PROCEEDINGS OF SCIENCE

A study of systematic uncertainties within the MSHT PDF Framework

Matthew Reader[a,*]

[a] *University College London,*
Gower Street, London, WC1E 6BT, United Kingdom

E-mail: matthew.reader.21@ucl.ac.uk

Experimental errors are now incredibly precise, and are often dominated by the systematic uncertainties. Therefore the errors obtained in the Parton Distribution Functions that are extracted from this data will also be dominated by these experimental systematic errors, as well as the systematic errors embedded in the theoretical calculations. However, as is well known, there are often significant uncertainties in these systematic errors, and so to determine precisely the errors in the Parton Distribution Functions, we need to be thoughtful about the uncertainties in the errors themselves. In this paper, we discuss an approach where these "errors on errors" can be incorporated into a χ^2 calculation, and investigate how such a model behaves and what it tells us about the resulting errors. Also we look at two data sets, ATLAS W,Z Data [3] and the ATLAS 7 TeV Inclusive Jet Distribution Data [4] and investigate the information that this model implies about these two data sets.

31st International Workshop on Deep Inelastic Scattering (DIS2024)
8–12 April 2024
Grenoble, France

*Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

PoS(DIS2024)059

1. Introduction

Experimental errors are becoming extremely precise and are now dominated by systematic uncertainties. However, there are often significant errors in the determination of these systematic errors. Therefore, it is becoming increasingly important that these "errors on errors" are incorporated into our Parton Distribution framework such that the extracted errors incorporate this extra layer of uncertainty. In this short document, we demonstrate how it is possible to achieve this.

2. Derivation of the Model

Consider a set of data, $\mathbf{y}$. The probability of $\mathbf{y}$ can be written $P(\mathbf{y}|\mu,\theta)$, where μ are parameters of interest and θ are nuisance parameters that are required for the correctness of the model. If we let $\theta = (\theta_1, ..., \theta_N)$ be independent Gaussian distributed values $u = (u_1, ..., u_N)$, with standard deviations $\sigma_u = (\sigma_{u_1} ..., \sigma_{u_N})$, then the Likelihood function can be written as:

$$\begin{aligned} L(\mu,\theta) &= P(\mathbf{y},\mathbf{u}|\mu,\theta) = P(\mathbf{y}|\mu,\theta)P(\mathbf{u}|\theta) \\ &= P(\mathbf{y}|\mu,\theta)\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_{u_i}}e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2} \end{aligned} \tag{1}$$

However, σ_{u_i} maybe uncertain. One way to incorporate this uncertainty in σ_{u_i} has been proposed in [1]. In this proposal we model the estimated variances, v_i, of $\sigma_{u_i}^2$, as Gamma distributed, which allows us to rewrite equation 1 as:

$$L(\mu,\theta,\sigma_{u_i}^2) = P(y|\mu,\theta)\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_{u_i}}e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}\frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}v_i^{\alpha_i-1}e^{-\beta_i v_i} \tag{2}$$

where $\alpha_i = 1/(4r_i^2)$, $\beta_i = 1/(4r_i^2\sigma_{u_i}^2)$ and r_i is defined as the relative uncertainty in the estimate of the systematic error. The parameters r_i can therefore be referred to as the "error on errors". This model can be reinterpreted as a Student's t-distribution, once we make a small change of variables:

$$L(\mu,\theta,\sigma_{u_i}^2) = P(y|\mu,\theta)\prod_{i=1}^{N}\frac{\Gamma(\frac{\nu_i+1}{2})}{\sqrt{\nu_i\pi}\Gamma(\nu_i/2)}\left(1+\frac{t_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}} \tag{3}$$

where $t_i = \frac{u_i-\theta_i}{\sqrt{v_i}}$ and $\nu_i = \frac{1}{2r_i^2}$. Therefore, we can treat our nuisance parameters as t-distributed!

Now we want to extend this model to incorporate correlated systematic errors. So consider:

$$y_i = d_i + errors = d_i + \sigma_i z_i + \sigma_{u_i} t_{u_i} + \sum_{j=1}^{M}\beta_{ij}t'_j \tag{4}$$

where for each observable y_i we have one statistical error σ_i, with a z_i that is a Normally distributed fluctuating variable, one uncorrelated systematic error σ_{u_i} with a t_{u_i} that is a t-distributed fluctuating variable with dof of $\nu = 1/2r_{\chi^2}^2$, and M correlated systematic errors, β_{ij}, each with a fluctuation t'_j that are t-distributed with degree of freedom of $\nu = 1/2r_{\chi^2}^2$.

If we treat all the t-distributions as independent [1], then the Log-likelihood function, once we have maximized with respect to z_i, can be written up to some constants as:

[1] Note that if we treated the t-distributions as a Multi-variate t-distribution with zero correlation between all the t_u and the t', then the likelihood function would be different.

PoS(DIS2024)059

$$-2LnL = \sum_{i=1}^{N}\left(\frac{m_i - d_i - \sigma_{u_i}t_{u_i} - \sum_j \beta_{ij}t'_j}{\sigma_i}\right)^2$$
$$+(\nu+1)\sum_{i=1}^{N} Ln\left(1+\frac{t_{u_i}^2}{\nu}\right) + (\nu+1)\sum_{j=1}^{M} Ln\left(1+\frac{t'^2_j}{\nu}\right) \equiv \chi^2 \quad (5)$$

where we can define this to be a χ^2 once we have minimized simultaneously with respect to both t_{u_i} and t'_j.

3. The case of Statistical and uncorrelated systematic errors only

Let's initially consider the case of only statistical and uncorrelated systematic errors. In this case we can write $y_i = d_i + \sigma_i z_i + \sigma_{u_i} t_{u_i}$, where $z_i \sim N(0,1)$, $t_{u_i} \sim t(0, \nu = 1/2r^2_{Dist})$. That is, we are drawing our y from a distribution where the statistical errors are normally distributed and the uncorrelated systematic errors are t-distributed.

Using numerical integration we can investigate the expectation of χ^2, $E[\chi^2]$, and the Variance of χ^2, $Var[\chi^2]$. In Figure 1 we plot the $E[\chi^2]$ as a function of r_{Dist}, where the $E[\chi^2]$ has been calculated at various different r^2_χ. As can be seen, the expectation is a growing function of r_{Dist}, even if $r_{\chi^2} = r_{Dist}$. In Figure 2 we plot $Var[\chi^2]/2$ as a function of r_{Dist}. This plot shows that the $Var[\chi^2]/2$ is a similarly increasing function of r_{Dist}, even as r_{χ^2} is increased.

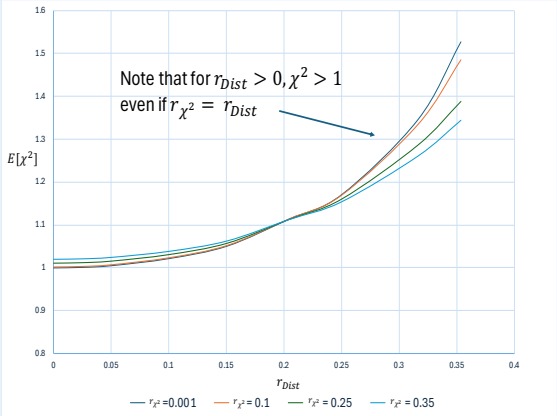

Figure 1: Graph of $E[\chi^2]$ as a Function of r_{Dist} for 4 different r_{χ^2} ($\sigma_i = \sigma_{u_i} = 1$)

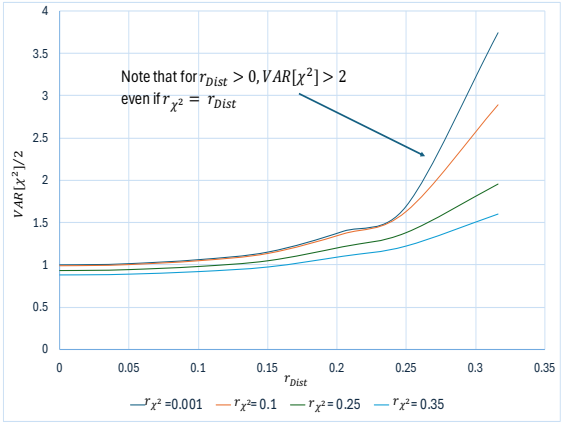

Figure 2: Graph of $Var[\chi^2]/2$ as a Function of r_{Dist} for 4 different r_{χ^2} ($\sigma_i = \sigma_{u_i} = 1$)

In the case of only normally distributed statistical and t-distributed uncorrelated systematic errors, the standard deviation of the simple mean, $y_{mean} = \sum_{i=1}^{N} y_i/N$, is given by:

$$\sigma_{Mean} \approx \frac{\sqrt{\sum_{i=1}^{N}\sigma_i^2 + \sigma_{u_i}^2 \nu/(\nu-2)}}{N} = \frac{\sqrt{\sum_{i=1}^{N} E[\chi_i^2(r_{\chi^2}\to 0)](\sigma_i^2+\sigma_{u_i}^2)}}{N} \quad (6)$$

In Table 1, we show how σ_{Mean} behaves as a function of r and N (with $\sigma_i = \sigma_{u_i} = 1$). As can be seen for all the cases of N, the ratio of $\sigma_{r_{Dist}=0.408}/\sigma_{r_{Dist}=0.0001} \approx 1.4$.

The next question we want to ask is what happens if we minimize the χ^2, calculated with $r_{\chi^2} = r_{Dist}$, with respect to our mean? That is, what is the standard deviation of the fitted mean,

PoS(DIS2024)059

r_{Dist}	N=2	N=10	N=500	Ratio
0.001	0.995	0.449	0.064	1.000
0.100	0.991	0.452	0.064	1.005
0.250	1.092	0.481	0.069	1.077
0.300	1.122	0.504	0.071	1.108
0.408	1.417	0.637	0.089	1.393

Table 1: Table showing σ_{mean} as a function of r and N (with $\sigma_i = \sigma_{u_i} = 1$). The last column given the Ratio $\frac{\sigma_{r_{Dist}}}{\sigma_{(r_{Dist}=0.001)}}$ for the case of N=500.

r_{Dist}	N=2	N=10	N=500	Ratio
0.001	0.995	0.449	0.064	1.000
0.100	0.991	0.452	0.064	1.004
0.250	1.092	0.479	0.068	1.069
0.300	1.122	0.493	0.069	1.087
0.408	1.417	0.547	0.076	1.393

Table 2: Table showing σ_{FIT} as a function of r and N (with $\sigma_i = \sigma_{u_i} = 1$). The last column given the Ratio $\frac{\sigma_{FIT}}{\sigma_{(r_{Dist}=r_{\chi^2}=0.001)}}$ for the case of N=500.

N	M	r	ν	$E[\chi^2(d_i)]$	σ_{χ^2}	$\frac{\sigma_{\varphi_{MEAN}}}{\sigma_{FIT(r_{\chi^2}=0.001)}}$	$\sigma_{\varphi_{FIT}}$	$\sigma_{\varphi_{FIT}}/\sigma_{\varphi_{MEAN}}$	$\frac{\sigma_{\varphi_{FIT}}}{\sigma_{\varphi_{FIT}(r_{Dist}=r_{\chi^2}=0.001)}}$
2	2	0.001	500000	1.99949	2.05734	1.58114	1.58114	1.000	1.000
2	2	0.25	8	2.28779	2.31064	1.78103	1.78970	1.005	1.132
2	2	0.40824829	3	2.87634	2.98990	2.51644	2.32738	0.925	1.472
5	5	0.001	500000	4.99873	3.20452	2.28036	2.28036	1.000	1.000
5	5	0.25	8	5.42717	3.44951	2.62217	2.62208	1.000	1.150
5	5	0.40824829	3	6.53625	4.49232	3.81179	3.51314	0.922	1.541
10	5	0.001	500000	9.99746	4.58094	2.25832	2.25833	1.000	1.000
10	5	0.25	8	10.47021	4.68864	2.64291	2.63296	0.996	1.166
10	5	0.40824829	3	11.61824	5.53776	4.08220	3.48161	0.853	1.542
10	10	0.001	500000	9.99746	4.53088	3.17806	3.17806	1.000	1.000
10	10	0.25	8	10.37782	4.69332	3.67337	3.64109	0.991	1.146
10	10	0.40824829	3	11.91221	5.81006	5.40928	4.72917	0.874	1.488

Table 3: Table showing how the expectation of the χ^2, the standard deviation of the χ^2, the standard deviation of the simple mean, the standard deviation of the fitted mean behave as function of $r = r_{Dist} = r_{\chi^2}$, the number of observables, N and the number of correlated systematic errors, M.

σ_{FIT}, if $r_{Dist} = r_{\chi^2}$? We show the results of this in Table 2 again with $\sigma_i = \sigma_{u_i} = 1$. As can be seen from the table, we have very similar behaviour to that in Table 1, apart from the fact that as N increases the standard deviation of the fitted mean, σ_{FIT}, initially starts to increase more slowly as a function of r, compared to the standard deviation of the simple mean.

4. Expectation and Variance of χ^2 as a Function of r for the Case of Statistical and Correlated Systematic Errors

Let's now consider the case of only statistical and correlated systematic errors. This is, let's consider the case of N observables each with a Gaussian statistical and M t-distributed correlated systematic errors : $y_i = d_i + \sigma_i z_i + \sum_{j=1}^{M} \beta_{ij} t'_j$, where $z_i \sim N(0,1)$, and $t'_j \sim t(0, \nu = 1/2r^2_{Dist})$.

In the case where $r = r_{Dist} = r_{\chi^2}$, and $\sigma_i = \beta_{ij} = 1$, we obtain the data shown in Table 4. This figure shows how the expectation of the χ^2, $E[\chi^2]$, the standard deviation of the χ^2, σ_{χ^2}, the standard deviation of the simple mean, $\sigma_{\varphi_{MEAN}}$, and the standard deviation of the fitted mean, $\sigma_{\varphi_{FIT}}$, behave as function of $r = r_{Dist} = r_{\chi^2}$, the number of observables, N, and the number of correlated systematic errors, M. The behaviour is very similar to what we saw in Section 3 for the case of uncorrelated errors, in that the expectation and variance increase as a function of r, the standard deviation of the simple mean grows more quickly than the standard deviation of the fitted

PoS(DIS2024)059

mean as r increases. The ratio in the last column of Figure 4, increases to about 1.5 in all cases compared to 1.4 in the uncorrelated systematic error case.

5. ATLAS W,Z Data analysis [3]

This very precise data gives a strong constraint on the strange quark. However, the fit quality using the MSHT20 (NNLO) PDF set is relatively poor giving a $\chi^2 \sim 120$. This data set consists of 61 data points, each with 1 statistical error, 1 uncorrelated systematic error, and 131 correlated systematic errors. In Figure 3 we show how the expectation of the χ^2, calculated in the Gaussian limit (i.e. a $r \to 0$), varies as a function of the underlying distributional r, i.e r_{Dist}. As can be seen, the χ^2 starts at 61, as expected, and increases as we increase the underlying distributional r in the simulation. It can also be seen from this Figure that $E[\chi^2]$ reaches 120 point when the underlying distribution has a $r \approx 0.4$.

Using the experimental data, we obtain the graph shown in Figure 4 when we calculate the χ^2 using equation 5, where we have optimized with respect to r_{u_i} and r'_j, as a function of r_{χ^2}. Once we include for this decrease in χ^2 with increasing r shown in Figure 4, we can infer that some of the inflated χ^2 for this data set is due to the error on errors.

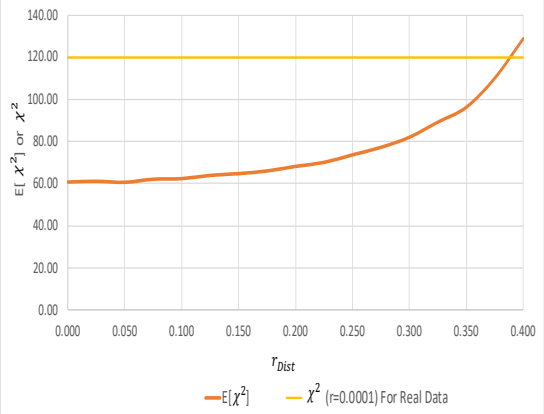

Figure 3: Graph shows the Expectation of $E[\chi^2]$, calculated using $r = 0.001$, as a function of relative error, r_{Dist} of the simulated underlying systematic errors.

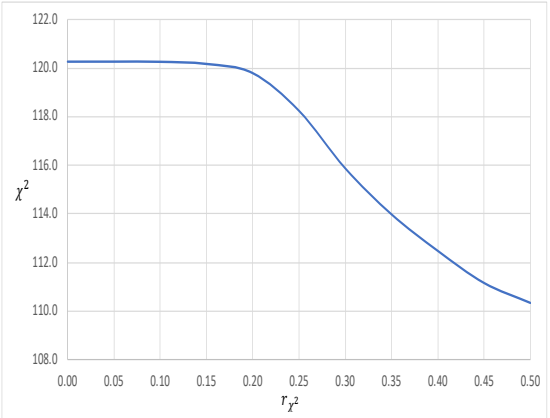

Figure 4: Graph shows the χ^2 as a function of relative error, r_{χ^2}, at which the χ^2 is calculated

6. ATLAS 7 TeV Inclusive Jet Distributions Analysis [4]

This ATLAS data, combined with the availability of NNLO corrections provides constraints on the Gluon PDF at high x. The data set consists of 140 data points, each with 1 correlated systematic error and 70 correlated systematic errors. Using the MSHT20 NNLO PDF set, the fit quality to this data set is relatively poor with a $\chi^2 \approx 280$. In order to improve the fit quality MSHT20 [2] use a de-correlation process which reduces the $\chi^2 \approx 243$.

In Figure 5, we show a similar graph to that in Figure 3 for this data set. The graph shows the expectation of χ^2, $E[\chi^2]$, calculated with $r_{\chi^2} = 0.00001$, where the systematic errors are sampled from t-distribution with d.o.f $1/2r^2_{Dist}$. As can be seen $E[\chi^2]$ crosses the $\chi^2 = 243$ line at an r_{Dist} of about 0.45. In Figure 6, we show how the χ^2 varies as a function of r_{χ^2} for the cases of just

PoS(DIS2024)059

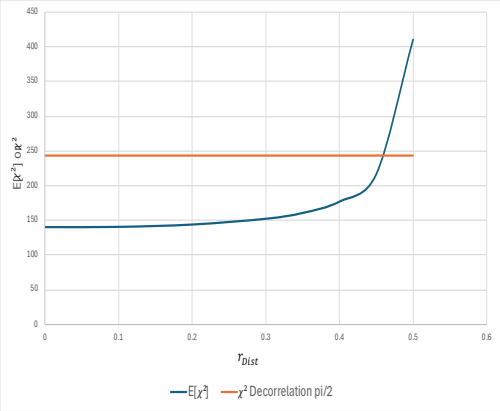

Figure 5: Graph shows $E[\chi^2]$, calculated with $r_{\chi^2} = 0.00001$, where systematic errors are sampled from t-distribution with d.o.f $1/2r^2_{Dist}$. Line at 243.43 is χ^2 calculated using $r_{\chi^2} = 0.0001$ for de-correlated data.

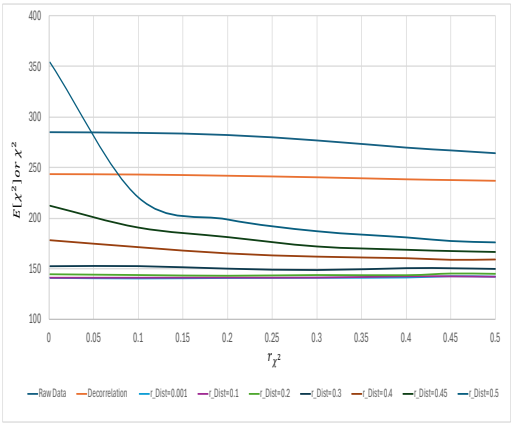

Figure 6: Graph shows the χ^2 or $E[\chi^2]$ as a function of relative error, r_{χ^2}. Raw data refers to just the raw data provided by ATLAS. De-correlation refers to χ^2 calculated with 3 of the "two point" systematic uncertainties de-correlated. Other lines show the $E[\chi^2]$ using pseudo data produced using various r_{Dist}.

the raw data, and for the case of the MSHT20 de-correlation procedure (labelled "Decorrelation"). Also in Figure 6, we show how the expectation of χ^2, $E[\chi^2]$, behaves as a function of r_{χ^2} for different values of r_{Dist}. Suggestive as it is, making the assertion that this data has an $r \approx 0.45$, would neglect not only the decreasing behaviour of χ^2 as a function of r, but also the theoretical uncertainties and the choice of de-correlation process used.

7. Conclusions

In this document we have shown how we can incorporate Errors on Errors into the calculation of a χ^2. We have also shown that the Expected χ^2 and Variance of χ^2 increase as the relative errors of the systematic errors increase. We have noted that for both data sets analysed $r \approx 0.4$. We have also observed that the ratio of the expected standard deviation of the mean, using $r_{\chi^2} = 0.001$ and $r_{Dist} = 0.4$, to that calculated using $r_{\chi^2} = 0.001$ and $r_{Dist} = 0.001$ is approximately $1.2 - 1.5$. This is suggestive of using a Tolerance, T^2, in the region of $1.5 - 2$ in these test cases.

References

[1] G. Cowan, Eur. Phys. J. C **79** (2019) no.2, 133 doi:10.1140/epjc/s10052-019-6644-4 [arXiv:1809.05778 [physics.data-an]].

[2] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin and R. S. Thorne, Eur. Phys. J. C **81** (2021) no.4, 341 doi:10.1140/epjc/s10052-021-09057-0 [arXiv:2012.04684 [hep-ph]].

[3] M. Aaboud *et al.* [ATLAS], Eur. Phys. J. C **77** (2017) no.6, 367 doi:10.1140/epjc/s10052-017-4911-9 [arXiv:1612.03016 [hep-ex]].

[4] G. Aad *et al.* [ATLAS], JHEP **02** (2015), 153 [erratum: JHEP **09** (2015), 141] doi:10.1007/JHEP02(2015)153 [arXiv:1410.8857 [hep-ex]]. v 2023

PoS(DIS2024)059