Absorbing systematic effects to obtain a better background model in a search for new physics

Sascha Caron[1], Glen Cowan[2], Eilam Gross[3],
Stephan Horner[1] & Jan Erik Sundermann[1]

[1]Physikalisches Institut, University of Freiburg
[2]Physics Department, Royal Holloway, University of London
[3]Dep. of Particle Physics, Weizmann Institute of Science, Rehovot

ACAT Workshop, February 23rd, 2010

For details please see: S Caron et al 2009 JINST 4 P10009, arXiv:0909.3718v2

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

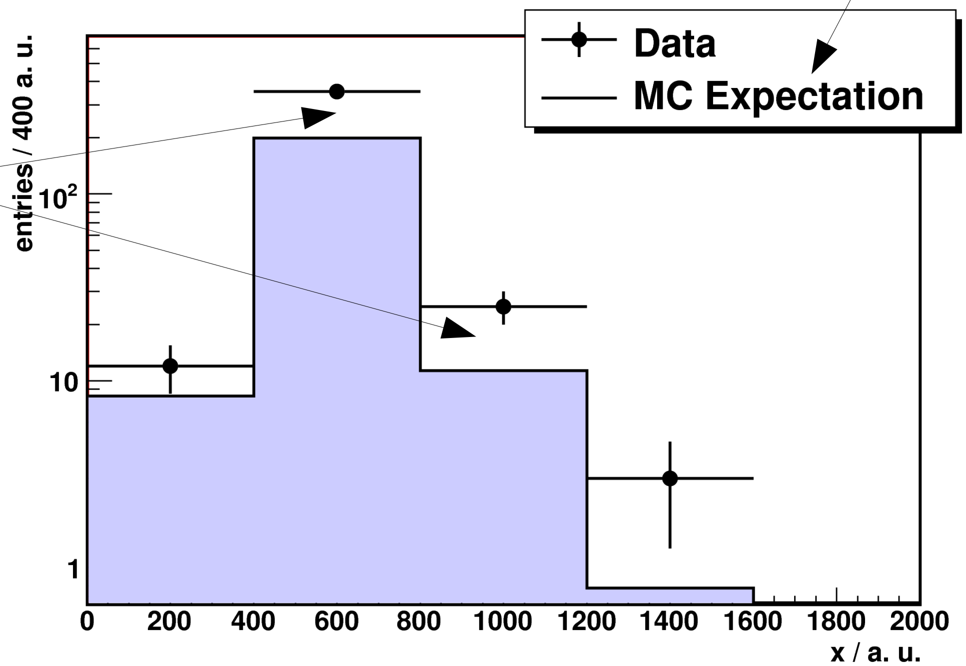
Physikalisches Institut
Albert-Ludwigs-
Universität Freiburg

Introduction

Sketch of a measurement (counting experiment):

prediction from theory

New physics or systematic effect?

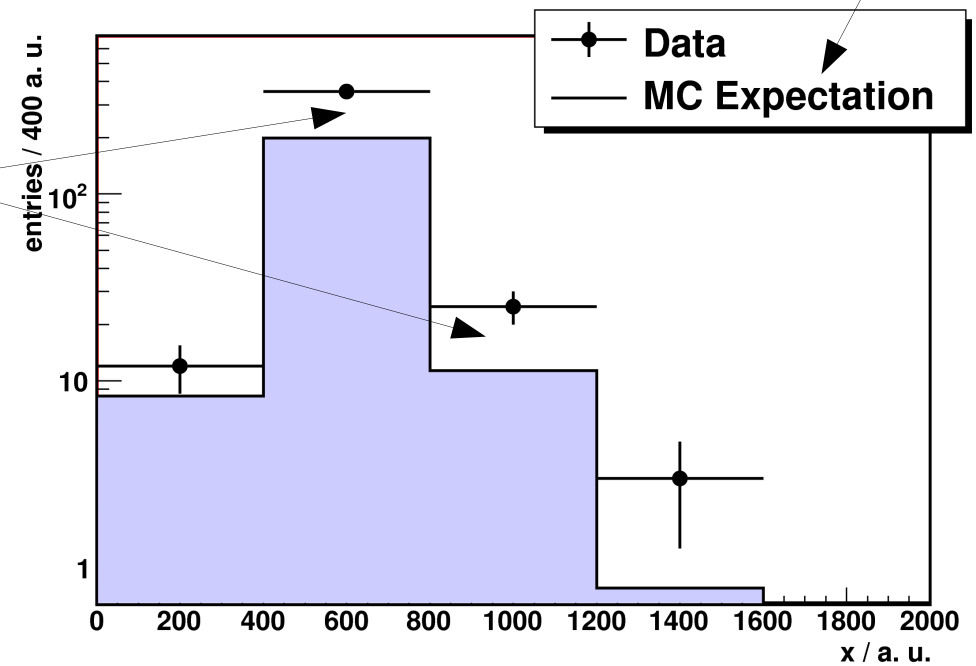


Introduction

Sketch of a measurement (counting experiment):

prediction from theory

New physics or systematic effect?



- The systematic effect can arise from shortcomings in modelling (both in theory and detector simulation).
- Therefore, the Monte Carlo (MC) prediction needs to be verified with data.

Introduction

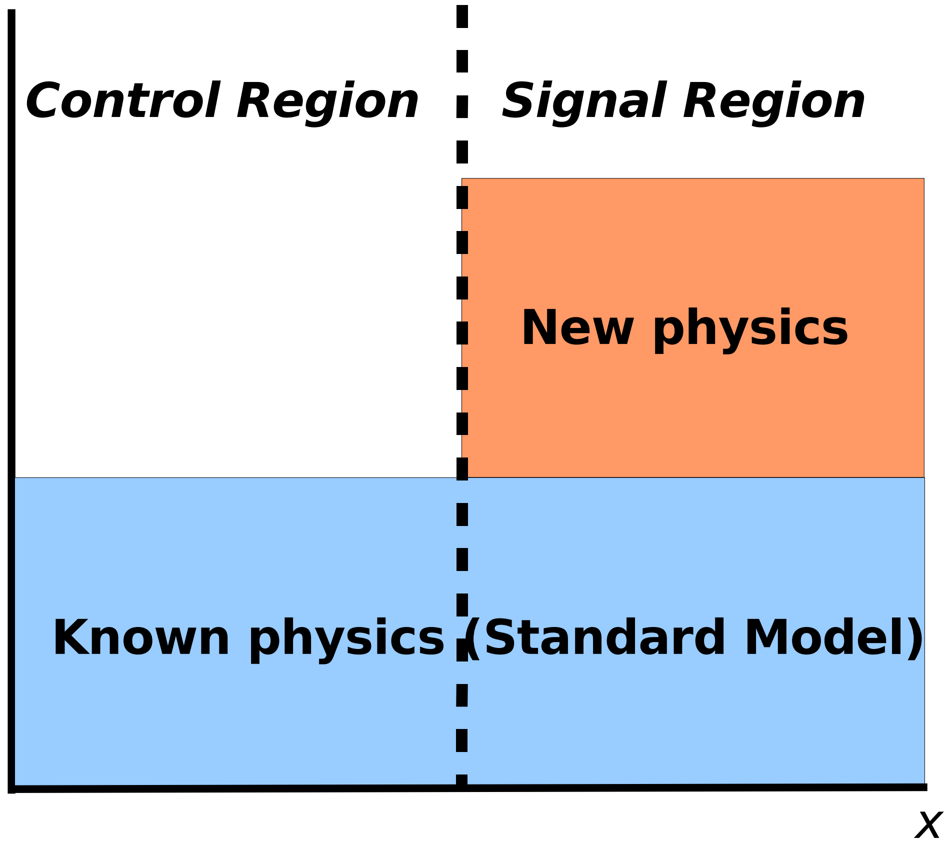
- To verify Monte Carlo find region in phase space, Control Region, satisfying:

  - ideally only known physics (Standard Model) present
  - observable of interest *x*: similar physical meaning and dependence on systematic effects in Control and Signal Region ("same" *x*)

Introduction

- To verify Monte Carlo find region in phase space, Control Region, satisfying:

  - ideally only known physics (Standard Model) present
  - observable of interest *x*: similar physical meaning and dependence on systematic effects in Control and Signal Region ("same" *x)*



Desired scenario:

- new physics can appear in Signal Region only
- Same background (known physics) in Control and Signal Region

Introduction

Common approaches to obtain a background prediction for the Signal Region:

a) Use data from Control Region (CR) as model for Signal Region (SR)

Drawbacks: - data fluctuations induce bias
- shapes in CR & SR must be the same

Introduction

Common approaches to obtain a background prediction for the Signal Region:

a) Use data from Control Region (CR) as model for Signal Region (SR)

Drawbacks: - data fluctuations induce bias
- shapes in CR & SR must be the same

b) Divide data by MC template in CR and use ratio as correction for SR

Drawbacks: - data fluctuations induce bias
- correct each bin in SR independently

Introduction

Common approaches to obtain a background prediction for the Signal Region:

a) Use data from Control Region (CR) as model for Signal Region (SR)

Drawbacks: - data fluctuations induce bias
- shapes in CR & SR must be the same

b) Divide data by MC template in CR and use ratio as correction for SR

Drawbacks: - data fluctuations induce bias
- correct each bin in SR independently

c) Fit function to data in CR and rescale it for SR

Drawbacks: - can be difficult to get shape right
- shapes in CR & SR must be the same

Introduction

Common approaches to obtain a background prediction for the Signal Region:

a) Use ... on (SR)

D...

Our proposal: Modify MC template with a correction function

- Use MC expectation as starting point, since it is best estimate when no systematics present

- Assume that systematic effects can be described by simple functions

b) Div ... n for SR

D...

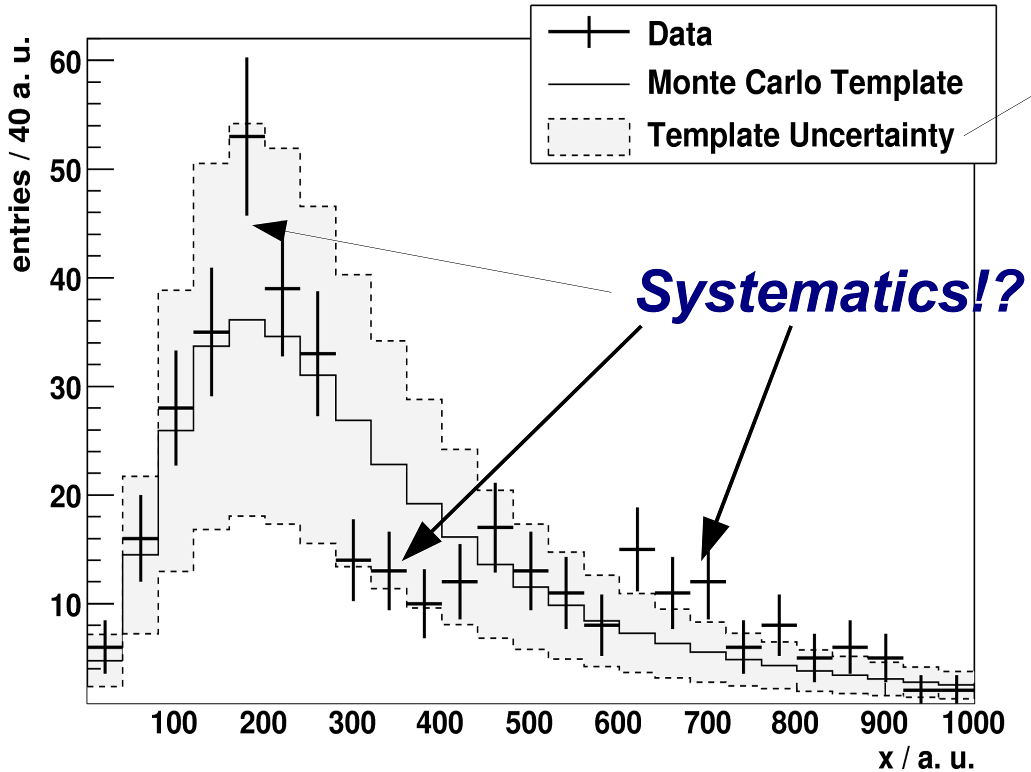
c) Fit ...

Draw...

- shapes in CR & SR must be the same

Introducing the method

Toy example of a measurement in a control region:



determined by varying known systematic sources

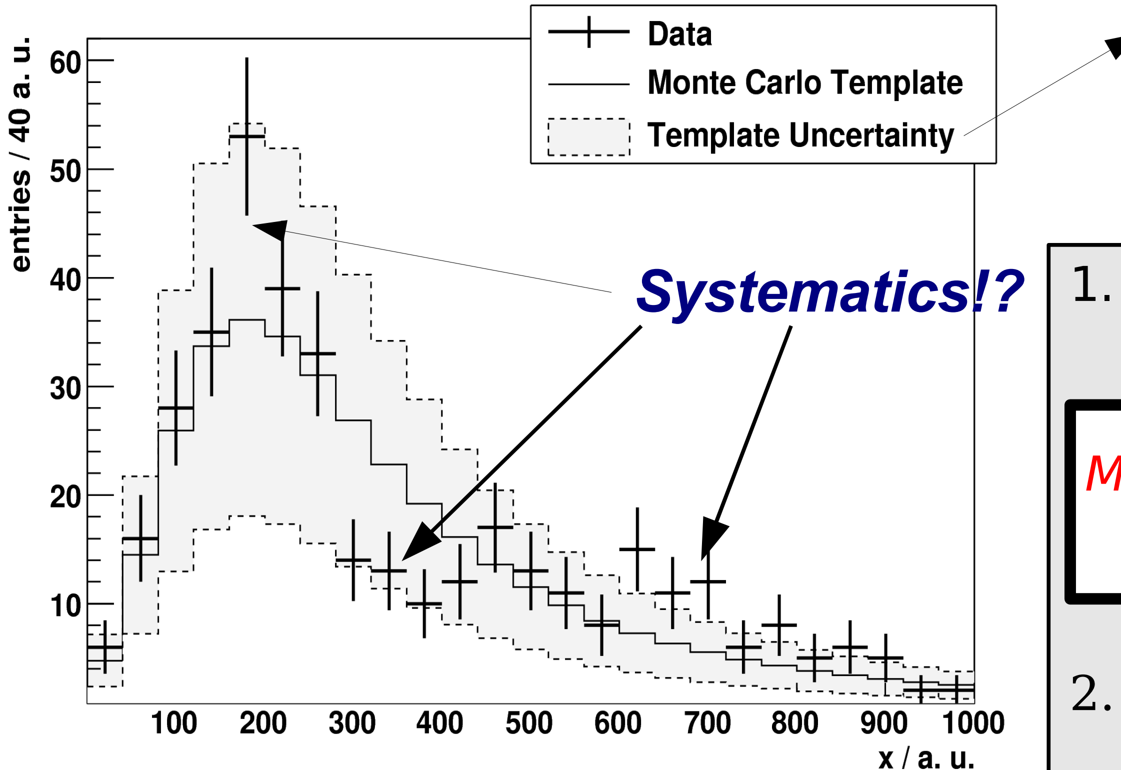
Compatibility with central prediction:

Probability *p = 0.002*

(Probability to observe such data or data less likely if MC template is true model)

Introducing the method

Toy example of a measurement in a control region:



1. Multiply the MC template with a correction function

Model_x = *Template* * *Polynomial* with *x parameters*

2. Fit the modified template to the data to determine parameters

3. Use successively more complex correction functions until satisfactory goodness-of-fit is reached (*p*-Value)

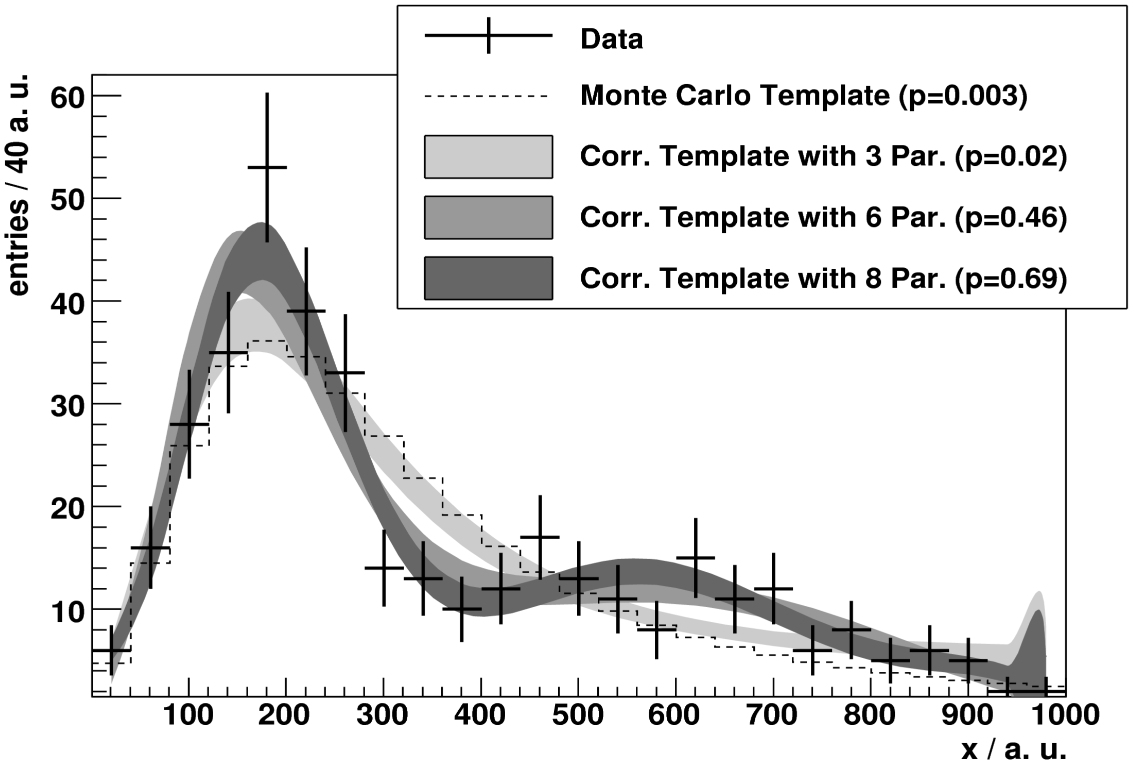
Compatibility with central prediction:

Probability $p = 0.002$

(Probability to observe such data or data less likely if MC template is true model)

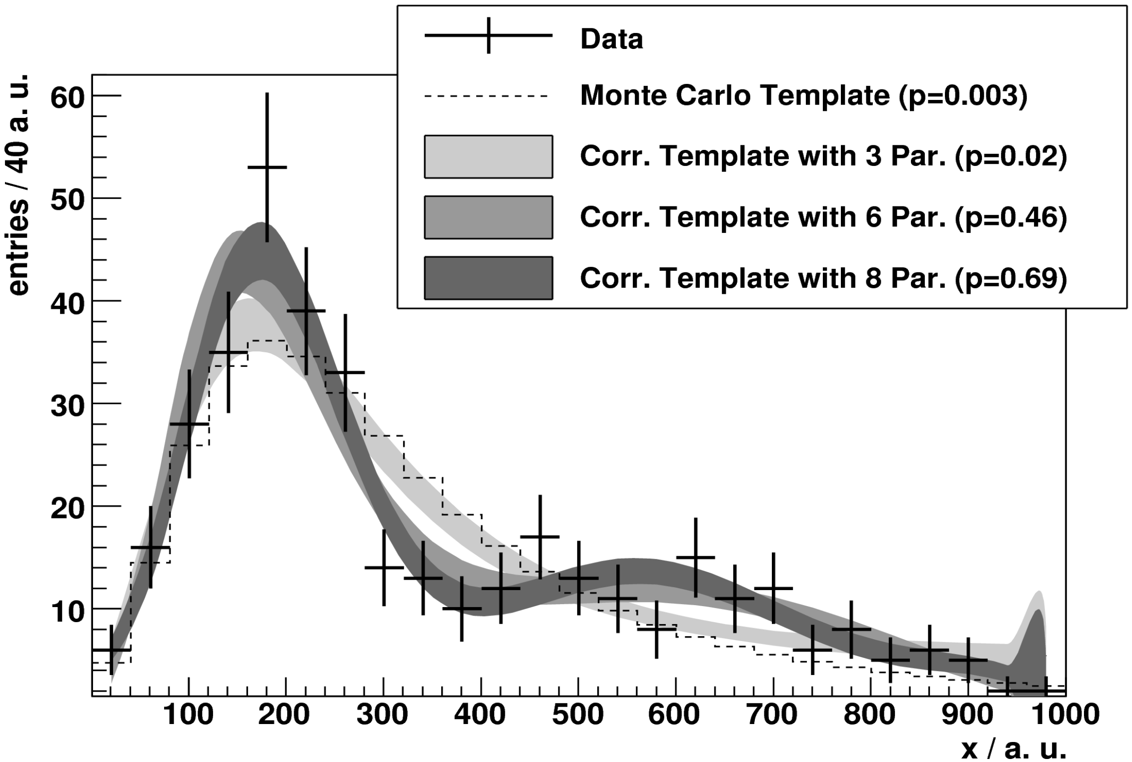
Selecting a better model

Ordinary polynomials as correction functions:
Model_x = Template * Polynomial with x parameters



Selecting a better model

Ordinary polynomials as correction functions:
Model_x = Template * Polynomial with x parameters



Absolute goodness-of-fit:

p(Model_0) = 0.0027
p(Model_1) = 0.0033
p(Model_5) = 0.33
p(Model_7) = 0.46
p(Model_8) = 0.69
p(Model_9) = 0.63

Relative goodness-of-fit:

p(Model_0 | Model_1) = 0.15
p(Model_7 | Model_8) = 0.04
p(Model_8| Model_9) = 0.80

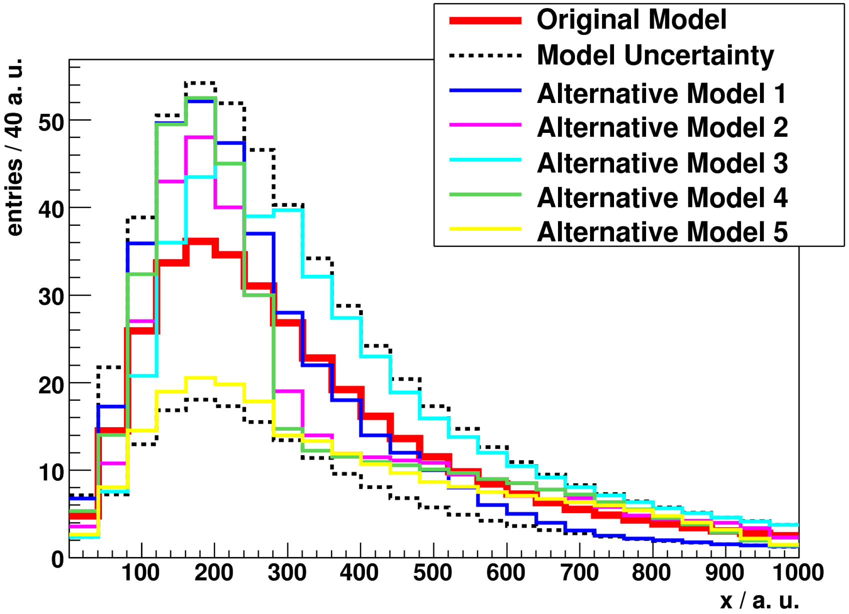
low number indicates improvement when going to the next model (see backup)

In this case several parameters needed
due to large systematic effects (see next slide)

Shape uncertainty in starting template

<u>In real case:</u> vary Monte Carlo prediction according to known systematic effects to obtain alternative starting templates.

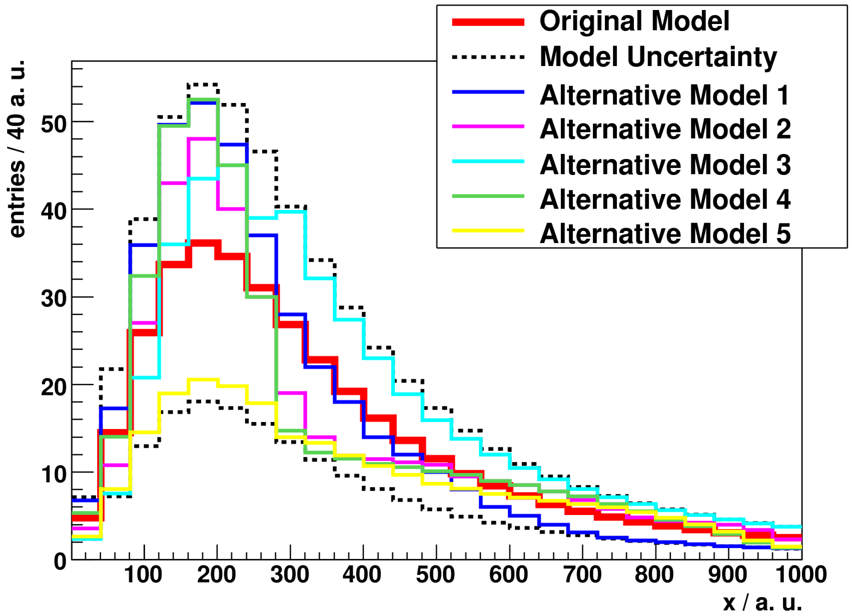
Before correction:

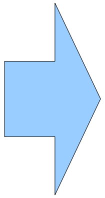


Shape uncertainty in starting template

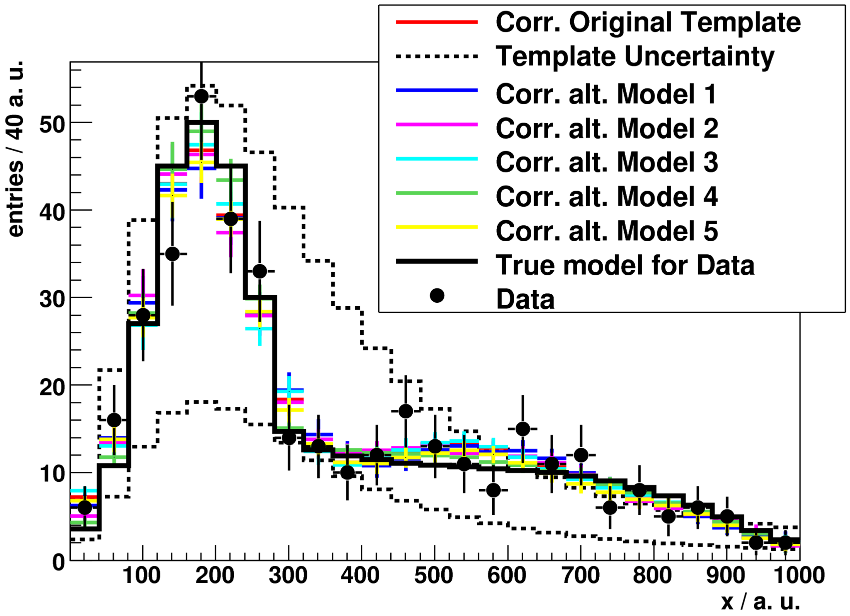
In real case: vary Monte Carlo prediction according to known systematic effects to obtain alternative starting templates.

Before correction:





After correction:



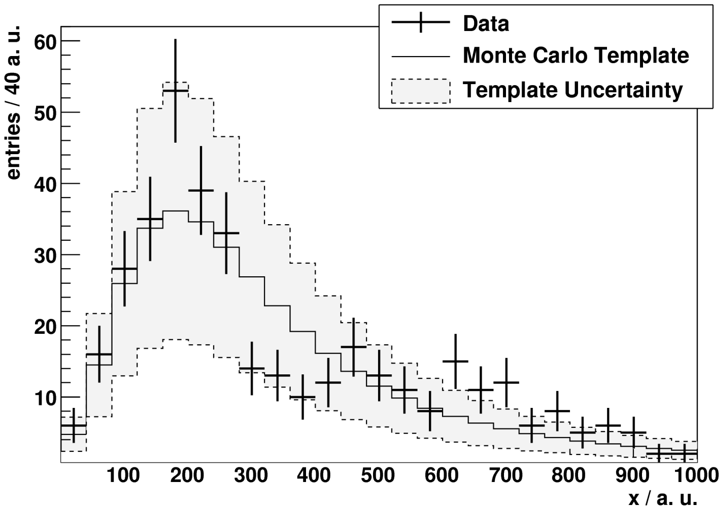
✗ True model has large systematic deviations from original MC template, but they are absorbed into the new improved model

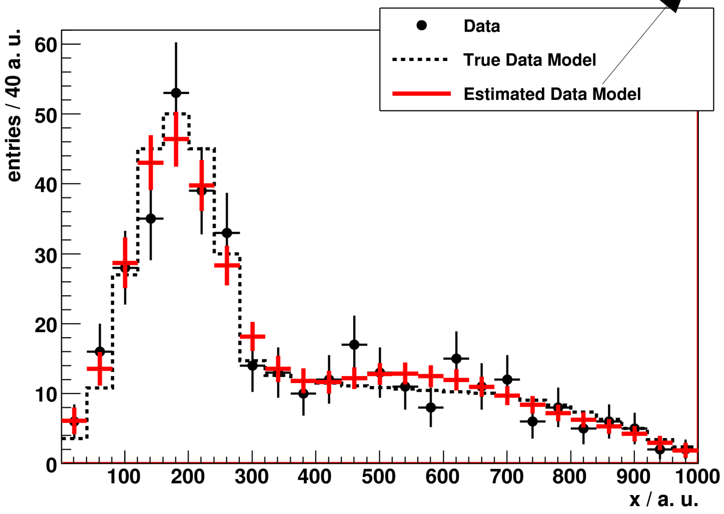
✗ Furthermore, choice of the starting template has only little influence.

Average corrected models to obtain a best estimate ⟶

Proposed Method applied:

Errors determined using toy data sets generated from Estimated Model

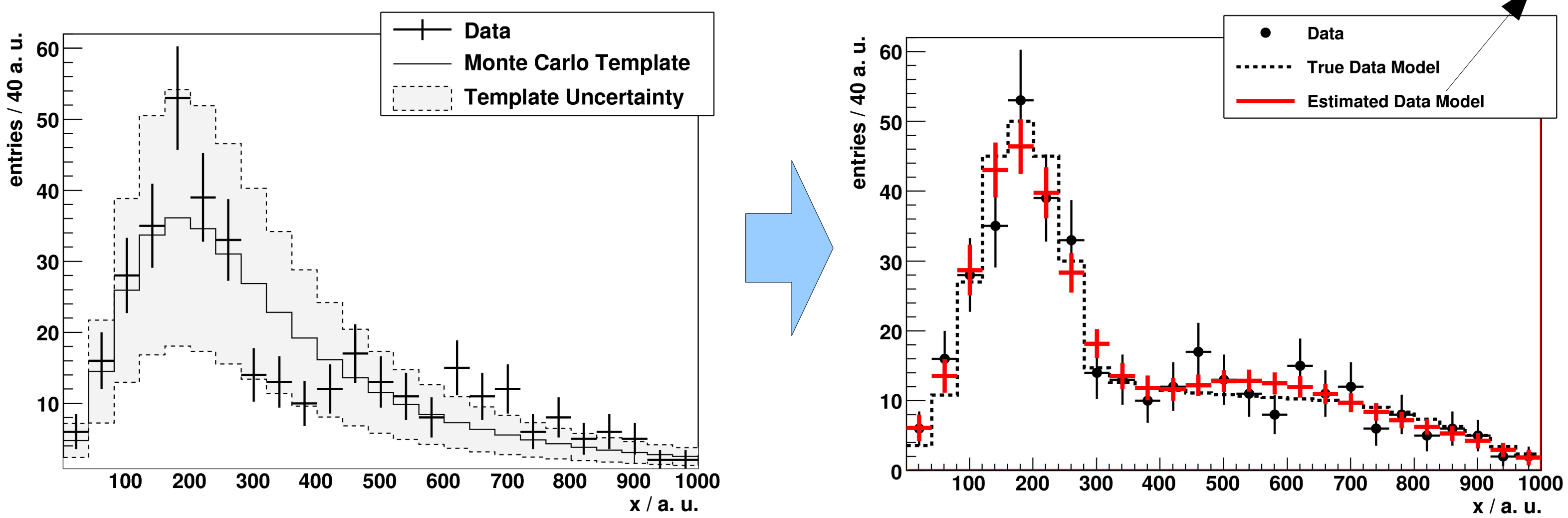




Large systematics absorbed and uncertainty reduced!

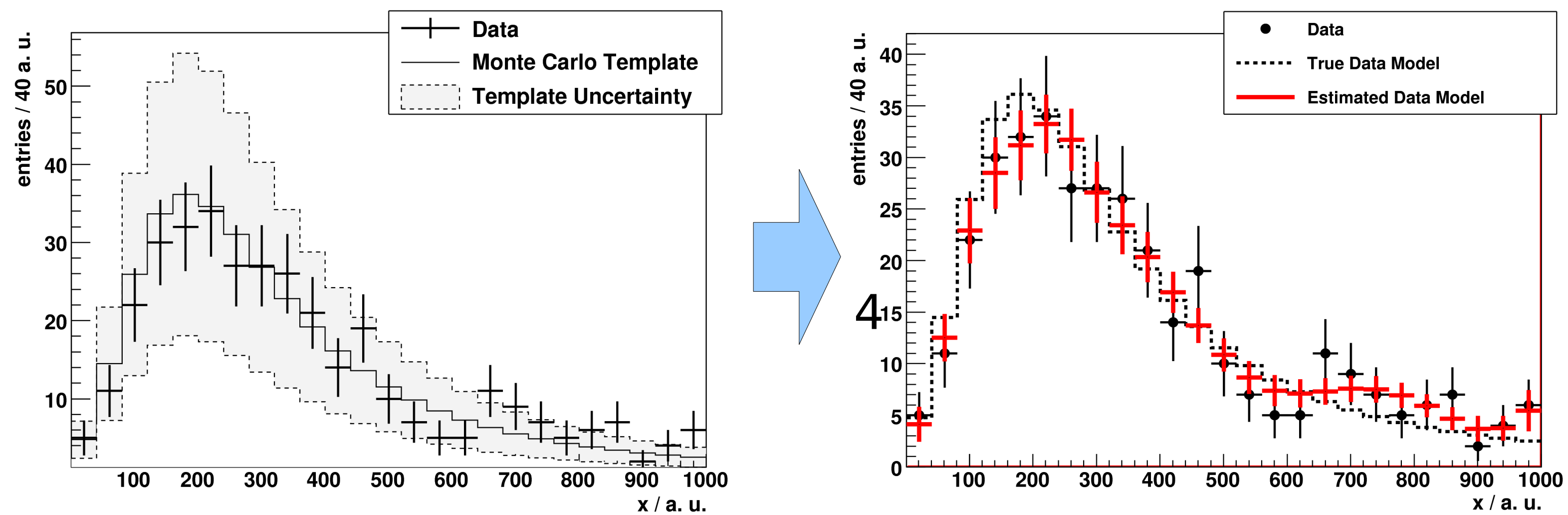
Proposed Method applied:

Errors determined using toy data sets generated from Estimated Model



Large systematics absorbed and uncertainty reduced!

Special test case: no systematic effects included



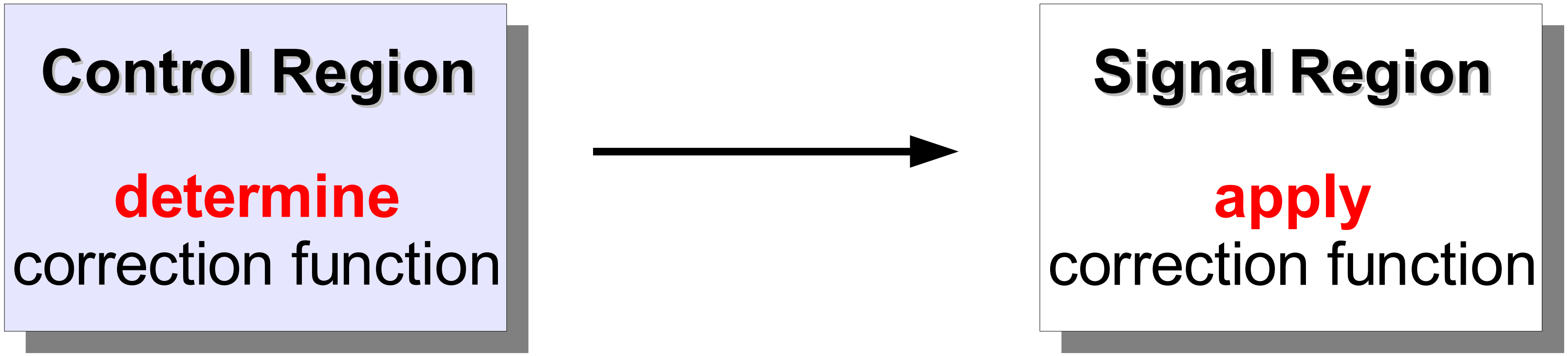
True model (= original MC prediction) reproduced!

Transfer to Signal Region:

After form of correction determined in Control Region, apply on Monte Carlo template for Signal Region

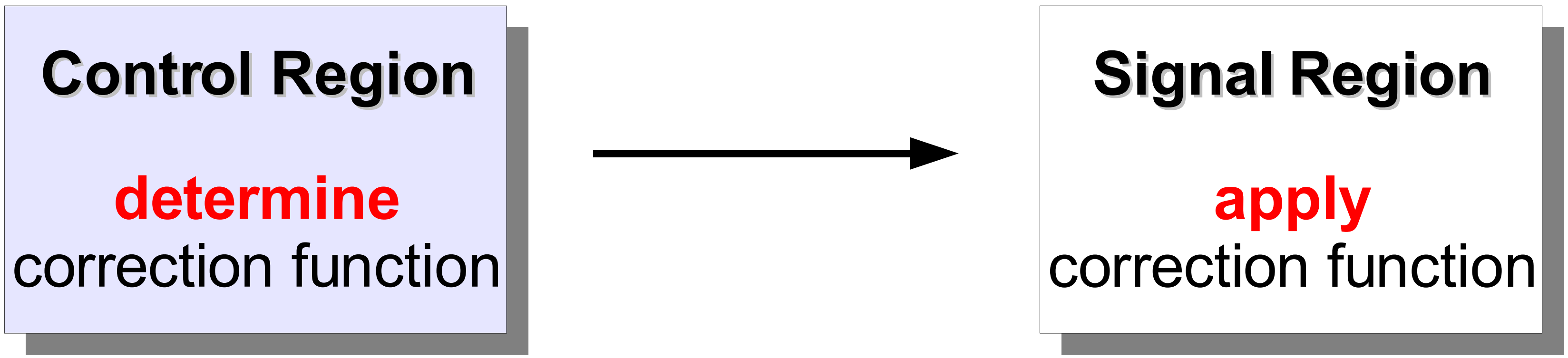
Transfer to Signal Region:

After form of correction determined in Control Region, apply on Monte Carlo template for Signal Region



Transfer to Signal Region:

After form of correction determined in Control Region, apply on Monte Carlo template for Signal Region



<u>Advantage of proposed method:</u>

Data distributions don't need to have the same shapes in signal and control regions. Only the systematics have to affect them similarly.

Now look at Signal Region

Consider simple case:

✗ Shapes of MC templates in both regions the same

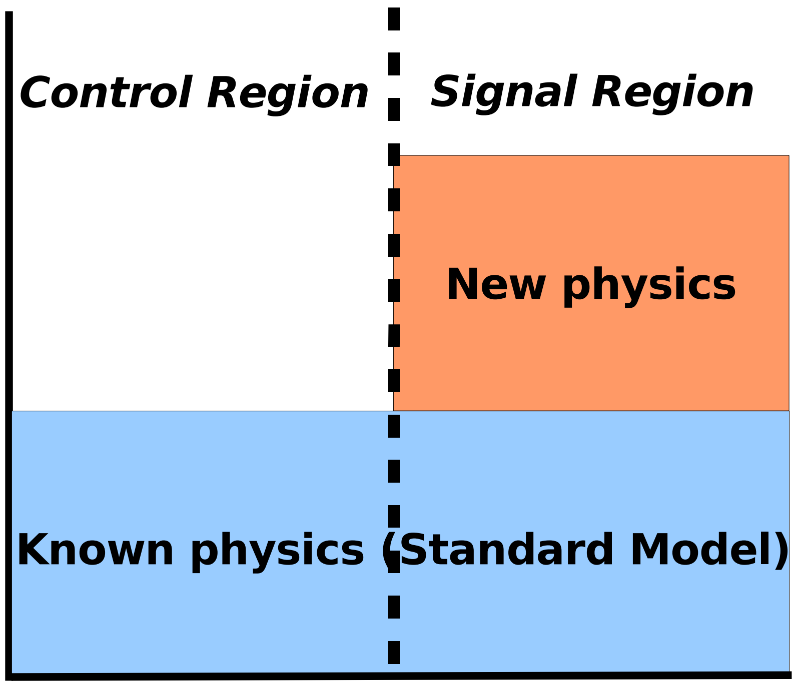
✗ Event efficiency of Signal to Control Region taken to be unity

Now look at Signal Region

Consider simple case:

✗ Shapes of MC templates in both regions the same

✗ Event efficiency of Signal to Control Region taken to be unity



→ *consider scenario with no systematic effects as a limiting case (original MC expectation = correct model) → next slide*

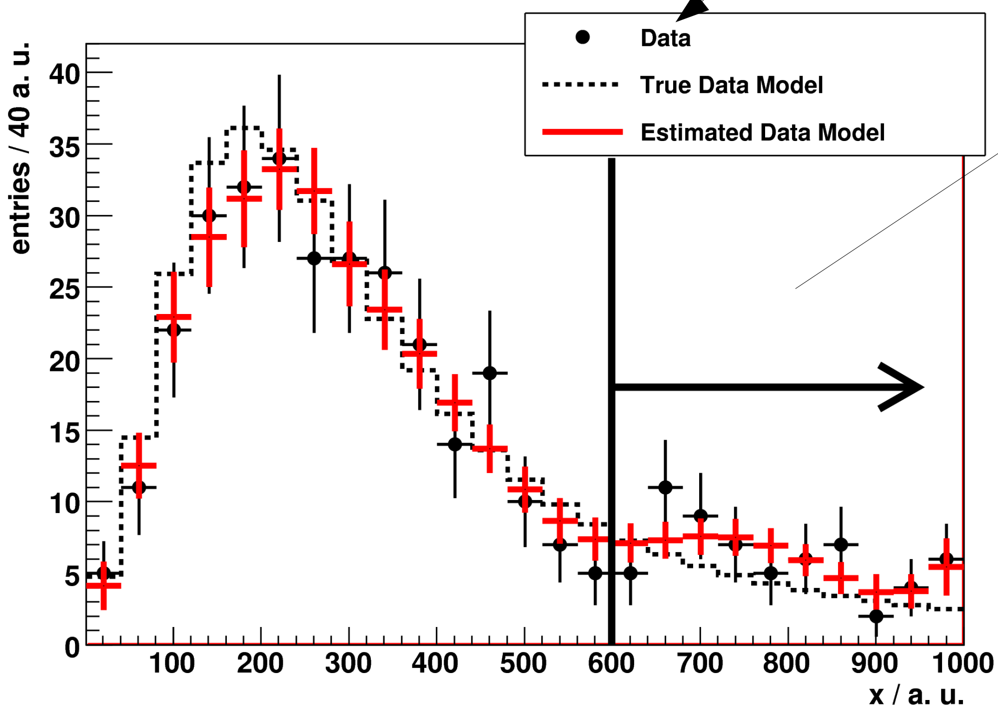
NOT accounted for here:

Systematic effects may affect regions differently → additional uncertainty

Expected background events in Signal Region

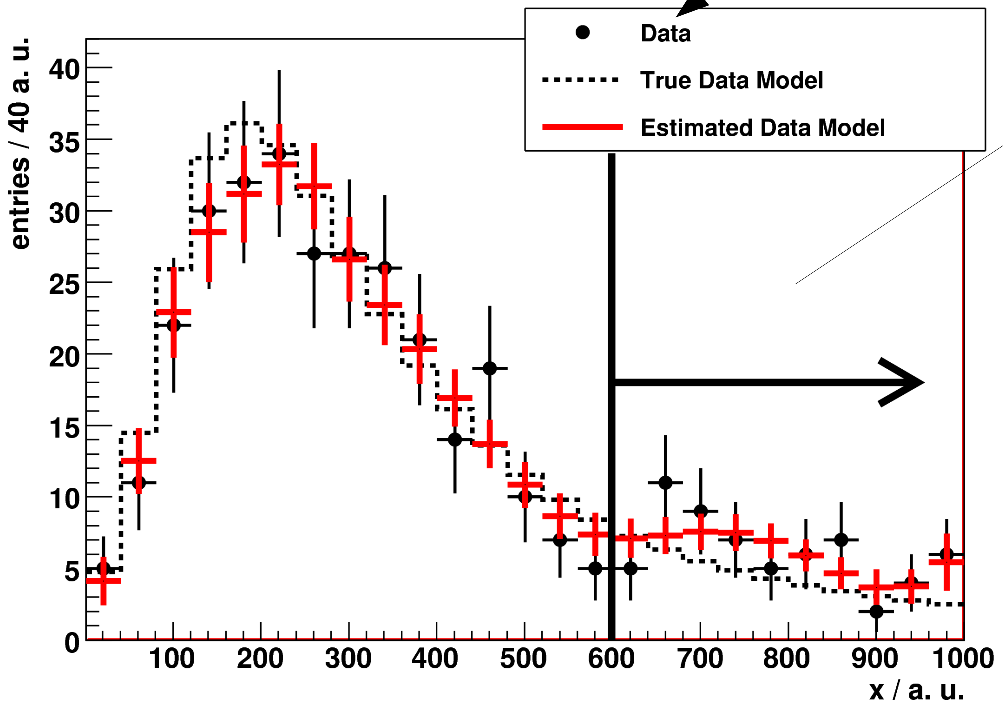
from Control Region!

Region of interest to look for new physics (x > 600 a. u.)



Expected background events in Signal Region

from Control Region!



Region of interest to look for new physics (x > 600 a. u.)

- ✗ Sum up bins taking into account the correlation
- ✗ Compare with simply using the data from Control Region

Model	Number of expected events	Relative error
Original prediction (MC template)	43.9 ± 21.9	50%
Corrected model	59.9 ± 7.6	12.7%
Data as model	62.0 ± 7.9	12.7%

But in general error of corrected model smaller than data error. →

Considering many experiments

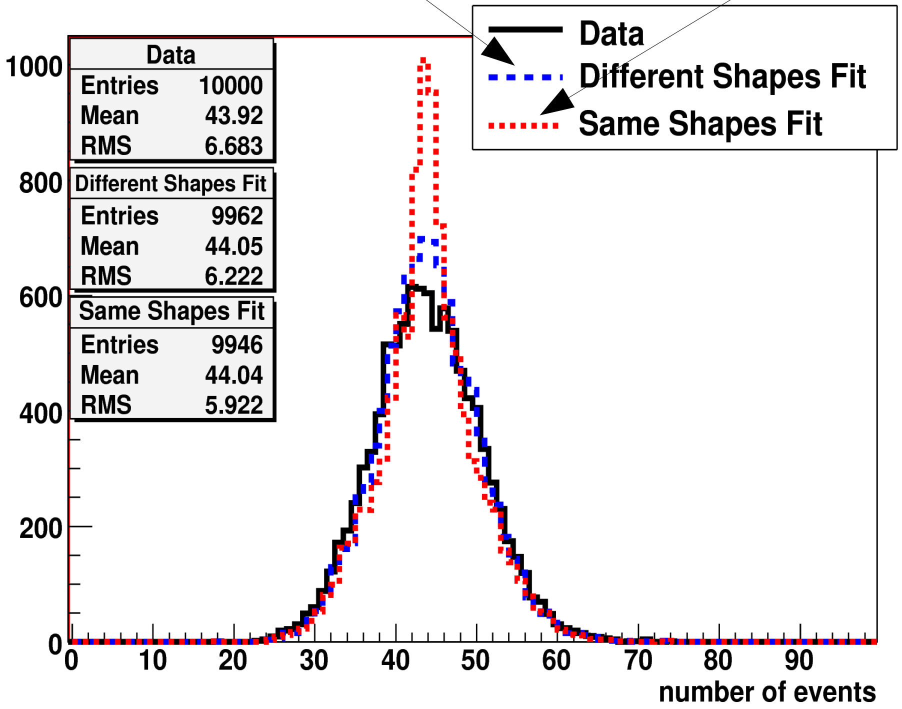
- ✗ Generate 10.000 toy data sets from true model and apply method

Considering many experiments

✗ Generate 10.000 toy data sets from true model and apply method

Same starting templates as before

Templates differ from true model by scale only

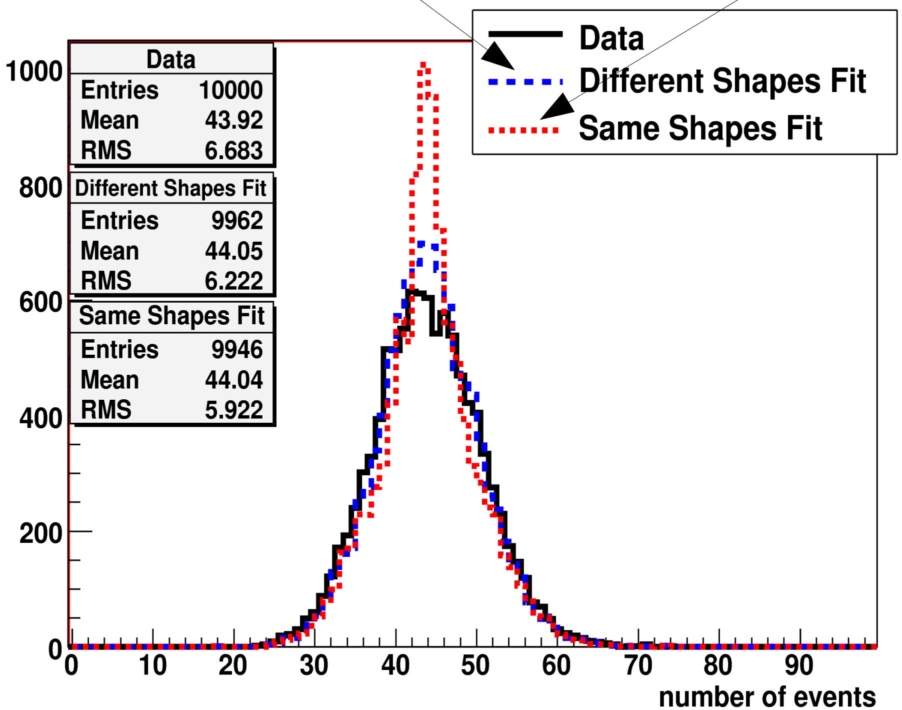


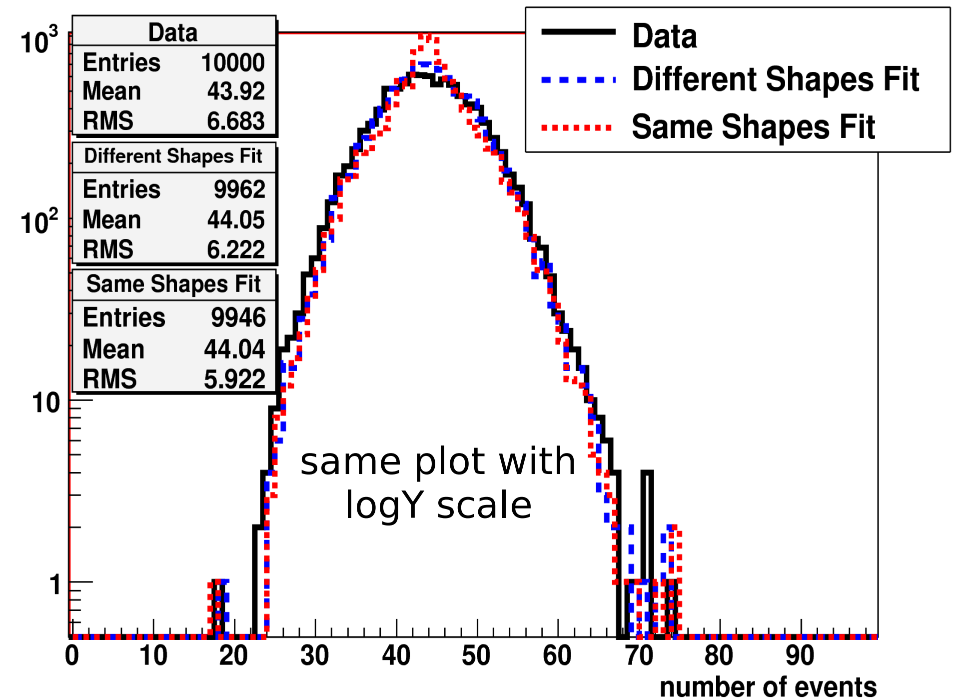
Considering many experiments

✗ Generate 10.000 toy data sets from true model and apply method

Same starting templates as before

Templates differ from true model by scale only





Method has smaller uncertainty than using the data as a model and reproduces true mean (43.89) within 2.6% of quoted error

Discovery Significance

Significance: convolute Poisson probability of a measurement with Gaussian priors for the background expectation (using the uncertainties from the previous slide):

Discovery Significance

Significance: convolute Poisson probability of a measurement with Gaussian priors for the background expectation (using the uncertainties from the previous slide):

Assume the following measurements

x > 600 a. u.: 99 events counted

	Bgrd predicted: (true value 43.89)	**Significance:**
Data	43.92 ± 6.683	5.01
Different Shapes	44.05 ± 6.222	5.15
Same Shapes	44.04 ± 5.922	5.25

Equivalent to 4% luminosity increase

Discovery Significance

Significance: convolute Poisson probability of a measurement with Gaussian priors for the background expectation (using the uncertainties from the previous slide):

Assume the following measurements

	x > 600 a. u.: 99 events counted		x > 800 a. u. : 52 events counted	
	Bgrd predicted: (true value 43.89)	**Significance:**	**Bgrd predicted:** (true value 15.61)	**Significance:**
Data	43.92 ± 6.683	5.01	15.62 ± 3.933	5.10
Different Shapes	44.05 ± 6.222	5.15	15.57 ± 3.596	5.29
Same Shapes	44.04 ± 5.922	5.25	15.53 ± 3.446	5.38

Equivalent to 4% luminosity increase   **12% lumi increase**

Improvement wrt. Data model even in this "optimal" scenario (no systematic effects, shapes in CR & SR identical)

Summary:

1. We propose to modify Monte Carlo predictions with correction functions to account for systematic effects.

2. Successively more complex functions are used until sufficient compatibility with data is reached.

Summary:

1. We propose to modify Monte Carlo predictions with correction functions to account for systematic effects.

2. Successively more complex functions are used until sufficient compatibility with data is reached.

3. Data distributions don't need to have the same shapes in signal and control regions.
   Only the systematics have to affect them similarly.

4. Method not restricted to High Energy Physics!

Thank you for your attention

Backup slides

Statistical tests to determine the best model

Employ 2 likelihood ratios to assess the compatibility with data:

1. Absolute goodness-of-fit

Compare model i (polynomial i * template) with most flexible model where each bin can vary independently and will therefore take on the data values:

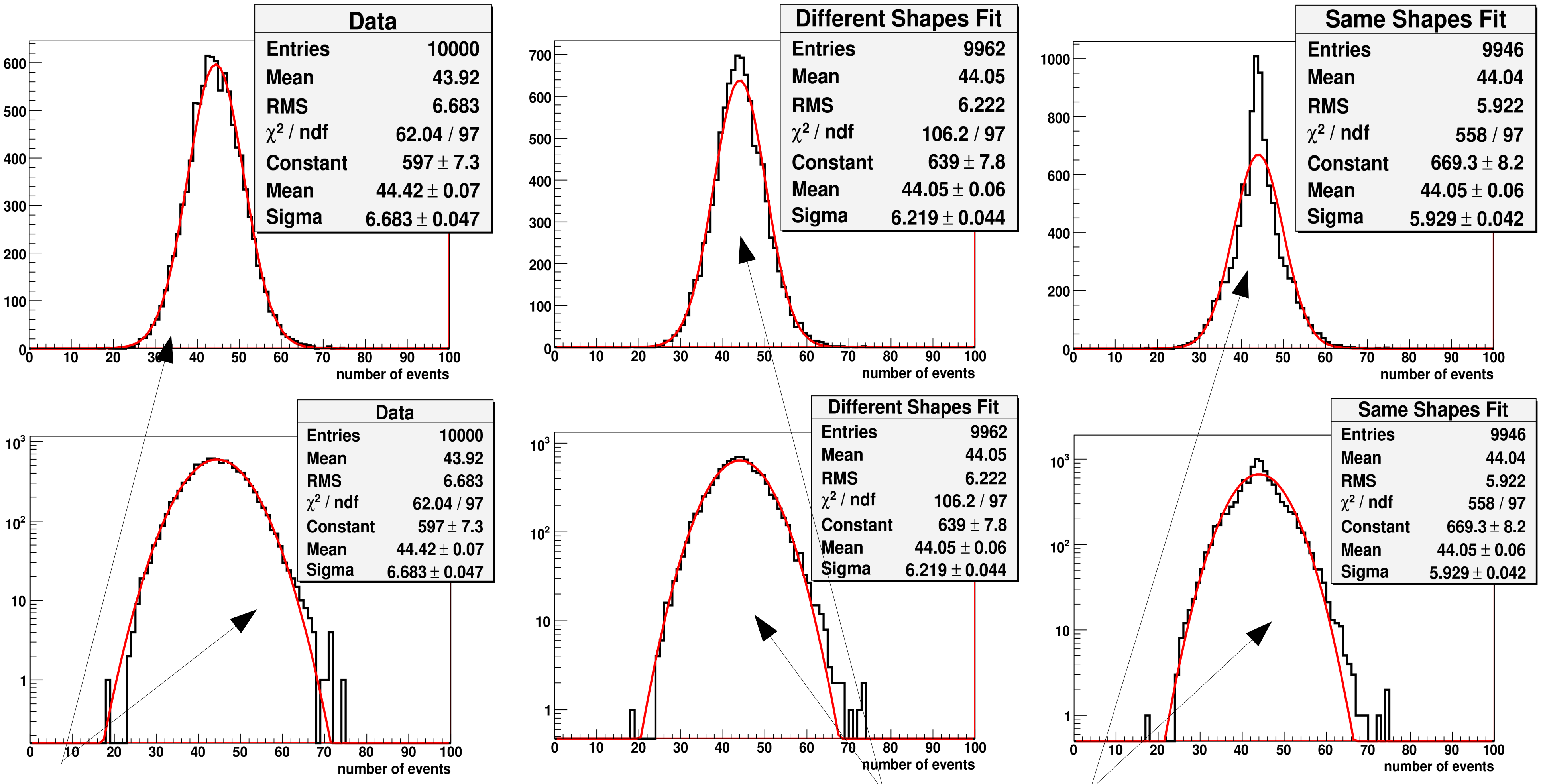
$$q_{abs} = -2 \ln \frac{\text{LH (Data | Model i)}}{\text{LH (Data | most flex. model = Data)}} \sim \chi^2$$

2. Does the next best model significantly improve the data description?

Compare model i with model i+1:

$$q_{rel} = -2 \ln \frac{\text{LH (Data | Model i)}}{\text{LH (Data | Model i+1)}} \sim \chi^2$$

Considering many experiments - Gaussian fits



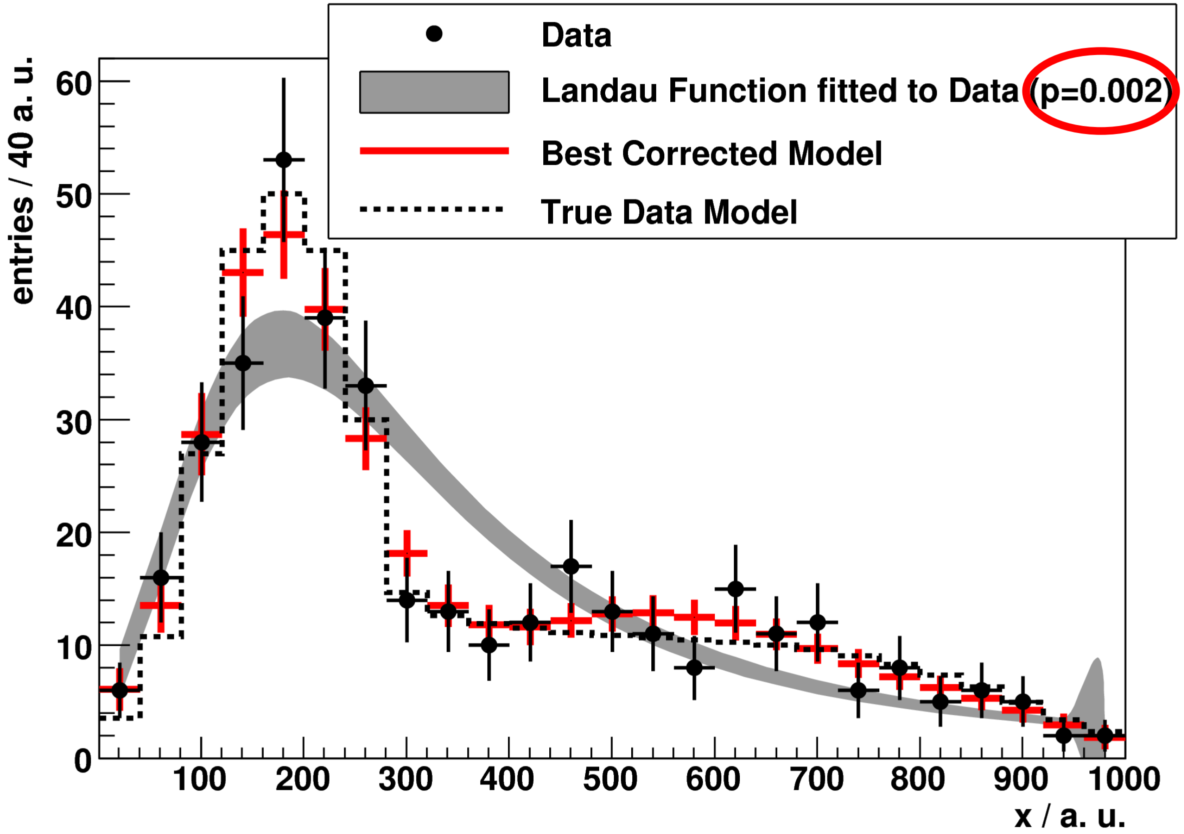
expect exact Poisson dist

Gaussian behavior desired

Expect Gaussian behavior to improve when including uncertainty for transfer from Control to Signal region.

Compare with: Parametrizing the Monte Carlo Template

Fit a function inspired by the MC to the data in the control region
(Original Template is a Landau Function)



This example: If systematics can't be compensated by adjustment of parameters data won't be nicely described.