CAPITULO V

EL METODO DE LA VEROSIMILITUD MAXIMA

Estimación de parámetros: limitaciones

En las previas láminas describimos una lógica "intuitiva" o "caso a caso" de la estimación de parámetros. Eso, por supuesto, no es generalizable (ni robusto) : de manera general, y más allá de los ejemplos sencillos

- los estimadores sencillos de los momentos de la función característica están sometidos a sesgos
- una descripción completa de la PDF require a priori un número infinito de momentos

Una descripción más general del problema es la siguiente :

- tenemos una muestra compuesta por N realizaciones independientes de variables aleatorias $\vec{x}$
- suponemos que esas realizaciones resultan de muestrear una PDF n-paramétrica

$$P\left(\vec{x}; \theta_1, \ldots, \theta_n\right) \; ,$$

- suponemos también que la dependencia funcional de la PDF es conocida, y que solamente ignoramos los valores numéricos de los parámetros

Fisher, 1921: el *teorema de la verosimilitud máxima* (maximum likelihood) es una herramienta poderosa para la estimación de parámetros $\theta_1, \ldots, \theta_n$ de nuestra PDF.

El teorema de verosimilitud máxima (I)

Definimos la función de verosimilitud $\mathcal{L}$, evaluada sobre una muestra compuesta por N eventos :

$$\mathcal{L}\left(\theta_1, \ldots, \theta_n\right) = \prod_{i=1}^{N} P\left(\vec{x}_i; \theta_1, \ldots, \theta_n\right) .$$

Teorema : los valores $\hat{\theta}_1, \ldots, \hat{\theta}_n$ que maximizan la función $\mathcal{L}$ son estimadores de los parámetros $\theta_1, \ldots, \theta_n$ de nuestra PDF,

$$\mathcal{L}\left(\hat{\theta}_1, \ldots, \hat{\theta}_n\right) = \max|_{\theta} \left\{\mathcal{L}\left(\theta_1, \ldots, \theta_n\right)\right\} ,$$

con varianzas $\hat{\sigma}_\theta$ que se extraen a partir de la matriz de covarianza de $\mathcal{L}$ alrededor de su máximo.
En palabras intuitivas: para una muestra dada, el MLE corresponde a los valores que maximizan la probabilidad de realizar esa muestra !
No es un pleonasmo : la función $\mathcal{L}$ debe satisfacer ciertas condiciones:

- ser derivable al menos dos veces con respecto a los parámetros $\theta_1, \ldots, \theta_n$,
- ser (asintóticalmente) no sesgada y eficiente (condición llamada "Cramer-Rao bound"),
- seguir una distribución (asintóticamente) multi-normal,

$$f\left(\hat{\vec{\theta}}, \vec{\theta}, \mathbf{\Sigma}\right) = \frac{1}{\sqrt{2\pi}\,|\mathbf{\Sigma}|} \exp\left\{-\frac{1}{2}\left(\hat{\vec{\theta}}_i - \vec{\theta}_i\right) \Sigma_{ij}^{-1} \left(\hat{\vec{\theta}}_j - \vec{\theta}_j\right)\right\} .$$

- seguir una distribución (asintóticamente) multi-normal,

$$f\left(\hat{\vec{\theta}}, \vec{\theta}, \mathbf{\Sigma}\right) = \frac{1}{\sqrt{2\pi}\,|\mathbf{\Sigma}|} \exp\left\{-\frac{1}{2}\left(\hat{\vec{\theta}}_i - \vec{\theta}_i\right) \Sigma_{ij}^{-1} \left(\hat{\vec{\theta}}_j - \vec{\theta}_j\right)\right\} .$$

donde la matriz de covarianza $\mathbf{\Sigma}$ es

$$\mathbf{\Sigma}_{ij}^{-1} = -E\left[\frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}\right] .$$

Al alejarnos del máximo, el valor de la función $\mathcal{L}$ disminuye, a una tasa que depende de los elementos de la matriz de covarianza :

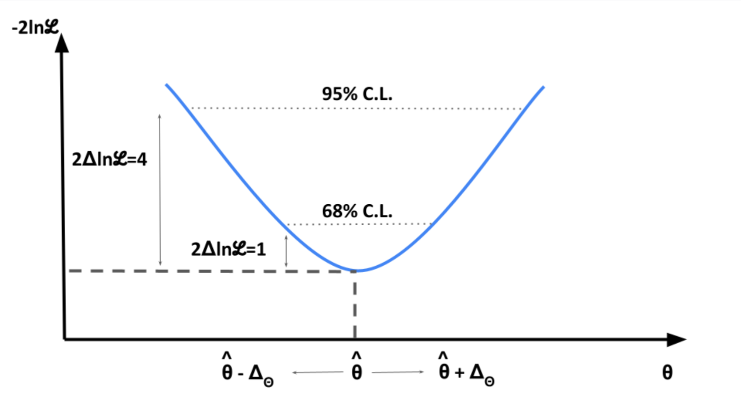
$$-2\Delta \ln \mathcal{L} = -2\left[\ln \mathcal{L}(\vec{\theta}) - \ln \mathcal{L}(\hat{\vec{\theta}})\right] = \sum_{i,j} \left(\theta_i - \hat{\theta}_i\right) \mathbf{\Sigma}_{ij}^{-1} \left(\theta_i - \hat{\theta}_j\right) .$$

En otras palabras: la matriz de covarianza define mapas de contorno alrededor de su máximo, que corresponden a *intervalos de confianza*.
En el caso de una $\mathcal{L}$ con un parámetro único $\mathcal{L}(\theta)$, el intervalo contenido dentro de $-2\Delta \ln \mathcal{L} < 1$ alrededor de $\hat{\theta}$ define un intervalo de confianza a $68\,\%$ que corresponde a un rango $-\Delta_\theta \leq \theta - \hat{\theta} \leq \Delta_\theta$ alrededor del punto máximo.
Por ello el resultado de MLE se escribe a menudo como $\left(\hat{\theta} \pm \hat{\Delta}_\theta\right)$.

El teorema de verosimilitud máxima (III)

En el caso de una $\mathcal{L}$ con un parámetro único $\mathcal{L}(\theta)$, el intervalo contenido dentro de $-2\Delta\ln\mathcal{L} < 1$ alrededor de $\hat{\theta}$ define un intervalo de confianza a $68\,\%$ que corresponde a un rango $-\Delta_\theta \leq \theta - \hat{\theta} \leq \Delta_\theta$ alrededor del punto máximo.
Por ello el resultado de MLE se escribe a menudo como $\left(\hat{\theta} \pm \hat{\Delta}_\theta\right)$.

(y de la misma manera, el intervalo $-2\Delta\ln\mathcal{L} < 4$ define un intervalo de confianza a $95\,\%$ corresponde a un rango $-2\Delta_\theta \leq \theta - \hat{\theta} \leq 2\Delta_\theta$, etc...)



Una muestra compuesta por N realizaciones de una única variable aleatoria x, que suponemos sigue una distribución Gaussiana de media μ y anchura σ. El teorema MLE nos permite estimar los parámetros así:

$$\mathcal{L}(\mu,\sigma) = \prod_{i=1}^{N} P_{\text{Gaus}}(x_i;\mu,\sigma) \;; -\ln\mathcal{L} = N\ln\sigma + \sum_{\text{i}=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\;\;(+\text{constante}).$$

Los estimadores $\hat{\mu}$ y $\hat{\sigma}$ son los ceros de las primeras derivadas de $-\ln\mathcal{L}$ con respecto a μ y σ :

$$\frac{\partial}{\partial\mu}(-\ln\mathcal{L})\bigg|_{\hat{\mu},\hat{\sigma}} = 0 \quad\longrightarrow\quad \hat{\mu} = \frac{1}{N}\sum_{\text{i}=1}^{N} x_i\ .$$

$$\frac{\partial}{\partial\sigma}(-\ln\mathcal{L})\bigg|_{\hat{\sigma},\hat{\mu}} = 0 \quad\longrightarrow\quad \hat{\sigma}^2 = \frac{1}{N}\sum_{\text{i}=1}^{N}(x_i-\mu)^2\ .$$

Las segundas derivadas nos dan los elementos de la matriz de covarianza:

$$\frac{\partial^2}{\partial\mu^2}(-\ln\mathcal{L})\bigg|_{\hat{\mu},\hat{\sigma}} = \frac{N}{\hat{\sigma}^2} \quad\longrightarrow\quad \Sigma_{\mu\mu} = \frac{\hat{\sigma^2}}{N} \quad\longrightarrow\quad \hat{\Delta}_{\mu} = \frac{\hat{\sigma}}{\sqrt{N}}\ ,$$

$$\frac{\partial^2}{\partial\sigma^2}(-\ln\mathcal{L})\bigg|_{\hat{\mu},\hat{\sigma}} = \frac{2N}{\hat{\sigma}^2} \quad\longrightarrow\quad \Sigma_{\sigma\sigma} = \frac{\hat{\sigma^2}}{2N} \quad\longrightarrow\quad \hat{\Delta}_{\sigma} = \frac{\hat{\sigma}}{\sqrt{2N}}\ ,$$

(y los términos no diagonales de la covarianza son ambos cero).

El MLE extrae de manera formal y robusta los parámetros y errores $(\hat{\mu}\pm\hat{\Delta}_{\mu})$ y $(\hat{\sigma}\pm\hat{\Delta}_{\sigma})$ de una Gaussiana.

Una muestra compuesta por N realizaciones de una única variable aleatoria x, que suponemos sigue una distribución exponencial con parámetro de forma ξ. El teorema MLE nos permite estimar ese parámetro xi analíticamente, al menos para el caso en que la variable x cubre el rango $[0; +\infty]$.

$$\mathcal{L}(\xi) = \prod_{i=1}^{N} P_{\text{exponential}}\left(x_i; \xi\right) = \prod_{i=1}^{N} \frac{1}{\xi} e^{-x_i/\xi} \ .$$

De manera análoga al ejemplo anterior, determinamos el NLL :

$$-\ln \mathcal{L} = N \ln \xi - \frac{1}{\xi} \sum_{i=1}^{N} x_i \ ,$$

y evaluamos sus primera y segunda derivadas para el valor $\hat{\xi}$ que anula la primera derivada :

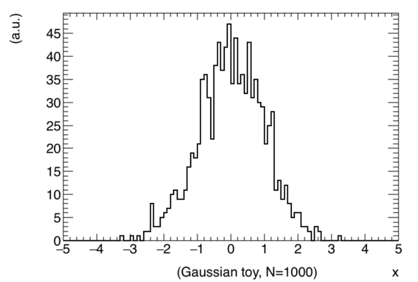
$$\left. \frac{\partial}{\partial \xi}(-\ln \mathcal{L}) \right|_{\hat{\xi}} = 0 \longrightarrow \hat{\xi} = \frac{1}{N} \sum_{i=1}^{N} x_i \ ,$$

$$\left. \frac{\partial^2}{\partial \xi^2} \right|_{\hat{\xi}} = \frac{N}{\hat{\xi}^2} \longrightarrow \Sigma_{\xi\xi} = \frac{\hat{\xi}^2}{N} \longrightarrow \hat{\Delta}_{\xi} = \frac{\hat{\xi}}{\sqrt{N}} \ .$$

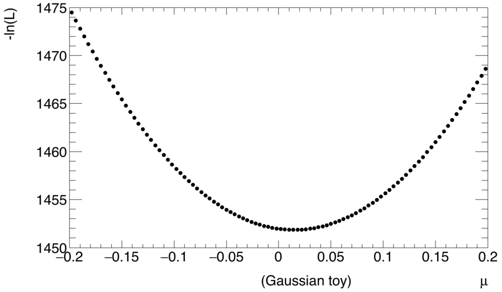
Con lo que el MLE nos da en este caso también una solución analítica para la estimación del parámetro de forma exponencial y su correspondiente error $(\hat{\xi} \pm \hat{\Delta}_{\xi})$.
Nota: en este caso como el anterior, los errores $\hat{\Delta}$ escalan como $1/\sqrt{N}$. Esa es una propiedad muy general.

Un “ejemplo de juguete” (I)

- Generamos una realización aleatoria con N=1000 eventos (“toy sample”) a partir de una PDF Gaussiana reducida, con $\mu = 0$, $\sigma = 1$.
- Evaluamos (el negativo del logaritmo de) la función de verosimilitud para esa muestra, para diferentes valores de μ y σ (“escaneamos” los parámetros de interés)



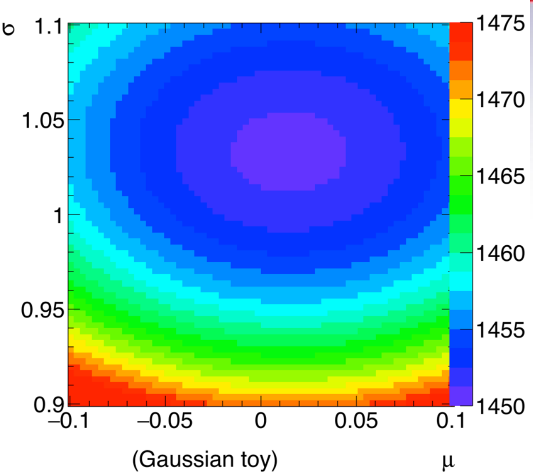
$$-\ln \mathcal{L}(\mu, \sigma) = N \ln \sigma + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \ .$$



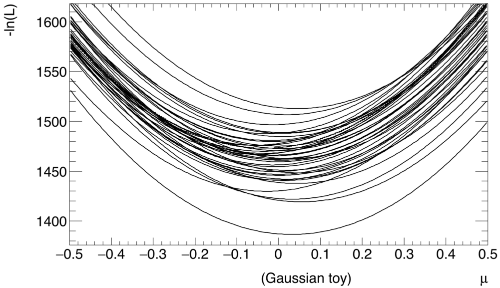
- La forma de $-\ln \mathcal{L}$ en función de μ (para σ fijo) sigue un perfil parabólico, y el mínimo coincide con el promedio empírico
- Por inspección alrededor del mínimo, se observa que el NLL aumenta en $0{,}5$ unidades para $\Delta x \sim \pm 0{,}03$, que corresponde aproximadamente a $\sigma/\sqrt{N}$ para $N = 1000$
- La lámina siguiente muestra el gráfico del escaneo en 2D de μ y σ

Un "ejemplo de juguete" (II)

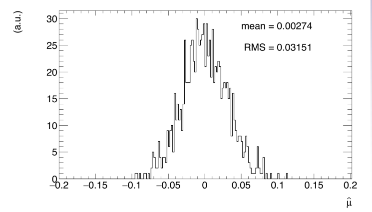
- El perfil bidimensional de $-\ln\mathcal{L}$ es el de un paraboloide, con semi-ejes diferentes y ortogonales
- El mínimo coincide con la posición del promedio y el RMS empíricos
- El semi-eje a lo largo de σ es más estrecho que el de μ, como se espera de la relación $1/\sqrt{2N}$ vs. $1/\sqrt{N}$

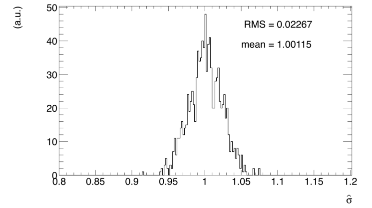


Para otras realizaciones aleatorias independientes, la posición y la anchura de los mínimos cambia :



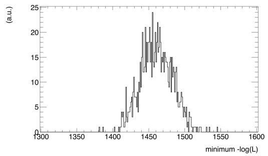
Un "ejemplo de juguete" (III)





Si realizamos el mismo estudio sobre un ensamble de muestras generadas con la misma PDF (aquí N=1000, $\mu = 0$, $\sigma = 1$) obtenemos los resultados siguientes :

- los mínimos de $\hat{\mu}$ fluctúan alrededor de su valor verdadero $\mu = 0$ con una dispersión de $\pm 3{,}1\,\%$, que corresponde a $\sigma/\sqrt{N}$
- los mínimos de $\hat{\sigma}$ fluctúan alrededor de su valor verdadero $\sigma = 1$ con una dispersión de $\pm 2{,}3\,\%$, que corresponde a $\sigma/\sqrt{2N}$
- los intervalos con $-\Delta \ln \mathcal{L} < 0{,}5$ alrededor del mínimo corresponden bien a las regiones que cubren $68{,}3\,\%$ de la dispersión
- el teorema MLE nos da una definición rigurosa y precisa de la incertidumbre estadística

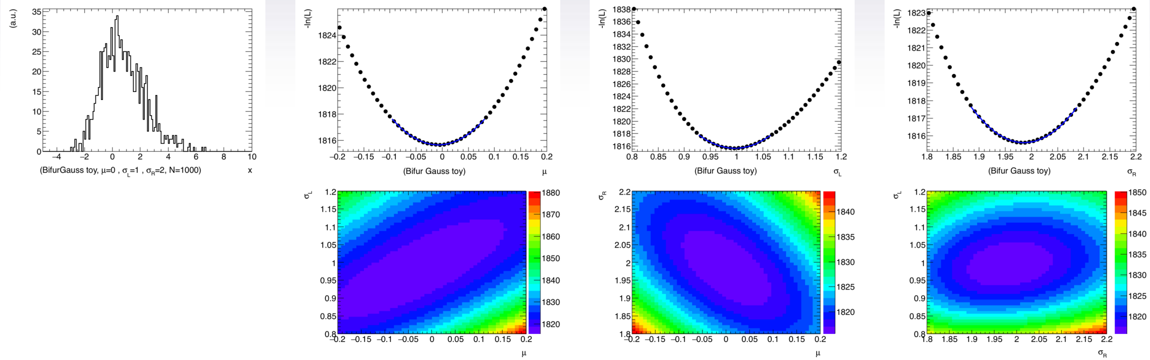


La distribución del $-\ln \mathcal{L}$ no siempre sigue una forma precisa (aquí parece bastante Gaussiana) pero permite hacer una estimación del "goddness-of-fit" :

- si el $-\ln \mathcal{L}$ observado en una muestra se aleja significativamente del intervalo cubierto en un "estudio de juguete", la calidad del modelo es sospechosa...

Otro “ejemplo de juguete” (I)

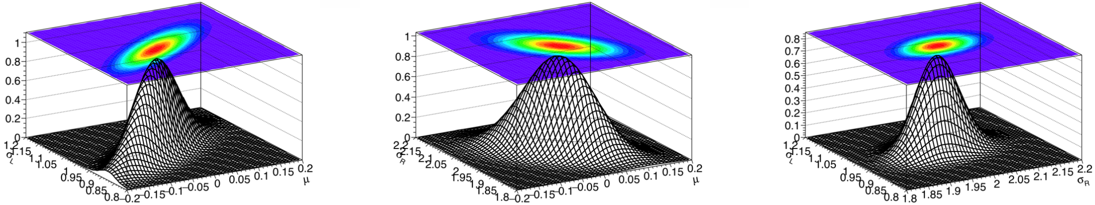
Consideremos ahora una Gaussiana bifurcada, PDF con tres parámetros : μ, σ_L y σ_R. Los mínimos del $-\log \mathcal{L}$ no son del todo parabólicos, y hay correlaciones importantes entre los parámetros.



Otro "ejemplo de juguete" (II)

Los parámetros de la PDF son ellos mismos variables aleatorias:

- si efectuamos otras realizaciones aleatorias a partir de la misma PDF, obtendremos valores diferentes de los parámetros
- y éstos se distribuirán siguiendo la matriz de covariancia de los parámetros (y por tanto tomando en cuenta las correlaciones)



Correlación entre μ y σ_L : +79 %   Correlación entre μ y σ_R : -59 %   Correlación entre σ_L y σ_R : +19 %

Nota: se trata por supuesto de una PDF en 3 dimensiones: por eso se muestra 3 proyecciones bidimensionales.

Situaciones más generales: algoritmos de minimización

Los ejemplos de juguete discutidos previamente son casos sencillos, con 2 o 3 parámetros, y pudimos explorar el espacio bi- o tri-dimensional con bucles sencillas.
En situaciones más generales, el número de parámetros puede ser significativamente superior, así que la aproximación de "escanear" los parámetros para identificar el mínimo del $-\log \mathcal{L}$ no es eficaz (y se vuelve rápidamente imposible).
Por ello se utilizan algoritmos de minimización numérica, que optimizan la búsqueda del mínimo de una función. El proceso se llama *ajuste numérico* o "fit". El algoritmo más usado en altas energías es MINUIT diseñado en el CERN en los años 1970. MINUIT está implementado en `ROOT`, en la clase `TMinuit`.

MINUIT

From Wikipedia, the free encyclopedia

MINUIT, now **MINUIT2**, is a numerical minimization computer program originally written in the FORTRAN programming language[1] by CERN staff physicist Fred James in the 1970s. The program searches for a minimum in a user-defined function with respect to one or more parameters using several different methods as specified by the user. In addition to that it can compute confidence intervals for the parameters by scanning the function around the minimum.

The original FORTRAN code was later ported to C++ by the ROOT project; both the FORTRAN and C++ versions are in use today. The program is very widely used in particle physics, and thousands of published papers cite use of MINUIT.[2] In the early 2000s, Fred James started a project to implement MINUIT in C++ using object-oriented programming. The new MINUIT is an optional package (minuit2) in the ROOT release. As of October 2014 the latest version is 5.34.14, released on 24 January 2014.[3] There is also a Java port[4] as well as a Python frontend to the C++ code.[5]

MINUIT is not a program that can be distributed as an executable binary to be run by a relatively unskilled user: the user must write and compile a subroutine defining the function to be optimized, and oversee the optimization process.

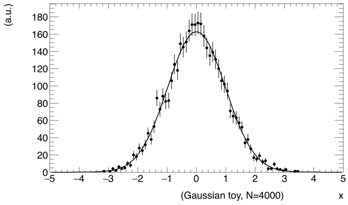
Diferencia entre un fit NLL y un fit de χ^2

Muchos paquetes y programas contienen algoritmos de minimización más sencillos, del tipo "mínimos cuadrados" o χ^2. Estos difieren de un ajuste de verosimilitud máxima: aquí la muestra se compone de un conjunto de n puntos y_i con sus incertidumbres σ_i, y la función a minimizar es:

$$\chi^2(\theta) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i;\theta)}{\sigma_i} \right)^2 ,$$

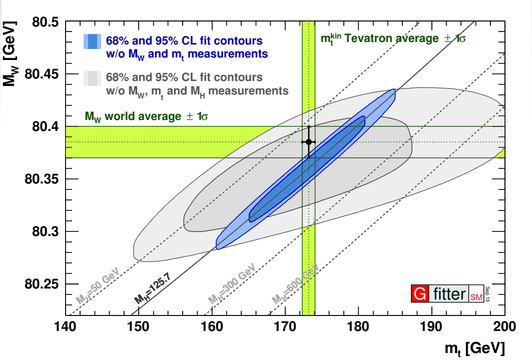
donde $f(x;\theta)$ es la función que pretende describir una dependencia funcional $y = f(x)$. La minimización del χ^2 provee los valores $\hat{\theta}$ estimados por el ajuste.

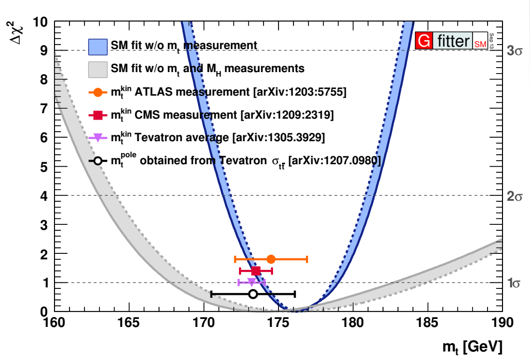
El conjunto de puntos puede provenir de medidas individuales, o ser una reducción por histogrameo de una muestra completa: en el ejemplo aquó se tiene una muestra compuesta de 4000 valores de una variable aleatoria x, histogrameada con 100 *bines* de anchura $\delta x = 0{,}1$ en el intervalo $-5 \leq x \leq +5$. El equivalente a y_i es el número de eventos con valores contenidos en el bin i, y la incertidumbre asociada es $\sigma_i = \sqrt{(y_i)}$.



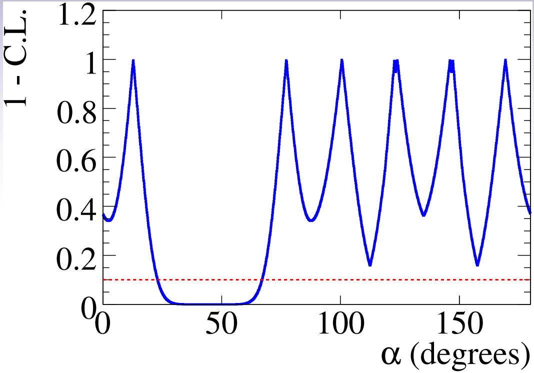
El ajuste de este histograma a una función Gaussiana nos da $\mu = (0{,}008 \pm 0{,}016)$ y $\sigma = (1{,}007 \pm 0{,}011)$.
Nota: si se hubiera usado otro "binning" el resultado puede ser un poco diferente!
Nota: en cambio, el ajuste por verosimilitud máxima a esta misma muestra nos dará exactamente los resultados teóricos: la media empírica y el RMS empírico, con sus incertidumbres $\mathrm{RMS}/\sqrt{n}$ y $\mathrm{RMS}/\sqrt{2n}$.

Los intervalos de confianza de Gfitter

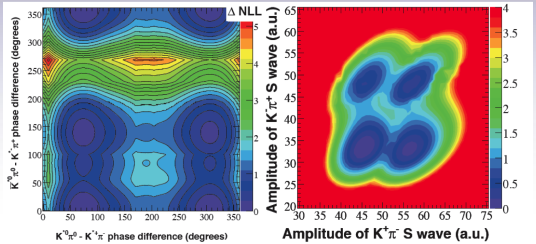




Ejemplos de funciones de verosimilitud complicadas



Un elemento importante del programa científico del experimento *BABAR* (y en general, de la *física de sabores*): la medida del ángulo α de la matriz CKM: $\alpha \neq 0 \longrightarrow$ violación de la simetría CP.
Problema : el observable físico es la asimetría dependiente del tiempo $B^0/\overline{B}^0 \to \pi^+\pi^-$, que es una función de $\sin 2(\alpha - \delta)$: ambigüedad octuple.



Situación más complicada: el perfil de interferencias entre las amplitudes de decaimiento
$B^0/\overline{B}^0 \to K^+\pi^-\pi^0$;

- ambigüedades múltiples, sin forma analítica precisa...
- numerosas amplitudes resonantes intermedias: $K^{\star\pm}\pi^\mp$, $K^{\star 0}\pi^0$, $\rho^\pm\pi^\mp$, varios estados excitados $K^\star$, ρ, otras resonancias... todo para B^0 y para $\overline{B}^0$
- la función de verosimilitud contenía unos 60 parámetros de interés: amplitudes y fases...

Ejemplos de funciones de verosimilitud complicadas

TABLE VIII. Full correlation matrix for the isobar parameters of solution I. The entries are given in percent. Since the matrix is symmetric, all elements above the diagonal are omitted.

		$\lvert c\rvert$							$\lvert\bar{c}\rvert$							
		ρ^0	K^*	S	f_2	f_X	NR	χ	f_0	ρ^0	K^*	S	f_2	f_X	NR	χ
$\lvert c\rvert$	ρ^0	100.0														
	K^*	51.9	100.0													
	S	54.0	65.0	100.0												
	f_2	8.4	2.8	21.0	100.0											
	f_X	14.9	23.2	32.2	22.7	100.0										
	NR	5.2	35.0	24.4	12.6	39.3	100.0									
	χ	6.4	9.9	7.8	2.0	7.4	6.1	100.0								
$\lvert\bar{c}\rvert$	f_0	31.3	30.3	39.9	25.2	36.7	31.3	8.0	100.0							
	ρ^0	20.6	48.6	51.2	8.0	27.7	27.5	5.6	17.3	100.0						
	K^*	44.7	73.5	56.3	−4.8	24.9	22.0	9.5	22.6	43.4	100.0					
	S	59.6	71.9	79.7	21.8	39.3	26.9	11.3	35.2	49.4	57.7	100.0				
	f_2	2.4	−10.1	6.3	−56.1	−1.5	3.9	−0.3	10.7	−6.2	−21.5	5.0	100.0			
	f_X	14.5	34.1	12.5	16.1	−23.0	12.4	2.5	34.5	7.3	8.3	12.9	−6.2	100.0		
	NR	17.8	57.6	41.7	12.7	10.1	49.7	2.4	40.0	32.1	25.0	31.7	7.5	46.2	100.0	
	χ	18.9	27.0	30.6	5.8	11.8	9.5	−84.2	21.5	17.8	24.1	27.8	0.8	8.1	20.2	100.0
arg(c)	ρ^0	−11.2	13.3	4.0	−16.1	−2.9	−2.1	−0.5	−0.2	24.1	16.3	3.2	−3.3	8.9	2.1	4.2
	K^*	25.0	8.6	−3.2	−0.2	−15.7	−9.7	6.3	−10.4	−3.9	5.5	16.0	3.8	6.3	−6.5	−3.2
	S	33.0	19.6	3.4	−4.7	−17.3	−16.5	6.2	−9.6	1.0	18.7	21.3	−4.2	9.6	−4.2	1.1
	f_2	12.1	−0.6	−9.8	−2.6	−23.1	−27.4	0.9	−16.7	−7.2	2.2	1.1	−10.6	7.2	−14.1	−2.6
	f_X	25.0	10.2	5.4	−0.5	−11.4	−11.8	1.0	−0.8	2.6	8.5	11.8	−3.8	15.6	2.4	0.4
	NR	31.6	17.0	39.3	1.0	−27.1	−31.7	−6.7	11.3	12.8	14.5	19.0	3.3	21.5	19.6	14.2
	χ	8.6	1.8	9.8	0.6	−9.9	−8.9	−7.9	2.8	3.8	1.3	4.2	3.5	7.3	8.9	12.4
arg($\bar{c}$)	f_0	32.2	11.7	18.9	3.5	−20.3	−26.2	−1.6	−3.6	−6.9	7.3	18.2	1.8	20.3	−7.1	4.3
	ρ^0	14.5	18.0	14.6	−17.3	−13.4	−21.0	−0.7	−8.7	14.3	19.8	13.4	1.7	7.2	−4.4	5.4
	K^*	17.1	7.1	22.0	5.2	−13.5	−17.3	−2.1	5.0	7.2	6.5	13.8	8.1	12.8	29.5	9.6
	S	22.5	15.9	25.2	−3.2	−16.9	−21.6	−0.5	4.2	10.6	17.7	16.1	1.7	14.1	28.8	10.8
	f_2	15.1	4.9	15.5	−5.0	−15.5	−17.9	−2.1	10.0	−2.5	3.9	2.9	11.1	15.7	18.6	7.5
	f_X	8.1	2.7	12.3	−0.6	16.5	−20.4	−0.9	12.2	6.1	3.4	4.8	1.4	−14.6	4.7	6.5
	NR	15.3	4.1	14.5	−3.0	−22.6	−20.8	0.8	1.7	8.2	1.8	5.2	2.6	20.0	15.1	3.2
	χ	10.9	1.1	12.8	0.7	−13.9	−18.0	−4.7	2.1	3.3	0.6	3.9	5.9	9.8	13.4	8.2

		arg(c)							arg($\bar{c}$)							
		ρ^0	K^*	S	f_2	f_X	NR	χ	f_0	ρ^0	K^*	S	f_2	f_X	NR	χ
arg(c)	ρ^0	100.0														
	K^*	10.4	100.0													
	S	18.2	90.9	100.0												
	f_2	19.6	54.1	61.8	100.0											
	f_X	25.5	49.3	56.9	58.1	100.0										
	NR	24.3	17.2	29.9	31.6	47.8	100.0									
	χ	5.0	6.7	7.9	10.2	17.6	30.8	100.0								
arg($\bar{c}$)	f_0	18.0	34.3	42.0	39.8	52.9	55.6	23.8	100.0							
	ρ^0	55.3	22.2	32.4	25.7	36.6	42.2	17.4	58.8	100.0						
	K^*	4.0	21.5	28.0	23.2	36.1	53.9	31.3	46.8	33.5	100.0					
	S	9.6	23.7	35.1	27.8	41.2	60.7	33.3	53.4	42.7	90.9	100.0				
	f_2	5.5	6.4	12.4	1.5	29.3	46.4	23.5	44.1	36.7	56.7	60.8	100.0			
	f_X	1.7	0.0	5.4	13.8	15.5	36.4	19.5	22.2	22.5	42.1	44.8	39.4	100.0		
	NR	7.2	19.2	27.5	28.9	42.3	55.5	32.9	47.3	37.9	63.2	72.5	48.1	48.4	100.0	
	χ	4.1	8.9	13.3	15.5	27.1	43.3	35.9	38.0	26.9	55.9	58.9	40.0	33.6	52.1	100.0

TABLE IX. Full correlation matrix for the isobar parameters of solution II. The entries are given in percent. Since the matrix is symmetric, all elements above the diagonal are omitted.

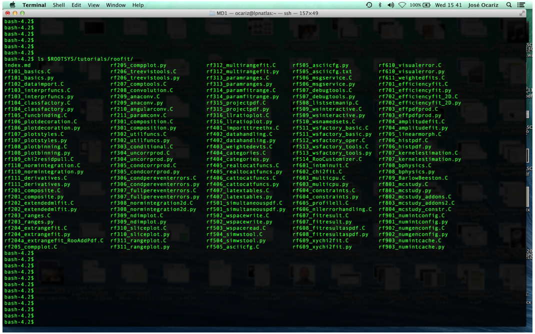
		$\lvert c\rvert$							$\lvert\bar{c}\rvert$							
		ρ^0	K^*	S	f_2	f_X	NR	χ	f_0	ρ^0	K^*	S	f_2	f_X	NR	χ
$\lvert c\rvert$	ρ^0	100.0														
	K^*	46.9	100.0													
	S	49.1	68.2	100.0												
	f_2	8.7	7.7	25.4	100.0											
	f_X	16.8	40.3	38.5	26.6	100.0										
	NR	−8.4	30.2	21.2	9.4	49.9	100.0									
	χ	5.5	11.7	9.3	3.4	12.1	9.1	100.0								
$\lvert\bar{c}\rvert$	f_0	29.2	42.1	50.2	31.5	57.9	34.1	10.0	100.0							
	ρ^0	61.5	68.1	40.4	6.9	20.6	6.4	6.0	31.6	100.0						
	K^*	39.8	75.7	59.8	0.3	33.1	25.3	10.9	33.2	36.3	100.0					
	S	50.6	75.2	83.2	25.4	49.9	33.4	13.1	51.6	46.0	61.4	100.0				
	f_2	0.8	−6.1	9.6	−53.9	6.0	13.3	0.2	14.7	5.3	−18.5	10.4	100.0			
	f_X	10.0	−3.3	−0.9	−10.6	−68.7	−17.8	−5.2	−18.4	6.3	−4.0	−4.9	2.2	100.0		
	NR	23.1	68.8	44.7	13.5	39.3	34.4	5.8	45.6	58.3	32.8	45.4	14.7	−13.8	100.0	
	χ	22.3	33.5	37.8	9.8	19.3	9.9	−79.2	31.3	20.7	30.2	36.1	3.3	−2.6	23.3	100.0
arg(c)	ρ^0	−23.1	13.7	5.5	−11.4	8.0	5.2	0.0	9.0	−11.9	14.5	6.3	−0.2	0.3	3.8	6.9
	K^*	30.6	2.0	−2.2	−6.3	−16.1	−28.1	−0.0	−15.2	14.3	−1.4	6.1	2.0	19.4	−10.3	0.5
	S	38.1	8.9	1.8	−10.1	−17.9	−39.5	−0.1	−15.8	17.4	9.4	7.7	−8.2	19.7	−12.1	3.7
	f_2	18.1	−10.0	−13.7	−7.4	−15.4	−41.3	−2.4	−18.6	1.0	−6.2	−10.2	−12.7	10.7	−21.6	−2.7
	f_X	26.2	−7.8	−12.2	−5.9	−7.7	−35.9	−1.7	−14.5	7.8	−5.7	−8.8	−9.9	12.2	−15.2	−3.6
	NR	32.4	−0.4	21.4	0.5	−29.5	−65.2	−10.4	−4.2	12.0	0.2	0.4	−6.4	21.2	−8.1	10.2
	χ	15.4	−2.2	0.2	−1.6	−9.9	−18.3	−4.9	−5.6	5.6	−3.0	−0.2	−0.8	9.2	−5.6	4.0
arg($\bar{c}$)	f_0	30.1	−8.0	−2.3	−0.9	−13.2	−43.0	−2.8	−16.7	12.1	−7.2	−5.5	−4.9	10.4	−18.7	−1.6
	ρ^0	7.6	11.4	5.8	−7.5	−1.8	−24.7	0.6	−7.5	4.1	15.1	5.5	−12.6	1.3	−7.0	4.0
	K^*	27.0	0.8	7.6	5.6	2.8	−27.8	0.6	−2.0	9.1	1.5	7.1	3.2	−6.9	13.9	4.1
	S	32.6	8.0	8.4	−1.1	0.6	−31.3	2.1	−4.1	12.6	12.1	7.6	−5.6	−4.4	12.2	4.7
	f_2	18.7	1.7	6.6	10.1	9.8	−22.9	0.7	7.6	8.6	3.5	0.6	−5.6	−21.6	9.3	4.6
	f_X	21.9	1.8	4.4	9.6	−0.7	−30.2	0.1	−5.0	8.1	2.8	4.0	−17.3	1.0	−6.6	−0.2
	NR	27.7	−1.9	−3.0	3.9	−0.5	−30.7	2.8	−13.3	7.8	−1.5	−1.2	−13.8	−7.2	−3.7	−5.0
	χ	19.7	−5.0	−0.5	2.3	−4.4	−27.6	2.7	−6.1	6.2	−4.1	−2.5	−1.6	−0.2	−0.1	−2.9

		arg(c)							arg($\bar{c}$)							
		ρ^0	K^*	S	f_2	f_X	NR	χ	f_0	ρ^0	K^*	S	f_2	f_X	NR	χ
arg(c)	ρ^0	100.0														
	K^*	2.9	100.0													
	S	7.4	90.6	100.0												
	f_2	9.9	56.6	65.5	100.0											
	f_X	5.9	57.0	64.4	69.5	100.0										
	NR	10.1	37.0	50.3	44.4	46.6	100.0									
	χ	2.6	39.3	40.3	29.1	31.3	28.6	100.0								
arg($\bar{c}$)	f_0	−0.6	45.8	53.5	47.1	61.0	51.9	27.4	100.0							
	ρ^0	41.3	29.5	39.2	31.2	39.1	33.0	16.5	54.9	100.0						
	K^*	−11.6	35.2	39.7	30.4	42.7	30.0	17.6	56.0	32.9	100.0					
	S	−8.7	38.8	47.7	36.1	49.1	33.7	19.5	62.4	41.1	91.1	100.0				
	f_2	−5.4	12.2	17.9	7.0	28.5	27.2	9.9	52.8	42.0	59.3	61.6	100.0			
	f_X	−7.0	23.2	28.6	28.0	34.4	29.9	15.4	34.6	30.2	43.3	47.1	41.5	100.0		
	NR	−9.0	41.4	47.9	44.2	59.5	30.9	25.2	68.9	48.1	68.6	77.6	54.6	55.8	100.0	
	χ	−7.3	29.3	33.3	28.8	38.8	29.9	8.8	47.1	26.9	54.7	58.0	38.6	35.8	54.6	100.0

El paquete RooFit

- Desarrollado a partir del 2005 en el experimento *BABAR*
- Viene incluido en ROOT desde Root4
- Utilizado en muchos análisis experimentales de altas energías
- Ventajas :
  - Minimización por MINUIT
  - Normalización automática de las PDFs, analítica si la conoce, numérica de lo contrario
  - Numerosas clases y ejemplos

- Inconvenientes :
  - Documentación parcial y no actualizada
  - Normalización automática de las PDFs...



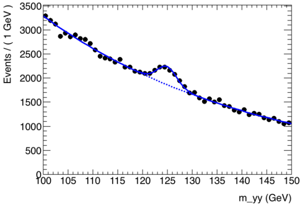
En un escenario típico, un proceso aleatorio puede tener contribuciones de origen diferente.
Para ser específicos, consideremos que los eventos que componen la muestra provienen de dos "especies", llamadas de manera genérica "señal" y "fondo" (la generalización a más de dos especies es sencilla).
Cada especie se realiza a partir de su propia densidad de probabilidad.
Si los rangos de las variables aleatorias no son totalmente disyuntos, es imposible saber evento a evento a cuál de las especies pertenece. Pero el MLE permite efectuar una *separación estadística*: la PDF subjacente es a combinación de mas PDFs de señal y fondo,

$$\mathcal{L}\left(f_{\text{sig}}, \theta; \vec{x}\right) = \prod_{i=1}^{N} \left[f_{\text{sig}} P_{\text{sig}}(\vec{x};\theta) + (1 - f_{\text{sig}}) P_{\text{bkg}}(\vec{x};\theta) \right] ,$$

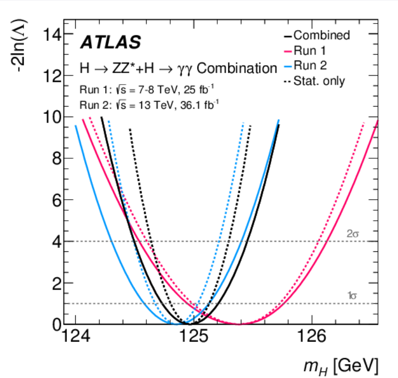
donde P_{sig} y P_{bkg} son las PDFs de señal y fondo, respectivamente, y la fracción de señal f_{sig} es el parámetro que cuantifica la pureza de la muestra : $0 \leq f_{\text{sig}} \leq 1$.

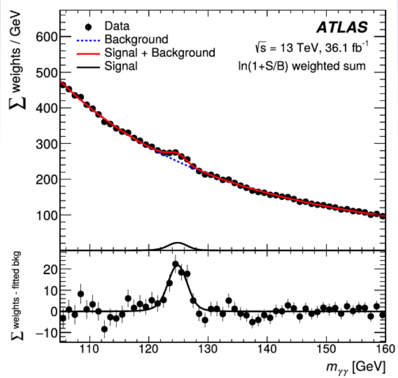
Ejemplo inspirado de la búsqueda del bosón de Higgs en el canal difotón :
Con ROOT instalado, la macro `H_yy.cc` debe correr sin problema, haciendo
`prompt> root -l H_yy.cc`

```
RooFitResult: minimized FCN value: 386054, estimated distance to minimum: 2.29788e-05
              covariance matrix quality: Full, accurate covariance matrix
              Status : MIGRAD=0 HESSE=0

  Constant Parameter    Value
  --------------------  ------------
               sig_s    2.0000e+00

  Floating Parameter  InitialValue    FinalValue +/-  Error     GblCorr.
  --------------------  ------------  --------------------------  --------
               bkg_a   -2.2500e-02   -2.2780e-02 +/-  2.31e-04  <none>
               sig_f    2.0000e-02    1.8732e-02 +/-  1.33e-03  <none>
               sig_m    1.2500e+02    1.2494e+02 +/-  1.92e-01  <none>
```



Figuras tomadas de M. Aaboud *etal*, the ATLAS Collaboration, Phys.Lett.B 784 (2018) 345-366





La figura que representa $-2\Delta(\ln\mathcal{L})$ en función de m_H ilustra claramente la interpretación del MLE en términos de *intervalos de confianza*:

- el rango de m_H que corresponde a $-2\Delta(\ln\mathcal{L}) < 1$, "un sigma", cubre 68 % de los resultados que se obtendrían repitiendo el experimento ATLAS numerosas veces;
- idem para el rango de m_H correspondiendo a $-2\Delta(\ln\mathcal{L}) < 4, 9, \cdots$ ("dos sigma", "tres sigma", etc...)

MLE en situaciones más elaboradas (III)

En experimentos de conteo de eventos, el número de eventos observados puede ser un parámetro de interés. Para el caso de una especie única, esto corresponde a “extender” la verosimilitud,

$$\mathcal{L}(\lambda, \theta; \vec{x}) = \frac{\lambda^N e^{-\lambda}}{N!} \prod_{i=1}^{N} P(\vec{x}_i; \theta) \ .$$

dónde el término multiplicativo adicional, corresponde a la distribución de Poisson (el término $N!$ en el denominador es irrelevante; es un factor global sin impacto sobre la forma de la verosimilitud)

Es fácil verificar que la verosimilitud es máxima cuand $\hat{\lambda} = N$, tal como se espera; ahora, si algunas de las PDFs dependen también de λ, el valor $\hat{\lambda}$ que maximiza $\mathcal{L}$ puede diferir.

La generalización a más de una especie es sencilla; para cada especie, un término multiplicativo de Poisson se incluye en la verosimilitud extendida, y las PDFs de cada especie son ponderadas por su fracción relativa de eventos.

Para el caso de dos especies, la versión extendida de la verosimilitud es

$$\mathcal{L}(N_{\mathrm{sig}}, N_{\mathrm{bkg}}\theta; \vec{x}) = (N_{\mathrm{sig}} + N_{\mathrm{bkg}})^N e^{-(N_{\mathrm{sig}}+N_{\mathrm{bkg}})} \prod_{i=1}^{N} \left[N_{\mathrm{sig}} P_{\mathrm{sig}}(\vec{x}; \theta) + N_{\mathrm{bkg}} P_{\mathrm{bkg}}(\vec{x}; \theta)\right] \ .$$

Estimar eficiencias a partir de ajustes MLE

Consideremos de nuevo el caso de un proceso aleatorio con dos resultados posibles: "yes" y "no". El estimador intuitivo de la eficiencia ε es el cociente entre el número de realizaciones de cada tipo, n_{yes} y n_{no}, y su varianza $V[\hat{\varepsilon}]$ viene dada por :

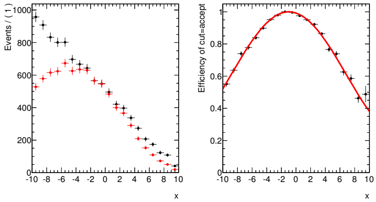
$$\hat{\varepsilon} = \frac{n_{\text{yes}}}{n_{\text{yes}} + n_{\text{no}}} \, , \; V[\hat{\varepsilon}] = \frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{n} \, ,$$

donde $n = n_{\text{yes}} + n_{\text{no}}$ es el número total de realizaciones. (ejercicio: reproducir este resultado)
Este estimador ingenuo $\hat{\varepsilon}$ claramente falla para pequeños valores de n, y en las situaciones de gran (in-)eficiencia.
La técnica del MLE provee una solución robusta para la estimación de eficiencias: si la muestra contiene una variable x sensible a la eficiencia (es decir $\varepsilon(x)$), que sigue la PDF $P(x;\theta)$, entonces la inclusión de una nueva variable aleatoria discreta y bivariada $c = \{\text{yes}, \text{no}\}$, nos da un modelo más elaborado:

$$P(x, c; \theta) = \delta(c - \text{yes})\varepsilon(x,\theta) + \delta(c - \text{no})\left[1 - \varepsilon(x,\theta)\right] \, .$$

- la función $\varepsilon(x)$ ha sido correctamente normalizada, para ser también una PDF
- la eficiencia ya no es un valor único, sino una función de x
- (y de otros parámetros θ que sean necesarios para caracterizar su forma)

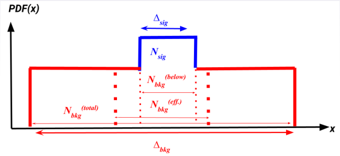
($ROOTSYS/tutorials/roofit/rf701_efficiencyfit.C)



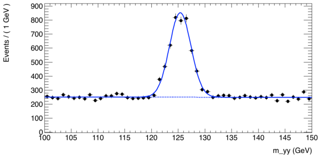
Sobre el fondo efectivo

De manera general, la zona que contiene la señal también está contaminada por fondo(s). El impacto de esos fondos se traduce en una *dilución* o disminución de la precisión en la medida de los parámetros de interés.
Ejercicio : Consideremos un escenario compuesto por dos especies: una señal, distribuida de manera uniforme en un intervalo $\Delta(\mathrm{sig})$, y un fondo distribuido de manera uniforme en un intervalo $\Delta(\mathrm{bkg})$ más amplio que cubre ambos lados del intervalo de señal ("sidebands").

- Generar una realización aleatoria, eligiendo valores particulares de los intervalos $\Delta(\mathrm{sig})$ y $\Delta(\mathrm{bkg})$, y del número de eventos de señal y fondo N_{sig} y N_{bkg}.
- El parámetro de interés es el número de eventos de señal N_{sig}. Estimar su valor y error $\hat{N}_s \pm \hat{\sigma_s}$ por verosimilitud máxima.
- Si el fondo fuera nulo, se tendría $\hat{\sigma_s} = \sqrt{N_s}$.
- Definir el "fondo efectivo" como el causante del aumento en el error, $\hat{\sigma_s} = \sqrt{N_s + N_{\mathrm{bkg}}^{\mathrm{eff}}}$.
- Comparar $N_{\mathrm{bkg}}^{\mathrm{eff}}$ al fondo "debajo" de la señal, $N_{\mathrm{bkg}}^{\mathrm{below}}$.
- Repitiendo el ejercicio para diferentes valores de N_{sig}, verificar que el fondo efectivo es siempre el mismo.
- Repitiendo el ejercicio para varios valores crecientes de $\Delta(\mathrm{bkg})$, verificar que el fondo efectivo tiende a $N_{\mathrm{bkg}}^{\mathrm{below}}$.
- Interpretar.



Ejercicio: Para una señal Gaussiana, realizar un ejercicio similar, con dos parámetros de interés adicionales: la posición del pico y su anchura.

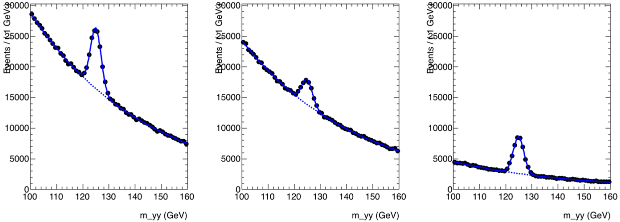


Categorización

En ocasiones, la muestra de análisis puede descomponerse en dos o más submuestras ("categorías"), cada una de ellas con sus propias PDFs y purezas.

Cuando las características de cada especie son razonablemente diferentes, puede ser de interés descomponer la función de verosimilitud de tal manera que cada categoría utilice sus propias PDFs y purezas. El resultado combinado sobre un parámetro común de interés tendrá una significación superior a la que se obtendría de un análisis "inclusivo", es decir usando PDFs y purezas promedio sobre la muestra completa.

Ejemplo sencillo (y ejercicio): supongamos que la muestra $H \to \gamma\gamma$ se descompone en dos categorías: una "limpia" con excelente cociente señal/fondo, y una "sucia" en la que el fondo es ampliamente dominante. Si el parámetro de interés es la masa del Higgs, la ventaja de realizar un análisis en categorías puede ser significativo.



Aquí las dos categorías difieren en el cociente señal/fondo, grande para la "limpia", y pequeño para la "sucia". Otra posibilidad es aprovechar diferencias en resolución.

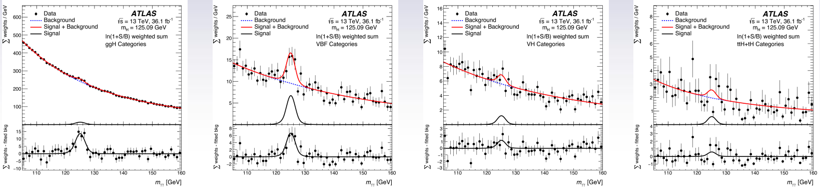
Error en el análisis inclusivo: $\sigma(m_H) = \pm 2{,}25\,\%$

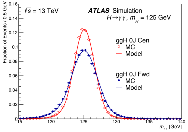
Errores en las categorías sucia y limpia : $\pm 3{,}70\,\%$ y $\pm 1{,}85\,\%$, respectivamente.

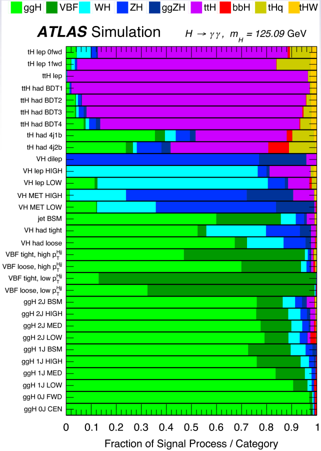
Error en la combinación de ambas categorías: $\pm 1{,}72\,\%$! Equivale a aumentar en $70\,\%$ la estadística inclusiva !

El uso de categorías en el análisis $H \to \gamma\gamma$

ATLAS separa su muestra de candidatos difotón en 31 categorías, definidas en función de varios criterios: el modo de producción del Higgs, diferencias en la resolución experimental en masa, diferencias en la pureza.







En el 2012, la categorización fue crucial para alcanzar los 5 σ de significancia en la observación del Higgs...