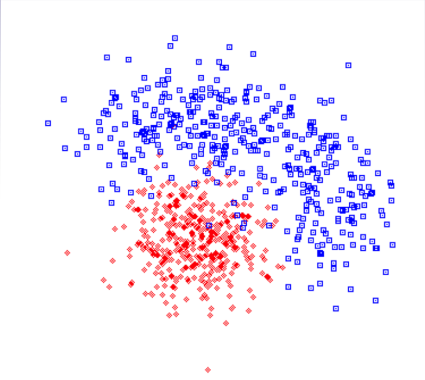
# CAPITULO VIII

# ANALISIS MULTIDIMENSIONAL

# Análisis multidimensional (I)

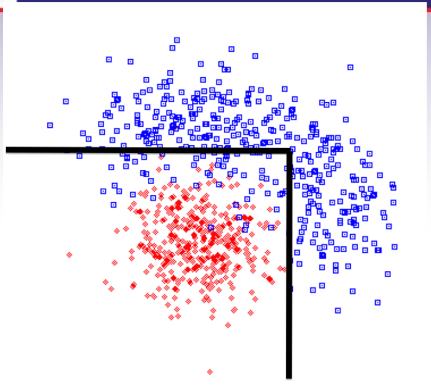


Puntos azules : muestra de control finita, distribuída como el fondo
Puntos rojos : muestra de control finita, distribuída como la señal

A menudo, hay grandes regiones del espacio de la muestra en las cuales los fondos son ampliamente dominantes, o donde la densidad de la señal es nula o despeciable.
Si se reduce la muestra a subconjuntos "enriquecidos en señal" del espacio completo, la pérdida de información puede ser mínima, y otras ventajas pueden compensar esas posibles pérdidas:

- para muestras multidimensionales, puede ser difícil caracterizar las formas de las PDFs en las regiones de baja densidad de eventos
- el reducir el tamaño de la muestra puede aliciar el consumo de memoria y CPU en las partes numéricas del análisis (p.e. la minimización)

# Análisis multidimensional (II)



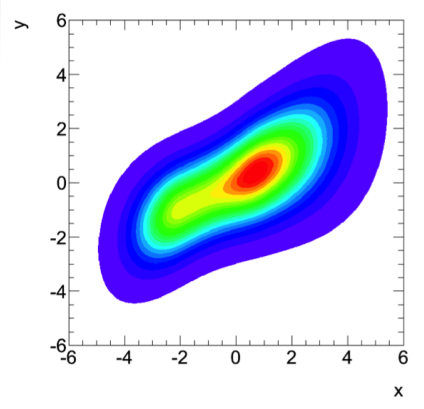
El método más sencillo de reducción de una muestra es restringiendo las variables, una a una, a intervalos finitos.
En la práctica, esas selecciones "cut-based" aparecen a varios niveles de la definición del espacio de muestreo: umbrales en las decisiones en línea (triggers), filtros a varios niveles posteriores del proceso de adquisición, eliminación de datos a partir de criterios de calidad...
Pero ya en etapas más avanzadas del análisis de datos, esas selecciones "accept-reject" convienen ser reemplazadas por procedimientos más sofisticados. Estos son llamados de manera genérica *técnicas multivariadas*.

Las líneas negras indican una selección "cut-based", definida de manera de conservar cerca de $100\,\%$ de la señal.

(para ser más precisos, la selección es $100\,\%$ eficaz sobre la muestra de control de señal, pero la caracterización precisa de la eficiencia de selección require de estimar la densidad de probabilidad de la señal fuera de la región seleccionada (y para caracterizar con precisión el nivel de fondo se require estimar la densidad de probabilidad del fondo dentro de ella)

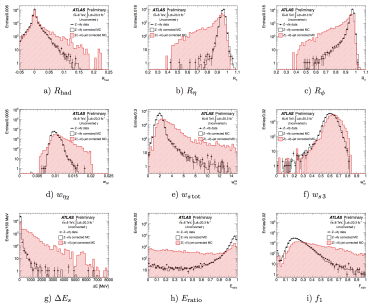
# Análisis multidimensional (III)

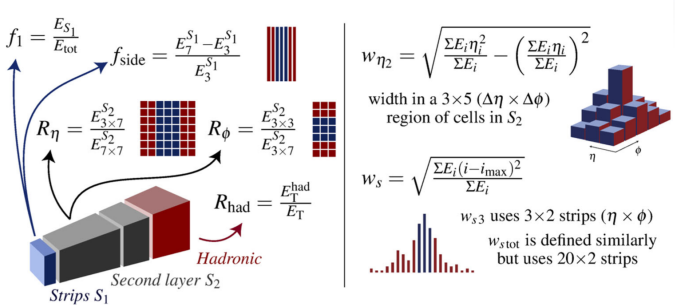
Consideremos un conjunto de $n$ variables aleatorias $\vec{x} = \{x_1, x_2 \ldots, x_n\}$. Si todas las variables son no-correlacionadas, Las PDFs $n$-dimensionales están completamente determinadas por el producto directo de sus $n$ PDFs uni-dimensionales.
si algunas de las variables están correlacionadas, y si sus patrones de correlacíon son completamente lineales, es posible definir un nuevo conjunto de variables $\vec{y}$, que son combinaciones lineales de $\vec{x}$, obtenidas diagonalizando el inverso de la matriz de covarianza.




En ocasiones, cuando los patrones de correlación son no-lineales, es tal vez posible en algunos casos definir una descripción analítica: por ejemplo, el perfil de correlación que incluye una (ligera) componente no lineal representado aquí, fue producido con el paquete `RooFit` aplicando la opción `Conditional` en `RooProdPdf` para producir un producto de PDFs: la anchura de $y$ varía de manera no-lineal con $x$.
En la práctica, ésta solución elegante no puede extenderse fácilmente a más de dos dimensiones, y no hay garantía que se puedan reproducir patrones no lineales complicados.
Frente a tales escenarios, un protocolo de *reducción dimensional* puede a menudo producir resultados más efectivos que un intento de descripción analítica de correlaciones.

Un escenario típico para considerar la reducción dimensional es cuando varias variables arrastran en gran parte información común (y por tanto exhiben correlaciones fuertes), pero contienen también algumos elementos (diluídos pero importantes) de información independiente.
Un ejemplo: la caracterización de cascadas en calorímetros segmentados. Las señales depositadas en celdas vecinas están fuertemente correlacionadas (tienen un origen común), pero permiten reconstruir detalles precisos del desarrollo de la cascada. Esas correlaciones son utilizadas para caracterizar las "formas de cascadas":



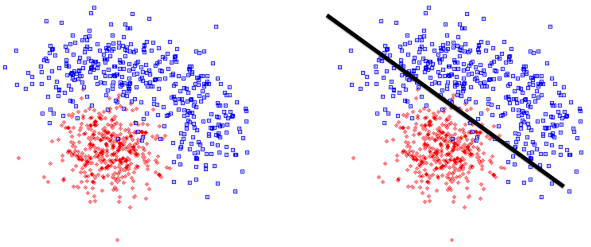

El resultado de combinar las informaciones de todas las "shower shape variables" es la *clasificación* del candidato en dos especies: cascadas electromagnéticas (fotones, electrones) vs. cascadas hadrónicas ("jets").

# Discriminantes lineales

El algoritmo más sencillo de reducción dimensional es el discriminante de Fisher: es una función lineal de las variables, con coeficientes definidos a partir de un criterio de optimización de la separación entre especies (c.f. próxima lámina).

- Fisher es un caso particular de los llamados PCA (principal component analysis);
- un análisis MLE es también un PCA: reduce un problema multidimensional a un problema en 1 dimensión: el comportamiento del estadístico de test $\lambda = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)}$.

Por construcción, los discriminantes lineales son óptimos para variables multinormales, por tanto con correlaciones perfectamente lineales.



Ejercicio:

- dos variables aleatorias, $x_1$, $x_2$;
- dos especies: señal y fondo;
- la PDF de la señal es una bigaussiana, con $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $c_{12} = 0$ ;
- la PDF del fondo es una bigaussiana, con $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$, $c_{12} = 0{,}5$ ;

$\rightarrow$ Determinar los coeficientes de Fisher

$\rightarrow$ Representar gráficamente las distribuciones de $\mathcal{F}$ para las dos especies

$\rightarrow$ Evaluar la separación $\langle s^2 \rangle$ de $\mathcal{F}$

Tenemos dos especies, "señal" y "fondo":

- distinguimos las especies con un índice $c = 1, 2$.

Tenemos $n$ variables discriminantes $\vec{x} = x_1, x_2, \cdots x_n$:

- caracterizamos las PDFs de cada especie usando muestras de control compuestas por $N_c$ eventos cada una, formando las llamadas "n-tuplas" $\mathbf{x}^c$.

Algunas definiciones necesarias:

- la "tupla-media" $\overline{\mathbf{x}}^c$ de cada especie, y la tupla especie-promediada $\overline{\mathbf{x}}$:

$$\overline{\mathbf{x}}^c = \frac{1}{N_c} \sum_{k=1}^{N_c} \mathbf{x}_k^c \quad , \quad \overline{\mathbf{x}} = \frac{1}{N_1 + N_2} \sum_{c=1}^{2} \sum_{k=1}^{N_c} \mathbf{x}_k^c \ ,$$

- las matrices de covarianza *intra-especie* $W_{ij}$ ("within") e *inter-especie* $B_{ij}$ ("between"):

$$W_{ij} = \frac{1}{N_1 + N_2} \sum_{c=1}^{2} \sum_{k=1}^{N_c} \left( x_i^{ck} - \overline{x}_i^c \right) \left( x_j^{ck} - \overline{x}_j^c \right) \quad , \quad B_{ij} = \frac{1}{N_1 + N_2} \sum_{c=1}^{2} N_c \left( \overline{x}_i^c - \overline{x}_i \right) \left( \overline{x}_j^c - \overline{x}_j \right) \ ,$$

- la matriz de covariancia especie-promediada $T_{ij} = B_{ij} + W_{ij}$. Los índices $i, j = 1, 2, \cdots, n$.

# El discriminante de Fisher (II)

El *discriminante de Fisher* $\mathcal{F}$, es una combinación lineal de las variables aleatorias $\vec{x} = x_1, x_2, \cdots x_n$ dada por

$$\mathcal{F} = f_0 + \sum_{i=1}^{n} f_i x_i \ .$$

donde los $f_i$ son los llamados coeficientes de Fisher-Mahalanobis, asignados a cada variable $i = 1, 2, \cdots, n$:

$$f_i = \frac{\sqrt{N_1 N_2}}{N_1 + N_2} \sum_{i=1}^{n} C_{ij}^{-1} \left( \overline{x}_i^1 - \overline{x}_i^2 \right) \quad , \quad f_0 = \sum_{i=1}^{n} f_i \left( \overline{x}_i^1 + \overline{x}_i^2 \right) \ ,$$

(aquí la matriz $C_{ij}^{-1}$ coresponde a $W_{ij}^{-1}$ para Fisher, y a $T_{ij}^{-1}$ para Mahalanobis)
El discriminante lineal de Fisher *reduce* la dimensionalidad de nuestro problema:

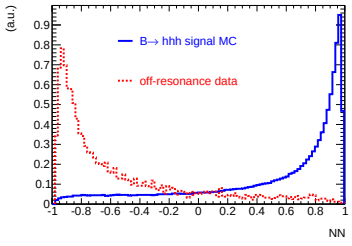
- teníamos $n$ variables aleatorias discriminantes $\vec{x} = x_1, x_2, \cdots x_n$,
- terminamos con una única variable aleatoria discriminante $\mathcal{F}$,
- $\mathcal{F}$ es una combinación lineal de las variables iniciales.

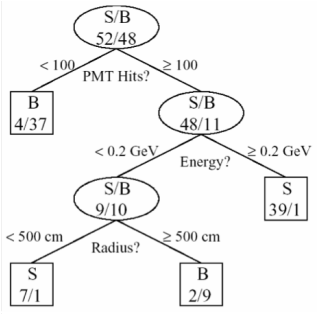
De manera general, el discriminante de Fisher

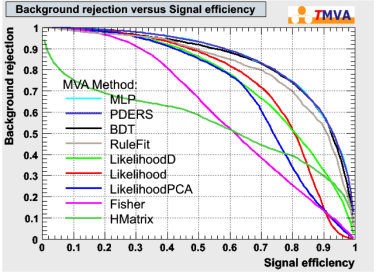
- produce una discriminación *óptima* para variables multinormales
  - (que comportan por tanto únicamente correlaciones perfectamente lineales)
- una discriminación subóptima pero elevada para distribuciones con correlaciones únicamente lineales
- puede ser claramente subóptimo, e incluso fallar totalmente, en caso de correlaciones altamente no-lineales.

# Discriminantes no lineales (I)

Para tratar correlaciones no lineales y con perfiles complejos, existe una variedad de técnicas y herramientas.
El paquete `TMVA` es una implementación popular en HEP de varios algoritmos de reducción dimensional: además de una biblioteca de discriminantes lineales y basados en likelihood, incluye métodos de entrenamiento y prueba con redes de neuronas artificiales y árboles de decisión (boosted decision trees), que forman parte de los más utilizados en HEP.
Idea general: un análisis multivariado utiliza una colección de variables de entrada, que se combinan de manera optimizada, a partir de un algoritmo "entrenado" sobre dos muestras independientes (correspondientes a señal y fondo), con dos protocolos: entrenamiento (training) y prueba (test). El performance del algoritmo entrenado se evalúa sobre las muestras de prueba, para evitar el efecto de "over-training".

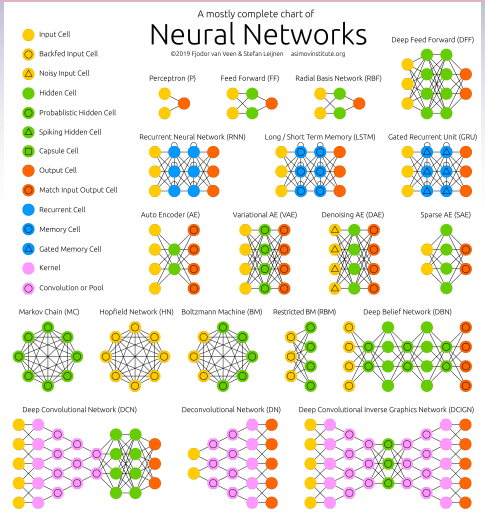


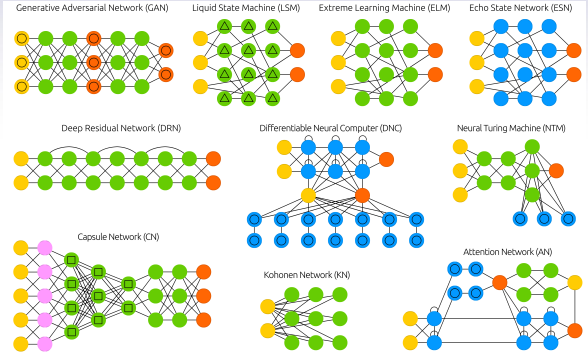




"ROC-curve":
Receiver Operating Characteristic

# Discriminantes no lineales (II)




(tomado de `https://www.asimovinstitute.org/neural-network-zoo/`)

En resumen, un *análisis multivariado* produce una reducción dimensional, proyectando un espacio de $n$ variables aleatorias $\vec{x} = \{x_1, x_2 \ldots, x_n\}$, sobre una variable final $\mathcal{Z}$. Esta variable puede ser

- una combinación lineal de las $\vec{x}$ (discriminante de Fisher) ;
- un estadístico de prueba más elaborado (por ejemplo el cociente de verosimilitudes $\lambda = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)}$) ;
- la salida de un algoritmo de *Machine Learning*. Ejemplos: *Multilayer Perceptron* (red de neuronas artificiales en sus múltiples apelaciones, tipo NN, CNN, ANN, GNN...), *Boosted Decision Tree* (árbol de decisiones), *algoritmos genéticos*, y otros más.

De manera general, los algoritmos de ML obtienen *performaces* superiores que los algoritmos más sencillos, pero es crucial verificar la robustez de esos performances con respecto a la calidad de las muestras de control, o con respecto al algoritmo de entrenamiento. Dependiendo del objetivo del análisis se podrá preferir la robustez al performance, o se privilegiará una combinación entre ambas...
Dos citas sacadas de WikiPedia:

- Trees can be very non-robust. A small change in the training data can result in a large change in the tree and consequently the final predictions.
- Decision-tree learners can create over-complex trees that do not generalize well from the training data. (This is known as overfitting.)

Estos tópicos serán tratados desde varias perspectivas complementarias en el curso siguiente: “Tópicos avanzados en ciencia de datos”.
¡ Feliz continuación !