Smooth, parametric form of the QCD CR systematic

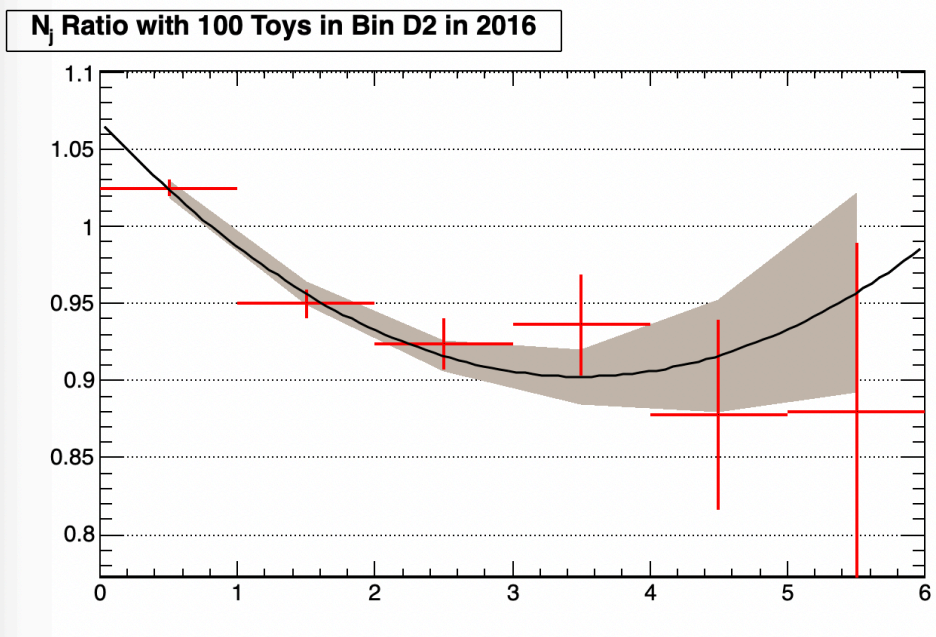
Owen Long, UC Riverside

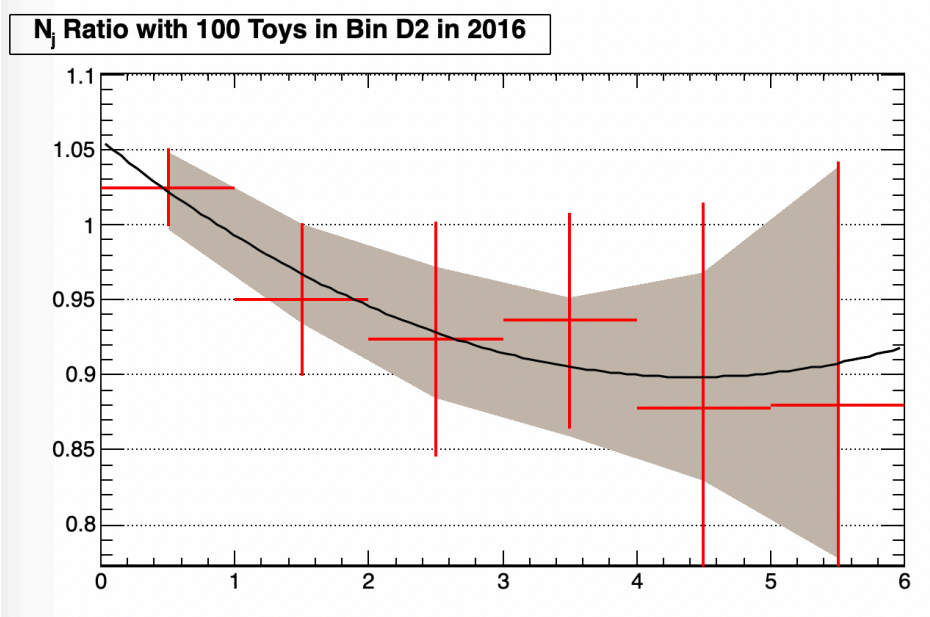
June 27, 2019

Issues with the current fit

- The uncorrelated bin-by-bin uncertainties in the QCD CR systematic make the Poisson pulls too small.
  - We need a plot of the fit in the paper. We can't currently make a normal plot.
  - If we made a histogram of the point pull values, it would not have a mean of zero and a width of one. The width would be much smaller than 1.
- The shape of the QCD CR systematic is not smooth
  - It includes statistical fluctuations present in the evaluation of the shape.
  - It can be smoothed with the bin-by-bin adjustments, but that's probably expensive in terms of delta log likelihood.
- A possible solution is to fit the shape histogram and make the shape flexible using the fit uncertainty.
  - We decided a 2^{nd}-order polynomial works pretty well.
  - A smooth function will probably solve the too-small-pulls problem.

The 2nd-order polynomial fit

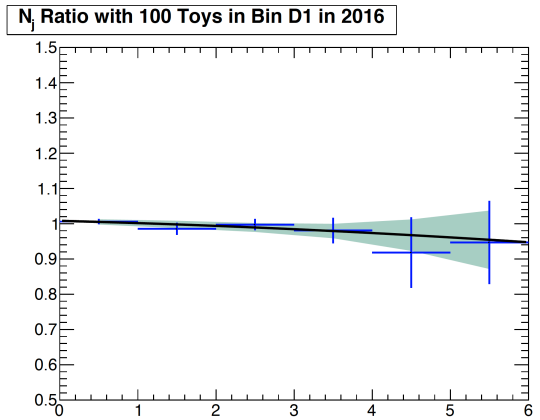


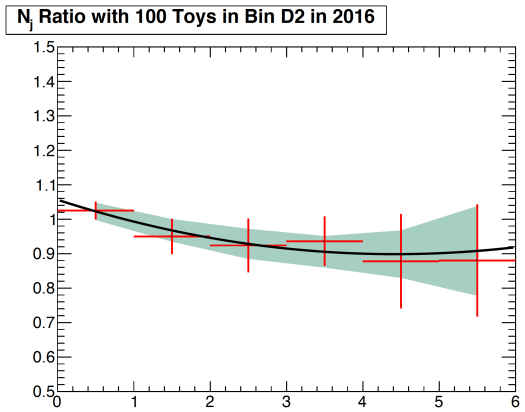


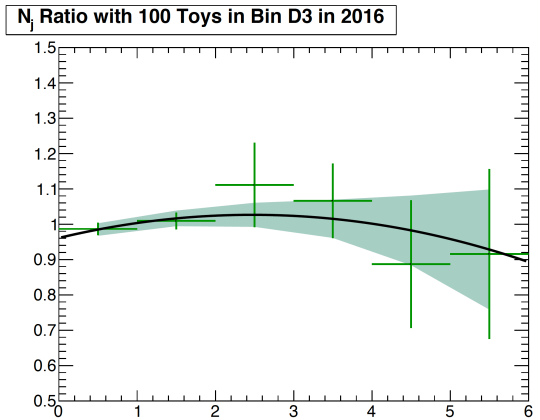
Error bars are only the statistical errors from the weights derived in the comparison of the QCD CR and the ttbar MC.

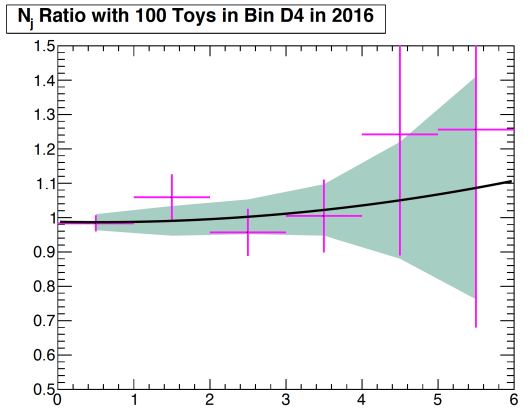
Errors include the full deviation from 1 in quadrature with the stat errors.

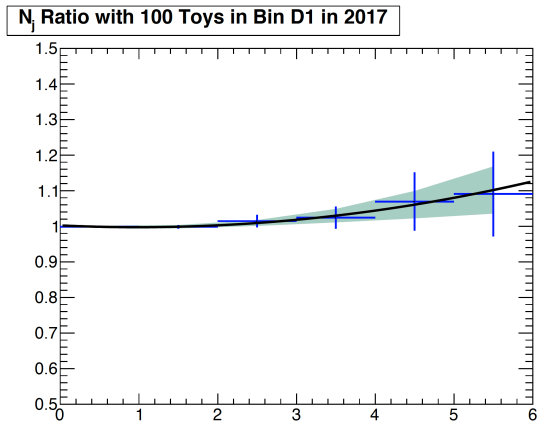
More fit examples

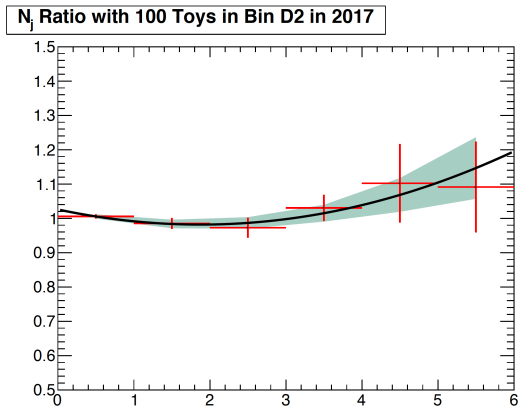


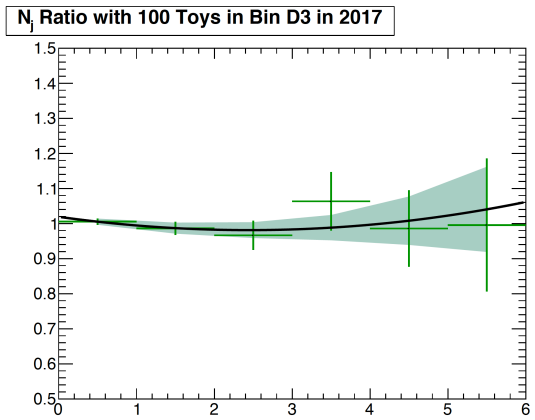


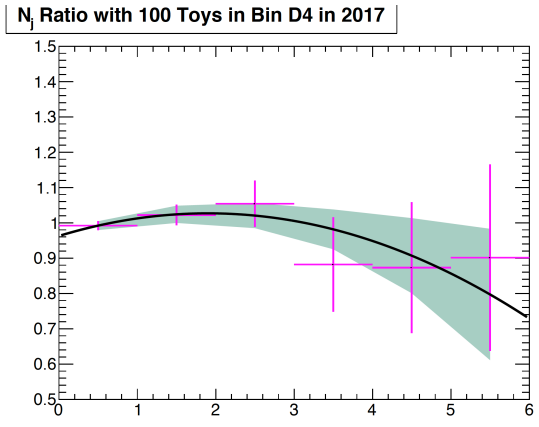






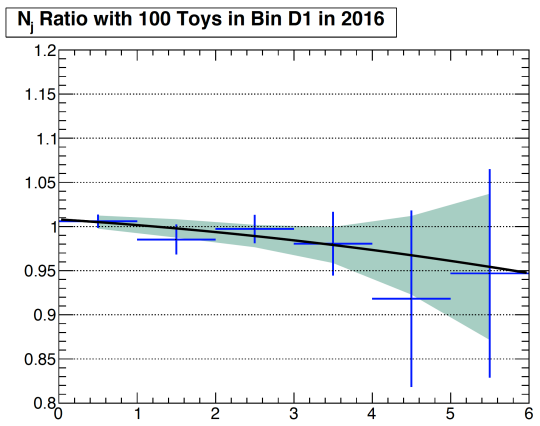


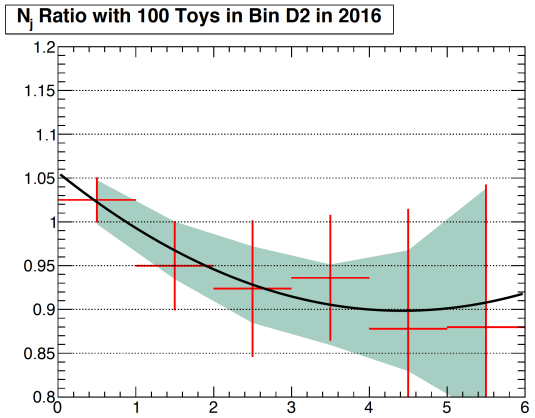


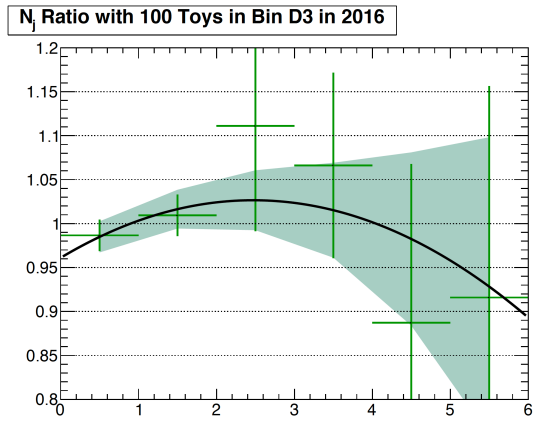


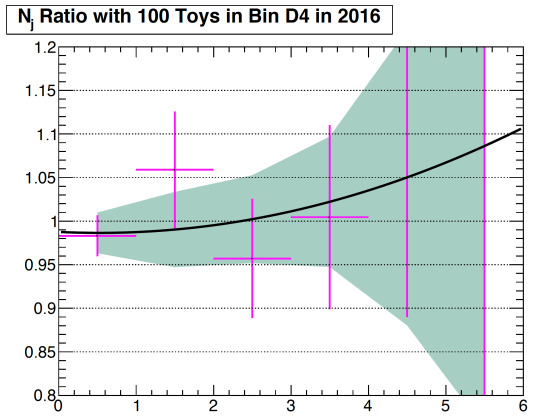
These include the full deviation from one in the uncertainties.

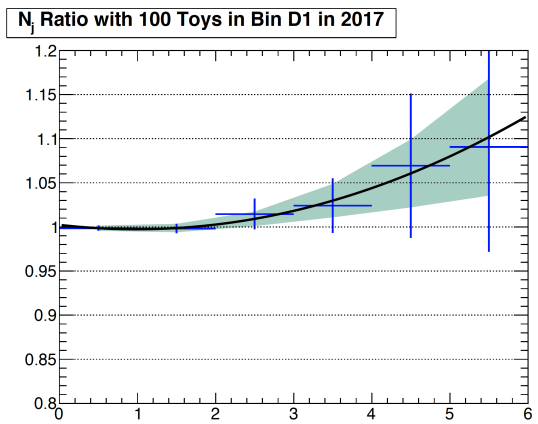
More fit examples

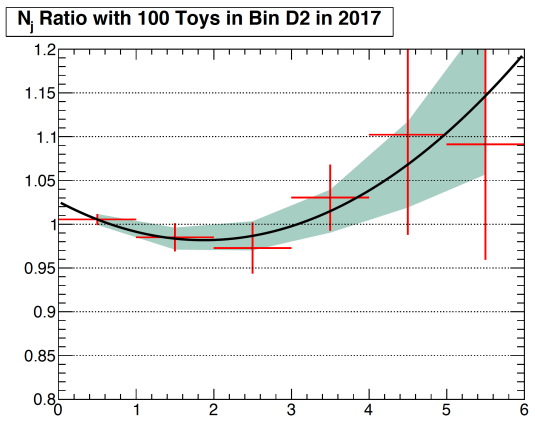


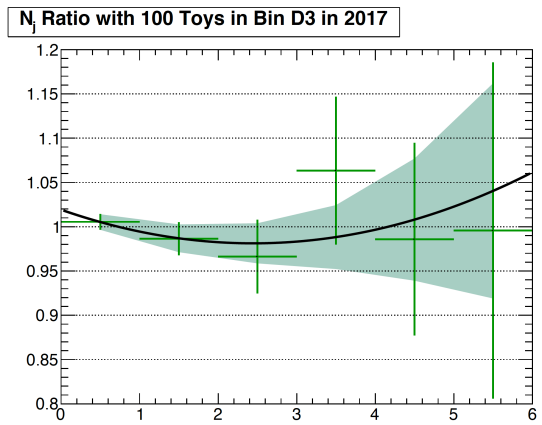


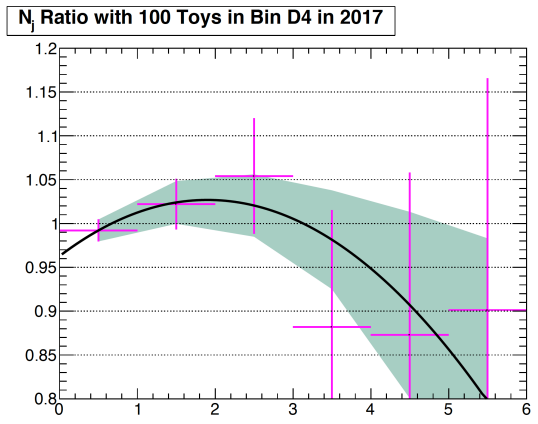






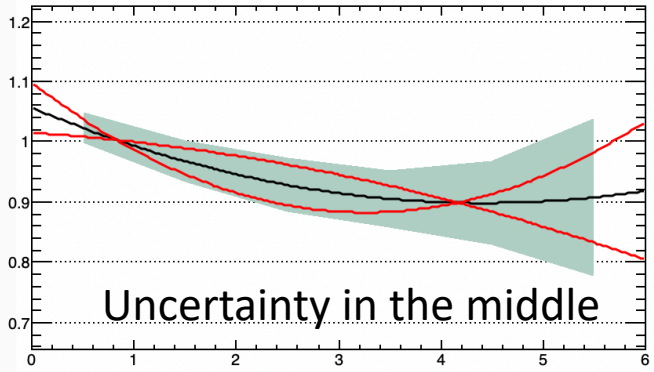


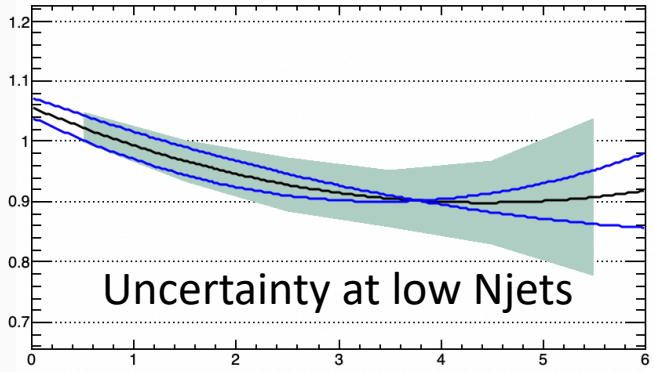


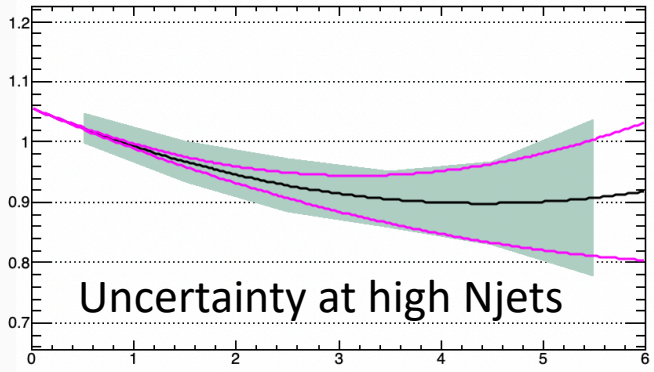


These include the full deviation from one in the uncertainties.

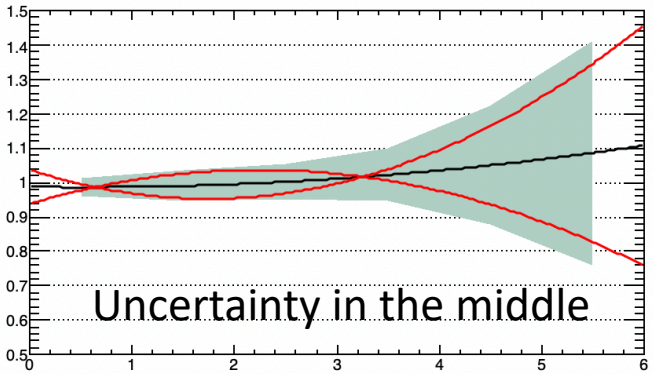
2016, D2

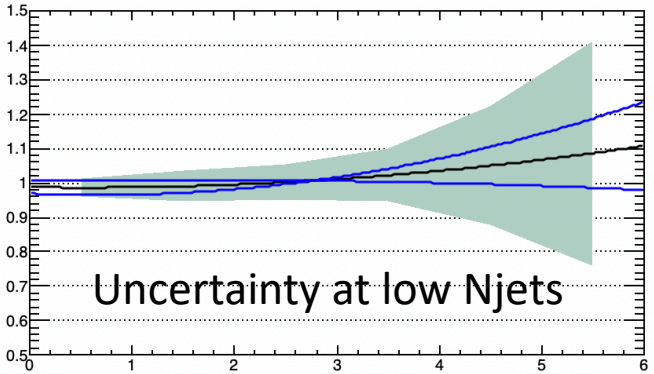


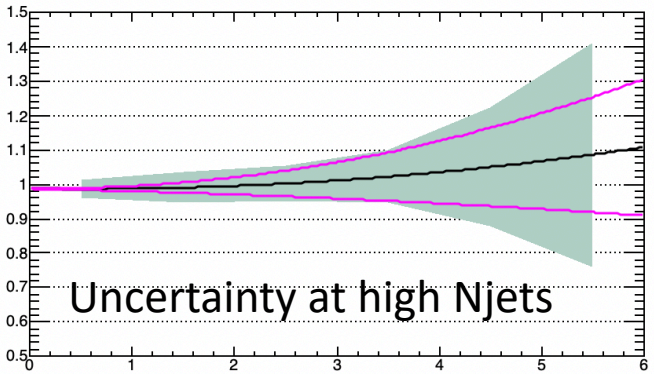




2016, D4







What the uncorrelated parameters do

The uncorrelated parameters control the uncertainty in specific ways, which will be helpful when interpreting the post-fit nuisance pulls.

Colored graphs show the curve after changing one of the uncorrelated parameters by +/- 1 sigma.

The parameters are sorted by the size of the eigenvalue, which is the square of the uncertainty on the uncorrelated parameter. Top is largest, bottom is smallest.

How we put this into Combine

The p0, p1, p2 parameters from the fit are correlated, so we can't use those directly as constrained nuisance parameters.

We can diagonalize the covariance matrix to transform to a parameter space where the three parameters are uncorrelated. The transformation matrix is composed of the normalized eigenvectors of the covariance matrix.

(transformation matrix) (p0, p1, p2) → (p0', p1', p2')

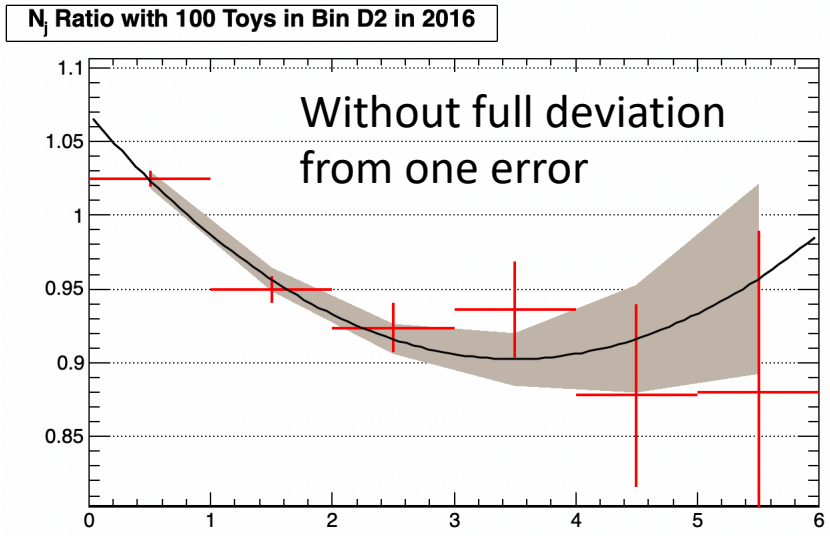
We create, for each fit, three uncorrelated, Gaussian constrained, nuisance parameters (p0', p1', p2'). The uncertainties on these parameters are the square root of the eigenvalues of the cov matrix.

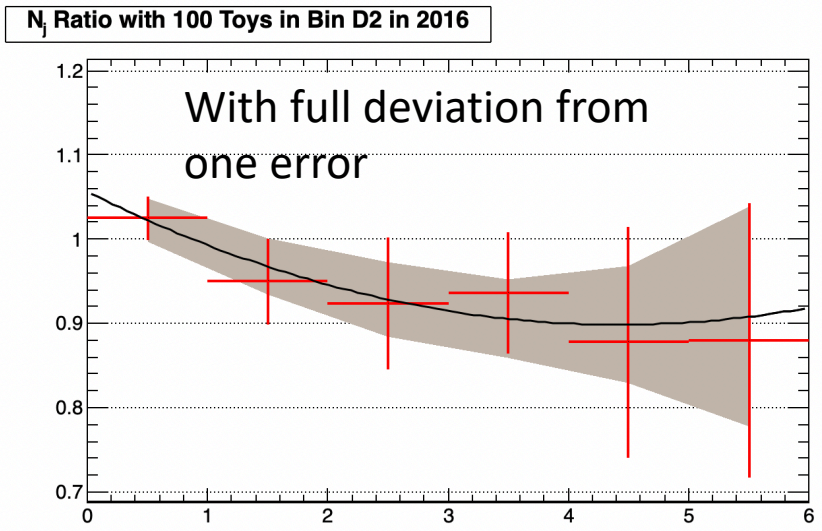
We can undo the transformation to translate the uncorrelated parameters to the three original polynomial parameters where we put the systematic into the ttbar pdf.

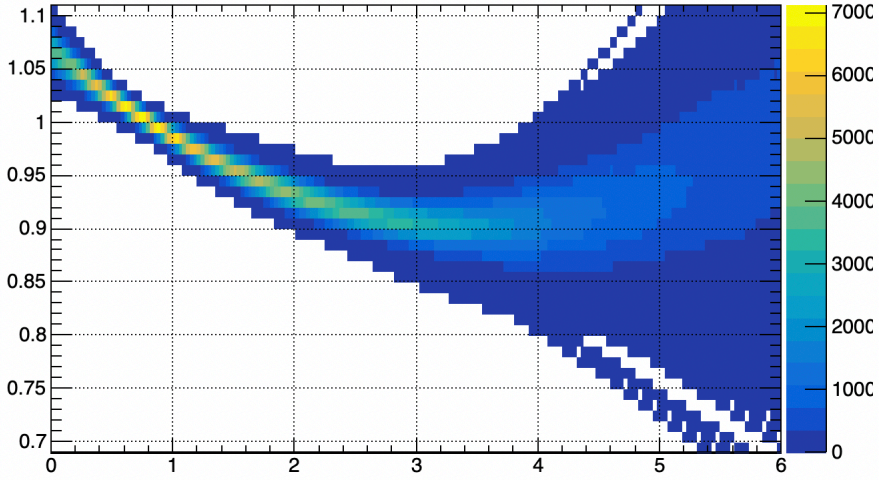
$(\text{transformation matrix})^{-1}$ (p0', p1', p2') → (p0, p1, p2)

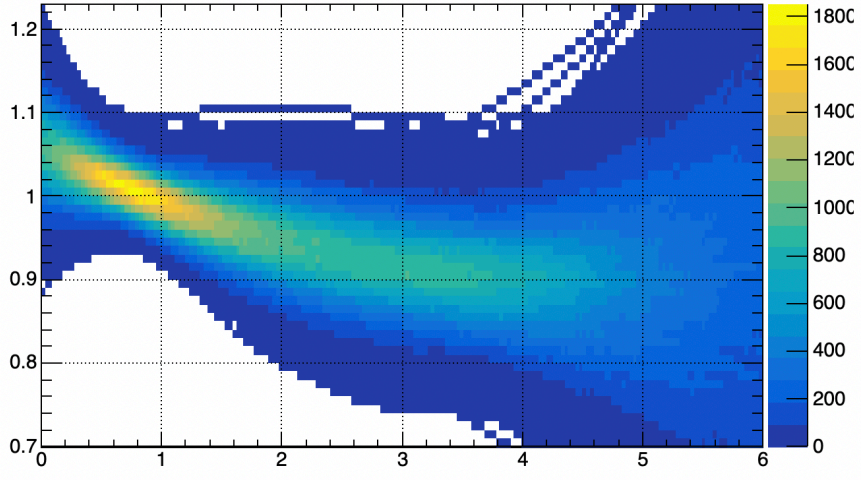
The details are given in the writeup at the end.

Validation of the concept



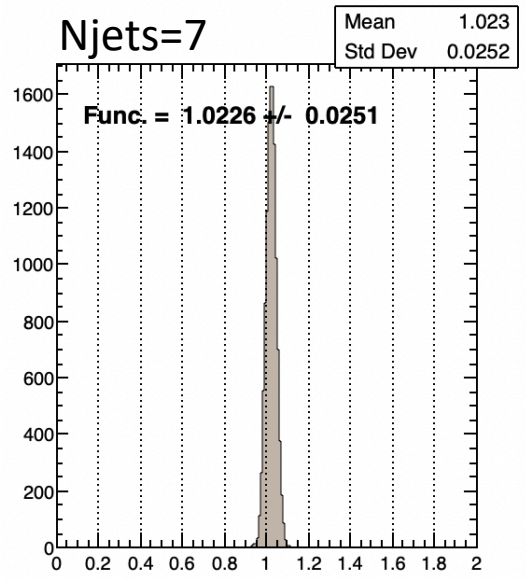


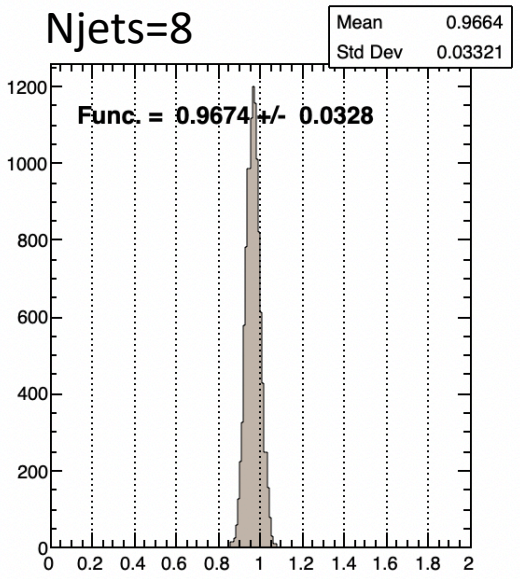


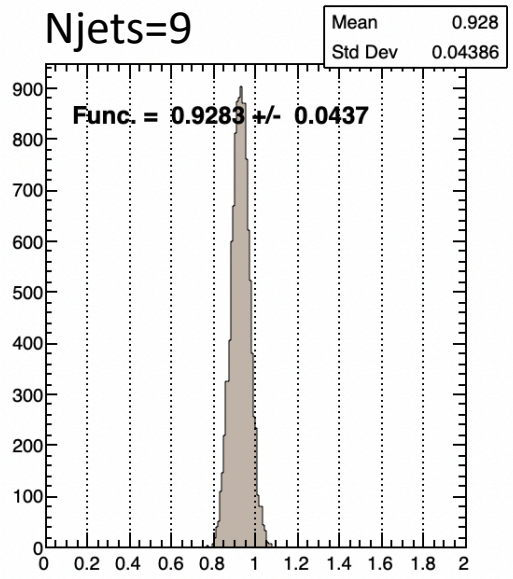


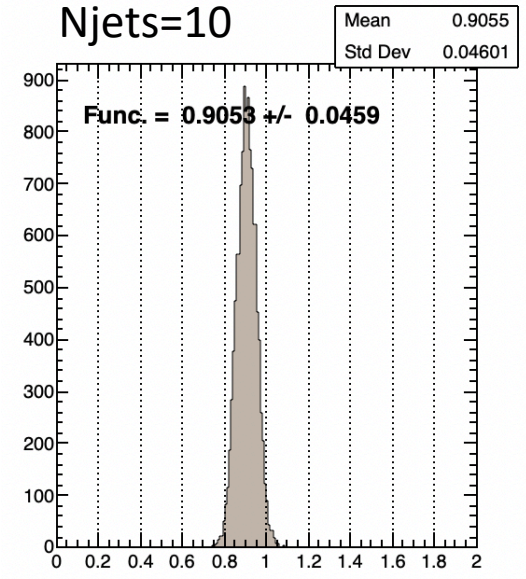
The bottom plots are a 2D histogram of 10000 curves generated by generating the three uncorrelated parameters (p0', p1', p2') for each curve and translating them to the original polynomial parameters (p0, p1, p2).

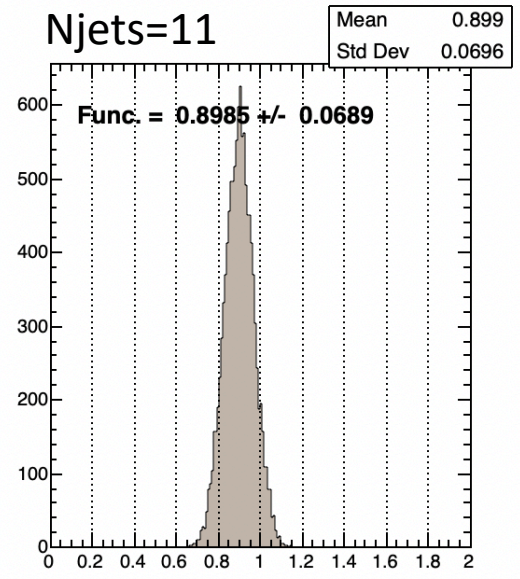
Validation of the concept

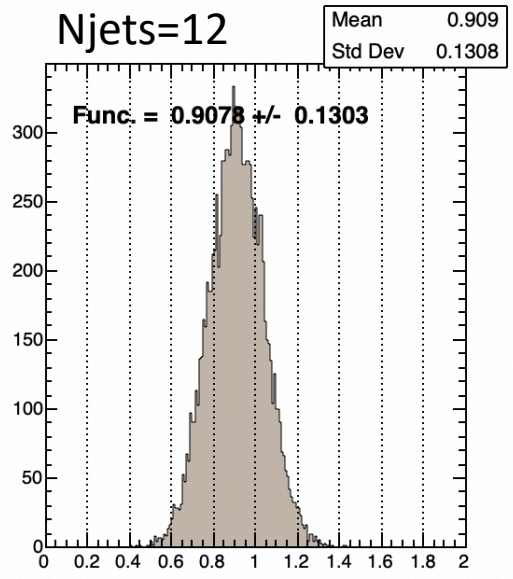










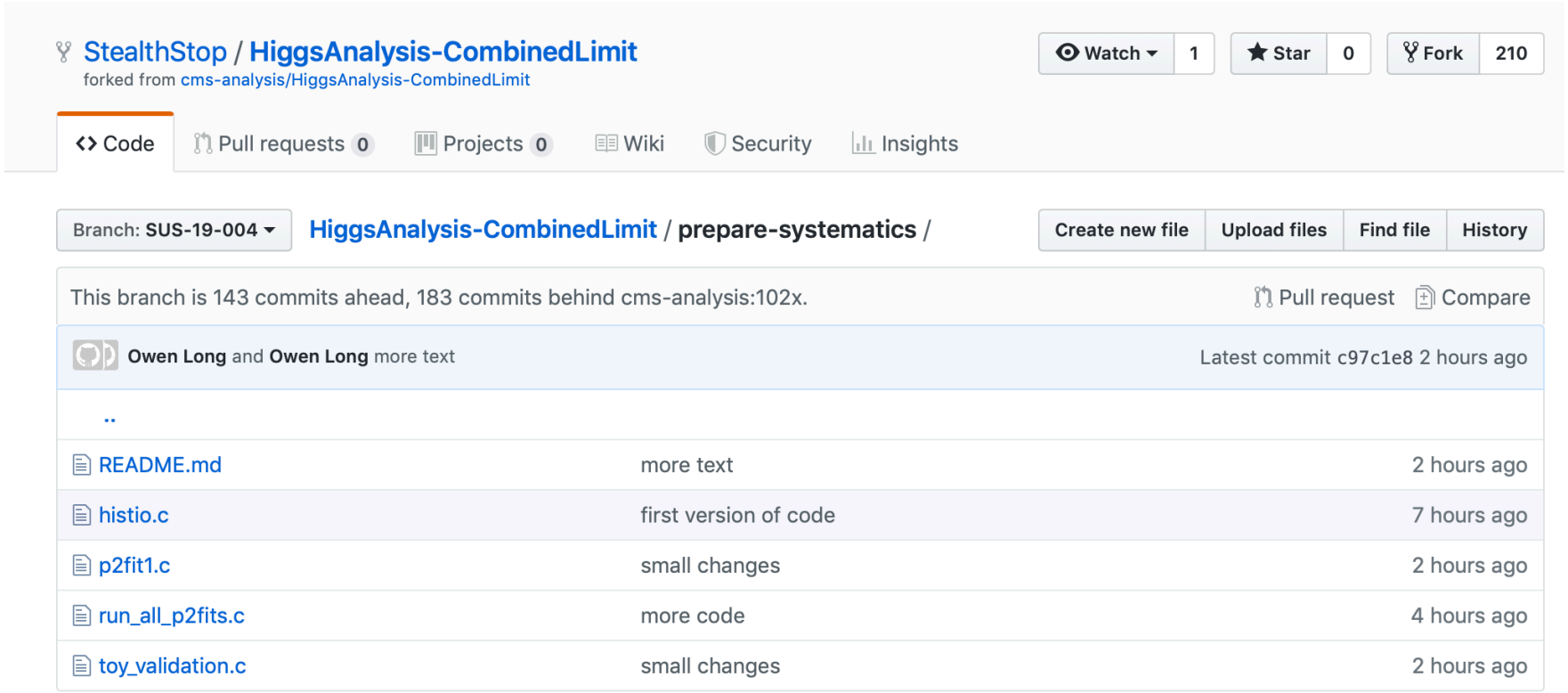


The histograms are slices of the 2D histograms on the bottom right of the previous slide at the 6 values of Njets.

The inset text is the value and uncertainty of the fit (width of the shaded band) at that value of Njets.

There's very good agreement between the histogram mean and RMS with the fit value and uncertainty, showing that the uncorrelated parameters are capable of correctly representing the fit.

Code



StealthStop / HiggsAnalysis-CombinedLimit
forked from cms-analysis/HiggsAnalysis-CombinedLimit

Watch 1 | Star 0 | Fork 210

Code | Pull requests 0 | Projects 0 | Wiki | Security | Insights

Branch: SUS-19-004 | HiggsAnalysis-CombinedLimit / prepare-systematics / | Create new file | Upload files | Find file | History

This branch is 143 commits ahead, 183 commits behind cms-analysis:102x. | Pull request | Compare

Owen Long and Owen Long more text — Latest commit c97c1e8 2 hours ago

..		
README.md	more text	2 hours ago
histio.c	first version of code	7 hours ago
p2fit1.c	small changes	2 hours ago
run_all_p2fits.c	more code	4 hours ago
toy_validation.c	small changes	2 hours ago

Chris and I discussed how to implement this over the last couple of days and agreed on the format of the inputs.

The code is now in github.

The input is the ttbar_systematics.root files for 2016 and 2017.

The code produces a new root file that will be an input for Combine.

It also produces plots of the fits.

Fit function is 2nd-order polynomial

$$f(x) = p_0 + p_1 x + p_2 x^2$$

If the fit parameters p_0, p_1, p_2 were uncorrelated, the error on the function at a particular x value would be

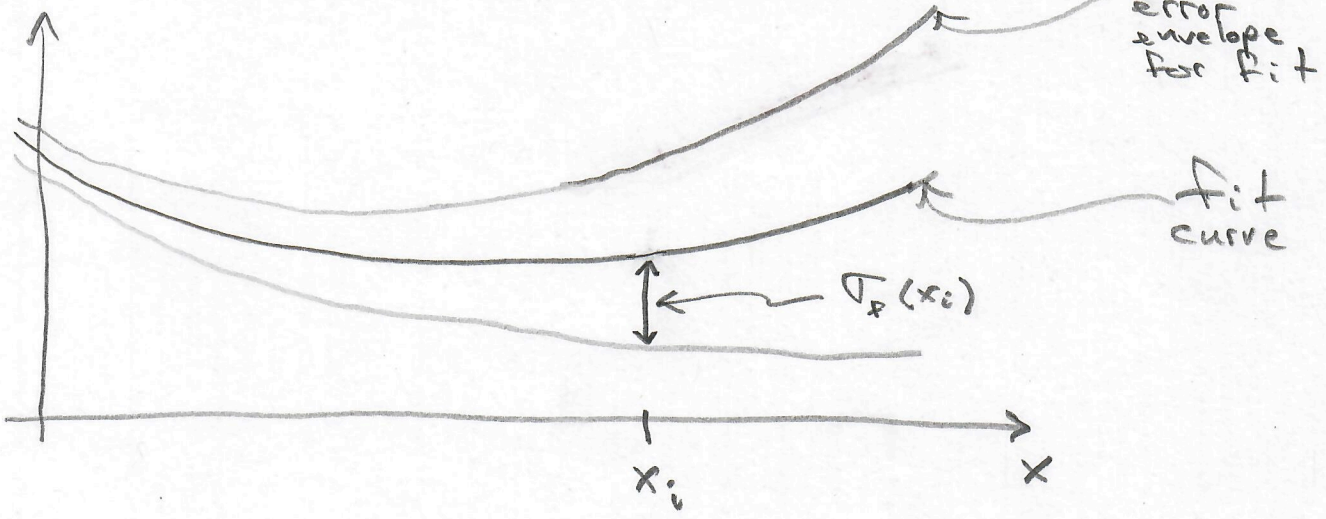
$$\sigma_f^2(\text{at } x) = \sum_i \left(\frac{\partial f}{\partial p_i}\right)_x^2 \sigma_{p_i}^2 \qquad \text{(no correlations)}$$

If they are correlated, and they are, then the error is

$$\sigma_f^2(\text{at } x) = \left(\frac{\partial f}{\partial p_0}, \frac{\partial f}{\partial p_1}, \frac{\partial f}{\partial p_2}\right)_x \begin{pmatrix} \text{Cov} \\ \text{mat} \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial p_0} \\ \frac{\partial f}{\partial p_1} \\ \frac{\partial f}{\partial p_2} \end{pmatrix}_x$$

where the partial derivatives of f are evaluated at the value of x for which you want the error on the fit.

This is how the shaded error band for the fit is calculated



$$\sigma_f^2(x_i) = (1, x_i, x_i^2) \begin{pmatrix} \text{Cov} \\ \text{mat} \end{pmatrix} \begin{pmatrix} 1 \\ x_i \\ x_i^2 \end{pmatrix}$$

We can transform from the 3-dimensional space of the correlated fit parameters P_0, P_1, P_2 to a 3D space where the parameters are uncorrelated using the eigenvalues and eigenvectors of the covariance matrix.

Using vector notation, the correlated polynomial parameters are

$$\vec{P} = P_0 \hat{u}_0 + P_1 \hat{u}_1 + P_2 \hat{u}_2 = \begin{pmatrix} P_0 \\ P_1 \\ P_2 \end{pmatrix}$$

where the $\hat{u}_i$s are orthogonal unit vectors $(\hat{u}_i \cdot \hat{u}_j = \delta_{ij})$

The eigenvalue equation for the covariance matrix is

$$\underline{C}\, \hat{u}_i' = \lambda_i \hat{u}_i'$$

where $\underline{C}$ is the cov. mat. and $\hat{u}_i'$ is the i^{th} eigen vector and λ_i is the i^{th} eigen value.

The eigen vectors define the transformation matrix between the correlated polynomial parameters $\vec{P}$ and the uncorrelated parameters $\vec{P}'$.

transformation matrix

$$\underline{D} = \left(\left| \hat{u}_0' \right| \left| \hat{u}_1' \right| \left| \hat{u}_2' \right| \right)$$

the 3 columns are the three normalized eigen vectors $\hat{u}_i'$ $(\hat{u}_i' \cdot \hat{u}_j' = \delta_{ij})$

Note that $\underline{D}^{-1} = \underline{D}^T$

or $\underline{D}^T \underline{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

The diagonalized covariance matrix is

$$\underline{C}' = \underline{D}^T \underline{C} \underline{D} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

so the interpretation of λ_i is $\sigma_i'^2$
or λ_i is the square of the uncertainty on the parameters $\vec{p}$ projected on to the $\hat{u}_i'$ axis.

The vector of fit parameters in the uncorrelated parameter space is

$$\vec{p}' = \underline{D}^T \vec{p} = \begin{pmatrix} p_0' \\ p_1' \\ p_2' \end{pmatrix} = p_0' \hat{u}_0' + p_1' \hat{u}_1' + p_2' \hat{u}_2'$$

To go from the uncorrelated parameters to the correlated fit polynomial parameters,

$$\vec{p} = \underline{D} \vec{p}'$$

In Combine, we need to use uncorrelated nuisance parameters so the 2nd-order polynomial fit, including the fit uncertainty, will be described by three uncorrelated constrained nuisance parameters, p_0', p_1', p_2'. To implement the systematic, these need to be transformed to the correlated polynomial parameters p_0, p_1, p_2.

$$\begin{pmatrix} p_0 \\ p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} D_{00} & D_{01} & D_{02} \\ D_{10} & D_{11} & D_{12} \\ D_{20} & D_{21} & D_{22} \end{pmatrix} \begin{pmatrix} p_0' \\ p_1' \\ p_2' \end{pmatrix}$$

$$= \begin{pmatrix} (D_{00}p_0' + D_{01}p_1' + D_{02}p_2') \\ (D_{10}p_0' + D_{11}p_1' + D_{12}p_2') \\ (D_{20}p_0' + D_{21}p_1' + D_{22}p_2') \end{pmatrix}$$

The function f at $x = N_j$ is $f(N_j) = p_0 + p_1 N_j + p_2 N_j^2$

$$F(N_j) = (D_{00}p_0' + D_{01}p_1' + D_{02}p_2') + (D_{10}p_0' + D_{11}p_1' + D_{12}p_2')N_j + (D_{20}p_0' + D_{21}p_1' + D_{22}p_2')N_j^2$$

$$f(N_j) = (D_{00} + D_{10}N_j + D_{20}N_j^2)p_0' + (D_{01} + D_{11}N_j + D_{21}N_j^2)p_1' + (D_{02} + D_{12}N_j + D_{22}N_j^2)p_2'$$

If we want the nuisance parameters to each be constrained by a Gaussian with mean zero, width one (like the rest) we can write $p_i' = p_{i,mean}' + \theta_i \sigma_i'$ where θ_i is the nuisance parameter and

$$p_{i\,mean}' = \underline{D}^T_{ij} p_j \quad , \quad \sigma_i' = \sqrt{\lambda_i}$$

↑ polynomial pars from the fit.

$$F(N_j) = (D_{00} + D_{10} N_j + D_{20} N_j^2)(p_{0,mean}' + \theta_0 \sigma_0')$$
$$+ (D_{01} + D_{11} N_j + D_{21} N_j^2)(p_{1,mean}' + \theta_1 \sigma_1')$$
$$+ (D_{02} + D_{12} N_j + D_{22} N_j^2)(p_{2,mean}' + \theta_2 \sigma_2')$$

$$F(N_j) = k_c + k_0 \theta_0 + k_1 \theta_1 + k_2 \theta_2$$

where

$$k_c = (D_{00} + D_{10} N_j + D_{20} N_j^2) p_{0,mean}'$$
$$+ (D_{01} + D_{11} N_j + D_{21} N_j^2) p_{1,mean}'$$
$$+ (D_{02} + D_{12} N_j + D_{22} N_j^2) p_{2,mean}'$$

$$k_0 = (D_{00} + D_{10} N_j + D_{20} N_j^2) \sigma_0'$$

$$k_1 = (D_{01} + D_{11} N_j + D_{21} N_j^2) \sigma_1'$$

$$k_2 = (D_{02} + D_{12} N_j + D_{22} N_j^2) \sigma_2'$$

The four ks for a given N_j bets (N_j value) are constants, so those four numbers are set when the PDF is prepared for the workspace.

Note that the k_c term can be written as

$$k_c = p_0 + p_1 N_j + p_2 N_j^2$$

In other words, k_c is just the value of the 2nd-order polynomial fit at that N_j value, which makes sense, since that's what it should be if the nuisance parameters θ_0, θ_1, θ_2 are not pulled.

All of this is implemented and validated in the code in

github.com/StealthStop/HiggsAnalysis-CombinedLimit/prepare-systematics

Note that the notation in the TTree produced by the code is different that what's used in this note

This note	TTree branch
k_c	coef 0
k_0	coef 1
k_1	coef 2
k_2	coef 3

Owen Long, UC Riverside, June 29, 2019