UNIVERSITAT DE BARCELONA

Departament de Física Aplicada i Òptica

Programa de Micro i Optoelectrònica Física

Bienni 1994-96

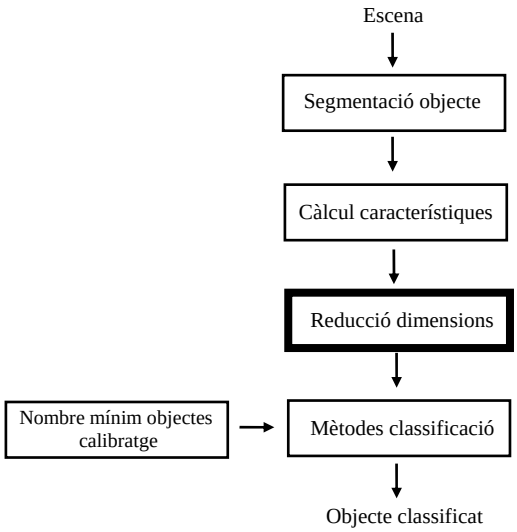
DISSENY D'UN PROTOCOL NUMÈRIC PER A LA CLASSIFICACIÓ INVARIANT D'IMATGES APLICANT TÈCNIQUES MULTIVARIANTS

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



3. Reducció de dimensions

En el capítol 2 s'han explicat diferents algorismes de càlcul de característiques discriminants. En un procés qualsevol de classificació d'imatges digitalitzades, el nombre de característiques que s'acostuma a calcular pot ser relativament elevat (en general inferior a 100, però). Suposem que *m* és el nombre de característiques utilitzades en un determinat problema de classificació. Moltes vegades, a causa d'una banda, que les característiques discriminants no són independents entre si i, per tant, aporten informació redundant, i de l'altra, que part de la seva informació conté soroll i, per tant, aporten confusió al problema, és convenient transformar adequadament el conjunt de *m* característiques discriminants en un conjunt m^* (amb $m^* < m$) més reduït de variables independents que ens permeti afrontar el problema amb més garanties d'èxit i amb un estalvi important de temps de càlcul en les etapes posteriors. Aquest procés s'anomena *reducció de dimensions (feature extraction).* A partir d'ara, quan parlem de variables farem referència a les magnituds resultants de transformar les característiques discriminants. Aquestes variables seran, en general, una combinació lineal de les característiques i, en el cas ideal, les variables seran linealment independents i, per tant, no contindran la informació redundant i tampoc el soroll que aporten les característiques. Les variables, una a una, aporten més informació que les característiques discriminants.

Resumint, quan es parla de reduir les dimensions d'un problema es fa referència a trobar, a partir de les *m* característiques discriminants, un nombre inferior *m** de variables que continguin el màxim possible d'informació sobre el problema. Per tant, el procés de reducció de dimensions té dues etapes. La primera consisteix a transformar les característiques en variables (matemàticament es pot interpretar com un canvi de variables). La segona etapa es basa a determinar el nombre *m** de variables que s'han de *retenir*, és a dir, el subconjunt *m** del conjunt total de variables que s'utilitzaran per caracteritzar tots els objectes. Aquesta segona etapa és necessària perquè els algorismes de reducció de dimensions generalment ordenen les variables, concentrant la major part de la informació en les primeres i deixant pel final les restants, que contindran molt soroll. Per tant, és necessari rebutjar part de les variables i quedar-se només amb les *m** primeres.

En els apartats 3..1 al 3.4. es detallen quatre algorismes diferents de reducció de dimensions i en l'apartat 3.6. s'expliquen diferents mètodes de selecció del nombre *m** de variables que cal retenir.

La figura 3.1. explica gràficament la transformació de les característiques en un nombre més reduït de variables:

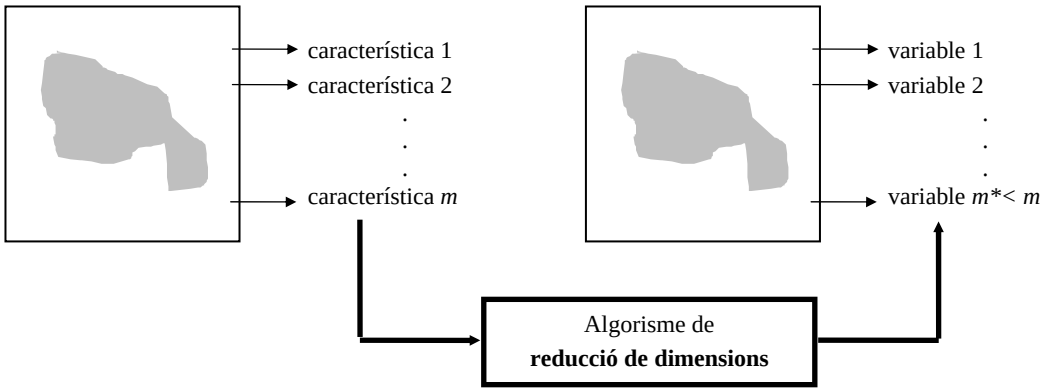


Figura 3.1.

En aquest treball no es treballa amb un altre conjunt de tècniques ben estudiades, anomenades s*elecció de característiques (feature selection)*, que consisteixen a realitzar una selecció de les millors característiques i a rebutjar les restants. Aquestes tècniques s'utilitzen, sobretot, quan es tenen moltes característiques. Vegeu [Jai97] i [Kud00].

Els algorismes de reducció de dimensions proporcionen informació de les característiques discriminants que tenen més pes en les variables. Cal tenir present que les variables són

combinacions lineals de les característiques. Per tant, les característiques que són poc discriminants tindran poc pes en les variables. Es veu, doncs, que una reducció de dimensions ja comporta, indirectament, una selecció de característiques. D'altra banda, si es vol, es pot fer una selecció de característiques eliminant les que tinguin pesos molt baixos. A més, el procés de reducció de dimensions és, en general, molt ràpid. És per aquests motius que s'ha preferit una reducció de dimensions a una selecció de característiques.

3.1 Anàlisi de components principals (Principal Components Analysis, PCA)

La PCA és una tècnica multivariable, la finalitat de la qual és simplificar l'estructura de les dades del problema. Aquest algorisme permet passar d'un problema de m característiques a un problema de m^* variables (amb $m^* \leq m$), anomenades *variables* o *components principals* (*PC*). Aquest fet, a més de reduir la dimensionalitat del problema (nombre de variables que cal considerar), permet reduir el cost computacional de les etapes posteriors. Aquesta tècnica també s'anomena *Karhunen-Loève Expansion*.

La PCA és una tècnica àmpliament utilitzada en diverses àrees del coneixement, com ara la medicina, la psicologia, la química, l'economia, etc. Consulteu [Fuk72], [Krz79], [Cua81], [Eas82], [Seb84], [Bee87], [Mur87], [Dun89], [Kow91], [Joh92], [Bas94] i [Esb94].

La PCA és l'algorisme que transforma les característiques en components principals. Aquests són combinacions lineals de les característiques, són ortogonals entre si i tenen variància màxima. Hi ha tants PC com característiques (m) i estan definits de manera que el primer d'aquests (PC_1) explica la màxima variància, el segon (PC_2) explica la màxima variància residual i així successivament (fins al PC_m).

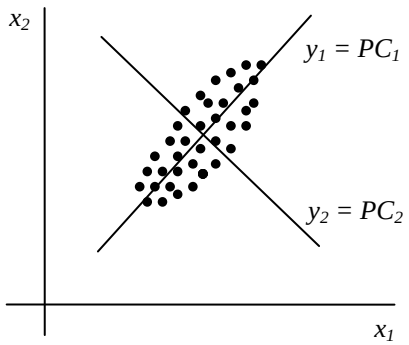


Figura 3.2.

La figura 3.2. mostra un conjunt de n objectes als quals s'han calculat $m = 2$ característiques, x_1 i x_2. Si es representen els n objectes en l'espai de les característiques (els eixos horitzontal i vertical seran, respectivament, les característiques x_1 i x_2), s'obté un núvol de punts. El PC_1 és l'eix que explica la màxima variància del núvol de punts i per tant, està dirigit en la direcció de màxima dispersió, mentre que el PC_2 és perpendicular a aquest i explica la màxima variància residual.

Si es disposa d'un conjunt de n objectes a cada un dels quals s'han calculat m característiques discriminants, tota aquesta informació es pot concentrar en la matriu de característiques $X = \{x_{ij}\}$, amb $i = 1,2...,n$ i $j = 1,2...,m$. Les m característiques defineixen un espai $\Re^m$. En aquest espai es poden representar els n punts corresponents als n objectes. La PCA busca el conjunt de m eixos ortogonals y_i (PC) que millor s'ajusti al conjunt de n punts d'aquest espai, per tal de reemplaçar els m eixos originals x_i d'aquest.

Potencialment es pot calcular un màxim de m PC (tants com característiques té el problema). La direcció del PC i-èssim respecte a l'espai vectorial X definit per les característiques ve donada pel vector columna v_i (els v_i són vectors propis de la matriu $X^T.X$). Els PC són ortonormals (compleixen: $v_i^T.v_j = 0, \forall i \neq j$ i $v_i^T.v_i = 1$) i incorrelacionats (compleixen: $v_i^T.S.v_j = 0, \forall i \neq j$, sent S la matriu global de covariància) i les seves variàncies són iguals als valors propis ($s_{v_i}^2 = \lambda_i$) corresponents als vectors propis v_i. Com que els PC són ortogonals, es poden considerar el resultat d'una rotació més una translació dels eixos originals, per tal d'alinear-los amb els eixos naturals del núvol de punts (en l'espai de les característiques, cada punt representa un objecte).

Un dels objectius bàsics de la PCA és concentrar la màxima variància (informació) en els primers components principals. D'aquesta manera, els últims components només contindran soroll. Quan es tenen en compte només els primers PC, d'una banda s'elimina el soroll i de l'altra se simplifica el problema, ja que s'aconsegueix reduir el nombre de variables i, per tant, les dimensions.

El primer component principal o PC_1 és l'eix que millor s'ajusta al núvol de punts i compleix la condició que la suma de distàncies de tots els punts (objectes) a ell és mínima. També es pot interpretar que la suma dels quadrats de les projeccions dels punts en aquest eix és màxima (és equivalent a maximitzar la variància dels punts quan es projecten en aquest eix).

La figura 3.3. explica les afirmacions anteriors:

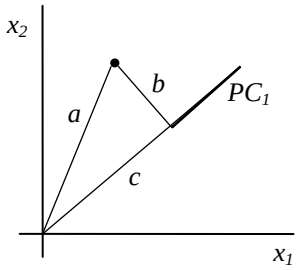


Figura 3.3.

a: posició del punt en l'espai $\Re^m$ (aquesta posició és un valor constant)

b: distància del punt al nou eix

c: projecció del punt sobre el nou eix

Es compleix: $a^2 = b^2 + c^2$ (3.1)

Com que *a* és constant, minimitzar *b* és equivalent a maximitzar *c*.

S'anomena v_i el vector director de l'*i*-èssim PC. El quadrat de la projecció de la matriu *X* de característiques en el nou eix es calcula com:

$$(X.v)^T.(X.v) = v^T.X^T.X.v = v^T.S.v \quad (3.2)$$

$$\text{on} \quad S = X^T.X \quad (3.3)$$

S'obté la forma quadràtica $v^T.S.v$. Perquè els valors d'aquesta no siguin molt grans ni molt petits, s'acostuma a imposar que el vector *v* sigui unitari:

$$v^T.v = 1 \quad (3.4)$$

Si es compleix la condició anterior, també es complirà la següent:

$$v^T.S.v = v^T.S.v - \lambda.(v^T.v - 1) \quad (3.5)$$

Perquè l'expressió anterior sigui màxima, es deriva i s'obté:

$$2.S.v - 2.\lambda.v = 0 \Rightarrow S.v = \lambda.v \quad (3.7)$$

Resulta com a solució: $\begin{cases} v_1 : \text{vector propi de } S = X^T.X \\ \lambda_1 : \text{valor propi associat a } v_1 \end{cases}$

Per tant, v_1 és el primer eix principal i λ_1, el primer valor propi. El valor propi és un factor de mèrit de l'eix, ja que dóna la variància total explicada per aquest.

El segon PC ha de ser ortogonal al primer i també, suposem, el seu vector director unitari:

$$v^T.S.v - \lambda.(v^T.v - 1) - \mu.(v^T.v_1), \quad (3.8)$$

on derivant s'obté:

$$2.S.v - 2.\lambda.v - \mu.v_1 = 0 \tag{3.9}$$

Multiplicant l'expressió anterior per v_1^T resulta que $\mu = 0$, de manera que:

$$2.S.v - 2.\lambda.v = 0 \Rightarrow S.v = \lambda.v. \tag{3.10}$$

Resulta com a solució: $\begin{cases} v_2 : \text{vector propi de } S=X^T.X \\ \lambda_2 : \text{valor propi associat a } v_2 \end{cases}$

Aquest procediment es pot repetir successivament fins calcular els *m* vectors i valors propis. Els valors propis que s'obtenen decreixen en valor: $\lambda_1 > \lambda_2 > ... > \lambda_m$. Els valors propis λ_i representen la variància de l'eix principal v_i.

La projecció dels valors de *X* sobre l'*i*-èssim vector propi s'anomena *score* ($y_i = X.v_i$, és un vector columna). Els *scores* són els valors de les variables i es faran servir en l'etapa de classificació. Es poden calcular tants *scores* com característiques tingui el problema. L'agrupació de tots els vectors columna y_i en una matriu forma la matriu de *scores Y*. Si agafem els *m* vectors columna v_i es forma la matriu *V*, complint-se:

$$Y_{(n,m)} = X_{(n,m)}.V_{(m,m)} \tag{3.11}$$

És molt útil representar gràficament els *scores* més representatius dels *n* objectes. En qualsevol problema de classificació és molt desitjable que els objectes pertanyents a la mateixa classe quedin agrupats i que les diferents classes quedin separades entre si (problema linealment separable).

Un mètode molt comú de calcular la matriu de vectors propis *V* consisteix a fer la descomposició en valors singulars (*svd*) de la matriu *X:*

$$svd(X) = U_{(n,m)}.\Sigma_{(m,m)}.V^T_{(m,m)} \tag{3.12}$$

Les matrius anteriors compleixen:

U: matriu ortogonal ($U^T.U = I_m$). Els seus vectors columna són vectors propis de $X.X^T$.

V: matriu ortogonal ($V^T.V = I_m$). Els seus vectors columna són vectors propis de $X^T.X$ (les seves columnes s'anomenen *loadings*).

Σ: matriu diagonal. La seva diagonal conté els valors singulars $s_j = \sqrt{\lambda_j}$, ordenats de major a menor.

La descomposició en valors singulars és aplicable a tota matriu $X_{(n,m)}$ amb $n \geq m$. En cas que $n < m$, es té que $s_j = 0$, per $j = n+1...,m$. Les corresponents columnes de U també seran nul·les.

L'algorisme i el programa en codi C per realitzar la descomposició en valors singulars d'una matriu es troben detallats en [Pre92].

Algorisme PCA

L'algorisme PCA és el següent:

1.- S'agafa la matriu base $X_{(n,m)}$.

2.- La matriu anterior se centra o s'autoescala, de manera que s'obté la matriu transformada $X^*_{(n,m)}$.

3.- Es realitza la descomposició en valors singulars de la matriu X^*:

$$svd(X^*) = U.\Sigma.V^T$$

4.- Es calculen les noves coordenades dels punts o *scores* corresponents a la matriu X^*, com:

$$Y_{(n,m)} = X^*_{(n,m)}.V_{(m,m)}$$

3.2 Anàlisi de variables canòniques (Canonical Variables Analysis, CVA)

La CVA és una tècnica la idea original de la qual va ser aportada per Fisher [Fis36], [Fis38]. Més tard va ser desenvolupada en el cas de dues dimensions per Foley i Sammon [Sam70], [Fol75]. Posteriorment ha estat àmpliament emprada per a dues i més dimensions. Vegeu [Dud73], [Lac75], [Che92], [Joh92], [Kii92], [Ren92], [Krz93] i [Krz94].

La CVA és una tècnica molt útil amb vista a reduir el nombre de característiques (dimensionalitat del problema), cosa que permet obtenir un problema més manejable. A més, és un mètode desenvolupat per accentuar les diferències entre les diferents classes d'objectes. L'avantatge principal d'aquest algorisme és que està basat en criteris de discriminació o separació de classes, i no en criteris d'ajust de punts (criteris de regressió), com seria el cas de la PCA.

La figura 3.4. mostra que el nombre de característiques es pot reduir de dues a una simplement projectant les dades de dues dimensions sobre una línia. Tot i que els objectes es trobin ben separats en l'espai de dues dimensions, pot ser que en projectar-los sobre aquesta línia es produeixi un cert grau de confusió. Girant aquesta línia, però, podem trobar una direcció en la qual la confusió sigui mínima (màxima separació entre les diferents classes).

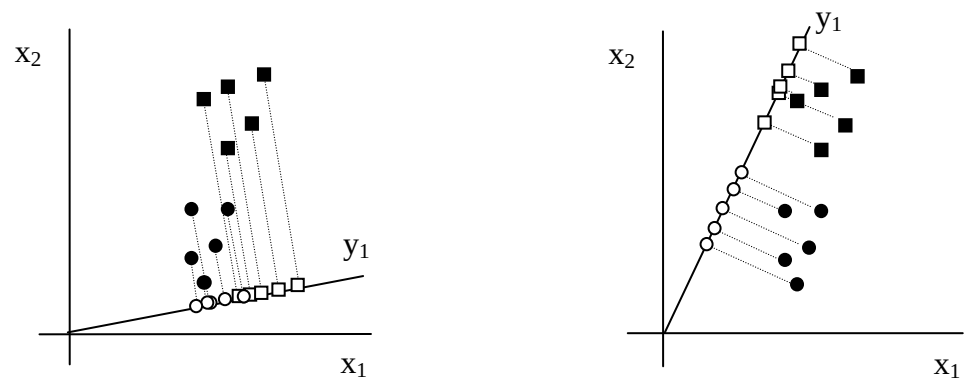


Figura 3.4.

En el cas de dues dimensions, la CVA pretén trobar la línia que separi millor els valors mitjans (centres de gravetat) de cada classe. En el cas de *m* dimensions i *c* classes, la CVA ens porta a $s = mínim[(m-1,c),(m,c-1)]$, eixos ortogonals en la mètrica *S* (matriu global de covariància). Aquests eixos són òptims per separar els valors mitjans de les classes i s'anomenen *variables*, *variables canòniques*, *funcions discriminants*, *coordenades discriminants* o, simplement, *discriminants*.

Considerem un problema de classificació multiclasse. Tenim *c* classes d'objectes amb n_i objectes pertanyents a la classe *i*-èssima. El nombre total d'objectes ve donat per $n = \sum_{i=1}^{c} n_i$, on cada objecte queda definit per *m* característiques. El vector de característiques d'un objecte qualsevol vindrà donat per $x^T = (x_1, x_{2...}, x_m)$.

Cas 1. Problema amb *c* = 2 classes

Fisher [Fis38] va suggerir transformar les observacions *m*-dimensionals *x* (característiques) en observacions univariades *y* (variables), de manera que les *y* es trobessin el màxim de separades possible. Fisher no va suposar que les classes tinguessin distribució normal, però va suposar la hipòtesi que tenien idèntiques matrius de covariàncies. Va pensar que seleccionant combinacions lineals adequades de les característiques *x* podria aconseguir les

variables canòniques *y* òptimes per separar les dues poblacions. Va suggerir maximitzar el quocient:

$$\phi = \frac{\text{diferència entre els valors mitjans de les } y}{\text{variància conjunta de les dues classes en l'espai de les } y}$$

Es defineixen:

Vector de característiques: $x^T = (x_1,x_2...,x_m)$ (3.13)

Vector dels coeficients de la combinació lineal: $v^T = (v_1,v_2...,v_m)$
(3.14)

Vector de variables: $y = v^T.x$
(3.15)

Matriu conjunta de covariàncies de les dues classes:

$$S_1 = S_2 = S = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T \qquad (3.16)$$

Com que es compleix la relació següent:

$$S_y = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (v^T.x_{ij} - v^T.\bar{x}_i)(v^T.x_{ij} - v^T\bar{x}_i)^T = v^T \left[\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T \right].v ,$$

resulta que:

$$S_y = v^T.S.v \qquad (3.17)$$

El quocient que s'ha de maximitzar resulta ser:

$$\phi = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y} = \frac{(vT \ .\bar{x}_1 - v^T.\bar{x}_2)^2}{v^T.S.v} . \qquad (3.18)$$

Ara cal maximitzar ϕ:

$$\frac{\partial \phi}{\partial v} = 0 \ \Rightarrow \qquad 0 = \frac{2(v^T.\bar{x}_1 - v^T.\bar{x}_2)(\bar{x}_1 - \bar{x}_2)v^T.S.v - (v^T.\bar{x}_1 - v^T.\bar{x}_2)^2.2.S.v}{(v^T.S.v)^2}$$

D'aquí resulta que:

$$v = \text{ct}.S^{-1}.(\bar{x}_1 - \bar{x}_2). \qquad (3.19)$$

La constant normalment s'agafa igual a 1, de manera que s'obté:

$$y = v^T.x = (\bar{x}_1 - \bar{x}_2)^T.\ S^{-1}.x \qquad (3.20)$$

Cas 2. Problema amb $c \geq 2$ classes

El procés que s'ha de realitzar és idèntic a l'anterior, però generalitzat a *c* classes. Vegeu [Joh92].

Es defineixen les matrius següents:

Matriu de dispersió entre classes (Between-Groups Covariance Matrix):

$$B_{(m,m)} = \sum_{i=1}^{c} n_i.(\bar{x}_i - \bar{x}) \cdot (\bar{x}_i - \bar{x})^T \tag{3.21}$$

Matriu de dispersió interna de les classes (Within-Groups Covariance Matrix):

$$W_{(m,m)} = \sum_{i=1}^{c} (n_i - 1).S_i = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \cdot (x_{ij} - \bar{x}_i)^T \tag{3.22}$$

Matriu de dispersió total (Overall Covariance Matrix):

$$S_{(m,m)} = \frac{1}{n-c} \cdot \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \cdot (x_{ij} - \bar{x}_i)^T \tag{3.23}$$

Com que se suposen totes les classes amb igual variabilitat, es té:

$$S_1 = S_2 = ... = S_c = S \tag{3.24}$$

D'aquí resulta que:

$$W = (n\text{-}c).S, \quad \text{amb} \quad n = \sum_{i=1}^{c} n_i \tag{3.25}$$

Per tant, *W i S* només difereixen en un escalar.

Per tal de calcular uns nous eixos que permetin separar òptimament les diferents classes, Fisher va suggerir maximitzar el quocient següent:

$$\phi = \frac{\sum (\text{Distància de les classes a la mitjana global en l'espai de les } y \text{ al quadrat})}{\text{variància conjunta de les } y} \tag{3.26}$$

D'aquí resulta que:

$$\phi = \frac{\sum_{i=1}^{c} |\bar{y}_i - \bar{y}|^2}{v^T.S.v} \Rightarrow \phi = ct \cdot \frac{\sum_{i=1}^{c} (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T}{v^T.W.v} = ct. \frac{\sum_{i=1}^{c} (v^T.\bar{x}_i - v^T.\bar{x})(v^T.\bar{x}_i - v^T.\bar{x})^T}{v^T.W.v}$$

$$\phi = ct. \frac{v^T.\left[\sum_{i=1}^{c} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T\right].v}{v^T.W.v} = ct'. \frac{v^T.B.v}{v^T.W.v} \tag{3.27}$$

Maximitzant ϕ s'obté:

$$\frac{\partial \phi}{\partial v} = 0 \;\Rightarrow\; (B - \lambda.W).v = 0 \;\Rightarrow\; (W^{-1}.B - \lambda.I).v = 0. \qquad (3.28)$$

En resulta un problema de valors i vectors propis. Si $m > c$, no hi ha més de $s = c-1$ solucions no nul·les, donades per:

λ_i: vectors propis de $W^{-1}.B$ $i = 1,2...,s$

v_i: vectors propis de $W^{-1}.B$ $i = 1,2...,s$

Resulta:

$$Y_{(n,s)} = X_{(n,m)}.V_{(m,s)}$$

Les columnes de la matriu V són els vectors propis v_i.

Es pot demostrar que: $v_i^T.W.v_j = v_j^T.W.v_i = 0 \quad \forall i \neq j$ (3.29)

Per tant, els vectors propis v_i són ortogonals en la mètrica definida per W.

Algorisme CVA

L'algorisme *CVA* és el següent:

1.- S'agafa la matriu base $X_{(n,m)}$.

2.- La matriu anterior se centra o s'autoescala, de manera que s'obté la matriu transformada $X^*_{(n,m)}$.

3.- Es calculen els vectors de mitjanes de cada classe: $\bar{x}_i = \frac{1}{n_i} \cdot \sum_{i=1}^{n_i} x^*_{ij}$ amb $i = 1,2,...c$)

4.- Es calcula $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{c} n_i.\bar{x}_i$, on $n = \sum_{i=1}^{c} n_i$.

5.- Es calcula la matriu de dispersió B: $B_{(m,m)} = \sum_{i=1}^{c} n_i.(\bar{x}_i - \bar{x}).(\bar{x}_i - \bar{x})^T$.

És una matriu simètrica i semidefinida positiva. El seu rang és, com a màxim, $c-1$, ja que el rang del producte extern de dos vectors és inferior o igual a 1.

6.- Es calcular la matriu de dispersió W: $W_{(m,m)} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x^*_{ij} - \bar{x}_i) \cdot (x^*_{ij} - \bar{x}_i)^T$.

És una matriu simètrica i semidefinida positiva i és proporcional a la matriu de covariàncies.

7.- Es compleix: $B.v_i = \lambda_i.W.v_i \;\Rightarrow\; (W^{-1}.B).v_i = \lambda_i.v_i$

En resulta un problema de valors propis. Però la matriu $(W^{-1}.B)$ no és simètrica, i per diagonalitzar-la cal realitzar abans la descomposició de *Cholesky* de la matriu W (vegeu [Pre92]), resultant: $W = L.L^T$, on L és una matriu que té zeros per sobre de la diagonal.

El problema queda expressat d'aquesta manera:

$$(L.L^T)^{-1}.B.v_i = \lambda_i.\ v_i \ \Rightarrow\ C.(L^T.v_i) = \lambda_i.(L^T.v_i),$$

$$\text{amb}\ \ C = L^{-1}.B.(L^{-1})^T.$$

La matriu C és simètrica i els seus valors propis són els mateixos que els del problema original.

7.1 Es fa la descomposició de *Cholesky* de W i es determinant la matriu L.

7.2 Es calcula la matriu $C = L^{-1}.B.(L^{-1})^T$.

7.3 Es determinen els s primers vectors propis normalitzats (a_i) de C, ordenats de manera que els seus valors propis estiguin en ordre decreixent. A partir d'aquests es construeix la matriu $A_{(m,s)}$ les columnes de la qual són els a_i.

8.- Com que $C.(L^T.v_i) = \lambda_i.(L^T.v_i)$, es veu que els vectors propis a_i de la matriu C compleixen:

$$a_i = L^T.v_i \quad \Rightarrow v_i = (L^T)^{-1}.a_i.$$

Ara cal construir la matriu de vectors propis del problema original:

$$V_{(m,s)} = (L^T)^{-1}{}_{(m,m)}.A_{(m,s)}.$$

9.- Les variables es calculen d'aquesta manera:

$$Y_{(n,s)} = X^*{}_{(n,m)}.V_{(m,s)}$$

Defectes que presenta el mètode CVA

La problemàtica que presenta l'algorisme CVA es pot resumir en els punts següents:

1. Suposa igual variància per a totes les classes. Assumeix que totes les classes estan disperses de la mateixa manera, fet que normalment no és així.
2. Les variables canòniques no són necessàriament ortogonals en l'espai definit per les característiques discriminants. En canvi, són ortogonals en l'espai definit per la matriu W.
3. La CVA no pot ser aplicada directament a conjunts en què el nombre de característiques sigui superior al nombre d'objectes menys el nombre de classes. Això es deu al fet que a llavors, la matriu W esdevé singular (i, per tant, no es pot invertir).

Relació entre els components principals i les variables canòniques

Krzanowski [Krz95] interpreta els components principals com combinacions lineals de les variables originals (característiques discriminants) de la forma: $y_i = v_i^T.x$, tals que maximitzen l'expressió $v^T.S.v$ sota les restriccions d'ortonormalitat:

$$v_i^T.v_i = 1\ (\forall i) \quad \text{i} \quad v_i^T.v_j = 0\ (\forall i \neq j) \ \text{amb}\ \ i,j = 1,2...,m. \tag{3.30}$$

Els vectors de coeficients v_i són els vectors propis corresponents als valors propis de S ordenats en valor decreixent. En el cas de centrar les dades, la matriu S es calcula de la manera següent:

$$S = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}) \cdot (x_{ij} - \bar{x})^T \tag{3.31}$$

Tal com s'acaba d'explicar, els components principals són ortogonals en l'espai definit per les característiques.

V és la matriu les columnes de la qual, estan formades pels vectors propis v_i, i compleix:

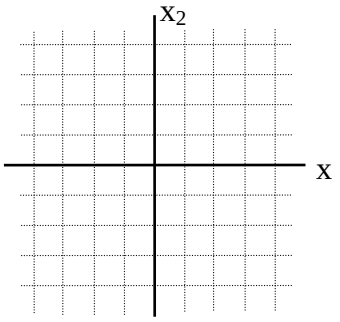
$$V.V^T = V^T.V = I. \tag{3.32}$$

Això ens indica que la transformació de les característiques x en les variables y (components principals), donada per l'expressió $Y = X.V$, és ortogonal. Geomètricament, significa que els PC formen 90º entre si i que els nous eixos OY_i s'originen en efectuar una rotació rígida dels eixos originals OX_i.

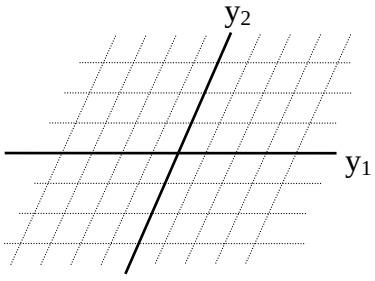
En canvi, Krzanowski interpreta les variables canòniques com les combinacions lineals donades per l'expressió $y_i = v_i^T.x$, tals que maximitzin el quocient $v^T.B.v / v^T.W.v$ sota les restriccions d'ortonormalitat en la mètrica W:

$$v_i^T.W.v_j = 1\ (\forall i) \quad i \quad v_i^T.W.v_j = 0\ \ (\forall i \neq j), \quad \text{amb}\ \ i,j = 1,2...,s. \tag{3.33}$$

Els vectors v_i, amb i=1,2..,s, són els vectors propis corresponents als valors propis de la matriu $W^{-1}.B$, ordenats en valor decreixent.



Graella original



Graella deformada

Figura 3.5.

En el cas de les variables canòniques, els eixos que aquestes originen no tenen perquè ser ortogonals en l'espai de les variables originals (característiques). Per tant, no formaran 90º entre si. Geomètricament, els nous eixos OY_i no es poden veure com una simple rotació dels eixos originals OX_i, ja que els nous eixos no seran ortogonals entre si. Això implica que la graella de referència definida pels eixos OX_i apareix deformada en l'espai definit pels eixos Oy_i. En la figura 3.5. es pot veure la deformació de la graella original.

Interpretació geomètrica dels components principals i de les variables canòniques

La figura 3.6. mostra tres classes d'objectes (*a, b* i *c),* representades en l'espai de les característiques:

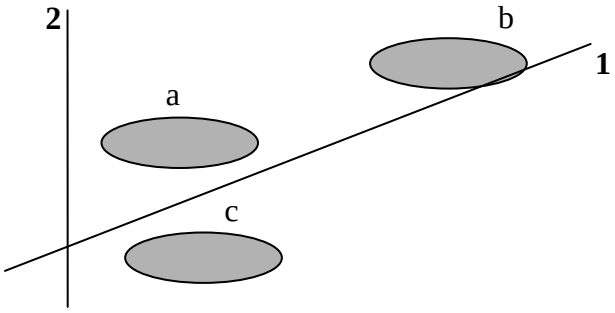


Figura 3.6.

Eix 1: correspon al primer component principal. És l'eix que millor ajusta el conjunt global de punts, sense diferenciar les diferents classes. Els components principals són òptims per ajustar les dades a un model de regressió. L'eix 1 correspon a la matriu *S.*

Eix 2: correspon a la primera variable canònica. És l'eix en el qual les diferents classes apareixen màximament separades. Les variables canòniques són òptimes per separar les diferents classes, són òptimes per classificar. L'eix 2 correspon a la matriu *B.*

3.3. Anàlisi de components principals discriminants (Discriminant Principal Components Analysis, DPCA)

La DPCA és una tècnica desenvolupada per Yendle i Macfie [Yen89]. Quan es tenen pocs objectes, la CVA estàndard és inviable (perquè la matriu *W* esdevé singular) i en aquests casos moltes vegades s'aplica la PCA seguida de la CVA. La tècnica DPCA ve a ser una fusió entre la PCA i la CVA i també és aplicable en el cas de tenir pocs objectes.

Els objectius fonamentals de la DPCA són:

1. Accentuar al màxim la distància entre classes.

2. Trobar variables discriminants que siguin ortogonals en l'espai de les característiques.

3. Aconseguir que el mètode serveixi en el cas de tenir pocs objectes de cada classe, fins i tot quan la matriu *W* sigui singular.

Per assolir tots els objectius anteriors, Yendle i Macfie van proposar trobar unes noves direccions v_i que siguin ortogonals en l'espai definit per les característiques *x*, tals que maximitzin l'expressió:

$$v_i^T .B^* .v_i , \tag{3.34}$$

on *B** és una versió escalada de la matriu de dispersió *Between-Group B.*

Per accentuar al màxim la distància entre classes diferents, cal basar-se en la matriu *B*. Però les variables que tenen un valor elevat de la variància *Between-Groups* no tenen perquè ser les més discriminants, ja que poden presentar al mateix temps un valor elevat de la variància *Within-Group.* En el camp de la classificació és molt conegut que el poder discriminatori està governat pel quocient:

$$\frac{\text{variància } \textit{between - groups}}{\text{variància } \textit{within - group}} \tag{3.35}$$

La DPCA esmena aquest problema reescalant cada característica de la matriu original *X* per la desviació estàndard *Within-Groups*. Això significa que la característica *i* de l'objecte *j* es divideix per $\sqrt{w_{ii}}$. Així, x_{ij} passa a transformar-se en $x_{ij} / \sqrt{w_{ii}}$, fet totalment equivalent a recalcular la matriu $B = (b_{ij})$ per la matriu $B^* = (b^*_{ij})$, on:

$$b^*_{ij} = b_{ij}/(\sqrt{w_{ii}} .\sqrt{w_{jj}}) \tag{3.36}$$

Després, s'aplica la *PCA* a la matriu *B**, obtenint-se les noves direccions que seran els vectors propis de *B** i compliran les condicions d'ortonormalitat:

$$v_i^T.v_i = 1 \;\; i \;\; v_i^T.v_j = 0 \quad \forall i \neq j = 1,2,...,m \tag{3.37}$$

Algorisme *DPCA*

L'algorisme *DPCA* és el següent:

1.- S'agafa la matriu base $X_{(n,m)}$.

2.- La matriu anterior es centra o s'autoescala, de manera que s'obté la matriu transformada $X^*_{(n,m)}$.

3.- Es calculen les matrius de dispersió *Between-Groups* i *Within-Groups*, donades respectivament per $B_{(m,m)}$ *i* $W_{(m,m)}$.

4.- Es calcula la matriu transformada B^*, que té per components:

$$b^*_{ij} = \frac{b_{ij}}{\sqrt{w_{ii}}.\sqrt{w_{jj}}}$$

5.- Es realitza la *PCA* sobre la matriu B^*. S'ha de calcular la descomposició en valors singulars de la matriu B^*:

$$svd(B^*) = U.\Sigma.V^T$$

6.- Les variables es calculen de la manera següent:

$$Y_{(n,m)} = X^*_{(n,m)}.V_{(m,m)}$$

3.4. Anàlisi de variables canòniques ortogonals (Ortogonal Canonical Variables Analysis, OCVA)

La OCVA és una tècnica que va ser desenvolupada per Krzanowski [Krz95] amb la finalitat de realitzar un algorisme matemàtic que fos al més similar possible a la CVA, però de tal manera que proporcionés un conjunt de direccions perpendiculars entre si en l'espai definit per les característiques.

Per poder aplicar l'algorisme OCVA cal tenir calculada la matriu de característiques $X_{(n,m)}$. A partir d'aquesta es calcularan les direccions successives $v_1, v_2 \ldots, v_m$ que maximitzin el quocient següent:

$$\phi = \frac{v_i^T.B.v_i}{v_i^T.W.v_i}$$

A aquestes direccions, se'ls imposen les restriccions d'ortonormalitat:

$$v_i'.v_i = 1 \quad i \quad v_i'.v_j = 0 \quad \forall i \neq j.$$

Per trobar els valors òptims dels v_i cal disposar d'un algorisme que ens permeti minimitzar una funció multivariable. En aquest cas, s'ha elegit l'algorisme de minimització de funcions de Powell per raons d'eficiència i perquè és un dels mètodes més emprats. Per a detalls sobre l'algorisme, vegeu [Pre92].

Primer s'ha de calcular la matriu X. Per tal de trobar la primera direcció v_1, es comença la cerca del mínim de la funció ϕ^{-1}. A l'algorisme de Powell, cal donar-li un valor inicial per començar la cerca del mínim. Aquest pot ésser:

$$v_{1o}^T = (1/\sqrt{m} ,...,1/\sqrt{m}) \tag{3.38}$$

A partir d'aquí, l'algorisme determina el vector v_1 que minimitzi el quocient ϕ^{-1}. Després aquest vector s'ha de normalitzar. Un cop es té v_1, s'ha de trobar la direcció v_2 que produeixi el següent valor més baix de ϕ^{-1} i que sigui ortogonal a v_1. Per fer això cal projectar les dades en el subespai ortogonal a v_1. Aquest fet s'aconsegueix construint la matriu:

$$X_{(2)} = X - X.v_1.v_1^T. \tag{3.39}$$

La matriu $X_{(2)}$ substituirà la matriu X en el càlcul de ϕ^{-1}. Es tornarà a aplicar el mètode de Powell agafant com a vector inicial:

$$v_{2o} = v_{1o} - (v_{1o}^T.v_1).v_1 \tag{3.40}$$

D'aquí resultarà el vector v^* que s'haurà de fer ortogonal a v_1 mitjançant el procés d'ortogonalització de *Gram-Schmidt*. Això es fa de la següent manera:

$$v_2 = v^* - (v_1^T.v^*).v_1. \tag{3.41}$$

Després, el vector v_2 es normalitzarà. En els passos successius, i situats en el pas t-èssim, per calcular el valor mínim de ϕ^{-1} es farà servir la matriu:

$$X_{(t)} = X_{(t-1)} - X.v_{t-1}.v_{t-1}^T \tag{3.42}$$

i s'aplicarà el mètode de minimització de Powell, agafant com a vector inicial:

$$v_{t_o} = v_{1o} - \sum_{i=1}^{t-1} (v_{1o}^T.v_i).v_i . \tag{3.43}$$

L'algorisme de minimització ens proporcionarà el vector v^*, que es farà ortogonal en els vectors v_1, v_2 ..., v_{t-1} aplicant el procés d'ortogonalització de Gram-Schmidt:

$$v_t = v^* - \sum_{i=1}^{t-1} (v_i^T.v^*).v_i . \tag{3.44}$$

Un cop fet aquest procés *m* vegades, s'obté la matriu $V_{(m,m)}$, les columnes de la qual estaran formades pel sistema de vectors ortonormals v_i, amb i =1,2...,*m*.

Algorisme OCVA

L'algorisme OCVA és el següent:

1.- S'agafa la matriu base $X_{(n,m)}$.

2.- La matriu anterior se centra o s'autoescala, de manera que s'obté la matriu transformada $X^*_{(n,m)}$.

A partir d'aquí, de manera iterativa es realitza el càlcul dels vectors ortogonals que minimitzen la funció ϕ^{-1}:

Pas 1.- S'agafa $X_{(1)} = X$ i s'aplica el mètode de minimització de Powell a la funció ϕ^{-1}, prenent el vector inicial $v_{1o}^T = (1/\sqrt{m} ...,1/\sqrt{m})$. D'aquí resulta el vector v_1, que s'ha de normalitzar.

Pas 2.- Es calcula $X_{(2)} = X_{(1)} - X.v_1.v_1^T$. Després s'aplica el mètode de minimització de Powell a la funció ϕ^{-1}. Cal prendre com a llavor el vector inicial $v_{2o} = v_{1o} - (v_{1o}^T.v_1).v_1$ i l'algorisme de Powell proporcionarà el vector v^*. Després es calcula $v_2 = v^* - (v_1^T.v^*).v_1$ i es normalitza.

Pas t.- Es calcula $X_{(t)} = X_{(t-1)} - X.v_{t-1}.v_{t-1}^T$ i s'aplica el mètode de minimització de Powell a ϕ^{-1} prenent el vector inicial $v_{t_0} = v_{1o} - \sum_{i=1}^{t-1}(v_{1o}^T.v_i).v_i$. D'aquí resulta el vector v^*. Es calcula $v_t = v^* - \sum_{i=1}^{t-1}(v_i^T.v^*).v_i$ i es normalitza.

Del tram d'algorisme anterior resulten els *m* vectors ortonormals.

3.- Es construeix la matriu $V_{(m,m)}$, les columnes de la qual estaran formades pels *m* vectors v_i, ortonormals entre si.

4.- Es calculen les noves coordenades dels punts corresponents a la matriu X^* aplicant la fórmula següent:

$$Y_{(n,m)} = X^*_{(n,m)}.V_{(m,m)}$$

3.5. Pretractament de les dades

Com que les característiques sota anàlisi sovint poden presentar valors molt diferents, és comú fer un pretractament de les dades (matriu *X*) abans d'aplicar els algorismes de reducció de dimensions. Els dos tipus més freqüents de pretractament són el centrat i l'estandardització (o autoescalat) de les dades.

És molt freqüent passar els algorismes de reducció de dimensions (PCA, CVA, DPCA i OCVA) per triplicat: a la matriu *X* sense cap mena de pretractament, a la matriu *X* centrada i a la matriu *X* autoescalada. Després es comprova quin dels tres pretractaments proporciona resultats més satisfactoris.

Centrat de les dades

Quan les diferents característiques del problema tenen valors mitjans força diferents, és convenient centrar-les respecte al seu valor mitjà. El centrat de les dades es fa columna a columna. A l'element *i*-èssim de la *j*-èssima columna de la matriu *X* de dades, se li aplica la següent transformació de centrat:

$$x'_{ij} = x_{ij} - \bar{x}_j , \tag{3.45}$$

on $\bar{x}_j$ és el valor mitjà de la *j*-èssima columna de la matriu *X* de dades.

Estandardització o autoescalat de les dades

Quan les diferents característiques del problema tenen valors mitjans força diferents i/o variabilitats bastant diferents, és convenient estandarditzar-les. Aquesta transformació de les dades es realitza columna a columna. A l'element *i*-èssim de la *j*-èssima columna de la matriu *X* de dades, se li aplica la següent transformació d'estandardització:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} , \tag{3.46}$$

on $\bar{x}_j$ és el valor mitjà de la columna *j*-èssima de la matriu *X* de dades i s_j és la seva desviació estàndard.

3.6. Nombre òptim de variables que cal retenir

En els apartats 3.1. a 3.4. s'han explicat els diferents mètodes de transformació de les característiques discriminants en variables. Aquest apartat analitza quin és el nombre òptim

*m** de variables que cal retenir, és a dir, el subconjunt *m** del conjunt total de variables que s'utilitzaran per caracteritzar tots els objectes.

Els algorismes de selecció del nombre de variables que cal retenir serveixen per als mètodes PCA, CVA, DPCA i OCVA ja que els valors propis resultants d'aquests mètodes tenen interpretacions semblants. Com més gran sigui el valor propi, més separació entre classes genera aquest en els casos de la CVA ,la DPCA i l'OCVA, i més variància explica en el cas de la PCA.

Un resum dels diferents mètodes utilitzats per obtenir el nombre de variables que s'han de retenir es troba a [Krz93]. Els mètodes són els següents:

- **Diagrama de caigudes (Scree Diagram).** Consisteix a fer un gràfic del valor numèric dels valors propis λ_i respecte a *i*. Quan el pendent es fa pla, vol dir que afegint més variables, aquestes pràcticament no aporten més informació. Es retenen les *m** primeres variables corresponents a la regió on la corba no és plana. Vegeu [Krz93] i [Esb94].

  La figura 3.7. mostra un exemple d'aplicació d'aquest mètode. En aquest cas particular, com que per més de tres variables el pendent de la corba resulta molt pla, ens quedarem amb les tres primeres variables.

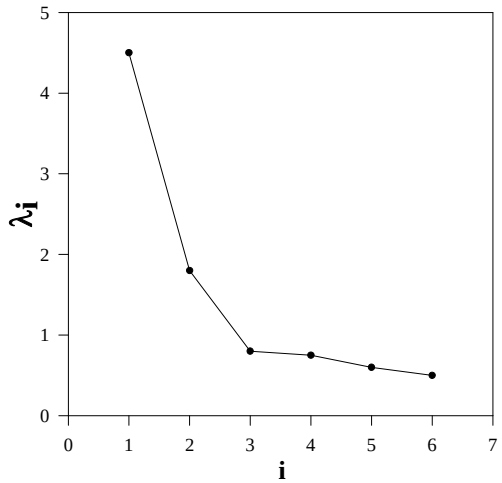


Figura 3.7.

- **Valor relatiu de les *m** primeres variables**. Es retenen les primeres *m** variables que compleixen que la relació següent:

$$P_{\%} = 100.\sum_{i=1}^{m^*}\lambda_i \,/\, \sum_{i=1}^{m}\lambda_i \qquad (3.47)$$

sigui superior al 80 %, al 90 % o al 95 %, depenent dels autors. Consulteu [Krz93].

- **Valors propis superiors a la unitat**. Molts autors apliquen el mètode de retenir totes les variables els valors propis associats a les quals tenen un valor superior a la unitat. Això implica que aquestes variables expliquen com a mínim tanta variància (informació) com les característiques del problema. Consulteu [Krz93].

- **Distribució esfèrica**. Si les dades es trobessin distribuïdes uniformement en tots els eixos, sense presentar direccions privilegiades, tindríem una distribució esfèrica i tots els valors propis serien iguals i de valor:

$$\bar{\lambda} = \sum_{i=1}^{m} \lambda_i \text{ /m.} \tag{3.48}$$

  Per tant, suposarem importants els valors propis de valor numèric superior a $\bar{\lambda}$ i menyspreables els restants. Consulteu [Krz93].

- **Diagrama de caigudes de l'error de classificació calculat aplicant validació creuada.** La validació creuada (*leave-one-out,* vegeu [Lac67]) és una tècnica iterativa que consisteix a classificar un a un tots els objectes del conjunt de calibratge. Es basa a treure el primer objecte del conjunt de calibratge i predir la classe a la qual pertany. Després es reintrodueix l'objecte anterior en el conjunt de calibratge, es retira el segon i es passa a predir aquest últim. Aquest procés es repeteix per a tots els objectes del conjunt de calibratge.

  Aquest criteri de selecció del nombre de variables que s'han de retenir s'aplica un cop es té a punt l'algorisme de predicció o de classificació. Consisteix a anar aplicant iterativament l'algorisme de classificació al conjunt d'objectes de calibratge, retenint en el primer pas només la primera variable per a cada objecte, en el segon pas es retenen les dues primeres variables i així successivament fins a agafar-les totes. Una vegada fet això, es fa un gràfic del valor de la funció error de classificació respecte al nombre de variables retingudes. Els valors de la funció error de classificació es calculen aplicant validació creuada a les dades de calibratge. Quan el pendent es fa pla, vol dir que encara que afegim més variables, aquestes pràcticament no aporten més informació. Ens quedarem amb les m^* primeres variables corresponents a la regió on la corba no és plana.

- **Degeneració dels valors propis** Aquest mètode, el van desenvolupar North, Bell i Callahan, vegeu [Nor82]. En l'article es defineix el concepte *de degeneració dels valors propis*. El valor propi λ_ι (amb $\lambda_i > \lambda_{i+1}$) serà degenerat si compleix:

$$\lambda_i - \alpha.\lambda_i < \lambda_{i+1} + \alpha.\lambda_{i+1} \quad \Rightarrow \quad \lambda_i < \lambda_{i+1} \cdot \frac{1+\alpha}{1-\alpha}, \tag{3.49}$$

on $\alpha \approx \sqrt{\frac{2}{n}}$ i n: nombre total d'objectes.

Es retenen les primeres variables corresponents als valors propis λ_i no degenerats.

- **Mètode d'Eastment i Krzanowski** Aquest mètode es troba desenvolupat en [Eas82] i s'aplica sobretot en processos de predicció (regressió multivariable) després d'haver calibrat el model. Per seleccionar el nombre òptim de variables parteix del càlcul de l'expressió següent:

$$W(r) = \left(\frac{PRESS(r-1) - PRESS(r)}{D_r} \right) \cdot \left(\frac{D_L}{PRESS(r)} \right), \tag{3.50}$$

on

$$D_r = n + m - 2.r \,, \quad D_L = n.m - m - D_r.$$

r és el nombre de variables que s'han de retenir.

El valor òptim de r és el valor més gran de r que compleixi que $W(r) > 1$.

El PRESS (Predictive Residual Error Sum of Squares) és una mesura de l'exactitud de la resposta proporcionada pel model matemàtic i es defineix de la manera següent:

$$PRESS = \sum_{i=1}^{n'} \sum_{j=1}^{m} (y_{ij} - y'_{ij})^2, \tag{3.51}$$

on y_{ij} és la sortida teòrica (*1* o *0*) del conjunt d'objectes de calibratge i y'_{ij} és la sortida d'aquests proporcionada pel model matemàtic.

Com s'acaba de veure, hi ha múltiples criteris per determinar el nombre de variables que s'han de retenir, i en molts dels casos no tots portaran al mateix resultat. Això ens indica que no hi ha cap mètode universal de selecció del nombre òptim de variables que cal retenir i que la decisió depèn força del criteri de l'investigador. Moltes vegades s'escull el criteri que indica un nombre més alt de variables que cal retenir.