Statistical Tools in Collider Experiments

Multivariate analysis in high energy physics

Lecture 3

Pauli Lectures - 08/02/2012

Nicolas Chanon - ETH Zürich

ETH Institute for Particle Physics

Outline

1. Introduction
2. Multivariate methods
3. Optimization of MVA methods
4. Application of MVA methods in HEP
5. Understanding Tevatron and LHC results

Lecture 3. Optimization of multivariate methods

Outline of the lecture

Optimization of the multi-variate methods

- Mainly tricks to **improve the performance**
- Check that the performance is **stable**
- These are possibilities that have to be tried, no recipe which would work in all cases

Systematic uncertainties

- **How to estimate systematics** on a multivariate method output ?
- It depends on how it is used in the analysis
- If **control samples** are available
- Depends a lot on the problem

Optimization

The problem.

- Once a multi-variate method is trained (say a NN or BDT), how do we know that the **best performance** is reached ?
- How to test that the results are **stable** ?

- Optimization is an **iterative process**, there is no recipe to make it work out of the box
- There are many things that one has to be careful of

- Possibilities for improvement :

- Number of variables
- Preselection
- Classifier parameters
- Training error / overtraining
- Weighting events
- Choosing a selection criterion on the output

Number of variables

Optimizing the number of variables :

- How to know if the **set of variables used** for the training is the **optimal** one ?
- This is a difficult question which depends a lot on the problem
- What is more manageable is to know if among all the variables, some are unuseful.

Variable ranking :

- Variable ranking **in TMVA is NOT satisfactory!!**
- Importance of input variables in MLP in TMVA depends on the **mean** of the variable and the **sum of the weights** for the **first layer**
- Imagine with variables having values with different orders of magnitudes.....

$$I_i = \bar{x_i} \sum_j^{n_1} w_{ij}^{l_1}$$

- A **more meaningful estimate** of the importance was proposed
- Does not depend on the variable mean
- Is a relative fraction of importance (all importance sums up to 1)

$$SI_i = \frac{\sum_j^{n_1} \left| w_{ij}^{l_1} \right|}{\sum_i^N \sum_j^{n_1} \left| w_{ij}^{l_1} \right|}$$

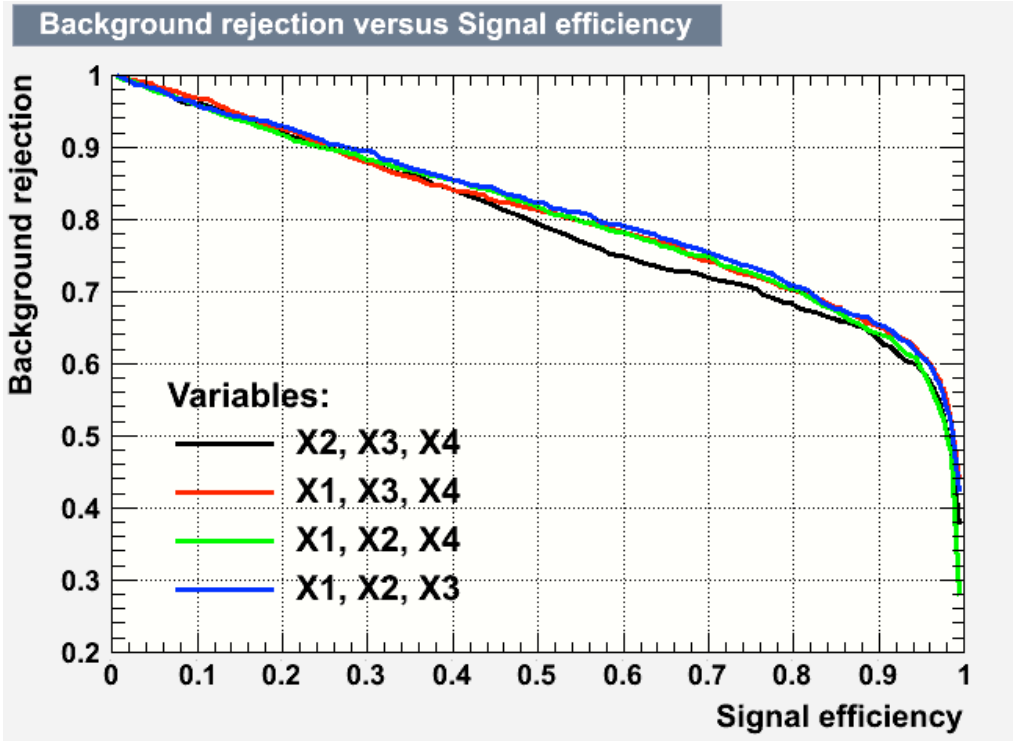
- **Problem : again rely only on the first layer**. What happens if more hidden layers ?

Number of variables

Proposed procedure (A. Hoecker) :
N-1 iterative procedure

- Start with a set of variables
- **Remove variables one by one**, keeping all the remaining as input. Check the performance
- The removed variables which **worsens the more** the performance is the **best** variable.
- **Remove** this variable definitively from the set.
- Repeat the operation until all variables have been removed => Get a **ranking** of the variables

But : This ignores if a smaller set of correlated variables would have performed better if used together



Removing X1 gives the worst performance

Selection

How to deal with ‘difficult’ events ?

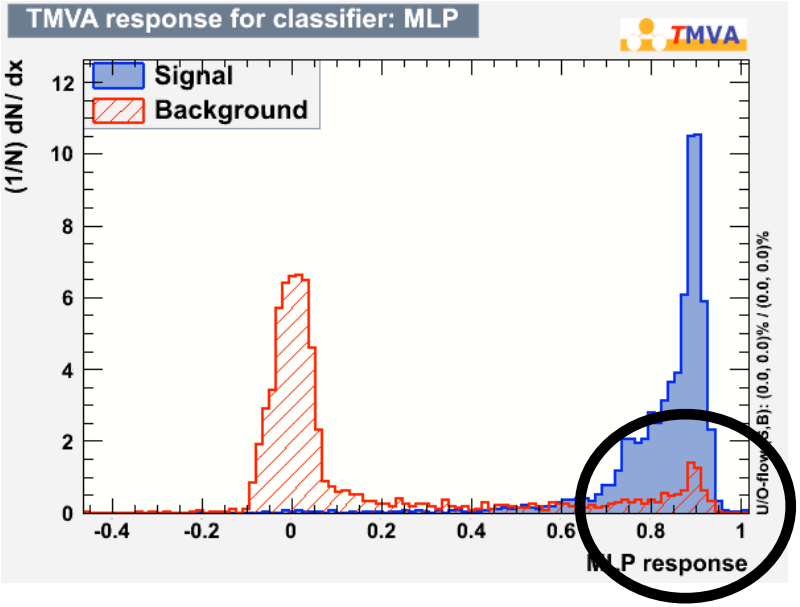
- E.g. events in a sample with high weight (difficult signal-like event in background sample with large cross-section)
- If including, might decrease the performance (few statistics)
- If excluding, the output on test sample can be random...

Tightness of the preselection

- Generally speaking, multivariate methods **performs better if a large phase-space is available**
- On the other hand applying relatively tight cuts before training might help to focus on some small region of the phase-space where discrimination is difficult...

Vetoing signal events in background samples

- Try to have only signal event in signal samples (etc)



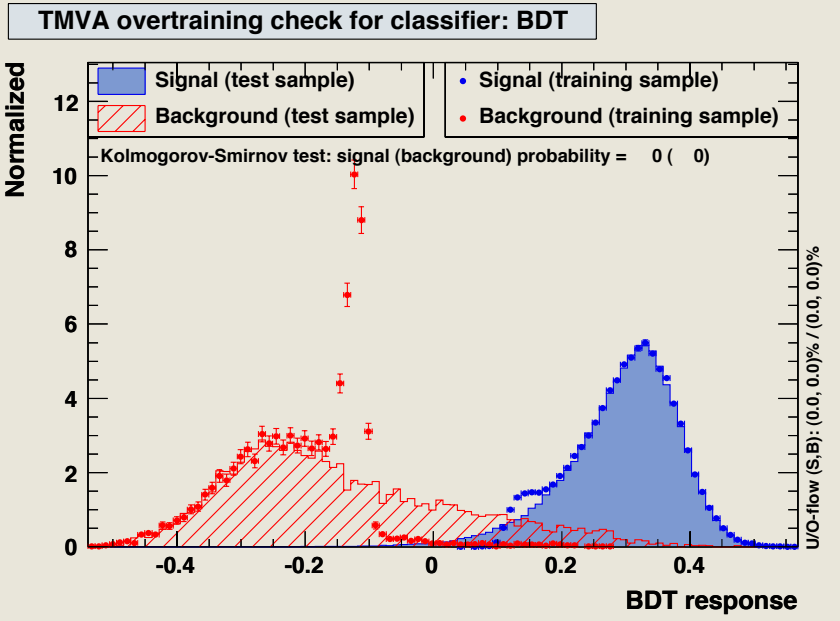
Variables definition

Variables with different orders of magnitude :

- Not a problem for BDT
- Normalizing them can help for NN

Undefined values for some events.

- BDT has problems if putting arbitrary numbers for those ones. How to cut on a value which is meaningless ?
- This is how BDT can be overtrained...
- Example : distance of a photon with respect to the closest track in a cone 0.4, in events where no track is there



Classifier parameters

Neural network parameters optimization :

- Vary number of **neurons**, and **hidden layers** : TMVA authors recommend one hidden layers with N+5 neurons for MLP
- Vary number of **epochs** (although performance might stabilize)
- Different activation function should give same performance

BDT parameters optimization

- Vary number of **cycles**
- Vary the **tree depth**, number of cuts on one variable
- Different decision function should give same performance
- Combination of boosting/bagging/random forest : TMVA authors recommend to boost simple trees with small depth

Preparing training samples

- Training and test samples have to be different events

Number of events in training samples :

- Sometime good to have as many events in the signal and the background.
- Number of events is shaping the output.
- A asymmetric number of events can lead to the same discrimination power, BUT at the price of more events needed => lower significance

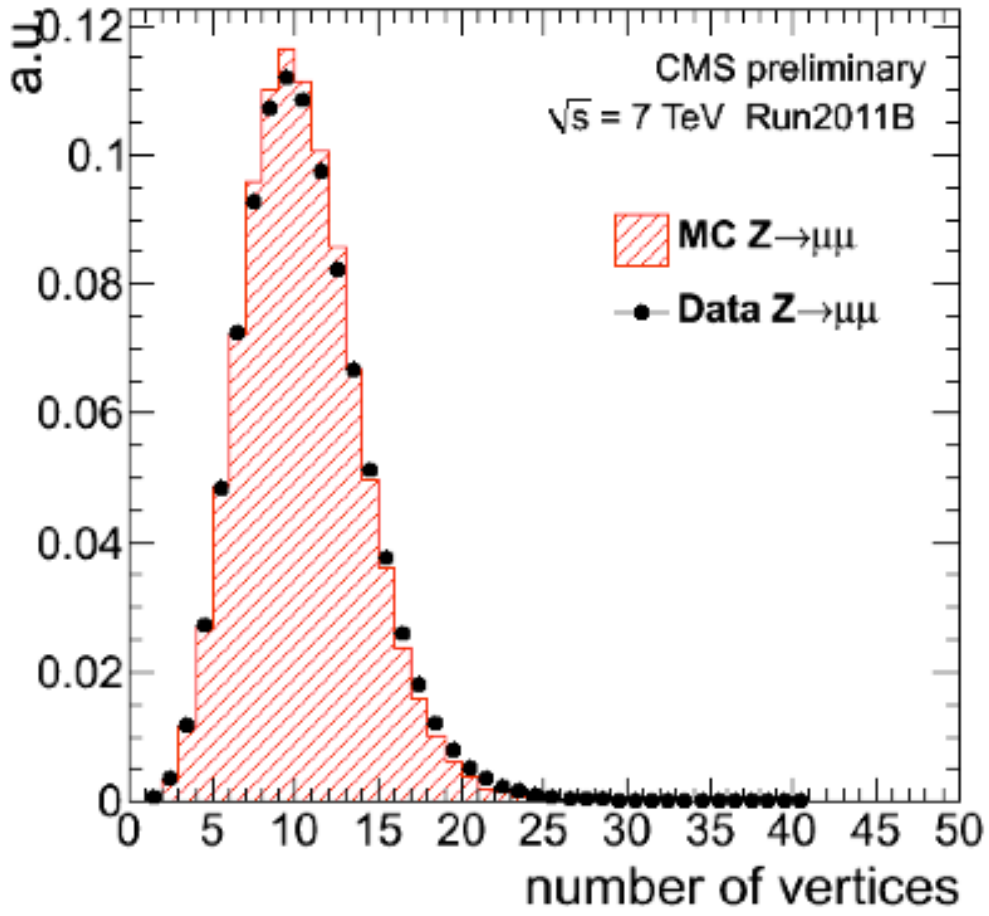
Using samples with different (fixed) weights :

- It is clearly not optimal, but sometimes we can not do otherwise
- If one sample with too few events and large weight, better to drop it

Weighting events

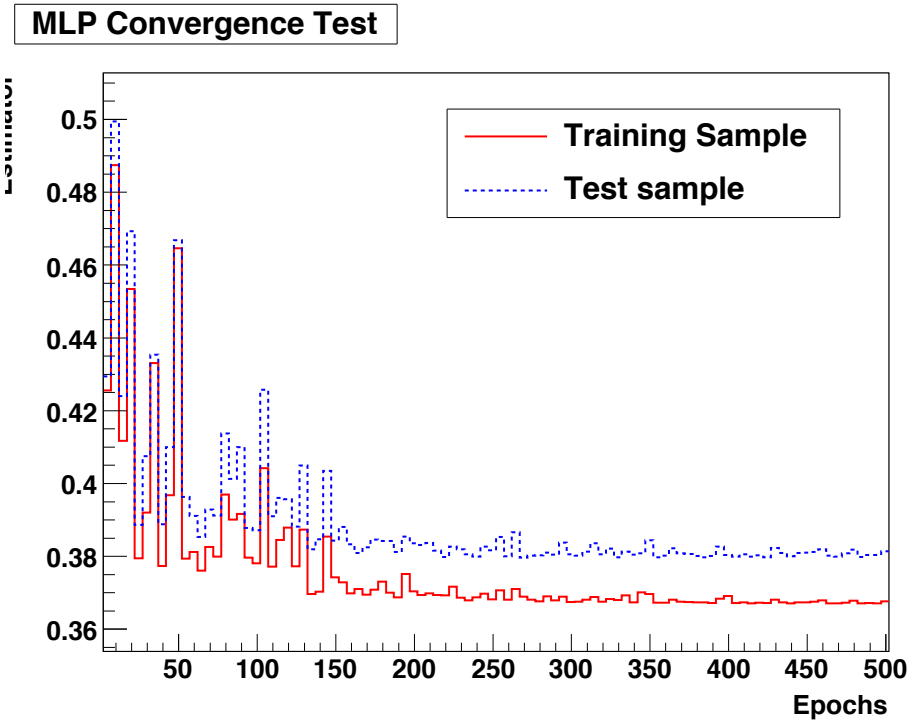
Weighting events for particular purposes :

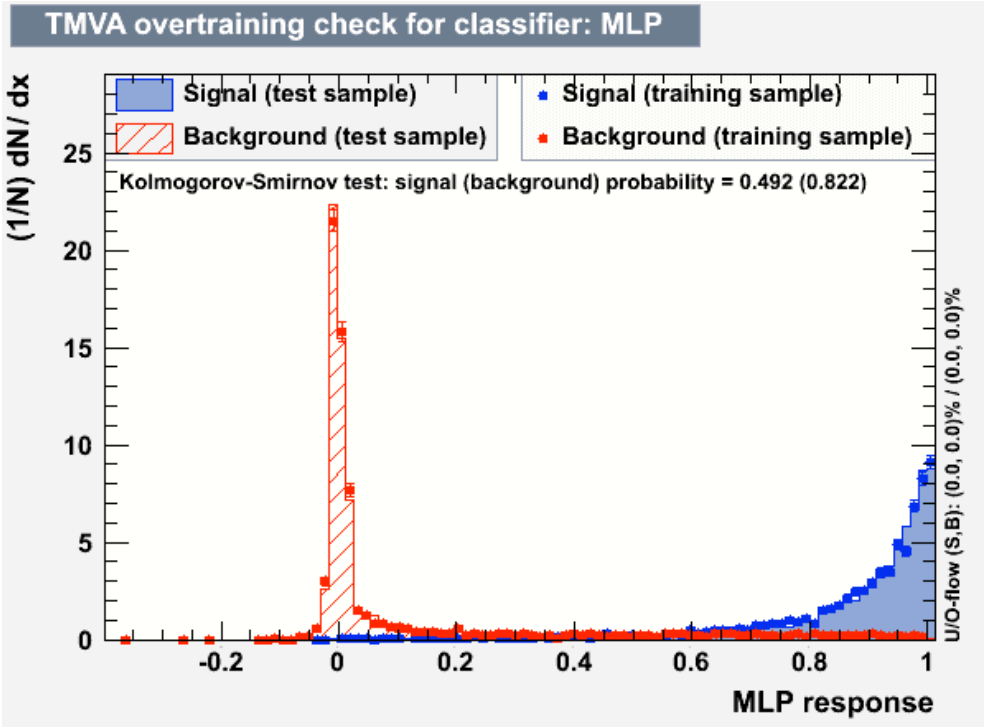
- One can weight events to improve the performance on some region of the phase-space
- E.g. : events with high pile-up or with high energy resolution



Error and overtraining

- Overtraining has to be checked





Using the output

- The multivariate discriminant is trained. How to use it in the analysis ?

Selection criteria :

- On the performance curve, choose a working point for a given s/b or background rejection
- Choose the working point maximizing S/sqrt(S+B) (approximate significance)
- Maximize significance or exclusion limits

If two values per event, which one to use ?

- E.g. for particle identification
- min, max value of the output ?
- Leading/subleading ? Both ?

Optimization : example

MiniBoone [arxiv:0408124v2]

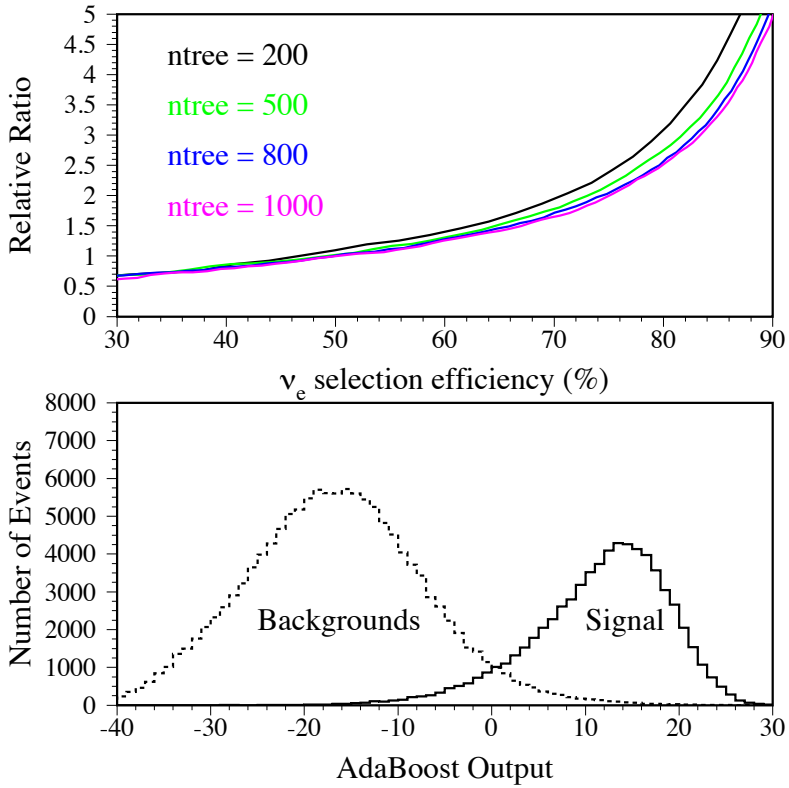


FIG. 3: Top: the number of background events kept divided by the number kept for 50% intrinsic ν_e selection efficiency and $N_{tree} = 1000$ versus the intrinsic ν_e CCQE selection efficiency. Bottom: AdaBoost output, All kinds of backgrounds are combined for the boosting training.

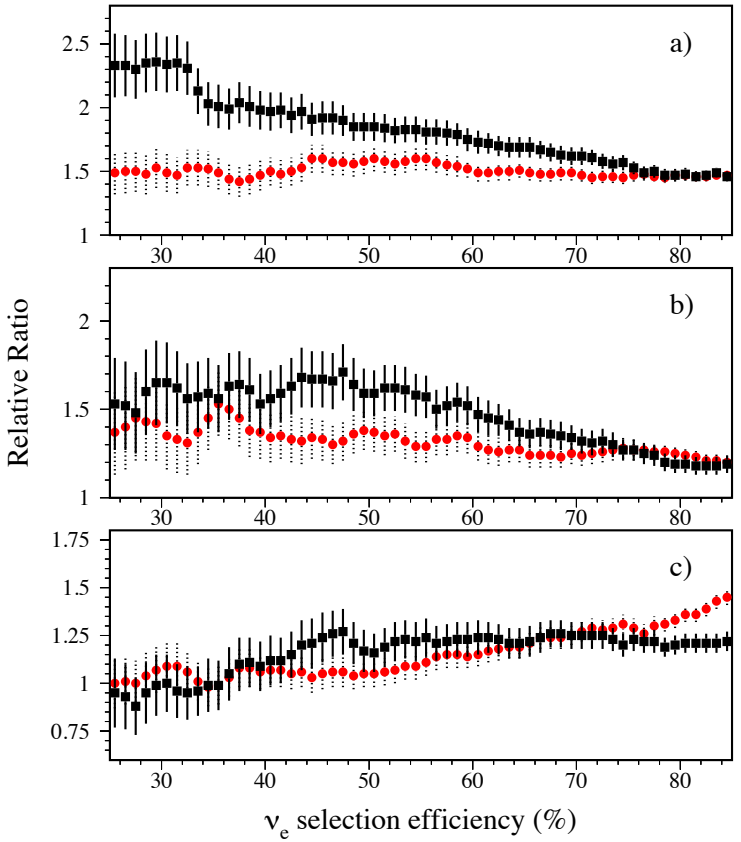


FIG. 4: Comparison of ANN and AdaBoost performance for test samples. Relative ratio(defined as the number of background events kept for ANN divided by the events kept for AdaBoost) versus the intrinsic ν_e CCQE selection efficiency. a) all kinds of backgrounds are combined for the training against the signal. b) trained by signal and neutral current π^0 background. c) relative ratio is re-defined as the number of background events kept for AdaBoost with 21(red)/22(black) training variables divided by that for AdaBoost with 52 training variables. All error bars shown in the figures are for Monte Carlo statistical errors only.

Systematic uncertainties

How to deal with systematics in an analysis using multivariate methods ?

- Usual cases of the signal/background discrimination :
  - Cut on the MVA output
  - Categories
  - Using the shape

- Systematic on the training ? On the application ?

- Importance of the **control samples**.

Training systematics ?

Should we consider systematic uncertainties due to the training ?

- General answer : No.

- If the classifier is overtrained, better redo the training properly (redo the optimization phase)

- Imagine a complicated expression for an observable with many *fixed* parameters. Would you move the parameters within some uncertainties if the variables is used in the analysis ? Generally speaking, no.
- This is the same for classifiers. The MVA is one way of computing a variable. One should not change the definition of the variable.

- Sometimes found in the litterature : remove one variable, redo the training, check the output, derive the uncertainty. BUT : it is changing the definition of the classifier output. Furthermore, too much variation if changing the input variables

Control samples

A **control sample** is a **data sample** used to :
- **Validate** the variables **modeling**
- Estimate the **systematic** uncertainties
- It should be **independent from the signal region** looked at in the analysis

=> **Crucial for classifiers** validation and systematics !

Data/MC agreement is fundamental to show that we understand the classifier behavior

(But if the mismodeling is “small”, it means the correlations are wrong, it would just lead to a non-optimal result, as long as the background is estimated from data)

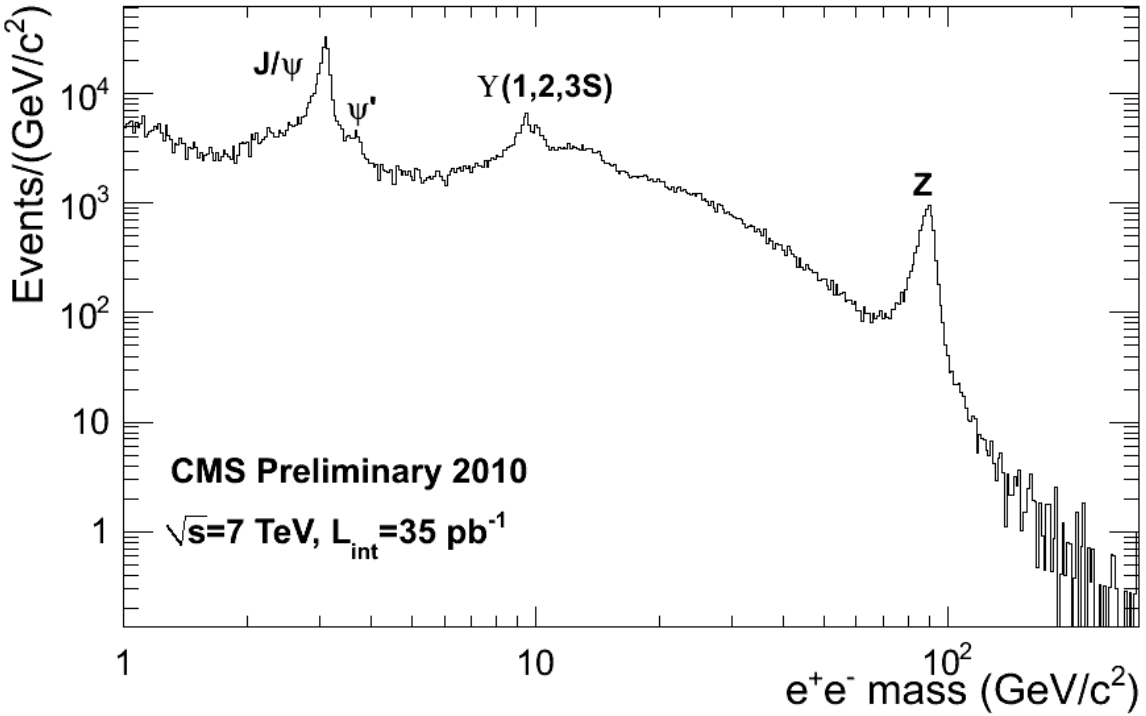
How to build a control sample ?
- Depending on the observable and the process, it can be easier to build control sample for the signal or the background
- This is really analysis dependent but there are some general rules
- One still have to rely on the Monte-Carlo to go from the control sample to the region of interest

Control samples : signal

Control samples for particle identification:

Signal control sample :

- **Usually use a resonance. Apply high quality cuts.**
- Electrons : Z→ee
- Photons : Z→ee (electrons / photons are somehow similar), Z→μμγ
- Muons : Z→μμ
- b-jets : top events



Control samples : background

Control samples for experimental particle identification:

Background control sample :

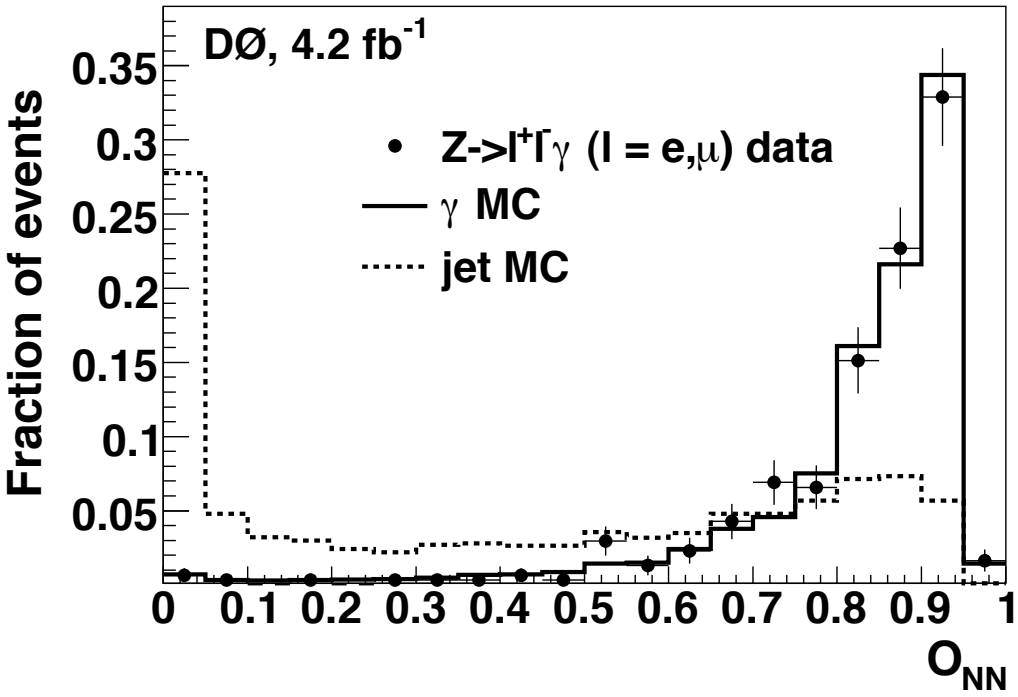
- **Cut inversion** to enrich the sample in background events (sideband method)
- Revert isolation cut
- Revert cuts on the shape of the electromagnetic energy deposit in the ECAL

Cut	Signal region	Sideband region
	Photon conversion method	
H/E	< 0.05	< 0.05
$\mathrm{Iso}_{\mathrm{TRK}}$ (GeV)	$< (2.0 + 0.001 E_{\mathrm{T}})$	$(2.0 + 0.001 E_{\mathrm{T}}) - (5.0 + 0.001 E_{\mathrm{T}})$
$\mathrm{Iso}_{\mathrm{ECAL}}$ (GeV)	$< (4.2 + 0.003 E_{\mathrm{T}})$	$< (4.2 + 0.003 E_{\mathrm{T}})$
$\mathrm{Iso}_{\mathrm{HCAL}}$ (GeV)	$< (2.2 + 0.001 E_{\mathrm{T}})$	$< (2.2 + 0.001 E_{\mathrm{T}})$
barrel: $\sigma_{\eta\eta}$	< 0.010	$0.010 - 0.015$
endcap: $\sigma_{\eta\eta}$	< 0.030	$0.030 - 0.045$
	Isolation method	
H/E	< 0.05	< 0.05
barrel: $\sigma_{\eta\eta}$	< 0.010	$0.0110 - 0.0115$
endcap: $\sigma_{\eta\eta}$	< 0.028	> 0.038

Control samples : examples

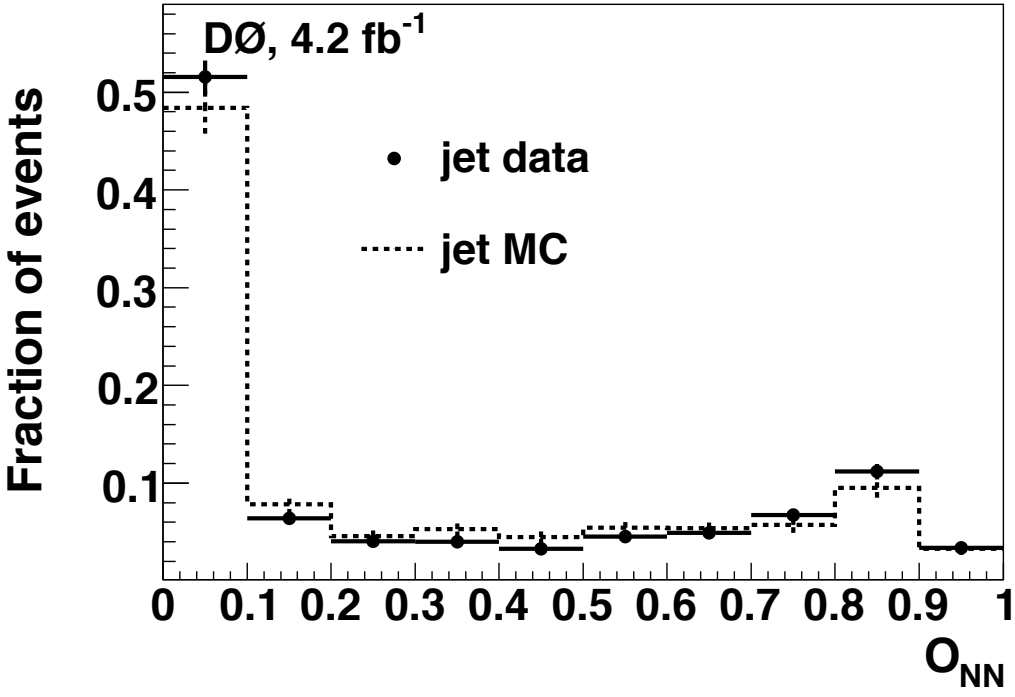
D0 photon identification with NN

Photon control sample



- Z→llγ selection

Jet control sample



- Photon selection +
Isolation cut inverted

Estimating systematics

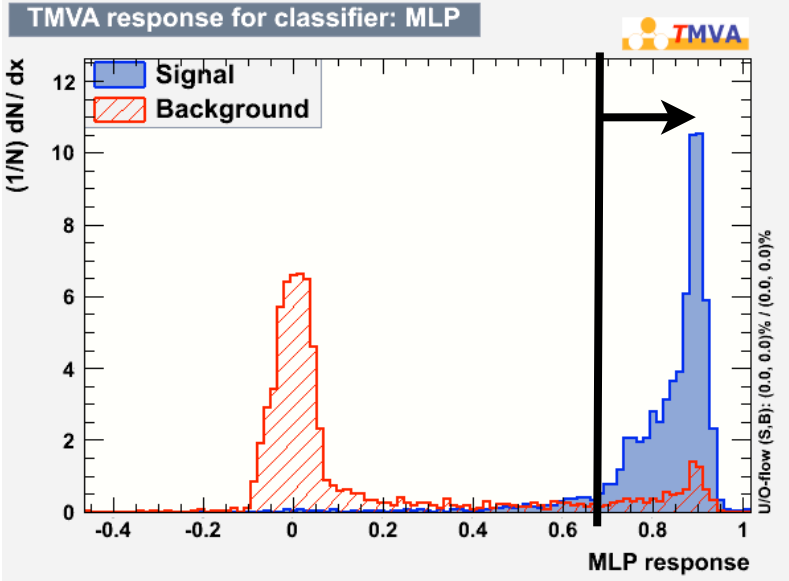
- Perform the training. This defines the classifier (set of weights, input variables)

- **Usual cases of the signal/background discrimination :**
  - **Cut** on the MVA output
  - **Categories**
  - Using the **shape**
- At each time a different way of dealing with systematics

- For **particle identification**, systematics are usually estimated from a control sample in data
- For **kinematics,** control samples can be checked but are rarely used to estimate the systematics. Indeed : what sample to use for e.g. Higgs kinematics ?

- Systematic uncertainty estimated from control samples turn out to be statistical uncertainty on this control sample

Uncertainties : cut on MVA output



The simplest use of a classifier is to **cut on the output**

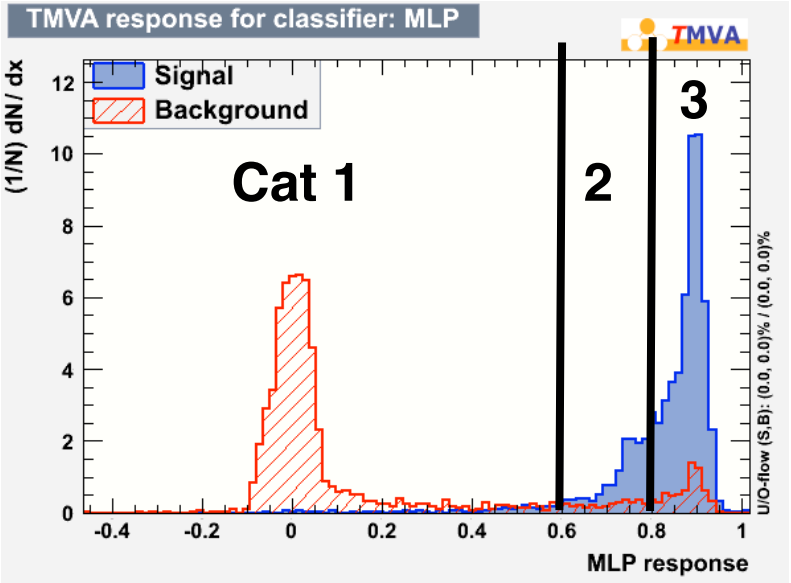
- To select the “signal region”, enhances s/b ratio
- The uncertainty comes only from this cuts : **uncertainty on selection efficiency for signal (and background)**

- To estimate the uncertainty, e.g. for particle identification one can use **control samples**.
- E.g. for photon identification. Use Z→ee in data and MC. Difference is used to correct the efficiency from data. Systematic is the signal efficiency difference between Z→ee and Photon MC.
- The same can be done for the background with jets faking photons (not obvious to build a non-biased control sample however...)

Uncertainties : categories

Categories :

- Events are divided in several categories
- E.g.: NNoutput<0.6,
  0.6<NNoutput<0.8,
  NNoutput>0.8
- Extension of cut (cut can be seen as one category)



Uncertainty for categorization :

- **Category migration** : possible migration of events in data from the bin where it is expected in MC to another because of mismodeling.
- Category migration depends on the slope of the distribution at the cut
- Estimated by varying up and down parameters => changes input distributions => impact the output and the selection efficiency in each bin
- Alternatively, control samples can be used to give ‘low’ and ‘high’ distributions

Uncertainties : output shape

What do we call shape ?

- Categories can be seen as binned shapes. Usually we select this category and then look at other observable to compute the sensitivity.
- But the whole (unbinned) shape, is used if 1) the classifier is the input of another classifier 2) if the classifier output is used to compute the analysis sensitivity (CLs method, exclusion or discovery)

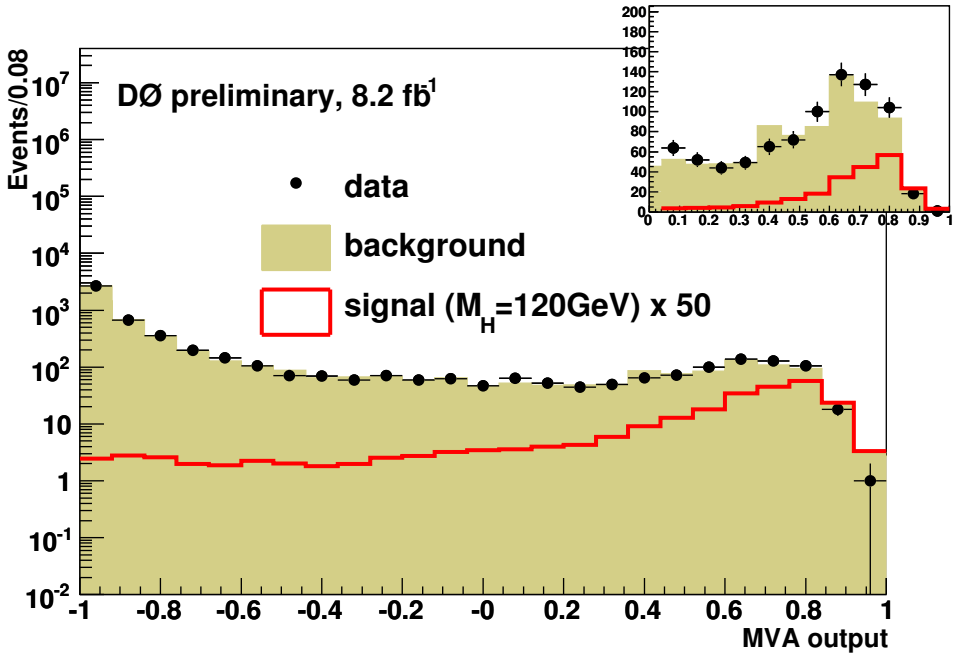
- Estimating the **uncertainty on a shape** is not an easy task
- Solution commonly accepted : varying the input distributions according to reasonable or meaningful values of parameters
- One obtains different output distributions

- **Experimental** uncertainties : **control samples**
- **Theory uncertainties**. Varying the renormalization/factorization scales => vary the shapes of the kinematical variables

Note on the signal region

Extra-care is needed for the signal region!

- Especially for kinematics MVA, generally no control sample
- This region drives the analysis sensitivity
- E.g. in the case of D0 H->2photons searches, the background shape is measured from the sidebands.



(c) $M_H = 120$ GeV