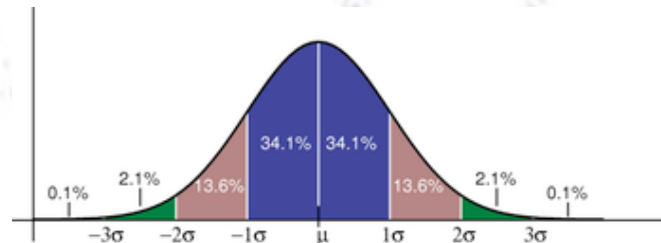


Applied Statistics

Advanced fitting



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Defining the Chi-Square

Problem Statement: Given N data points (x, y) , adjust the parameter(s) θ of a model, such that it fits data best.

The best way to do this, given uncertainties σ_i on y_i is by minimising:

$$\chi^2(\theta) = \sum_i^N \frac{(y_i - f(x_i, \theta))^2}{\sigma_i^2}$$

The power of this method is hard to overstate!

Not only does it provide a simple, elegant and unique way of fitting data, but more importantly it provides a **goodness-of-fit measure**.

This is the Chi-Square test!

Chi-Square probability interpretation

The Chi-Square probability can roughly be interpreted as follows:

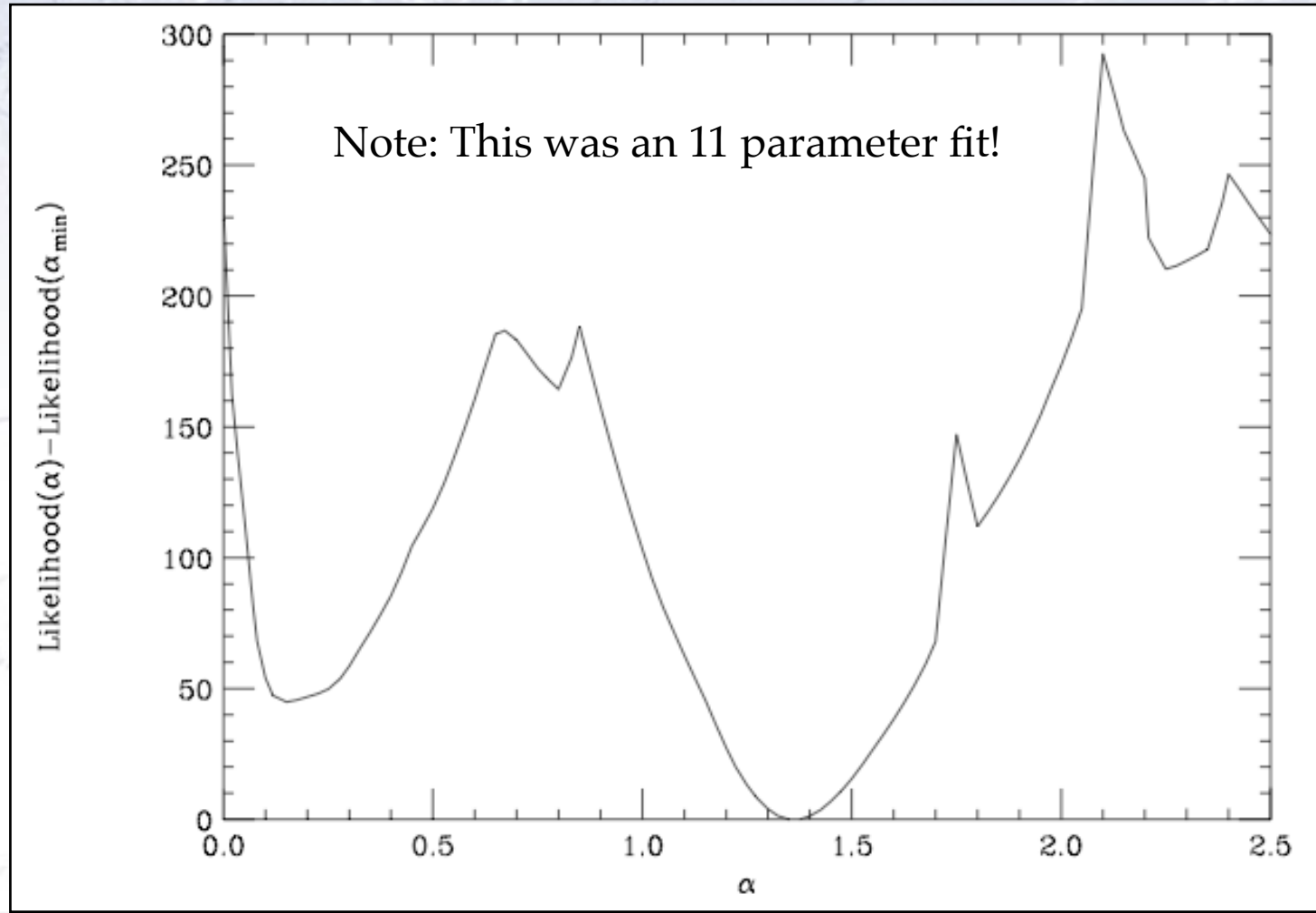
- If $\chi^2 / \text{Ndof} \approx 1$ or more precisely if $0.01 < p(\chi^2, \text{Ndof}) < 0.99$, then all is good.
- If $\chi^2 / \text{Ndof} \gg 1$ or more precisely if $p(\chi^2, \text{Ndof}) < 0.01$, then your fit is bad, and your hypothesis is probably not correct.
- If $\chi^2 / \text{Ndof} \ll 1$ or more precisely if $0.99 < p(\chi^2, \text{Ndof})$, then your fit is TOO good and you probably overestimated the errors.

If the statistics behind the plot is VERY high (great than 10^6), then you might have a hard time finding a model, which truly describes all the features in the plot (as now tiny effects become visible), and one hardly ever gets a good Chi-Square probability.

However, in this case, one should not worry too much, unless very high precision is wanted.

Anyway, the Chi-Square still allows you to compare several models, and determine which one is the better.

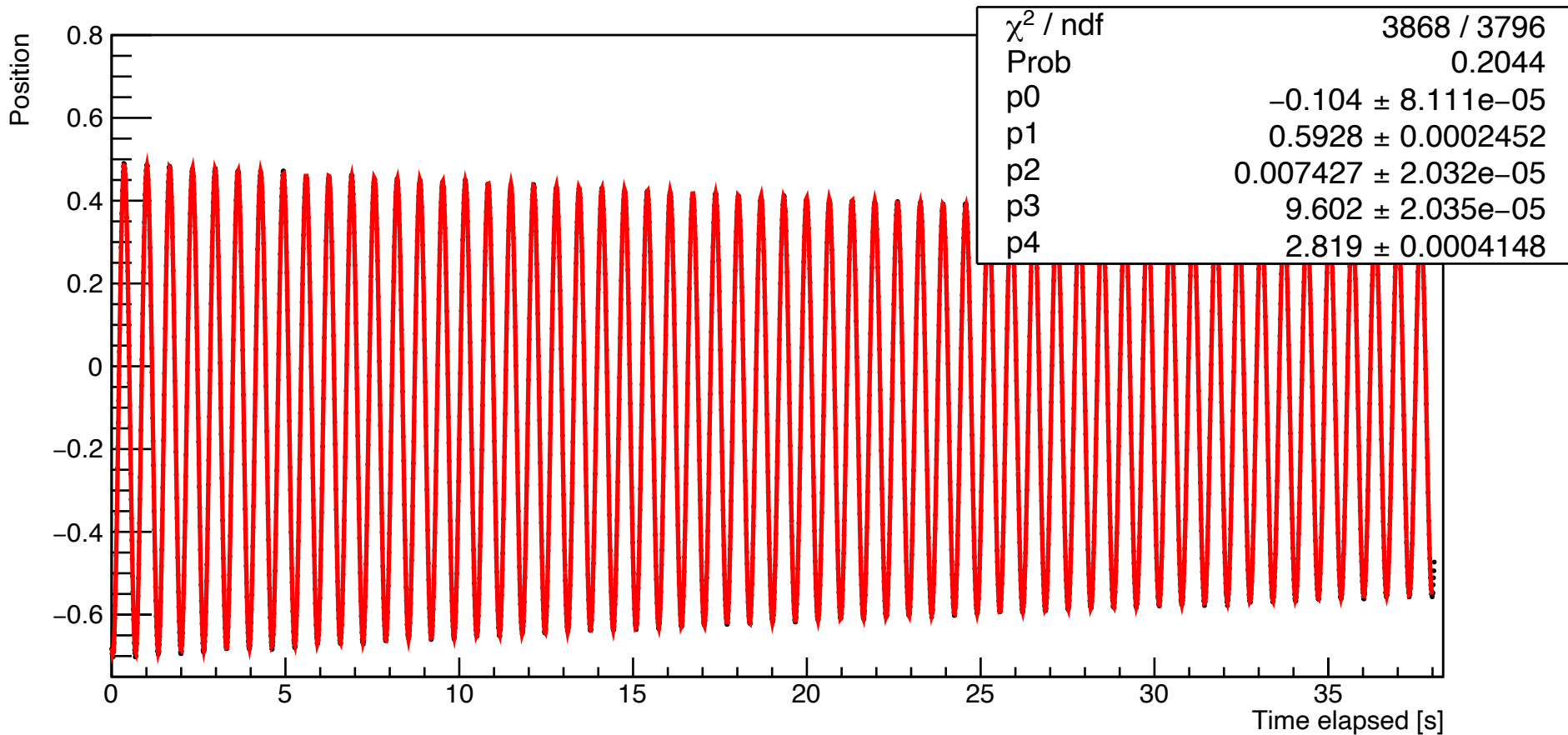
Example likelihood “landscape”



The fact that there are several minima makes fitting difficult/uncertain!
Always give good starting values!!!

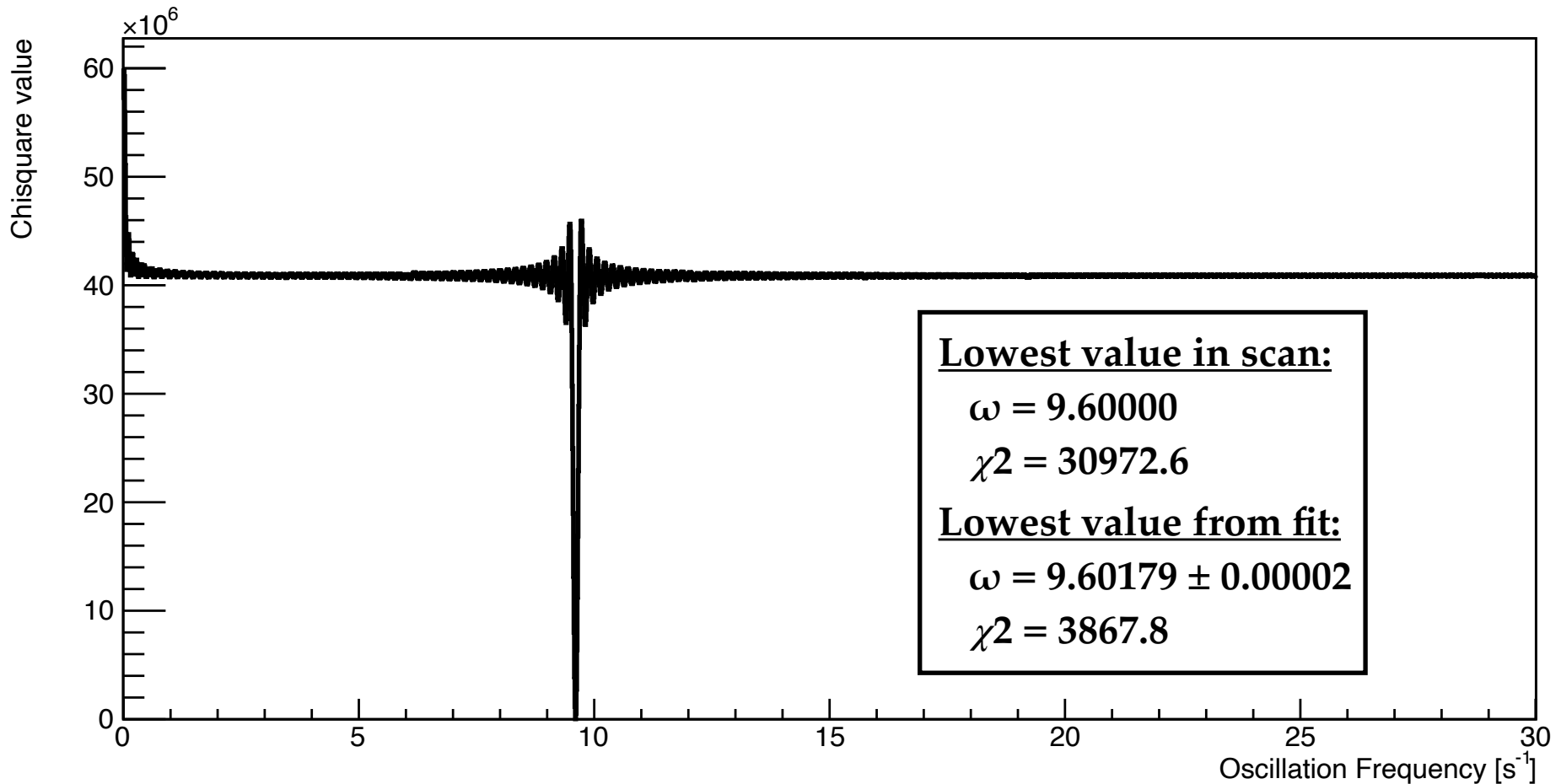
Example of Chi-Square

Especially fitting oscillatory data requires a good starting value for omega.
Even a small offset may result in a Chi2, which is not sensitive to omega!



Example Chi-Square landscape

Especially fitting oscillatory data requires a good starting value for omega.
Even a small offset may result in a Chi2, which is not sensitive to omega!





When to use what type of fit?

When to use what type of fit?

Fitting a set of points:

When fitting a set of points, each with values for x , y , σ_y (and possibly σ_x), I would **always choose a ChiSquare fit**:

- 1) It is equivalent to a likelihood fit (errors are Gaussian) and thus optimal.
- 2) It yields a goodness-of-fit measure and thus the essential p-value.

Fitting distribution of values (i.e. histogram):

When fitting a histogram with high statistics, the situation reduces to that above, since the bins will have Gaussian errors. However, care has to be taken to the choice of binning.

However, if the statistics is (very?) small, the likelihood is preferable. If possible, one should use the **unbinned likelihood fit**. The binned likelihood is “only” to be used in one of the following cases:

- 1) When you don't have unbinned data!
- 2) When the data is not continuous, but categorical i.e. binned (e.g. integers).

It is very often smart to start with a ChiSquare, as this has better convergence.



Tricks in fitting

Correlations between parameters

The fit parameters should have as little correlation as possible. If two (or more) are very correlated, then they **represent the same feature** in the model, and one should possibly be fixed or a relation made between them.

Example 1 - Fitting two Gaussians with common mean:

The “naive” approach would probably be:

$$N_1 G_1(x, \mu, \sigma_1) + N_2 G_2(x, \mu, \sigma_2)$$

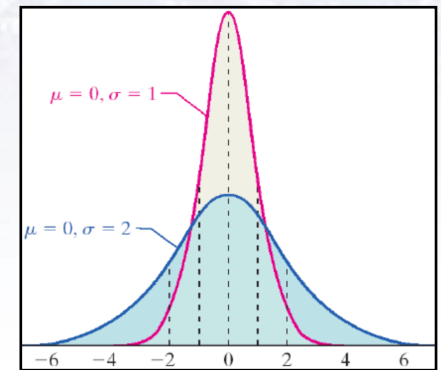
But here N_1 and N_2 will be *very* correlated, avoided by:

$$N \times (f G_1(x, \mu, \sigma_1) + (1 - f) G_2(x, \mu, \sigma_2))$$

Now, N represents the overall number of events, while f is the fraction of G_1 .

In any case, f , σ_1 and σ_2 will be correlated, which can not be avoided. If any knowledge about their values or relation is known, this can with great advantage be included, e.g.

- Fixing one of the parameter values, i.e. $f = 0.8$.
- Fixing the relation between σ_1 and σ_2 , i.e. $\sigma_2 = 3 \times \sigma_1$.



Correlations between parameters

The fit parameters should have as little correlation as possible. If two (or more) are very correlated, then they **represent the same feature** in the model, and one should possibly be fixed or a relation made between them.

Example 2 - Fitting two exponential functions:

The “naive” approach could be:

$$N_1 \exp(\tau_1) + N_2 \exp(\tau_2)$$

Again, N_1 and N_2 will be correlated, and due to the lacking normalisation, N_1 and N_2 will also (unnecessarily) be correlated with τ_1 and τ_2 . Instead use:

$$N(f/\tau_1 \cdot \exp(-t/\tau_1) + (1 - f)/\tau_2 \cdot \exp(-t/\tau_2))$$

Now, N represents the overall number of events, while f is the fraction of Exp1. There will still be strong correlations between especially τ_1 and τ_2 , and again it might be worthwhile to put in a relation/constraint between them or fixing f .

Global fits

Occasionally, one has several samples to be fitted, which have overlapping parameters. In this case, one can make a “global” fit of all parameters. Both for a ChiSquare and a Likelihood fit.

The advantage is that all parameters are determined simultaneously, thus **automatically including all correlations and their effects.**

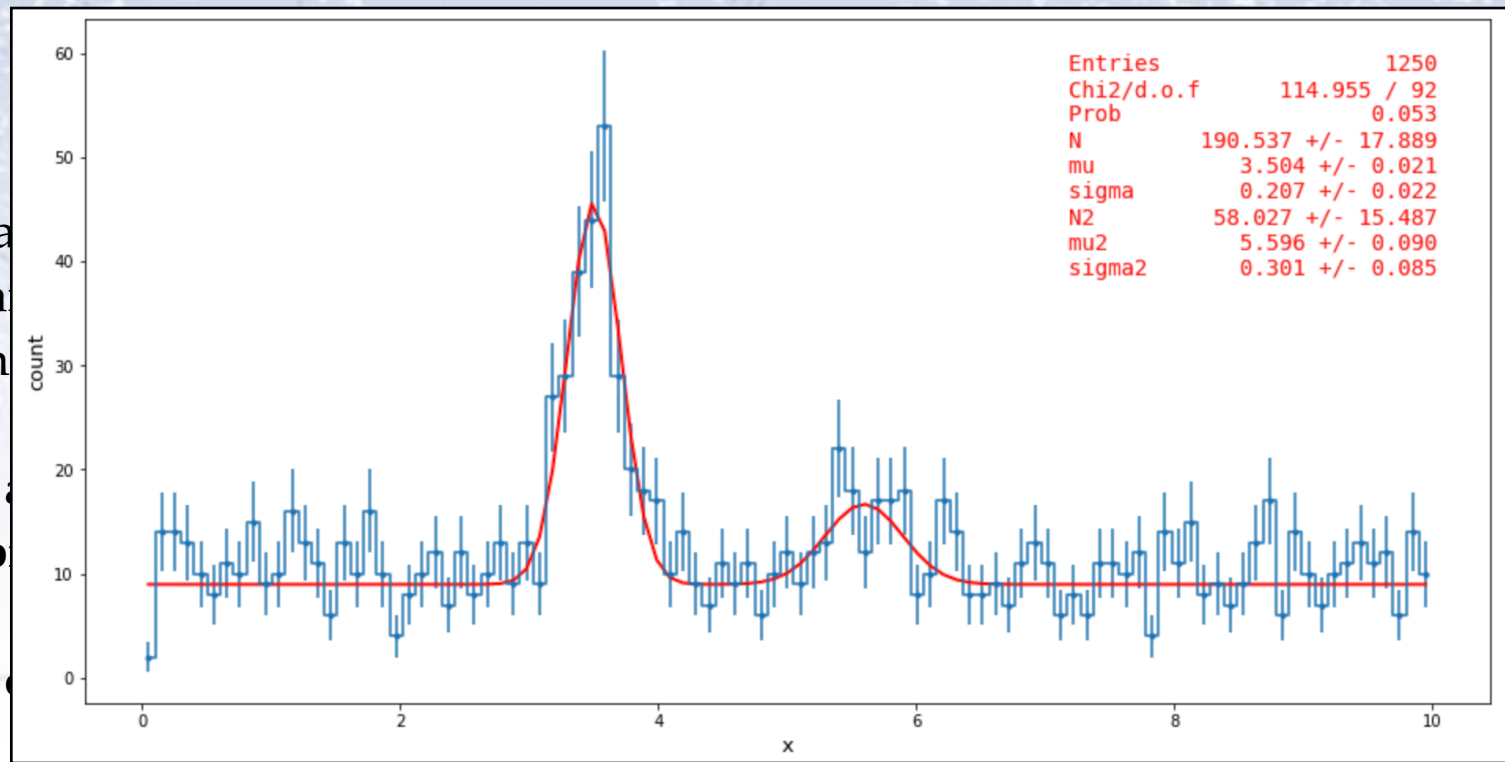
The down side is that fit gets more complicated (see advanced example later).

An example could be, that you have two samples or features/peaks:

1. Signal sample (i.e. the one of interest).
2. A larger calibration sample, which also includes some of the parameters to be determined in the signal sample.

In this case, including the calibration sample in the fit will help constrain the parameters, which the calibration sample are sensitive to. Example: Large calibration (C) peak next to a much smaller signal (S) peak, where $\sigma_C = \sigma_S$.

Occa
para
a Ch
The
auto
The



for
r).

An example could be, that you have two samples or features/peaks:

1. Signal sample (i.e. the one of interest).
2. A larger calibration sample, which also includes some of the parameters to be determined in the signal sample.

In this case, including the calibration sample in the fit will help constrain the parameters, which the calibration sample are sensitive to. Example: Large calibration (C) peak next to a much smaller signal (S) peak, where $\sigma_C = \sigma_S$.

ChiSquare penalty terms

If a parameter θ_n is known ($\theta_n = X$), it should (of course) be fixed:

$$\chi^2(\boldsymbol{\theta}) = \sum_i^N \frac{(y_i - f(x_i, \boldsymbol{\theta}))^2}{\sigma_i^2}, \quad \theta_n = X$$

However, what to do, if a parameter θ_n is partially known ($\theta_n = X \pm \sigma_X$) from some other source?

ChiSquare penalty terms

If a parameter θ_n is known ($\theta_n = X$), it should (of course) be fixed:

$$\chi^2(\boldsymbol{\theta}) = \sum_i^N \frac{(y_i - f(x_i, \boldsymbol{\theta}))^2}{\sigma_i^2}, \quad \theta_n = X$$

However, what to do, if a parameter θ_n is partially known ($\theta_n = X \pm \sigma_X$) from some other source?

The trick is to include a “penalty” term, which penalises the fit for choosing a value of θ_n far from the known range:

$$\chi^2(\boldsymbol{\theta}) = \sum_i^N \frac{(y_i - f(x_i, \boldsymbol{\theta}))^2}{\sigma_i^2} + \frac{(\theta_n - X)^2}{\sigma_X^2}$$



Fitting in multiple dimensions

4 components in 3 dimensions

The study of B-mesons (particles consisting of a quark and an anti-quark, where one is a b-quark) is a fields where fitting in multiple dimensions is often used.

At the BaBar experiment at Stanford, B-mesons were studied using three almost uncorrelated variables:

- Invariant mass of the decay particles (Gaussian around true B mass).
- Energy difference between beam and B-meson (Gaussian around 0).
- Shape of decay products (approximately asymmetrically Gaussian).

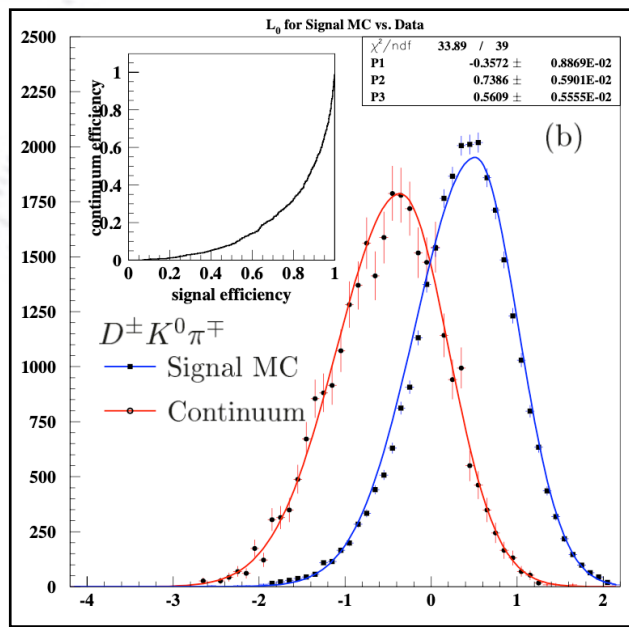
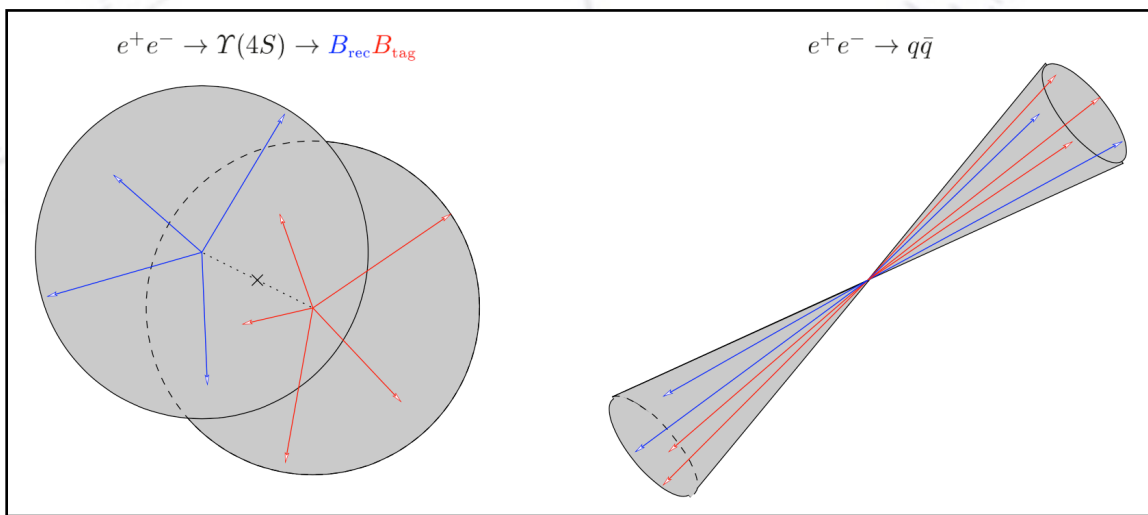
4 components in 3 dimensions

The study of B-mesons (particles consisting of a quark and an anti-quark, where one is a b-quark) is a fields where fitting in multiple dimensions is often used.

At the BaBar experiment at Stanford, B-mesons were studied using three almost uncorrelated variables:

- Invariant mass of the decay particles (Gaussian around true B mass).
- Energy difference between beam and B-meson (Gaussian around 0).
- Shape of decay products (approximately asymmetrically Gaussian).

$$\mathcal{F} \equiv c_0 + c_1 L_0 + c_2 L_2 + c_3 |\cos \theta_{(\vec{p}_B, \vec{z})}| + c_4 |\cos \theta_{(\vec{T}_B, \vec{z})}|$$



4 components in 3 dimensions

The study of B-mesons (particles consisting of a quark and an anti-quark, where one is a b-quark) is a field where fitting in multiple dimensions is often used.

At the BaBar experiment at Stanford, B-mesons were studied using three almost uncorrelated variables:

- Invariant mass of the decay particles (Gaussian around true B mass).
- Energy difference between beam and B-meson (Gaussian around 0).
- Shape of decay products (approximately asymmetrically Gaussian).

The challenge is that despite best efforts, there are several background sources mixed with the signal in the selected data sample.

One advantage is that good estimates of all distributions are known from simulated data. This allows one to fix some parameters in the (large) fit.

4 components in 3 dimensions

The study of B-mesons (particles consisting of a quark and an anti-quark, where one is a b-quark) is a fields where fitting in multiple dimensions is often used.

At the BaBar experiment at Stanford, B-mesons were studied using three almost uncorrelated variables:

- Invariant mass of the decay particles (Gaussian around true B mass).
- Energy difference between beam and B-meson (Gaussian around 0).
- Shape of decay products (approximately asymmetrically Gaussian).

Component	Signal	Continuum	$B\bar{B}$	$B\bar{B}$ peak	Total
m_{ES}	G_1	Argus ₁	Argus ₂	G_2	7
ΔE	GG	$P1_1$	$P1_2$	Exp	7
\mathcal{F}	BG_1	BG_2	BG_3	BG_3	9
N parameters	9	6	6	6	23

Table 10.1: PDFs used for signal, continuum, and $B\bar{B}$ background (non-peaking and non-degenerate peaking). Four parameters are common among components, which has to included when summing the bottom line of the table. The abbreviations are G = Gaussian, GG = double Gaussian, $P1$ = polynomial of first degree, Exp = exponential, and BG = Bifurcated Gaussian. The color code is Black: Fixed in fit, Magenta: fixed from MC, and Green: Free.

4 components in 3 dimensions

Once the PDFs to be used had been determined, all that remained was to write up an unbinned maximum likelihood fit, minimise it, and see the result...

The events are fitted with an extended unbinned maximum likelihood fit containing the variables m_{ES} , ΔE , and \mathcal{F} . A probability product, $P_j(m_{\text{ES},i}, \Delta E_i, \mathcal{F}_i) = P_j(m_{\text{ES},i}) \cdot P_j(\Delta E_i) \cdot P_j(\mathcal{F}_i)$, is assigned to each event, i , and an unbinned likelihood is constructed:

$$\ln \mathcal{L} = \sum_i \ln \left(\sum_{j=\text{comp.}} N_j P_j(m_{\text{ES},i}, \Delta E_i, \mathcal{F}_i) \right) - \sum_{j=\text{comp.}} N_j, \quad (10.1)$$

where the sum j is over the four components of the sample (signal, continuum, combinatorial and peaking background), and N_j is the number of events in each component.

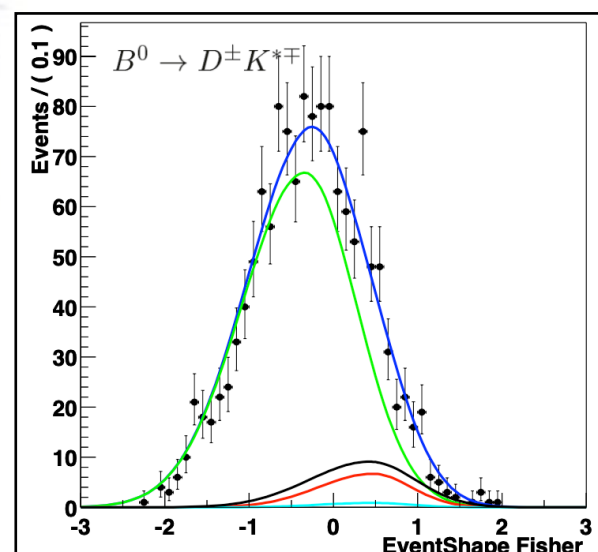
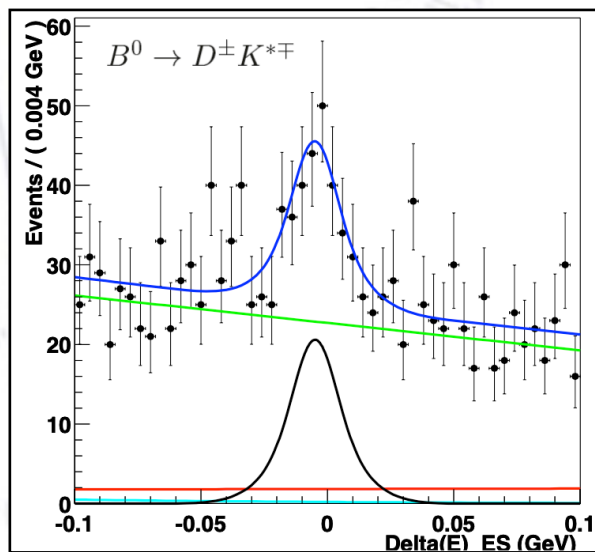
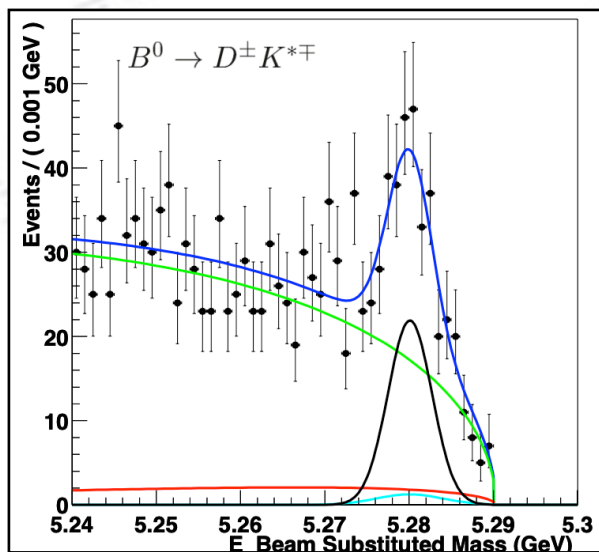
4 components in 3 dimensions

Once the PDFs to be used had been determined, all that remained was to write up an unbinned maximum likelihood fit, minimise it, and see the result...

The events are fitted with an extended unbinned maximum likelihood fit containing the variables m_{ES} , ΔE , and \mathcal{F} . A probability product, $P_j(m_{\text{ES},i}, \Delta E_i, \mathcal{F}_i) = P_j(m_{\text{ES},i}) \cdot P_j(\Delta E_i) \cdot P_j(\mathcal{F}_i)$, is assigned to each event, i , and an unbinned likelihood is constructed:

$$\ln \mathcal{L} = \sum_i \ln \left(\sum_{j=\text{comp.}} N_j P_j(m_{\text{ES},i}, \Delta E_i, \mathcal{F}_i) \right) - \sum_{j=\text{comp.}} N_j, \quad (10.1)$$

where the sum j is over the four components of the sample (signal, continuum, combinatorial and peaking background), and N_j is the number of events in each component.



Component	Parameter	Fixed	Value	Unit	Source
Signal	m_{ES} Mean	Yes	5.2801 ± 0.0001	GeV	$D^\pm a_1$
	m_{ES} Width	Yes	2.612 ± 0.060	MeV	$D^\pm a_1$
	ΔE Mean	Yes	-4.89 ± 0.30	MeV	$D^\pm a_1$
	ΔE Width ₁	Yes	8.18 ± 1.18	MeV	$D^\pm a_1$
	ΔE Width ₂	Yes	15.55 ± 2.11	MeV	$D^\pm a_1$
	ΔE f_{G1}	Yes	0.443 ± 0.186	–	$D^\pm a_1$
	\mathcal{F} Mean	Yes	0.432 ± 0.025	–	$D^{*\pm} \pi$
	\mathcal{F} σ_{Left}	Yes	0.716 ± 0.017	–	$D^{*\pm} \pi$
	\mathcal{F} σ_{Right}	Yes	0.539 ± 0.016	–	$D^{*\pm} \pi$
Continuum	m_{ES} Shape	No	-19.2 ± 3.3	–	
	m_{ES} Endpoint	Yes	5.2903 ± 0.0001	GeV	$D^\pm a_1$
	ΔE Slope	No	-1.74 ± 0.29	ev/ GeV	
	\mathcal{F} Mean	Yes	-0.339 ± 0.066	–	Off-Resonance
	\mathcal{F} σ_{Left}	Yes	0.744 ± 0.043	–	Off-Resonance
	\mathcal{F} σ_{Right}	Yes	0.613 ± 0.041	–	Off-Resonance
$B\bar{B}$	m_{ES} Shape	No	-13.3 ± 9.3	–	
	ΔE Slope	No	-0.46 ± 0.72	ev/ GeV	
	\mathcal{F} Mean	Yes	0.468 ± 0.018	–	$B\bar{B}$ Generic MC
	\mathcal{F} σ_{Left}	Yes	0.640 ± 0.012	–	$B\bar{B}$ Generic MC
	\mathcal{F} σ_{Right}	Yes	0.454 ± 0.011	–	$B\bar{B}$ Generic MC
$B\bar{B}$ peak	m_{ES} Mean	Yes	5.2801 ± 0.0008	GeV	$B\bar{B}$ Generic MC
	m_{ES} Width	Yes	3.95 ± 0.56	MeV	$B\bar{B}$ Generic MC
	ΔE Exp. Coef.	No	-8.6 ± 6.3	GeV^{-1}	



Very complicated fitting

Very advanced fitting

Sometimes, one has several samples (signal, control, validation, background, etc.) to fit, and one would like to include the full information in all of these.

If one (input) sample dominates in size, then there is no reason not to fit this sample separately, and then just fix the parameters to the result of this fit.

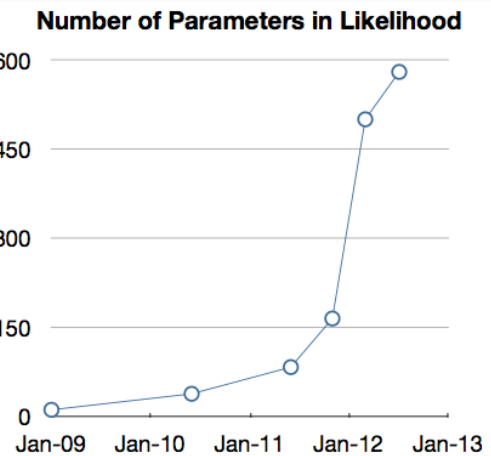
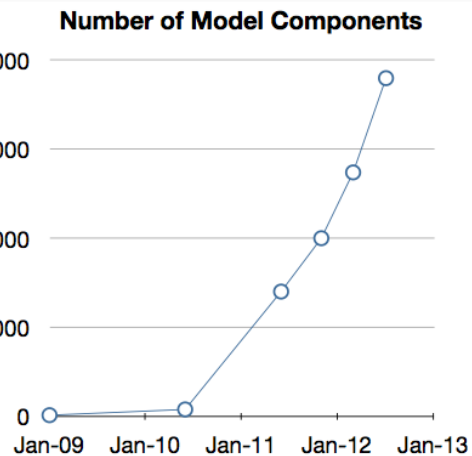
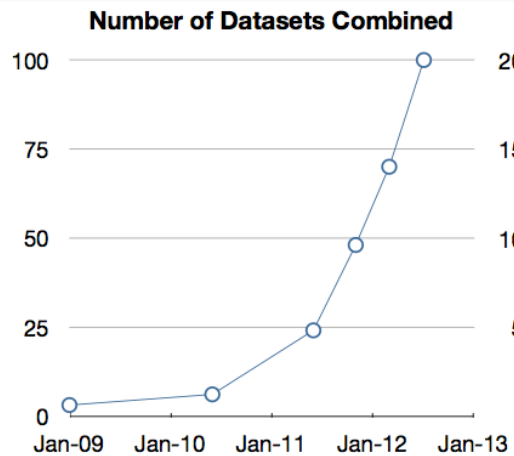
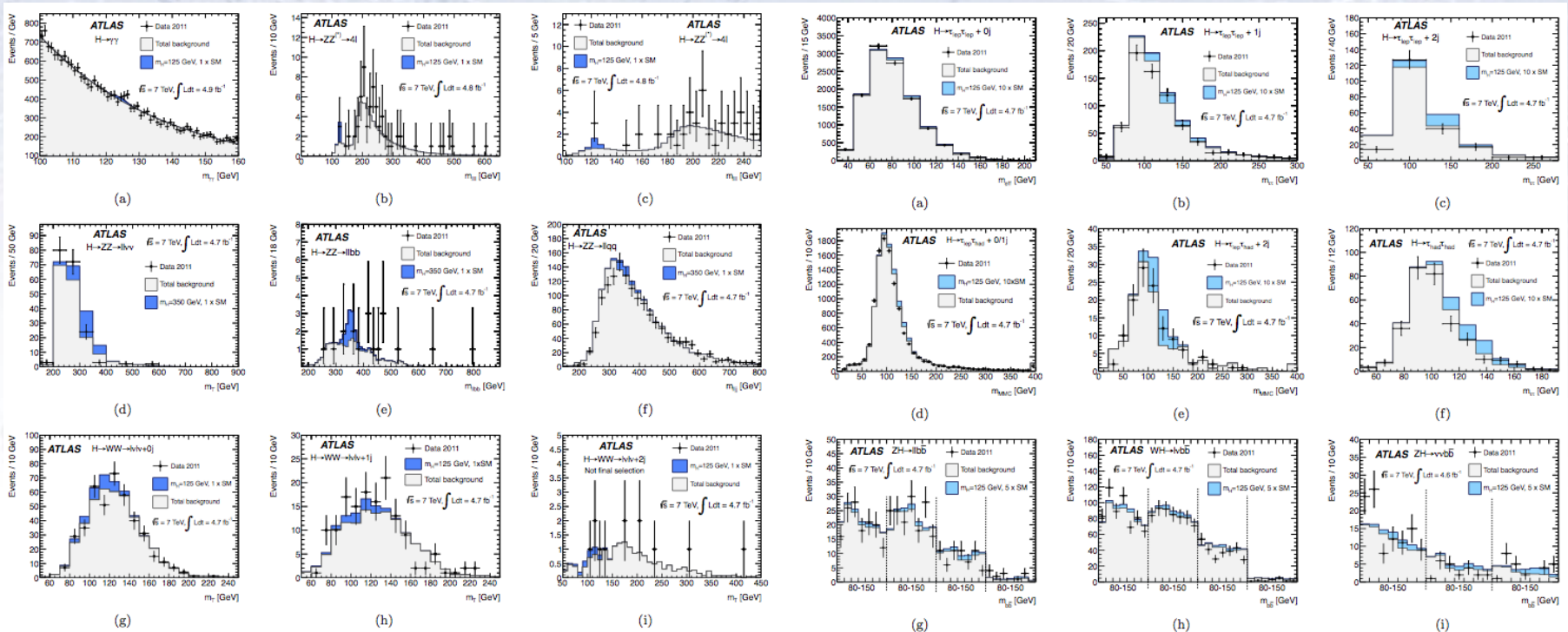
However, if they are of similar size, then a simultaneous fit of the samples is the optimal solution. This can grow quite large...

In 2002 the BaBar collaborations fit for CP-violation ($\sin 2\beta$) included **98 floating parameters** applied to four datasets.

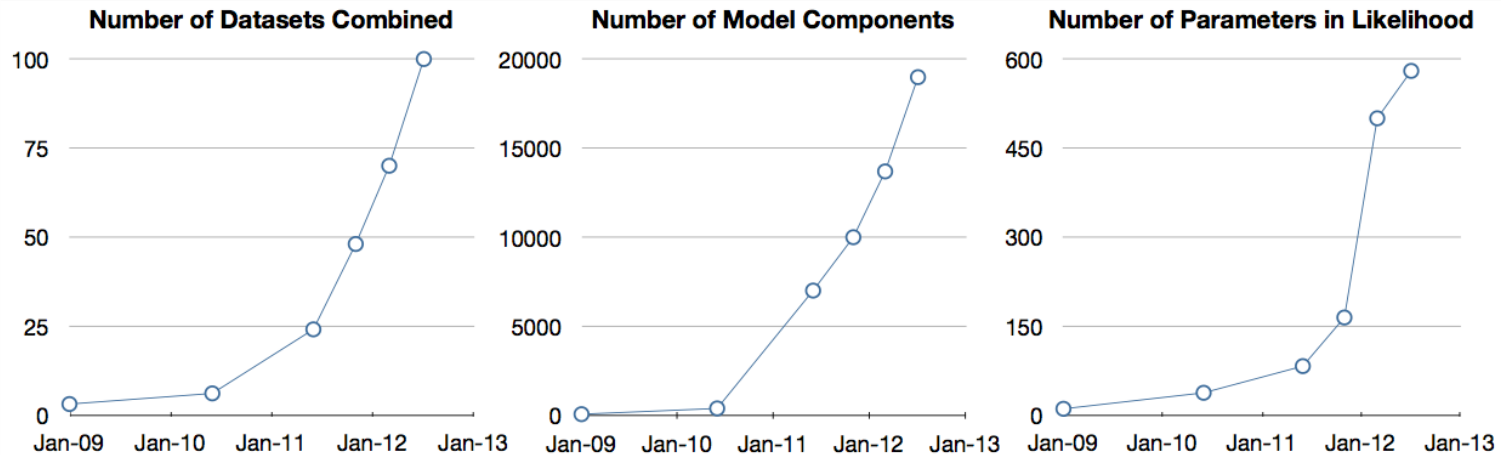
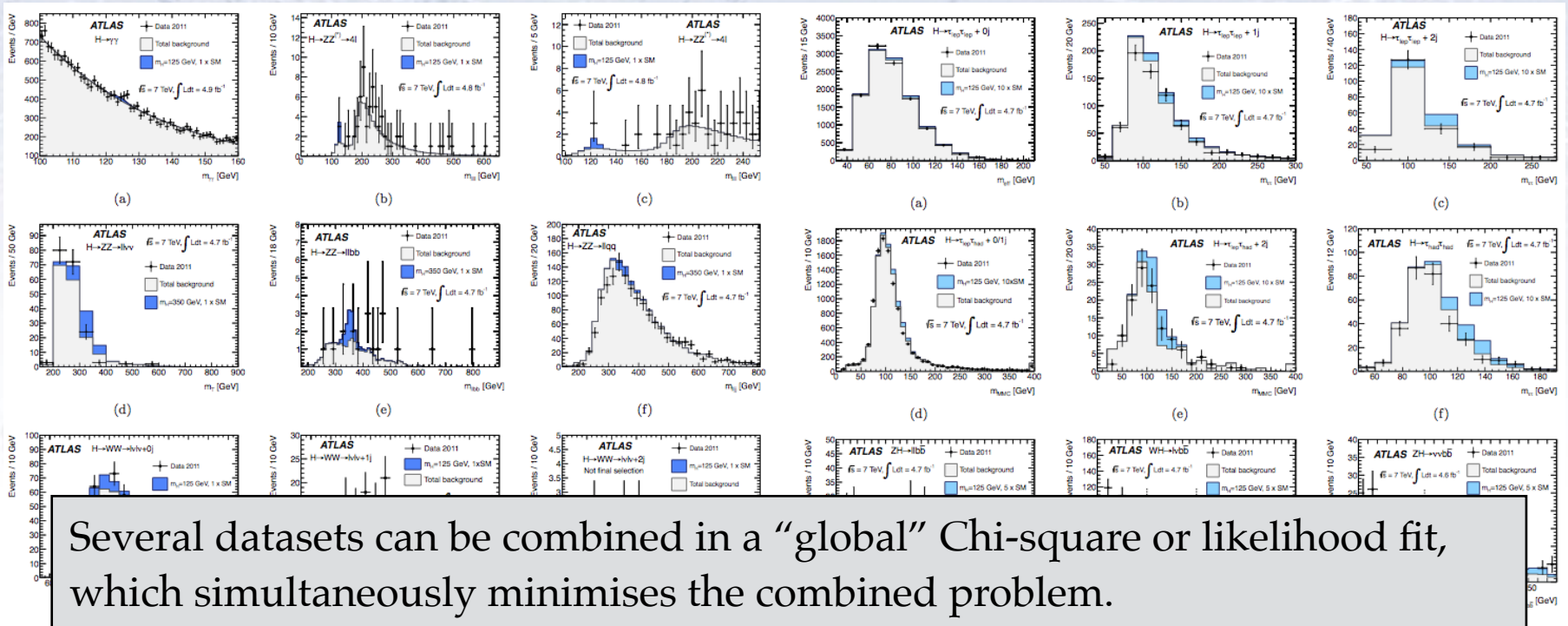
In 2005 the BaBar collaborations fit for mixing in the D0-system included **120 floating parameters** applied to six datasets.

But that was nothing compared to...

The ATLAS Higgs discovery fit



The ATLAS Higgs discovery fit



Fitting tips & tricks

There are a few tips & tricks that will make your (fitting) life a bit easier:

- **Always give good initial values!!!**
- Never start with an advanced fit - make a simple one work and expand!
- Try to make your parameters as little correlated as possible.
- Let the parameters represent the quantities of interest.
- Start with a ChiSquare fit, as these usually has better convergence.

When a fit refuses to work, try the following:

- Draw function on top of data to check formula and quality of initial values.
- Check the correlations between the parameters.
- Try to fix one or more parameter to a value you find reasonable.

Even with all of this advice, there is no guaranty that your fit will work.

It is after all a bit of an art....

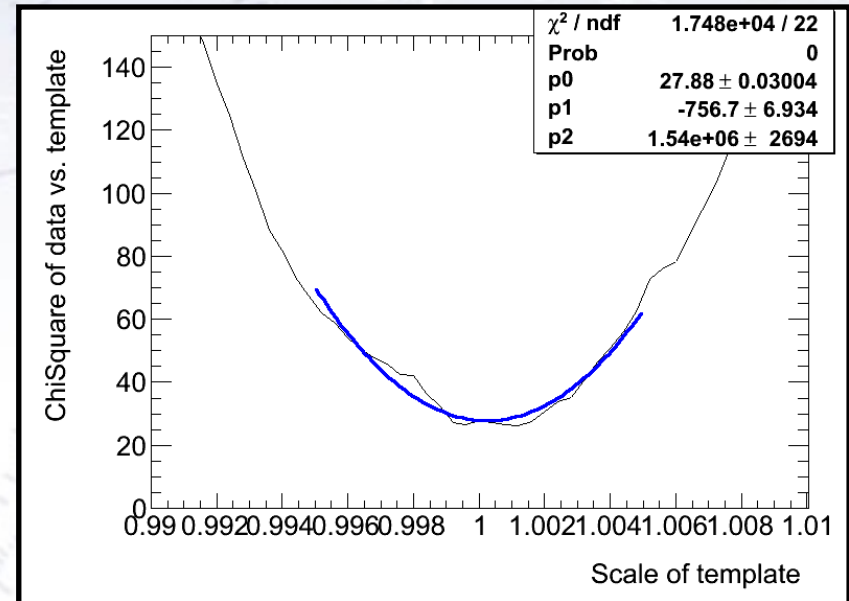
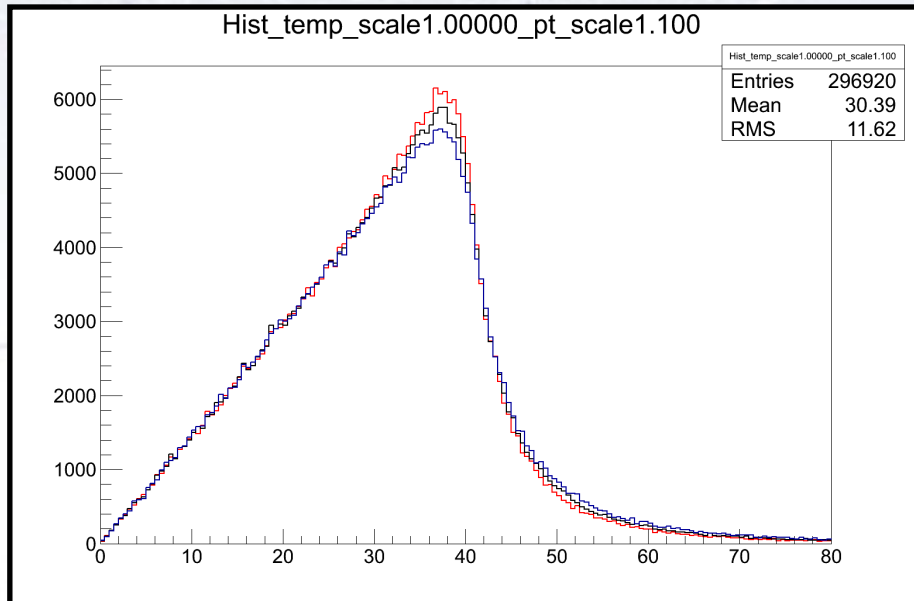


Fitting equations/constraints

Examples...

Fitting with templates

Sometimes, the shape to be fitted can not be expressed as a function, but obtained through a histogram from simulation/ data.



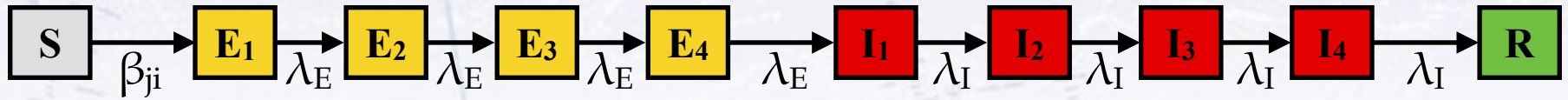
For each template, one calculates the ChiSquare between the data and the template. You then repeat it for all templates, and subsequently obtain a parabola with a minimum (central value) and a curvature (uncertainty).

Fitting with a model

Occasionally, the model is not a function, but a more advanced model. One example could be the SEIR-model for epidemics, here with 4 E and I stages:

Exposed phase:

Infected phase:



Fitting with a model

Occasionally, the model is not a function, but a more advanced model. One example could be the SEIR-model for epidemics, here with 4 E and I stages:



```
##### #
# SEIR model with time variation in beta: #
##### #

# SEIR model, including modelling of time delays and varying beta:
def func_SEIRmodel(x, dayLockdown, nI0, beta0, beta1, lambdaE, lambdaI) :

    # Initial numbers:
    N_tot = 5800000
    S = N_tot - nI0

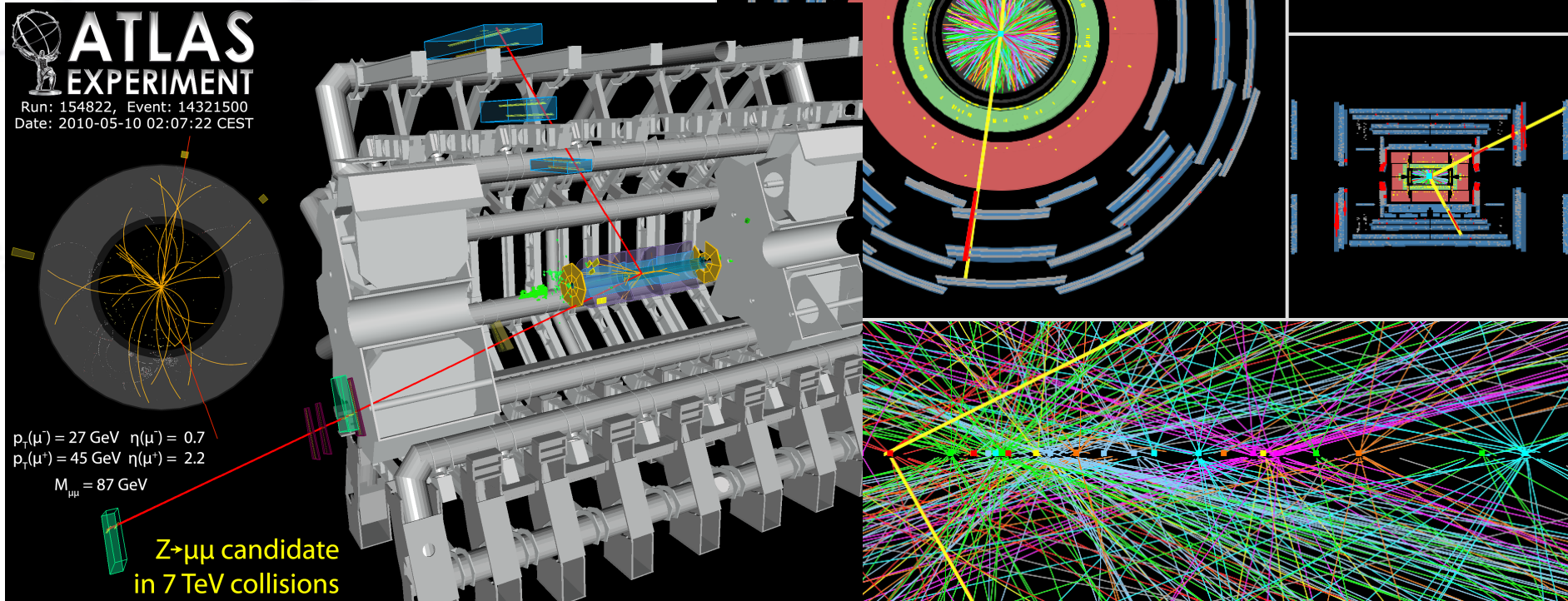
    # The initial number of exposed and infected are scaled to match beta0. Factors in front are ad hoc!
    Norm = np.exp(0.8*lambdaE * beta0) + np.exp(0.7*lambdaE * beta0) + np.exp(0.6*lambdaE * beta0) + np.exp(0.5*lambdaE * beta0) + \
           np.exp(0.4*lambdaI * beta0) + np.exp(0.3*lambdaI * beta0) + np.exp(0.2*lambdaI * beta0) + np.exp(0.1*lambdaI * beta0)
    E1 = nI0 * np.exp(0.8*lambdaE * beta0) / Norm
    E2 = nI0 * np.exp(0.7*lambdaE * beta0) / Norm
    E3 = nI0 * np.exp(0.6*lambdaE * beta0) / Norm
    E4 = nI0 * np.exp(0.5*lambdaE * beta0) / Norm
    I1 = nI0 * np.exp(0.4*lambdaI * beta0) / Norm
    I2 = nI0 * np.exp(0.3*lambdaI * beta0) / Norm
    I3 = nI0 * np.exp(0.2*lambdaI * beta0) / Norm
    I4 = nI0 * np.exp(0.1*lambdaI * beta0) / Norm
    #####
    E1 = nI0/8
    E2 = nI0/8
    E3 = nI0/8
    E4 = nI0/8
    I1 = nI0/8
    I2 = nI0/8
    I3 = nI0/8
    I4 = nI0/8
    #####
    R = 0
    Tot = S + E1+E2+E3+E4 + I1+I2+I3+I4 + R
```

So this is a model,
not a “classic” function!

Setting the momentum scale

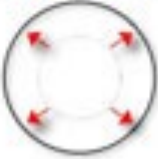




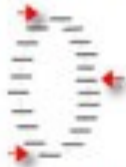


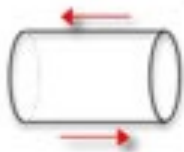
A well known type of event in particle physics (at the LHC accelerator) is the $Z \rightarrow \mu^+ \mu^-$ decay, which has two muons flying out through the detector.

As the Z-mass is well known, these events can be used to check and correct the reconstruction.



Setting the momentum scale

There are many ways in which this reconstruction can be biased due to the detector not being aligned correctly.

	ΔR	$\Delta\phi$	ΔZ
R	Radial Expansion (distance scale) 	Curl (Charge asymmetry) 	Telescope (COM boost) 
ϕ	Elliptical (vertex mass) 	Clamshell (vertex displacement) 	Skew (COM energy) 
Z	Bowing (COM energy) 	Twist (CP violation) 	Z expansion (distance scale) 

Charge dependent fit

The approach follows ATLAS-CONF-2012-141:

There is a charge independent (radial):

$$p \rightarrow p (1 + \delta_{\text{radial}})$$

The charge and p_T dependent (sagitta):

$$q/p \rightarrow q/p (1 + qp_T \delta_{\text{sagitta}})$$

The improvement is that the charge dependent should take the momentum into account, as the effect changes with momentum.

I chose the eta binning of the fit to have 24 or 51 bins.

The fit thus provides 24/51 values for δ_r and 24/51 for δ_s .

Introduction to idea

In order to use all $Z \rightarrow \ell\ell$ events, the idea is to:

- Divide leptons into bins in eta, pt, phi, charge.
Bin definition should match variables we are interested in [here: 51 in eta, 2 in charge]
- For lepton pair of bins ij , plot Z mass for data and MC (i.e. N^2 plots).
The N^2 plots limits number of bins to about 100. Not all ij -values are filled, which is OK.
- For each ij data-MC pair, determine value α_{ij} :

$$\alpha_{ij} = m_Z^{MC} / m_Z^{data}$$

The problem is: “from which lepton i or j ” does the bias come from?

- Non-unit values are from lepton bins i and j . Obtain them, by minimizing:

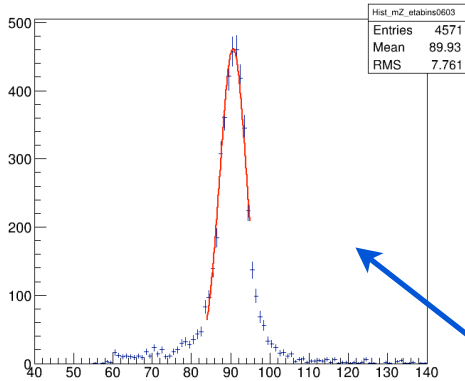
$$\chi^2(\beta_i) = \sum_{ij} \left(\frac{\alpha_{ij} - \sqrt{\beta_i \beta_j}}{\sigma(\alpha_{ij})} \right)^2$$

From this we obtain the momentum/energy scales β for each bin i .
There are N^2 α_{ij} values (perhaps a bit less) to constrain the N β values.

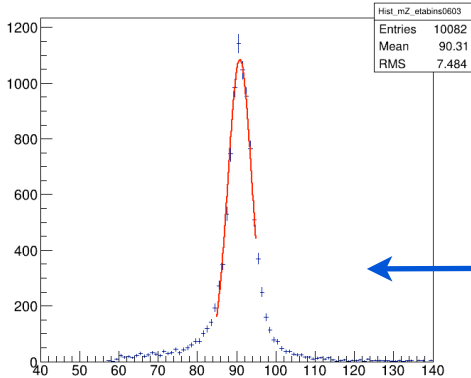
Outline of the method

For each bin in $\eta+\eta-$,
fit the mass peak:

Data: $m_Z = 89.93 \pm 0.13$ GeV

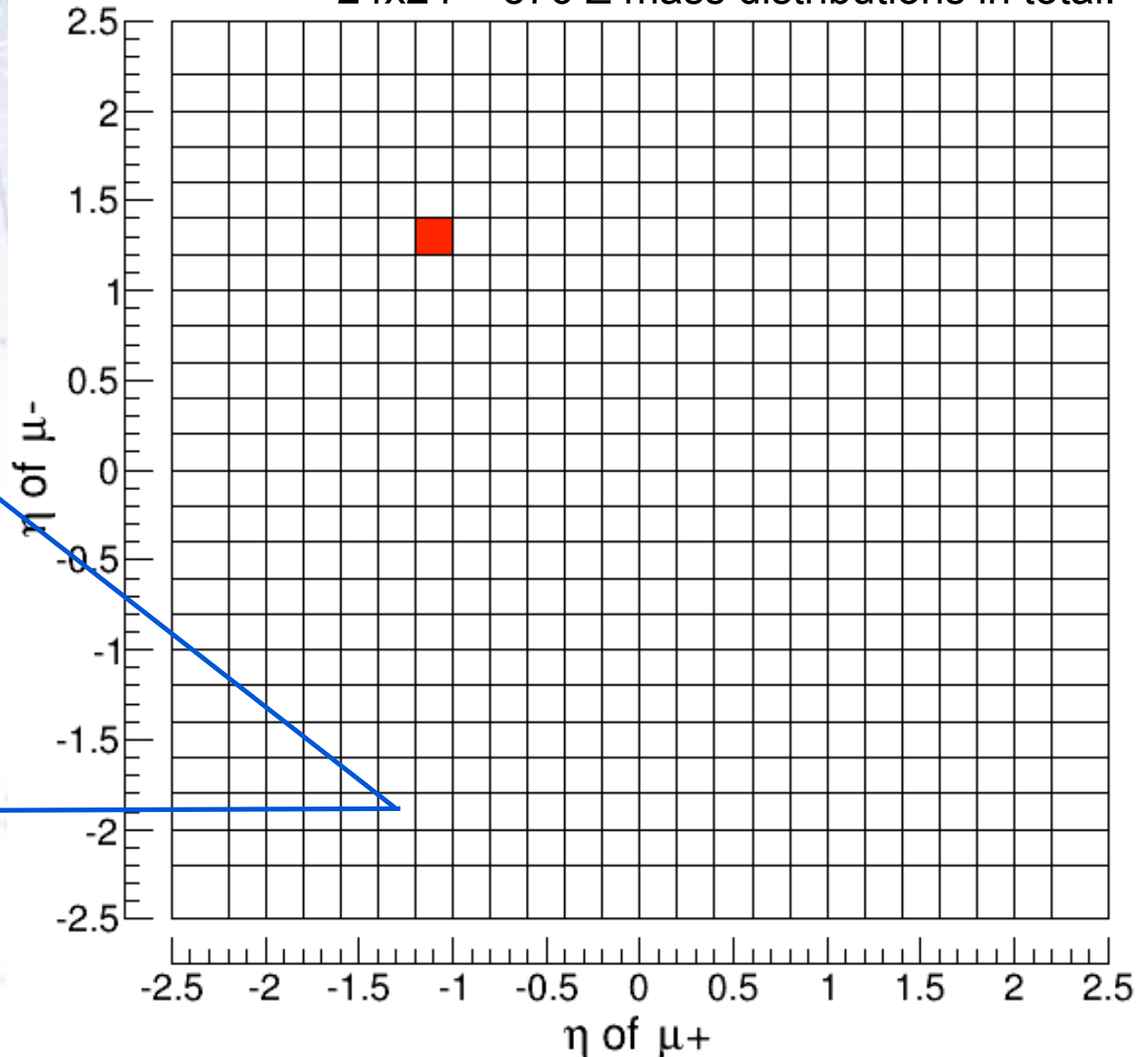


MC: $m_Z = 90.31 \pm 0.09$ GeV



Ratio: $= 0.9971 \pm 0.0009$ GeV

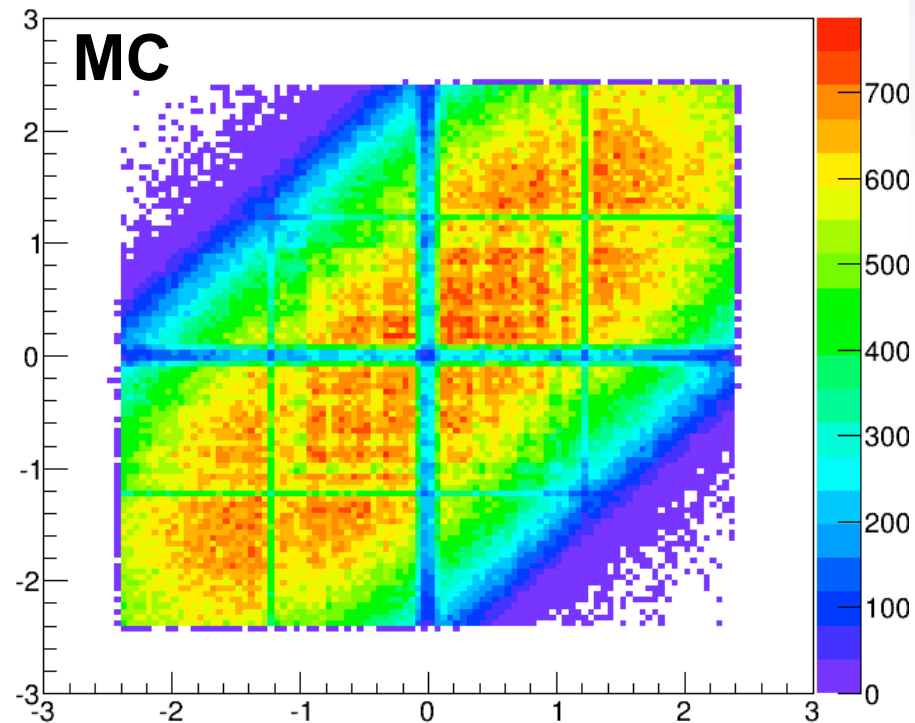
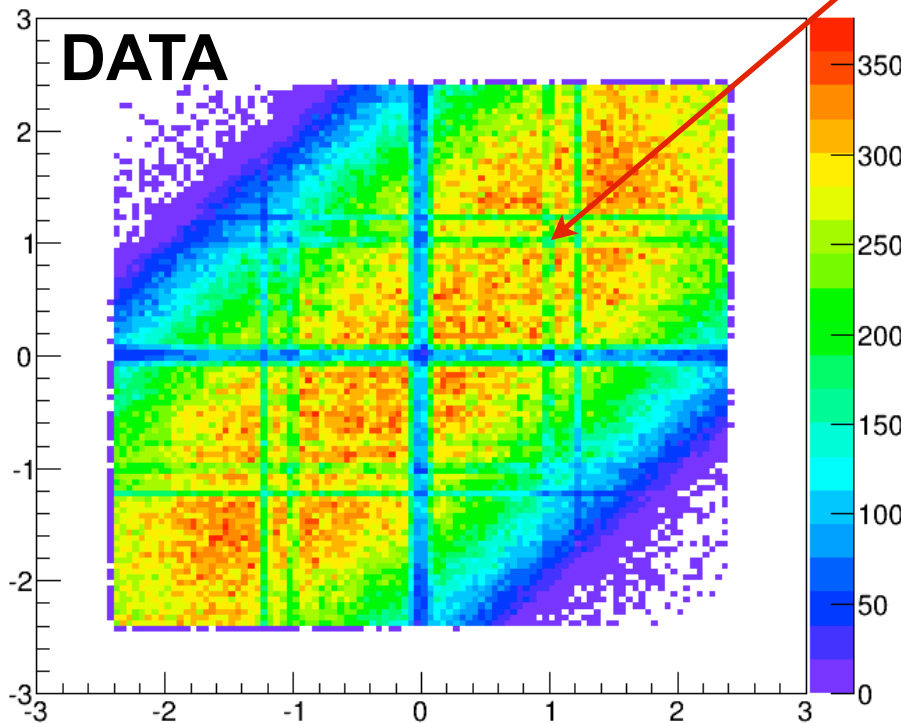
24x24 = 576 Z mass distributions in total!



Inspection of eta distribution

Data: 1.64M events, MC: 4.7M events

At first sight, they seem pretty consistent. However, a closer look reveals differences. What is causing the difference at $|\eta| = 0.95$?



504 out of the 576 $\eta+\eta^-$ has enough data ($N > 25$) and a good fit in both data and MC. These values are then used in the further fit.

Charge dependent fit

The 504 ratios of Z masses
looks like this:

0	0	0.9936	0.0028
0	1	0.9980	0.0023
0	2	0.9974	0.0021
0	3	0.9940	0.0017
0	4	0.9946	0.0017
0	5	0.9931	0.0017
0	6	0.9957	0.0018
0	7	0.9977	0.0017
0	8	0.9980	0.0016

$$p \rightarrow p(1 + \delta_{\text{radial}})$$

$$q/p \rightarrow q/p(1 + qp_T\delta_{\text{sagitta}})$$

These values are used in a χ^2 as follows:

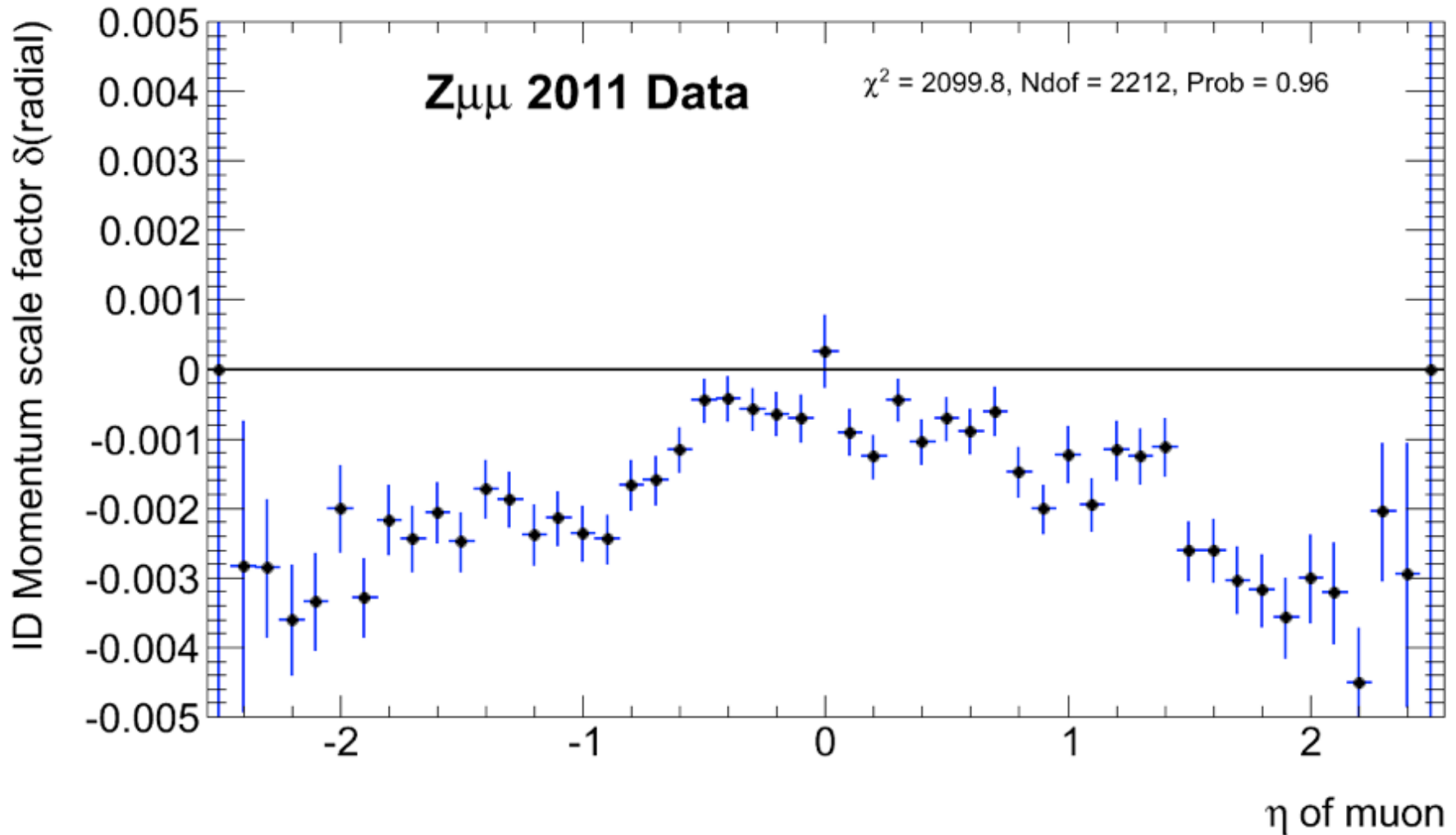
$$\chi^2 = \sum_{ij} \left(\frac{R(m_Z)_{ij} - \sqrt{(1 + \delta r_i)(1 + p_T \delta s_i)} \times (1 + \delta r_j)(1 - p_T \delta s_j)}{\sigma(R(m_Z)_{ij})} \right)^2$$

1	0	1.0006	0.0025
1	1	0.9979	0.0019
1	2	0.9963	0.0016
1	3	0.9968	0.0014
1	4	0.9960	0.0013
1	5	0.9969	0.0014
1	6	0.9971	0.0013
1	7	0.9962	0.0013
1	8	0.9959	0.0012
1	9	0.9965	0.0012
1	10	0.9980	0.0012
1	11	0.9986	0.0015

In short, the dr and ds values should minimise the expected difference between data and MC Z mass ratio.

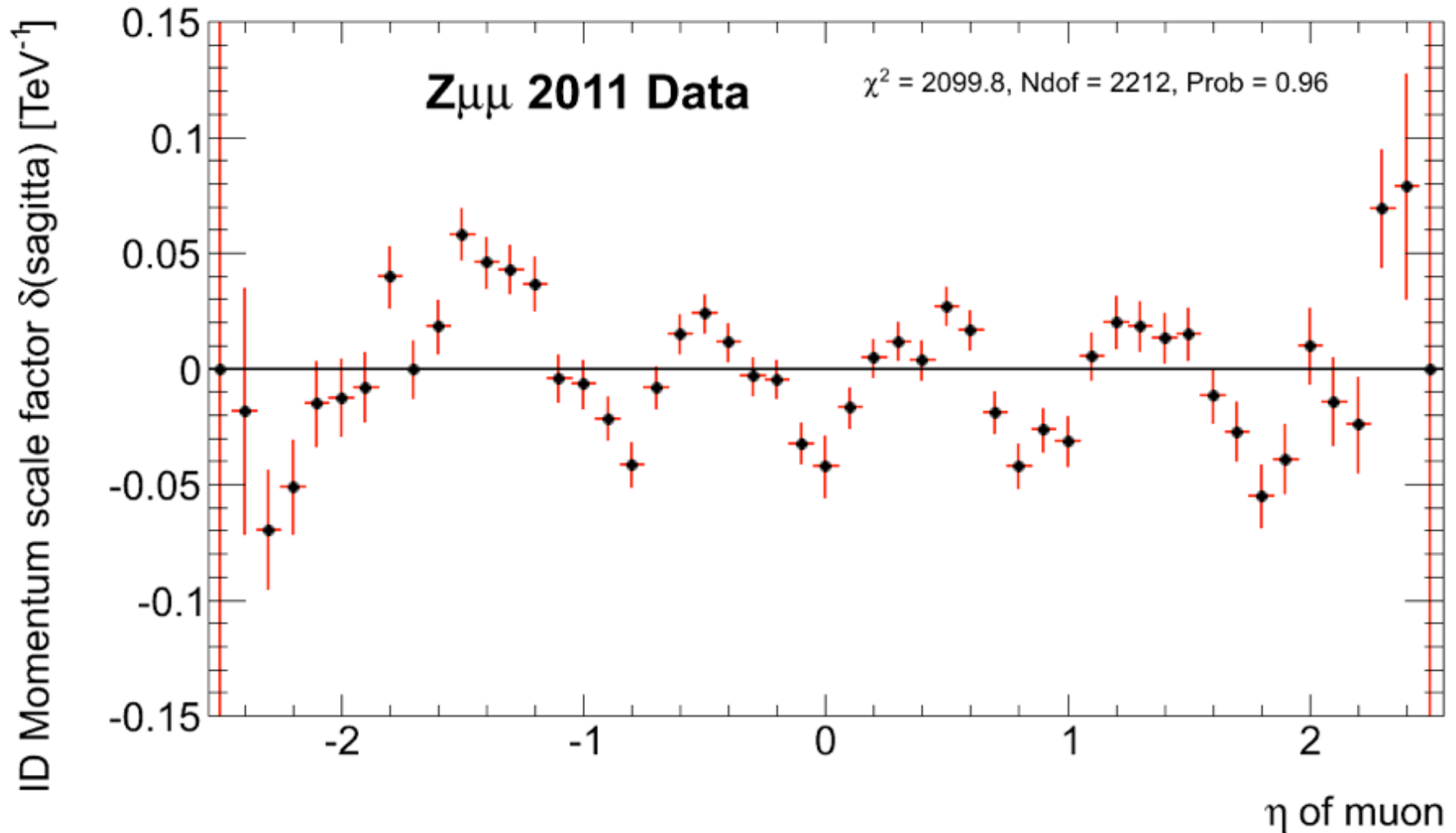
Result for Inner Detector

This is what out detector looks like for Inner Detector Muons:



Result for Inner Detector

This is what our detector looks like for Inner Detector Muons:



Corrected result

Rerunning with the correction applied to MC, I get:

