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

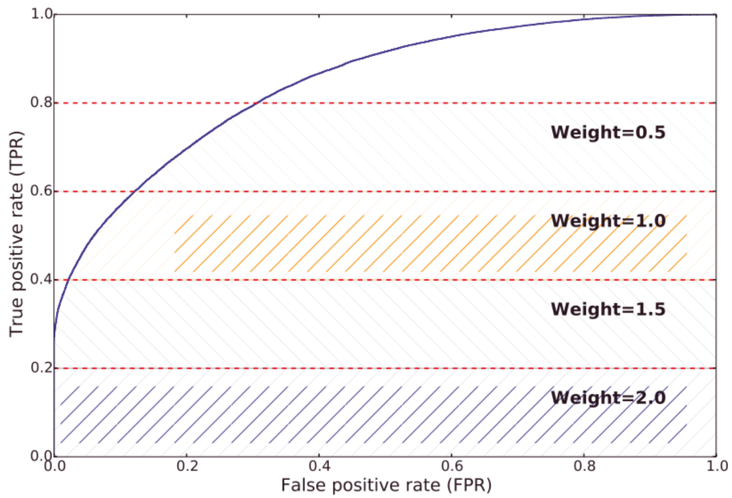


Figure 3: Weights assigned to the different segments of the ROC curve for the purpose of submission evaluation. The x axis is the False Positive Rate (FPR), while the y axis is True Positive Rate (TPR).

T. Blake at al., *Flavours of Physics: the machine learning challenge for the search of $\tau \rightarrow \mu\mu\mu$ decays at LHCb* (2015, unpublished).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"

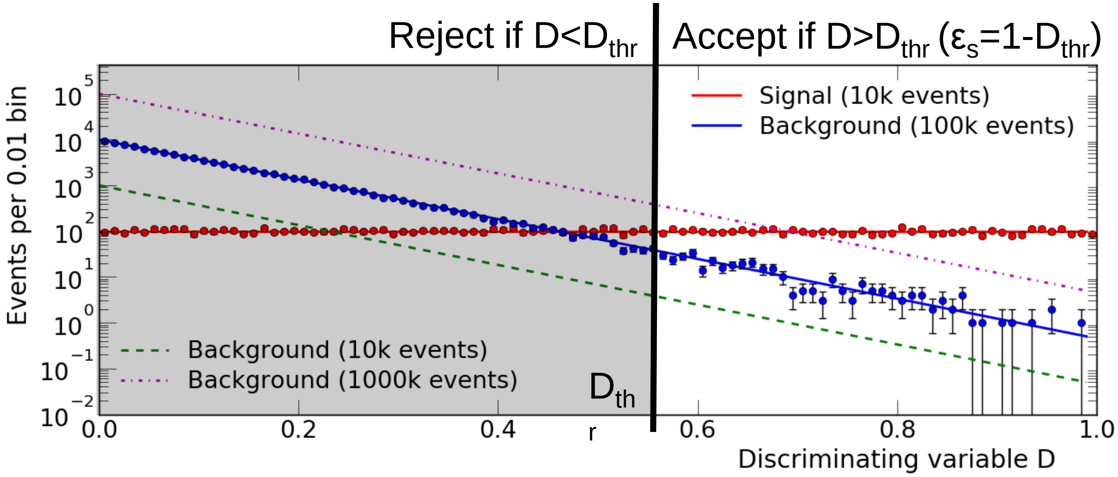
I will cover three domains:

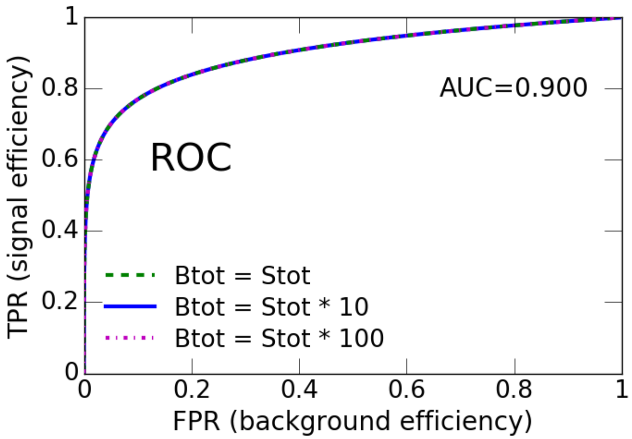
- **Medical Diagnostics (MED)**
  does Mr. A. have cancer?
- **Information Retrieval (IR)**
  Google documents about "ROC"
- **HEP event selection (HEP)**
  select Higgs event candidates

MED: prevalence

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

Discrete vs. Scoring classifiers – ROC curves



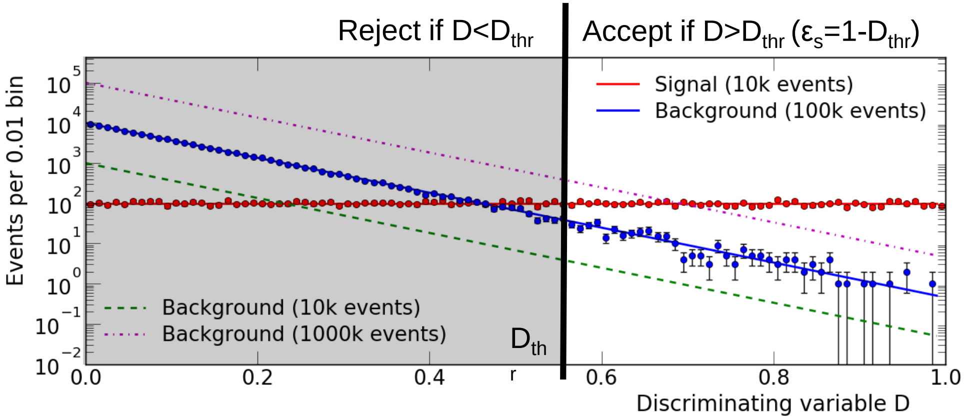


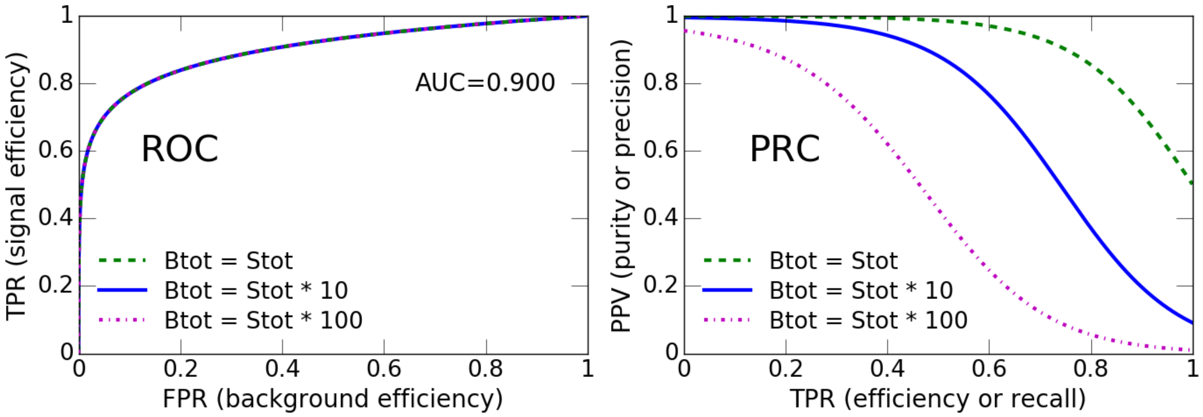
- Discrete classifiers → either select or reject → confusion matrix
- Scoring classifiers → assign score D to each event (e.g. BDT)
  - ideally related to likelihood that event is signal or background (Neyman-Pearson)
  - from scoring to discrete: choose a threshold → classify as signal if D>Dthr
- ROC 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)

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

  – 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)

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 <u>highly</u> 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

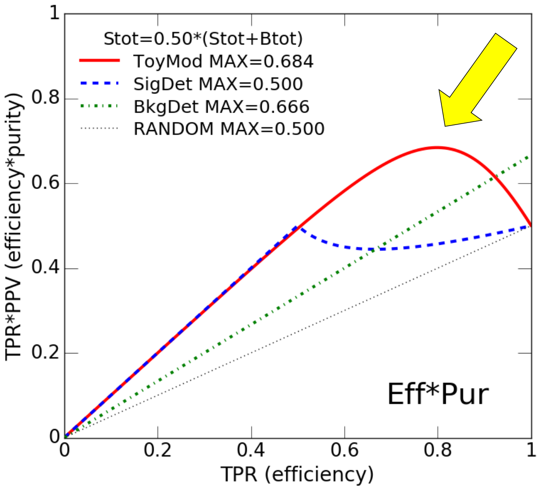
- HEP event selection (HEP)
select Higgs event candidates

- *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



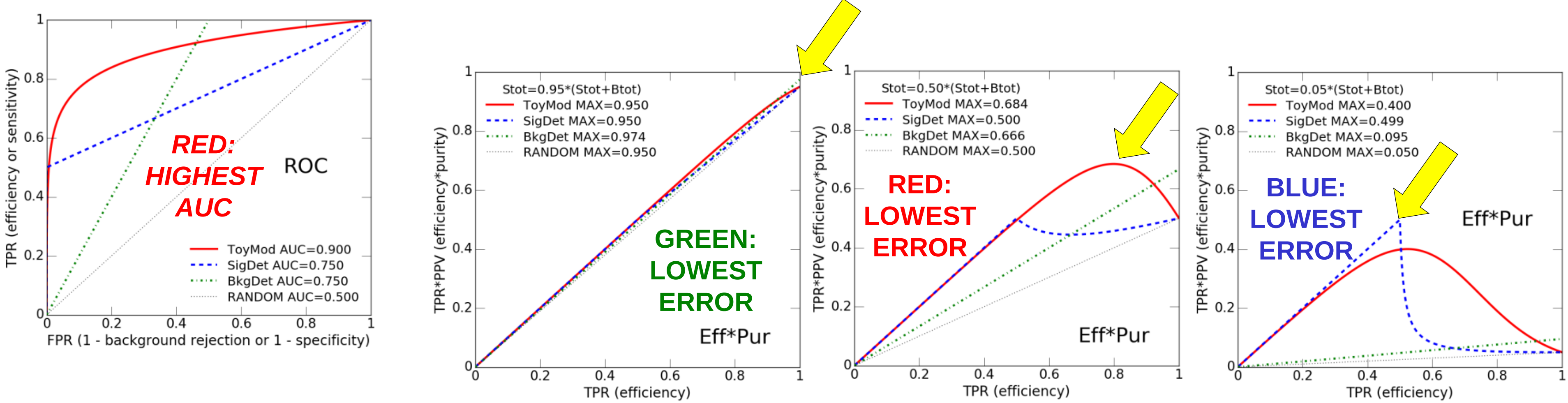
- ε_s*ρ: metric between 0 and 1
  - qualitatively relevant: the higher, the better
  - numerically: fraction of Fisher information (1/error^2) available after selecting
  - *correct metric only for σ_s by counting!* → table with more cases on a next slide

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!

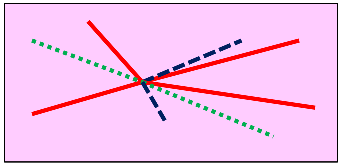
- Example: cross-section by counting
  - maximise product $\varepsilon_s \rho \rightarrow$ i.e. minimise the statistical error $\Delta\sigma^2$
  - depending on S_{tot}/B_{tot}, a different classifier (green, red, blue) should be chosen
  - in two out of three scenarios, the classifier with the highest 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

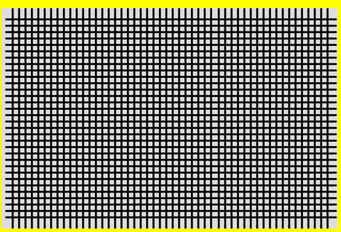
- **HEP event selection (HEP)**
  select Higgs event candidates

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)*

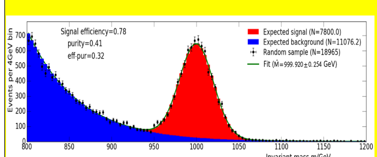
Instances: tracks within one event, created by earlier reconstruction stage.
→ P = real tracks, N = fake tracks (ghosts) → goal: keep real tracks, reject ghosts
→ TN = fake tracks identified as such and rejected: ***TN are relevant*** (IIUC...)
[Optimisation: should translate tracking metrics into measurement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

Property \ Domain	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 $\sum_i s_{tot,i}*\varepsilon_{s,i}*\rho_i*\left(\frac{1}{s_{tot,i}}\frac{\partial s_{tot,i}}{\partial\theta}\right)^2$ Partition in bins of equal $\rho_i*\left(\frac{1}{s_{tot,i}}\frac{\partial s_{tot,i}}{\partial\theta}\right)$
	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:

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

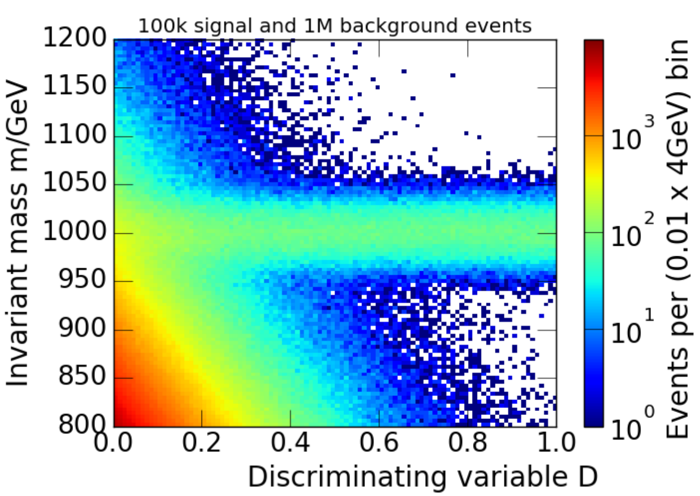
  - *ε_i and ρ_i → local signal efficiency and purity in the i^{th} bin*
- Define a binary classifier metric as information fraction to ideal classifier:
  - in [0,1] → 1 if keep all signal and reject all backgrounds
  - higher is better → maximise IF
  - interpretation: $(\Delta\hat{\theta}^{(\text{real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

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

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

Numerical tests with a toy model

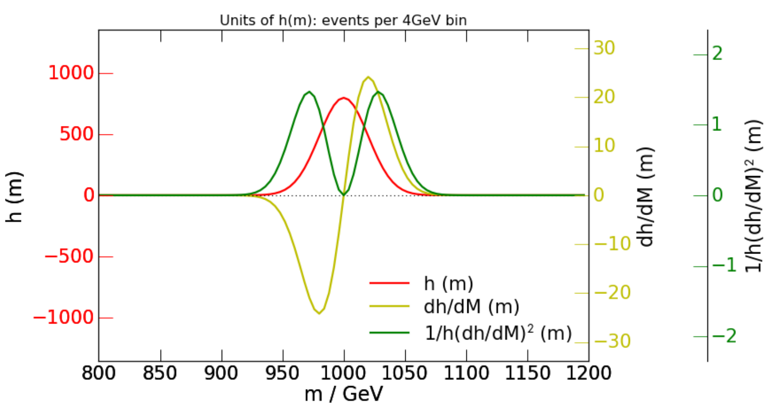
- I used a simple toy model to make some numerical tests
  - Verify that my formulas are correct – and also illustrate them graphically
  - Two-dimensional distribution (m,D) $\rightarrow$ signal Gaussian, background exponential
- Two measurements:
  - total cross-section measurement by counting and 1-D or 2-D fit
  - mass measurement by 1-D or 2-D fits
- Details in the backup slides

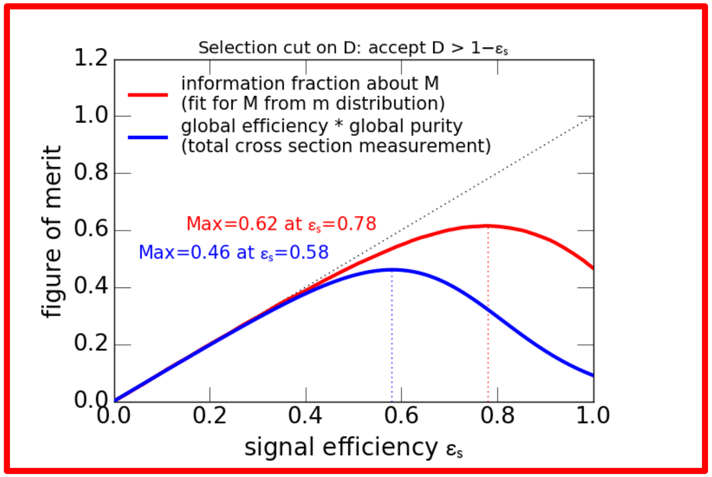


Using scipy / matplotlib / numpy and iminuit in Python from SWAN

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 $\rightarrow$ local ε_i, ρ_i relevant!

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

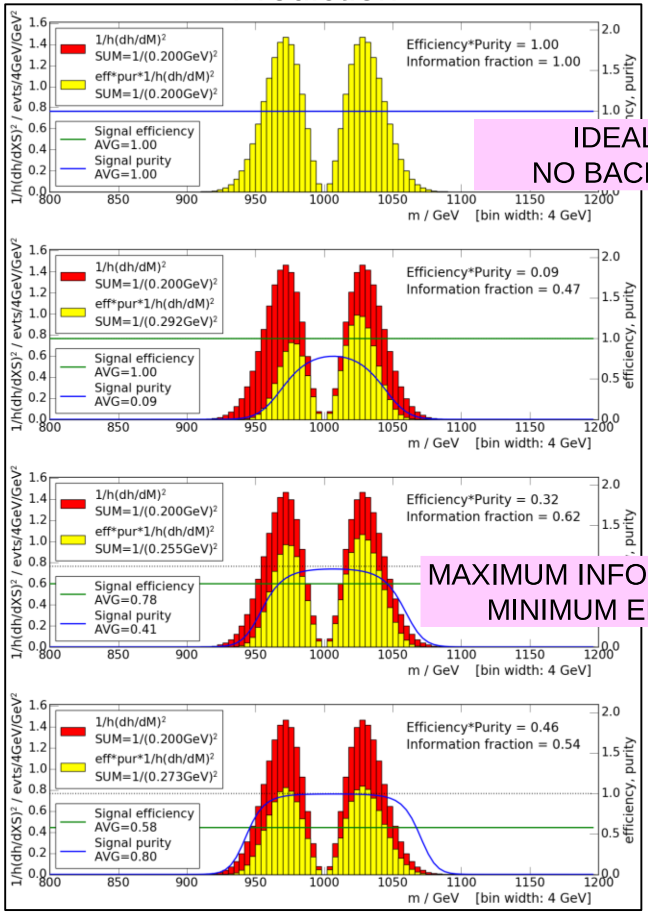
Red histogram: information per bin, ideal case $\frac{1}{s_i}\left(\frac{\partial si}{\partial M}\right)^2$

Blue line: local purity in the bin, ρ_i

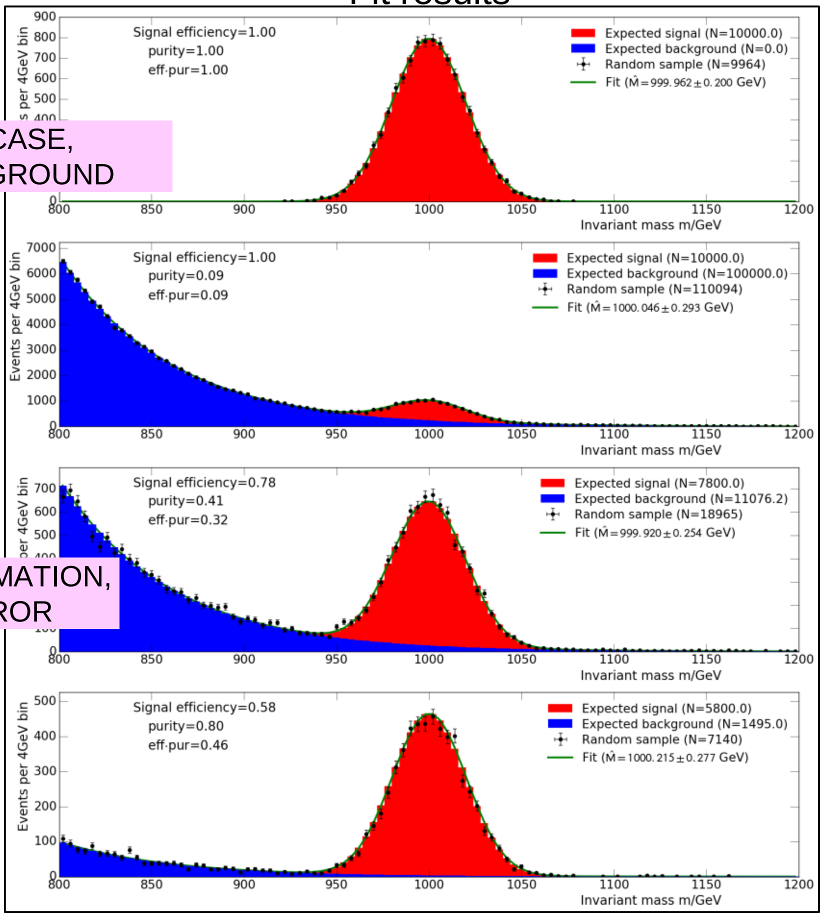
Green line: local efficiency in the bin, ε_i

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



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).

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

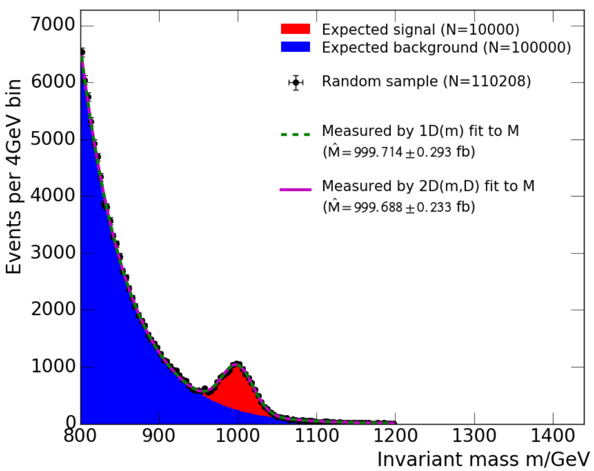
  - information increases (errors on parameters decrease) if $\frac{1}{w_i}\frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i}\frac{\partial z_i}{\partial \theta}$
  - effect of the classifier → information increases if $\rho_w \frac{1}{s_w}\frac{\partial s_w}{\partial \theta} \neq \rho_z \frac{1}{s_z}\frac{\partial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

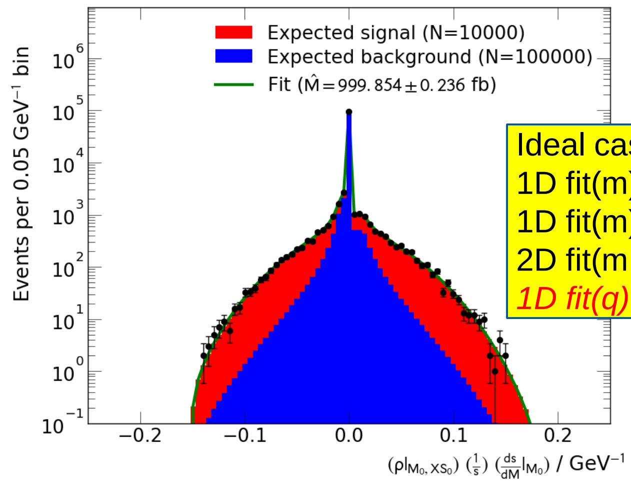
- In summary: **try to partition the data into bins of equal $\rho_i \frac{1}{s_i}\frac{\partial s_i}{\partial \theta}$**
  - for cross-section measurements (and searches?): split into bins of equal ρ_i
    - "use the scoring classifier D to partition the data, not to reject events"

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

Statistical error in binned fits

- Observed data: event counts n_i in m bins of a (multi-D) distribution f(x)
  - the expected 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 $\frac{1}{(\Delta\hat{\theta})^2_{(\mathrm{bin}\ dx)}} = \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{1}{(\Delta f)^2} = \left(\frac{\partial f}{\partial\theta}\right)^2 \left(\frac{\sqrt{dx}}{\sqrt{f}}\right)^2 = \left(\frac{\partial f}{\partial\theta}\right)^2 \frac{dx}{f}$
  - 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 $(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$

Effect of realistic classifiers on fits

- Previous slide: variance on estimated $\hat{\theta}$ is $(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$ where $\mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i} \left(\frac{\partial y_i}{\partial \theta} \right)^2$
- With an *ideal classifier*, all signal events and only signal events are selected, i.e. $y_i = S_i$, hence: $\mathcal{I}_\theta^{(\text{ideal classifier})} = \sum_{i=1}^{m} \frac{1}{S_i} \left(\frac{\partial S_i}{\partial \theta} \right)^2$
- With a realistic classifier, only a fraction of all available signal events are selected, as well as some background events: $y_i(\theta) = \epsilon_i S_i(\theta) + b_i$
  - here ε_i is the *local signal efficiency* in bin x_i
  - note that $\frac{1}{y_i} = \rho_i \frac{1}{\epsilon_i S_i}$ where the *local signal purity* is defined as $\rho_i = \frac{s_i}{s_i + b_i}$
  - the available information is therefore reduced to $\mathcal{I}_\theta^{(\text{real classifier})} = \sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i} \left(\frac{\partial S_i}{\partial \theta} \right)^2$
- In summary, with respect to an ideal classifier, a realistic classifier leads to a higher error on the fitted parameter, $(\Delta\hat{\theta}^{(\text{real classifier})})^2 \geq \frac{1}{\mathrm{IF}} (\Delta\hat{\theta}^{(\text{ideal classifier})})^2$
- "IF" is the "information fraction" available after cuts: $$\mathrm{IF} = \frac{\mathcal{I}_\theta^{(\text{real classifier})}}{\mathcal{I}_\theta^{(\text{ideal classifier})}} = \frac{\sum_{i=1}^{m} \epsilon_i \rho_i \times \frac{1}{S_i} \left(\frac{\partial S_i}{\partial \theta} \right)^2}{\sum_{i=1}^{m} \frac{1}{S_i} \left(\frac{\partial S_i}{\partial \theta} \right)^2}$$

Information fraction vs. AUC

- "IF" is a figure of merit between 0 and 1 (like the AUC...)

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

  - it depends on efficiency and purity (PRC rather than ROC)
    - True Negatives are irrelevant...
  - it depends on local efficiencies and purities
    - but also applies to counting experiments (1 single "bin") – see examples
  - it depends on the choice of a point on the PRC/ROC (a threshold on D)
    - but one can also use it in a fit to the full distribution of D – see examples
  - it is qualitatively (higher is better) and quantitatively ($\Delta\hat{\theta} \sim 1/\mathrm{IF}$) relevant

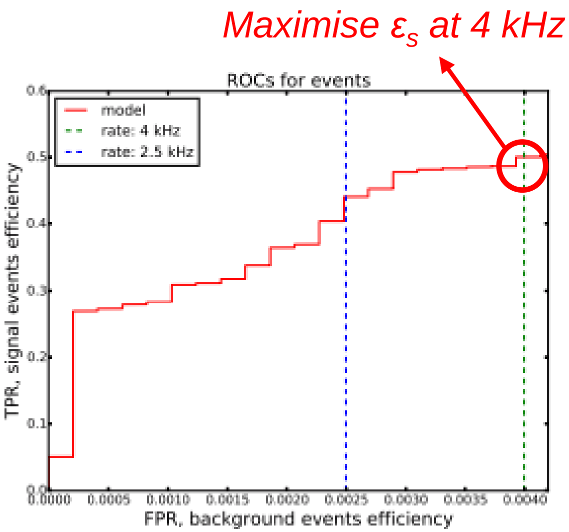
- *A different figure of merit is needed for every different problem!*
  - I derived this for statistical errors in parameter fits (precision measurements)
  - A similar f.o.m. can certainly be derived for optimizing searches
    - "combining" the different bins of the distribution is done slightly differently...
  - Systematic errors need to be handled differently...

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*

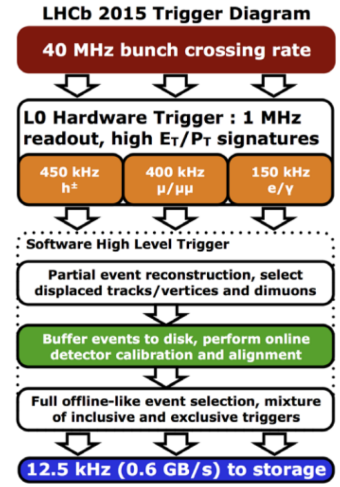


Trigger

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

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

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)

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

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

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

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

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

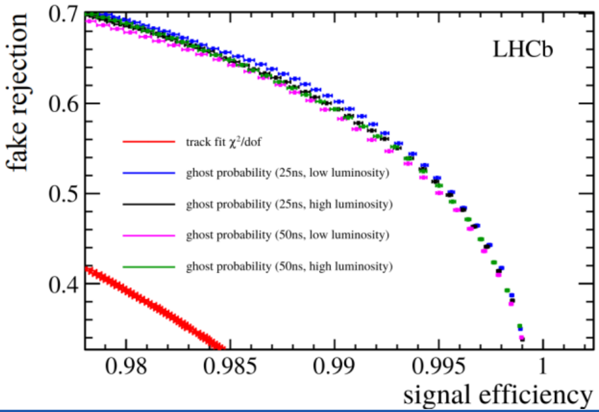
G. Punzi, *Sensitivity of searches for new signals and its optimization*, Proc. PhyStat2003, Stanford, USA (2003). arXiv:physics/0308063v2 [physics.data-an]
G. Cowan, E. Gross, *Discovery significance with statistical uncertainty in the background estimate*, ATLAS Statistics Forum (2008, unpublished). 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

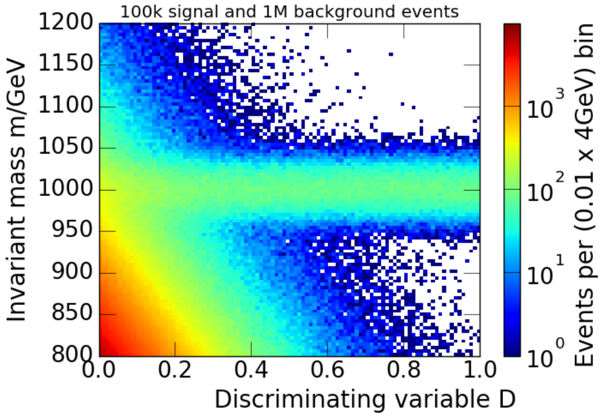
- Several other interesting open questions → *beyond the scope of this talk*
  - optimization of systematics? → e.g. see AMS1 in Higgs ML challenge
  - predict significance in a binned fit? → integral over Z^2 (=sum of log likelihoods)?

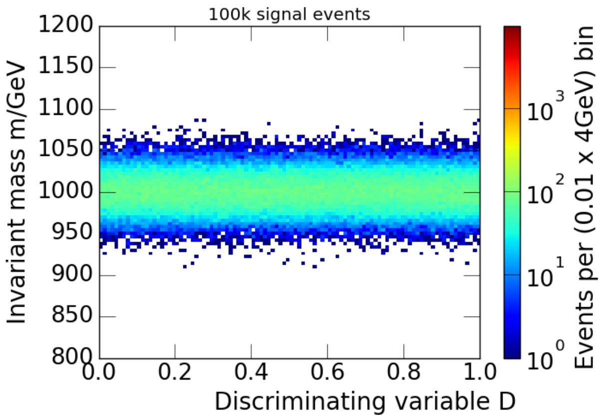
Tracking and particle-ID

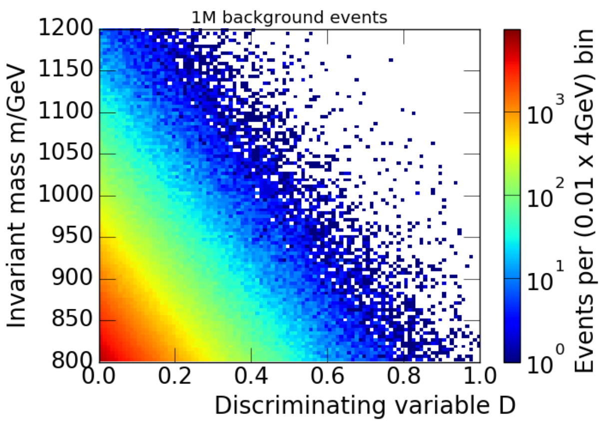
- ROCs irrelevant in event selection → but relevant in other HEP problems
- Event reconstruction and particle identification
  - Binary classifiers on a set of components of one event → not on a set of events
- Example: fake track rejection in LHCb
  - data set within one event: “track” objects created 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







Simple toy model

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

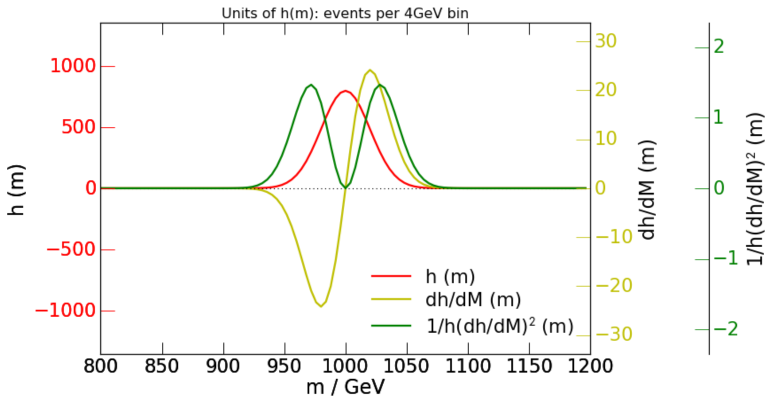
- Two independent observables → f(m,D)=g(D)*h(m)
  - discriminating variable D → scoring classifier
  - invariant mass m → used to fit signal mass M
- Signal (XS=100 fb): Gaussian peak in m, flat in D
  - mass M=1000 GeV, width W=20 GeV
  - flat in D → $\varepsilon_s=1-D_{thr}$ if accept events with $D>D_{thr}$
- 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

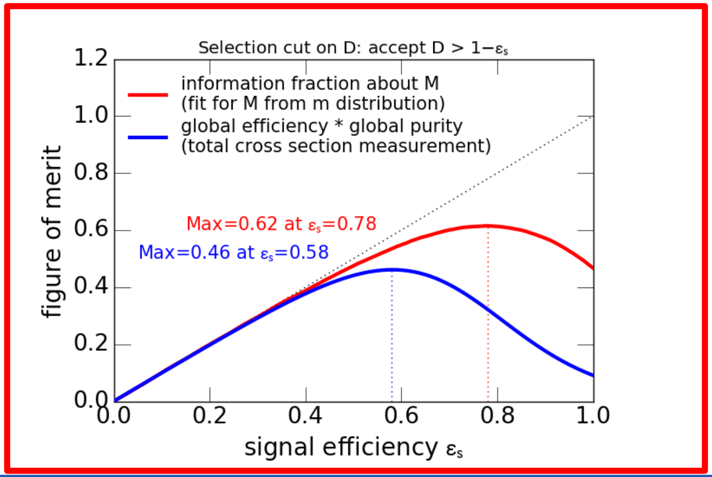
M by 1D fit to m – optimizing the classifier

- Goal: fit true mass M from invariant mass m distribution after a cut on D
  - Vary $\varepsilon_s=1-D_{thr}$ by varying cut $D_{thr} \rightarrow$ compute information fraction on M for $\varepsilon_s \rightarrow$ 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 $\rightarrow$ different metrics $\rightarrow$ different optimizations
  - maximum of information for fit to M $\rightarrow$ 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 $\rightarrow$ ε_s*ρ=0.46 at $\varepsilon_s=0.58$

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

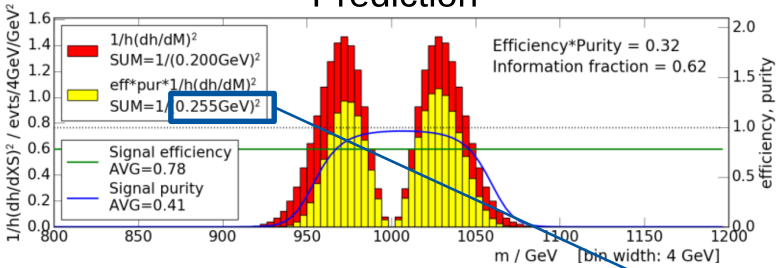




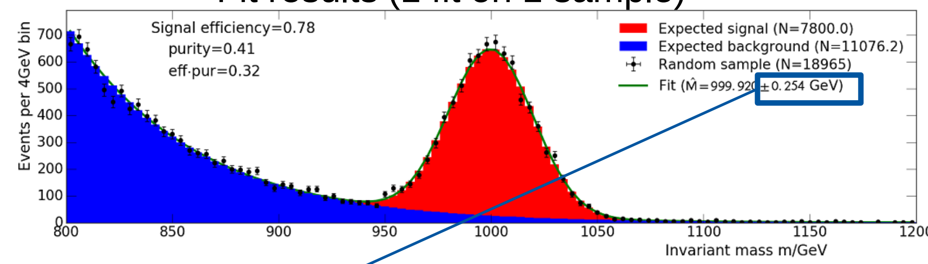
M by 1D fit to m – cross-check

- 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

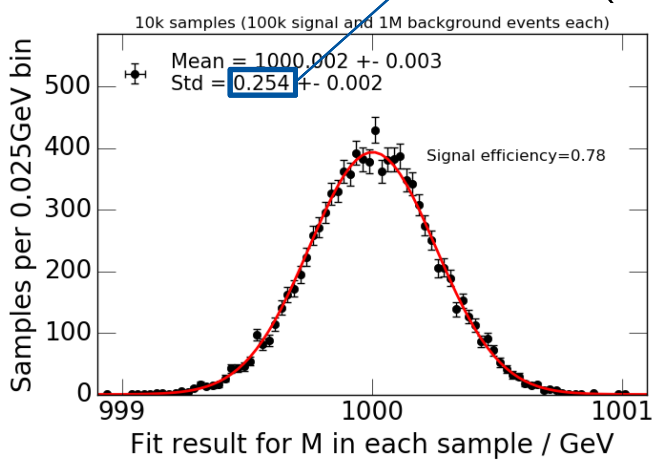


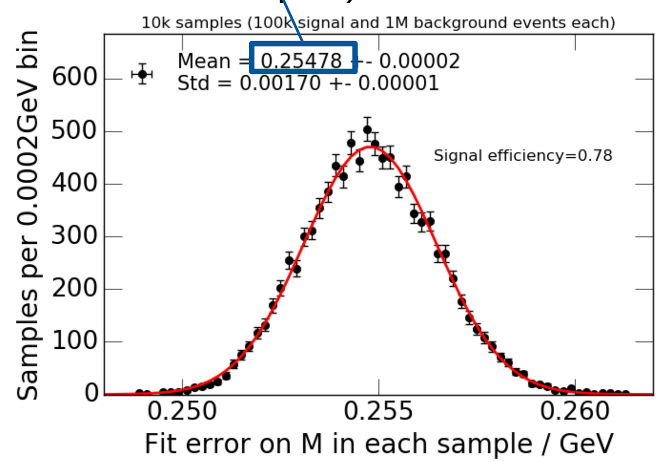
Fit results (1 fit on 1 sample)



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

Fit results (10k fits on 10k samples)





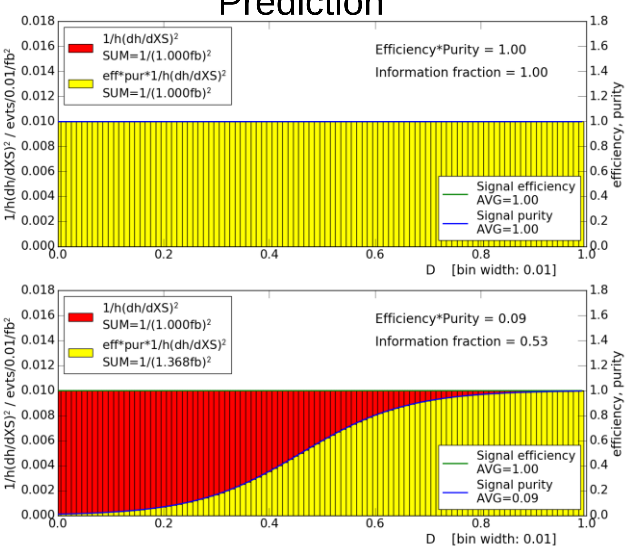
Cross-section by 1D fit to D

i.e. the common practice of "BDT fits"

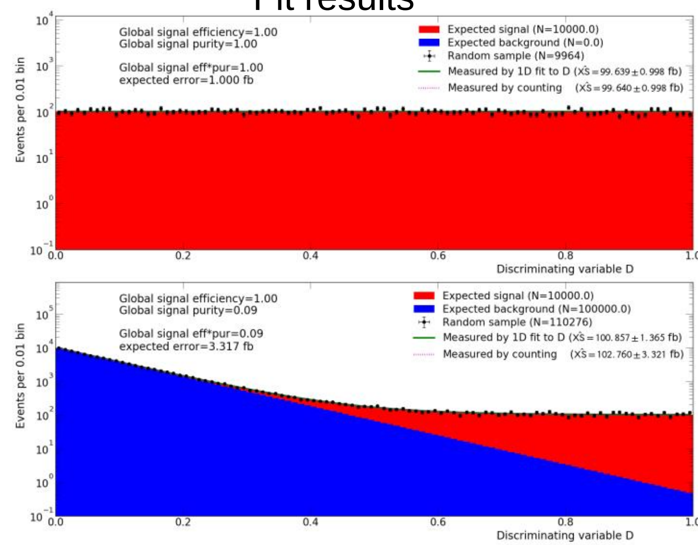
- Cross-section fits analogous to mass fits but simpler
  - Differential cross-section proportional to total cross-section
  - $\frac{1}{s_i}\frac{\partial si}{\partial \sigma_s} = \frac{1}{\sigma_s}$ is constant $\rightarrow \sum_i \frac{1}{s_i}\left(\frac{\partial s_i}{\partial \sigma_s}\right)^2 * \varepsilon_i * \rho_i = \sum_i s_i * \varepsilon_i * \rho_i$
    - 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 $\rightarrow$ next slides

Prediction

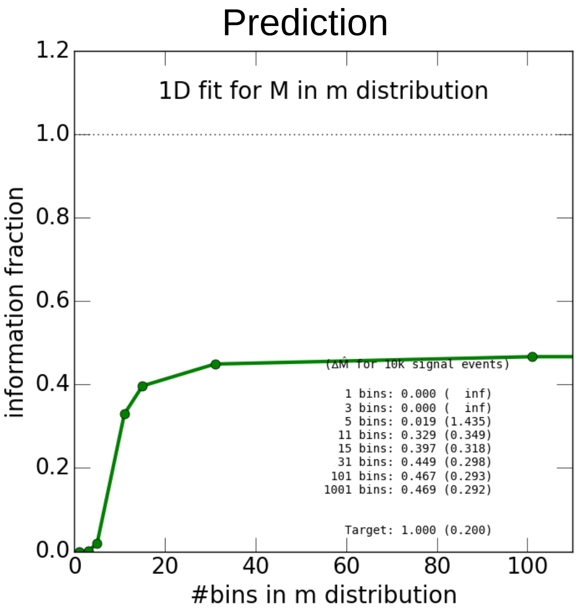


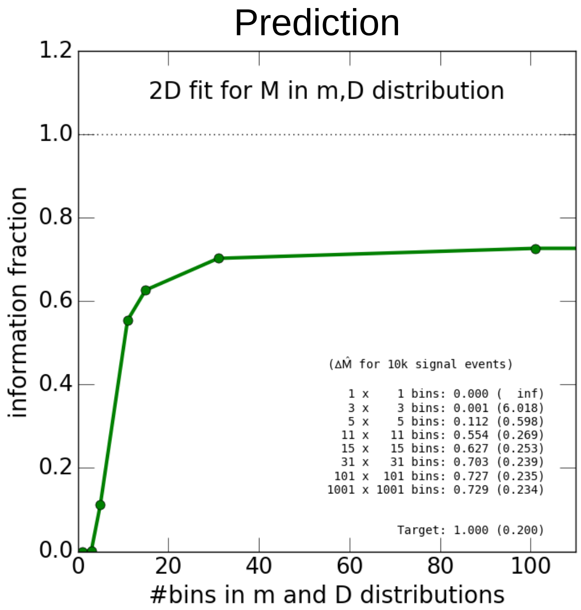
Fit results

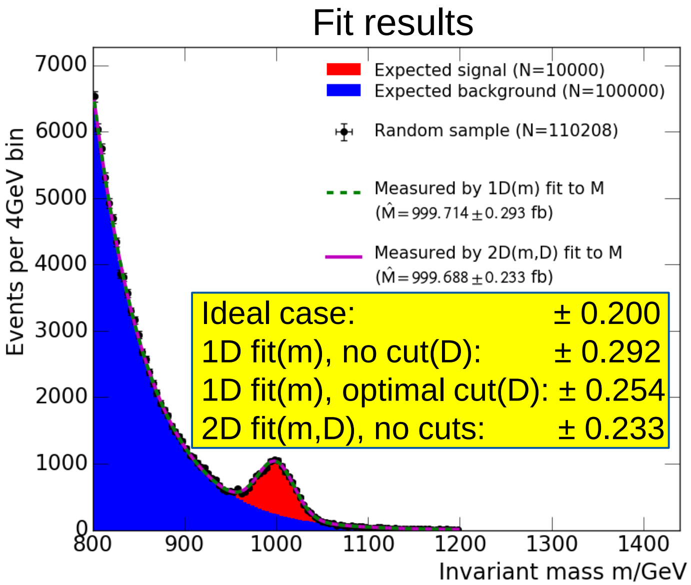


M by 2D fit – use classifier to partition, not to cut

- 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...

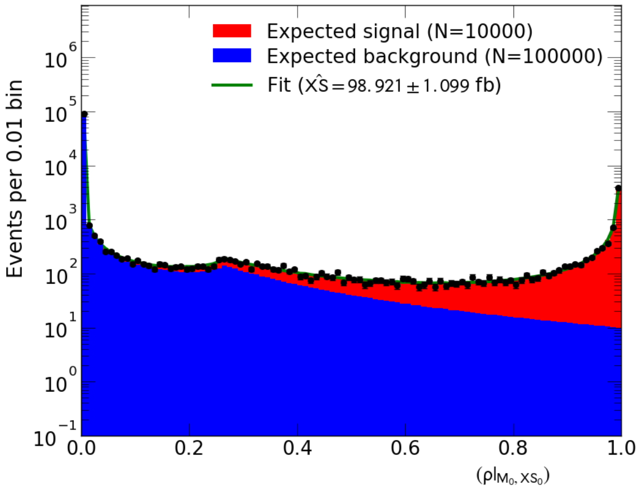


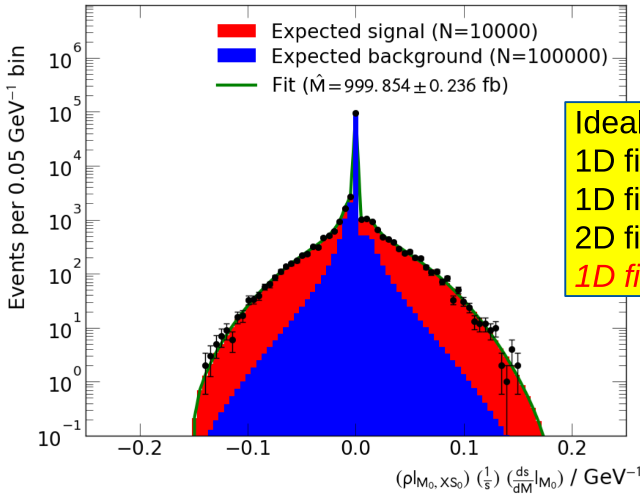




Optimal partitioning – optimal variables

- How to partition the data into bins of equal $\rho_i \frac{1}{s_i}\frac{\partial si}{\partial \sigma_s}$?
  - as a proof of concept → also made a 1D fit for M against this one variable "q"
  - not surprisingly, the precision is the same as that of the 2D fit on m,D





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(optimal q): :	*± 0.236*

- In practice: train one ML variable to reproduce $\frac{1}{s_i}\frac{\partial si}{\partial \sigma_s}$?

- Same general idea as 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

OLDER SLIDES

HEP event selection properties

- **HEP event selection (HEP)**
  select Higgs event candidates

- Binary classifier optimisation goal: maximise physics reach at given budget
  - Trigger and computing → maximise signal event throughput within constraints
  - Physics analyses → maximise physics information from available data sets

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

- I will attempt a systematic analysis of properties:
  - 1. Qualitative class imbalance → signal relevant, background irrelevant
    - TN irrelevant and ill-defined (preselection, generator cuts) → only TP, FP, FN matter
  - 2. Extreme quantitative class imbalance → signal events swamped in background
  - 3. Prevalence largely constant in time → fixed by quantum physics cross section
    - Prevalence: known in advance for precision measurements; unknown for searches.
  - 4. Scale invariance (with two exceptions) → optimization based on 2 ratios ε_s, ρ
    - Exception: trigger rate → constraint on throughput of FP(+TP) per unit time
    - Exception: total error (statistical + systematic) minimization also depends on scale L
  - 5. Fits to differential distributions → local ε_s, ρ relevant (global ε_s, ρ ~irrelevant)

- More details and examples in the following slides

Medical diagnostics (1) – accuracy

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

- Binary classifier optimisation goal: maximise "diagnostic accuracy"
  - not obvious: many different specific goals → many different possible definitions
    - patient's perspective → minimise diagnostic impact and impact of no/wrong treatment
    - society's perspective: ethical and economic → allocate healthcare with limited budget
    - physician's perspective → get knowledge of patient's condition, manage patient

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

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

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

True Positives (TP) (correctly diagnosed as ill)	False Positives (FP) (truly healthy, but diagnosed as ill)
False Negatives (FN) (truly ill, but diagnosed as healthy)	True Negatives (TN) (correctly diagnosed as healthy)

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

where "prevalence" is $\pi_s = \frac{S_{\mathrm{tot}}}{S_{\mathrm{tot}} + B_{\mathrm{tot}}}$

- Symmetric → all patients important, both truly ill (TP) and truly healthy (TN)

Medical diagnostics (2) – from ACC to ROC

- ACC metric → widely used in medical diagnostics in the 1980-'90s (still now?)
  - Also "by far the most commonly used metric" in ML in the 1990s

- Limitation: ACC depends on relative prevalence
  - issue for imbalanced problems → diagnostic accuracy for rare diseases
  - issue if prevalence unknown or variable over time → disease epidemics

- Since the '90s → shift from ACC to ROC in MED and ML fields
  - TPR (sensitivity) and TNR (specificity) studied separately
    - reminder: all patients important, both truly ill (TP) and truly healthy (TN)

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

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

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

- Evaluation often based on the AUC → two advantages *for medical diagnostics*:
  - 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)

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 (3) – from ROC to PRC?

- ROC and AUC metrics → currently widely used in medical diagnostics and ML
- Limitation: ROC-based evaluation questionable for *highly imbalanced data sets*
  - – ROC may provide an overly optimistic view of performance with highly skewed data sets
- 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

Simplest HEP example – total cross-section

- Total cross-section measurement in a counting experiment

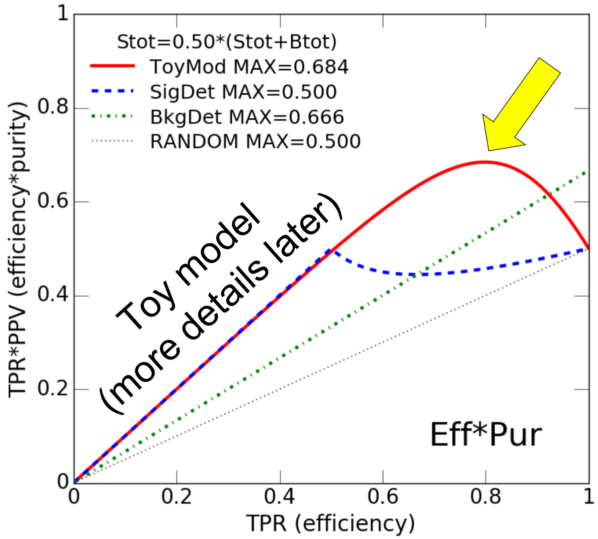
- To minimize statistical errors: *maximise efficiency*purity ε_s*ρ*
  - well-known since decades
  - *global* efficiency $\varepsilon_s = S_{sel}/S_{tot}$ and *global* purity $\rho = S_{sel}/(S_{sel}+B_{sel})$ – "1 single bin"

$$(\sigma_s)_{\mathrm{meas}} = \frac{N_{\mathrm{meas}} - \mathcal{L}\epsilon_b\sigma_b}{\mathcal{L}\epsilon_s}$$

$$\Delta\sigma_s = \frac{\Delta N_{\mathrm{meas}}}{\mathcal{L}\epsilon_s} = \frac{1}{\mathcal{L}\epsilon_s}\sqrt{N_{\mathrm{exp}}}$$

$$N_{\mathrm{exp}} = S_{\mathrm{sel}} + B_{\mathrm{sel}} = \frac{S_{\mathrm{sel}}}{\rho} = \frac{\mathcal{L}\sigma_s\epsilon_s}{\rho}$$

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



- ε_s*ρ: metric between 0 and 1
  - qualitatively relevant *(only for this specific use case!)*: the higher, the better
  - numerically: fraction of Fisher information (1/error^2) available after selecting

Predict and optimize statistical errors in binned fits

- Observed data: event counts n_i in m bins of a (multi-D) distribution f(x)
  - expected counts $y_i = f(x_i,\theta)dx \rightarrow$ 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]

- Easy to show (backup slides) that minimum variance achievable is:

$(\Delta\hat{\theta})^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$ (Cramer-Rao lower bound), where $\mathcal{I}_\theta = \sum_{i=1}^{m} \frac{1}{y_i}\left(\frac{\partial y_i}{\partial\theta}\right)^2 = \int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx$ (Fisher information)

- With an ideal classifier (or no background) → $y_i=S_i$ and $\mathcal{I}_\theta^{(\text{ideal classifier})} = \sum_{i=1}^{m} \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2$

- With a realistic classifier → $y_i(\theta) = \epsilon_i S_i(\theta) + b_i$ and $\mathcal{I}_\theta^{(\text{real classifier})} = \sum_{i=1}^{m} \epsilon_i\rho_i \times \frac{1}{S_i}\left(\frac{\partial S_i}{\partial\theta}\right)^2$
  - *ε_i and ρ_i → local signal efficiency and purity in the i^{th} bin*

- Binary classifier optimization → maximise
  - higher is better
  - interpretation: $(\Delta\hat{\theta}^{(\text{real classifier})})^2 \geq \frac{1}{\mathrm{IF}}(\Delta\hat{\theta}^{(\text{ideal classifier})})^2$

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

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 $\frac{1}{w_i}\frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i}\frac{\partial z_i}{\partial \theta}$
- Both w_i and z_i can be written as $f = \epsilon s + b = \frac{\epsilon s}{\rho} \rightarrow \frac{\partial f}{\partial \theta} = \epsilon \frac{\partial s}{\partial \theta} \rightarrow \frac{1}{f}\frac{\partial f}{\partial \theta} = \rho \frac{1}{s}\frac{\partial s}{\partial \theta}$
- In summary: **try to partition the data into bins of equal $\rho_i \frac{1}{s_i}\frac{\partial si}{\partial \sigma_s}$**
  - 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”*
    - the BDT normally tries to represent a signal likelihood – i.e. ultimately the real ρ_i