Statistical methods in LHC data analysis *part I.1*

Luca Lista

INFN Napoli

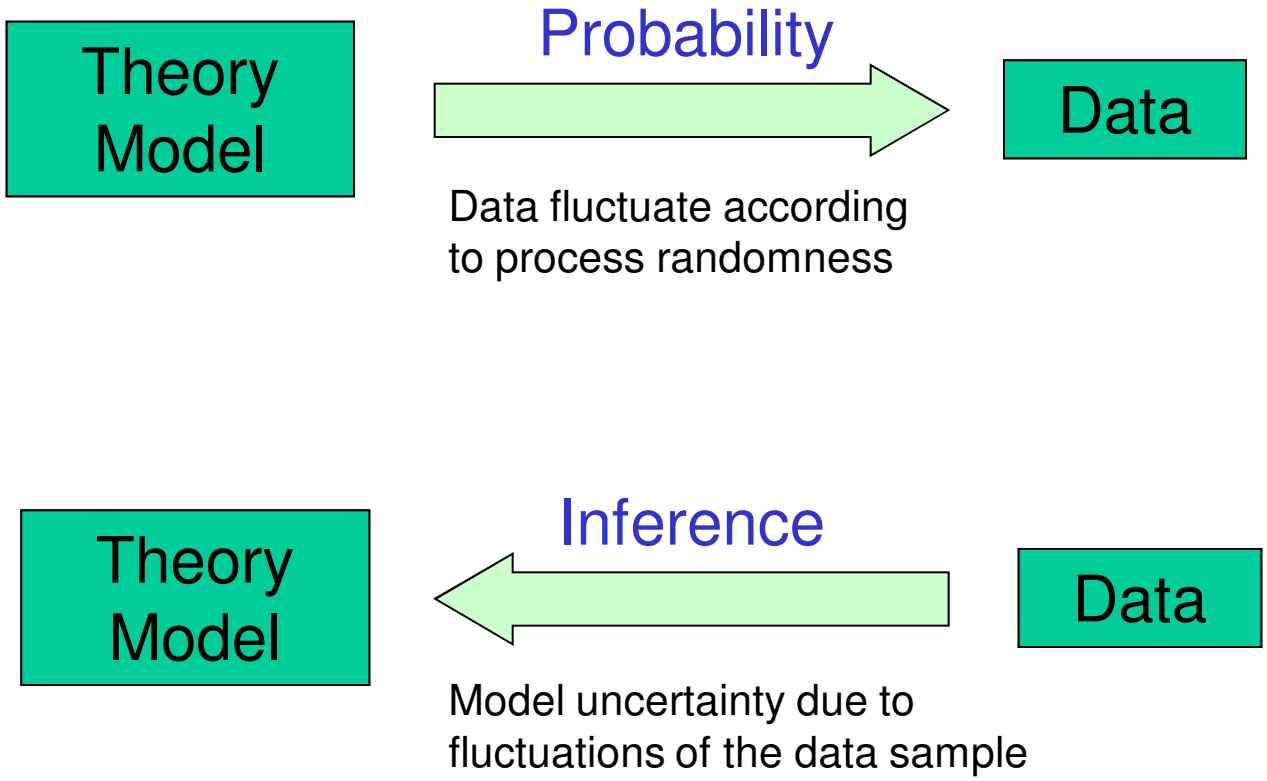
Contents

- Parameter estimates
- Likelihood function
- Maximum Likelihood method
- Problems with asymmetric errors

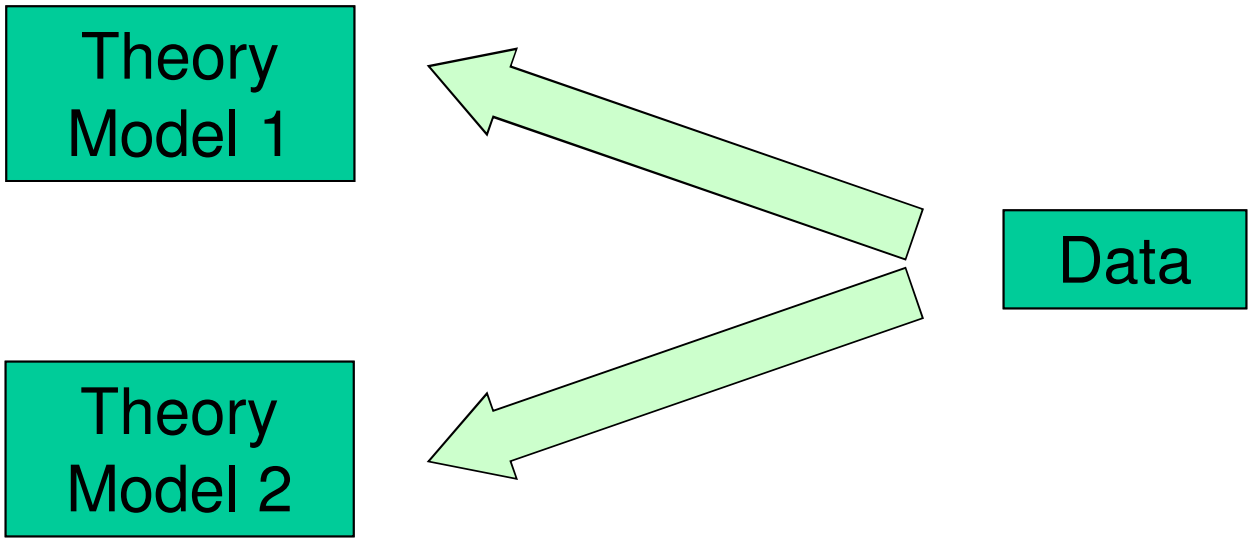
Meaning of parameter estimate

- We are interested in some physical unknown parameters
- Experiments provide samplings of some PDF which has among its parameters the physical unknowns we are interested in
- Experiment's results are statistically "related" to the unknown PDF
  - PDF parameters can be determined from the sample within some approximation or uncertainty
- Knowing a parameter within some error may mean different things:
  - **Frequentist**: a large fraction (68% or 95%, usually) of the experiments will contain, in the limit of large number of experiments, the (fixed) unknown true value within the quoted confidence interval, usually $[\mu - \sigma, \mu + \sigma]$ ('coverage')
  - **Bayesian**: we determine a degree of belief that the unknown parameter is contained in a specified interval can be quantified as 68% or 95%
- We will see that there is still some more degree of arbitrariness in the definition of confidence intervals…

Statistical inference



Hypothesis tests



Which hypothesis is the most consistent with the experimental data?

Parameter estimators

- An estimator is a function of a given sample whose statistical properties are known and related to some PDF parameters
  - "Best fit"
- Simplest example:
  - Assume we have a Gaussian PDF with a *known* σ and an *unknown* μ
  - A single experiment will provide a measurement x
  - We estimate μ as $\mu^{est} = x$
  - The distribution of μ^{est} (repeating the experiment many times) is the original Gaussian
  - 68.27%, *on average*, of the experiments will provide an estimate within: $\mu - \sigma < \mu^{est} < \mu + \sigma$
- We can determine: $\mu = \mu^{est} \pm \sigma$

Likelihood function

- Given a sample of N events each with variables $(x_1, \ldots, x_n)$, the likelihood function expresses the probability density of the sample, as a function of the unknown parameters:

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

- Sometimes the used notation for parameters is the same as for conditional probability:

$$f(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m)$$

- If the size N of the sample is also a random variable, the extended likelihood function is also used:

$$L = p(N; \theta_1, \cdots, \theta_m) \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)$$

  - Where p is most of the times a Poisson distribution whose average is a function of the unknown parameters

- In many cases it is convenient to use $-\ln L$ or $-2\ln L$: $\prod_i \to \sum_i$

Maximum likelihood estimates

- ML is the widest used parameter estimator
- The "best fit" parameters are the set that maximizes the likelihood function
  - "Very good" statistical properties, as will be seen in the following
- The maximization can be performed analytically, for the simplest cases, and numerically for most of the cases
- Minuit is historically the most used minimization engine in High Energy Physics
  - F. James, 1970's; rewritten in C++ recently

Extended likelihood

- In case of Poissonian signal and background processes:

$$L(x_i; s, b, \theta) = \frac{(s+b)^n e^{-(s+b)}}{n!} \prod_{i=1}^{n} (f_s P_s(x_i;\theta) + f_b P_b(x_i;\theta))$$

$$f_s = \frac{s}{s+b} \qquad f_b = \frac{b}{s+b}$$

$$= \frac{e^{-(s+b)}}{n!} \prod_{i=1}^{n} (s P_s(x_i;\theta) + b P_b(x_i;\theta))$$

- We can fit simultaneously s, b and θ minimizing:

constant!

$$-\ln L = s + b + \sum_{i=1}^{n} \ln(s P_s(x_i,\theta) + b P_b(x_i,\theta)) - \ln n!$$

Gaussian approximation

- If we have n measurements whose PDFs are identical and Gaussian, we have:

$$-2\ln L = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} + n\left(\ln 2\pi + 2\ln\sigma\right)$$

- Performing an analytical minimization on μ (assuming σ is known) of $-2\ln L$ we obtain:

$$\mu^{\text{est}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- If σ^2 is also unknown, the ML estimate of σ^2 is:

$$\sigma^{2\,\text{est}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu^{\text{est}}\right)^2$$

- The above estimate can be demonstrated to have a *bias*

Estimator properties

- Consistency
- Bias
- Efficiency
- Robustness

Estimator consistency

- The estimator converges to the true value (in probability)

$$\forall \varepsilon \quad \lim_{n \to \infty} P\left(\left| \theta_n^{\text{est}} - \theta \right| < \varepsilon \right) = 1$$

- ML estimators are consistent

Bias of parameter estimators

- The bias is the average deviation of the estimate from the true parameter:

$$b(\theta) = \left\langle \theta^{\text{est}} - \theta \right\rangle = \left\langle \theta^{\text{est}} \right\rangle - \theta$$

Bias of the parameter θ

- ML estimators may have a bias, but the bias decreases as n increases (if the fit model is correct…!)
- E.g.: in the case of the estimate of a Gaussian's σ^2 the un-biased estimate is the well known:

$$\sigma^{2\,\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu^{\text{est}})^2$$

Efficiency of the estimator

- The variance of any consistent estimator is subject a lower bound (Cramér-Rao bound):

$$V(\theta^{\text{est}}) \geq \frac{\left(1 + \frac{\partial b(\theta)}{\partial \theta}\right)^2}{\left\langle \left(\frac{\partial \ln L(x_1, \cdots x_n; \theta)}{\partial \theta} \right)^2 \right\rangle}$$

Fisher information

- Efficiency = ratio of Cramér-Rao bound over variance
- Efficiency for ML estimators is asymptotically 1
  - No asymptotically unbiased estimator has asymptotic mean squared error smaller than the ML est.

Robustness

- If the sample distribution has (slight?) deviations from the theoretical PDF model, some estimators may deviate more or less than others from the true value
  - E.g.: unexpected tails ("outliers")
- The median is a robust estimate of a distribution average, while the mean is not
- Trimmed estimators: removing n extreme values
- Evaluation of estimator robustness:
  - Breakdown point: max. fraction of *incorrect* measurements that above which the estimate may be arbitrary large
    - Trimmed observations at $x\%$ have a break point of x
    - The median has a break point of 0.5
  - Influence function:
    - Deviation of estimator if one measurement is replaced by an arbitrary (incorrect measurement)

- Details are beyond the purpose of this course…

Errors with maximum likelihood fits

- Two approaches to the determination of parameter error:
- Local error: 2nd order partial derivatives w.r.t. fit parameters around the minimum:

$$C_{ij}^{-1} = -\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}$$

MIGRAD/HESSE

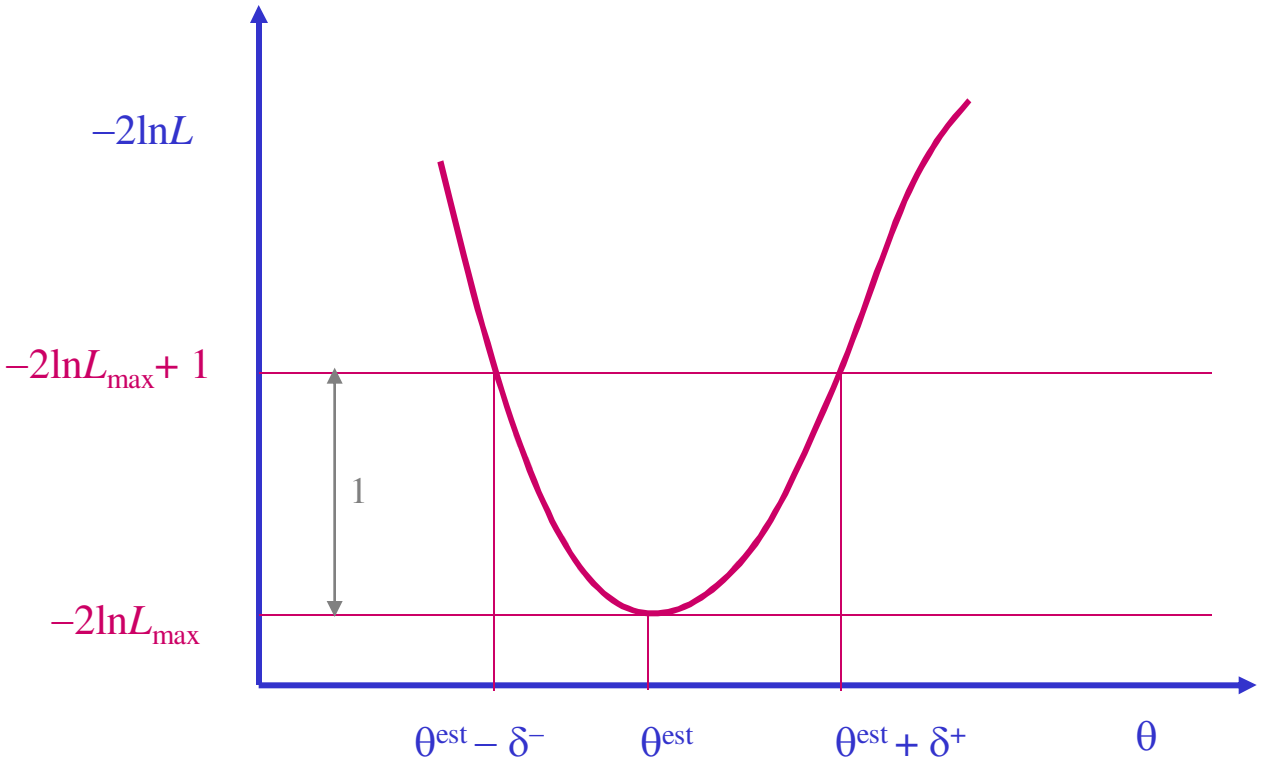
  - Under Gaussian approximation equal to covariance matrix
  - May lead to underestimate for finite samples
- Look at excursion of $-2\ln L$ around maximum of L
  - Leads to usual error matrix in a Gaussian model
  - May lead to asymmetric errors

in MINUIT

MINOS

Asymmetric error determination

- If $-2\ln L$ is close to a parabolic shape the derivatives can be approximated by parameter excursion ranges
- Error ($n\sigma$) determined by the range around the Likelihood maximum for which $-2\ln L$ increases by *one* (n^2)



- Errors can be asymmetric
  - Be careful about interpretation!
- Identical to PDF's σ for Gaussian models

Error of the (Gaussian) average

- We have the previous log-likelihood function:

$$-2 \ln L = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} + n \left(\ln 2\pi + 2 \ln \sigma \right)$$

- The error on μ is given by:

$$\frac{1}{\sigma_\mu^2} = \frac{\partial^2 (-\ln L)}{\partial \mu^2} = \frac{n}{\sigma^2}$$

- I.e.: the error on the average is:

$$\sigma_\mu = \frac{\sigma}{\sqrt{n}}$$

Exercise

- Assume we have n independent measurements from an exponential PDF:

$$f(t) = \lambda e^{-\lambda t}$$

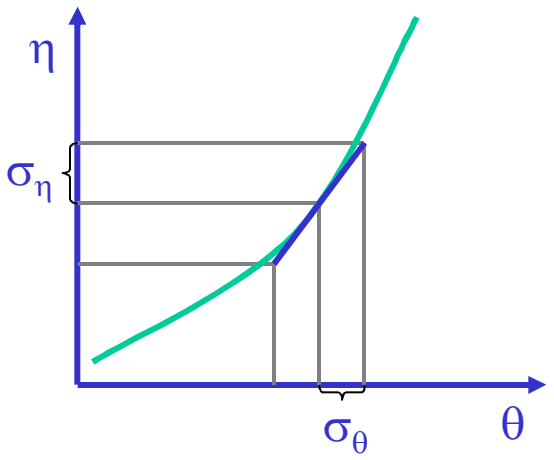
- How can we estimate by ML λ and its error?

Error propagation

- Assume we estimate from a fit the parameter set: $\theta = (\theta_1, \ldots, \theta_n)$.
- We want to determine a new set of parameters that are functions of θ: $\eta = (\eta_1, \ldots, \eta_m)$.
- We can do a linear approximation around the central values of θ by Taylor expansion, using the corresponding error matrix, Θ_{ij}, as:

$$H_{ij} = \sum_{k,l} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \Theta_{kl}$$

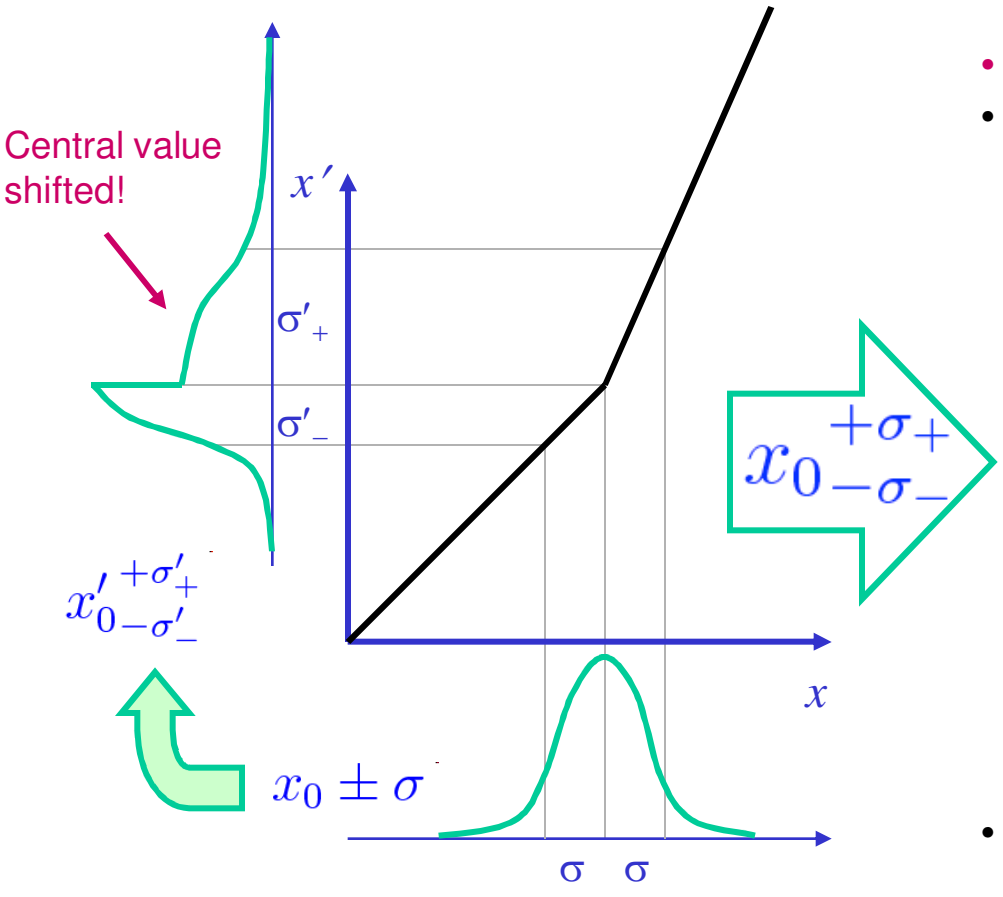
$$H = A^T \Theta A \quad \text{, in matrix form}$$



Care with asymmetric errors

- Be careful about:
  - Asymmetric error propagation
  - Combining measurements with asymmetric errors
  - Difference of "most likely value" w.r.t. "average value"
- Naïve quadrature sum of σ_+ and σ_- lead to wrong answer
  - Violates the central limit theorem: the combined result should be more symmetric than the original sources!
  - A model of the non-linear dependence may be needed for quantitative calculations
  - Biases are very easy to achieve (depending on $\sigma_+ - \sigma_-$, and on the non-linear model)
- Much better to know the original PDF and propagate/combine the information properly!
  - Be careful about interpreting the meaning of the result
- Average value and Variance propagate linearly, while most probable value (mode) does not add linearly
- Whenever possible, use a single fit rather than multiple cascade fits, and quote the final asymmetric errors only

Non linear models



- Mean, variance and skewness add linearly when doing convolution
- Not the most probable values (→fit)!
- For this model:

$$\mu = x_0 + \frac{1}{\sqrt{2\pi}}(\sigma_+ - \sigma_-)$$

$$V = (\frac{\sigma_+ + \sigma_-}{2})^2 + (\frac{\sigma_+ - \sigma_-}{2})^2(1 - \frac{2}{\pi})$$

$$\begin{aligned}\gamma &= \langle x^3 \rangle - 3\langle x \rangle \langle x^2 \rangle + 2\langle x \rangle^3 \\ &= \frac{1}{2\pi}\left[2(\sigma_+^3 - \sigma_-^3) - \frac{3}{2}(\sigma_+ - \sigma_-)(\sigma_+^2 + \sigma_-^2)\right. \\ &+ \left.\frac{1}{\pi}(\sigma_+ - \sigma_-)^3\right]\end{aligned}$$

- Online calculator (R. Barlow): http://www.slac.stanford.edu/~barlow/java/statistics1.html

See: R. Barlow, PHYSTAT2003

A concrete fit example (I)

Study of $\mathcal{B}(B^+ \to J/\psi\pi^+) / \mathcal{B}(B^+ \to J/\psi K^+)$
in BaBar

Fitting $B(B^+ \to J/\psi \pi^+) / B(B^+ \to J/\psi K^+)$

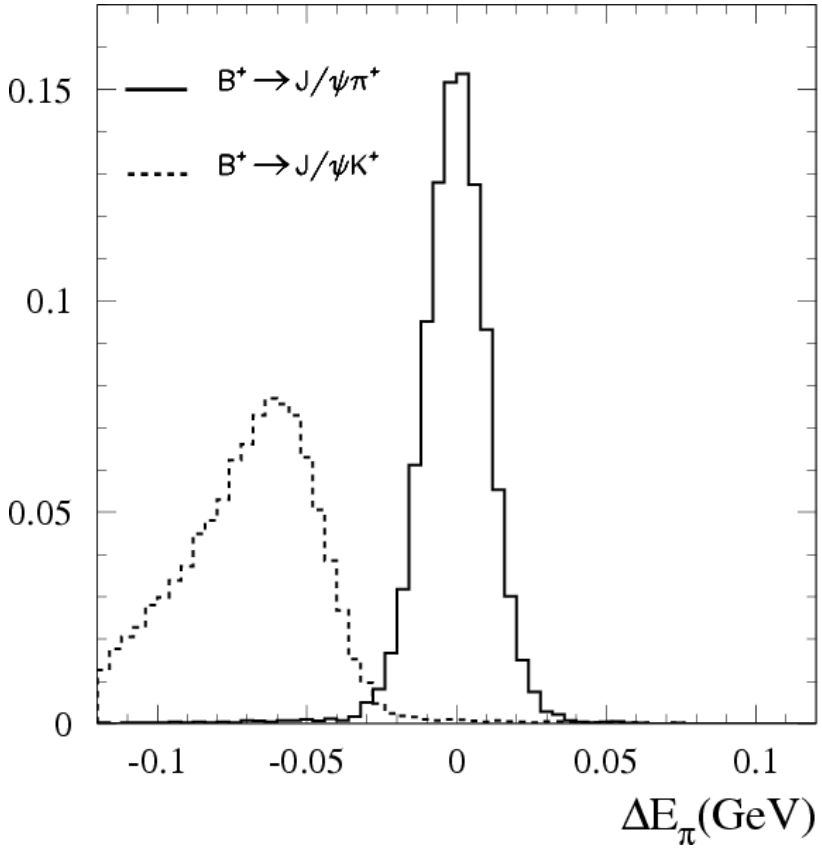
- Four variables:
  - m = B reconstructed mass as J/ψ + charged hadron invariant mass

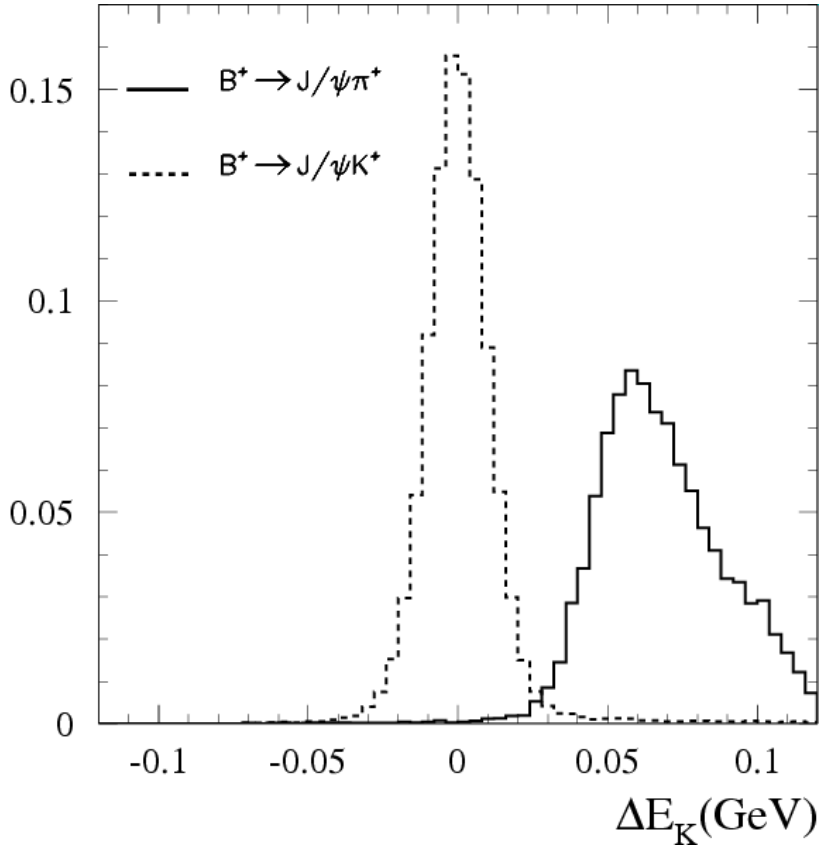
$$m_{ES} = \sqrt{E^2_{\text{beam}} - p^2_B}$$

  - ΔE_π = Beam – B energy in the π^+ mass hypothesis
  - ΔE_K = Beam – B energy in the K^+ mass hypothesis
  - q = B meson charge
- Two samples:
  - $J/\psi \to \mu^+\mu^-$, $J/\psi \to e^+e^-$
- Simultaneous fit of:
  - Total yield of $B^+ \to J/\psi\pi^+$, $B^+ \to J/\psi K^+$ and background
  - Resolutions separately for $J/\psi \to \mu^+\mu^-$, $J/\psi \to e^+e^-$
  - Charge asymmetry (direct CP violation)

ΔE_π and ΔE_K

Depend on charged hardron mass hypothesis!





Extended Likelihood function

- To extract the ratio of BR:

$$
\begin{aligned}
-\ln L &= n_\pi + n_K + n_{bkg} \\
&= -\sum_i \ln \Big[n_\pi P_\pi(\Delta E_{\pi i}, \Delta E_{Ki}, m_i) \\
&\quad + n_K P_K(\Delta E_{\pi i}, \Delta E_{Ki}, m_i) \\
&\quad + n_{bkg} P_{bkg}(\Delta E_{\pi i}, \Delta E_{Ki}, m_i)
\end{aligned}
$$

Poisson term

$B \to J/\psi\pi$

$B \to J/\psi K$

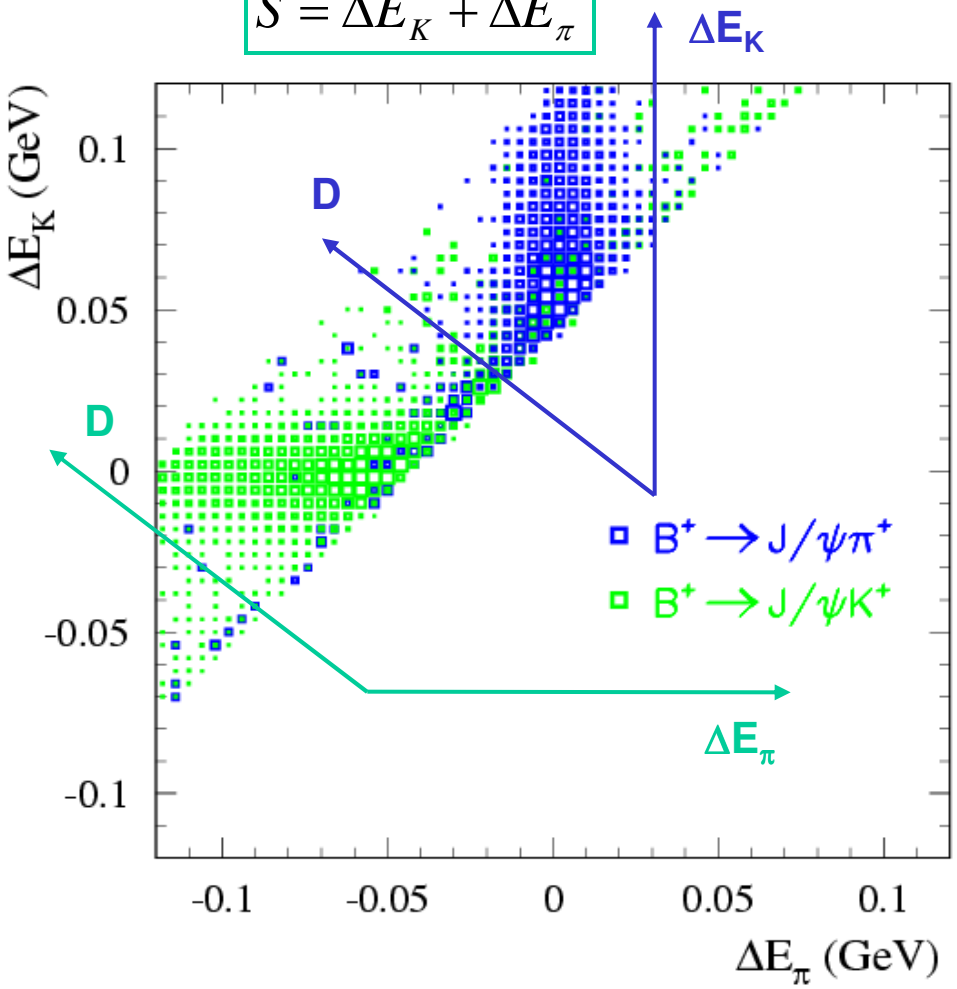
Background

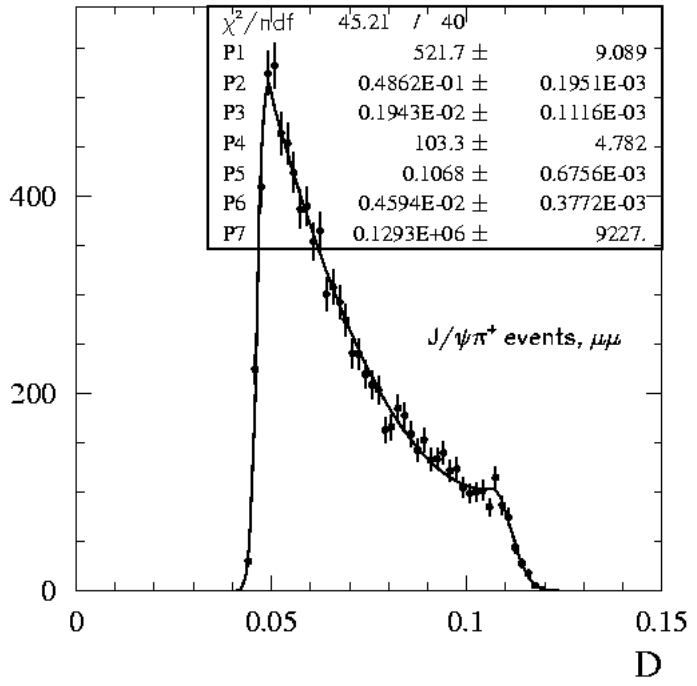
- Likelihood can be written separately, or combined for ee and $\mu\mu$ events
- Fit contains parameters of interest (mainly n_π, n_K) plus uninteresting *nuisance* parameters
- Separating $q = +1 / -1$ can be done adding A_{CP} as extra parameter

Model for independent PDFs

$$D = \Delta E_K - \Delta E_\pi$$
$$S = \Delta E_K + \Delta E_\pi$$

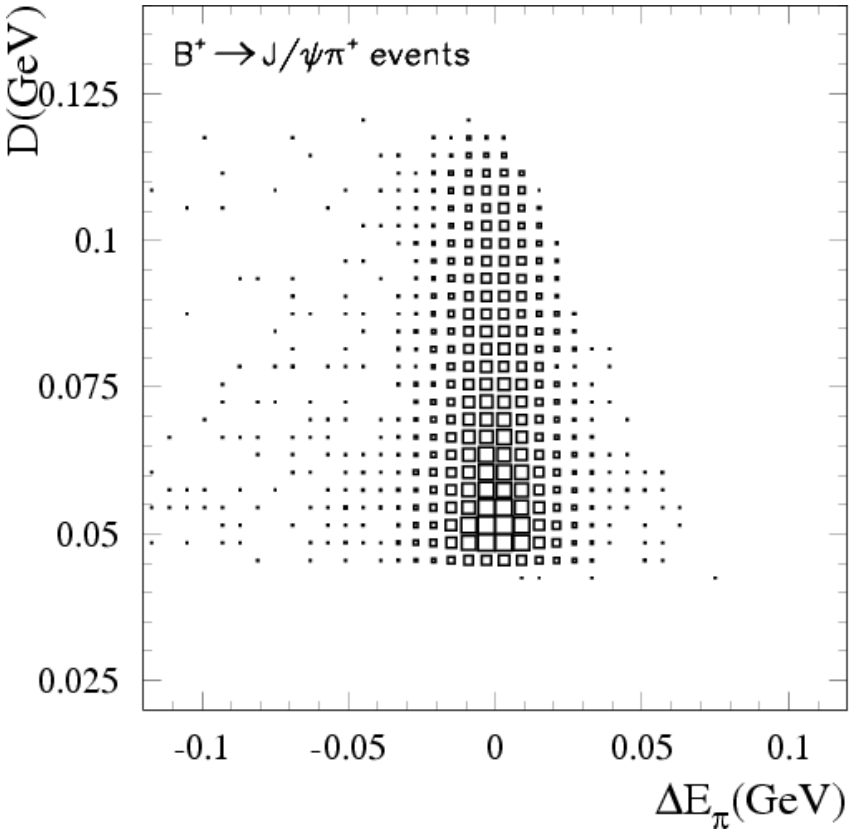
$$P_\pi(\Delta E_\pi, \Delta E_K, m) = f_\pi(\Delta E_\pi) g_\pi(D) h_\pi(m)$$
$$P_K(\Delta E_\pi, \Delta E_K, m) = f_K(\Delta E_K) g_K(D) h_K(m)$$
$$P_{bkg}(\Delta E_\pi, \Delta E_K, m) = 2 f_{bkg}(S, D) h_{bkg}(m)$$

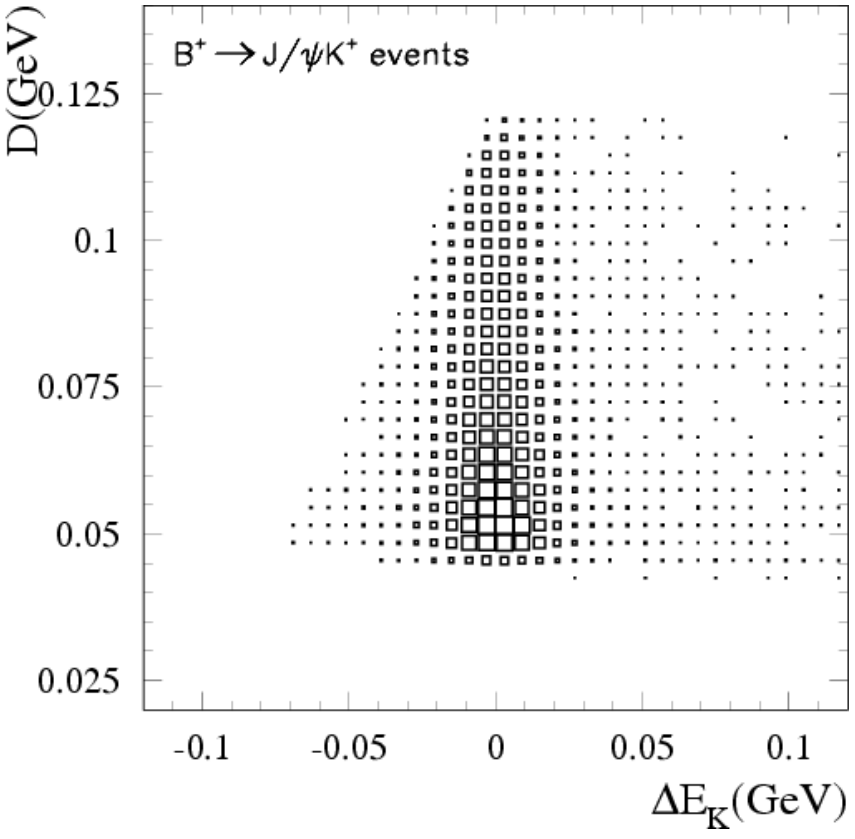




Signals PDFs in new variables

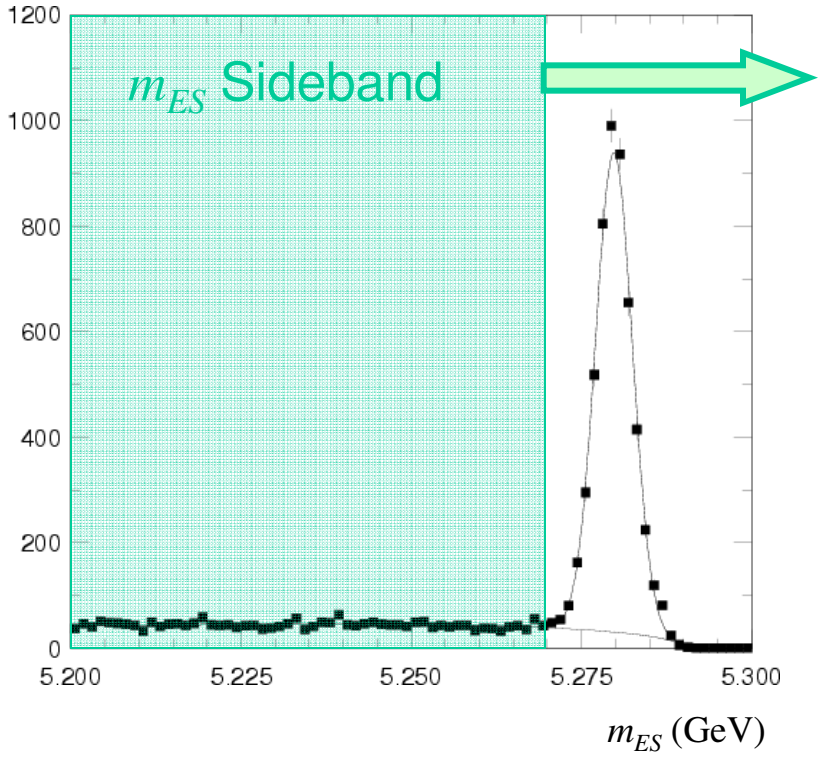
- $(\Delta E_\pi, \Delta E_K) \rightarrow (\Delta E_\pi, \Delta E_K - \Delta E_\pi), (\Delta E_K, \Delta E_K - \Delta E_\pi)$

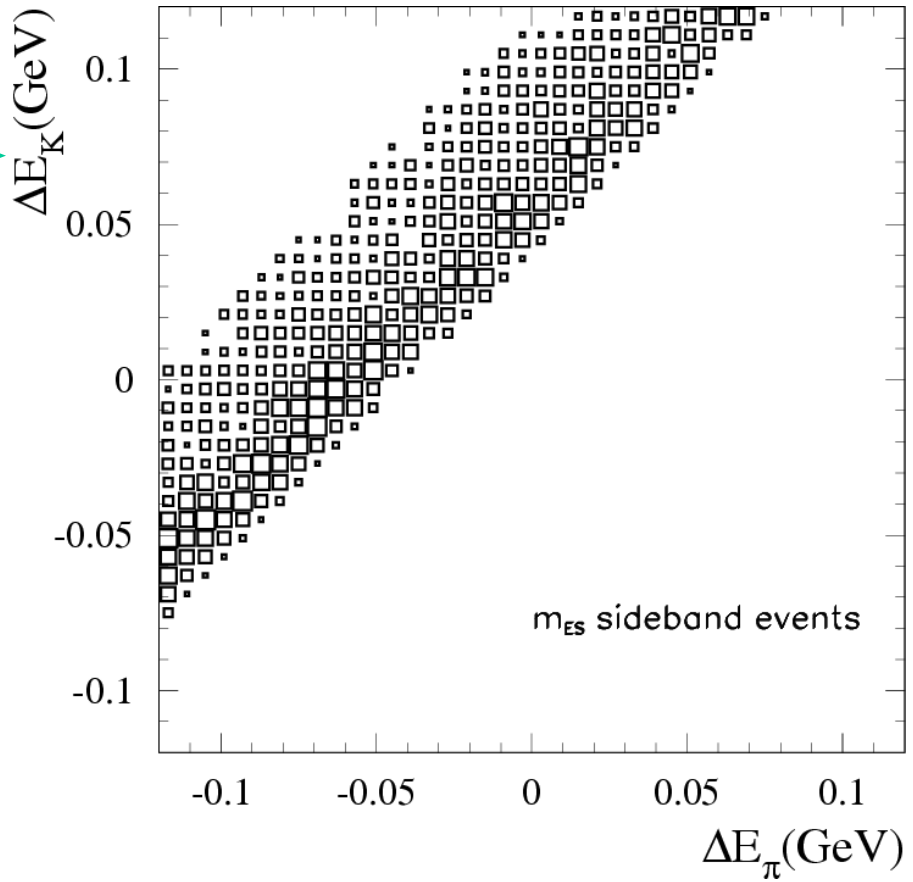




Background PDF

- Background shape is taken from events in the m_{ES} sideband ($m_{ES} < 5.27$ GeV)





Dealing with kinematical pre-selection

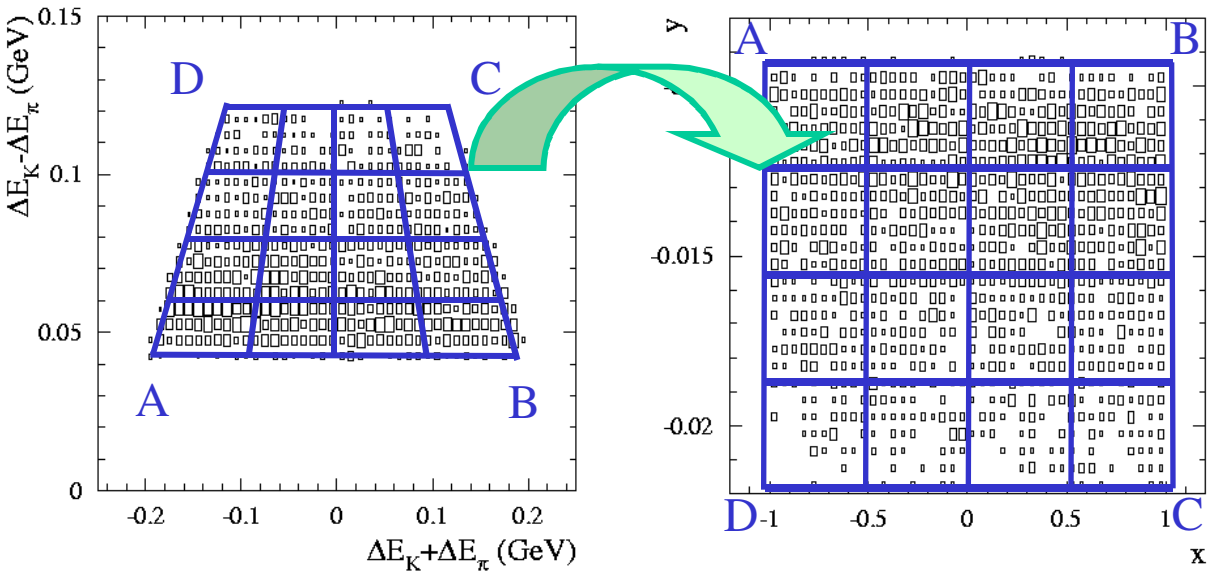
$$P_{bkg}(\Delta E_\pi, \Delta E_K, m) = 2 f_{bkg}(S,D) h_{bkg}(m)$$
$$f_{bkg}(S,D) = f_X(X) g_Y(Y)$$

$$X = \frac{S}{D - 2\cdot 120}$$

$$Y = D \cdot \left[\frac{D}{2} - 240 \right]$$

$$\left| \frac{\partial(X,Y)}{\partial(S,D)} \right| = 1$$

-120 MeV < ΔE_π, ΔE_K < 120MeV



The area is preserved after the transformation

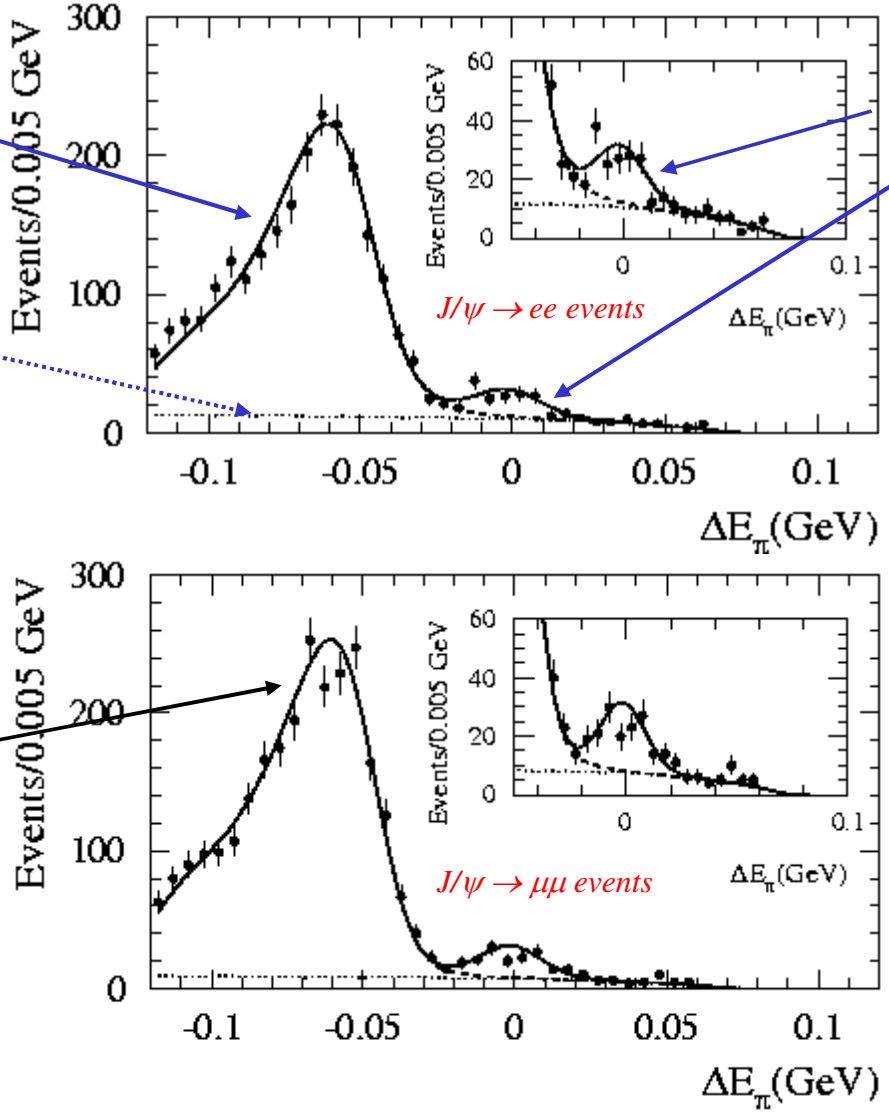
Signal extraction

$B^+ \rightarrow J/\psi K^+$

$B^+ \rightarrow J/\psi \pi^+$

Background

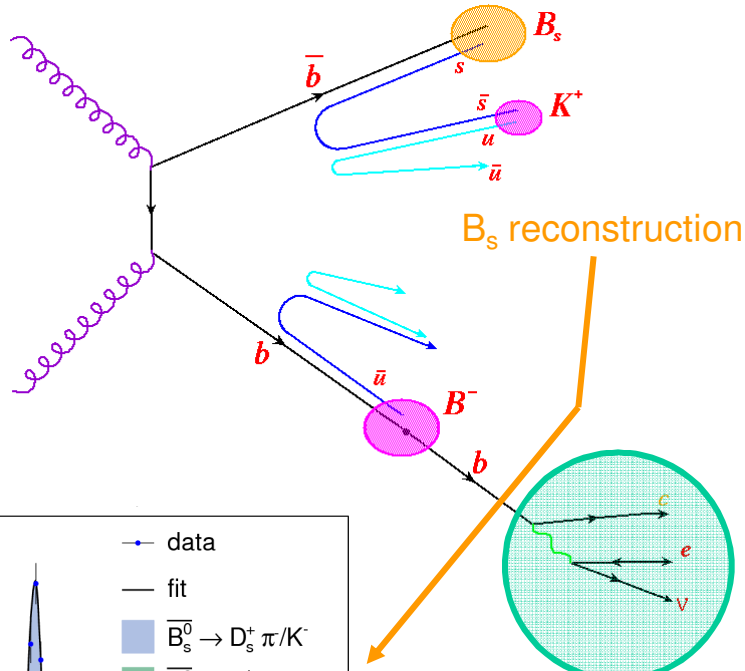
Likelihood projection

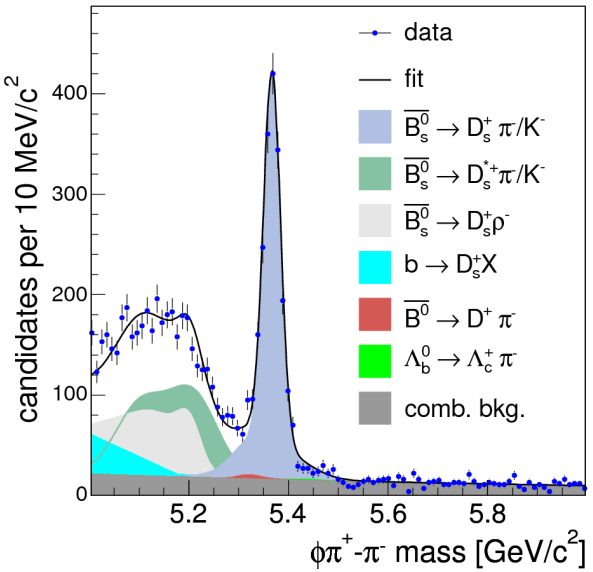


A concrete fit example (II)

Measurement of Δm_s by CDF

B production at the TeVatron





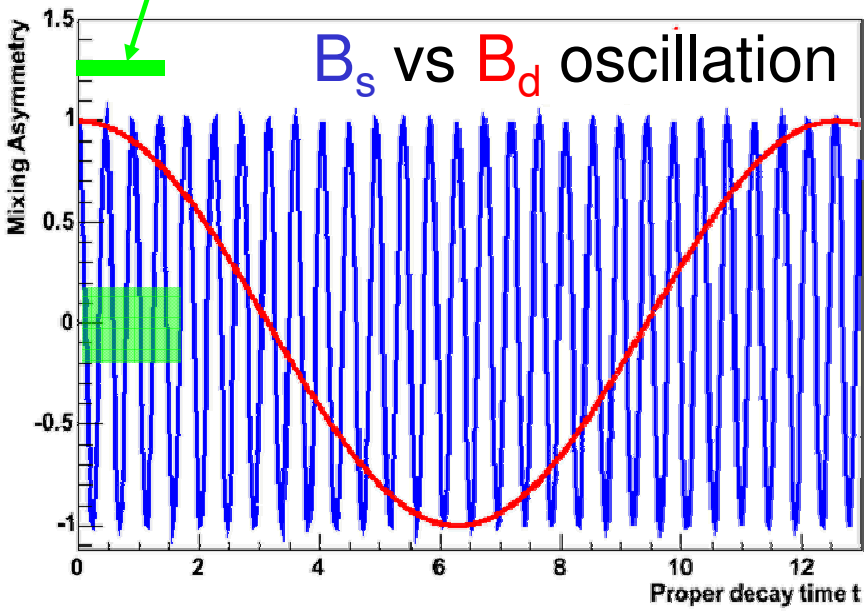
- Production: gg→bb
- NO QM coherence, unlike B factories
- Opposite flavor at production→one of the b quarks can be used to tag the flavor of the other at production
- Fragmentation products have some memory of b flavor as well

B_s vs B_d Mixing

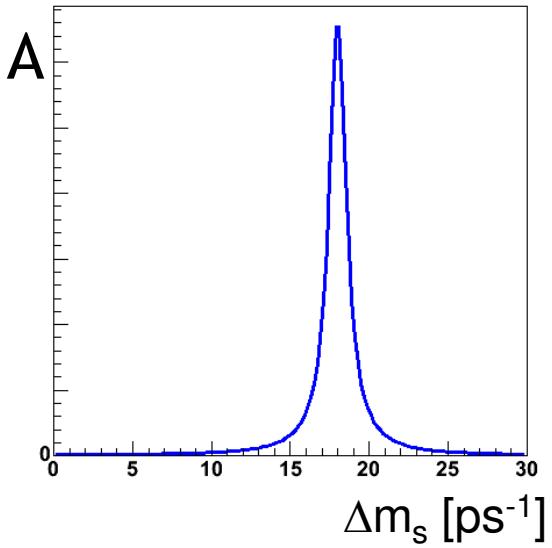
$$A = \frac{N_{unmix} - N_{mix}}{N_{unmix} + N_{mix}} \propto \cos(\Delta m_s \, t)$$

- $\Delta m_s >> \Delta m_d$
- Different oscillation regime → Amplitude Scan

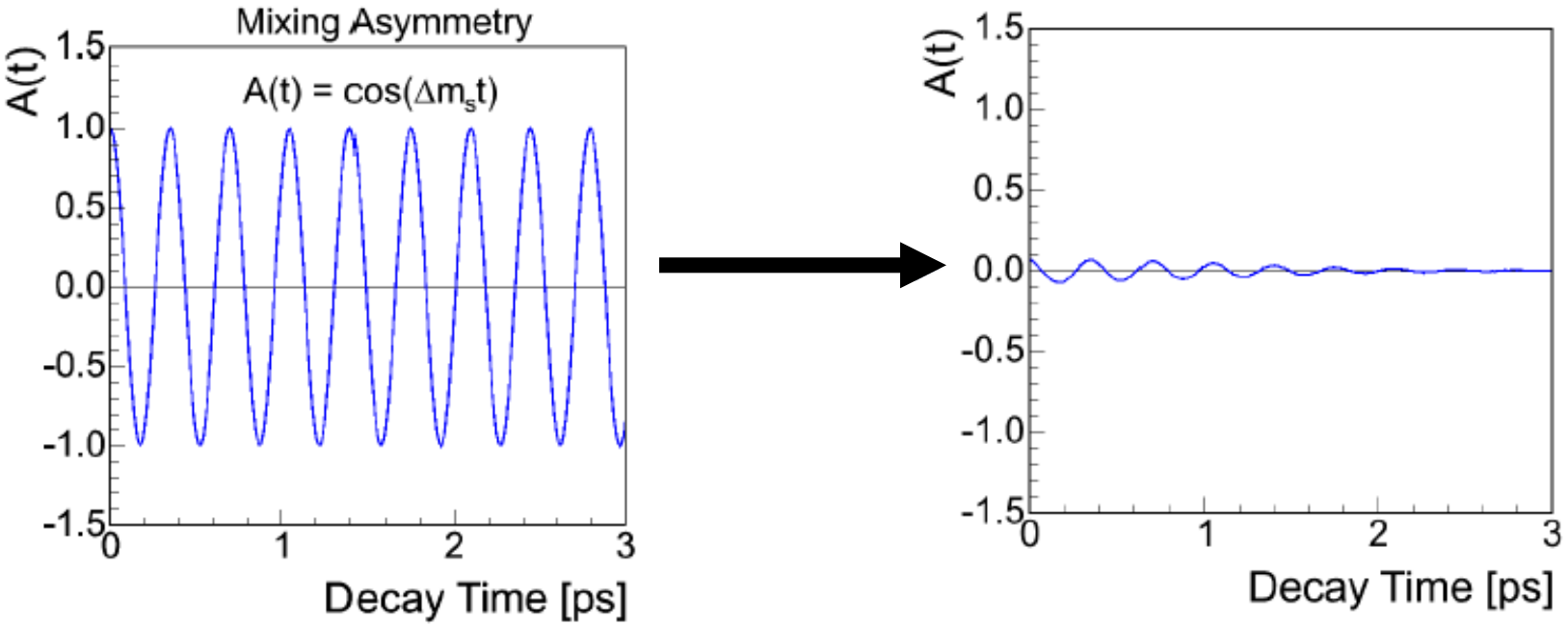
B lifetime



Perform a 'Fourier transform' rather than fit for frequency



Mixing in the real world



Flavor tagging power (1.5%) Proper time resolution (0.1÷0.4 ps)

$$\text{Significance} = \sqrt{\frac{S\varepsilon D^2}{2}} e^{-\Delta m_s \sigma_t^2/2} \sqrt{\frac{S}{S+B}}$$

Likelihood definition

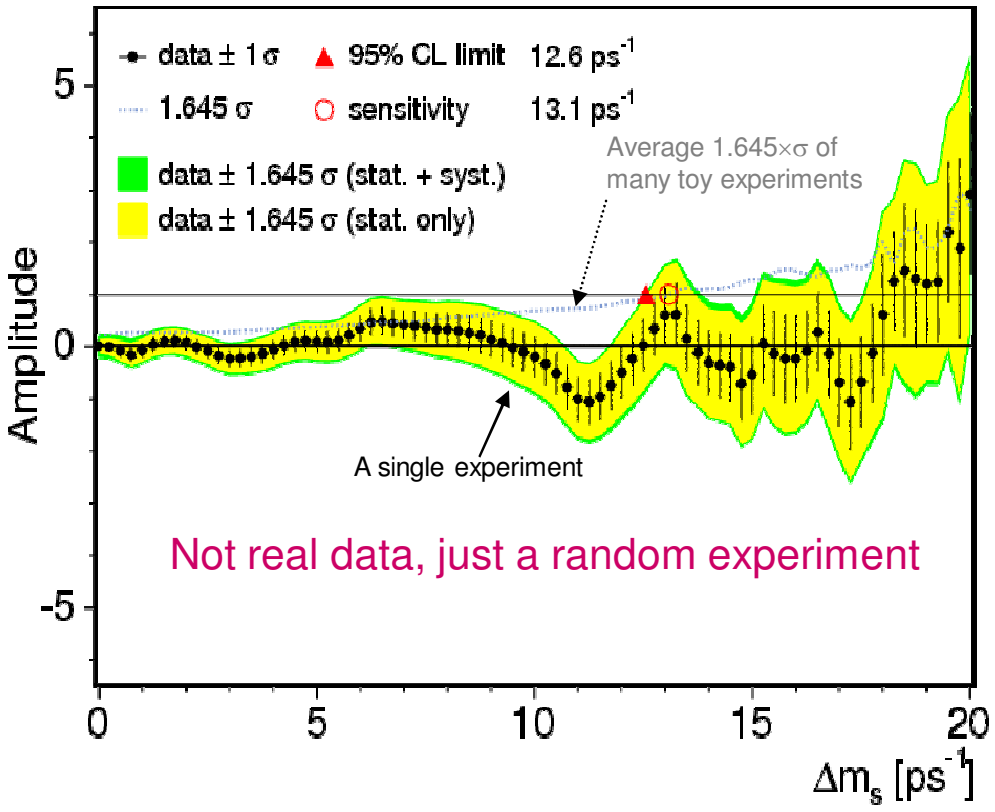
- Product of PDF (for i^{th} event) defined as:

$$\mathcal{S}_{\pm}(t_i, \sigma_{t_i}, \mathcal{D}_i) = \varepsilon(t_i) \int \frac{\Gamma_s}{2} e^{-\Gamma_s t'} [1 \pm \mathcal{A}\mathcal{D}_i \cos(\Delta m_s t')] \times \mathcal{G}(t_i - t', \sigma_{t_i}) dt',$$

- $+/-$ = same/opposite b flavor
- A = amplitude (=1 for right Δm_s, 0 for wrong value)
  - fitted for each point of the scan at fixed Δm_s
- D = dilution factor
  - = $1-2w$ (w = wrong tag fraction)
- ε = trigger + selection efficiency
  - depends on t, taken from MC
- G = resolution function
  - Gaussian, with resolution σ, estimated event by event
- Γ_s = decay width of the B_s = inverse of decay time

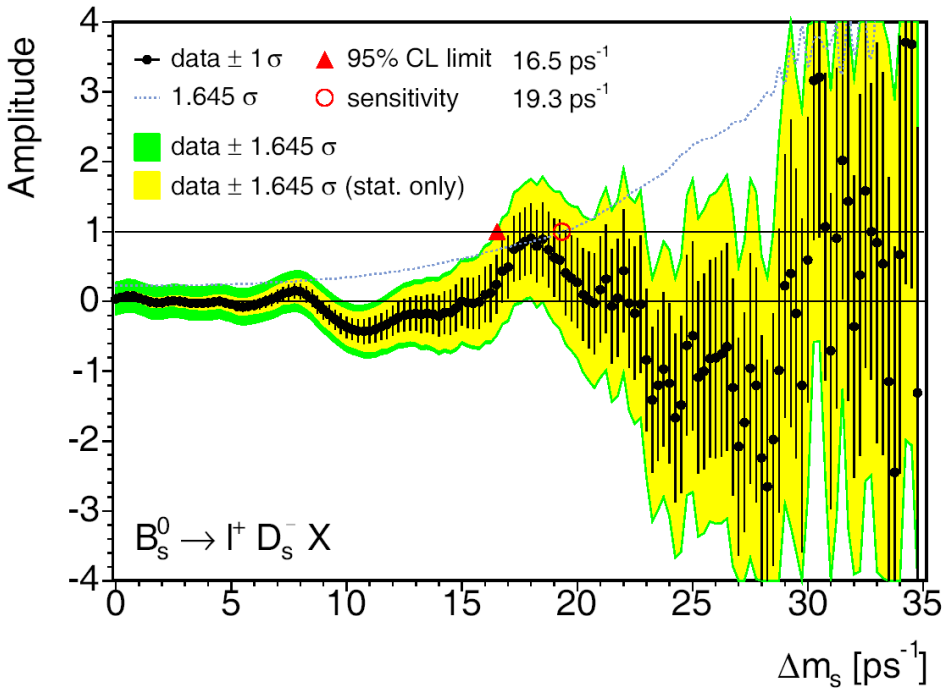
Bs mixing: method

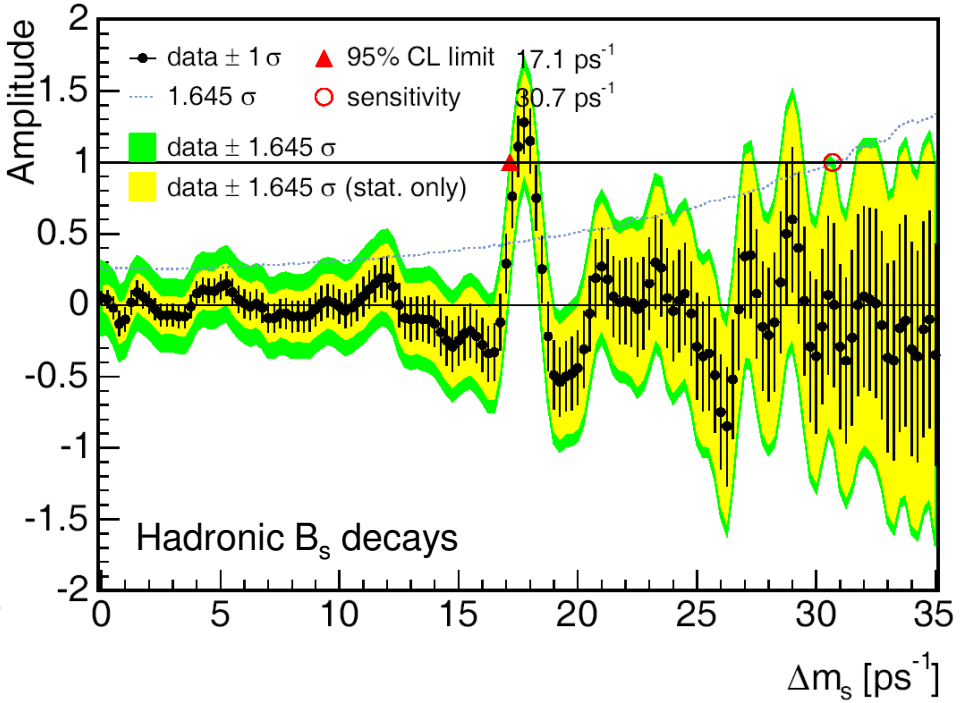
- Mixing amplitude A fitted for each (fixed) value of Δm
  - On average $A=0$ for every wrong Δm
  - $A=1$ for right value of Δm
- Green band below 1 $\Rightarrow$ A=1 is excluded at $\geq$95% C.L.
  - $1.645\sigma = 95\%$C.L.
  - Exclude range where green area is below the $A=1$ line
- Actual limit for *a single* experiment defined by the systematic band centered at the measured asymmetry
- "Sensitivity": Δm for which the average 1.645×r.m.s. of many "toy" experiments [with $A = 0$] reaches 1



- Combining experiments as easy as averaging points!

B_s Mixing: Hadronic vs semilept.





$\Delta m_s > 16.5$ @ 95% CL

Sensitivity: 19.3 ps^{-1}

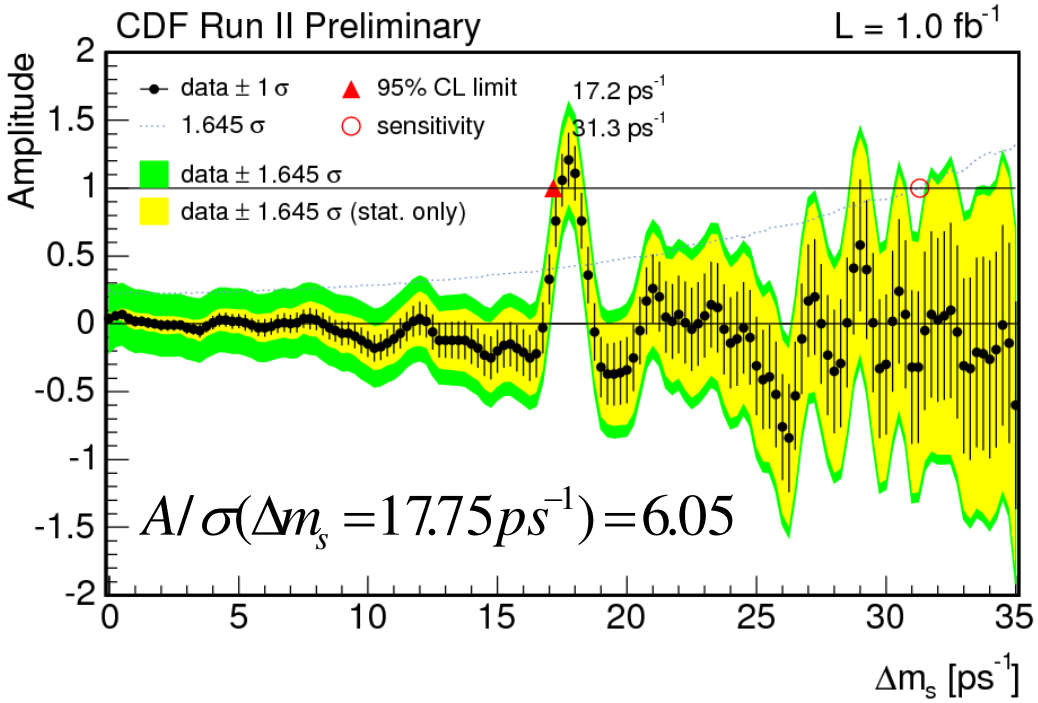
Reach at large Δm_s limited by incomplete reconstruction (σ_{ct})!

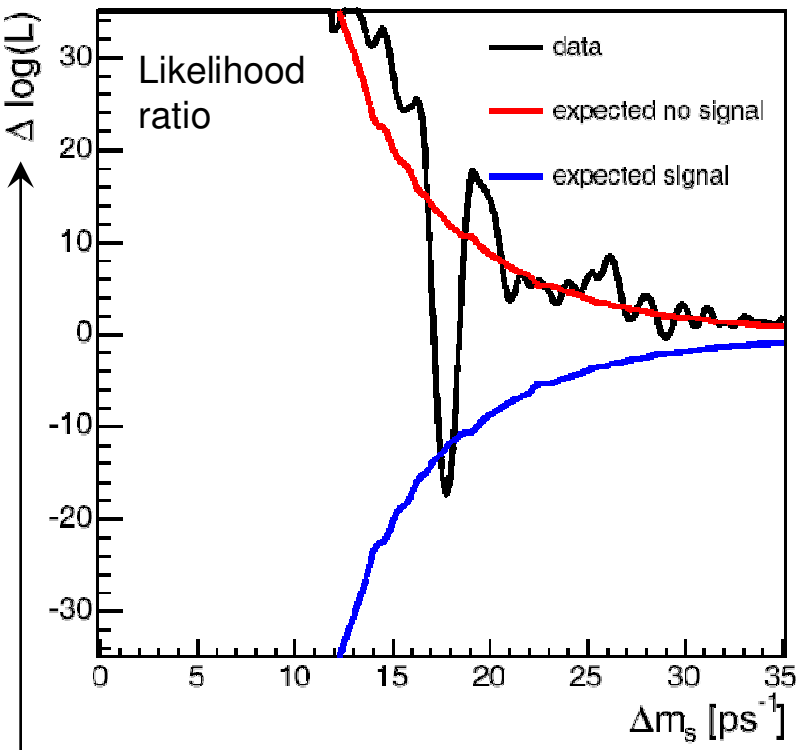
$\Delta m_s > 17.1$ @ 95% CL

Sensitivity: 30.7 ps^{-1}

This looks a lot like a signal!

B_s Mixing: combined CDF result





$\Delta m_s > 17.2\ ps^{-1}$ @ 95% CL

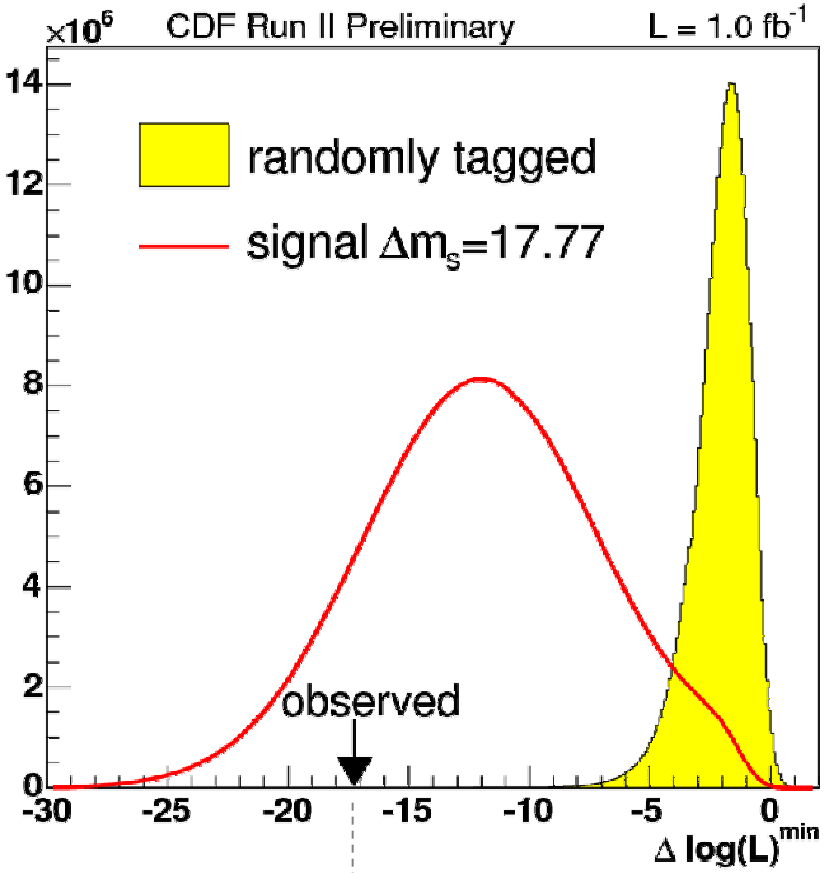
Sensitivity: 31.3 ps^{-1}

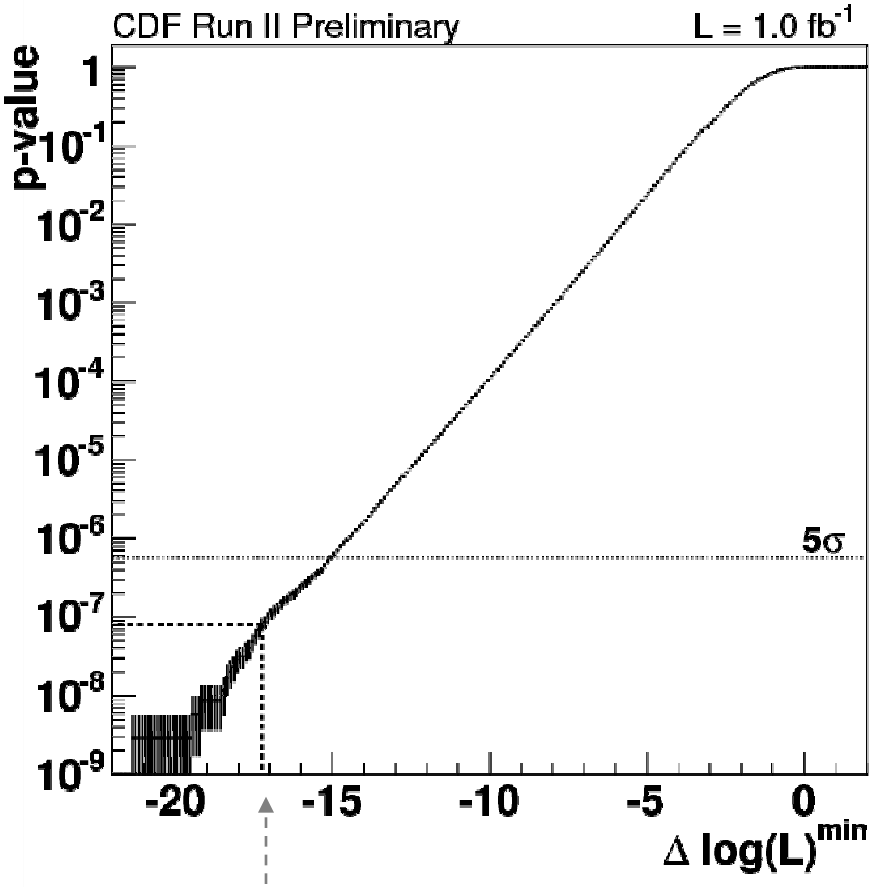
$$-\log\left(\frac{L(A=1)}{L(A=0)}\right)$$

Log-likelihood ratio preferred to log-likelihood:
Will be discussed in next slides

Minimum: −17.26

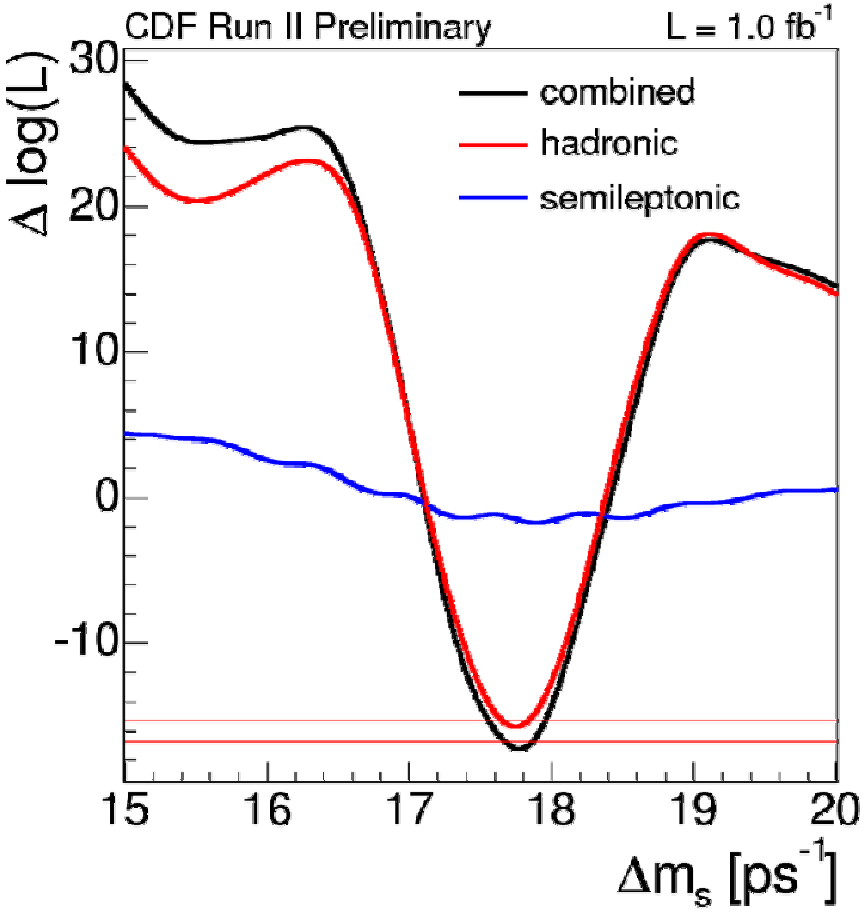
Likelihood Ratio





$$p \approx 8 \times 10^{-8} \quad \Rightarrow \quad 5.4\,\sigma$$

Likelihood Ratio



$$\Delta m_s = 17.77 \pm 0.10(\text{stat}) \pm 0.07(\text{syst})$$

PRL **97,** 062003 (2006)

References

- F.James
  - CERN Program Library Long Writeup D506
    http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html
- Robust statistics (introduction with some references):
  - http://en.wikipedia.org/wiki/Robust_statistics
- Use of chi-square and Likelihood for binned samples
  - S. Baker and R. Cousins, Clarification of the Use of Chi-Square and Likelihood Functions in Fits to Histograms, NIM 221:437 (1984)
- Unified approach:
  - G.Feldman, R.D. Cousins *Phys. Rev. D* 57, 3873 (1988)
  - G.Feldman, Fermilab Colloquium: Journeys of an Accidental Statistician
    http://www.hepl.harvard.edu/~feldman/Journeys.pdf
- Asymmetric error treatment
  - R. Barlow, proceedings of PHYSTAT2003
  - R. Barlow, arXiv:physics/0406120v1
  - G. D'Agostini, Asymmetric Uncertainties Sources, Treatment and Potential Dangers, arXiv:physics/0403086
- Lep Electro-Weak Working Group
  - http://lepewwg.web.cern.ch/LEPEWWG/
  - ZFITTER: Comput. Phys. Commun. 174 (2006) 728-758. hep-ph/0507146
- Fitting $B(B^+ \to J/\psi \pi^+) / B(B^+ \to J/\psi K^+)$
  - BaBar collaboration
    Phys.Rev.Lett.92:241802,2004, hep-ex/0401035
    Phys.Rev.D65:091101,2002, hep-ex/0108009
  - F.Fabozzi, L.Lista:
    BaBar Analysis Document (BAD) 93, 574
- B_s mixing by CDF
  - CDF collaboration
    Phys.Rev.Lett.97**:**062003,2006, hep-ex/0609040
  - Presentation by Alessandro Cerri (2006)