4.3.3: Continuación del Método ICA

La matriz de covarianza: Sea M una matriz $n \times l$, tomamos el promedio de cada columna x_j de M notado por $\overline{x_j}$; con estos promedios formamos un vector que denotamos por $\overline{X}$. Se calcula cada entrada de la matriz de covarianza Q de M como:

$$q_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ki} - \overline{x_k})$$

donde x_{ij} es cada una de las entradas de M, y n es el número de componentes en cada columna. En términos de cada columna X_j de M y del vector promedio de los datos, podemos escribir la matriz de covarianza Q como:

$$Q = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

El resultado es una matriz Q cuadrada $n \times n$ simétrica que contiene sobre su diagonal la varianza de cada una de las columnas de la matriz y en las entradas q_{ij} contiene la covarianza entre la columna i y la columna j. El valor de covarianza entre dos variables, nos da un parámetro de correlación entre éstas; si la covarianza es mayor que 0 entonces se dice que las variables están directamente correlacionadas, esto quiere decir que si una de las variables aumenta, la otra también lo hace. Si la covarianza es menor que 0 entonces se dice que las variables están inversamente correlacionadas, esto quiere decir que si una de las variables aumenta, la otra disminuye. Si la covarianza es cero, las dos variables no tienen una relación lineal.

Para nuestro caso, el cálculo de la matriz de covarianza se reduce a:

$$Q = Cov(X) = \frac{XX^T}{n}$$

porque se trabaja con datos centrados, por lo tanto el vector promedio de cada señal es nulo. Realizamos ahora la descomposición espectral de la matriz de covarianza para obtener $Q = EDE^T$, eliminamos la covarianza entre las señales mezcladas calculando la inversa de la raíz cuadrada de la matriz de covarianza,

$$\widetilde{Q} = Q^{-1/2} = ED^{-1/2}E^T$$

Al multiplicar nuestro vector de señales mezcladas por la matriz $D^{-\frac{1}{2}}E^T$, es decir $\widetilde{X} = D^{-\frac{1}{2}}E^TX$ eliminamos la covarianza entre cada par de señales que tenemos; si calculamos la matriz de covarianza de los datos blanqueados obtenemos una matriz diagonal:

$$Cov(\widetilde{X}) = \frac{\widetilde{X}\widetilde{X}^T}{n}$$

$$Cov(\widetilde{X}) = \frac{D^{-\frac{1}{2}}E^TXX^TED^{-\frac{1}{2}}}{n}$$

$$
\begin{aligned}
&= D^{-\frac{1}{2}} E^T \frac{XX^T}{n} E D^{-\frac{1}{2}} \\
&= D^{-\frac{1}{2}} E^T Cov(X^T) E D^{-\frac{1}{2}} \\
&= D^{-\frac{1}{2}} E^T E D E^T E D^{-\frac{1}{2}} \\
&= I_n
\end{aligned}
$$

Sabemos que la covarianza es una medida de correlación, con lo anterior se garantiza que cada una de las señales observadas está no correlacionada con las demás. Aplicando el proceso de blanqueo a los datos de la $figura1 - 9$ obtenemos el siguiente resultado:

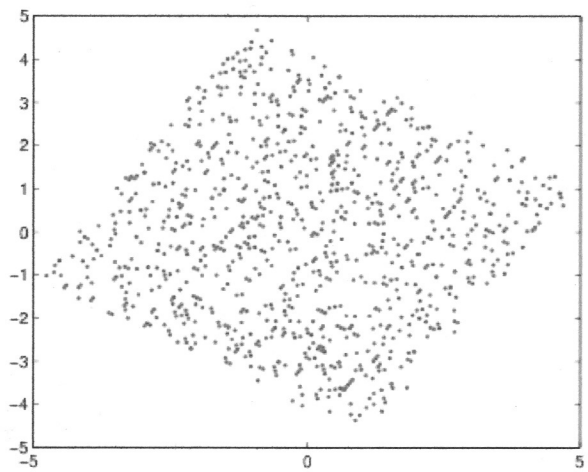


Figura 1-12: Datos blanqueados.

1.4.2. Teorema del límite central

[14]Como mencionamos anteriormente, la idea principal para el método ICA es la búsqueda de independencia estadística entre las variables, el siguiente teorema nos da una relación entre la medida de gaussianidad para variables aleatorias y la independencia estadística.

Teorema del Limite Central: Sea $x_1, ..., x_n$ una muestra aleatoria de tamaño n, es decir, una secuencia de variables aleatorias independientes e idénticamente distribuidas con valores esperados dados por μ y varianzas finitas dadas por σ^2. Supongamos que estamos interesados en el promedio de la muestra

$$S_n = \frac{(x_1 + ... + x_n)}{n}$$

de estas variables aleatorias. Por la ley de los grandes números, los promedios de la muestra convergen en probabilidad y casi seguramente al valor esperado μ cuando n tiende a infinito.

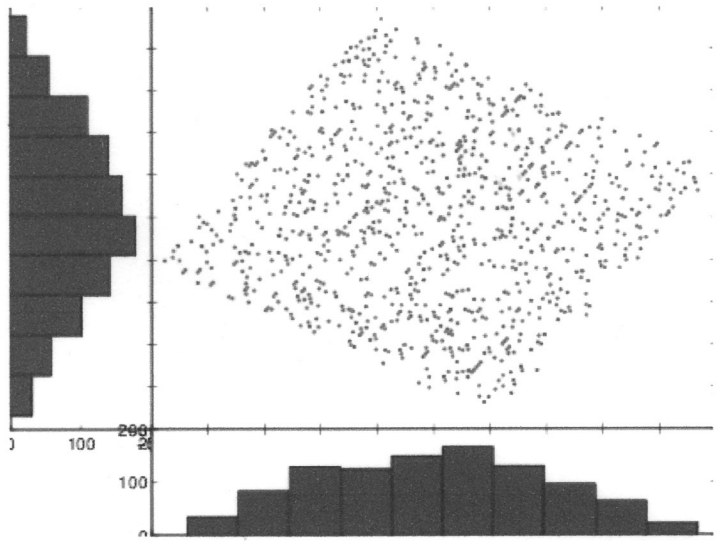


Figura 1-13: Funciones de densidad de las señales mezcladas.

El teorema del límite central afirma que a medida que n se hace más grande, la distribución de la diferencia entre la media de la muestra S_n y su límite μ, cuando n es muy grande, se aproxima a la distribución normal con media 0 y varianza σ^2.
Por el teorema del limite central, nuestras señales mezcladas $\mathbf{x}$ siempre van a ser mas gaussianas que las señales independientes originales $\mathbf{s}$, la *figura 1-13* muestra las funciones de densidad de probabilidad para los datos de la *figura 1-12*, podemos observar la forma de campana gaussiana que tienen estas funciones de densidad *(figura 1-14)*.
El trabajo del método consistirá entonces en tratar de reducir la gaussianidad que presentan las señales mezcladas para encontrar las señales independientes, geométricamente, es casi como "girar"los datos hacia los ejes que determinan las componentes independientes y de esta forma minimizar la gaussianidad entre las señales.

1.4.3. Negentropia y medida de Nogaussianidad

¿Cómo logramos entonces reducir la gaussianidad?, la medida de Gaussianidad está relacionada con la independencia estadística entre las variables, entre mayor sea la Gaussianidad [3], mayor información comparten las variables y son menos independientes entre si. Definamos primero el concepto de Entropía H para un vector aleatorio $\mathbf{y} = [y_1, ..., y_n]$:

$$H(\mathbf{y}) = -\int f(\mathbf{y}) log f(\mathbf{y}) d\mathbf{y}$$

La entropía se puede considerar como una medida de cantidad de información mutua que contienen los elementos del vector. En base a la entropía H podemos definir la función de negentropía J como:

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y})$$

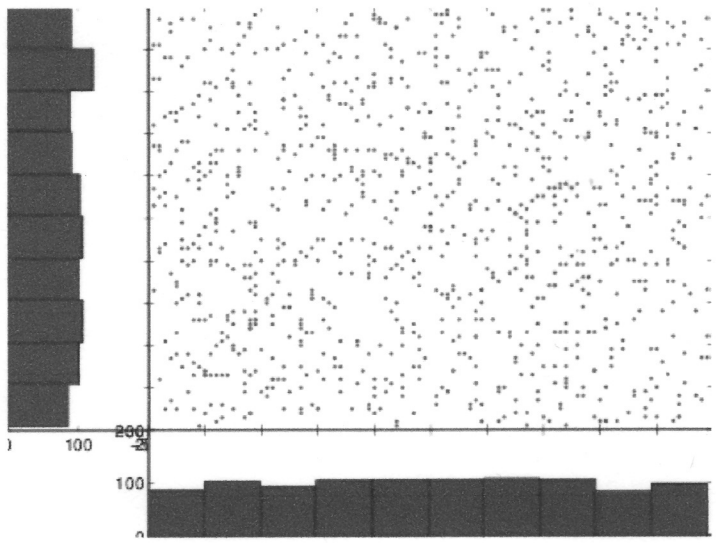


Figura 1-14: Funciones de densidad de las señales.

donde $\mathbf{y}_{gauss}$ es un vector aleatorio Gaussiano con la misma matriz de covarianza de $\mathbf{y}$. Si la entropía mide la cantidad de información mutua en el vector, la negentropía determina la cantidad de información no mutua entre los elementos del vector, es decir a mayor negentropía, menor es la Gaussianidad de los elementos del vector. La negèntropía la podemos considerar como una medida de Nogaussianidad.

Dado que la esperanza de una variable aleatoria se define como: $E(X) = \int xf(x)dx$, podemos usar una aproximación de la negentropía a través de la esperanza del vector aleatorio de la siguiente forma[3]:

$$J(y_i) \approx c[E\{G(y_i)\} - E\{G(V)\}]$$

donde G es una función llamada función de contraste, c es una constante irrelevante y v es una variable Gaussiana con media cero y varianza uno. Esta aproximación de la función de negentropía nos ayuda a definir el objetivo del método ICA, si hacemos $y_i = w^T x$, donde x es una de las componentes mezcladas y w^T es el vector de desmezclado para este x, nuestra meta será maximizar la función J dada por:

$$J(w) = E\{G(w^T x)\} - E\{G(v)\}$$

donde w es el vector $m - dimensional$ de mezcla de forma que $E\{(w^T x)^2\} = 1$. Esta ultima condición es simplemente de escala, para que los vectores no crezcan arbitrariamente.

Ahora podemos generalizar la función anterior para toda la matriz de desmezclado $\mathbf{W}$, teniendo en cuenta que se maximizan cada una de las negentropías cuando la suma de todas

ellas es máxima. El problema final de maximización será el siguiente:

$$max \sum_{i=1}^{m} J(w_i)$$

$$s.a. \quad E[(w_k^T x)(w_j^T x)] = \delta_{jk}$$

La restricción del problema de maximización garantiza que cada una de las componentes de desmezclado w_j no esté correlacionada con las demás. Dos vectores aleatorios y_i, y_j están no correlacionados si $E[(y_i)(y_j)] - E(y_i)E(y_j) = 0$; Dado el preproceso de blanqueo de los datos, la media de cada vector es cero, por consiguiente $E(y_i) = 0$ y la condición de no correlación se reduce a

$$E[(y_i)(y_j)] = \delta_{jk}$$

donde

$$\delta_{jk} = \begin{cases} 1 & si \quad j = k \\ 0 & si \quad j \neq k \end{cases}$$

1.4.4. Algoritmo en una dimensión

Observemos que el óptimo de $J(w)$ se obtiene en un óptimo de $E\{G(w^T x)\}$. Aplicando el metodo de los multiplicadores de Lagrange, el óptimo de $E\{G(w^T x)\}$ bajo las condiciones $E\{G(w^T x)^2\} = 1$, $\|w\|^2 = 1$ se obtiene cuando

$$\nabla(E\{G(w^T x)\}) = \beta \nabla(\|w\|^2)$$

$$E\{xg(w^T x)\} - \beta w = 0$$

donde $g = G'$, β será entonces:

$$\beta = w_0^T E\{xg(w_0^T x)\}$$

$$\beta = E\{w_0^T xg(w_0^T x)\}$$

con w_0 el valor óptimo de w.
Para solucionar la ecuación $E\{xg(w^T x)\} - \beta w = 0$ y encontrar el optimo W_0, aplicaremos el Método de Newton, si notamos el lado izquierdo de la ecuación anterior como F, obtenemos el Jacobiano JF como

$$JF(w) = E\{xx^T g'(w^T x)\} - \beta I$$

Para simplificar la expresión, aproximamos el término $E\{xx^T g'(w^T x)\}$, como $E\{xx^T\}E\{g'(w^T x)\} \simeq E\{g'(w^T x)\}I$, esto debido a que se garantiza independencía estadística entre los datos. Ahora la matriz Jacobiana es diagonal y puede representarse como:

$$JF(w) = E\{g'(w^T x)\}I - \beta I$$

NOTA: Para mayor profundidad en el tema, se tiene el libro de Aapo Hyvarien.

$$= (E\{g'(w^T x)\} - \beta)I$$

lo anterior permite evitar el calculo de la inversa de la matriz Jacobiana al aplicar el método de Newton, porque la inversa de JF se reduce a:

$$JF(w) = \frac{1}{E\{g'(w^T x)\} - \beta} I$$

Aplicando el método de Newton se obtiene:

$$w^+ = w - [E\{xg(w^T x)\} - \beta w]/[E\{g'(w^T x)\} - \beta]$$

$$w^* = w^+/\|w^+\|$$

donde w^* denota el nuevo valor de w, podemos simplificar el algoritmo multiplicando la ecuación anterior por $\beta - E\{g'(w^T x)\}$.

$$w^+ = E\{xg(w^T x)\} - E\{g'(w^T x)\}w$$

$$w^* = w^+/\|w^+\|$$

1.4.5. Algoritmo en varias dimensiones

El algoritmo anterior, puede ser usado para construir un sistema de m vectores para m señales, así se obtienen los vectores de desmezclado $w_1, ..., w_n$. Para ello se aplica el algoritmo anterior encontrando cada componente independiente una por una. Si ya se han encontrado p componentes independientes, podemos encontrar el $p+1$ vector de mezcla de la siguiente forma:

1. Partiendo de un vector aleatorio w_{p+1} se realiza la resta de las proyecciones de los vectores $w_1, w_2, ..., w_p$ sobre el nuevo vector de la forma:

$$w_{p+1} = w_{p+1} - \sum_{j=1}^{p} w_{p+1}^T C w_j w_j$$

$$w_{p+1} = w_{p+1}/\sqrt{w_{p+1}^T C w_{p+1}}$$

   donde C es la matriz de correlación.

2. El procedimiento anterior garantiza que el nuevo vector aleatorio que se usará no tenga relación con los vectores de desmezclado que ya se han encontrado. Con este nuevo vector se procede a aplicar nuevamente en algoritmo planteado en la sección anterior para encontrar una nueva componente independiente.

4.3.4

LDA- ANÁLISIS DISCRIMINANTES LINEALES

Funciones discriminantes

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + \mathbf{w}_0$$

$\mathbf{w}$: vector de pesos

w_0 : bias o threshold

asigno x a c_1 si $y(\mathbf{x}) > 0$ y c_2 en otro caso.

La superficie de decisión corresponde a un hiperplano de dimensión d - 1 si el espacio de entrada es de dimensión d.

LDA- Análisis Discriminante Lineal
Discriminante Lineal de Fisher

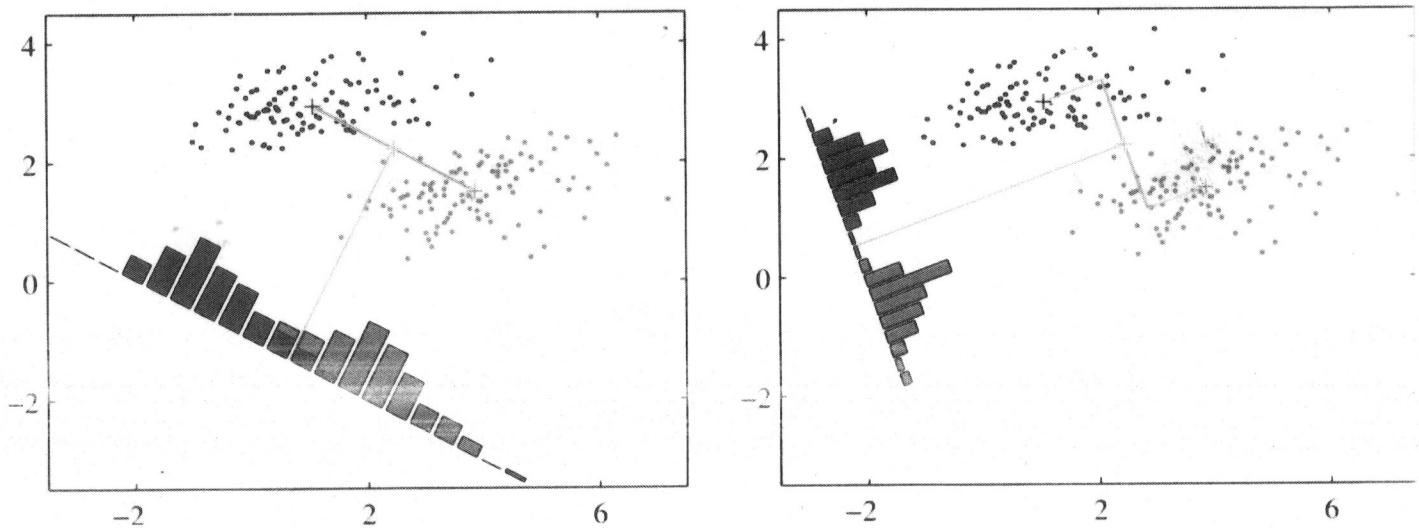
- Objetivo: Seleccionar una proyección que maximice separabilidad entre clases. Método supervisado.
- Ejemplo- 2 clases: c_1 y c_2 con N_1 y N_2 patrones.
- Medias:

$$\mathbf{m_1} = \frac{\sum_{n \in c_1} \mathbf{x_n}}{N_1} \qquad \mathbf{m_2} = \frac{\sum_{n \in c_2} \mathbf{x_n}}{N_2}$$

- Si consideramos la proyección sobre una dimensión, una medida de separabilidad es la separación de las medias luego de proyectar:

$$\boxed{y = \mathbf{w}^{\mathbf{T}}\mathbf{x} \qquad m_2 - m_1 = \mathbf{w}^{\mathbf{T}}(\mathbf{m_2} - \mathbf{m_1})}$$

- Restringiendo a los **w** con módulo 1 y usando multiplicadores de Lagrange: **w** proporcional $(\mathbf{m}_2 - \mathbf{m}_1)$.



- Si las matrices de covarianza fuertemene no lineal, no son diagonales puede existir solapamiento considerable luego de proyectar y a pesar de ser linealmente separables.
- Discriminante de Fisher busca maximizar separación de medias proyectadas, pequeña varianza dentro de cada clase, minimizar solapamiento entre clases.

Varianza intra- clase:

$$s_k^2 = \sum_{n \in c_k} (y_n - m_k)^2 \qquad y_n = \mathbf{w}^T \mathbf{x}_n$$

Criterio de Fisher:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

Varianza entre clases

Varianza intra clases

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

con $\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$

$$\mathbf{S}_w = \sum_{n \in c_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in c_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

Diferenciando respecto a $\mathbf{w}$: $J(\mathbf{w})$ máximo si:

$$(\mathbf{w}^T \mathbf{S}_B \mathbf{w}) \mathbf{S}_w \mathbf{w} = (\mathbf{w}^T \mathbf{S}_w \mathbf{w}) \mathbf{S}_B \mathbf{w}$$

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$$

- $\mathbf{S}_w$: simétrica y semidefinida positiva, usualmente no singular si n>p.

$$\boxed{\mathbf{S}_w^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}}$$

- En este caso particular no es necesario resolver el problema de vectores y valores propios ya que :

$$\mathbf{S}_B\mathbf{w} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$$

$$\boxed{\mathbf{w} \propto \mathbf{S}_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)}$$

- No importa la magnitud de **w**. Determino una dirección para la proyección de los datos proyectados.

4.3.5: MLDA

Análisis de Discriminantes Lineales Múltiples

- La generalización a un problema de C clases involucra c-1 funciones discriminantes. Se proyecta de un espacio de dimensión p a un espacio de dimensión c-1, asumiendo p≥c.
- Generalización de $\mathbf{S}_W$ es directa:

$$\mathbf{S}_w = \sum_{i=1}^{c} \mathbf{S}_i \qquad \mathbf{S}_i = \sum_{n \in c_i} (\mathbf{x}_n - \mathbf{m}_i)(\mathbf{x}_n - \mathbf{m}_i)^T \quad \mathbf{m}_i = \frac{1}{N_i} \sum_{n \in c_i} \mathbf{x}_n$$

- $\mathbf{S}_B$: consideramos media y covarianza global de los datos

$$\mathbf{m} = \frac{1}{N} \sum_j \mathbf{x}_j = \frac{1}{N} \sum_{i=1}^{c} N_i \mathbf{m}_i \qquad \mathbf{S}_T = \sum_j (\mathbf{x}_j - \mathbf{m})(\mathbf{x}_j - \mathbf{m})^T$$

LDA Múltiples Generalización

$$\mathbf{S}_T = \sum_{i=1}^{c} \sum_{x \in c_i} (\mathbf{x} - \mathbf{m_i} + \mathbf{m_i} - \mathbf{m})(\mathbf{x} - \mathbf{m_i} + \mathbf{m_i} - \mathbf{m})^T$$

$$= \sum_{i=1}^{c} \sum_{x \in c_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m_i})^T + \sum_{i=1}^{c} \sum_{x \in c_i} (\mathbf{m_i} - \mathbf{m})(\mathbf{m_i} - \mathbf{m})^T$$

$$\mathbf{S}_T = \mathbf{S}_w + \sum_{i=1}^{c} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T = \mathbf{S}_w + \mathbf{S}_B$$

LDA Múltiples Generalizado

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2 \ldots\ldots \mathbf{w}_{c-1}] \quad \mathbf{W} : c-1 \times p \quad \mathbf{y} = \mathbf{W}^T \mathbf{x}$$

$$\text{Def}: \quad \boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{n \in c_n} \mathbf{y}_n \qquad \boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{c} N_i \boldsymbol{\mu}_i$$

$$\tilde{\mathbf{S}}_{\mathbf{W}} = \sum_{i=1}^{c} \sum_{n \in c_n} (\mathbf{y} - \boldsymbol{\mu}_i)(\mathbf{y} - \boldsymbol{\mu}_i)^T \quad \tilde{\mathbf{S}}_B = \sum_{i=1}^{c} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

$$\tilde{\mathbf{S}}_B = \mathbf{W}^T \mathbf{S}_B \mathbf{W} \qquad \tilde{\mathbf{S}}_{\mathbf{W}} = \mathbf{W}^T \mathbf{S}_{\mathbf{W}} \mathbf{W}$$

- Necesitamos un medida de la razón entre la varianza entre clases y la varianza intra-clases. Una medida escalar simple es el determinante de la matriz otra la traza de las matrices.

$$J(\mathbf{w}) = \frac{\det(\tilde{\mathbf{S}}_B)}{\det(\tilde{\mathbf{S}}_\mathbf{W})} = \frac{\det(\mathbf{W}^T\mathbf{S}_B\mathbf{W})}{\det(\mathbf{W}^T\mathbf{S}_\mathbf{W}\mathbf{W})}$$

$$\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{S}_\mathbf{W}\mathbf{w}_i$$

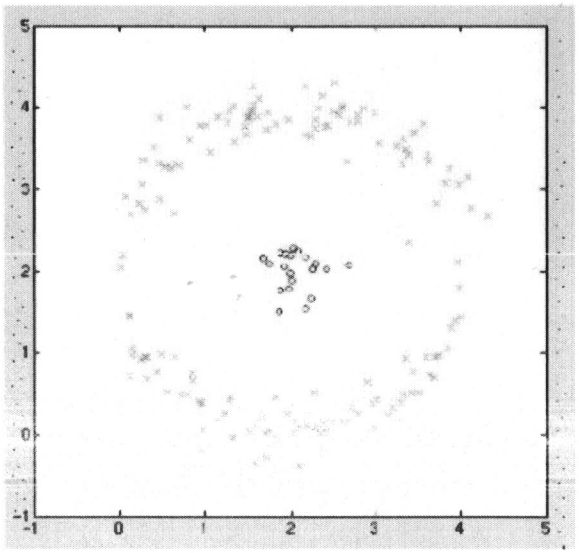
- Problema de vectores propios generalizado, las columnas de **W** son los vectores propios generalizados correspondientes a los valores propios mayores de:
- Si $\mathbf{S}_\mathbf{w}$ no es singular:

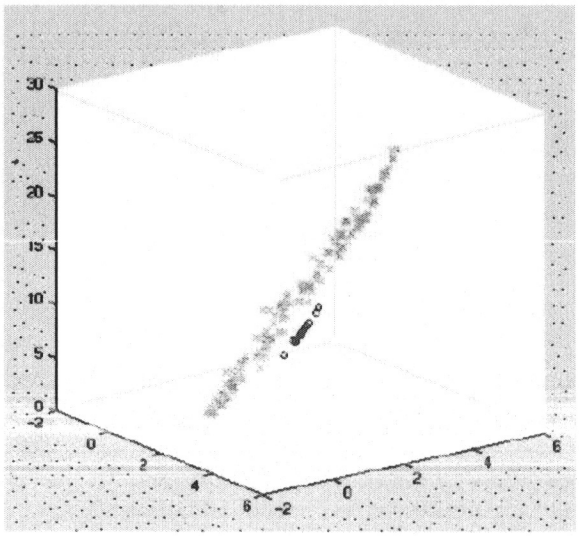
$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{w}_i$$

LDA Múltiples Generalizados

- Nota: $\mathbf{S}_B$ es la suma de c matrices que son el producto de dos vectores por ende de rango 1, solo (c-1) son independientes. Por lo tanto $\mathbf{S}_B$ es a lo sumo de rango c-1 y consecuentemente tengo a lo sumo c-1 vectores propios no nulos.
- Obs: Las clases no tienen porqué ser linealmente separables. Alternativa SVM: mapeo no lineal y discriminantes lineales.

La magia de la alta dimensión





$$\Phi : \mathbf{R}^2 \to \mathbf{R}^3$$

$$(x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2)$$

La teoría de VC (Vapnik-Chervonenkis) dice que a menudo un mapeo a un espacio de mayor dimensión tiene mayor poder de clasificación que el espacio de entrada