Statistical techniques for incorporating systematic/theory uncertainties

Theory/Experiment Interplay at the LHC

RHUL, 8 April, 2010

Glen Cowan
Physics Department,
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

General statistical formalism for a search at the LHC

Mainly frequentist

Significance test using profile likelihood ratio

Distributions of profile likelihood ratio in large sample limit

(with E. Gross, O. Vitells, K. Cranmer)

General strategy for dealing with systematics

Improve model by including additional parameters

Example 1: tau hadronic mass distribution

Example 2: $b \rightarrow s\gamma$

Prototype analysis

Search for signal in a region of phase space; result is histogram of some variable *x* giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\rm tot} \int_{\mathrm{bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx \,, \qquad b_i = b_{\rm tot} \int_{\mathrm{bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx \,.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

(N.B. here m = number of counts, not mass!)

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_{\rm s}$, $\boldsymbol{\theta}_{\rm b}$,$b_{\rm tot}$)

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Can base significance test on the profile likelihood ratio:

maximizes L for specified μ

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

maximize L

The likelihood ratio gives optimum test between two point hypotheses (Neyman-Pearson lemma).

Should be near-optimal in present analysis with variable μ and nuisance parameters $\boldsymbol{\theta}$.

Test statistic for discovery

Try to reject background-only ($\mu = 0$) hypothesis using

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. only regard upward fluctuation of data as evidence against the background-only hypothesis.

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p-value for an observed $q_{0,\text{obs}}$ is

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0)\, dq_0$$

will get formula for this later

Test statistic for upper limits

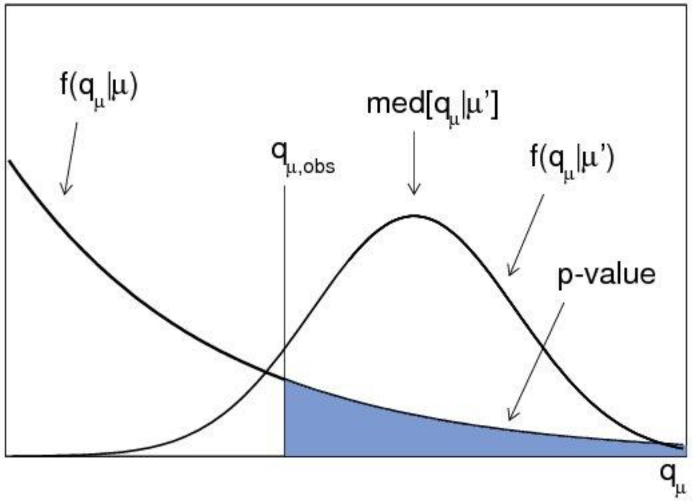
For purposes of setting an upper limit on μ use

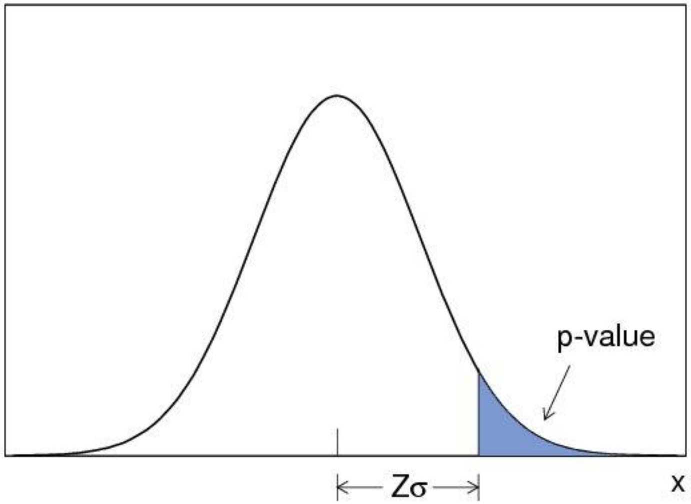
$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Note for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ.

p-value / significance of hypothesized μ

Test hypothesized μ by giving p-value, probability to see data with $\leq$ compatibility with μ compared to data observed:





Equivalently use significance, Z, defined as equivalent number of sigmas for a Gaussian fluctuation in one direction:

$$Z = \Phi^{-1}(1 - p)$$

Using p-value for discovery/exclusion

Carry out significance test of various hypotheses (background-only, signal plus background, …)

Result is p-value.

Exclude hypothesis if p-value below threshold:

Discovery: test of background-only hypothesis. Exclude if

$p < 2.9 \times 10^{-7}$ (i.e. Gaussian signif. $Z = \Phi^{-1}(1-p) > 5$)

Limits: test signal (+background) hypothesis. Exclude if

$p < 0.05$ (i.e. 95% CL limit)

Wald approximation for profile likelihood ratio

To find *p*-values, we need: $f(q_0|0), \quad f(q_\mu|\mu)$

For median significance under alternative, need: $f(q_\mu|\mu')$

Use approximation due to Wald (1943)

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N})$$

$$\hat{\mu} \sim \text{Gaussian}(\mu', \sigma)$$

sample size

$$\text{i.e.,} \quad E[\hat{\mu}] = \mu'$$

σ from covariance matrix V, use, e.g.,

$$V^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

Distribution of q_0

Assuming the Wald approximation, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \Phi\left(\frac{\mu'}{\sigma}\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \sqrt{\Lambda}\right)^2\right]$$

The special case $\mu' = 0$ is a "half chi-square" distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of *q*0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The *p*-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance *Z* is simply

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$

Distribution of q_μ

Similar results for q_μ

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{(\mu - \mu')}{\sigma}\right)^2\right]$$

$$f(q_\mu|\mu) = \frac{1}{2}\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}e^{-q_\mu/2}$$

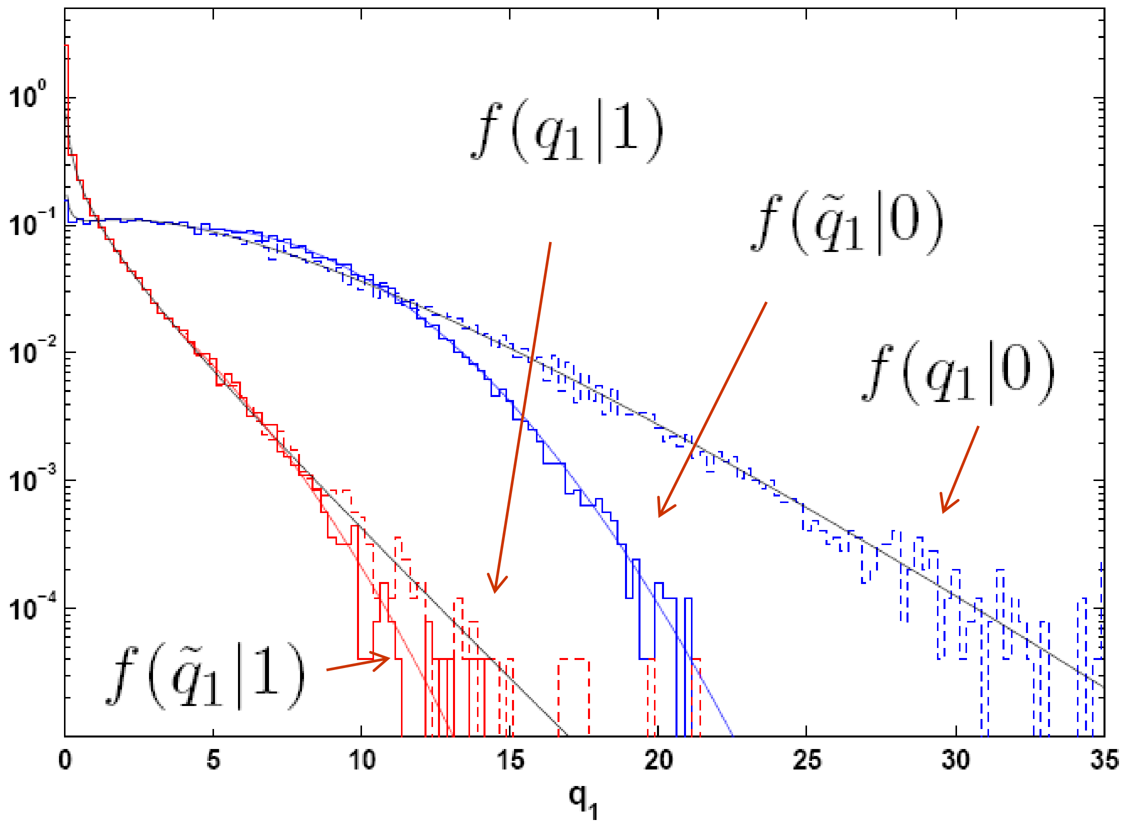
$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{(\mu - \mu')}{\sigma}\right)$$

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

An example

$$n \sim \text{Poisson}(\mu s + b) \qquad s = 50,\ b = 100,\ \tau = 1$$

$$m \sim \text{Poisson}(\tau b)$$

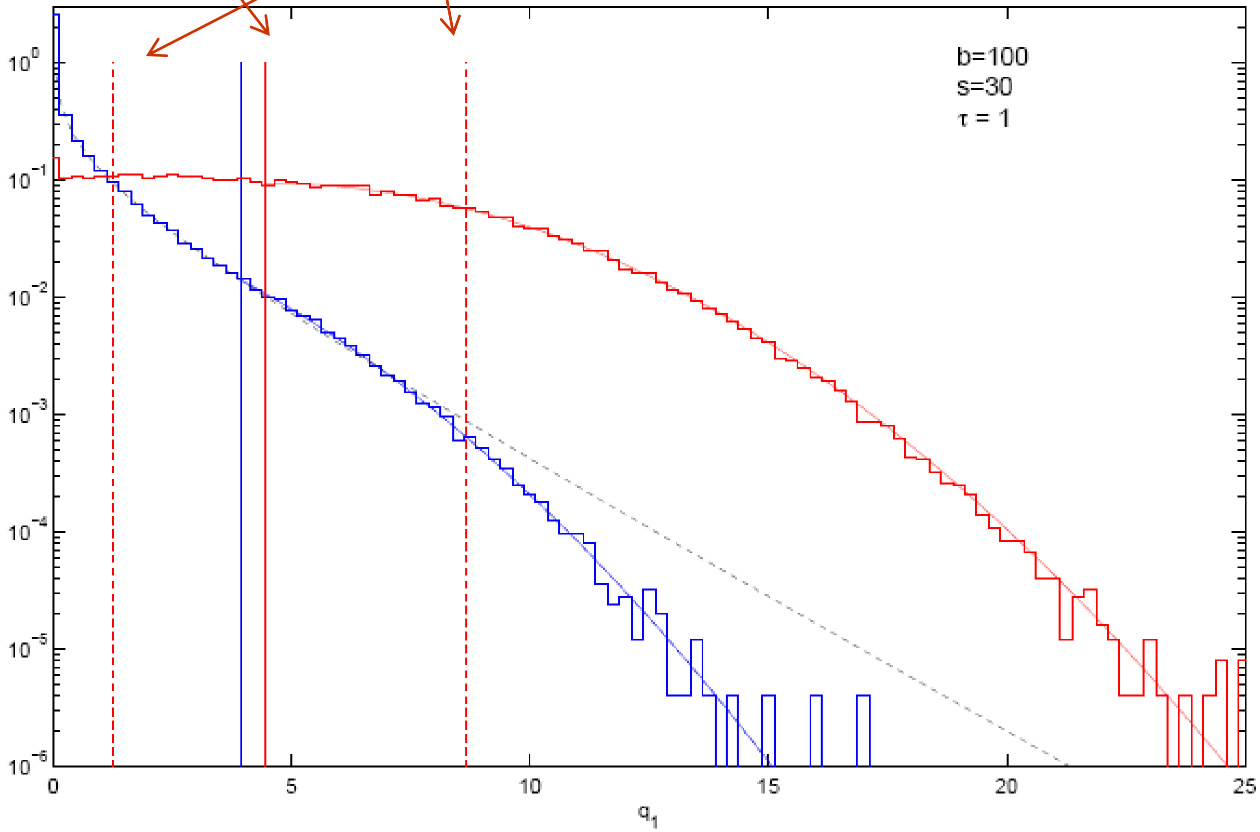


Error bands

Because of the monotonic relation between $\hat{\mu}$ and q_0, q_μ, $\tilde{q}_\mu$, we can easily derive e.g. $\pm 1\sigma$ error bands error bands for the median significance.

$\mathrm{med}[\tilde{q}_1|0]$

$\pm 1\sigma$



Dealing with systematics

S. Caron, G. Cowan, S. Horner, J. Sundermann, E. Gross, 2009 JINST 4 P10009

Suppose one needs to know the shape of a distribution.
Initial model (e.g. MC) is available, but known to be imperfect.

Q: How can one incorporate the systematic error arising from use of the incorrect model?

A: Improve the model.

That is, introduce more adjustable parameters into the model so that for some point in the enlarged parameter space it is very close to the truth.

Then use profile the likelihood with respect to the additional (nuisance) parameters. The correlations with the nuisance parameters will inflate the errors in the parameters of interest.

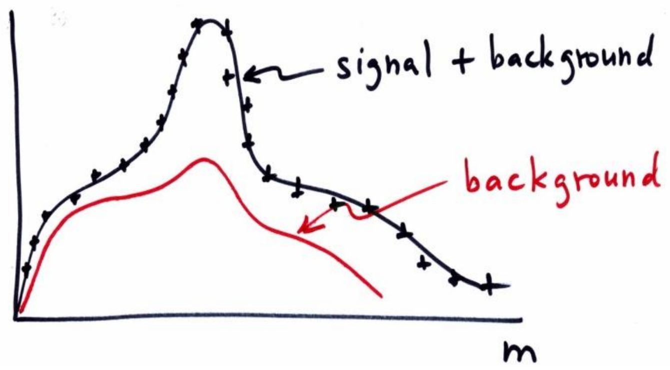
Difficulty is deciding how to introduce the additional parameters.

Example of inserting nuisance parameters

Fit of hadronic mass distribution from a specific τ decay mode.

Important uncertainty in background from non-signal τ modes.

Background rate from other measurements, shape from MC.



Want to include uncertainty in rate, mean, width of background component in a parametric fit of the mass distribution.

Number of events in bin i, $n_i \sim \text{Poisson}(s_i(\boldsymbol{\theta}) + b_i)$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(s_i(\boldsymbol{\theta}) + b_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta})+b_i)}$$

fit from MC

Step 1: uncertainty in rate

Scale the predicted background by a factor r: $b_i \rightarrow rb_i$

Uncertainty in r is σ_r

Regard $r_0 = 1$ ("best guess") as Gaussian (or not, as appropriate) distributed measurement centred about the true value r, which becomes a new "nuisance" parameter in the fit.

New likelihood function is:

$$L(\boldsymbol{\theta}, r) = \prod_{i=1}^{N} \frac{(s_i(\boldsymbol{\theta}) + rb_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta})+rb_i)} \frac{1}{\sqrt{2\pi}\sigma_r} e^{-(r-r_0)^2/2\sigma_r^2}$$

For a least-squares fit, equivalent to

$$\chi^2(\boldsymbol{\theta}) \rightarrow \chi^2(\boldsymbol{\theta}) + \frac{(r - r_0)^2}{\sigma_r^2}$$

Dealing with nuisance parameters

Ways to eliminate the nuisance parameter *r* from likelihood.

1) Profile likelihood:

$$L_{\rm p}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{\hat{r}}), \text{ where } \hat{\hat{r}} \text{ is value of } r \text{ that maximizes } L \text{ for the given } \boldsymbol{\theta}.$$

2) Bayesian marginal likelihood:

$$L_{\rm m}(\boldsymbol{\theta}) = \int \prod_{i=1}^{N} \frac{(s_i(\boldsymbol{\theta}) + r b_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + r b_i)} \frac{1}{\sqrt{2\pi}\sigma_r} e^{-(r - r_0)^2/2\sigma_r^2} \, dr$$

$L(n_1, \ldots, n_N | \boldsymbol{\theta}, r)$ $\pi(r)$ (prior)

Profile and marginal likelihoods usually very similar.

Both are broadened relative to original, reflecting the uncertainty connected with the nuisance parameter.

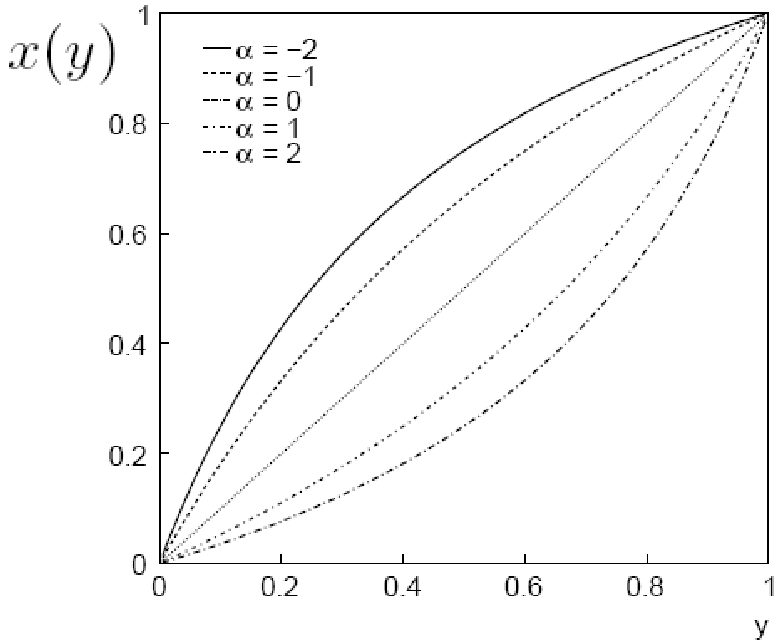
Step 2: uncertainty in shape

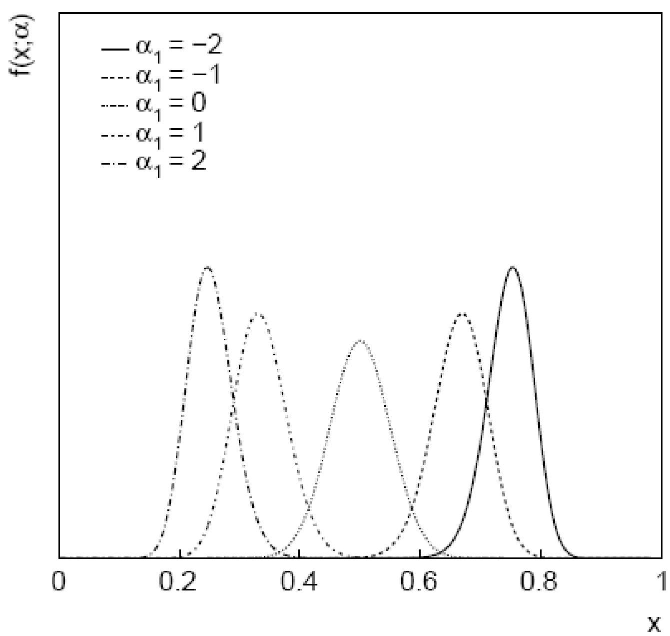
Key is to insert additional nuisance parameters into the model.

E.g. consider a distribution $g(y)$. Let $y \rightarrow x(y)$,

$$x(y) = \begin{cases} \frac{y}{1+\alpha(1-y)} & \alpha \geq 0 , \\ \frac{(1-\alpha)y}{1-\alpha y} & \alpha < 0 . \end{cases}$$

$$f(x) = g(y(x)) \left| \frac{dy}{dx} \right|$$



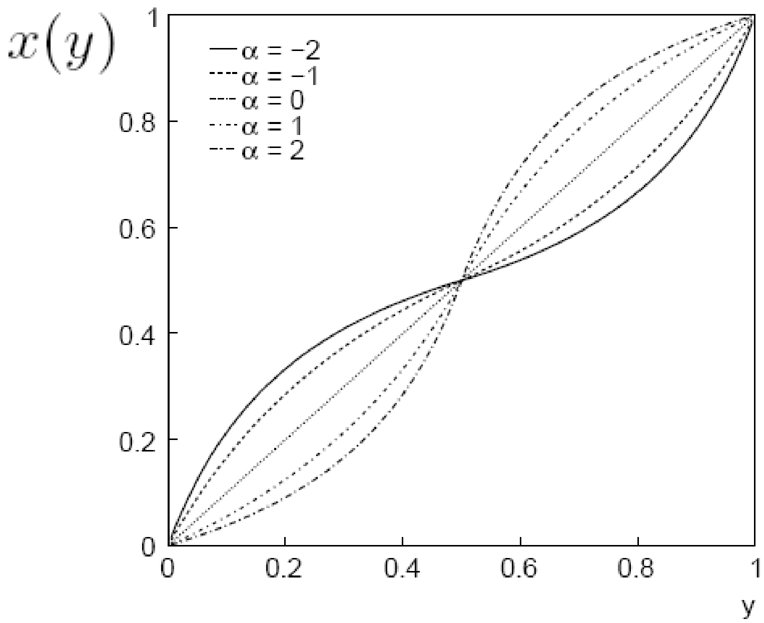


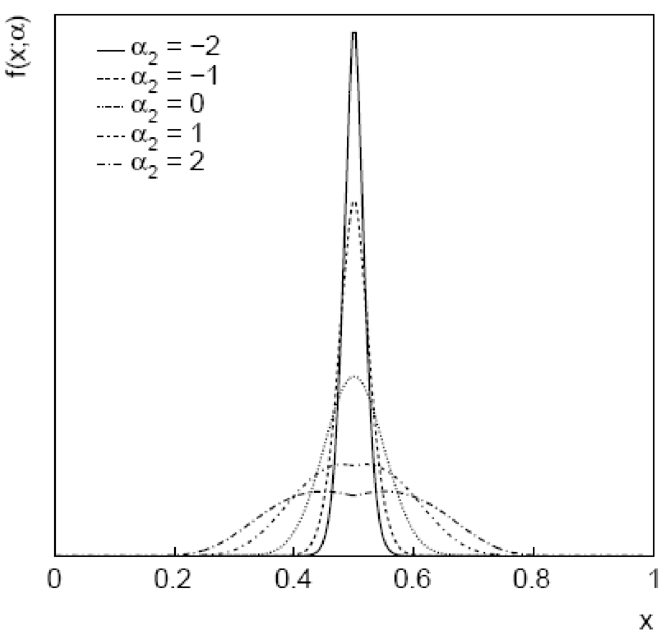
More uncertainty in shape

The transformation can be applied to a spline of original MC histogram (which has shape uncertainty).

Continuous parameter α shifts distribution right/left.

Can play similar game with width (or higher moments), e.g.,





A sample fit (no systematic error)

Consider a Gaussian signal, polynomial background, and also a peaking background whose form is take from MC:

True mean/width of signal:

$$\mu_{\mathrm{s}} = 0.5,\ \sigma_{\mathrm{s}} = 0.1$$

True mean/width of back-ground from MC:

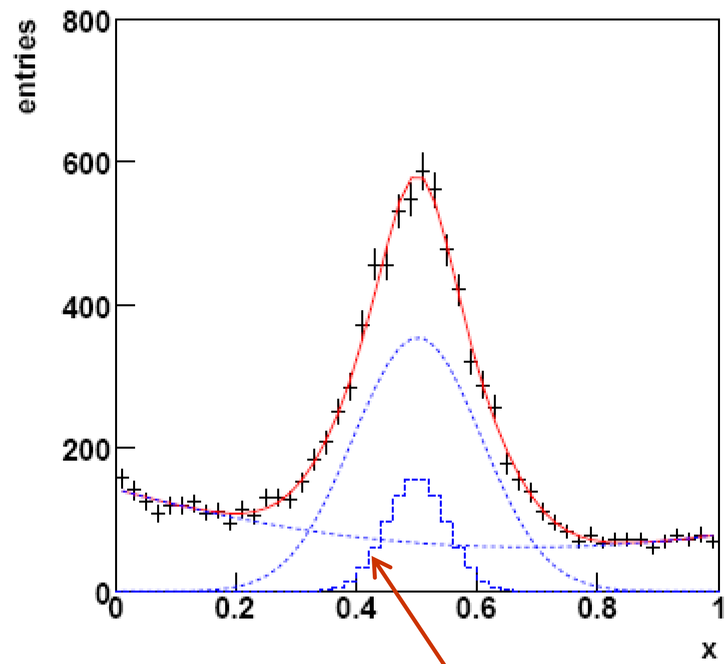
$$\mu_{\mathrm{b}} = 0.5,\ \sigma_{\mathrm{b}} = 0.05$$

Fit result:

$$\hat{\mu}_{\mathrm{s}} = 0.50025 \pm 0.00232$$

$$\hat{\sigma}_{\mathrm{s}} = 0.10578 \pm 0.00325$$

$$\chi^2 = 30.6 \text{ with } 44 \text{ degrees of freedom}$$



Template from MC

Sample fit with systematic error

Suppose now the MC template for the peaking background was systematically wrong, having

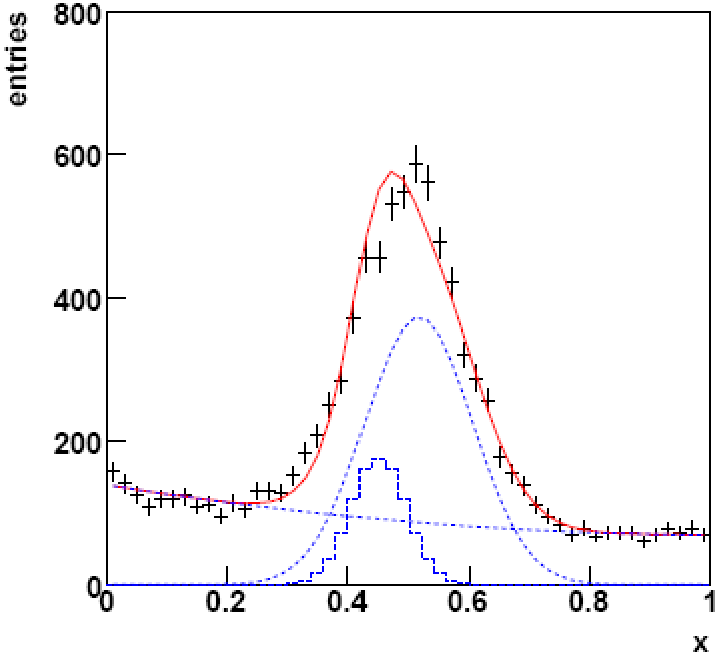
$$\mu_{\rm b} = 0.45, \ \sigma_{\rm b} = 0.045$$

Now fitted values of signal parameters wrong, poor goodness-of-fit:

$$\hat{\mu}_{\rm s} = 0.51676 \pm 0.00226$$

$$\hat{\sigma}_{\rm s} = 0.08933 \pm 0.00308$$

$\chi^2 = 91.2$ for 44 degrees of freedom



Sample fit with adjustable mean/width

Suppose one regards peak position and width of MC template to have systematic uncertainties:

$$\sigma_{\mu_{\rm b}} = 0.05 \qquad\qquad \sigma_{\sigma_{\rm b}} = 0.005$$

Incorporate this by regarding the nominal mean/width of the MC template as measurements, so in LS fit add to χ^2 a term:

altered mean
of MC template

orignal mean
of MC template

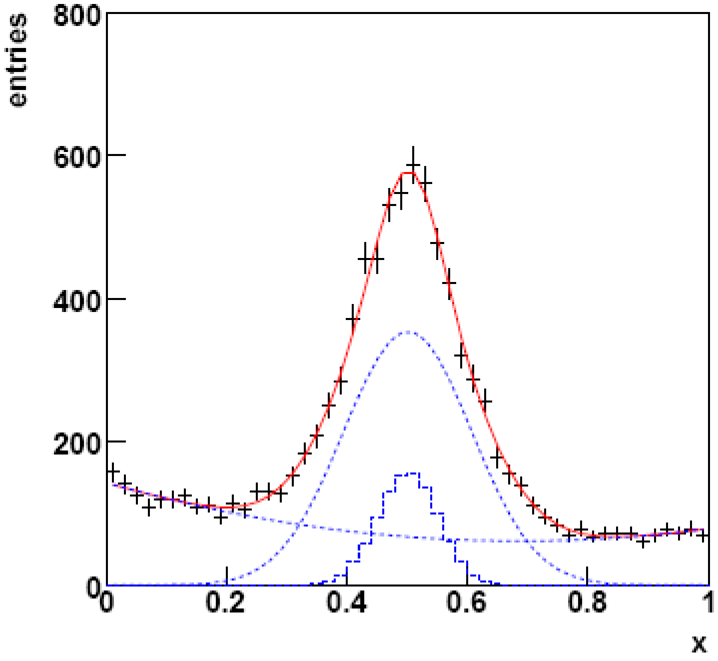
$$\left(\frac{\mu_{\rm b}(\boldsymbol{\alpha}) - \mu_{\rm b}(0)}{\sigma_{\mu_{\rm b}}} \right)^2 + \left(\frac{\sigma_{\rm b}(\boldsymbol{\alpha}) - \sigma_{\rm b}(0)}{\sigma_{\sigma_{\rm b}}} \right)^2$$

Sample fit with adjustable mean/width (II)

Result of fit is now "good":

$$\hat{\mu}_{\rm s} = 0.50014 \pm 0.00290$$
$$\hat{\sigma}_{\rm s} = 0.10582 \pm 0.00347$$

$\chi^2 = 32.1$ for 44 degrees of freedom



In principle, continue to add nuisance parameters until data are well described by the model.

Systematic error converted to statistical

One can regard the quadratic difference between the statistical errors with and without the additional nuisance parameters as the contribution from the systematic uncertainty in the MC template:

$$\sigma_{\hat{\mu},\mathrm{sys}} = \sqrt{0.00290^2 - 0.00226^2} = 0.00182$$

$$\sigma_{\hat{\sigma},\mathrm{sys}} = \sqrt{0.00347^2 - 0.00308^2} = 0.00160$$

Formally this part of error has been converted to part of statistical error (because the extended model is ~correct!).

Systematic error from "shift method"

Note that the systematic error regarded as part of the new statistical error (previous slide) is much smaller than the change one would find by simply "shifting" the templates plus/minus one standard deviation, holding them constant, and redoing the fit. This gives:

$$\Delta\hat{\mu}_{\rm sys} = |0.50025 - 0.51676| = 0.01651$$

$$\Delta\hat{\sigma}_{\rm sys} = |0.10578 - 0.08933| = 0.01645$$

This is not necessarily "wrong", since here we are not improving the model by including new parameters.

But in any case it's best to improve the model!

Issues with finding an improved model

Sometimes, e.g., if the data set is very large, the total χ^2 can be very high (bad), even though the absolute deviation between model and data may be small.

It may be that including additional parameters "spoils" the parameter of interest and/or leads to an unphysical fit result well before it succeeds in improving the overall goodness-of-fit.

Include new parameters in a clever (physically motivated, local) way, so that it affects only the required regions.

Use Bayesian approach -- assign priors to the new nuisance parameters that constrain them from moving too far (or use equivalent frequentist penalty terms in likelihood).

Unfortunately these solutions may not be practical and one may be forced to use ad hoc recipes (last resort).

Summary and conclusions

Key to covering a systematic uncertainty is to include the appropriate nuisance parameters, constrained by all available info.

Enlarge model so that for at least one point in its parameter space, its difference from the truth is negligible.

In frequentist approach can use profile likelihood (similar with integrated product of likelihood and prior -- not discussed today).

Too many nuisance parameters spoils information about parameter(s) of interest.

In Bayesian approach, need to assign priors to (all) parameters.

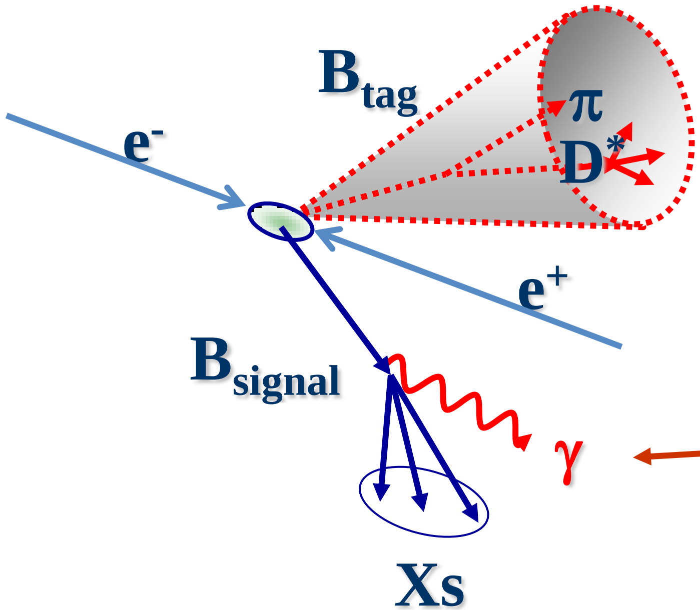
Can provide important flexibility over frequentist methods.
Can be difficult to encode uncertainty in priors.
Exploit recent progress in Bayesian computation (MCMC).

Finally, when the LHC announces a 5 sigma effect, it's important to know precisely what the "sigma" means.

Extra slides

Fit example: $b \to s\gamma$ (BaBar)

B. Aubert et al. (BaBar), Phys. Rev. D 77, 051103(R) (2008).



"recoil method"

high-energy γ

Decay of one B fully reconstructed (B_{tag}).

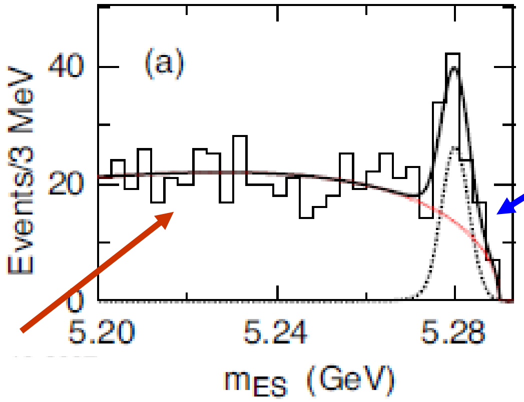
Look for high-energy γ from rest of event.

Signal and background yields from fit to m_{ES} in bins of E_γ.

$$m_{\mathrm{ES}} = \sqrt{E_{\mathrm{beam}}^{*2} - p_{\mathrm{tag}}^2} \quad (\approx m_{\mathrm{B}} \text{ for signal})$$

Fitting m_{ES} distribution for b → sγ

Fit m_{ES} distribution using superposition of Crystal Ball and Argus functions:



$$c(m;\alpha,\beta,\mu,\sigma) = \begin{cases} N e^{-(m-\mu)^2/2\sigma^2} & (m-\mu)/\sigma > -\alpha\,, \\ N\left(\frac{\beta}{|\alpha|} - |\alpha| - \frac{m-\mu}{\sigma}\right)^{-\beta}\left(\frac{\beta}{|\alpha|}\right)^{\beta} e^{-\alpha^2/2} & \text{otherwise.} \end{cases}$$

$$a(m;\xi) = \begin{cases} Nm\sqrt{1-\left(\frac{m}{m_{\max}}\right)^2}\,\exp\left[-\xi\left(1-\left(\frac{m}{m_{\max}}\right)^2\right)\right] & 0 < m \le m_{\max}\,, \\ 0 & \text{otherwise,} \end{cases}$$

log-likelihood: $$\ln L(\underbrace{\nu_c, \nu_a}_{\text{rates}}, \underbrace{\alpha,\beta,\mu,\sigma,\xi}_{\text{shapes}}) = \sum_{i=1}^{N} (n_i \ln \nu_i - \nu_i)$$

obs./pred. events in *i*th bin

Simultaneous fit of all m_{ES} distributions

Need fits of m_{ES} distributions in 14 bins of E_γ:

At high E_γ, not enough events to constrain shape, so combine all E_γ bins into global fit:

$$\ln L(\vec{\nu}_{\mathrm{c}}, \vec{\nu}_{\mathrm{a}}, \vec{\alpha}, \vec{\beta}, \vec{\mu}, \vec{\sigma}, \vec{\xi}) = \sum_{i=1}^{M} \ln L(\nu_{\mathrm{c},i}, \nu_{\mathrm{a},i}, \alpha_i, \beta_i, \mu_i, \sigma_i, \xi_i)$$

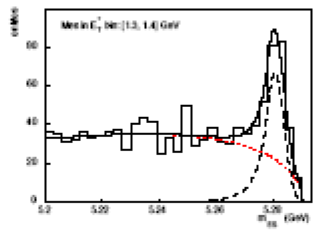
Shape parameters could vary (smoothly) with E_γ.

So make Ansatz for shape parameters such as

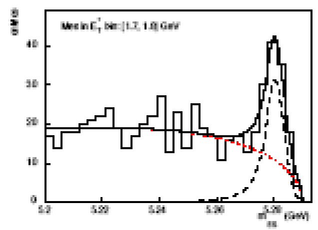
$$\alpha(E) = \alpha_0 + \alpha_1 E + \alpha_2 E^2 + \ldots$$

Start with no energy dependence, and include one by one more parameters until data well described.

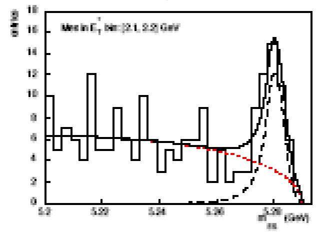
$1.3\ \mathrm{GeV} < E_\gamma < 1.4\ \mathrm{GeV}$



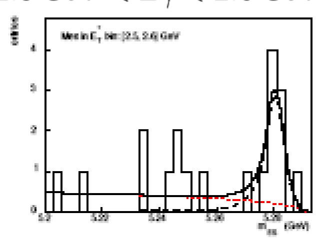
$1.7\ \mathrm{GeV} < E_\gamma < 1.8\ \mathrm{GeV}$



$2.1\ \mathrm{GeV} < E_\gamma < 2.2\ \mathrm{GeV}$



$2.5\ \mathrm{GeV} < E_\gamma < 2.6\ \mathrm{GeV}$



Finding appropriate model flexibility

Here for Argus ξ parameter, linear dependence gives significant improvement; fitted coefficient of linear term -10.7 ± 4.2.

	fit option	χ^2	degrees of freedom
(1)	no E dependence	389.70	387
(2)	linear for Argus ξ	386.22	386
(3)	quadratic for Argus ξ	385.61	385
(4)	linear for ξ and α	386.29	385
(5)	linear for ξ and σ	386.42	385
(6)	linear for ξ and μ	386.12	385
(7)	linear for ξ, α, σ, μ	385.59	383

D. Hopkins, PhD thesis, RHUL (2007).

$\chi^2(1) - \chi^2(2) = 3.48$
p-value of (1) = 0.062
$\rightarrow$data want extra par.

Inclusion of additional free parameters (e.g., quadratic E dependence for parameter ξ) do not bring significant improvement.

So including the additional energy dependence for the shape parameters converts the systematic uncertainty into a statistical uncertainty on the parameters of interest.