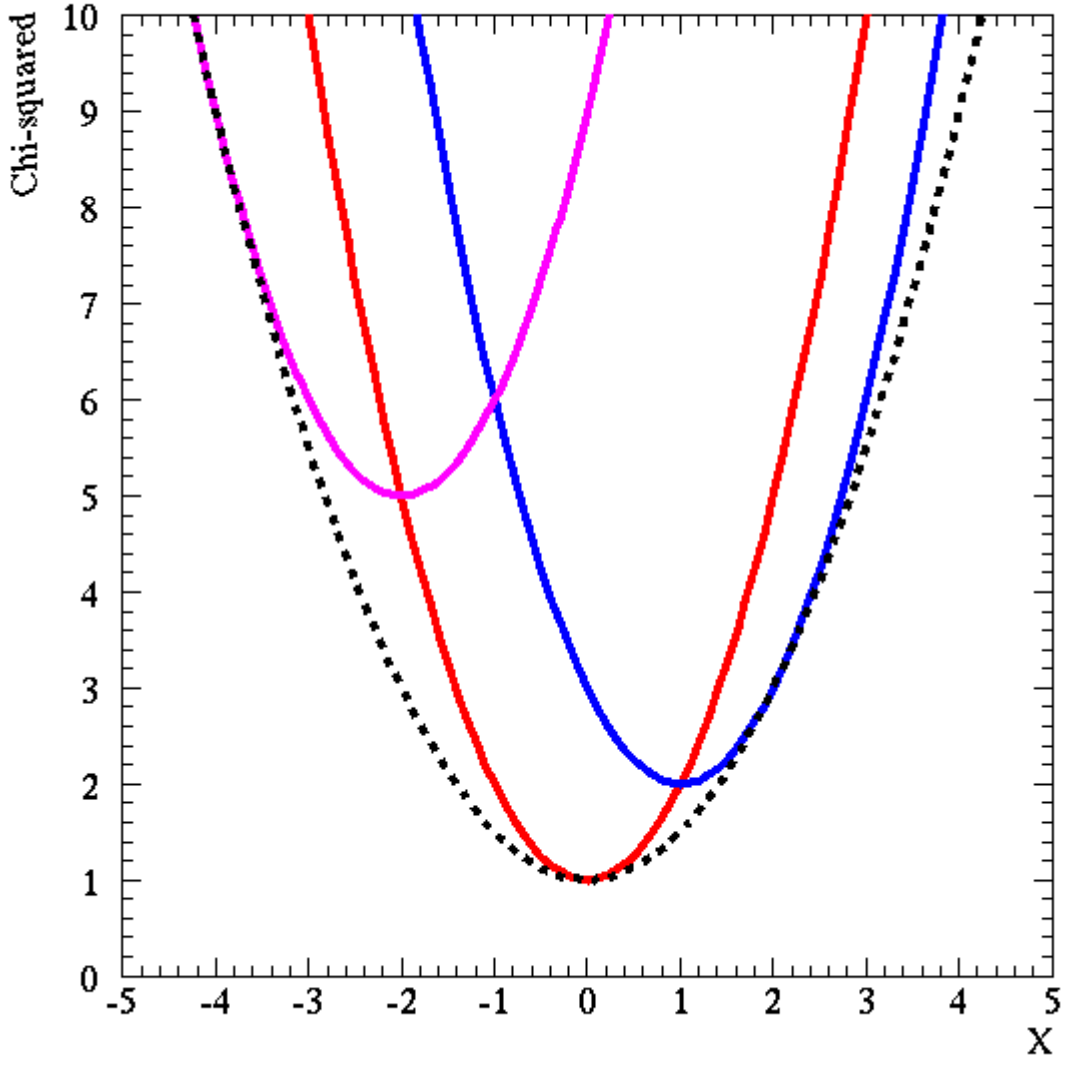
Physics 509: Propagating Systematic Uncertainties

Scott Oser
Lecture #12



Additive offset model

Suppose we take N measurements from a distribution, and wish to estimate the true mean of the underlying distribution.

Our measuring apparatus might have an offset *s* from 0. We attempt to calibrate this. Our systematic error model consists of:

1) There is some additive offset *s* whose value is unknown.
2) It affects each measurement identically by $x_i \rightarrow x_i + s$.
3) The true mean is estimated by:

$$\hat{\mu} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) - \hat{s}$$

4) Our calibration is $s = 2 \pm 0.4$

Covariance solution to additive offset model 1

We start by assembling the covariance matrix of the data:

$$\mathrm{cov}(x_i,x_j) = \delta_{ij}\sigma_i^2 + \sigma_s^2$$

$$V_{ij} = \begin{bmatrix} \sigma_1^2+\sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2+\sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_3^2+\sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_4^2+\sigma_s^2 \end{bmatrix}$$

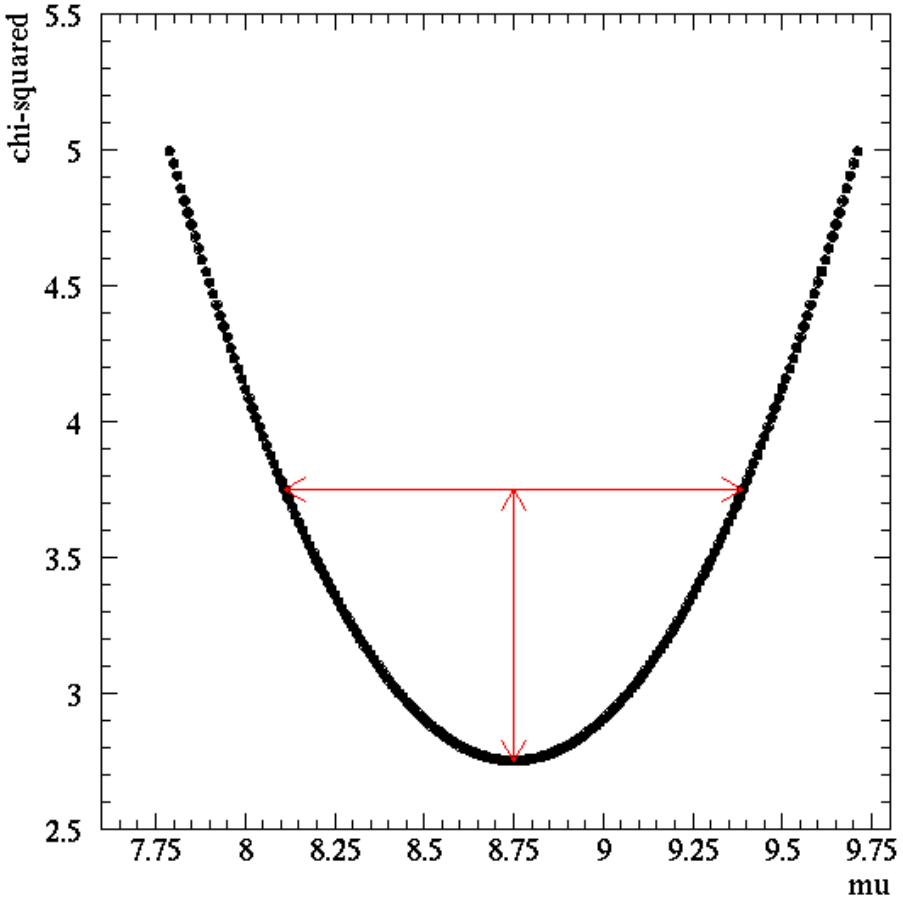
Next we form the likelihood (or χ^2) using this covariance matrix:

$$\chi^2(\mu) = \sum_i \sum_j (x_i - \mu - \hat{s}) V_{ij}^{-1} (x_j - \mu - \hat{s})$$

The ML (LS) estimator is that $\hat{\mu}$ which minimizes the χ^2.

Covariance solution to additive offset model 2

$$\chi^2(\mu) = \sum_i \sum_j (x_i - \mu - \hat{s}) V_{ij}^{-1} (x_j - \mu - \hat{s})$$



I calculated this for

x_i = (10,12,10,11)

σ_i = 1 σ_s = 0.4 $\hat{s} = 2$

Result = 8.75 ± 0.64

You've got to be kidding me ...

By now most of you must have realized that nobody calculates errors this way. The more usual approach is:

1) Calculate the central value and its statistical error:

$$\hat{\mu} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) - \hat{s} = 8.75 \pm \frac{1.0}{\sqrt{4}} = 8.75 \pm 0.5$$

2) Calculate the systematic error on the parameter due to the uncertainty in s. Since we said s=2.0±0.4, and we can think of $\hat{\mu} = \hat{\mu}(s|x_i)$ and so apply error propagation to get the systematic.

$$\hat{\mu} = 8.75 \pm 0.5 \pm 0.4 = 8.75 \pm \sqrt{0.5^2 + 0.4^2} = 8.75 \pm 0.64$$

Systematic error propagation by the "Δ" method

All of these suggests a simple way to propagate uncertainties for a parameter estimator. Let $\hat{g}=\hat{g}(x_i|s_1, s_2, ...)$ be a function of the data x_2 and a set of systematics parameters.

- Calculate the parameter estimate and its statistical uncertainty, keeping all of the systematics fixed, by the usual methods.
- Vary each systematic (nuisance parameter) one by one by ±1σ, and note the change in $\hat{g}$. In other words, calculate

$$\Delta\hat{g}_1=\hat{g}(x_i|s_1\pm\sigma_1, s_2, ...) - \hat{g}(x_i|s_1, s_2, ...)$$

- Combine the various uncertainties using error propagation:

$$g = \hat{g} + \Delta\hat{g}_{stat} + \Delta\hat{g}_1 + \Delta\hat{g}_2 + ...$$

- Finally, add up systematics in quadrature to get

$$\Delta\hat{g}_{sys}=\sqrt{(\Delta\hat{g}_1)^2+(\Delta\hat{g}_2)^2+...}$$

(If any systematics are correlated, include correlation coefficients as well.)

$$\Delta\hat{g}_{sys}=\sqrt{(\Delta\hat{g}_1)^2+(\Delta\hat{g}_2)^2+2\rho(\Delta\hat{g}_1)(\Delta\hat{g}_2)...}$$

An example of the Δ method: linear fit with distorting systematics

Suppose our model predicts y=mx+b, and we wish to estimate m and b. Each measurement has dy=1. For a particular set of data:

Best fit: m = 0.46±0.12, b = 4.60±0.93 (statistical errors only)

Now suppose there is a systematic bias in the measurement of y given by $\Delta y = ax + cx^2$. We believe a=0.00±0.05, and c=0.00±0.01

We make a spreadsheet:

a	c	m	b	Δm	Δb
0	0	0.464	4.602	0.000	0.000
0.05	0	0.414	4.602	-0.050	0.000
-0.05	0	0.514	4.602	0.050	0.000
0	0.01	0.604	4.228	0.140	-0.374
0	-0.01	0.324	4.975	-0.140	0.373

An example of the Δ method: linear fit with distorting systematics

a	c	m	b	Δm	Δb
0	0	0.464	4.602	0.000	0.000
0.05	0	0.414	4.602	-0.050	0.000
-0.05	0	0.514	4.602	0.050	0.000
0	0.01	0.604	4.228	0.140	-0.374
0	-0.01	0.324	4.975	-0.140	0.373

Since our knowledge of the nuisance parameters a and c are independent, we treat the systematics as uncorrelated and add them in quadrature:

Best fit: $m = 0.46 \pm 0.12(\text{stat}) \pm 0.15(\text{sys}) = 0.46 \pm 0.19$

$b = 4.60 \pm 0.93(\text{stat}) \pm 0.37(\text{sys}) = 4.60 \pm 1.00$

Pull method: linear fit with distorting systematics

$$\chi^2(m,b,a,c)=\sum_i \left(\frac{y_i-mx_i-b-ax_i-cx_i^2}{1.0}\right)^2+\left(\frac{a-0}{0.05}\right)^2+\left(\frac{c-0}{0.01}\right)^2$$

Find confidence intervals on m, b by marginalizing over all other parameters

Both a and c fixed:
m=0.46 ± 0.12
b= 4.64 ± 0.93

Minimizing over both a and c:
m=0.45 ± 0.19
b= 4.65 ± 0.99

Minimizing over a only (c fixed):
m=0.45 ± 0.13
b= 4.65 ± 0.93

Minimizing over c only (a fixed):
m=0.45 ± 0.18
b= 4.65 ± 0.99

To get "systematic error", subtract statistical error from total error in quadrature:

$$m=0.45\pm 0.12\,(stat)\pm\sqrt{0.19^2-0.12^2}\,(sys)=0.45\pm 0.12\pm 0.15$$

$$b=4.65\pm 0.93\,(stat)\pm\sqrt{0.99^2-0.93^2}\,(sys)=0.45\pm 0.93\pm 0.34$$

Residuals of the nuisance parameters

The Δ method gave the same result as the pull method in this example.

It's instructive to look at the values of the nuisance parameters at the best fit point (floating both a and c):

Best fit: a = 0, c=0.0009

Recall that our constraint term specified a=0±0.05 and c=0±0.01

The fact that the nuisance parameter values at the best fit are consistent with the prior constraints indicates that the constraints on the nuisance parameters imposed by the data itself are consistent with those from our prior calibration.

If they were inconsistent (eg. if the best fit were c=0.05), that would tell us either that the calibration of c was wrong, that there is a problem with the data, that the model itself is wrong, or that we were unlucky.

Note: we do not expect the best fit values of the nuisance parameters to show much scatter (i.e. if we did the experiment many times, we would not find that a was scattered with an RMS of 0.05, even though that's our uncertainty on a.

What if the systematic uncertainties are larger?

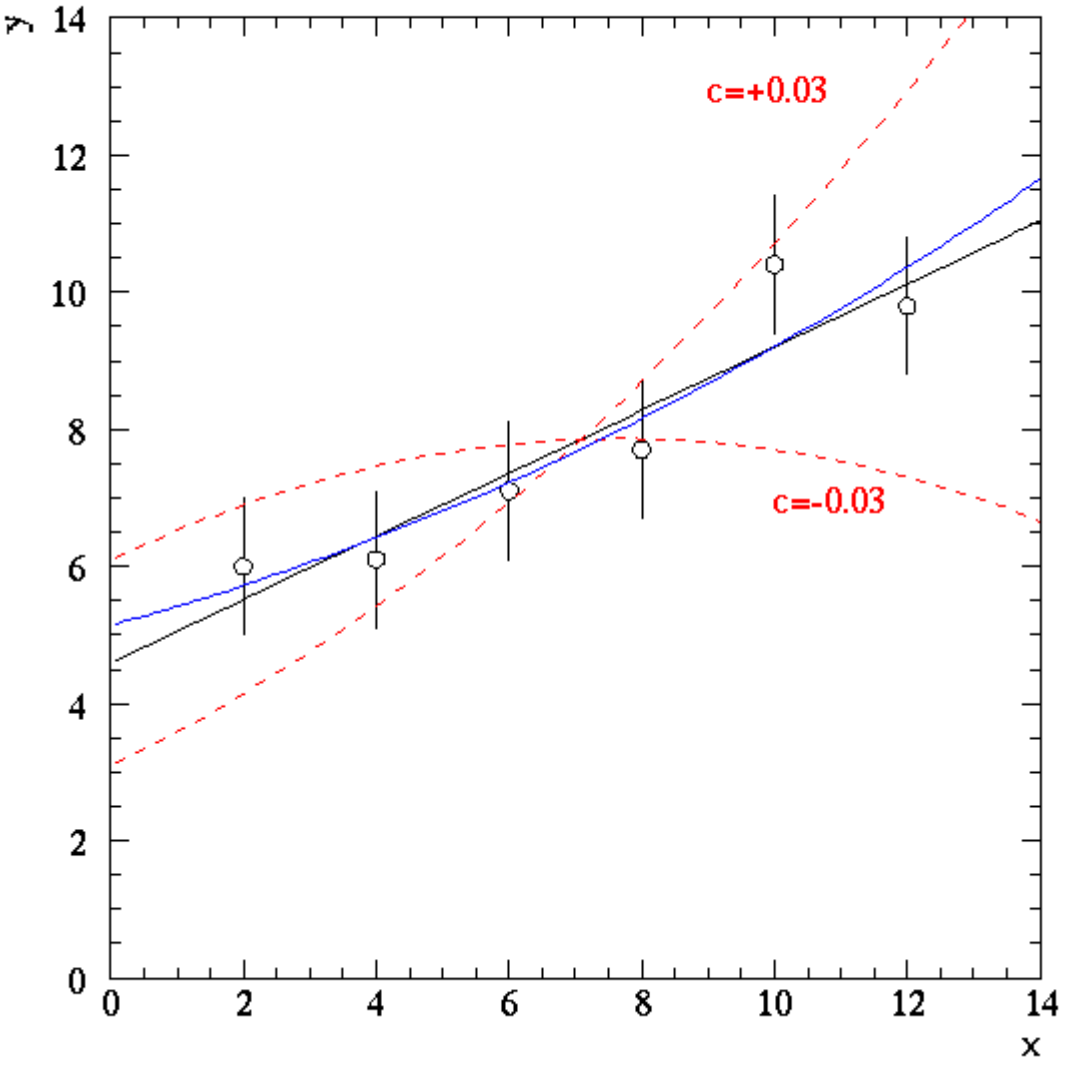
a	c	m	b	Δm	Δb
0	0	0.464	4.602	0.000	0.000
0.05	0	0.414	4.602	-0.050	0.000
-0.05	0	0.514	4.602	0.050	0.000
0	0.03	0.044	5.722	-0.420	1.120
0	-0.03	0.884	3.482	0.420	-1.120

Δ method:
m = 0.46 ± 0.12(stat) ± 0.42(sys) = 0.46 ± 0.44
b = 4.60 ± 0.93(stat) ± 1.12(sys) = 4.60 ± 1.46

Pull method:
m = 0.39 ± 0.36
b = 4.79 ± 1.29

How come the two methods gave identical results before, and now they don't?

A closer look at the pull method



Prior range on c was 0±0.03

However, the extreme ranges clearly don't fit the data (see red curves). We *know* c<0.03! Δ method ignores this fact.

In the pull method, the penalty term keeps c from straying too far from its prior range, but also lets the data keep c reasonable. Values of c which are terrible fits to data give large χ^2.

Advantages/disadvantages of the pull method

The pull method resulted in a reduced error, because it used the data itself as an additional calibration of the systematic.

Of course, this “calibration” is dependent on the correctness of the model you're fitting---fit the wrong model to the data, and you'll get spurious values for the nuisance parameters.

The pull method is easy to code up, but can be computationally intensive if you have many nuisance parameters---you have to minimize the likelihood over a multi-dim parameter space.

Also, if the fitted value of the nuisance parameter is very different from the prior value, stop and try to understand why---that's a clue that something is messed up in your analysis!

What if you don't care about using the data itself to improve your systematic uncertainty? Is it OK to just use the Δ method?

A morality tale about systematics

Suppose there exist two classes of events A and B. For each event we measure some property E, which can have two values (if you like, imagine that we measure E and then bin it into two bins). The two kinds of events have different distributions in the variable E:

A events: $P(E=1|A) = \alpha$
$P(E=2|A) = 1-\alpha$
B events: $P(E=1|B) = 0.5$
$P(E=2|B) = 0.5$

We don't know α perfectly, but estimate $\alpha = 0.80 \pm 0.02$.

We observe 200 events in total in one hour, and measure E for each. We want to estimate the mean rates of A events and B events by fitting the E distribution.

Setting up the ML fit

This problem is perfectly suited for the extended maximum likelihood fit:

$$-\ln L(\phi_A,\phi_B)=-(\phi_A+\phi_B)+\sum_{i=1}^{200}\ln(\phi_A P(E_i|A)+\phi_B P(E_i|B))$$

The ML parameter estimates for my data set are

ϕ_A = 100.0 ± 23.5 (stat)
ϕ_B = 100.0 ± 23.5 (stat)

To get the systematic uncertainty, I use the Δ method, and so repeat the fit for α=0.8±0.02.

$$\phi_A=100.0\pm 23.5\,(stat)\ {}^{+7.1}_{-6.3}(sys)\approx 100.0\pm 24.4$$
$$\phi_B=100.0\pm 23.5\,(stat)\ {}^{+6.3}_{-7.1}(sys)\approx 100.0\pm 24.4$$

Adding a second observable

Now suppose that in addition to measuring E for each event, we measure the value of a second parameter θ that also provides some A/B discrimination.

A events: $P(\theta=1|A) = \beta$
$P(\theta=2|A) = 1-\beta$
B events: $P(\theta=1|B) = 0.5$
$P(\theta=2|B) = 0.5$

Our calibration data tells us that $\beta=0.8 \pm 0.15$.

Notice that both θ and E provide basically the same discrimination power between A and B events, since the PDFs have the same form. But since they're independent measurements, we can improve our event separation by using both.

2D PDFs

We form 2D PDFs for A and B as a function of the discriminating variables θ and E.

A events: $P(E=1,\theta=1|A) = \alpha\beta = 0.64$

$P(E=1,\theta=2|A) = \alpha(1-\beta) = 0.16$

$P(E=2,\theta=1|A) = (1-\alpha)\beta = 0.16$

$P(E=2,\theta=2|A) = (1-\alpha)(1-\beta) = 0.04$

(This assumes that E and θ are independent.)

For events of type B, which have flat distributions in both variables, the probability for each of the four combinations is 0.25.

The log likelihood is just as before, but generalized to 2D PDFs

$$-\ln L(\phi_A,\phi_B) = -(\phi_A+\phi_B) + \sum_{i=1}^{200} \ln\left(\phi_A P(E_i,\theta_i|A) + \phi_B P(E_i,\theta_i|B)\right)$$

Event rates using both E and θ

When I fit the data set using both variables to improve the discrimination, I get:

$$\phi_A = 100.0 \pm 18.0 \text{ (stat)}$$
$$\phi_B = 100.0 \pm 18.0 \text{ (stat)}$$

Recall that the statistical error was 23.5 when using just E in the fit. Including the additional information provided further separation between the two classes of events.

(Side note: if I could somehow figure out how to perfectly separate the two kinds of events, the error bars would reduce to √N---this would just become a counting experiment! The fact that the errors are larger than √N reflects the fact that the separation between A and B is imperfect, and in fact the two rate estimates are negatively correlated.)

Systematics when using both E and θ

Now I'll propagate the systematics. I have uncertainties on both α and β. My calibrations of these quantities are independent, so we can safely take them to be independent. Therefore I propagate the systematics using the Δ method with error propagation:

$$(d\,\phi_A)^2 = \left|\frac{\partial \phi_A}{\partial \alpha}\right|^2 (d\,\alpha)^2 + \left|\frac{\partial \phi_A}{\partial \beta}\right|^2 (d\,\beta)^2$$

In practice I just vary α and β separately by their $\pm 1\sigma$ ranges, and add the results in quadrature. My final answer is:

$$\phi_A = 100.0 \pm 18.0\,(stat)\ {}^{+17.4}_{-26.7}\,(sys) \approx 100.0^{+25.0}_{-32.2}$$

$$\phi_B = 100.0 \pm 18.0\,(stat)\ {}^{+26.7}_{-17.4}\,(sys) \approx 100.0^{+32.2}_{-25.0}$$

This is nonsense!

Let's compare the error bars on ϕ_A for the case when I use only E in the fit, and for the case when I use both E and θ.

$$\text{E only:} \quad \phi_A = 100.0 \pm 23.5\,(stat)\ {}^{+7.1}_{-6.3}\,(sys) \approx 100.0 \pm 24.4$$

$$\text{E and } \theta: \quad \phi_A = 100.0 \pm 18.0\,(stat)\ {}^{+17.4}_{-26.7}\,(sys) \approx 100.0^{+25.0}_{-32.2}$$

Overall the uncertainty is larger when we include both E and θ than when we use E alone.

But this makes no sense---E and θ are independent and consistent pieces of information. I get the same flux values fitting with either, and comparable uncertainties. It's really like having two independent measures of ϕ_A and ϕ_B---the combination of the two should be more powerful, and should give smaller uncertainties, than either alone.

Nonsense revealed

The statistical errors did in fact get smaller when we use both E and θ. The problem is evidently in the systematic uncertainties.

$$\text{E only:} \quad \phi_A = 100.0 \pm 23.5\,(stat)\ {}^{+7.1}_{-6.3}\,(sys) \approx 100.0 \pm 24.4$$

$$\text{E and } \theta: \quad \phi_A = 100.0 \pm 18.0\,(stat)\ {}^{+17.4}_{-26.7}\,(sys) \approx 100.0^{+25.0}_{-32.2}$$

Imagine that our uncertainty on β were infinite---in that case we have infinite uncertainty on the θ PDF, and gain no information from including θ as a variable in the fit. (Adding infinitely uncertain information is the same as adding no information at all---the error bar shouldn't increase, it just doesn't get any smaller.)

We have a logical paradox---as long as the uncertainty on β is finite, then using as a θ as a variable in the ML fit should only improve the separation between events of type A and B, and the errors should get smaller.

What does the "floating systematic" approach yield?

The pull method is straightforward to implement:

$$-\ln L(\phi_A,\phi_B,\alpha,\beta)=(\phi_A+\phi_B)-\sum_{i=1}^{200}\ln(\phi_A P(E_i,\theta_i,\alpha,\beta|A)+\phi_B P(E_i,\theta_i|B))+\frac{1}{2}\left(\frac{\alpha-0.8}{0.02}\right)^2+\frac{1}{2}\left(\frac{\beta-0.8}{0.15}\right)^2$$

The hardest part is simply setting up the minimization over 4D and getting it to converge on the minimum, but even that's straightforward.

The best fit result:

ϕ_A = 100.0 ± 22.7
ϕ_B = 100.0 ± 22.7
α = 0.800 ± 0.0197
β = 0.800 ± 0.074

Let us contemplate this upon the Tree of Woe ...

Contemplation of the pull result

The best fit result was:

$\phi_A = 100.0 \pm 22.7$
$\phi_B = 100.0 \pm 22.7$
$\alpha = 0.800 \pm 0.0197$
$\beta = 0.800 \pm 0.074$

Note that the overall uncertainty on the rates is smaller than that obtained when fitting only with E (±23.5).

Note as well that the uncertainties on α and β are smaller than our prior estimates ($\alpha = 0.80 \pm 0.02$ and $\beta = 0.80 \pm 0.15$), especially the estimate on β.

The result from the floating systematics fit avoids the mathematical paradox we say in the Δ method of propagating systematics!

Why does the pull method work?

The pull method puts nuisance parameters on the same footing as other parameters---all are treated as unknowns. The penalty function ("constraint term") is none other than a frequentist version of the Bayesian prior on the nuisance parameter.

In our example, we had a tight constraint on α $(=0.80\pm0.02)$ and a weaker constraint on β $(=0.80\pm0.15)$. In principle both have equal discriminatory power between A and B events, except for the difference in systematics. Using E, which depends on α, gives a good constraint on the rates ϕ_A and ϕ_B. These in term can be turned back around to give a tighter constraint on β than we had *a priori*.

In this particular example it was important that the rates depended on both α and β---through their effects on the rates they constrain each other. If we had just one or the other, we wouldn't get any benefit from the pull method.

Effectively there's a correlation between the nuisance parameters

Even though our prior estimates of α and β were uncorrelated, the fit procedure itself introduces an effective correlation between the two.
(From the fit, cov(α,β)=0.144).

The data itself is saying that certain combinations of the two parameters that *a priori* you would have accepted are not good fits to the data, and should be rejected. For example, if α is on the large side, then β cannot be too small while still fitting the data. The data effectively prunes the range of reasonable α and β.

In contrast, the Δ method of propagating systematics gives an incorrect, paradoxical answer.

Conclusion: use the pull method!

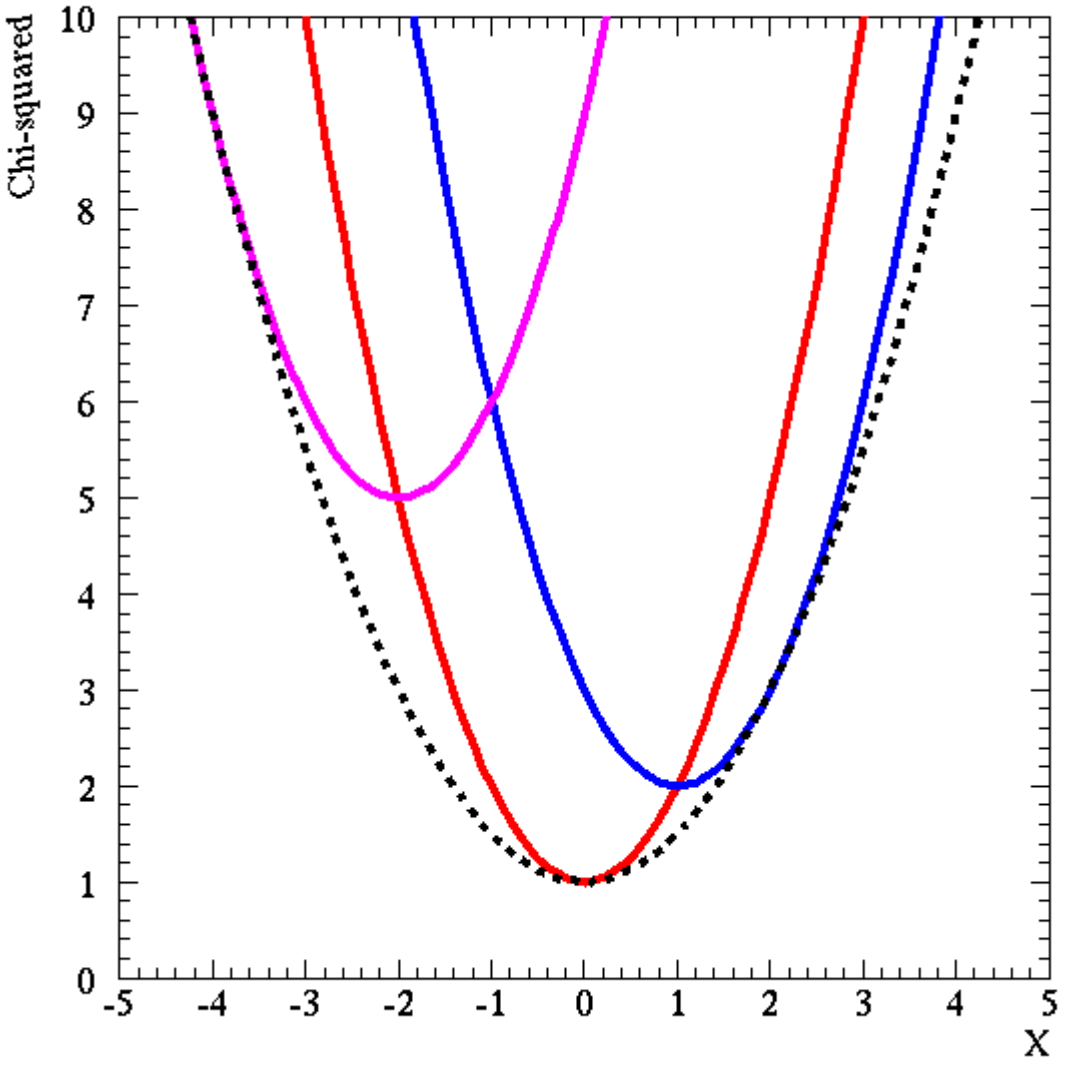
The pull method provides a self-consistent treatment of the data and uncertainties. Other methods do not, and will produce wrong results. You have two choices:

- add constraint terms directly to the likelihood, or
- calculate the covariance matrix between all the data points given the systematic uncertainties, and include that covariance matrix in the likelihood or χ^2

Generally either approach will give the correct answer, although the first is usually easier, and more obviously yields constraints on the systematics themselves. I recommend that approach.

Do NOT use the Δ method of just varying the systematics one at a time and adding up all of the resulting variations, in quadrature or otherwise. It might work in certain limited circumstances, but in others it is grossly incorrect.

A graphical interpretation



Consider a fit for one parameter X with one systematic whose nuisance parameter we'll call c.

The red line represents the χ^2 for the best fit value of c (including its constraint term).

The blue and purple curves are the χ^2 vs X curves for two other values of c that don't fit as well. The curves are shifted both up and sideways.

The black curve is the envelope curve that touches the family of χ^2 curves for all different possible values of c. It is the global χ^2 curve after marginalizing over c, and its width gives the total uncertainty on X (stat+sys).

A simple recipe that usually will work

1) Build a quantitative model of how your likelihood function depends on the nuisance parameters.

2) Form a joint negative log likelihood that includes both terms for the data vs. model and for the prior on the nuisance parameter.

3) Treat the joint likelihood as a multidimensional function of both physics parameters and nuisance parameters, treating these equally.

4) Minimize the likelihood with respect to all parameters to get the best-fit.

5) The error matrix for all parameters is given by inverting the matrix of partial derivatives with respect to all parameters:

$$V = \left(\frac{-\partial^2 \ln L(\vec{\theta})}{\partial \theta_i \partial \theta_j} \right)^{-1}$$