Statistical and other experience with new structures in hadron spectra

Kai Yi (University of Iowa)

(in Babar, CDF and CMS)

XIIth Quark Confinement and the Hadron Spectrum

from 28 August 2016 to 4 September 2016

Europe/Athens timezone

Outline

- Motivation
- A little on exotic mesons
    - X(3872)
    - $J/\psi\phi$ structures
- Significance estimation of Y(4140)
- Significance estimation on multiple peaks
- MC tests on error underestimation in low statistics
- MC modeling of significance when adding more data
- Significance of very significant peaks
- Summary

Motivation

- Many statistical applications are needed in hadron spectra
  - significance evaluation
  - estimation of expectations
  - estimation of uncertainties

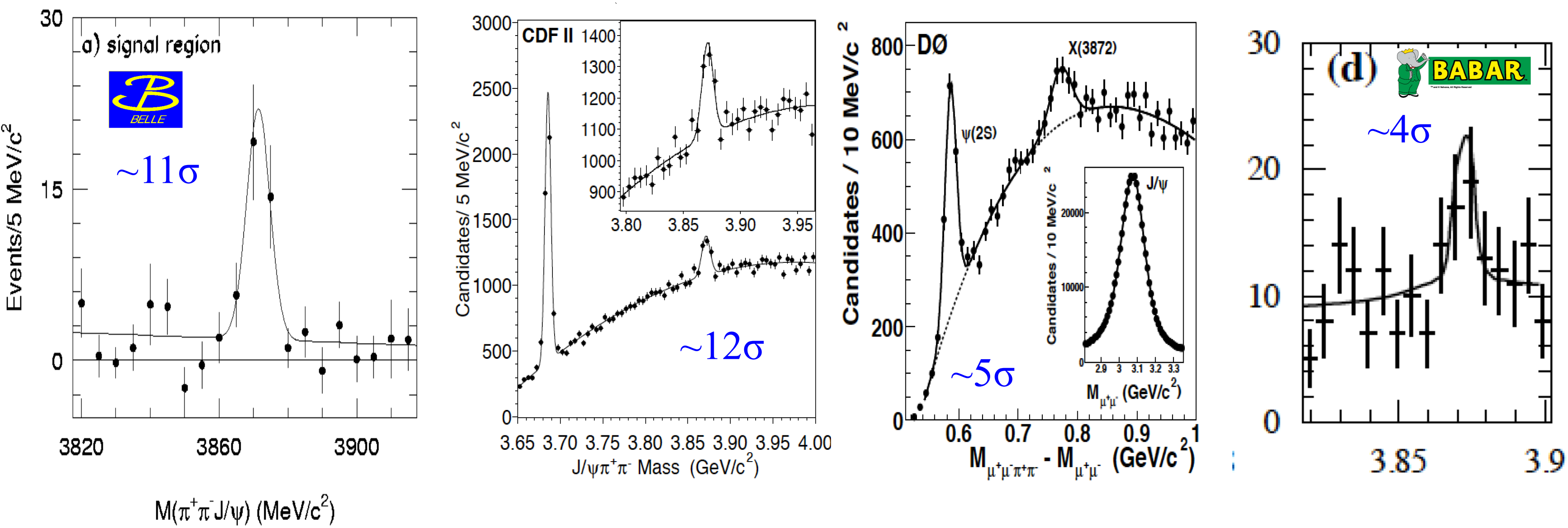
- Much first hand experience through various experiments

- Hope to be useful to other cases

- To receive feedback for further improvement

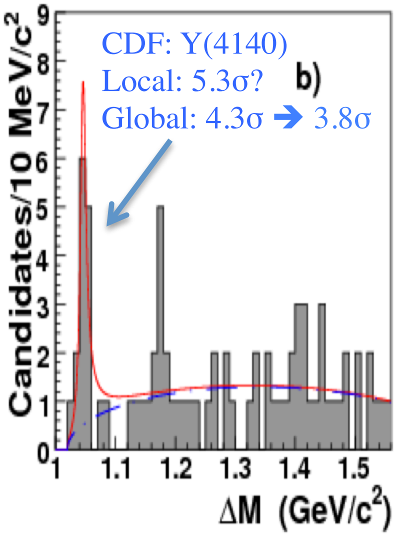
Note: lots of recent development @LHC via Higgs discovery

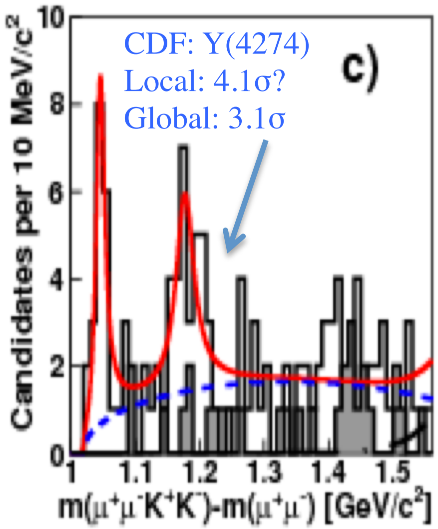
X(3872) (2003)—start of the recent exotic

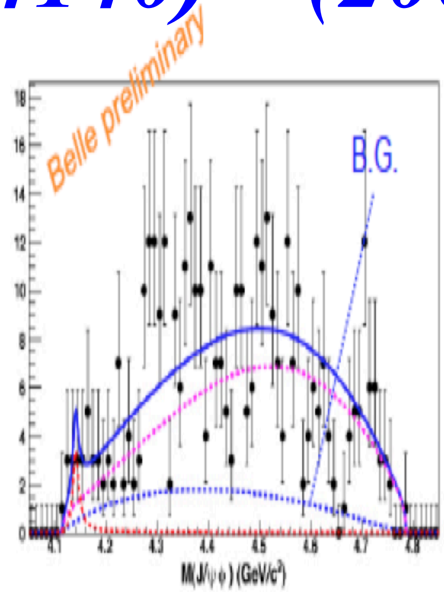


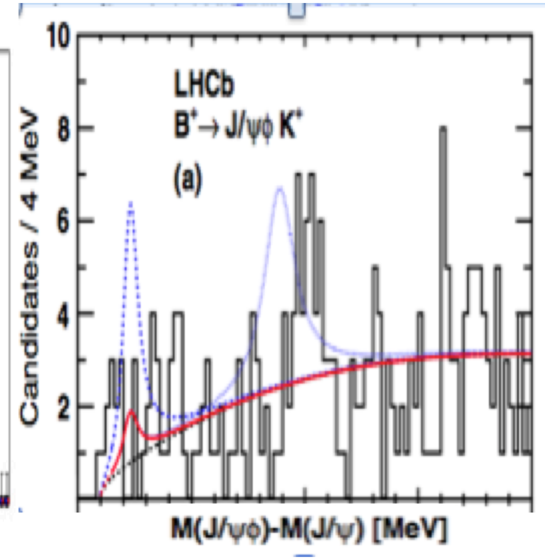
- *Significant signal first seen by Belle, quickly confirmed by CDF,D0,Babar*
- *Established very quickly, and revitalized the interest on exotic meson*
- *No Look-Elsewhere-Effect (LEE, trial factor) considered*
- *Y(4140) was the first X/Y/Z that addresses the LEE, and that is a long story…*

The Story of Y(4140)—(2009-2011)









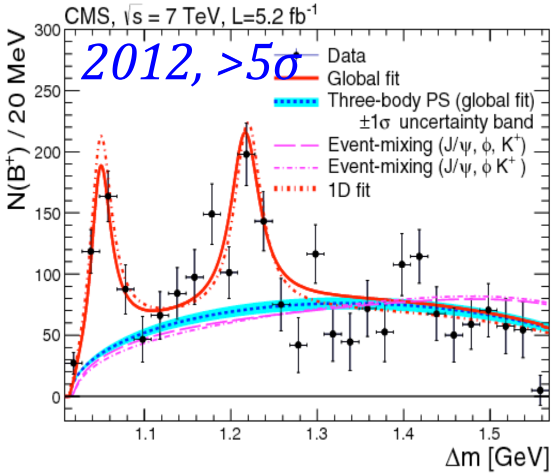
- *CDF (2009) addresses LEE explicitly for Y(4140), 5.3σ? ➔ 4.3σ ➔ 3.8σ*

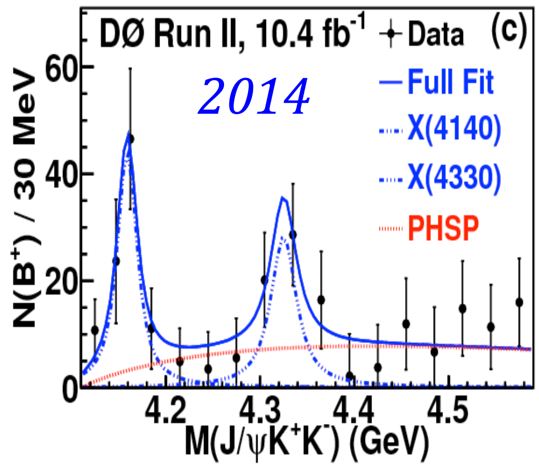
  LEE systematic
- *Belle (2010) cannot confirm or deny the existence of Y(4140)*
- *LHCb (2011) confirms neither of the structures, 2.4σ disagreement with CDF*
- *The existence of Y(4140) was in serious doubt*

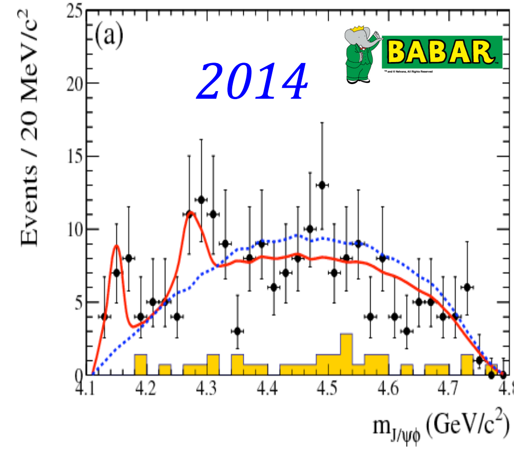
Tommaso Dorigo's blog summarizes the status in 2011

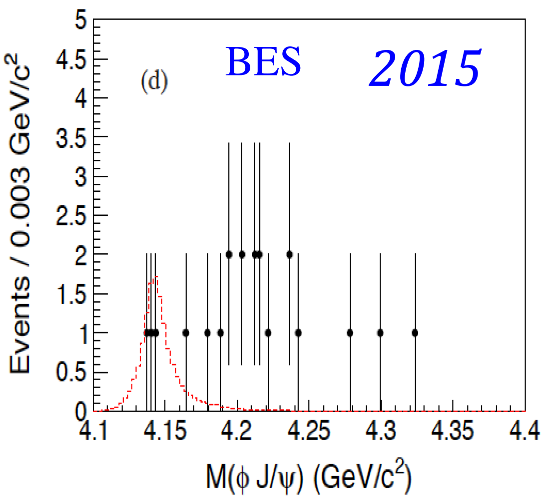
result. Note that, as reported in the figure, if the CDF signal were as estimated by CDF, LHCb would have been able to fit 39+-9+-6 events. The Y(4140) is on very shaky ground at the moment, and the new PDG will likely change its status in the particle zoo...

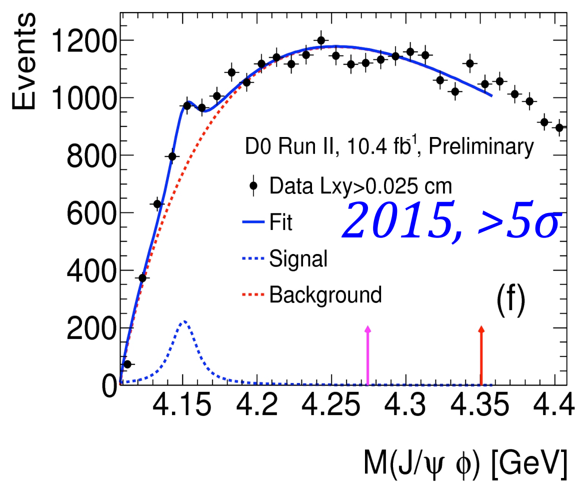
The Story of Y(4140)—(2012-2016)

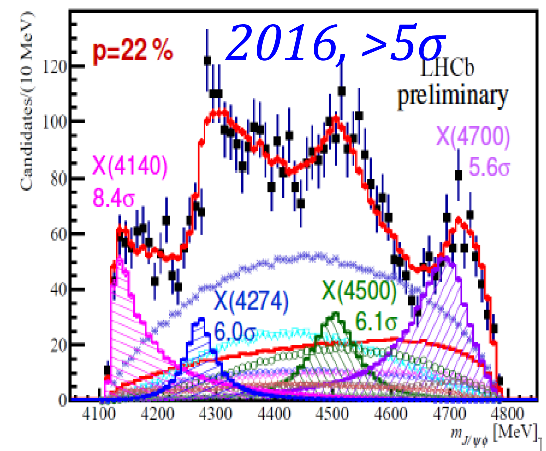








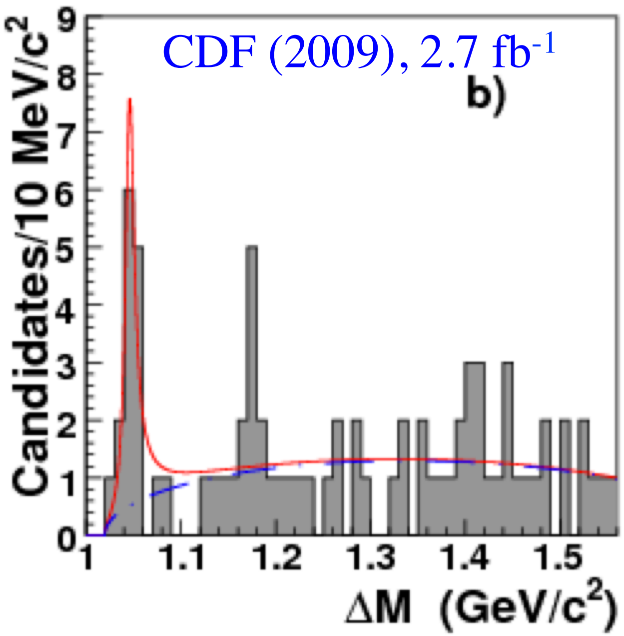




- *CMS, D0, LHCb finally confirm the existence of Y(4140) with significance >5σ (CMS also took LEE into account)*

- *This talk focuses on various Toy tests, other statistical treatments for the Y(4140)*

The significance of Y(4140)—background



background shape
3-body-PS
$q^3(Q^{1/2} + a_1 Q^{3/2})$
$q/(q + a_1 \times q)$
$q^3(Q^{1/2} + a_1 Q^{3/2})$
$q/(q + a_1 \times q)$

Phase Space

q is the ϕ momentum in the rest frame of $J/\psi\phi$ and $Q = m_{max} - m$, and $m_{max} = 1.69$ GeV

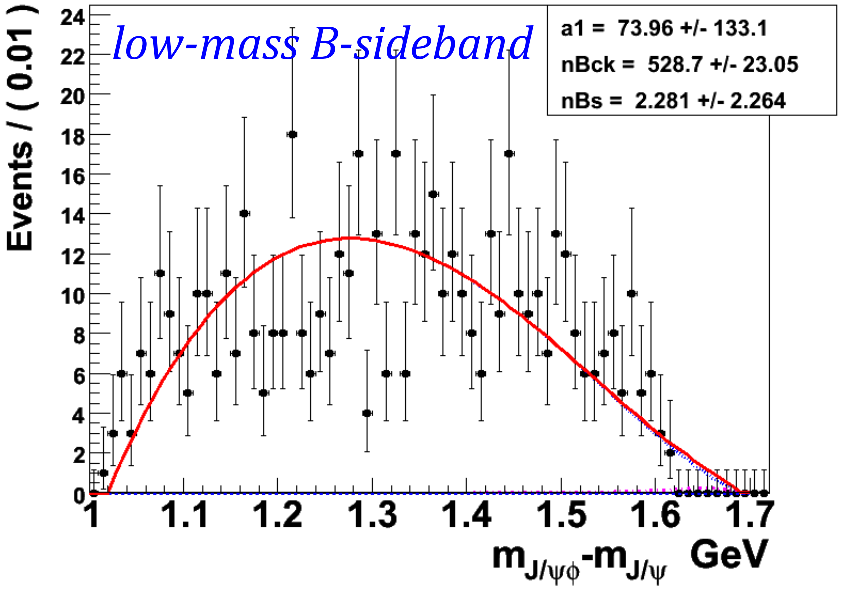
- *Once the Δm spectrum is obtained, a background a needed significance determination strongly rely on knowledge/modeling of a background*

- *Various shapes are considered, eventually 3-body-PS was chosen, because:*
  - *more defensible: $B^+ \rightarrow J/\psi\phi K^+$ is a 3-body decay assuming no substructures it also happens to be more conservative*

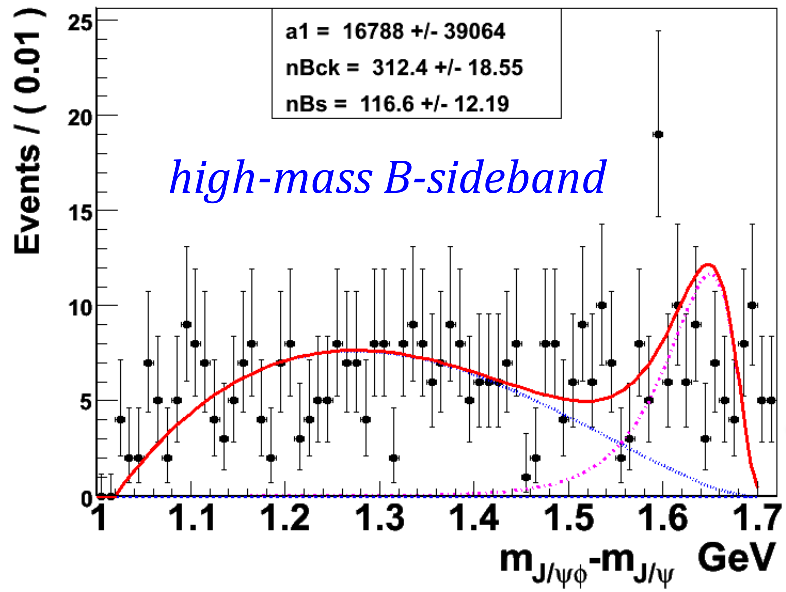
- *Why require $\Delta M <= 1.56$ GeV? A posterior decision due to B_s contamination—does not increase significance*

Fit range and signal model

CDF, 2.7 fb-1, m(J/ψϕK+) in [5.15,5.25] GeV

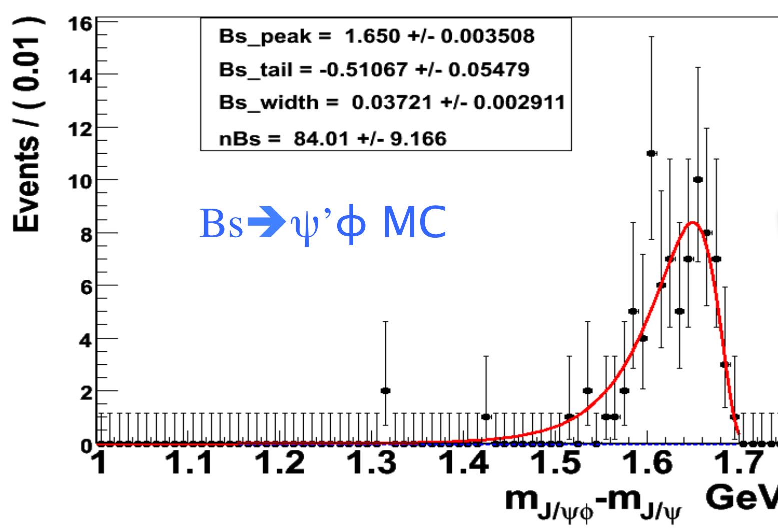


CDF, 2.7 fb^{-1}, m($J/\psi\phi K^+$) in [5.35,5.45] GeV



π/K

CDF, MC, Bs➔ψ'ϕ, ψ'➔J/ψ$\pi^+\pi^-$
in defined B mass window



- *Identified a component from Bs decays that concentrated at a region with Δm>1.56 GeV*
- *Thus fit range: Δm<1.56 GeV*
  Always need flexibility to face reality
- *Signal shape: relativistic Breit-Wigner convoluted with resolution*

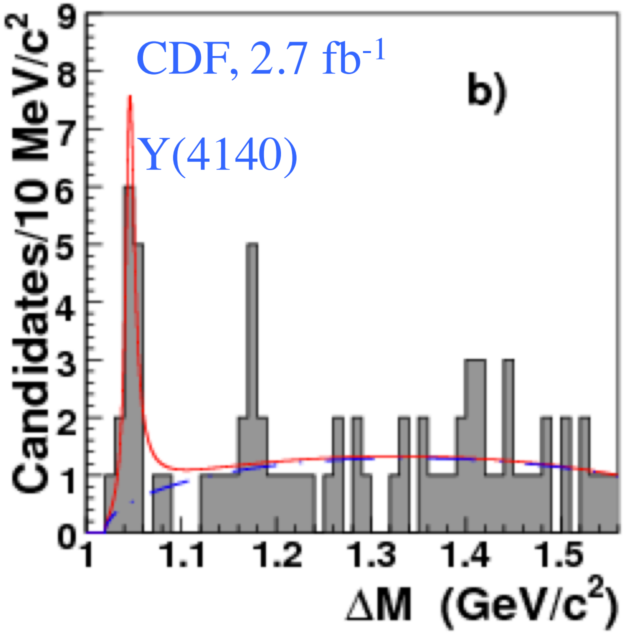
Elements for significance estimation

- *Range: 1.02<Δm<=1.56 GeV*

- *Background: 3-body-PS*

- *Signal shape: relativistic Breit-Wigner convoluted with resolution*

- *Another question: signal width?*
  - *not consistent with mass resolution (1.7 MeV)*
  - *it is one additional degree of freedom*
  - *it was allowed to float in the fit*



- *Likelihood ratio between null- and signal-hypothesis---2ln (L_{max}/L_0): 28.1 more robust way to evaluate significance comparing something like S/sqrt(S+B)*

What is the significance of the excess near threshold?
First ignore the possible second peak

This is the first X/Y/Z discovery at Tevatron

Simple significance estimation?

- *$2\ln(L_{max}/L_0)$=28.1*

- *Treat it as approximate χ^2 of d.o.f=1, 2, 3?*
  - *d.o.f=1, p=5.8×10^{-8}, 5.3σ if mass and width are known and fixed*
  - *d.o.f=2, p=4.0×10^{-7}, 4.9σ if mass or width is known and fixed*
  - *d.o.f=3, p=1.7×10^{-6}, 4.6σ if mass and width NOT known*

  Both mass and width are not known/fixed, wrong to use d.o.f=1 or 2
  still incorrect to use d.o.f=3, complicated, and Toy result will demonstrate

- *No way to have a significance number directly in this case, need simulation*
  Both mass and width are unknown, the mass/width range are the LEE

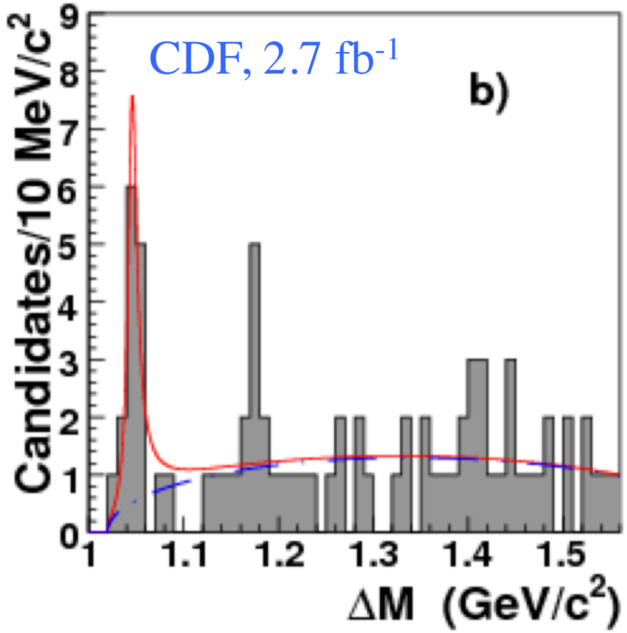
- *Full Toy MC was needed to determine a reasonable significance*
  *Recent Gross-Vitells method (*arXiv:1005.1891*) can reduce number of Toys*

Determine uncertain items to perform Toy MC—next page
Try to be reasonable, cannot be absolutely correct/accurate

Uncertain items for Toy--CDF

- *Mass range*
  - *I) Δm from 1 to 1.7 GeV—kinematically allowed?*
  - *II) Δm from 1 to 1.56 GeV—fit range?*
  - *Converged to use II) though it is a posterior thing, I) may sound more natural*

  (We do have a peak between 1.56 and 1.7 GeV but cannot claim due to contamination)



- *Width range*
  - *Lower bound*

    --mass resolution-1.7 MeV

    --we kept it as a constant for all masses

    (one could argue to use a varying lower bound at different masses and may be lowering the bound by 5% or 10%, but we ignore it)
  - *Higher bound*

    -- ¼ of the fit range 1.56 GeV; 10X of observed width. Both are arbitrary

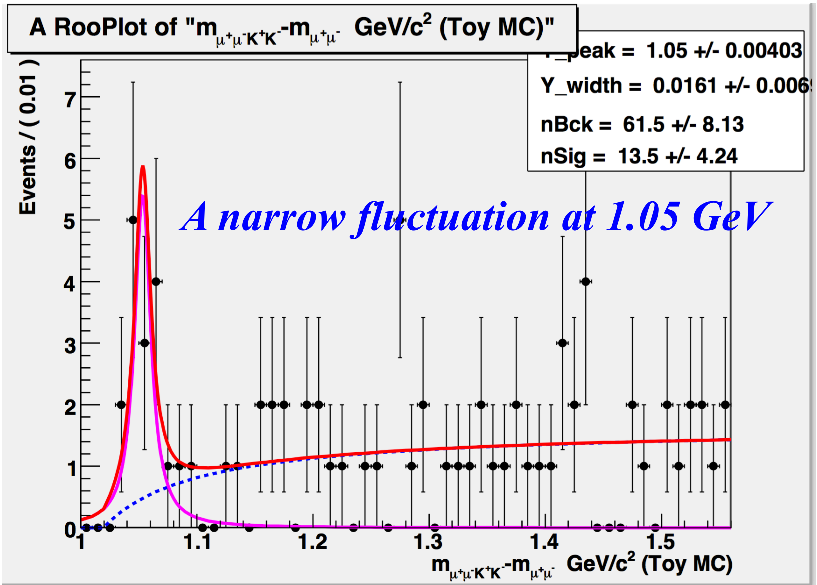
    --reasoning: be able to identify as a structure in the fit range

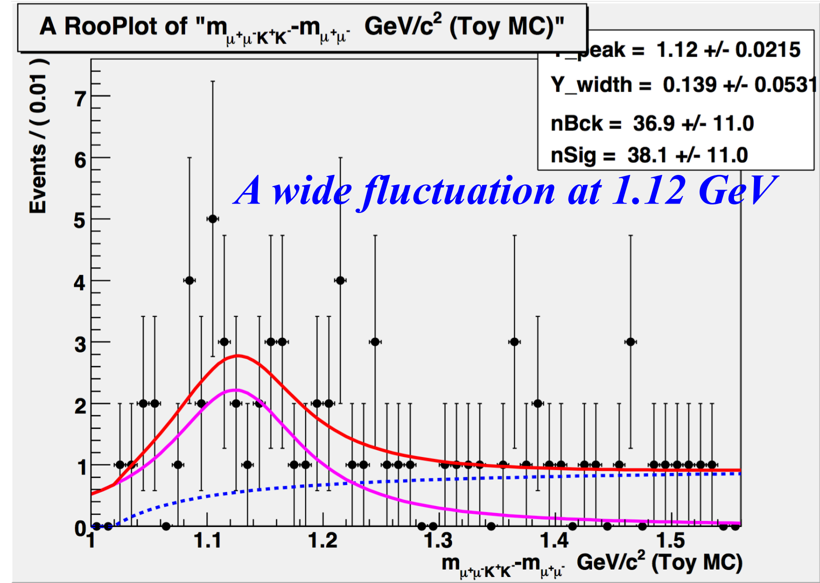
- *Use Voigt function (simple BW convoluted with Gaussian) for signal to save CPU*

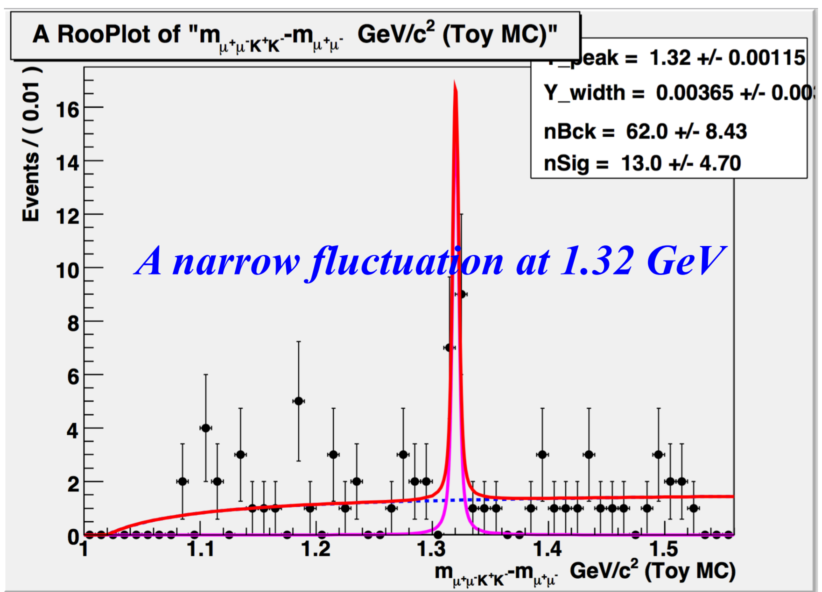
Toy MC procedure--CDF

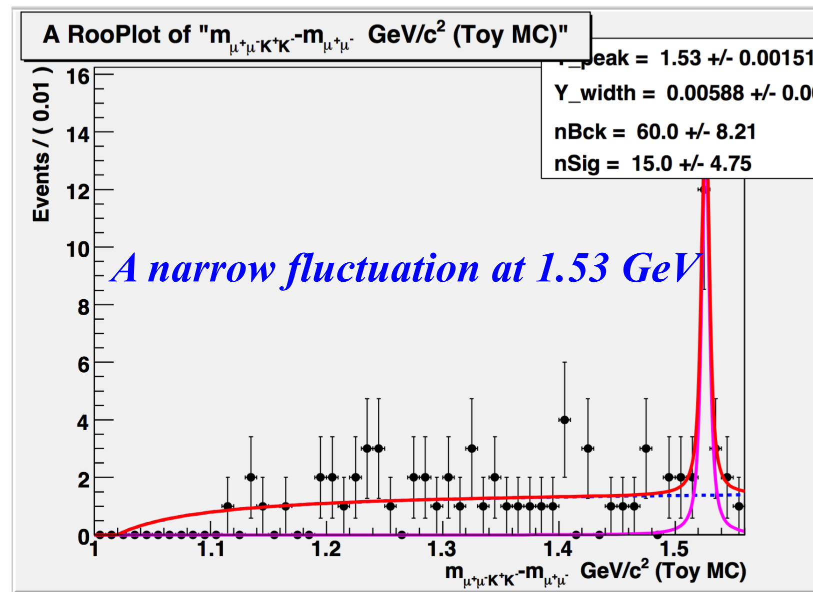
- *Determine significance from simulation (Toy MC):*
  - *Using Three-body decay Phase Space only to generate the Δm spectrum*
  - *Each time generated 75 events based on background shape. Varied the number events based on a Poisson distribution in a very sample, not difference*
  - *Find the most significant fluctuation in the Toy events for each trial anywhere in Δm between 1.02 and 1.56 GeV, and with a BW width between 1.7 MeV (resolution) and 120 MeV--10 times of the observed width. Ignore the resolution variation as a function of mass*
  - *Repeat the process for N_T times.*
  - *Count the number of times that a fluctuation with greater or equal to the $-2\ln(L_{max}/L_0)$ value in the data—N*
  - *Calculate a p-value= N/N_T, and convert it to a significance number*

Toy MC examples--CDF

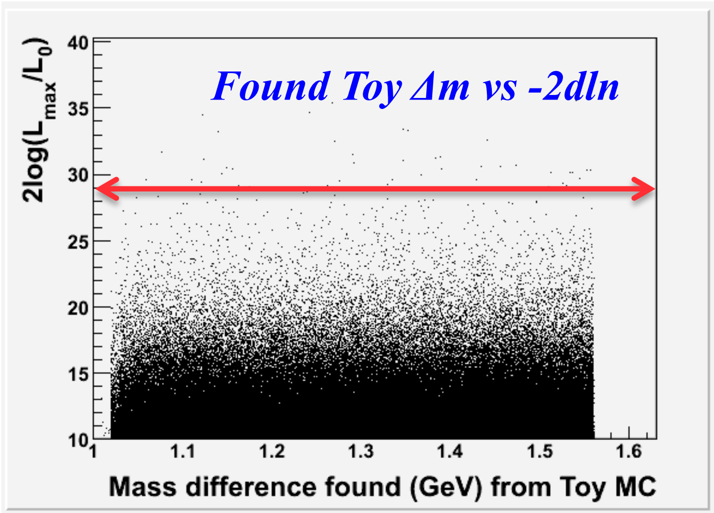






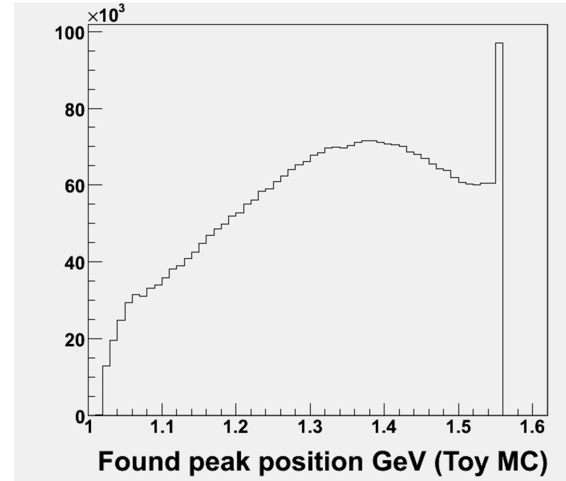


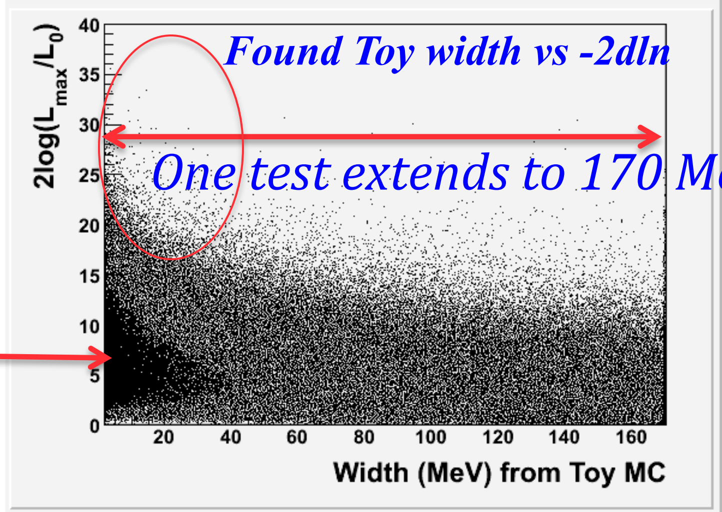
Fluctuation position and width distribution--CDF



Project to

Δm

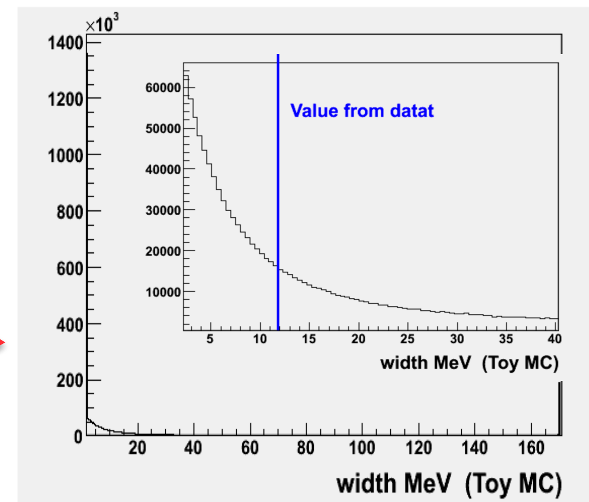




One test extends to 170 MeV

Project to

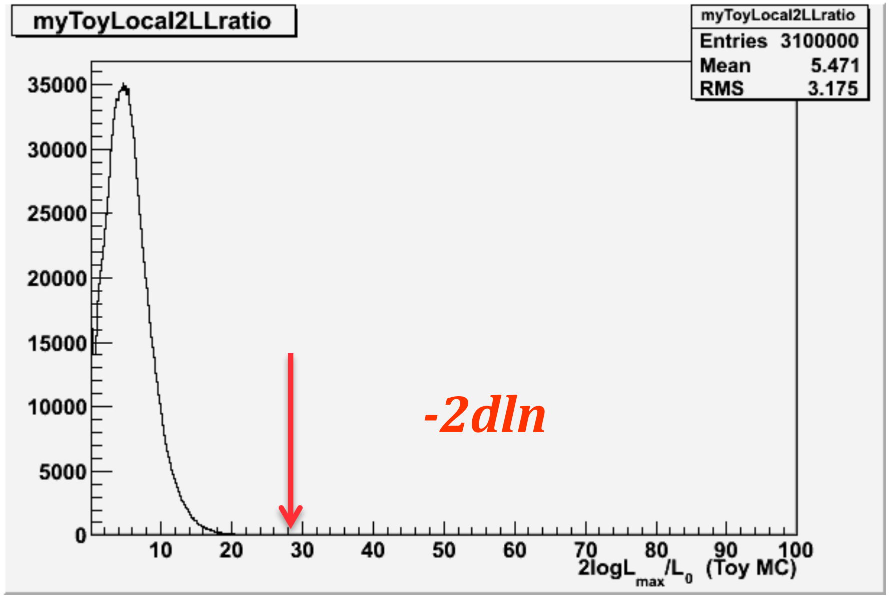
width



- *Most (significant) fluctuations happens at narrower width, i.e., 1.7 MeV an expected feature*

- *Fluctuation position distribution is flatter if width of the fluctuation is fixed*

Toy MC result--CDF



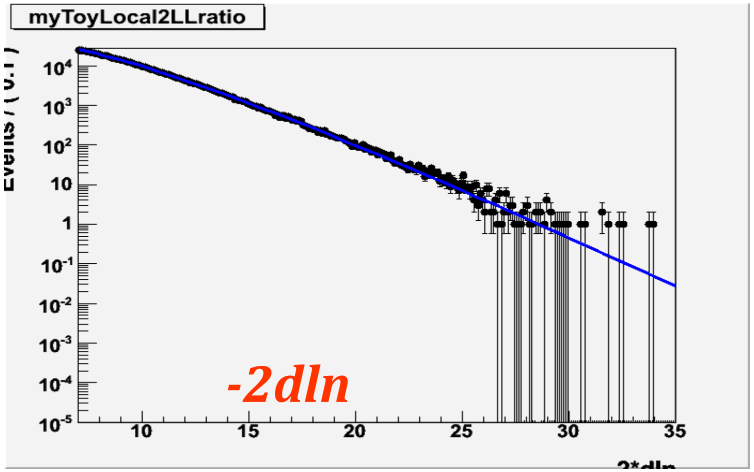
- *Each fit to Toy spectrum was done as the same as that in data —unbined loglikelihood fit*

- *Significance:*
  - *p-value from **counting**: (28/3.1million)=9.3X10^{-6}, corresponding to 4.3σ*
  - *drops from 5.3σ to 4.3σ due to LEE*
  - *A price due to **absence of a priori predictions for mass and width***
- *We further studied the -2dln distribution using χ^2 Probability Density Function*

Significance verification by χ^2-like PDF--CDF

χ^2 Probability Density Function (PDF)

$$f(z;n) = \frac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma(n/2)}\ ; \qquad z \geq 0$$



- *Significance from χ^2 PDF :*
  - *Fit -2dln (as z) distribution using χ^2 PDF to get n (d.o.f)*

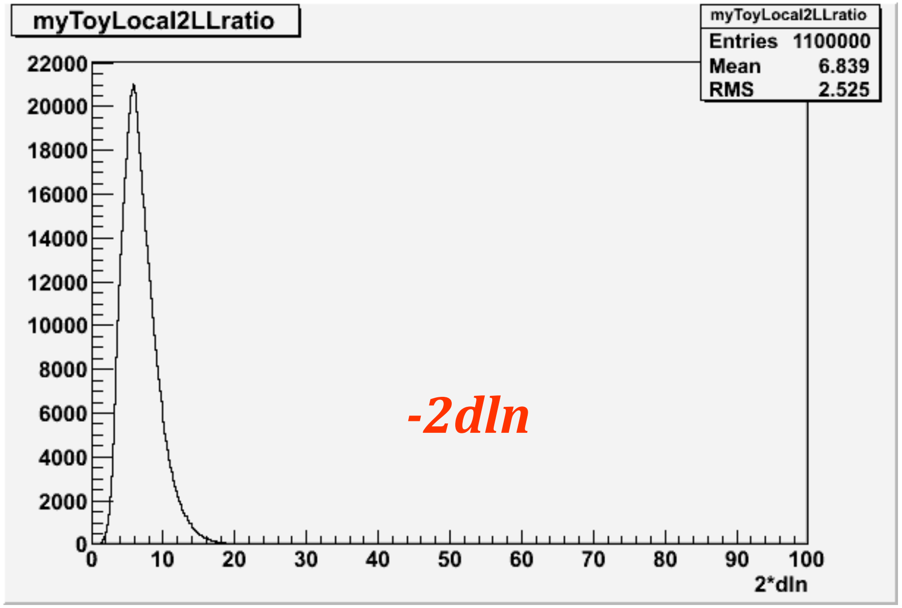
    --had to cut off the low -2dln part to have a converged fit, once it is converged, the actual cutoff position does not affect the final significance

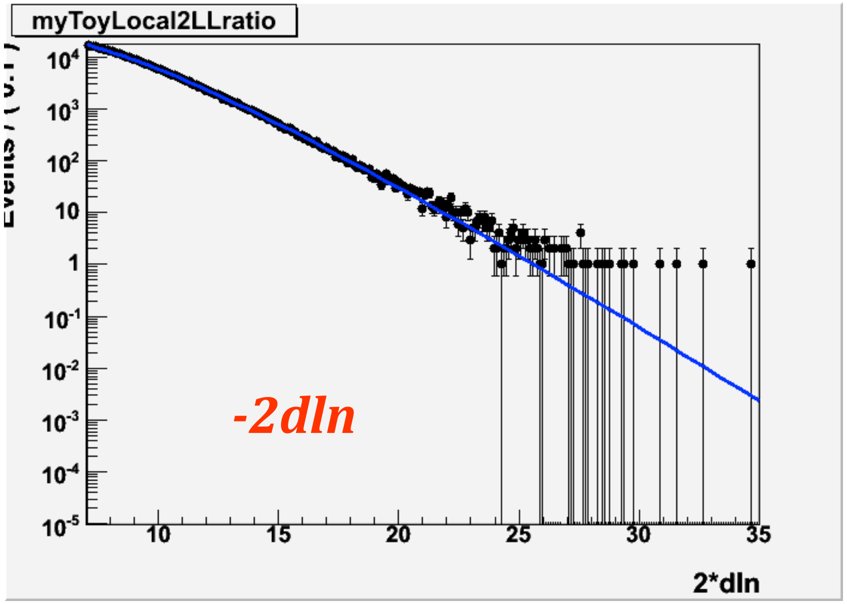
    --need to multiply a scale factor (float) to z in order to have a reasonable fit

    *just consider f(z*s,n) as a function to be able to fit the distribution*

    *--integrate f(z*s,n) function from 28.1 to infinity to get p-value*
  - *p-value from χ^2* ***PDF****: $6.5\mathrm{X}10^{-6}$, 4.3σ, consistent with counting*
  - *can extract p-values for high significant peaks from reasonable number of toys. 0.1σ means orders of magnitude difference in p-value for high significance*

Significance with another background--CDF

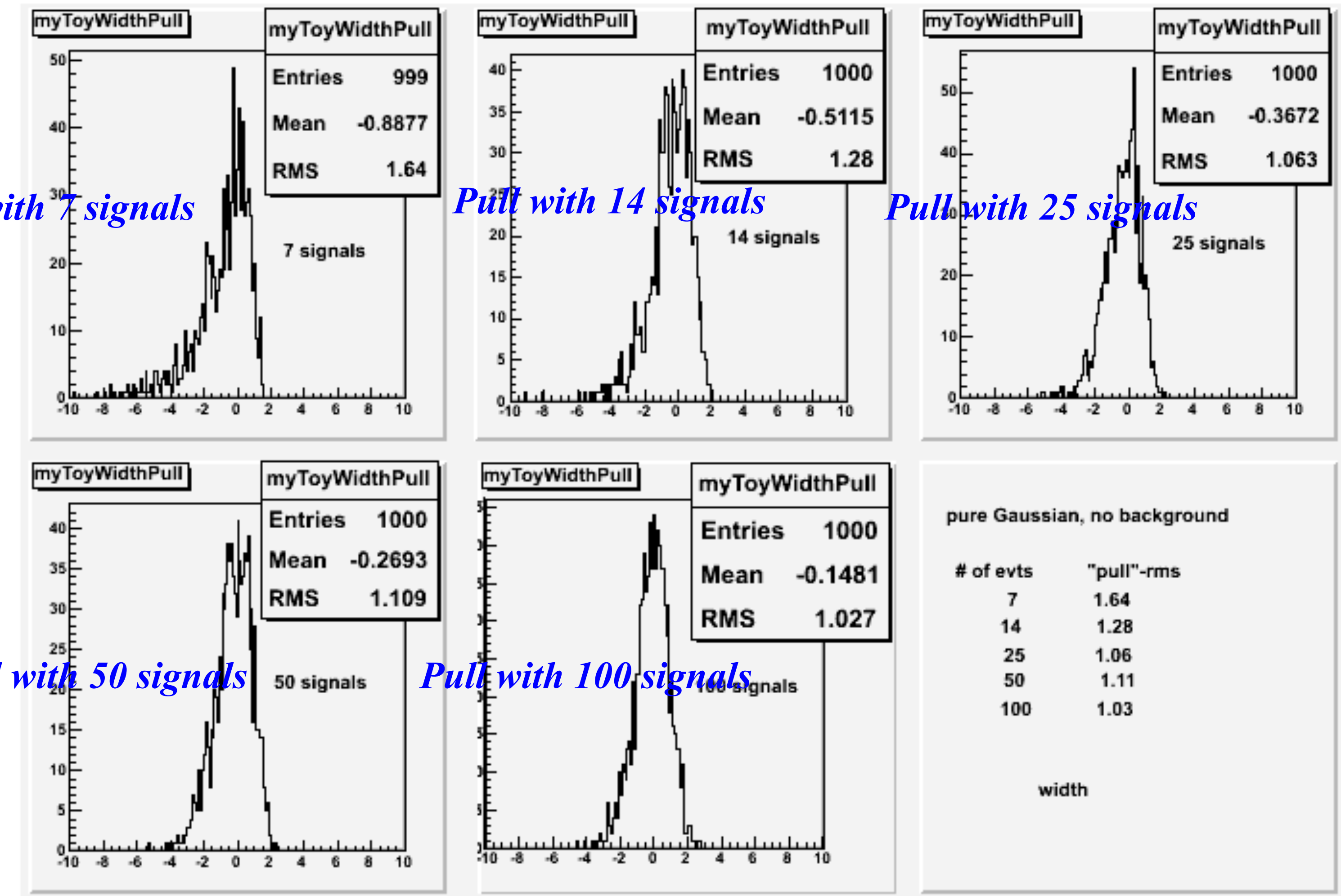




- *Significance with 3-body-PS and flat background for non-B component:*
  - *p-value from **counting**: (90/1.1million)=9.0X10^{-5}, corresponding to 3.8σ*
  - *p-value from χ^2 **PDF**: 2.4X10^{-5}, 3.8σ, consistent with counting*
  - *most conservative background but unphysical*

4.3σ could be the official significance result, this flat background was viewed as systematic consideration, that turns the official significance to be 3.8σ

Error underestimation with low statistics--CDF



- *Use pure signal to study the pull distributions of the fitted signal width and mass*
- *"pull" is defined as (observed-truth)/error-on-observed*

Error underestimation with low statistics--CDF

pure Gaussian, no background

# of evts	"pull"-rms
7	1.64
14	1.28
25	1.06
50	1.11
100	1.03

width

pure Gaussian, no background

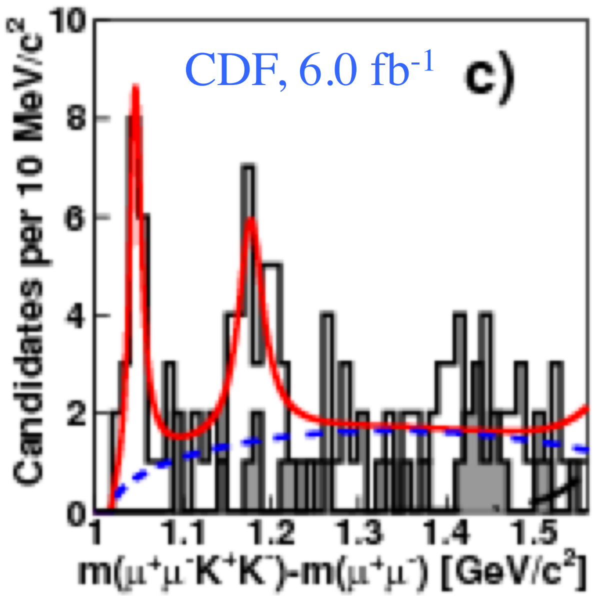
# of evts	"pull"-rms
7	1.34
14	1.14
25	1.08
50	1.08
100	0.98

mass

- *The estimated error depends on the observed value—an effect seen in low statistics, and it becomes negligible with higher statistics.*

- *However, it does not affect the significance number as the note stated in the paper: …*Such an underestimate of the fitted parameter uncertainties does not influence the evaluation of the signal significance, which depends only on the background fluctuation probability.

Significance for multiple peaks--CDF



- *The issue: two peaks in the same spectrum*
- *do we consider the existence of another peak when we evaluate the significance of one peak?*

- *The converged recommendation:*

  A peak can be considered as existence—with its signal shape as part of the background when evaluating the significance of other peaks only if that peak's significance reached 5sigma without considering the existence of other peaks.

- *The 1st peak reached 5.0σ ignoring the existence of the second peak after repeating the same Toy process. Thus considered the existence of the first peak for the significance of the 2nd peak, 3.1σ. Toy treatment similar to single peak case, will not be reviewed here.*

Determine when to give an update--CDF

- *The issue: CDF continued to collect data after published the first paper, what is the right time to give an update? It is a complicated question considering the following situations:*
    - *A real signal--significance get increased/decreased by adding data*
    - *A fluctuation--significance get increased/decreased by adding data*

- *Need certain statistics to confirm or deny. Take an optimistic altitude, in the case of confirmation, we asked that how much data is needed to add in order to reach 5sigma? One way to avoid randomess and have a control on the process*

- *Decided a priori:*
    - *find out the approximate -2dln value--$2dln_{min}$ to reach 5sigma*
    - *perform Toy MC to find out the probability--p_{min} that 2dln to be equal or bigger than $2dln_{min}$ by combined x% of new data.*
    - *give an update if p_{min} >=75%*

  We estimate the approximate $2dln_{min}$ to be 35 from previous χ^2 PDF. Implicitly assumed the signal and background will behave the same as that in the previous data , considering the CDF stable data-taking situation: same energy, same instantaneous luminosity, same trigger, stable detector…

The probability to reach 5sigma by adding 60% data-CDF

- *Toy MC sample for each trial:*
  - *base data: the published CDF data--14±5 signal out of 75 events*

  - *adding 60% signal events: use a Gaussian function (mean=14*0.60, sigma=5*0.60) to generate a number n1, use a Poisson function (mean=n1) to generate a number n2. n2 events based on BW PDF were add to the basis*

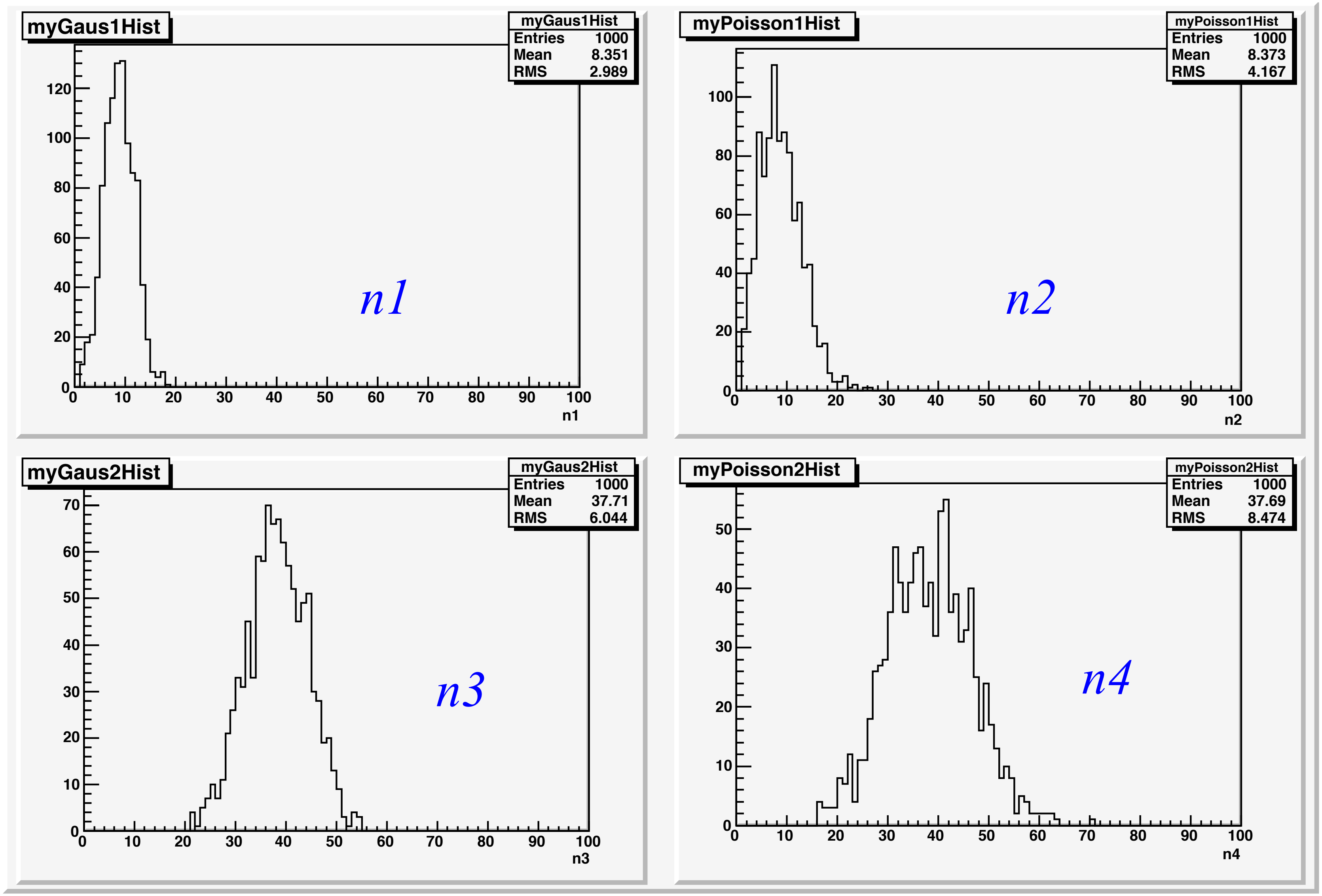
  - *adding 60% background events: assume a similar behavior for non-resonance B background and combinatorial background.   We expect total 37.6 background events. Use a Gaussian function (mean=37.6, sigma=sqrt(37.6)) to generate number n3, use a Poisson function (mean=n3) to generate a number n4. Then generate n4 events using 3-body phase space PDF and add to the basis*

  - *For each trial, we added n2 signal events and n4 background events to the original sample. A null- and signal-hypothesis fits were done to each mixed sample and all -2dln value were recorded*

note: assumed the same behavior for previous data and new data

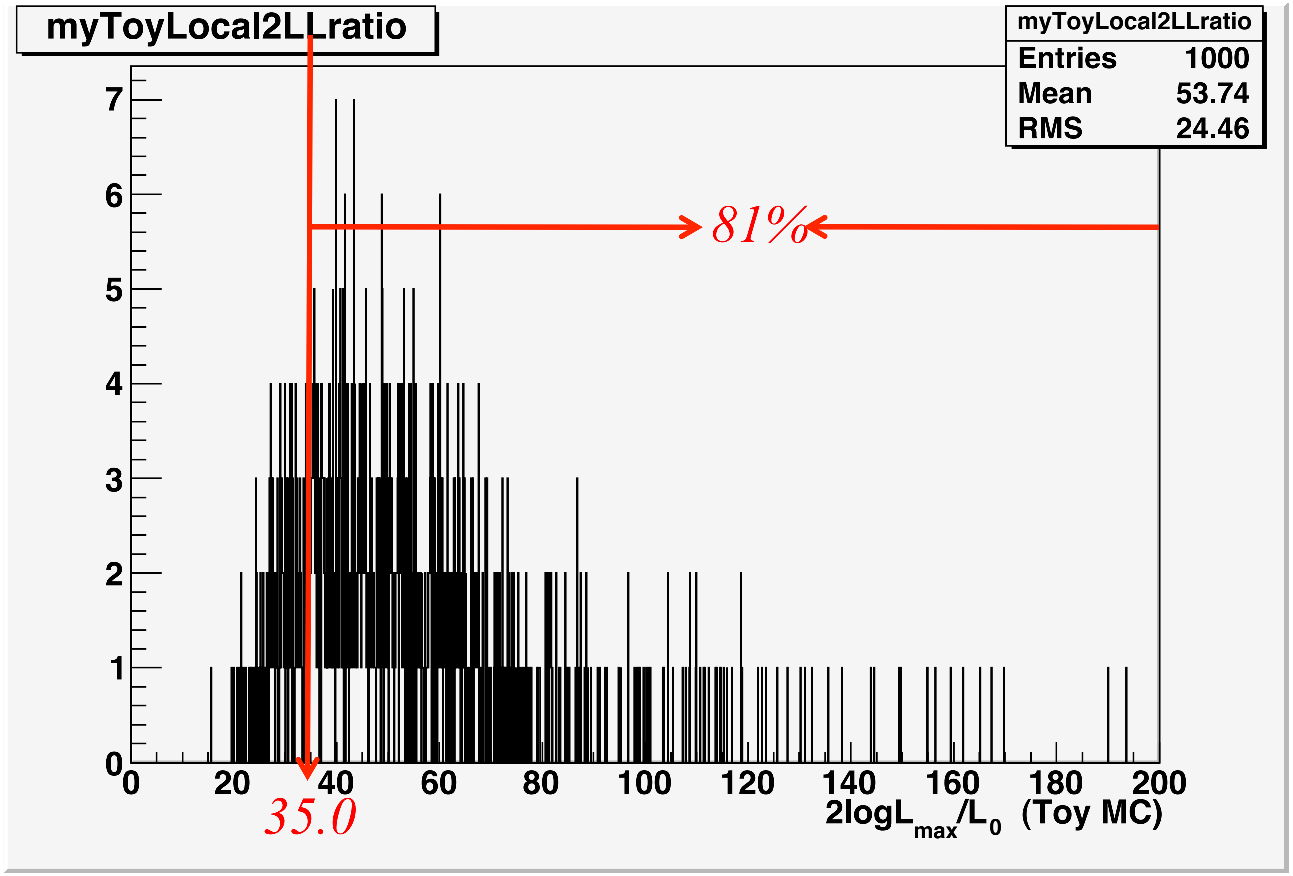
The n1, n2, n3, n4 distribution from Toy-CDF

- n1,n2,n3,n4 distribution from 1000 trials



The probability to reach 5sigma by adding 60% data-CDF

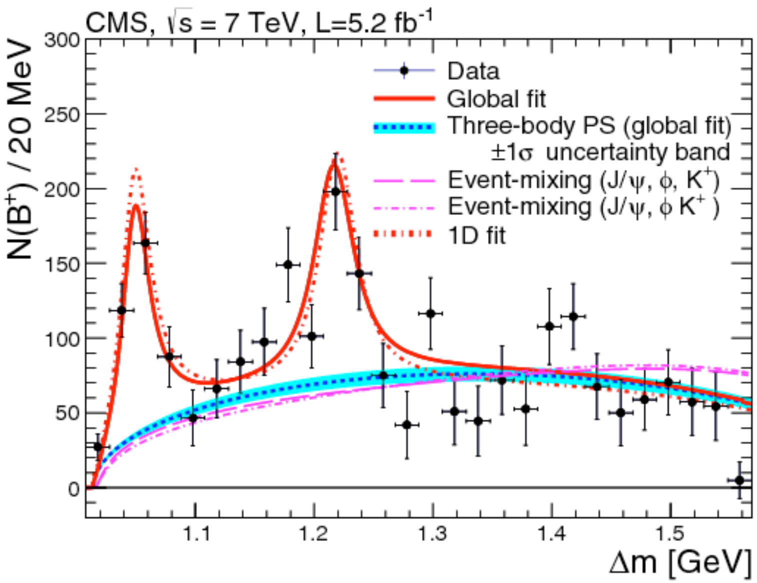
- *-2dln distribution from 1000 trials*

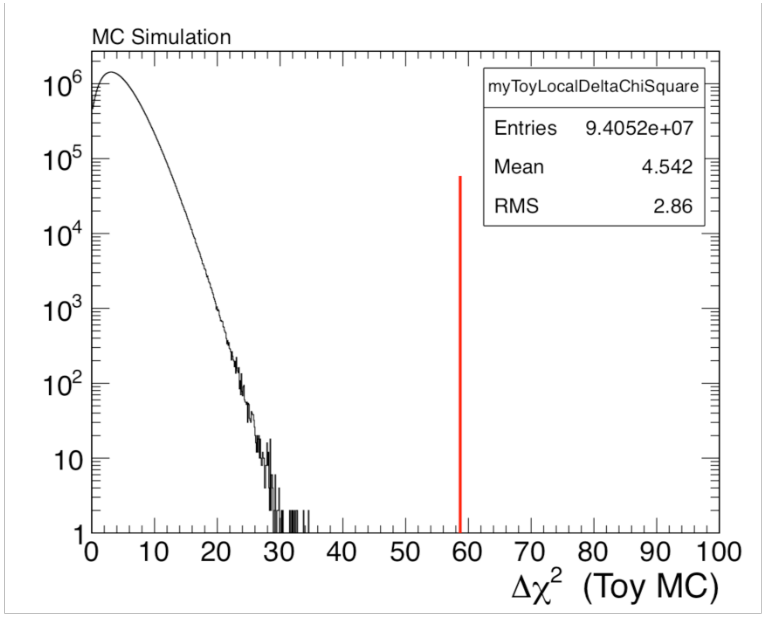


The probability to pass 5 sigma is 81% by adding 60% more data for Y(4140). We passed the threshold--75%, thus an update was given in 2010.

arXiv:1101.6058 [hep-ex]

Significance for very significant peaks--CMS





- *Basically repeat the Toy MC procedure in CDF for the first peak. LEE was considered giving the lack of confirmation at that time.*
- *Eventually performed 50+ million Toys, no single Toy with -2dln value to be greater than 58 (the -2dln value in date)*
- *Limited by CPU resource, we simply set significance lower limit >5.0 sigma based on p-value limit > 2x10^{-8}.*
- *It does not add anything by finding a specific significance number--huge amount of resource is needed.*

Summary

- *presented first-hand experience on statistical application in hadron spectra in different experiments*

- *The first explicit example to take LEE via Toys for unexpected X/Y/Z*
    - *A point on the fluctuation dependence on width*
    - *A point on error underestimation for low statistic case*

- *An example to use Toy to estimation the probability to reach certain significance by adding more data*

Thank you!