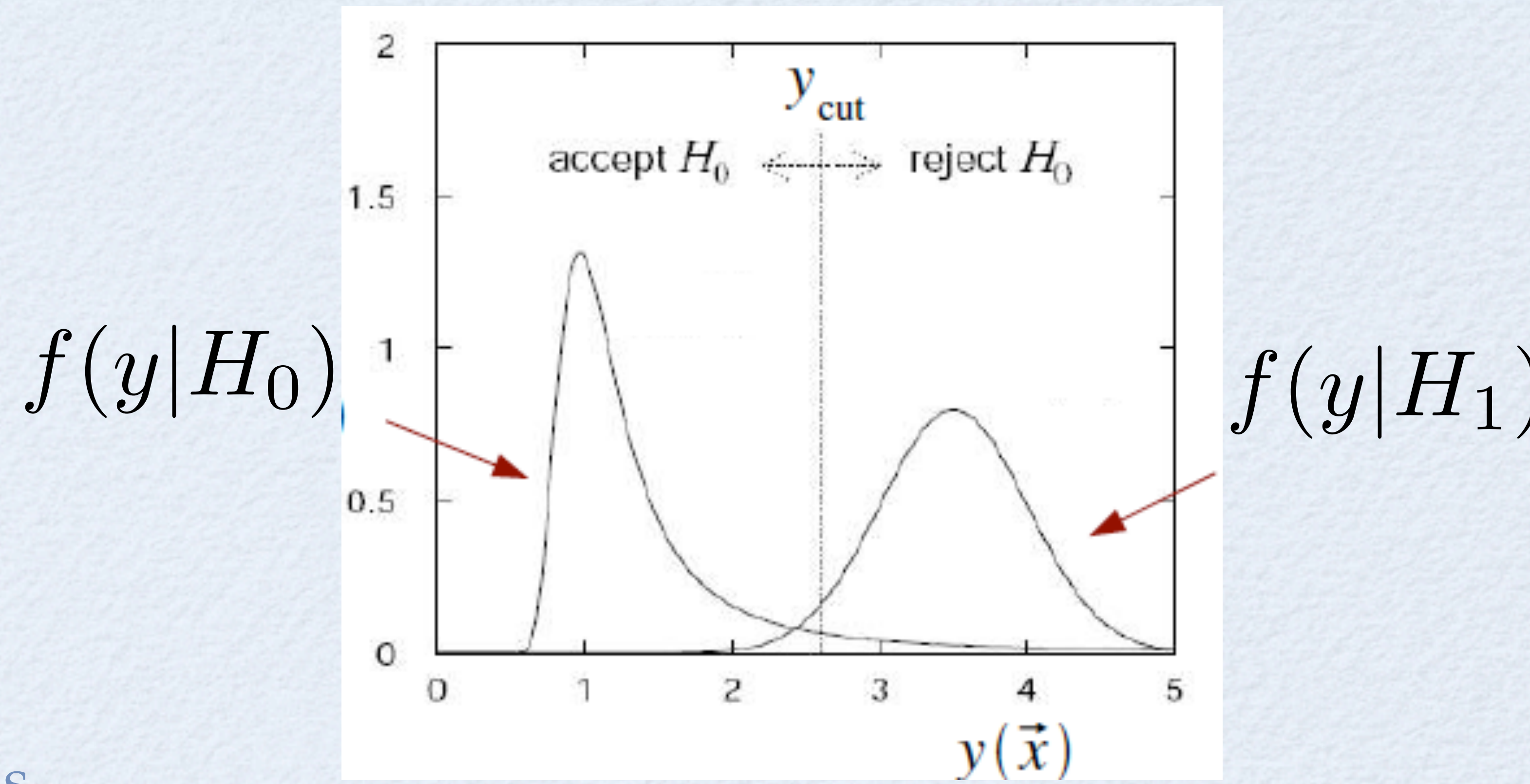
Event Classification Part II

Final Project

- Placed an ntuple on blackboard under the folder Final Project
- Simulated LHC data - your job is to convince me for evidence of particle production (of both known and unknown particles!)
- See instructions for detail - due one week from Tuesday

Discriminant

- Say we have two hypothesis for what an event is:
    - background - H0
    - signal - H1
- Some variable or discriminant which we can classify the event as more likely signal or background
- This can be a function of many variables - but now the question is how to optimally separate the two



R_0 (accept H_0)

R_1 (reject H_0)

DEFINITIONS

Significance Level:
Probability to reject H0 when it is true...

$$\alpha = \int_{R_1} f(y|H_0)dy$$

Power:
Probability to accept H0 when H1 is true...

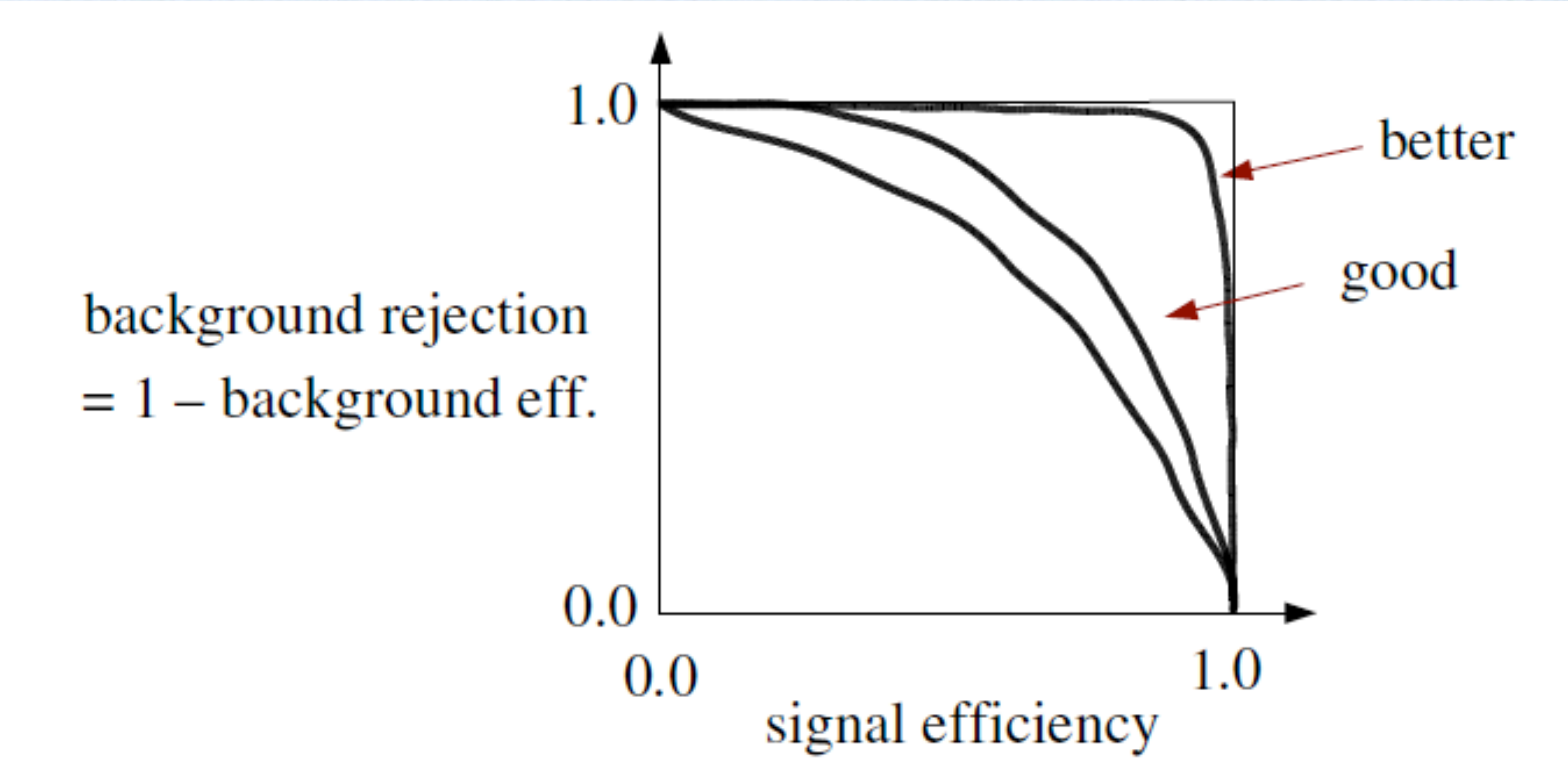
$$\beta = \int_{R_0} f(y|H_1)dy$$

Purity/ Misclassification Rate

- Consider the probability that an event is correctly assigned
- We can use Bayes' theorem to calculate this
- Note that purity depends on prior probability (i.e. s and b cross-sections)

$$P(s|x \in R_1) = \frac{P(x \in R_1|s)P(s)}{P(x \in R_1|s)P(s) + P(x \in R_0|b)P(b)}$$

Choosing one..



- Characterize the performance by plotting the efficiency versus rejection. Best to be in the upper right hand corner!
- Can use area under curve as a measure of quality

Neyman-Person Lemma

- The Neyman-Pearson lemma states: to obtain the highest background rejection for a given signal efficiency (highest power for a given significance level), choose the acceptance region for signal choosing c for a given efficiency

$$\frac{p(\vec{x}|s)}{p(\vec{x}|b)} > c$$

- Equivalently , it states that the optimal discriminant is given by the ratio of probabilities or the likelihood ratio

$$y(\vec{x}) = \frac{p(\vec{x}|s)}{p(\vec{x}|b)}$$

Prior odds

Posterior odds

$$\frac{p(s|\vec{x})}{p(b|\vec{x})} = \frac{p(\vec{x}|s)p(s)}{p(\vec{x}|b)p(b)}$$

Likelihood Ratio

Since we can't change the prior odds making a cut on the likelihood ratio is equivalent to maximizing the posterior odds of the event classification

So then what

- The problem is that we usually don't have explicit formulae for the pdfs p(x | s), p(x | b), so for a given x we can't evaluate the likelihood ratio.

- Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate $\vec{x} \sim p(\vec{x}|\mathrm{s}) \longrightarrow \vec{x}_1, \ldots, \vec{x}_{N_s}$

generate $\vec{x} \sim p(\vec{x}|\mathrm{b}) \longrightarrow \vec{x}_1, \ldots, \vec{x}_{N_b}$

"training data"
events of known type

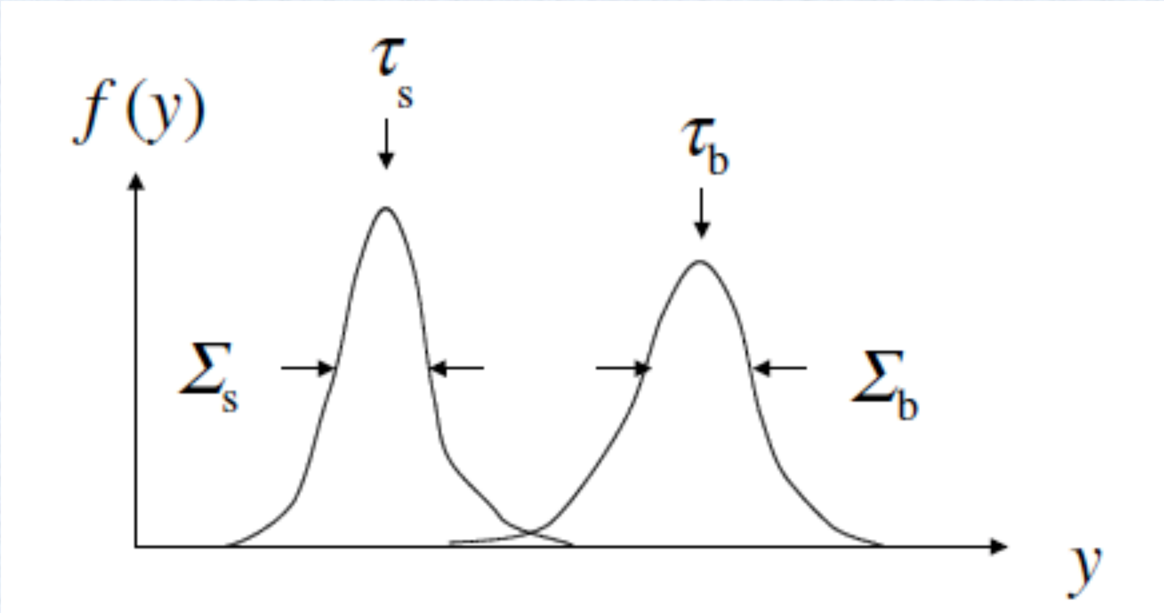
Compromise

- Simulate monte carlo events and get an approximate pdf
- Or form a test statistic from those approximate pdfs

Fisher Discriminant

$$y(\vec{x}) = \sum_{i=1}^{n} w_i x_i = \vec{w}^T \vec{x}$$

- Form a test statistic which is a weighted sum of the inputs
- Want large separation between mean value and small variance
- So maximize:



$$J(\vec{w}) = \frac{(\tau_s - \tau_b)^2}{\Sigma_s^2 + \Sigma_b^2}$$

CALCULATION

$$(\mu_k)_i = \int x_i \, p(\vec{x}|H_k)\, d\vec{x}$$

mean, covariance of ***x***

$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j \, p(\vec{x}|H_k)\, d\vec{x}$$

- Making the optimal selection reduces to the problem of finding the optimal weights of the variables

$$\tau_k = \int y(\vec{x})\, p(\vec{x}|H_k)\, d\vec{x} = \vec{w}^T \vec{\mu}_k$$

$$\Sigma_k^2 = \int (y(\vec{x}) - \tau_k)^s \, p(\vec{x}|H_k)\, d\vec{x} = \vec{w}^T V_k \vec{w}$$

CONTINUED...

$$(\tau_0 - \tau_1)^2 = \sum_{i,j=1}^{n} w_i w_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j$$

'between' classes

$$= \sum_{i,j=1}^{n} w_i w_j B_{ij} = \vec{w}^T B \vec{w}$$

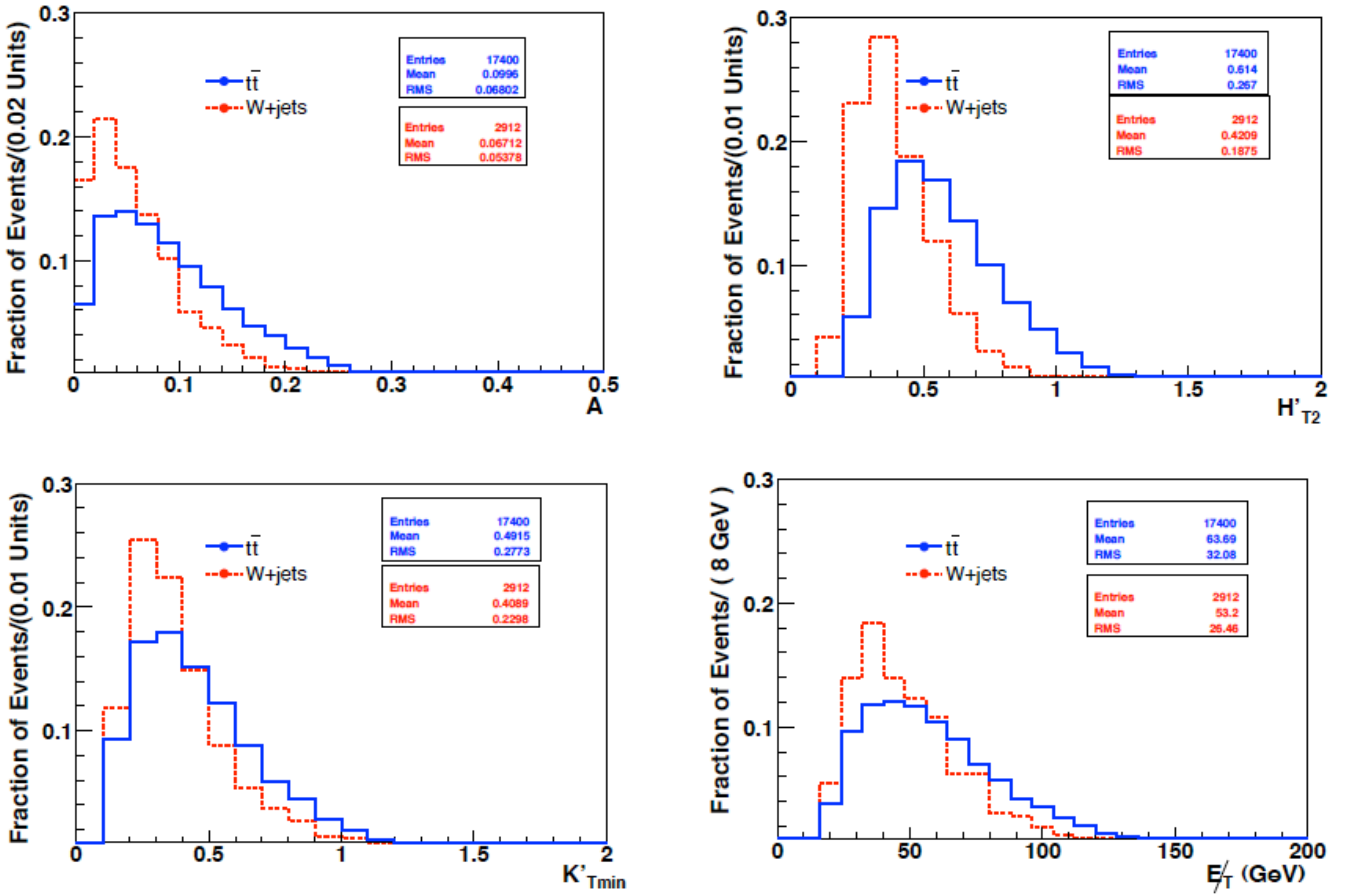
and the denominator is

'within' classes

$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} w_i w_j (V_0 + V_1)_{ij} = \vec{w}^T W \vec{w}$$

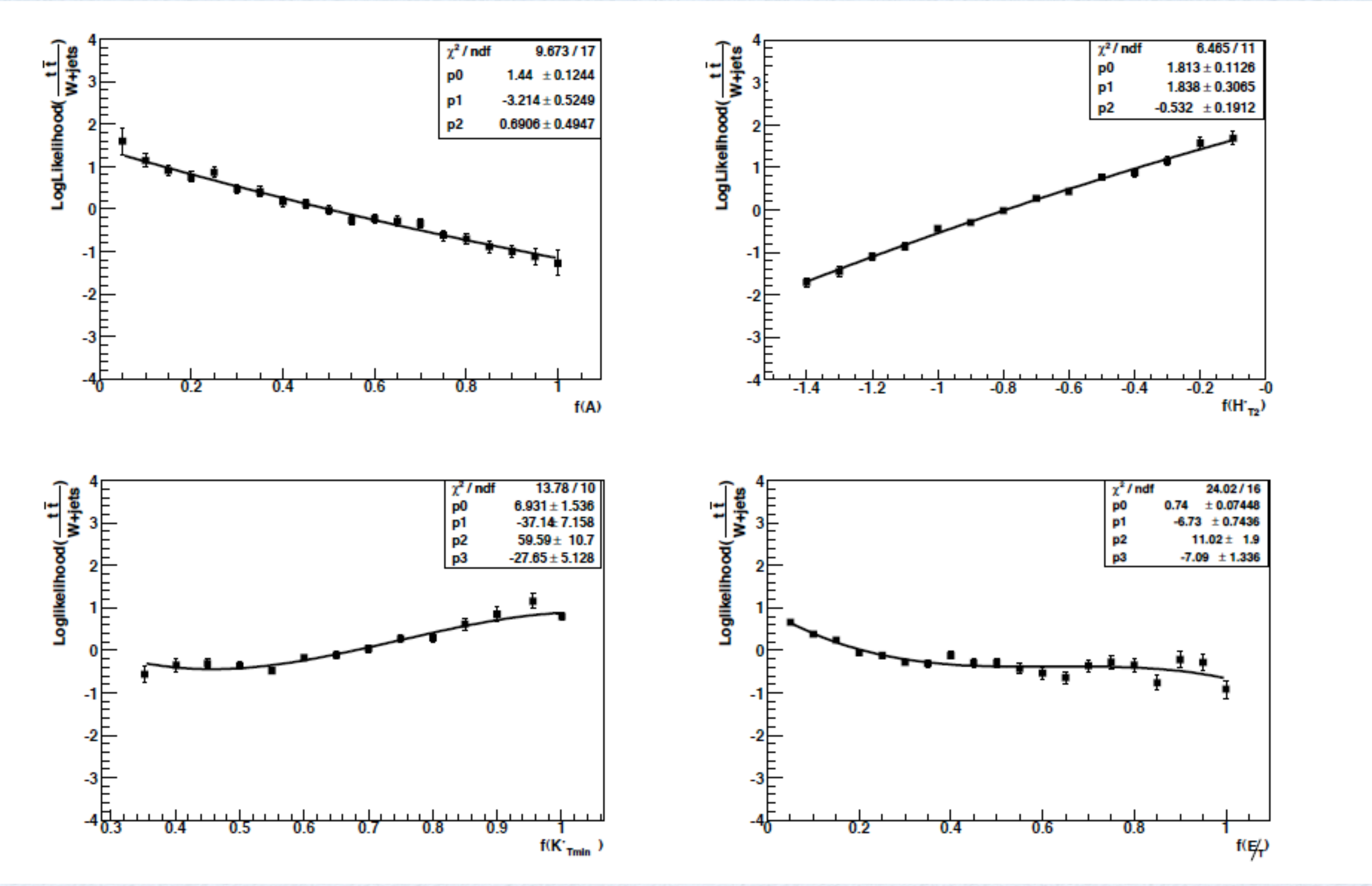
→ maximize $$J(\vec{w}) = \frac{\vec{w}^T B \vec{w}}{\vec{w}^T W \vec{w}} = \frac{\text{separation between classes}}{\text{separation within classes}}$$

EXAMPLE

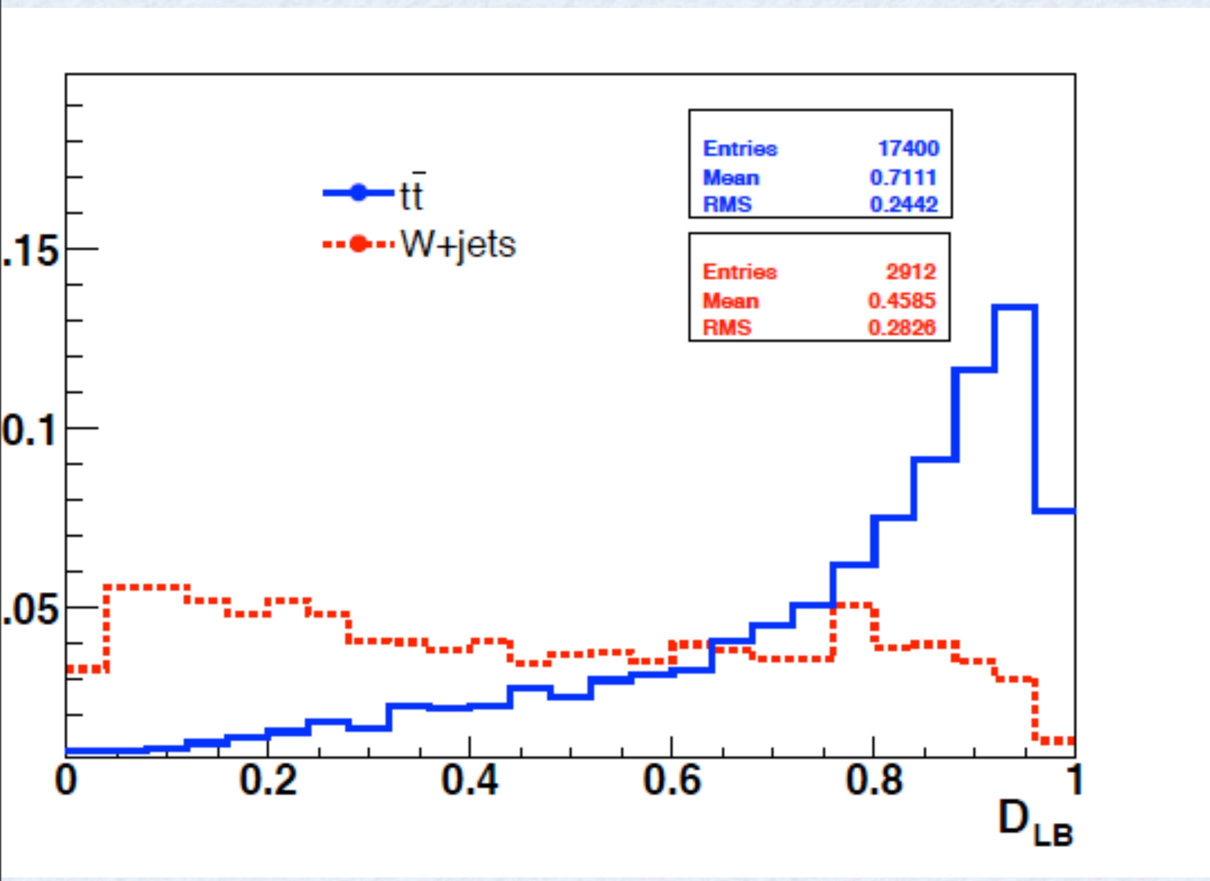


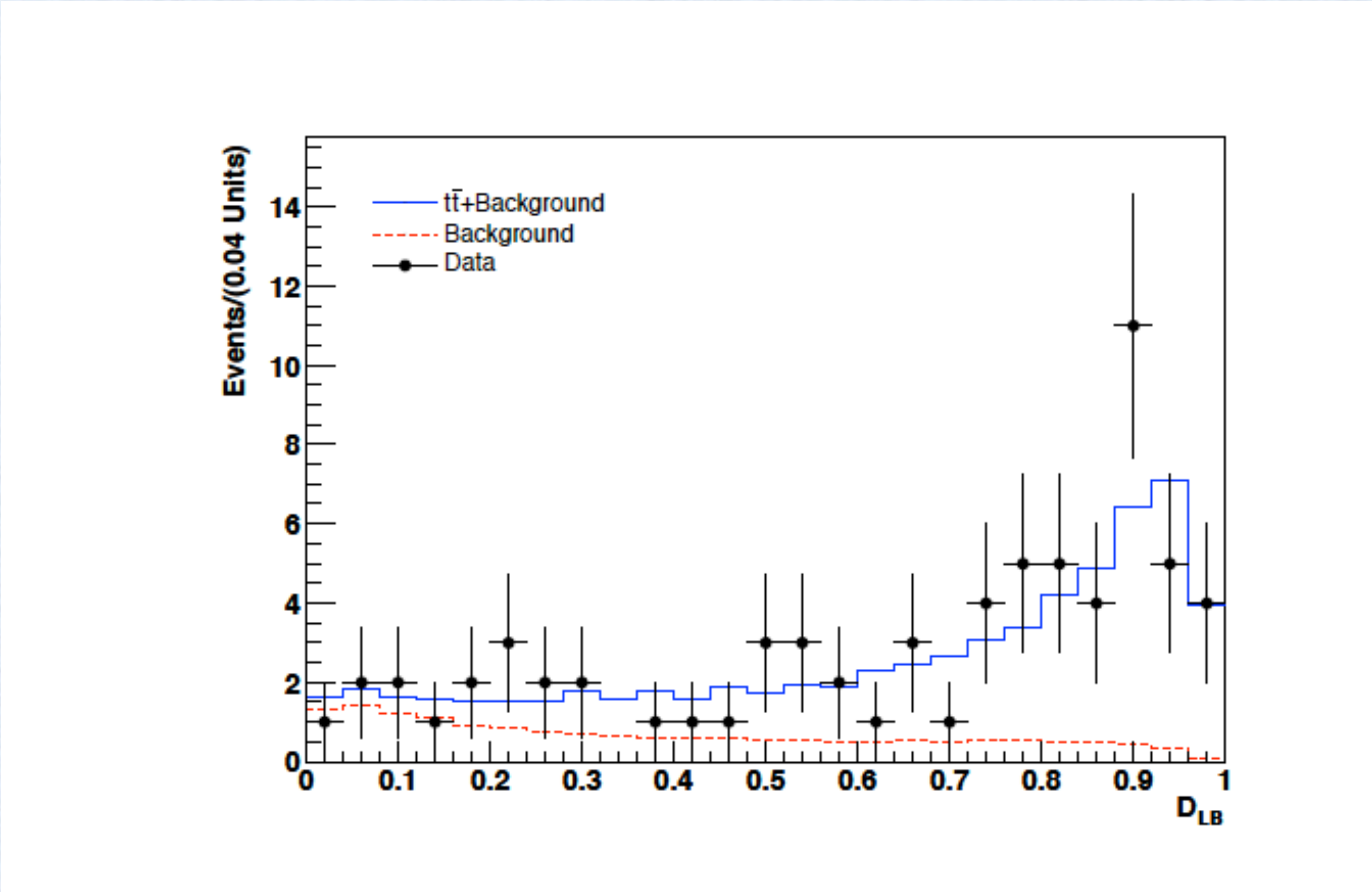
Fraction of Events/(0.02 Units)
A
Entries 17400
Mean 0.0996
RMS 0.06802
Entries 2912
Mean 0.06712
RMS 0.05378
tt̄
W+jets
Fraction of Events/(0.01 Units)
H'T2
Entries 17400
Mean 0.614
RMS 0.267
Entries 2912
Mean 0.4209
RMS 0.1875
Fraction of Events/(0.01 Units)
K'Tmin
Entries 17400
Mean 0.4915
RMS 0.2773
Entries 2912
Mean 0.4089
RMS 0.2298
Fraction of Events/ (8 GeV)
E̸T (GeV)
Entries 17400
Mean 63.69
RMS 32.08
Entries 2912
Mean 53.2
RMS 26.46

LOG LIKELIHOOD



FINALLY





$$L^{tb} = \sum_{i=1}^{4} L_i^{tb}(f_i)$$

$$\mathcal{D}_{LB} = \frac{1}{1 + \mathcal{P}e^{-L^{tb}}}$$