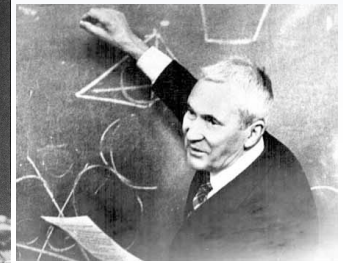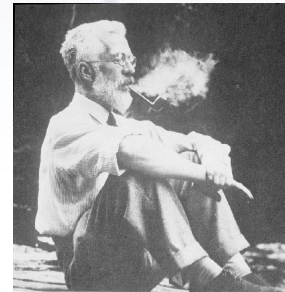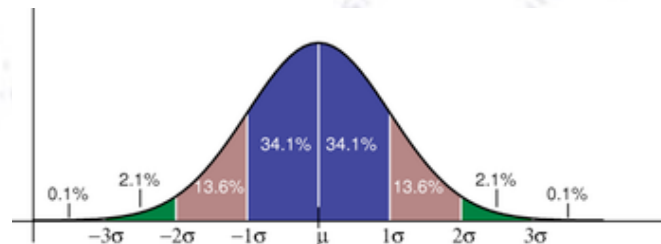# Applied Statistics
Advanced fitting

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 $\chi^2 / \mathbf{Ndof} \simeq 1$ or more precisely if $\mathbf{0.01 < p(\chi^2, Ndof) < 0.99}$, then all is good.
- If $\chi^2 / \mathbf{Ndof} \gg 1$ or more precisely if $\mathbf{p(\chi^2, Ndof) < 0.01}$, then your fit is bad, and your hypothesis is probably not correct.
- If $\chi^2 / \mathbf{Ndof} \ll 1$ or more precisely if $\mathbf{0.99 < p(\chi^2, Ndof)}$, 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 Chi-Square "landscape"
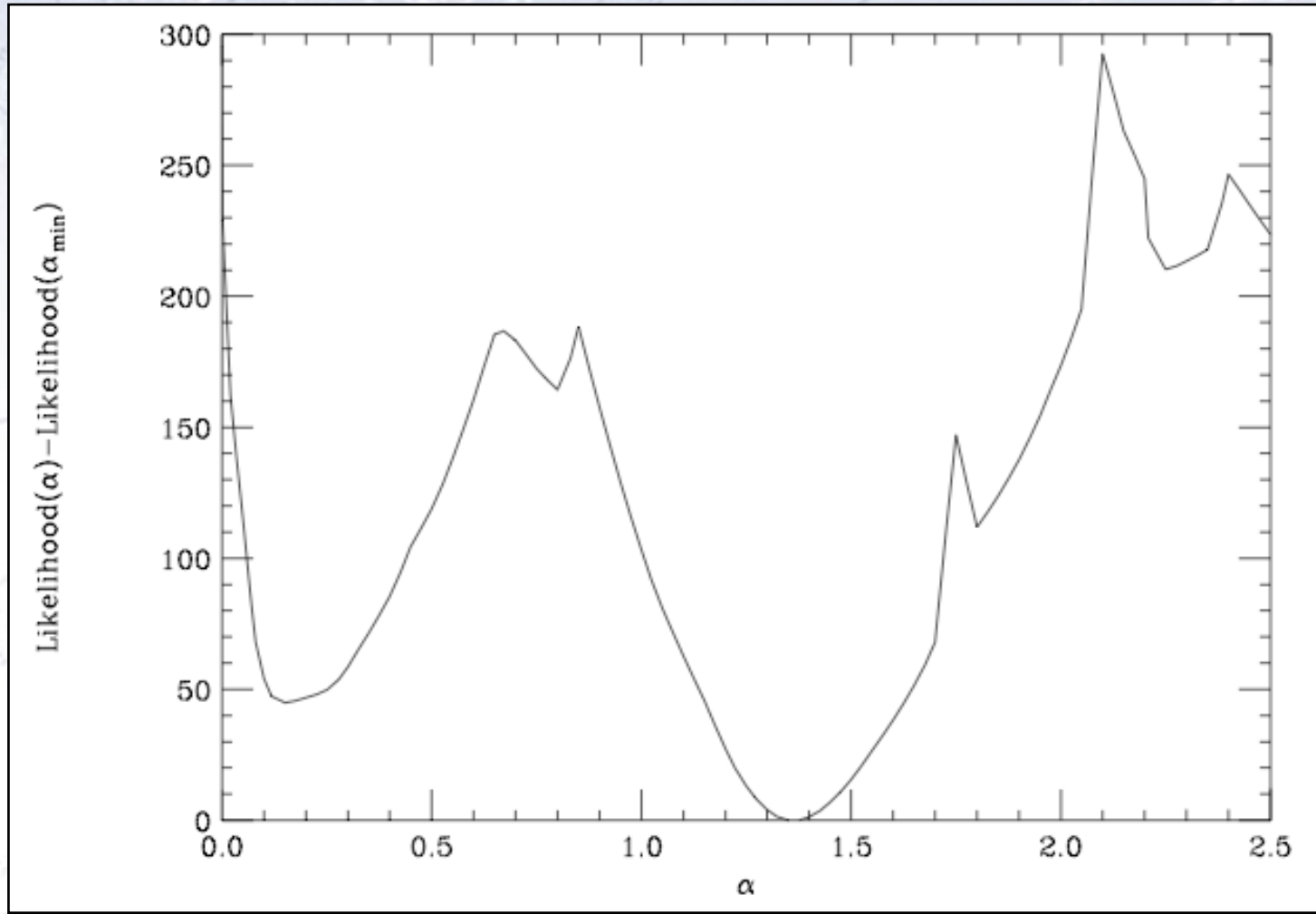
The fact that there are several minima makes fitting difficult/uncertain!
Always give good starting values!!!

4

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!

χ^2 / ndf	3868 / 3796
Prob	0.2044
p0	−0.104 ± 8.111e−05
p1	0.5928 ± 0.0002452
p2	0.007427 ± 2.032e−05
p3	9.602 ± 2.035e−05
p4	2.819 ± 0.0004148
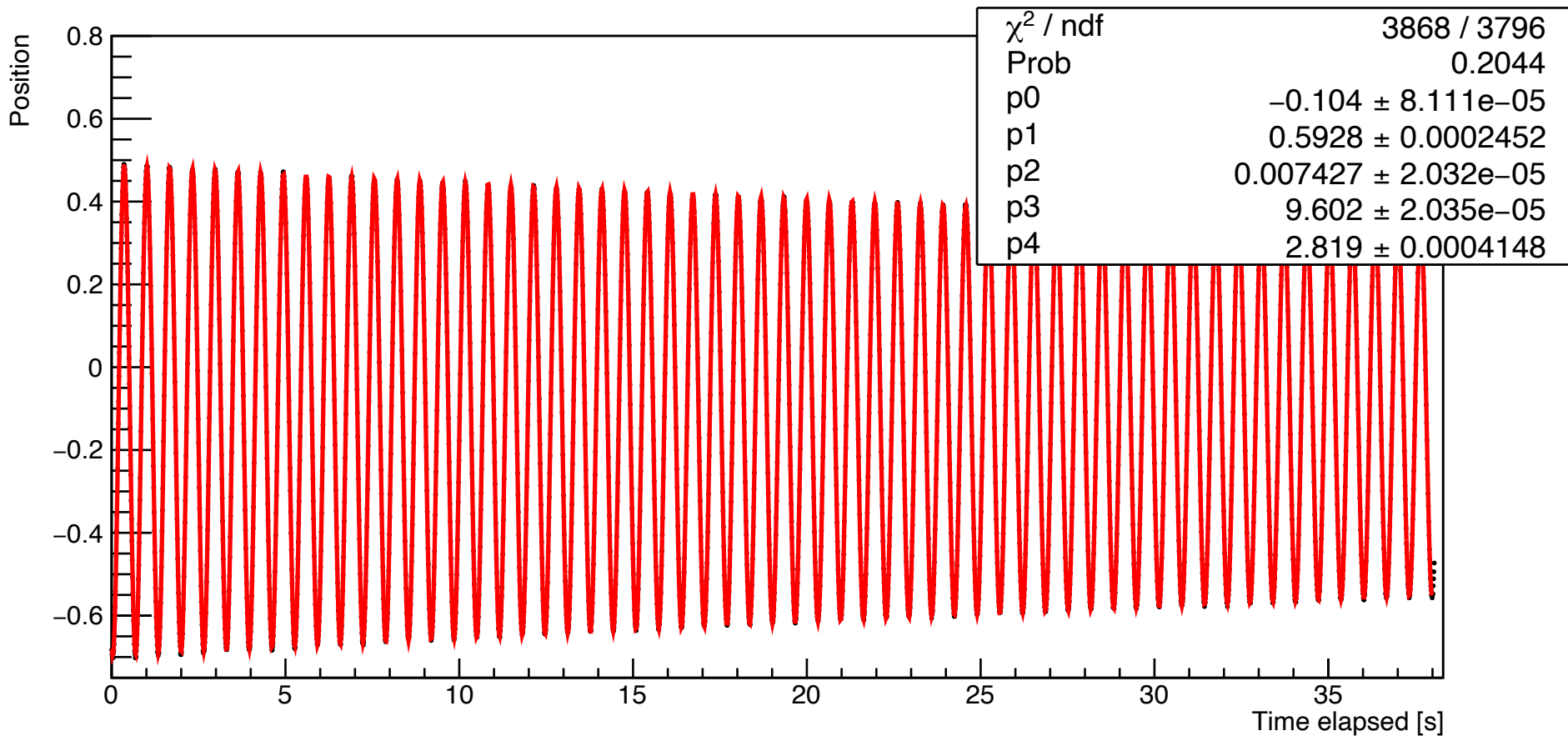
Example Chi-Square landscape
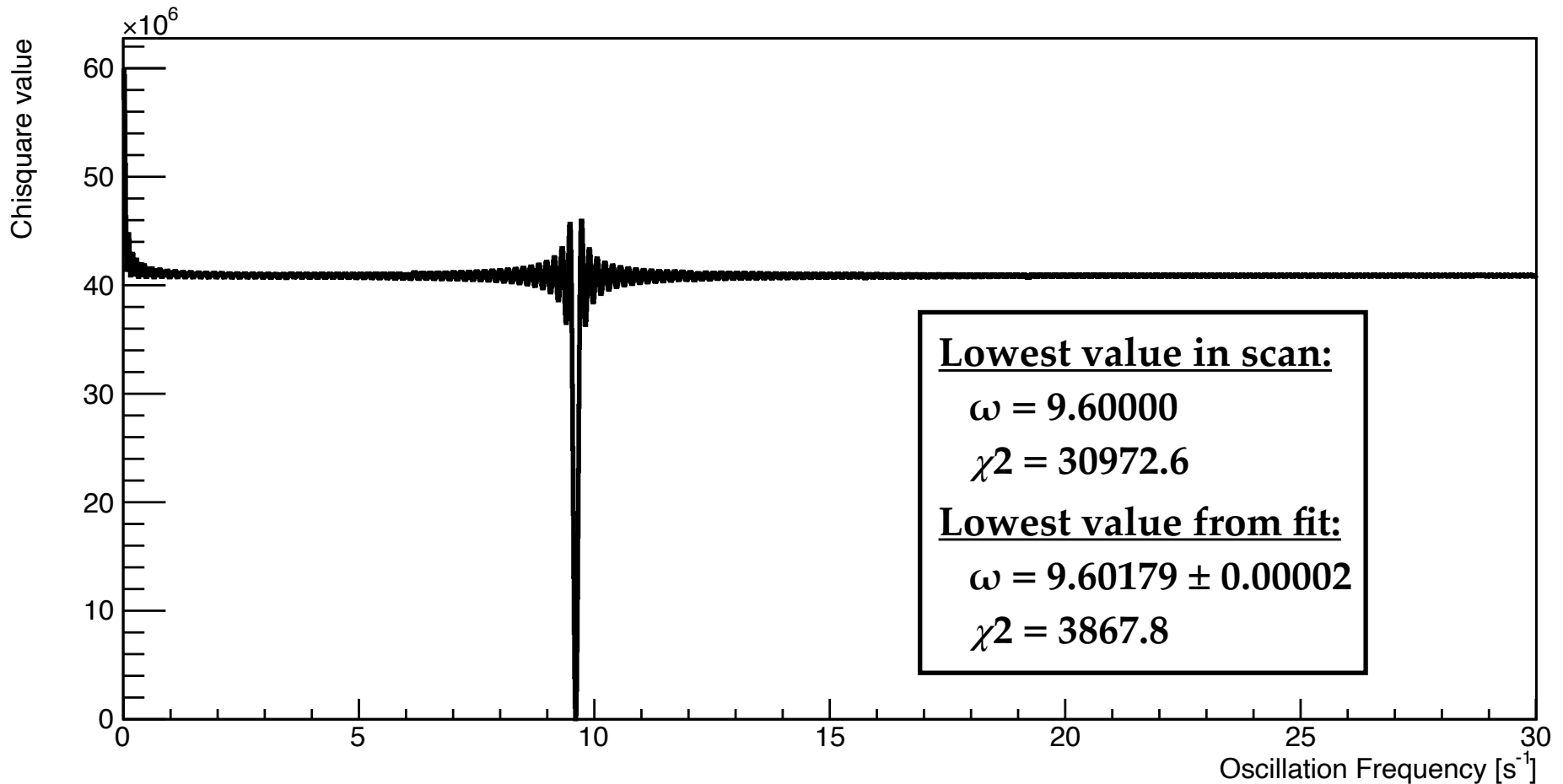
Especially fitting oscillatory data requires a good starting value for omega.
Even a small offset may result in a Chi2, which is not sensitive to omega!

Lowest value in scan:
  $\omega = 9.60000$
  $\chi 2 = 30972.6$
Lowest value from fit:
  $\omega = 9.60179 \pm 0.00002$
  $\chi 2 = 3867.8$

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:

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


$\mu = 0, \sigma = 1$
$\mu = 0, \sigma = 2$

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, sigma1 and sigma2 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.
a) Fixing one of the parameter values, i.e. f = 0.8.
b) Fixing the relation between sigma1 and sigma2, i.e. sigma2 = 3 x sigma1.
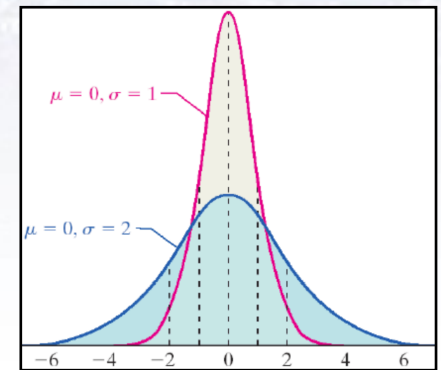
Correlations between parameters

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:

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

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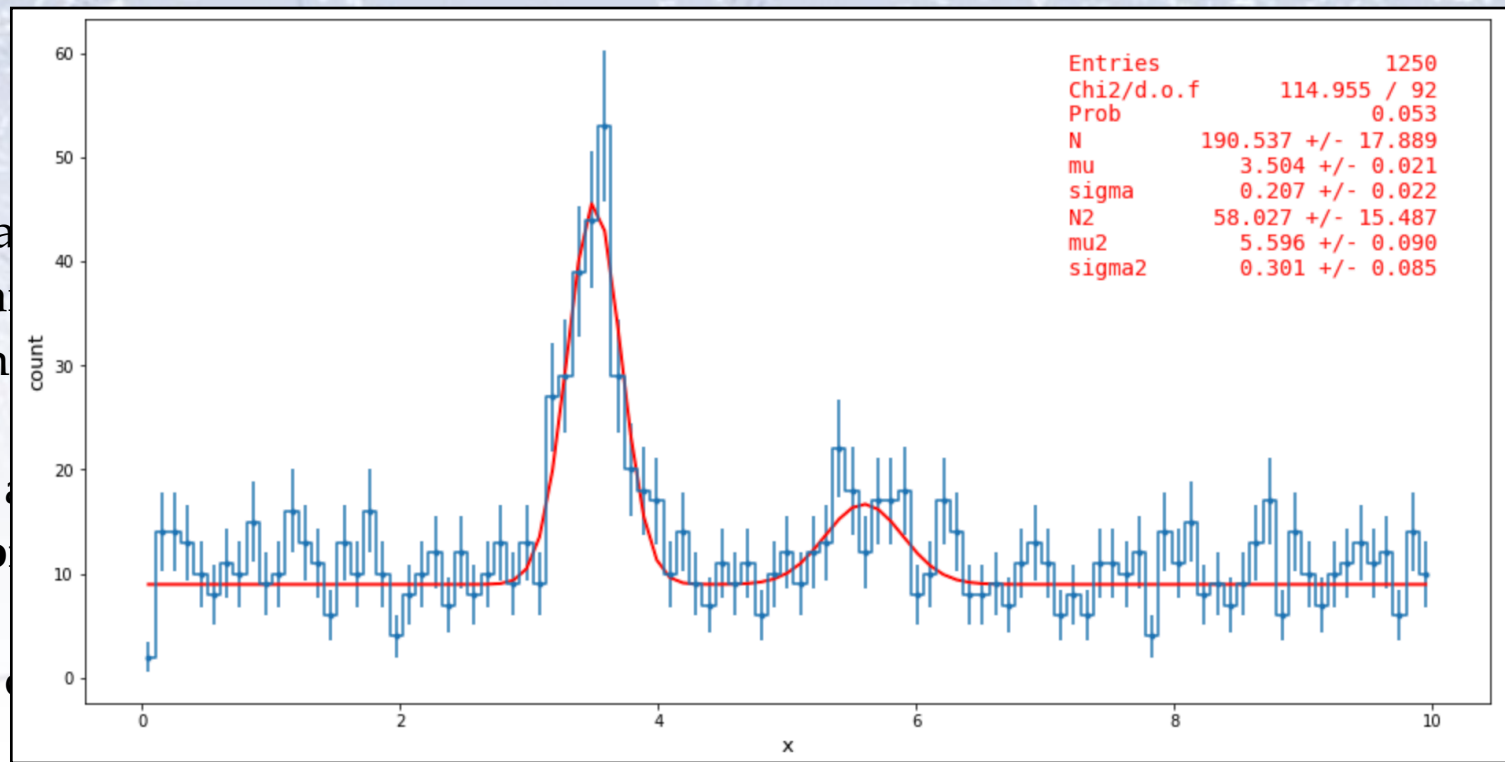
The down side is that fit gets more complicated (see advanced example later).

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 $\sigma_C = \sigma_S$.

Entries 1250
Chi2/d.o.f 114.955 / 92
Prob 0.053
N 190.537 +/- 17.889
mu 3.504 +/- 0.021
sigma 0.207 +/- 0.022
N2 58.027 +/- 15.487
mu2 5.596 +/- 0.090
sigma2 0.301 +/- 0.085

Occa
para   for
a Ch

The
auto

The   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:

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

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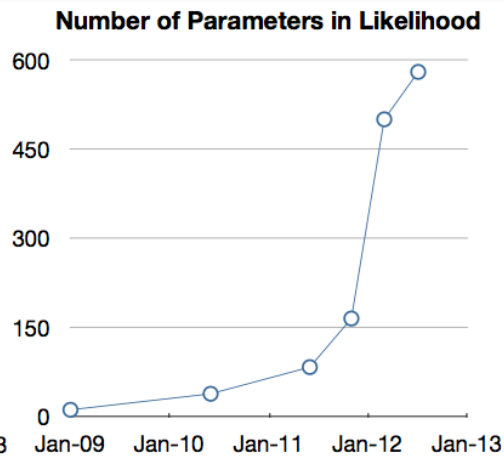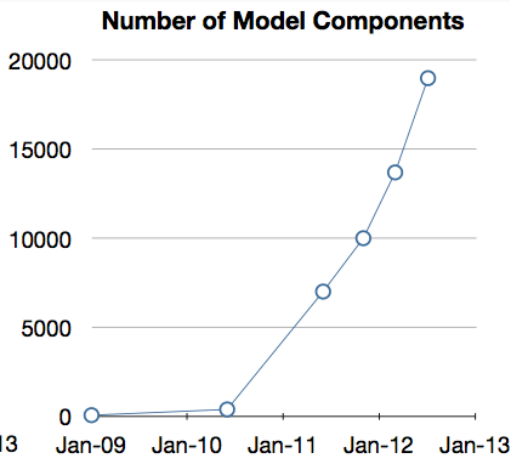
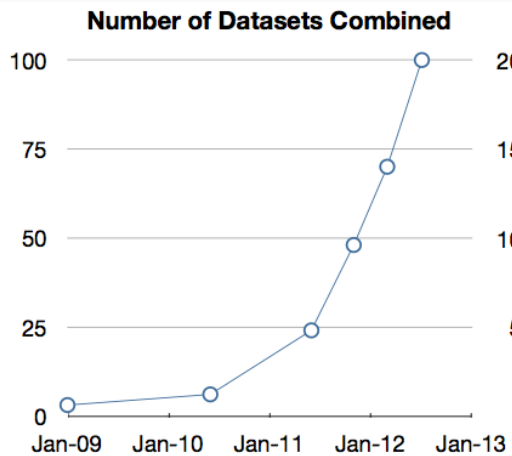However, if they are of similar size, then a simultaneous fit of the samples is the optimal solution. This can grow quite large…

In 2002 the BaBar collaborations fit for CP-violation ($\sin2\beta$) 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

18

The ATLAS Higgs discovery fit

Several datasets can be combined in a "global" Chi-square or likelihood fit, which simultaneously minimises the combined problem.

Number of Datasets Combined | Number of Model Components | Number of Parameters in Likelihood
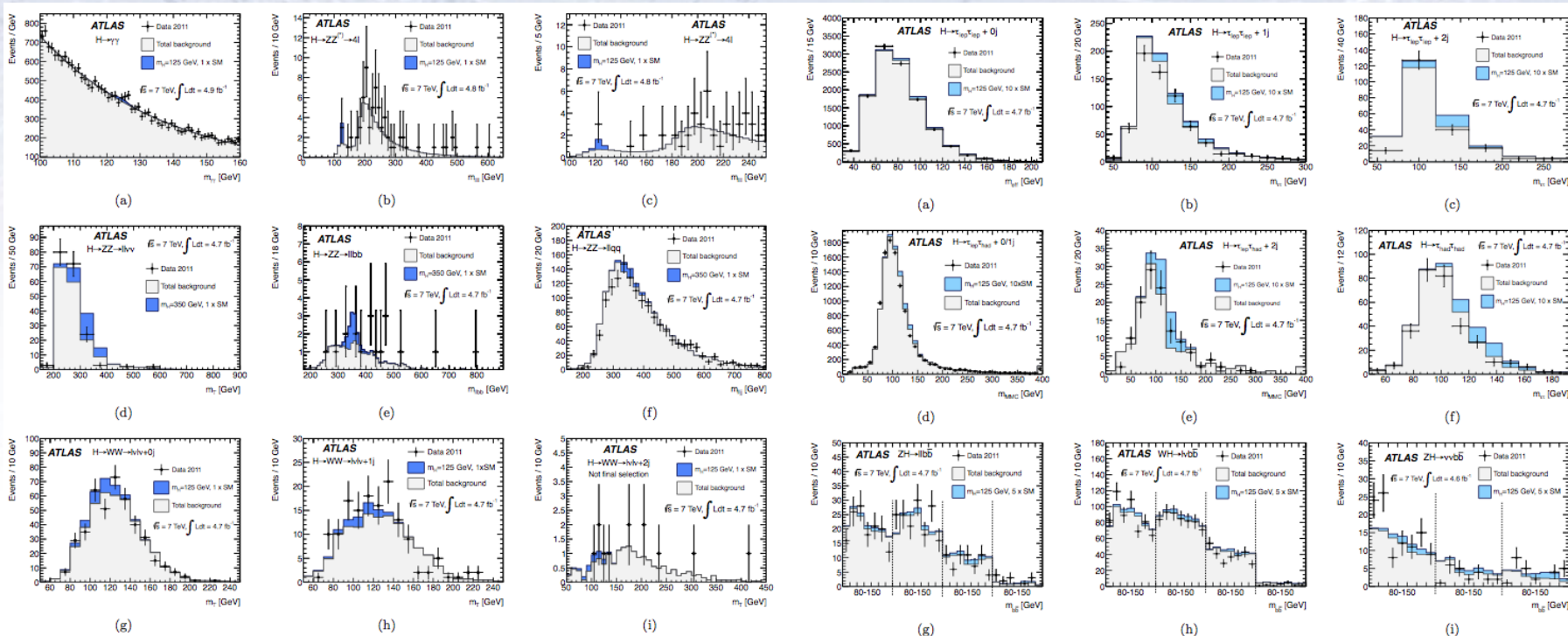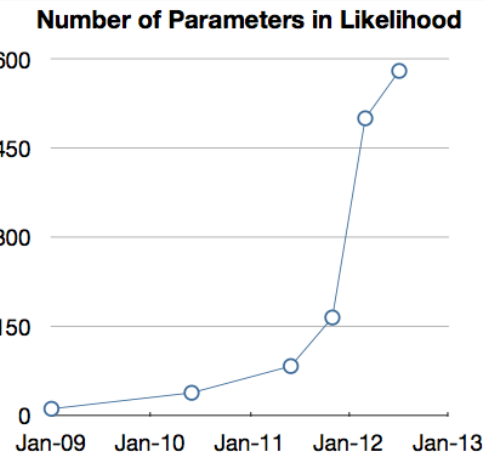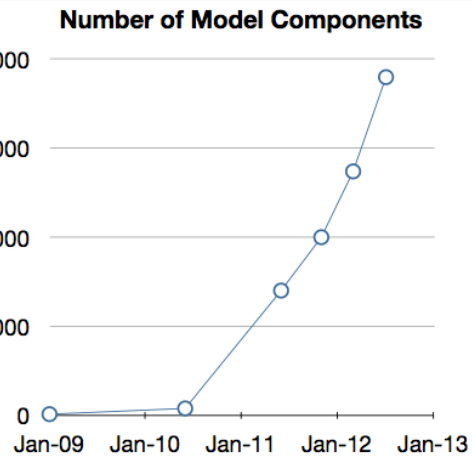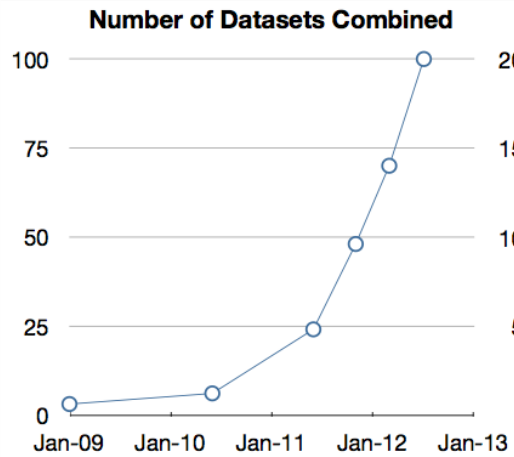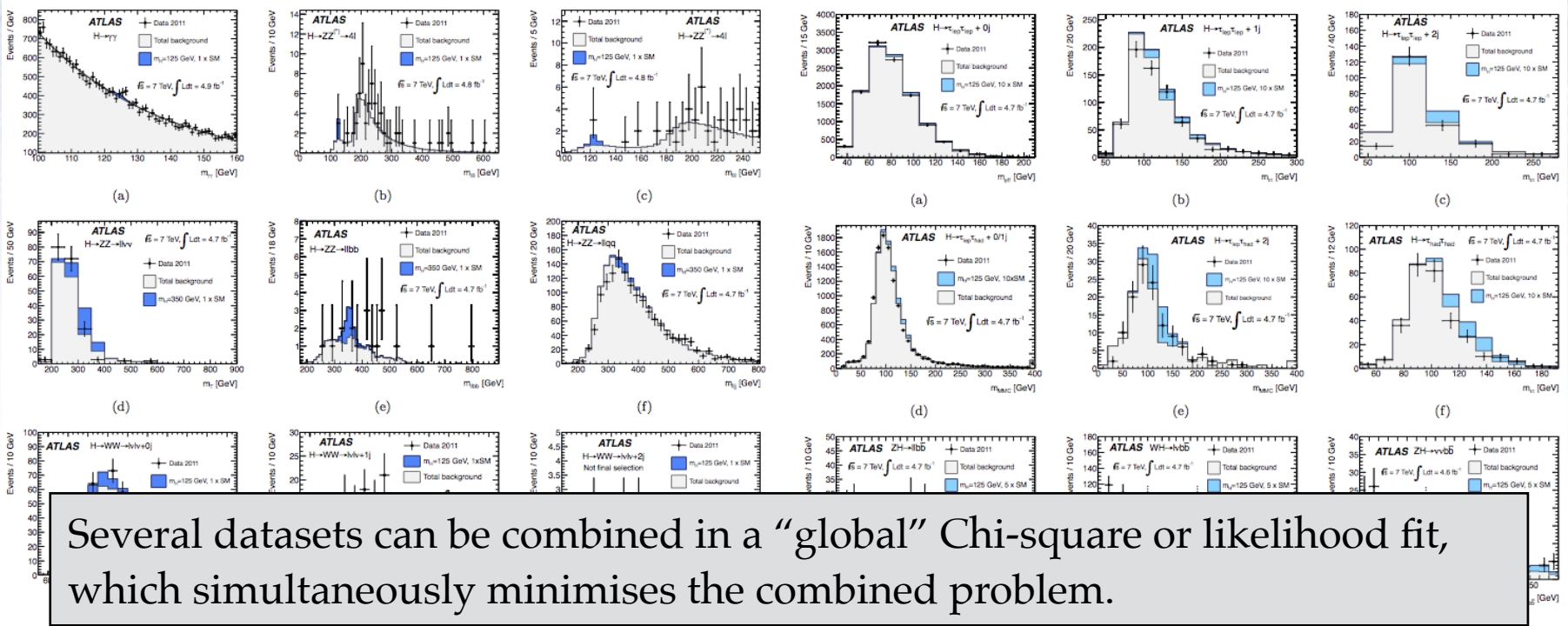
Fitting tips & tricks

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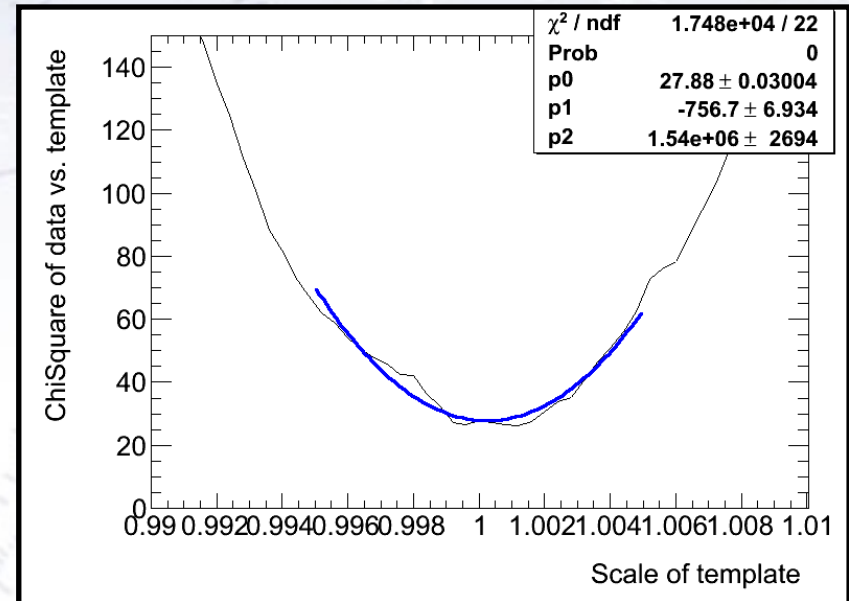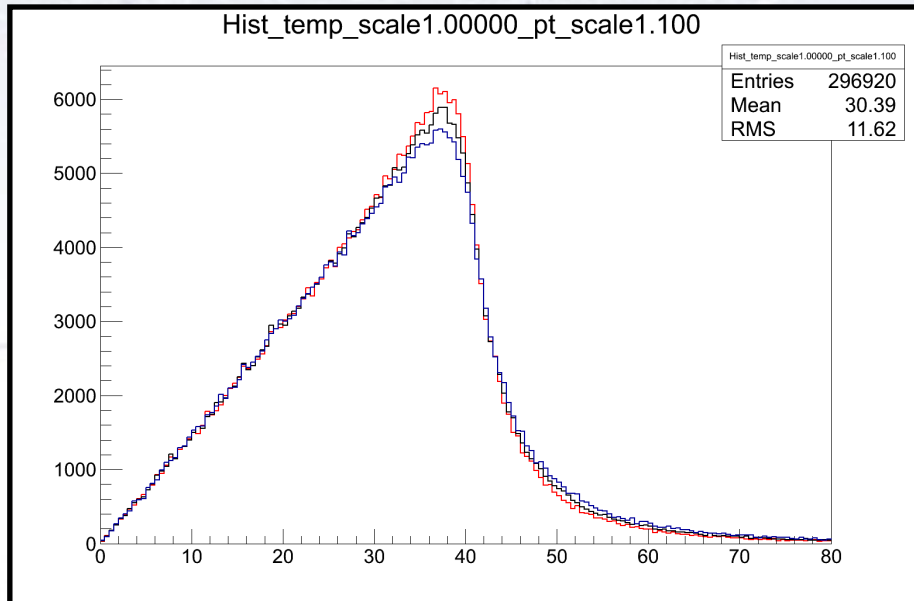
For each template, one calculates the ChiSquare between the data and the template. You then repeat it for all templates, and subsequently obtain a parabola with a minimum (central value) and a curvature (uncertainty).

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:

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:**

$$S \xrightarrow{\beta_{ji}} E_1 \xrightarrow{\lambda_E} E_2 \xrightarrow{\lambda_E} E_3 \xrightarrow{\lambda_E} E_4 \xrightarrow{\lambda_E} I_1 \xrightarrow{\lambda_I} I_2 \xrightarrow{\lambda_I} I_3 \xrightarrow{\lambda_I} I_4 \xrightarrow{\lambda_I} R$$

```python
# ---------------------------------------------------------------------- #
# 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
```
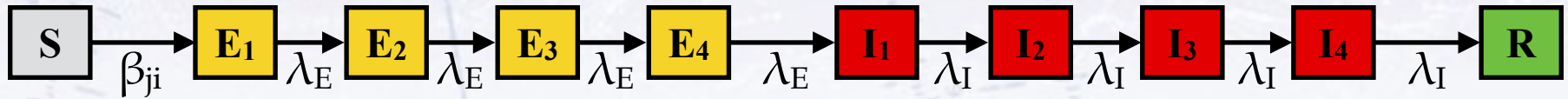
**So this is a model,
not a "classic" function!**

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:
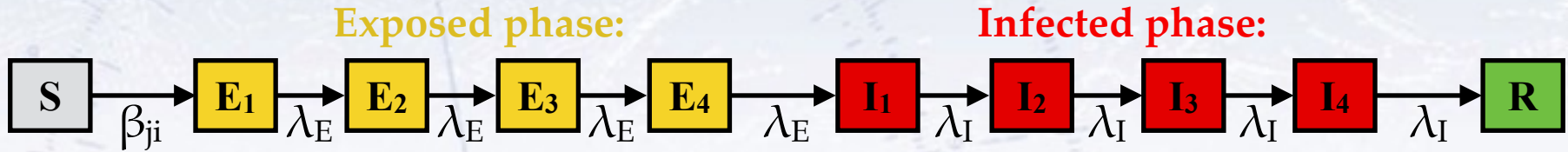
Exposed phase: **Infected phase:**

$$S \xrightarrow{\beta_{ji}} E_1 \xrightarrow{\lambda_E} E_2 \xrightarrow{\lambda_E} E_3 \xrightarrow{\lambda_E} E_4 \xrightarrow{\lambda_E} I_1 \xrightarrow{\lambda_I} I_2 \xrightarrow{\lambda_I} I_3 \xrightarrow{\lambda_I} I_4 \xrightarrow{\lambda_I} R$$

An example outbreak of Covid-19 in Denmark happened in Aarhus early August 2020:

The number of tests and positives are given, from which a "scaled" number of infected can be calculated and fitted.

The fit is exactly with the SEIR model above, fixing most parameters.

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:**

$$\boxed{S} \xrightarrow{\beta_{ji}} \boxed{E_1} \xrightarrow{\lambda_E} \boxed{E_2} \xrightarrow{\lambda_E} \boxed{E_3} \xrightarrow{\lambda_E} \boxed{E_4} \xrightarrow{\lambda_E} \boxed{I_1} \xrightarrow{\lambda_I} \boxed{I_2} \xrightarrow{\lambda_I} \boxed{I_3} \xrightarrow{\lambda_I} \boxed{I_4} \xrightarrow{\lambda_I} \boxed{R}$$

An example outbreak of Covid-19 in Denmark happened in Aarhus early August 2020:

The number of tests and positives are given, from which a "scaled" number of infected can be calculated and fitted.

The fit is exactly with the SEIR model above, fixing most parameters.

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:**

$$S \xrightarrow{\beta_{ji}} E_1 \xrightarrow{\lambda_E} E_2 \xrightarrow{\lambda_E} E_3 \xrightarrow{\lambda_E} E_4 \xrightarrow{\lambda_E} I_1 \xrightarrow{\lambda_I} I_2 \xrightarrow{\lambda_I} I_3 \xrightarrow{\lambda_I} I_4 \xrightarrow{\lambda_I} R$$

An example outbreak of Covid-19 in Denmark happened in Aarhus early August 2020:

The number of tests and positives are given, from which a "scaled" number of infected can be calculated and fitted.

The fit is exactly with the SEIR model above, fixing most parameters.

Legend:
- Fit med SEIR Model - uden tiltag
- Fit med SEIR Model - med tiltag
- Nye positive (skaleret til 3000 tests/day)

Y-axis: Nye positive pr. dag (skaleret)
X-axis: Dag (1. Marts er dag 0)

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:**

S — β_{ji} → E_1 — λ_E → E_2 — λ_E → E_3 — λ_E → E_4 — λ_E → I_1 — λ_I → I_2 — λ_I → I_3 — λ_I → I_4 — λ_I → R

An example outbreak of Covid-19 in Denmark happened in Aarhus early August 2020:

The number of tests and positives are given, from which a "scaled" number of infected can be calculated and fitted.

The fit is exactly with the SEIR model above, fixing most parameters.

Setting the momentum scale
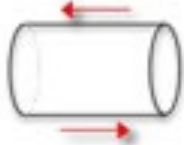
A well known type of event in particle physics (at he LHC accelerator) is the Z ➤ μ+ μ- decay, which has two muons flying out through the detector.

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

ATLAS EXPERIMENT
Run: 154822, Event: 14321500
Date: 2010-05-10 02:07:22 CEST

$p_T(\mu^-) = 27$ GeV $\eta(\mu^-) = 0.7$
$p_T(\mu^+) = 45$ GeV $\eta(\mu^+) = 2.2$
$M_{\mu\mu} = 87$ GeV

Z➤μμ candidate
in 7 TeV collisions

ATLAS EXPERIMENT
Run Number: 201289, Event Number: 24151616
Date: 2012-04-15 16:52:58 CEST
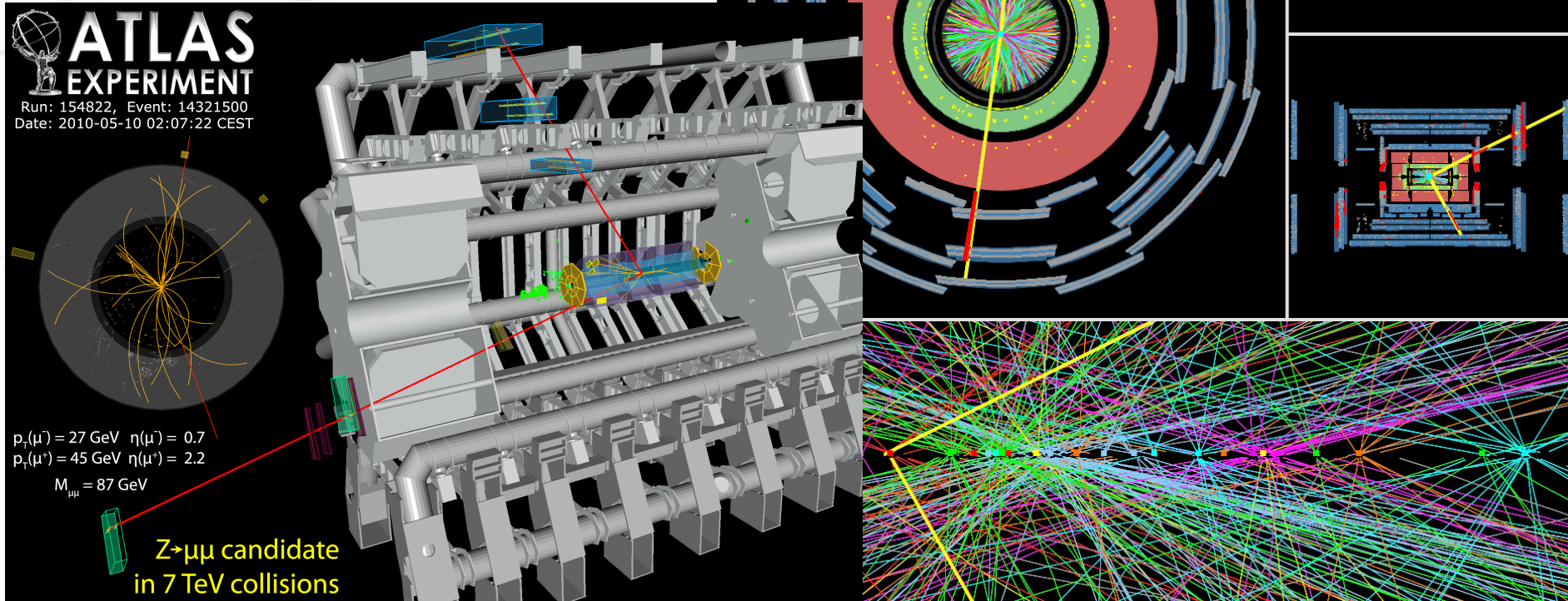
Setting the momentum scale

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

	ΔR	Δφ	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):

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

The charge and p_T dependent (sagitta):

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

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 δr and 24/51 for δs.

Introduction to idea

In order to use all Z ➤ ll 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. N² plots).

  The N² 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.

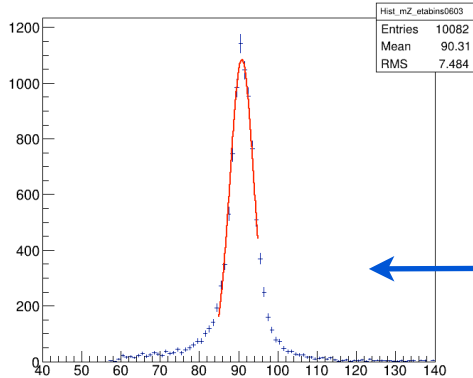There are N² α$_{ij}$ values (perhaps a bit less) to constrain the N β values.

Outline of the method

For each bin in η+η-,
fit the mass peak:

24x24 = 576 Z mass distributions in total!

Data: mZ = 89.93 ± 0.13 GeV

Hist_mZ_etabins0603	
Entries	4571
Mean	89.93
RMS	7.761

MC: mZ = 90.31 ± 0.09 GeV

Hist_mZ_etabins0603	
Entries	10082
Mean	90.31
RMS	7.484

η of μ-

η of μ+
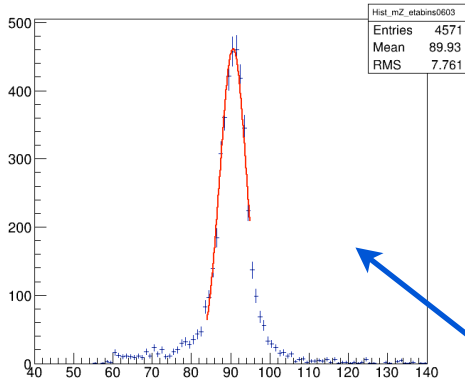
Ratio: = 0.9971 ± 0.0009 GeV
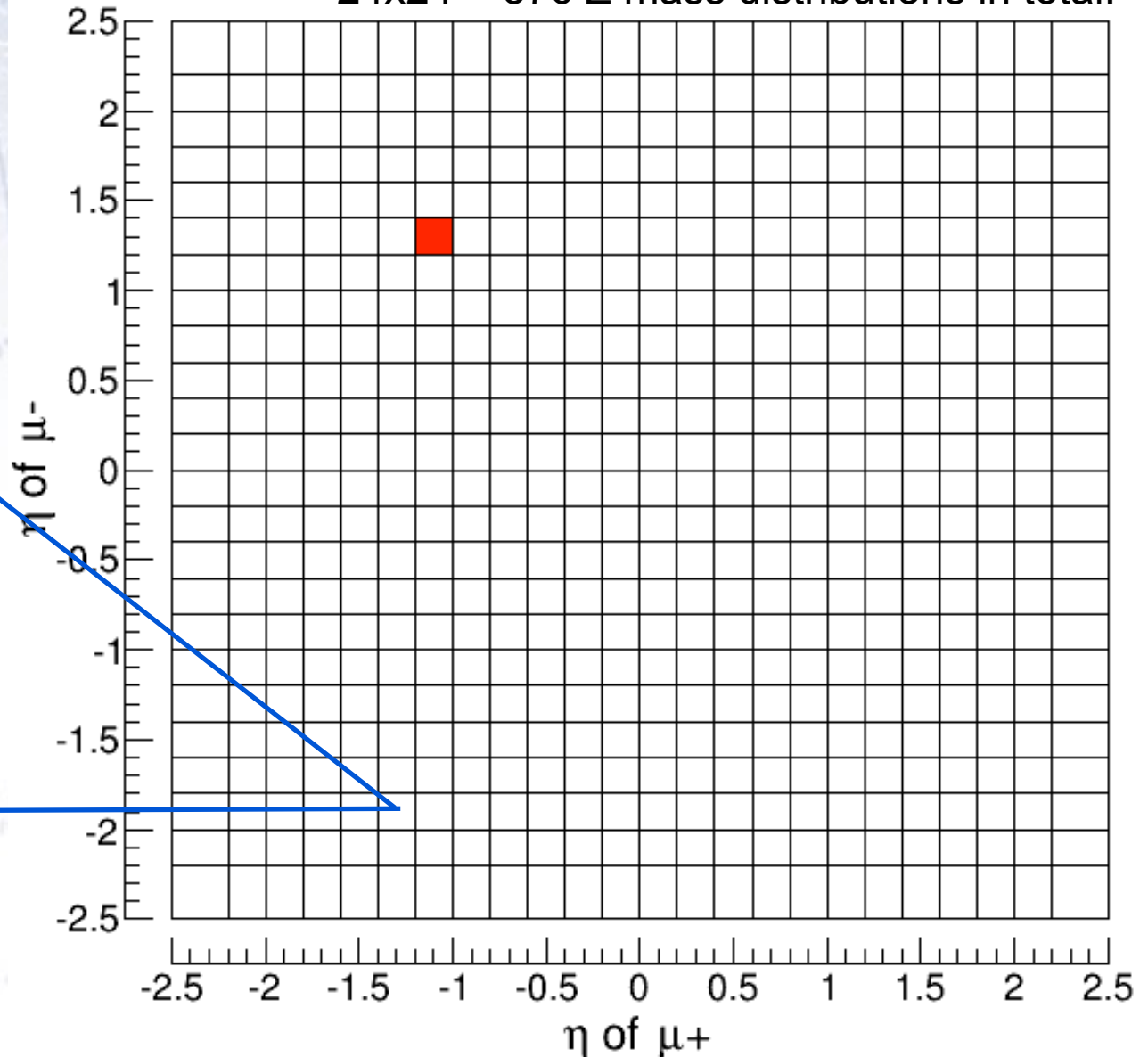
Inspection of eta distribution

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 $|\eta| = 0.95$?
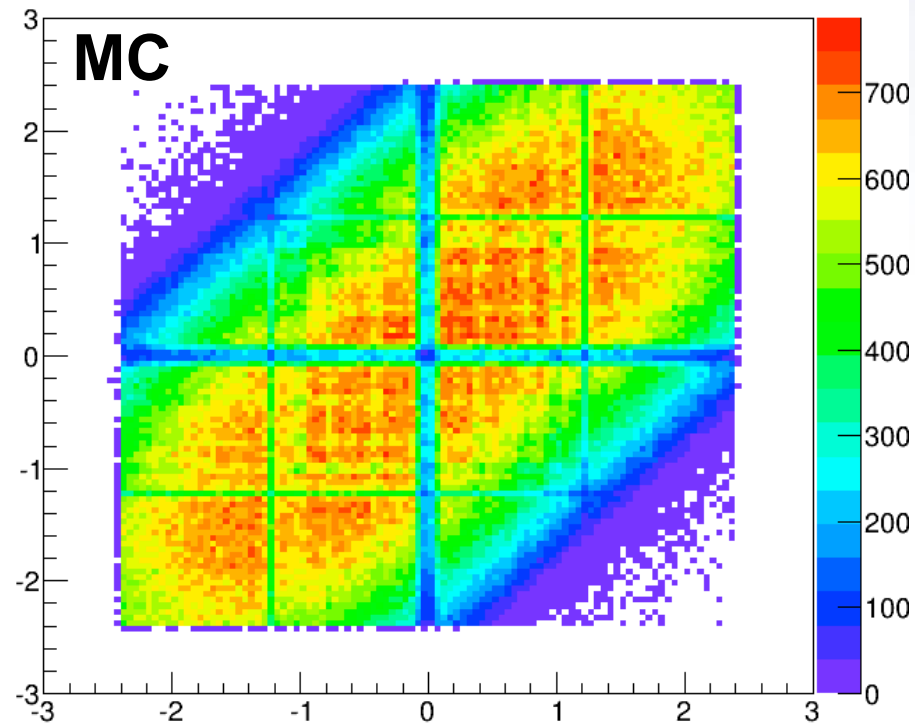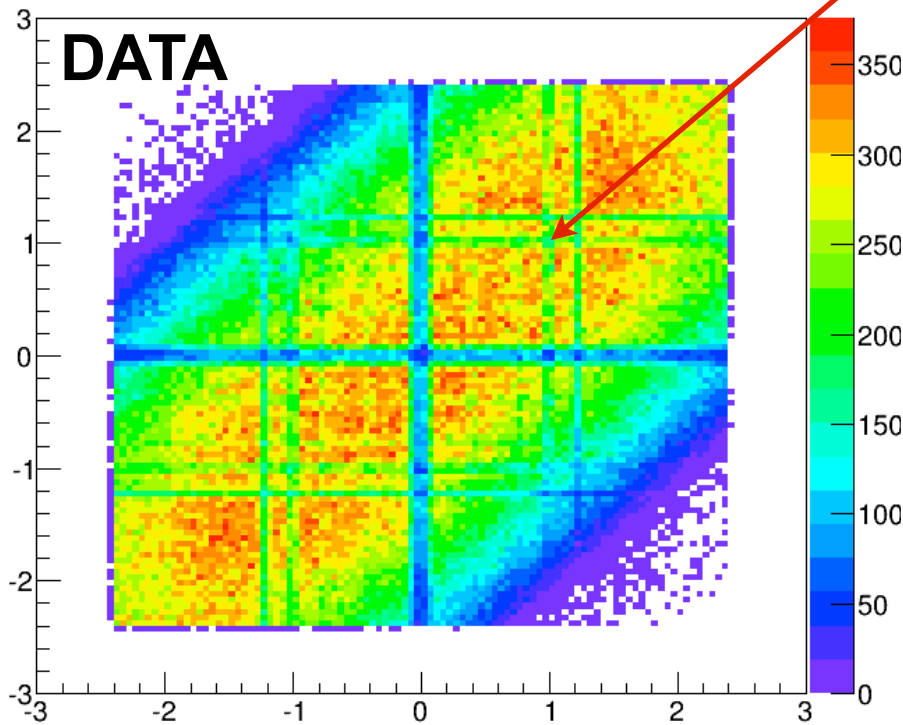
504 out of the 576 $\eta + \eta -$ has enough data ($N > 25$) and a good fit in both data and MC. These values are then used in the further fit.

Charge dependent fit

The 504 ratios of Z masses looks like this:

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

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

```
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
```

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.
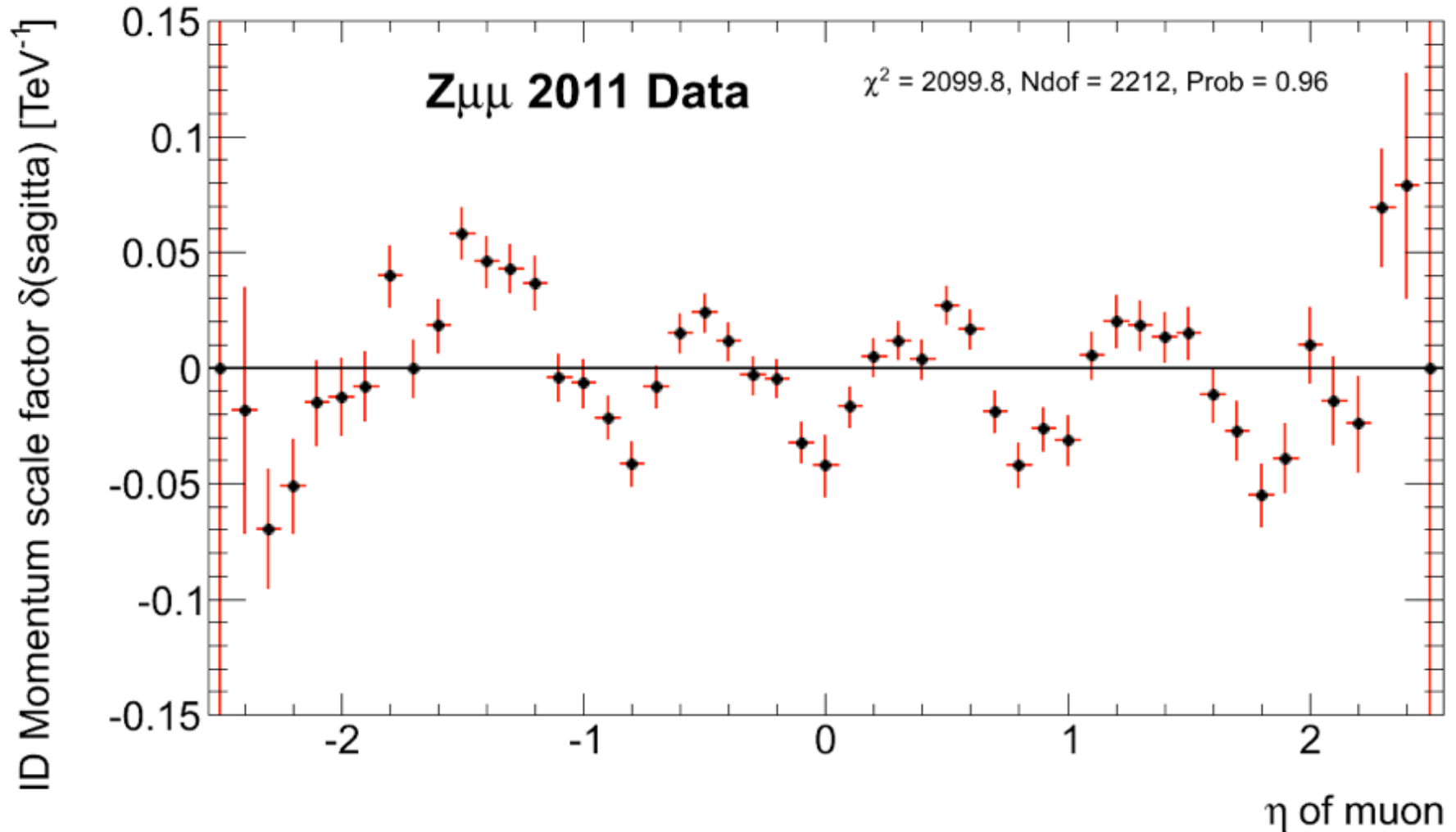
Result for Inner Detector

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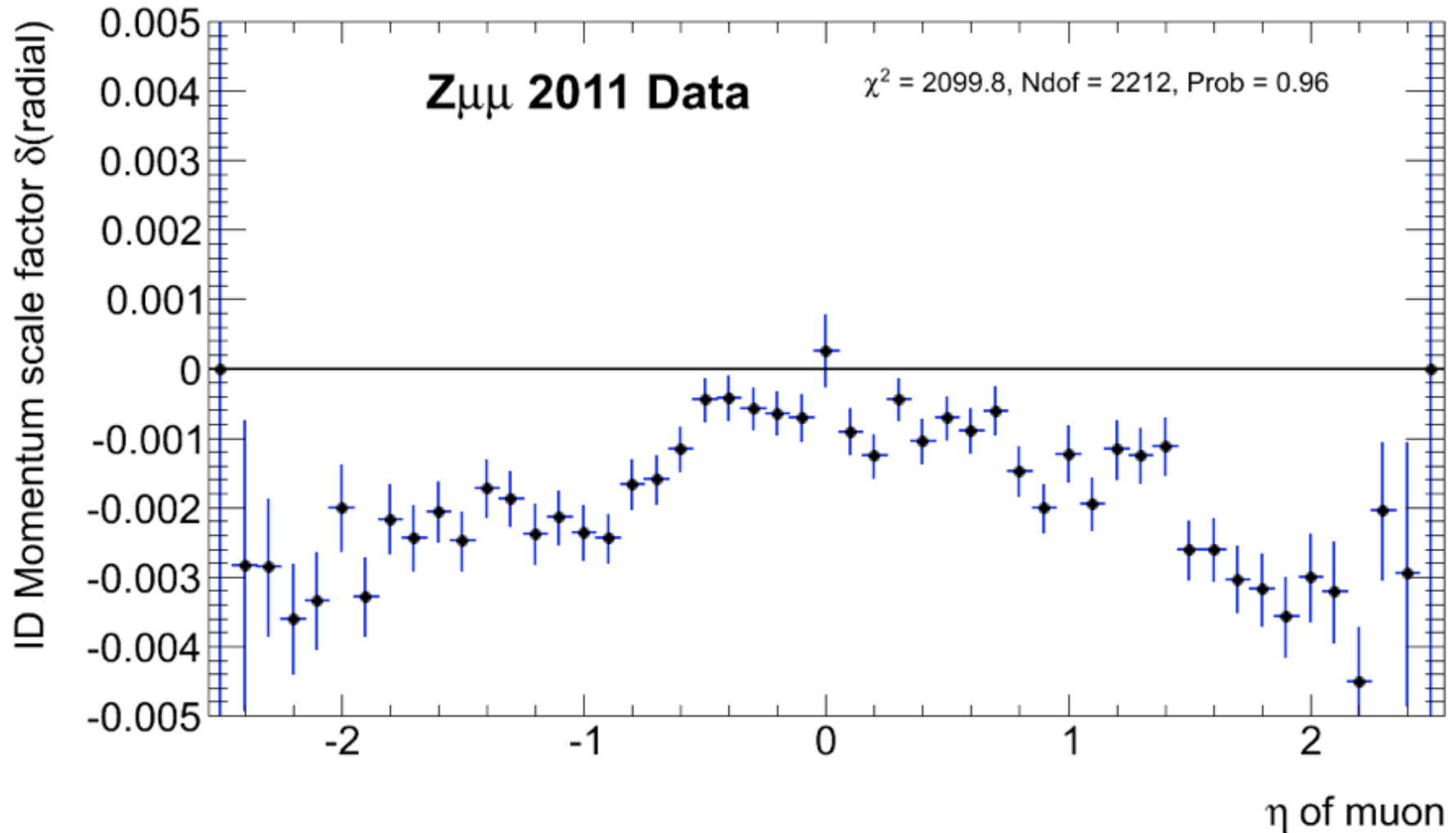
Result for Inner Detector

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

Corrected result

Rerunning with the correction applied to MC, I get: