Theory / Hour-6

MAXIMUM LIKELIHOOD METHOD for FITTING

MAXIMUM LIKELIHOOD Method

In HEP practise the most frequently adopted ***parameter estimation method*** is based on the construction of the **combined probability distribution of all measurements** in our data sample, called ***likelihood function***.

The estimate of the parameters we want to determine is obtained by finding the parameter set that corresponds to the maximum value of the likelihood function. This approach takes the name of ***maximum likelihood method***.

The procedure is also called ***best fitting*** because it determines (estimates) the parameters for which the theoretical pdf model *best fits* the experimental data sample.

Maximum likelihood fits are every day used - in HEP data analysis - because of the very good statistical properties characterizing the maximum likelihood estimators. It is better than the *chi-squared method* that has the limitation to deal only with *binned* data (not *unbinned* as the ML method can also do (*)) and does not behave well when in some bins there are few entries. The ***chi-squared method*** was used in **`ROOT`** before the development of **`RooFit`** .

(*) Remember that - in **`RooFit`** - the ML method can work on both two classes of data:
`RooDataHist` (histograms i.e. binned data) & **`RooDataSet`** (unbinned data).

LIKELIHOOD Function - I

The ***likelihood function*** is the function that, for given values of the unknown parameters, returns the value of the pdf evaluated at the observed data sample.

If the measured values of n r.v.s are $(x_1, \dots, x_n)$

and our pdf model depends on $\mathbf{m}$ unknown parameters $(\theta_1, \dots, \theta_m)$

... the likelihood function is: $L(x_1, \dots, x_n; \theta_1, \dots, \theta_m) = f(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$

JOINT pdf of the r.v.s $(x_1, \dots, x_n)$

The ***maximum likelihood estimator*** of the unknown parameters $(\theta_1, \dots, \theta_m)$ is the function that returns the values of the parameters (called ***ML best estimates***) $(\hat{\theta}_1, \dots, \hat{\theta}_m)$ for which the likelihood function, evaluated at the measured sample is maximum!

LIKELIHOOD Function - II

If we have N repeated measurements, each consisting of the n values of the random variables $(x_1, \dots, x_n)$, the likelihood function is the probability density corresponding to the total sample $\vec{x} = \{(x_1^{N=1}, \dots, x_n^{N=1}), \dots, (x_1^N, \dots, x_n^N)\}$.

If the observations are independent of each other (*), the **likelihood function** of the total sample consisting of the N events recorded by our experiment can be written as the product of the pdfs corresponding to the measurement of each single event:

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{N} f\,(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m)$$

(*) In physics often the word ***event*** is used with a different meaning w.r.t. statistics and it refers to a collection of measurements of observable quantitie $(x_1, \dots, x_n)$ corresponding to a physical phenomenon, like a collision of particles at an accelerator, or the interaction of a particle, or a shower of particles from cosmic rays, in a detector.

Measurements performed at different events are tipically uncorrelated and each sequence of variables taken from N different events can be considered a sampling of independent and identically distributed random variables.

LIKELIHOOD Function - III

Often the logarithm of the likelihood function is computed so that the product of many terms can be transformed into a sum of the logarithm. Moreover, instead of maximizing the likelihood function, it is often more convenient to minimize the ***Negative Log-Likelihood***:

$$NLL \equiv -lnL(\vec{x};\vec{\theta}) = -\sum_{i=1}^{N} lnf(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m)$$

[sometimes $NLL \equiv -2lnL(\vec{x};\vec{\theta})$]

Its minimization can be performed analitically only in the simplest cases.

In most of the realistic cases the NLL minimization requires numerical methods implemented as computer algorithms.

The software **MINUIT** (F. James *et al.*) [*] is one of the most widely used minimization tool in the HEP field since the 1970s.

The minimization is based on the steepest descent direction in the parameter space, which is determined based on a numerical evalution of the gradient of (logarithm of) the likelihood function.

MINUIT has been re-implemented from the original Fortran version into C^{++} and is available in the **ROOT** software tookit.

[*] CERN Program Library Long Writeup D506
http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html

Example: Gaussian Likelihood Function

For N repeated measurements of a r.v. $\mathbf{x}$ distributed according to a Gaussian function with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, twice the NLL function can be written as:

$$2NLL \equiv -2lnL\left(\vec{x} \equiv (x_1, \dots, x_i, \dots x_N; \vec{\theta} \equiv (\mu, \sigma^2)\right) = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} + N[\ln(2\pi) + 2\ln(\sigma)]$$

Its minimization can be performed analitically by finding the zeros of the first partial derivatives of -2lnL w.r.t. the 2 parameters (*)!

The following maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can be thus obtained: $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ & $\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$

The maximum likelihood estimate $\widehat{\sigma^2}$ is affected by a bias, in the sense that its mean deviates from the true $\boldsymbol{\sigma}^2$.

It can be calculated that the bias is $\frac{N-1}{N}$ thus vanishes in the limit $N \to \infty$ (asymptotically unbiased estimator).

A fully unbiased estimator can be then obtained by introducing the suitable correction factor: $\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$

(*) $\frac{\partial NLL}{\partial \mu} = 0$ & $\frac{\partial NLL}{\partial \sigma^2} = 0$

EXTENDED Likelihood Function - I

If the number of recorded events N is also a random variable that tipically follows a poissonian distribution (stocastic phenomenon) whose exp. value μ may also depend on the m unknown parameters, the ***extended likelihood function*** can be defined as:

$$L\left(\vec{x}, N; \vec{\theta}\right) = \boldsymbol{P(N; \vec{\theta})} \cdot \prod_{i=1}^{N} f\left(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m\right)$$

The extended likelihood function exploits the number of recorded events as information in order to determine the parameters' estimate, in addition to the data sample. By expliciting the poissonian function (with exp. value μ) we can write:

$$L\left(\vec{x}, N; \vec{\theta}\right) = \frac{e^{-\mu} \cdot \mu^N}{N!} \cdot \prod_{i=1}^{N} f\left(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m\right)$$

where $\mu = \mu(\vec{\theta})$

EXTENDED Likelihood Function - II

Let us consider the typical case where the pdf is a linear combination of two pdfs, one for the «signal», f_S , and one for the «background», f_B :

$$L(\vec{x}, N; s, b, \theta) = \frac{e^{-(s+b)} \cdot (s+b)^N}{N!} \cdot \prod_{i=1}^{N} [w_S f_S(x_i; \theta) + w_B f_B(x_i; \theta)]$$

where the **fractions of the signal & background** are: $w_S = \frac{s}{s+b}$ & $w_B = \frac{b}{s+b}$

Note that $w_S + w_B = 1$, hence $f = [w_S f_S + w_B f_B]$ is normalized, assuming that f_S & f_B are normalized.

More compactly:

$$L(\vec{x}, N; s, b, \theta) = \frac{e^{-(s+b)}}{N!} \cdot \prod_{i=1}^{N} [\boldsymbol{s} f_S(x_i; \theta) + \boldsymbol{b} f_B(x_i; \theta)]$$

…where s & b are **yields**!

The logarithm of the likelihod functioon provides a more convenient expression:

$$-lnL(\vec{x}, N; s, b, \theta) = \boldsymbol{s} + \boldsymbol{b} + \sum_{i=1}^{\boldsymbol{N}} \boldsymbol{ln}[s f_S(x_i; \theta) + b f_B(x_i; \theta)] - \cancel{ln(N!)}$$

can be omitted in the minimization (since it is constant w.r.t. the parameters)

Note the difference with the *non-extended case* for which : $L(\vec{x}, ; w_S, \theta) = \prod_{i=1}^{N} [w_S f_S(x_i; \theta) + (1 - w_S) f_B(x_i; \theta)]$

…where w_S is the **signal fraction**!

Example of application of an EXTENDED Likelihood Function - I

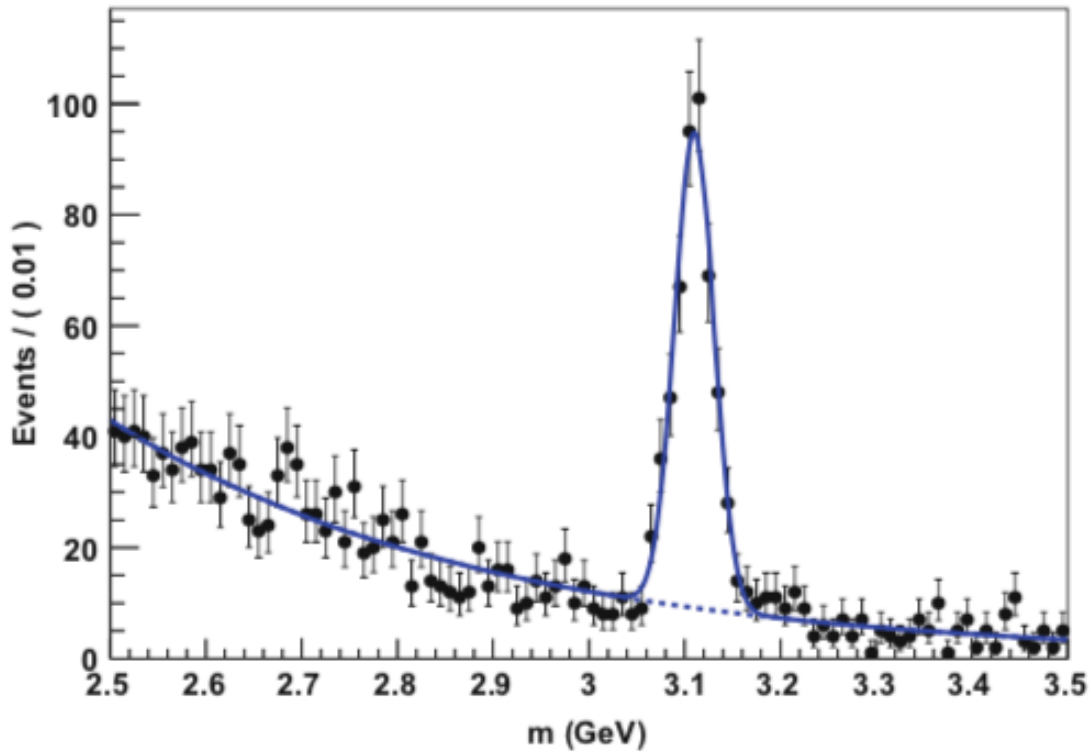


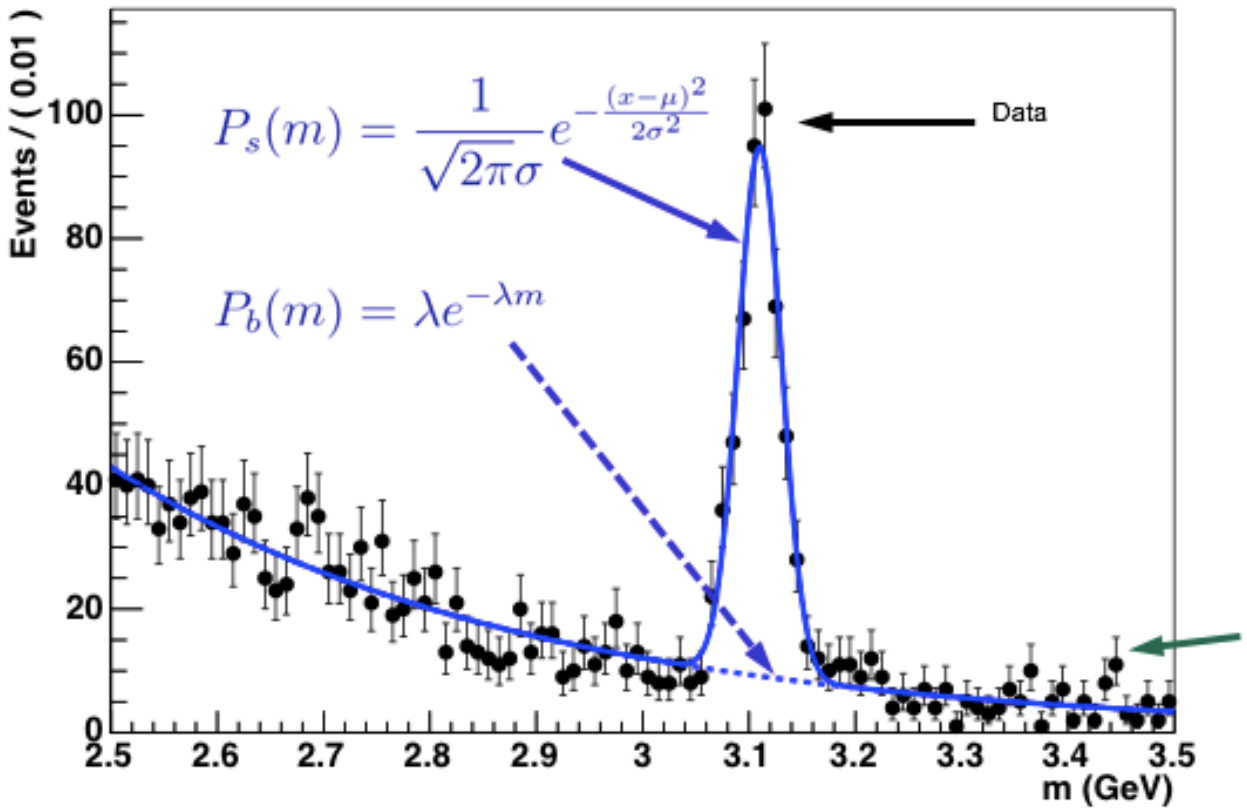
Fig. 5.2 Example of unbinned extended maximum likelihood fit of a simulated dataset. The fit curve is superimposed to the data points (*black dots with error bars*) and shown as *solid blue line*

by L.Lista

Example of application of an EXTENDED Likelihood Function - II

- $P_s(m)$: Gaussian peak
- $P_b(m)$: exponential shape

Exponential decay parameter λ, Guassian mean μ and standard deviation σ can be fit together with sig. and bkg. yields s and b.



The additional parameters, beyond the parameters of interest (s in this case), used to model background, resolution, etc. are examples of nuisance parameters

In the plot, data are accumulated into **bins** of a given width

Error bars usually represent uncertainty on each bin count (in this case: Poissonian)

by L.Lista

[see details in the hands-on sessions]

Variance of ML Estimators : the RCF bound (lowest variance) - I

It turns out in many applications to be too difficult to compute the variances analytically, and a Monte Carlo study usually involves a significant amount of work. In such cases one typically uses the **Rao–Cramér–Frechet (RCF) inequality**, also called the **information inequality**, which gives a lower bound on an estimator's variance. This inequality applies to any estimator, not only those constructed from the ML principle. For the case of a single parameter θ the limit is given by

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right], \qquad (6.16)$$

where b is the bias as defined in equation (5.4) and L is the likelihood function. A proof can be found in [Bra92]. Equation (6.16) is not, in fact, the most general form of the RCF inequality, but the conditions under which the form presented here holds are almost always met in practical situations (cf. [Ead71] Section 7.4.5). In the case of equality (i.e. minimum variance) the estimator is said to be **efficient**. It can be shown that if efficient estimators exist for a given problem, the maximum likelihood method will find them. Furthermore it can be shown that ML estimators are always efficient in the large sample limit, except when the extent of the sample space depends on the estimated parameter. In practice, one often assumes efficiency and zero bias. In cases of doubt one should check the results with a Monte Carlo study.

by G.Cowan

Variance of ML Estimators : the RCF bound (lowest variance) - II

The variance $V[\hat{\theta}]$ of any consistent estimator is subject to a lower bound due to Cramér [1] and Rao [2] which is given by:

$$V[\hat{\theta}] \geq V_{\mathrm{CR}}(\hat{\theta}) = \frac{\left(1 + \frac{\partial b(\hat{\theta})}{\partial \theta}\right)^2}{\left\langle \left(\frac{\partial \log L(x_1, \cdots, x_n; \theta)}{\partial \theta} \right)^2 \right\rangle} , \tag{5.6}$$

where $b(\hat{\theta})$ is the bias of the estimator (Eq. (5.5)) and the denominator is the *Fisher information*, already defined in Sect. 3.7.

The ratio of the Cramér–Rao bound to the estimator's variance is called estimator's *efficiency*:

$$\varepsilon(\hat{\theta}) = \frac{V_{\mathrm{CR}}(\hat{\theta})}{V[\hat{\theta}]} . \tag{5.7}$$

Any consistent estimator $\hat{\theta}$ has efficiency $\varepsilon(\hat{\theta})$ lower or equal to one, due to Cramér–Rao bound.

Uncertainties with the ML Method - I

Once the estimate $\hat{\theta}$ of a parameter θ is determined using the ML method, a confidence interval needs to be determined.

Two approximate methods to determine the parameters' uncertainties are presented (for both cases the coverage is only approximately ensured; the required coverage is, in most cases, equal to 68.27%, correspoding to 1σ).

1. Second Derivatives Matrix

- A parabolic approximation of $-2\ln L$ around the minimum is equivalent to a Gaussian approximation
  - Sufficiently accurate in many but not all cases

$$-2\ln L = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} + \text{const.}$$

- Estimate of the covariance matrix from 2nd order partial derivatives w.r.t. fit parameters at the minimum:

$$V_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\theta_k = \hat{\theta}_k}$$

- Implemented in Minuit as MIGRAD/HESSE function

this Covariance Matrix gives an n-dimensional elliptic confidence contour (having the correct coverage only if the pdf model is exactly Gaussian)

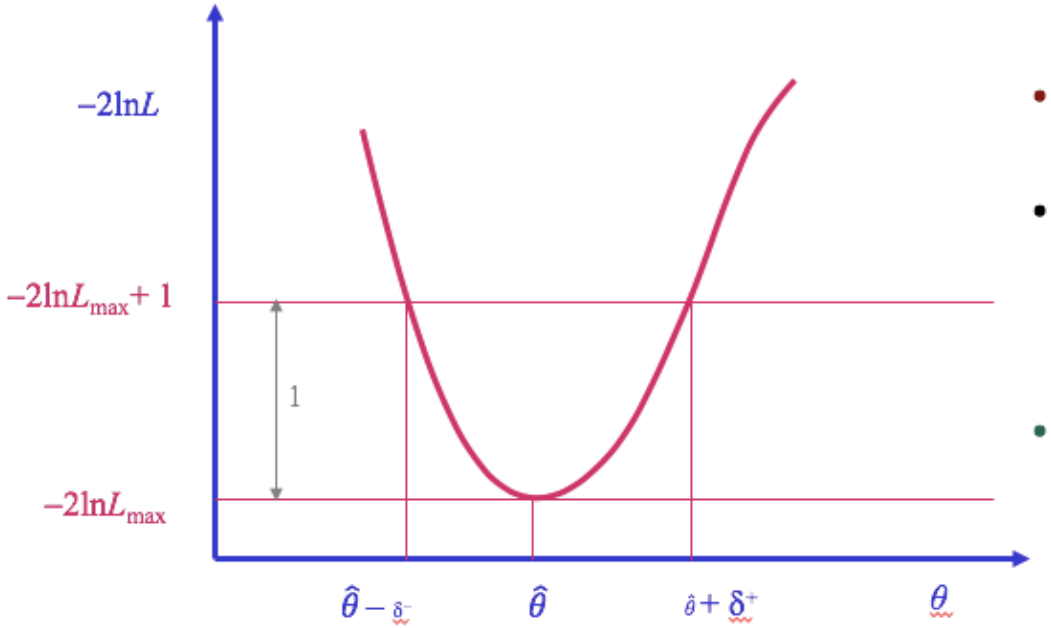
Uncertainties with the ML Method - II

2. Likelihood Scan

Another often used method is to consider a scan of -2lnL around the minimum value $-2\ln L_{max}$ (that is corresponding to the parameter set that maximizes L : $L_{MAX} = L\left(\vec{x}, \hat{\vec{\theta}}\right)$).

An interval corresponding to an increase of -2lnL by 1 unit w.r.t. its minimum value can be determined as graphically shown for a single parameter:

- Error ($n\sigma$) determined by the range around the maximum for which $-2\ln L$ increases by $+1$ ($+n^2$ for $n\sigma$ intervals)



- Errors can be asymmetric
- For a Gaussian PDF the result is identical to the 2nd order derivative matrix
- Implemented in Minuit as MINOS function

This method leads to identical errors as those in the covariance matrix only in the Gaussian case, in which -2lnL has an exact parabolic shape !

by L.Lista

Uncertainties with the ML Method - III

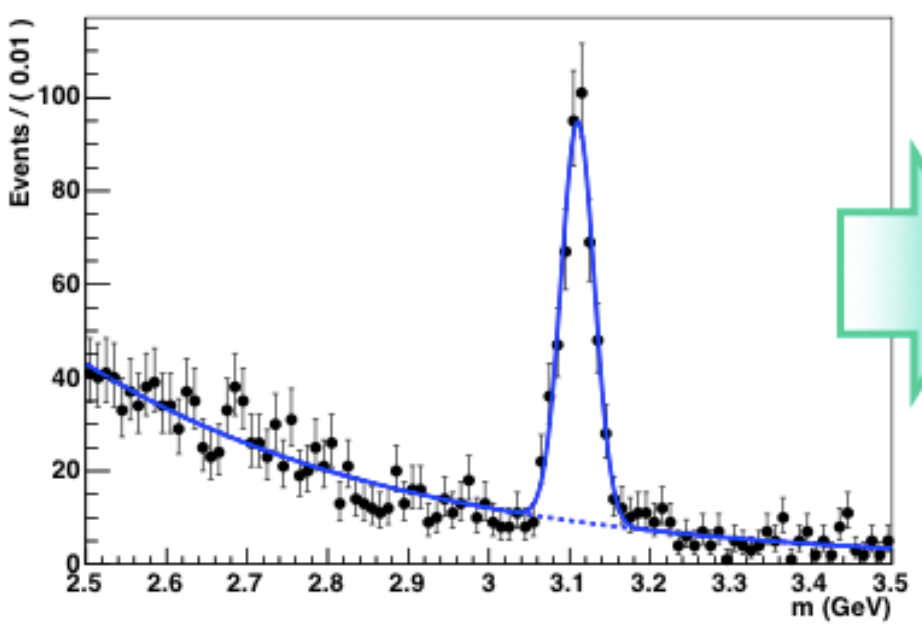
For more than one parameter, the ***error contour*** corresponds to the set of parameter values $\vec{\theta}$ such that:

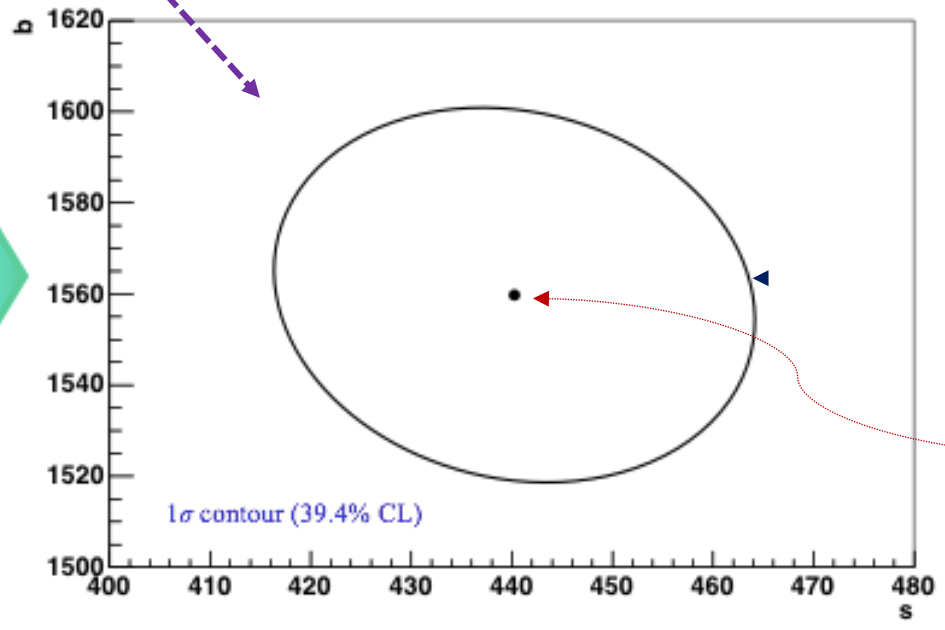
$$-2lnL(\vec{\theta}) = -2lnL_{MAX} + 1$$

$$-2lnL(\vec{\theta}) = -2lnL_{MAX} + Z^2 \text{ (for } Z\sigma)$$

2D *error contour* plot showing the 1σ uncertainty ellipse for s & b:

The contour shows for this case a mild correlation between s and b





$-2lnL_{MAX} + 1$

$-2lnL_{MAX}$

In case of very large number of measurements, computing the likelihood can be numerically unpractical: use binning !