

Numu CC - NC Separation in the Near Detector

N. Saoulidou, Fermilab, NC phone meeting 10-22-04

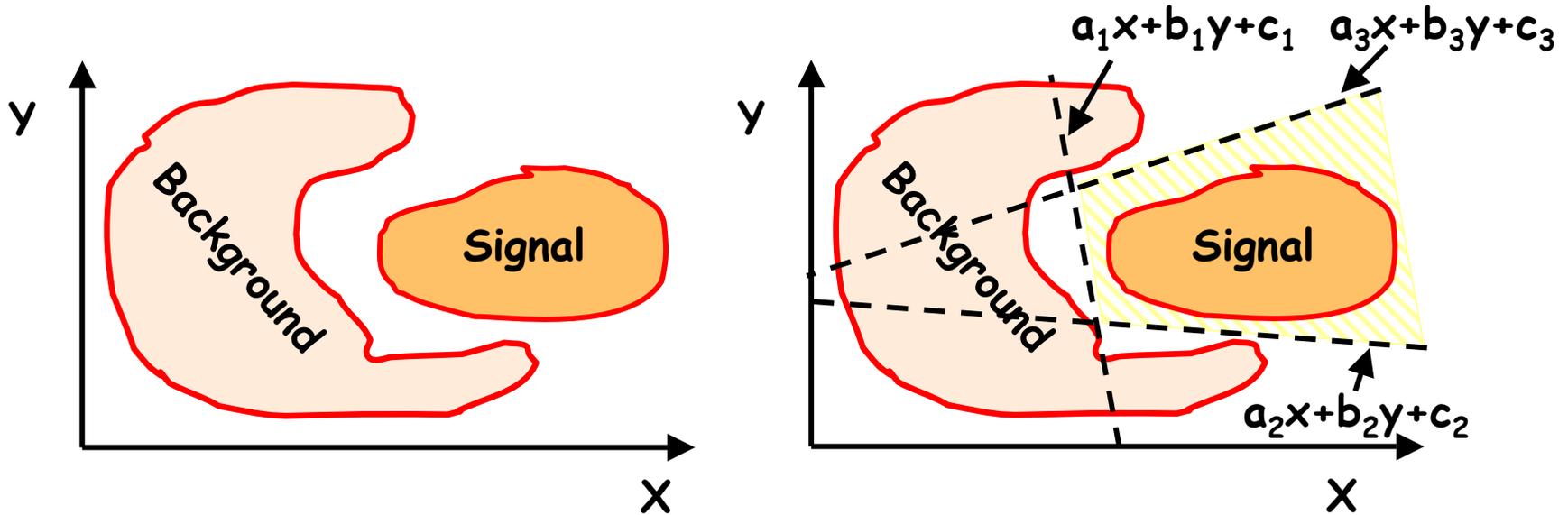
Outline

- Method (ANN)
- Events used
- Preliminary Results
- Summary - On going work

Method & Events used

- Reconstructed, using development release, 5 (so far) MDC ND overlaid files writing out NtpSR, NtpMC, NtpTH (**Jims latest truth variables are ABSOLUTELY essential for any kind of analysis that needs to know CC and NC events in overlaid files**)
- **The statistics is very poor!!**
- Used a plane cut to select an extremely pure sample of CC events (**60% of CC events**).
- Constructed an **ANN (MLPfit)** to classify the remaining CC and NC events with total length < 40 planes

ANN BASICS



- Event sample characterized by two variables X and Y (left figure)
- A linear combination of cuts can separate "signal" from "background" (right fig.)
- Define "step function" $S(ax + by + c) = \begin{cases} 0 & \text{"Signal (x, y)" OUT} \\ 1 & \text{"Signal (x, y)" IN} \end{cases}$
- Separate "signal" from "background" with the following function:

$$C(x, y) = S(S(a_1x + b_1y + c_1) + S(a_2x + b_2y + c_2) + S(a_3x + b_3y + c_3) - 2)$$

ANN BASICS

- Error function : $E = \sum_p E_p = \sum_{jp} (d_{pj} - t_{pj})^2$, where
 - p : runs over the events of the training set,
 - j : the index of an output neuron,
 - d_{pj} : the desired output of neuron j in event p ,
 - t_{pj} : the network output.

- All **minimization** methods use the computation of first order derivatives:

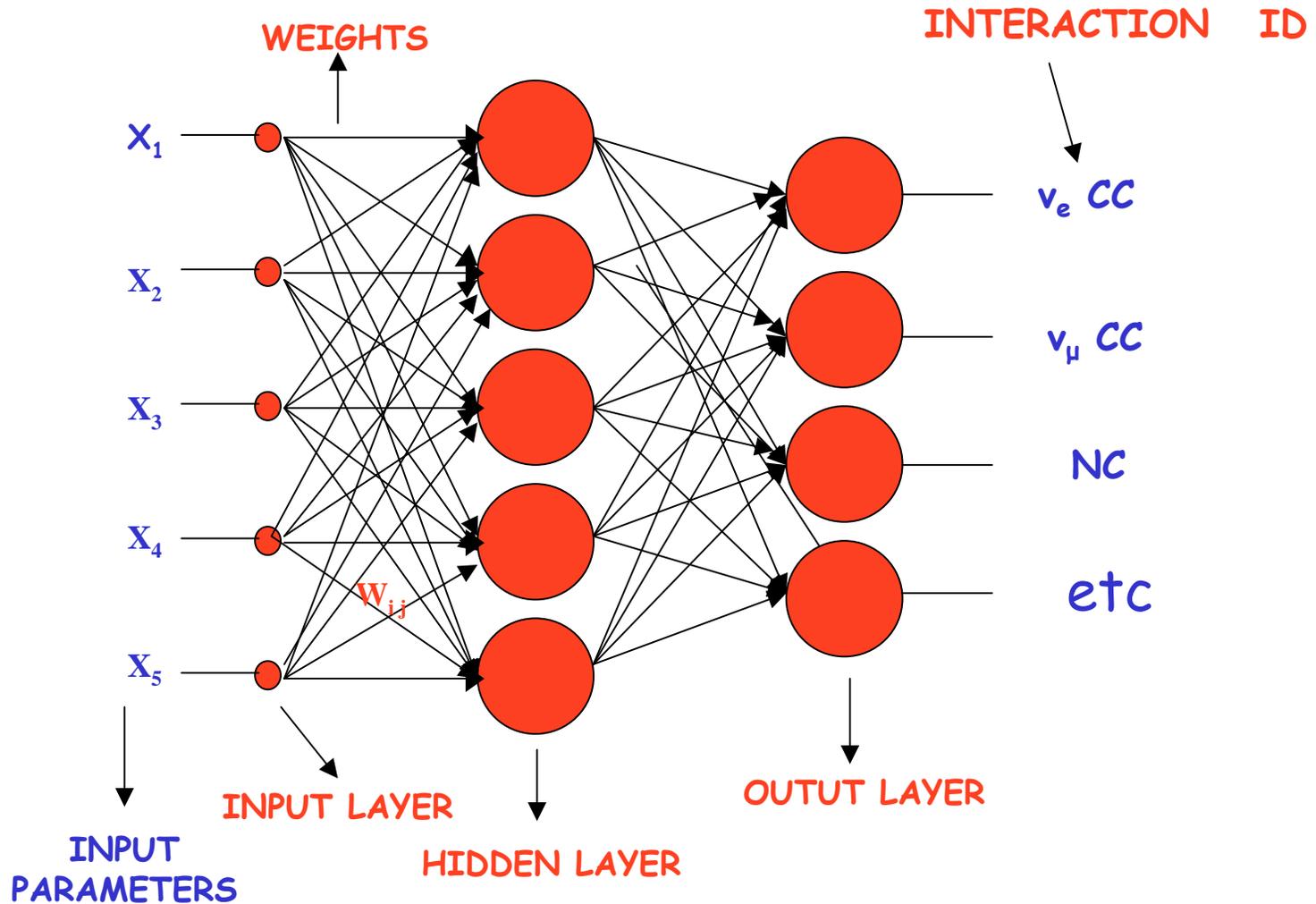
$$\frac{\partial E}{\partial w_{ji}} = \sum_p \frac{\partial E_p}{\partial w_{ji}}$$

- The description of **backpropagation** is that in each iteration :

$$\Delta_p w_{ji}(n+1) = -\varepsilon \frac{\partial E_p}{\partial w_{ji}} + \alpha \Delta_p w_{ji}(n) \text{ , where}$$

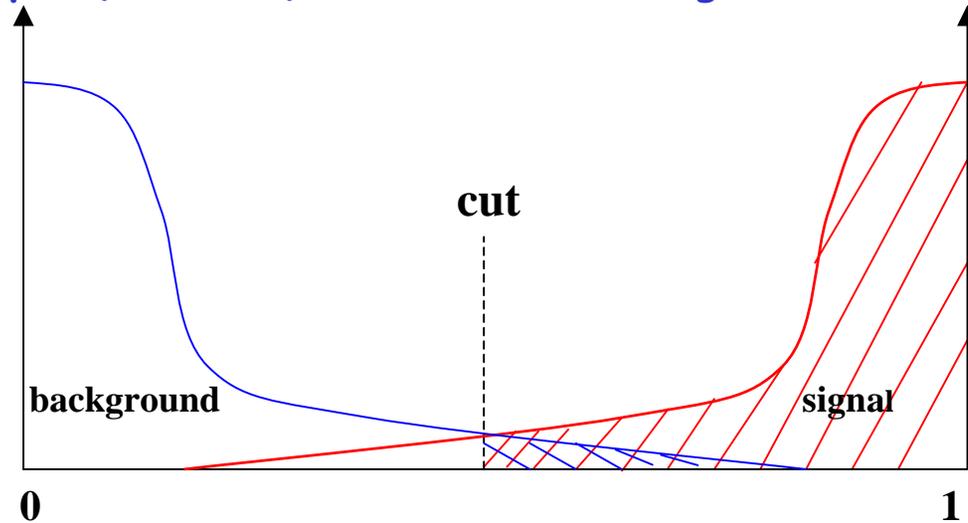
- $\Delta_p w_{ji}(n+1)$: the **change in w_{ji}** in iteration $n+1$,
- ε : the distance to move along the gradient (**learning coefficient**)
- α : a smoothing term (**"momentum"**)

ANN Schematic



ANN Parameters

Network output (selection) function for "background "and "signal" events



S = Total # Signal events

B = Total # Background events

S_C = Signal events above Cut

B_C = Background events above Cut

$$\text{efficiency} = \frac{S_C}{S}$$

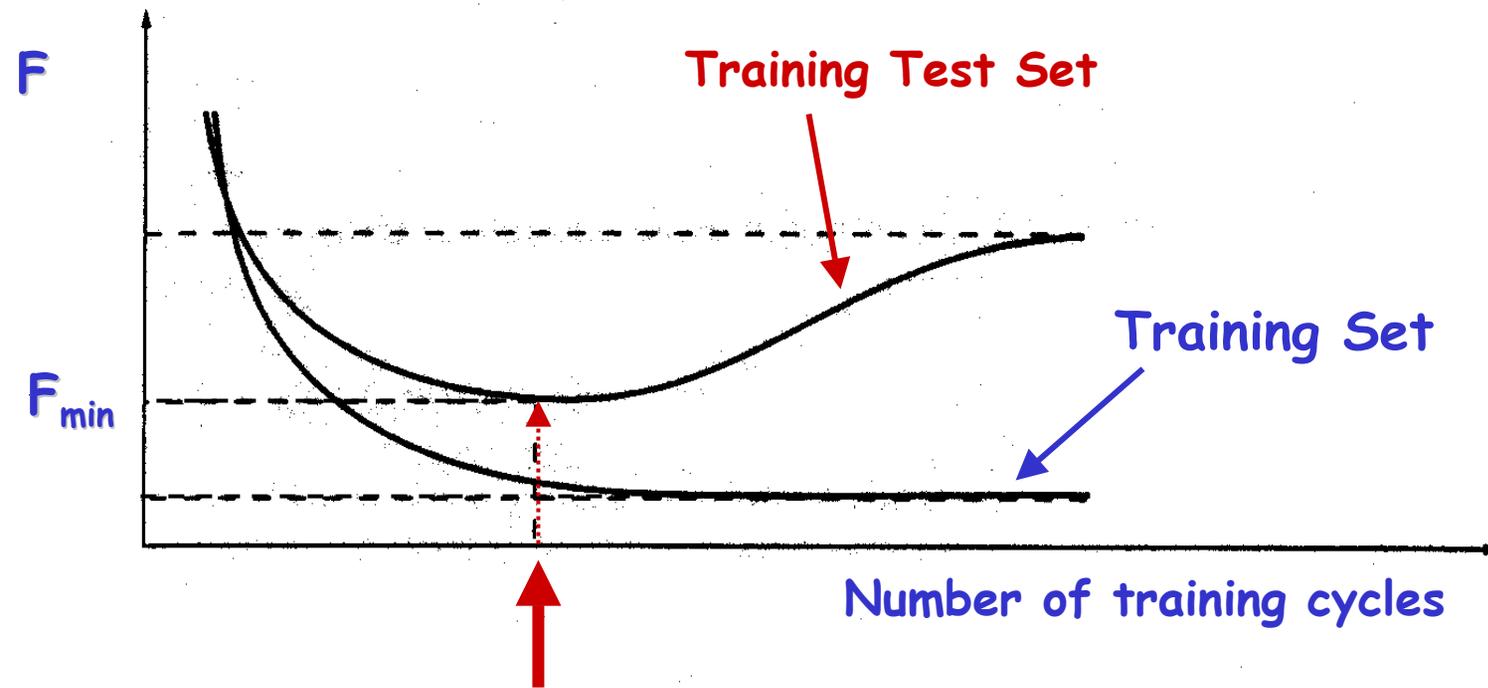
$$\text{purity} = \frac{S_C}{S_C + B_C}$$

$$\text{contamination} = \frac{B_C}{B}$$

Overtraining (Early stopping method)

F is the "cost function"

When an ANN get overstrained is loses its generalization ability and learns ONLY the specific training examples



**STOP
HERE**

ANN architecture & Input variables

ANN Architecture

13 inputs

1 hidden layer with 5 neurons

1 output

Input Variables :

Pulse height per plane

Pulse height per strip

Number of tracks

Track momentum from range

Track PH per plane

Percentage of track PH to total event PH

Shower number of strips

Shower number of planes

Shower PH

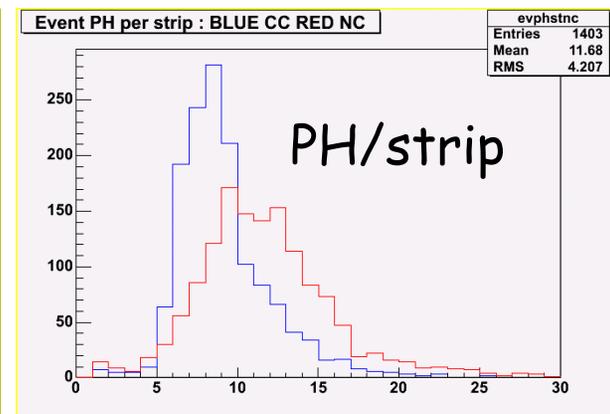
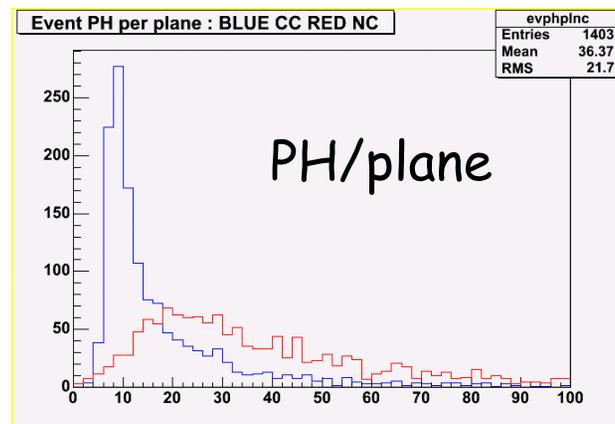
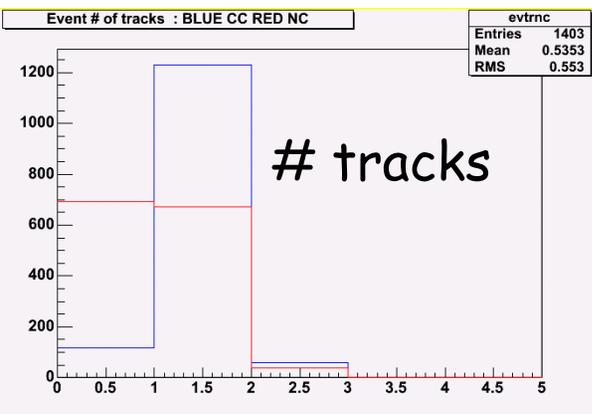
Shower number of digits

Percentage of shower energy to total event energy

Shower PH per plane

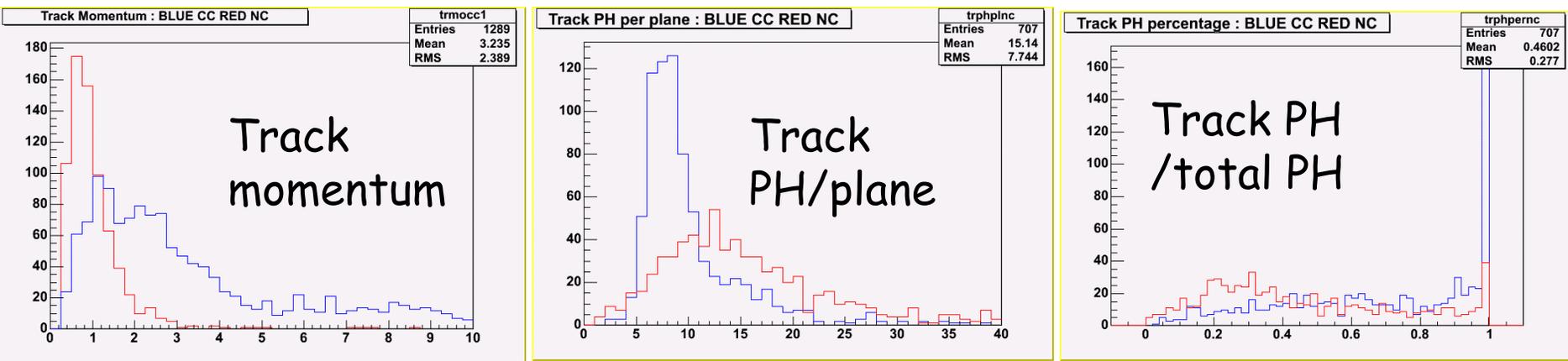
Shower PH per strip

Numu CC - NC variables (Event characteristics)



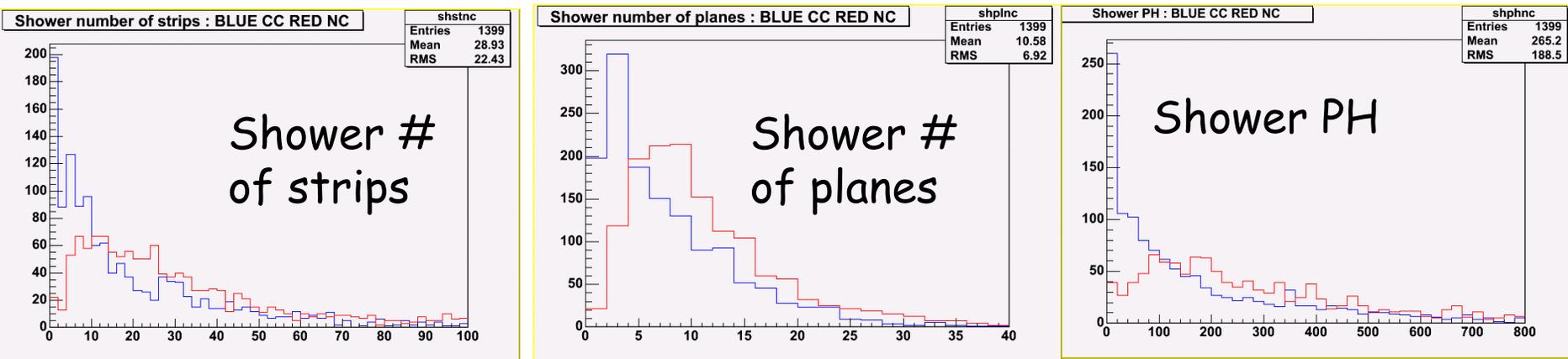
- These distributions are normalized to have \sim the same number of entries.
- In "MC reality" (!) CC events that survive the 40 plane cut are ~ 4.5 times more than NC events.
- These distributions so far show that CC events could be selected with a "decent" purity but NC events cannot...

Numu CC - NC variables (Track characteristics)



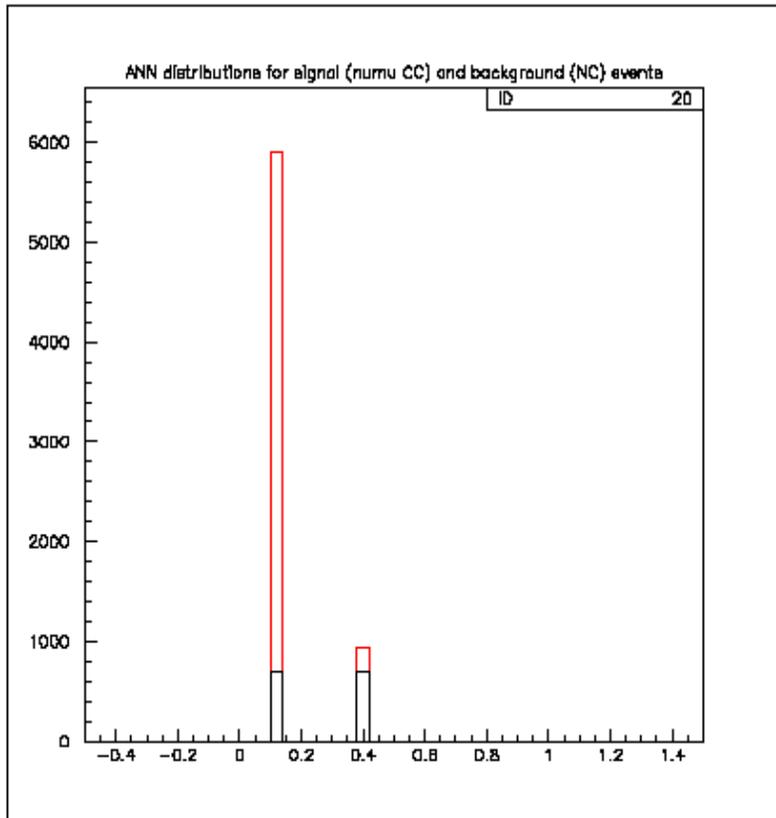
- Again these distributions so far show that CC events could be selected with a "decent" purity but NC events cannot...
- Also if the blue (CC) distributions increase by 4.5 times then the red (NC) distributions will be ~ completely swallowed.

Numu CC - NC variables (Shower characteristics just 3 representative variables)



- Again these distributions so far show that CC events could be selected with a "decent" purity but NC events cannot...
- Also if the blue (CC) distributions increase by 4.5 times then the red (NC) distributions will be ~ completely swallowed.

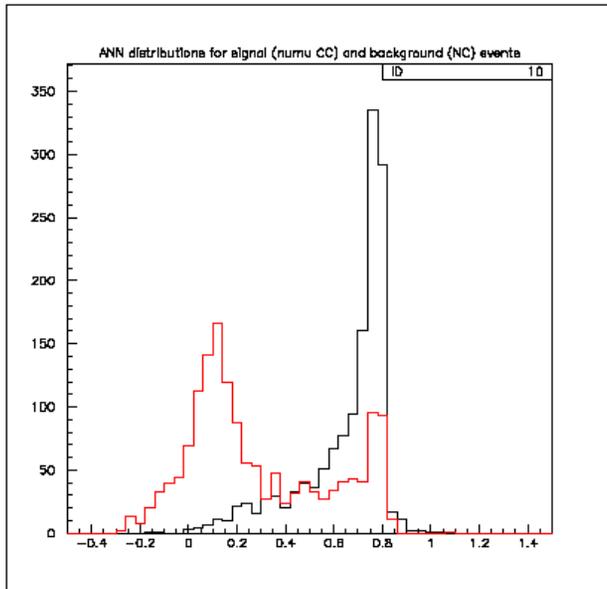
Sanity test with ANN



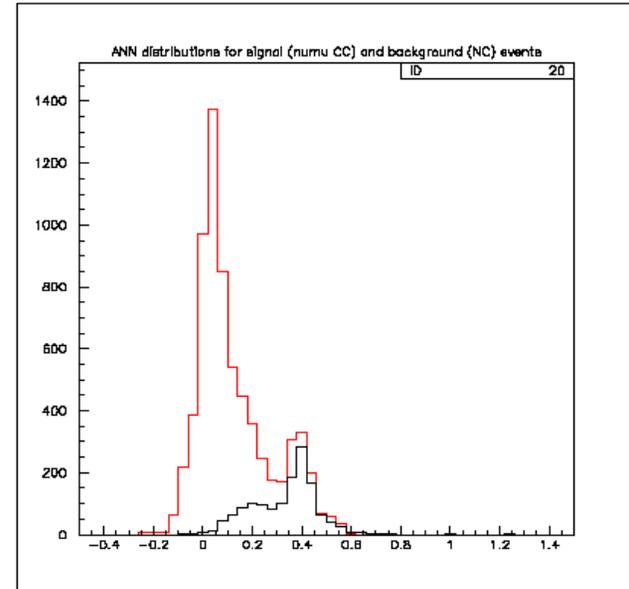
- Use just number of tracks to see in the worst possible case what I would get.
- What the ANN says is that in this case CC events (red) with zero tracks would be classified as NC and NC events (black) with one track as CC which is what we expect.
- The number of CC and NC events that you see are in the proportions expected.
- If the other variables don't add any significant information for the event classification then we expect to see ~ the same picture.

Final ANN Results

A priori probabilities 1:1



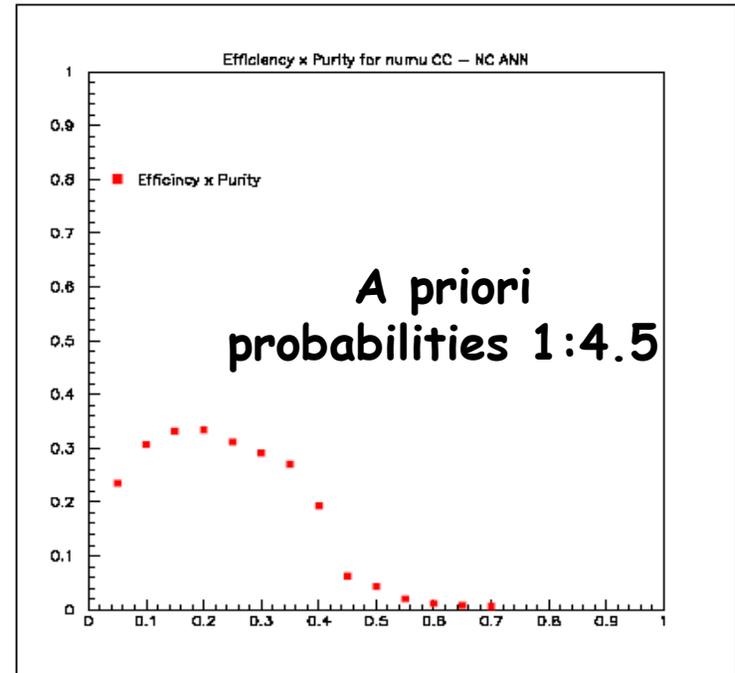
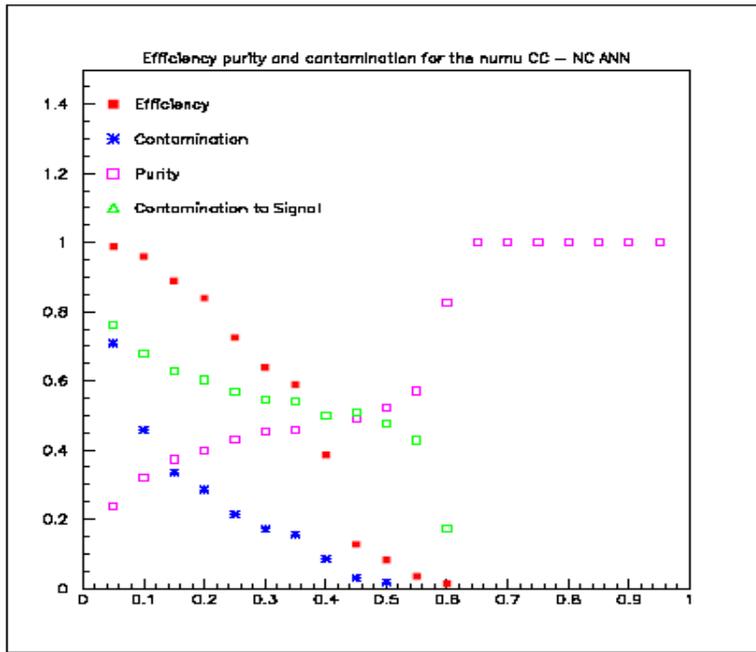
A priori probabilities 1:4.5



Event Probability

- The ANN performs as expected : Higher purity for CC selection and poor for NC selection especially using the ACTUAL a priori ratios.

Final ANN Results

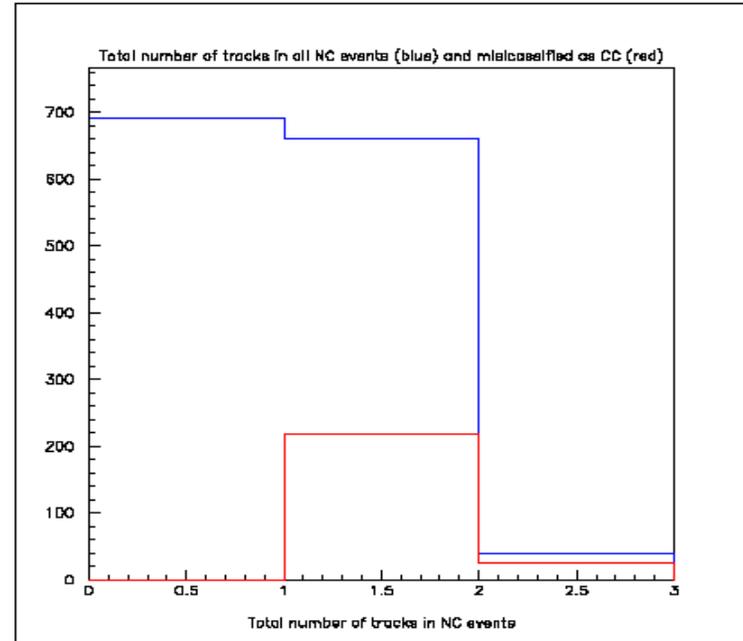
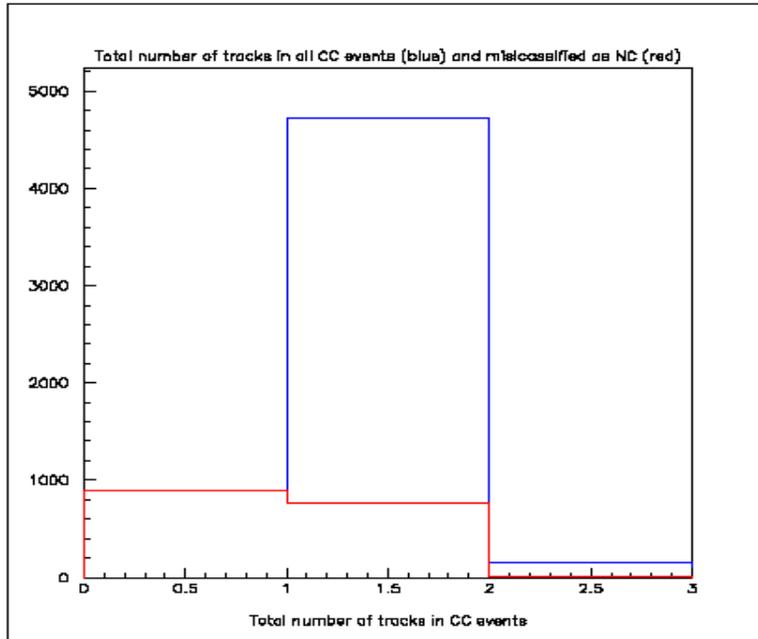


Efficiency (red) and purity (magenta) as a function of cut in the ANN output function for the signal (NC events)

Efficiency x Purity

- If we set the cut @ 0.2 (i.e.) we have an efficiency and a purity of 40 %...

ANN misclassification



Number of tracks for CC events.

Number of tracks for NC events.

Blue : Events correctly classified as CC. Blue : Events correctly classified as NC

Red : Events wrongly classified as NC Red : Events wrongly classified as CC

- The number of CC events that are misclassified as NC is **LARGER** than the total number of NC events...(a priori ratios are 4.5:1 CC:NC)
- The major problem seems to be CC events with no tracks ...

Summary / On going work

- In general we can select the sliced and reconstructed ND overlay CC events with a high efficiency and high purity.
- NC events seem hard (at this stage) to distinguish.
- I want to visually scan the events from both categories that were misclassified in order to :
 - Identify possible failures in slicing and/or reconstruction
- I will continue processing MDC files in order to my statistics that is very bad at the moment.
- I want to search for new variables that I could add in order to achieve a better classification.
- More results hopefully in the next meeting...