ROYAL HOLLOWAY UNIVERSITY OF LONDON

03/07/24

Uncertainties on systematics in PDF fits

Enzo Canonero

Glen Cowan

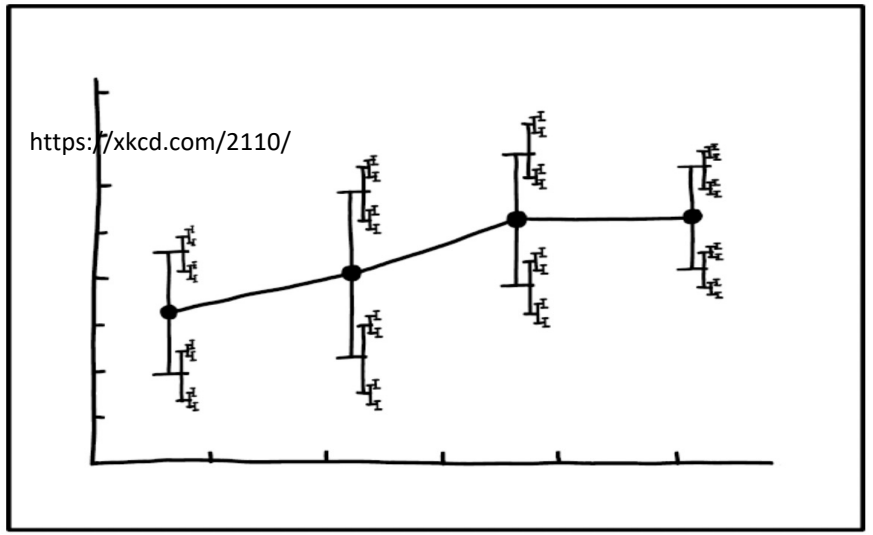
Motivation

1) *Some **systematic uncertainties*** can be well estimated:

- **Related to stat. error of control measurements**
- **Related to size of MC event sample**

2) But they can also be ***quite uncertain***:

- **Theory systematics**
- **Two points systematics ….**

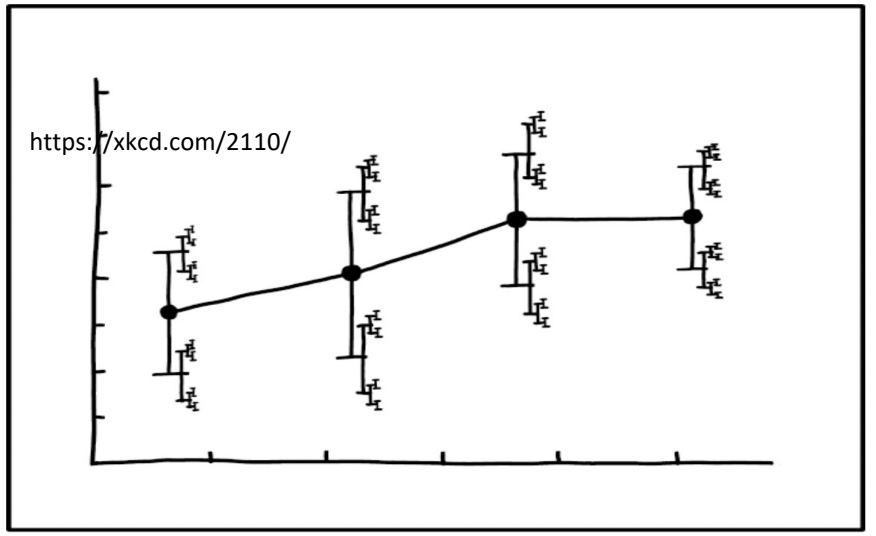


see: G. Cowan, Eur. Phys. J. C (2019) 79:133; arXiv:1809.05778,
Canonero, E., Brazzale, A.R. & Cowan, G. *Eur. Phys. J. C* **83**, 1100 (2023)

Motivation

1) *Some **systematic uncertainties*** can be well estimated:

- **Related to stat. error of control measurements**
- **Related to size of MC event sample**



2) But they can also be ***quite uncertain***:

- **Theory systematics**
- **Two points systematics ….**

see: G. Cowan, Eur. Phys. J. C (2019) 79:133; arXiv:1809.05778,
Canonero, E., Brazzale, A.R. & Cowan, G. *Eur. Phys. J. C* **83**, 1100 (2023)

- Non-trivial consequences:
  - **Fits are pulled less by incompatible data**
  - **Incompatible data are treated as an extra source of uncertainty resulting in inflated confidence intervals**

Formulation of the problem

- Suppose measurements $\boldsymbol{y}$ have a probability density $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta})$
  - $\boldsymbol{\mu}$ = Parameters of interest
  - $\boldsymbol{\theta}$ = Nuisance parameters

- Auxiliary Measurements $\boldsymbol{u}$ are used to provide info on nuisance parameters and are (often) assumed to be independently Gaussian distributed

- The resulting Likelihood is:

Can be a real measurement or just our best guess based on theoretical reasons

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta}) \times \prod_i \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}$$

Formulation of the problem

- Suppose measurements $\boldsymbol{y}$ have a probability density $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta})$
  - $\boldsymbol{\mu}$ = Parameters of interest
  - $\boldsymbol{\theta}$ = Nuisance parameters

- Auxiliary Measurements $\boldsymbol{u}$ are used to provide info on nuisance parameters and are (often) assumed to be independently Gaussian distributed

- The resulting Likelihood is:

Can be a real measurement or just our best guess based on theoretical reasons

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta}) \times \prod_i \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}$$

- And the log Likelihood:

$$\log L(\boldsymbol{\mu},\boldsymbol{\theta}) = \log P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta}) - \sum \frac{(\boldsymbol{u_i}-\boldsymbol{\theta_i})^2}{\boldsymbol{2\sigma_{u_i}^2}}$$

Formulation of the problem

- Suppose measurements $\boldsymbol{y}$ have a probability density $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta})$
  - $\boldsymbol{\mu}$ = Parameters of interest
  - $\boldsymbol{\theta}$ = Nuisance parameters

- Auxiliary Measurements $\boldsymbol{u}$ are used to provide info on nuisance parameters and are (often) assumed to be independently Gaussian distributed

- The resulting Likelihood is:

Can be a real measurement or just our best guess based on theoretical reasons

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta}) \times \prod_i \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}$$

- And the log Likelihood:

$$\log L(\boldsymbol{\mu},\boldsymbol{\theta}) = \log P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta}) - \sum \frac{(\boldsymbol{u_i}-\boldsymbol{\theta_i})^2}{2\boldsymbol{\sigma_{u_i}^2}}$$

Let systematic errors be potentially uncertain!

Gamma Variance Model (GVM)

- The original **quadratic terms** in the log likelihood replaced by **logarithmic terms:**

$$\sum_i \frac{(u_i - \theta_i)^2}{2\sigma_{u_i}^2} \longrightarrow \sum_i \left(1 + \frac{1}{2\varepsilon_i^2}\right) \log\left(1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2}\right)$$

ε = error-on-error parameter

ε = 0.3 means 30% uncertainty on σ

Gamma Variance Model (GVM)

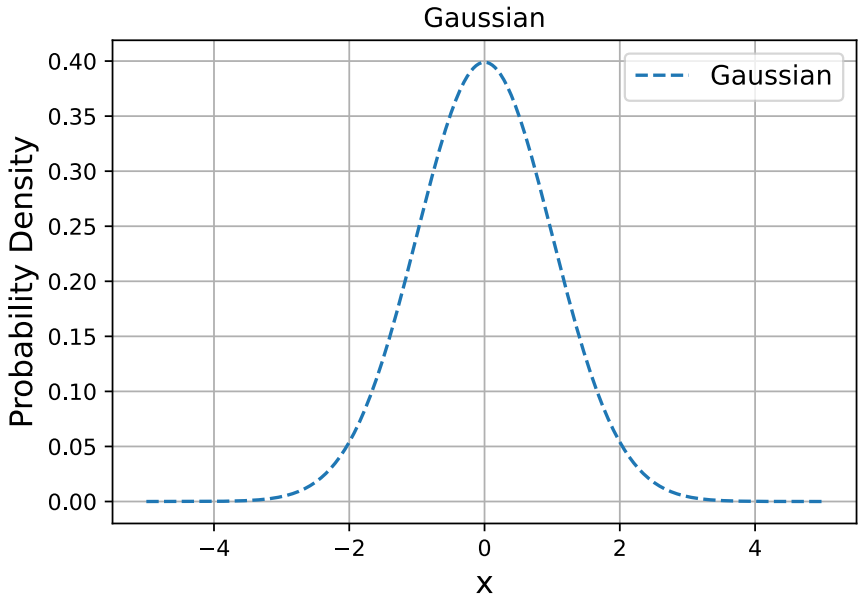
- The original **quadratic terms** in the log likelihood replaced by **logarithmic terms:**

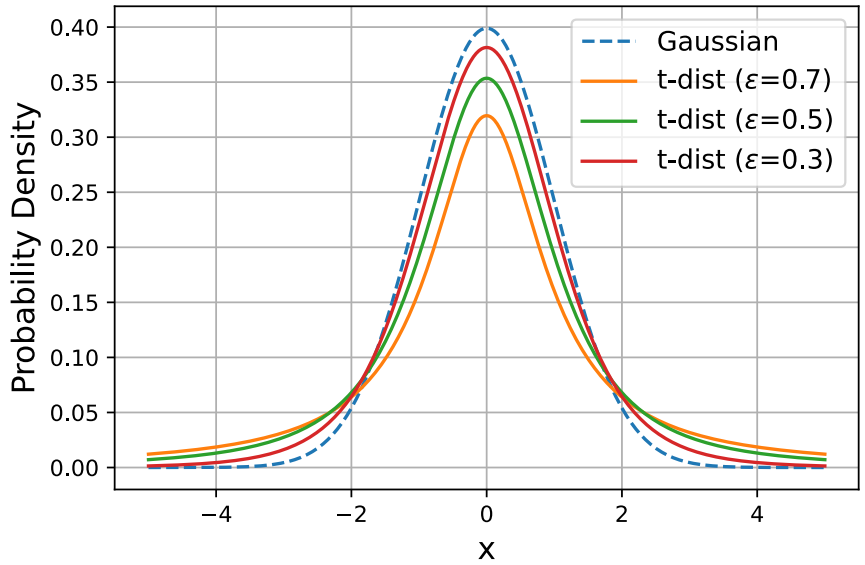
$$\sum_i \frac{(u_i - \theta_i)^2}{2\sigma_{u_i}^2} \longrightarrow \sum_i \left(1 + \frac{1}{2\varepsilon_i^2}\right) \log\left(1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2}\right)$$

ε = error-on-error parameter

ε = 0.3 means 30% uncertainty on σ

- Equivalent to switch from **Gaussian constraints** to **Student's t constraints** for systematics:



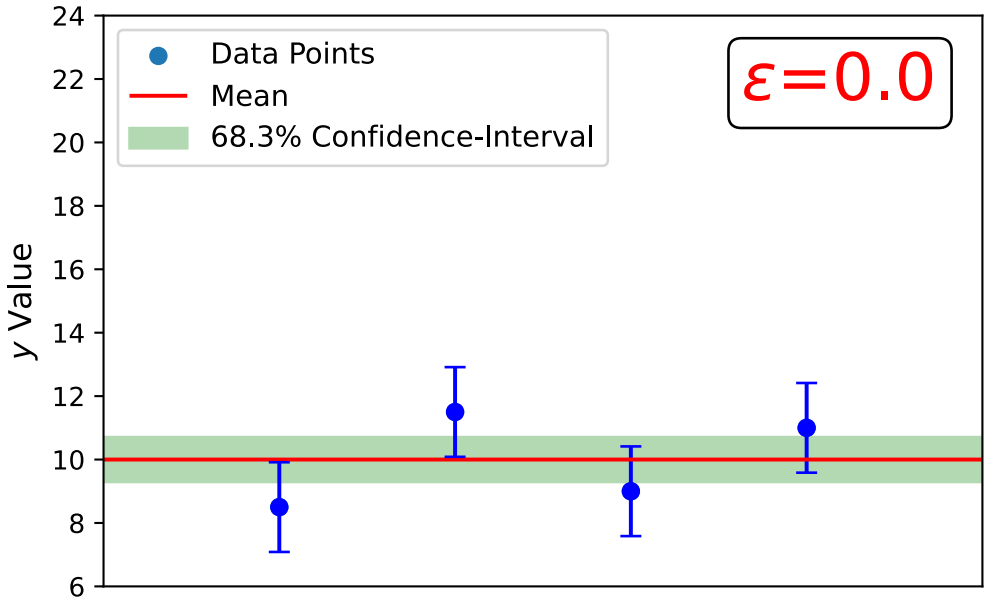


Sensitivity to outliers

- Suppose we want to average 4 measurements all with **statistical** and **syst errors** equal to **1**. Also assume they all have equal ***errors-on-errors*** $\boldsymbol{\varepsilon}$ (auxiliary measurements set to zero):

$$\log L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = -\frac{1}{2}\sum_i \frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2} - \frac{1}{2}\sum_i \left(1 + \frac{1}{2\boldsymbol{\varepsilon}_i^2}\right) \log\left(1 + 2\boldsymbol{\varepsilon}_i^2 \frac{{\theta_i}^2}{\sigma_{u_i}^2}\right)$$

Measurements internally compatible

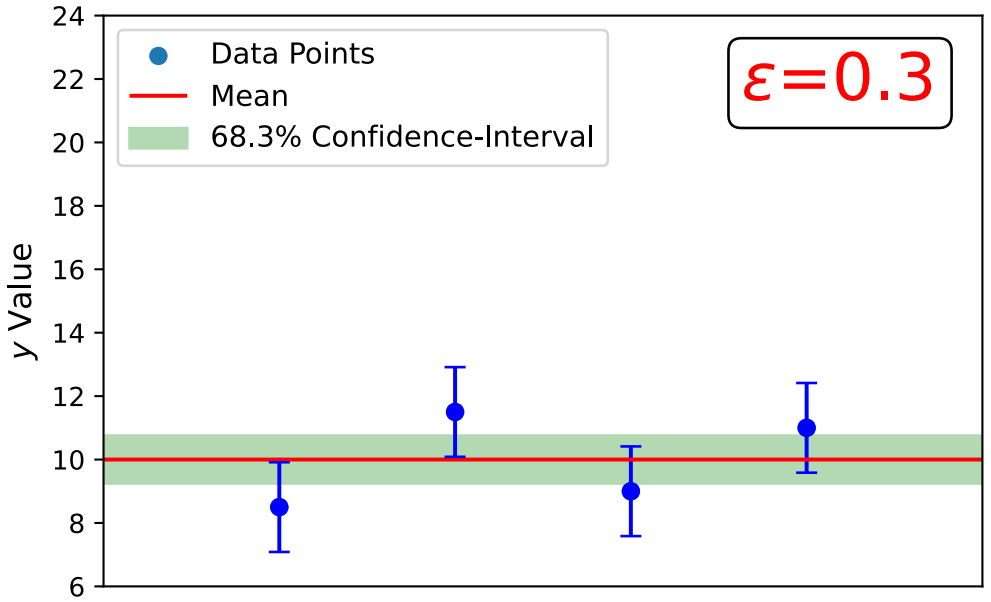


Sensitivity to outliers

- Suppose we want to average 4 measurements all with **statistical** and **syst errors** equal to **1**. Also assume they all have equal ***errors-on-errors*** $\boldsymbol{\varepsilon}$ (auxiliary measurements set to zero):

$$\log L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = -\frac{1}{2}\sum_i \frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2} - \frac{1}{2}\sum_i \left(1 + \frac{1}{2\boldsymbol{\varepsilon}_i^2}\right) \log\left(1 + 2\boldsymbol{\varepsilon}_i^2 \frac{{\theta_i}^2}{\sigma_{u_i}^2}\right)$$

Measurements internally compatible

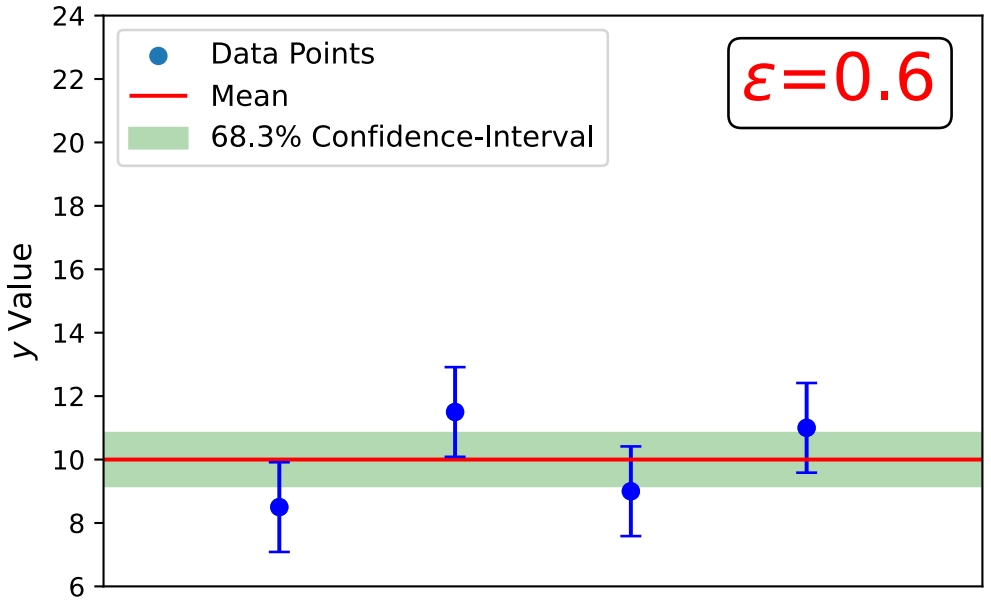


Sensitivity to outliers

- Suppose we want to average 4 measurements all with **statistical** and **syst errors** equal to **1**. Also assume they all have equal ***errors-on-errors*** $\boldsymbol{\varepsilon}$ (auxiliary measurements set to zero):

$$\log L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = -\frac{1}{2}\sum_i \frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2} - \frac{1}{2}\sum_i \left(1 + \frac{1}{2\boldsymbol{\varepsilon}_i^2}\right) \log\left(1 + 2\boldsymbol{\varepsilon}_i^2 \frac{{\theta_i}^2}{\sigma_{u_i}^2}\right)$$

Measurements internally compatible



Sensitivity to outliers

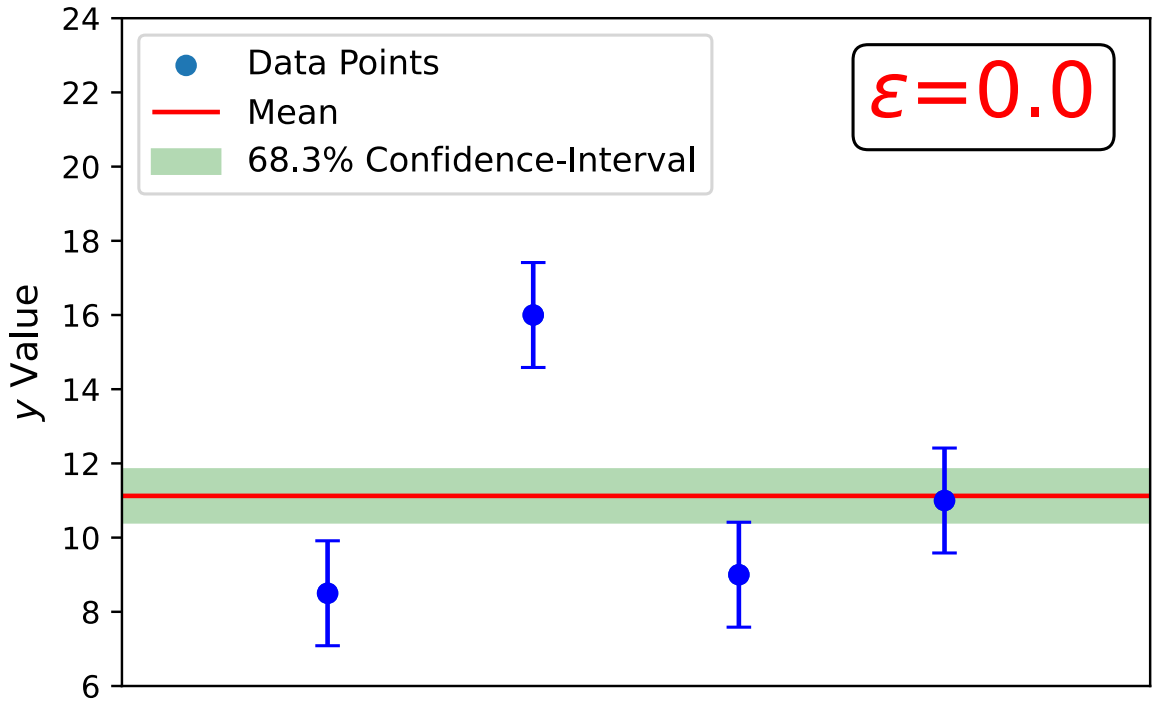
1. If data are internally compatible results are only slightly modified

2. The estimate of the mean does not change when we increase ε

3. The size of the confidence interval for the mean only slightly increases, reflecting the extra degree of uncertainty introduced by errors-on-errors

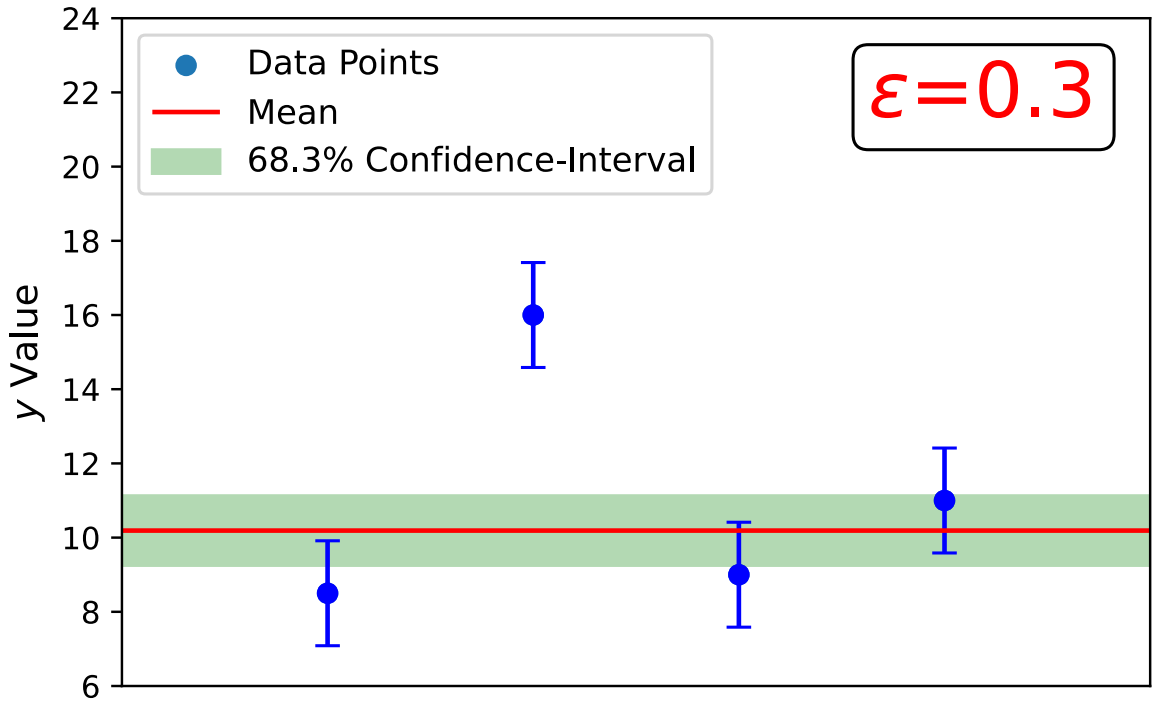
Sensitivity to outliers

- Suppose one of the measurements is an outlier
- If data are internally incompatible important changes can be observed



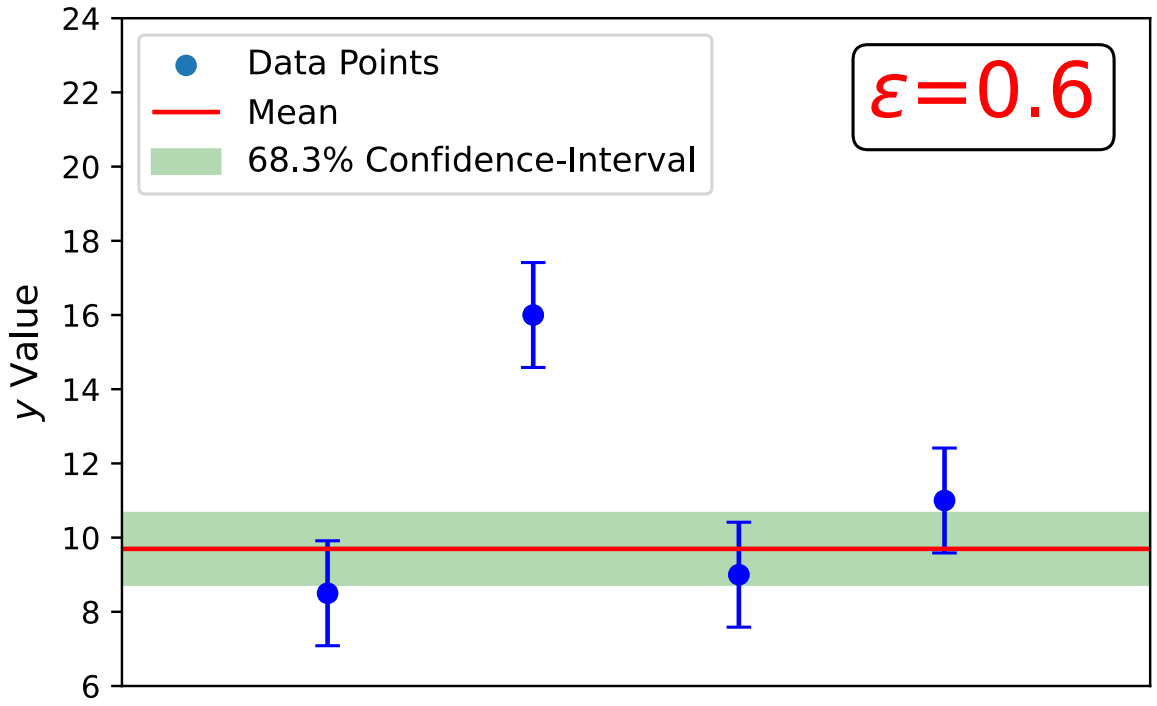
Sensitivity to outliers

- Suppose one of the measurements is an outlier
- If data are internally incompatible important changes can be observed

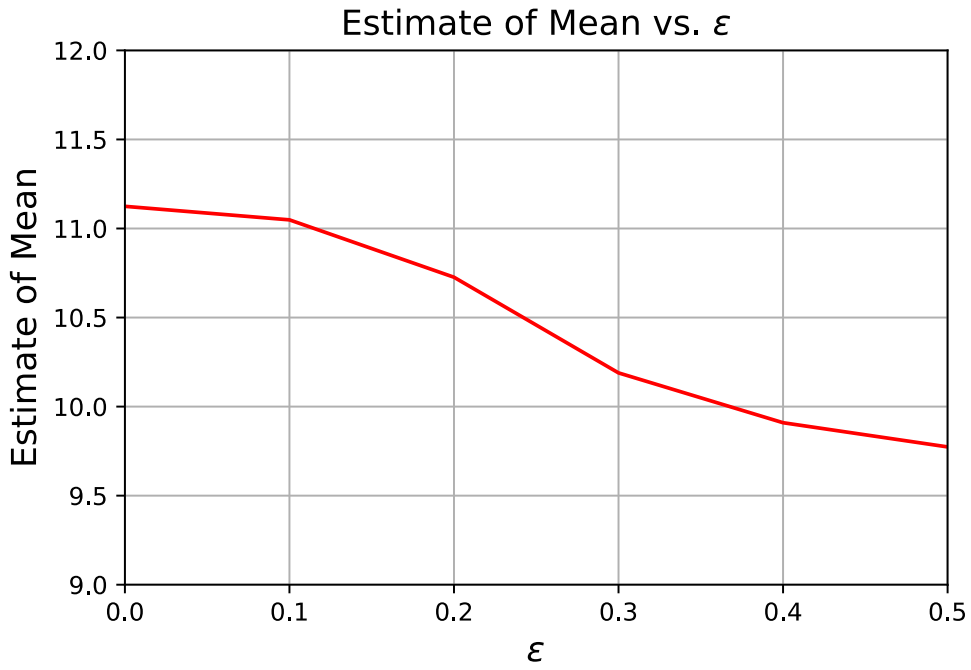


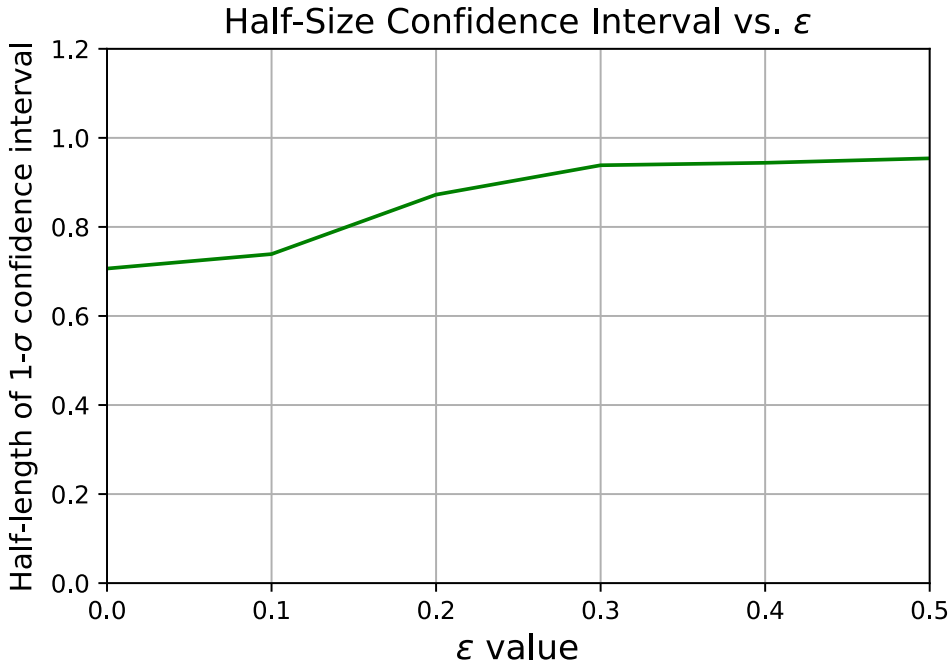
Sensitivity to outliers

- Suppose one of the measurements is an outlier
- If data are internally incompatible important changes can be observed



Sensitivity to outliers





1. With increasing ε, the estimate of mean is pulled less strongly by the outlier

2. The error bar grows more significantly: the GVM treats internal incompatibility as an additional source of uncertainty

3. The model is sensitive to internal compatibility of the data

Complete model

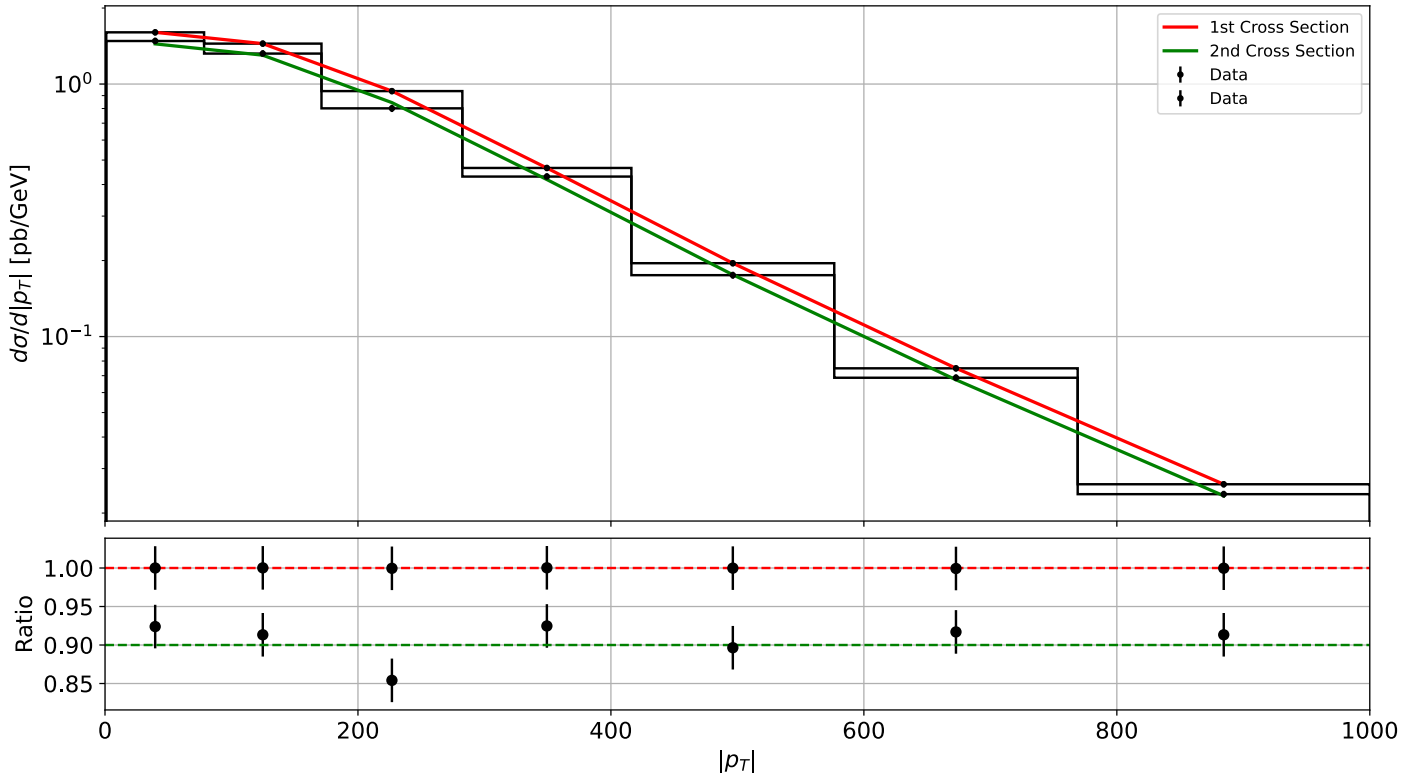
- Gamma Variance Model:

$$\chi^2 = \sum_i \frac{(y_i - f_i(\boldsymbol{a}) - \sum_s \Gamma_i^s \theta_s)^2}{\sigma_i^2} + \sum_s \left(1 + \frac{1}{2\boldsymbol{\varepsilon}_i^2}\right) \log\left(1 + 2\boldsymbol{\varepsilon}_i^2 \boldsymbol{\theta}_i{}^2\right)$$

- Appliable both to addictive and multiplicative systematics as only the systematic terms in the chi2 are being changed

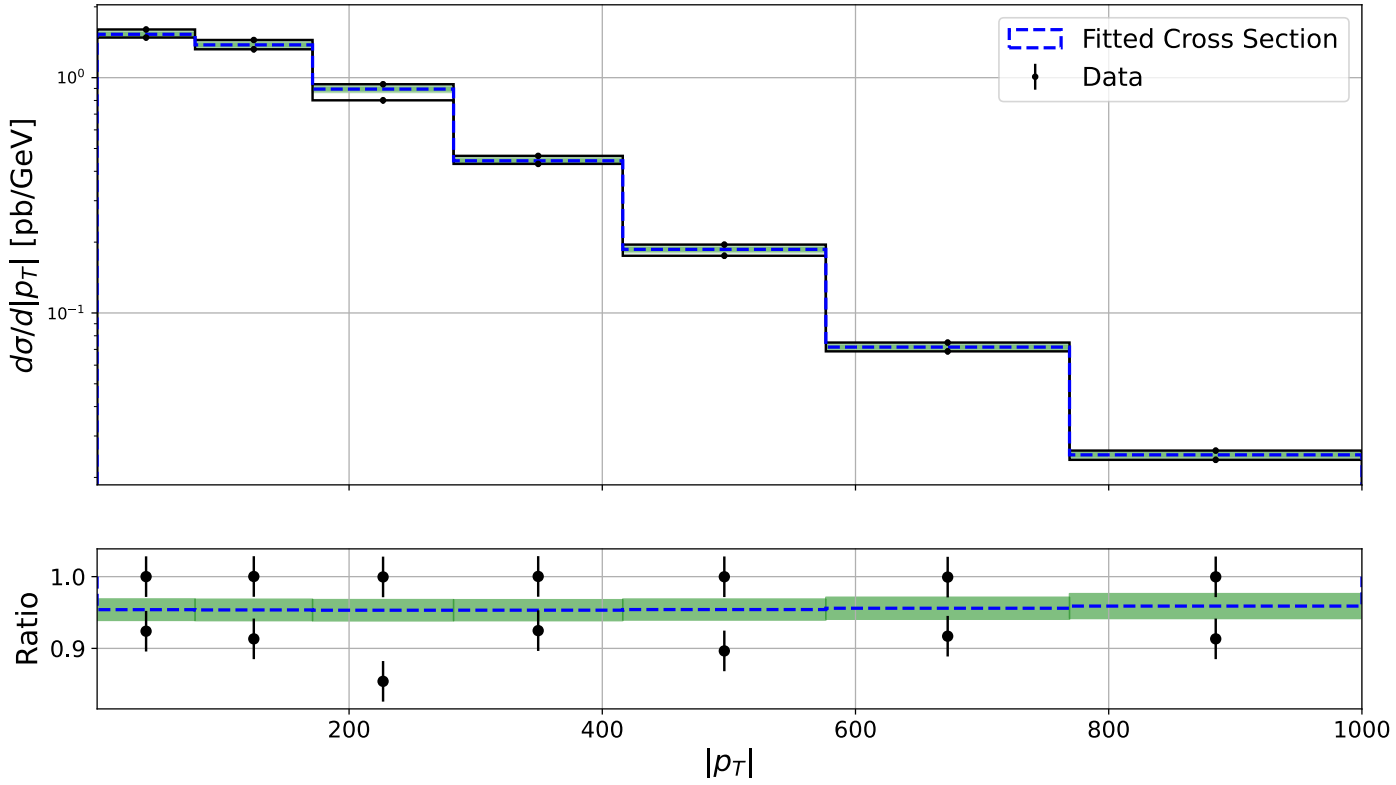
Two cross-sections example

1. Consider two measurements of the same distribution, analogous to results from two separate experiments.
2. Both distributions are subject to a normalization uncertainty, which is assumed to be itself uncertain.



Two cross-sections example

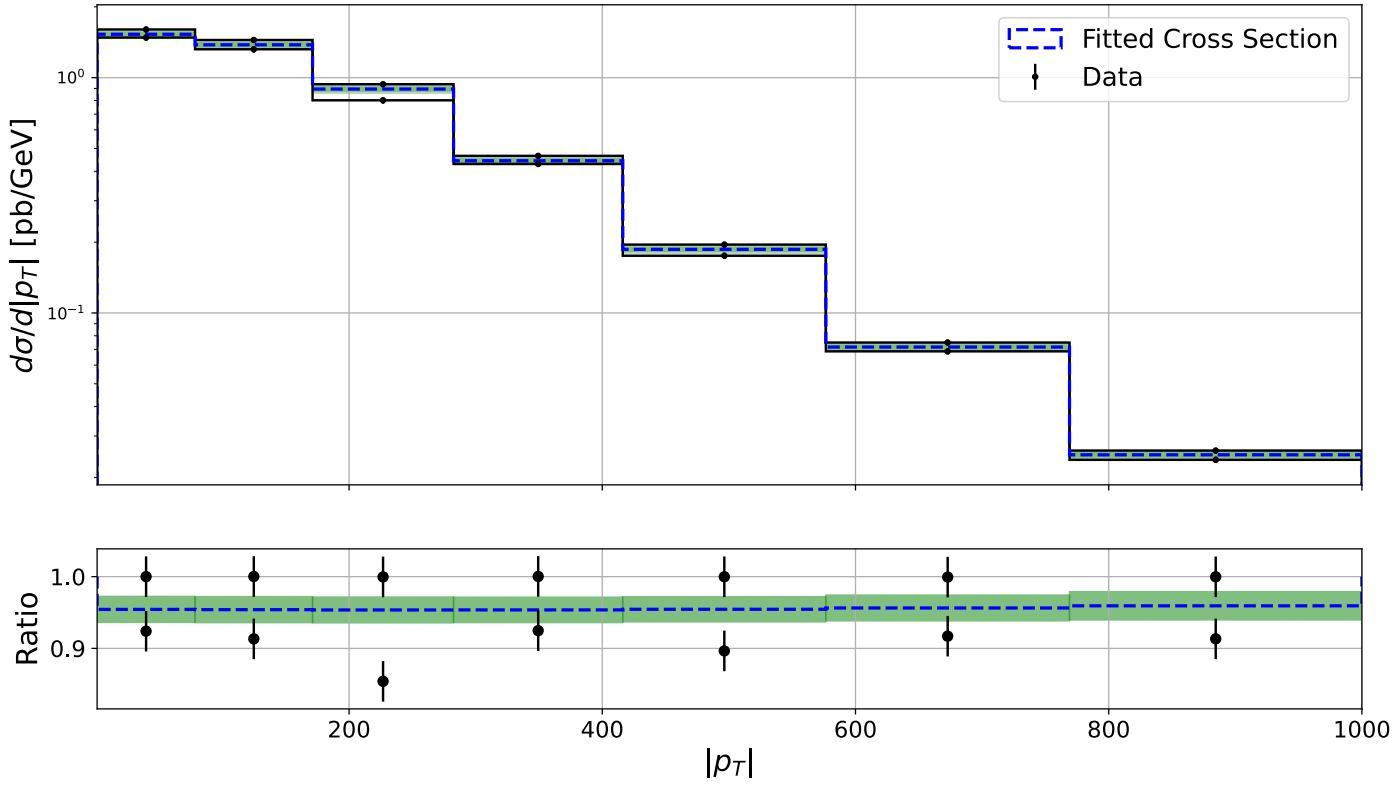
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 0%

Two cross-sections example

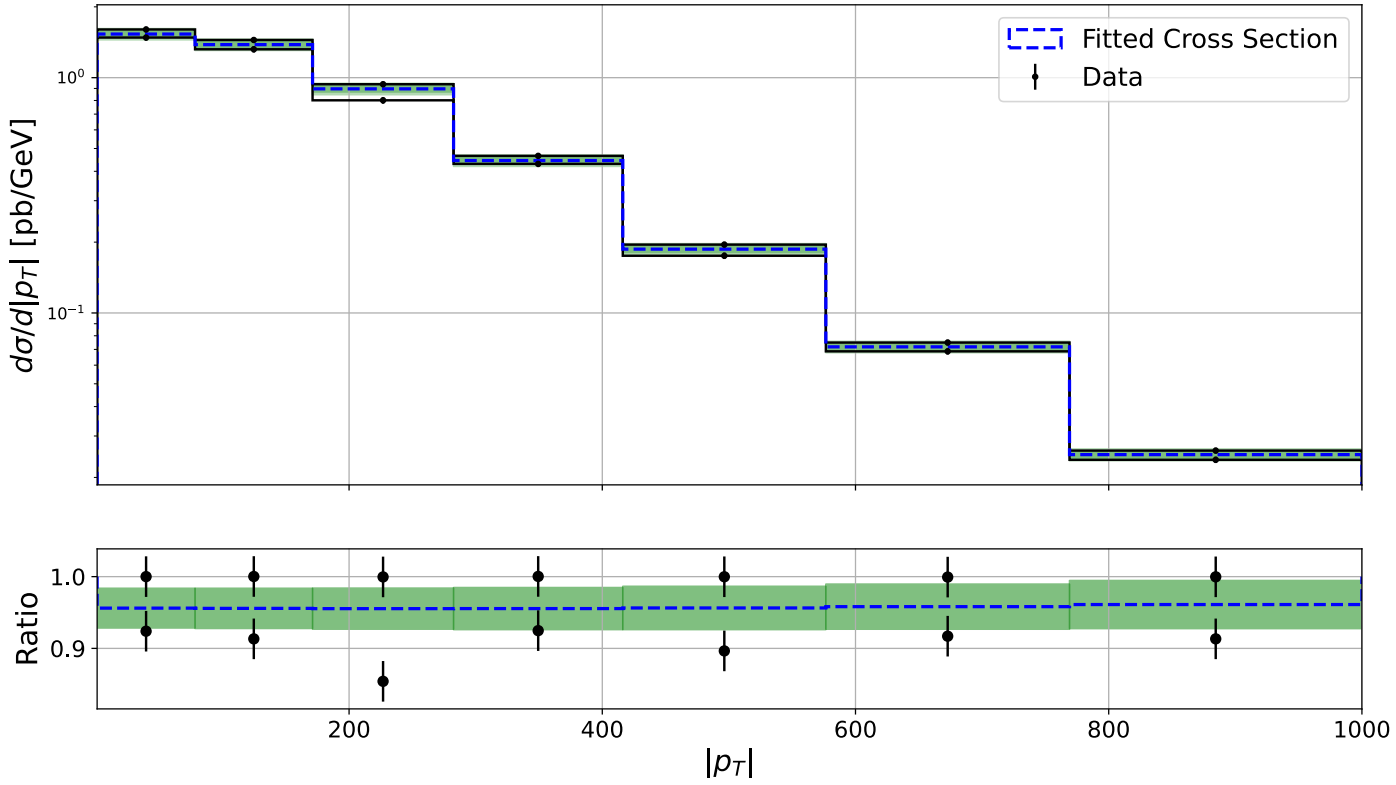
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 10%

Two cross-sections example

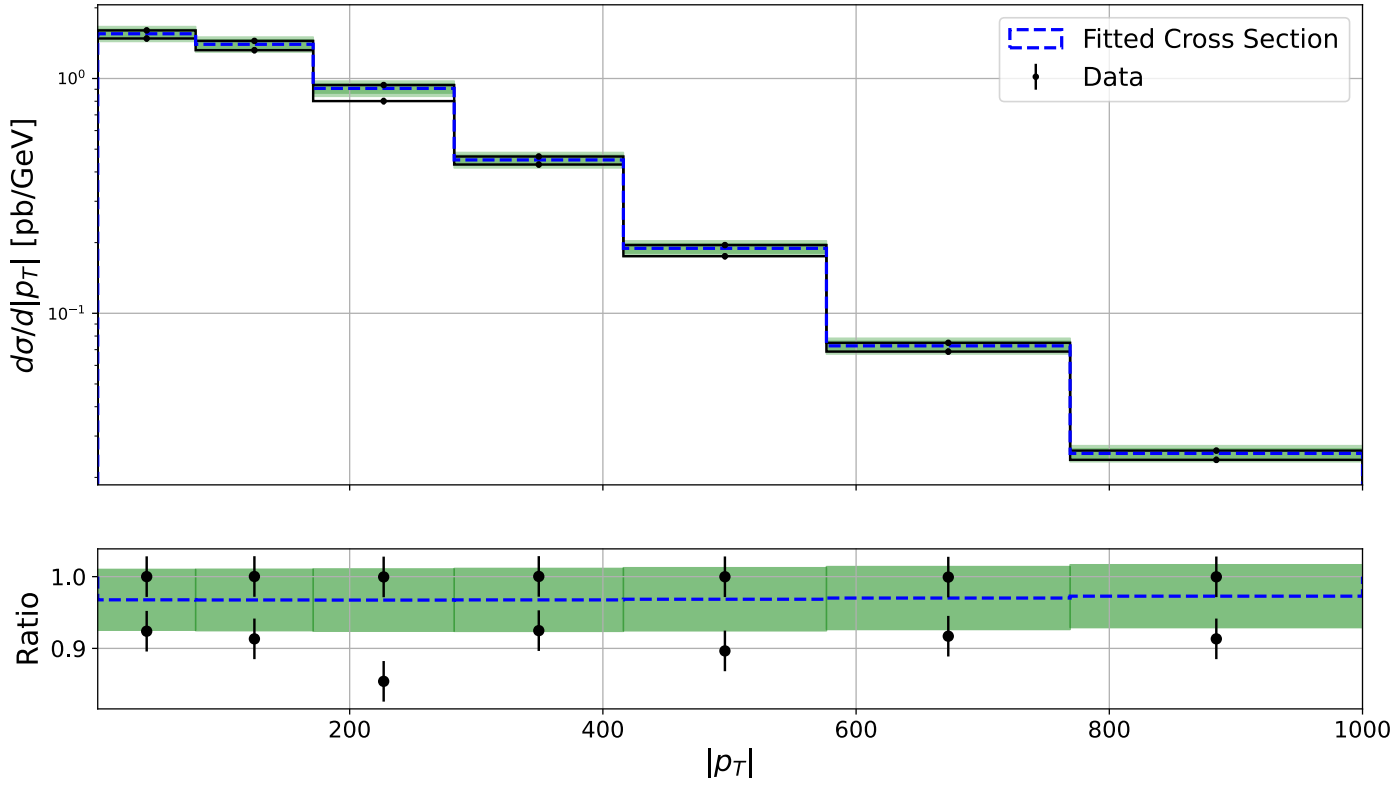
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 20%

Two cross-sections example

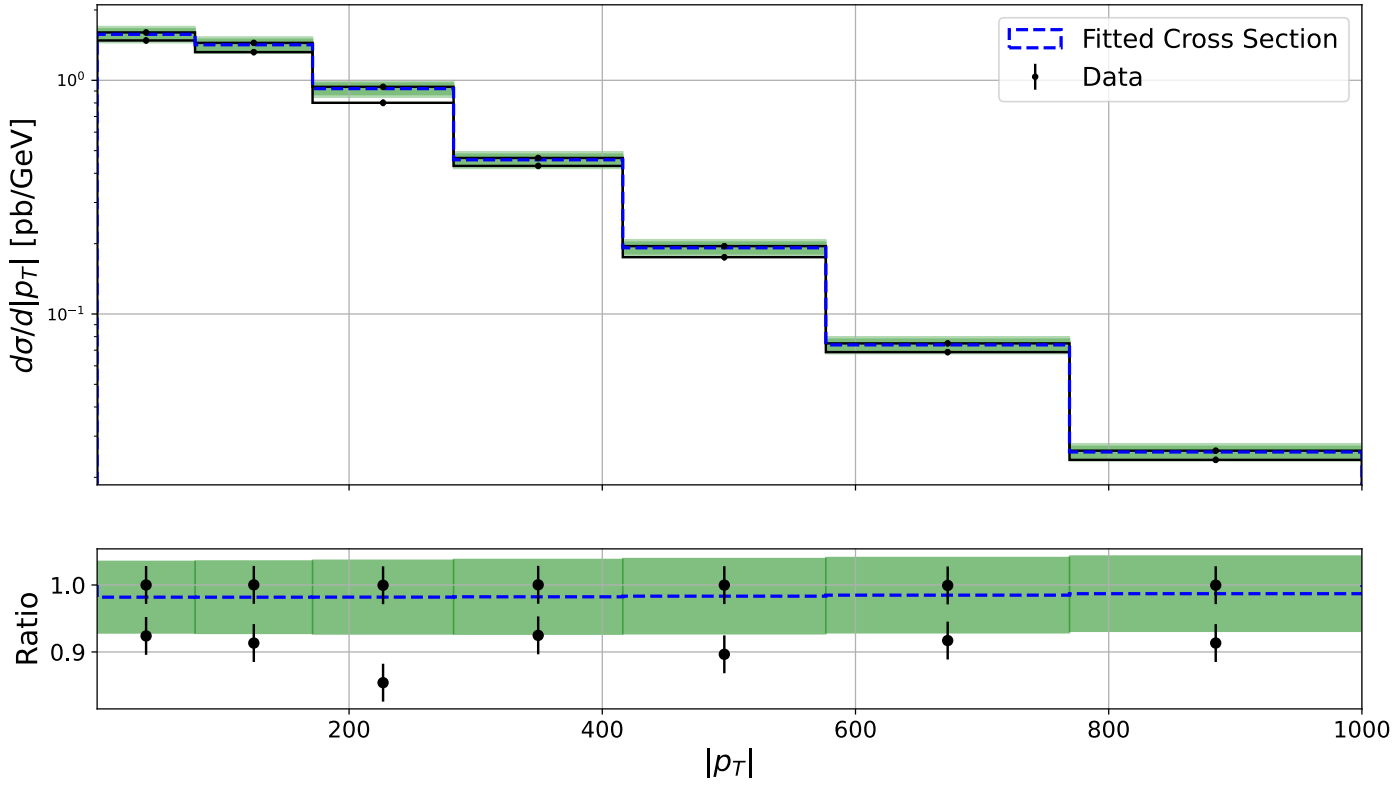
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 30%

Two cross-sections example

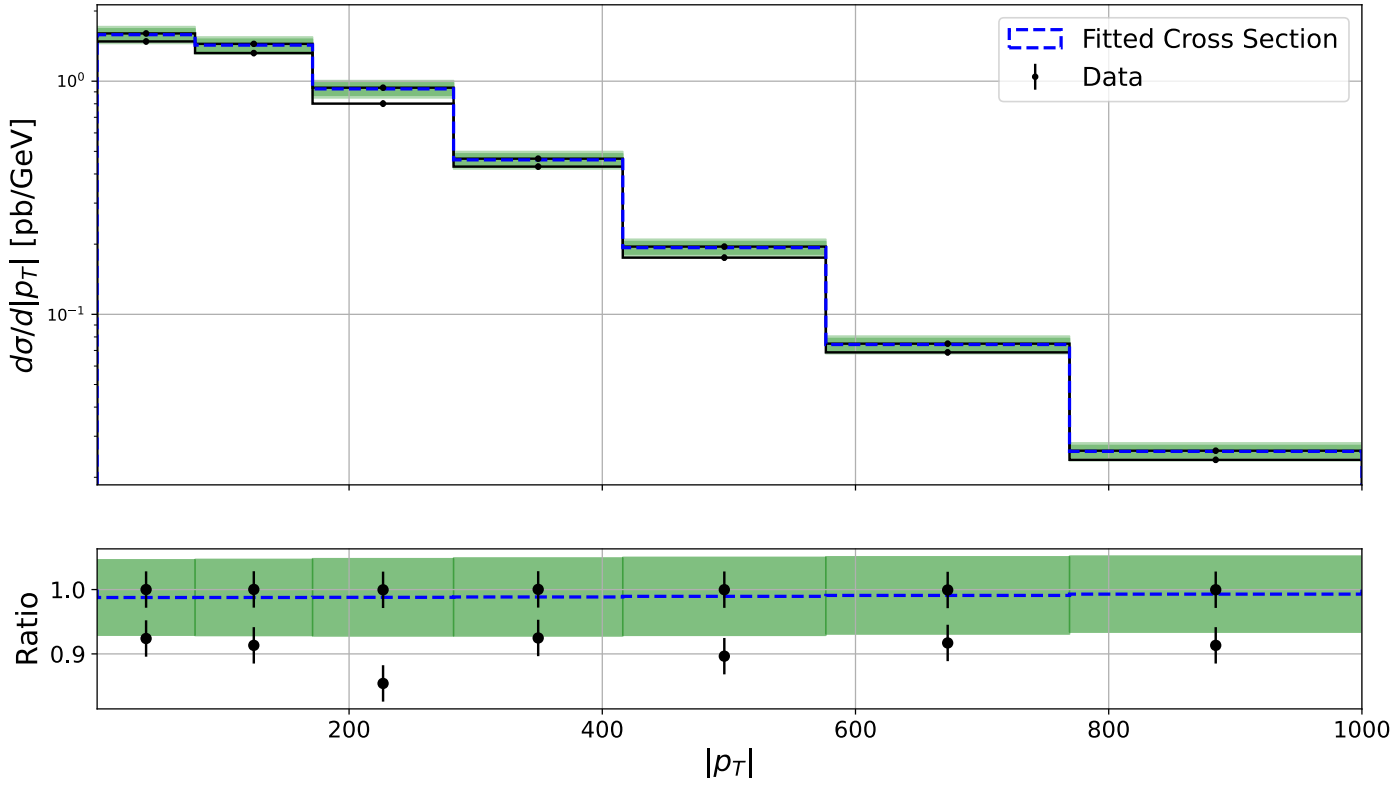
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 40%

Two cross-sections example

- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 50%

How to implement these changes in xFitter

- Treat uncertainties with ‘errors-on-errors’ as external parameters, since the minimization needs to be conducted numerically.

- Typically, only a few uncertainties will significantly impact the results if they are themselves uncertain, usually those with pulls greater than 1.

- Introduce an “eps parameter” input in the systematics namespace:
  - If eps = 0 or is not specified, use the standard quadratic term.
  - If eps > 0, treat it as an external parameter and modify the constraint in the chi2 accordingly.

Which file to modify?

For now I use this trivial approach (to be modified):

- Modify steering file

```
&Systematics
  epsilon_value = 0.61
  ListOfSources = 'sysHZComb1031:E', 'proc_tb21:E', 'sysHZComb1153:E', 'sysHZComb1098:E', 'sysHZComb1101:E'
&End
```

- Read epsilon value in read_steer.f

- Modify GetChisquare.f

```
C Correlated chi2 part:
      fcorchi2_in = 0.d0
      do k=1, NSys
      if (SysForm(k) .eq. isExternal) then
         print *, epsilon_value, epsilon_bias
         temp_val = 2 * epsilon_value**2 * (rsys_in(k) - epsilon_bias)**2 * SysPriorScale(k)
         fcorchi2_in = fcorchi2_in + (1 + 1.0D0 / (2 * epsilon_value**2)) * log(1 + temp_val)
      else
         fcorchi2_in = fcorchi2_in + rsys_in(k)**2 * SysPriorScale(k)
      endif
```

- What else should be modified?
- Am I forgetting some other bits of the code (ex. Routines to compute uncertainties, ExtraParConstr.cc, …)
- Other comments?

Outlook

- Test the impact of “errors-on-errors” on the Hera fit using the initial simple implementation (for framework testing purposes).
- Enhance the implementation method.
- Improve the computation of confidence intervals, as “errors-on-errors” needs simulations for precise confidence interval analysis.

Thank you for your attention

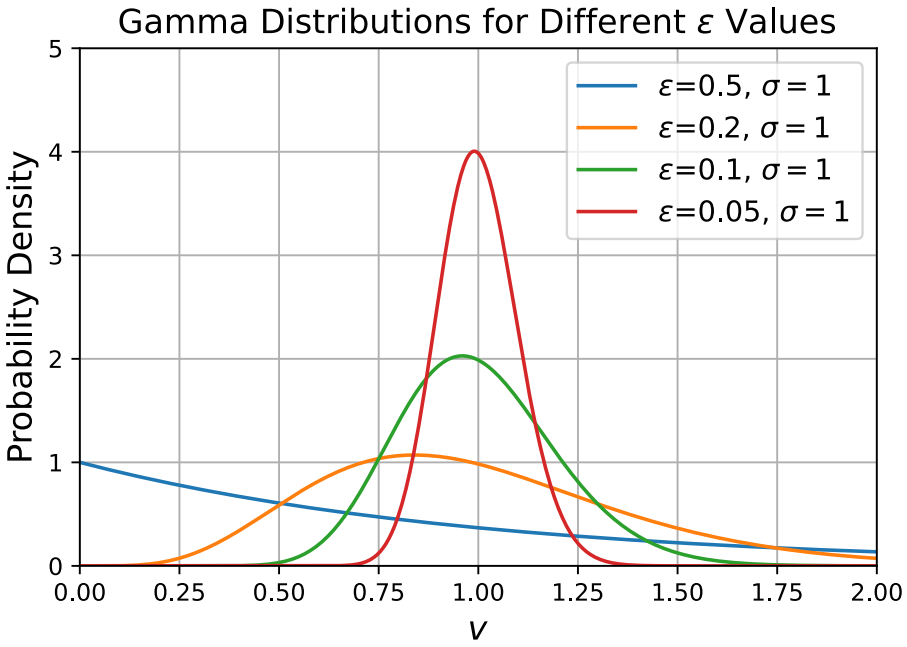
Back-up slides

Gamma Distributions

- Treat the systematic variances $\sigma_{u_i}^2$ are *adjustable parameters* (*nuisance parameter*).
- Suppose their best estimates v_i are gamma distributed:

$$v \sim \frac{\beta^{\alpha}}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}$$

$$\alpha = \frac{1}{4\varepsilon_i^2} \qquad \beta = \frac{1}{4\varepsilon_i^2 \sigma_{u_i}^2}$$



- σ_{u_i} Systematic Error
- $\varepsilon_i = \frac{1}{2}\frac{\sigma_{v_i}}{\sigma_{u_i}^2} \cong \frac{\sqrt{v_i}}{\sigma_{u_i}}$ relative error on σ_{u_i}: "***Error on error***"

Gamma Variance Model (GVM)

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- The likelihood is modified as follows:

$$L\left(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma_{u_i}^2}\right) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \times \prod_i \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\boldsymbol{\sigma_{u_i}^2}} \times \frac{\boldsymbol{\beta_i^{\alpha_i}}}{\boldsymbol{\Gamma(\alpha_i)}} \boldsymbol{v_i^{\alpha_i-1}} \boldsymbol{e^{-\beta_i v_i}}$$

- One can profile over $\boldsymbol{\sigma_{u_i}^2}$ in closed form:

$$\log L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = \log P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) - \frac{1}{2}\sum_i \left(\mathbf{1} + \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\varepsilon_i^2}}\right) \mathbf{log}\left(\mathbf{1} + \mathbf{2}\boldsymbol{\varepsilon_i^2} \frac{\boldsymbol{(u_i - \theta_i)^2}}{\boldsymbol{v_i}}\right)$$

- We call this model the Gamma Variance Model (GVM)

(see: G. Cowan, Eur. Phys. J. C (2019) 79:133; arXiv:1809.05778)

Motivation for the GVM

- Gamma distributions allow to parametrize distributions of positive defined variables (like estimates of variances)

- Using Gamma distributions it is possible to profile in close form over σ_i^2

Motivation for the GVM

- Gamma distributions include the case where the variance is estimate from a real dataset of control measurements:

$$v_i = \frac{1}{n_i - 1}\sum\left(u_{i,j} - \bar{u}_i\right)^2$$

- $(n-1)v_i/\sigma_{u_i}^2$ follows a χ_{n-1}^2 distribution and v_i a Gamma distribution with:

$$\alpha_i = \frac{n_i - 1}{2}$$

$$\beta_i = \frac{n_i - 1}{2\sigma_{u_i}^2}$$

From BLUE to the Gamma Variance Model

- BLUE (Best Linear Unbiased Estimators) approach to combinations:

$$\chi^2 = \sum_i \left(y_i - f(\boldsymbol{a})\right) V_{ij}^{-1} \left(y_j - f(\boldsymbol{a})\right)$$

$$V_{ij} = V_{ij}^{(stat)} + V_{ij}^{(syst)}$$

- $V_{ij}^{(stat)}$: Statistical covariance matrix.
- $V_{ij}^{(syst)}$: Covariance matrix induced by systematic source.
- $V_{ij}^{(syst)} = \sum_s V_{ij}^{(s)}$

From BLUE to the Gamma Variance Model

- Nuisance parameters approach:

Magnitude of the bias induced by the systematic source *s*

$$\chi^2 = \sum_i \frac{(y_i - f(\boldsymbol{a}) - \sum_s \Gamma_i^s \theta_s)^2}{\sigma_i^2} + \sum_s \theta_S^2$$

From BLUE to the Gamma Variance Model

- Nuisance parameters approach:

Magnitude of the bias induced by the systematic source *s*

$$\chi^2 = \sum_i \frac{(y_i - f(\boldsymbol{a}) - \sum_s \Gamma_i^s \theta_s)^2}{\sigma_i^2} + \sum_s \theta_S^2$$

- Connection:

$$V_{ij}^{(syst)} = \sum_s V_{ij}^{(s)}$$

$$V_{ij}^{(s)} = \Gamma_i^s \Gamma_j^s$$

- Gamma Variance Model:

$$\chi^2 = \sum_i \frac{(y_i - f(\boldsymbol{a}) - \sum_s \Gamma_i^s \theta_s)^2}{\sigma_i^2} + \sum_s \left(1 + \frac{1}{2\boldsymbol{\varepsilon}_i^2}\right) \log\left(1 + 2\boldsymbol{\varepsilon}_i^2 \boldsymbol{\theta}_i{}^2\right)$$

- What to do if we do not have access to the factors Γ_i^s (we only know $V_{ij}^{(syst)}$)?

✓ $V_{ij}^{(syst)} = \sum_s V_{ij}^{(s)}$ ✗

From BLUE to the Gamma Variance Model

- Switch to a nuisance parameters approach:

Proof is non-trivial!

$$\chi^2 = \sum_i \frac{(y_i - \mu - \theta_i)^2}{\sigma_i^2} + \sum_{ij} \theta_i C_{ij}^{-1} \theta_j$$

$$C_{ij} = V_{ij}^{(s)}$$

- Substitute quadratic term with log-constraint:

$$\sum_{ij} \boldsymbol{\theta_i C_{ij}^{-1} \theta_j} \longrightarrow \sum_i \left(\boldsymbol{N + \frac{1}{2\varepsilon_i^2}} \right) \boldsymbol{\log\left(1 + 2\varepsilon_i^2 \theta_i C_{ij}^{-1} \theta_j\right)}$$

Bartlett Correction

- The Hessian method is based on the assumption that the χ^2 follows a χ^2 distribution.

- Our "goodness-of-fit" statistics *q is* is not a χ^2 so will will not follow exactly a χ^2 for large values of ϵ^2

Large literature on the topic:

- *Bartlett, M. S. (1937) Proceedings of the Royal Society A, 160, 268–282)*
- *Applied Asymptotics Case Studies in Small-Sample Statistics by A. R. Brazzale, A. C. Davison and N. Reid)*
- Canonero, E., Brazzale, A.R. & Cowan, *Eur. Phys. J. C* **83**, 1100 (2023).

Bartlett Correction

- Modify the test statistic q so that its distribution is closer to a χ^2 :

$$q \longrightarrow q^* = q\frac{N_{dof}}{E[q]}$$

Bartlett Correction

- Modify the test statistic q so that its distribution is closer to a χ^2 :

$$q \longrightarrow q^* = q\frac{N_{dof}}{E[q]}$$

Expectation value in the asymptotic limit (degrees of freedom of χ^2)

Exact expectation value

Bartlett Correction

- Modify the test statistic q so that its distribution is closer to a χ^2 :

$$q \longrightarrow q^* = q\frac{N_{dof}}{E[q]}$$

Expectation value in the asymptotic limit (degrees of freedom of χ^2)

Exact expectation value

$$q \sim \chi^2 + \mathcal{O}(\epsilon^2)$$

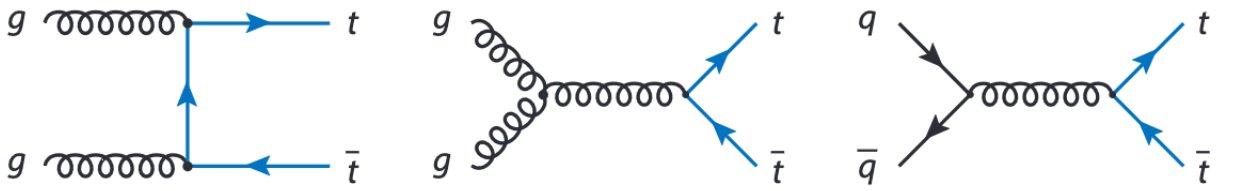
$$q^* \sim \chi^2 + \mathcal{O}(\epsilon^4)$$

Simplified Model (no real data)

GOAL:

- Construct a simplified toy model to test the implementations of errors-on-errors in a real PDF fit
- Choose a simple process that allows an easy and fast implementation.

$gg \to t\bar{t}$ **LO cross section:**

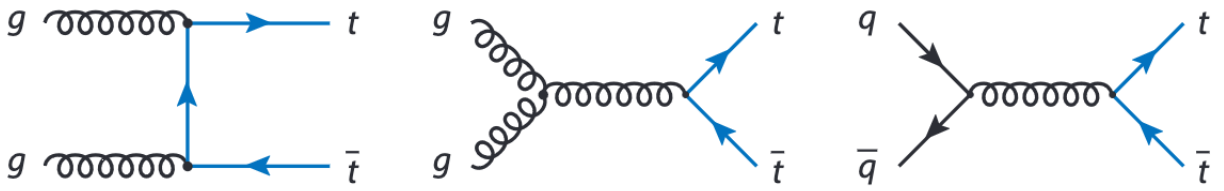


Simplified Model (no real data)

GOAL:

- Construct a simplified toy model to test the implementations of errors-on-errors in a real PDF fit
- Choose a simple process that allows an easy and fast implementation.

$gg \to t\bar{t}$ **LO cross section:**



$$\frac{d\hat{\sigma}}{dcos\theta} = \frac{\alpha_s^2}{32s}\sqrt{1-\frac{4m_t^2}{s}}\frac{7m_t^4 - 7m_t^2(t+u) + 4t^2 - tu + 4u^2}{3s^2(m_t^2-t)^2(m_t^2-u)^2}(tu(t^2+u^2)$$
$$- 6m_t^8 + m_t^4(3t^2 + 14tu + 3u^2) - m_t^2(t+u)(t^2+6tu+u^2))$$

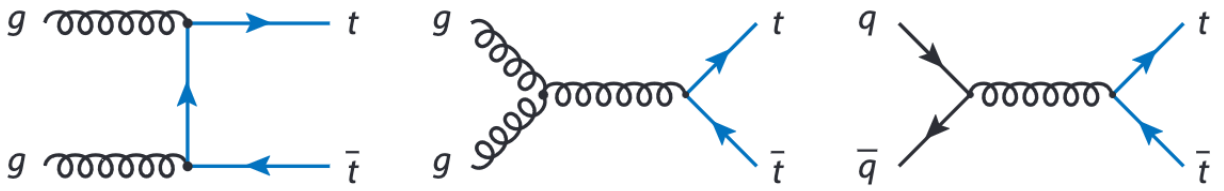
$$\frac{d\sigma_{pp}}{dx_1 dx_2 dcos\theta} = \mathrm{g}(\mathrm{x}_1)\mathrm{g}(\mathrm{x}_2)\frac{d\hat{\sigma}}{dcos\theta}$$

Simplified Model (no real data)

GOAL:

- Construct a simplified toy model to test the implementations of errors-on-errors in a real PDF fit
- Choose a simple process that allows an easy and fast implementation.

$gg \to t\bar{t}$ **LO cross section:**



$$\frac{d\hat{\sigma}}{dcos\theta} = \frac{\alpha_s^2}{32s}\sqrt{1-\frac{4m_t^2}{s}}\frac{7m_t^4 - 7m_t^2(t+u) + 4t^2 - tu + 4u^2}{3s^2(m_t^2-t)^2(m_t^2-u)^2}(tu(t^2+u^2)$$
$$- 6m_t^8 + m_t^4(3t^2 + 14tu + 3u^2) - m_t^2(t+u)(t^2+6tu+u^2))$$

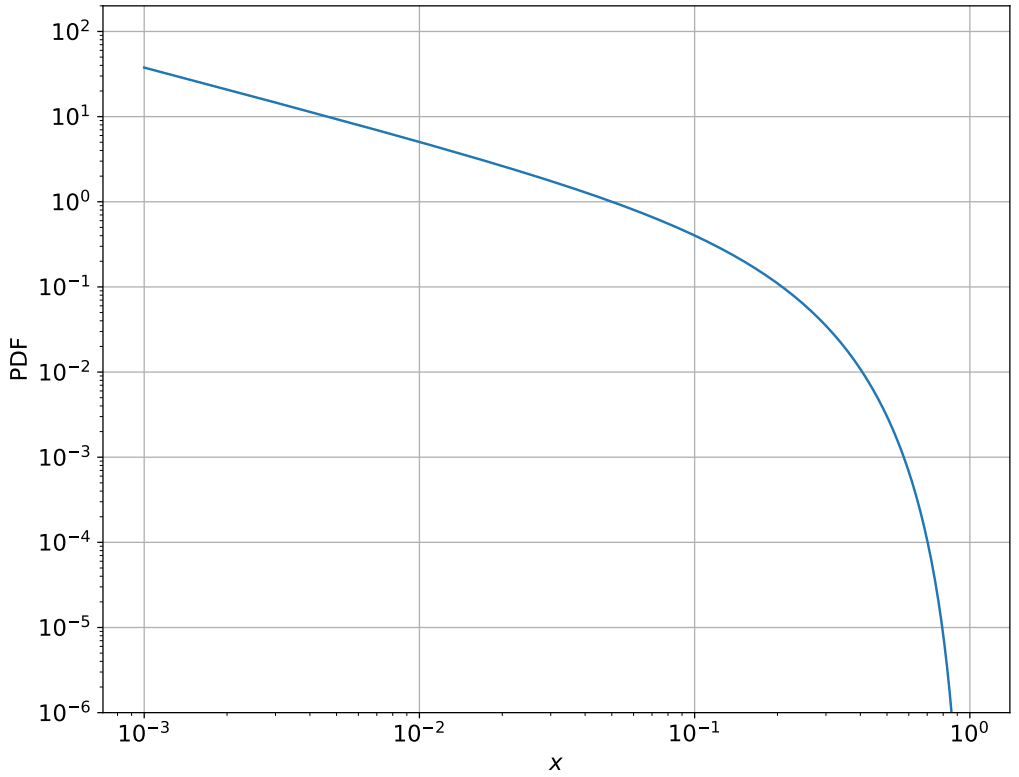
$$\frac{d\sigma_{pp}}{dx_1 dx_2 dcos\theta} = \mathrm{g}(\mathrm{x}_1)\mathrm{g}(\mathrm{x}_2)\frac{d\hat{\sigma}}{dcos\theta}$$

Use this to compute differential observables of the $t\bar{t}$ system.

Simplified Model

- The aim of the exercise is to fit the gluon PDF, using fictious data points.

- The gluon PDF is parametrized as follow
  - $g(x) = Cx^A(1-x)^B$
  - $\begin{cases} A = -0.85 \\ B = 6 \end{cases}$
  - $C : \int_0^1 g(x)dx = 1/2$



- We are assuming that this is the gluon PDF shape at Q^2 close to $t\bar{t}$ production scale.