Statistical methods (application to the single top search in D0)

Outline :

- Strategy for analysis
- Discrimination tools between Signal/Background
  - ANN
  - Decision Tree
  - Likelihood
- Limits setting : A bayesian method
- Conclusion

Analysis strategy

Depending on the physic goal :

- Measurement of an observable (e.g. Top mass with matrix element method)

-Search for process (New Physics, Single Top evidence)

BUT

A crucial need to provide efficient, stable statistical methods to discriminate S/B (ex: σ(single top)~3pb, bckg (Wjets+multijet+tt+fake)~O(100pb))

In the case of non-evidence (small S / large B):

necessity to use performant method in order to extract limits on S:

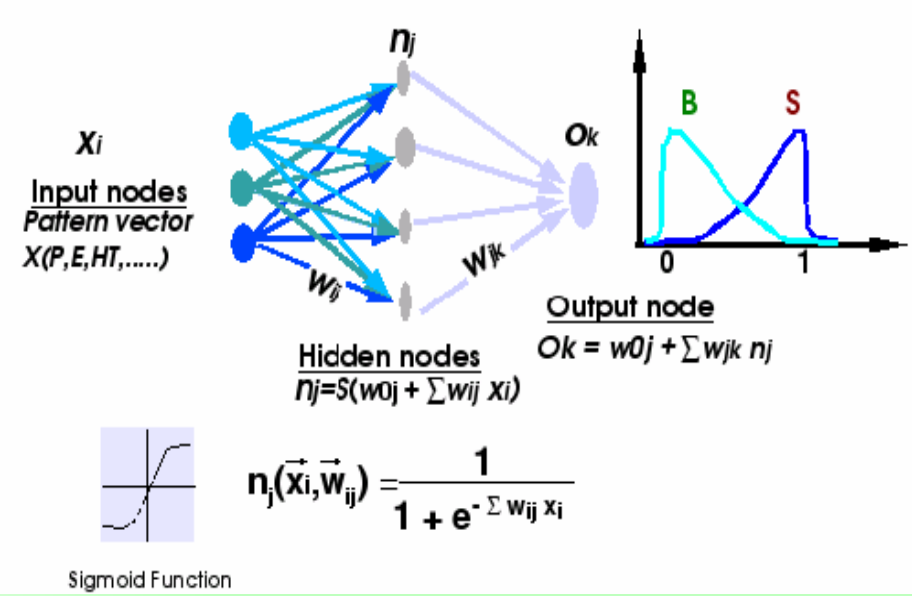
- use of the full information (counting+distribution shapes)
- statistical and systematic errors properly treated

Neural Networks (NN)

NN : An attempt to simulate the cognition apparatus at the most primitive level, the neuron.

Principle:

A set of variables is chosen that have discriminating power between S and B. The NN scans the phase-space of the variables in the sample by making many pairs of S and B events and derives function that separate the 2 event types (peaks around 1(0) for S(B)).



Composed of nodes connected by links (weights) to propagate the activation function designed to be active(1) (right inputs) or inactive(0) (wrong inputs).

Recursivity structure is the key to a NN functionnal flexibility →

A NN has to be trained (learning phase).

NN, cont'd

Learning in theory:

Consists of minimizing an Error Function defined in the space of the NN parameters (weights, threshold):

$$Error = \sum_{i=1}^{Nevt} W_i^2 (O_i^{observed} - O_i^{desired})^2$$

$O^{desired}$ = 1 (0) for S (B)

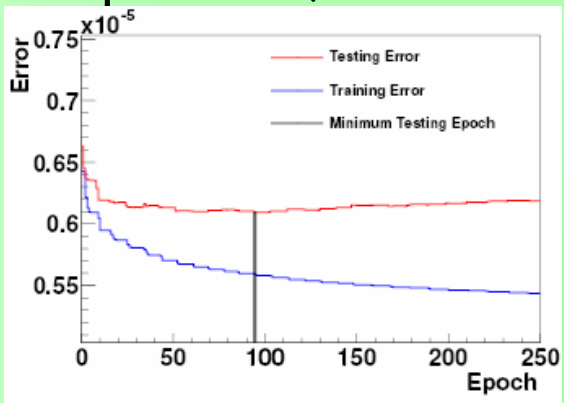
NN parameters are slightly adjusted to minimize E by using dedicated and tunable algorithm. Through successive epoch (iterations), NN is trained to separate the phase space of S and B events.

Learning in pratice:

Parameters to optimize:

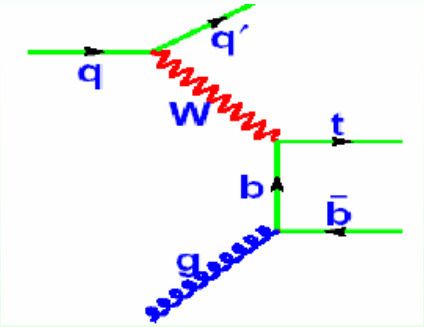
-Number and choice of input variables (good S/B separation, uncorrelation a.m.a.p., good agreement data/MC

- Number of hidden nodes and duration of training : Avoid overtraining by using a testing epoch →



Application of NN: The singletop example

Example of the single top t-channel :

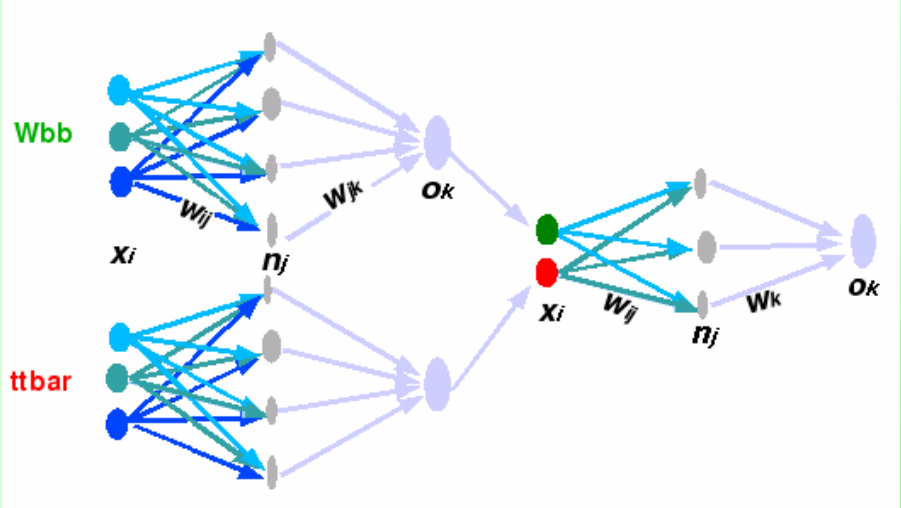


Discriminate against Wjj (j=light quark), Wbb, tt→lj,ll , multijet events.

+ 2 b-tagging schemes (=1 tag, ≥ 2 tags) + e and μ channels

→ 20 independant trained/optimzed NNs. D0 uses MLPfit NN (http://root.cern.ch/root/html/src/TMultiLayerPerceptron.cxx.html)

Final step: combine the discriminative power of the 5 NN into a 'super' NN (no gain in separation but summarize the output in a single distribution)



Application of NN, cont'd

Network optimization:

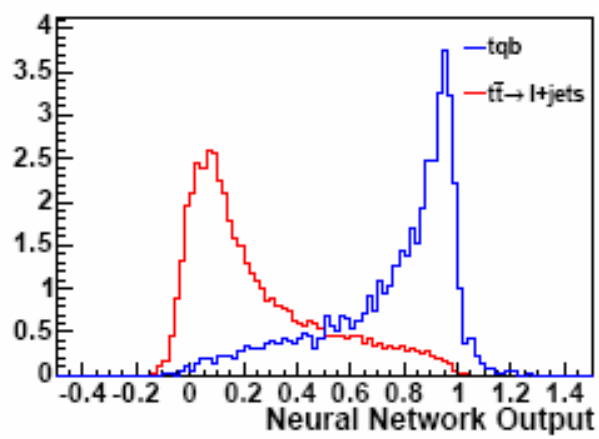
- Optimize the choice of input variables (using ~ 11 out of 40 variables).

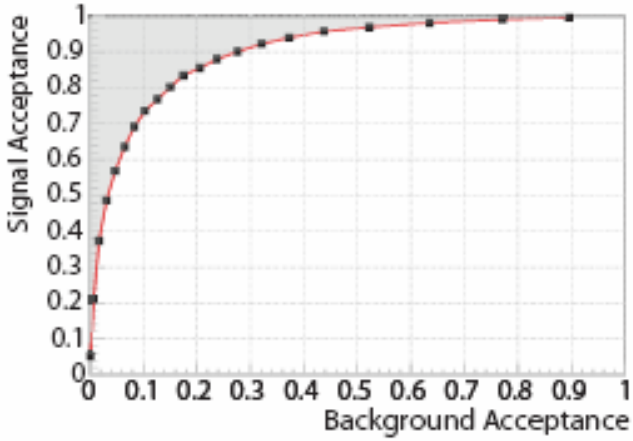
A Decision Tree is used to provide a ranking system for the importance of all variables (P_T of jets, masses, angular distributions, …)

- Optimize number of hidden nodes (found to be close to 30)

- Optimize number of training epochs (150-250)

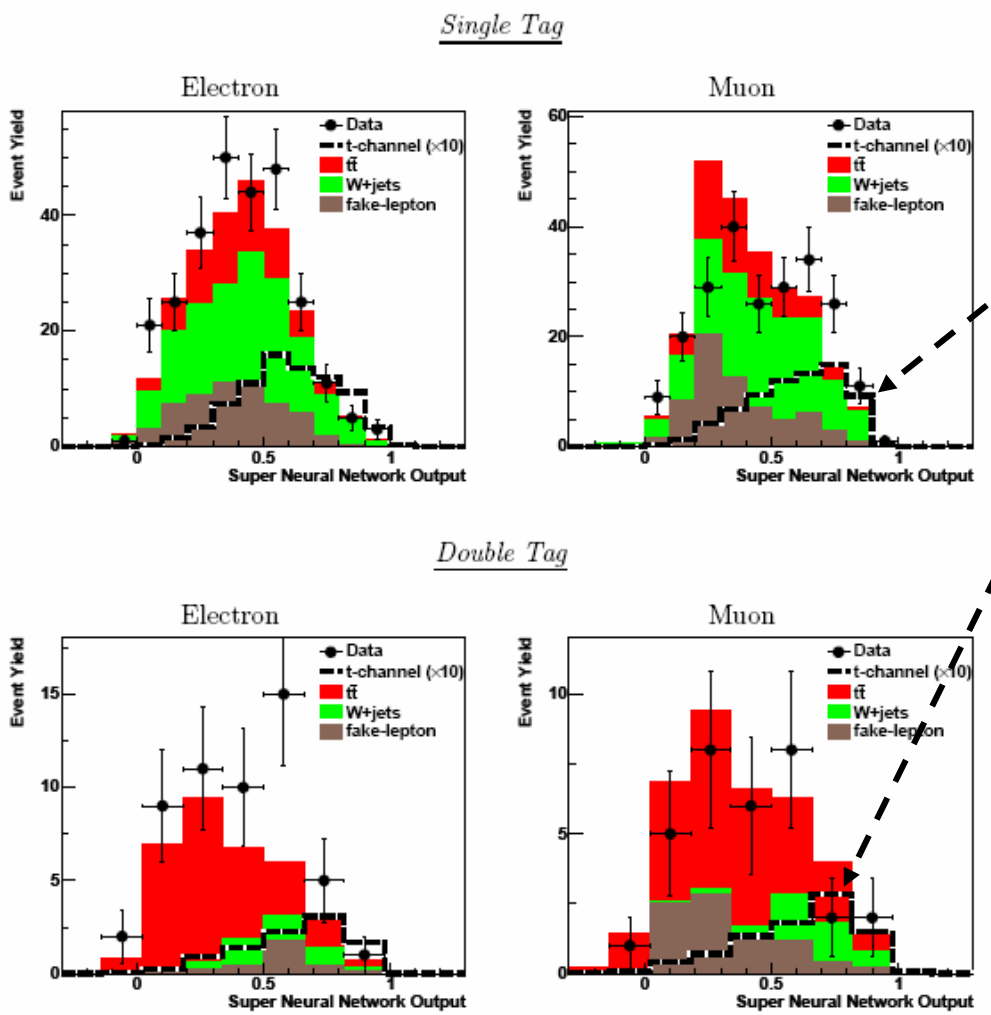
→ Figure of merite :





Application of NN, cont'd

Results for t-channel single top:



Signal x 10 !

→ output of super NN used to derive upper limits for single top cross section

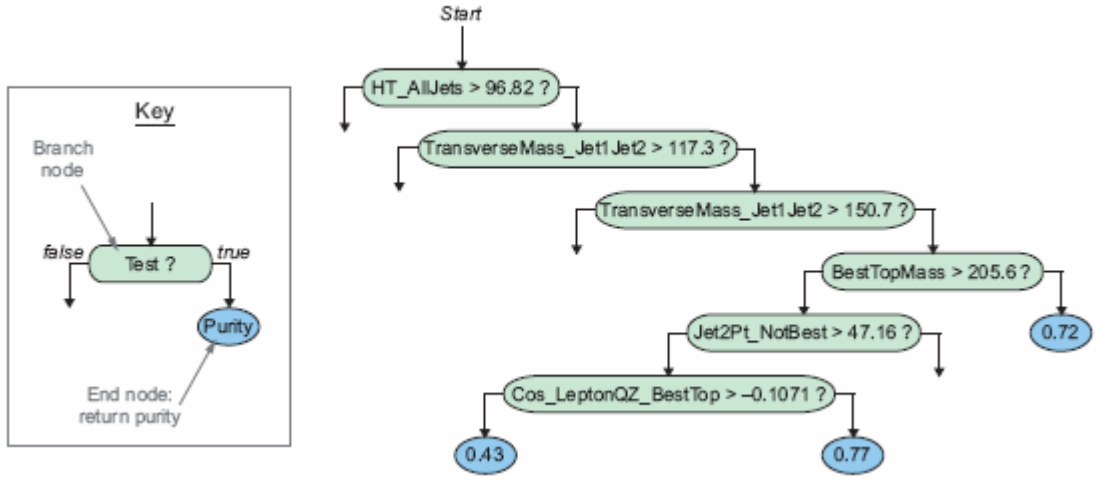
A few words about Decision Trees

Decision Tree = Multivariable analysis technique (learning) trying to recover events that fails criteria in cut-based analysis.

-Successive decision nodes to categorize events as S or B.

-Splitting is optimized at each node on a single variable with a single cut

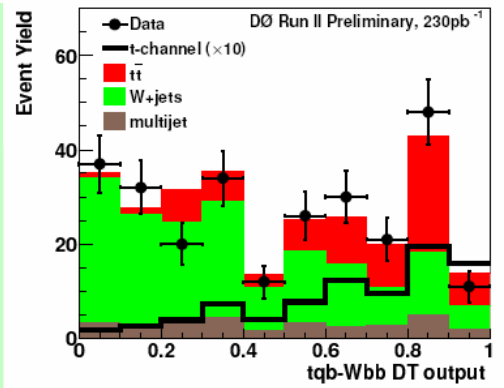
-Process repeated recursively



→ find nodes (leaves) with optimal separation

Advantages:

- not a black box, fast training, resistant to irrelevant variables, ranking of discriminative power



Defaults: unstable if not 'boosted', unknown in HEP community

The likelihood method

Most straightfoward multivariate analyzer !

Define likelihood ratio for a set $\vec{x}$ of variables:

$$\mathcal{L}(\vec{x}) = \frac{\mathcal{P}_{signal}(\vec{x})}{\mathcal{P}_{signal}(\vec{x}) + \mathcal{P}_{bckg}(\vec{x})}$$

$\mathcal{P}$: PDF for each class of events (S or B), $\mathcal{L} \rightarrow 1$ (0) for signal-like (background-like) events

For uncorrelated variables:

$$\mathcal{P}(\vec{x}) = \prod_{i}^{n} \mathcal{P}(x_i)$$

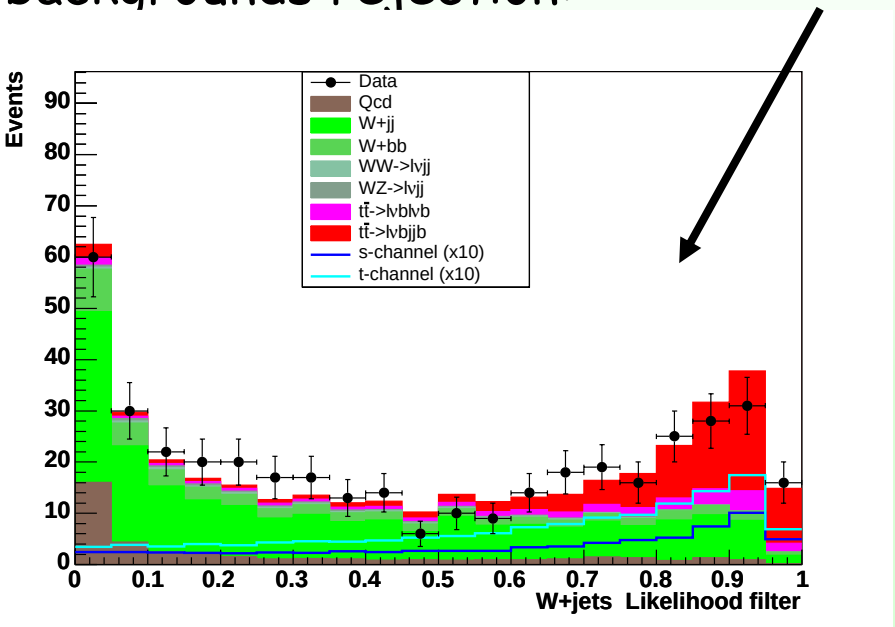
Likelihood ratio method is optimal

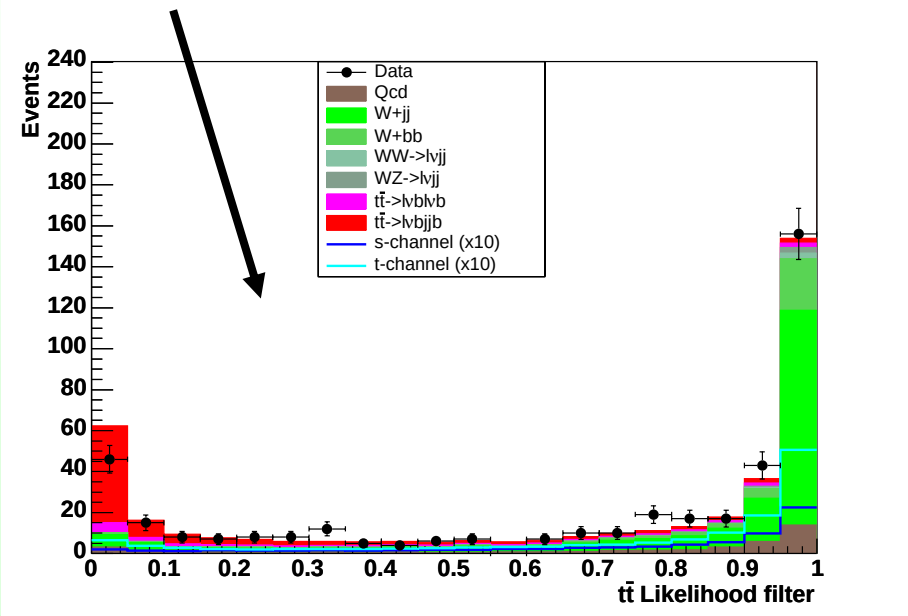
For p independant sources : $\mathcal{P}(\vec{x}) = \sum_{j=1}^{p} f_j \prod_{i} \mathcal{P}_j(x_i)$ f: relative aboundance for each source

BUT correlations !! Difficult to use directly P in a multidimensional space

The likelihood method (plots!)

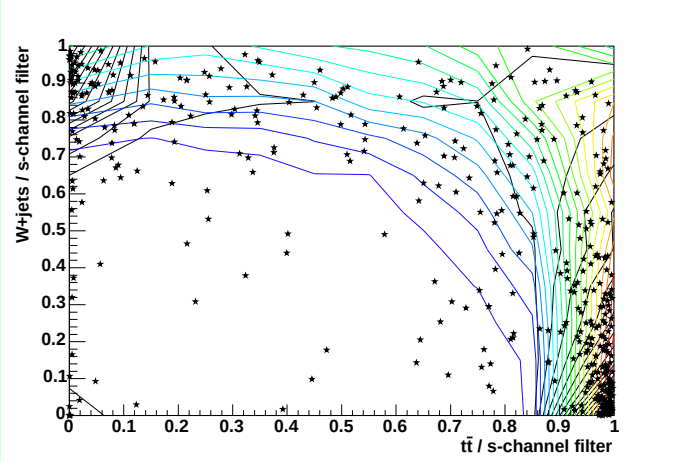
Two likelihoods, optimized for W+jets (7 variables) and tt (9 variables) backgrounds rejection:





Here for single-top s channel

→ combined :



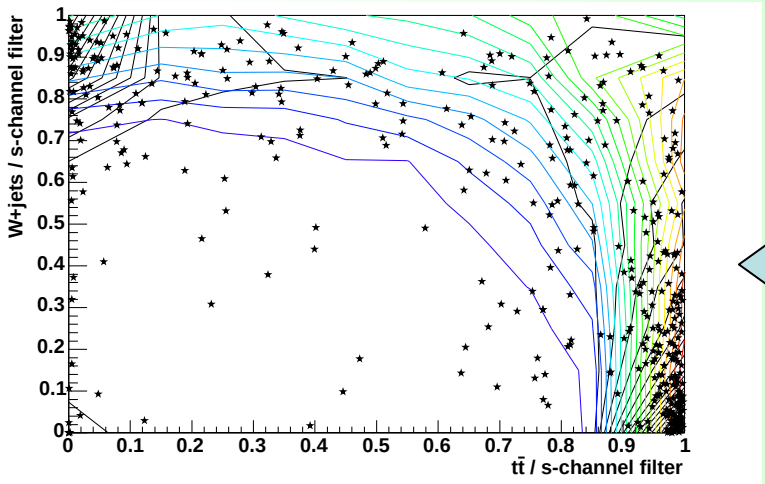
Setting a limit : Bayesian model applied to single top

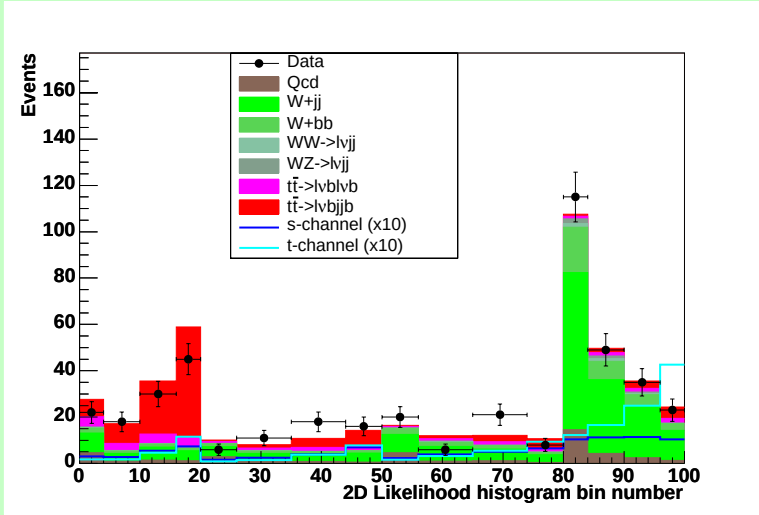
Different conceptual probabilistic approachs:

-Frequentist (frequency as a measurement of the probability (=unkown and constant quantity) that a phenomenon occured→ widely used at Lep (eg: Higgs mass search/limit)

-Bayesian (measure of the degree of belief depending on the status of information available → subjectivism)
→ natural, general (applied to any thinkable event), powerful (possibility to make a very general theory of uncertainty function of stat+sys errors)

Bayesian used in the statistical inference context or HOW TO DERIVE LIMITS FROM:





Setting a limit, cont'd

Use of the Bayes' Theorem:

Mutually exclusive hypothesis H_i which could condition an event E,

Conditionnal probability of H_i given E:

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{\sum_j P(E \mid H_j)P(H_j)}$$

$P(H_i)$ = Prior probability of H_i = proba of this H hypotesis with status of info avalaible before the knowledge that E has occured

$P(H_i \mid E)$ = posterior proba of H_i after the new information (WHAT WE ARE INTERESTED IN).

$P(E \mid H_i)$ = likelihood (WHAT CAN BE QUANTIFIED)

Setting a limit: Recipe for the single top

1/ Define the model

$$\mu = b + L\varepsilon\sigma$$

$$\mu = \mu(\sigma, \vec{x})$$

μ : number of expected events (data)

σ : signal X-section

L : integrated luminosity

ε : single top signal efficiency

b : number of expeted B events

$\vec{x}$ = nuisance parameters (to be estimated)

2/ Determine the likelihood

Probablity to observe k events given $\mu = \mu(\sigma, \vec{x})$: Poissonian probability

$$P(k \mid \mu) = \prod^{channels} \prod^{bins} \frac{e^{-\mu}\mu^k}{k!}$$

Recipe, cont'd

3/ Assign the prior probability for all parameters

No correlation between σ and $\vec{x}$ →

$$P_{rior}(\sigma, \vec{x}) = P_{rior}(\sigma) P_{rior}(\vec{x})$$

- Conventional use for $P_{rior}(\sigma)$: a flat PDF in a finite range:

$$P_{rior}(\sigma) \begin{cases} 1/\sigma_{max} & \text{if } 0 \leq \sigma \leq \sigma_{max} \text{ about 40 pb} \\ 0 & \text{otherwise} \end{cases}$$

- More difficult to determine $P_{rior}(\vec{x})$ (presence of sys. uncertainties (+/- correlated) of nuisance parameters) → $P_{rior}(\vec{x} + \delta\vec{x})$

$$\delta\vec{x} = \delta\vec{x}_{norm} + \delta\vec{x}_{shape}$$

$\delta\vec{x}_{norm}$: uncertainties affecting overall normalization (lumi, MC B, X-sec)

$\delta\vec{x}_{shape}$: uncer. affecting distribution shapes (JES, trigger turn on, MC TRF, JER, Jet ID)

Recipe, cont'd

4/ Apply Bayes' Theorem to find posterior PDF:

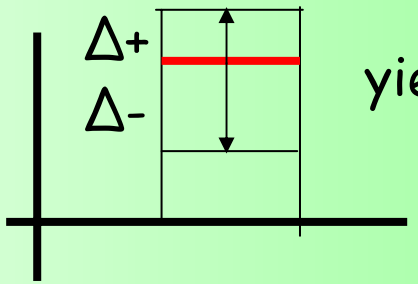
$$P_{ost} = \frac{\mathcal{L}(k\,|\,\sigma)}{\int_{\sigma'} \mathcal{L}(k\,|\,\sigma')d\sigma'} \quad \text{with} \quad \mathcal{L}(k\,|\,\sigma) = \int_{\vec{x}'} P(k\,|\,\sigma,\vec{x}')P_{rior}(\vec{x}')d\vec{x}'$$

Normalization term obtained by sampling computation:

$$\mathcal{L}(k\,|\,\sigma) = \frac{1}{N}\sum_{n=1}^{N} P(k\,|\,\sigma,\vec{x}_n + \vec{\delta x}_n)$$

$\vec{\delta x}_{n,norm}$: normalization effects modeled by sampling effective lumi and bkg by multidimensionnal gaussian N(0,1) (with covariance matrix)

$\vec{\delta x}_{n,shape}$: modeled by shifting one by one each source of uncertainty



yield→yield+Δ+ if δ>0
→yield+Δ- if δ<0

asymetric! δ from G(0,1)

Recipe, cont'd

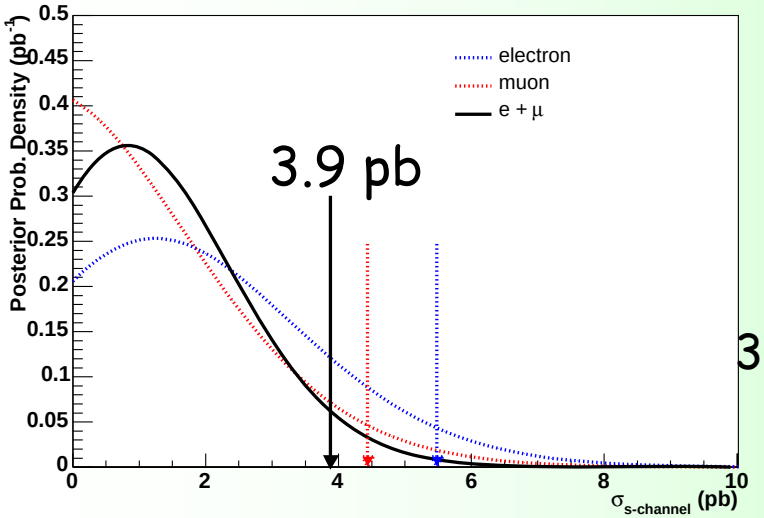
5/ Use resulting posterior to calculate quantity of interest

Posterior contains full information !

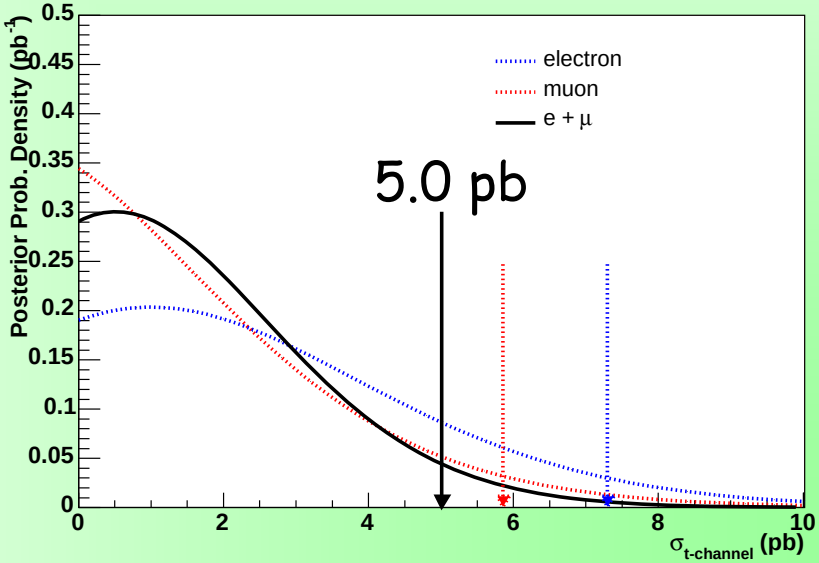
→ maximal value → measured X-section

→ Upper limit UL at 100CL% confidence level:

$$CL = \int_0^{\sigma_{UL}} P_{ost}(\sigma')d\sigma'$$



370 pb^{-1}



Conclusions

Many different analysis technics used in D0 but whatever the approach considered,

a good quality for inputs is predominant:

- good reconstruction of observables
- realistic agreement between simulation and real data

Advanced techniques using multivariate analyzer may be very useful for delicate processes (small signal against large background)

D0 has succesfully tested various approaches for the single-top study:

Likelihood, neural networks and promizing Decision Trees

Back up slides

Matrix element method (Mtop measurement, l+jets channel)

Principle: Compute an event-by-event probability as a function of Mtop making use of all kinematical information from reconstructed objects in the event.

Event probability: $P_{event} = f_{top} P_{sig}(x; m_{top}) + (1 - f_{top}) P_{bkg}(x)$

ftop: signal fraction, x: kinematic variable (reconstructed) , bkg: Wjets events

with $$P_{sig} = \frac{1}{\sigma(t\bar{t}, m_{top})} \int_{q_1, q_2, y} \sum_{q_1, q_2\, flavors} dq_1 dq_2 F(q_1) F(q_2) \times (2\pi)^4 \frac{|\mathcal{M}(t\bar{t} \to y)|^2}{q_1 q_2 s} \times W(x, y) d\phi$$

i.e. convolution of the transfer function W with the differential X-section. Y: partonic final state four momenta (generation level), F: partonic density function, dΦ: elementary phase space.

Similar definition for Pbkg: tt→Wjets

Transfer function W: Mapping from parton level variables (y) to reconstructed level variables(x). HARD CORE of the analysis !

Considering only jet variables (no l, neutrino) , 24 assignements of jets to parton →

$$W(x, y) = \frac{1}{24} \sum_{i=1}^{24} \prod_{j=1}^{4} W_{jet}(E_{jet\ j}, E_{quark\ i})$$

Matrix element method, cont'd

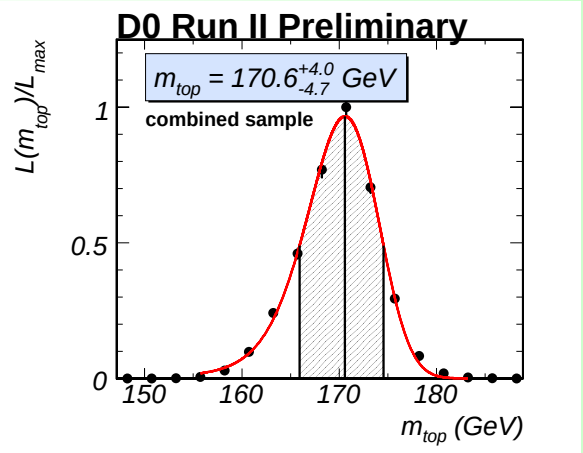
Transfer function $W_{jet}(E_j, E_q)$ is determined from simulated events and is parametrized by a gaussian form (depending on 120 parameters!)

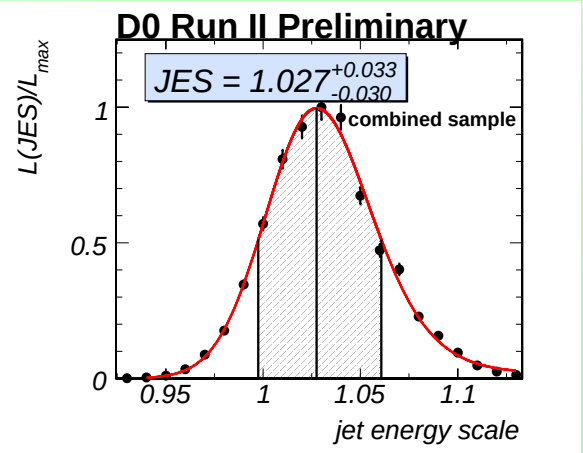
If b-tagging available, use modified W parametrization allowing better measurement to contribute more :

$$W(x,y) = \frac{\sum_{i=1}^{24} w_i \prod_{j=1}^{4} W_{jet}(E_{jet\ j}, E_{quark\ i})}{\sum_{i=1}^{24} w_i}$$

Wi: depending on the jet tagging efficiency

Extraction of Mtop by means of a maximum likelihood method (JES fitted as well) :





Setting limit: the CLs method

Test-statistic: function of the observables AND the model parameters of the known bckg and hypothetical signal. Constructed to increase monotonically for increasingly signal-like experiments → The confidence in the S+B hypothesis is given by the probability that the test-statistic Q is less or equl to the value observed in the experiment Qobs:

$$CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$$

Small CLs+b → poor compatibility with s+b hypothesis
Similary, confidence in the B hypothesis :

$$CL_b = P_b(Q \leq Q_{obs})$$

Small CLb → good compatibility with b hypothesis

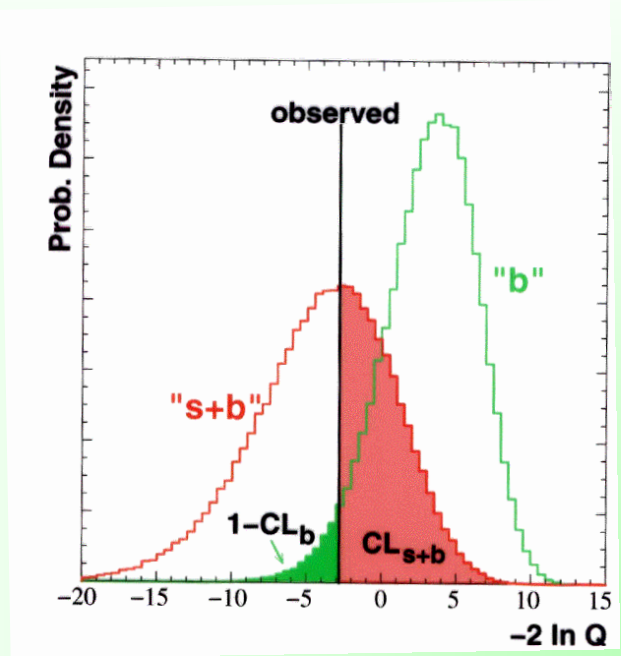
Modified Frequentist confidence level CLs: CLs+b/CLb

One appropriate test-statistic: the likelihood ratio test

$$Q = \prod_i^{bins} \frac{\frac{(s_i + b_i)^{n_i^{data}} e^{-(s_i+b_i)}}{n_i^{data}!}}{\frac{(b_i)^{n_i^{data}} e^{-(b_i)}}{n_i^{data}!}}$$

CLs method, cont'd

Ps+b(Q) and Pb(Q) pdf are derived from pseudoexperiments by varying n_i^{data} from a gaussian distribution:



In order to take into account uncertainties for S and B : integrate over these quantities →

$$P_{poisson}(n|s+b) \rightarrow \int_0^\infty ds' \int_0^\infty db' \frac{1}{2\pi\sigma_s\sigma_b} P_{poisson}(n|s'+b') e^{-\frac{(s-s')^2}{2\sigma_s^2}} e^{-\frac{(b-b')^2}{2\sigma_b^2}}$$

Improvements of basic likelihood method

Approximation : using projection of PDF

$$\mathcal{P}(x_i) \rightarrow \mathcal{Y}_i = \sqrt{2}Erf^{-1}(2F(x_i)-1) \quad , \quad \mathcal{Y} \sim G(0,1)$$

Cumulative distribution of $\mathcal{P}$

$$\mathcal{P}(\vec{x}) \approx |V|^{-1/2} \exp(-0.5\vec{\mathcal{Y}}^T (V^{-1}-1)\vec{\mathcal{Y}}) \prod_i \mathcal{P}(x_i)$$

V : covariance matrix for $\vec{\mathcal{Y}}$

But not really satisfactory in practice.

Kernel method computation :

$$\mathcal{P}(\vec{x}) = \frac{1}{nh^d} \sum_i^n K\left(\frac{\vec{x}-\vec{x}_i}{h}\right)$$

h : smoothing parameter, K : gaussian choice: $K(\vec{x}) = \frac{1}{(2\pi)^{d/2}} \exp(-0.5\|\vec{x}\|^2)$

Example of discri.variables (likelihood method)

	voie s-W+jets	voie s-$t\bar{t}$	voie t-W+jets	voie t-$t\bar{t}$
1. $p_T^{j_1}$	$\surd$	$\surd$	$\surd$	$\surd$
2. $p_T^{j_2}$	$\surd$	$\surd$	$\surd$	$\surd$
3. $p_T^{j_3}$	—	$\surd$	—	$\surd$
4. $METLep$	$\surd$	$\surd$	$\surd$	$\surd$
5. M_{jets}	$\surd$	$\surd$	$\surd$	$\surd$
6. M_T^W	$\surd$	$\surd$	$\surd$	$\surd$
7. M_{Wb}	$\surd$	—	$\surd$	—
8. ΔR_{min}^{jj}	—	$\surd$	—	$\surd$
9. $\cos(\widehat{j_2, \ell})_{top}$	$\surd$	—	$\surd$	—
10. *Sphéricité*	—	$\surd$	—	$\surd$
11. *Centralité*	—	$\surd$	—	$\surd$
12. $Q \times \eta$	—	—	$\surd$	$\surd$