Fisher information metrics for binary classifier evaluation and training

Event selection for HEP precision measurements

Andrea Valassi
(CERN IT-DI-LCG)

Overview – scope of this talk

Different domains (or different ML problems in a domain) → different metrics

*This talk: <u>event selection</u> to minimize <u>statistical errors</u> in HEP <u>point estimation</u> <u>analyses</u>**

(not tracking – not systematic errors – not searches for new physics – not trigger)

(e.g. cross-section measurements by counting or by distribution fits; mass measurements by distribution fits)

Metrics based on Fisher information are appropriate for this specific HEP problem

- directly related to the ultimate goal, statistical errors on parameter estimates

They also meet some more general specificities of the HEP domain

- focus only on the signal and treat the background as a nuisance
- can be used in fits of differential distributions

*I discussed other domains and other HEP problems in an IML talk I gave in January (see backup slides)

Outline

- **Evaluation** (for generic binary classifiers)
  - – ROC AUCs vs. *Fisher information metrics*

- **Training** (for Decision Trees)
  - – Gini impurity and Shannon entropy vs. *Fisher information metrics*

The same Fisher information metrics can be used for both evaluation and training

Binary classifier evaluation – reminder

Discrete classifiers: the confusion matrix

	true class: Positives (HEP: **signal Stot**)	*true class*: Negatives (HEP: **background Btot**)
classified as Positives (HEP: **selected**)	**True Positives (TP)** (HEP: selected signal **Ssel**)	**False Positives (FP)** (HEP: selected bkg **Bsel**)
classified as Negatives (HEP: **rejected**)	**False Negatives (FN)** (HEP: rejected signal **Srej**)	**True Negatives (TN)** (HEP: rejected bkg **Brej**)

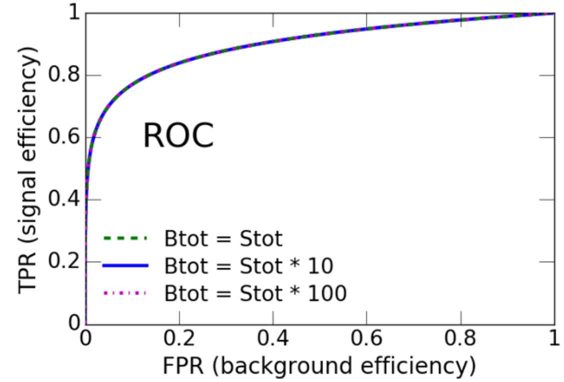
TP (S_{sel}), FP (B_{sel}), FN (S_{rej}), TN (B_{rej})	TP (S_{sel}), FP (B_{sel}), FN (S_{rej}), TN (B_{rej})	TP (S_{sel}), FP (B_{sel}), FN (S_{rej}), TN (B_{rej})
$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}$	$\mathbf{PPV} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}$	$\mathbf{TNR} = \frac{\mathbf{TN}}{\mathbf{TN}+\mathbf{FP}} = \mathbf{1-FPR}$
HEP: "efficiency" $\epsilon_s = \frac{S_{sel}}{S_{tot}}$	HEP: "purity" $\rho = \frac{S_{sel}}{S_{sel}+B_{sel}}$	HEP: "background rejection" $1-\epsilon_b = 1 - \frac{B_{sel}}{B_{tot}}$
IR: "recall"	IR: "precision"	—
MED: "sensitivity"	—	MED: "specificity"

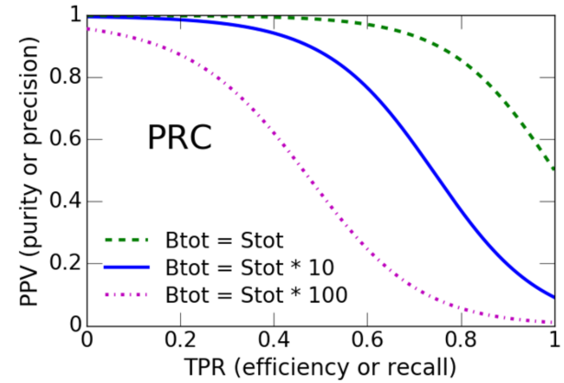
MED: prevalence
$$\pi_s = \frac{S_{tot}}{S_{tot}+B_{tot}}$$

Different domains
→ Focus on different concepts
→ Different terminologies

Examples from three domains:
- **Medical Diagnostics (MED)** *does Mr. A. have cancer?*
- **Information Retrieval (IR)** *Google documents about "ROC"*
- **HEP event selection (HEP)** *select Higgs event candidates*

Scoring classifiers: ROC and PRC curves





Purity can be studied using ROC only if prevalence is also known:
$$\rho = \frac{\epsilon_s S_{tot}}{\epsilon_s S_{tot} + \epsilon_b B_{tot}} = \frac{1}{1+\frac{\epsilon_b}{\epsilon_s}\frac{1-\pi_s}{\pi_s}}$$

Alternative: PN curve – TP vs FP (less used)

Binary classifier evaluation – HEP vs. other domains

- **Medical Diagnostics** $\rightarrow$ maximize diagnostic accuracy
  - qualitatively symmetric $\rightarrow$ all patients important, both truly ill (TP) and truly healthy (TN)
    - quantitatively: prevalence may be unknown, varying in time, from very balanced to extremely unbalanced
  - evaluation now based on ROC because insensitive to prevalence – now questioned for imbalanced data
    - simplest accuracy definition (ACC): "probability of correct test result" $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} = \pi_s \times \mathrm{TPR} + (1-\pi_s) \times \mathrm{TNR}$
    - *area under the ROC curve (ROC AUC):* "probability that test result of randomly chosen sick subject indicates greater suspicion than that of randomly chosen healthy subject" $\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$

- **Information Retrieval (IR)** $\rightarrow$ maximize effectiveness in retrieving relevant documents
  - qualitatively asymmetric $\rightarrow$ distinction between relevant and non-relevant documents
    - quantitatively: large class imbalance, irrelevant documents outnumber relevant documents
  - evaluation based on the PRC: precision and recall (purity and signal efficiency)
    - unranked: F-measures, e.g. F1-score
    - ranked: precision at k, Mean Average Precision, area under the PRC curve (AUCPR) $\mathrm{AUCPR} = \int_0^1 \rho \, d\epsilon_s$

- **HEP event selection** $\rightarrow$ minimize measurement errors
  - qualitatively asymmetric $\rightarrow$ only signal is important, background is a nuisance
    - quantitatively: large class imbalance, background outnumbers signal, prevalence fixed by physics cross-sections
  - *IMO evaluation metrics must include purity and prevalence (as in IR): TN and AUC are irrelevant*
  - *fits to differential distributions are largely a specificity of HEP – existing metrics do not describe them*

[FIP1] Simplest HEP example: cross-section by counting

- Counting experiment: measure a single number N_{meas}
- Well-known since decades: *maximize ε_s*ρ* to minimize statistical errors
  - *global* signal efficiency and *global* purity ("1 single bin")

$$(\sigma_s)_{\mathrm{meas}} = \frac{N_{\mathrm{meas}} - \mathcal{L}\epsilon_b\sigma_b}{\mathcal{L}\epsilon_s} \quad \rightarrow \quad \boxed{\frac{1}{(\Delta\sigma_s)^2} = \frac{1}{\sigma_s}\mathcal{L}\epsilon_s\rho = \frac{1}{\sigma_s^2}S_{\mathrm{tot}}\epsilon_s\rho}$$

- Relevant metric is ε_s*ρ *[NB: relevant only for σ_s by counting, should not be misused for other cases]*
  - metric in [0, 1] → 1 if keep all signal and no background
  - higher is better (qualitatively relevant)
  - directly related to $\Delta\hat{\sigma}$ (numerically relevant): ratio of $1/\Delta\hat{\sigma}^2$ to $1/\Delta\hat{\sigma}^2$ if background were 0
    - first example of Fisher Information Part metric: 'FIP1'
- Single "operating point" used (cut on scoring classifiers) – to compare classifiers:
  - find max ε_s*ρ for each classifier → chose classifier with highest max ε_s*ρ
  - from PRC: $\boxed{\mathrm{FIP1} = \max_{\epsilon_s} \epsilon_s\rho}$
  - from ROC (plus prevalence): $\boxed{\mathrm{FIP1} = \max_{\epsilon_s} \frac{\epsilon_s}{1 + \frac{\epsilon_b}{\epsilon_s}\frac{1-\pi_s}{\pi_s}}}$

More generally – Fisher Information Part metrics

- Fit θ from a binned multi-dimensional distribution
  - expected counts $y_i = f(x_i,\theta)dx = \varepsilon_i * S_i(\theta) + b_i$ → depend on parameter θ to fit

- Statistical error related to Fisher information (Cramer-Rao lower bound)

$$(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta} \quad \text{where} \quad \mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i}\left(\frac{\partial y_i}{\partial\theta}\right)^2 = \int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx$$

  - binned fit → combine independent measurements in each bin, weighted by information

- Compare classifier to "ideal classifier" that keeps all signal and rejects all background

$$\mathcal{I}_\theta^{(\text{ideal classifier})} = \sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2 \quad \text{vs.} \quad \mathcal{I}_\theta^{(\text{real classifier})} = \sum_{i=1}^{m} \epsilon_i\rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2$$

  - *ε_i and ρ_i → local signal efficiency and local purity in the i^{th} bin*

- **Fisher Information Part**: available information retained by the classifier
  - FIP in [0,1] → 1 if keep all signal and no background
  - higher is better → maximize FIP
  - directly related to $\Delta\hat{\theta}$: $(\Delta\hat{\theta}^{(\text{real classifier})})^2 = \frac{1}{\mathrm{FIP}}(\Delta\hat{\theta}^{(\text{ideal classifier})})^2$

$$\mathrm{FIP} = \frac{\mathcal{I}_\theta^{(\text{real classifier})}}{\mathcal{I}_\theta^{(\text{ideal classifier})}} = \frac{\sum_{i=1}^{m} \epsilon_i\rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2}{\sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2}$$

- Special case: cross-section measurements $\theta=\sigma_s$ → $\frac{1}{S_i}\frac{\partial S_i}{\partial\sigma_s} = \frac{1}{\sigma_s}$
  - global ε*ρ is the FIP ('FIP1') for measuring $\theta=\sigma_s$ in a 1-bin fit (counting experiment)

Optimal partitioning in binned fits – information inflow

- Information about θ in a binned fit $\mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i} \left(\frac{\partial y_i}{\partial \theta} \right)^2$

- Can I reduce the error $\Delta\hat{\theta}$ by splitting bin y_i into two separate bins? $y_i = w_i + z_i$
  - i.e. is the "information inflow" positive? $\frac{1}{w_i}\left(\frac{\partial w_i}{\partial \theta}\right)^2 + \frac{1}{z_i}\left(\frac{\partial z_i}{\partial \theta}\right)^2 - \frac{1}{w_i+z_i}\left(\frac{\partial (w_i+z_i)}{\partial \theta}\right)^2 = \frac{\left(w_i \frac{\partial z_i}{\partial \theta} - z_i \frac{\partial w_i}{\partial \theta}\right)^2}{w_i z_i (w_i+z_i)} \geq 0$
  - information increases (errors on parameters decrease) if $\frac{1}{w_i}\frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i}\frac{\partial z_i}{\partial \theta}$

- Effect of background: $y_i=\varepsilon_i S_i(\theta)+b_i \rightarrow \frac{1}{y_i}\frac{\partial y_i}{\partial \theta} = \rho_i \frac{1}{S_i}\frac{\partial S_i}{\partial \theta}$
  - information increases if $\rho_w \frac{1}{s_w}\frac{\partial s_w}{\partial \theta} \neq \rho_z \frac{1}{s_z}\frac{\partial s_z}{\partial \theta}$
  - therefore: **try to partition the data into bins of equal $\rho_i \frac{1}{s_i}\frac{\partial si}{\partial \theta}$**
    - for cross-section measurements, $\frac{1}{S_i}\frac{\partial S_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$: split into bins of equal ρ_i

- Two important practical consequences:
  - *1. use the scoring classifier to partition the data, not to reject events*
  - *2. information can be used also for training classifiers like decision trees*

Three examples – FIP1, FIP2, FIP3

- **[FIP1]** cross-section measurements by counting
  - Global event selection/cut $\rightarrow$ discrete classifier (one single "operating point" of the scoring classifier)
  - Counting experiment with one single bin $\rightarrow$ global efficiency and purity are relevant
  - Cross-section: $\frac{1}{S_i}\frac{\partial S_i}{\partial \theta} = \frac{1}{\sigma_S}$ $\rightarrow$ signal events all have the same weight, only event counts matter
  - *In this talk: described in the previous slides*

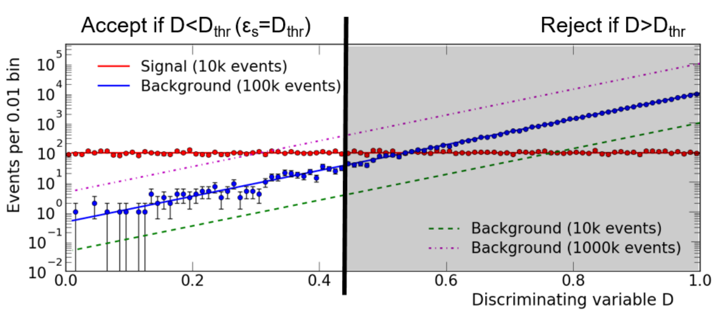
- **[FIP2]** cross-section measurements by fits to 1-D scoring classifier distributions
  - Keep all (preselected) events $\rightarrow$ scoring classifier partitions events into bins (use all "operating points")
  - Distribution fit $\rightarrow$ local purity in each bin is relevant (local efficiency = 1, keep all events)
  - Cross-section: $\frac{1}{S_i}\frac{\partial S_i}{\partial \theta} = \frac{1}{\sigma_S}$ $\rightarrow$ signal events all have the same weight, only event counts matter
  - ***In this talk: main focus of the following slides***

- **[FIP3]** other parameter measurements by fits to distributions
  - Keep all (preselected) events $\rightarrow$ scoring classifier partitions events into bins (use all "operating points")
  - Distribution fit $\rightarrow$ local purity in each bin is relevant (local efficiency = 1, keep all events)
  - Example: mass fit $\frac{1}{S_i}\frac{\partial S_i}{\partial M}$ varies bin by bin $\rightarrow$ signal events have different event-by-event weights
  - *In this talk: just a few comments at the end (work in progress)*

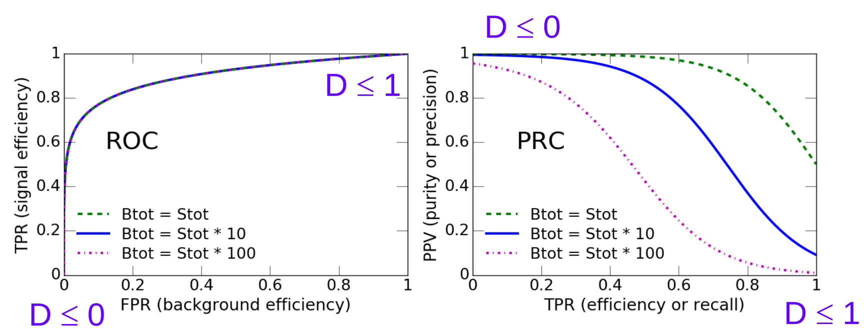
[FIP2] cross-section measurement by fitting the 1-D scoring classifier distribution

- Information and FIP in fit for σ_s of a (generic) binned distribution:
  - If all events are kept and partitioned into bins (local efficiency in each bin = 1): $y_i = n_i = s_i + b_i$
  - Cross-section measurement: $\frac{1}{s_i}\frac{\partial s_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$
  - Information: $\mathcal{I}_{\sigma_s} = \sum_{i=1}^{m} \frac{1}{y_i}\left(\frac{\partial y_i}{\partial \sigma_s}\right)^2 = \frac{1}{\sigma_s^2}\sum_{i=1}^{m}\frac{s_i^2}{n_i} = \frac{1}{\sigma_s^2}\sum_{i=1}^{m}\rho_i s_i = \frac{1}{\sigma_s^2}\sum_{i=1}^{m} n_i \rho_i^2$
  - Ratio to no-background case: $\mathrm{FIP2} = \frac{\sum_{i=1}^{m} s_i^2/n_i}{\sum_{i=1}^{m} s_i} = \frac{\sum_{i=1}^{m} \rho_i s_i}{\sum_{i=1}^{m} s_i}$

- These formulas are valid for σ_s fits irrespective of the variable used for binning
  - <u>If events are binned according to the scoring classifier D (FIP2): use the ROC and/or PRC!</u>
  - *By definition, ROCs (PRCs) describe how $\varepsilon_s/\varepsilon_b$ (ρ/ε_s) are related when varying the cut on D*
    - See details in the next slide



simple example: D distribution flat for signal



FIP2 from the ROC (+prevalence) or from the PRC

FIP2: integrals on ROC and PRC,
more relevant to HEP than AUC or AUCPR!
(well-defined meaning for distribution fits)

- From the previous slide: $\mathrm{FIP2} = \frac{\sum_{i=1}^{m} \rho_i s_i}{\sum_{i=1}^{m} s_i}$
- FIP2 from the ROC (+prevalence $\pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}} + B_{\mathrm{tot}}}$):

$$S_{\mathrm{sel}} = S_{\mathrm{tot}}\,\epsilon_s \qquad B_{\mathrm{sel}} = B_{\mathrm{tot}}\,\epsilon_b \;\Longrightarrow\; s_i = dS_{\mathrm{sel}} = S_{\mathrm{tot}}\,d\epsilon_s \qquad b_i = dB_{\mathrm{sel}} = B_{\mathrm{tot}}\,d\epsilon_b \;\Longrightarrow\; \rho_i = \frac{1}{1 + \frac{B_{\mathrm{tot}}}{S_{\mathrm{tot}}}\frac{d\epsilon_b}{d\epsilon_s}} \;\Longrightarrow\; \boxed{\mathrm{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s}\frac{d\epsilon_b}{d\epsilon_s}}}$$

Compare FIP2(ROC) to AUC

$$\mathrm{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$$

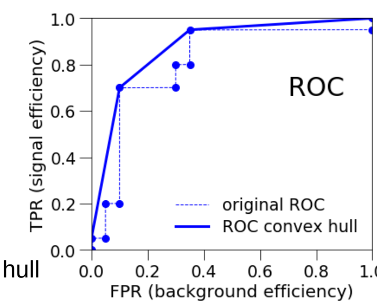
- FIP2 from the PRC:

$$S_{\mathrm{sel}} = S_{\mathrm{tot}}\,\epsilon_s \qquad B_{\mathrm{sel}} = S_{\mathrm{sel}}\left(\frac{1}{\rho} - 1\right) \;\Longrightarrow\; s_i = dS_{\mathrm{sel}} = S_{\mathrm{tot}}\,d\epsilon_s \qquad b_i = dB_{\mathrm{sel}} = S_{\mathrm{tot}}\left[d\epsilon_s\left(\frac{1}{\rho} - 1\right) - \epsilon_s\frac{d\rho}{\rho^2}\right] \;\Longrightarrow\; \rho_i = \frac{\rho}{1 - \frac{\epsilon_s}{\rho}\frac{d\rho}{d\epsilon_s}} \;\Longrightarrow\; \boxed{\mathrm{FIP2} = \int_0^1 \frac{\rho\, d\epsilon_s}{1 - \frac{\epsilon_s}{\rho}\frac{d\rho}{d\epsilon_s}}}$$

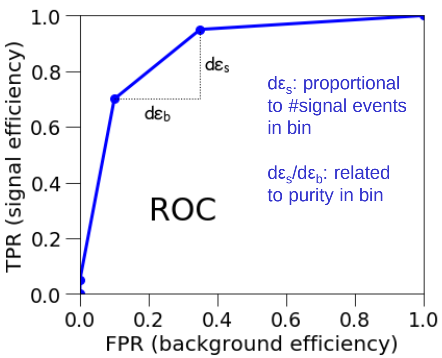
Compare FIP2(PRC) to AUCPR

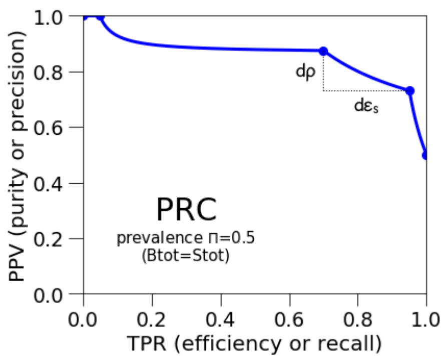
$$\mathrm{AUCPR} = \int_0^1 \rho\, d\epsilon_s$$

- Easier calculation and interpretation from ROC (+prevalence) than from PRC
  - *region of constant ROC slope** = region of constant signal purity*
  - decreasing ROC slope = decreasing purity
    - technicality (my Python code): convert ROC to convex hull** first



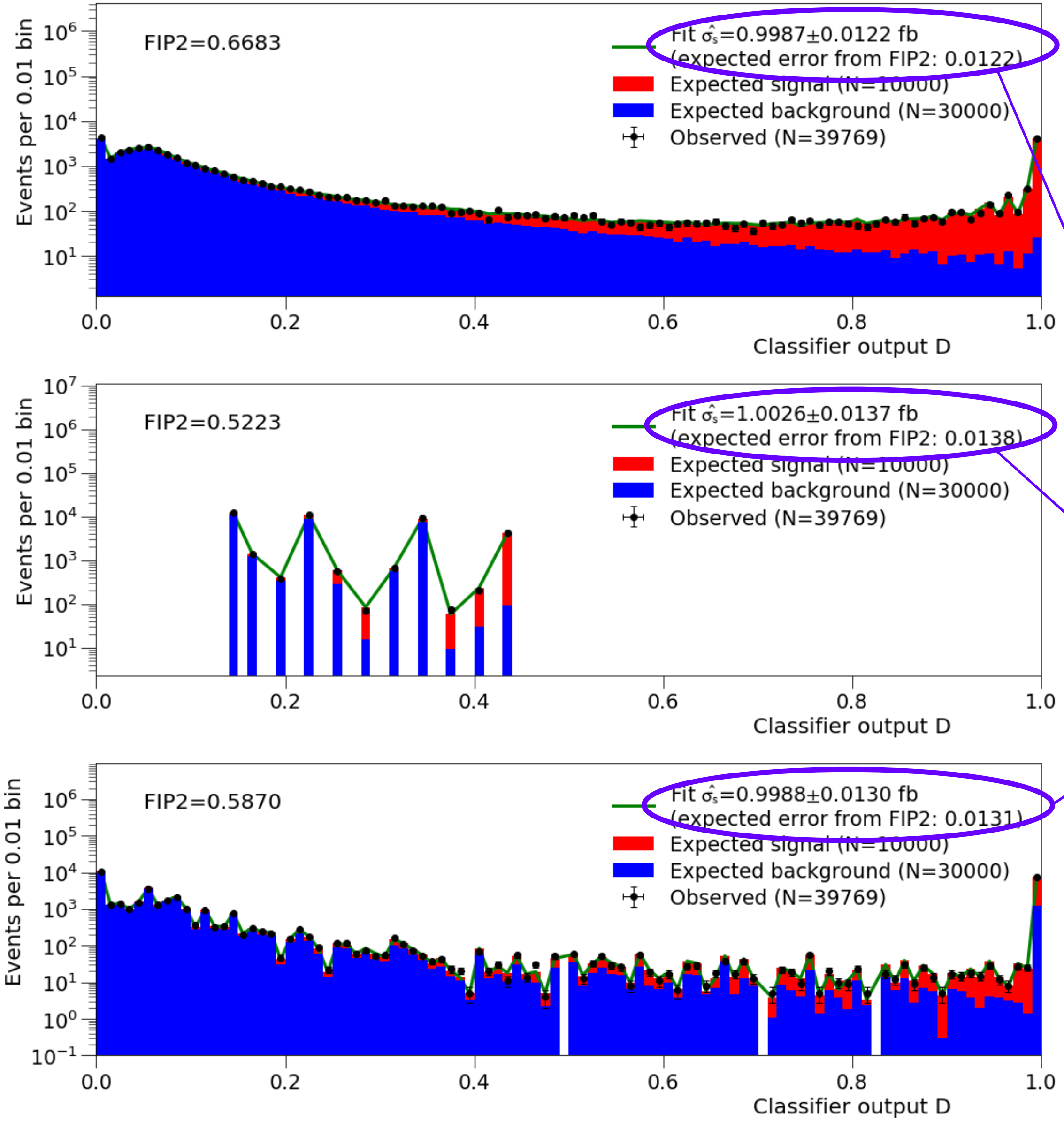
**Convert ROC to convex hull
- ensure decreasing slope
- avoid staircase effect that would artificially inflate FIP2
(bins of 100% purity: only signal or only background)





*ROC slopes are also discussed in medical literature in relation to diagnostic likelihood ratios [Choi 1998], but their use does not seem to be widespread(?)

Sanity check

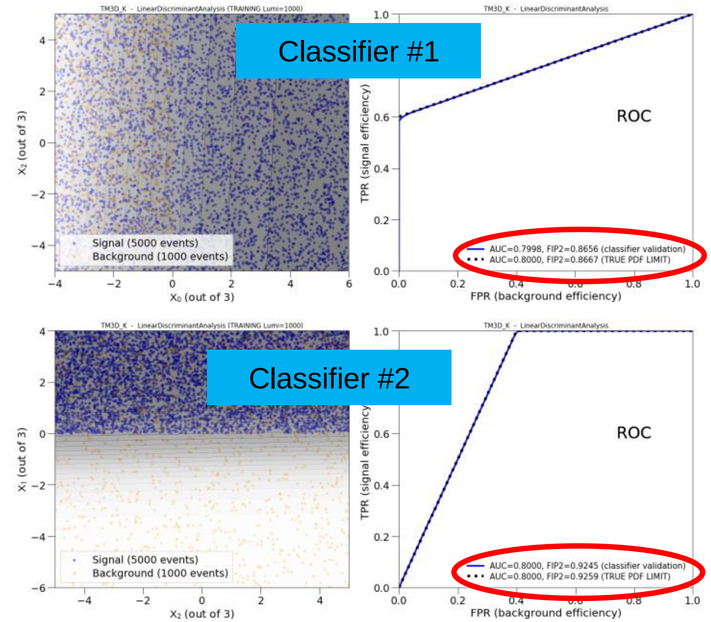


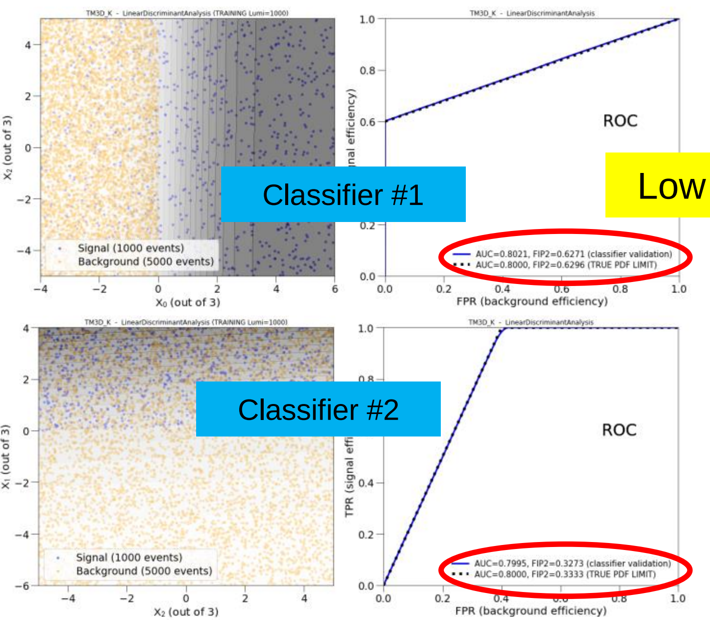
- Three random forests (on the same 2-D pdf)
  - reasonable
  - undertrained
  - overtrained
- Fit σ_s from the distribution of the classifier output
  - Errors consistent with FIP2

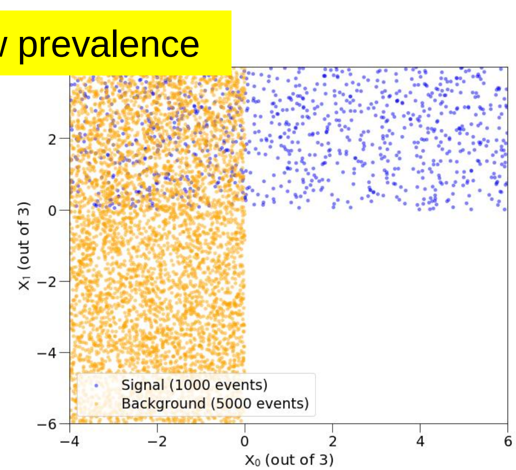
$$(\Delta\hat{\theta}^{(\text{real classifier})})^2 = \frac{1}{\text{FIP}}(\Delta\hat{\theta}^{(\text{ideal classifier})})^2$$

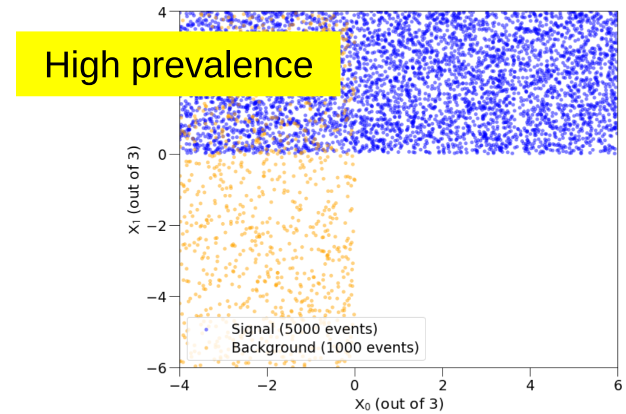
My development environment: SciPy ecosystem, iminuit and bits of rootpy, on SWAN at CERN. Thanks to all involved in these projects!

- Prepared a model just to show that AUC is misleading
  - pdf with two useful features and a third random one
  - two classifiers, each trained only one useful feature
  - two prevalence scenarios: S/B=5 and S/B=1/5
- Same AUC (0.80) in all four cases
  - it is well known that AUC is insensitive to prevalence
  - *ROC curves of the two classifiers cross*
- Low prevalence: FIP2 favors classifier #1 (0.63 > 0.33)
- High prevalence: FIP2 favors classifier #2 (0.87 < 0.93)
- Do not choose the best classifier based on AUC
  - not for a cross-section fit on the classifier output, nor in general!









FIP2 vs AUC

FIP2 for training decision trees

- Decision tree $\rightarrow$ recursively partition the training set into nodes of different purities

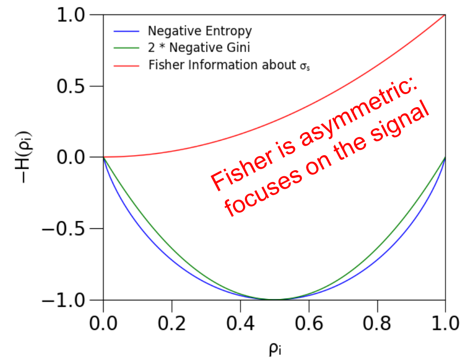
- Given a node (n,s) with n total events and s signal events:
  - (if I do decide to split it) **how do I best split node (n,s) into two nodes (n_L,s_L) + (n_R,s_R)?**
  - choose the Left/Right splitting that maximizes the gain in a appropriate figure of merit

- Two criteria are most often used (e.g. in sklearn):
  - Gini impurity – "Gini diversity index" in CART algorithm (Breiman et al. 1984)
    - derived from a metric for *economic inequality*, adapted for *ecological diversity* (Simpson-Gini index)
  - Shannon information (Shannon entropy) – a concept from information theory
  - Maximize loss of impurity or entropy at each split

- Fisher information metrics (e.g. FIP2) can also be used for training decision trees
  - Maximize the total information (about signal event cross-section) in the whole system
  - *Advantage: use the same metric for evaluation and training*
  - *Advantage: train the classifier to minimize measurement errors on physics parameters*
  - *Advantage: total sum over all bins is a well defined meaningful concept*

- Note a conceptual difference setting HEP apart (again): qualitative class asymmetry
  - *Gini and Shannon impurity/diversity/entropy indices consider all classes as equal*
  - *Fisher information (about a property of the signal) focuses only on the signal*

Training decision trees: FIP2 vs Gini vs entropy



- Information or negative impurity in one node (higher is better):
  - negative Gini impurity $\rightarrow$ $-n_i H(\rho_i) = n_i \times [-2\rho_i(1-\rho_i)]$
  - negative Shannon entropy $\rightarrow$ $-n_i H(\rho_i) = n_i \times [\rho_i \log_2 \rho_i + (1-\rho_i)\log_2(1-\rho_i)]$
  - Fisher information about σ_s $\rightarrow$ $-n_i H(\rho_i) = n_i \times [\rho_i^2]$
- The best split $(n,s)=(n_L,s_L)+(n_R,s_R)$ maximizes information gain (impurity loss):
  - information gain (higher is better) $\rightarrow$ $\Delta = -n_L H(\rho_L) - n_R H(\rho_R) + nH(\rho)$
- The shapes of the impurity functions look very different, but...
  - *...information gain is the same for Gini and Fisher!* (modulo a constant factor)

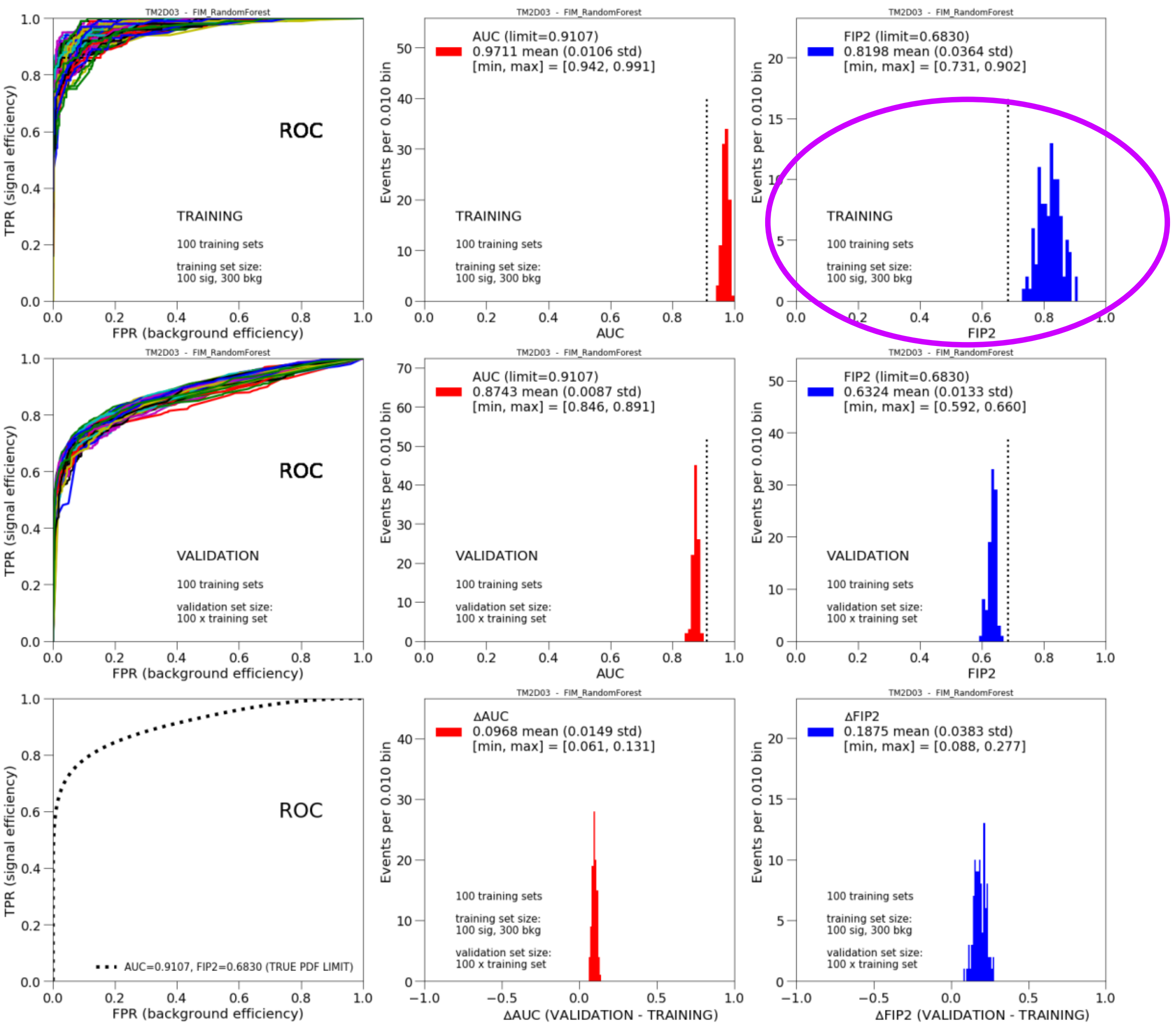
$$\Delta_{\text{Fisher}} = \frac{s_L^2}{n_L} + \frac{s_R^2}{n_R} - \frac{(s_L+s_R)^2}{n_L+n_R} = \frac{(s_L n_R - s_R n_L)^2}{n_L n_R (n_L+n_R)} \qquad \frac{\Delta_{\text{Gini}}}{2} = -s_L\left(1-\frac{s_L}{n_L}\right) - s_R\left(1-\frac{s_R}{n_R}\right) + (s_L+s_R)\left(1-\frac{s_L+s_R}{n_L+n_R}\right) = \Delta_{\text{Fisher}}$$

  - the interpretation is clearer for Fisher: extra reduction in measurement error on σ_s
    - unless this is overtraining (briefly discussed in the next slide)

Technicality: user-defined criteria for DecisionTree's will only be available in the next sklearn release

$\rightarrow$ *I implemented a DecisionTree from scratch, heavily reusing the excellent iCSC notebooks by Thomas Keck (many thanks!)*

FIP2: same metric for evaluation and training



OVERTRAINING example – random forests with min_samples_leaf=1

- Using the same metric for evaluation and training eases the interpretation of results
- Example: overtraining
  - FIP2 from training is systematically above the theoretical limit of the pdf
  - *you may trace back every increase in FIP2 from training to one node split in the tree*
    - splitting a node (n,s) gives in average an information gain:

$$\Delta_{\mathrm{expected}}(n,s) = \frac{s(n-s)}{n(n-1)}$$

  - *Note: what really matters is that FIP2 from validation is as close as possible to the limit*
    - *some overtraining (a value of FIP2 from training higher than the limit) is necessary*

[FIP3] other parameter fits – just a few ideas

- The general ideas for σ_s fits apply to fits for other parameters θ, e.g. mass fits

- The difference is that different events have *different event-by-event sensitivities to θ*
  - for instance, should compute $\frac{1}{w_\alpha}\frac{\partial w_\alpha}{\partial\theta} = \frac{1}{|\mathcal{M}|^2_\alpha}\frac{\partial|\mathcal{M}|^2_\alpha}{\partial\theta}$ from the MC generator for each event α
    - this can be positive or negative (e.g. left and right of a mass peak)
  - remember, ***partition the data into bins of equal ρ_i $(\frac{1}{s_i}\frac{\partial s_i}{\partial\theta})$***
    - for unweighted MC events $s_i = \sum_{\alpha\in\mathrm{bin\,i}} w_\alpha = \sum_{\alpha\in\mathrm{bin\,i}} 1$ and this is equal to $\rho_i\frac{1}{s_i}\frac{\partial s_i}{\partial\theta} = \frac{s_i}{n_i}\frac{1}{s_i}\frac{\partial s_i}{\partial\theta} = \frac{1}{n_i}\sum_{\alpha\in\mathrm{bin\,i}}\frac{\partial w_\alpha}{\partial\theta}$

- For instance, perform a 2-D fit for θ on the distributions of $(\frac{1}{s_i}\frac{\partial s_i}{\partial\theta})$ and ρ_i
  - train a regression tree for $(\frac{1}{s_i}\frac{\partial si}{\partial\theta})$ to partition signal events in bins of $(\frac{1}{s_i}\frac{\partial s_i}{\partial\theta})$
  - train a classification tree for ρ_i to partition signal and background events in bins of ρ_i $(\frac{1}{s_i}\frac{\partial s_i}{\partial\theta})$
    - taking into account the event-by-event $\frac{\partial w_\alpha}{\partial\theta}$ when computing the separation gain at each node split

- In summary: the distinction between classification and regression blurs even further
  - not simply “select signal, reject background”
  - keep all events, in different partitions according to *signal purity <u>and</u> sensitivity to θ*

Conclusions

- ***Different domains (or different ML problems in a domain) need different metrics***

- I discussed some general properties of HEP event selection – two, in particular:
  - signal is relevant, background is a nuisance: use asymmetric metrics, TN and AUC are irrelevant
  - we use distribution fits: need (the right) integrals over all operating points of scoring classifiers, e.g. FIP2

- I discussed Fisher information metrics relevant to statistical errors in HEP point estimation
  - qualitatively (higher is better) and numerically (related to parameter errors) relevant – unlike AUC
  - can be used both for evaluation and training

- Distribution fits are a specialty of HEP – decision trees are their natural ML companions
  - we could probably gain by developing and using the right metrics for evaluating and training them

- *More generally, it would be useful IMO to do more research on ML fundamentals for HEP*
  - define the ultimate quantitative goals first, then choose metrics for evaluation, and possibly training too
  - which relevant ML metrics should be used for searches, for systematic errors, for event reconstruction...

I am preparing a paper on this, thank you for your feedback on this presentation!

Backup slides

Backup – statistical error in binned fits

- Data: *observed event counts n_i* in m bins of a (multi-D) distribution f(x)
  - *expected event counts y_i* = $f(x_i,\theta)dx$ depend on a parameter θ that we want to fit
  - [NB here f is a differential cross section, it is not normalized to 1 like a pdf]

- Fitting θ is like combining the independent measurements in the m bins
  - expected error on n_i in bin x_i is $\Delta n_i = \sqrt{y_i} = \sqrt{f(xi,\theta)\,dx}$
  - expected error on $f(x_i,\theta)$ in bin x_i is $\Delta f = f * \Delta n_i / n_i = \sqrt{f / dx}$
  - expected error on estimated $\hat{\theta}_i$ in bin x_i is $\frac{1}{(\Delta\hat{\theta})^2_{(\text{bin } dx)}} = \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{1}{(\Delta f)^2} = \left(\frac{\partial f}{\partial\theta}\right)^2 \left(\frac{\sqrt{dx}}{\sqrt{f}}\right)^2 = \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{dx}{f}$
  - expected error on estimated $\hat{\theta}$ by combining the m bins is $\left(\frac{1}{\Delta\hat{\theta}}\right)^2 = \int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx$

- A bit more formally, joint probability for observing the n_i is $P(\mathbf{n};\theta) = \prod_{i=1}^{m} \frac{e^{-y_i} y_i^{n_i}}{n_i!}$
  - Fisher information on θ from the data available is then

$$\mathcal{I}_\theta = E\left[\frac{\partial \log P(\mathbf{n};\theta)}{\partial\theta}\right]^2 \quad \text{i.e.} \quad \mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i}\left(\frac{\partial y_i}{\partial\theta}\right)^2 = \int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx$$

  - The minimum variance achievable (Cramer-Rao lower bound) is $(\Delta\hat{\theta})^2 = \text{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$

Slides from the January IML talk

https://indico.cern.ch/event/679765/contributions/2814562

ROC curves, AUC's and alternatives in HEP event selection and in other domains

Andrea Valassi (IT-DI-LCG)

Inter-Experimental LHC Machine Learning WG – 26th January 2018

Disclaimer: I last did physics analyses more than 15 years ago

(mainly statistically-limited precision measurements and combinations – e.g. no searches)

Why and when I got interested in this topic

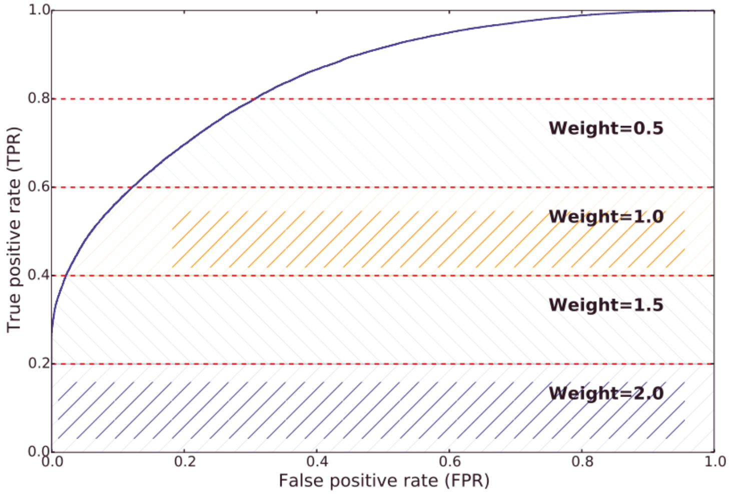


Figure 3: Weights assigned to the different segments of the ROC curve for the purpose of submission evaluation. The x axis is the False Positive Rate (FPR), while the y axis is True Positive Rate (TPR).

T. Blake at al., *Flavours of Physics: the machine learning challenge for the search of $\tau \to \mu\mu\mu$ decays at LHCb* (2015, unpublished). https://kaggle2.blob.core.windows.net/competitions/kaggle/4488/media/lhcb_description_official.pdf (accessed 15 January 2018)

The 2015 LHCb Kaggle ML Challenge
- Event selection in search for $\tau \rightarrow \mu\mu\mu$
- *Classifier wins if it maximises a weighted ROC AUC*
- Simplified for Kaggle – real analysis uses CLs

- First time I saw an *Area Under the Roc Curve (AUC)*
- My reaction: what is this? is this relevant in HEP?
  – try to understand why the AUC was introduced in other scientific domains
  – review *common knowledge* for optimizing several types of HEP analyses

Questions for you – How extensively are AUC's used in HEP, particularly in event selection? Are there specific HEP problems where it can be shown that AUC's are relevant?

Spoiler! – What I will argue in this talk

- Different disciplines / problems → different challenges → different metrics
  - Tools from other domains → assess their relevance before using them in HEP
- Most relevant metrics in HEP event selection: purity ρ and signal efficiency ε_s
  - "Precision and Recall" – HEP closer to Information Retrieval than to Medicine
  - "True Negatives", ROCs and AUCs irrelevant in HEP event selection*
    - AUCs → Higher not always better. Numerically, no relevant interpretation.
- HEP specificity: fits of differential distributions → binning / partitioning of data
  - local efficiency and purity in each bin → more relevant than global averages of ρ,ε_s
  - scoring classifiers → more useful for partitioning data than for imposing cuts
    - optimize statistical errors on parameter estimates → metrics based on local $\rho_i{}^*\varepsilon_{s,i}$
    - optimal partitioning: split into bins of uniform purity ρ_i and sensitivity $\frac{1}{S_i}\frac{\partial Si}{\partial\theta}$

* ROCs are relevant in particle-ID – but this is largely beyond the scope of this talk

Outline

- Introduction to binary classifiers: the confusion matrix, ROCs, AUCs, PRCs

- Binary classifier evaluation: domain-specific challenges and solutions
  - Overview of Diagnostic Medicine and Information Retrieval
  - A systematic analysis and summary of optimizations in HEP event selection

- Statistical error optimization in HEP parameter estimation problems
  - Information metrics and the effect of local efficiency and purity in binned fits
  - Optimal binning and the relevance of local purity

- Conclusions

Binary classifiers: the "confusion matrix"

- Data sample containing instances of two classes: Ntot = Stot + Btot
  - HEP: signal Stot = Ssel + Srej
  - HEP: background Btot = Bsel + Brej
- Discrete binary classifiers assign each instance to one of the two classes
  - HEP: classified as signal and selected Nsel = Ssel + Bsel
  - HEP: classified as background and rejected Nrej = Brej + Srej

	true class: **P**ositives **+** (HEP: **signal**)	***true class***: **N**egatives **-** (HEP: **background**)
classified as: positives (HEP: **selected**)	**True Positives (TP)** (HEP: selected signal **Ssel**)	**False Positives (FP)** (HEP: selected bkg **Bsel**)
classified as: negatives (HEP: **rejected**)	**False Negatives (FN)** (HEP: rejected signal **Srej**)	**True Negatives (TN)** (HEP: rejected bkg **Brej**)

T. Fawcett, *Introduction to ROC analysis*, Pattern Recognition Letters 27 (2006) 861. doi:10.1016/j.patrec.2005.10.010

I will not discuss multi-class classifiers (useful in HEP particle-ID)

The confusion matrix about the confusion matrix...

Different domains → focus on different concepts → different terminologies

TP (S_{sel}) FP (B_{sel}) **FN** (S_{rej}) TN (B_{rej})	**TP** (S_{sel}) **FP** (B_{sel}) FN (S_{rej}) TN (B_{rej})	TP (S_{sel}) **FP** (B_{sel}) FN (S_{rej}) **TN** (B_{rej})
$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}$	$\mathbf{PPV} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}$	$\mathbf{TNR} = \frac{\mathbf{TN}}{\mathbf{TN}+\mathbf{FP}} = \mathbf{1} - \mathbf{FPR}$
HEP: "efficiency" $\epsilon_s = \frac{S_{\mathrm{sel}}}{S_{\mathrm{tot}}}$	HEP: "purity" $\rho = \frac{S_{\mathrm{sel}}}{S_{\mathrm{sel}}+B_{\mathrm{sel}}}$	HEP: "background rejection" $1-\epsilon_b = 1 - \frac{B_{\mathrm{sel}}}{B_{\mathrm{tot}}}$
IR: "recall"	IR: "precision"	—
MED: "sensitivity"	—	MED: "specificity"

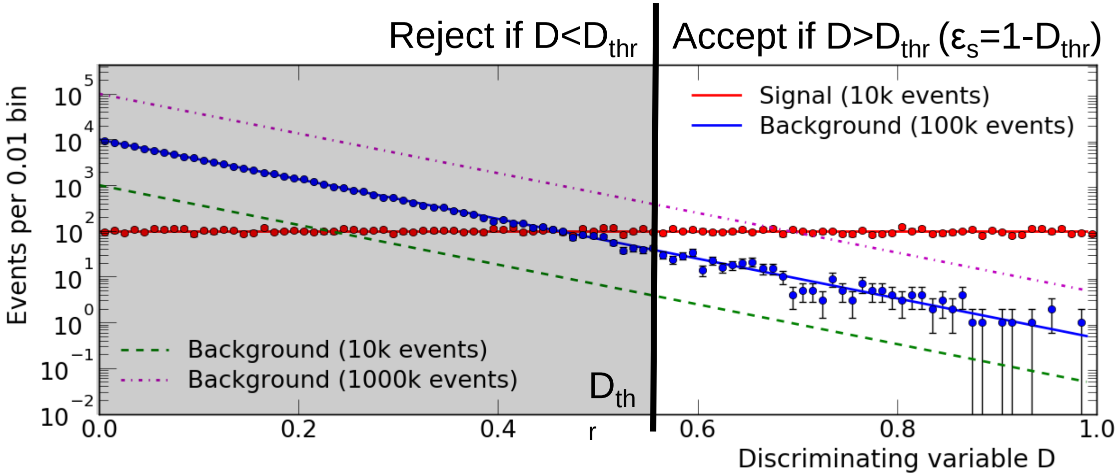
I will cover three domains:

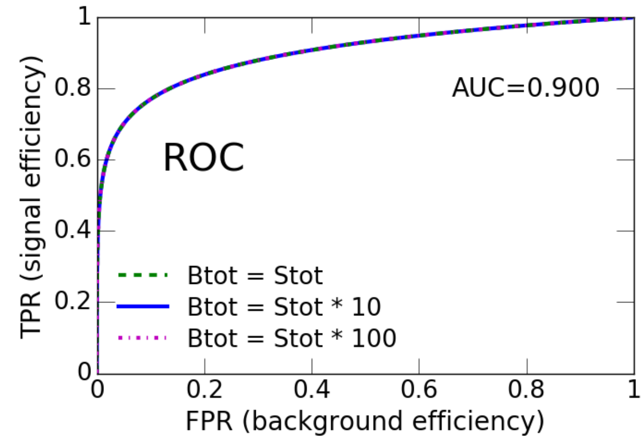
- **Medical Diagnostics (MED)**
  does Mr. A. have cancer?
- **Information Retrieval (IR)**
  Google documents about "ROC"
- **HEP event selection (HEP)**
  select Higgs event candidates

MED: prevalence

$$\pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}}+B_{\mathrm{tot}}}$$

Discrete vs. Scoring classifiers – ROC curves



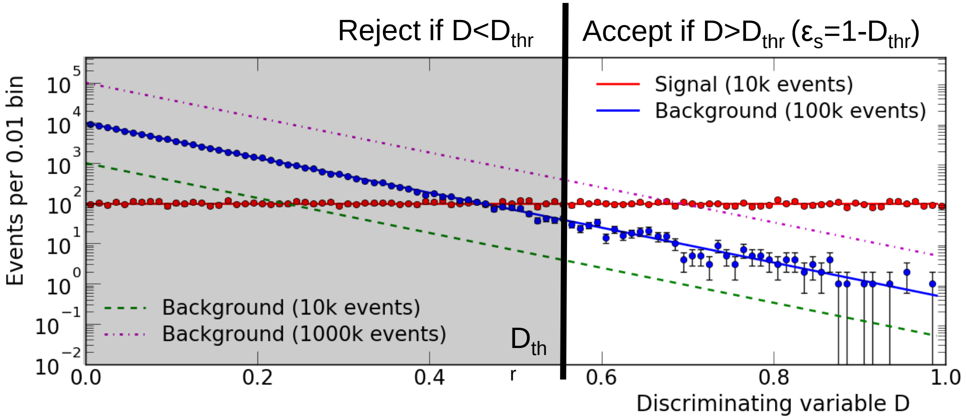


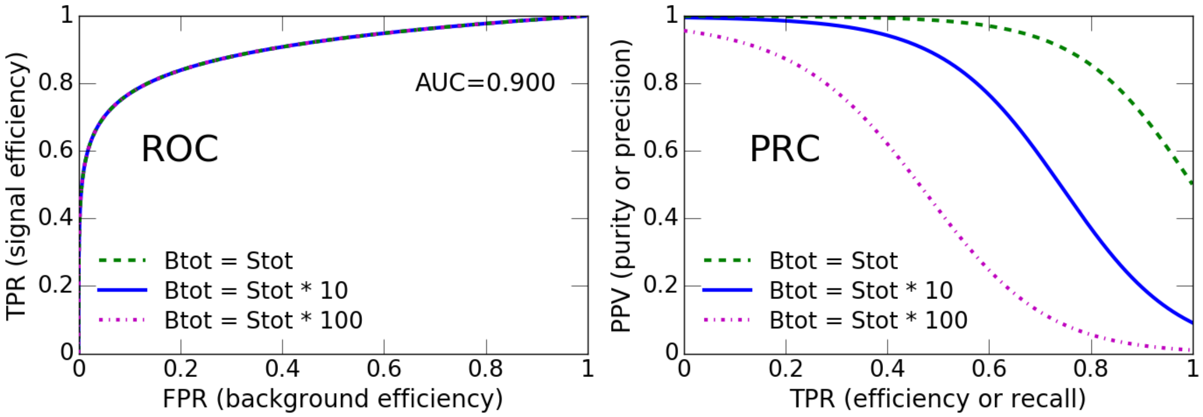
- Discrete classifiers → either select or reject → confusion matrix
- Scoring classifiers → assign score D to each event (e.g. BDT)
  - ideally related to likelihood that event is signal or background (Neyman-Pearson)
  - from scoring to discrete: choose a threshold → classify as signal if D>Dthr
- ROC curves describe how FPR(ε_b) and TPR(ε_s) are related when varying Dthr
  - used initially in radar signal detection and psychophysics (1940-50's)

W. W. Peterson, T. G. Birdsall, W. C. Fox, *The theory of signal detectability*, Transactions of the IRE Professional Group on Information Theory 4 (1954) 171. doi:10.1109/TIT.1954.1057460
W. P. Tanner, J. A. Swets, *A decision-making theory of visual detection*, Psychological Review 61 (1954), 401. doi:10.1037/h0058700
J. A. Swets, *Is There a Sensory Threshold?*, Science 134 (1961) 168. doi:10.1126/science.134.3473.168
J. A. Swets, W. P. Tanner, T. G. Birdsall, *Decision processes in perception*, Psychological Review 68 (1961) 301. doi:10.1037/h0040547

ROC and PRC (precision-recall) curves

- Different choice of ratios in the confusion matrix: $\varepsilon_s, \varepsilon_b$ (ROC) or ρ, ε_s (PRC)
- When Btot/Stot ("prevalence") varies → PRC changes, ROC does not





Understanding domain-specific challenges

- Many domain-specific details → but also general cross-domain questions:
  - **1. Qualitative imbalance?**
    - Are the two classes equally relevant?
  - **2. Quantitative imbalance?**
    - Is the prevalence of one class much higher?
  - **3. Prevalence known? Time invariance?**
    - Is relative prevalence known in advance? Does it vary over time?
  - **4. Dimensionality? Scale invariance?**
    - Are all 4 elements of the confusion matrix needed?
    - Is the problem invariant under changes of some of these elements?
  - **5. Ranking? Binning?**
    - Are all selected instances equally useful? Are they partitioned into subgroups?

M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002

- Point out properties of MED and IR, attempt a systematic analysis of HEP

Medical diagnostics (1)

and ML research

H. Sox, S. Stern, D. Owens, H. L. Abrams, *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions*, The National Academies Press (1989). doi:10.17226/1432

X. H. Zhou, D. K. McClish, N. A. Obuchowski, *Statistical Methods in Diagnostic Medicine* (Wiley, 2002). doi:10.1002/9780470317082

- **Medical Diagnostics (MED)**
does Mr. A. have cancer?

- Binary classifier optimisation goal: maximise "diagnostic accuracy"
  - patient / physician / society have different goals → many possible definitions

- Most popular metric: "accuracy", or "probability of correct test result":

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} = \pi_s \times \mathrm{TPR} + (1-\pi_s) \times \mathrm{TNR}$$

TP (correctly diagnosed as ill)	**FP** (truly healthy, but diagnosed as ill)
FN (truly ill, but diagnosed as healthy)	**TN** (correctly diagnosed as healthy)

  - Symmetric → all patients important, both truly ill (TP) and truly healthy (TN)
  - Also "by far the most commonly used metric" in ML research in the 1990s

F. J. Provost, T. Fawcett, *Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions*, Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, USA (1997). https://aaai.org/Library/KDD/1997/kdd97-007.php

L. B. Lusted, *Signal Detectability and Medical Decision-Making*, Science 171 (1971) 1217 doi:10.1126/science.171.3977.1217

J. A. Swets, *Measuring the accuracy of diagnostic systems*, Science 240 (1988) 1285. doi:10.1126/science.3287615

- Since the '90s → shift from ACC to ROC in the MED and ML fields
  - TPR (sensitivity) and TNR (specificity) studied separately
    - solves ACC limitations (imbalanced or unknown prevalence – rare diseases, epidemics)
  - Evaluation often AUC-based → two perceived advantages *for MED and ML fields*
    - *AUC interpretation: "probability that test result of randomly chosen sick subject indicates greater suspicion than that of randomly chosen healthy subject"*
    - ROC comparison without prior D_{thr} choice (prevalence-dependent D_{thr} choice)

F. J. Provost, T. Fawcett, R. Kohavi, *The Case against Accuracy Estimation for Comparing Induction Algorithms*, Proc. 15th Int. Conf. on Machine Learning (ICML '98), Madison, USA (1998). https://www.researchgate.net/publication/2373067

A. P. Bradley, *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition 30 (1997) 1145. doi:10.1016/S0031-3203(96)00142-2

J. A. Hanley, B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology 143 (1982) 29. doi:10.1148/radiology.143.1.7063747

Medical diagnostics (2)
and ML research

- ROC and AUC metrics → currently widely used in the MED and ML fields
  - Remember: moved because *ROC better than ACC with imbalanced data sets*
- Limitation: evidence that *ROC not so good for highly imbalanced data sets*
  - may provide an overly optimistic view of performance
  - PRC may provide a more informative assessment of performance in this case
    - PRC-based reanalysis of some data sets in life sciences has been performed
- Very active area of research → other options proposed (CROC, cost models)
  - Take-away message: *ROC and AUC not always the appropriate solutions*

J. Davis, M. Goadrich, *The relationship between Precision-Recall and ROC curves*, Proc. 23rd Int. Conf. on Machine Learning (ICML '06), Pittsburgh, USA (2006). doi:10.1145/1143844.1143874

C. Drummond, R. C. Holte, *Explicitly representing expected cost: an alternative to ROC representation*, Proc. 6th Int. Conf. on Knowledge Discovery and Data Mining (KDD-00), Boston, USA (2000). doi:10.1145/347090.347126

D. J. Hand, *Measuring classifier performance: a coherent alternative to the area under the ROC curve*, Mach Learn (2009) 77: 103. doi:10.1007/s10994-009-5119-5

S. J. Swamidass, C.-A. Azencott, K. Daily, P. Baldi, *A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval*, Bioinformatics 26 (2010) 1348. doi:10.1093/bioinformatics/btq140

D. Berrar, P. Flach, *Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them)*, Briefings in Bioinformatics 13 (2012) 83. doi:10.1093/bib/bbr008

H. He, E. A. Garcia, *Learning from Imbalanced Data*, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263. doi:10.1109/TKDE.2008.239

T. Saito, M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*, PLoS One 10 (2015) e0118432. doi:10.1371/journal.pone.0118432

Information Retrieval

- **Information Retrieval (IR)**
  Google documents about "ROC"

- Qualitative distinction between "relevant" and "non-relevant" documents
  - also a very large quantitative imbalance

- Binary classifier optimisation goal: make users happy in web searches
  - minimise # relevant documents not retrieved → maximise "recall" i.e. efficiency
  - minimise # of irrelevant documents retrieved → maximise "precision" i.e. purity
  - retrieve the more relevant documents first → ranking very important
  - maximise speed of retrieval

- IR-specific metrics to evaluate classifiers based on the PRC (i.e. on ε_s, ρ)
  - unranked evaluation → e.g. F-measures $F_\alpha = \frac{1}{\alpha/\varepsilon_s + (1-\alpha)/\rho}$
    - $\alpha \in [0,1]$ *tradeoff between recall and precision* → equal weight gives $F1 = \frac{2\varepsilon_s\rho}{\varepsilon_s + \rho}$
  - ranked evaluation → precision at k documents, mean average precision (MAP), ...
    - MAP approximated by the Area Under the PRC curve (AUCPR)

C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, 2008).
https://nlp.stanford.edu/IR-book

NB: Many different of meanings of "Information"!
IR (web documents), HEP (Fisher), Information Theory (Shannon)...

First (simplest) HEP example

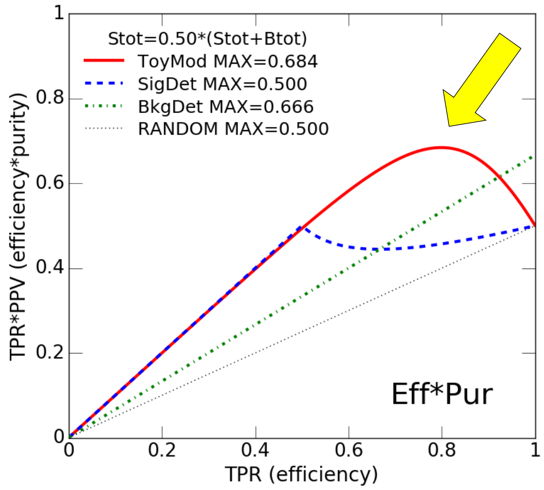
- HEP event selection (HEP)
select Higgs event candidates

- *Measurement of a total cross-section σ_s in a counting experiment*

- To minimize statistical errors: *maximise ε_s*ρ* (well-known since decades)
  - *global* efficiency $\varepsilon_s=S_{sel}/S_{tot}$ and *global* purity $\rho=S_{sel}/(S_{sel}+B_{sel})$ – “1 single bin”

$$\frac{1}{(\Delta\sigma_s)^2} = \frac{1}{\sigma_s}\mathcal{L}\epsilon_s\rho = \frac{1}{\sigma_s^2}S_{\mathrm{tot}}\epsilon_s\rho$$

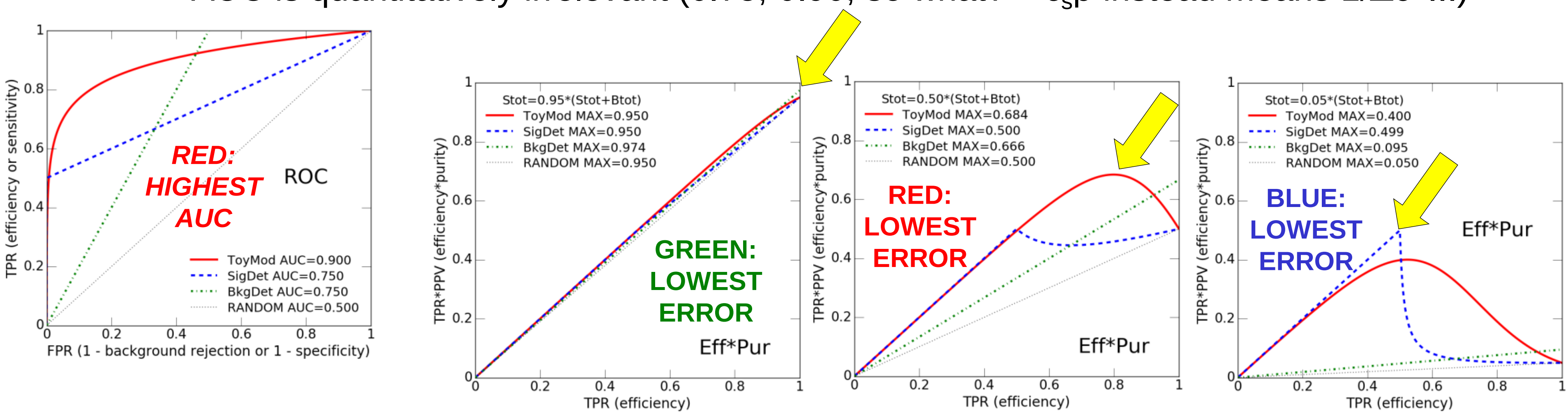
- To compare classifiers (red, green, blue, black):
  - in each classifier → vary Dthr cut → vary ε_s and ρ
    → find maximum of ε_s*ρ (choose “operating point”)
  - chose classifier with maximum of ε_s*ρ out of the four



- ε_s*ρ: metric between 0 and 1
  - qualitatively relevant: the higher, the better
  - numerically: fraction of Fisher information (1/error^2) available after selecting
  - *correct metric only for σ_s by counting!* → table with more cases on a next slide

Examples of issues with AUCs – crossing ROCs

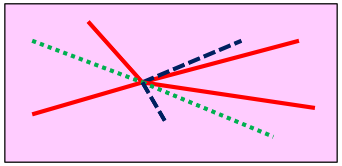
- Choice of classifier easy if one ROC “dominates” another (higher TPR ∀FPR)
  - PRC “dominates” too, then – and of course AUC is higher, too
- Choice is less obvious if ROCs cross!
- Example: cross-section by counting
  - maximise product $\varepsilon_s\rho \rightarrow$ i.e. minimise the statistical error $\Delta\sigma^2$
  - depending on S_{tot}/B_{tot}, a different classifier (green, red, blue) should be chosen
  - in two out of three scenarios, the classifier with the highest AUC is not the best
    - AUC is qualitatively irrelevant (higher is not always better)
    - AUC is quantitatively irrelevant (0.75, 0.90, so what? – $\varepsilon_s\rho$ instead means $1/\Delta\sigma^2$...)



Binary classifiers in HEP

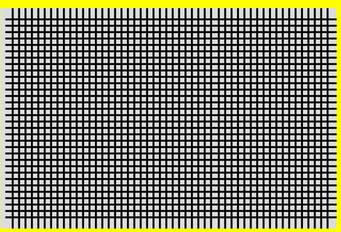
- HEP event selection (HEP)
select Higgs event candidates

Binary classifier optimisation goal: maximise physics reach at a given budget



Tracking and particle-ID (event reconstruction) – e.g. fake track rejection
→ maximise identification of particles *(all particles within each event are important)*

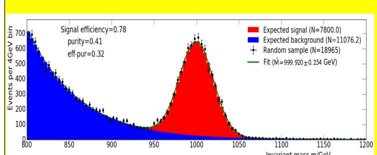
Instances: tracks within one event, created by earlier reconstruction stage.
→ P = real tracks, N = fake tracks (ghosts) → goal: keep real tracks, reject ghosts
→ TN = fake tracks identified as such and rejected: ***TN are relevant*** (IIUC...)
[Optimisation: should translate tracking metrics into measurement errors in physics analyses]



Trigger → maximise signal event throughput, within the computing budget – e.g. HLT

Instances: events, from the earlier trigger stage (e.g. L0 hardware trigger)
→ P = signal events, N = background events [per unit time: trigger rates]
→ goal: ***maximise retained signal efficiency*** TP/(TP+FN) at a given trigger rate FP (as TP ≪ FP)
→ TN = background events identified as such and rejected: ***TN are irrelevant***
→ constraint: max HLT rate (from HLT throughput), whatever the input L0 rate is: ***TN are ill-defined***

EVENT SELECTION – I WILL FOCUS ON THIS IN THIS TALK



Physics analyses → maximise the physics reach, given the available data sets

Instances: events, from pre-selected data sets
→ P = signal events, N = background events
→ goal: ***minimise measurement errors*** or maximise significance in searches
→ TN = background events identified as such and rejected: ***TN are irrelevant***
→ physics results independent of pre-selection or MC cuts: ***TN are ill-defined***

TP = S_{sel}	FP = B_{sel}
FN = S_{rej}	~~TN = B_{rej}~~

Domain / Property	Medical diagnostics	Information retrieval	HEP event selection
Qualitative class imbalance	**NO.** Healthy and ill people have "equal rights". *TN are relevant.*	**YES.** "Non-relevant" documents are a nuisance. *TN are irrelevant.*	**YES.** Background events are a nuisance. *TN are irrelevant.*
Quantitative class imbalance	**From small to extreme.** From common flu to very rare disease.	**Generally very high.** Only very few documents in a repository are relevant.	**Generally extreme.** Signal events are swamped in background events.
Varying or unknown prevalence π	**Varying and unknown.** Epidemics may spread.	**Varying and unknown** in general (e.g. WWW).	**Constant in time** (quantum cross-sections). **Unknown** for searches. **Known** for precision measurements.
Dimensionality and invariances M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002	**3 ratios ε_s, ε_b, π + scale.** New metrics under study because ROC ignores π. Costs scale with N_{tot}.	**2 ratios ε_s, ρ + scale.** ε_s, ρ enough in many cases. Costs and speed scale with N_{tot}. Show only N_{sel} docs in one page. *TN are irrelevant.*	**2 ratios ε_s, ρ + scale.** ε_s, ρ enough in many cases. Lumi is needed for: trigger, syst. vs stat., searches. *TN are irrelevant.*
Different use of selected instances	**Binning – NO.** **Ranking – YES?** Treat with higher priority patients who are more likely to be ill?	**Binning – NO.** **Ranking – YES.** Precision at k, R-precision, MAP all involve *global* precision-recall ("top N_{sel} documents retrieved)	**Binning – YES.** Fits to distributions: *local ε_s, ρ in each bin* rather than global ε_s, ρ.

Different HEP problems → Different metrics

Binary classifiers for HEP event selection (signal-background discrimination)				
Statistical error minimization (or statistical significance maximization)	**Cross-section (1-bin counting)**	Only 2 or 3 global/local variables – *TN, AUC irrelevant*	2 variables: global ε_s, ρ (given S_{tot})	Maximise $S_{tot}*\varepsilon_s*\rho$ (at any S_{tot})
	Searches (1-bin counting)		Simple and CCGV – 2 variables: global S_{sel}, B_{sel} (or equivalently ε_s, ρ)	Maximise $\frac{S_{sel}}{\sqrt{S_{sel}+Bsel}}$ (i.e. $\sqrt{S_{tot}*\varepsilon_s*\rho}$)
				Maximise $\sqrt{2((S_{sel}+Bsel)\log\left(1+\frac{S_{sel}}{B_{sel}}\right)-Ssel)}$
			HiggsML – 2 variables: global S_{sel}, B_{sel}	Maximise $\sqrt{2((S_{sel}+Bsel+K)\log\left(1+\frac{S_{sel}}{B_{sel}+K}\right)-Ssel)}$
			Punzi – 2 variables: global ε_s, B_{sel}	Maximise $\frac{\varepsilon_s}{A/2+\sqrt{B_{sel}}}$
	Cross-section (binned fits)		2 variables: local $\varepsilon_{s,i}$ and ρ_i in each bin (given $s_{tot,i}$ in each bin)	Maximise $\sum_i s_{tot,i}*\varepsilon_{s,i}*\rho_i$ Partition in bins of equal ρ_i
	Parameter estimation (binned fits)			Maximise $\sum_i s_{tot,i}*\varepsilon_{s,i}*\rho_i*(\frac{1}{s_{tot,i}}\frac{\partial s_{tot,i}}{\partial\theta})^2$ Partition in bins of equal $\rho_i*(\frac{1}{s_{tot,i}}\frac{\partial s_{tot,i}}{\partial\theta})$
	Searches (binned fits)		3 variables: local s_{sel}, s_{tot}, s_{sel} in each bin (2 counts or ratios enough?)	Maximise a sum? *
Statistical + Systematic error minimization			3 variables: ε_s, ρ, lumi (lumi: tradeoff stat. vs. syst.)	No universal recipe * (may use local S_{sel}, B_{sel} in side band bins)
Trigger optimization			2 variables: global B_{sel}/time, global ε_s	Maximise ε_s at given trigger rate
Binary classifiers for HEP problems other than event selection				
Tracking and Particle-ID optimizations			All 4 variables? * *(NB: TN is relevant)*	ROC relevant – is AUC relevant? *
Other? *			? *	? *

* Many open questions for further research

Predict and optimize statistical errors in binned fits

- Fit θ from a binned multi-dimensional distribution
  - expected counts $y_i = f(x_i,\theta)dx = \varepsilon_i{*}s_i(\theta)+b_i \rightarrow$ depend on parameter θ to fit
- Statistical error related to Fisher information $(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$ (Cramer-Rao)
  - binned fit → combine measurements in each bin, weighed by information
- Easy to show (backup slides) that Fisher information in the fit is:

$$\mathcal{I}_\theta^{(\mathrm{real\ classifier})} = \sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2 \qquad \mathcal{I}_\theta^{(\mathrm{ideal\ classifier})} = \sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2$$

  - *ε_i and ρ_i → local signal efficiency and purity in the i^{th} bin*
- Define a binary classifier metric as information fraction to ideal classifier:
  - in [0,1] → 1 if keep all signal and reject all backgrounds
  - higher is better → maximise IF
  - interpretation: $(\Delta\hat{\theta}^{(\mathrm{real\ classifier})})^2 \geq \frac{1}{\mathrm{IF}}(\Delta\hat{\theta}^{(\mathrm{ideal\ classifier})})^2$

$$\mathrm{IF} = \frac{\mathcal{I}_\theta^{(\mathrm{real\ classifier})}}{\mathcal{I}_\theta^{(\mathrm{ideal\ classifier})}} = \frac{\sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}{\sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial \theta}\right)^2}$$

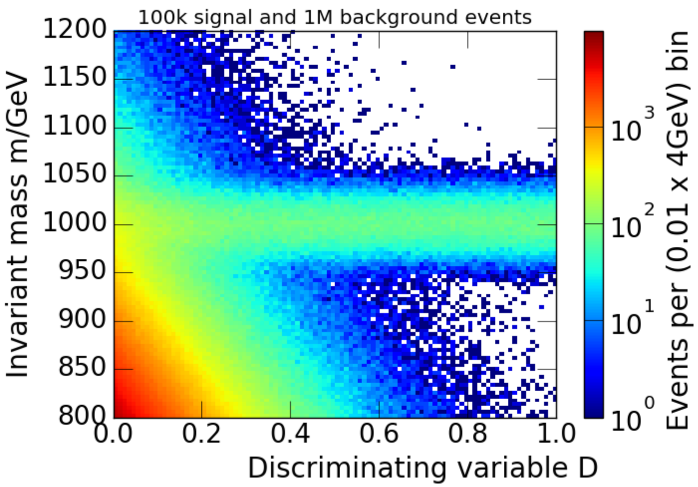
*NB: global ε*ρ <u>is</u> the IF for measuring $\theta=\sigma_s$ in a 1-bin fit (counting experiment)!*

Numerical tests with a toy model

- I used a simple toy model to make some numerical tests
  - Verify that my formulas are correct – and also illustrate them graphically
  - Two-dimensional distribution (m,D) $\rightarrow$ signal Gaussian, background exponential

- Two measurements:
  - total cross-section measurement by counting and 1-D or 2-D fit
  - mass measurement by 1-D or 2-D fits

- Details in the backup slides

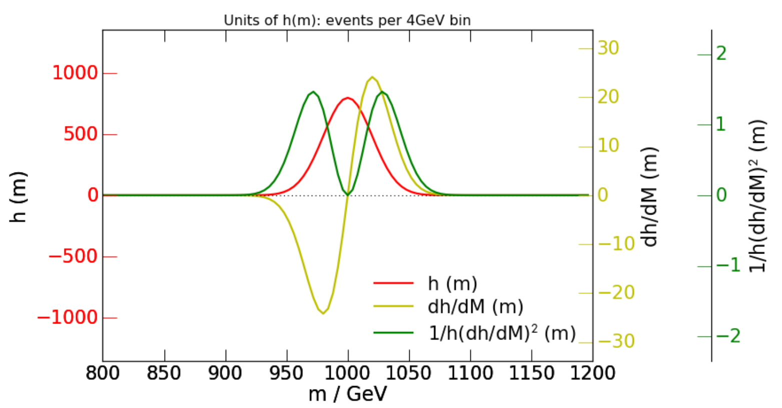


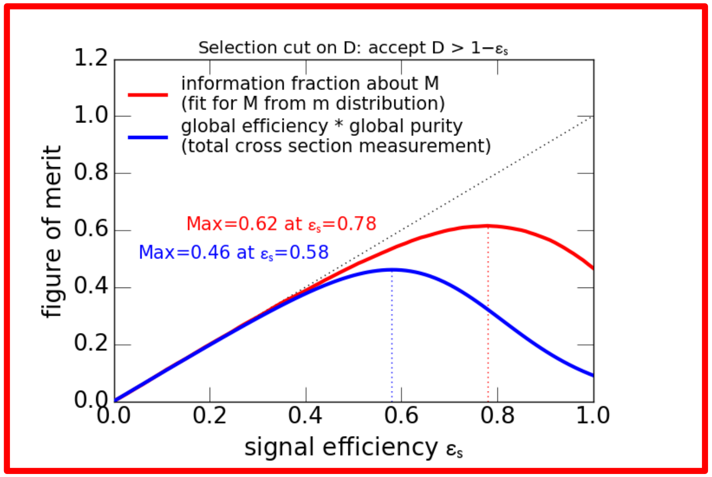
Using scipy / matplotlib / numpy and iminuit in Python from SWAN

M by 1D fit to m – optimizing the classifier

- Choose operating point D_{thr} optimizing information fraction for θ=M in m-fit
  - NB: different to operating point maximising ε*ρ (IF for $\theta=\sigma_s$ in a 1-bin fit)

- To compute IF as sum over bins → need average $\frac{1}{s}\frac{\partial s}{\partial \theta}$ in each bin
  - proof-of-concept → integrate by toy MC with *event-by-event weight derivatives*
    - in a real MC, could save $\frac{1}{|\mathcal{M}|^2}\frac{\partial |\mathcal{M}|^2}{\partial \theta}$ for the matrix element squared $|\mathcal{M}|^2$





M by 1D fit to m – visual interpretation

- Information after cuts: $\sum_i \frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2 * \varepsilon_i * \rho_i \rightarrow$ show the 3 terms in each bin i
  - fit = combine N different measurements in N bins → local $\varepsilon_{i,}$ ρ_i relevant!
  - important thing is: maximise purity, efficiency in bins with highest sensitivity!

Ideal case - yellow histogram (after cuts) coincides with and covers red histogram (ideal)

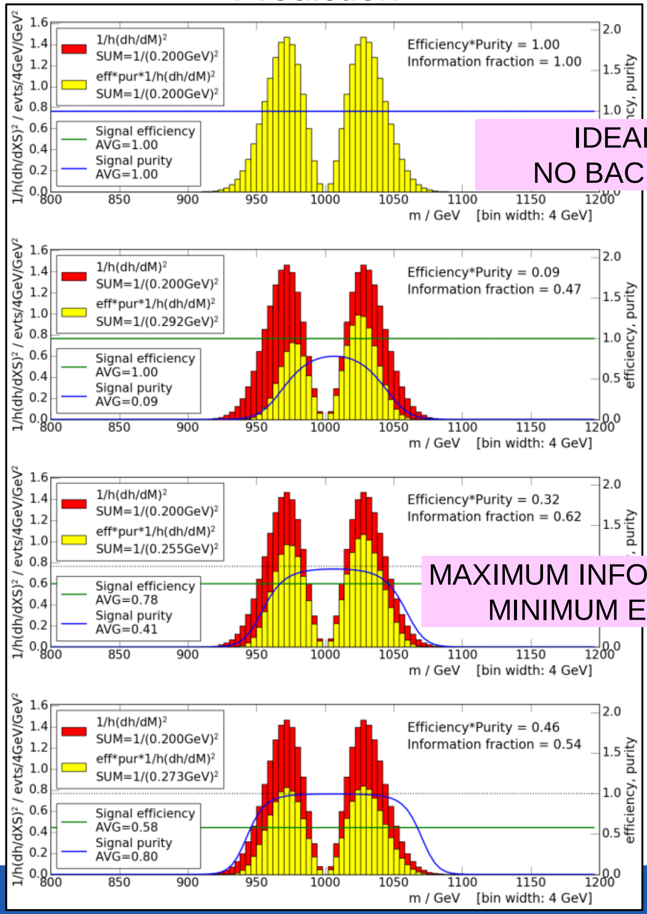
Red histogram: information per bin, ideal case $\frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2$

Blue line: local purity in the bin, ρ_i

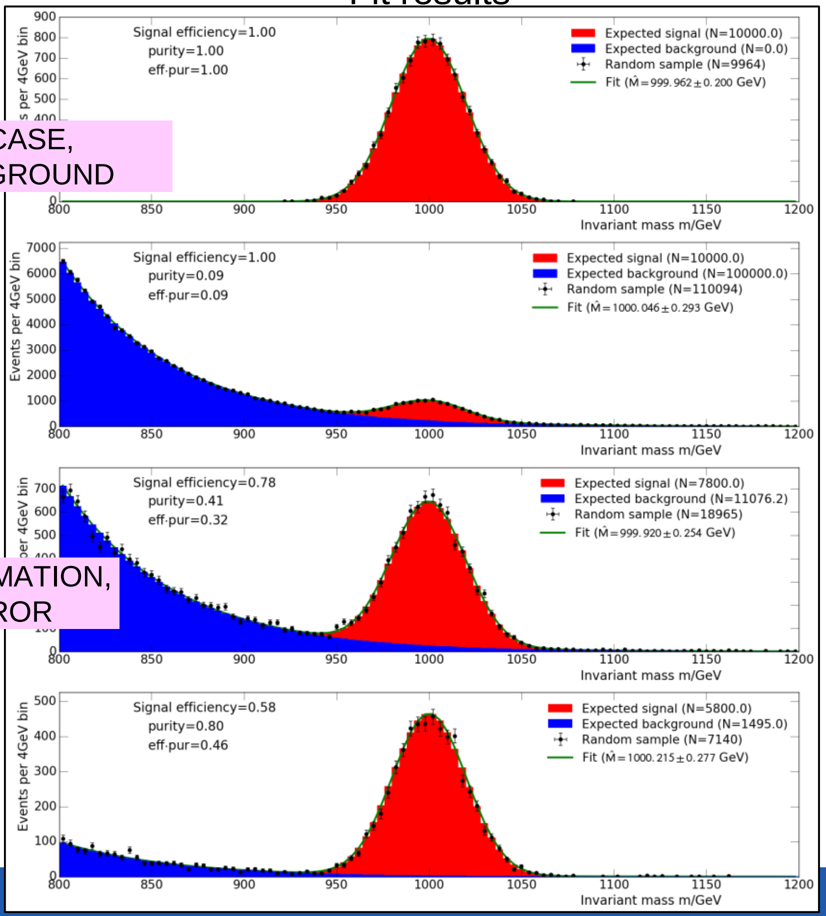
Green line: local efficiency in the bin, ε_i

Yellow histogram: information per bin, after cuts $\varepsilon_i * \rho_i * \frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2$

Prediction



Fit results



IDEAL CASE, NO BACKGROUND

MAXIMUM INFORMATION, MINIMUM ERROR

Optimal partitioning – information inflow

- Information about θ in a binned fit → $\mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i} \left(\frac{\partial y_i}{\partial \theta} \right)^2$
- Do I gain anything by splitting bin y_i into two separate bins? $y_i = w_i + z_i$
  - i.e. is the "information inflow"* positive?

*A. van den Bos, *Parameter Estimation for Scientists and Engineers* (Wiley, 2007).

$$\frac{1}{w_i}\left(\frac{\partial w_i}{\partial \theta}\right)^2 + \frac{1}{z_i}\left(\frac{\partial z_i}{\partial \theta}\right)^2 - \frac{1}{w_i+z_i}\left(\frac{\partial (w_i+z_i)}{\partial \theta}\right)^2 = \frac{\left(w_i\frac{\partial z_i}{\partial \theta} - z_i\frac{\partial w_i}{\partial \theta}\right)^2}{w_i z_i (w_i+z_i)} \geq 0$$

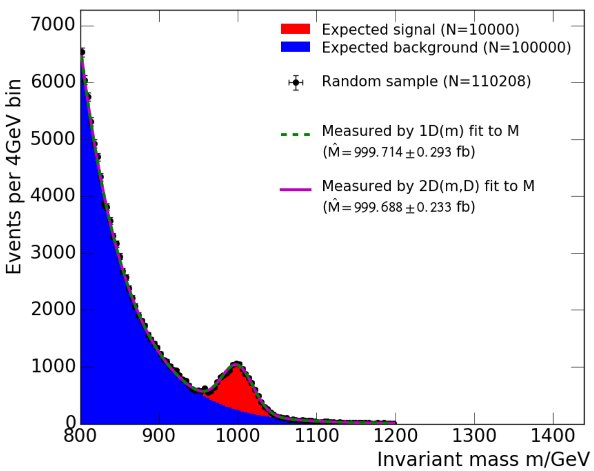
  - information increases (errors on parameters decrease) if $\frac{1}{w_i}\frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i}\frac{\partial z_i}{\partial \theta}$
  - effect of the classifier → information increases if $\rho_w \frac{1}{s_w}\frac{\partial s_w}{\partial \theta} \neq \rho_z \frac{1}{s_z}\frac{\partial s_z}{\partial \theta}$

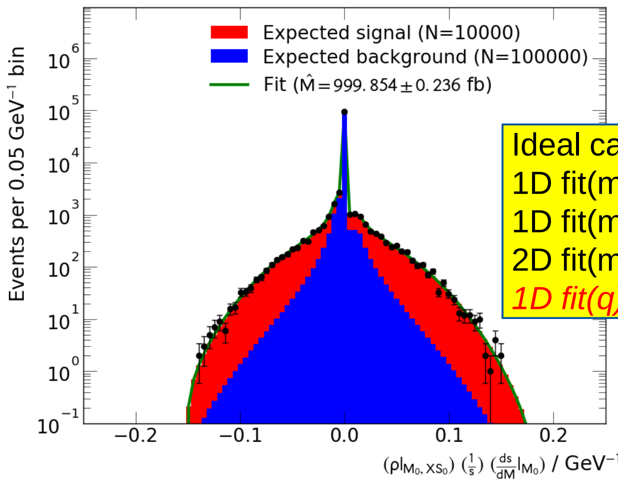
- In summary: **try to partition the data into bins of equal $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$**
  - for cross-section measurements (and searches?): split into bins of equal ρ_i
    - "use the scoring classifier D to partition the data, not to reject events"

Optimal partitioning – optimal variables

- The previous slide implies that $q = \rho \frac{1}{s}\frac{\partial s}{\partial \theta}$ is an optimal variable to fit for θ
  - proof of concept → 1-D fit of q has the same precision on M as 2-D fit of (m,D)
  - closely related to the "optimal observables" technique

M. Davier, L. Duflot, F. LeDiberder, A. Rougé, *The optimal method for the measurement of tau polarization*, Phys. Lett. B 306 (1993) 411. doi:10.1016/0370-2693(93)90101-M
M. Diel, O. Nachtmann, *Optimal observables for the measurement of three-gauge-boson couplings in $e^+e^- \to W^+W^-$*, Z. Phys. C 62 (1994) 397. doi:10.1007/BF01555899
O. Nachtmann, F. Nagel, *Optimal observables and phase-space ambiguities*, Eur. Phys. J. C40 (2005) 497. doi:10.1140/epjc/s2005-02153-9





Ideal case:	± 0.200
1D fit(m), no cut(D):	± 0.292
1D fit(m), optimal cut(D):	± 0.254
2D fit(m,D), no cuts:	± 0.233
1D fit(q):	*± 0.236*

- In practice: train one ML variable to reproduce $\frac{1}{s}\frac{\partial s}{\partial \theta}$?
  - not needed for cross-sections or searches (this is constant)

Conclusion and outlook

- *Different disciplines / problems → different challenges → different metrics*
  - there is no universal magic solution – and the AUC definitely is not one
  - I proposed a systematic analysis of many problems in HEP event selection only

- True Negatives, ROCs & AUCs are irrelevant in HEP event selection
  - PRC approach (like IR, unlike MED) more appropriate → purity ρ, efficiency ε_s

- Binning in HEP analyses → global averages of ρ, ε_s irrelevant in that case
  - FOM integrals that are relevant to HEP use local ρ, ε_s in each bin
  - AUC is an integral of global ρ, ε_s → one more reason why it is irrelevant
  - optimal partitioning exists to minimise statistical errors on fits

- What am I proposing about ROCs and AUCs, essentially?
  - stop using AUCs and ROCs in HEP event selection
    - ROCs confusing → they make you think in terms of the wrong metrics
  - identify the metrics most appropriate to your specific problem
    - I summarized many metrics that exist for some problems in event selection
    - *more research needed* in other problems (e.g. pID, systematics in event selection...)

I am preparing a paper on this – thank you for your feedback in this meeting!

Backup slides of the January IML talk

Systematic errors

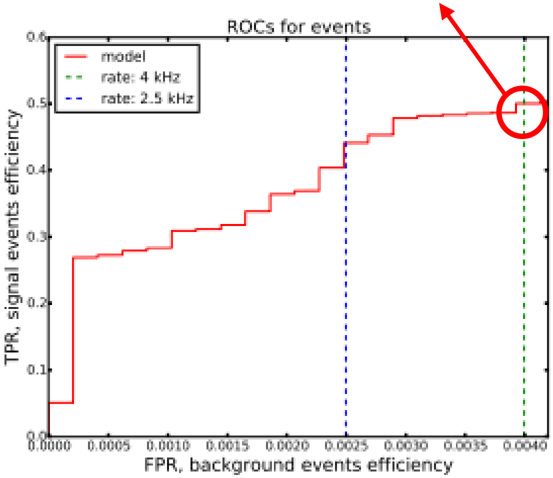
- Statistical errors $\propto \frac{1}{\sqrt{N}}$ $\rightarrow$ systematics become more relevant as N grows
  - Minimise statistical errors at low N $\rightarrow$ only depends on ε_s, ρ
  - Minimise stat+syst errors at high N $\rightarrow$ also depends on luminosity scale (S_{tot})
    - i.e. need all three numbers TP, FP, FN $\rightarrow$ but TN remains irrelevant

- Simple example $\rightarrow$ measure σ_s by counting, 1% relative uncertainty in σ_b
  - systematic error is lower than statistical error if $\left(\frac{1-\rho}{\sqrt{\rho}}\right) \leq \frac{1}{\sqrt{\epsilon_s S_{\mathrm{tot}}}} \times \frac{1}{\Delta\sigma_b/\sigma_b}$
  - optimizing total systematic + statistical error is a tradeoff involving ε_s, ρ, S_{tot}

- Complex problem, no universal recipe $\rightarrow$ interesting problem to work on!
  - more in-depth discussion is *beyond the scope of this talk*

Trigger

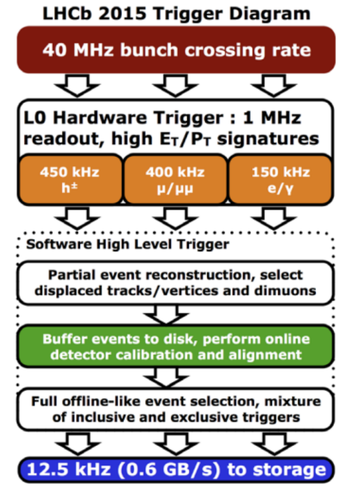
Maximise ε_s at 4 kHz



T. Likhomanenko et al., *LHCb Topological Trigger Reoptimization*, Proc. CHEP 2015, J. Phys. Conf. Series 664 (2015) 082025. doi:10.1088/1742-6596/664/8/082025

Figure 2. Trigger events ROC curve. An output rate of 2.5 kHz corresponds to an FPR of 0.25%, 4 kHz — 0.4%. Thus to find the signal efficiency for a 2.5 kHz output rate, we take 0.25% background efficiency and find the point on the ROC curve that corresponds to this FPR.

IIUC, 4kHz is ε_b (FPR) = 0.4% of 1 MHz L0 hw rate



F. Dordei, *LHCb detector and trigger performance in Run II*, Proc. 5th Int. Conf. on New Frontiers in Physics (ICNFP 2016), EPJ Web of Conferences 164, 01016 (2017). doi:10.1051/epjconf/201716401016

- Different meaning of absolute numbers in the confusion matrix
  - Trigger → events per unit time i.e. trigger rates
  - (Physics analyses → total event sample sizes i.e. total integrated luminosities)

- Binary classifier optimisation goal: maximise ε_s for a given B_{sel} per unit time
  - i.e. maximise TP/(TP+FN) for a given FP → TN irrelevant

- Relevant plot → ε_s vs. B_{sel} per unit time (i.e. *TPR vs FP*)
  - ROC curve (*TPR vs. FPR*) confusing and irrelevant
  - e.g. maximise ε_s for 4 kHz trigger rate, whether L0 rate is 1 MHz or 2MHz

Event selection in HEP searches

- Statistical error in searches by counting experiment → "significance"
  - several metrics → but optimization always involves ε_s, ρ alone → TN irrelevant

C. Adam-Bourdarios et al., *The Higgs Machine Learning Challenge*, Proc. NIPS 2014 Workshop on High-Energy Physics and Machine Learning (HEPML2014), Montreal, Canada, PMLR 42 (2015) 19. http://proceedings.mlr.press/v42/cowa14.html

$$Z_0 = \frac{S_{\rm sel}}{\sqrt{S_{\rm sel}+B_{\rm sel}}} \Longrightarrow (Z_0)^2 = S_{\rm tot}\epsilon_s\rho$$

Z_0 – Not recommended? (confuses search with measuring σ_s once signal established)

Z_2 – Most appropriate? (also used as "AMS2" in Higgs ML challenge)

$$Z_2 = \sqrt{2\left((S_{\rm sel}+B_{\rm sel})\log(1+\frac{S_{\rm sel}}{B_{\rm sel}}) - S_{\rm sel}\right)} \Longrightarrow (Z_2)^2 = 2S_{\rm tot}\epsilon_s\left(\frac{1}{\rho}\log(\frac{1}{1-\rho})-1\right) = S_{\rm tot}\epsilon_s\rho\left(1+\frac{2}{3}\rho+\mathcal{O}(\rho^2)\right)$$

$$Z_3 = \frac{S_{\rm sel}}{\sqrt{B_{\rm sel}}} \Longrightarrow (Z_3)^2 = S_{\rm tot}\epsilon_s\frac{\rho}{1-\rho} = S_{\rm tot}\epsilon_s\rho\left(1+\rho+\mathcal{O}(\rho^2)\right)$$

Z_3 ("AMS3" in Higgs ML) – Most widely used, but strictly valid only as an approximation of Z_2 as an expansion in $S_{sel}/B_{sel} \ll 1$?

$$\frac{S_{\rm sel}}{B_{\rm sel}} = \frac{\rho}{1-\rho} = \rho\left(1+\rho+\mathcal{O}(\rho^2)\right)$$

Expansion in $\rho \ll 1$? – use the expression for Z_2 if anything

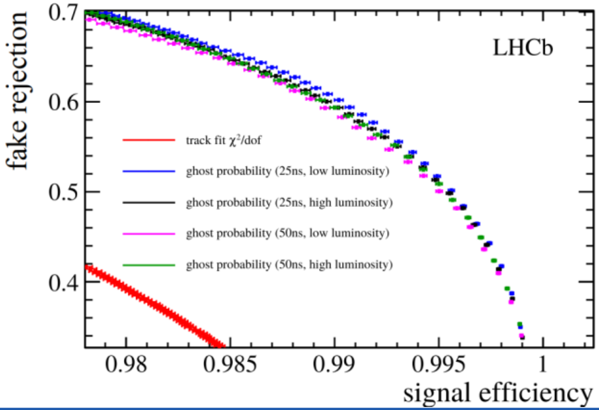
G. Punzi, *Sensitivity of searches for new signals and its optimization*, Proc. PhyStat2003, Stanford, USA (2003). arXiv:physics/0308063v2 [physics.data-an]
G. Cowan, E. Gross, *Discovery significance with statistical uncertainty in the background estimate*, ATLAS Statistics Forum (2008, unpublished). http://www.pp.rhul.ac.uk/~cowan/stat/notes/SigCalcNote.pdf (accessed 15 January 2018)

R. D. Cousins, J. T. Linnemann, J. Tucker, *Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process*, Nucl. Instr. Meth. Phys. Res. A 595 (2008) 480. doi:10.1016/j.nima.2008.07.086
G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C 71 (2011) 15. doi:10.1140/epjc/s10052-011-1554-0

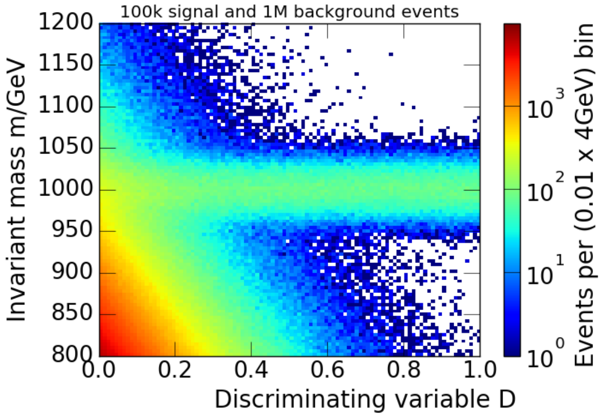
- Several other interesting open questions → *beyond the scope of this talk*
  - optimization of systematics? → e.g. see AMS1 in Higgs ML challenge
  - predict significance in a binned fit? → integral over Z^2 (=sum of log likelihoods)?

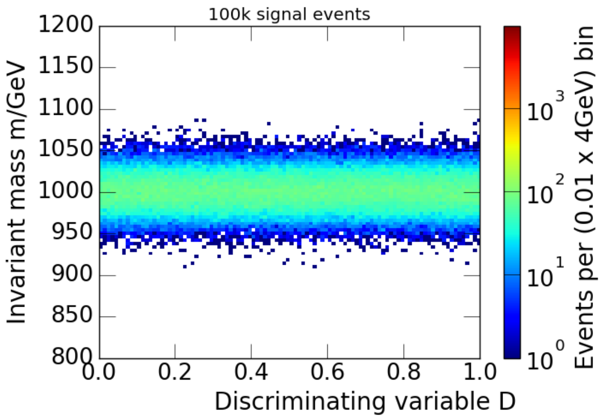
Tracking and particle-ID

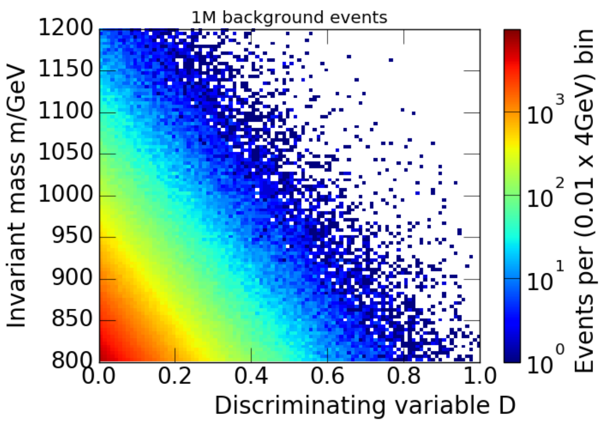
- ROCs irrelevant in event selection → but relevant in other HEP problems
- Event reconstruction and particle identification
  - Binary classifiers on a set of components of one event → not on a set of events
- Example: fake track rejection in LHCb
  - data set within one event: "track" objects created by the tracking software
    - True Positives: tracks that correspond to a charged particle trajectory in MC truth
    - True Negatives: tracks with no MC truth counterpart → relevant and well defined
- Binary classifier evaluation: ε_s and ε_b both relevant → ROC curve relevant
  - is AUC relevant? maximise physics performance? what if ROC curves cross?
  - these questions are *beyond the scope of this talk*



M. De Cian, S. Farry, P. Seyfert, S. Stahl, *Fast neural-net based fake track rejection in the LHCb reconstruction*, LHCb Public Note LHCb-PUB-2017-011 (2017). https://cds.cern.ch/record/2255039







Using scipy / matplotlib / numpy and iminuit in Python from SWAN

Simple toy model

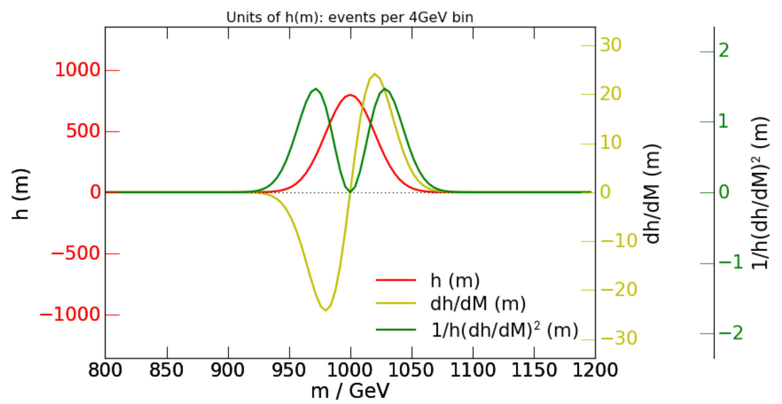
- Two independent observables → f(m,D)=g(D)*h(m)
  - discriminating variable D → scoring classifier
  - invariant mass m → used to fit signal mass M
- Signal (XS=100 fb): Gaussian peak in m, flat in D
  - mass M=1000 GeV, width W=20 GeV
  - flat in D → $\varepsilon_s = 1 - D_{thr}$ if accept events with $D > D_{thr}$
- Background (XS=1000 fb): exponential in both m and D
  - cross-section 1000 fb → B_{tot}=100k
- Two measurements (lumi=100 fb^{-1} → S_{tot}=10k, B_{tot}=100k)
  - mass fit → estimate $\widehat{M}$ (assuming XS, W)
  - cross section fit → estimate $\widehat{XS}$ (assuming M, W)
  - counting, 1D and 2D fits, with/without cuts on D
- Compare binary classifier to ideal case (no bkg):
  - ideal case → $\Delta\widehat{M} = W/\sqrt{S_{tot}}$ = 0.200 GeV
  - ideal case → $\Delta\widehat{XS} = XS/\sqrt{S_{tot}}$ = 1.00 fb

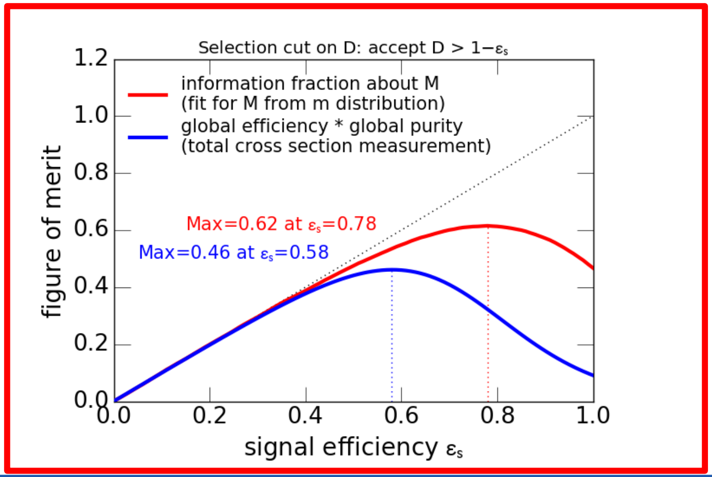
M by 1D fit to m – optimizing the classifier

- Goal: fit true mass M from invariant mass m distribution after a cut on D
  - Vary $\varepsilon_s = 1 - D_{thr}$ by varying cut $D_{thr} \to$ compute information fraction on M for $\varepsilon_s \to$ maximum of information fraction: IF=0.62 ($\Delta\widehat{M} = 0.254 = \frac{0.200}{\sqrt{0.62}}$) at $\varepsilon_s = 0.78$

- Different measurements $\to$ different metrics $\to$ different optimizations
  - maximum of information for fit to M $\to$ IF=0.62 ($\Delta\widehat{M} = 0.254 = \frac{0.200}{\sqrt{0.62}}$) at $\varepsilon_s = 0.78$
  - maximum of information for XS by counting $\to$ $\varepsilon_s * \rho = 0.46$ at $\varepsilon_s = 0.58$

- To compute IF as sum over bins $\to$ need average $\frac{1}{h}\frac{\partial h}{\partial M}$ in each bin
  - proof-of-concept $\to$ integrate by toy MC with event-by-event weight derivatives

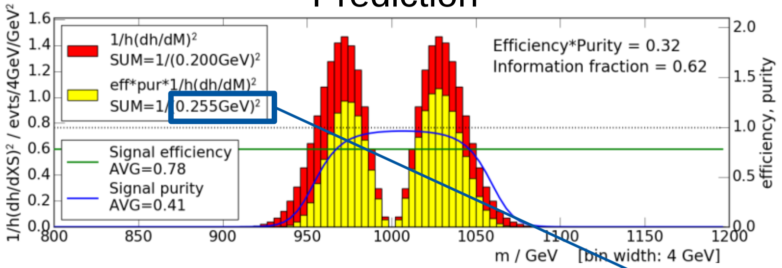




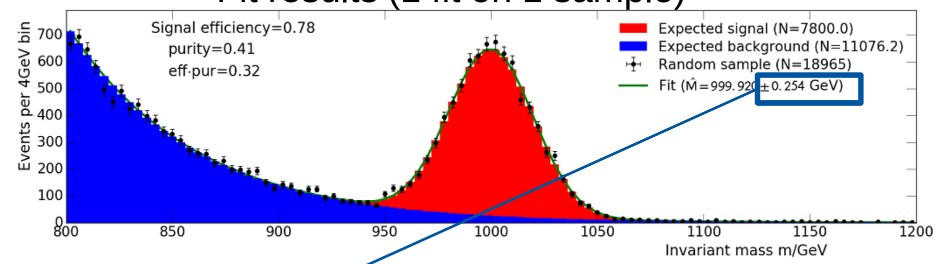
M by 1D fit to m – cross-check

- Cross-check fit error returned by iminuit → repeat fit on 10k samples
  - check this only at the point of max information → ε_s=0.78 and $\Delta\widehat{M}$=0.254

Prediction

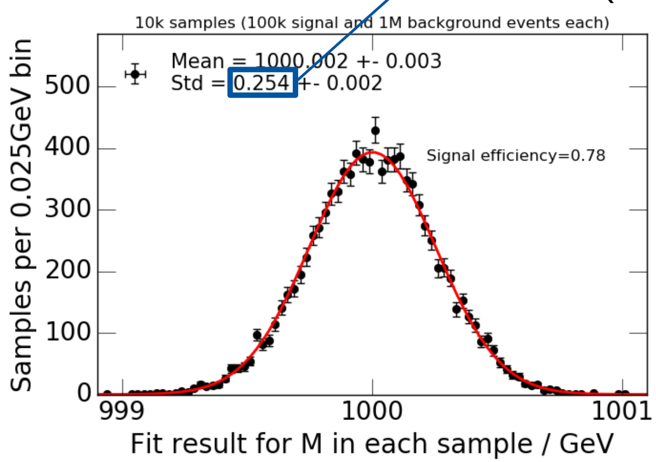


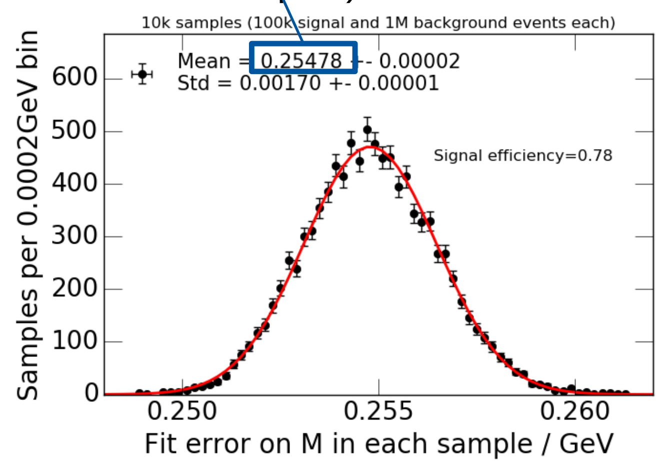
Fit results (1 fit on 1 sample)



OK! $\Delta\widehat{M}$=0.254 consistently

Fit results (10k fits on 10k samples)





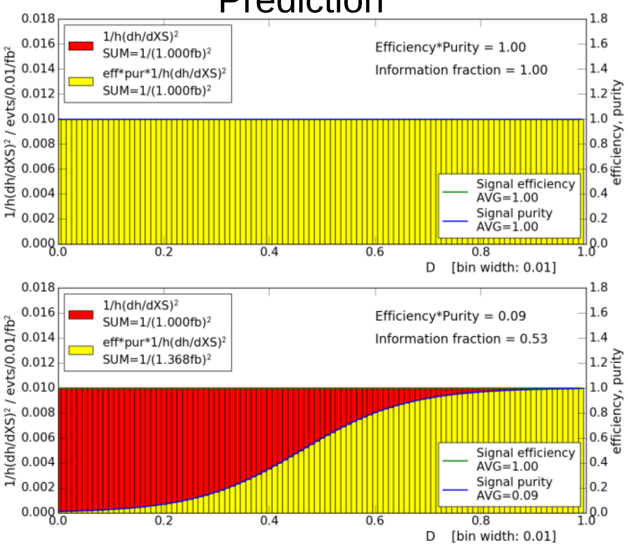
Cross-section by 1D fit to D

i.e. the common practice of “BDT fits”

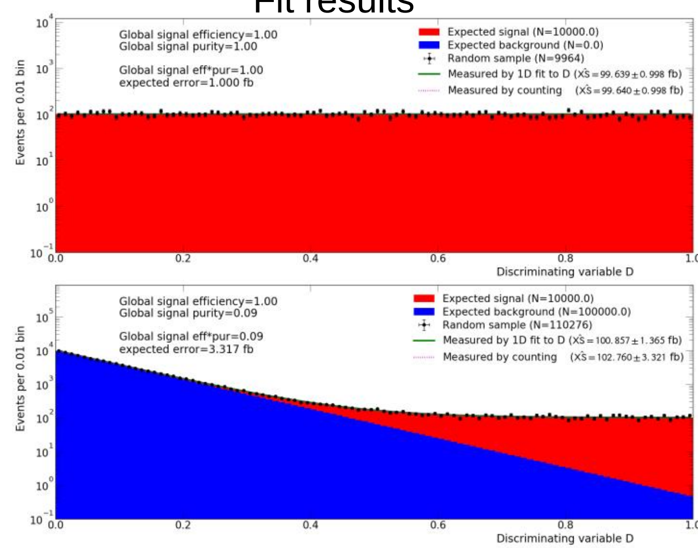
- Cross-section fits analogous to mass fits but simpler
  - Differential cross-section proportional to total cross-section
  - $\frac{1}{s_i}\frac{\partial si}{\partial \sigma_s} = \frac{1}{\sigma_s}$ is constant $\rightarrow \sum_i \frac{1}{s_i}\left(\frac{\partial s_i}{\partial \sigma_s}\right)^2 * \varepsilon_i * \rho_i = \sum_i s_i * \varepsilon_i * \rho_i$
    - special case : for a single bin (counting experiment) $S_{tot} * \varepsilon * \rho \rightarrow$ maximise global $\varepsilon * \rho$

- For simplicity show only fit in D (could fit m, or m and D) and no cuts
  - binning improves precision, also without cuts on D
  - use the scoring classifier D to partition data, not to reject events → next slides

Prediction



Fit results



M by 2D fit – use classifier to partition, not to cut

- Showed a fit for M on m, after a cut on D → can also fit in 2-D with no cuts
  - again, use the scoring classifier D to partition data, not to reject events
- Why is binning so important, especially using a discriminating variable?
  - next slide...