设定触发或过滤阈值的一个方法

臧石磊
slz2008@gmail.com

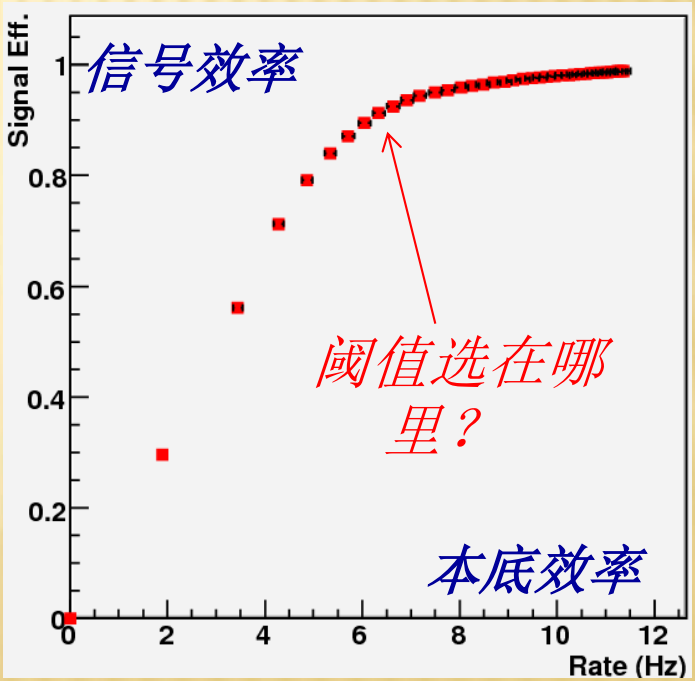
南京大学

粒子物理计算软件与技术研讨会
山东·威海
2013年7月4日

优化触发/过滤阈值

- 问题：在触发(trigger)，过滤(skim or filter)，粒子鉴别（PID），物理分析中，如何设定阈值或cut值？
- 本报告：几年前的结果，基于CMS实验的模拟数据，优化超对称物理分析中的光子触发阈值(HLT)和分析中的光子选择条件。
- 本报告目的：和大家交流，希望大家尝试和检验这个方法。

- 信号效率 vs. 本底效率 → 怎样用一个客观的方法优化阈值？



问题来源

- 在触发、过滤、粒子鉴别、多变量分析和物理分析中，用一些变量来压制本底同时保留信号。原因一：资源限制（电子学、存储空间、计算能力CPU等）；原因二：通过压低本底得到清晰信号。
- 对于原因一，一般通过研究信号效率vs本底效率，用眼观察设定一个合理的阈值。
- 对于原因二，在传统物理分析中，为了得到清晰信号，会采用：
  - 选择条件或cut条件的设置，用和物理量相关的统计量来优化:
    分支比的误差最小；结果的显著性最大; 90% 置信上限最好;
    等等。即:

$$\text{Significance} \equiv \sqrt{2lnQ} \approx 2(\sqrt{N_S + N_B} - \sqrt{N_B}) \approx N_S/\sqrt{N_B} \approx N_S/\sqrt{N_S + N_B}$$

  - 在质量测量中（如top质量测量），*最小化测量质量的误差.*
  - → *缺点：测量结果被人为地调试得更好。测量结果不全是盲（blind）分析。结果有人为倾向性。*

信息理论(I)

- 1948年香农(C. Shannon)提出信息理论
- N 个数（事例数）：信息量: $log_2 N$.
- 信号和本底的信息量: $log(N_S)$, $log(N_B)$
- 一个cut带来的信号效率 ε 和本底效率 b
- Cut之后的信息量: $log(N_S\,\varepsilon)$, $log(N_B\,b)$
- 信息量的减少: $-log(\varepsilon)$, $-log(b)$

✓ *信号和本底信息量减少的比值:* $log(\varepsilon)/\,log(b)$

- 假设两个cut条件减少的信息量相同，即

  $-log(\varepsilon)-log(b) = -log(\varepsilon')-log(b')$,

  如果$log(\varepsilon)/\,log(b) < log(\varepsilon')/\,log(b')$ → 那么*cut* 比*cut'*好
- $log(\varepsilon)/\,log(b)$*越小越好*
- $log(\varepsilon)/\,log(b) < a$ → $\varepsilon > b^a$ $(0< \varepsilon, b, a \leq 1)$.

  → *统计量* $log(\varepsilon)/\,log(b)$ *做为优化标准*

信息理论(II)

- N 是信使数量，物理结果是N个信使携带的消息的意义。
- 对于分支比，信使数量就是信使携带的消息的意义。
- 对于宽度、质量、极化等等物理量，消息的意义不仅仅是信使数量，还有信使所携带的其他信息（例如：动量、能量、角度、顶点等等）。

➢ *优点：* **盲分析！**

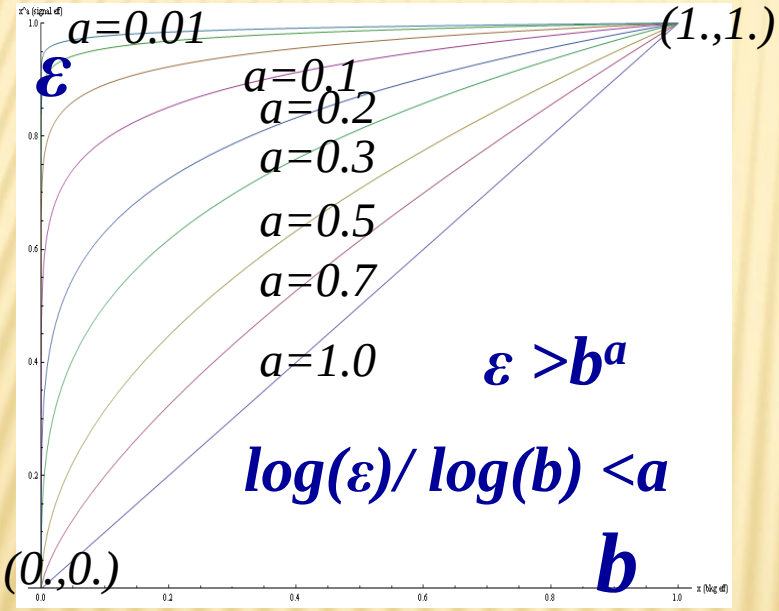
log(ε)/ log(b) 只涉及信息量，不涉及消息的意义。

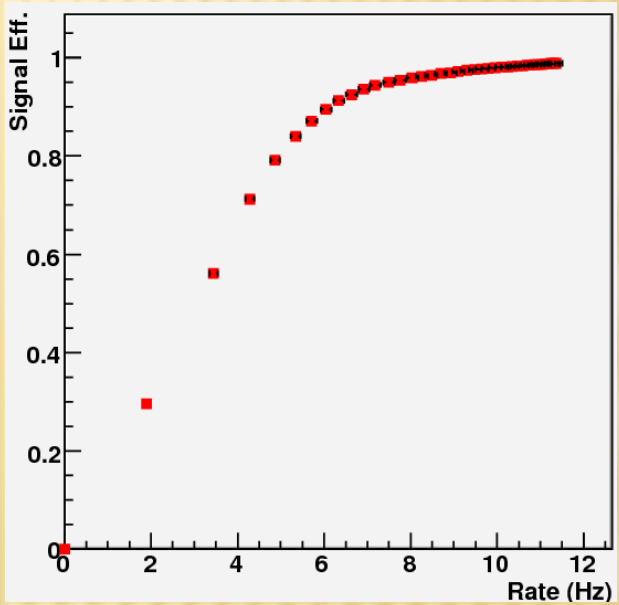
➢ *注意:* 在其他cut条件都使用的情况下，用 *log(ε)/ log(b)*优化一个cut条件。

- 信息量的定义满足如下属性： $\log_2(xy)=\log_2(x)+\log_2(y)$,并且是满足这个属性的唯一解。
- 这个属性对***log(ε)/ log(b)***方法非常重要，假设有两个cut条件，$\log(\varepsilon_1\,\varepsilon_2)/\log(b_1b_2)=(\log(\varepsilon_1)+\log(\varepsilon_2))/(\log(b_1)+\log(b_2))$ → 两个cut分开优化和合起来优化产生一致的结果。

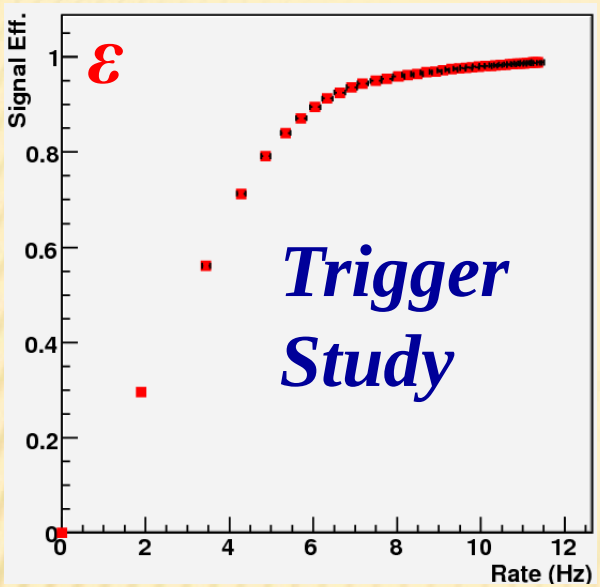
log(ε)/log(b)方法

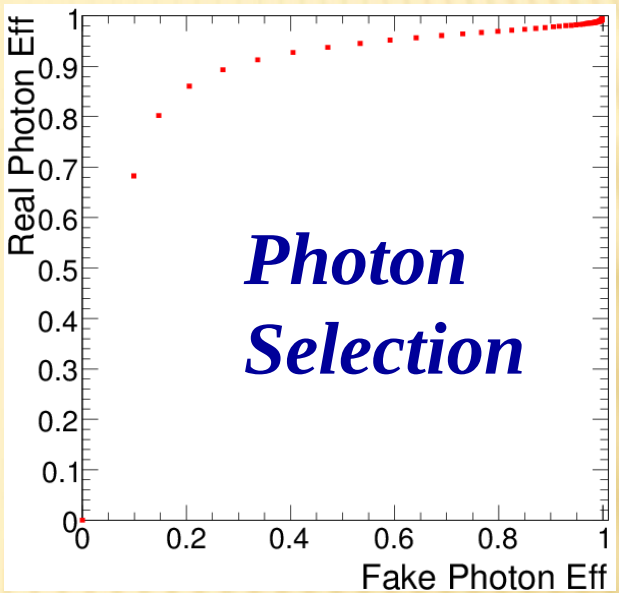
- 信号和本底信息量减少的比值log(ε)/ log(b) 做为优化标准
- log(ε)/ log(b)越小, 越好. → 在资源限制下 (electronic readout, storage, CPU, etc.).

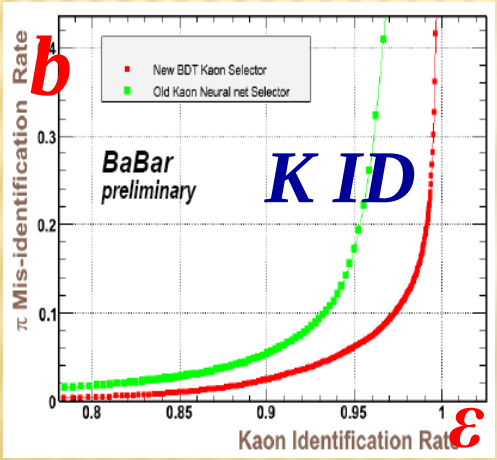


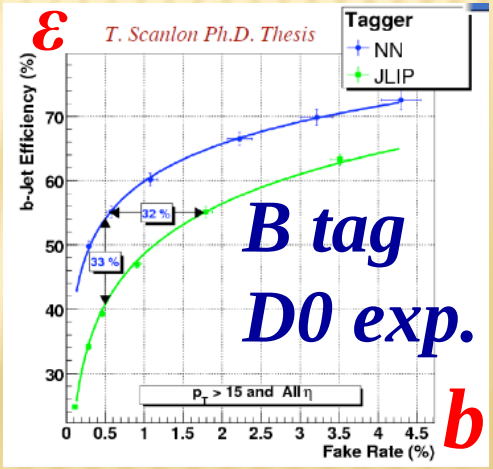


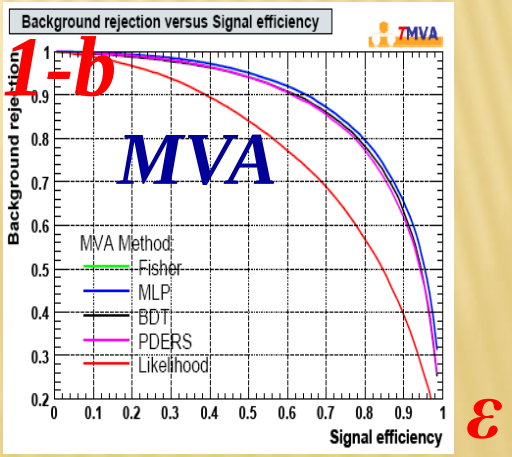
- log(ε)/ log(b) <a → ε > b^a (0< ε, b, a ≤1).


Signal Eff.
ε
Trigger
Study
b
Rate (Hz)

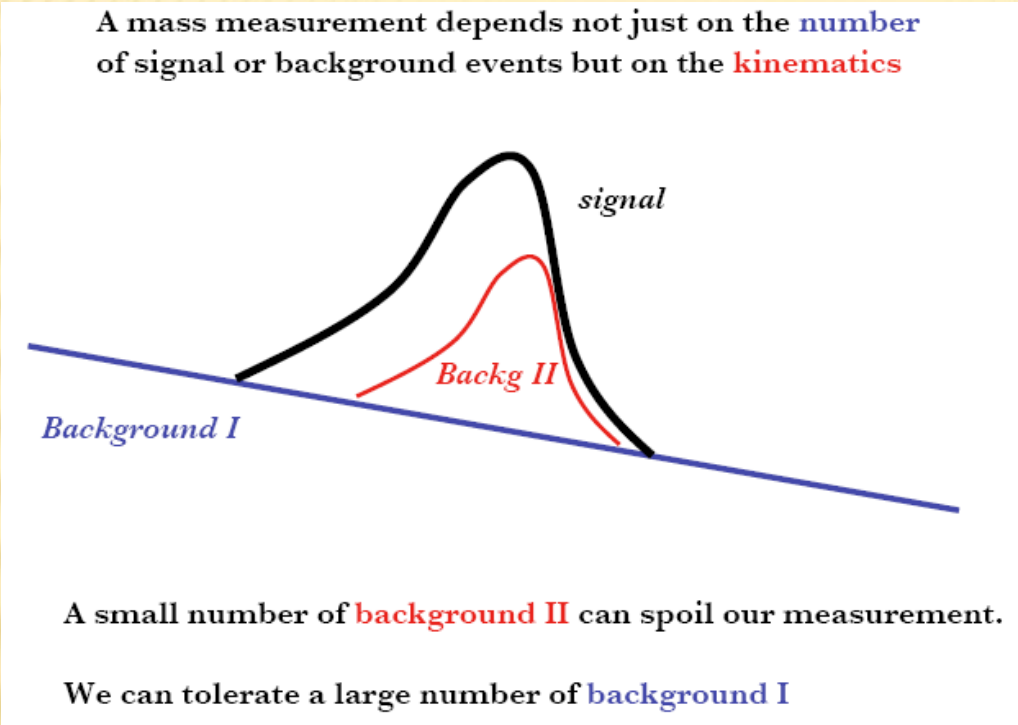

ε
Real Photon Eff
Photon
Selection
b
Fake Photon Eff

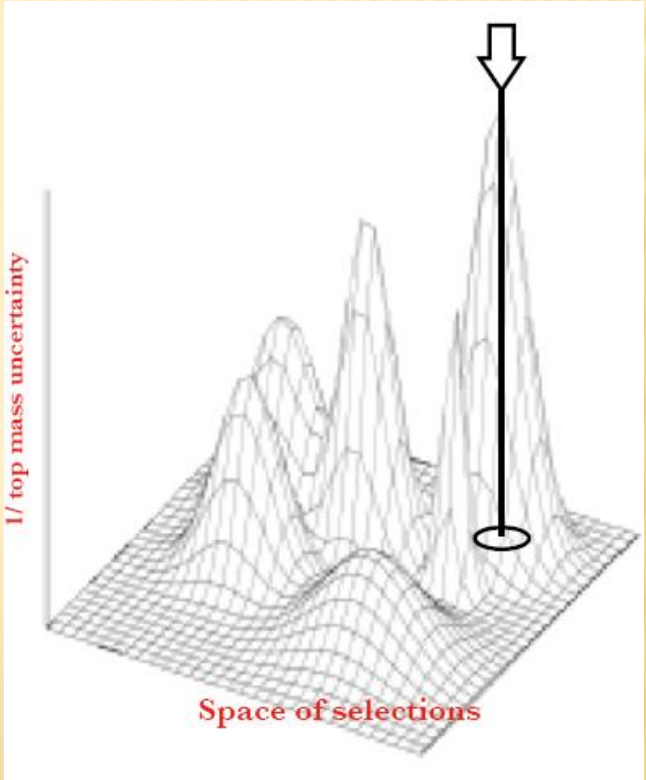

b
π Mis-identification Rate
New BDT Kaon Selector
Old Kaon Neural net Selector
BaBar
preliminary
K ID
Kaon Identification Rate
ε


ε
T. Scanlon Ph.D. Thesis
Tagger
NN
JLIP
b-Jet Efficiency (%)
B tag
D0 exp.
32 %
33 %
p_T > 15 and All η
Fake Rate (%)
b


Background rejection versus Signal efficiency
TMVA
1-b
MVA
MVA Method:
Fisher
MLP
BDT
PDERS
Likelihood
Background rejection
Signal efficiency
ε

top质量测量中可能人为造成测量误差偏小





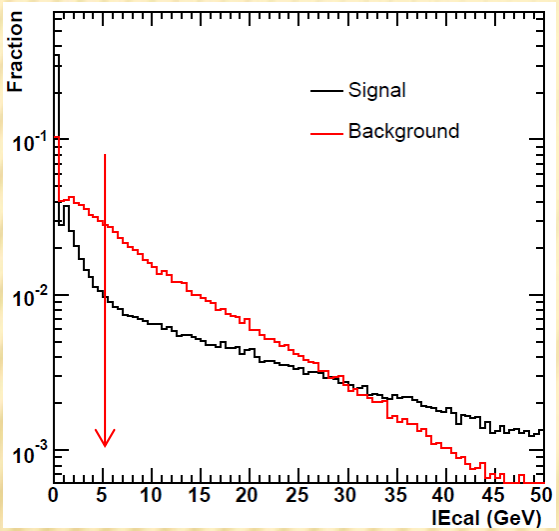
- *将测量误差作为优化标准，当然测量误差人为造成偏小。*
- *在对测量结果满意的限制下（比如清晰地看到top信号，对测量误差满意），将log(ε)/ log(b)作为优化标准，就不会有上面的问题。*
- *有待检验。*

优化的过程

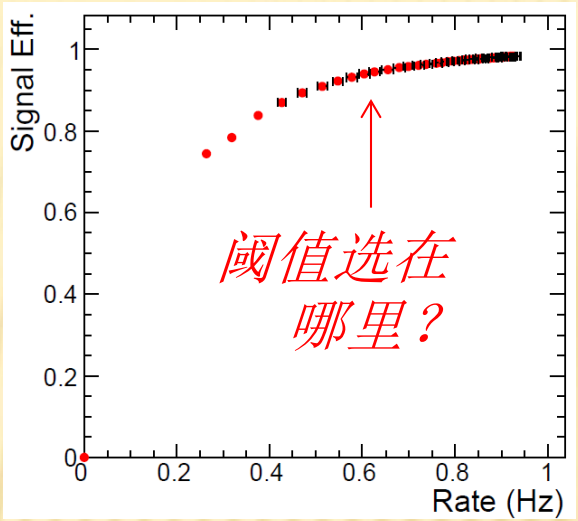
- 对每一个cut， $a_i = log(\varepsilon_i) / log(b_i)$ ，得到一个最小值a_i^{min}；假设 $a_1^{min} < a_2^{min} < a_3^{min} < \ldots < a_k^{min}$
- 如果对效率不满意(太小)，放松a_k，去掉 a_k，放松a_{k-1}，去掉 a_{k-1}，…，直到我们对效率满意
- 如果对本底不满意 (太多)，将a_1 从 a_1^{min} 变化到a_2^{min}，然后将$a_1 = a_2$ 从a_2^{min} 变化到a_3^{min}，…，直到我们对本底满意.

- → *对于物理分析，用上面的过程优化选择条件，直到对MC预测的物理结果满意*

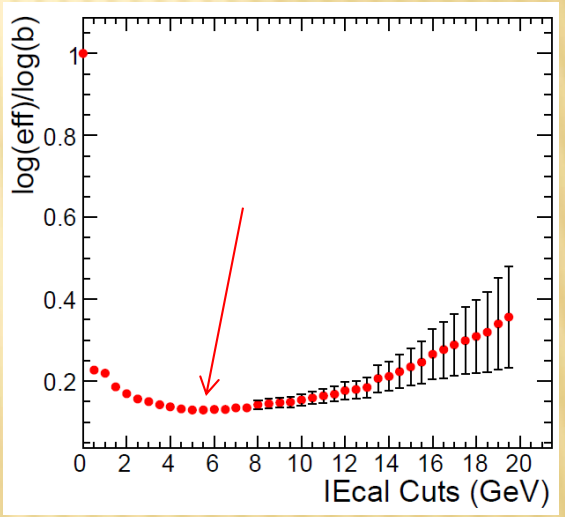
研究一：触发阈值



信号和本底的
触发变量分布



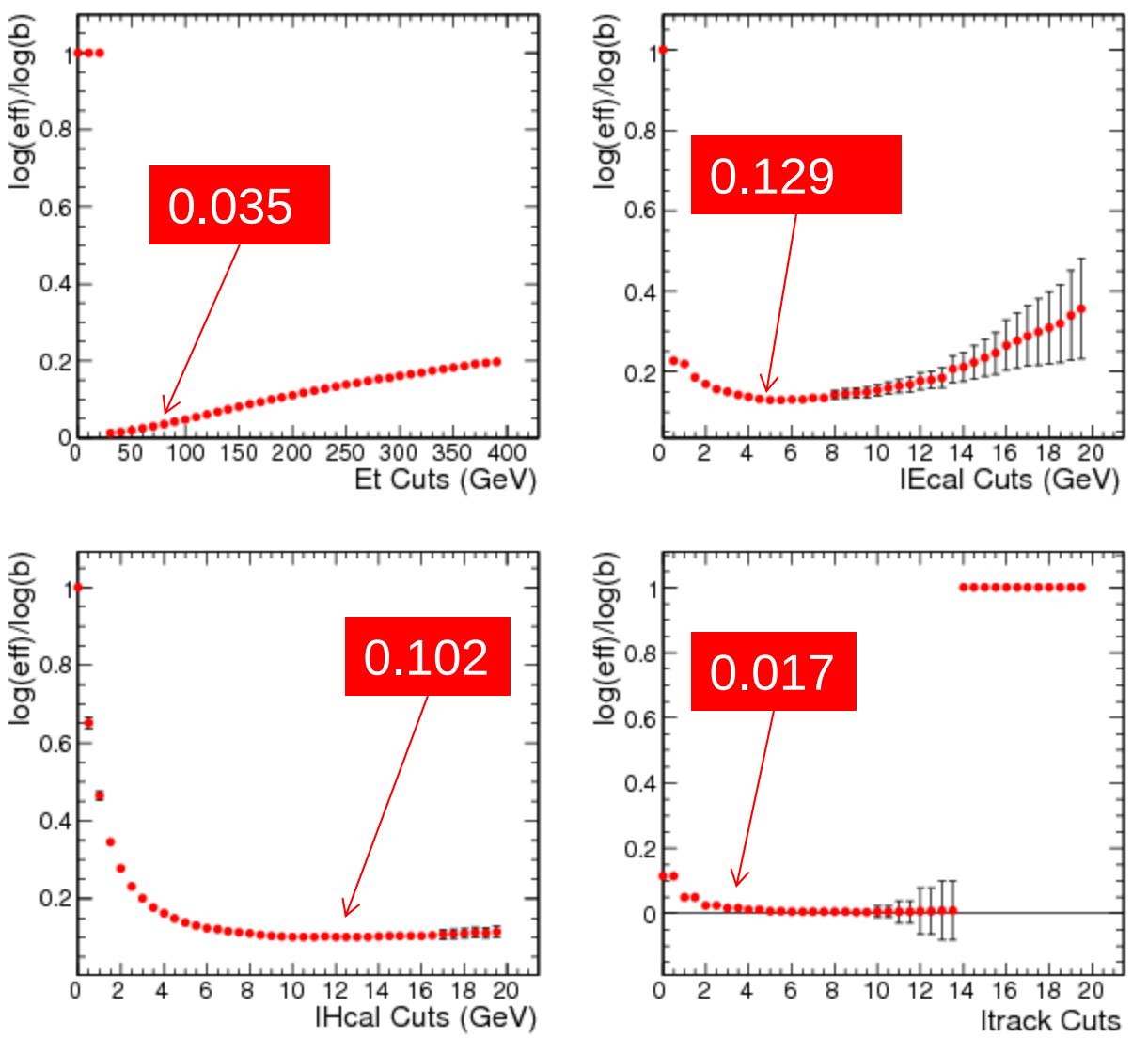
信号效率 vs.
本底效率



log(ε)/ log(b) vs.
触发阈值

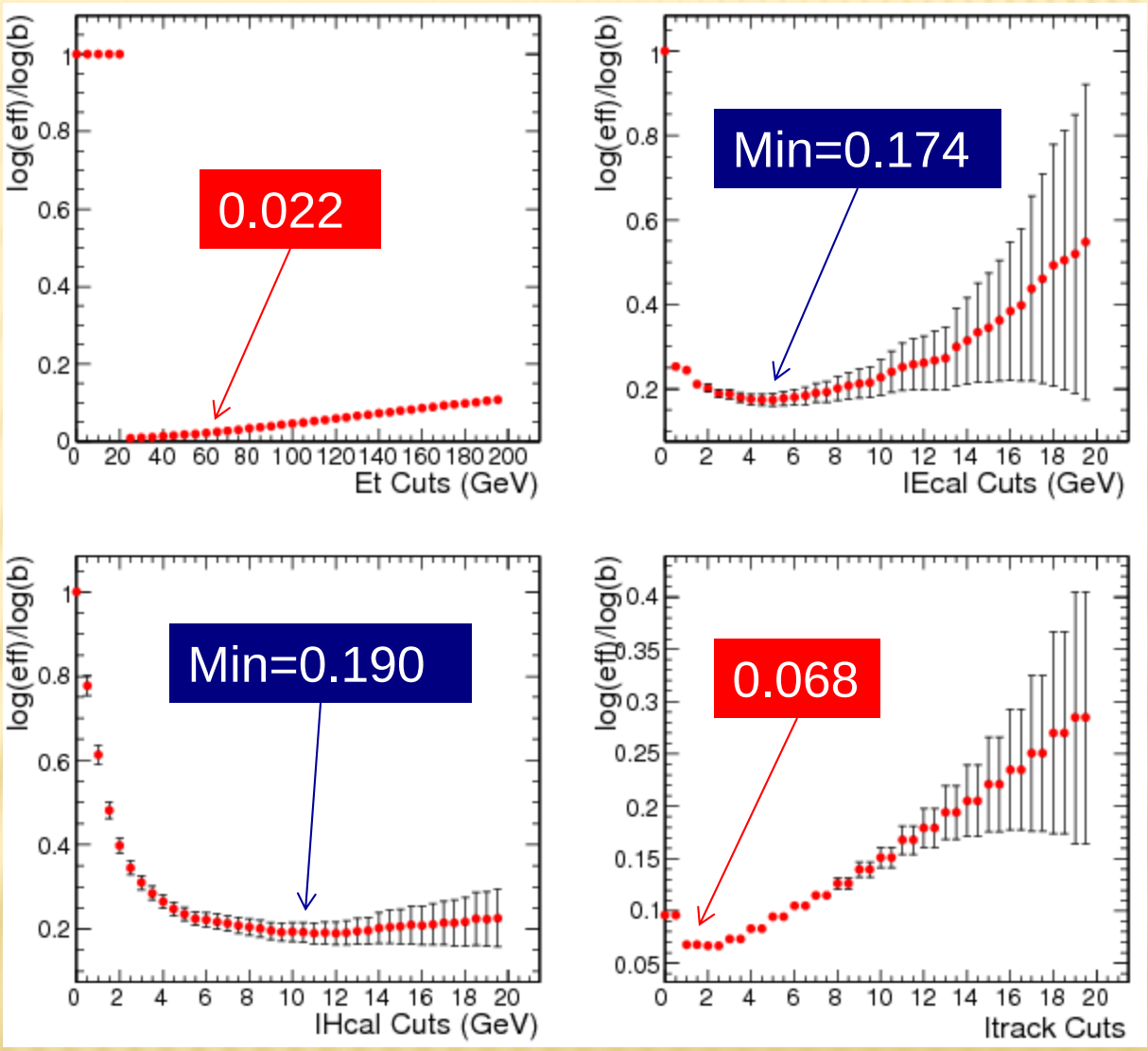
log(ε)/log(b) vs. Cuts (原始EM-High-Et)

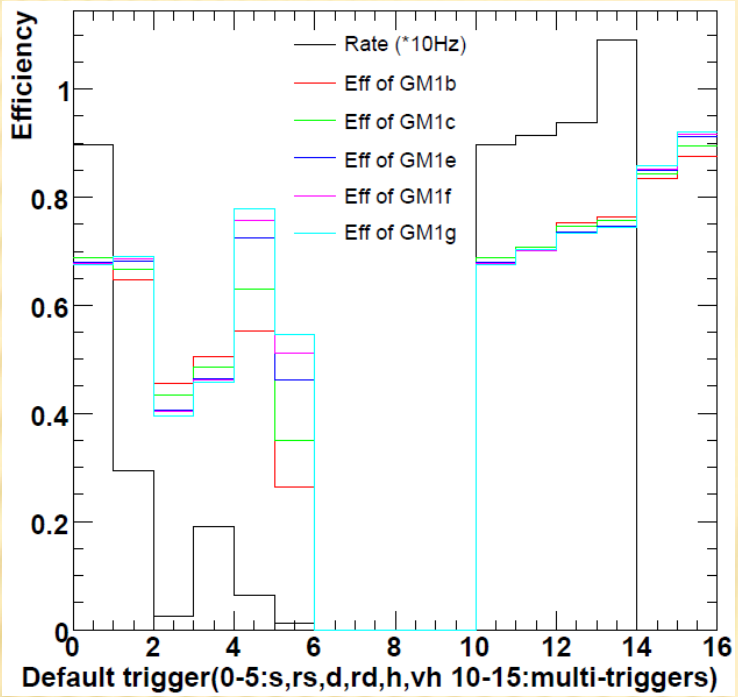
- ***Et>80***
  Iecal<5
  Ihcal<12
  Itrack<4
- ***ITrack比IEcal和IHcal好***

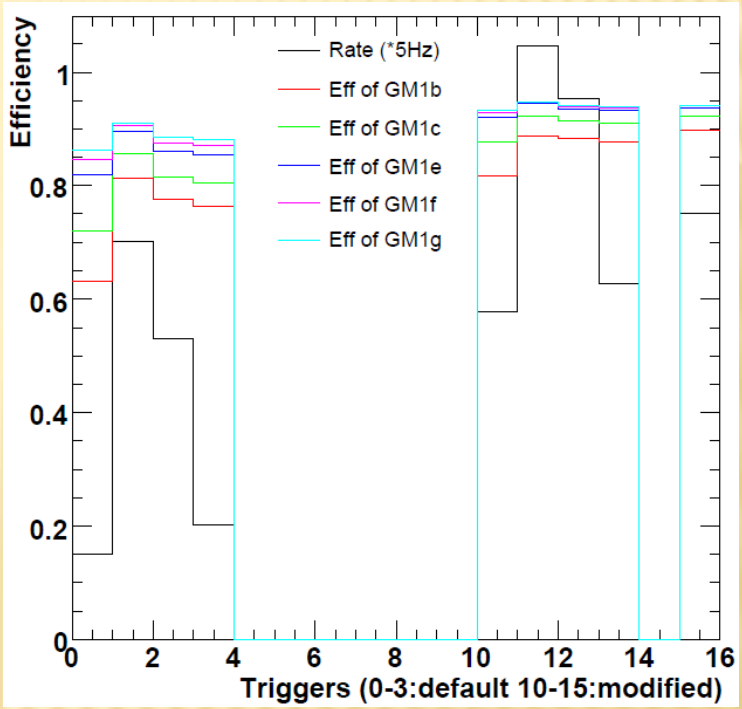


log(ε)/log(b) vs. Cuts (优化后EM-High-Et)

- **Et>60; ITrack<2**
- ***只有当 Itrack<5 时它才比较好***



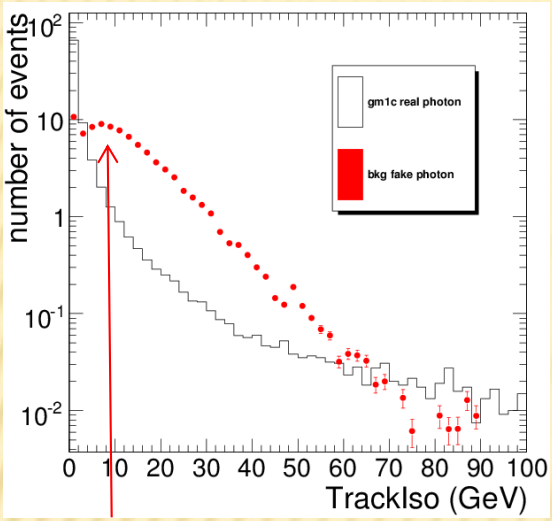




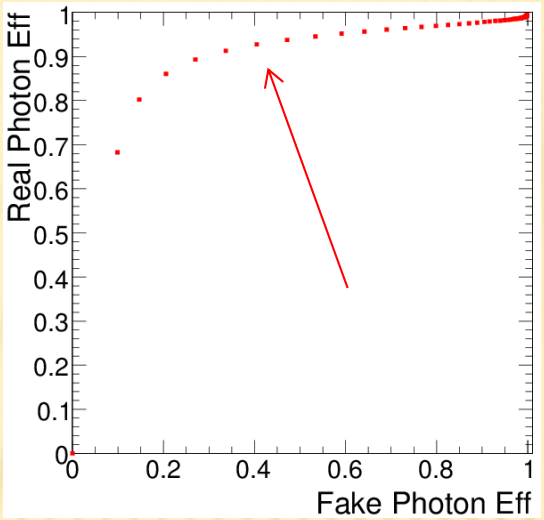
优化前各个触发path的信号效率（彩色直方图）和本底rate率（黑色直方图）

优化后

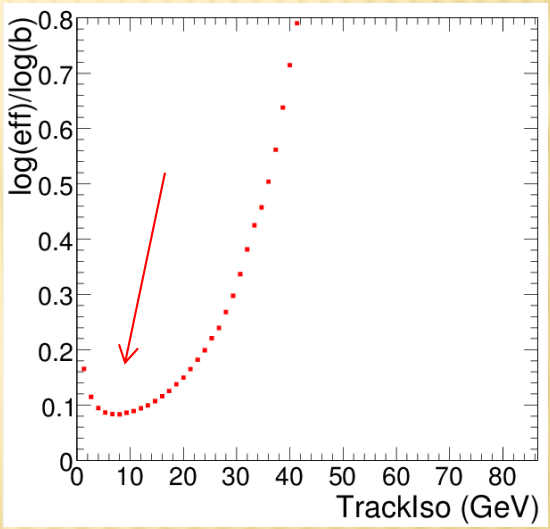
研究二：光子选择



信号和本底的Cut变量分布



信号效率 vs. 本底效率

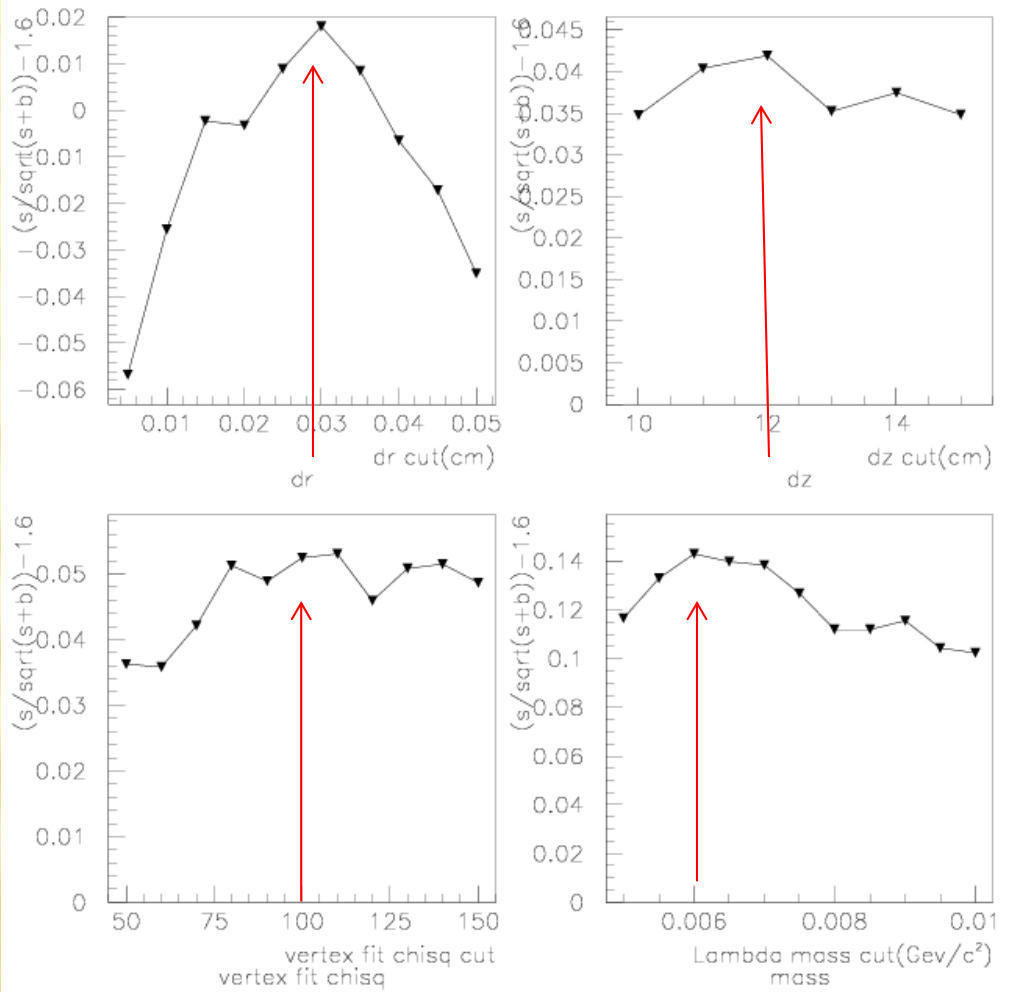


log(ε)/ log(b) vs. Cut条件

- 优化出的选择条件，被物理分析采用：

1) sigmaEtaEta < 0.01
2) HadOverEM < 0.05
3) EcalIso-0.15Pt < 3GeV
4) TrackIso < 8GeV （上面的图）

$N_S/\sqrt{N_S+N_B}$ 优化选择条件（tight cuts）



结论

- 提出了一个新的方法，使用信号和本底信息量减少的比值 log(ε)/ log(b) 做为优化标准，来优化触发、过滤、粒子鉴别、物理分析、多变量分析等等中的阈值设定。
- 方法有信息学和数理逻辑的支撑。
- 方法是客观的，盲分析，减少了人为倾向性。
- 在触发阈值设定和事例选择的研究中被证明是可行的或有用的。
- 在粒子鉴别、物理分析、多变量分析，以及其他领域（比如股市中的收益vs风险），值得尝试应用，预计有所帮助。

谢谢！

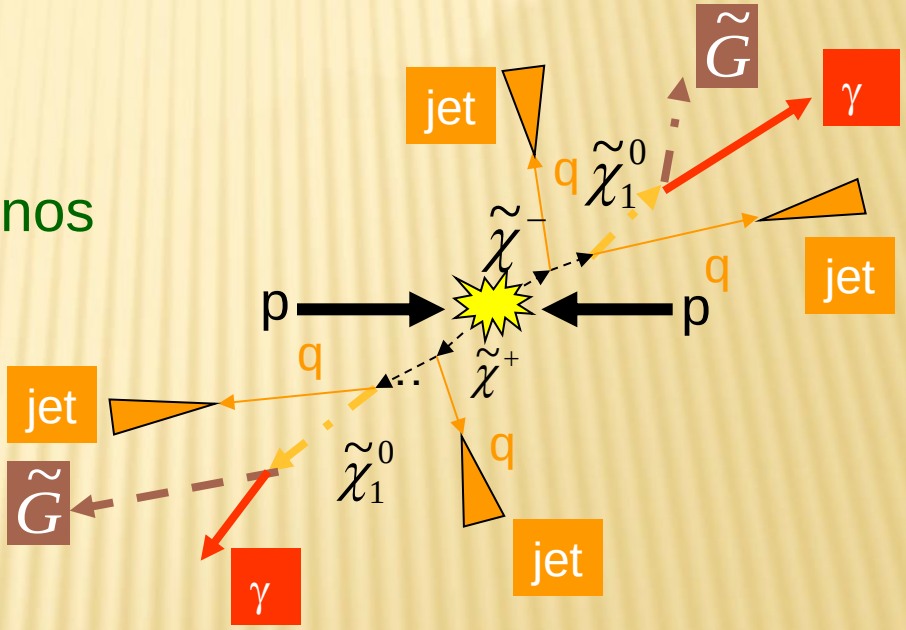
Backup Slides

数据

- 2007 夏季样本, Pythia6，在14TeV模拟产生
- *GMSB 信号：*
  - *Five GMSB samples with different Λ parameter*
- *本底:*
  - *Photon+jets (all pT_hat bins)*
  - *QCD jets (all pT_hat bins)*
  - *W→enu, Z→ee*
- *CMSSW_1_6_7;* 优化到亮度$10^{32}\ cm^{-2}s^{-1}$.

GMSB 信号中的末态光子

- NLSP → LSP + photon $\tilde{\chi}_1^0 \to \tilde{G}\gamma$
- Prompt decay
- Experimental signature
  - high pT photons
  - large MET due to gravitinos
  - multi-jets



Trigger variables for photons

- *L1Match:* Reconstructed super-cluster in the ECAL is required to match L1 energy deposit in some eta and phi windows.
- *Et :* Et of super-cluster in the ECAL is required to exceed a threshold.
- *IEcal:* ECAL isolation, total Et of all clusters with $\Delta R<0.3$ around the photon candidate, excluding those belonging to the super-cluster itself.
- *IHcal:* HCAL isolation, total Et of hadron calorimeter towers with $\Delta R<0.3$ around the photon candidate.
- *ITrack:* Track isolation, number of tracks with Pt>1.5GeV inside a cone $\Delta R<0.3$ of photon candidate.

Number of background events processed for rate estimation

QCD Jets	*N*	QCD Jets	*N*	Photon Jets	*N*	Bkg	*N*
0_15	*295,613*	_470	*88,086*	0_15	*500,000*	Wenu	*205,707*
15_20	*1,255,976*	_600	*55,000*	15_20	*509,825*	Zee	*162,219*
20_30	*2,513,934*	_800	*21,974*	20_30	*606,680*		
30_50	*2,416,441*	_1000	*33,330*	30_50	*510,094*		
50_80	*2,451,439*	_1400	*5,299*	50_80	*169,741*		
_120	*1,161,823*	_1800		_120	*164,000*		
_170	*499,389*	_2200		_170	*69,993*		
_230	*428,888*	_2600		_300	*24,993*		
_300	*172,619*			_500	*15,554*		
_380	*82,998*			_7000	*6,666*		