Model Uncertainty in the Measurement of the Gluon-Jet Fraction

S. Shulha
JINR, Dubna

ML4Jets 2026
Vienna,
September, 14-18, 2026

Physics motivation

1. Physics significance

The gluon-jet fraction provides an indirect link between the hard QCD process and the observed gluon-jet population:

$$\text{Hard QCD} \longrightarrow \text{Parton shower} \longrightarrow \text{Hadronization} \longrightarrow \text{q/g-jets} \longrightarrow \boxed{\text{g-jet fraction}}$$

2. Practical motivation

Precision measurements of **q- and g-jet properties** require a reliable determination of the **q/g-jet fractions**

$$\text{Hard QCD} \longrightarrow \text{Parton shower} \longrightarrow \boxed{\text{Hadronization} \longrightarrow \text{q/g-jets}} \longrightarrow \text{g-jet fraction}$$

$$\downarrow$$

$$\boxed{\text{q/g-jet properties}}$$

Objectives

Standard template fit [1]

$$\text{Data} \approx \alpha^{g} \cdot \text{g-template} + (1-\alpha^{g}) \cdot \text{q-template}$$

where the templates are obtained from a Monte Carlo model.

Bin representation

Data: p_i vs. Model: $\pi_i(\alpha^{g}, a, b)$

where

$$\pi_i = \alpha^{g} \cdot p_i^{g}(a) + (1-\alpha^{g}) \cdot p_i^{q}(b)$$

p_i	: normalized content of bin i in data
p_i^{g}, p_i^{q}	: templates from a MC model
α^{g}	: g-jet fraction
a, b	: template modification parameters

The problem: the templates are model predictions that may deviate from the true ones.

Therefore, **even if the statistical uncertainty of the fit is small**, the extracted gluon fraction can still be biased if the templates do not describe the true q/g distributions

Goal of this talk

Quantify the model uncertainty of the extracted g-jet fraction

[1] S.S. e.a. Methodology for measuring gluon jet fraction and characteristics of quark and gluon jets for hadron-hadron collisions. Phys.Part.Nucl.Lett.18(2021),no.2

Model Uncertainties

Jet observables [2]:

mult: jet constituent multiplicity

a_2: minor angular spread of the $p_{T\,k}$ weighted constituent distribution

$$\mathrm{ptD} \equiv \frac{\sqrt{\sum_k p_{T\,k}^2}}{\sum_k p_{T\,k}}$$: momentum-sharing/fragmentation variable

We use these observables as independent ways of extracting the same underlying g-jet fraction

Type-II m.u.

– spread of g-fraction between observables.

Internal estimate: within a single MC model

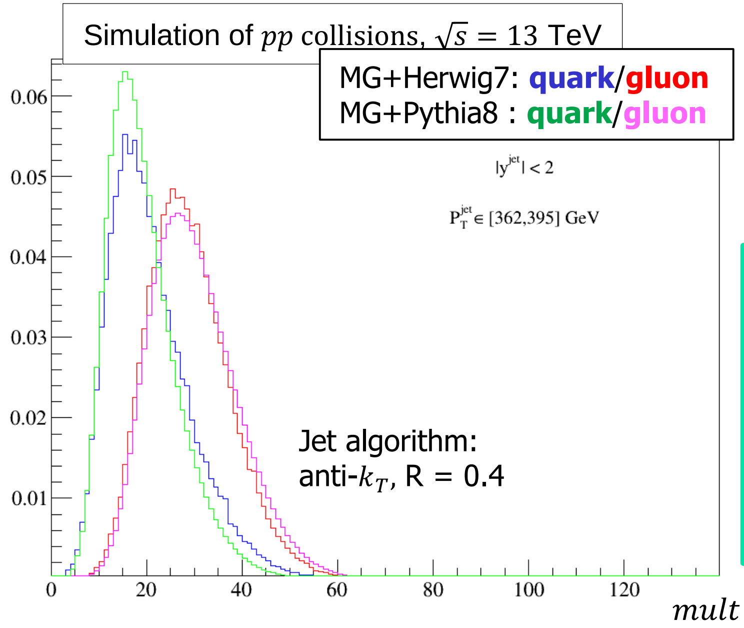


Fig.: Illustrative q/g templates

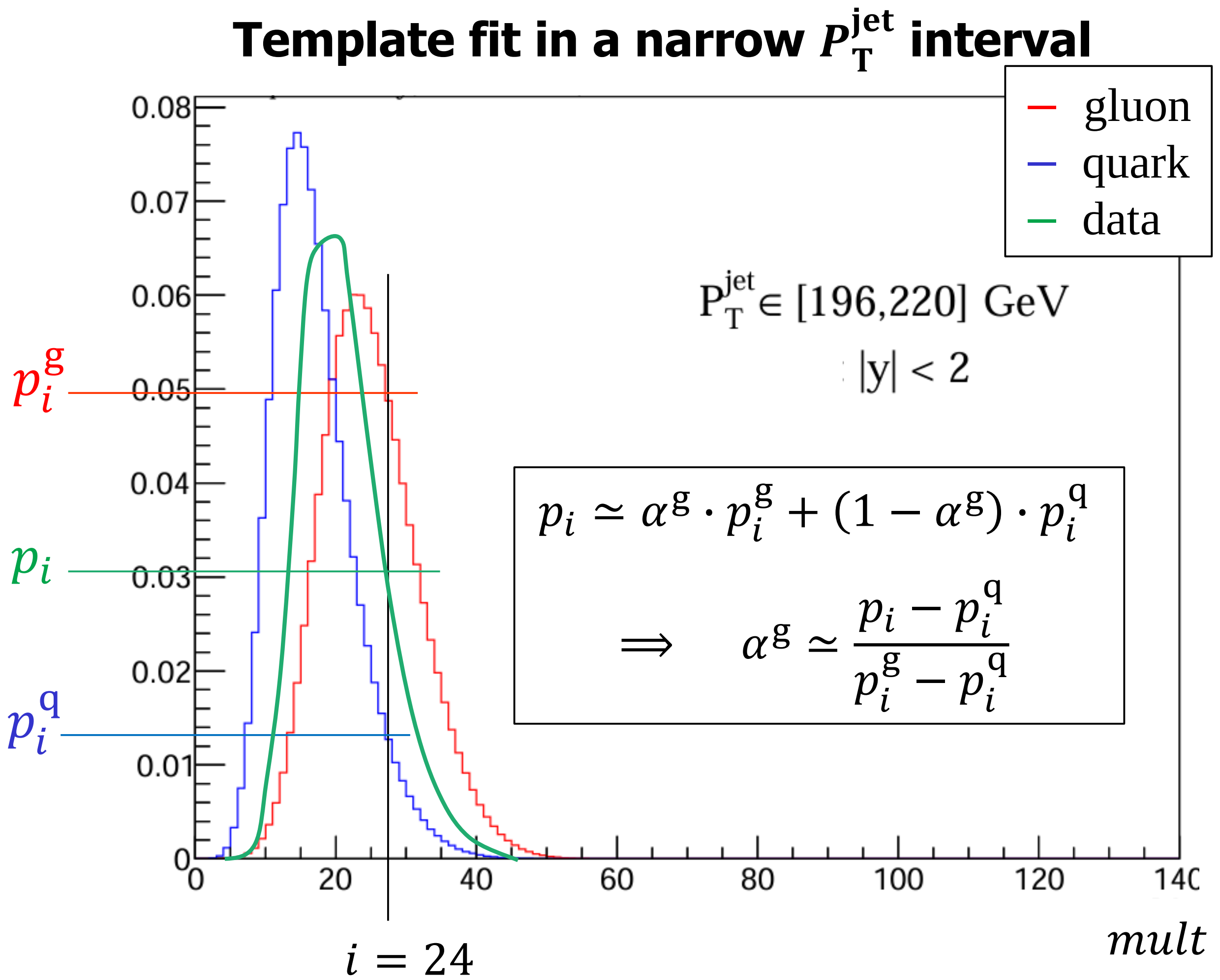
Type-III m.u.

– spread of g-fraction between models.

External estimate: based on measurements with different models

[2] CMS PAS JME-13-002
PAS JME-16-003

Type-I Model Uncertainty [3]



$$p_i \simeq \alpha^g \cdot p_i^g + (1-\alpha^g) \cdot p_i^q$$

$$\Longrightarrow \quad \alpha^g \simeq \frac{p_i - p_i^q}{p_i^g - p_i^q}$$

The g-jet fraction is determined independently in each bin of the jet-observable distribution.

Type-I m.u.
– spread of g-fraction between bins i.

Internal estimate: within a single MC model, within a single jet observable

Goal of this talk

Quantify the Type-I m.u.

[3] S.S. Model Uncertainty in Measuring the Gluon Jet Fraction at the Hadron Collider. *Phys. Atom. Nuclei* **87**, 509–512 (2024)

Model Uncertainties

- Thus, the measured fraction of g-jets can exhibit three levels of g-fraction spread:
  - Bin-to-bin spread – does the model correctly describe the shape of the templates? **Type-I m.u.**
  - Observable-to-observable spread – does the model correctly describe the mechanism of jet formation and its spatial structure? **Type-II m.u.**
  - Model-to-model spread – how strongly does the result depend on the choice of the underlying generator model? **Type-III m.u.**
- Only the "**Type-I m.u.**" requires a dedicated statistical construction. Its determination requires a statistical construction that relates the **global** g-fraction, $\hat{\alpha}^{g}_{(V)}$, to the **local** g-fractions $\hat{\alpha}^{g}_{(V)\,i}$, obtained in individual bin i of a jet observable V.
- The statistical framework shown below is therefore primarily driven by the requirements of the "**Type-I m.u.**".
- The Type-II and Type-III uncertainties can subsequently be determined from comparisons of the corresponding **global** measurements.

Where Are We Now?

STAGE 1 — MODEL DIAGNOSTICS

Current focus: $\boxed{\text{Local g–fraction measurement} \rightarrow \text{Model uncertainty}}$

- **Quantify** the model-induced spread
- **Parameterize** template modifications
- **Reduce** the model-induced spread
- Provide feedback for **q/g-jet model tuning**

WE ARE HERE: Stage 1

Three relevant scales: From the local measurements $\alpha^{g}_{(V)i}$, we can assess:

$\sigma_{\text{inside V}}$ $\sigma_{\text{between V}}$ σ_{stat}

Model effects dominate → keep observables ***separate***

Initial situation: $\sigma_{\text{inside V}} > \sigma_{\text{stat}}$ and $\sigma_{\text{between V}} \gg \sigma_{\text{stat}}$

After model improvement: $\sigma_{\text{inside V}}, \sigma_{\text{between V}} \lesssim \sigma_{\text{stat}}$

Only then → STAGE 2 — GLOBAL COMBINATION

$\boxed{mult \oplus ax2 \oplus ptD \longrightarrow \textbf{Global fit}}$

- Gain **statistical precision**

$$\underbrace{\text{Stage 1: Model uncertainty}}_{\text{NOW}} \longrightarrow \underbrace{\text{Stage 2: Statistical precision}}_{\text{LATER}}$$

So the question becomes: **what statistical framework** naturally provides the chain of "Stage 1" — local measurements, statistical and model uncertainty decomposition, model diagnostics, and eventually the global combination

Objectives

Model Uncertainties

Analysis strategy

Statistical Framework

G-fraction: Mean vs. Local

Statistical uncertainty

Total Variance S_{w}^{2}

Model

Statistical Properties of Weighted Jet Counts

i : histogram bin
n_i : unnormalized bin count
μ_i : model prediction

Unweighted jets

For independent jet counts in the i-bins:

$$n_i \sim \mathrm{Poiss}(\mu_i)$$
$$\mathrm{E}(n_i) = \mathrm{Var}(n_i) = \mu_i$$

Weighted jets

jet : jet
v_{jet}: jet weight

The bin counts is

$$n_i = \sum_{\text{jet} \in i} v_{\text{jet}}$$

with $\mathrm{E}(n_i) = \mu_i$ and $\mathrm{Var}(n_i) = \sum_{\text{jet} \in i} v_{\text{jet}}^2$

$$\Longrightarrow \boxed{\mathrm{Var}(n_i) \neq \mathrm{E}(n_i)}$$

Key consequence:

- ✓ So the key point is very simple: weighted jet counts are not Poisson
- ✓ This means that we cannot directly apply the usual Poisson likelihood to the weighted histogram. We therefore need an appropriate statistical construction for the weighted sample
- ✓ This construction leads us to an effective Poisson description, then to a multinomial model, and finally to the asymptotic Gaussian model

From Weighted Events to the Asymptotic Gaussian Model

Statistical construction

Weighted Jet counts → Effective Poisson → Multinomial → Asymptotic Gaussian

Weights → effective stat. power N_{eff} → normalized bin contents → large-N_{eff} limit

We can transform the weighted sample to an effective Poisson representation, formulate the normalized bin contents as a multinomial problem, and then use the large-N_{eff} limit, where the multinomial becomes multivariate Gaussian

(Details see in Backup, p.20-22)

Asymptotic Gaussian Model

For large N, the multivariate CLT gives:

$$\boldsymbol{p} = \left\{\frac{n_i}{N}\right\}$$

has multivariate Gaussian distribution

$$\boldsymbol{p} \xrightarrow[N_{\text{eff}} \gg 1]{} \mathbb{N}\left(\boldsymbol{\pi}, \frac{1}{N_{\text{eff}}}\left[diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^{\text{T}}\right]\right),$$

$$n_i = \sum_{\text{jet} \in i} v_{\text{jet}} \qquad N = \sum_{\text{jet}=1}^{N_{\text{jets}}} v_{\text{jet}}$$

$$\boldsymbol{\pi} = \{\pi_i\}$$

$$\pi_i \equiv \alpha^{\text{g}} \cdot p_i^{\text{g}}(a) + (1 - \alpha^{\text{g}}) \cdot p_i^{\text{q}}(b)$$

$$N_{\text{eff}} = \frac{\left(\sum_{\text{jet}=1}^{N_{\text{jets}}} v_{\text{jet}}\right)^2}{\sum_{\text{jet}=1}^{N_{\text{jets}}} v_{\text{jet}}^2}$$

The resulting asymptotic Negative Log-Likelihood:

$$\boxed{-2\ln(L) \simeq N_{\text{eff}} \cdot \sum_{i=1}^{K} \frac{[p_i - \pi_i(\alpha^{\text{g}}, a, b)]^2}{p_i} + \text{const}}$$

- ✓ This is the statistical model that we will use
- ✓ Now we can ask what this likelihood implies for the measurements of the mean g-jet fraction?
- ✓ **Answer**: the mean = result of fit with asymptotic Gaussian Likelihood can be related analytically to the local g-fraction measurements

Analytical profiling over α^g

Minimizing NLL:

$$-2\ln[L(\alpha^{\mathrm{g}})] = N_{\mathrm{eff}} \cdot \sum_i \frac{[p_i - \pi_i(\alpha^{\mathrm{g}})]^2}{p_i} + const$$

with respect to α^{g} gives

$$\boxed{\hat{\alpha}^{\mathrm{g}} = \sum_i w_i \cdot \alpha_i^{\mathrm{g}}}$$

$$\boxed{\pi_i = \alpha^{\mathrm{g}} \cdot p_i^{\mathrm{g}} + (1-\alpha^{\mathrm{g}}) \cdot p_i^{\mathrm{q}}}$$

where local g-fraction is defined by

$$\boxed{\alpha_i^{\mathrm{g}} \stackrel{\mathrm{def}}{=} \frac{p_i - p_i^{\mathrm{q}}}{\Delta_i},}$$

$$\boxed{\Delta_i = p_i^{\mathrm{g}} - p_i^{\mathrm{q}}}$$

Separation of q and g templates

weight is defined by

$$\boxed{w_i = \frac{\mathrm{W}_i}{\sum_{i=1}^{K} \mathrm{W}_i},}$$

$$\boxed{\mathrm{W}_i = \frac{\Delta_i^2}{p_i}}$$

$w_i \sim$ squared separation of q and g templates

This analytical relation between the global estimator and the local measurements, derived from the MLE, will be essential for what follows. It allows us to use the dispersion of the local estimates as a diagnostic of model uncertainty (Type-I m.u.).

Statistical uncertainty

Variance of the Mean g-fraction

$$\hat{\alpha}^{g} = \sum_i w_i \cdot \alpha_i^{g}$$

$$\Longrightarrow \mathrm{Var}(\hat{\alpha}^{g})_{\mathrm{stat}} = \sum_i w_i^2 \cdot \mathrm{Var}(\alpha_i^{g})_{\mathrm{stat}}$$

$$\Longrightarrow \boxed{\mathrm{Var}(\hat{\alpha}^{g})_{\mathrm{stat}} = \frac{1}{N_{\mathrm{eff}} \cdot \sum_i \mathrm{W}_i}}$$

Variance of the Local g-fraction

$$\mathrm{Var}(\alpha_i^{g})_{\mathrm{stat}} = \frac{1}{N_{\mathrm{eff}} \cdot \mathrm{W}_i}$$

$$\mathrm{W}_i \equiv \frac{\Delta_i^2}{p_i}$$

See Backup p.23

$\Longrightarrow$ **Statistical uncertainty** of $\hat{\alpha}^{g}$: $\quad \Delta\hat{\alpha}^{g}_{\mathrm{stat}} = \sqrt{\dfrac{1}{N_{\mathrm{eff}} \cdot \sum_i \mathrm{W}_i}}$

See Backup p.30 – how to add template modification parameters "a, b"

- ✓ So this uncertainty has a very transparent interpretation: it is determined by the effective statistical size of the sample, N_{eff}, and by the **total** q/g separation across all bins, $\sum_i \mathrm{W}_i$
- ✓ At this stage, importantly, the templates are fixed. We are quantifying only the statistical fluctuations of the data
- ✓ Now we keep the data fixed and ask the complementary question: what happens if the q/g templates are not exactly the true templates?

The local g-fraction provides a particularly useful diagnostic of the q/g templates.

$$\boxed{\alpha_i^{\mathrm{g}} = \frac{p_i - p_i^{\mathrm{q}}}{\Delta_i}}$$

If the q/g templates are correct,

$$\boxed{\begin{aligned} p_i &= (1 - \alpha_0) \cdot p_i^{\mathrm{q}} + \alpha_0 \cdot p_i^{\mathrm{g}}, \\ \Longrightarrow \quad \alpha_i^{\mathrm{g}} &= \alpha_0, \qquad \forall \; i \end{aligned}}$$

What does observed bin-to-bin spread mean?

Heuristic quantification of the observed spread using MLE-derived weights (!):

$$\hat{\alpha}^{\mathrm{g}} = \sum_{i=1}^{K} w_i \cdot \alpha_i^{\mathrm{g}}, \qquad S_w^2 \stackrel{\text{def}}{=} \sum_{i=1}^{K} w_i \cdot \left(\alpha_i^{\mathrm{g}} - \hat{\alpha}^{\mathrm{g}}\right)^2$$

The observed variance S_w^2 contains both statistical fluctuations and authentic model effects. The next step is therefore to separate these two contributions

Observed local spread contains two contributions

Natural assumption: statistical and model-induced fluctuations are independent:

$$\boxed{E(S_{w}^2) = E(S_{w}^2)_{\mathrm{stat}} + E(S_{w}^2)_{\mathrm{model(I)}}}$$

$$E(S_{\mathrm{w}}^2)_{\mathrm{stat}} \backsimeq (K-1)\cdot \mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{stat}}$$

$$E(S_{\mathrm{w}}^2)_{\mathrm{model(I)}} \backsimeq (K_{\mathrm{eff}}-1)\cdot \mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{model(I)}}$$

see Backup p.24-26

with

$$K_{\mathrm{eff}} = \frac{[\sum_{i=1}^{K} w_i]^2}{\sum_{i=1}^{K} w_i^2} \equiv \frac{1}{\sum_{i=1}^{K} w_i^2}$$

$$w_i = \frac{\mathrm{W}_i}{\sum_{i=1}^{K}\mathrm{W}_i}$$

$$\mathrm{W}_i = \frac{\Delta_i^2}{p_i}$$

Key point: $\boxed{E(S_{w}^2) \rightarrow \mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{model(I)}}}$

Therefore, once the total observed spread and the statistical contribution are known, we can extract the model variance of the mean g-fraction

Type-I Model Uncertainty

Observed local spread: $S_w^2 = \sum_{i=1}^{K} w_i \cdot (\alpha_i^g - \hat{\alpha}^g)^2$

$$\Longrightarrow E(S_w^2) = (K-1) \cdot \mathrm{Var}(\hat{\alpha}^g)_{\mathrm{stat}} + (K_{\mathrm{eff}} - 1) \cdot \mathrm{Var}(\alpha^g)_{\mathrm{model(I)}}$$

Model variance

$$\Longrightarrow \quad \mathrm{Var}(\hat{\alpha}^g)_{\mathrm{model(I)}} = \frac{S_w^2 - (K-1) \cdot \mathrm{Var}(\hat{\alpha}^g)_{\mathrm{stat}}}{(K_{\mathrm{eff}} - 1)}$$

In the final estimator, the observed S_w^2 is used as an estimate of its expectation value $E(S_w^2)$.

Final model uncertainty

$$\Delta\hat{\alpha}^g_{\mathrm{model(I)}} = \sqrt{\mathrm{Var}(\hat{\alpha}^g)_{\mathrm{model(I)}}}$$

- ✓ An important point is that the statistical variance used here is the full statistical variance, including the statistical uncertainty associated with the fitted template modification parameters a and b
- ✓ So the central idea is quite simple: ***we measure the total local spread, subtract what can be explained by statistical fluctuations, and interpret the remaining excess as a diagnostic estimate of template-model uncertainty***
- ✓ I emphasize once again that this is a heuristic model-uncertainty estimate. It relies on some working assumptions (see Backup p.26-27)

Current situation

Type-II Model Uncertainty

$$\Delta\hat{\alpha}^{\mathrm{g}}_{\mathrm{model(I)}} > \Delta\hat{\alpha}^{\mathrm{g}}_{\mathrm{stat}}$$

Statistical optimization is therefore not yet justified.

Global g-fraction — current-model estimate

$$\hat{\alpha}^{\mathrm{g}} = \frac{1}{3}\sum_{V}\hat{\alpha}^{\mathrm{g}}_{\mathrm{(V)}} \qquad V = \{mult, ax2, ptD\}$$

The dominant uncertainty is the **spread between observables**:

$$\Delta\hat{\alpha}^{\mathrm{g}}_{\mathrm{model(II)}} = \frac{1}{2}\left[\max_{\mathrm{V}}\hat{\alpha}^{\mathrm{g}}_{\mathrm{(V)}} - \min_{\mathrm{V}}\hat{\alpha}^{\mathrm{g}}_{\mathrm{(V)}}\right]$$

$$\Delta\hat{\alpha}^{\mathrm{g}}_{\mathrm{model(II)}} \gg \Delta\hat{\alpha}^{\mathrm{g}}_{\mathrm{stat}}$$

The dominant uncertainty is the spread between the three observables

The important conclusion is: the measurement is currently limited by model uncertainty, not by statistics.

This tells us what the **next step should be**: improve the q/g-jet models rather than further optimize the statistical combination

Conclusions

- The standard **template-fit approach** gives us a measurement of the g-jet fraction and its **statistical** uncertainty. This uncertainty decreases with the size of the data sample and, in the asymptotic limit, vanishes.
- But this does not mean that the measurement becomes exact. A q/g template bias remains even with an infinitely large data sample.
- To address this problem, we introduced local g-fraction measurements in individual bins. Bin-to-bin variation of local measurements provides a direct diagnostic of model quality.
- The main methodological conclusion is therefore that a complete measurement of the g-jet fraction should not report only its statistical uncertainty. It should also include an intrinsic estimate of the model uncertainty associated with the q/g templates, which does not vanish in the large-statistics limit.

Backup

From Local Measurements to a Global g-Fraction

For each jet observable we can determine a **local** g-fraction in every bin.

Local g-fractions $\alpha^{g}_{(V)i}$ $\xrightarrow[\text{bin-by-bin}]{\text{average over bins}}$ One g-fraction per observable: $\alpha^{g}_{(mult)}, \alpha^{g}_{(ax2)}, \alpha^{g}_{(ptD)}$ $\xrightarrow{\text{combine observables}}$ **Global** α^{g}

$V = \{mult, ax2, ptD\}$

Alternative: A multi-parameter, multi-stage fit is also possible

All observables $\longrightarrow$ Global α^{g}

But there is a potential problem with doing this from the beginning - the **model uncertainty is completely lost**. Model uncertainty is always present and does not disappear, even in the large-statistics limit.

Thus, the measurement strategy should start with local measurements, followed by averaging over bins for each observable, and only then proceed to the global g-fraction, averaged across all observables

Weighted Poisson → Effective Poisson → Multinomial → Gaussian

[4] *G.Bohm, G.Zech,* **NIMA** 691 *(2012) 171-177.*

Weighted i-bin $\quad n_i = \sum_{\mathrm{jet} \in i} v_{\mathrm{jet}}$

Define $\quad S_{1,i} = \sum_{\mathrm{jet} \in i} v_{\mathrm{jet}}$ and $S_{2,i} = \sum_{\mathrm{jet} \in i} v_{\mathrm{jet}}^2$

So that $\quad \hat{\mu}_i = S_{1,i}\,,\ \widehat{\mathrm{Var}(n_i)} = S_{2,i}$

Böhm–Zech transformation [4]: $\boxed{c_i = \frac{S_{1,i}}{S_{2,i}},\ \tilde{n}_i \equiv c_i \cdot n_i,\ \ \tilde{\mu}_i \equiv c_i \cdot \mu_i}$

Define the **effective count**

$$n_{\mathrm{eff},i} = \frac{S_{1,i}^2}{S_{2,i}}$$

"Local" effective number of jets in i-bin = "Equivalent number of unweighted jets" in i-bin [4]

Then, at the level of data-based estimates,

$$\Longrightarrow \boxed{\widehat{\mathrm{Var}(\tilde{n}_i)} = E(\tilde{n}_i) \simeq n_{\mathrm{eff},i}}$$

Key result: Transformed weighted bin has Poisson mean–variance relation.

Effective Poisson likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{K} \frac{\tilde{\mu}_i^{\tilde{n}_i}}{\tilde{n}_i!} \cdot e^{-\tilde{\mu}_i}$$

Effective Poisson NLL:

$$-\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{K} [\tilde{\mu}_i - \tilde{n}_i \cdot \ln \tilde{\mu}_i] + const$$

Key assumption

$$\boxed{c_i \simeq C}$$

approximately independent of the bin i.

Therefore

$$\tilde{n}_i \equiv C \cdot n_i, \qquad \tilde{\mu}_i \equiv C \cdot \mu_i$$

with

$$n_i = N \cdot p_i \text{ and } \mu_i = N \cdot \pi_i$$

$$\boxed{N = \sum_{\text{jet}=1}^{N_{\text{jets}}} v_{\text{jet}}}$$

and

$$\pi_i \equiv (1 - \alpha^{\text{g}}) \cdot p_i^{\text{q}} + \alpha^{\text{g}} \cdot p_i^{\text{g}}$$

We obtain Multinomial NLL:

$$\boxed{-\ln L(\alpha^g, a, b) = -N_{\text{eff}} \cdot \sum_{i=1}^{K} [p_i \cdot \ln \pi_i(\alpha^{\text{g}}, a, b)] + const,}$$

where

$$\boxed{N_{\text{eff}} = C \cdot N = \frac{\left(\sum_{\text{jet}=1}^{N_{\text{jets}}} v_{\text{jet}}\right)^2}{\sum_{\text{jet}=1}^{N_{\text{jets}}} v_{\text{jet}}^2}}$$

Key result: Weighted Effective Poisson NLL → Multinomial NLL with N_{eff} as the statistical scale

Weighted Poisson → Effective Poisson → Multinomial → Gaussian

For large N_{eff}, the multivariate CLT gives

$$\boldsymbol{p} = \left\{\frac{\tilde{n}_i}{N_{\mathrm{eff}}}\right\} = \left\{\frac{n_i}{N}\right\}$$

has multivariate Gaussian distribution

$$\boldsymbol{p} \xrightarrow[N_{\mathrm{eff}} \gg 1]{} \mathbb{N}\left(\boldsymbol{\pi}, \frac{1}{N_{\mathrm{eff}}}\left[diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}}\right]\right),$$

$$\pi_i \equiv \alpha^{\mathrm{g}} \cdot p_i^{\mathrm{g}}(a) + (1 - \alpha^{\mathrm{g}}) \cdot p_i^{\mathrm{q}}(b)$$

Asymptotic Gaussian NLL:

$$-2\ln(L) \simeq N_{\mathrm{eff}} \cdot \sum_{i=1}^{K} \frac{[p_i - \pi_i(\alpha^{\mathrm{g}}, a, b)]^2}{\pi_i(\alpha^{\mathrm{g}}, a, b)} + \mathrm{const}$$

Asymptotic plug-in: for large N_{eff}

$$\pi_i(\alpha^{\mathrm{g}}, a, b) \simeq p_i$$

Therefore,

$$\boxed{-2\ln(L) \simeq N_{\mathrm{eff}} \cdot \sum_{i=1}^{K} \frac{[p_i - \pi_i(\alpha^{\mathrm{g}}, a, b)]^2}{p_i} + \mathrm{const}}$$

Key result: Multinomial → Gaussian χ^2-like form with N_{eff} as the statistical scale

Statistical uncertainty

Mean g-fraction

$$\hat{\alpha}^{g} = \sum_i w_i \cdot \alpha_i^{g}$$

$$\mathrm{Var}(\hat{\alpha}^{g})_{\mathrm{stat}} = \sum_i w_i^2 \cdot \mathrm{Var}(\alpha_i^{g})$$

$$\Longrightarrow \boxed{\mathrm{Var}(\hat{\alpha}^{g})_{\mathrm{stat}} = \frac{1}{N_{\mathrm{eff}} \cdot \sum_i \mathrm{W}_i}}$$

$\Longrightarrow$ Statistical uncertainty of $\hat{\alpha}^{g}$:

$$\Delta\hat{\alpha}^{g}_{\mathrm{stat}} = \sqrt{\frac{1}{N_{\mathrm{eff}} \cdot \sum_i \mathrm{W}_i}}$$

Local g-fraction

$$\alpha_i^{g} = \frac{p_i - p_i^{q}}{\Delta_i} \qquad \Delta_i = p_i^{g} - p_i^{q}$$

For statistical fluctuations of the data:

$$\mathrm{Var}(N_{jets} \cdot p_i) = \mu_i$$

$$\Longrightarrow \mathrm{Var}(p_i) = \frac{\pi_i}{N_{\mathrm{eff}}} \approx \frac{p_i}{N_{\mathrm{eff}}}$$

If α_i are independent (Poisson model):

$$\Longrightarrow \boxed{\mathrm{Var}(\alpha_i^{g}) = \frac{1}{N_{\mathrm{eff}} \cdot \mathrm{W}_i}}$$

with $\mathrm{W}_i \equiv \frac{\Delta_i^2}{p_i}$

$\Delta\hat{\alpha}^{g}_{\mathrm{stat}}$ takes into account stat. fluctuations of **data** p_i with **fixed model** $p_i^{q/g}$

Statistical variance $E(S_{\mathrm{w}}^2)_{\mathrm{stat}}$

For an **exact** model, all local estimators have the same expectation:

$$E(\alpha_i^{\mathrm{g}}) = \alpha$$

The weighted global estimator is

$$\hat{\alpha}^{\mathrm{g}} = \textstyle\sum_i w_i \cdot \alpha_i^{\mathrm{g}}, \quad w_i = \frac{\mathrm{W}_i}{\sum_i \mathrm{W}_i}$$

$$\mathrm{W}_i = \frac{\Delta_i^2}{p_i}$$

$$w_i = \frac{\mathrm{W}_i}{\sum_i \mathrm{W}_i}$$

Define the observed weighted spread of local estimates:

$$S_w^2 \stackrel{\mathrm{def}}{=} \sum_i w_i \cdot (\alpha_i^{\mathrm{g}} - \hat{\alpha}^{\mathrm{g}})^2$$

$$\Longrightarrow \quad E(S_w^2) = \sum_i w_i \cdot E(\alpha_i^{\mathrm{g}} - \hat{\alpha}^{\mathrm{g}})^2$$

$$E(S_w^2) = \sum_i w_i \cdot \underbrace{E\left[\left(\alpha_i^{\mathrm{g}} - \alpha\right)^2\right]}_{\boxed{\mathrm{Var}(\alpha_i^{\mathrm{g}})}} + \underbrace{E[(\alpha - \hat{\alpha}^{\mathrm{g}})^2]}_{\boxed{\mathrm{Var}(\hat{\alpha}^{\mathrm{g}}) = \sum_i w_i^2 \mathrm{Var}(\alpha_i^{\mathrm{g}})}} + \sum_i 2 \cdot w_i \cdot \underbrace{E[(\alpha_i^{\mathrm{g}} - \alpha)(\alpha - \hat{\alpha}^{\mathrm{g}})]}_{\boxed{-\mathrm{Cov}(\alpha_i^{\mathrm{g}}, \hat{\alpha}^{\mathrm{g}}) = w_i \cdot \mathrm{Var}(\alpha_i^{\mathrm{g}})}}$$

with $\mathrm{Cov}\left(\alpha_i^{\mathrm{g}}, \alpha_j^{\mathrm{g}}\right) \simeq 0 \ (i \neq j)$

$$\Longrightarrow \quad \boxed{E(S_{\mathrm{w}}^2)_{\mathrm{stat}} = \sum_i w_i \cdot (1 - w_i) \cdot \mathrm{Var}(\alpha_i^{\mathrm{g}})_{\mathrm{stat}} = (K-1) \cdot \mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{stat}}} \quad \otimes$$

$$\mathrm{Var}(\alpha_i^{\mathrm{g}})_{\mathrm{stat}} = \frac{1}{N_{\mathrm{eff}} \cdot \mathrm{W}_i}$$

$$\mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{stat}} = \frac{1}{N_{\mathrm{eff}} \cdot \sum_i \mathrm{W}_i}$$

Statistical variance $E(S_{\mathrm{w}}^2)_{\mathrm{stat}}$

Remark for p.24

$$\mathrm{Cov}\left(\alpha_i^{\mathrm{g}}, \alpha_j^{\mathrm{g}}\right) \simeq 0 \ (i \neq j)$$

In the asymptotic limit, the multinomial distribution approaches a multivariate Gaussian. For sufficiently small bin probabilities p_i (sufficiently large number of bins), the off-diagonal multinomial correlations are suppressed, allowing the covariance matrix to be approximated as diagonal.

It is the presence of the condition $p_i \ll 1$ that makes the approximation of independent α_i^{g} justified.

Model Contribution to $E(S_w^2)$

Working assumptions

Fixed observed data p_i ; imperfect q/g templates:

$$\mathrm{Var}(\alpha_i^g)_{\mathrm{model}} \simeq \sigma_{\mathrm{model}}^2 \qquad \forall\ i$$

$$\mathrm{Cov}\left(\alpha_i^g, \alpha_j^g\right) = 0 \qquad (i \neq j)$$

The bin-to-bin covariance of the model-induced variations is unknown.

Model variance of the mean

$$\mathrm{Var}(\hat{\alpha}^g)_{\mathrm{model}} \simeq \sum_i w_i^2 \cdot \sigma_{\mathrm{model}}^2$$

$$\boxed{\mathrm{Var}(\hat{\alpha}^g)_{\mathrm{model}} \simeq \frac{\sigma_{\mathrm{model}}^2}{K_{\mathrm{eff}}}}$$

$$K_{\mathrm{eff}} = \frac{1}{\sum_{i=1}^{K} w_i^2}$$

Contribution to the observed spread

$$E(S_w^2)_{\mathrm{model}} = \left(1 - \frac{1}{K_{\mathrm{eff}}}\right) \sigma_{\mathrm{model}}^2$$

or

$$\boxed{E(S_w^2)_{\mathrm{model}} = (K_{\mathrm{eff}} - 1) \cdot \mathrm{Var}(\alpha^g)_{\mathrm{model}}}$$

Derivation in Backup, p.27-29

Model variance $E(S_w^2)_{\text{model}}$

We consider only the model variation, without statistical fluctuations. Let δ_i be the model deviation of the local estimate in the i-th bin. The notation is a conceptual model, not a directly computable decomposition:

$$\alpha_i^g = \alpha + \delta_i$$

We only observe

$$\alpha_1^g, \alpha_2^g, \ldots$$

And that is why we propose to use the observed spread

$$\boxed{S_w^2 \stackrel{\text{def}}{=} \sum_{i=1}^{K} w_i \cdot (\alpha_i^g - \hat{\alpha}^g)^2}$$

as an estimate of the characteristic scale σ_{model}^2.

Our assumptions:

$$E(\delta_i) = 0$$
$$\text{Var}(\delta_i) = E\left(\delta_i^2\right) = \sigma_{\text{model}}^2$$
$$E\left(\delta_i, \delta_j\right) = 0, \qquad i \neq j$$

The latter is our working assumption that there are no correlations between model deviations.

$$\hat{\alpha}^g = \sum_i w_i \cdot \alpha_i^g = \sum_i w_i \cdot (\alpha + \delta_i) = \alpha + \delta_G, \qquad \delta_G = \sum_i w_i \cdot \delta_i$$

$$\Longrightarrow \quad \alpha_i^g - \hat{\alpha}^g = \delta_i - \delta_G$$

$$\boxed{S_w^2 = \sum_{i=1}^{K} w_i \cdot (\delta_i - \delta_G)^2}$$

Objec…

Mode… Uncer…

Analy… strate…

Statist… Framework

G-fraction: Mean vs. Local

Statistical uncertainty

Total Variance S_w^2

Model Uncertainty-I

Model Uncertainty-II

Conclusions

Model variance $E(S_w^2)_{\mathrm{model}}$

$$\hat{\alpha}^g = \sum_i w_i \cdot \alpha_i^g = \sum_i w_i \cdot (\alpha + \delta_i) = \alpha + \delta_G, \qquad \delta_G = \sum_i w_i \cdot \delta_i$$

$$\Longrightarrow \quad \alpha_i^g - \hat{\alpha}^g = \delta_i - \delta_G$$

$$S_w^2 = \sum_{i=1}^{K} w_i \cdot (\delta_i - \delta_G)^2$$

$$(\delta_i - \delta_G)^2 = \delta_i^2 - 2\delta_i\delta_G + \delta_G^2.$$

$$S_w^2 = \sum_i w_i\delta_i^2 - 2\sum_i w_i\delta_i\delta_G + \sum_i w_i\delta_G^2.$$

$$\sum_i w_i\delta_i\delta_G = \delta_G \sum_i w_i\delta_i.$$

$$\sum_i w_i\delta_i = \delta_G.$$

$$\sum_i w_i\delta_G^2 = \delta_G^2.$$

$$\Longrightarrow \quad S_w^2 = \sum w_i\delta_i^2 - 2\delta_G^2 + \delta_G^2,$$

$$\Longrightarrow \quad E(S_w^2) = E\left(\sum_i w_i\delta_i^2\right) - E(\delta_G^2).$$

$$E(\delta_i^2) = \sigma_{\mathrm{model}}^2, \quad \Longrightarrow \quad \boxed{E\left(\sum_i w_i\delta_i^2\right) = \sigma_{\mathrm{model}}^2.} \quad \boxed{E(\delta_G^2) = \sigma_{\mathrm{model}}^2 \sum_i w_i^2.}$$

$$\boxed{E(S_w^2)_{\mathrm{model}} = \left(1 - \sum_i w_i^2\right) \cdot \sigma_{\mathrm{model}}^2}$$

Model variance $E(S_w^2)_{\text{model}}$

$$\boxed{E(S_w^2)_{\text{model}} = \left(1 - \sum_i w_i^2\right) \cdot \sigma^2_{\text{model}} = \sigma^2_{\text{model}} \cdot \left(1 - \frac{1}{K_{eff}}\right)}$$

$$\hat{\alpha}^{\text{g}} = \sum_i w_i \cdot \alpha_i^{\text{g}} \qquad \text{Var}(\hat{\alpha}^{\text{g}}) = \sum_i w_i^2 \,\text{Var}(\alpha_i^{\text{g}}) = \frac{\sigma^2_{\text{model}}}{K_{eff}}$$

$$\text{Var}(\alpha_i^{\text{g}}) = \text{Var}(\delta_i) = \sigma^2_{\text{model}}$$

$$\Rightarrow \quad \boxed{E(S_w^2)_{\text{model}} = (K_{\text{eff}} - 1) \cdot \text{Var}(\hat{\alpha}^{\text{g}})_{\text{model}}} \quad \otimes$$

Statistical Uncertainty of the Mean g-Jet Fraction with Free Template Modification Parameters

Fixed templates parameters (a, b):

$$\mathrm{Var}(\hat{\alpha}^{\mathrm{g}}(a,b))_{\mathrm{stat}} = \frac{1}{N_{\mathrm{eff}} \cdot \sum_i \mathrm{W}_i}$$

Free template parameters

The template parameters (a, b) *are fitted from data:*

$$\mathrm{Cov}(a,b)_{\mathrm{stat}} = V_{a,b}$$

Since

$$\hat{\alpha}^{\mathrm{g}}(a,b)$$

their statistical fluctuations contribute to the uncertainty of $\hat{\alpha}^{\mathrm{g}}$.

Using first-order error propagation,

$$\mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{stat}} = \mathrm{Var}(\hat{\alpha}^{\mathrm{g}}(a,b))_{\mathrm{stat}} + \begin{Bmatrix} \frac{\partial\hat{\alpha}^{\mathrm{g}}}{\partial a} & \frac{\partial\hat{\alpha}^{\mathrm{g}}}{\partial b} \end{Bmatrix} \begin{Bmatrix} V_{a,a} & V_{b,a} \\ V_{a,b} & V_{b,b} \end{Bmatrix} \begin{Bmatrix} \frac{\partial\hat{\alpha}^{\mathrm{g}}}{\partial a} \\ \frac{\partial\hat{\alpha}^{\mathrm{g}}}{\partial b} \end{Bmatrix}$$

Therefore

$$\Delta\hat{\alpha}^{\mathrm{g}}_{\mathrm{stat}} = \sqrt{\mathrm{Var}(\hat{\alpha}^{\mathrm{g}})_{\mathrm{stat}}}$$

With free templates, the statistical uncertainty of $\hat{\alpha}^{\mathrm{g}}$ includes the statistical uncertainty of the fitted template parameters (a, b).