Pythia Week
30/04/24

ROYAL HOLLOWAY UNIVERSITY OF LONDON

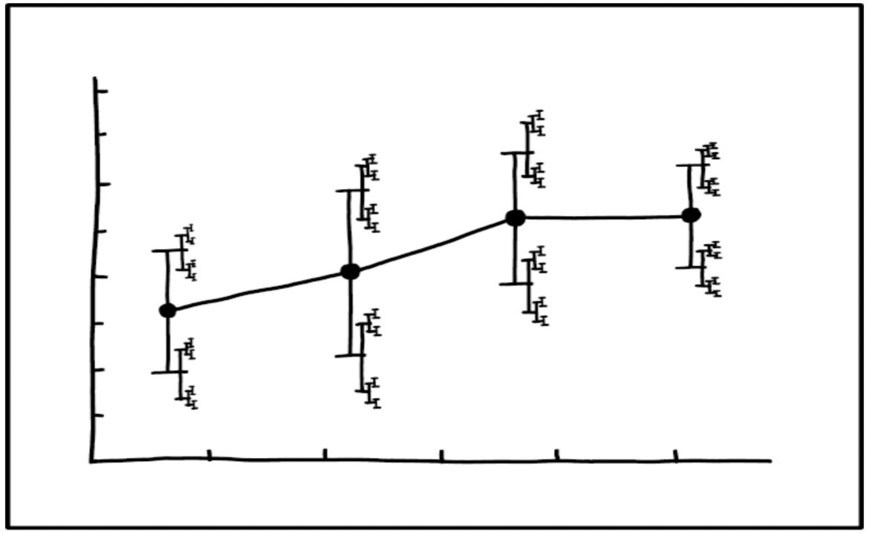
Uncertain systematics

Enzo Canonero

Glen Cowan

Motivation

https://xkcd.com/2110/



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**
- **…**

Goal: Show how uncertain systematics can be implemented in a fit.

Non-trivial consequences!

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)

Formulation of the problem

ROYAL HOLLOWAY UNIVERSITY OF LONDON

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

$$\mathrm{E}[\boldsymbol{y}] = f(\boldsymbol{\mu}) + \sum \theta_i$$

Formulation of the problem

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

$$\mathrm{E}[\boldsymbol{y}] = f(\boldsymbol{\mu}) + \sum \theta_i$$

- Nuisance parameters are used to model systematic effects and are constrained by auxiliary measurements $\boldsymbol{u}$

- The us are assumed to be independently Gaussian distributed

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

Formulation of the problem

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

$$\mathrm{E}[\boldsymbol{y}] = f(\boldsymbol{\mu}) + \sum \theta_i$$

- Nuisance parameters are used to model systematic effects and are constrained by auxiliary measurements $\boldsymbol{u}$
- The us are 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

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

- So, if the likelihood is

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

- The resulting log Likelihood will be:

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

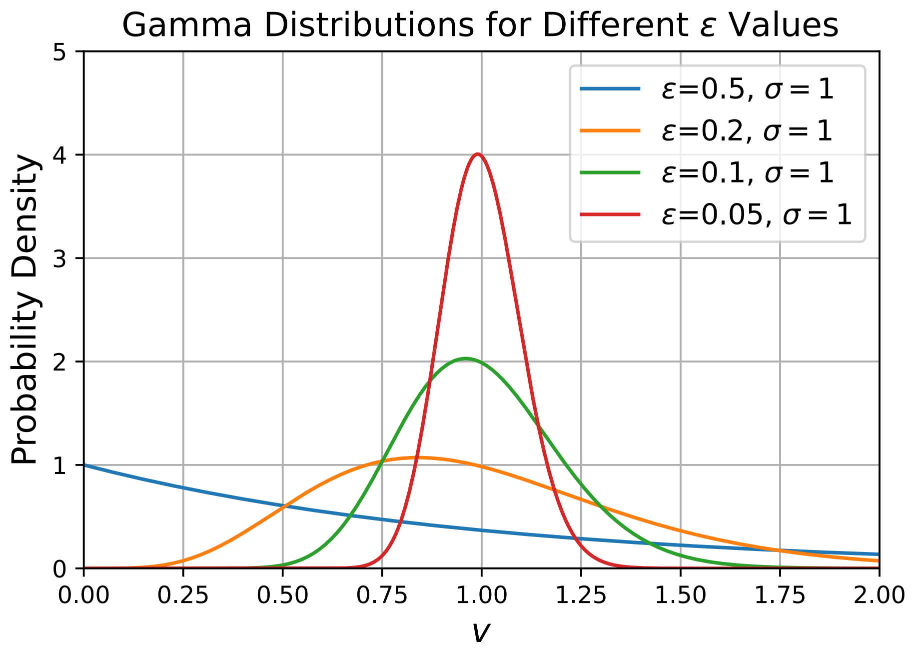
Let systematic errors be potentially uncertain!

Gamma distribution

To implement "errors-on-errors" suppose the systematic variances $\boldsymbol{\sigma_{u_i}^2}$ are *adjustable parameters*, and their best estimates $\boldsymbol{v_i}$ are gamma distributed:

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

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



- $\boldsymbol{\sigma_{u_i}^2}$ Expectation value of $\boldsymbol{v_i}$
- $\boldsymbol{\varepsilon_i}$: relative error on $\boldsymbol{\sigma_{u_i}}$: "***Error on error***"*

*$\boldsymbol{\varepsilon}$ used to be r in previous references

Gamma Variance Model (GVM)

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- The likelihood is modified as follows:

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

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

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

Gamma Variance Model (GVM)

- The likelihood is modified as follows:

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{\boldsymbol{u}_i}^2) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \times \prod_i \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\boldsymbol{\sigma}_{\boldsymbol{u}_i}^2} \times \frac{\boldsymbol{\beta}_i^{\boldsymbol{\alpha}_i}}{\boldsymbol{\Gamma}(\boldsymbol{\alpha}_i)} \boldsymbol{v}_i^{\boldsymbol{\alpha}_i - \mathbf{1}} \boldsymbol{e}^{-\boldsymbol{\beta}_i \boldsymbol{v}_i}$$

- One can profile over $\boldsymbol{\sigma}_{\boldsymbol{u}_i}^2$ in closed form:

$$\log L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = \log P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) - \frac{1}{2} \sum_i \left(\mathbf{1} + \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\varepsilon}_i^2}\right) \mathbf{log} \left(\mathbf{1} + \mathbf{2}\boldsymbol{\varepsilon}_i^2 \frac{(\boldsymbol{u}_i - \boldsymbol{\theta}_i)^2}{\boldsymbol{v}_i}\right)$$

- Profiling means computing

$$L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = L\left(\boldsymbol{\mu}, \boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\sigma}_{\boldsymbol{u}_i}^2}}\right), \qquad \widehat{\widehat{\boldsymbol{\sigma}_{\boldsymbol{u}_i}^2}} = argmax_{\sigma_{u_i}^2}\left(L(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{\boldsymbol{u}_i}^2)\right)$$

Gamma Variance Model (GVM)

ROYAL HOLLOWAY UNIVERSITY OF LONDON

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

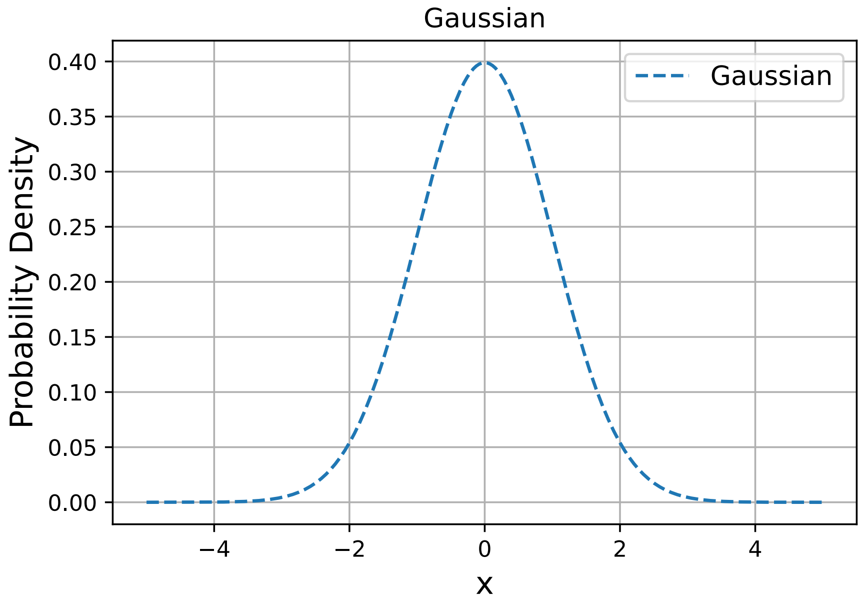
$$\sum_i \frac{(u_i - \theta_i)^2}{2\sigma_{u_i}^2} \longrightarrow \sum_i \frac{1}{2}\left(1 + \frac{1}{2\varepsilon_i^2}\right)\log\left(1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right)$$

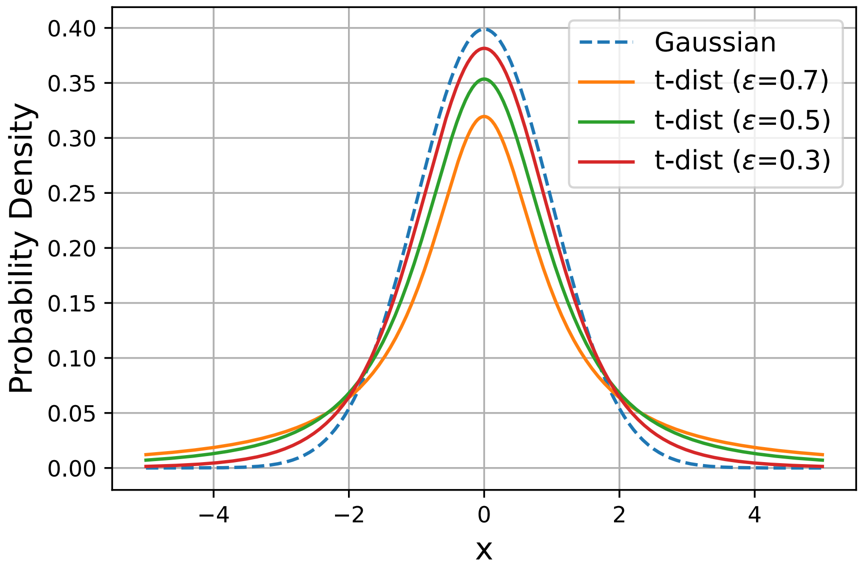
Gamma Variance Model (GVM)

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

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

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



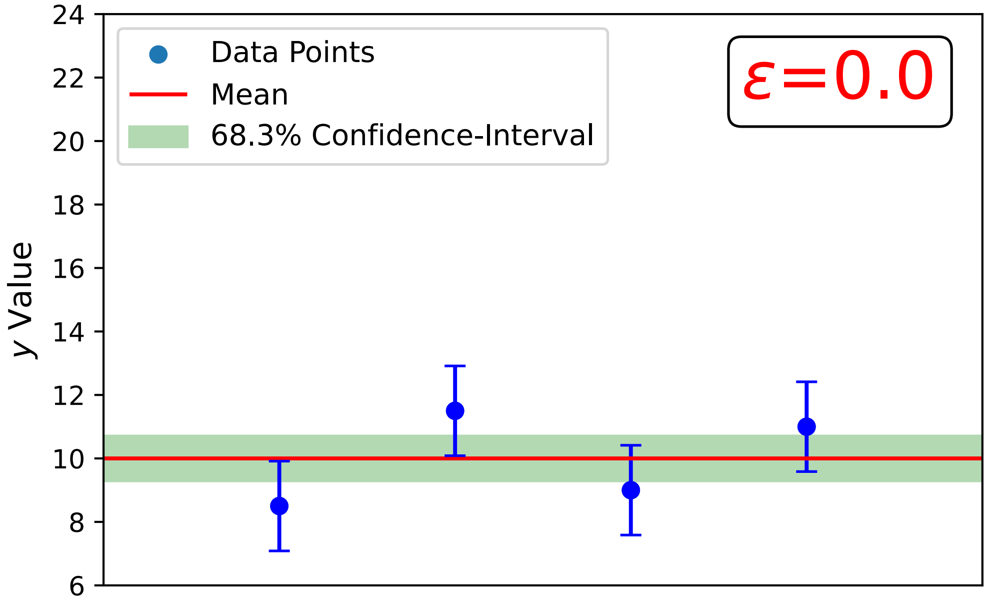


Sensitivity to outliers

- 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

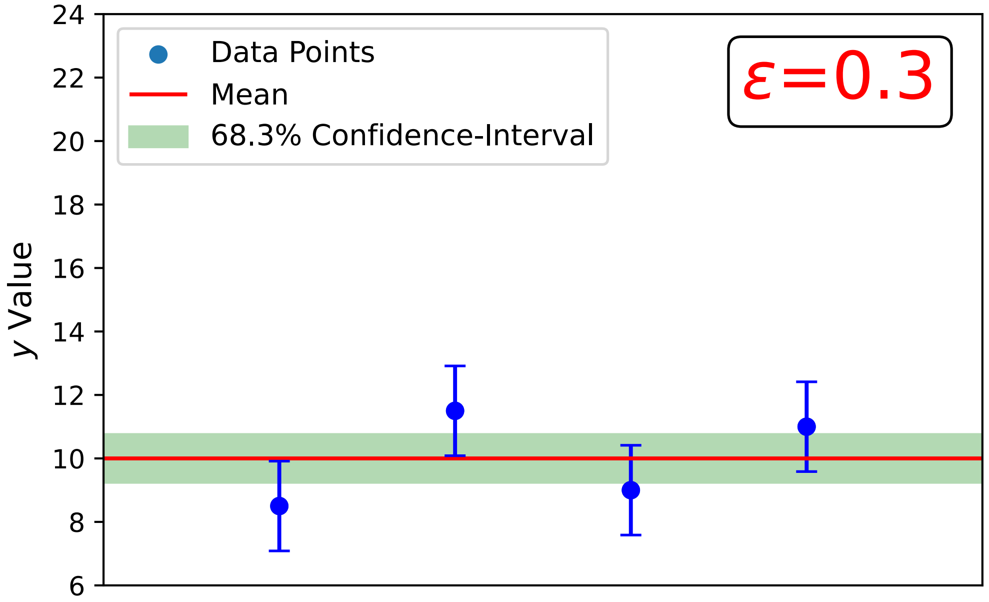


Sensitivity to outliers

- 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

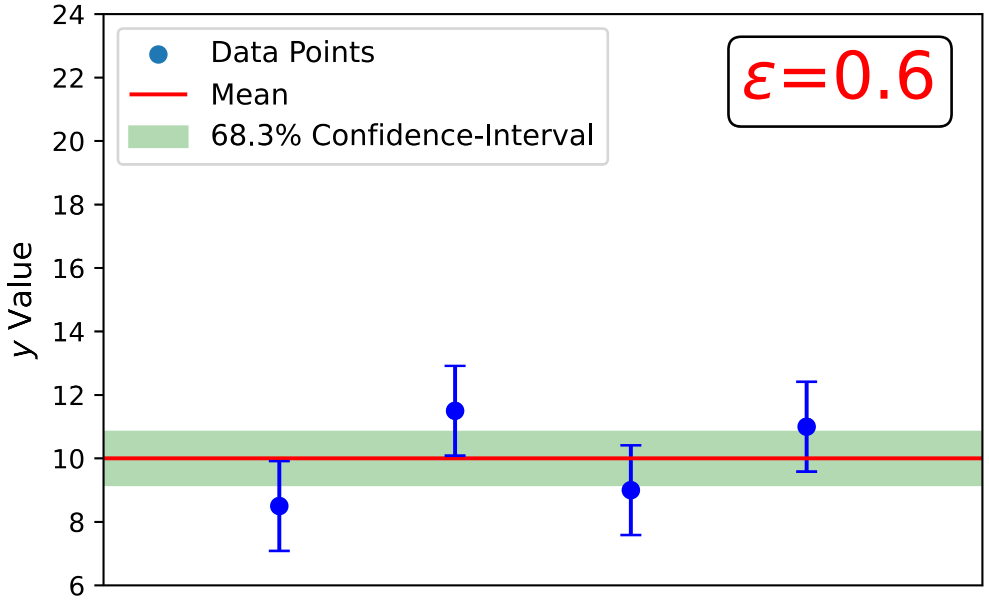


Sensitivity to outliers

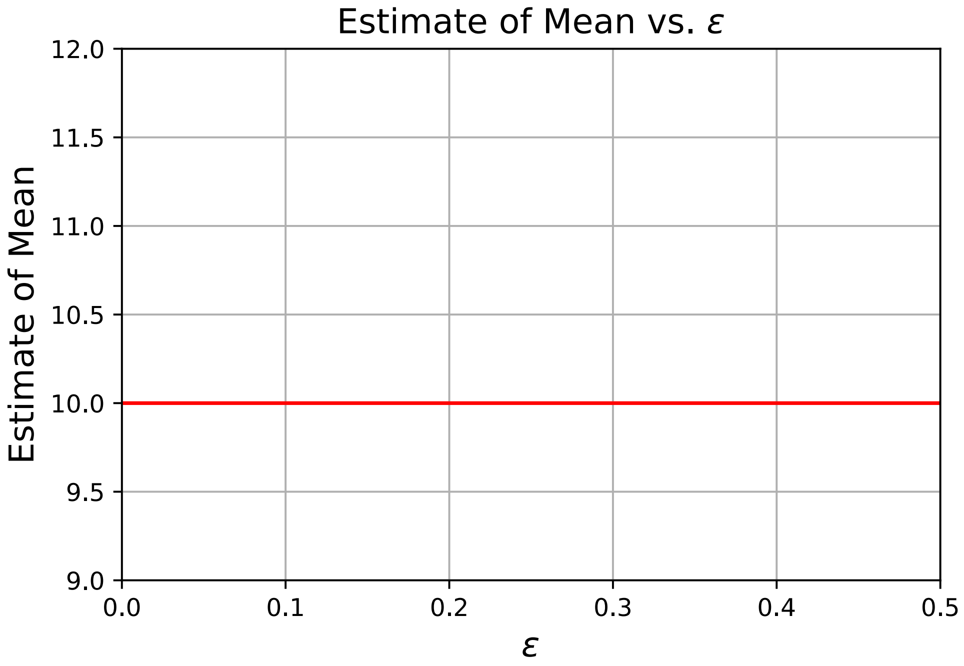
- Suppose we want to average 4 measurements all with **statistical** and **syst errors** equal to **1**. Also assume they all have equal ***errors-on-errors*** $\boldsymbol{\varepsilon}$ (auxiliary measurements set to zero):

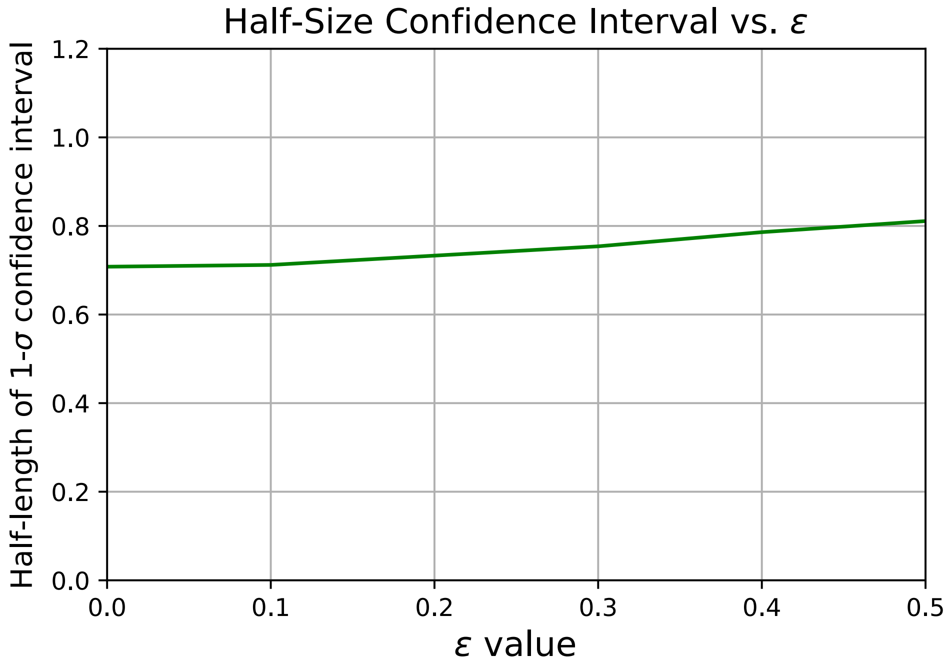
$$\log L_P(\boldsymbol{\mu}, \boldsymbol{\theta}) = -\frac{1}{2}\sum_i \frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2} - \frac{1}{2}\sum_i \left(1 + \frac{1}{2\boldsymbol{\varepsilon}_i^2}\right) \log\left(1 + 2\boldsymbol{\varepsilon}_i^2 \frac{{\theta_i}^2}{\sigma_{u_i}^2}\right)$$

Measurements internally compatible



Sensitivity to outliers

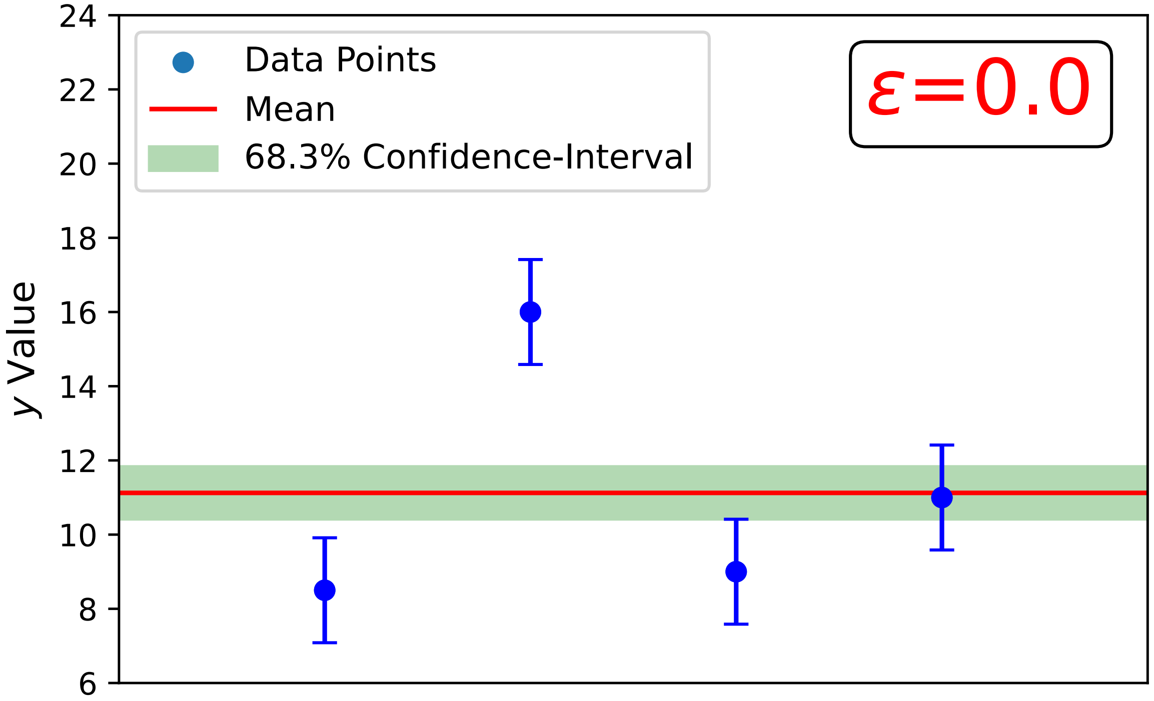




1. The estimate of the mean does not change when we increase ε
2. The size of the confidence interval for the mean only slightly increases, reflecting the extra degree of uncertainty introduced by errors-on-errors
3. If data are internally compatible results are only slightly modified

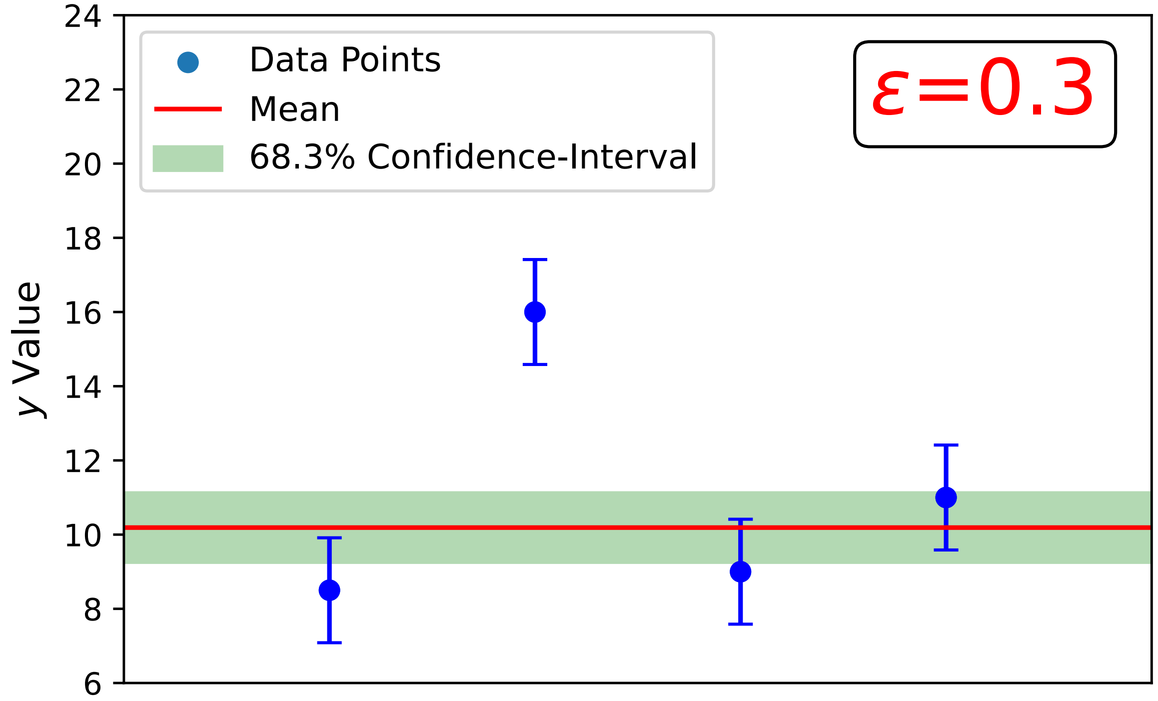
Sensitivity to outliers

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



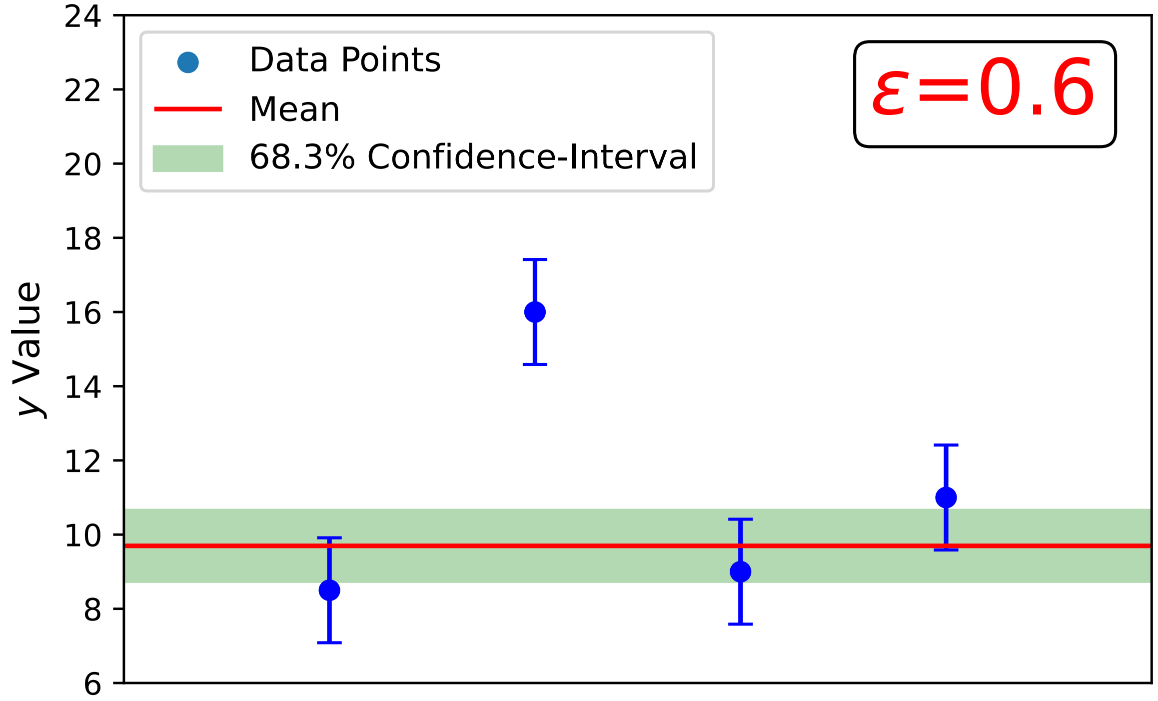
Sensitivity to outliers

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

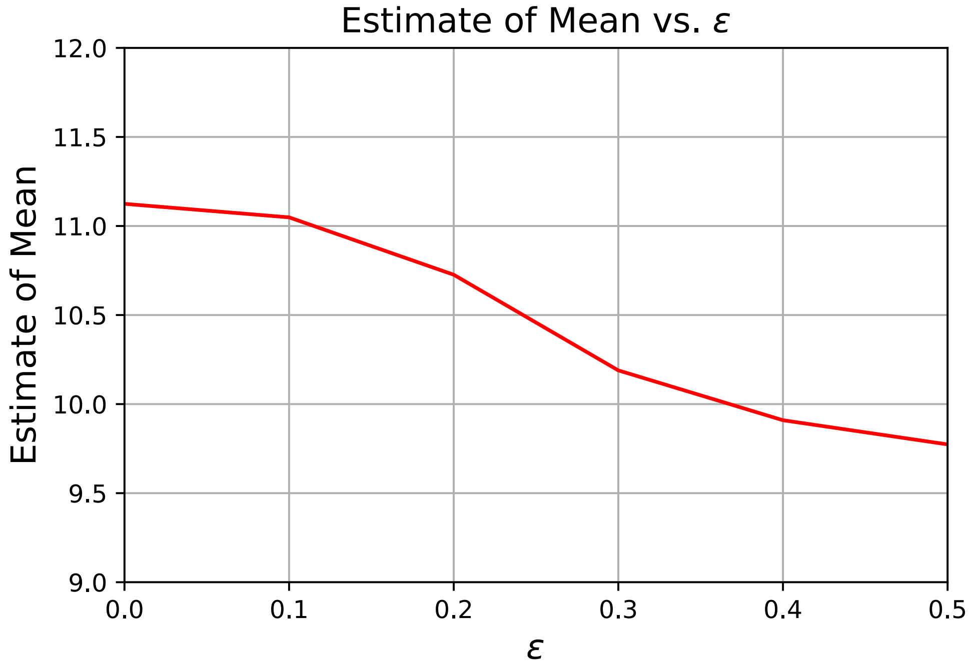


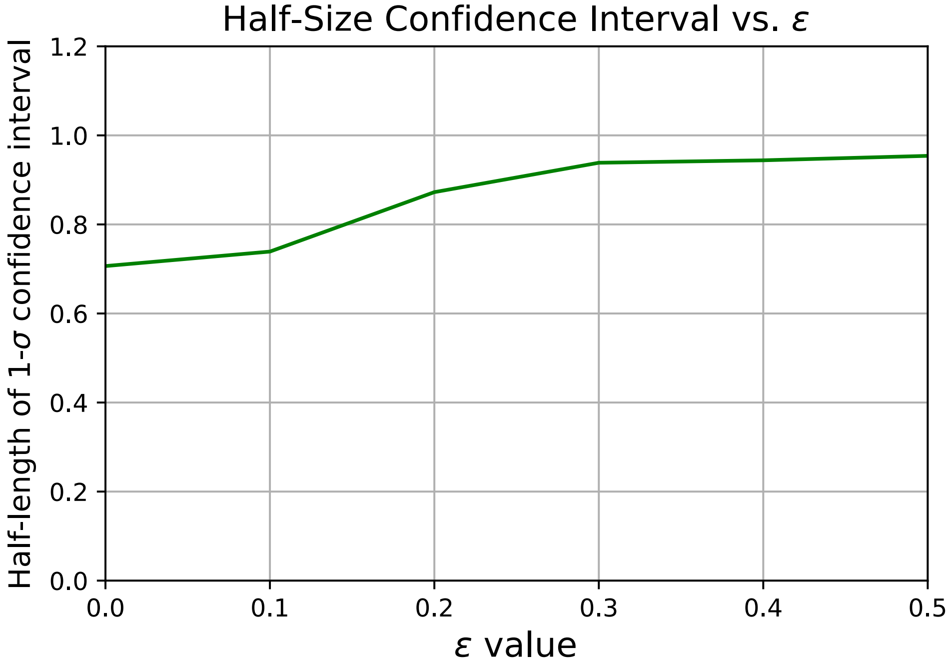
Sensitivity to outliers

- 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

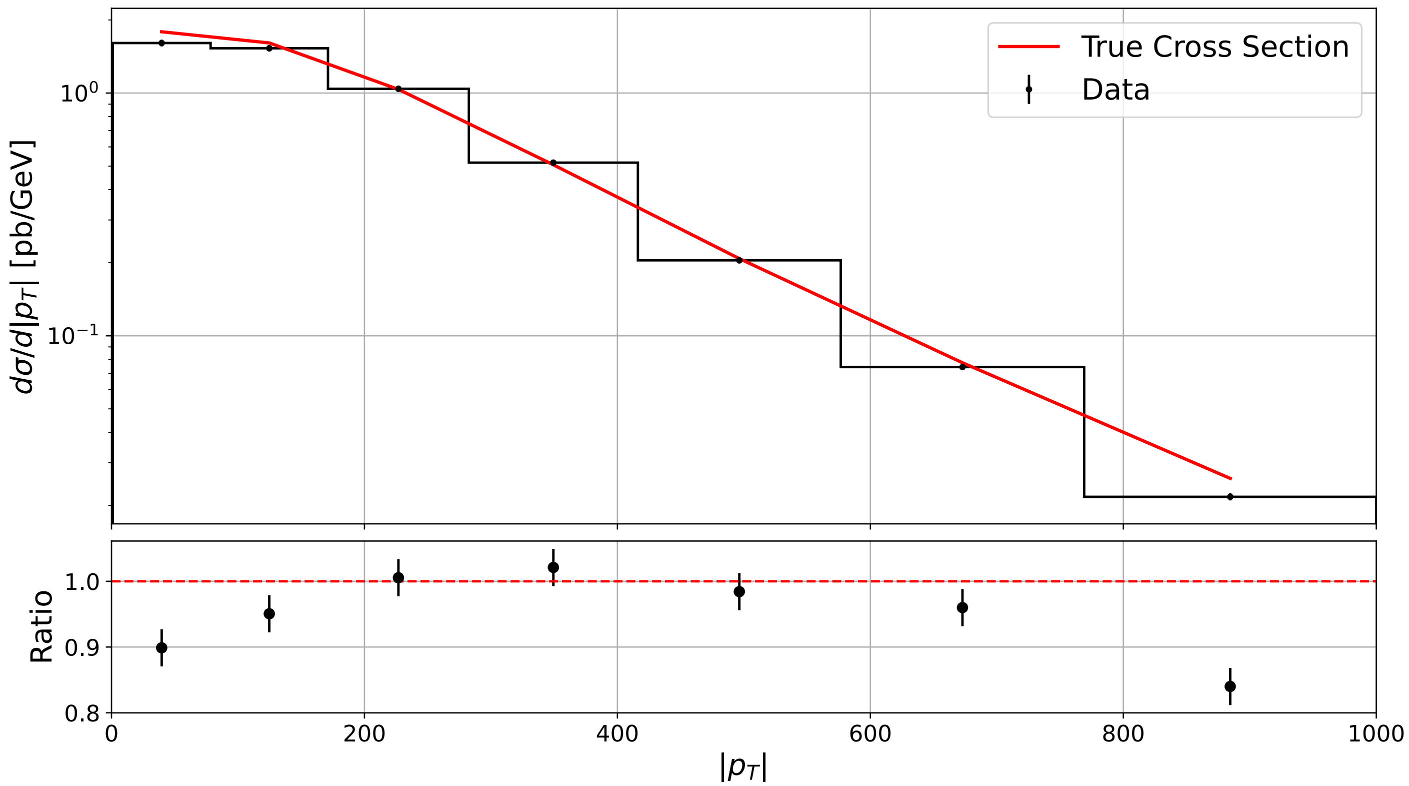
2. The error bar grows more significantly: the GVM treats internal incompatibility as an additional source of uncertainty

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

Realistic fit example (from PDF fitting)

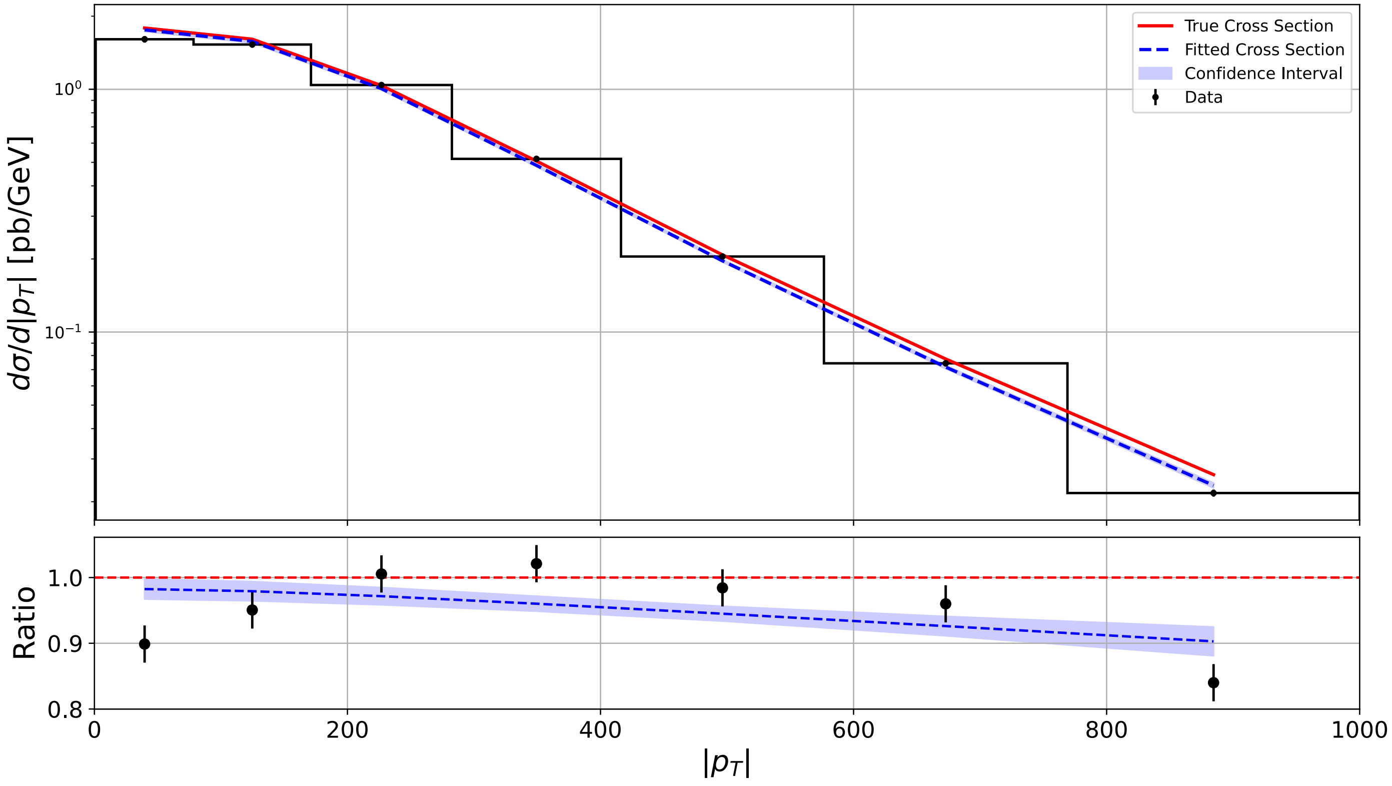
Goal: fit the parameters of a complicated non-linear function using a differential distribution.
(a differential cross-section from a PDF fit example)

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



Realistic fit example (from PDF fitting)

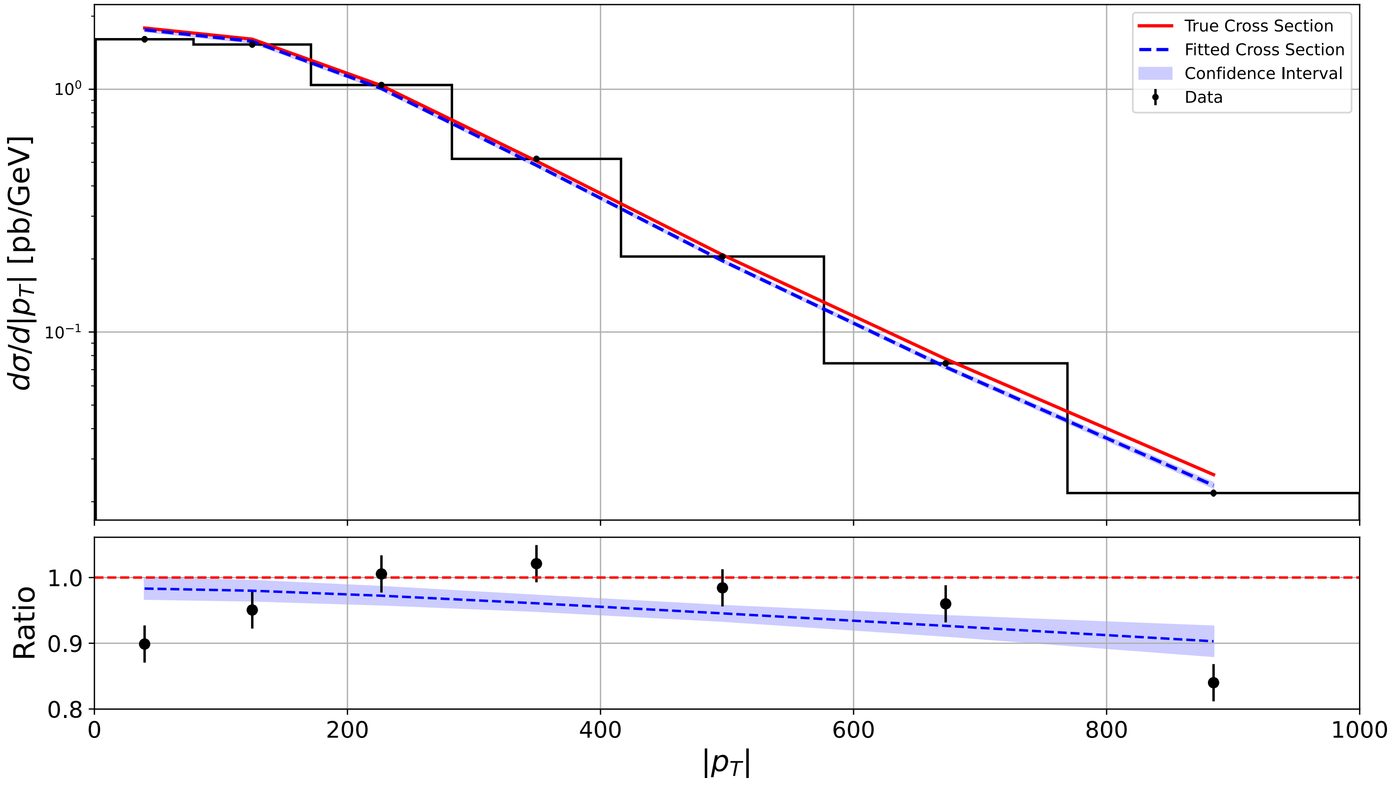
- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements



Errors-on-errors: 0%

Realistic fit example (from PDF fitting)

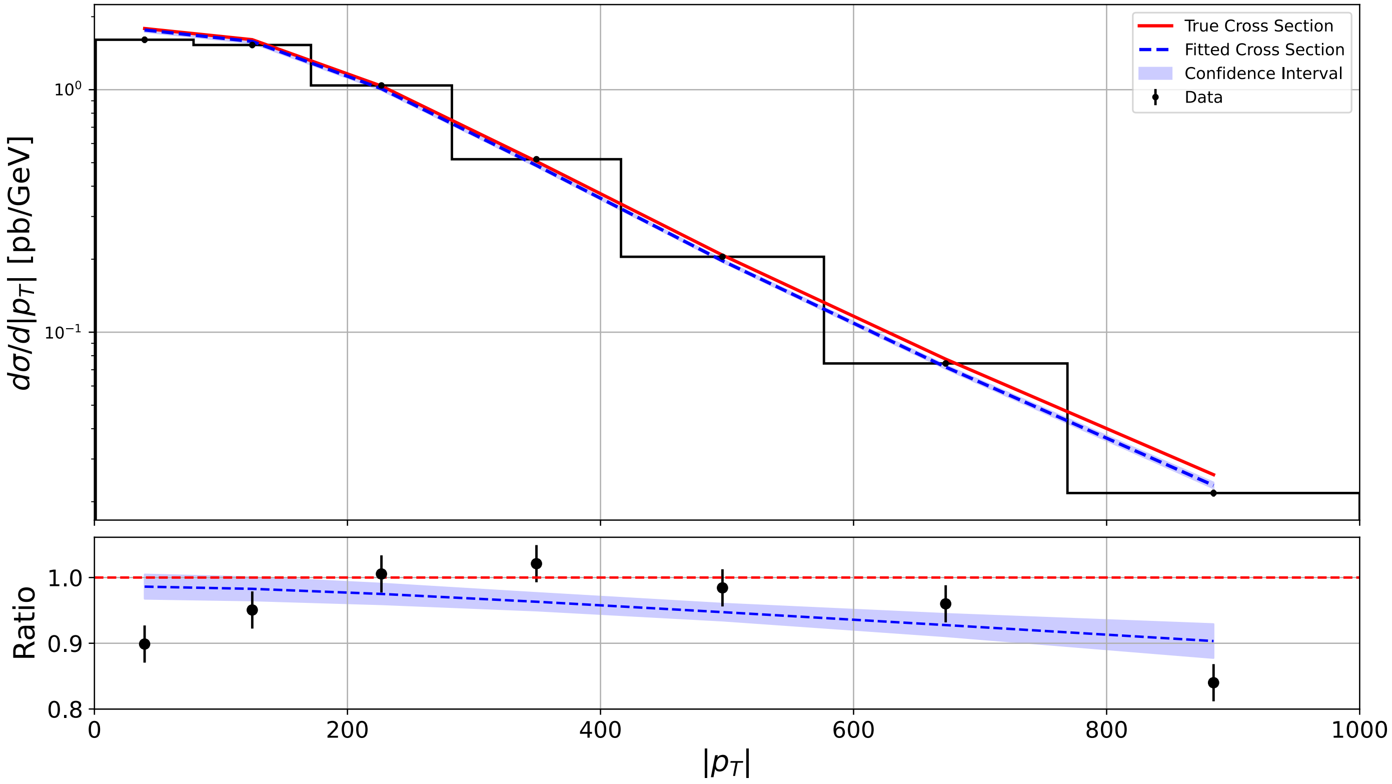
- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements



Errors-on-errors: 10%

Realistic fit example (from PDF fitting)

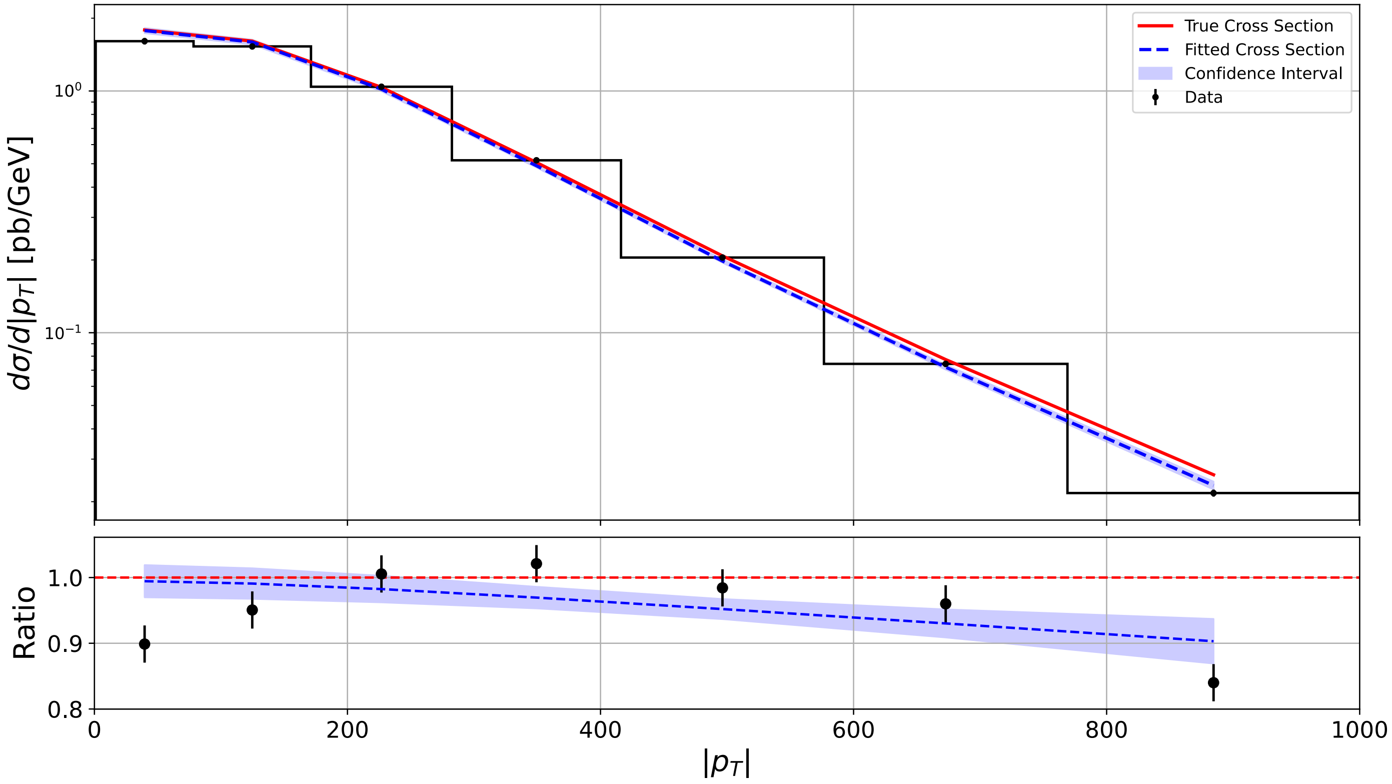
- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements



Errors-on-errors: 20%

Realistic fit example (from PDF fitting)

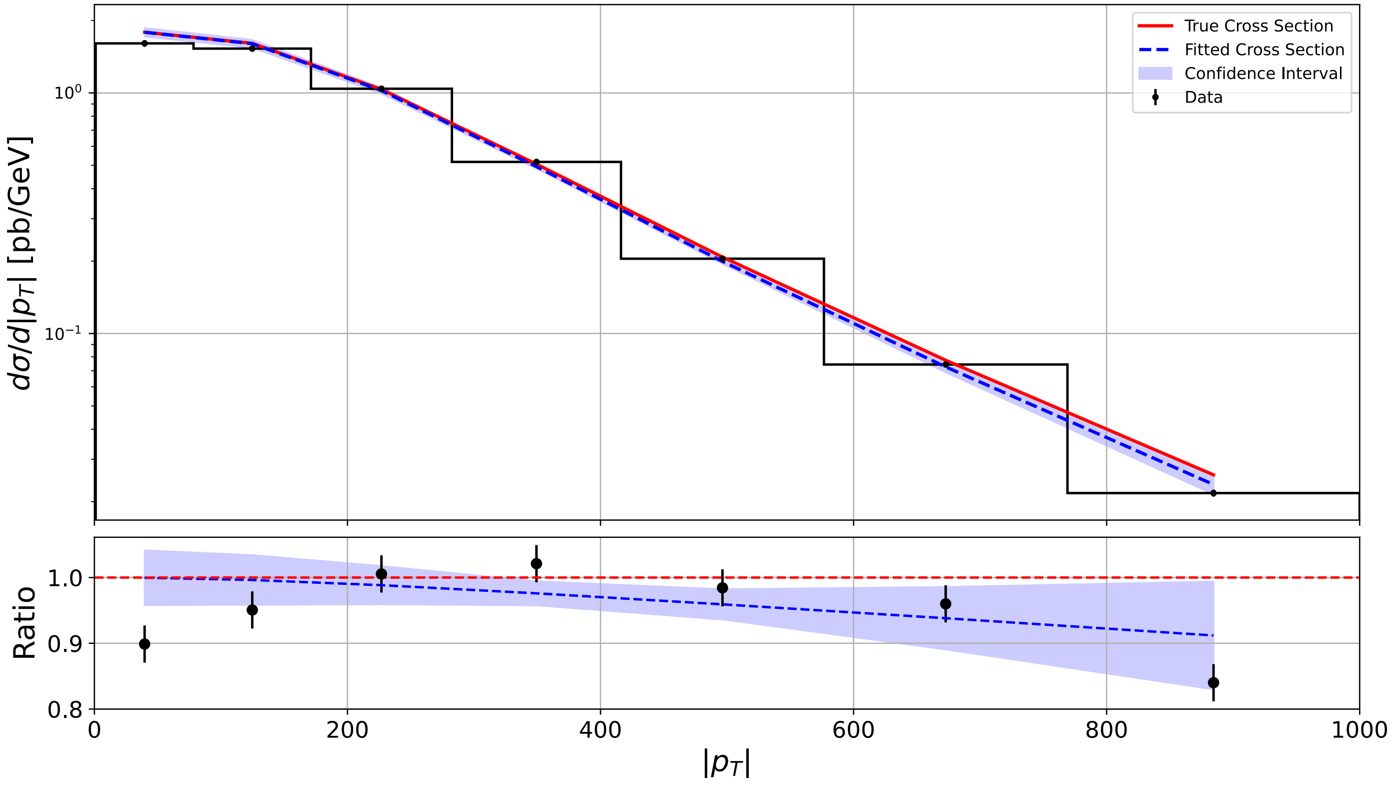
- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements



Errors-on-errors: 30%

Realistic fit example (from PDF fitting)

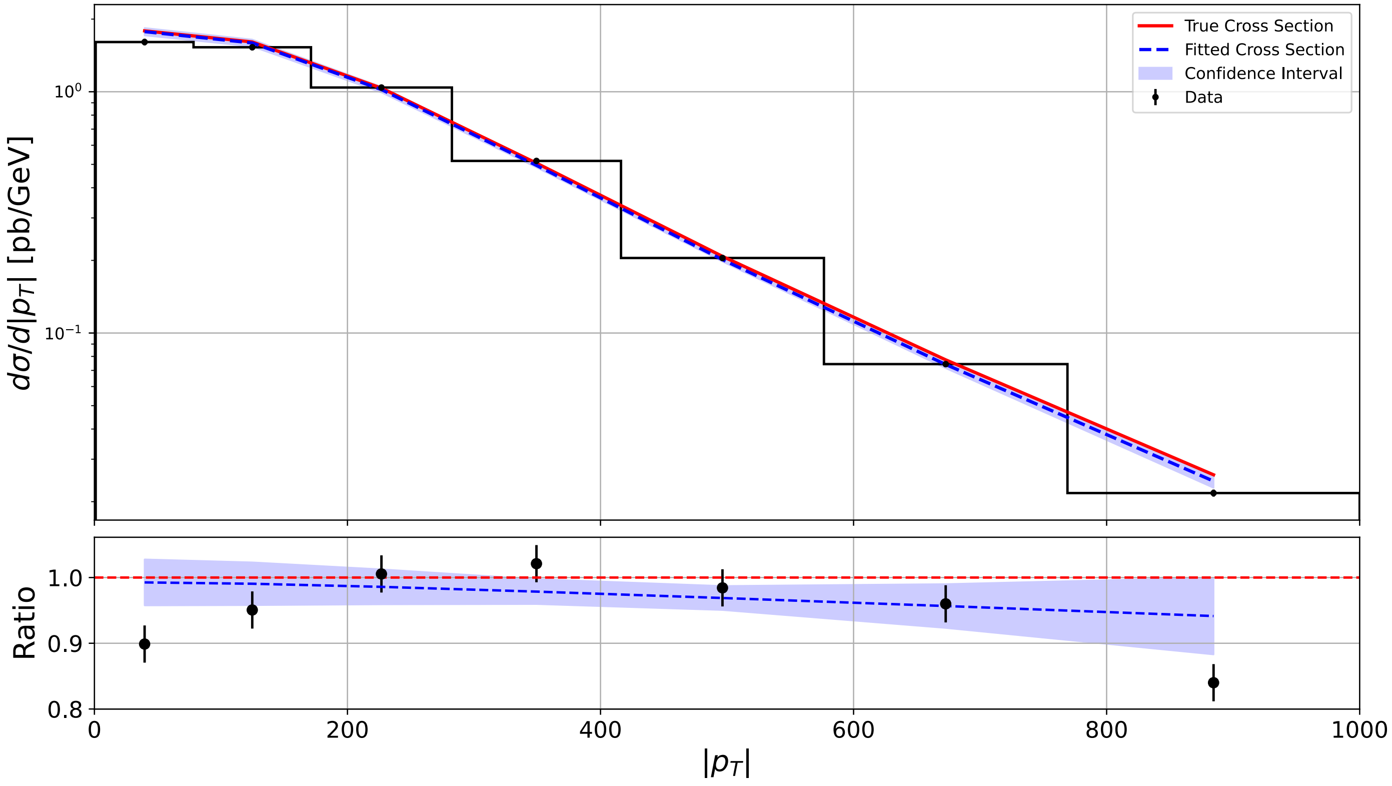
- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements



Errors-on-errors: 40%

Realistic fit example (from PDF fitting)

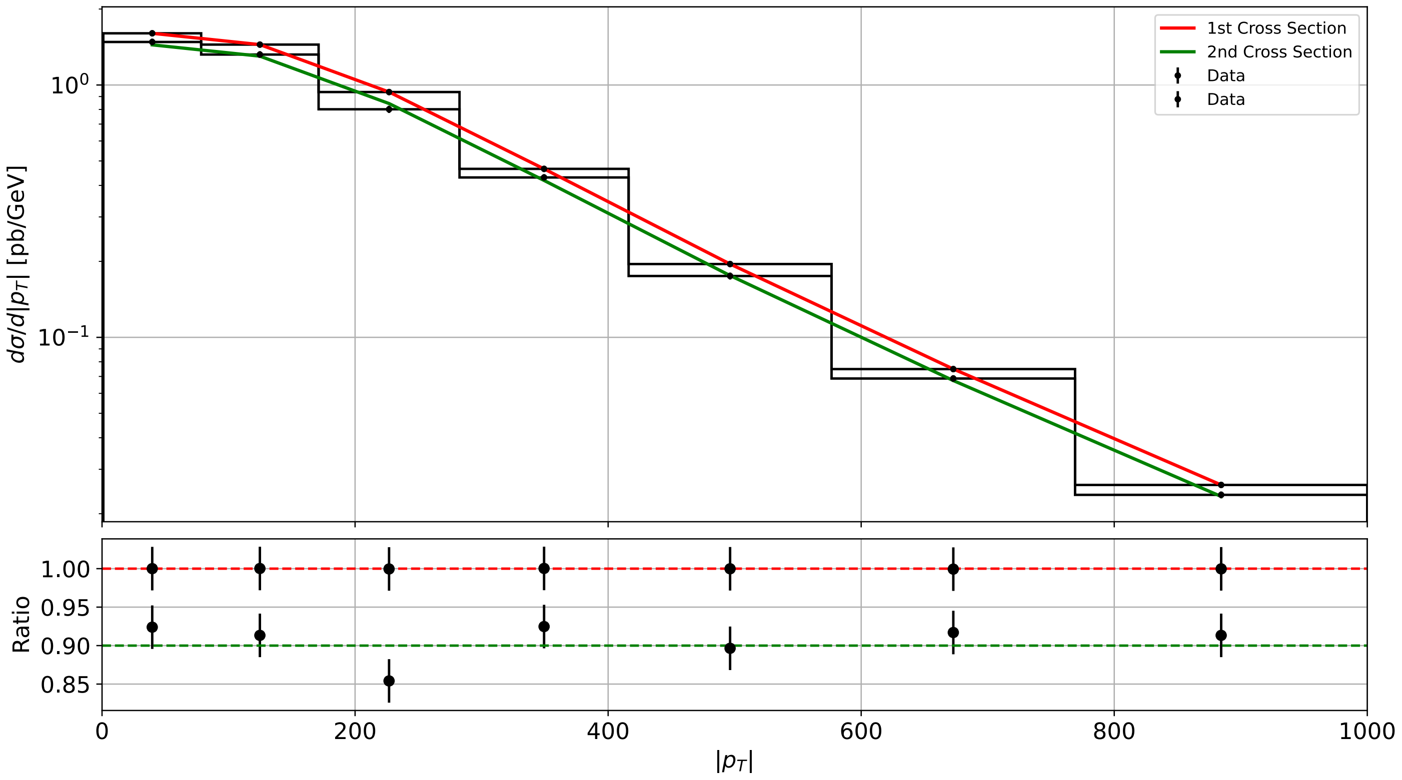
- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements



Errors-on-errors: 50%

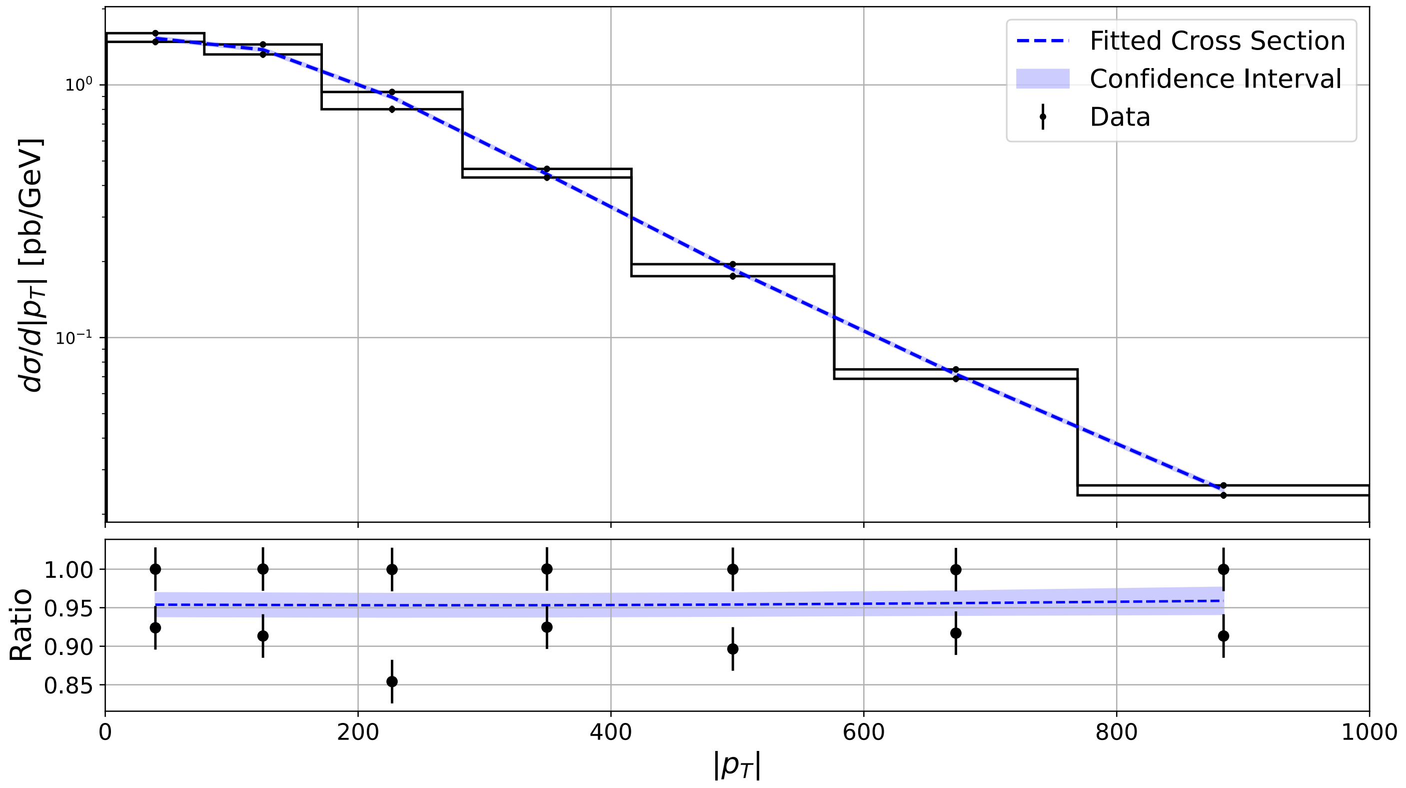
Scale factor example

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.



Scale factor example

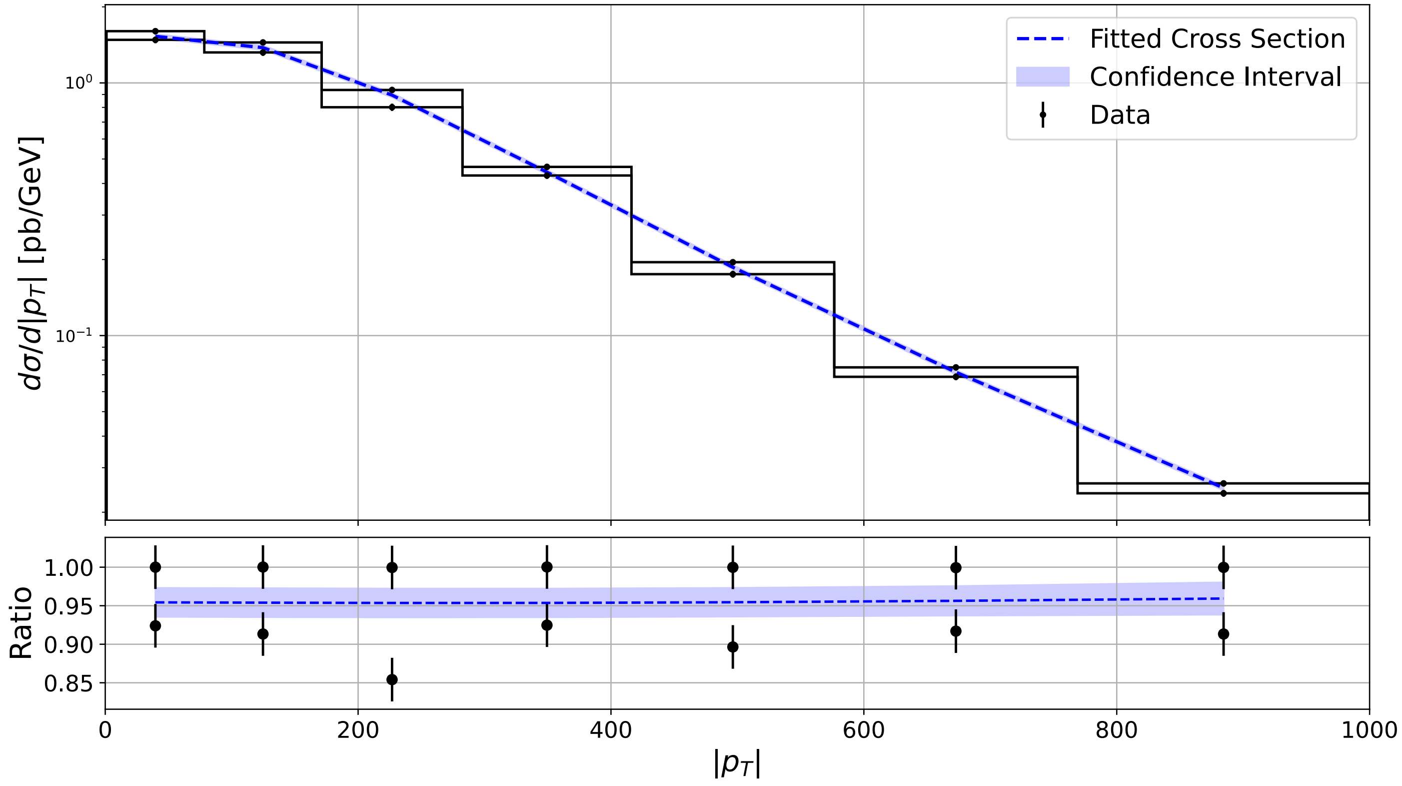
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 0%

Scale factor example

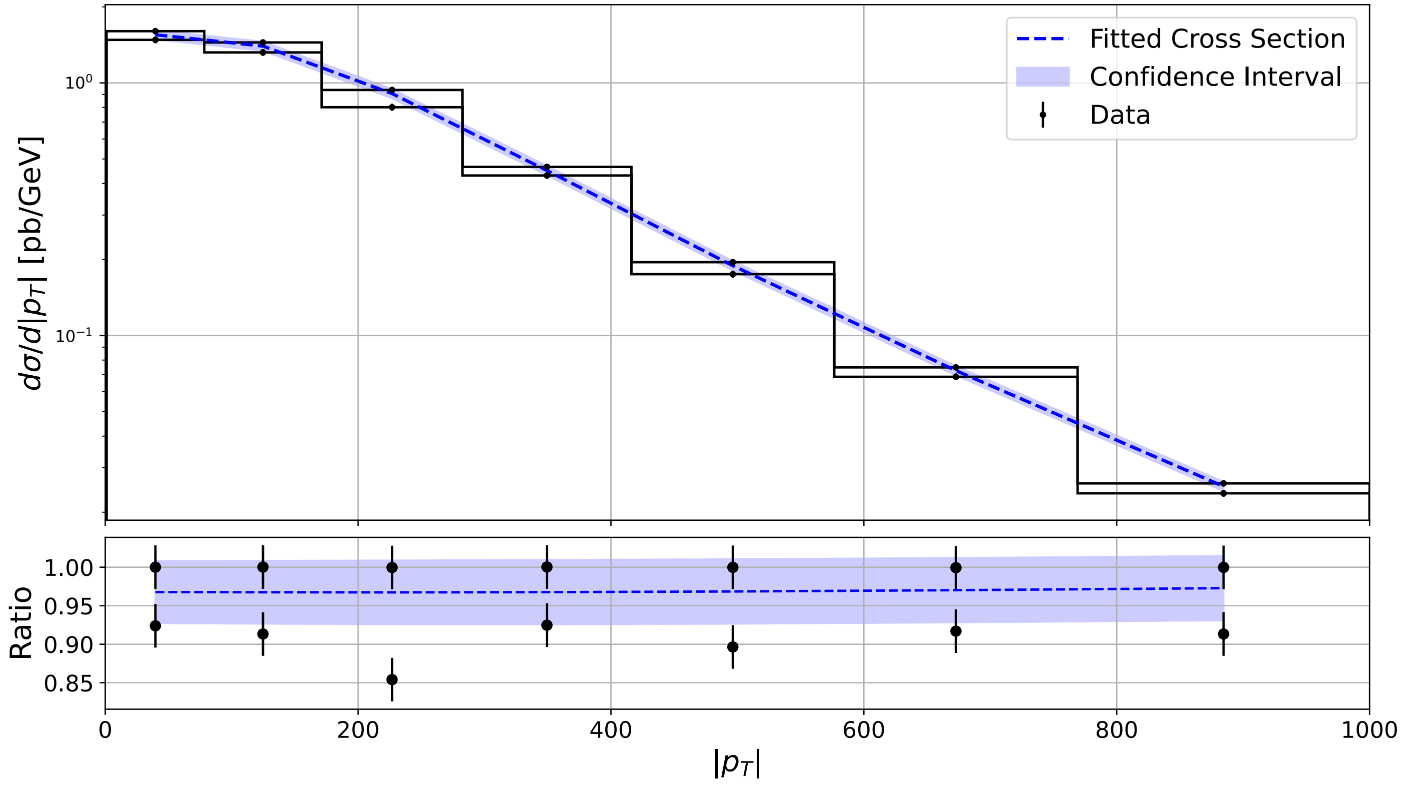
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 10%

Scale factor example

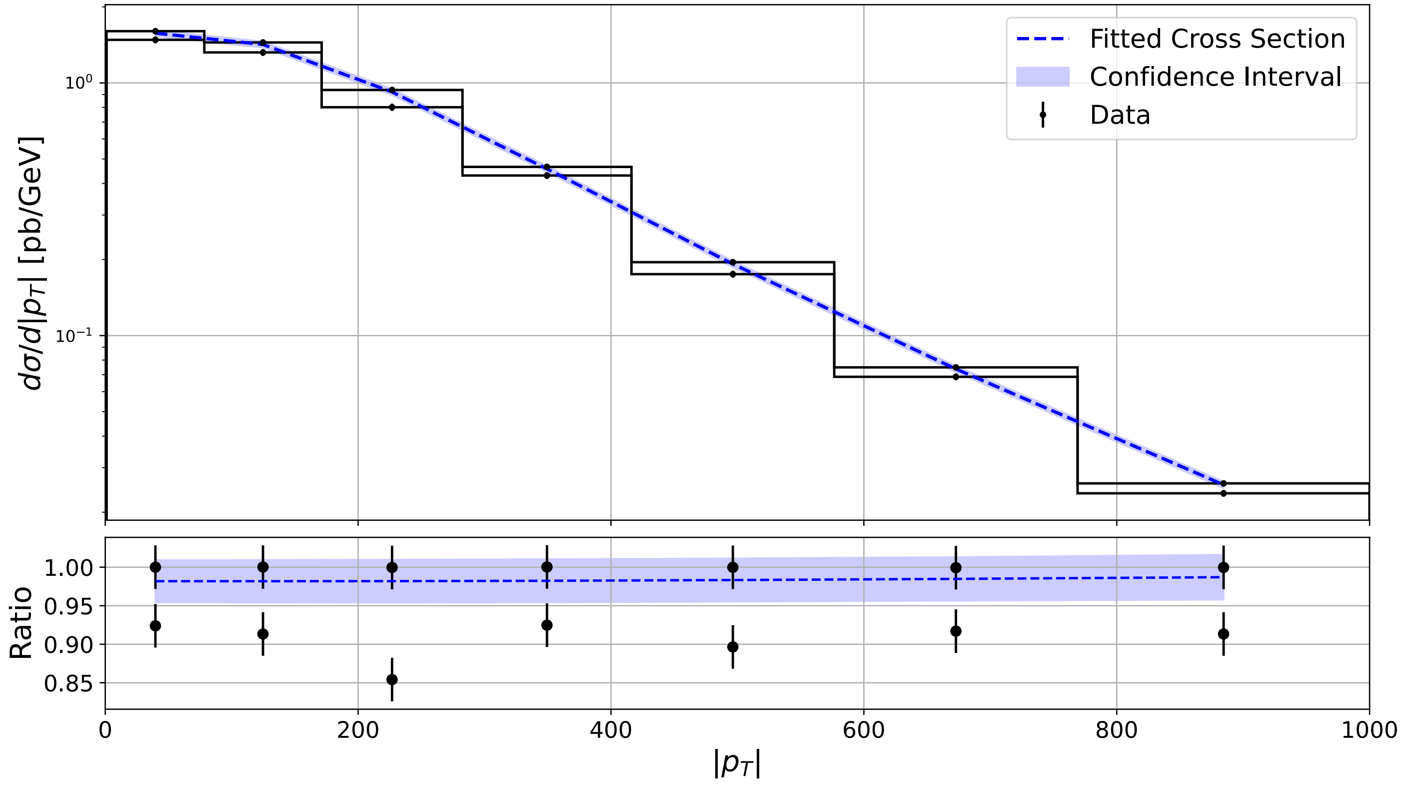
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 20%

Scale factor example

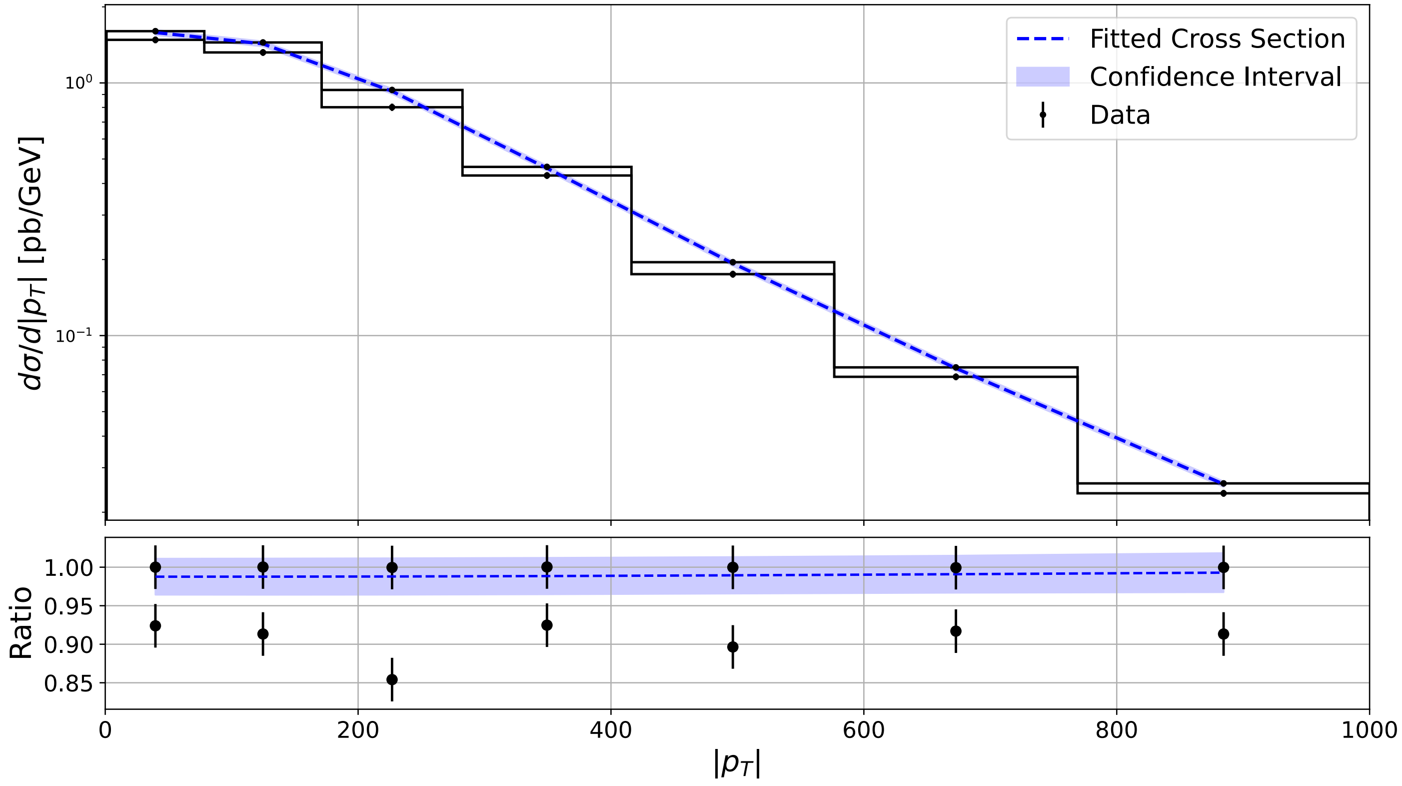
- When considering errors-on-errors, the model gives greater weight to the more internally consistent distribution in the fit.
- The confidence interval is inflated to reflect the uncertainty coming from the conflicting scale factors.



Errors-on-errors: 30%

Scale factor 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%

Outlook and conclusions

- The Gamma Variance Model allows for more meaningful inference in contexts where the procedures used to assign systematic errors are themselves uncertain.

- The primary advantage of this approach is that it reduces the sensitivity of the fits to outliers and data that are incompatible.

- The presence of incompatible data is reflected by inflated error bars on the final results.

- The values of the error-on-error parameters are fixed parameters of the model.
    - they can be assigned using expert knowledge
    - they can be varied on meaningful ranges to study the dependence of results on different assumptions

Thank you for your attention

Back-up slides

Motivation for the GVM

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- 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:

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

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

Calculation of confidence intervals

- The likelihood function can be used to construct the profile likelihood ratio test statistic:

$$w_{\boldsymbol{\mu}} = -2ln\frac{L\left(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}\right)}{L(\hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\theta}})}$$

- Use the p-value:

$$p_{\boldsymbol{\mu}} = \int_{w_{\boldsymbol{\mu},obs}}^{\infty} f\left(w_{\boldsymbol{\mu}}|\boldsymbol{\mu}\right) dw_{\boldsymbol{\mu}}$$

- Include $\boldsymbol{\mu}$ such that:

$$p_{\boldsymbol{\mu}} < \alpha$$

Calculation of confidence intervals

- Modify the likelihood ratio w directly so that its distribution is closer to the asymptotic form:

$$w_{\mu} \longrightarrow w^*_{\mu} = w_{\mu}\frac{M}{E[w]}$$

To compute confidence intervals, rescale the results obtained with Standard methods, such as the Hessian method, by $\frac{M}{E[w]}$

$$\boxed{\begin{aligned} w &\sim \chi^2_M + \mathcal{O}(\boldsymbol{n}^{-1}) \\ w^* &\sim \chi^2_M + \mathcal{O}(\boldsymbol{n}^{-2}) \end{aligned}}$$

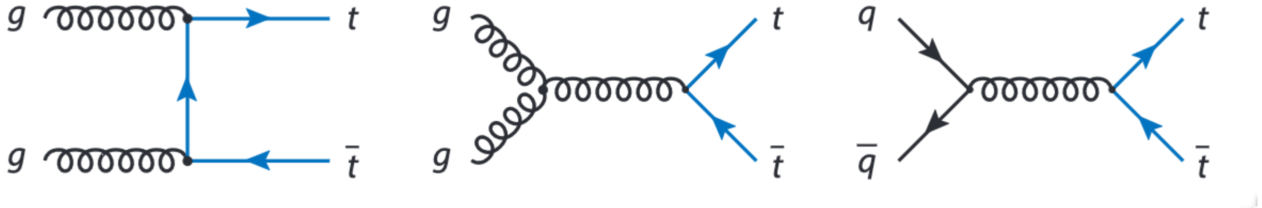
Simplified Model (no real data)

ROYAL HOLLOWAY UNIVERSITY OF LONDON

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.

$gg \rightarrow t\bar{t}$ LO cross section:



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

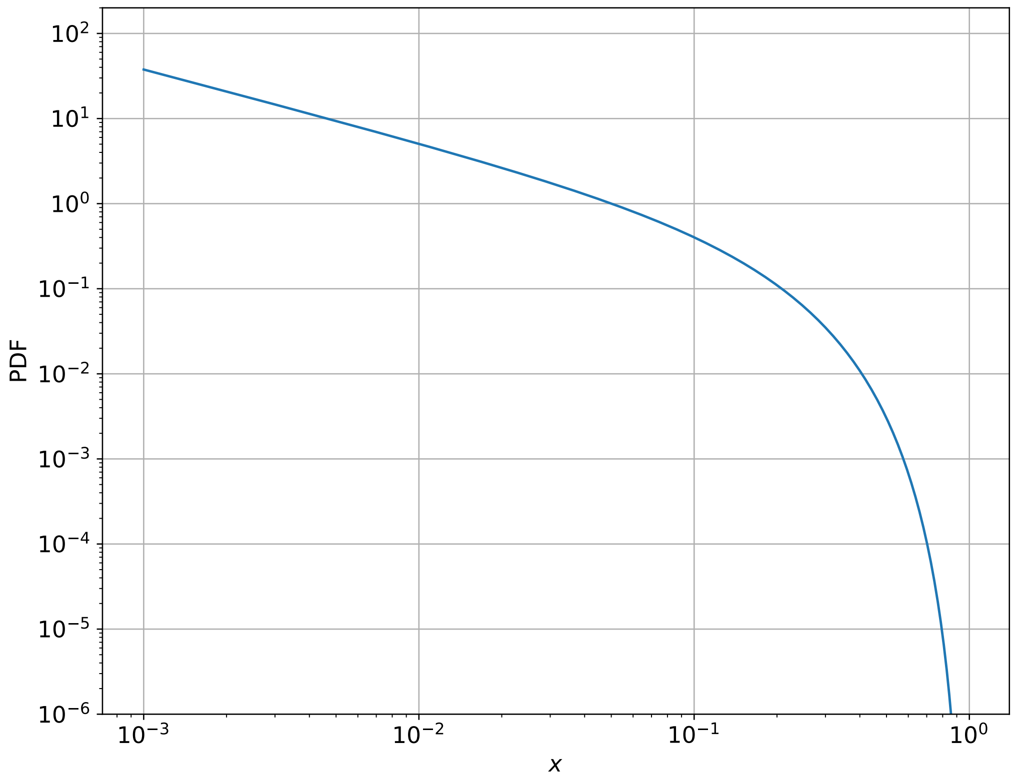
$$\frac{d\sigma_{pp}}{dx_1 dx_2 dcos\theta} = \mathrm{g}(\mathrm{x}_1)\mathrm{g}(\mathrm{x}_2)\frac{d\hat{\sigma}}{dcos\theta}$$

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.

Simplified Model – Outlier Example

ROYAL HOLLOWAY UNIVERSITY OF LONDON

To fit the Gluon PDF I will use the $|p_T|$ differential cross-section (Other cross-sections could have been used as-well)

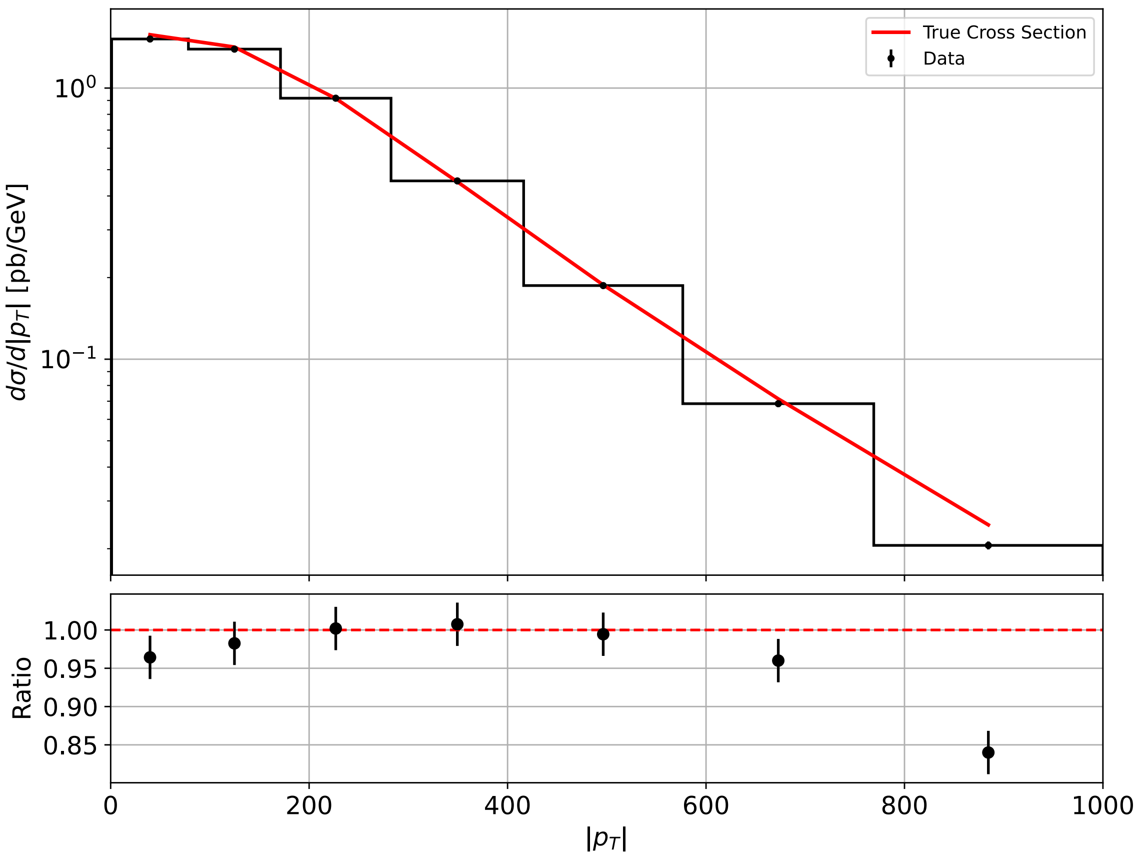
1. Compute the predicted cross-section value in each bin, using the chosen PDF parameter values.

2. Generate Gaussian data points around the predicted values.

3. Shift the last data point at high $|p_T|$ to simulate the presence of an outlier.

4. The uncertainties are made by a statistic and systematic component of equal sizes

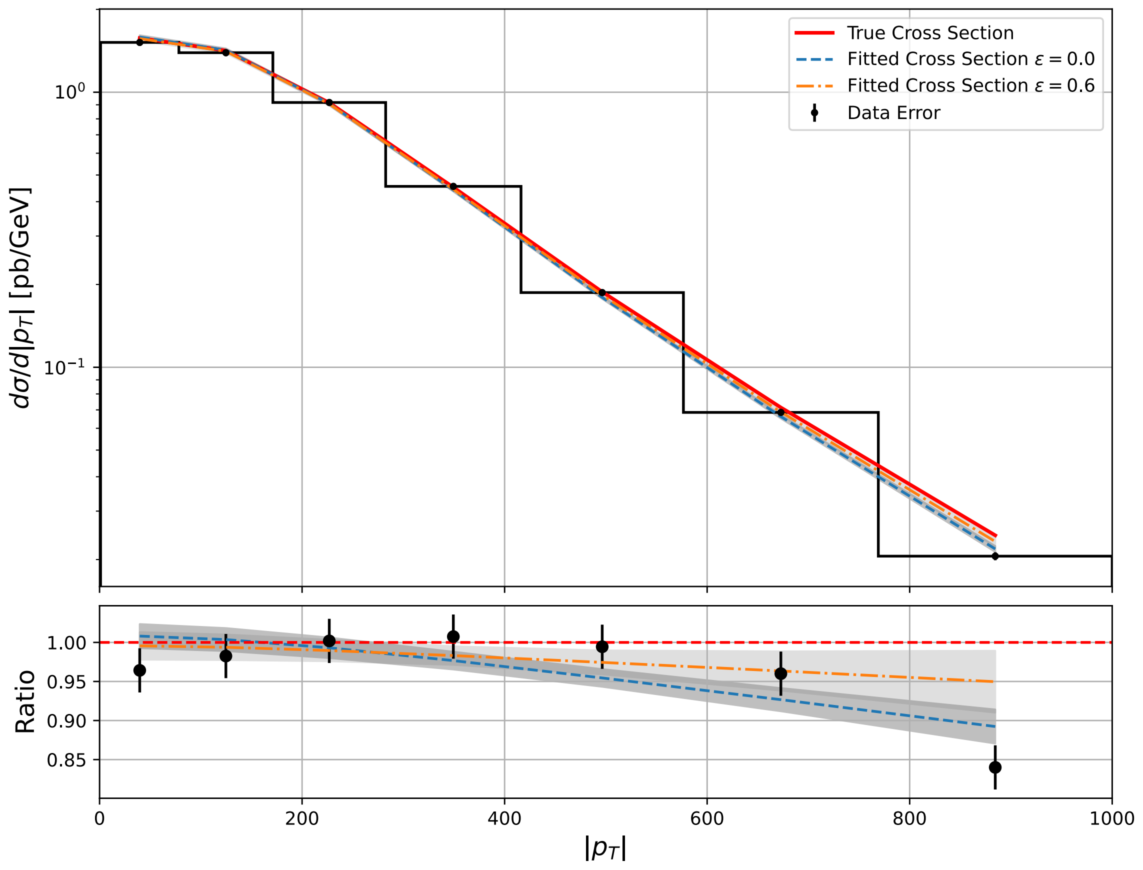
5. Assume the systematic component is itself uncertain

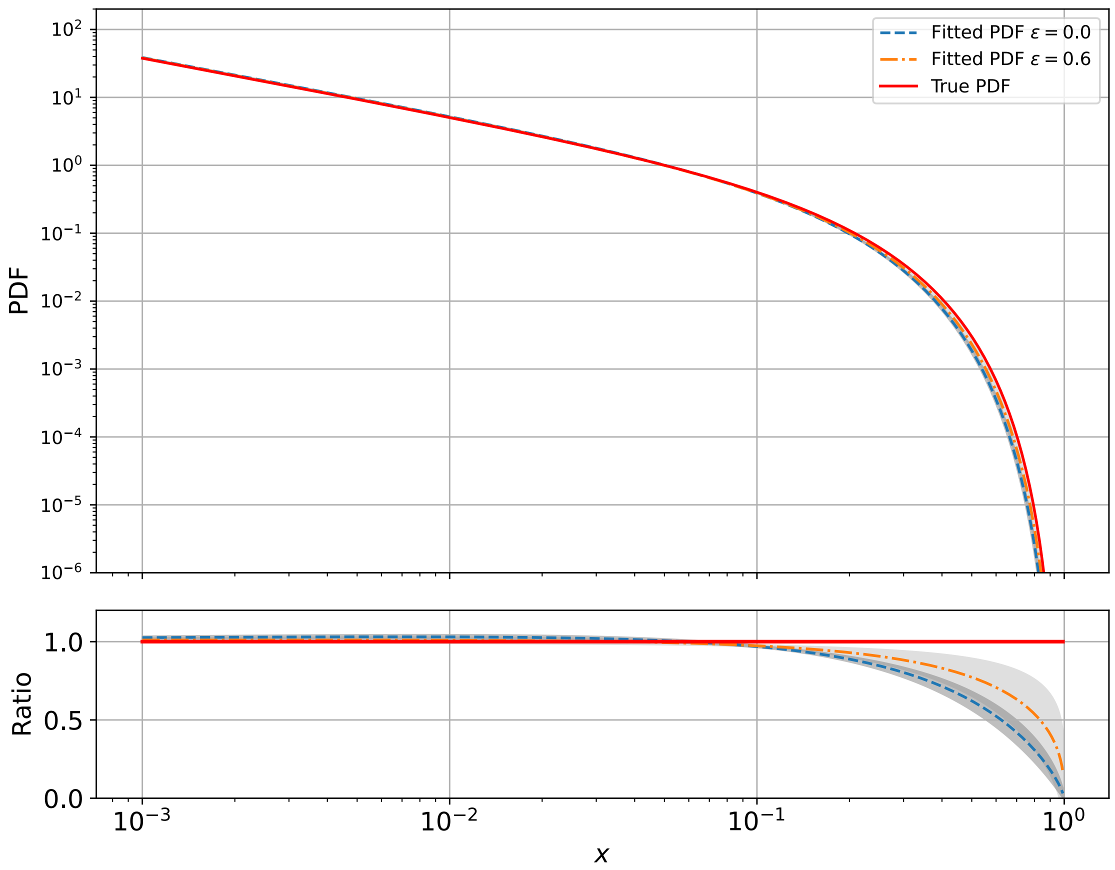


Simplified Model – Outlier Example

ROYAL HOLLOWAY UNIVERSITY OF LONDON

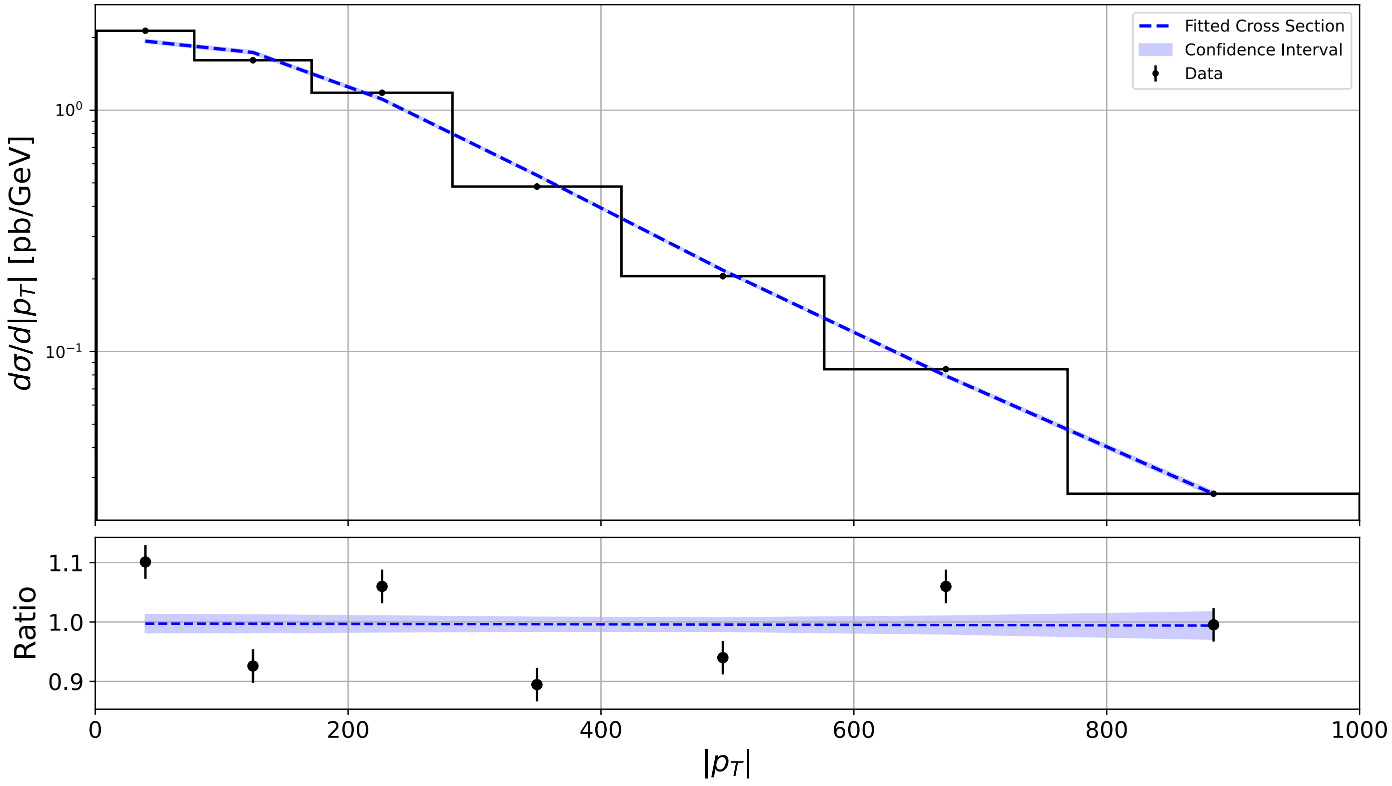
- When considering errors-on-errors, the bias introduced by the outlier is reduced.
- The confidence interval is adjusted to reflect the increased uncertainty in the region affected by the outlier.





Realistic fit example (from PDF fitting)

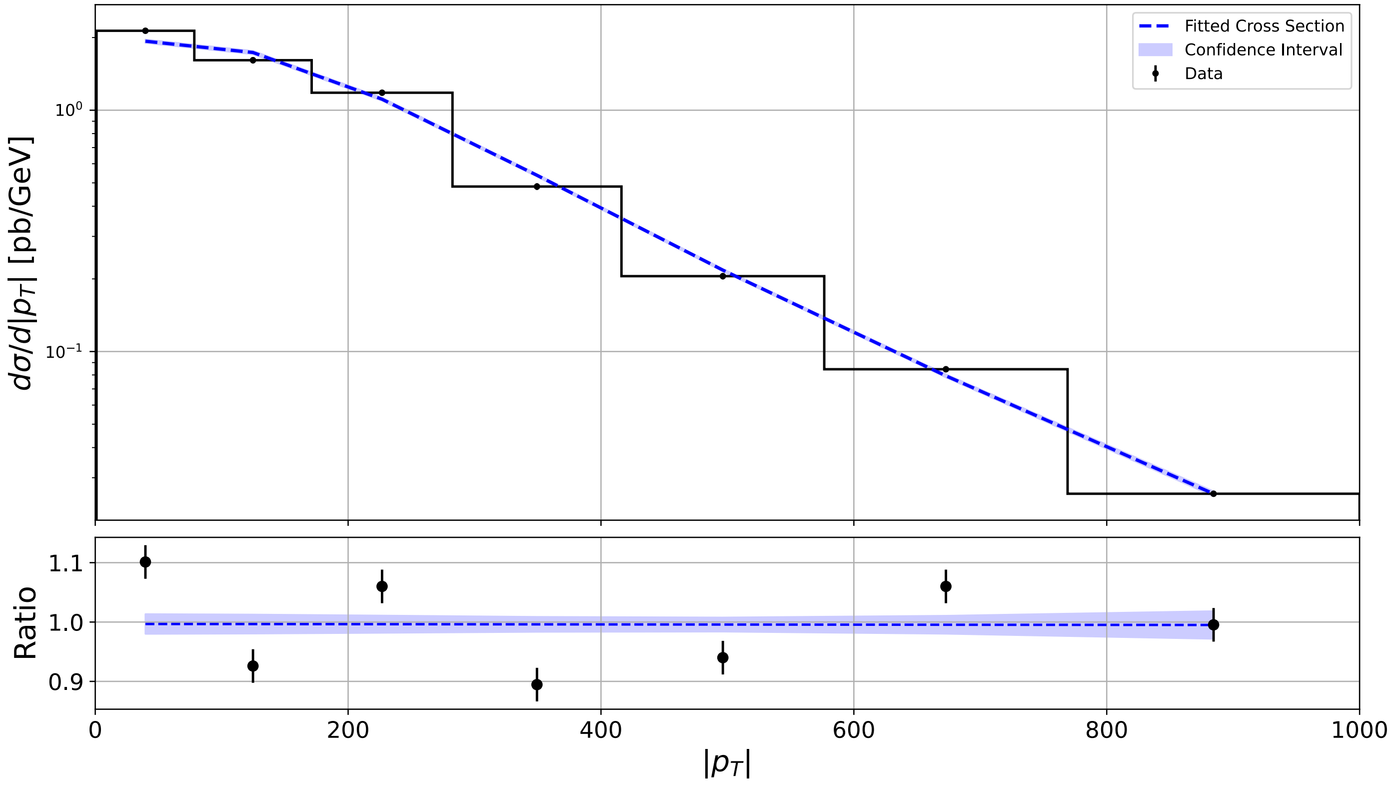
- As errors-on-errors increase, the model fits the set of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements and the fit result



Errors-on-errors: 0%

Realistic fit example (from PDF fitting)

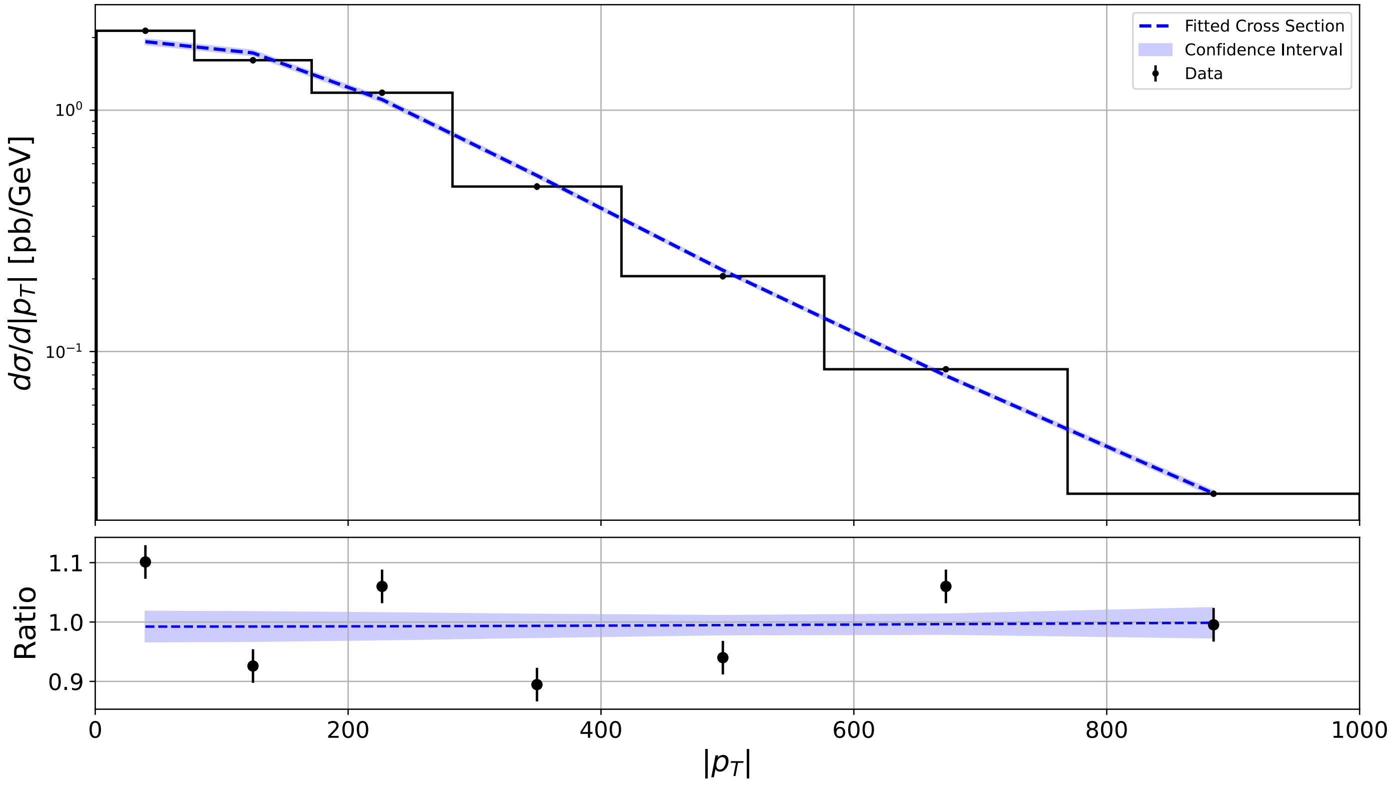
- As errors-on-errors increase, the model fits the set of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements and the fit result



Errors-on-errors: 10%

Realistic fit example (from PDF fitting)

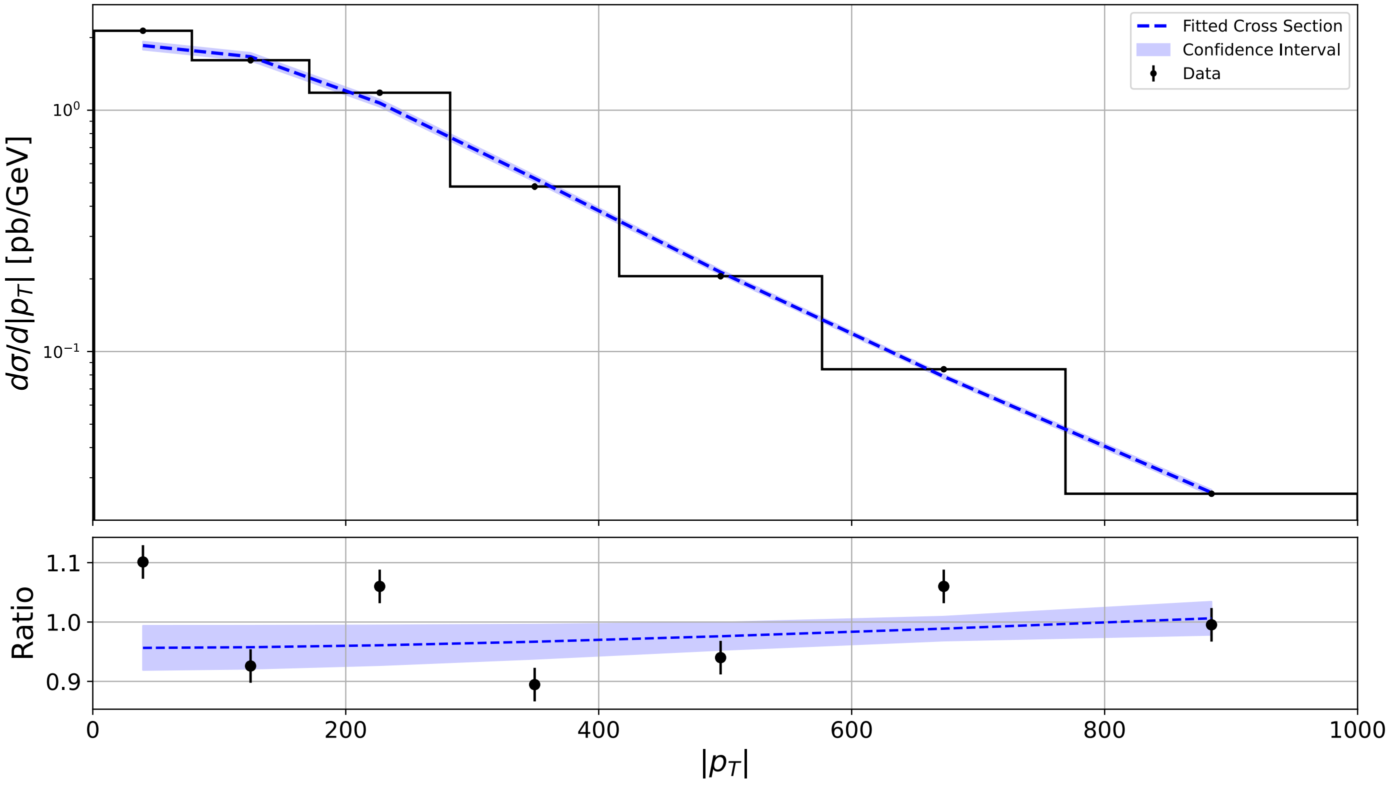
- As errors-on-errors increase, the model fits the set of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements and the fit result



Errors-on-errors: 20%

Realistic fit example (from PDF fitting)

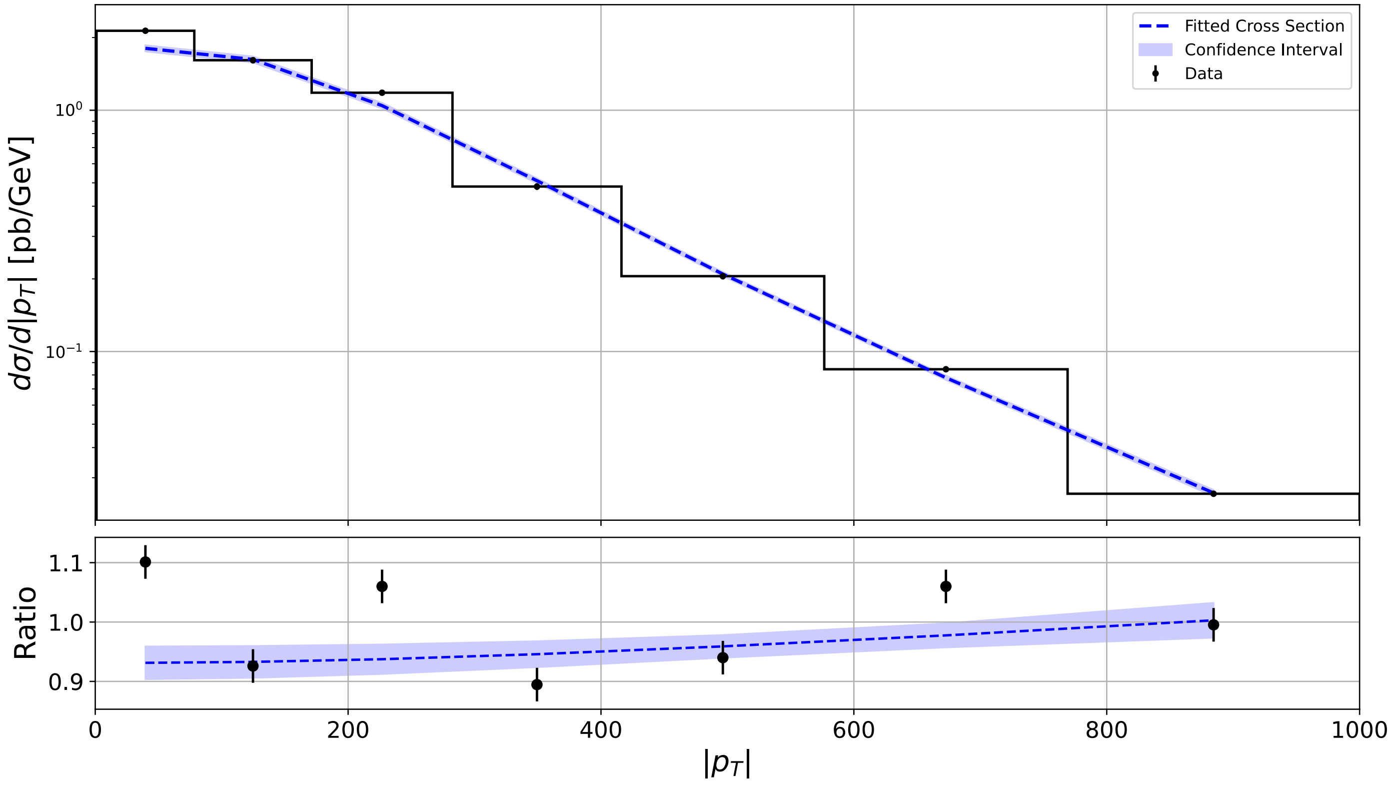
- As errors-on-errors increase, the model fits the set of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements and the fit result



Errors-on-errors: 30%

Realistic fit example (from PDF fitting)

- As errors-on-errors increase, the model fits the subset of data that have the highest degree of internal compatibility
- The confidence interval is adjusted to reflect the degree of uncertainty arising from inconsistencies within the measurements and the fit result



Errors-on-errors: 40%