

sent n el nombre total d'objectes de calibratge.

Aplicant aquest mètode resulten no degenerats els dos primers valors propis; per tant, cal retenir només les dues primeres variables.

- **Criteri de la distribució esfèrica**

Aquest criteri aconsella retenir els valors propis que tenen un valor superior al valor mitjà de tots els valors propis.

Com que en aquest cas el valor mitjà de tots els valors propis és 0,3290, segons aquest cal retenir només les dues primeres variables.

S'acaba de comprovar que els diferents mètodes no porten a un resultat únic, sinó que produeixen una diversificació dels resultats. En aquest cas, cal ser una mica conservador i, com que el nombre total de variables és petit –fet que indica que el cost computacional no serà massa elevat–, és millor seguir el criteri que indiqui el nombre més gran de variables que s'han de retenir. Aquest criteri és el del 95 % de la variància total explicada, que fa que s'hagin de retenir totes les variables. Per tant, d'ara en endavant sempre es treballarà amb les cinc variables canòniques resultants d'aplicar la CVA.

Quan es treballi amb tècniques de regressió multivariable (PLS, PCR, etc.), com que es basen en la PCA, no tindrà sentit aplicar la CVA. Per tant, en aquest cas s'haurà de determinar el nombre òptim de components principals que cal retenir (vegeu l'apartat 7.5).

7.4. Nombre mínim d'objectes de calibratge

En el capítol 4 es van tractar tècniques desenvolupades per calcular el nombre mínim d'objectes de calibratge necessaris per poder tractar amb garanties un problema de classificació. En l'apartat 4.2. es van desenvolupar dues tècniques *a posteriori*, amb les quals es treballarà a continuació.

La primera d'aquestes tècniques es basa en el càlcul de la distància computada. Primer s'aplica la CVA (per ser la tècnica que ha proporcionat millors resultats ens) sobre les 300 dades de calibratge. S'obtindrà una matriu de transformació de l'espai definit per les característiques discriminants a l'espai de les variables. Sempre amb aquesta matriu de transformació, es calcularan les coordenades, en l'espai de les variables, de subgrups d'objectes de calibratge, agafats aleatòriament, de 270, 240, 210, 180, 150, 120, 90 i 60 objectes, respectivament.

El dos gràfics següents (figures 7.8. i 7.9.) mostren el resultat d'aplicar el criteri de la distància computada (DC) entre el centre del grup total d'objectes de calibratge i els centres dels diferents subgrups. En el primer d'aquests, la DC es calcula respecte a les cinc primeres variables canòniques (resultants de la CVA), mentre que en el segon es calcula respecte les quaranta-vuit característiques discriminants.

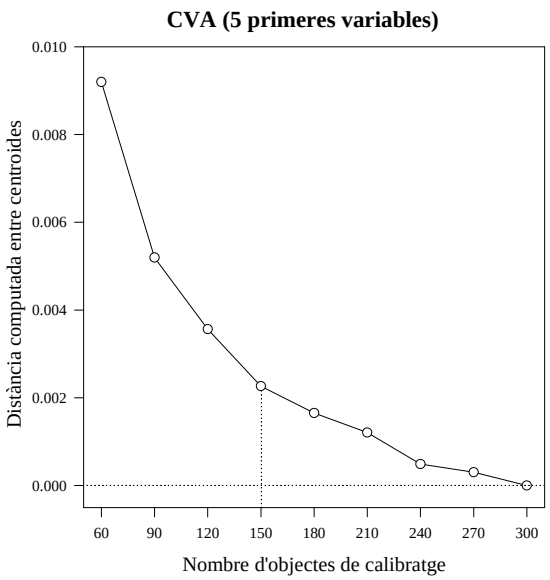


Figura 7.8.

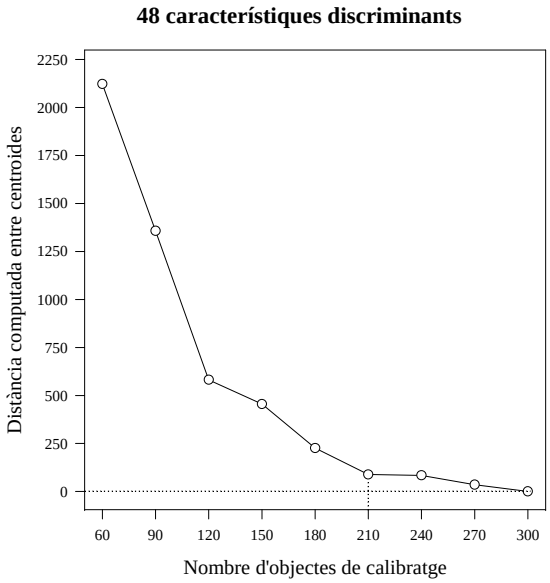


Figura 7.9.

La figura 7.8. mostra que a partir de 150 objectes de calibratge el pendent de la corba és menys pronunciat. Per tant, sembla que 150 objectes de calibratge (25 per classe) han de ser suficients. Aquest resultat és concordant amb la hipòtesi de Jain i Chandrasekaran explicada en l'apartat 4.1. ($n > 5.c.m^*$ = 5.6.5 = 150).

En canvi, quan el problema s'ataca directament amb les característiques discriminants (figura 7.9.), sense haver aplicat cap tècnica de reducció de dimensions, es requereix un nombre més elevat d'objectes de calibratge. Aquest resultat indica que les característiques discriminants són menys eficients que les variables, fins i tot pel que fa al nombre mínim d'objectes de calibratge.

La segona tècnica ha estat proposada en aquest treball i és molt similar a l'anterior, però canviant la distància computada per la taxa d'errors de classificació. Després d'aplicar la CVA sobre les 300 dades de calibratge, i sempre amb la mateixa matriu de transformació, es calculen les variables canòniques de subgrups escollits a l'atzar de 270, 240, 210, 180, 150, 120, 90 i 60 objectes, respectivament. Tot seguit s'apliquen tècniques ràpides i

eficients de classificació, com són la LDA i la QDA i s'avalua la taxa d'errors en funció del nombre d'objectes de calibratge. En aquest cas, cal fixar-se en l'últim canvi abrupte del pendent de la corba.

El dos gràfics següents (figures 7.10. i 7.11.) mostren el resultat d'aplicar el criteri del nombre d'errors de classificació en classificar diferents subgrups del grup total dels 300 objectes de calibratge. En la figura 7.10. el mètode de classificació utilitzat ha estat la LDA, mentre que en la figura 7.11. ha estat la QDA.

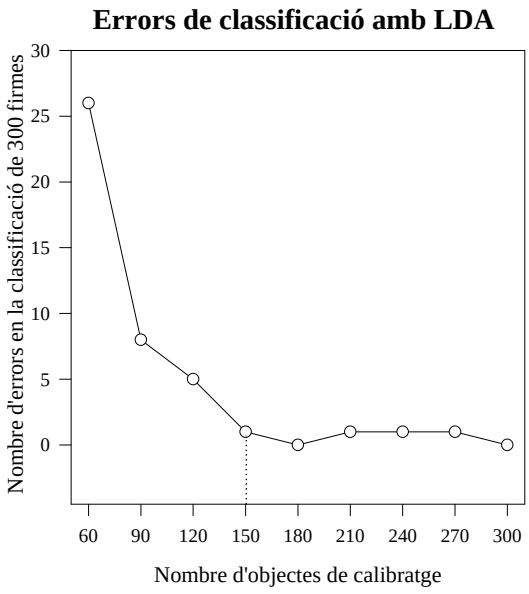


Figura 7.10.

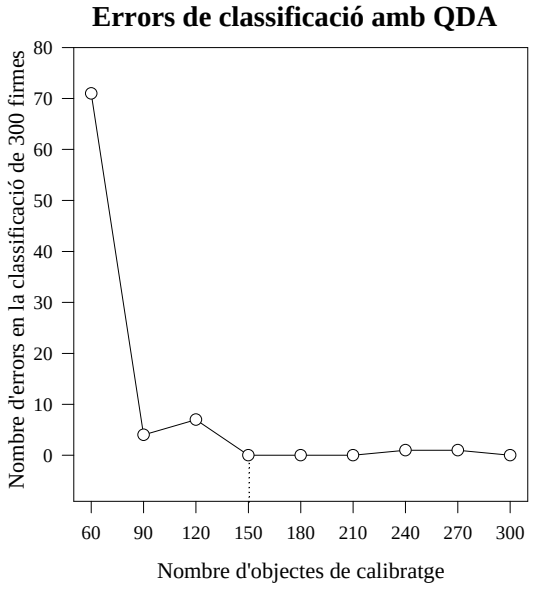


Figura 7.11.

En les figures 7.10. i 7.11. es comprova que, en el cas del problema de les firmes, amb 150 objectes de calibratge per classe n'hi ha d'haver prou. Aquests resultats tornen a confirmar la hipòtesi apuntada per Jain i Chandrasekaran ($n > 5.c.m^*$ = 5.6.5 = 150).

7.5. Classificació aplicant tècniques de regressió

En el capítol 5 es va veure que una de les possibilitats a l'hora d'afrontar un procés de classificació és aplicar tècniques multivariables de regressió. Les tècniques de regressió que es compararan són la PLS, la PCR, la LRR i la MLR.

Les figures 7.12. a 7.15. mostren, per a les diferents tècniques de regressió aplicades, el PRESS (mesura de l'error de predicció; vegeu l'apartat 5.3.) resultant de la classificació

dels 300 objectes del conjunt de calibratge enfront del nombre de variables retingudes. Totes les tècniques de regressió han estat assajades amb les dades sense cap mena de pretractament, amb les dades centrades i amb les dades autoescalades.

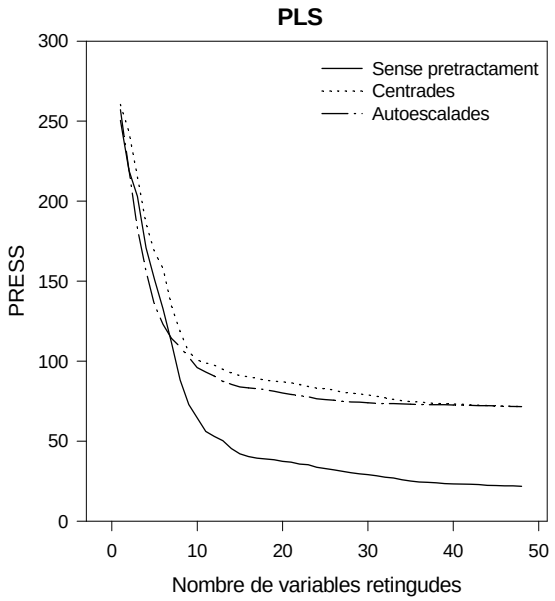


Figura 7.12.

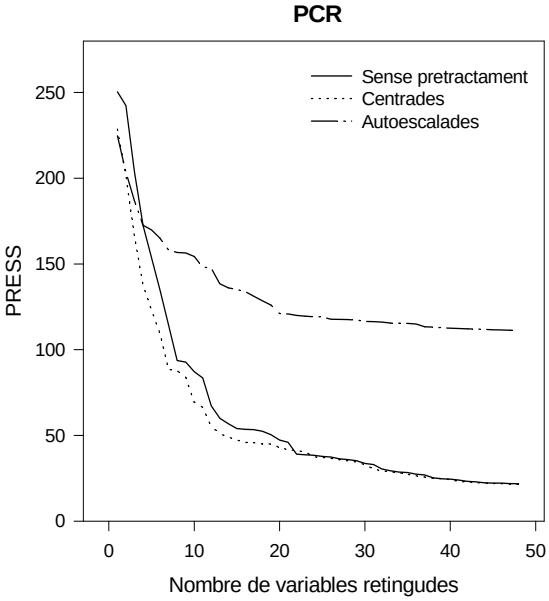


Figura 7.13.

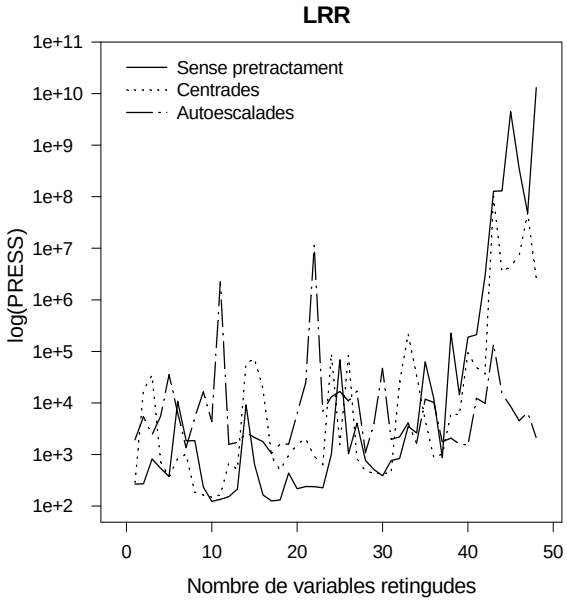


Figura 7.14.

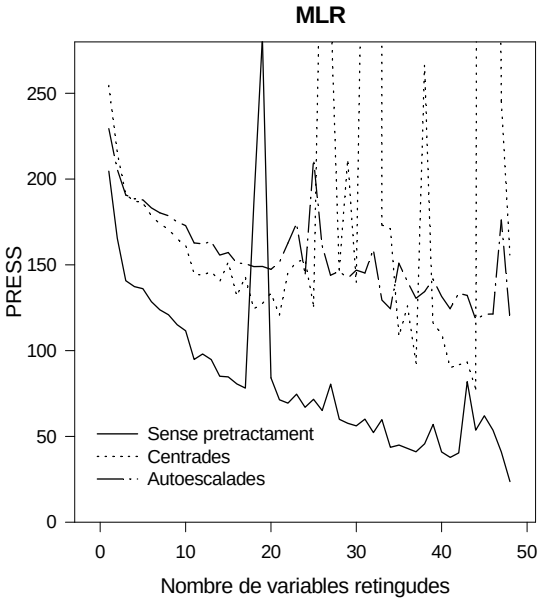


Figura 7.15.

Les figures 7.14. i 7.15. indiquen que els resultats proporcionats per la LRR i la MLR no són estables. Per tant, aquestes dues tècniques es deixaran de banda amb vista a solucionar aquest problema de classificació. Això fa que per a la classificació d'objectes de test ens quedem amb la PCR i la PLS.

7.5.1. Resultats de la classificació amb tècniques de regressió

En aquest apartat es realitza la classificació del conjunt de 300 objectes de test. Els algorismes utilitzats són la PLS i la PCR, perquè en l'apartat anterior s'ha vist que eren les tècniques de regressió que funcionen millor en aquest problema concret.

Per escollir el nombre de variables que s'han de retenir (PCR i PLS) fem servir els criteris de selecció del nombre de variables apropiades per la PCA, com són el del 95 % de la variància, el de W(r) i el de retenir les variables associades a valors propis superiors a la unitat. Com que el cost computacional d'aquests mètodes de regressió és relativament baix, ens quedarem amb el criteri que proporcioni el nombre més elevat de variables que cal retenir (recordeu que quan s'apliquen tècniques de regressió no es pot treballar amb les variables resultants d'aplicar la CVA).

PLS amb les dades sense pretractament

En el cas de la PLS apliquem les dades sense pretractament perquè en el gràfic 7.12. es veu que és el cas que proporciona resultats millors. Tots els mètodes de classificació proporcionen tantes sortides com classes tingui definides el problema. En el problema de les signatures es tenen c = 6 classes i, com que es classifiquen 300 objectes, hi haurà un total de $6 \times 300 = 1.800$ sortides.

Criteri	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
95 % variància	6	141	7,83 %	138,76
W(r)	17	15	0,83 %	47,10
$\lambda_i > 1$	28	8	0,44 %	46,21

Taula 7.4.

Els resultats d'aplicar els tres criteris de selecció del nombre de variables sobre els 300 objectes de test es reflecteixen en la taula 7.4. Aquests resultats són sobre objectes de test,

objectes no inclosos en els objectes de calibratge del model; per tant, són resultats definitius.

En la taula 7.4. també es comprova que la hipòtesi que s'havia apuntat anteriorment, de quedar-se amb el criteri que obligui a retenir un major nombre de variables, és bona. Per tant, en el cas de la PLS cal agafar les vint-i-vuit primeres variables i les dades sense pretractament, i s'obtindrà una taxa de sortides errònies de 0,44 %.

PCR amb les dades centrades

En el cas de la PCR apliquem les dades centrades perquè en el gràfic 7.13. es veu que és el cas que proporciona resultats millors. Els resultats d'aplicar els tres criteris de selecció del nombre de variables sobre els 300 objectes de test es reflecteixen en la taula següent:

Criteri	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
95 % variància	9	54	3,00 %	86,46
W(r)	10	38	2,11 %	72,49
$\lambda_i > 1$	28	11	0,61 %	44,41

Taula 7.5.

De la taula anterior també es comprova que la hipòtesi que s'havia apuntat anteriorment, de quedar-se amb el criteri que obligui a retenir un major nombre de variables, és vàlida. Per tant, en el cas de la PCR cal agafar les vint-i-vuit primeres variables i les dades centrades, i s'obtindrà una taxa de sortides errònies de 0,61 %.

Els dos gràfics següents mostren els resultats de la classificació dels 300 objectes del conjunt de test. Per poder-los interpretar cal tenir en compte que els algorismes de classificació proporcionen tantes sortides com classes d'objectes té definides el problema (en aquest cas, sis). Per tant, la resposta dels algorismes a l'entrada d'un objecte de test seran sis sortides, una per classe. Un objecte serà associat a una classe determinada si la sortida de l'algorisme de classificació corresponent a aquesta classe és superior o igual a 0,5, i no hi serà associat en cas contrari.

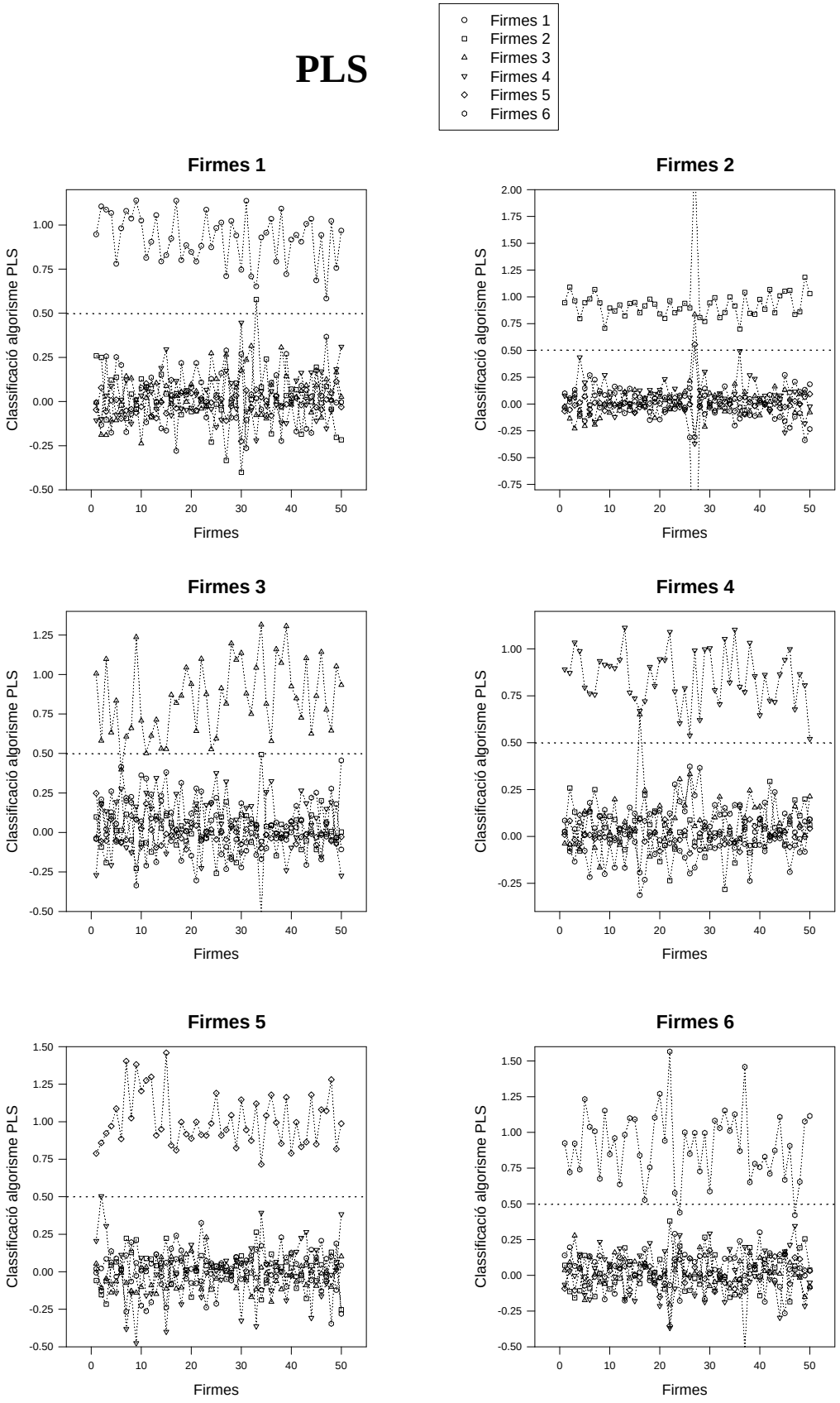

PLS
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació algorisme PLS
Firmes

Figura 7.16.

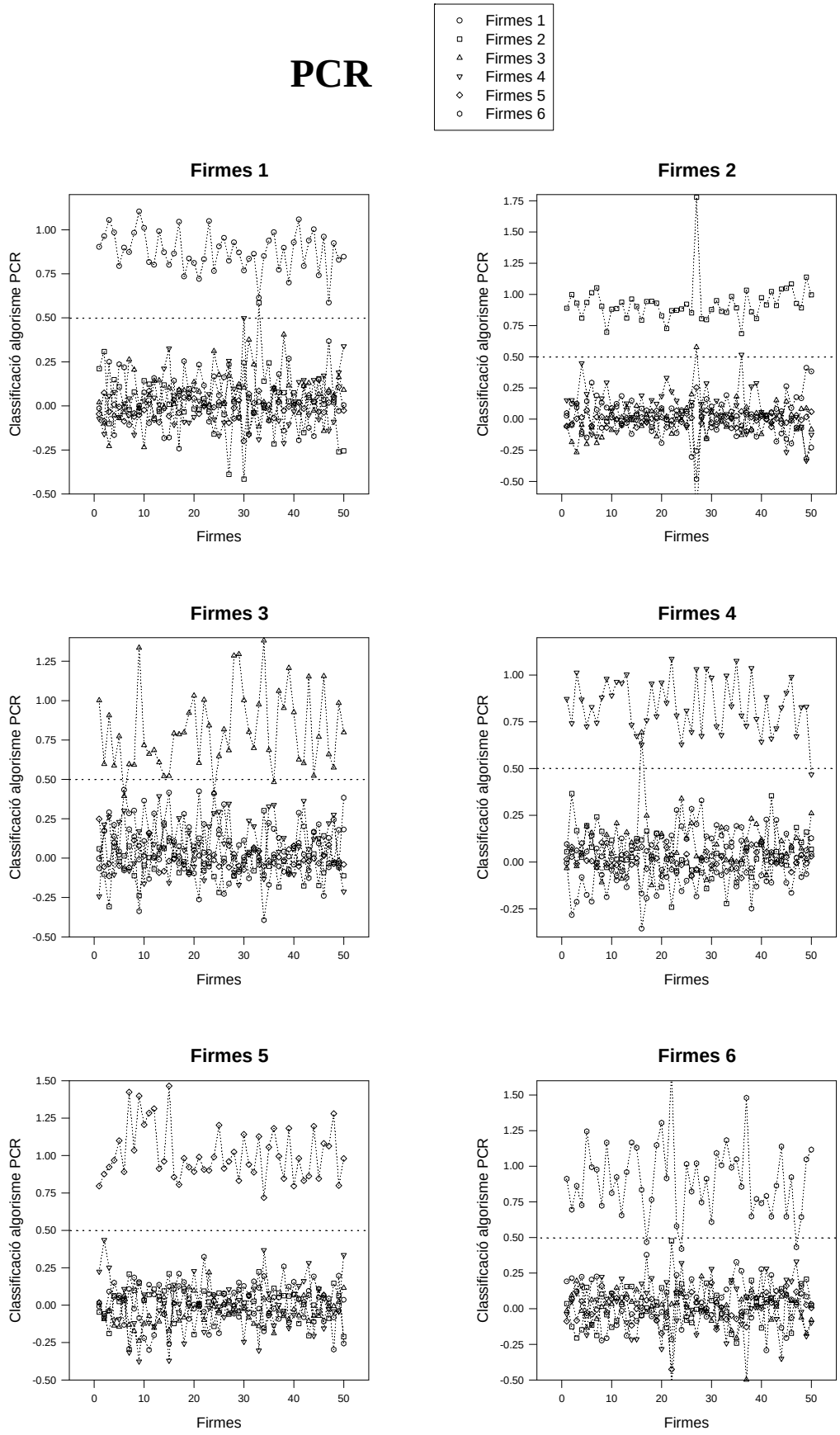

PCR
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació algorisme PCR
Firmes

Figura 7.17.

7.6. Classificació amb anàlisi discriminant

En aquest apartat s'apliquen la LDA i la QDA, que, respectivament, suposen igual i diferent matriu de variàncies-covariàncies per a cada classe. Primer s'han agafat les dades (sense haver fet cap pretractament), després s'ha realitzat la CVA i, agafant les cinc primeres variables canòniques, s'han aplicat la LDA i la QDA. La taula 7.6. mostra els resultats de classificació d'aquestes dues tècniques sobre els 300 objectes de test.

Criteri	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
LDA	5 de la CVA	3	0,17 %	3,00
QDA	5 de la CVA	5	0,28 %	5,00

Taula 7.6.

Les sortides dels algorismes LDA i QDA, per ser tècniques de classificació, només presenten dos valors enters possibles: 1 o 0. Per tant, l'algorisme presenta per a cada objecte tantes sortides com classes tingui el problema, i cada sortida tindrà el valor 1 o 0. Per a un objecte donat, per decidir si la sortida corresponent a una classe determinada ha de ser 1 o 0 cal haver definit prèviament un valor llindar o *threshold*. En aquest treball es proposa agafar els llindars següents per a cada classe:

$$T_{LDA,i} = [m_i + M_i]/2$$

$$T_{QDA,i} = \sqrt{m_i.M_i}\ , \qquad (7.2)$$

on m_i és el valor mínim que proporciona la gaussiana corresponent a la classe *i*-èssima a tots els objectes de calibratge pertanyents a aquesta classe, i M_i és el valor màxim que proporciona la gaussiana corresponent a la classe *i*-èssima a tots els objectes de calibratge que no pertanyen a aquesta classe.

Les següents figures mostren els resultats de classificació de la LDA i la QDA sobre els 300 objectes de test.

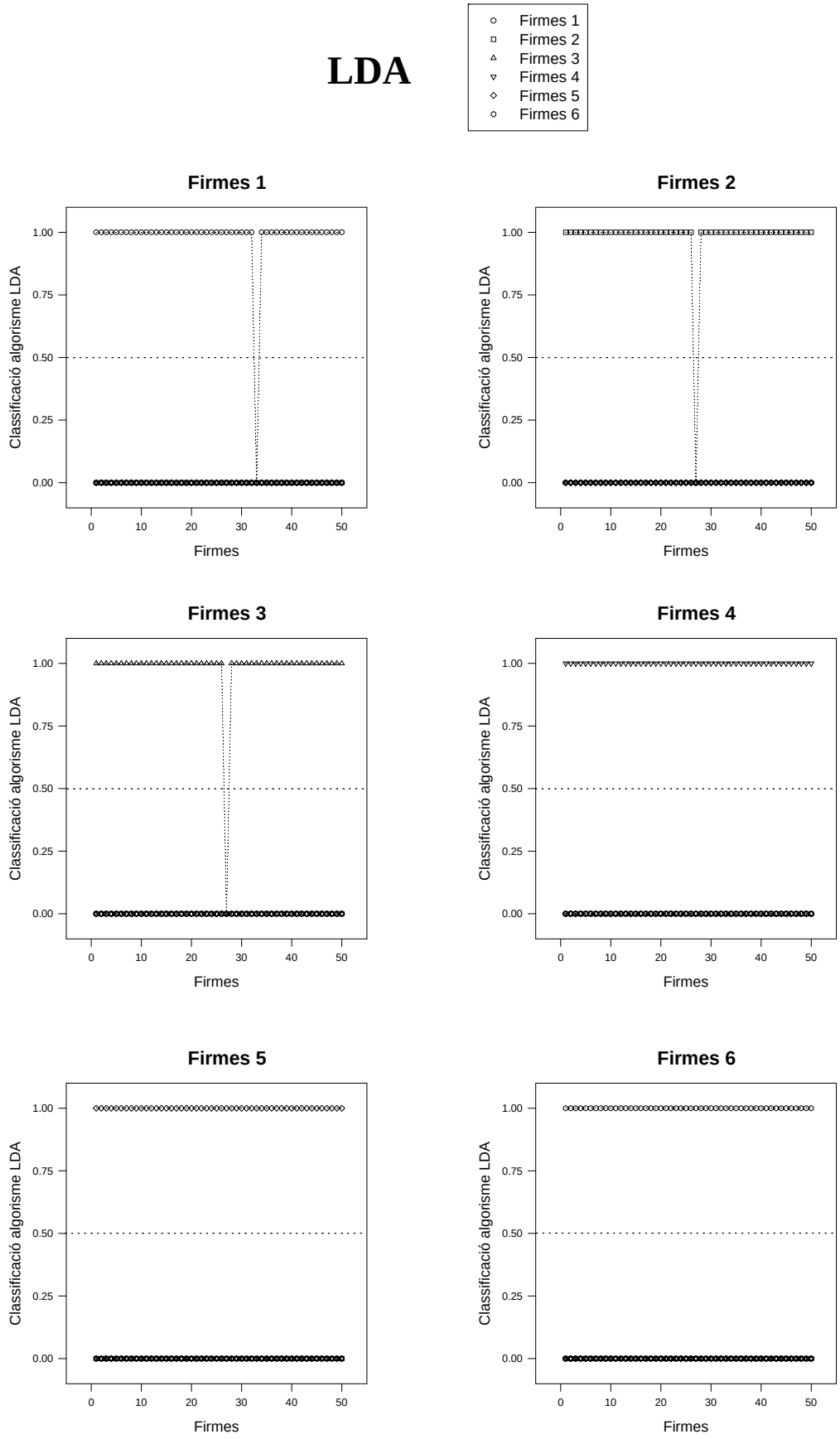

LDA
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació algorisme LDA
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.18.

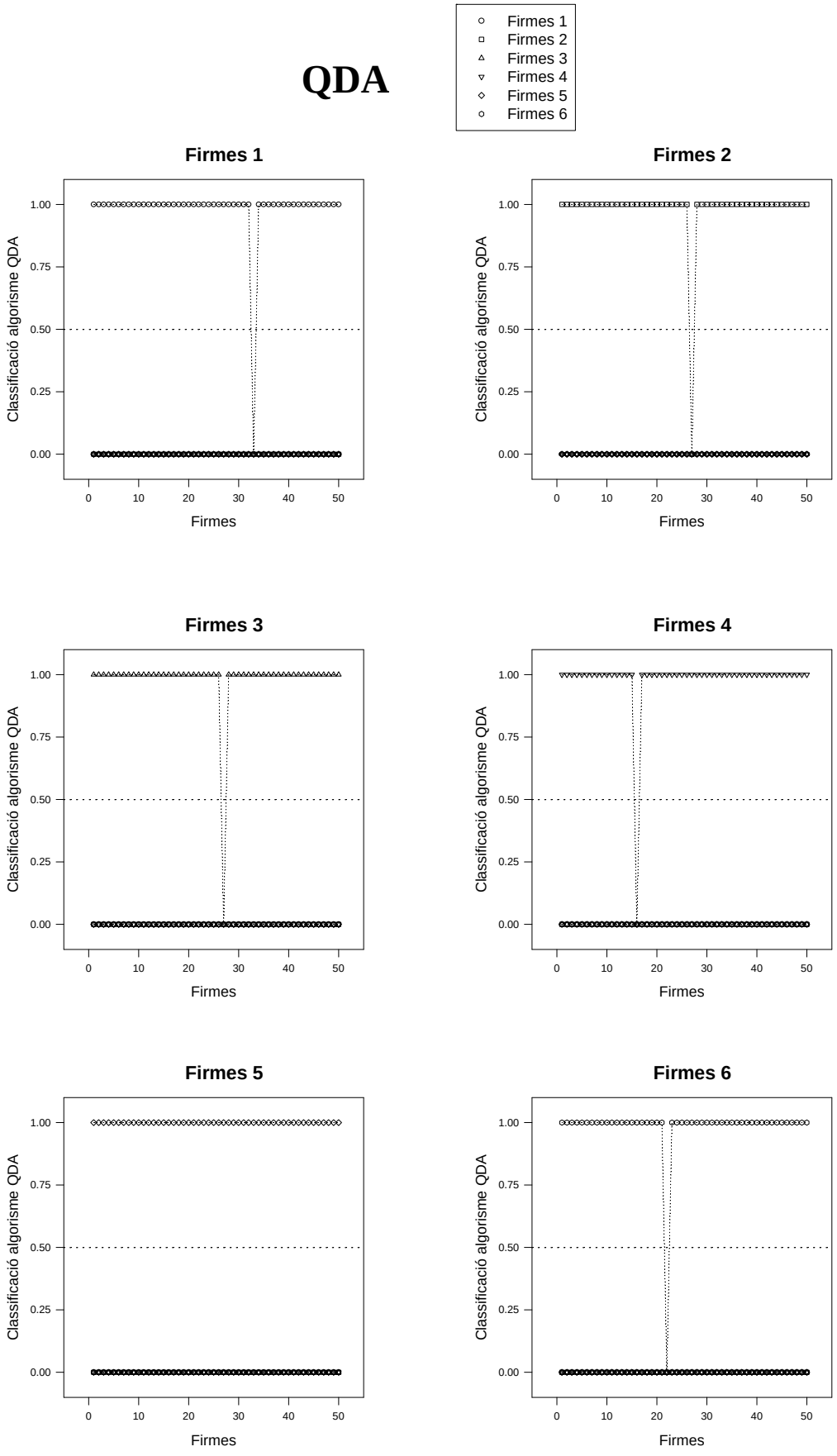

QDA
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació algorisme QDA
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.19.

7.7. Classificació amb SIMCA i DASCO

Tant SIMCA com DASCO són tècniques de classificació força utilitzades. Es treballarà amb les cinc variables canòniques resultants d'efectuar la CVA. La taula següent mostra els resultats de classificació dels 300 objectes de calibratge amb les dades resultants de la CVA sense pretractament, amb les dades centrades i amb les dades autoescalades.

Tècnica	Pretractament	Sortides errades de 1.800	Error total	PRESS
SIMCA	Cap	19	1,06 %	19,00
SIMCA	Centrat	19	1,06 %	19,00
SIMCA	Autoescalat	20	1,11 %	20,00
DASCO	Cap	5	0,28 %	5,00
DASCO	Centrat	5	0,28 %	5,00
DASCO	Autoescalat	13	0,72 %	13,00

Taula 7.7.

El nombre de variables que automàticament retenen per classe els algorismes SIMCA i DASCO, en el cas dels resultats anteriors, es reflecteix en la taula següent:

	Nombre de variables retingudes per classe					
Pretractament	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Cap	1	1	1	1	1	1
Centrat	1	1	1	1	1	1
Autoescalat	1	1	1	1	1	1

Taula 7.8.

Els resultats anteriors indiquen que cal tenir en compte tant el model SIMCA com el model DASCO amb les dades centrades o sense cap pretractament. Per comoditat s'ha escollit la segona possibilitat.

La taula següent mostra els resultats d'aquestes dues tècniques sobre els 300 objectes del conjunt de test.

Criteri	Sortides errades de 1.800	Error total	PRESS
SIMCA sense pretract.	47	2,61 %	47,00
DASCO sense pretract.	19	1,06 %	19,0

Taula 7.9.

Les sortides dels algorismes SIMCA i DASCO, per ser tècniques de classificació, només presenten dos valors enters possibles: 1 i 0. Per tant, l'algorisme presenta per a cada objecte tantes sortides com classes tingui el problema, i cada sortida tindrà el valor 1 o 0. En l'apartat 5.2.1. s'expliquen els algorismes d'aquests dos mètodes. Per a un objecte donat, per decidir si la sortida corresponent a una classe determinada ha de ser 1 o 0 cal haver definit prèviament un valor de tall o *threshold*. En aquest treball es proposa agafar els llindars següents per cada classe:

$$T_i = \sqrt{m_i . M_i} \,, \tag{7.5}$$

on m_i és el valor mínim de la distància de Mahalanobis al centre de la classe *i*-èssima de tots els objectes de calibratge que pertanyen a aquesta classe mentre que M_i és el valor màxim de la distància de Mahalanobis al centre de la classe *i*-èssima de tots els objectes de calibratge que no pertanyen a aquesta classe.

Les figures 7.20. i 7.21. mostren els resultats d'efectuar SIMCA i DASCO sobre els 300 objectes de test.

SIMCA (sense pretract.)

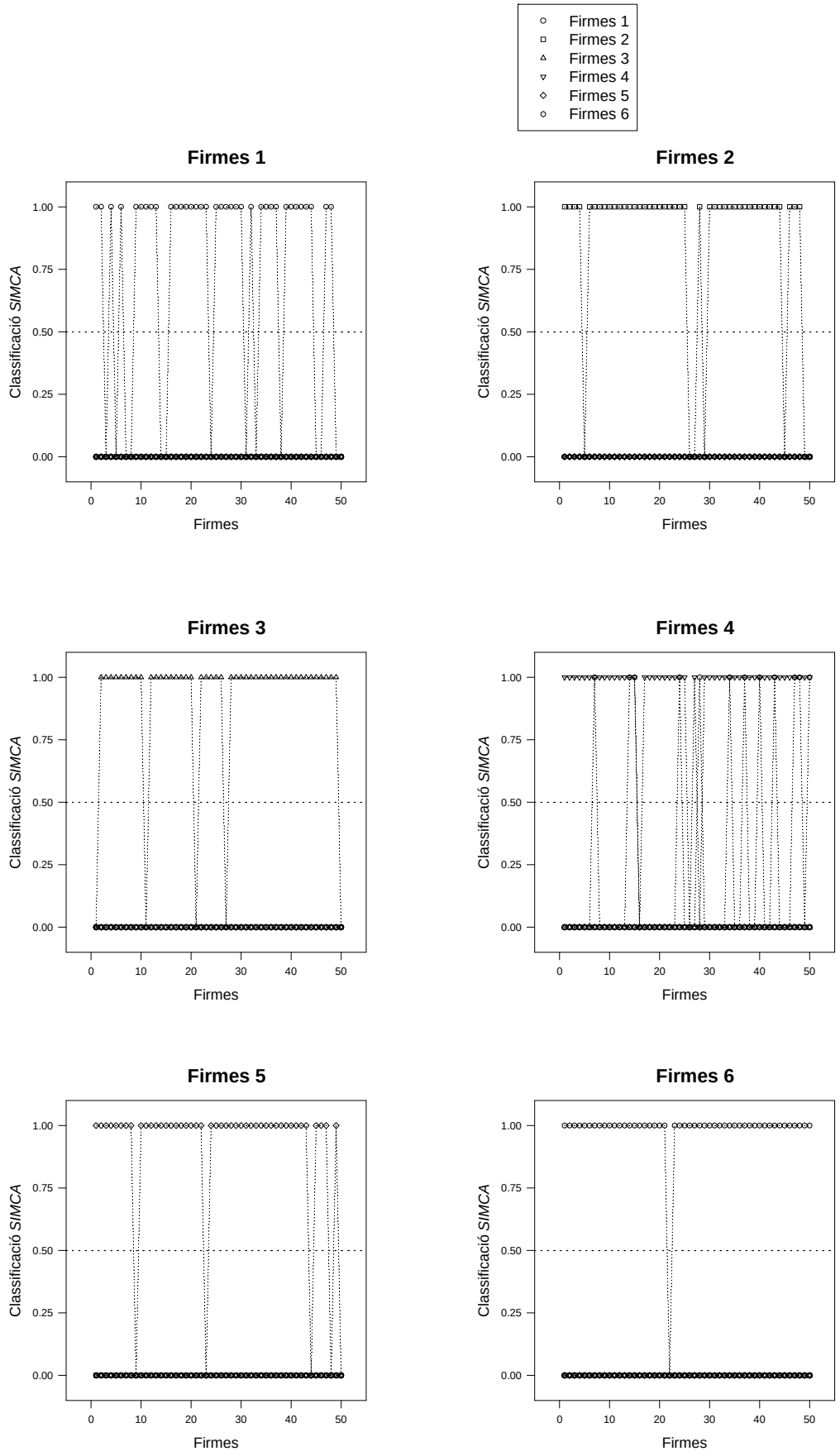

Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació SIMCA
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.20.

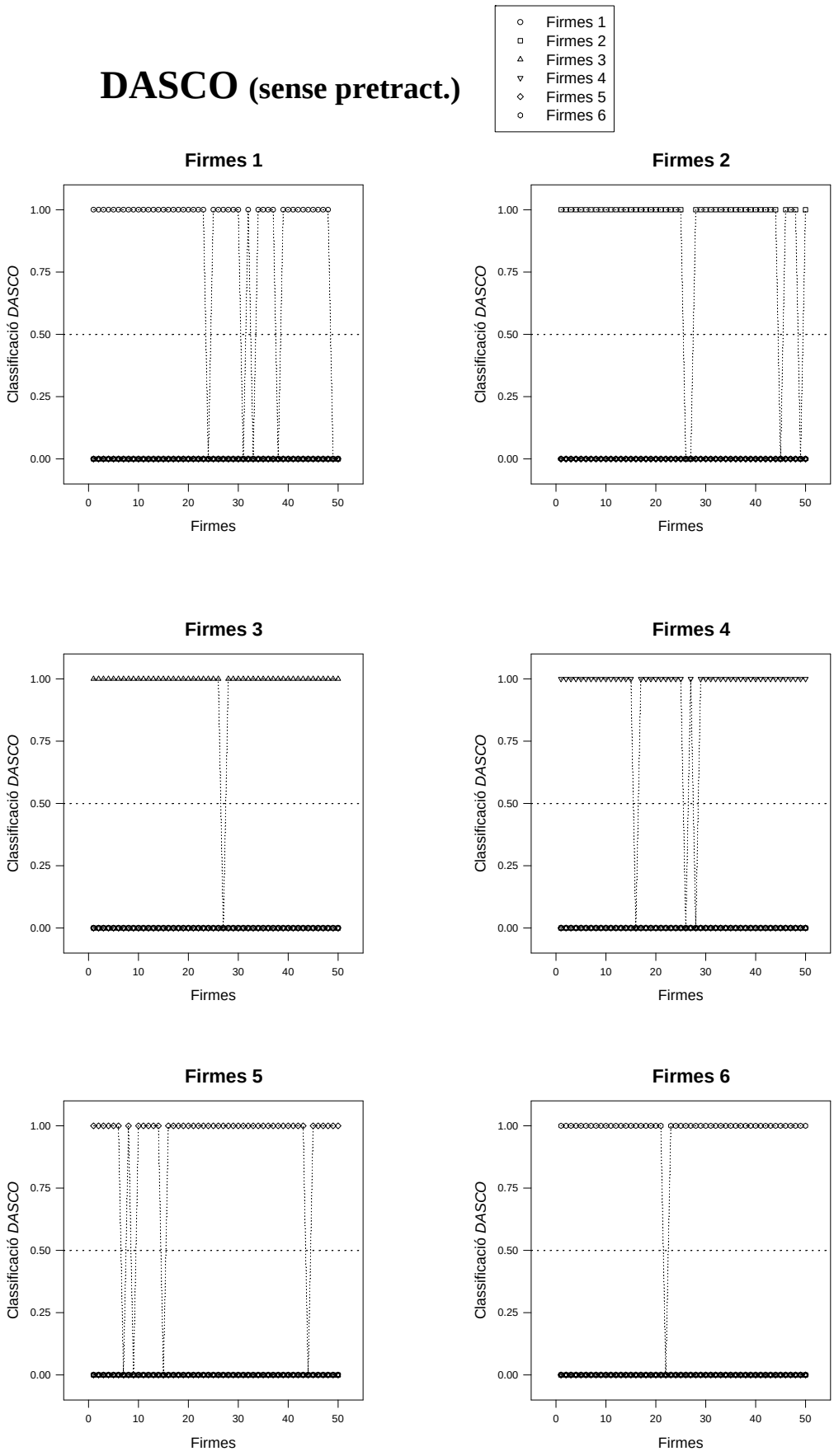

DASCO (sense pretract.)
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació DASCO
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.21.

7.8. Classificació amb *k*NN

Una tècnica molt emprada en classificació és la dels *k* veïns propers o *k*NN (vegeu el capítol 5). La bibliografia suggereix agafar entre tres i cinc veïns per raons de cost computacional.

Primer s'han agafat les dades (sense pretractament), després s'ha aplicat la CVA i, agafant les cinc primeres variables canòniques, s'ha aplicat aquesta tècnica. La taula següent mostra els resultats de classificació produïts per la *k*NN sobre els 300 objectes del conjunt de calibratge en variar el nombre de veïns de dos a vuit.

Nombre de veïns	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
2	5 de la CVA	0	0,00 %	0,00
3	5 de la CVA	0	0,00 %	0,00
4	5 de la CVA	0	0,00 %	0,00
5	5 de la CVA	0	0,00 %	8,89e-03
6	5 de la CVA	0	0,00 %	1,81e-02
7	5 de la CVA	0	0,00 %	2,30e-02
8	5 de la CVA	0	0,00 %	2,47e-02

Taula 7.10.

Els resultats anteriors confirmen els suggeriments bibliogràfics d'agafar entre tres i cinc veïns. En aquest cas farem la classificació amb tres i quatre veïns, resultats que mostra la taula següent:

Nombre de veïns	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
3	5 de la CVA	2	0,11 %	2,22
4	5 de la CVA	2	0,11 %	2,18

Taula 7.11.

Les figures següents mostren els resultats de classificació dels 300 objectes de test aplicant la *k*NN amb 3 i 4 veïns.

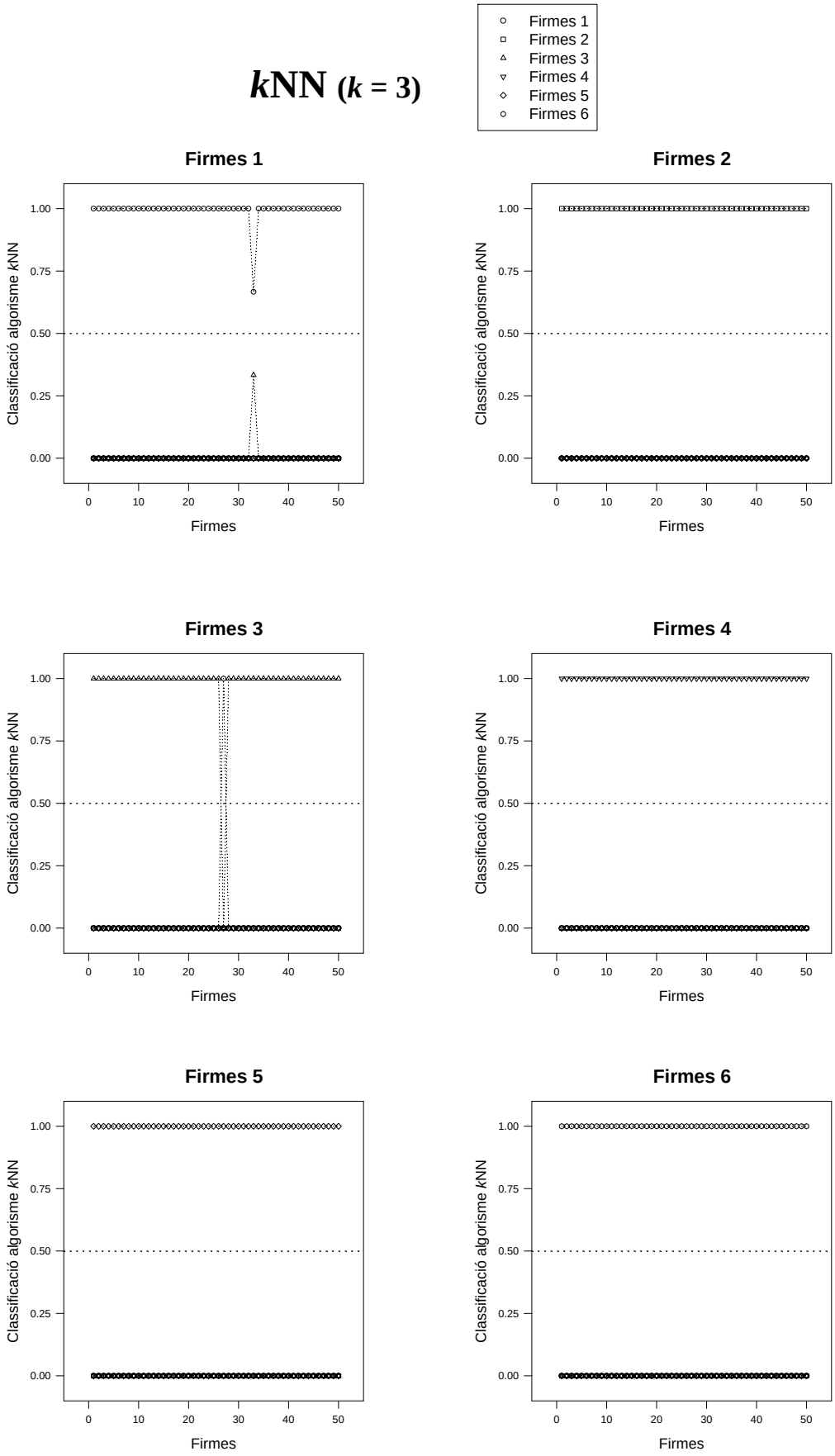

kNN (k = 3)
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació algorisme kNN
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.22.

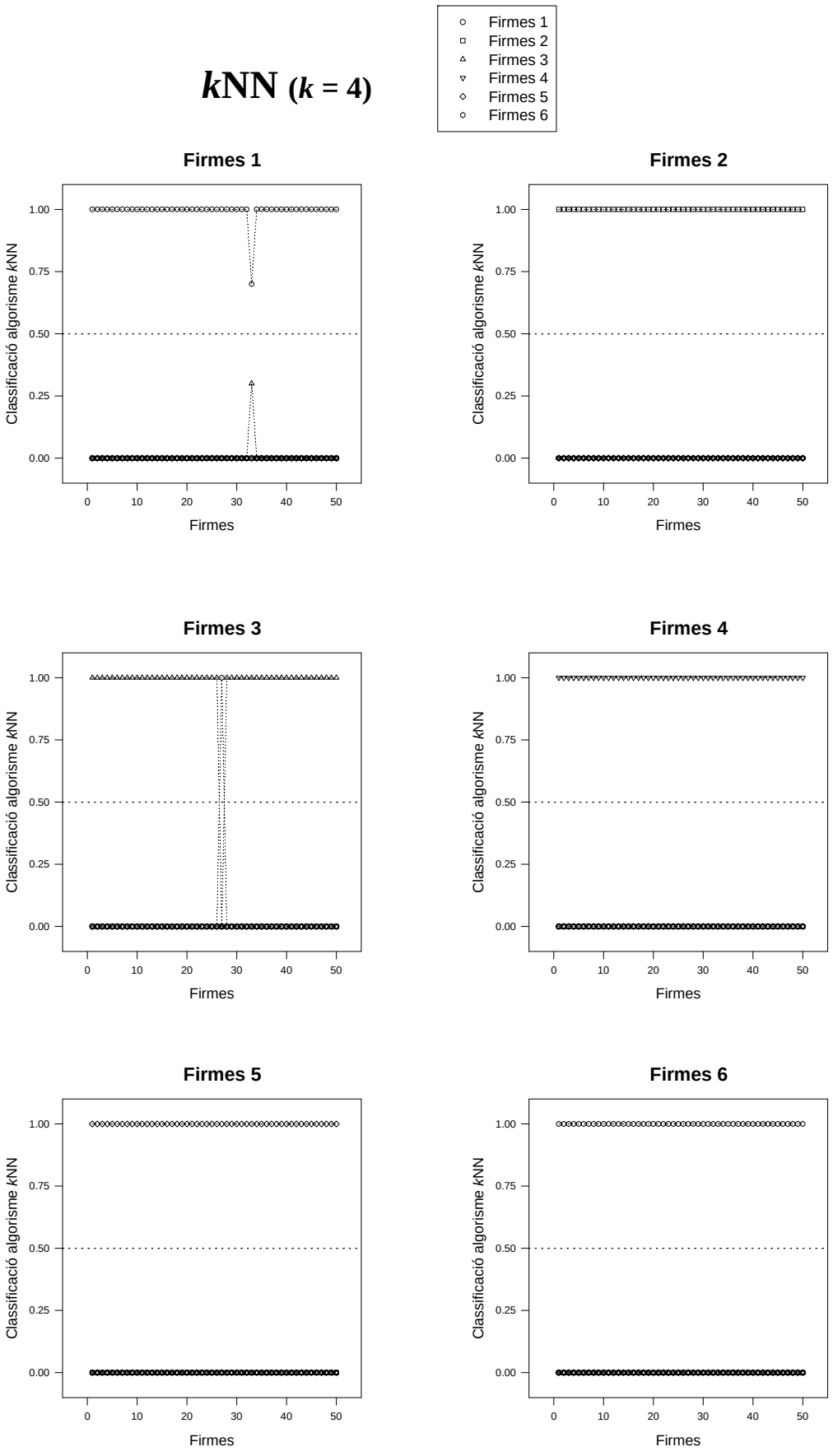

kNN (k = 4)
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació algorisme kNN
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.23.

7.9. Classificació amb el paral·lelepípede que conté els objectes de calibratge

Aquesta tècnica és una proposta nova d'aquest treball. És una tècnica supervisada i no paramètrica (no pressuposa cap tipus de distribució de les dades del problema) que té un cost computacional molt baix.

Primer s'han agafat les dades (sense haver fet cap pretractament), després s'ha aplicat la CVA i, a la matriu resultant, li ha estat aplicada directament aquesta tècnica de classificació. S'ha fet un programa que variï iterativament el nombre de variables retingudes entre 1 i 5, així com els valors del factor *k*, que modula la velocitat de decreixement de les funcions exponencials decreixents de la fórmula 7.6 de 0,5 a 2,00, en increments successius de 0,05 (vegeu l'apartat 5.2.5.).

$$g_{1ij} = \exp\left(-k.\left|\frac{x - min_{ij}}{max_{ij} - min_{ij}}\right|\right) \quad , \quad g_{2ij} = \exp\left(-k.\left|\frac{x - max_{ij}}{max_{ij} - min_{ij}}\right|\right) \tag{7.6}$$

La figura següent indica la influència del valor del factor *k* de l'exponencial decreixent:

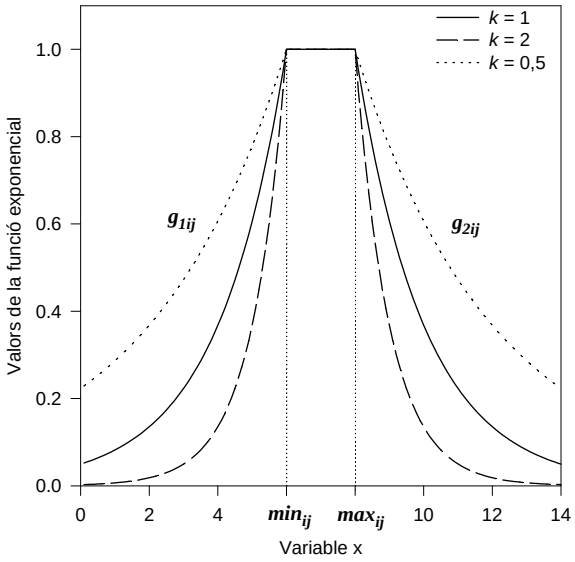


Figura 7.24.

S'ha dividit el conjunt de calibratge en dos subconjunts: un de 240 objectes i un altre de 60 objectes. El primer serveix per generar múltiples models de classificació i el segon serveix per determinar quin d'aquests és el que proporciona millors resultats.

La taula següent mostra els resultats de classificació dels 60 objectes (separats del conjunt de calibratge) segons diferents models de classificació generats a partir dels 240 objectes de calibratge:

Nombre de variables retingudes	Factor exponencial	Sortides errades de 360	Error total	PRESS
5 de la CVA	1,05	1	0,28 %	3,86
5 de la CVA	1,10	0	0,00 %	3,46
5 de la CVA	1,15	0	0,00 %	3,13
5 de la CVA	1,20	0	0,00 %	2,84
5 de la CVA	1,25	0	0,00 %	2,61
5 de la CVA	1,30	0	0,00 %	2,41
5 de la CVA	1,35	0	0,00 %	2,24
5 de la CVA	1,40	1	0,28 %	2,10

Taula 7.12.

Els resultats de la taula anterior suggereixen que el millor model és el que reté cinc variables i pren un valor del factor de l'exponencial decreixent d'1,35 (és el model que menys sortides errades proporciona i té alhora un PRESS més baix).

Els resultats de classificació dels 300 objectes de test (agafant els 300 objectes de calibratge, les cinc variables i un *factor* = 1,80) estan expressats en la taula següent:

Nombre de variables retingudes	Factor exponencial	Sortides errades de 1.800	Error total	PRESS
5 de la CVA	1,35	9	0,50 %	17,27

Taula 7.13.

La figura següent mostra els resultats de classificació dels 300 objectes del conjunt de test.

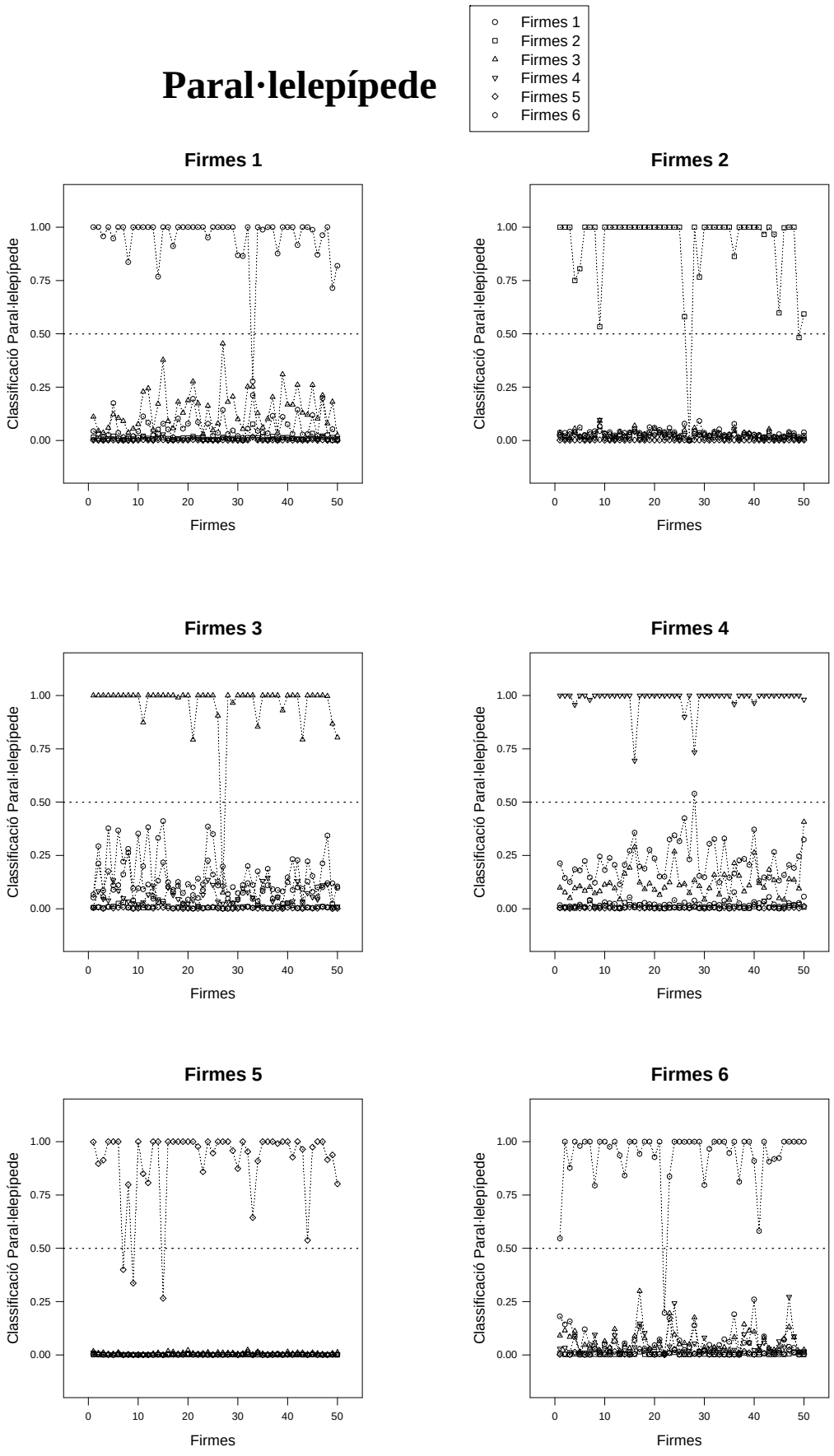

Paral·lelepípede
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació Paral·lelepípede
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.25.

7.10. Classificació amb xarxes neuronals

Les xarxes neuronals s'han creat i entrenat amb el Neural Network Toolbox del Matlab. Per entendre'n el funcionament cal consultar les referències [Mat93] i [Mat94].

Primer s'han agafat les dades (sense haver fet cap pretractament), després s'ha aplicat la CVA i, a la matriu resultant, li ha estat aplicada directament aquesta tècnica (tenint en compte les cinc variables canòniques).

En el cas de les xarxes neuronals tipus *Backpropagation* s'ha creat un model de xarxa amb una única capa oculta de quinze neurones amb funció de transferència *logsig* i una capa de sortida formada per sis neurones amb funció de transferència lineal. El fet que siguin quinze neurones i no un altre nombre es deu al fet que l'entrenament (calibratge de la xarxa) es comença amb un nombre baix de neurones ocultes, que s'augmenta fins que s'aconsegueix fer baixar l'error. En aquest cas, amb quinze neurones s'ha pogut entrenar satisfactòriament la xarxa i no ha calgut emprar un nombre més gran de neurones en la capa oculta. Per entrenar la xarxa, primer s'ha utilitzat la funció *trainbp* del Matlab i quan l'error ha disminuït fins a 1 s'ha aplicat la funció *trainlm* fins a assolir un error global inferior a 1e-4.

La taula següent mostra els resultats de classificació dels 300 objectes de test proporcionats per la xarxa *Backpropagation*:

Criteri	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
Backprop.	5 de la CVA	5	0,28 %	11,02

Taula 7.14.

En el cas de les xarxes neuronals Radial Basis Functions, en la fase d'entrenament (calibratge) el Matlab exigeix triar el valor d'un factor d'aprenentatge. Segons el valor triat d'aquest paràmetre la xarxa s'entrenarà de forma diferent i, per tant, els resultats de classificació no coincidiran. Per això s'ha dividit el conjunt de calibratge en dos subconjunts, de 240 i 60 objectes. El primer d'aquests (240 objectes) ha servit per entrenar les diferents xarxes i el segon (60 objectes) ha servit per determinar quina d'aquestes és la que proporciona millors resultats. Després de provar diferents valors del factor d'aprenentatge, s'ha comprovat que, en aquest cas, amb un factor d'aprenentatge de 0,25 s'obtenien els errors més petits.

Aquest tipus de xarxes són molt sensibles al nombre de neurones de la capa oculta i, per tant, és molt important determinar quin és el nombre òptim de neurones ocultes.

La taula següent mostra els resultats de classificació dels 60 objectes del conjunt de calibratge en crear diferents models de xarxes diferenciats pel nombre de neurones de la capa oculta:

Factor	Nombre de variables retingudes	Nombre de neurones	Sortides errades de 360	Error total	PRESS
0.25	5 de la CVA	125	0	0,00 %	0,2904
0.25	5 de la CVA	150	0	0,00 %	0,2609
0.25	5 de la CVA	175	0	0,00 %	0,1988
0.25	5 de la CVA	200	0	0,00 %	0,2180
0.25	5 de la CVA	240	0	0,00 %	0,2206

Taula 7.15.

De tots els models anteriors, ha estat seleccionat el de 175 neurones, ja que és el que presenta el PRESS més baix. Per tant, amb aquest mateix model s'ha efectuat la classificació dels 300 objectes de test. Els resultats obtinguts es mostren en la taula següent:

Nombre de variables retingudes	Nombre de neurones	Sortides errades de 1800	Error total	PRESS
5 de la CVA	175	2	0,11 %	4,36

Taula 7.16.

Les figures 7.26. i 7.27. mostren els resultats de classificació proporcionats per les xarxes tipus *Backpropagation* i *Radial Basis* sobre els 300 objectes de test:

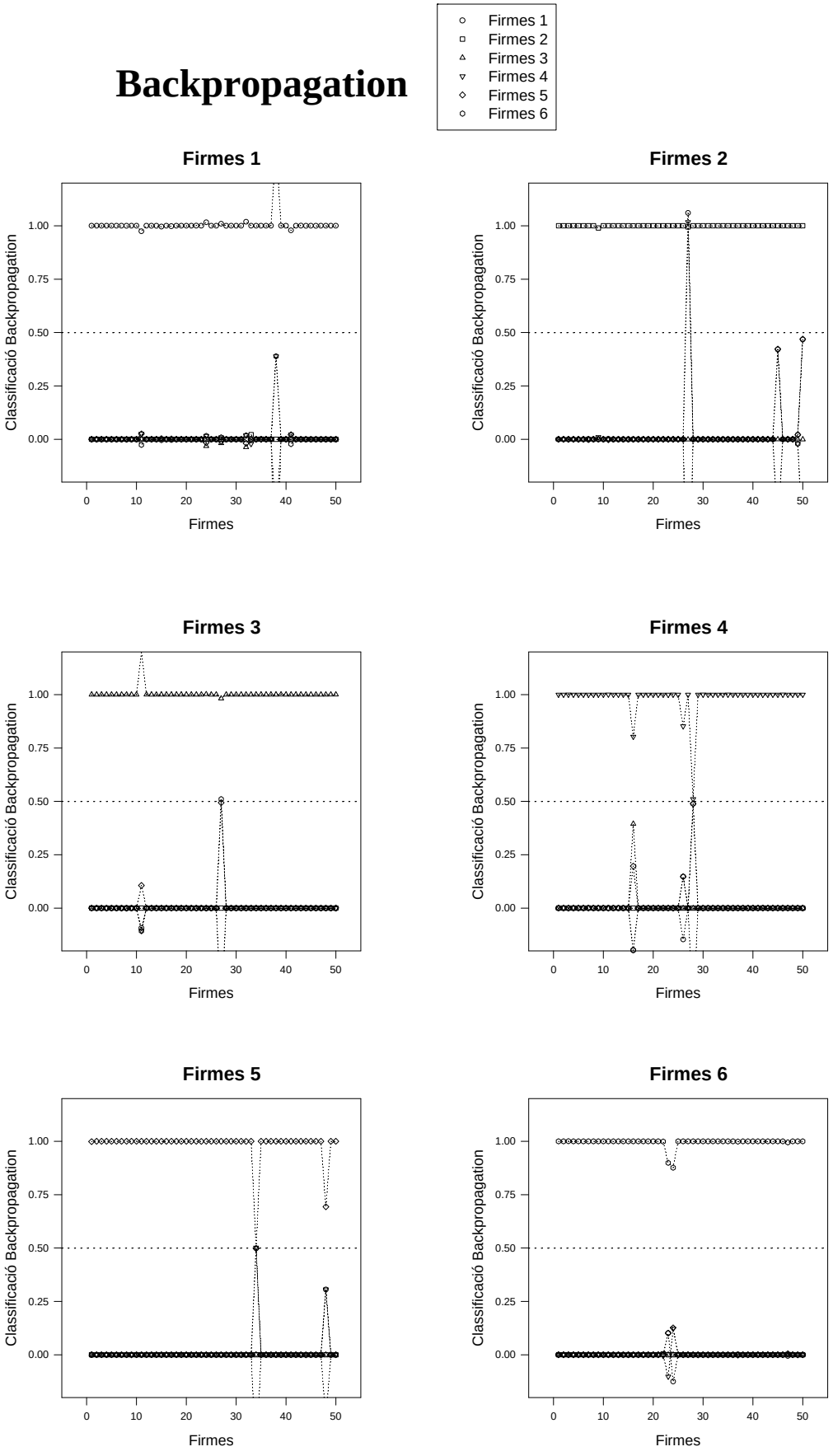

Backpropagation
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació Backpropagation
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.26.

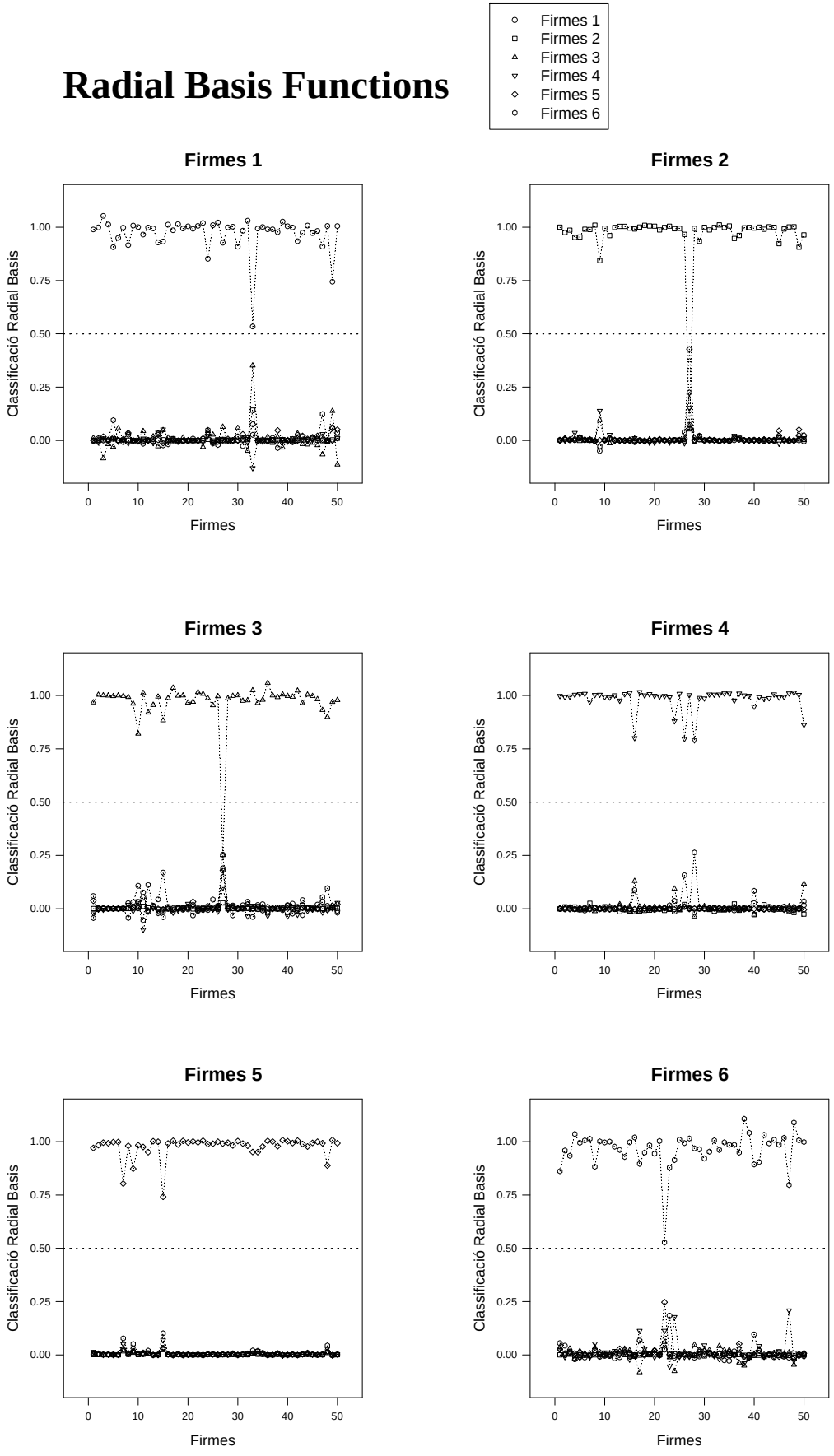

Radial Basis Functions
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació Radial Basis
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.27.

7.11. Classificació amb lògica difusa

Primer s'han agafat les dades (sense haver fet cap pretractament), després s'ha aplicat la CVA i, a la matriu resultant, li ha estat aplicada directament aquesta tècnica de classificació. La lògica difusa és una tècnica supervisada. Primer es calcula el valor mitjà i la desviació típica de totes les variables de cada una de les classes dels objectes de calibratge.

En aquest treball s'ha optat per definir 3 conjunts difusos (bo, regular i dolent) –tot i que se'n podrien definir més– de forma triangular per a cada variable de cada una de les classes. Per a cada variable *x* de cada classe es fa la transformació $x' = |x - \bar{x}|$, on $\bar{x}$ és el valor mitjà de la variable per a la classe en qüestió. També s'ha cregut convenient centrar els tres conjunts difusos (bo, regular i dolent) en els punts *0*, $factor.\sigma$ i $2.factor.\sigma$, respectivament (σ és la desviació típica de la variable considerada). La figura següent mostra com queden definits els tres conjunts difusos per a cada variable de cada una de les diferents classes:

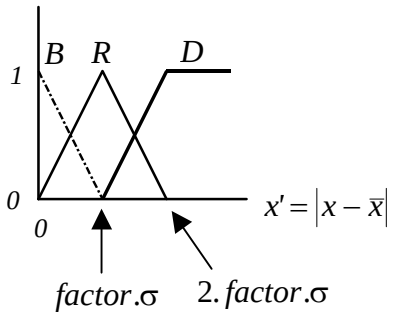


Figura 7.28.

Per trobar el valor òptim del factor s'ha fet un programa que permeti variar els valors d'aquest des de 0,1 fins a 3 en increments de 0,05. També s'ha variat el nombre de variables retingudes entre 1 i 5. S'ha dividit el conjunt de calibratge en dos subconjunts: un de 240 objectes i un altre de 60 objectes. El primer serveix per generar múltiples models de classificació i el segon serveix per determinar quin d'aquests és el que proporciona millors resultats.

La taula següent mostra els resultats de classificació dels 60 objectes (separats del conjunt de calibratge) segons els diferents models de classificació generats a partir dels 240 objectes de calibratge:

Nombre de variables retingudes	Factor	Sortides errades de 360	Error total	PRESS
5 de la CVA	1,65	8	2,22 %	22,38
5 de la CVA	1,70	5	1,39 %	22,92
5 de la CVA	1,75	5	1,39 %	23,56
5 de la CVA	1,80	4	1,11 %	24,20
5 de la CVA	1,85	5	1,39 %	24,89
5 de la CVA	1,90	6	1,67 %	25,65
5 de la CVA	1,95	6	1,67 %	26,48
5 de la CVA	2,00	8	2,22 %	27,35

Taula 7.17.

Els resultats de la taula anterior suggereixen que el millor model és el que reté cinc variables canòniques i agafa un factor d'1,80 (és el model que proporciona menys sortides errades).

Els resultats de classificació dels 300 objectes de test (agafant els 300 objectes de calibratge, les cinc variables i un factor = 1,80) estan expressats en la taula següent:

Nombre de variables retingudes	Factor	Sortides errades de 1.800	Error total	PRESS
5 de la CVA	1,80	33	1,83 %	129,24

Taula 7.18.

Els resultats de classificació dels 300 objectes del conjunt de test es mostren en la figura següent.

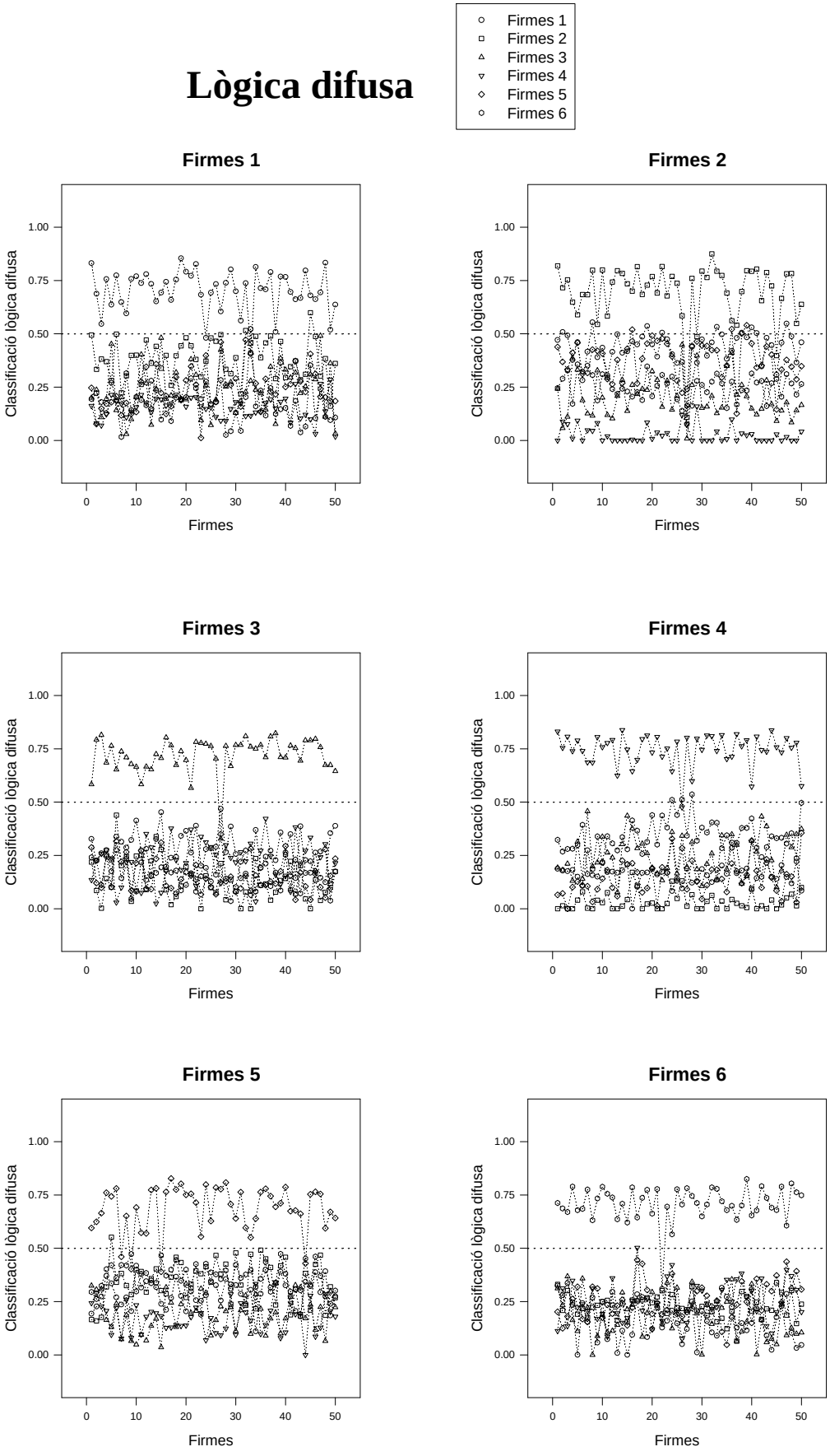

Lògica difusa
Firmes 1
Firmes 2
Firmes 3
Firmes 4
Firmes 5
Firmes 6
Classificació lògica difusa
Firmes
0.00
0.25
0.50
0.75
1.00
0
10
20
30
40
50

Figura 7.29.

7.12. Comparativa dels diferents mètodes

La següent taula mostra un resum dels resultats de classificació dels 300 objectes del conjunt de test obtinguts aplicant tots els mètodes utilitzats en aquest capítol:

Tècnica	Nombre de variables retingudes	Sortides errades de 1.800	Error total	PRESS
PCR	28	11	0,61 %	44,41
PLS	28	8	0,44 %	46,21
LDA	5 de la CVA	3	0,17 %	3,00
QDA	5 de la CVA	5	0,28 %	5,00
SIMCA	1 per classe	47	2,61 %	47,00
DASCO	1 per classe	19	1,06 %	19,00
*k*NN (*k* = 4)	5 de la CVA	2	0,11 %	2,18
Paral·lelepípede	5 de la CVA	9	0,50 %	17,27
Backpropagation	5 de la CVA	5	0,28 %	11,02
Radial Basis	5 de la CVA	2	0,11 %	4,36
Lògica difusa	5 de la CVA	33	1,83 %	129,24

Taula 7.19.

La taula anterior indica que, per a aquest problema concret, les tècniques que proporcionen millors resultats són la *k*NN (amb tres o quatre veïns) i la xarxa neuronal tipus *Radial Basis Functions*, ja que ambdues només fallen en dues de les 1.800 sortides. Després, per ordre, les segueixen la LDA, la QDA i la xarxa neuronal tipus *Backpropagation*.