PANDA Quality Measures for PID Classification Problems

PID Workshop GSI

May 2017

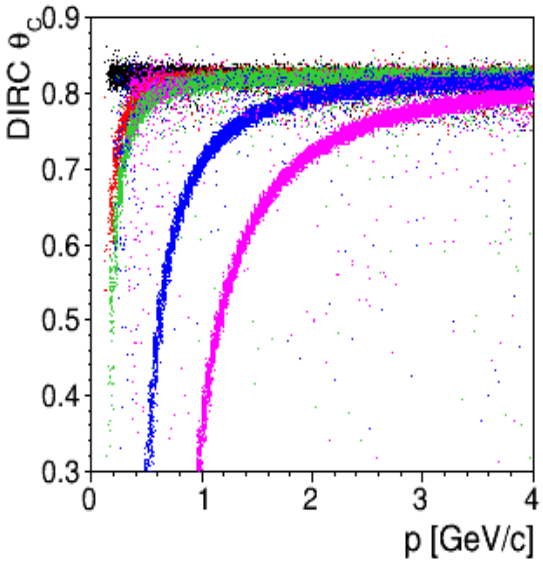
K. Götzen

GSI Darmstadt

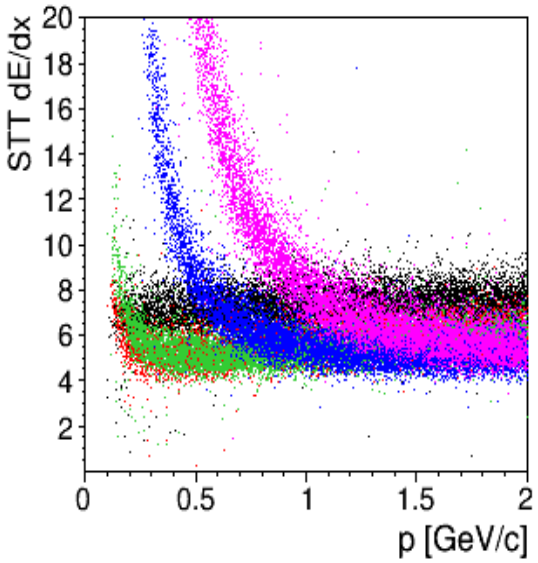
Quality Measures for Classification

- Particle identification = classification problem
- Q: How well are (two or more) particle types separable based on combined detector output?
  - Ideally independent of a concrete classification involving efficiency, purity or fraction of mis-identification (mis-ID)

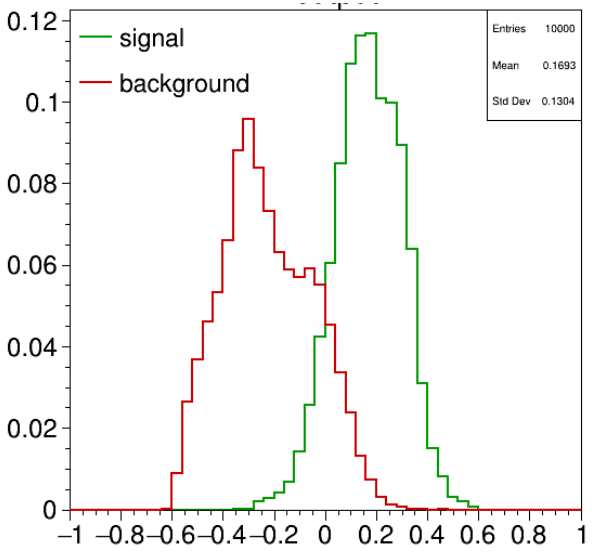
FULL: DIRC θ_C



FULL: STT dE/dx



combined information



Standard Quantity Definitions

Ingredients:

- True signal S_0, accepted signal S
- True background B_0, accepted background B

$$\text{efficiency} = \frac{S}{S_0} = \frac{\text{correctly identified signals}}{\text{all signal events}}$$

$$\text{mis\_id} = \frac{B}{B_0} = \frac{\text{wrongly accepted background}}{\text{all background events}}$$

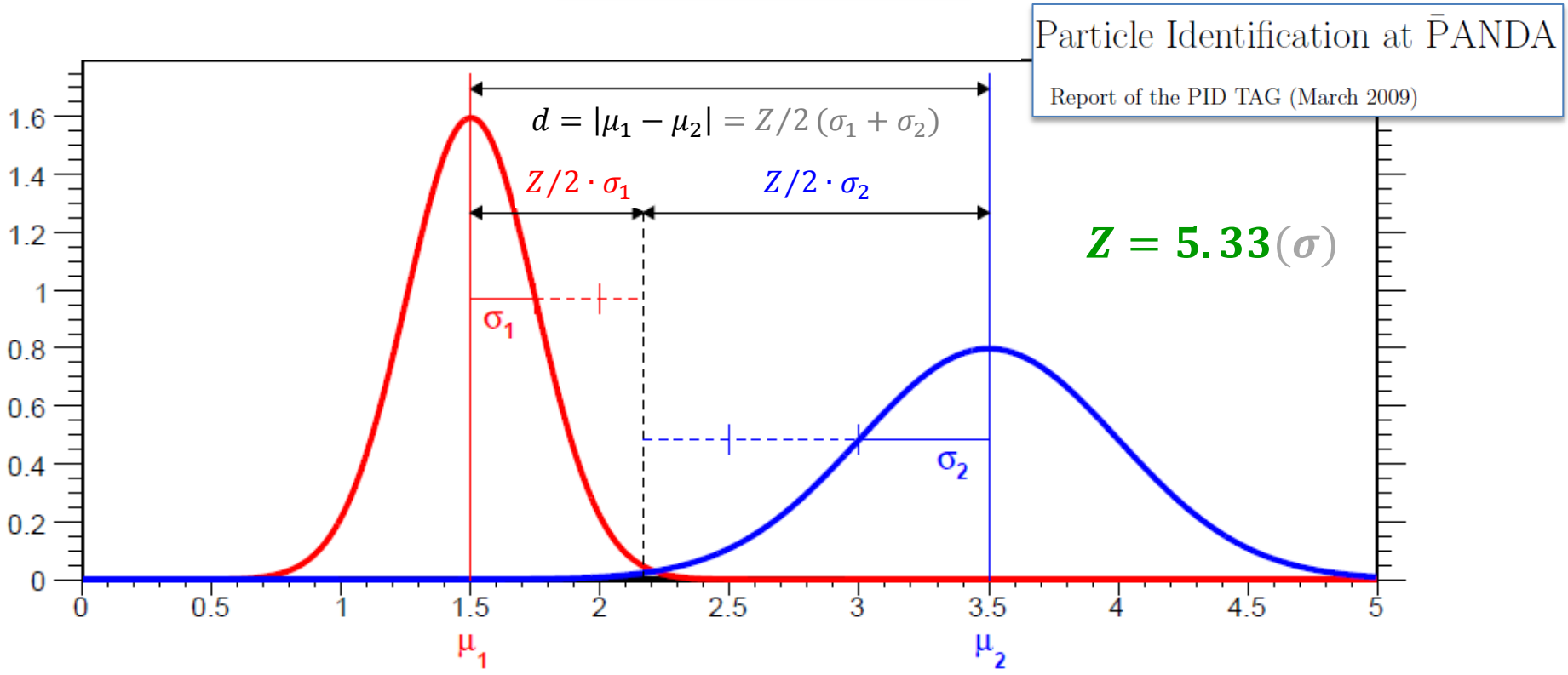
$$\text{purity} = \frac{S}{S+B} = \frac{\text{correctly identified signals}}{\text{all selected events}}$$

depends on relative fluxes → not used here

Note: All above need a concrete selection (e.g. a cut) to be computed!

Separation Power defined in PID TAG Report

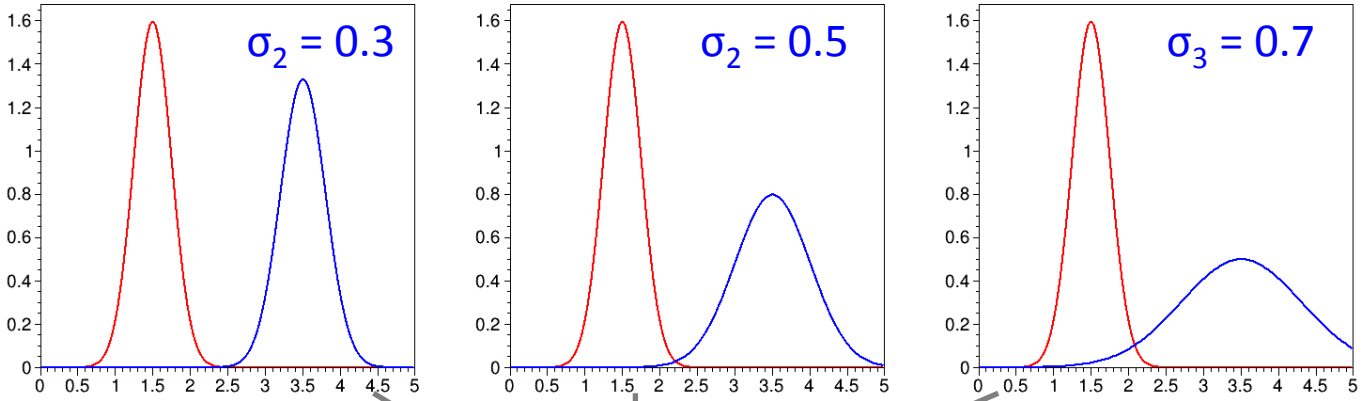
- Separation power Z is the distance of two Gaussian likelihoods in units of standard deviations
- PANDA agreed on $$Z = \frac{|\mu_1 - \mu_2|}{0.5(\sigma_1 + \sigma_2)}$$

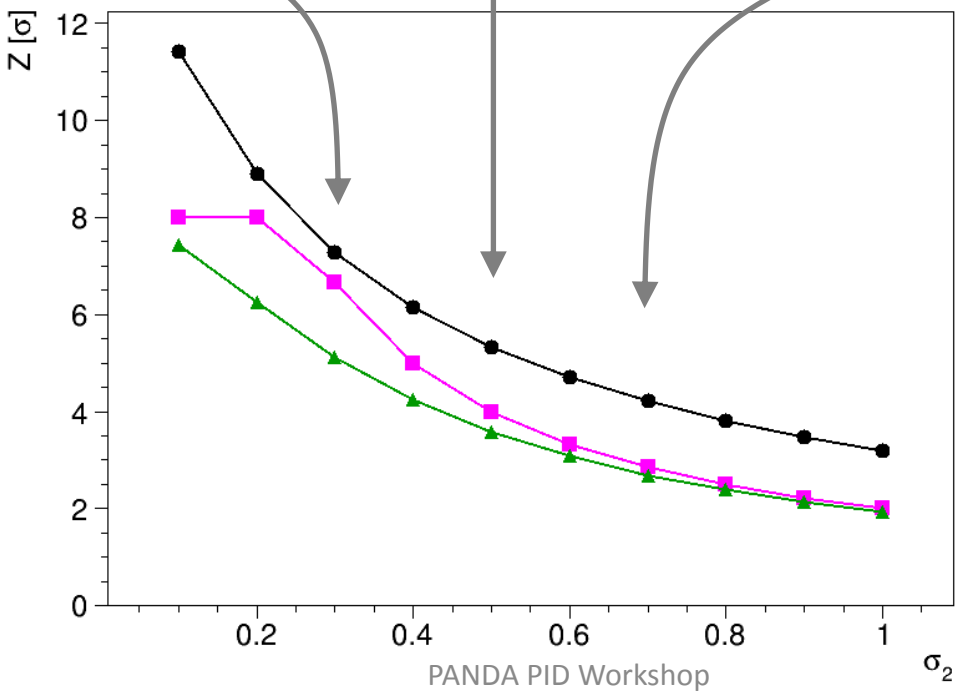


Alternative Definitions

- There are other possibilities as well

$\sigma_1 = 0.25$

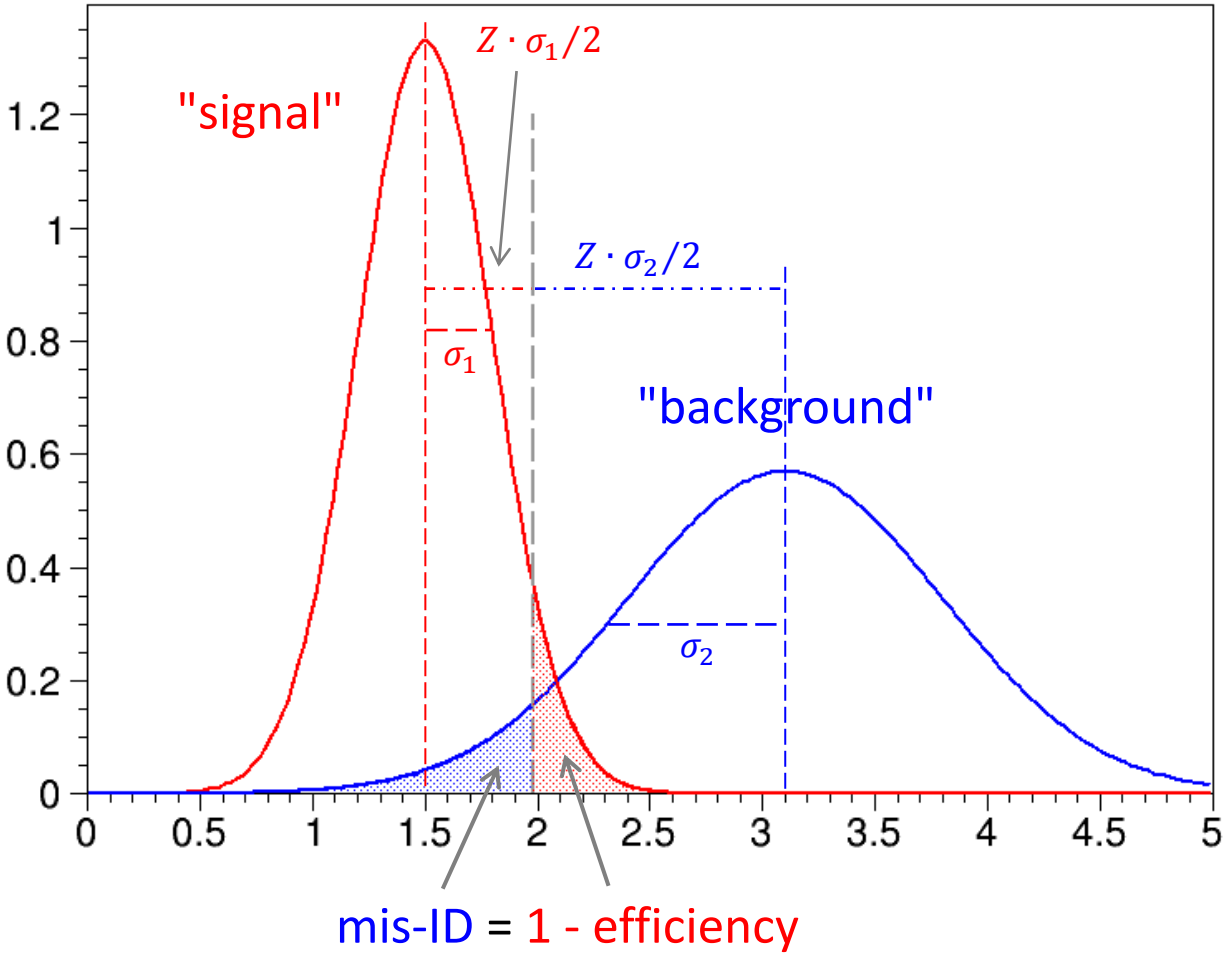




- $Z = \dfrac{|\mu_1 - \mu_2|}{0.5(\sigma_1 + \sigma_2)}$
- $Z = \dfrac{|\mu_1 - \mu_2|}{\max(\sigma_1, \sigma_2)}$
- $Z = \dfrac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$

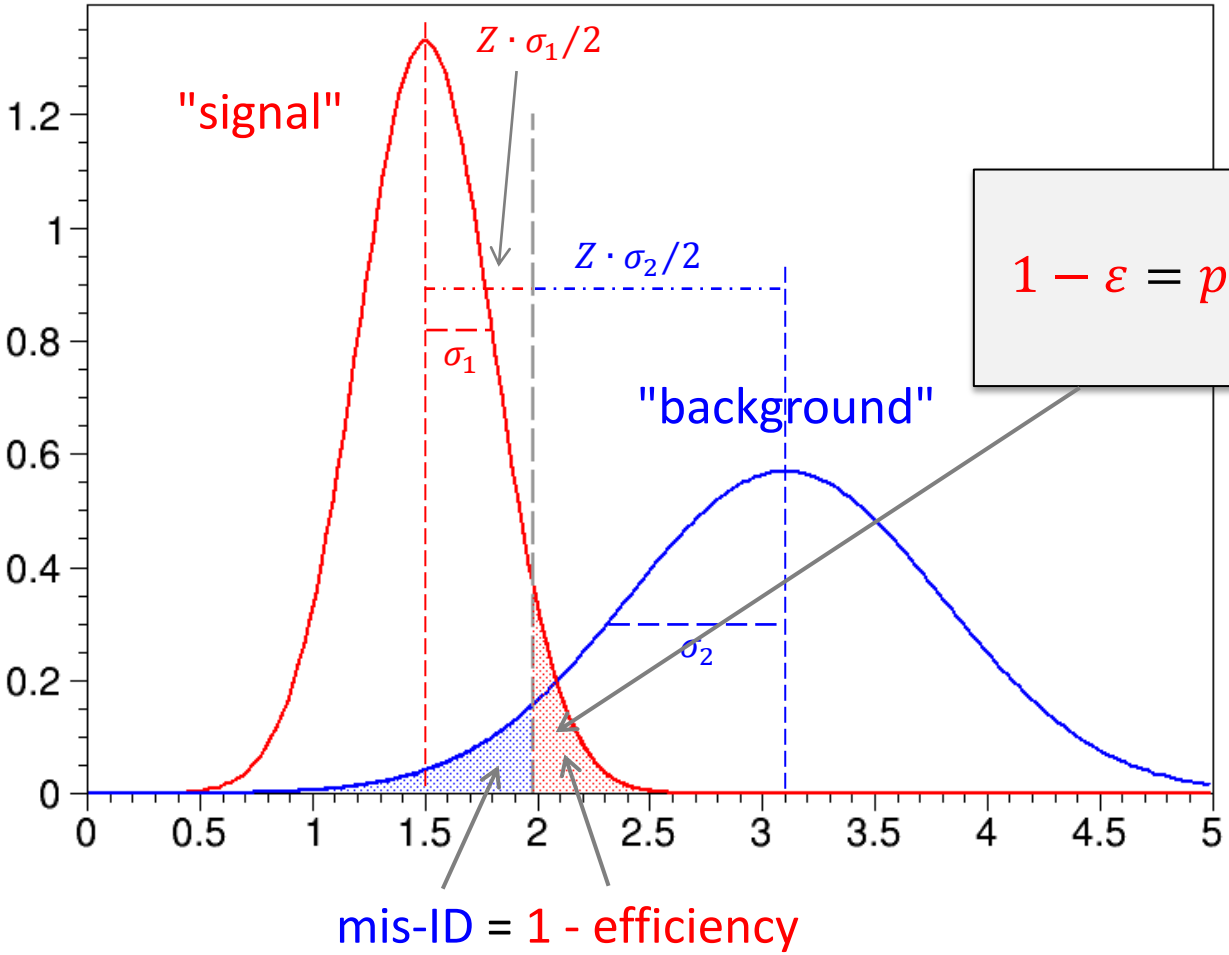
Separation Power and p-Values

- Start thinking in p-values rather than mean value distances!



Separation Power and p-Values

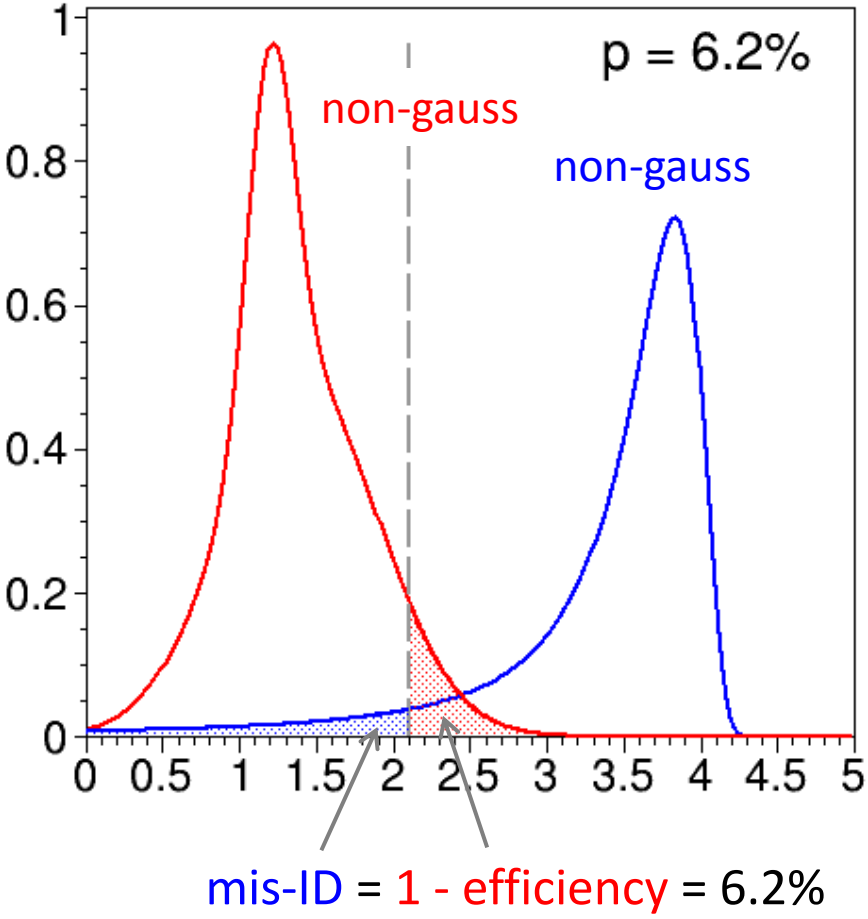
- Start thinking in p-values rather than mean value distances!
- Separation power can be mapped to a certain p-value

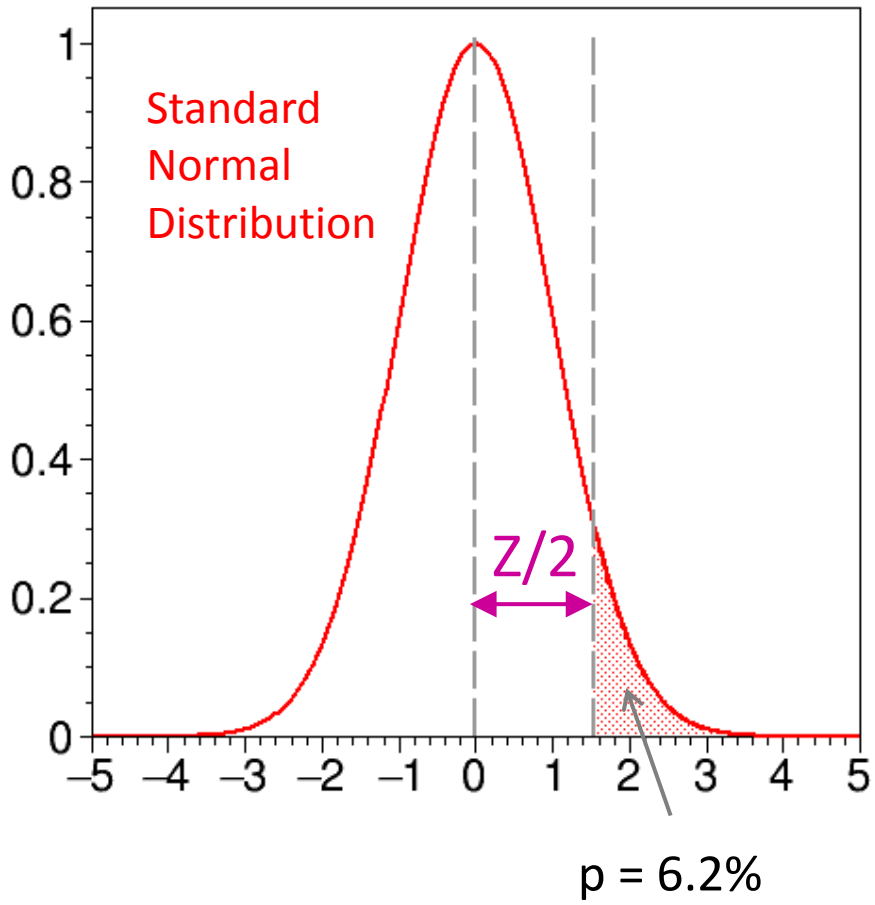


$$1 - \varepsilon = p(Z) = \int_{Z/2}^{\infty} N(x; 0, 1)\, dx$$

How treat non-Gaussian Likelihoods?

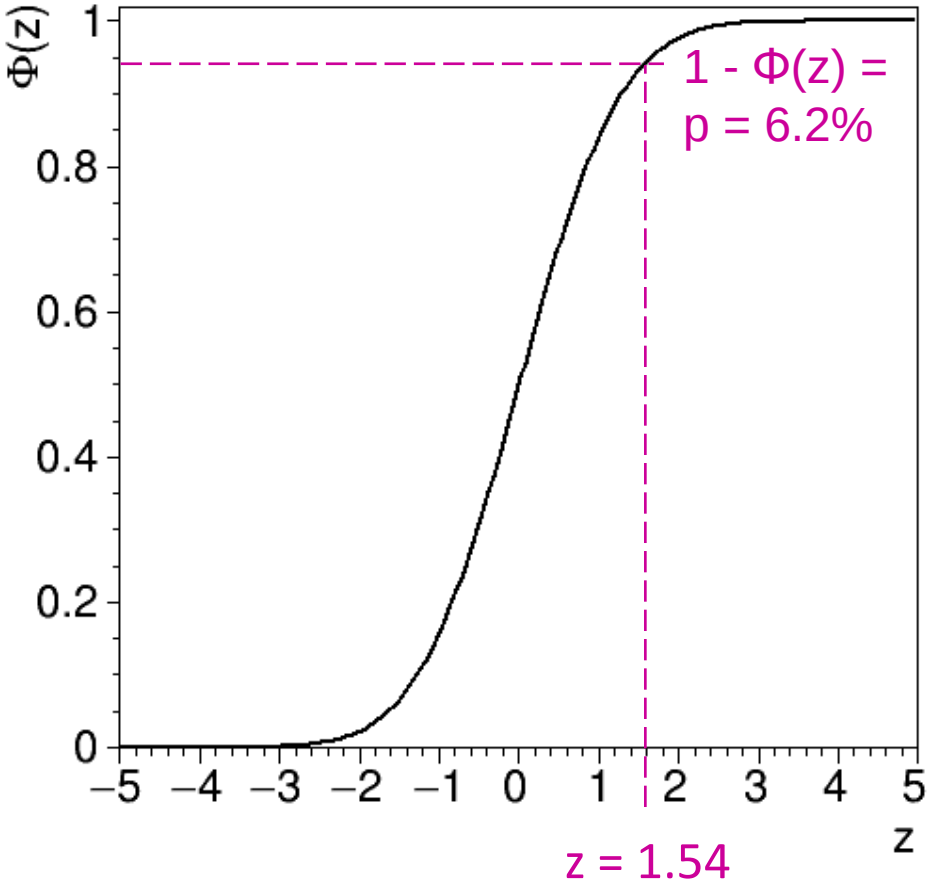
- Find cut with mis-ID = 1 - efficiency = p-value → find Gaussian quantile → compute Z = 2 · quantile of standard Gauss

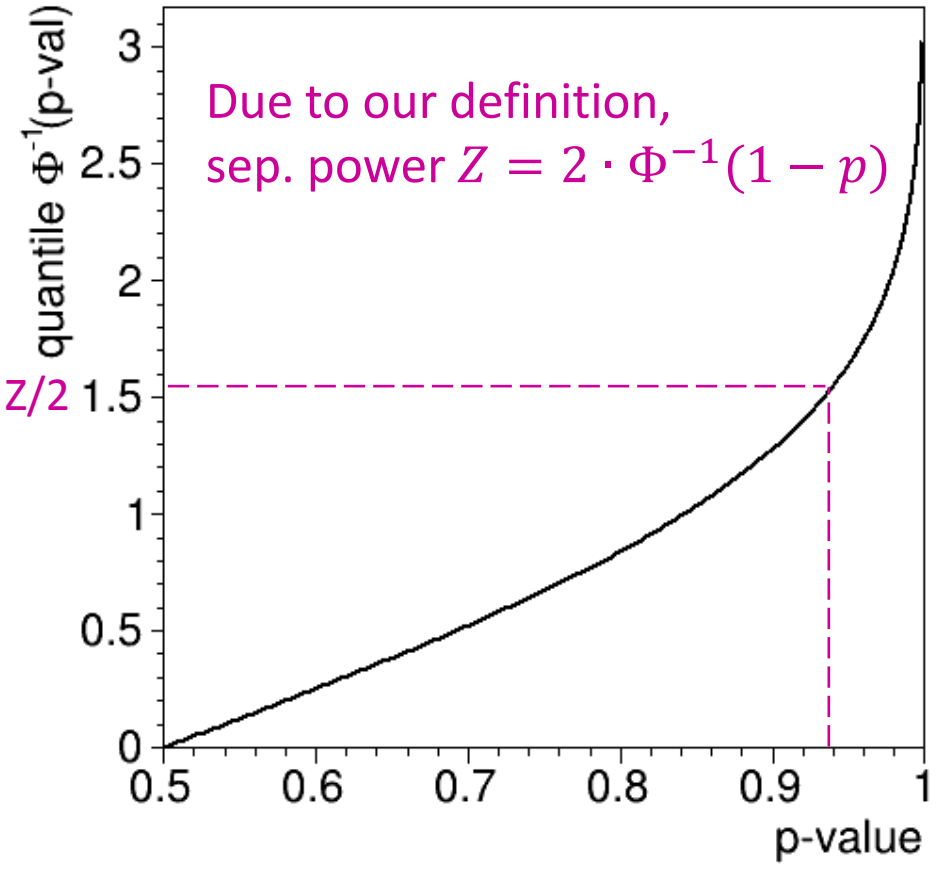




How treat non-Gaussian Likelihoods?

Gaussian quantile is inverse of distribution function $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2}\, dz$



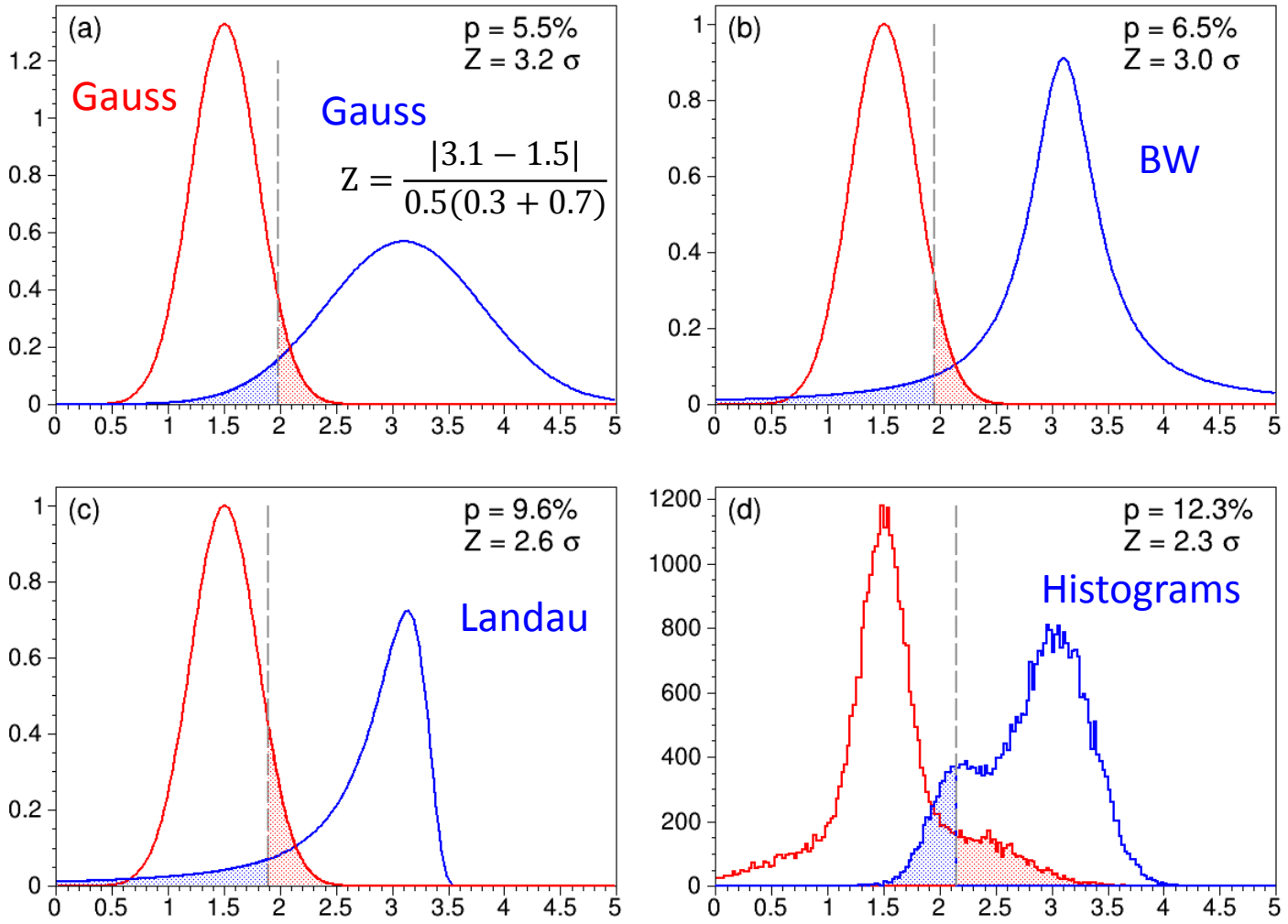


`ROOT::Math::gaussian_cdf(z)`

`ROOT::Math::gaussian_quantile_c(p,1)`

How treat non-Gaussian Likelihoods?

- Method can be applied to all kind of distributions

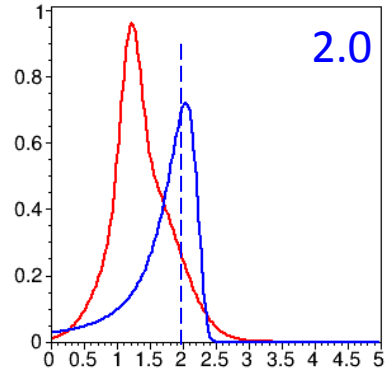


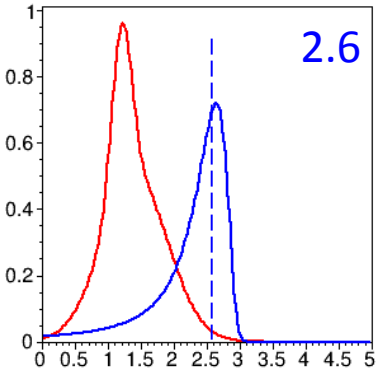
FoM Alternatives to Separation Power

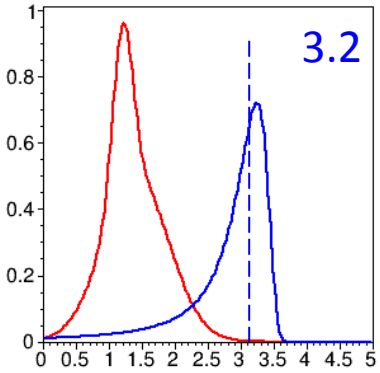
Two obvious alternatives as Figure-of-Merit

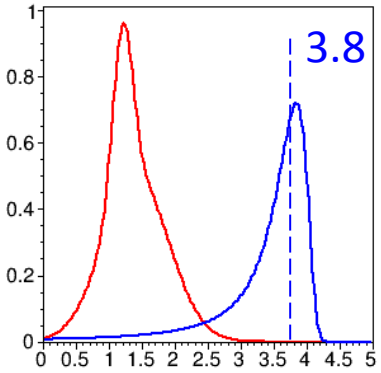
- **Mis-ID**
  - Specify mis-ID level for a fixed efficiency
- **Efficiency**
  - Specify efficiency for fixed mis-ID
- **Be aware**: While separation power is symmetric in signal and background LH, the upper two are not!

Comparison Figure-of-Merit

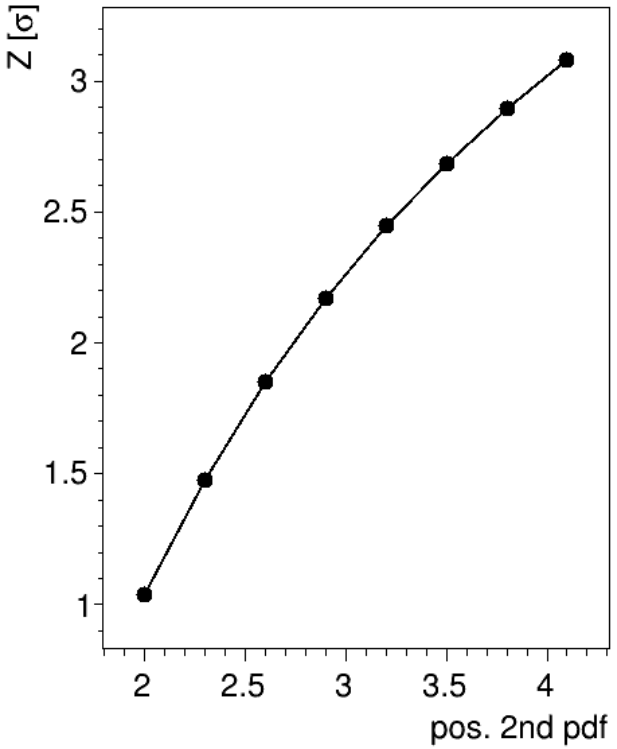


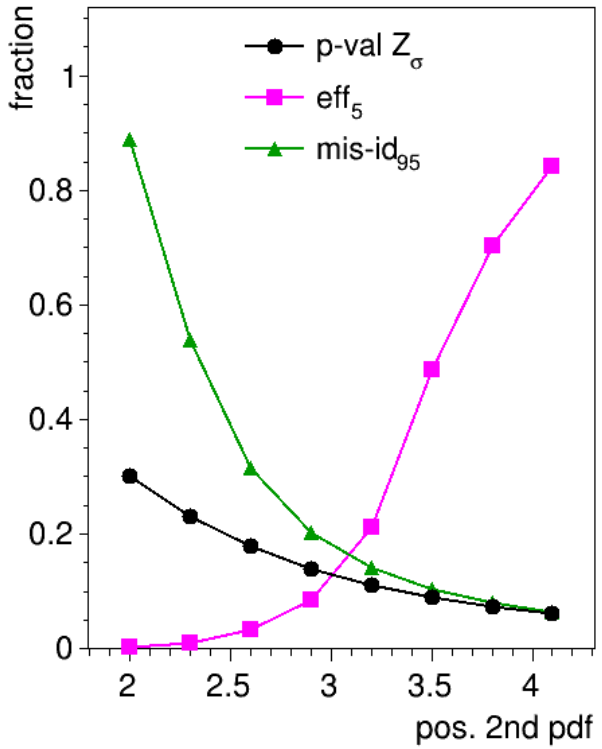


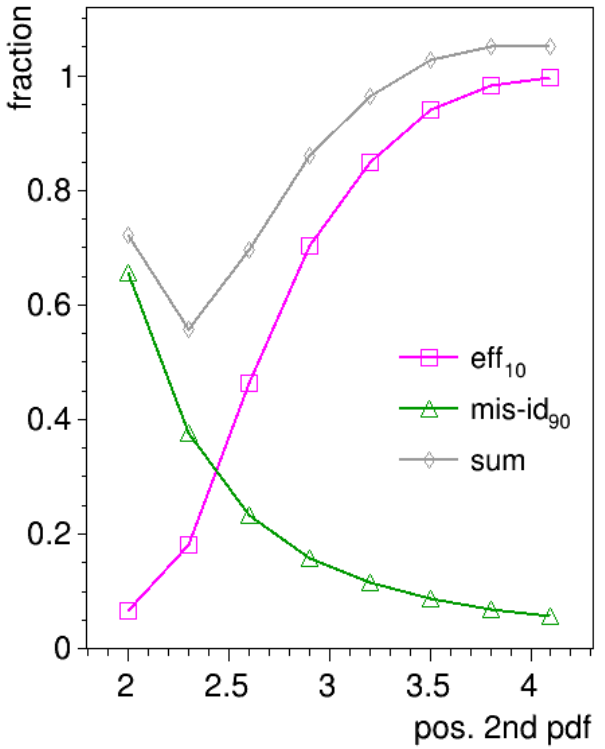




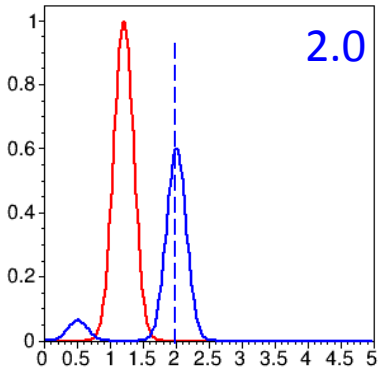
pos 2nd pdf

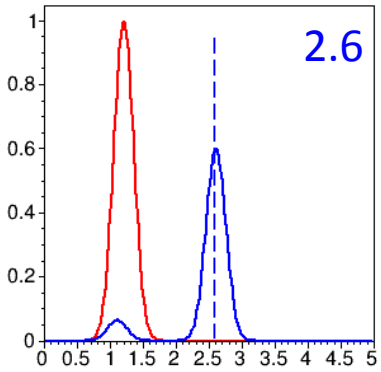


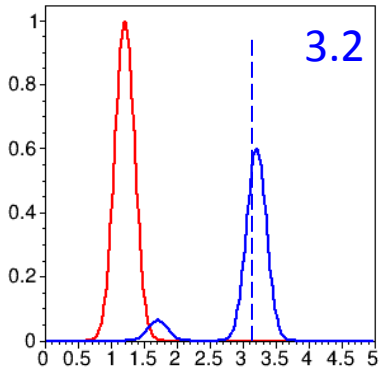


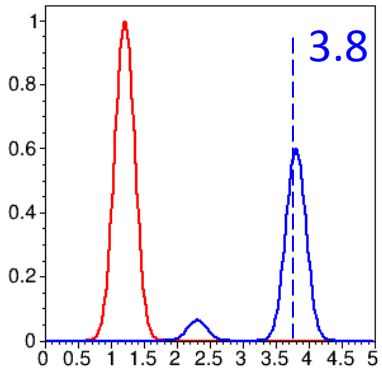


Comparison Figure-of-Merit

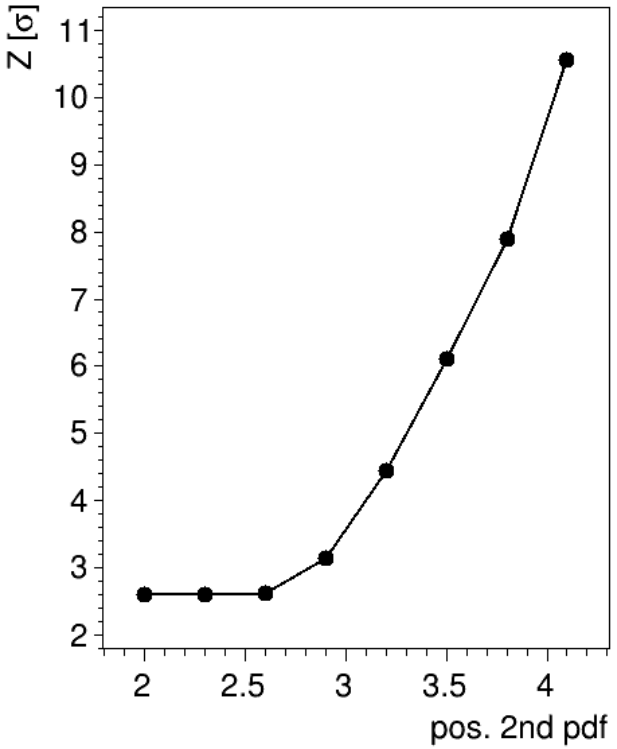


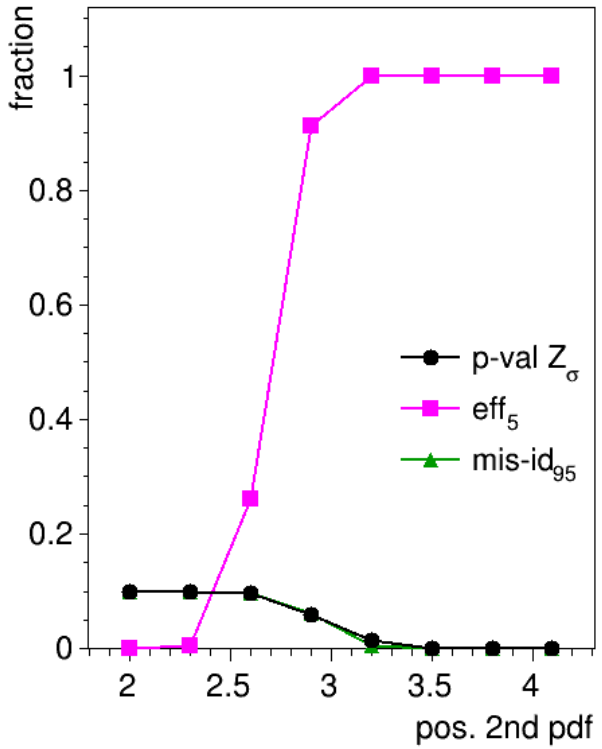


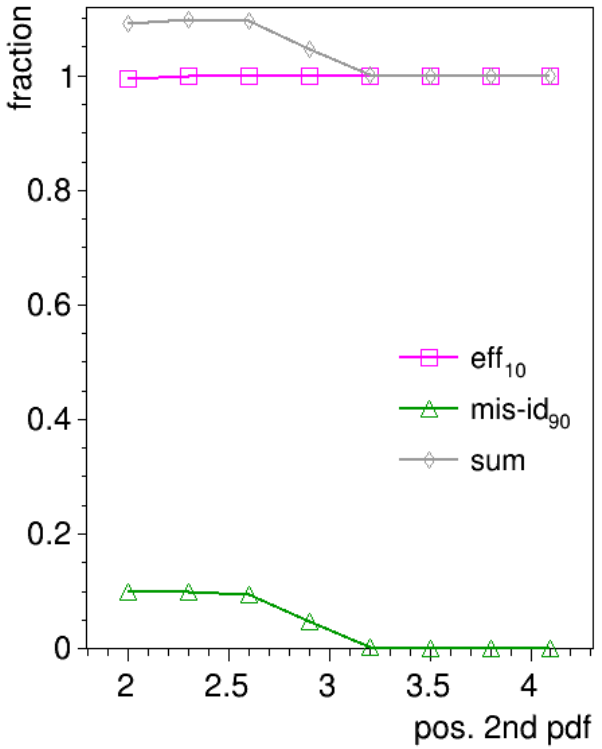




pos 2nd pdf

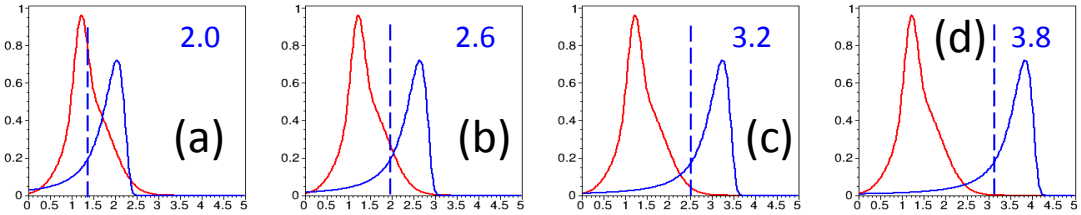


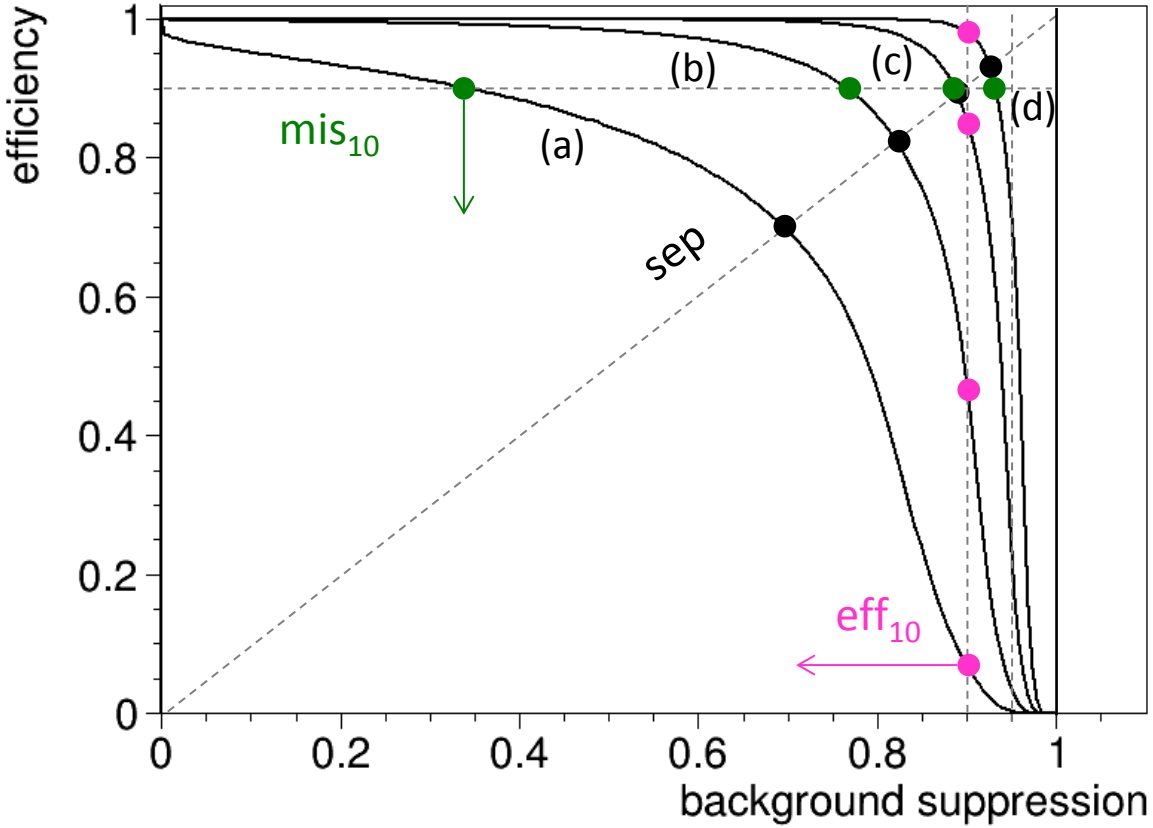




The ROC Curve: Efficiency and MisID at once

- ROC = Receiver Operating Charateristic
- Efficiency as function of background suppression (= 1 - misID)





Integral close to 1.0
⇒ Good separation

Integral of ROC curves
(a) 0.70
(b) 0.86
(c) 0.93
(d) 0.95

Conclusion

- Separation power = measure for potential classification quality
- Only properly defined for Gaussian PDFs
- If transformed to p-value
  - Also applicable for non-Gaussian PDFs

- Alternatives
  - Efficiency for fixed mis-ID level
  - Mis-ID for fixed efficiency level

- Complete description by ROC curve (characteristic of classification)

BACKUP