Después de sopesar los datos haciendo XD, la función a optimizar se transforma en:

$$J_D = \frac{\left|W^T D\Sigma_B DW\right|}{|W^T D\Sigma_W DW|}. \tag{3.8}$$

3.2.3. Variables ponderadas y criterio de relevancia

En las subsecciones precedentes se definieron algunas transformaciones lineales sopesadas; ahora el interés es proyectar los datos a un espacio de dimensión f. Tal dimensión depende del criterio de rotación escogido; por ejemplo, si se tiene un problema bi-clase y se quiere probar WRDA, la dimensión fija debe ser $f = 1$, con el fin de asegurar la convergencia.

Para la evaluación de la relevancia de una proyección sopesada a una dimensión fija, se utiliza una medida de separabilidad. El parámetro a ser optimizado es la matriz de peso D, y el criterio seleccionado es el cociente de trazas de las matrices de dispersión entre-clases e intra-clases. Este criterio es conocido como J_4 [82]. Para los datos proyectados y ponderados, esta medida es:

$$J_4(D, \Phi) = \frac{traza(\Phi^T D\Sigma_B D\Phi)}{traza(\Phi^T D\Sigma_W D\Phi)}, \tag{3.9}$$

donde Φ puede ser U o W. El tamaño de Φ es $c \times f$ y f denota la dimensión fija, correspondiente al número de vectores de proyección Φ tal que $\Phi = (\phi_1, \phi_2, ..., \phi_f)$.

Reescribiendo la función D como un vector columna d, y utilizando el producto de *Hadamard* de matrices (denotado como $\circ$), las trazas de la Ecuación 3.7 se pueden reescribir como:

$$traza(\Phi^T D\Sigma_B D\Phi) = \sum_{i=1}^{f} d^T(\Sigma_B \circ \phi_i\phi_i^T)d = d^T\left(\sum_{i=1}^{f} \Sigma_B \circ \phi_i\phi_i^T\right)d. \tag{3.10}$$

Entonces, la Ecuación 3.7 se transforma en:

$$J_4(d) = \frac{d^T \left(\sum_{i=1}^{f} \Sigma_B \circ \phi_i \phi_i^T \right) d}{d^T \left(\sum_{i=1}^{f} \Sigma_W \circ \phi_i \phi_i^T \right) d}. \tag{3.11}$$

Esta función es similar, en naturaleza, a la función de análisis discriminante lineal LDA. Por consiguiente, la solución de d con la norma de L_2, será encogida por el "eigenvector"principal dado por:

$$\left(\sum_{i=1}^{f} \Sigma_W \circ \phi_i \phi_i^T \right)^{-1} \left(\sum_{i=1}^{f} \Sigma_B \circ \phi_i \phi_i^T \right). \tag{3.12}$$

Note que este tipo de descripción supone los elementos de Φ como estáticos. Este problema se supera intercalando el cálculo de d y Φ hasta la convergencia de ambos. En cuanto a la interpretabilidad de los pesos, generalmente se requiere que sean positivos para definir la dispersión. No obstante, en el contexto de la función de relevancia usada en este trabajo, pueden obtenerse signos negativos; por esto se toma el valor absoluto de d.

3.3. Reducción ponderada WPCA, WRDA, LPCA1 y LPCA2.

En [78] se presenta convergencia por potencias del método Q-α, cuya función objetivo es similar a J_4. Dicha prueba fue realizada para un caso particular, donde la función objetivo debe poseer una matriz escasa, es decir existe un subconjunto de características que induzca un "cluster" coherente y la función positiva, lo que en general no se puede cumplir. Por esta razón, en esta sección se hace la prueba de la convergencia de la función objetivo en WPCA, explicita en el lema 1 considerando $\delta = 0$ obtiene el problema de reducción $WPCA$, mientra que con $\delta \neq 0$ se obtiene $WRDA$.

El siguiente lema garantiza la convergencia de la función objetivo. Además, cualquier método de búsqueda converge al mismo límite.

Lema 1 *La función objetivo*

$$J_4(W,D) = d^T \left(B \cdot \sum_{i=1}^{K} w_i w_i^T + \delta I d \right)^{-1} A \cdot \sum_{i=1}^{k} w_i w_i^T d$$

converge.

Donde d es la diagonal de D, $A = \Sigma_B$, $B = \Sigma_W$ son las matrices de varianza inter e intra clases, w_i es la columna i de W, δ coeficiente regularizante y k es la dimensión de reducción.

Prueba 1 (Demostración 1. (versión analítica)) *La función objetivo se puede representar como:*

$$J(W,D) = \langle d, \Gamma_W(d) \rangle ,$$

donde

$$\Gamma_w(d) = \left(B \cdot \sum_{i=1}^{k} w_i w_i^T + \delta I d \right)^{-1} A \cdot \sum_{i=1}^{k} w_i w_i^T d.$$

Sea el conjunto $C = \{\langle d, \Gamma_W(d) \rangle : \|d\| = 1\}$. *Este conjunto es no vacío, ya que cualquier vector propio* $d = \beta_i$ *satisface la condición.*

Ahora, se quiere demostrar que el conjunto posee supremo. Al tomar una base de vectores propios $\{\beta_1, \beta_2, ..., \beta_n\}$ *asociada con la transformación* Γ_W*, entonces para cualquier vector* d *visto como combinación lineal de los* β_i*, se tiene que* $d = \sum_{i=1}^{n} c_i \beta_i$*, tal que:*

$$J(W,D) = \langle d, \Gamma_W(d) \rangle = \left\langle \sum\nolimits_{i=1}^{n} c_i \beta_i, \Gamma_W \left(\sum\nolimits_{i=1}^{n} c_i \beta_i \right) \right\rangle = \sum\nolimits_{i=1}^{n} \lambda_i c_i^2,$$

donde $\|d\|^2 = \|\sum_{i=1}^{n} c_i^2\| > 1$. *Luego* C *está acotado superiormente. Por el axioma del supremo existe* $\sup(C)$. *Al considerar*

$$\|d\|^2 = \sum\nolimits_{i=1}^{n} c_i^2 = 1 \; y \; \hat{\lambda} = \max_{i \leq i \leq n} \{\lambda_i\} ,$$

implica que $sup(C) = \hat{\lambda}$. *Por esta razón, si* $d = \hat{\beta}$ *el vector propio asociado al mayor valor propio* $\hat{\lambda}$*, maximiza la función objetivo.*

Prueba 2 (Demostración 2. (versión algebraica)) *Se maximiza la función objetivo*

$$J(w,D) = traza(W^T DADW)$$

sujeta a las restricciones: $W^TW = Id$, $traza(W^TDBDW) = 1$ *y* $\|d\| = 1$. *Los datos originalmente se encuentran almacenados en la matriz* X*; de ella se analiza su covarianza* X^TX*, la cual se descompone en* $X^TX = A + B$. *Tanto* A *como* B *son matrices simétricas y semidefinidas positivas, las cuales se pueden sustituir por descomposición de Cholesky:* $X^TX = A_1^TA_1 + B_1^TB_1$. *Multiplicando por* W_TD *a la izquierda y por* DW *a la derecha, y tomando las trazas se obtiene:*

$$J(W,D) = traza\left(W^TDA_1^TA_1DW\right) = traza\left(DA_1^TA_1DWW^T\right),$$

Usando la propiedad de ortogonalidad de W*, se llega a:*

$$J(W,D) = traza\left(DA_1^TA_1D\right) = \|A_1D\|_F^2.$$

Así,

$$J(W,D) = \left\|[d_1A_1(1,:), d_2A_1(2,:), \cdots, d_nA_1(n,:)]^T\right\|_2^2.$$

Su representación matricial es:

$$J(W,D) = \left\| \begin{bmatrix} A_1(:,1) & 0 & \cdots & 0 \\ 0 & A_1(:,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_1(:,n) \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} \right\|_2^2.$$

Por otro lado $traza(W^TDB_1^TB_1DW) = 1$ *implica que* $\sum d_i^2B_1(:,i)^TB_1(:,i) = 1$*, lo que se puede transformar en* $\sum \tilde{d}_i = 1$ *donde* $\tilde{d}_i = d_iy_i$*, tal que* $y_i = B_1(:,i)^TB_1(:,i)$. *Por lo anterior* D *puede verse de manera matricial como:*

$$J(W,D) = \left\| \begin{bmatrix} 1/y_1 & 0 & \cdots & 0 \\ 0 & 1/y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/y_n \end{bmatrix} \begin{bmatrix} \tilde{d}_1 \\ \tilde{d}_2 \\ \vdots \\ \tilde{d}_n \end{bmatrix} \right\|_2^2.$$

La anterior expresión se puede transformar en la siguiente función a maximizar: $J(W,D) = \left\|M\tilde{d}\right\|_2^2$*, sujeto a que* $\left\|\tilde{d}\right\| = 1$*, la cual, como es evidente, tiene como máximo* $\|M\|$.

A continuación se presenta el algoritmo WPCA y su convergencia.

3.3.1. Algoritmo WPCA y demostración de su convergencia

La naturaleza iterativa y la convergencia de la estimación de parámetros de E-M y PCA probabilístico son usados para obtener la convergencia del método WPCA, el cual se describe como sigue, donde r es el índice de iteración:

i. Normalice cada vector de características para obtener media cero y norma euclídea uno ($||x||^2 = 1$).

ii. Inicie con algún conjunto ortonormal de vectores $U^{(0)}$.

iii. Calcule D^r de la solución dada en la Ecuación 3.12 y pondere los datos.

iv. Calcule el $paso - E$ y el $paso - M$, a partir de la Ecuación 3.4, y de la Ecuación 3.5 respectivamente. Normalice las columnas de C para obtener $||C(:,i)||_2 = 1$.

v. Si $||C^{(r)} - C^{(r-1)}||_2 > \varepsilon$, regrese al paso iii.

vi. Ortonormalice el subespacio obtenido, hallando su descomposición en valores singulares (SVD), como sigue: $\mathrm{SVD}(C^T D X^T X D C) = A S A^T, C_{final} = A^T C$, donde A, S son los elementos obtenidos en la descomposición SVD.

Convergencia de WPCA

Como se mencionó en la sección anterior, al ponderar las características en la integración con el método EM, se pretende garantizar la convergencia de los pesos D, y así asegurar la obtención de las características relevantes. De la Ecuación 3.4 se tiene la siguiente relación en el algoritmo:

$$D^{(r)} X = C^{(r)} Z^{(r)} + V^{(r)}. \tag{3.13}$$

Se recuerda que r corresponde a la iteración. A medida que se aplica EM (aumenta r) la perturbación $V^{(r)}$ decrece, debido a que D^r se está aproximando a los ejes más discriminantes. Esto es, si $r \to \infty$, entonces $||V^{(r)}|| \to 0$, lo que garantiza la convergencia.

Teorema 1 *Si $C^{(r)} \to \hat{C}$ y $Z^{(r)} \to \hat{Z}$, entonces $D^{(r)} \to \hat{D}$.*

Prueba 3 *Dada la Ecuación 3.13 para las iteraciones r y $r+1$, la resta produce:*

$$(D^{(r+1)} - D^{(r)})X = (C^{(r+1)}Z^{(r+1)} - C^{(r)}Z^{(r)}) + (V^{(r+1)} - V^{(r)}).$$

Aplicando, a la relación anterior, cualquier tipo de norma, se obtiene:

$$\left\|(D^{(r+1)} - D^{(r)})X\right\| \leq \left\|C^{(r+1)}Z^{(r+1)} - C^{(r)}Z^{(r)} + V^{(r+1)} - V^{(r)}\right\|,$$

Sabemos que si $r \to \infty$, $||V^{(r)}|| \to 0$, entonces:

$$\left\|(D^{(r+1)} - D^{(r)})X\right\| \leq \left\|C^{(r+1)}Z^{(r+1)} - C^{(r)}Z^{(r)}\right\|.$$

Al sumar y restar $C^{(r+1)}Z^{(r)}$, al miembro derecho se tiene:

$$\left\|(D^{(r+1)} - D^{(r)})X\right\| \leq \left\|C^{(r+1)}(Z^{(r+1)} - Z^{(r)}) + (C^{(r+1)} - C^{(r)})Z^{(r)}\right\|.$$

Aplicando la desigualdad triangular y la propiedad multiplicativa se llega a:

$$\left\|(D^{(r+1)} - D^{(r)})\right\| \|X\| \leq \left\|C^{(r+1)}\right\| \left\|(Z^{(r+1)} - Z^{(r)})\right\| + \left\|(C^{(r+1)} - C^{(r)})\right\| \left\|Z^{(r)}\right\|,$$

Cuando $r \to \infty$, entonces

$$\left\|(D^{(r+1)} - D^{(r)})\right\| \|X\| \leq \left\|\hat{C}\right\| \left\|(Z^{(r+1)} - Z^{(r)})\right\| + \left\|(C^{(r+1)} - C^{(r)})\right\| \left\|\hat{Z}\right\|.$$

Por hipótesis X es la matriz de los datos originales, luego $||X|| > 0$. En un espacio normado, si una sucesión es convergente, es de Cauchy. Por lo tanto, si $r \to \infty$, luego $\left\|C^{(r+1)} - C^{(r)}\right\| \to 0$, $\left\|Z^{(r+1)} - Z^{(r)}\right\| \to 0$, y $\left\|D^{(r+1)} - D^{(r)}\right\| \to 0$. Así se obtiene la convergencia de los pesos al pertenecer a un espacio de Banach.

3.3.2. Algoritmo WRDA

Para este algoritmo no son importantes los errores producidos por la rotación con los datos ponderados, ya que tanto la función que calcula la rotación como la función que pondera tienen direcciones similares. El algoritmo se describe a continuación :

i. Fije la dimensión $k - 1$, siendo k el número de clases.

ii. Normalice cada vector de características para que tenga media cero y norma euclídea uno.

iii. Inicie con cualquier conjunto ortonormal de vectores $W^{(0)}$.

iv. Calcule d^r de la solución dada en la Ecuación 3.12 que pondera los datos.

v. Calcule los W^r a partir de las Ecuaciones 3.7 y 3.8.

vi. Si $\left\|W^{(r)} - W^{(r-1)}\right\|_2 > \varepsilon$, vaya al paso iii, con ε es un error fijo en el proceso.

La función objetivo es justamente la misma del **Lema 1** cuando $\delta \neq 0$. por esta razón la convergencia del método $WRDA$ queda implícitamente garantizada.

3.4. Métodos de reducción por líneas LPCA1, LPCA2

Los métodos de reducción mencionados anteriormente, tienen la desventaja de converger en un número grande de pasos. Por esta razón, al cambiar al proceso de búsqueda de la función objetivo J_4 se pretende reducir el tiempo de convergencia. Los algoritmos propuestos, llamados análisis de componentes principales por líneas LPCA, fueron desarrollados al cambiar la búsqueda del óptimo usando búsqueda en línea de allí, su nombre.

La teoría que se presenta del método de búsqueda en línea en espacio Euclídeos a lo largo de la siguiente sección es tomada de [85] y sus extensiones a espacios de Banach es inmediata. Posteriormente, se hace una descripción detallada de los algoritmos propuestos y en la última subsección se realiza la extensión de los métodos de búsqueda en líneas a espacios de Banach. En el apéndice A se encuentra una fundamentación teórica y la extensión de los métodos de optimización de búsqueda en líneas complementaria para funciones con dominios en subconjuntos de espacios de Banach.

3.4.1. Métodos de búsqueda en líneas en espacios euclideos

a. Fundamentos para los métodos de búsqueda en líneas

Sea la función objetivo a minimizar $f(x)$, $f : \Re^n \to \Re$ continuamente diferenciable. Los métodos de búsqueda para hallar los mínimos o máximos presentan la forma general:

$$x_{k+1} = x_k + \alpha_k d_k, \tag{3.14}$$

donde x_k es el punto de búsqueda, d_k es la dirección de búsqueda y α_k es el tamaño del paso en la iteración k. La dirección de búsqueda debe satisfacer la relación dada en:

$$\nabla f(x_k)^T d_k < 0, \tag{3.15}$$

la cual garantiza el movimiento en dirección descendente de $f(x)$ desde el punto x_k. Las siguientes hipótesis son necesarias para garantizar que los métodos de línea lleguen al mínimo (Análogamente como un problema de dualidad llegar el máximo).

Hipótesis 1. La función f debe poseer cotas inferiores sobre el conjunto $\Gamma = \{x \in \Re : f(x) \leq f(x_0)\}$, para x_0 un punto inicial.

Hipótesis 2. El gradiente $\nabla f(x)$ es continuo y Lipschitz en un conjunto abierto y convexo B, el cual contiene a Γ , esto es: existe $L > 0$, tal que:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \forall x, y \in B.$$

El objetivo principal de los métodos de búsqueda en línea es hallar el tamaño de paso α_k. A continuación se presentan diferentes métodos de búsqueda:

(a) *Minimización.* Para cada iteración, α_k es seleccionado tal que

$$f(x_k + \alpha_k d_k) = \min_{\alpha > 0} f(x_k + \alpha_k d_k). \tag{3.16}$$

(b) *Minimización Aproximada.* Para cada iteración, α_k es seleccionado tal que

$$\alpha_k = \min \{\alpha : \langle g(x_k + \alpha d_k), d_k \rangle = 0, \alpha > 0\}. \tag{3.17}$$

(c) *Armijo.* Sean escalares s_k, β y σ con $s_k = -\frac{\langle g_k, d_k \rangle}{\|d_k\|^2}$, $\beta \in (0,1)$ y $\sigma \in \left(0, \frac{1}{2}\right)$. Sea $\alpha_k = \beta^{m_k}.s_k$, donde m_k es el primer entero no negativo m para el cual

$$f_k - f(x_k + \beta^m s_k d_k) \geq -\sigma \beta^m s_k \langle g_k, d_k \rangle. \tag{3.18}$$

i,e; $m = 0, 1, ...$ sucesivamente hasta que la desigualdad anterior se satisfaga para $m = m_k$.

(d) *Minimización límitada.* Sea $s_k = -\frac{\langle g_k, d_k \rangle}{\|d_k\|^2}$, tal que

$$f(x_k + \alpha_k d_k) = \min_{\alpha \in [0, s_k]} f(x_k + \alpha_k d_k). \tag{3.19}$$

(e) *Goldstein.* Un escalar fijo $\sigma \in \left(0, \frac{1}{2}\right)$ es seleccionado, y α_k es escogido tal que

$$\sigma \leq \frac{f(x_k + \alpha_k d_k) - f_k}{\alpha_k \langle g_k, d_k \rangle} \leq 1 - \sigma. \tag{3.20}$$

Es posible demostrar que si f es acotada, existe un intervalo de tamaño de paso α_k que cumpla la relación anterior, y los α_k son bastante sencillos de encontrar a través de un número finito de operaciones aritméticas.

(f) *Wolfe fuerte.* α_k es escogido para que satisfaga simultáneamente,

$$f_k - f(x_k + \alpha_k d_k) \geq -\sigma \alpha_k \langle g_k, d_k \rangle, \tag{3.21}$$

y

$$|\langle g(x_k + \alpha_k d_k), d_k \rangle| \leq -\beta \langle g_k, d_k \rangle, \tag{3.22}$$

donde σ y β son escalares tal que $\sigma \in \left(0, \frac{1}{2}\right)$ y $\beta \in (\sigma, 1)$.

(q) *Wolfe.* α_k es escogido para que satisfaga la Ecuación 3.22 y

$$\langle g(x_k + \alpha_k d_k), d_k \rangle \geq \beta \langle g_k, d_k \rangle. \tag{3.23}$$

Teorema 2 *Si $f(x)$ satisface las dos hipótesis anteriores y $\nabla f(x_k)^T d_k < 0$ y si además, cada uno de los siete métodos en línea generan una sucesión infinita x_k, entonces: $\lim_{k \to \infty} \left(-\nabla f(x_k)^T d_k / \|d_k\|\right)^2 = 0$.*

Definición 1 *Sea $\{x_k\}$ una sucesión generada por un método de línea $x_{k+1} = x_k + \alpha_k d_k$. Se dice que d_k es uniformemente gradiente relacionada con x_k, si para cada subsucesión convergente $\{x_{k_i}\}_{i \in K}$, para la cual $\lim_{k_i \to \infty, k_i \in K} \nabla f(x_k) \neq 0$, se tiene también que: $\lim_{k_i \to \infty, k_i \in K} \nabla f(x_k)^T d_k > 0$ y $\limsup_{k_i \to \infty, k_i \in K} |d_k| < \infty$.*

Lema 2 *Suponga que $f(x)$ es acotada inferiormente y sea $\{x_k\}$ generada por algún método de líneas $x_{k+1} = x_k + \alpha_k d_k$, y suponga que $\{d_k\}$ es uniformemente gradiente relacionada con $\{\alpha_k\}$, entonces $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$.*

La prueba de este lema es consecuencia directa del Teorema 2. Una consecuencia del anterior lema es que si d_k satisface $-\nabla f(x_k)^T d_k \geq c_1 \|\nabla f(x_k)\|^2$ y $\|d_k\| \leq c_2 \max_{0\leq j\leq k} \|\nabla f(x_k)\| = 0$, donde existen $c_1, c_2 > 0$ tal que $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$. Sin embargo garantizar la existencia de $c_1, c_2 > 0$ no es directa, por esta razón, el trabajo se simplifica si se usa la siguiente caracterización.

Hipótesis 3. Sea $f(x)$ continuamente diferenciable y uniformemente convexa en $\Re$.

Lema 3 *Si $f(x)$ satisface la Hipótesis 3, entonces también satisface la Hipótesis 1 y la Hipótesis 2. Además $f(x)$ posee un único punto mínimo $x*$, y existen $m, M \in \Re, 0 < m \leq M$, tal que: $m\|y\|^2 \leq y^T \nabla^2 f(x) y \leq M\|y\|^2, \forall x, y \in \Re^n$, entonces $\frac{m}{2}\|x - x*\|^2 \leq f(x) - f(x*) \leq \frac{M}{2}\|x - x*\|^2, \forall x \in \Re^n$, tomando la derivada se llega a:*

$$m\|x-y\|^2 \leq \left(\nabla^2 f(x) - \nabla^2 f(x*)\right)^2 (x-y) \leq M\|x-y\|^2, \forall x, y \in \Re^n.$$

Usando el teorema del valor medio y la desigualdad de Cauchy-Schwartz, la condición que garantiza la dirección descendente puede ser sustituida por:

Condición 1: Para $0 < \tau \leq 1$, $\frac{\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\|\|d_k\|} > \tau$.

En el siguiente teorema se obtiene la base teórica para garantizar la convergencia del método propuesto.

Teorema 3 *Si se satisfacen las tres hipótesis precedentes y la **Condición 1**, y cualquiera de los métodos de búsqueda en línea genera una sucesión infinita $\{x_k\}$, entonces: $\lim_{k\to\infty} \left\|\nabla f(x_k)^T\right\| = 0$. Además, $\{x_k\}$ converge a $x*$ por lo menos linealmente.*

b. Algoritmo de búsqueda en línea

Con los desarrollos previos, se plantean los algoritmos por métodos de búsqueda en línea de la siguiente manera:

i. Elegir un punto inicial x_1 y hacer $k = 0$.

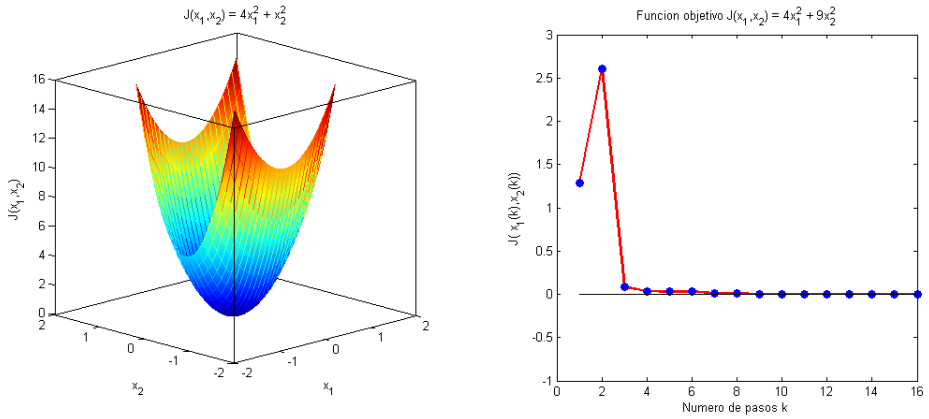
ii. Si $\|\nabla f(x_k)\| = 0$, terminar.

iii. Calcular $x_{k+1} = x_k + \alpha_k d_k$, usando α_k determinado por cualquiera de los 7 métodos de búsqueda citados anteriormente.

iv. Seleccionar aleatoriamente d_k que satisfaga la Condición 1. Hacer $k = k+1$ y regresar al literal i.

A continuación, con un ejemplo se pretende aclarar el proceso de optimización del método de búsqueda en líneas. Considere el paraboloide definido por (Figura 3.1(a)):

$J_1(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 4x_1^2 + 9x_2^2$. Del cálculo clásico se conoce su valor mínimo $J_1(x_1, x_2) = 0$, el cual se obtiene cuando $(x_1, x_2) = (0, 0)$. En la Figura 3.1(b), se muestran la función objetivo y el comportamiento del proceso de búsqueda al cabo de 16 pasos con un umbral de error de 0,001. El proceso se inicia aleatoriamente en $\vec{x}_0 = \begin{bmatrix} 0,3055 & -0,1559 \end{bmatrix}$ y termina en el paso $k = 16$ en $\vec{x}_{16} = \begin{bmatrix} 0,0001 & -0,0015 \end{bmatrix}$ con un valor de $J(\vec{x}_{16}) = 0,00002$.



(a) Paraboloide $J(x_1, x_2) = 4x_1 + x_2$ (b) Convergencia

Figura 3.1: Método de convergencia en líneas aplicado a un paraboloide

3.4.2. Búsqueda en líneas en espacios de Banach

Los métodos de búsqueda en líneas se introdujeron en la década de los 80 para alcanzar los extremos de una función, teniendo un amplio uso en la ingeniería y en la matemática. Sin embargo, estos métodos se emplean en funciones de dominio en un subconjunto de espacios euclideos. En[85], se proponen varios métodos de búsqueda relacionados con la forma como se halla el tamaño del paso realizando las pruebas de su convergencia.

Generalización de la búsqueda en líneas. Problema de minimización

$$\text{mín } f(x); \qquad x \in \mathbb{E}, \tag{3.24}$$

donde $\mathbb{E}$ es un espacio de Banach, $f : \mathbb{E} \to \mathbb{R}$ función diferenciable y continua en $\mathbb{E}$. El método de línea para resolver 3.24 está dado por:

$$x_{k+1} = x_k + \alpha_k d_k, \tag{3.25}$$

donde x_k es un punto iterativo, d_k la búsqueda direccional y α_k entero positivo llamado el tamaño del paso. Denotamos $\nabla f(x_k)$ por g_k, $f(x_k)$ por f_k, $f(x*)$ por $f*$, respectivamente. Sea $x*$ el minimizador de 3.24, así $g(x*) = 0 \in \mathbb{E}$. La búsqueda de la dirección d_k requiere que el vector satisfaga la condición descendente

$$\langle g_k, d_k \rangle < 0, \tag{3.26}$$

lo cual garantiza que d_k siga una dirección similar a la del gradiente negativo de $f(x)$ en x_k. Pruebas relacionadas con teoremas extendidos a espacios de Banach y la optimización del método de reducción LPCA1, se presentan en el **apéndice A.**

3.4.3. Algoritmos LPCA1, LPCA2

c. Extensión del algoritmo de búsqueda en línea a los métodos propuestos

En los métodos de reducción de dimensionalidad WPCA y WRDA su principales inconvenientes son la estabilidad y la convergencia hacia el valor óptimo. Para obviar este inconveniente, se propone cambiar el proceso de optimización a métodos de búsqueda en líneas sobre la función objetivo J_4, llegando así a los métodos propuestos: LPCA1 y LPCA2. Es importante destacar que los métodos

de optimización en líneas usan funciones de dominio en subespacios de $\mathbb{R}^n$ y codominio $\mathbb{R}$; sin embargo, el aporte de la generalización propuesta en esta investigación es que la función objetivo propuesta tiene dominio el espacio de las matrices semidefinidas positivamente y su recorrido es $\mathbb{R}$.

Sean A, B matrices cuadradas semidefinidas positivamente y simétricas de tamaño $d \times d$ que representaran las matrices de varianza y covarianza entre clases e intra clases respectivamente, sea x matriz de tamaño $m \times n$, que representara el subespacio óptimo al que queremos proyectar los datos. Al considerar en la función objetivo $J_4 = \frac{traza\big((DW)^T A(DW)\big)}{traza((DW)^T B(DW))}$ proveniente de $WPCA\ x = DW$, minimizar la función objetivo J_4 equivale dejar fijo el denominador y minimizar el numerador por ser J_4 un cociente. Esta optimización se traduce también a problema de maximización, si dejamos fijo el numerador y maximizamos el denominador. Por lo tanto, nuestro problema de optimización entonces equivale a minimizar la función dada en la siguiente definición.

Definición 2 *Maximizar la función*

$$J(x) = traza\left(x^T B x\right) \tag{3.27}$$

sujeta a: $traza\left(x^T A x\right) = 1$ *y* $x^T x = Id_{n\times n}$.

Definición 3 *Minimizar la función*

$$J(x) = traza\left(x^T A x\right) \tag{3.28}$$

sujeta a: $traza\left(x^T B x\right) = 1$ *y* $x^T x = Id_{n\times n}$.

Teorema 4 *Cuando B es simétrica y semidefinida positivamente, $J(x)$ satisface la Hipótesis 3, sobre el espacio de matrices de tamaño $m \times n$, tomando la norma $\|\cdot\|_B^2$ y el producto interior $\langle ., . \rangle_B$.*

Prueba 4 (Demostración:) *$J(x)$ es continuamente diferenciable puesto que su derivada es*

$$\frac{d}{dx} J(x) = (B + B^T)x.$$

$J(x)$ es uniformemente convexa, ya que existe $0 < c \leq 2$ tal que

$$J\left(\alpha x + (1-\alpha)y\right) \leq \alpha J(x) + (1-\alpha)J(y) - \frac{c}{2}\alpha(1-\alpha)\left\|x-y\right\|_B^2.$$

Esto último se debe a la distributibidad del $\langle ., . \rangle_A$.

d. Algoritmo de búsqueda en línea LPCA1

i. Normalice cada vector de características para obtener media cero y norma Euclídea uno ($||x||_2 = 1$).

ii. Inicie con algún conjunto ortonormal de vectores $X = U^0$.

iii. Calcule $A = \Sigma_B$ y $B = \Sigma_W$.

iv. Inicie el proceso de búsqueda en líneas con la función objetivo J en la Ecuación (16), para hallar X.

e. Algoritmo de búsqueda en línea LPCA2

i. Normalice cada vector de características para obtener media cero y norma euclídea uno ($||x||_2 = 1$).

ii. Inicie con algún conjunto ortonormal de vectores $X = U^0$.

iii. Calcule $A = \Sigma_B$ y $B = \Sigma_W$.

iv. Inicie el proceso de búsqueda en líneas con la función objetivo J en la Ecuación (16), para hallar X.

v. $X = D * X$, donde $D = diag(X * X^T)$.

La convergencia de los métodos LPCA1 y LPCA2 se obtiene por dos principales razones:

a. La función objetivo es uniformemente convexa.

b. Los métodos usan un parámetro de tamaño de salto el cual justamente se reduce conforme aumenta la iteración de acuerdo con la pruebas en [85].

3.5. Aplicación de los métodos de reducción ponderados

Diversos ejemplos de los diferentes métodos de reducción ponderados sobre varios tipos de datos, que muestran la potencia de los métodos propuestos son presentados en esta sección. Utilizando un clasificador máquina de vectores de soporte (MSV) y evaluando su desempeño a través de dos procesos: usando curvas ROC o hipersuperficies ROC y empleando el error de clasificación, se estudia el comportamiento de las técnicas de reducción de dimensionalidad PCA, PPCA, WPCA, WRDA y las propuestas LPCA1 y LPCA2. Se utilizaron datos artificiales n-dimensionales, con clases solapadas (Figura 3.2(a)) y clases esféricas concéntricas (Figura 3.3(b)).

También se emplearon datos reales generados a través de características geométricas tales como: áreas, perímetros, orientaciones, dispersión, centroides, y distintos momentos estadísticos obteniendo 70 características aplicadas a imágenes capilares (Figura 3.4) e imágenes de peatones (Figura 3.5). En las Figuras 3.2 (a), 3.3 (a), 3.4 (a) y 3.5 (a) se muestran las clases diferenciadas por colores. Las Figuras 3.2 (b), 3.3 (b), 3.4 (b) y 3.5 (b) presentan el desempeño frente a los diferentes métodos de reducción tratados (WRDA, PCA, PPCA, WPCA LPCA1 y LPCA2). El pico más alto corresponde al método LPCA1 con mejor desempeño.

Por último, las Figuras 3.2 (c), 3.3 (c), 3.4 (c) y 3.5 (c), muestran el error de clasificación; el pico más bajo lo obtuvo el método LPCA1. Se observó que el método propuesto LPCA1 presenta el mejor comportamiento poque obtuvo valores mayores en las curvas de desempeño y valores menores en las curvas de error.

Los resultados fueron sintetizados en la Tabla 3.1; allí se puede observar que el método propuesto LPCA1 fue mejor que los otros métodos.

La evaluación de los métodos de reducción, se obtiene desde dos puntos:

a. Usando el error de clasificación curva azul, ver Figura 3.5 (c) .

b. Usando el hipervolumen asociado con las curvas ROC generadas por la clasificación. Ver curva roja en la ver Figura 3.5 (b). El calculo del hipervolumen se hizo por Montecarlo, por limitaciones de memoria al usar Matlab.