# Análisis discriminante lineal

## Idea intuitiva

El Análisis Discriminante Lineal o *Linear Discrimiant Analysis (LDA)* es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos *a priori* y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, *LDA* estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, $P(Y = k|X = x)$. Finalmente se asigna la observación a la clase *k* para la que la probabilidad predicha es mayor.

Es una alternativa a la regresión logística cuando la variable cualitativa tiene más de dos niveles. Si bien existen extensiones de la regresión logística para múltiples clases, el *LDA* presenta una serie de ventajas:

- Si las clases están bien separadas, los parámetros estimados en el modelo de regresión logística son inestables. El método de *LDA* no sufre este problema.
- Si el número de observaciones es bajo y la distribución de los predictores es aproximadamente normal en cada una de las clases, *LDA* es más estable que la regresión logística.

# Análisis discriminante lineal

Cuando se trata de un problema de clasificación con solo dos niveles, ambos métodos suelen llegar a resultados similares.

El proceso de un análisis discriminante puede resumirse en 6 pasos:

- Disponer de un conjunto de datos de entrenamiento (*training data*) en el que se conoce a que grupo pertenece cada observación.
- Calcular las probabilidades previas (*prior probabilities*): la proporción esperada de observaciones que pertenecen a cada grupo.
- Determinar si la varianza o matriz de covarianzas es homogénea en todos los grupos. De esto dependerá que se emplee *LDA* o *QDA*.
- Estimar los parámetros necesarios para las funciones de probabilidad condicional, verificando que se cumplen las condiciones para hacerlo.
- Calcular el resultado de la función discriminante. El resultado de esta determina a qué grupo se asigna cada observación.
- Utilizar validación cruzada (*cross-validation*) para estimar las probabilidades de clasificaciones erróneas.

# Análisis discriminante lineal

## Teorema de Bayes para clasificación

Considérense dos eventos $A$ y $B$, el teorema de Bayes establece que la probabilidad de que $B$ ocurra habiendo ocurrido $A$ $(B|A)$ es igual a la probabilidad de que $A$ y $B$ ocurran al mismo tiempo $(AB)$ dividida entre la probabilidad de que ocurra $A$.

$$P(B|A) = \frac{P(AB)}{P(A)}$$

Supóngase que se desea clasificar una nueva observación en una de las *K* clases de una variable cualitativa *Y*, siendo $K \geq 2$, a partir de un solo predictor *X*. Se dispone de las siguientes definiciones:

- Se define como *overall*, *prior probability* o probabilidad previa $(\pi_k)$ la probabilidad de que una observación aleatoria pertenezca a la clase *k*.
- Se define $f_k(X) \equiv P(X = x|Y = k)$ como la función de densidad de probabilidad condicional de *X* para una observación que pertenece a la clase *k*. Cuanto mayor sea $f_k(X)$ mayor la probabilidad de que una observación de la clase *k* adquiera una valor de $X \approx x$.
- Se define como *posterior probability* o probabilidad posterior $P(Y = k|X = x)$ la probabilidad de que una observación pertenezca a la clase *k* siendo *x* el valor del predictor.

# Análisis discriminante lineal

Aplicando del teorema de Bayes se pueden conocer la *posterior probability* para cada clase:

$$\mathrm{P}(\text{pertenecer a la clase k} \mid \text{valor x observado}) =$$

$$\frac{\mathrm{P}(\text{pertenecer a la clase k y observar x})}{\mathrm{P}(\text{observar x})}$$

Si se introducen los términos, definidos anteriormente, dentro la ecuación se obtiene:

$$P(Y=k|X=x) = \frac{\pi_k P(X=x|Y=k)}{\sum_{j=1}^{K} \pi_j P(X=x|Y=j)} = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$$

La clasificación con menor error (clasificación de Bayes) se consigue asignando la observación a aquel grupo que maximice la *posterior probability*. Dado que el denominador $\sum_{j=1}^{K} \pi_j f_j(x)$ es igual para todas las clases, la norma de clasificación es equivalente a decir que se asignará cada observación a aquel grupo para el que $\pi_k f_k(x)$ sea mayor.

Para que la clasificación basada en Bayes sea posible, se necesita conocer la probabilidad poblacional de que una observación cualquiera pertenezca a cada clase $(\pi_k)$ y la probabilidad poblacional de que una observación que pertenece a la clase *k* adquiera el valor *x* en el predictor $(f_k(X) \equiv P(X=x|Y=k))$. En la práctica, raramente se dispone de esta información, por lo que los parámetros tienen que ser estimados a partir de la muestra. Como consecuencia, el clasificador *LDA* obtenido se aproxima al clasificador de Bayes pero no es igual.

# Análisis discriminante lineal

## Estimación de $\pi_k$ y $f_k(X)$

La capacidad del *LDA* para clasificar correctamente las observaciones depende de cómo de buenas sean las estimaciones de $\pi_k$ y $f_k(X)$. Cuanto más cercanas al valor real, más se aproximará el clasificador *LDA* al clasificador de Bayes. En el caso de la *prior probability* $(\pi_k)$ la estimación suele ser sencilla, la probabilidad de que una observación cualquiera pertenezca a la clase *k* es igual al número de observaciones de esa clase entre el número total de observaciones $\hat{\pi}_k = \frac{n_k}{N}$.

La estimación de $f_k(X)$ no es tan directa y para conseguirla se requiere de ciertas asunciones. Si se considera que $f_k(X)$ se distribuye de forma normal en las *K* clases, entonces se puede estimar su valor a partir de la ecuación:

$$f_k(X) = P(Y = k|X = x) = \frac{1}{\sqrt{2\pi}\sigma_k} exp(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2)$$

Donde $\mu_k$ y $\sigma_k^2$ son la media y la varianza para la clase *k*.

# Análisis discriminante lineal

Si además se asume que la varianza es constante en todos los grupos $\sigma_1^2 = \sigma_2^2 \ldots = \sigma_k^2 = \sigma^2$, entonces, el sumatorio $\sum_{j=1}^{K} \pi_j f_j(x)$ se simplifica en gran medida permitiendo calcular la *posterior probability* según la ecuación:

$$P(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{1}{2\sigma^2}(x - \mu_k)^2)}{\sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{1}{2\sigma^2}(x - \mu_j)^2)}$$

Esta ecuación se simplifica aun más mediante una transformación logarítmica de sus dos términos:

$$\hat{\delta}_k(x) = log(P(Y = k|X = x)) = x \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

El término lineal en el nombre *Análisis discriminante lineal* se debe al hecho de que la función discriminatoria es lineal respecto de *X*.

# Análisis discriminante lineal

En la práctica, a pesar de tener una certeza considerable de que *X* se distribuye de forma normal dentro de cada clase, los valores $\mu_1 \dots \mu_k$, $\pi_1 \dots \pi_k$ y $\sigma^2$ se desconocen, por lo que tienen que ser estimados a partir de las observaciones. Las estimaciones empleadas en *LDA* son:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_1} x_i$$

$$\hat{\sigma}_k = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i:y_1} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{N}$$

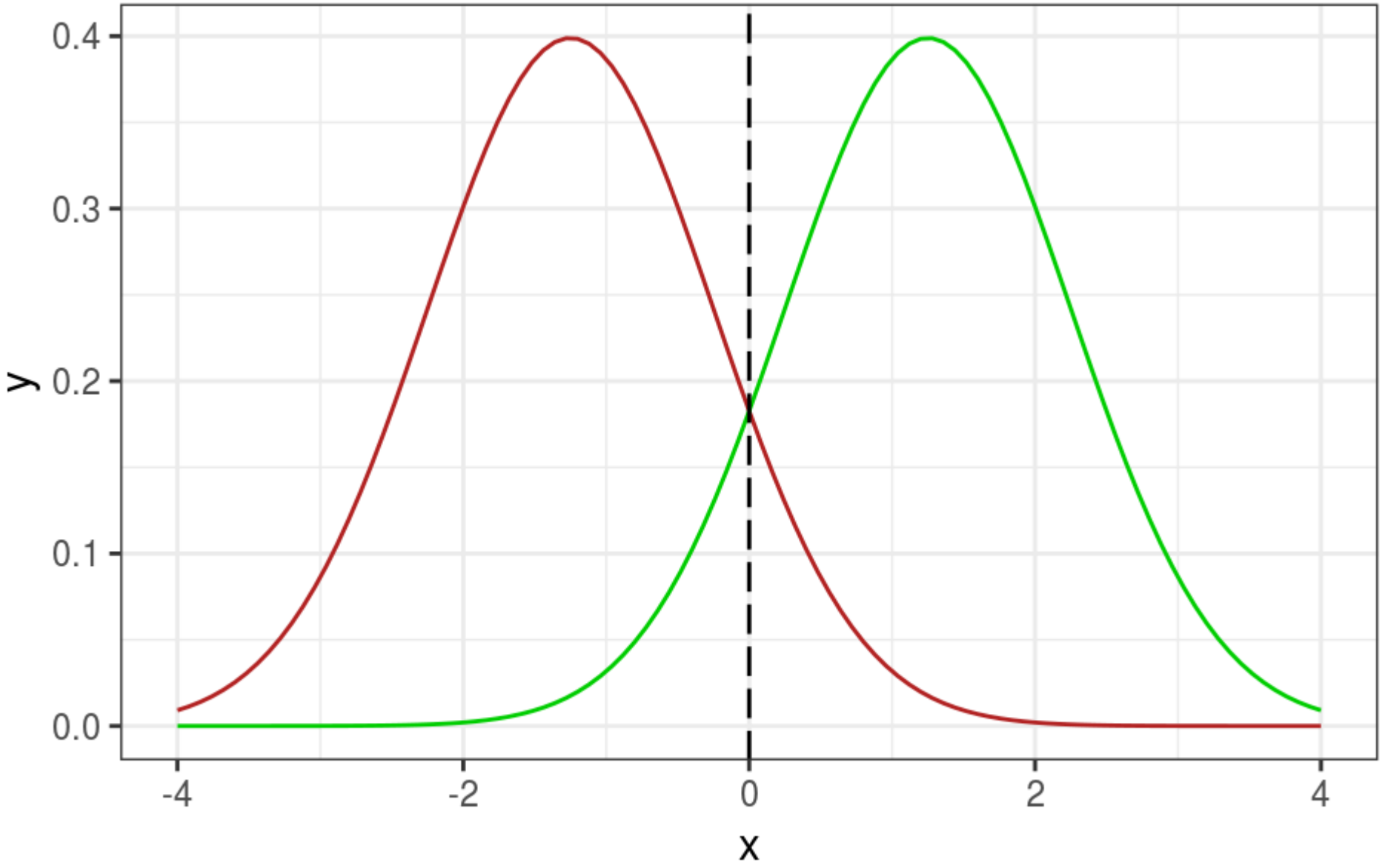
$\hat{\mu}_k$ es la media de las observaciones del grupo *k*, $\hat{\sigma}_k$ es la media ponderada de las varianzas muestrales de las *K* clases y $\hat{\pi}_k$ la proporción de observaciones de la clase *k* respecto al tamaño total de la muestra.

# Análisis discriminante lineal

La clasificación de Bayes consiste en asignar cada observación *X* = *x* a aquella clase para la que $P(Y = k|X = x)$ sea mayor. En el caso particular de una variable cualitativa *Y* con solo dos niveles, se puede expresar la regla de clasificación como un ratio entre las dos *posterior probabilities*. Se asignará la observación a la clase 1 si $\frac{P(Y=1|X=x)}{P(Y=2|X=x)}$>1, y a la clase 2 si es menor. En este caso particular el límite de decisión de Bayes viene dado por $x = \frac{\mu_1+\mu_2}{2}$.

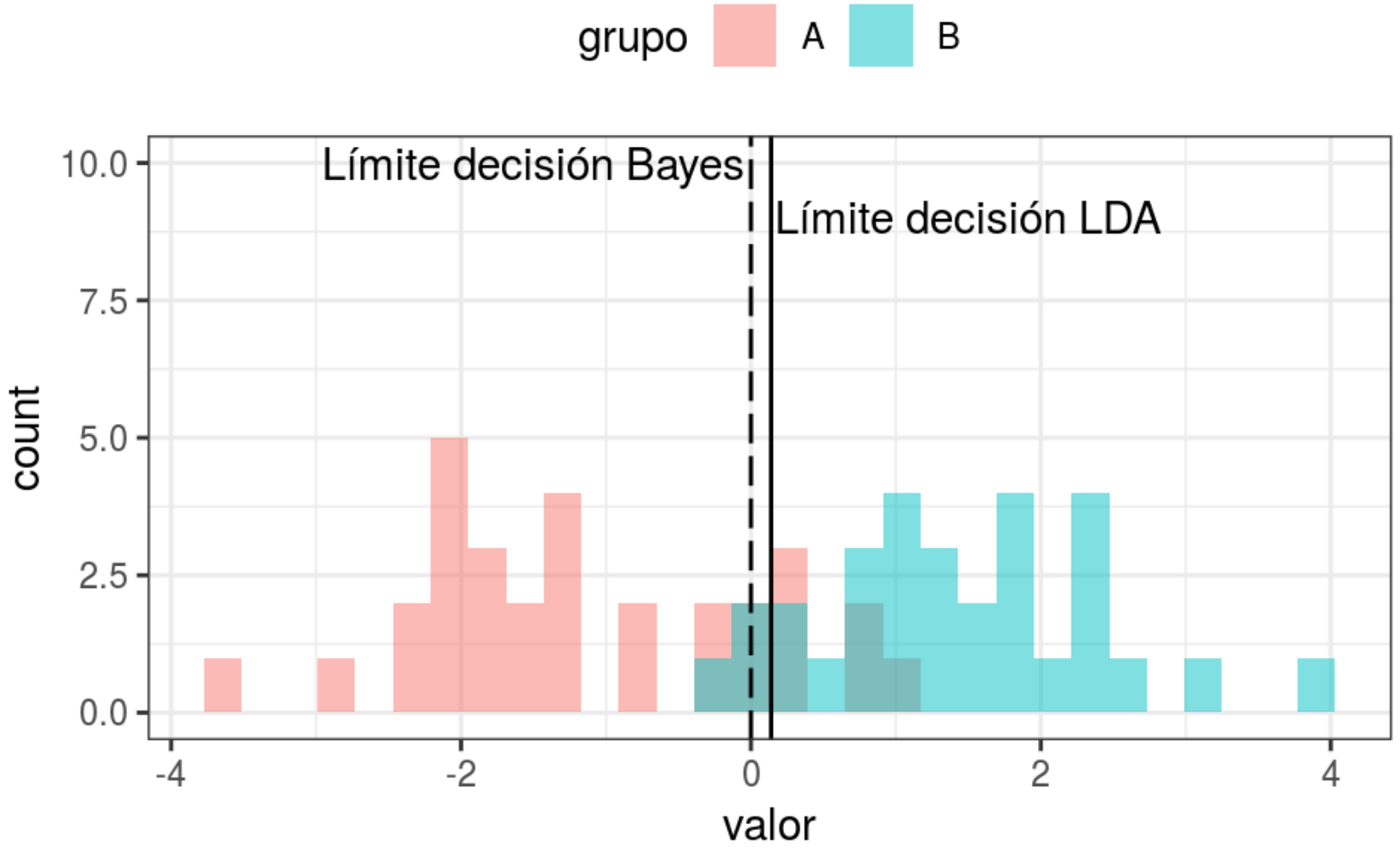
La siguiente imagen muestra dos grupos distribuidos de forma normal con medias $\mu_1$ = -1.25, $\mu_2$ = 1.25 y varianzas $\sigma^2_1 = \sigma^2_2 =$ 1. Dado que se conoce el valor real de las medias y varianzas poblacionales (esto en la realidad no suele ocurrir), se puede calcular el límite de decisión de Bayes $x = \frac{-1.25+1.25}{2}$ = 0 (linea discontinua).

# Análisis discriminante lineal



# Análisis discriminante lineal

Si en lugar de conocer la verdadera distribución poblacional de cada grupo solo se dispone de muestras, escenario que suele ocurrir en los casos reales, el límite de decisión *LDA* se aproxima al verdadero límite de decisión de Bayes pero no es exacto. Cuanto más representativas sean las muestras mejor la aproximación.



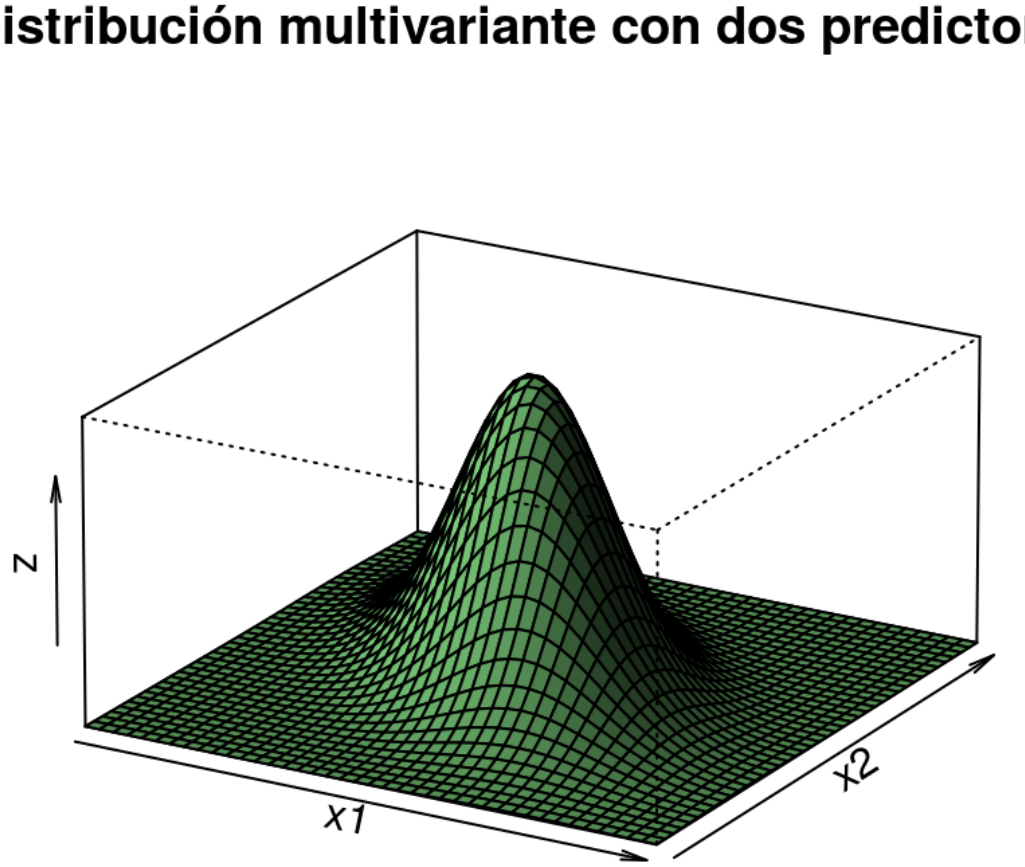
# Análisis discriminante lineal

## Extensión del LDA para múltiples predictores

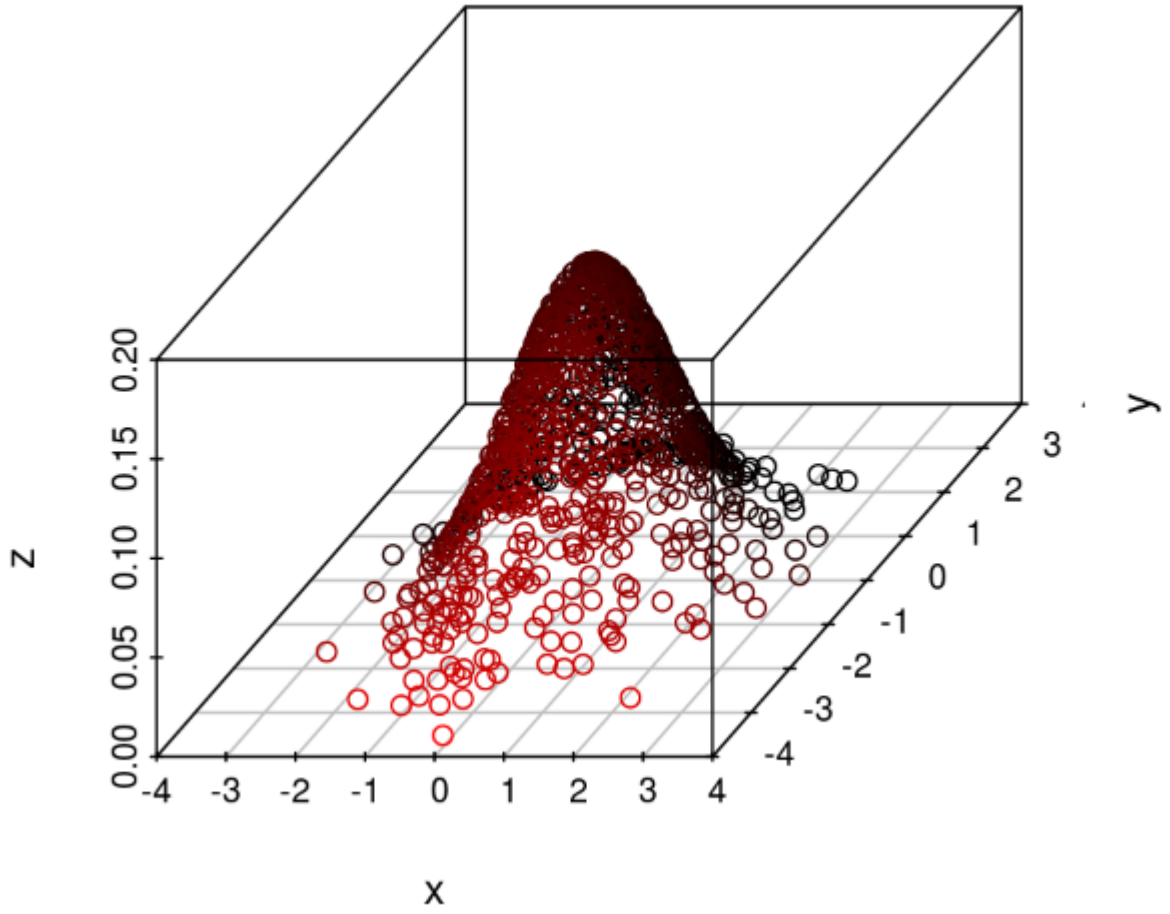
Los conceptos anteriormente descritos empleando un único predictor pueden generalizarse para introducir múltiples predictores en el modelo. La diferencia reside en que *X*, en lugar de ser un único valor, es un vector formado por el valor de *p* predictores $X = (X_1, X_2, \ldots, X_3)$ y que, en lugar de proceder de una distribución normal, procede de una distribución normal multivariante.

Un vector sigue una distribución *k*-normal multivariante si cada uno de los elementos individuales que lo forman sigue una distribución normal y lo mismo para toda combinación lineal de sus *k* elementos. Las siguientes imágenes muestran representaciones gráficas de distribuciones normales multivariante de 2 elementos (distribución normal bivariante).

# Análisis discriminante lineal



# Análisis discriminante lineal



# Análisis discriminante lineal

Para indicar que una variable aleatoria *p*-dimensional *X* sigue una distribución normal multivariante se emplea la terminología $X \sim N(\mu, \sum)$. Donde $\mu$ es el vector promedio de *X* y $\sum$ es la covarianza de *X*, que al ser un vector con *p* elementos, es una matriz *p x p* con la covarianza de cada par de predictores. La ecuación que define la función de densidad de una distribución normal multivariante es:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} \, exp\,(-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu))$$

Si se sigue el mismo procedimiento que el mostrado para *LDA* con un solo predictor, pero esta vez con la ecuación de multivariante normal, y se asume que la matriz de covarianzas $\sum$ es igual para las *K* clases, se obtiene que el clasificador de Bayes es:

$$\hat{\delta}_k(x) = log(P(Y=k|X=x)) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + log(\pi_k)$$

Cuando los parámetros poblacionales se desconocen, no se puede calcular el límite de decisión de Bayes exacto, por lo que se recurre a la estimación de $\mu_1 \ldots, \mu_k$, $\pi_1 \ldots \pi_k$ y $\sum$ para obtener los límites de decisión de *LDA*.

# Análisis discriminante lineal

## Condiciones de LDA

Las condiciones que se deben cumplir para que un Análisis Discriminante Lineal sea válido son:

- Cada predictor que forma parte del modelo se distribuye de forma normal en cada una de las clases de la variable respuesta. En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases.
- La varianza del predictor es igual en todas las clases de la variable respuesta. En el caso de múltiples predictores, la matriz de covarianza es igual en todas las clases. Si esto no se cumple se recurre a Análisis Discriminante Cuadrático (*QDA*).

Cuando la condición de normalidad no se cumple, el LDA pierde precisión pero aun así puede llegar a clasificaciones relativamente buenas. *Using discriminant analysis for multi-class classification: an experimental investigation (Tao Li, Shenghuo Zhu, Mitsunori Ogihara)*.

# Análisis discriminante lineal

## Dos aproximaciones a LDA: Bayes y Fisher

Existen varios enfoques posibles para realizar un *LDA*. La aproximación descrita anteriormente está basada en el clasificador de Bayes, y utiliza todas las variables originales para calcular las probabilidades posteriores de que una observación pertenezca a cada grupo.

Antes de que el clasificador de Bayes fuese introducido en el *LDA*, Fisher propuso una aproximación en la que el espacio $p$-dimensional (donde $p$ es el número de predictores originales) se reduce a un subespacio de menos dimensiones formado por las combinaciones lineales de los predictores que mejor explican la separación de las clases. Una vez encontradas dichas combinaciones se realiza la clasificación en este subespacio. Fisher definió como subespacio óptimo a aquel que maximiza la distancia entre grupos en términos de varianza. Los términos de *discriminante lineal de Fisher* y *LDA* son a menudo usados para expresar la misma idea, sin embargo, el artículo original de Fisher describe un discriminante ligeramente diferente, que no hace algunas de las suposiciones del *LDA* como la de una distribución normal de las clases o covarianzas iguales entre clases.

La aproximación de Fisher se puede ver como un proceso con dos partes:

- Reducción de dimensionalidad: Se pasa de $p$ variables predictoras originales a $k$ combinaciones lineales de dichos predictores (variables discriminantes) que permiten explicar la separación de los grupos pero con menos dimensiones ($k < p$).
- Clasificación de las observaciones empleando las variables discriminantes.

# Análisis discriminante lineal

Los resultados de clasificación obtenidos mediante el método de Fisher son iguales a los obtenidos por el método de Bayes cuando:

- En el método de Bayes se asume que la matriz de covarianzas en igual en todos los grupos y se emplea como estimación la *pooled within-class covariance matrix*.
- En el método de Fisher, todos los discriminantes lineales se utilizan para la clasificación. El número máximo de discriminantes obtenido tras la reducción de dimensionalidad es *número grupos*-1.

*Bayes Optimality in Linear Discriminant Analysis Onur C. Hamsici and Aleix M. Martinez*

*Generalizing Fisher's linear discriminant analysis via the SIR approach, Chapter 14*

# Análisis discriminante lineal

## Precisión del LDA

Una vez que las normas de clasificación se han establecido, se tiene que evaluar como de buena es la clasificación resultante. En otras palabras, evaluar el porcentaje de aciertos en las clasificaciones.

Las matrices de confusión son una de las mejores formas de evaluar la capacidad de acierto que tiene un modelo *LDA*. Muestran el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. El método *LDA* busca los límites de decisión que más se aproximan al clasificador de Bayes, que por definición, tiene el menor ratio de error total de entre todos los clasificadores (si se cumple la condición de normalidad). Por lo tanto, el *LDA* intenta conseguir el menor número de clasificaciones erróneas posibles, pero no diferencia entre falsos positivos o falsos negativos. Si se quiere intentar reducir el número de errores de clasificación en una dirección determinada (por ejemplo, menos falsos negativos) se puede modificar el límite de decisión, aunque como consecuencia aumentará el número de falsos positivos.

Cuando para evaluar el error de clasificación se emplean las mismas observaciones con las que se ha creado el modelo, se obtiene lo que se denomina el *training error*. Si bien esta es una forma sencilla de estimar la precisión en la clasificación, tiende a ser excesivamente optimista. Es más adecuado evaluar el modelo empleando observaciones nuevas que el modelo no ha visto, obteniendo así el *test error*. En el capítulo Validación de modelos de regresión se describen diferentes estrategias para estimar el *test error*.

# Análisis discriminante lineal

# Análisis discriminante lineal

# Análisis discriminante lineal

# Análisis discriminante lineal