Robust Measures of Resolution

Should we use rms_{90} in ILD performance measurements?

Graham W. Wilson

University of Kansas

October 20, 2018

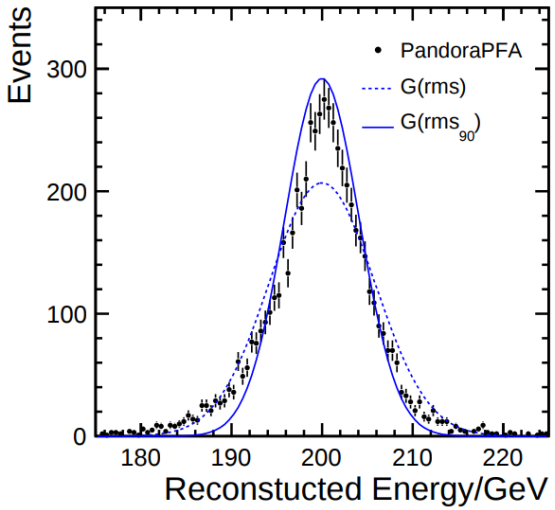


Figure 7: The total reconstructed energy from reconstructed PFOs in 200 GeV $Z \to uds$ events for initial quark directions within the polar angle acceptance $|\cos\theta_{qq}| < 0.7$. The dotted line shows the best fit Gaussian distribution with an rms of 5.8 GeV. The solid line shows a Gaussian distribution, normalised to the same number of events, with standard deviation

History

PFA performance metric

- For more than a decade, rms_{90} used as the basic statistical estimator of jet energy resolution in LC detector studies
- rms_{90} is defined as the rms of the smallest interval with ≥ 90 % of the events
- Implemented in PandoraAnalysis using binned histograms (0.05 GeV bins)

Pros of rms_{90}

1. Does not depend on a parametrization or assumption of a Gaussian etc
2. Tolerant of asymmetric distribution
3. Tolerant of outliers (robust)

Cons of rms_{90}

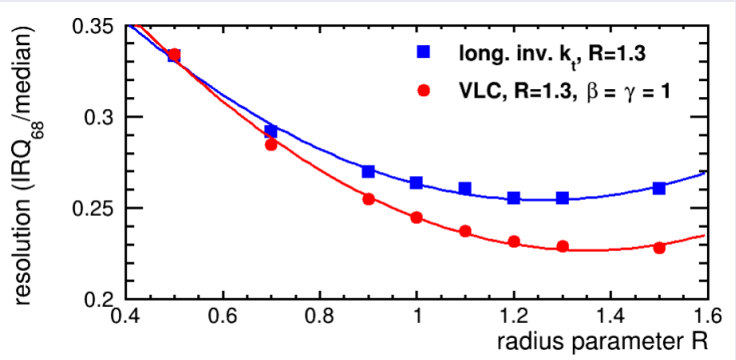
1. Severely biased for Gaussian (if no consistency factor applied)
2. Usually implemented with binned data (loss of information)
3. May not be the most statistically efficient
4. Insensitive to high kurtosis (fat tails) or low kurtosis (peaked)

Robust Measures of Scale

Resolution Measurement

- There are a number of metrics in the robust statistics literature that are well suited to resolution measurement and presentation
- We are all familiar with robust measures of location such as the median that reduce the sensitivity to outliers.
- There are also "robust measures of scale" suited to resolution measurement

Simple Examples

1. IQR. Inter-Quartile Range. $Q_3 - Q_1$ ie. $x_{75\%} - x_{25\%}$.
2. MAD. Median of the absolute deviations from the median.
3. Inter-Quantile Range. For example, $x_{95\%} - x_{5\%}$ (IQR90) etc.
4. 

5. CLIC HH study uses IQR68 ($x_{84\%} - x_{16\%}$), arXiv:1607.05039. CMS also.

Trimming/Winsorization Examples

Many different robust estimators have been developed. The primary focus is often ultra-robustness/break-down point rather than our usage case of being a parameter estimator.

Trimmed RMS

- Most intuitive is to trim symmetrically. For example discarding the first 5% and last 5% of the distribution.
- Then compute the rms of the remaining 90% $(x_{0.05}, x_{0.95})$.
- Compute a correction factor for consistency with a Gaussian.
- Performance very similar to rms90.

Winsorized RMS

- Similar to a symmetric trimmed RMS, but each outlier is not discarded.
- Each is replaced by the trimmed minimum, $x_{0.05}$, or the trimmed maximum, $x_{0.95}$, as appropriate.
- Then compute the rms from all the entries. (100% $(x_{0.05}, x_{0.95})$).
- Compute a correction factor for consistency with a Gaussian.
- Performance of Wrms90 is better than rms90.

More Complex Examples

Based on pairwise estimators.

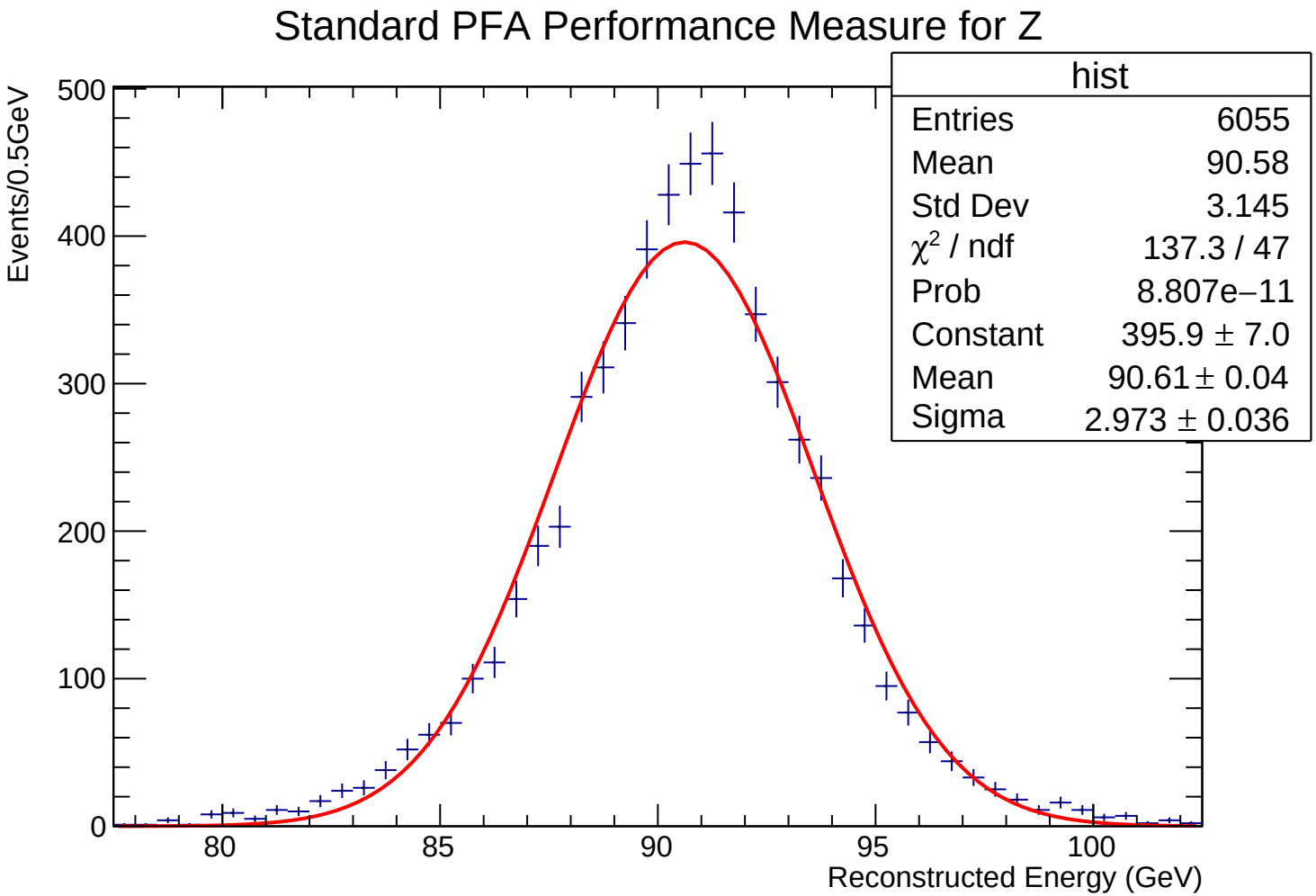
1. S_n of Rousseeuw, Croux. Defined as $\mathrm{med}_i(\mathrm{med}_j(|x_i - x_j|))$.
2. Q_n of Rousseeuw, Croux. Defined as the first quartile of all pair-wise absolute differences.
3. P_n of Tarr, Müller, Weber. Defined as the IQR (IQR50) of the pair-wise means.

Why use these? Robust and much better statistical efficiency than MAD and IQR.

Q_n could easily be generalized to something that is more sensitive to the tails by choosing a different quantile.

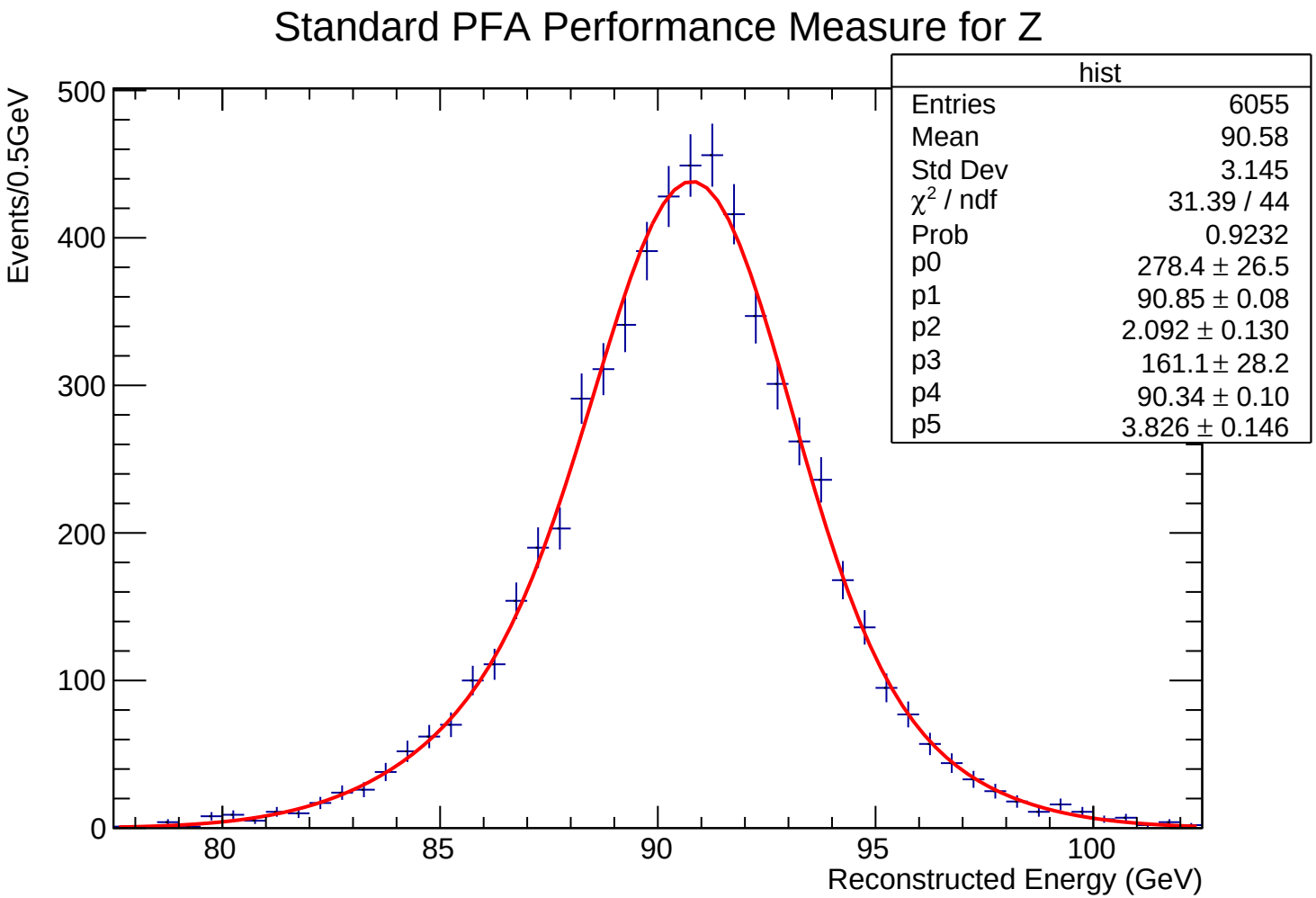
P_n could easily be generalized by choosing other quantiles for the range.

$q\bar{q}(q = u, d, s)$ at $\sqrt{s} = 91$ GeV



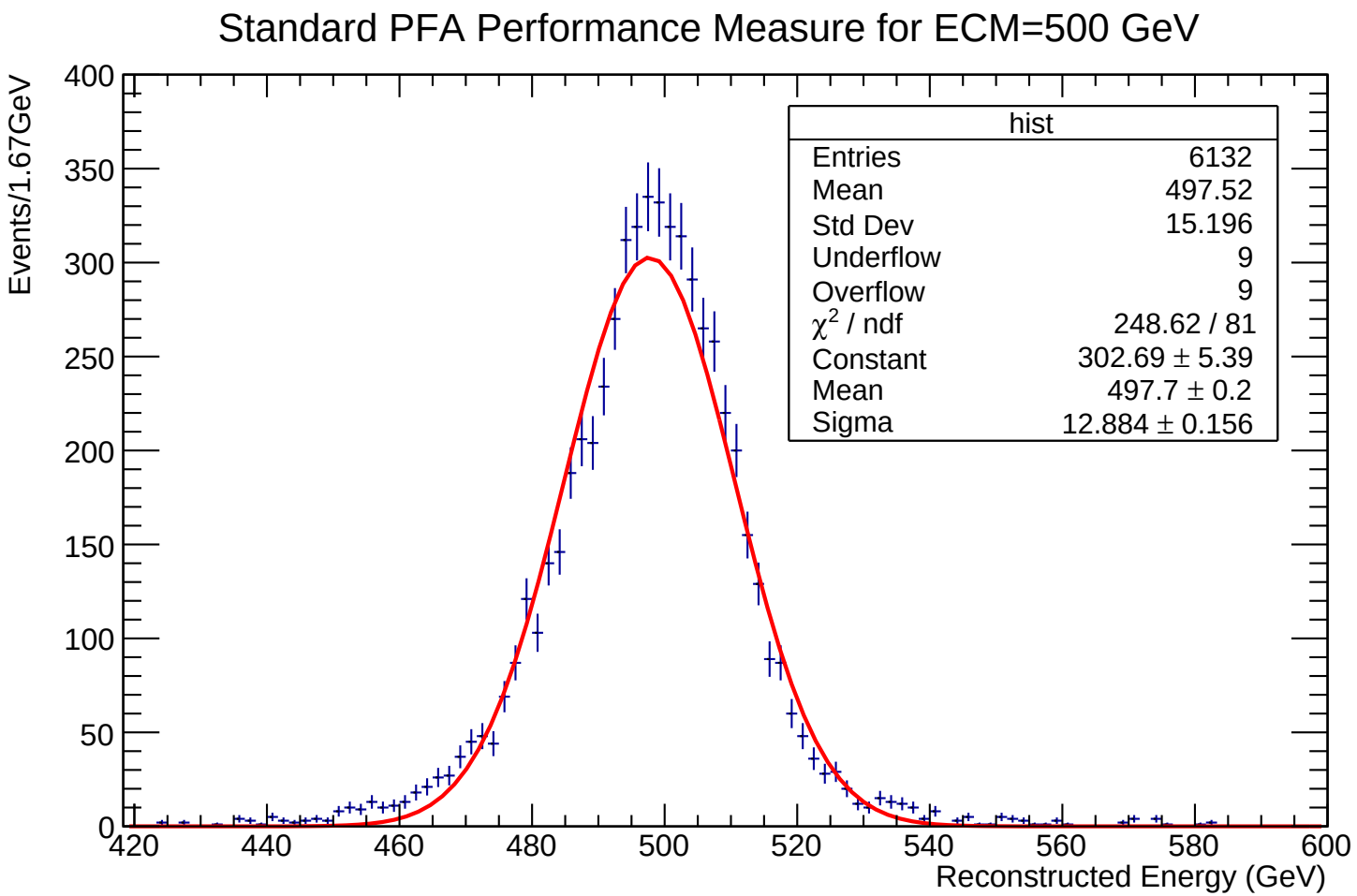
Distribution is not so non-Gaussian.

$q\bar{q}(q = u, d, s)$ at $\sqrt{s} = 91$ GeV



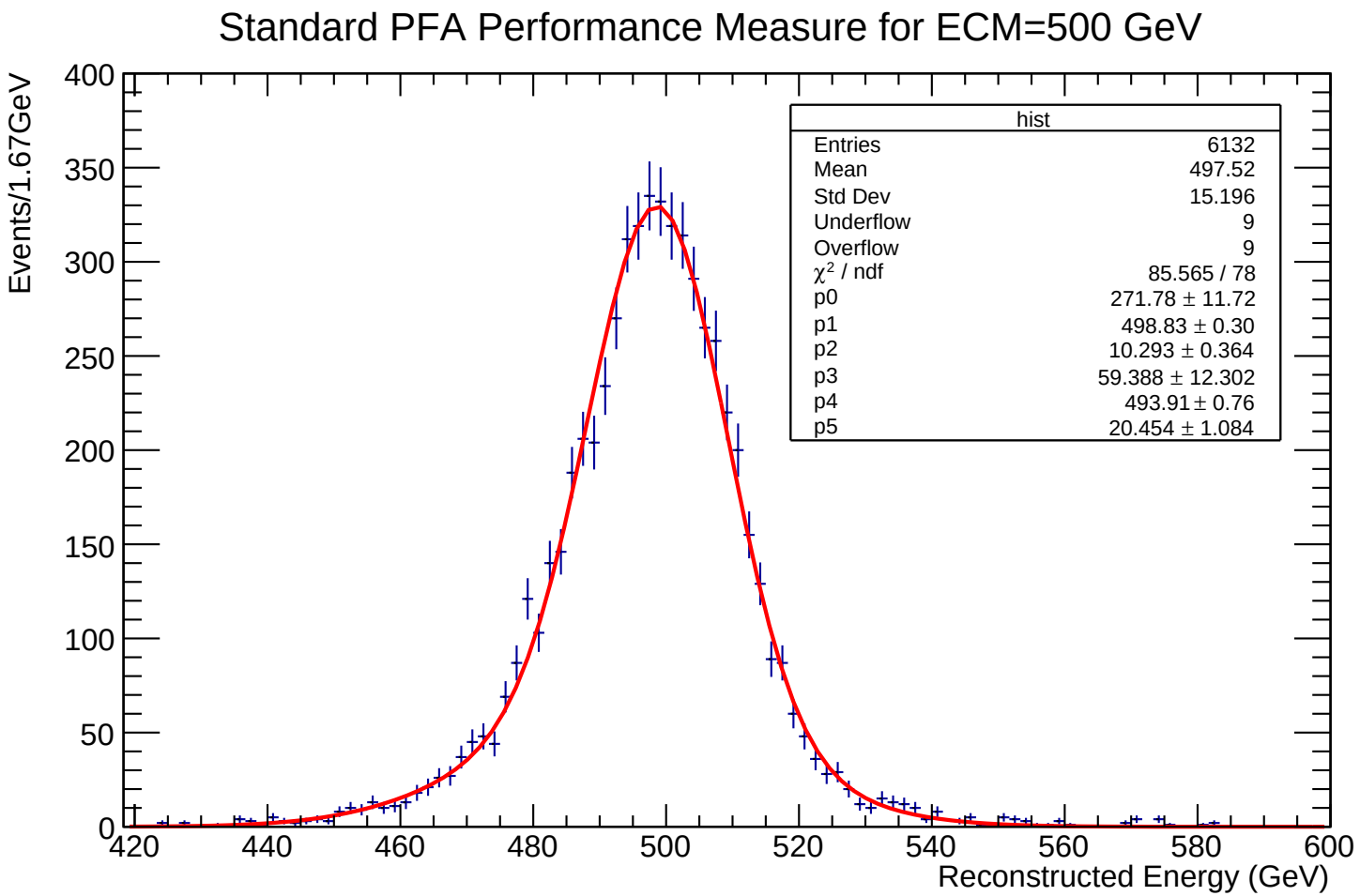
At $\sqrt{s} = 91$ GeV can be described very adequately with a double Gaussian with 5 shape parameters $(f_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$. Note small asymmetry: $\mu_1 \neq \mu_2$. The corresponding 5-parameter unbinned likelihood fit to this range is used in later toy MC studies.

$q\bar{q}(q = u, d, s)$ at $\sqrt{s} = 500$ GeV



Distribution is less Gaussian

$q\bar{q}(q = u, d, s)$ at $\sqrt{s} = 500$ GeV



But again, now at $\sqrt{s} = 500$ GeV, very adequately described with a double Gaussian with 5 shape parameters $(f_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$. Note small asymmetry: $\mu_1 \neq \mu_2$. The corresponding 5-parameter unbinned likelihood fit to this range is used in later toy MC studies.

Desirable Statistical Properties of Resolution Estimators

Mathematical statistics literature discusses ...

High Breakdown Point Insensitive to outliers

Efficiency $\varepsilon \equiv \mathrm{var}_{MVU}/\mathrm{var}(X)$ where var_{MVU} is the minimum variance unbiased estimator related to the Cramer-Rao bound (Fisher information ...).

Asymptotic Gaussian Efficiency For the Gaussian case the MVU estimator is known (close to the SD). So efficiency of estimator X is defined as $\mathrm{SD}^2/\mathrm{var}(X)$

Consistency Should give close to the Gaussian number if the distribution is Gaussian.

Deals with Skewness

See for example, Huber. Robust Statistics.

The statistics considered

1. SD
2. rms90 (binned)
3. Rms90* (discrete and calibrated)
4. Wrms90 (Winsorized)
5. IQR90
6. MAD
7. S_n
8. Q_n
9. P_n

All can be calibrated with toy MC to give unbiased estimates for a true Gaussian

- Gaussian consistency factor
- Small sample size factor
- Uncertainty estimate (related to asymptotic efficiency)

For emphasis and consistent with current practice, the Gaussian consistency factor was not applied to rms90. Rms90* is calibrated.

Resolution Estimators Implementation

Naive

- The pairwise estimators, S_n, Q_n and P_n, consider all $\frac{1}{2}n(n-1)$ distinct pairs in a sample of size n.
- The straightforward - but expensive in CPU and memory - naive algorithm is to store all these, sort them and find the requisite quantiles
- Pretty bad scaling though: $\mathcal{O}(n^2 \log n)$

Less Naive

- Implement $\mathcal{O}(n \log n)$ algorithms from Croux, Rousseeuw for S_n and Q_n (translated from f77 to C^{++}).
- Validate vs naive algorithms. (Note original code was limited to $n \leq 500$)
- Discovered that STL has powerful "nth_element" selection method in vectors that makes sorting unnecessary. Use this too.

Overall now have relatively painless implementation for $n \leq 10000$. Works for $n = 40000$. For $n = 50000$, exceeds 8GB RAM.

Toy MC Studies

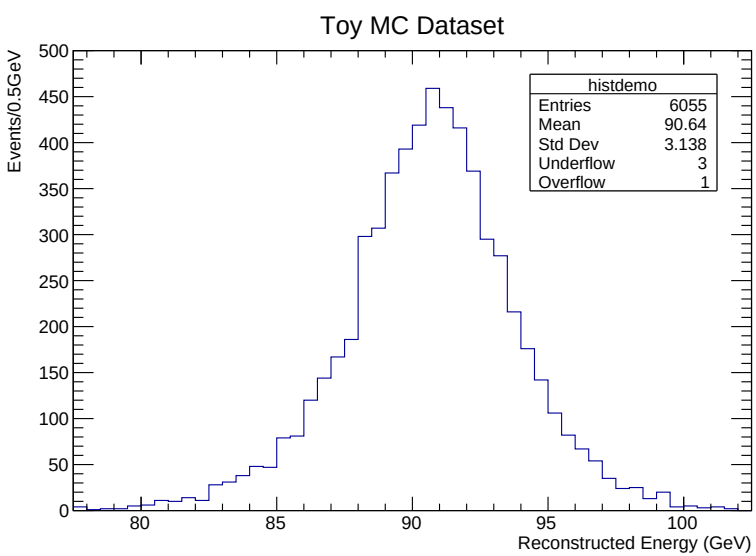
1. "Calibration"

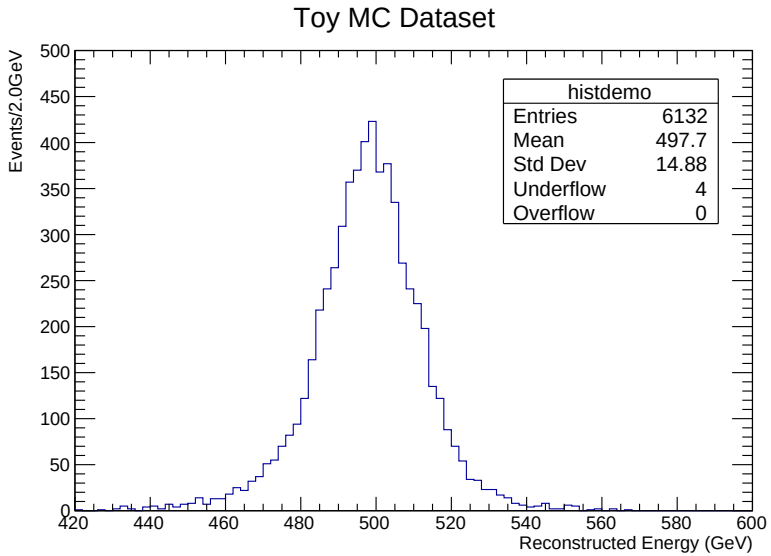
Generate Gaussian data-sets

2. PFA Response

Use fitted double Gaussian parametrizations shown earlier to generate data-sets from these underlying distributions

Some examples

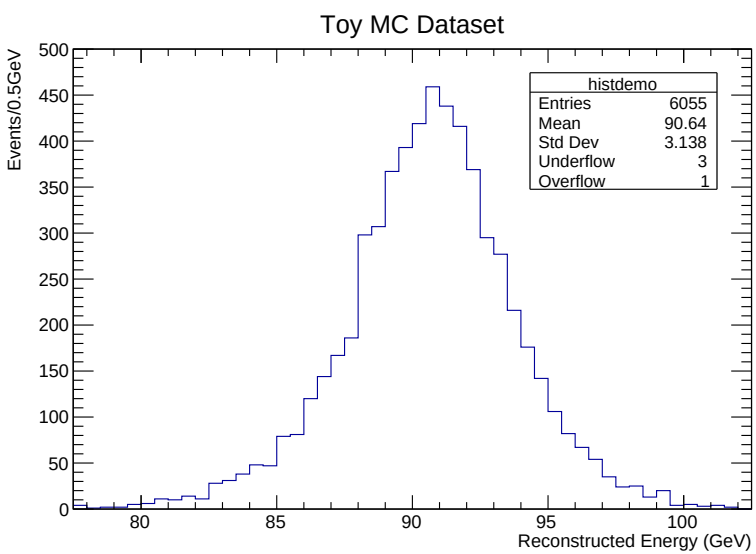


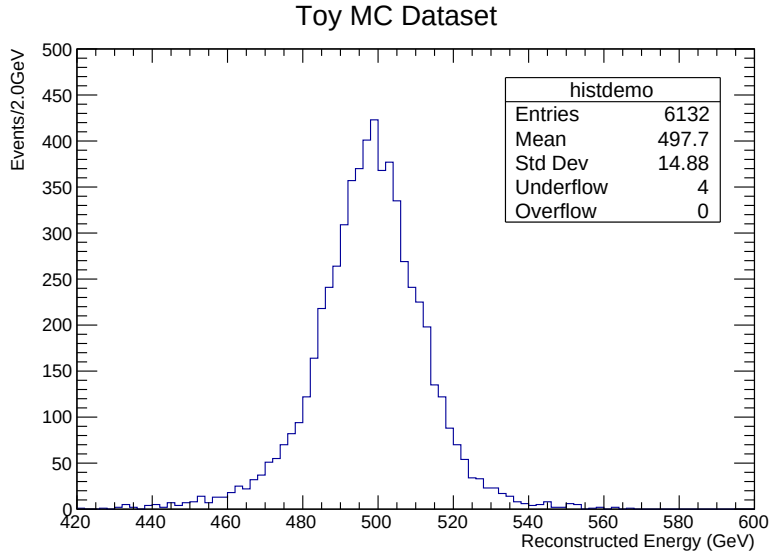


We'll see shortly how well the various statistical estimators work.
What would be the "correct" resolution estimate?

"Correct Resolution Estimate"

Base on how well one can measure the mean.





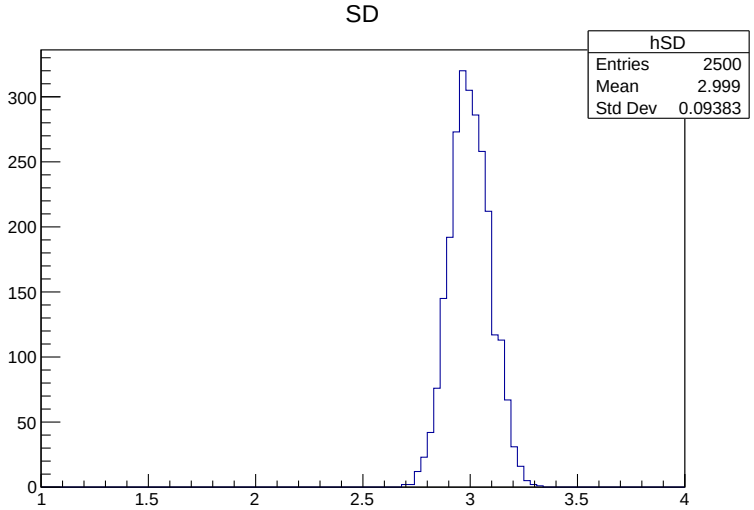
Use 1M events generated from each pdf. Do unbinned likelihood fit for the overall mean (all other parameters fixed to true values - assumes shape known exactly).
Use 1000 $\times$ the error on the mean to infer the per event effective resolution.
Results

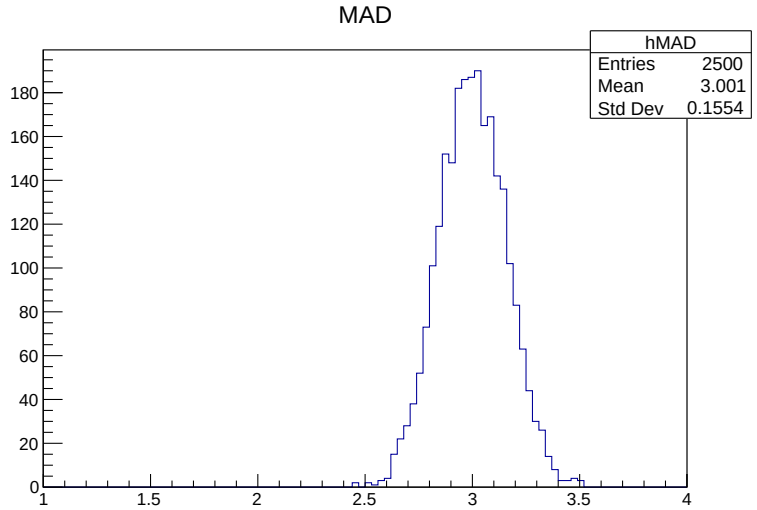
- $\sigma_{91} = 2.981 \pm 0.002$ GeV.
- $\sigma_{500} = 13.39 \pm 0.01$ GeV.

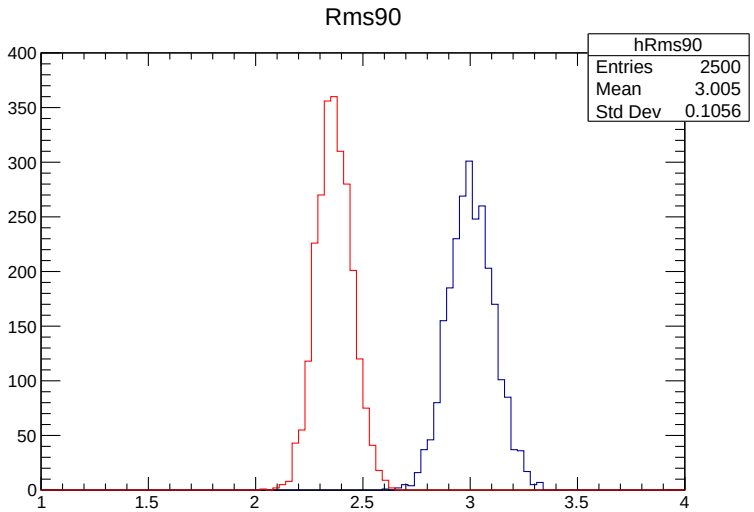
Not that different from the sample rms values ...
This corresponds I think to evaluating the Fisher information.

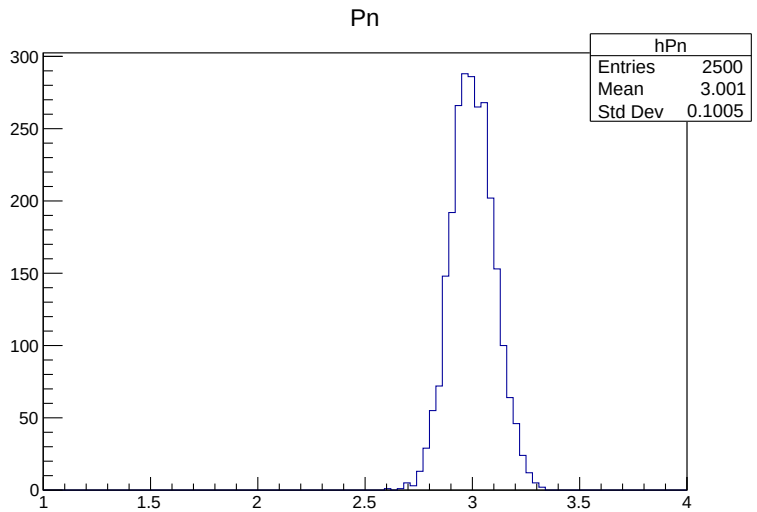
Gaussian Model with $\sigma = 3$ GeV and $\sqrt{s} = 91$ GeV

2500 repetitions with 500 events in each experiment









rms90 severely biased. For Gaussian, SD is by definition best in "efficiency" (100%). MAD is poor (36%). Rms90* is 77%. Wrms90 is 84%, P_n is 87%.

Gaussian Model Results ($\mu = 91.0$, $\sigma = 3.0$)

Toy MC

Estimator	Sample Mean (GeV)	Efficiency (%)
SD	2.996 ± 0.002	100.0
rms90	2.362 ± 0.002	77.3
Rms90*	3.000 ± 0.002	77.3
Wrms90	2.995 ± 0.002	84.2
IQR90	3.013 ± 0.002	64.3
MAD	2.994 ± 0.003	36.2
S_n	2.995 ± 0.002	57.5
Q_n	2.999 ± 0.002	80.4
P_n	2.997 ± 0.002	85.5

- As expected, rms90 underestimates resolution of a true Gaussian. Needs to be scaled up by a factor of 1.27.
- All other estimators have good accuracy (by design).
- rms90 and Rms90* are basically 100% correlated (+99.996%).
- Rms90*, Wrms90, Q_n and P_n stand out as having the best efficiency (ie precision for Gaussian) of the robust estimators.

91 GeV Double Gaussian Model Results

Toy MC ("Correct answer": $\sigma_{91} = 2.981 \pm 0.002$ GeV)

Estimator	Sample Mean (GeV)	Sample S.D. (MeV)
SD	3.139 ± 0.002	118
rms90	2.320 ± 0.002	93
Rms90*	2.946 ± 0.002	118
Wrms90	3.011 ± 0.002	121
IQR90	3.165 ± 0.003	153
MAD	2.806 ± 0.003	152
S_n	2.896 ± 0.003	135
Q_n	2.942 ± 0.002	118
P_n	2.975 ± 0.002	116

This time, rms90 underestimates resolution by 1/1.285. SD, MAD, IQR90 not very accurate.

500 GeV Double Gaussian Model Results

Toy MC ("Correct answer": $\sigma_{500} = 13.39 \pm 0.01$ GeV)

Estimator	Sample Mean (GeV)	Sample S.D. (MeV)
SD	15.077 ± 0.014	703
rms90	10.451 ± 0.009	428
Rms90*	13.273 ± 0.011	543
Wrms90	13.637 ± 0.012	599
IQR90	14.486 ± 0.017	834
MAD	12.703 ± 0.014	684
S_n	13.089 ± 0.012	605
Q_n	13.318 ± 0.011	548
P_n	13.503 ± 0.011	551

This time, rms90 underestimates resolution by 1/1.28. SD, MAD, IQR90 not very accurate.

Actual Detector Simulation Results

Old 91 GeV sample (v01-17-08), and recent 500 GeV one (ILD_l5_o1_v02) with v02-00

Estimator	σ_{91} (GeV)	σ_{500} (GeV)
"correct"	(2.981)	(13.39)
SD	3.243 ± 0.029	24.86 ± 0.22
rms90	2.331 ± 0.021	10.50 ± 0.09
Rms90*	2.960 ± 0.028	13.34 ± 0.13
Wrms90	3.023 ± 0.027	13.76 ± 0.12
IQR90	3.166 ± 0.046	14.52 ± 0.21
MAD	2.819 ± 0.042	12.82 ± 0.19
S_n	2.893 ± 0.034	13.09 ± 0.15
Q_n	2.954 ± 0.030	13.37 ± 0.13
P_n	2.989 ± 0.029	13.62 ± 0.13

Rms90*, Wrms90, Q_n and P_n give a consistent picture. rms90 underestimates actual resolution by factors of $1/1.28$.

Conclusions To Date

- Four estimators stand out as very reasonable for the distributions studied both in precision and accuracy. Namely, Rms90*, Wrms90, Q_n and P_n.
- The distributions are now a lot more Gaussian than I had naively thought. So all this is a bit moot/overkill.
- Note if the distributions were to be more different in shape the picture may not be so stable.

Correlation Coefficients (%)

Estimators	Gaussian	91 GeV	500 GeV
(Rms90*, Wrms90)	+96.2	+94.7	+93.5
(Rms90*, Q_n)	+98.3	+97.0	+97.5
(Rms90*, P_n)	+98.5	+98.5	+98.6
(Wrms90, Q_n)	+95.0	+89.2	+88.5
(Wrms90, P_n)	+96.4	+94.2	+93.7
(Q_n, P_n)	+98.9	+97.9	+98.3

Very large positive correlations $\Longrightarrow$ not much more information from Wrms90, Q_n and P_n that is not already contained in Rms90*.

Summary and Recommendations

- Investigated a number of more standard robust measures of resolution.
- rms90 - a specific example of a trimmed rms, behaves well, but very biased and unmotivated. (Usually they are trimmed symmetrically or "Winsorized").
- Rms90* (corrected rms90), Wrms90, and two alternative measures, Q_n and P_n, based on pairs of data-points, are very similar. (Latter three are better but not decisively so).
- If we continue to use rms90, let's be honest and quote rms^*_{90}. At a minimum, it just needs a simple multiplicative factor of 1.27.
- There may be something to be said for also including measures more sensitive to the central region and also the tails.
- Maybe not much to gain beyond applying the consistency factor to rms90 especially if one insists on one unique statistic to summarize the distribution (more conventional measures such as IQR68 are insensitive to the tails).

Code is on github at ILDbench_WWqqlnu\graham\RobustResolution

References I


C. Croux and P. Rousseeuw, "Time-efficient algorithms for two highly robust estimators of scale", Computational Statistics, 1 (1992) 411-428.


G. Tarr, S. Müller, N. Weber, "A robust scale estimator based on pairwise means", Journal of Nonparametric Statistics 24 (2012) p187-199

Backup Slides