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

6. Resultats experimentals. Segells

En aquest capítol es fa un reconeixement invariant de quatre tipus diferents de segells, d'idèntiques dimensions reals. Per tal d'efectuar el reconeixement es posaran en pràctica quasi totes les tècniques explicades en els capítols anteriors. Es disposa de cent cinquanta mostres (objectes) de cada un dels segells, cinquanta de les quals formaran part del conjunt de calibratge. Les cent restants seran el conjunt de test. Els segells han estat digitalitzats un a un amb el mateix digitalitzador, a una resolució de 128 x 128 píxels i amb 256 nivells de gris.

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

1.- Binarització de cada objecte, prenent com a nivell de gris de tall el 128.

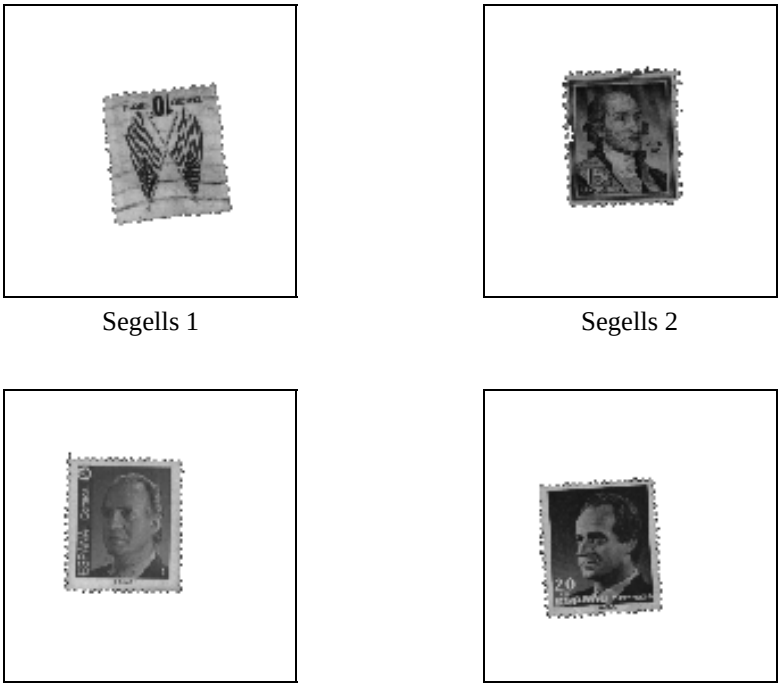
2. Càlcul de trenta característiques discriminants per objecte. Això genera una matriu de calibratge de cinquanta files i quaranta-vuit columnes per classe. Aquest punt es desenvolupa en l'apartat 6.1.
3. Estudi i selecció de la millor tècnica de reducció de dimensions. S'apliquen la CVA, la DPCA, l'OCVA i la PCA. Aquest punt es desenvolupa en l'apartat 6.2.
4. Selecció del nombre òptim de variables que s'han de retenir. Aquest punt es desenvolupa a l'apartat 6.3.
5. Determinació del nombre mínim d'objectes de calibratge. Aquest punt es desenvolupa en l'apartat 6.4.

6. Estudi comparatiu dels diferents mètodes de predicció i de classificació. Aquest punt es desenvolupa en l'apartat 6.5.

Les diferents mostres de segells han estat digitalitzades en diferents posicions, mides i orientacions dintre de l'escena. S'ha fet així per poder realitzar un reconeixement invariant a transformacions lineals dels objectes. La figura 6.1. mostra diferents posicions, orientacions i mides d'un mateix segell:

Figura 6.1.

La figura següent mostra els quatre tipus diferents de segells utilitzats:



Segells 1 Segells 2

Segells 3 Segells 4

Figura 6.2.

6.1. Càlcul de les característiques

De totes les característiques invariants explicades en el capítol 2, en el cas dels segells no té massa sentit calcular les que depenen del contorn o de l'esquelet perquè tots els segells tenen el mateix contorn exterior. Per tant, calcularem les característiques restants.

Després d'haver fet una selecció de les característiques que funcionen millor, s'ha optat per calcular per a cada segell les trenta característiques següents:

- Els següents vuit moments calculats de l'histograma:

  $$m_1, m_2, m_3, m_4, \mu_1, \mu_2, \mu_3 \text{ i } \mu_4.$$

- Els primers quatre moments invariants de Hu:

  $$\varphi[1], \varphi[2], \varphi[3] \text{ i } \varphi[4].$$

- 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 factor de circularitat *C*.

- Els sis moments de les projeccions següents:

  $$\eta_2^H, \eta_2^V, \eta_4^H, \eta_4^V, K^H \text{ i } K^V.$$

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

  $$D, a, b, E = b/a, \varepsilon = c/a \text{ i } A_1.$$

No s'han calculat els descriptors de Fourier perquè en el capítol 2 s'ha vist que, a més de requerir un càlcul molt intensiu, en molts casos no presenten un nivell de discriminació massa elevat. Això es deu al fet que es basen en la transformada de Fourier, la qual a part de requerir un nombre important d'iteracions per al seu còmput, és molt sensible a petites variacions de les freqüències fonamentals dels objectes.

Excepte en el cas dels moments calculats a partir de l'histograma (aquestes característiques es basen en l'histograma i, per tant, en els nivells de gris de l'objecte), per al còmput de totes les altres característiques, les imatges han estat binaritzades amb un valor de tall de 128, ja que així la dispersió de les característiques obtingudes és menor que treballant amb els 256 nivells de gris. En aquest problema en concret només s'ha realitzat una binarització perquè les característiques obtingudes ja aportaven prou informació sobre les diferents classes, però si fos necessari, es podrien realitzar diverses binaritzacions de les imatges. En cadascuna d'aquestes binaritzacions s'hauria de prendre un nivell de gris de tall diferent i

s'hauria de realitzar el càlcul de totes les característiques excepte de les basades en l'histograma.

Amb un ordinador *Pentium II* a 233 MHz i amb programari (*software*) compilat a 32 bits, el temps de càlcul mitjà de les trenta característiques és de 0,13 segons per segell, però amb un PC d'última generació aquest temps es pot reduir substancialment.

6.2. Selecció de la millor tècnica de reducció de dimensions

En aquest treball es tracten diferents mètodes per reduir el nombre de variables (reducció de dimensions), que són la CVA, la DPCA, l'OCVA i la PCA. Si el problema tingués dues o tres dimensions i dues o tres classes, fent una representació gràfica de tots els objectes similar a la que mostra la figura 6.3., visualment es podria determinar quin és el sistema de reducció de dimensions que funciona millor. Però quan la complexitat del problema és gran, aquest mètode deixa de ser viable, a més de ser un mètode subjectiu.

Com que es disposa de 4 classes de segells, cinquanta segells de calibratge per classe i trenta característiques per segell, a causa de la gran complexitat del problema seria útil disposar d'un índex que permetés mesurar el grau de dispersió de les diferents classes. Duda i Hart [Dud73] proposen diferents criteris per calcular aquest índex, però el que s'ha triat es calcula de la manera següent:

$$I_M = traça(W^{-1}.B) = \sum_{i=1}^{m} \lambda_i \qquad (6.1)$$

on la traça d'una matriu quadrada és la suma dels elements de la seva diagonal i els λ_i són els valors propis de la matriu. Les matrius *W* i *B* són, respectivament, la matriu de dispersió interna de les classes (Within-Group Dispersion Matrix) i la matriu de dispersió entre classes (Between-Group Dispersion Matrix), i es calculen de la manera següent:

$$W_{(m,m)} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \cdot (x_{ij} - \bar{x}_i)^T \qquad (6.2)$$

$$B_{(m,m)} = \sum_{i=1}^{c} n_i . (\bar{x}_i - \bar{x}) \cdot (\bar{x}_i - \bar{x})^T \qquad (6.3)$$

La matriu *W* informa de la dispersió entre els objectes d'una mateixa classe, mentre que la matriu *B* informa de la dispersió o el distanciament entre classes diferents.

La igualtat de l'expressió 6.1 és certa perquè la traça d'una matriu simètrica $-W^{-1}.B-$ és igual a la suma dels seus valors propis.

L'índex I_M és una magnitud invariant a transformacions lineals de les dades, ja que els valors propis ho són. El valor numèric d'aquest índex és una mesura del quocient entre la dispersió entre classes i la dispersió interna de les classes (intuïtivament s'està fent *B/W*) en la direcció dels vectors propis. Per tant, com més distanciades es trobin les diferents classes més alt serà el valor d'aquest índex. Consulteu [Sta82].

L'índex I_M proporciona un valor que es calcula a partir d'una realitat multidimensional molt complexa. La importància d'aquest índex no és el valor numèric concret que proporciona, sinó el coneixement de si aquest valor es troba per sobre o per sota del proporcionat per una altra situació. Per tant, la interpretació d'aquest índex ha de ser més qualitativa que quantitativa.

La figura següent mostra dues situacions diferents. En la primera el valor de l'índex I_M serà més gran que en la segona, perquè les diferents classes es troben més separades.

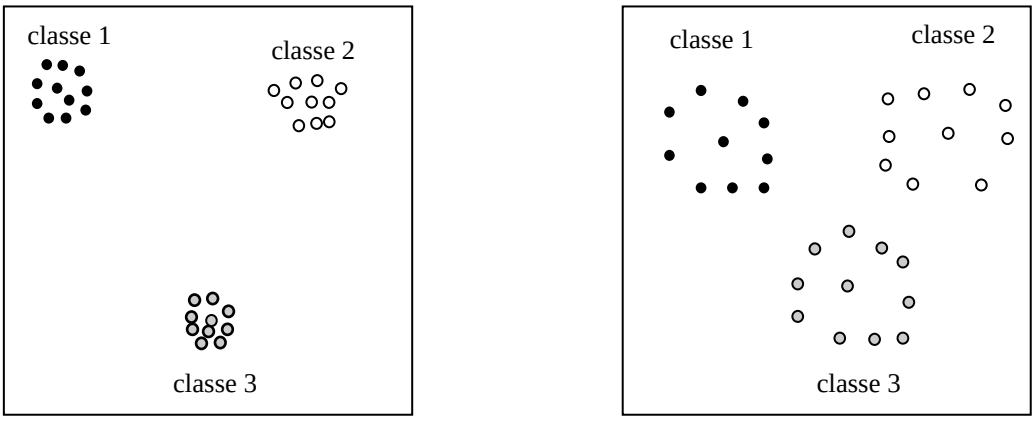


Figura 6.3.

L'índex I_M, en calcular la separació entre els centres de les diferents classes, té en compte tant la dispersió interna dels objectes pertanyents a una mateixa classe com la dispersió entre classes. Si les diferents classes estan molt separades, això significa que les distàncies entre els centres de les classes són grans en comparació amb les dispersions internes de cadascuna i, per tant, el valor de l'índex I_M serà elevat. En el cas contrari, si les diferents classes es troben encavalcades, això significa que les distàncies entre els centres de les classes són petites en comparació amb les dispersions internes de les classes, fet que fa que el valor de l'índex I_M sigui petit.

L'índex I_M serà útil per determinar quin dels mètodes de reducció de variables (CVA, DPCA, OCVA i PCA) proporciona màxima separació entre classes. S'ha de seleccionar el mètode de reducció de variables que, amb un nombre reduït de variables, maximitzi el valor de l'índex I_M.

La taula 6.1. mostra els valors proporcionats per l'índex I_M per al problema dels 200 segells (50 x 4 = 200) del conjunt de calibratge. Hem considerat quatre tècniques de reducció de dimensions i tres situacions per a cadascuna d'aquestes: dades sense cap mena de pretractament, centrat i autoescalat previ de les dades.

Índex $\boldsymbol{I_M}$	Nombre variables retingudes	**CVA**	**DPCA**	**OCVA**	**PCA**
Dades sense pretractament	1	12,11	1,47	16,56	1,47
	2	13,02	9,28	23,88	7,66
	3	18,24	**15,00**	25,65	**15,40**
Dades centrades	1	10,13	1,47	16,92	1,47
	2	37,59	9,28	24,12	7,67
	3	39,59	14,98	30,34	15,38
Dades autoescalades	1	33,39	2,20	28,24	0,94
	2	45,67	2,99	29,66	1,11
	3	**50,00**	3,18	**35,22**	1,75

Taula 6.1.

De la taula anterior es dedueix que la separació entre classes augmenta en fer-ho el nombre de variables que retenim. La taula 6.1. ens mostra que la tècnica que amb un nombre reduït de variables proporciona major separació de les diferents classes és la CVA (amb les dades prèviament autoescalades), seguida a certa distància de l'OCVA (amb les dades també autoescalades). A partir d'ara es treballarà amb les variables procedents de la CVA (amb les dades autoescalades), perquè és la tècnica que separa més les diferents classes. Això confirma que en augmentar el nombre de variables retingudes, augmenta el grau de separació entre les diferents classes.

Les figures següents (6.4. a 6.7.) mostren una representació dels 200 objectes del conjunt de calibratge en l'espai de les variables.

La figura 6.4. mostra els resultats proporcionats per la CVA amb les dades autoescalades:

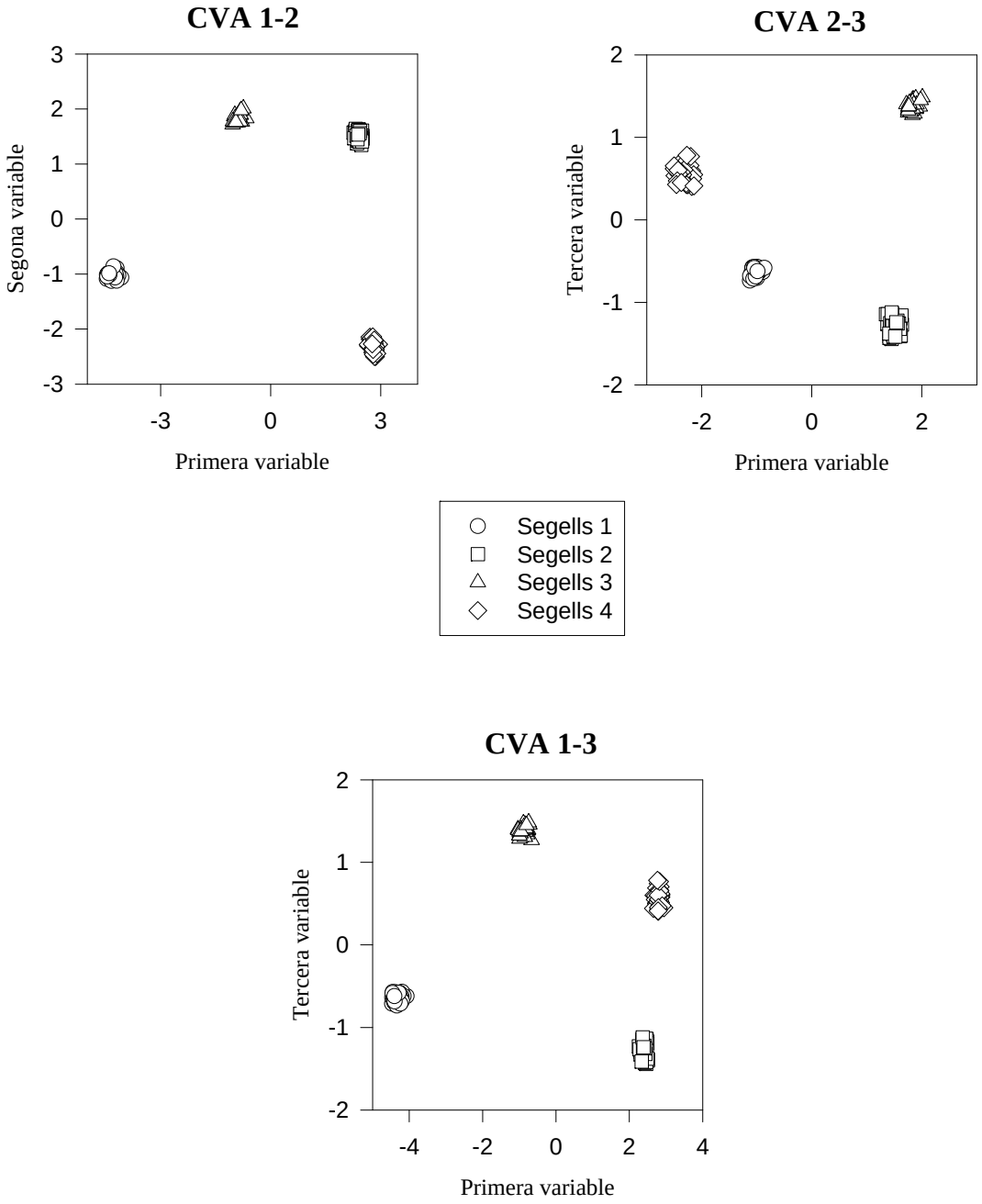


Figura 6.4.

La figura 6.5. mostra els resultats proporcionats per la DPCA sense haver efectuat cap pretractament de les dades:

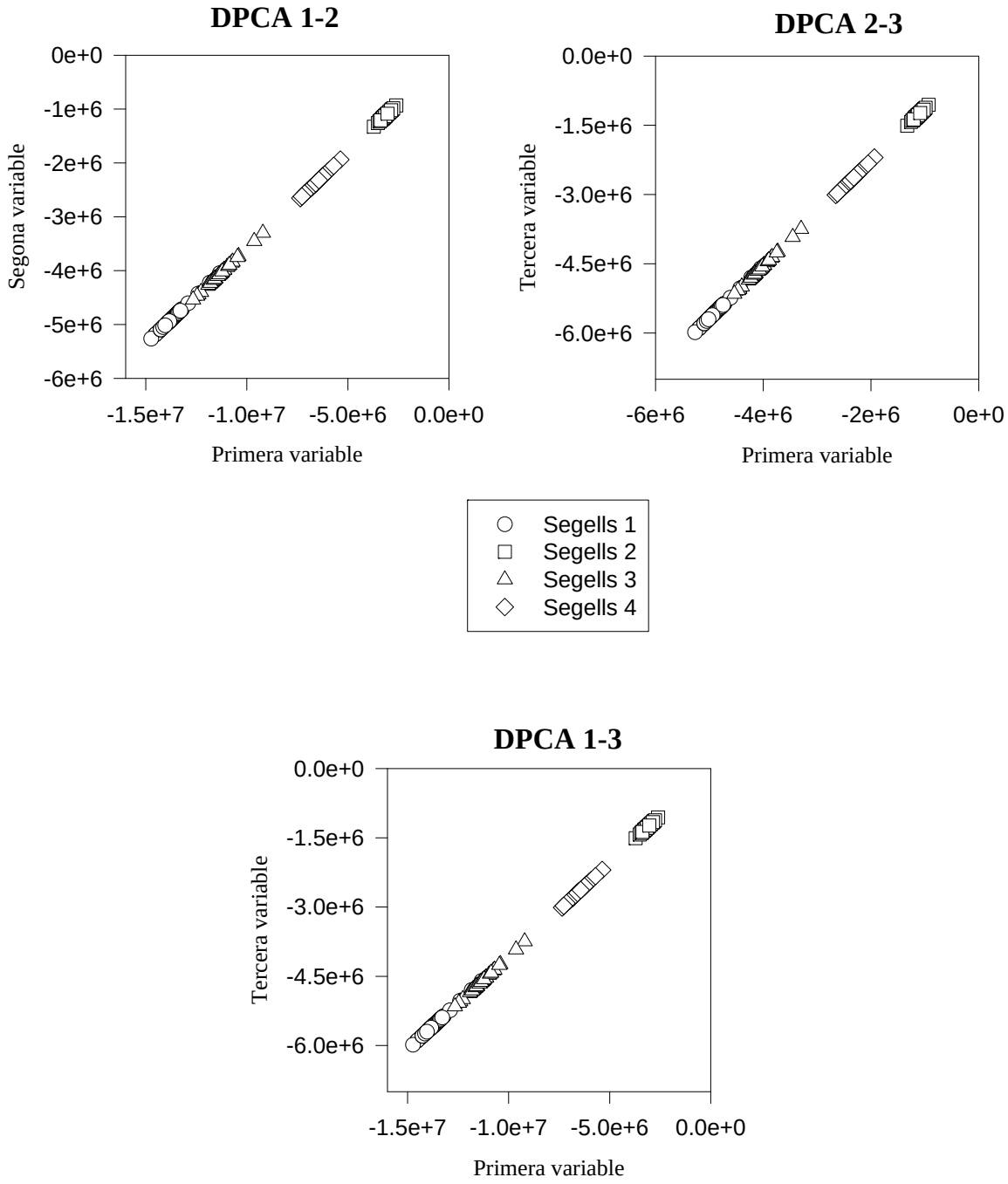


Figura 6.5.

La figura 6.6. mostra els resultats proporcionats per l'OCVA havent autoescalat prèviament les dades:

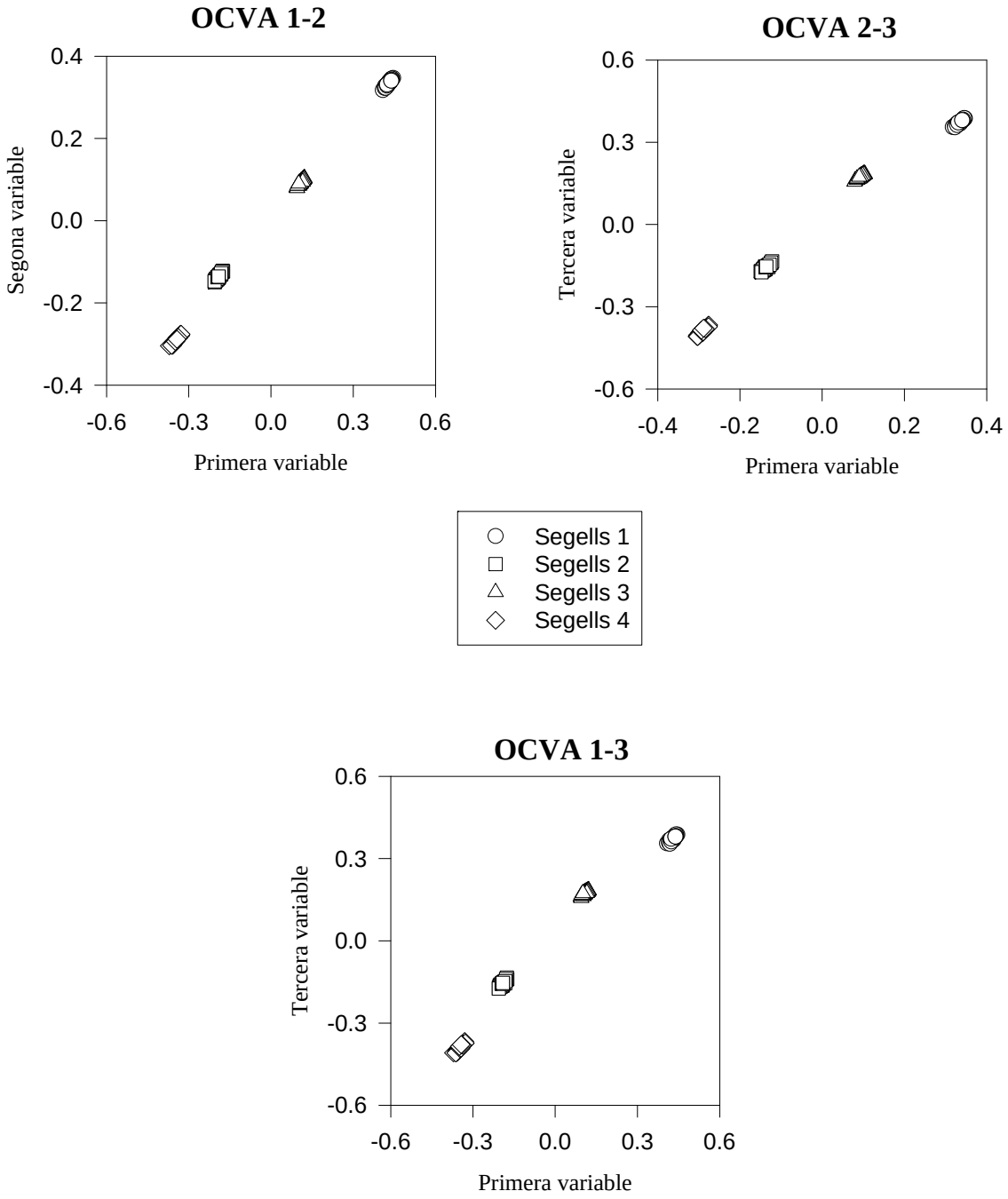


Figura 6.6.

La figura 6.7. mostra els resultats proporcionats per la PCA sense haver fet cap pretractament de les dades:

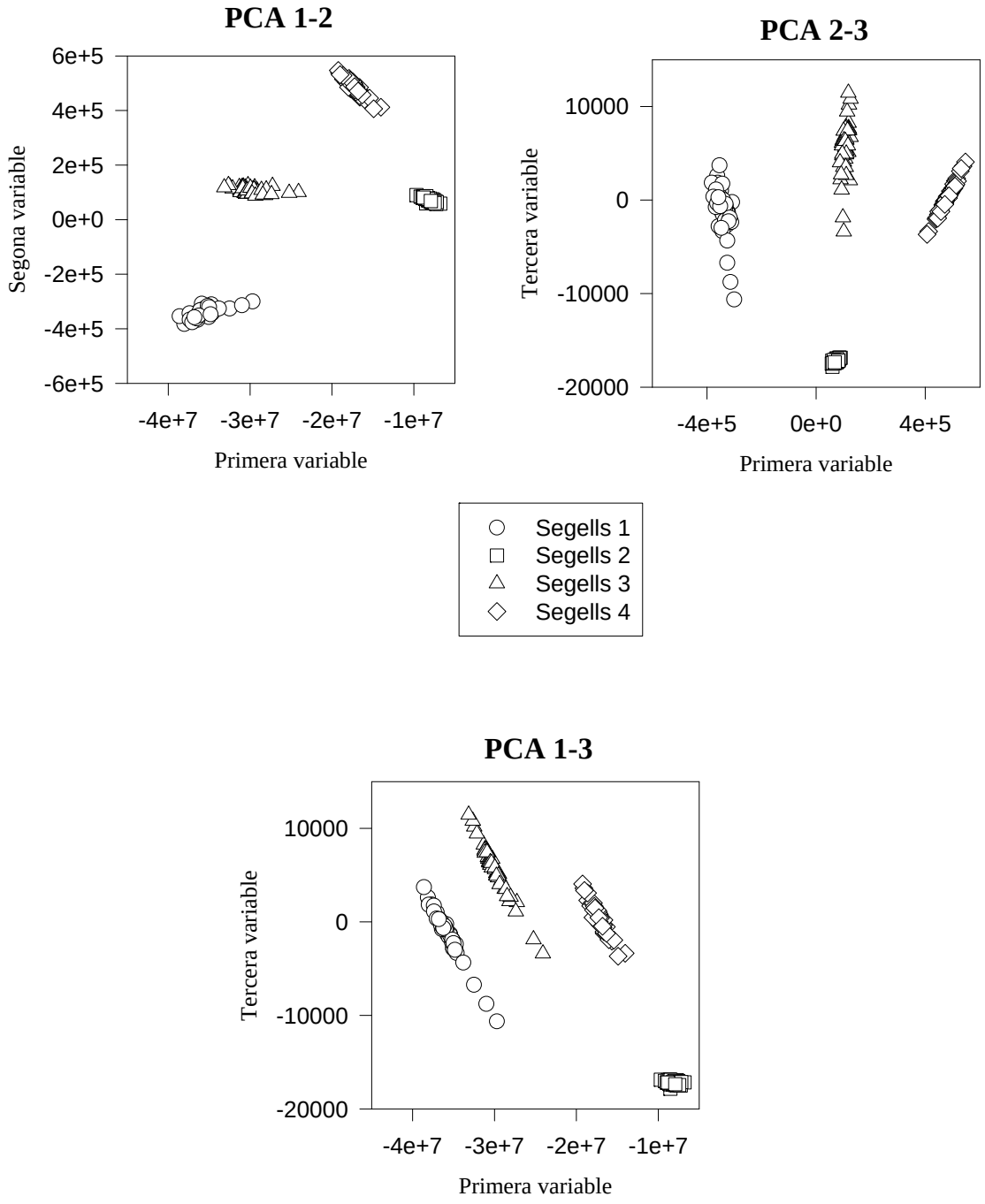


Figura 6.7.

Les figures 6.4. a 6.7. confirmen que l'algorisme de reducció de dimensions que proporciona millors resultats amb les primeres variables és la CVA.

6.3. Selecció del nombre òptim de variables que cal retenir

En l'apartat anterior s'ha vist que, en el cas dels segells, la tècnica de reducció de dimensions que proporciona millors resultats és la CVA aplicada a les dades autoescalades. Per tant, una vegada s'ha aplicat la CVA, es fa necessari decidir quin és el nombre òptim de variables (en aquest cas variables canòniques) que s'han de retenir. En l'apartat 3.6. s'exposen les tècniques per seleccionar el nombre de variables que s'han de retenir.

Els valors propis associats a les tres variables canòniques de la CVA amb les dades autoescalades són els següents:

Valors propis	Valor numèric dels valors propis
λ_1	32,9207
λ_2	11,9551
λ_3	4,2521

Taula 6.2.

En aquest cas, per haver-hi quatre classes només tenim tres variables canòniques i, a la vista de la distribució d'aquestes classes en l'espai de les variables (vegeu la figura 6.6.), queda clar que aquestes tres variables seran suficients per poder afrontar el problema amb garanties d'èxit. Per tant, tot i que amb menys variables retingudes el problema també seria linealment separable, i com que el cost computacional de retenir-ne només tres és força baix, d'ara en endavant tots els algorismes de classificació o de predicció que s'aplicaran per solucionar el problema tindran en compte totes les variables canòniques.

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 s'han de retenir (vegeu l'apartat 6.5.).

6.4. Nombre mínim d'objectes de calibratge

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

La primera d'aquestes tècniques es basa en el càlcul de la DC (distància computada). Primer s'aplica la CVA (per ser la tècnica que millors resultats ens ha proporcionat) sobre les 200 dades de calibratge i s'obté una matriu de transformació de l'espai definit per les característiques discriminants a l'espai de les variables (variables canòniques). 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 180, 160, 140, 120, 100, 80, 60 i 40 objectes, respectivament.

El dos gràfics següents (figures 6.8. i 6.9.) mostren el resultat 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 gràfic, la DC es calcula respecte a les tres primeres variables canòniques (resultants de la CVA amb les dades autoescalades), mentre que en el segon la DC es calcula directament respecte de les trenta característiques discriminants.

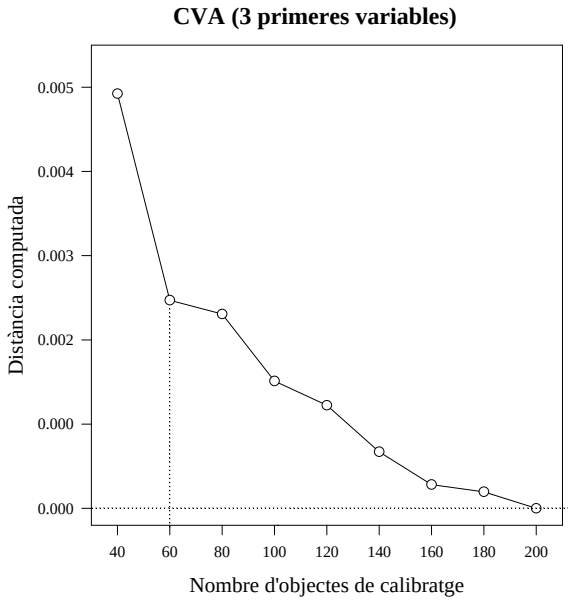


Figura 6.8.

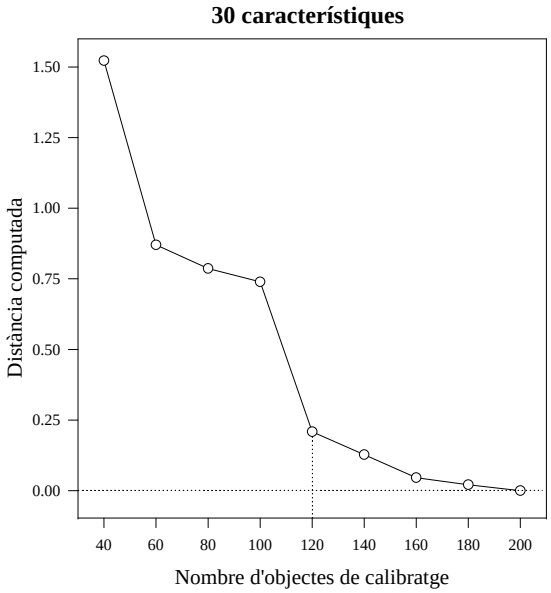


Figura 6.9.

La figura 6.8. mostra que a partir de seixanta objectes de calibratge el pendent de la corba és menys pronunciat. Per tant, sembla que seixanta objectes de calibratge (quinze per classe) són els mínims necessaris per poder afrontar aquest problema amb garanties d'èxit. Aquest resultat concorda amb la hipòtesi de Jain i Chandrasekaran explicada en l'apartat 4.1. ($n > 5.c.m^*$ = 5.4.3 = 60).

En canvi, quan el problema s'ataca directament amb les característiques discriminants (vegeu la figura 6.9.) es requereix un nombre més elevat d'objectes de calibratge i, a més,

la distància computada presenta valors més grans que en el cas de les variables. Aquests resultats indiquen que les característiques són menys eficients que les variables també pel que fa al nombre mínim d'objectes de calibratge.

La segona tècnica és una proposta nova d'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 200 dades de calibratge i sempre amb la mateixa matriu de transformació, es calculen les variables canòniques de subgrups, escollits a l'atzar, de 180, 160, 140, 120, 100, 80, 60 i 40 objectes, respectivament. Ara s'apliquen tècniques ràpides 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, però, tots aquests subgrups porten a una taxa d'errors de classificació del 0,0 %, cosa que fa que no sigui viable l'aplicació d'aquestes tècniques en aquest problema concret. Això es deu al fet que el problema és linealment separable.

6.5. Classificació aplicant tècniques de regressió

En el capítol 5 es va veure que una de les possibilitats per afrontar un procés de classificació és aplicar tècniques multivariables de regressió. Les tècniques de regressió que compararem són la PCR, la PLS, la LRR i la MLR. Com que aquestes tècniques tenen com a base la PCA, avaluarem els valors propis de la PCA amb les dades sense pretractament, centrades i autoescalades, tal com es mostra en la figura següent:

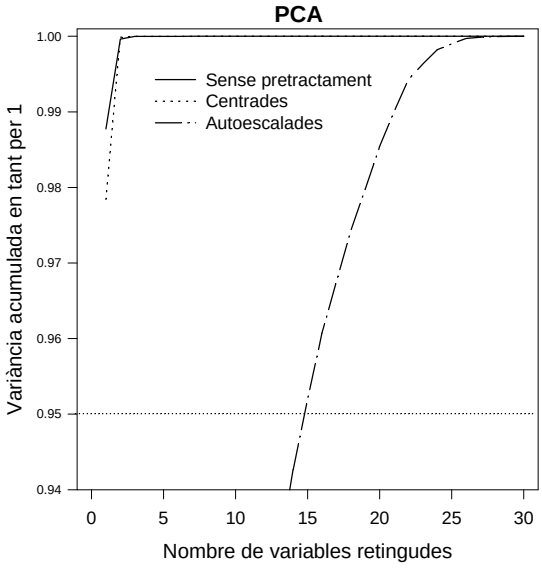


Figura 6.10.

La figura 6.10. indica que, amb les dades sense cap mena de pretractament o amb les dades centrades, és quan menys nombre de variables de la PCA s'han de retenir (cal recordar que quan s'apliquen tècniques de regressió no es pot treballar amb les variables resultants d'aplicar la CVA).

Les figures següents 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 predicció dels 200 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.

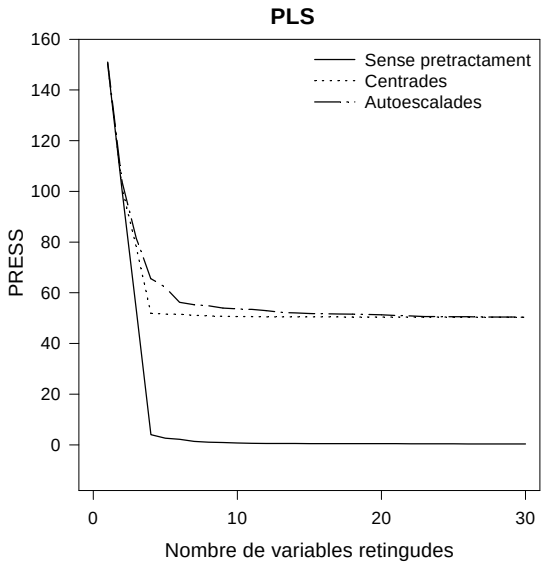


Figura 6.11.

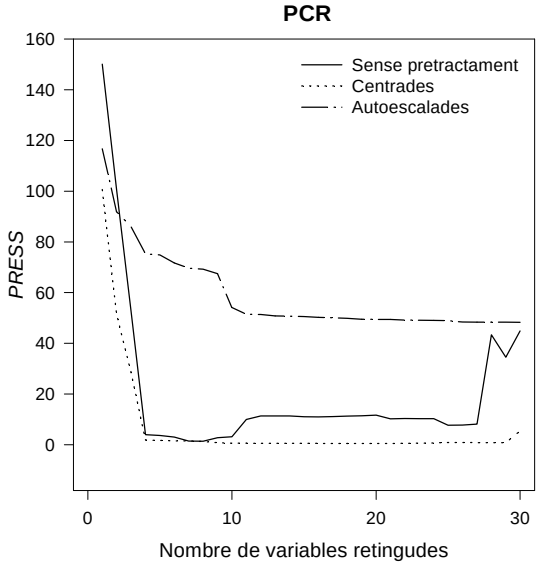


Figura 6.12.

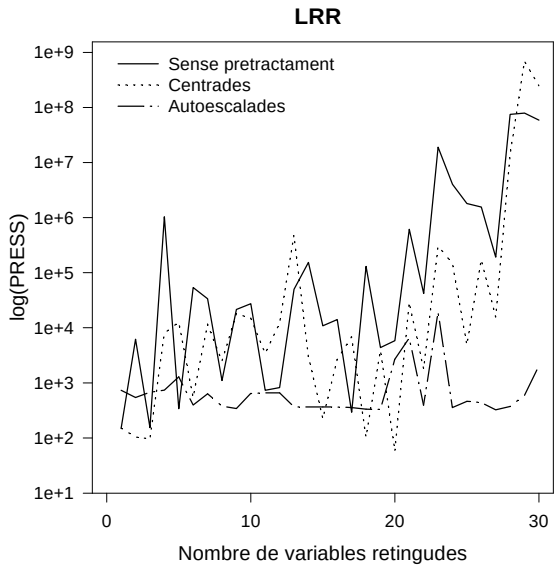


Figura 6.13 .

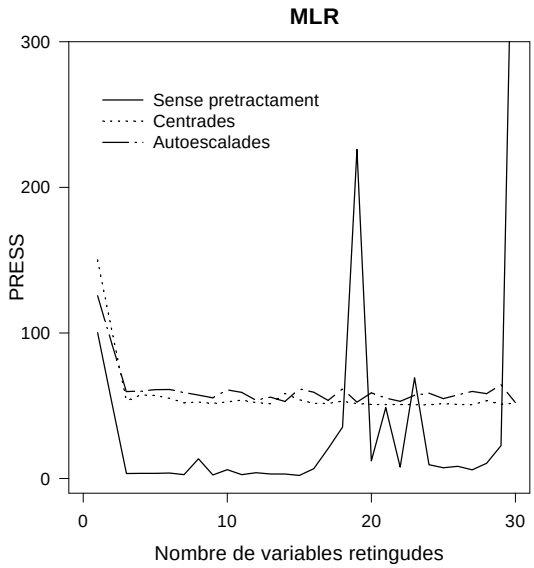


Figura 6.14.

En els gràfics anteriors s'observa que els resultats proporcionats per la LRR i la MLR no són massa estables. Per tant, deixarem de banda aquestes dues tècniques amb vista a solucionar el nostre problema. Això vol dir que per a la predicció d'objectes de test ens quedarem amb la PCR i la PLS.

6.5.1. Resultats de la classificació amb tècniques de regressió

En aquest apartat es classifica el conjunt dels 400 objectes de test.

Per escollir el nombre de variables que s'han de retenir (PCR i PLS) s'apliquen els criteris de selecció del nombre de variables (vegeu l'apartat 3.6.) apropiats per a la PCA, com són el del 95 % de la variància, el criteri d'Eastment i Krzanowski (*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, s'aconsella quedar-se amb el criteri que proporcioni el nombre més elevat de variables que cal retenir.

Treballarem amb 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.

PLS amb les dades sense pretractament

En el cas de la PLS apliquem les dades sense pretractament perquè en el gràfic 6.11. es veu que és el cas que proporciona millors resultats. Tots els mètodes de classificació proporcionen tantes sortides com classes tingui definides el problema. En aquest cas, $c = 4$ i, com que es classifiquen 400 objectes, hi haurà un total de $4 \times 400 = 1600$ sortides. Els resultats d'aplicar els tres criteris de selecció del nombre de variables sobre els 400 objectes de test es reflecteixen en la taula següent:

Criteri	Nombre de variables retingudes	Sortides errades de 1.600	Error total	PRESS
95 % variància	1	352	22,00 %	300,31
W(r)	10	0	0,00 %	2,06
$\lambda_i > 1$	18	0	0,00 %	1,81

Taula 6.3.

Els resultats de la taula anterior són sobre objectes de test, objectes no inclosos en el conjunt de calibratge del model; per tant, són resultats definitius. En 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 PLS cal agafar les divuit primeres variables i les dades sense pretractament, i s'obtindrà una taxa de sortides errònies del 0,00 %.

PCR amb les dades centrades

En el cas de la PCR apliquem les dades centrades perquè en el gràfic 6.12. es veu que és el cas que proporciona millors resultats. Els resultats d'aplicar els tres criteris de selecció del nombre de variables sobre els 400 objectes de test es reflecteixen en la taula següent:

Criteri	Nombre de variables retingudes	Sortides errades de 1.600	Error total	PRESS
95 % variància	1	211	13,19 %	201,84
W(r)	8	0	0,00 %	3,19
$\lambda_i > 1$	18	0	0,00 %	2,02

Taula 6.4.

En la taula anterior 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 PCR, cal agafar les divuit primeres variables i les dades centrades, i s'obtindrà una taxa de sortides errònies del 0,00 %.

Els dos gràfics següents (figures 6.15. i 6.16.) mostren els resultats de la classificació dels 400 objectes de test. Per poder-los interpretar cal tenir en compte que l'algorisme de classificació proporciona tantes sortides com classes d'objectes es tracten (en aquest cas quatre). Per tant, la resposta de l'algorisme a l'entrada d'un objecte de test seran quatre sortides, una per classe. Un objecte serà associat a una classe si la sortida corresponent a aquesta classe és superior o igual a 0,5 i no hi serà associat en cas contrari. Per tant, podria ser que en cas d'error un objecte fos associat a més d'una classe a la vegada o a cap classe. Els gràfics següents es troben dividits en sis apartats cadascun. Cada apartat conté separadament els cent segells de test de cada classe; així és fàcil de veure ràpidament el grau de fiabilitat de les respostes dels algorismes.

PLS

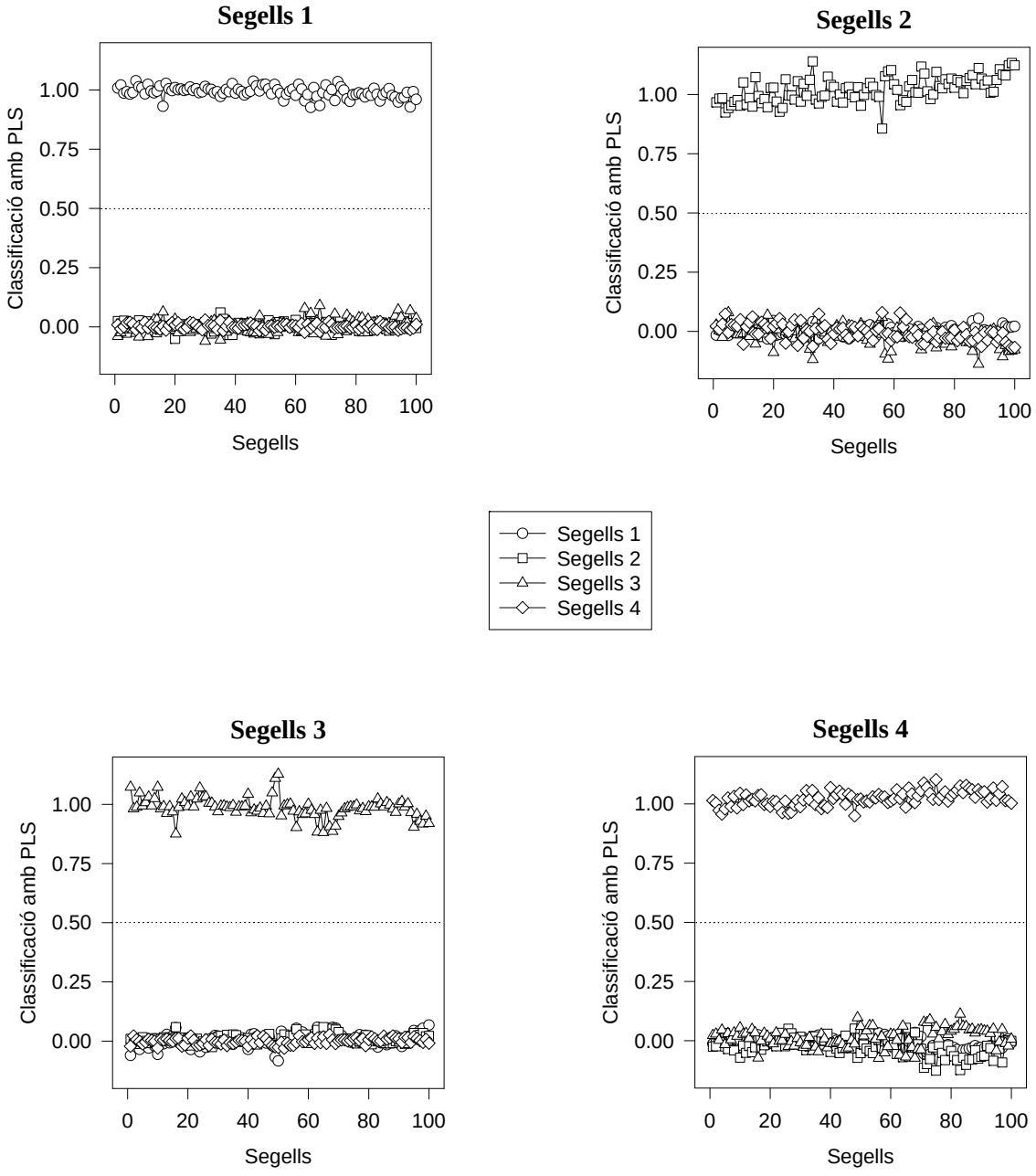


Figura 6.15.

PCR

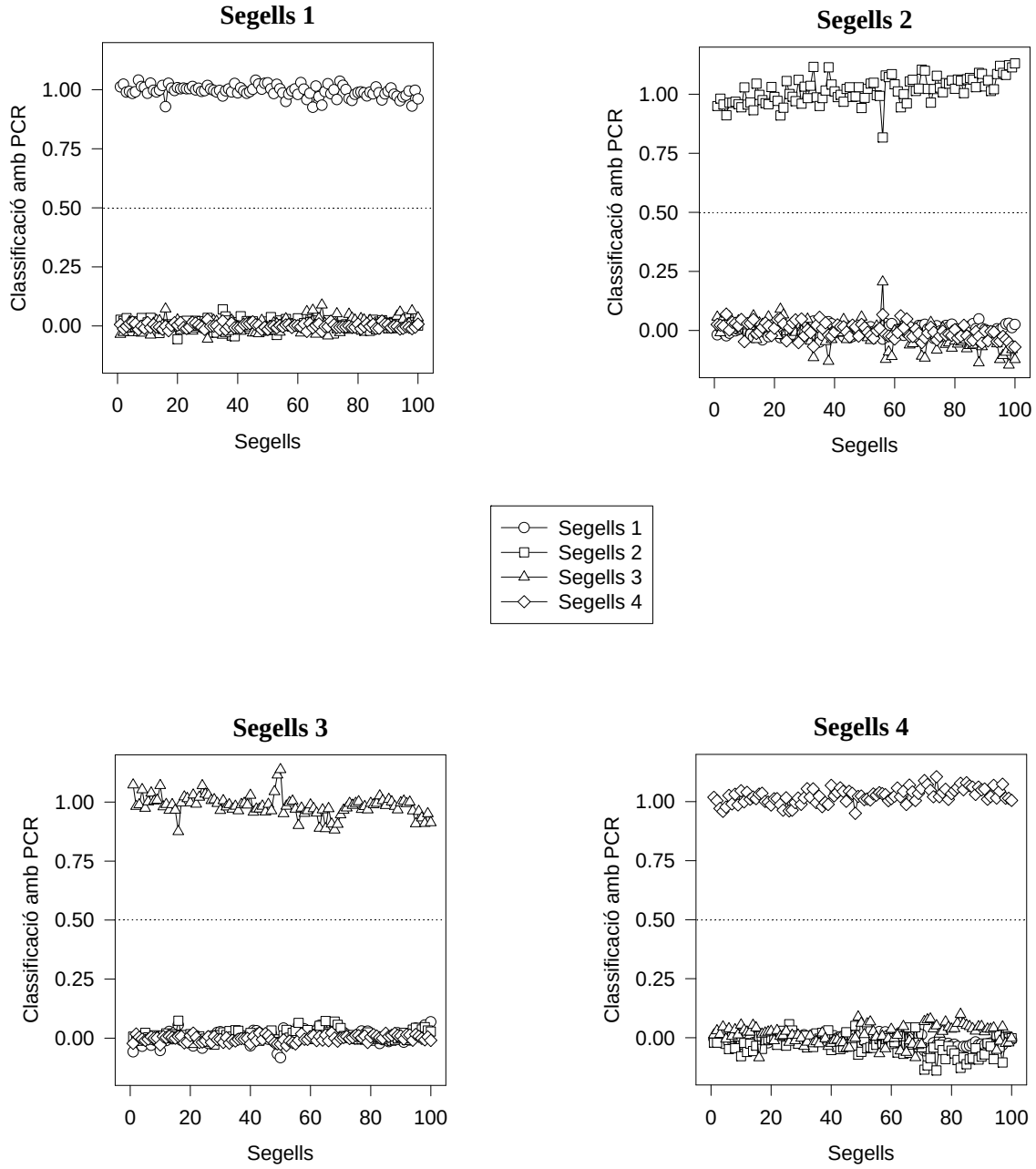


Figura 6.16.

6.6. Classificació amb anàlisi discriminant

Una de les tècniques de classificació més eficients que hi ha és l'anàlisi discriminant. En aquest apartat s'apliquen la LDA (suposa igual matriu de variàncies-covariàncies per a cada classe) i la QDA (suposa diferent matriu de variàncies-covariàncies per a cada classe). Primer s'han agafat les dades autoescalades, després s'ha aplicat la CVA i retenint les tres primeres variables canòniques, s'han aplicat la LDA i la QDA. La taula següent mostra els resultats d'aquestes dues tècniques sobre els 400 objectes del conjunt de test.

Criteri	Nombre de variables retingudes	Sortides errades de 1.600	Error total	PRESS
LDA	3 de la *CVA*	0	0,00 %	0,00
QDA	3 de la *CVA*	0	0,00 %	0,00

Taula 6.5.

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, on 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 llindar o *threshold*. El valor d'aquest és força crític per l'èxit dels algorismes. En aquest treball es proposa agafar els llindars següents:

$$T_{LDA,i} = [m_i + M_i]/2 \qquad T_{QDA,i} = \sqrt{m_i.M_i}\ , \tag{6.4}$$

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 figures 6.17. i 6.18. mostren els resultats d'efectuar la LDA i la QDA sobre els 400 objectes de test.

LDA

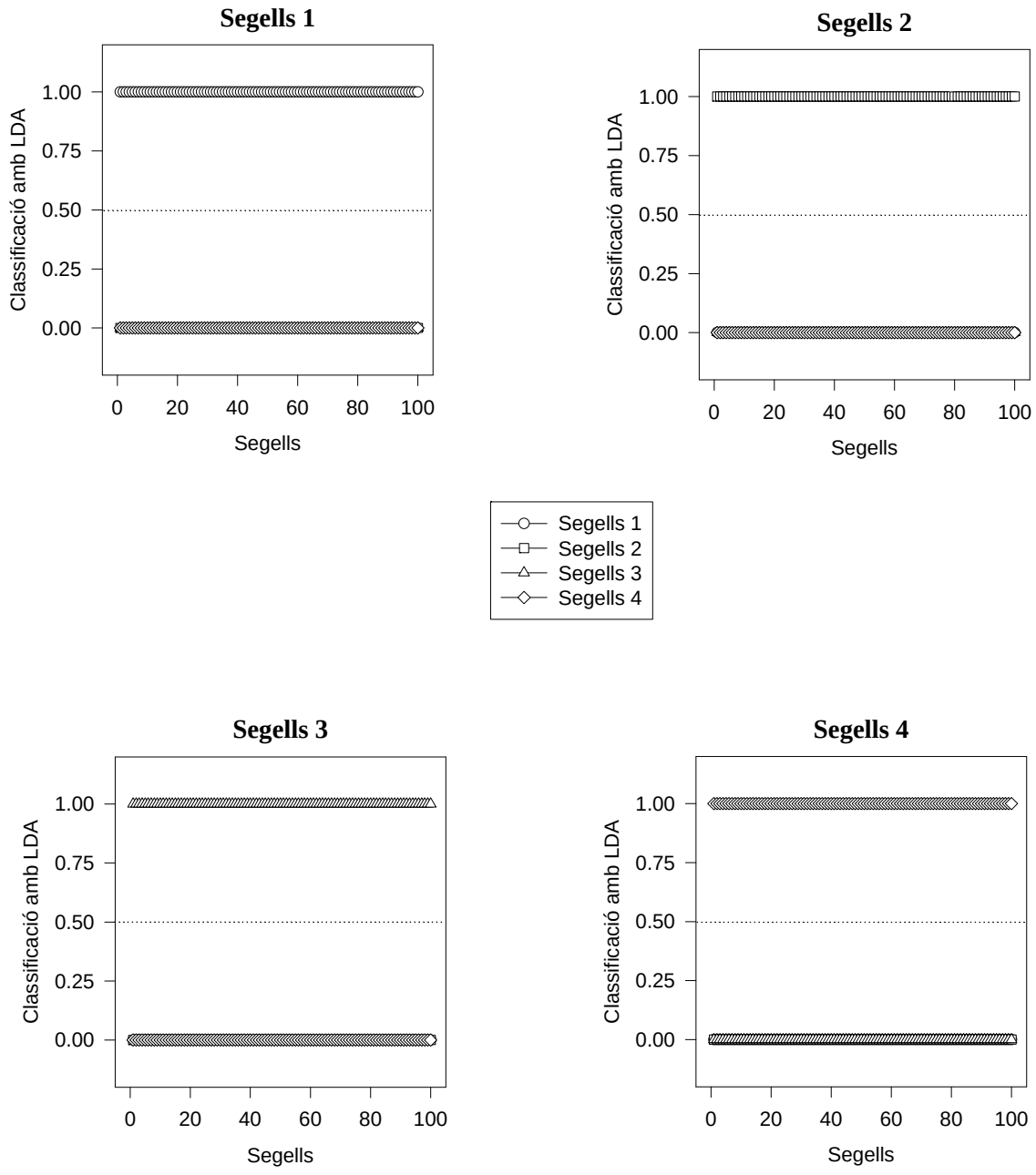


Figura 6.17.

QDA

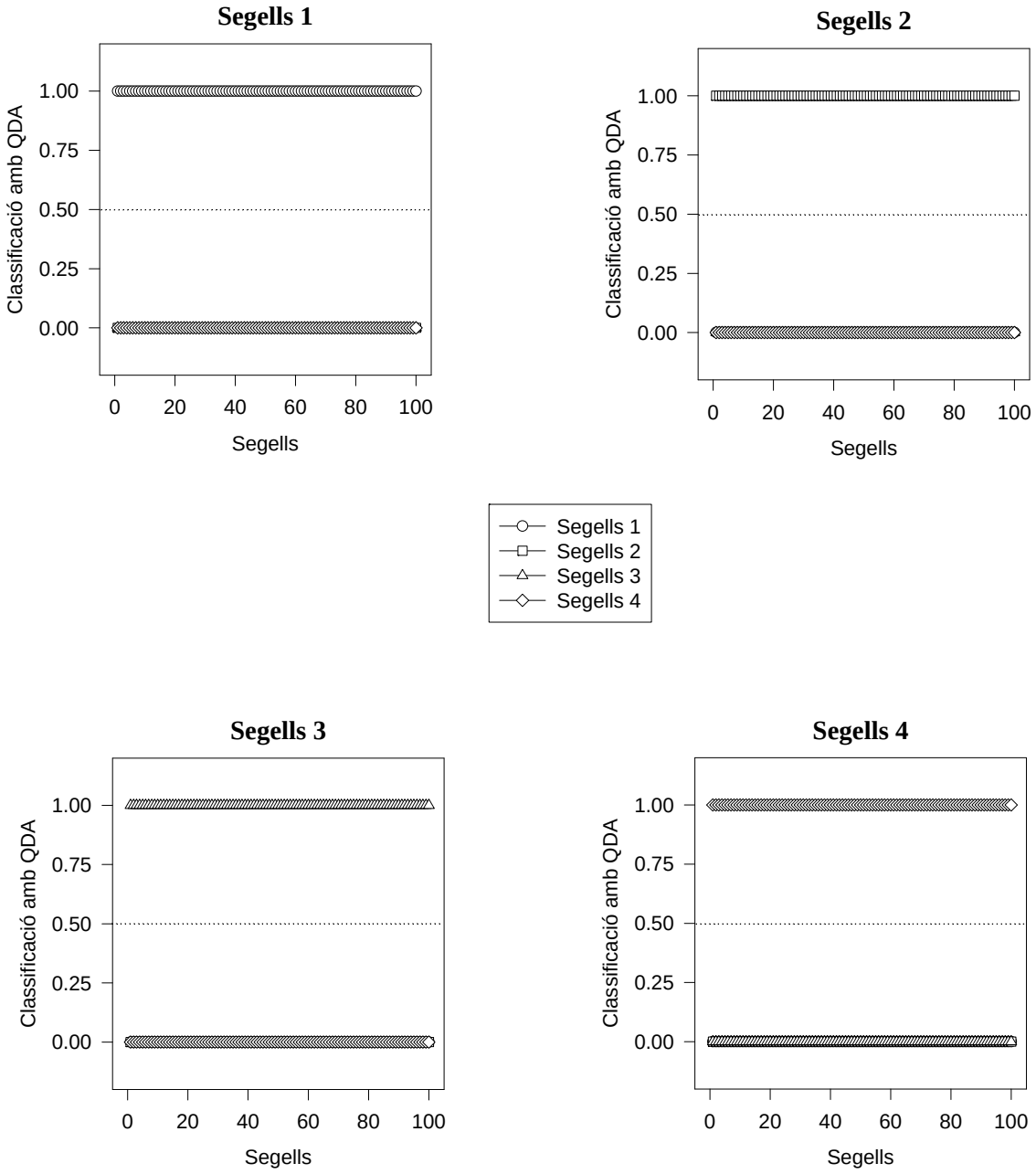


Figura 6.18.

6.7. Classificació amb SIMCA i DASCO

Tant SIMCA com DASCO són tècniques de classificació força utilitzades. Treballarem amb les tres variables canòniques resultants d'efectuar la CVA. La taula següent mostra els resultats de classificació dels 200 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 800	Error total	PRESS
SIMCA	Cap	1	0,13 %	1,00
SIMCA	Centrat	2	0,25 %	2,00
SIMCA	Autoescalat	1	0,13 %	1,00
DASCO	Cap	2	0,25 %	2,00
DASCO	Centrat	0	0,00 %	0,00
DASCO	Autoescalat	1	0,13 %	1,00

Taula 6.6.

Els resultats anteriors indiquen que, en el cas de SIMCA, cal tenir en compte les dades sense cap pretractament o autoescalades i, el cas de DASCO, les dades s'han de centrar.

La taula següent reflecteix el nombre de variables que automàticament retenen per classe els algorismes SIMCA i DASCO, depenent del tipus de pretractament de les dades:

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

Taula 6.7.

La taula següent mostra els resultats de classificació d'aquestes dues tècniques sobre els 400 objectes del conjunt de test.

Criteri	Sortides errades de 1.600	Error total	PRESS
SIMCA dades originals	9	0,56 %	9,00
DASCO dades centrades	2	0,13 %	2,00

Taula 6.8.

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, on 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 llindar o *threshold*. El valor d'aquest és força crític per l'èxit dels algorismes. El llindar proposat en aquest treball és:

$$T_i = \sqrt{m_i.M_i}\ , \tag{6.5}$$

on:

m_i: 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.

M_i: 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 6.19. i 6.20. mostren els resultats d'efectuar SIMCA i DASCO sobre els 400 objectes de test.

SIMCA

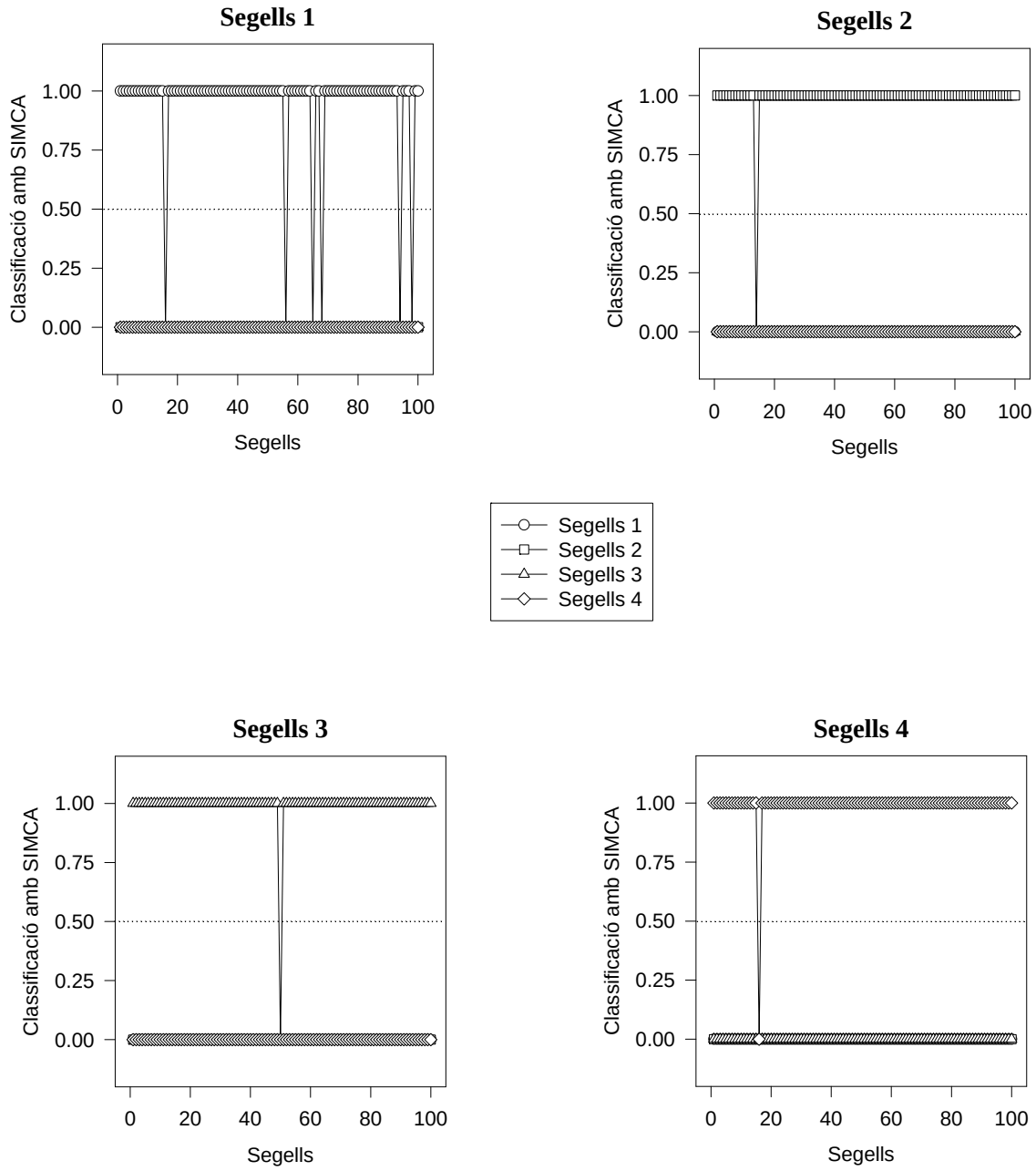


Figura 6.19.

DASCO

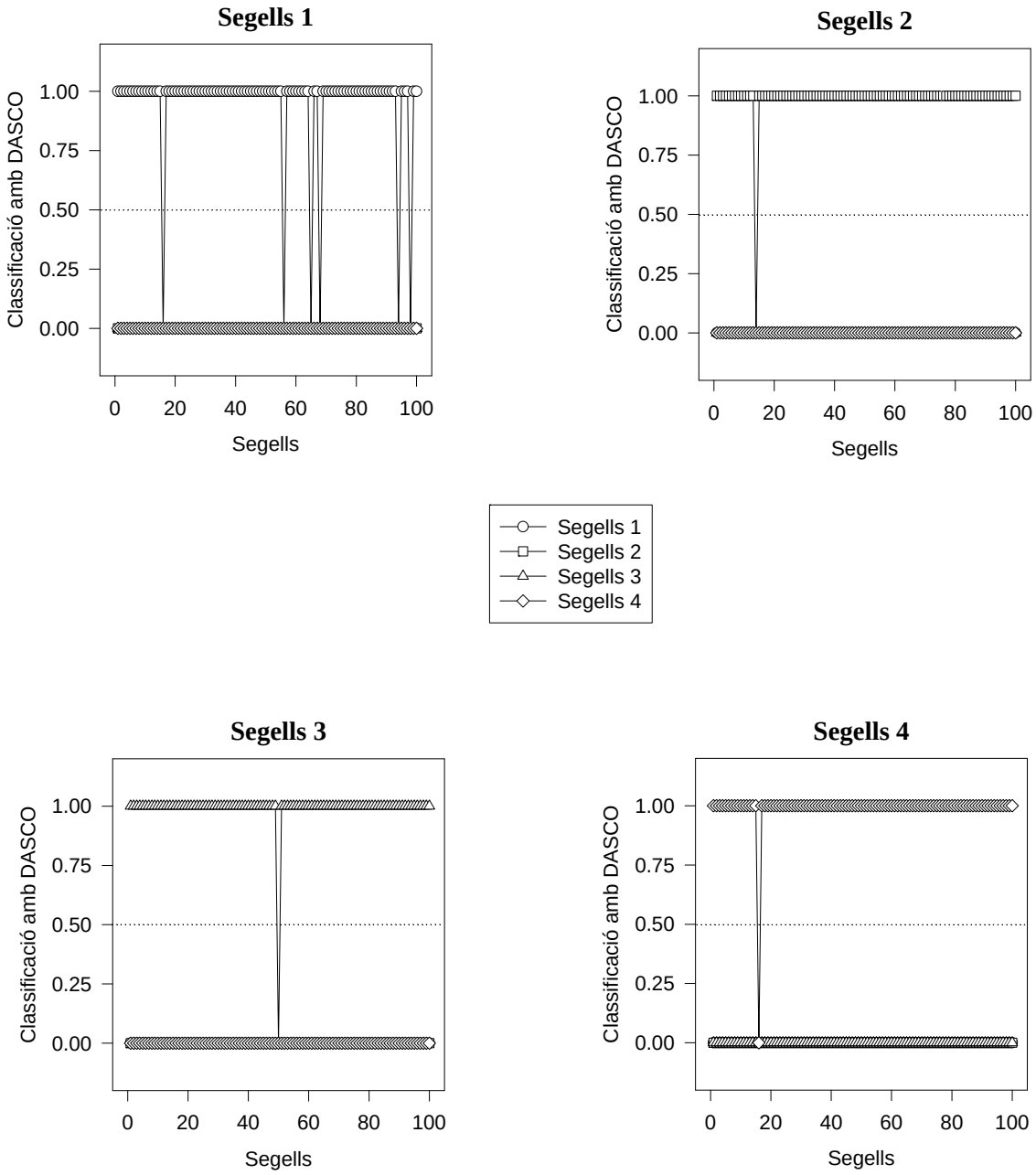


Figura 6.20.

6.8. 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 autoescalades, després s'ha aplicat la CVA i, retenint les tres primeres variables canòniques, s'ha aplicat aquesta tècnica. La taula següent mostra els resultats de classificació produïts per la *k*NN sobre els 200 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	3 de la CVA	0	0,00 %	0,00
3	3 de la CVA	0	0,00 %	0,00
4	3 de la CVA	0	0,00 %	0,00
5	3 de la CVA	0	0,00 %	0,00
6	3 de la CVA	0	0,00 %	0,00
7	3 de la CVA	0	0,00 %	0,00
8	3 de la CVA	0	0,00 %	0,00

Taula 6.9.

Dels resultats anteriors es dedueix que en aquest cas qualsevol valor de *k* comprès entre dos i vuit hauria de conduir a un bon model de classificació. En aquest cas, classificarem els 400 objectes del conjunt de test considerant tres i quatre veïns, resultats que mostra la taula següent:

Nombre de veïns	Nombre de variables retingudes	Sortides errades de 1.600	Error total	PRESS
3	3 de la CVA	0	0,00 %	0,00
4	3 de la CVA	0	0,00 %	0,00

Taula 6.10.

La figura següent mostra els resultats de classificació dels 400 objectes de test aplicant la *k*NN amb 3 i 4 veïns.

*k*NN (*k* = 3 o *k* = 4)

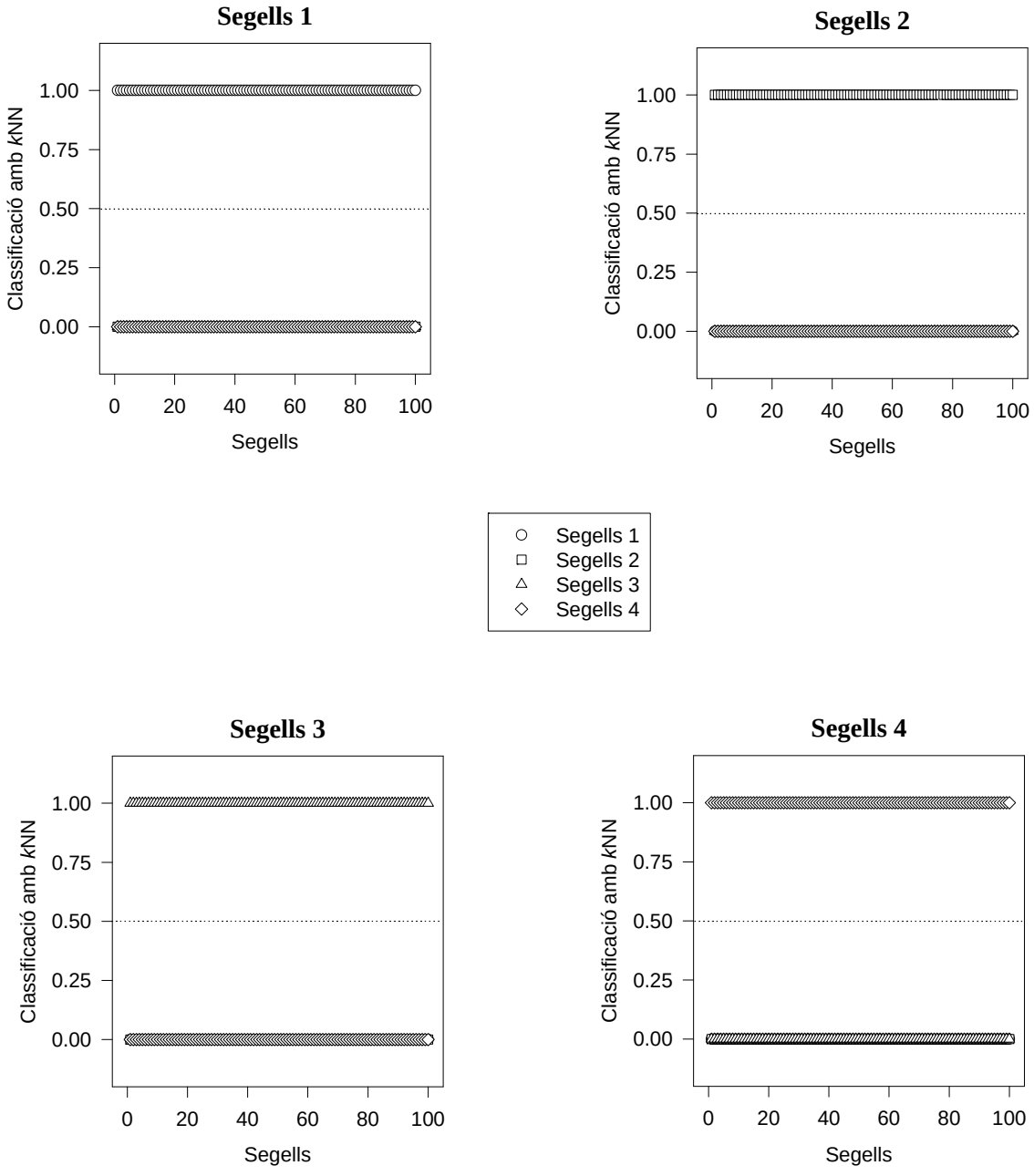


Figura 6.21.

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

Aquesta tècnica ha estat elaborada en 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 autoescalades i després s'ha aplicat la CVA directament a la matriu. S'ha fet un programa que variï iterativament el nombre de variables retingudes entre una i tres, així com els valors del factor *k*, que modula la velocitat de decreixement de les funcions exponencials decreixents de la fórmula 6.6. de 0,10 a 3,00 en increments successius de 0,05 (vegeu l'apartat 5.2.5.).

$$g_{1ij} = \exp\left(-k.\left|\frac{x - min_{ij}}{max_{ij} - min_{ij}}\right|\right),\ g_{2ij} = \exp\left(-k.\left|\frac{x - max_{ij}}{max_{ij} - min_{ij}}\right|\right). \qquad (6.6)$$

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

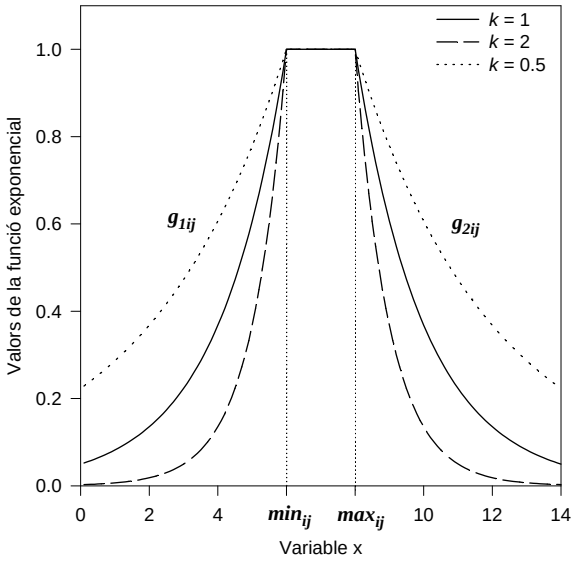


Figura 6.22.

S'ha dividit el conjunt de calibratge en dos subconjunts: un de 160 objectes i un altre de 40 objectes. El primer serveix per generar múltiples models de classificació i el segon d'aquests serveix per determinar quin és el que proporciona millors resultats.

La taula següent mostra els resultats de classificació dels 40 objectes (separats del conjunt de calibratge) segons diferents models de classificació generats a partir dels 160 objectes de calibratge:

Nombre de variables retingudes	Factor exponencial	Sortides errades de 160	Error total	PRESS ($x10^{-3}$)
3 de la CVA	0,25	0	0,00 %	5,557
3 de la CVA	0,30	0	0,00 %	1,452
3 de la CVA	0,35	0	0,00 %	0,819
3 de la CVA	0,40	0	0,00 %	0,854
3 de la CVA	0,45	0	0,00 %	1,037

Taula 6.11.

La taula 6.12 mostra els resultats de classificació dels 400 objectes del conjunt de test. S'ha escollit un valor del factor de decreixement de 0,35 i s'han retingut les tres primeres variables canòniques.

Nombre de variables retingudes	Factor exponencial	Sortides errades de 1.600	Error total	PRESS
3 de la *CVA*	0,35	1	0,06 %	2,15

Taula 6.12.

La figura següent mostra els resultats de classificació dels 400 objectes del conjunt de test.

Paral·lelepípede

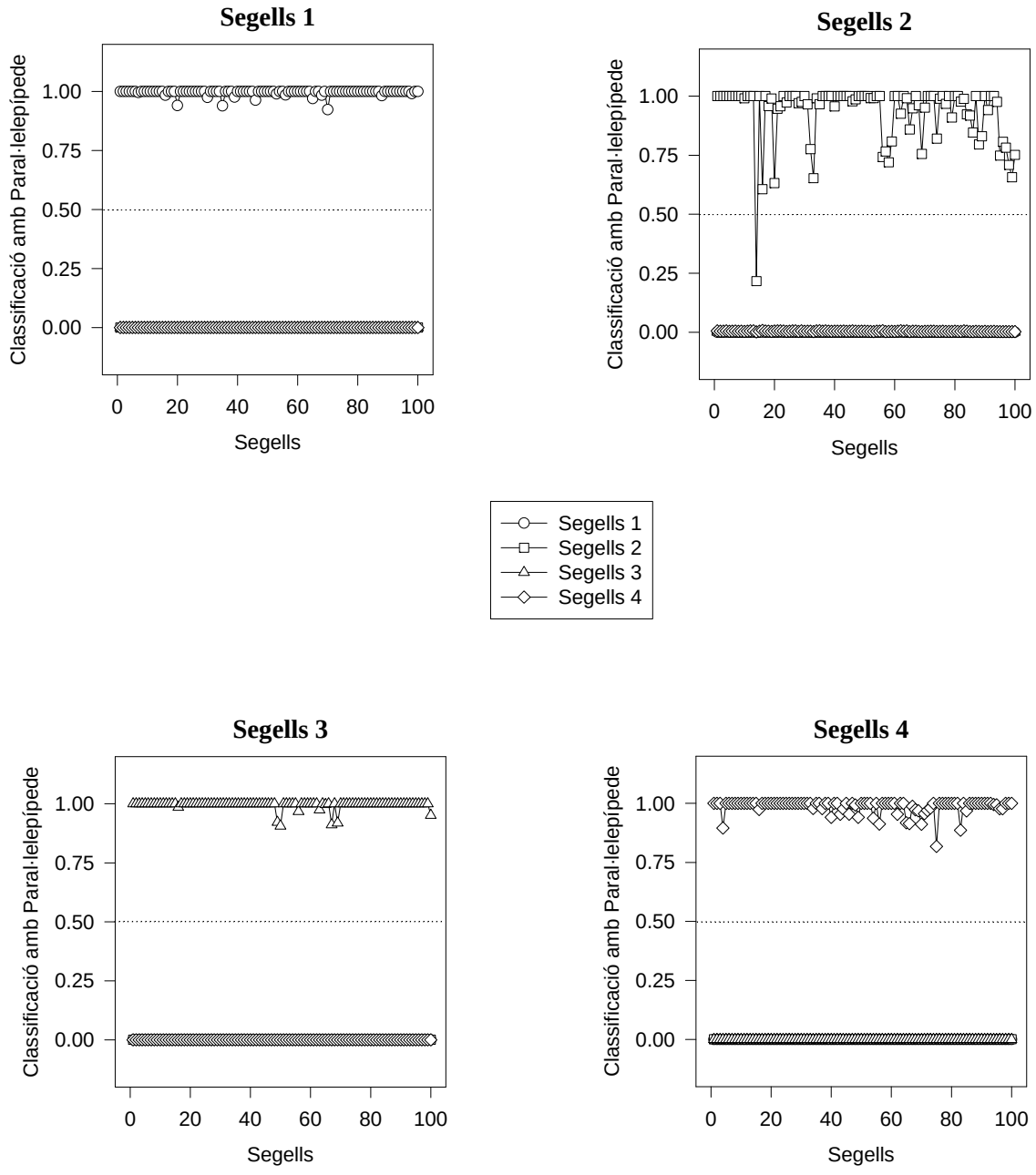


Figura 6.23.

6.10. Classificació amb xarxes neuronals

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

Primer s'han agafat les dades autoescalades, després s'ha aplicat la CVA i, a la matriu resultant, li ha estat aplicada directament aquesta tècnica (retenint les tres variables canòniques).

En el cas de les xarxes neuronals tipus Backpropagation, s'ha creat un model de xarxa amb una única capa oculta de deu neurones amb funció de transferència *logsig* i una capa de sortida formada per quatre neurones amb funció de transferència lineal. En aquest cas, amb deu neurones s'ha pogut entrenar satisfactòriament la xarxa i no ha calgut utilitzar 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 de 4e-5.

La taula següent mostra els resultats de classificació dels 400 objectes de test proporcionats per la xarxa Backpropagation:

Criteri	Nombre de variables retingudes	Sortides errades de 1.600	Error total	PRESS
Backprop.	3 de la CVA	0	0,00 %	2,17e-4

Taula 6.13.

En el cas de les xarxes neuronals Radial Basis Functions, en la fase de 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 160 i 40 objectes cadascun. El primer d'aquests (160 objectes) ha servit per entrenar les diferents xarxes i el segon (40 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 0,50 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 predicció dels 40 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 160	Error total	PRESS
0,50	3 de la CVA	60	0	0,00 %	6,28e-5
0,50	3 de la CVA	80	0	0,00 %	8,97e-6
0,50	3 de la CVA	100	0	0,00 %	6,18e-6
0,50	3 de la CVA	120	0	0,00 %	7,55e-6
0,50	3 de la CVA	140	0	0,00 %	1,47e-6
0,50	3 de la CVA	160	0	0,00 %	8,35e-7

Taula 6.14.

De tots els models anteriors ha estat seleccionat el de 160 neurones, ja que és el que presenta el PRESS més baix. Per tant, amb aquest mateix model s'ha efectuat la classificació dels 400 objectes de test. Els resultats obtinguts es mostren en la taula següent:

Nombre de variables retingudes	Nombre de neurones	Sortides errades de 1.800	Error total	*PRESS*
3 de la CVA	160	0	0,00 %	1,43e-3

Taula 6.15.

Les figures 6.24. i 6.25. mostren els resultats de classificació proporcionats per les xarxes tipus Backpropagation i Radial Basis sobre els 400 objectes de test:

Backpropagation

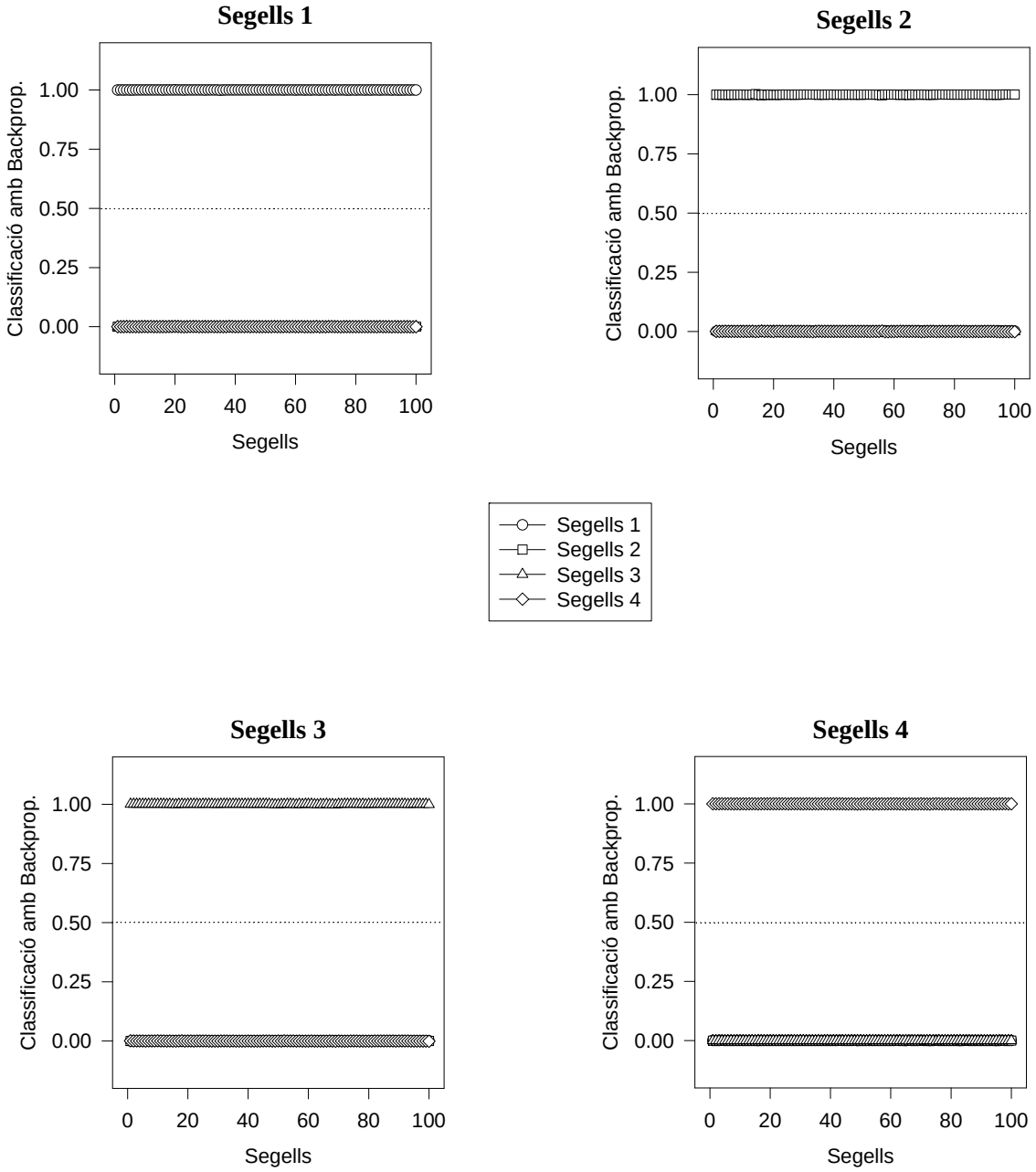


Figura 6.24.

Radial Basis Functions

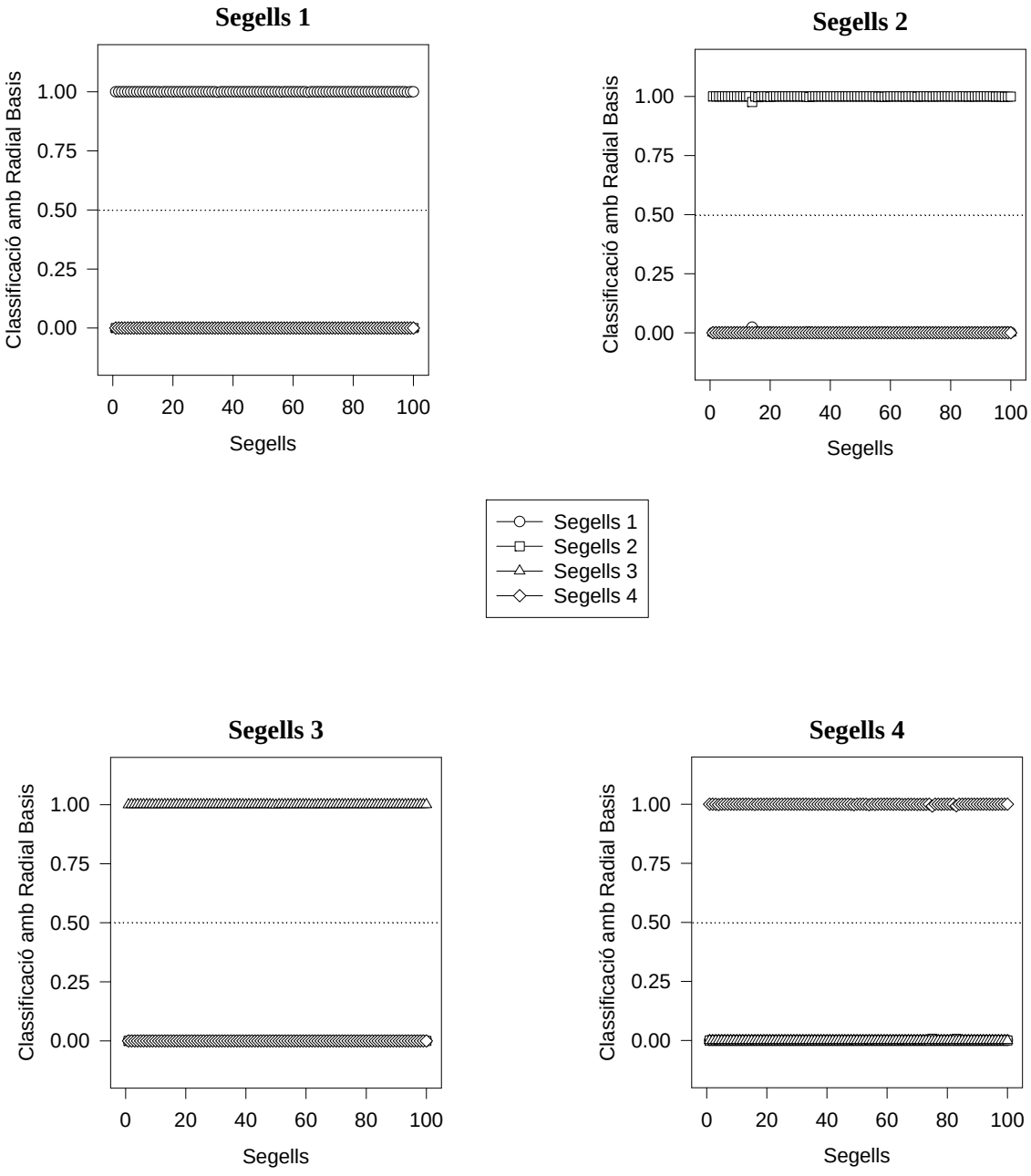


Figura 6.25.

6.11. 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 tres conjunts difusos (bo, regular i dolent) –tot i que se'n podrien definir més– de forma triangular per a cada variable de cadascuna 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 cadascuna de les diferents classes:

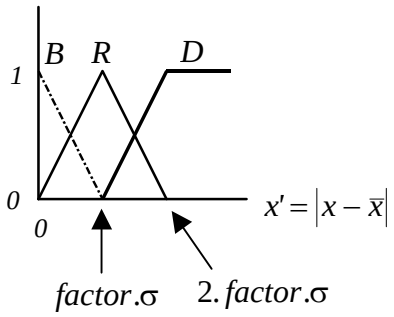


Figura 6.26.

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 3. S'ha dividit el conjunt de calibratge en dos subconjunts: un de 160 objectes i l'altre de 40 objectes. El primer serveix per generar múltiples models de classificació i el segon serveix per determinar quin d'aquests models és el que proporciona millors resultats.

La taula següent mostra els resultats de classificació dels 40 objectes (separats del conjunt de calibratge) segons els diferents models de classificació generats a partir dels 160 objectes de calibratge:

Nombre de variables retingudes	Factor	Sortides errades de 160	Error total	PRESS
1 de la CVA	3,30	0	0,00 %	1,0320
1 de la CVA	3,35	0	0,00 %	1,0184
1 de la CVA	3,40	0	0,00 %	1,0083
1 de la CVA	3,45	0	0,00 %	1,0018
1 de la CVA	3,50	0	0,00 %	0,9986
1 de la CVA	3,55	0	0,00 %	0,9988
1 de la CVA	3,60	0	0,00 %	1,0020
1 de la CVA	3,65	0	0,00 %	1,0080
1 de la CVA	3,70	0	0,00 %	1,1067

Taula 6.16.

Els resultats de la taula anterior suggereixen que el millor model és el corresponent a retenir només la primera variable canònica i agafar un factor de 3,50 (és el que no presenta cap sortida defectuosa i al mateix temps proporciona un valor del PRESS més baix).

Els resultats de classificació dels 400 objectes de test (agafant els 200 objectes de calibratge, la primera variable i un factor = 3,50) estan expressats en la taula següent:

Nombre de variables retingudes	Factor	Sortides errades de 1.600	Error total	PRESS
1 de la CVA	3,50	4	0,25 %	14,28

Taula 6.17.

Els resultats de classificació dels 400 objectes del conjunt de test es mostren en la figura següent.

Lògica difusa

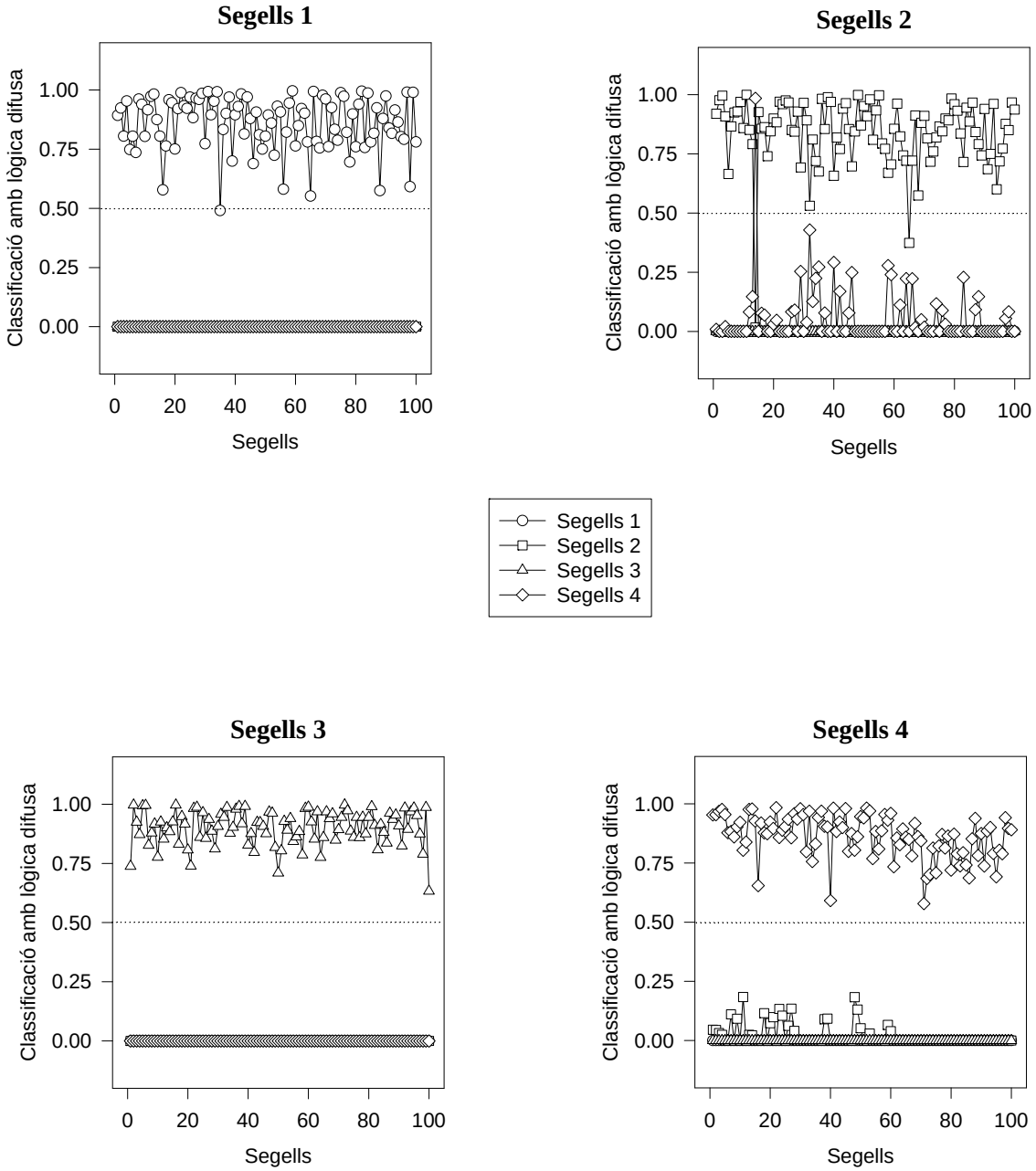


Figura 6.27.

6.12. Taula comparativa dels diferents mètodes

La taula següent mostra un resum dels resultats de classificació dels 400 objectes del conjunt de test obtinguts aplicant tots els mètodes utilitzats en aquest capítol:

Tècnica	Nombre de variables retingudes	Sortides errades de 1.600	Error total	PRESS
PCR	18	0	0,00 %	2,02
PLS	18	0	0,00 %	1,81
LDA	3 de la CVA	0	0,00 %	0,00
QDA	3 de la CVA	0	0,00 %	0,00
SIMCA	1 per classe	9	0,56 %	9,00
DASCO	1 per classe	2	0,13 %	2,00
*k*NN (*k* = 4)	3 de la CVA	0	0,00 %	0,00
Paral·lelepípede	3 de la CVA	1	0,06 %	2,15
Backpropagation	3 de la CVA	0	0,00 %	2,17e-4
Radial Basis	3 de la CVA	0	0,00 %	1,43e-3
Lògica difusa	1 de la CVA	4	0,25 %	14,28

Taula 6.18.

La taula anterior indica que, per a aquest problema concret, quasi totes les tècniques proporcionen resultats molt propers al 100 % d'èxit de classificació. Les que funcionen pitjor, per ordre, són: SIMCA, la lògica difusa i DASCO.