Alma Mater Studiorum - Università di Bologna

Facoltà di Scienze Matematiche Fisiche e Naturali
Dottorato di Ricerca in Fisica, XX Ciclo

PhD Thesis

Evaluation of a multi variated analysis for the selection of t-tbar multijet events in the current CMS implemented environment at LHC

Dott. **William Bacchi**

Advisor:
Chiar.mo Prof.
Paolo Capiluppi

PhD Coordinator:
Chiar.mo Prof.
Fabio Ortolani

Co-advisor:
Chiar.mo Prof.
Andrea Castro

FIS/01

Bologna, Italy, March 2008

Alma Mater Studiorum - Università di Bologna

Facoltà di Scienze Matematiche Fisiche e Naturali
Dottorato di Ricerca in Fisica, XX Ciclo

PhD Thesis

Evaluation of a multi variated analysis for the selection of t-tbar multijet events in the current CMS implemented environment at LHC

Dott. **William Bacchi**

Advisor:
Chiar.mo Prof.
Paolo Capiluppi

PhD Coordinator:
Chiar.mo Prof.
Fabio Ortolani

Co-advisor:
Chiar.mo Prof.
Andrea Castro

FIS/01

Bologna, Italy, March 2008

Contents

Introduction

The Standard Model has been verified deeply and accurately in many experiments in several laboratories.

According to the Standard Model the origin of particles mass is due to the Higgs Mechanism, which implies the existence of a new scalar boson, yet unobserved.

LHC is designed to collide protons at 14 TeV of center of mass energy with an instantaneous luminosity of $\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and will scan all the theoretical admitted mass range for the Higgs Boson.

LHC will also explore the energy region where Beyond the Standard Model (BSM) theories predict the existence of new particles.

The top quark, discovered at Fermilab in 1995, will be produced in large quantities at LHC, and this huge statistics (8 million $\mathrm{t\bar{t}}$ pairs per year) will permit measurements of top quark properties with a very high precision.

According to the Standard Model the top quark decays via $\mathrm{t} \longrightarrow \mathrm{W\,b}$ almost exclusively and the golden channel for the top physics is the leptonic or semileptonic channel where at least one of the W's decays in a lepton (e or μ) because leptons (especially muons) have a clear signature at CMS.

The fraction of top quarks decaying without leptons (46%) yields a large amount of events which are worth investigating, because in these events all the energy stops into the detector.

The limit of the fully hadronic decay channel is the overwhelming background from QCD multi-jet events in absence of a clear leptonic signature.

What we are going to do in this thesis is to calculate some kinematical variables and to use them to discriminate between QCD background and $\mathrm{t\bar{t}}$ signal. Discriminating two classes of event using several variables is the object of Multi Variate Analyses.

In the present study three classifiers will be trained using the event variables and their output will be used to calculate the statistical uncertainty expected on the cross section measurement of $t\bar{t}$ production.

Chapter 1

Top quark Physics

1.1 The Standard Model

The Standard Model (SM) is a quantum field theory that describes elementary particles and their interactions.

The world described by SM is populated by interacting fermions (table 1.1) (matter fields of spin[1] $\frac{1}{2}$.) and bosons (table 1.2) (gauge fields of spin 1).

Fermions are distinguished into leptons and quarks and are grouped in three generations with identical quantum numbers and different masses.

SM interactions include Electromagnetic interactions, Weak interactions and Strong interactions.

The remaining known interaction, the Gravitational one, is negligible in the interplay between elementary particles, except at very high energies, well beyond the accelerator scale.

The Electromagnetic interaction manifests itself through the exchange of a massless electrically neutral scalar boson (photon γ) between charged particles.

The Weak interaction manifests itself through the exchange of electrically neutral and charged massive vector bosons (W^+,W^-,Z^0) between fermions.

The Strong interaction manifests itself through the exchange of electrically neutral, massless vector bosons (gluons) carrying a strong charge, which can be of three types (colors), between quarks and gluons. Quarks can be grouped in color triplets, while gluons can be grouped in color octets.

[1]Spin values are in units of $\hbar$

Even if at present universe temperature Weak interactions and Electromagnetic interactions are very different, they have been recognized to be manifestations of a more general interaction, called the Electroweak interaction, present in early stages of the universe when temperature was much higher.

To explain this difference, we must first notice that in physics there are many examples of systems where the symmetry is broken by the presence of a field that acquires an expectation value and creates a non invariant vacuum state (Ferromagnetic materials, heated layers of fluid, etc.)

In SM this symmetry is broken by adding terms to the Lagrangian in a way that it remains locally gauge invariant (local gauge invariant theories have interesting properties like renormalizability), but the vacuum state acquires an expectation value which is not invariant.

The Standard Model Lagrangian can be split in three terms

$$L_{SM} = L_{gauge} + L_{Higgs} + L_{Yukawa} \tag{1.1}$$

where the gauge term describes the interactions, the Yukawa term contains the coupling between fermions, and the Higgs term

$$V(\Phi) = -\mu^2\Phi^+\Phi + \lambda(\Phi^+\Phi) \tag{1.2}$$

is responsible for the Spontaneous Symmetry breaking and for fermion and bosons masses.

In this way fundamental fermions and gauge bosons of weak interaction acquire a mass, while photons remain massless and last but not least a new term describing a massive scalar particle appears, the Higgs Boson.

The Higgs Boson observation is the main reason for the LHC construction, but at LHC energy and luminosity scale, there will be also a huge Top quark production that will permit to study this particle properties with a very high precision.

1.2 Top quark in Standard Model

The Top quark in the Standard Model is a spin 1/2 and charge 2/3 fermion; it has a color charge and it is the weak isospin partner of the bottom quark.

Family	Quarks			Leptons		
	name	symbol	charge	name	symbol	charge
I	up	u	$+\frac{2}{3}$	electronic neutrino	ν_e	0
	down	d	$-\frac{1}{3}$	electron	e^-	-1
II	charm	c	$+\frac{2}{3}$	muonic neutrino	ν_μ	0
	strange	s	$-\frac{1}{3}$	muon	μ^-	-1
III	top	t	$+\frac{2}{3}$	tauonic neutrino	ν_τ	0
	bottom	b	$-\frac{1}{3}$	tau	τ^-	-1

Table 1.1: Standard Model Fermions

Interaction	Boson (S = 1)	Mass (GeV/c^2)	electric charge (e)
electromagnetism	Photon γ	0	0
weak	Vector boson W^+	80.4	+1
	Vector boson W^-	80.4	-1
	Vector boson Z^0	91.19	0
strong	Gluons g	0	0

Table 1.2: Standard Model Bosons

Top quark decays with strange and down quarks are negligible respect to final states with bottom quarks because the corresponding elements of the Cabibbo-Kobayashi-Maskawa mixing matrix V_{ts} and V_{td} can be estimated to be less than 0.043 and 0.014.

The width predicted in Standard Model with decays dominated by the channel $t \longrightarrow Wb$ is :

$$\Gamma_t = \frac{G_F m_t^3}{8\pi\sqrt{2}} \left(1 - \frac{M_W^2}{M_t^2}\right)^2 \left(1 + 2\frac{M_W^2}{M_t^2}\right) \left[1 - \frac{2\alpha_s}{3\pi}\left(\frac{2\pi^2}{3} - \frac{5}{2}\right)\right] \tag{1.3}$$

As we can see, the Top quark width is related to the Top quark mass and increases with it (from 1.02 GeV/c^2 for a Top quark mass of 160 GeV/c^2 to 1.53 GeV/c^2 for a Top quark mass of 180 GeV/c^2).

Due to his width the Top quark has a very short lifetime of about 0.5×10^{-24} s and decays before hadronization.

The final states can be divided into three classes depending on the decay of the 2 W bosons:

$$\begin{array}{ll} t\bar{t} \longrightarrow W^+bW^-\bar{b} \longrightarrow q\bar{q}'bq''\bar{q}'''\bar{b} & \text{with BR} = 46.2\ \% \\ t\bar{t} \longrightarrow W^+bW^-\bar{b} \longrightarrow q\bar{q}'b\ell\bar{\nu}_\ell\bar{b} + \bar{\ell}\nu_\ell bq\bar{q}'\bar{b} & \text{with BR} = 43.5\ \% \\ t\bar{t} \longrightarrow W^+bW^-\bar{b} \longrightarrow \bar{\ell}\nu_\ell b\ell'\bar{\nu}_{\ell'}\bar{b} & \text{with BR} = 10.3\ \% \end{array}$$

where the quarks in final state evolve into jets of hadrons.

The three channels are called all-jets, lepton+jets (ℓ + jets) and dilepton ($\ell\ell$) channels.

Leptons and missing energy are a very good signature for $t\bar{t}$ events. This indicates leptonic channels as golden channels for Top quark analysis.

The all-jets channel, although challenging, is worth to be investigated because it is the most populated and has also the advantage that all the decay products should stop in the detector.

As I said, in the all-jets channel we do not have leptons or missing energy; the problem is then how to reject the huge QCD background with no clear signature.

What I am going to investigate in this thesis is how tools for multivariate analysis can help in enhancing the signal/background ratio.

1.3 Beyond the Standard Model

The Standard Model predictions have been deeply verified in the last years experiments.

Anyway some evidences on neutrino masses and other conceptual problems make many physicists believe that the Standard Model is just a low energy limit of a more general theory.

Recent evidences of neutrino flavor oscillations require different masses for neutrinos, which the Standard Model assume massless ([1],[2]).

The Standard Model is a theory with about 20 free parameters (too many for a fundamental theory) and three families of elementary particles (while ordinary matter contains only particles from the lighter family).

The charge of down-type quarks is exactly $\frac{1}{3}$ of the charge of electrons, making the standard matter stable, but the Standard Model cannot predict this value.

At a cosmological scale, there are astrophysical observations which cannot be explained just with the matter known in the Standard Model. This *Dark Matter* can be described only by new theories.

The gauge couplings of the three fundamental interactions depend on the energy at which the interaction occurs. The Weak and Strong interaction coupling constants decrease with the energy scale while the Electromagnetic coupling increases with the energy scale. Near the Plank Scale the three coupling constants are very close, but for the Standard Model they don't converge exactly and we cannot have the unification.

The *hierarchy problem* of the Standard Model is the fact that even if there are large quantum quadratic corrections to the square of the Higgs Boson Mass, this mass is not so huge, requiring a fine tuning of parameters to admit a precise cancellation of quadratic radiative corrections and bare mass.

The Standard Model does not include Gravitation.

All these reasons give us hope that something new will appear under our eyes: what we only have to do is to look in all directions ready to see what we were not expecting.

Chapter 2

The LHC Project and the CMS Detector

2.1 Large Hadron Collider

The Large Hadron Collider is the Proton - Proton Accelerator actually under construction at CERN. It is being built in a circular ring of 27 Km in circumference which is buried from 50 m to 175 m underground. Along the ring there are four interaction points where main LHC experiments are placed (Fig. 2.1):

CMS The Compact Muon Solenoid
ATLAS A large Toroidal LHC Apparatus
ALICE A Large Ion Collider Experiment
LHCb Large Hadron Collider beauty experiment

The first beams should circulate in May 2008 and the first collisions at high energy should happen in mid-2008.

Proton - Proton beams will collide at an energy of 14 TeV in the center-of-mass while heavy ions (Pb - Pb) beams will collide at an energy of 2.76 TeV per nuclear pair.

The beams of protons will circulate in separate beam pipes and they will cross at the interaction points every 25 ns, reaching a crossing frequency of 40 MHz.

The beams are organized in bunches which contains about 1.1×10^{11} protons

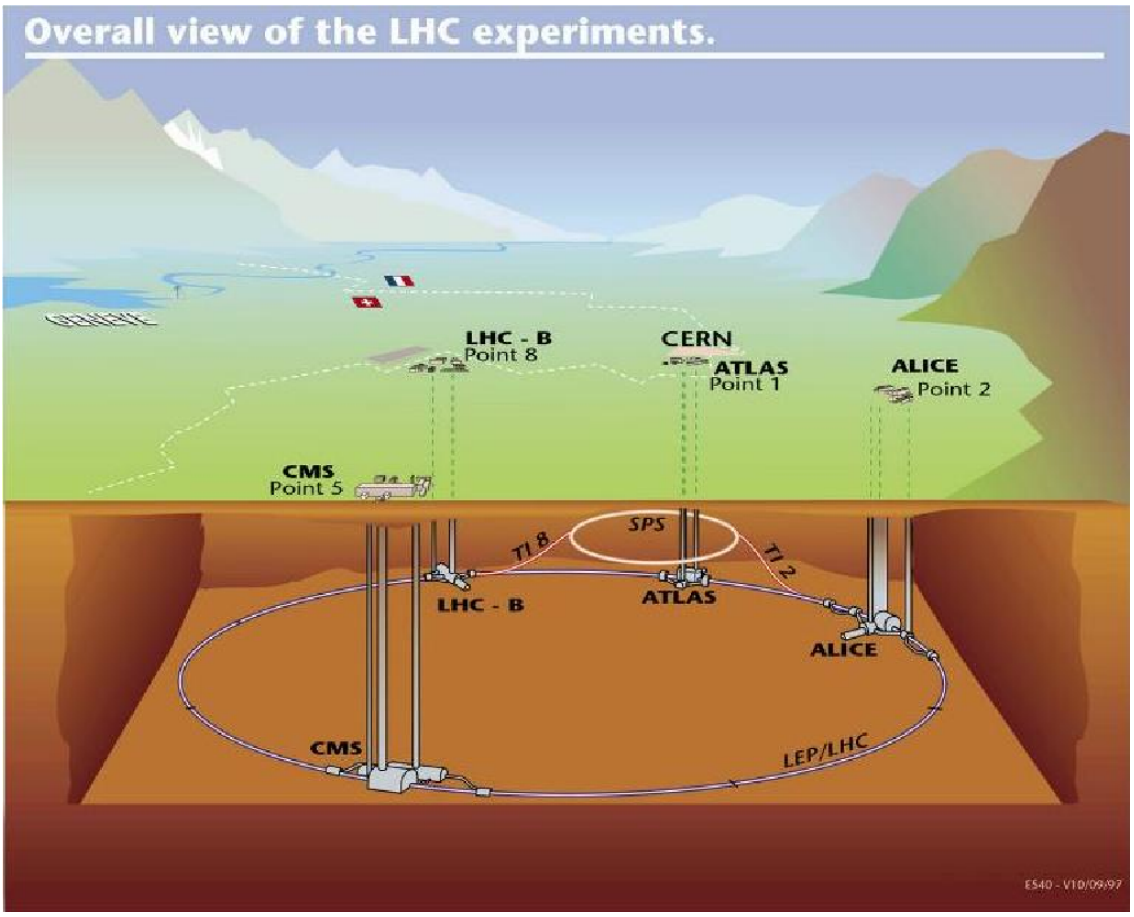


Figure 2.1: LHC and the four collision points.

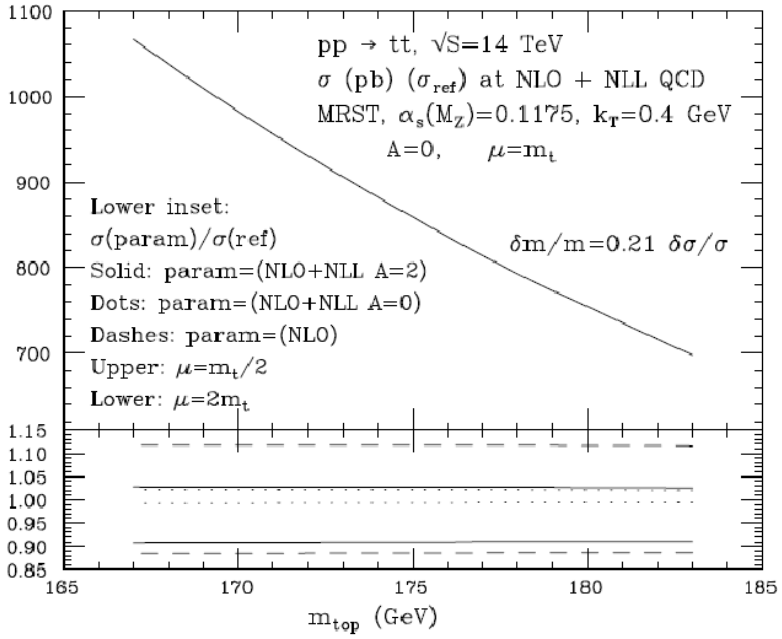


Figure 2.2: $t\bar{t}$ Total Cross Section scale dependence

and which have a transverse size of about 15 μm and a size along the beam axis of 53 mm.

For each bunch crossing at the low luminosity $\mathcal{L}$ of $10^{33} cm^{-2} s^{-1}$we expect an average of 4 concurrent interactions.

2.2 $t\bar{t}$ production at LHC

Since LHC is a Proton-Proton collider, the main source of top production is through gluons fusion($\approx 90\%$) and quark-antiquark annihilation($\approx 10\%$).
Theoretical uncertainties can be due to renormalization and factorization scale variations [4].

Fitting the distribution seen in Fig. 2.2, we can see that a 5% precision in cross section measurement is equivalent to a 1% precision in the mass measurement and so a 5% in the precision of the cross section measurement should be a minimal requirement.

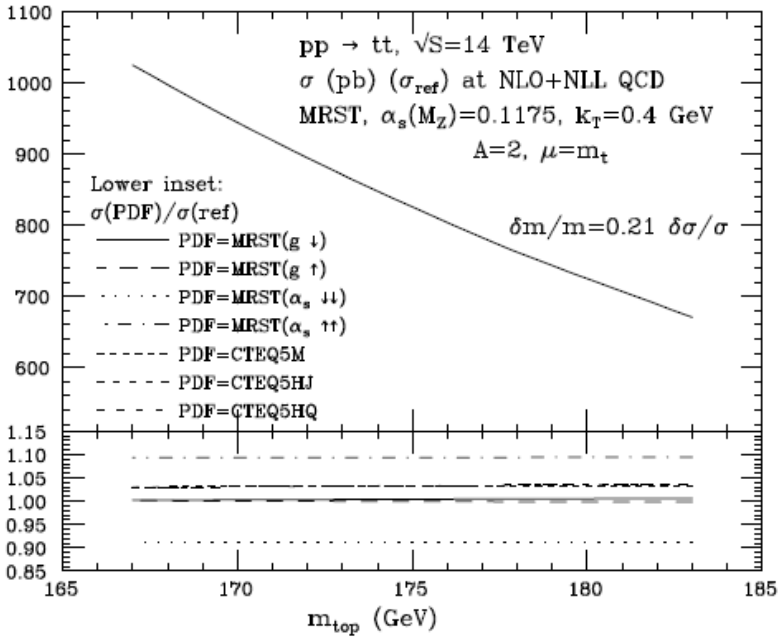


Figure 2.3: $t\bar{t}$ Total Cross Section PDF dependence

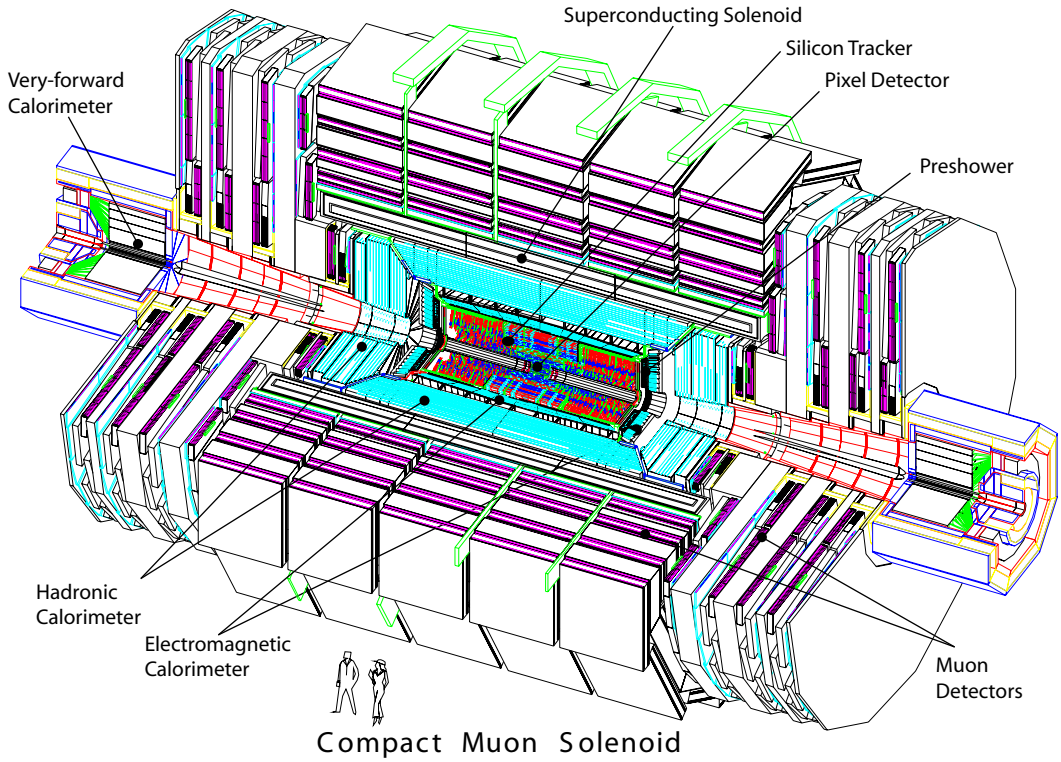
Another source of theoretical systematic uncertainty is the total cross section dependence on Parton Distribution Function(Fig. 2.3).
The total systematic theoretical uncertainty resulting from the combination of PDF dependence and scale dependence is of 12% corresponding to a 4 GeV/c^2 uncertainty on the determination of the top quark from the total cross section.

2.3 CMS Detector

The Compact Muon Solenoid (CMS) is a general purpose detector which will operate at LHC. It is hosted in a cave 100 m under the city of Cessy in France.

CMS is 21.6 m long, its diameter is of 14.6 m, and weights 12500 t. The core feature driving the CMS design is the high field solenoid needed to generate a 4 T magnetic field.

Such a high magnetic field is needed to bend charged high energy particles



enough to measure with a sufficient precision their momentum.

Inside the magnet coil is hosted the inner tracker and the calorimetry. In the outer iron return yoke are hosted muon stations consisting of drift tubes (DT) in the barrel region and of cathode strip chambers (CSC) in the endcap region completed with resistive plate chambers (RPC).

All these muons detectors ensure robustness and full geometric coverage.

2.3.1 Magnet

The CMS superconducting magnet is 12.5 m long and has a diameter of 6 m.

It has been designed to reach a 4 T field and it will store an energy of 2.6 GJ at full current. To reach a 4 T field the winding is made of 4 layers instead of the usual 1 or 2. The CMS coil is a thin coil (the radial extent of the coil is small compared with the radius). The CMS magnet uses a self supporting conductor to provide the

necessary hoop strength.

The iron return yoke is composed of 5 barrel wheels and 6 endcap disks, for a total weight of 10 000 t.

These yoke elements host the muon stations and the magnetic return flux in this region (magnetic field can reach 2 T inside muon detectors) is high enough to provide muon bending which can be used to perform a muon trigger based on the $\hat{p_T}$.

2.3.2 Tracker

The Tracker system will be hosted inside the magnet coil in a homogeneous magnetic field.

It is designed to provide a precise and efficient measurement of charged particle trajectories and a precise reconstruction of secondary vertexes.

It has a length of 5.8 m and a diameter of 2.5 m.

With about 200 m^2 of active silicon area, the CMS tracker is the largest silicon tracker ever built.

Inside the tracker there will be a very intense particle flux (about 1000 particles every 25 ns). The requirements of a tracker system with a high granularity, a high speed and a sufficient radiation hardness lead to a silicon based detector.

The tracker system is made of a pixel detector and of a silicon strip detector and will be cooled down to -10 °C to enhance radiation hardness.

The expected resolution in transverse momentum for this system is around 1 - 2 % up to $-1.6 < \eta < 1.6$, while the expected resolution in transverse impact parameter can reach 10 μm for high-p_T tracks.

Pixel detector

The pixel system is the closest detector to the interaction region. It is important for a good secondary vertex 3D resolution.

The pixel detector covers a pseudo rapidity range $-2.5 < \eta < 2.5$.

The pixel cell size is of 100×150 μm^2 and permits to have precise tracking points in $r - \phi$ and z.

The pixel detector consists of three barrel layers and two endcap disks.

The disposition of barrels and endcaps gives 3 tracking points over almost the full η range.

Silicon strips detector

The silicon strip detector system is made of 10 barrel detection layers and 3 + 9 endcaps disks.

The four inner barrels realize the Tracker Inner Barrel (TIB) closed with three Tracker Inner Disks (TID).

The outer six barrels realize the Tracker Outer Barrel (TOB) closed with nine disks (Tracker End Cap).

The minimal unit are 24 244 silicon strips, organized in 15 148 modules of varying shapes and sizes to match resolution and geometrical requirements.

2.3.3 Electromagnetic Calorimetry

The Electromagnetic Calorimeter of CMS is made of lead tungstate crystals.

Crystals cross-section is approximately of 0.0174×0.0174 in $\eta - \phi$ corresponding to 22×22 mm^2 at the front face of the crystal, and 26×26 mm^2 at the rear face. The crystal length is of 230 mm, corresponding to 25.8 X_0.

The Calorimeter is made of a central barrel (EB) part with 61 200 crystals and two endcaps (EE) of 7 324 crystals each. In front of the endcaps are placed preshower detectors for neutral pions rejection.

The EB covers $|\eta| < 1.479$ while the EE covers $1.479 < |\eta| < 3.0$.

The photodetectors used are the avalanche photodiodes for the barrel region and vacuum phototriods for the endcaps.

The use of high density crystals makes this calorimeter fast, with a fine granularity and good radiation resistance. Its good energy resolution (as provided by a homogeneous crystal calorimeter) enhances its chances to detect the decay of Higgs Boson to two photons.

The energy resolution for this Electromagnetic Calorimeter system can be described by equation

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2 \tag{2.1}$$

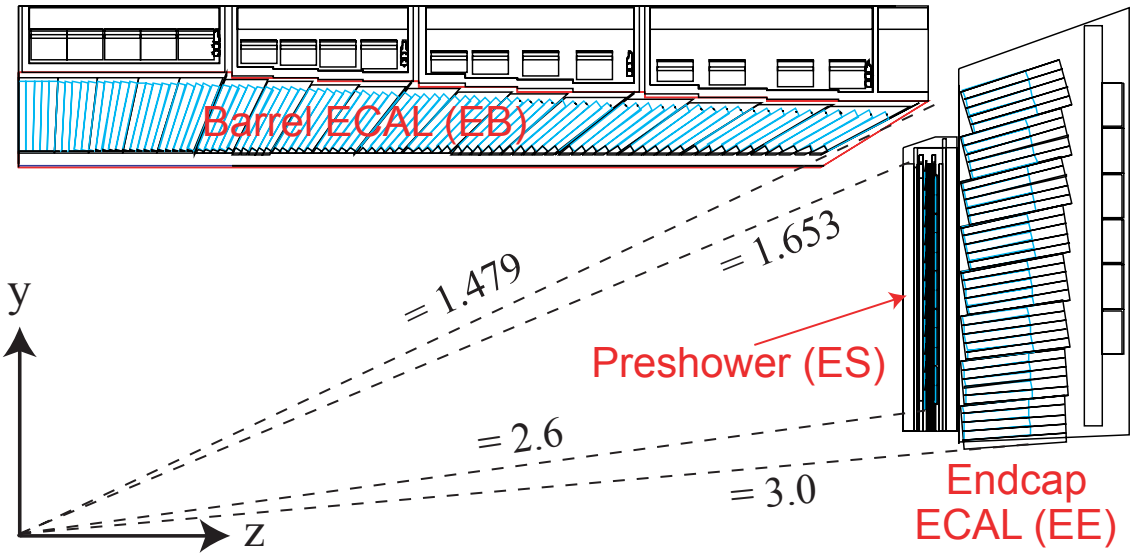


Figure 2.4: Transverse section of the ECAL

where S is a stochastic term which depends on photostatistics, lateral shower containment and fluctuation in energy deposited in the preshower absorber.

N is the noise term, with contributions coming from electronic noise, digitization noise and pileup.

C is a constant term due to non uniformity in the longitudinal light collection, intercalibration errors and leakage of energy from the back of the crystals.

2.3.4 Hadronic Calorimetry

The Hadron Calorimeters (HCAL) are very important for the measurement of hadron jets and transverse missing energy due to neutrinos or exotic particles.

The Hadron Calorimeter Barrel and Endcaps are placed between the Electromagnetic Calorimeter and the inner of the magnet coil.

This collocation limits the total amount of material which can be put to absorb the hadronic shower. This is the reason why an outer calorimeter is placed outside the solenoid.

The η range covered by the Hadron Barrel (HB) calorimeter is up to $|\eta| < 1.3$. To extend this range to $|\eta| < 5.2$ a Forward Hadronic calorimeter is added outside

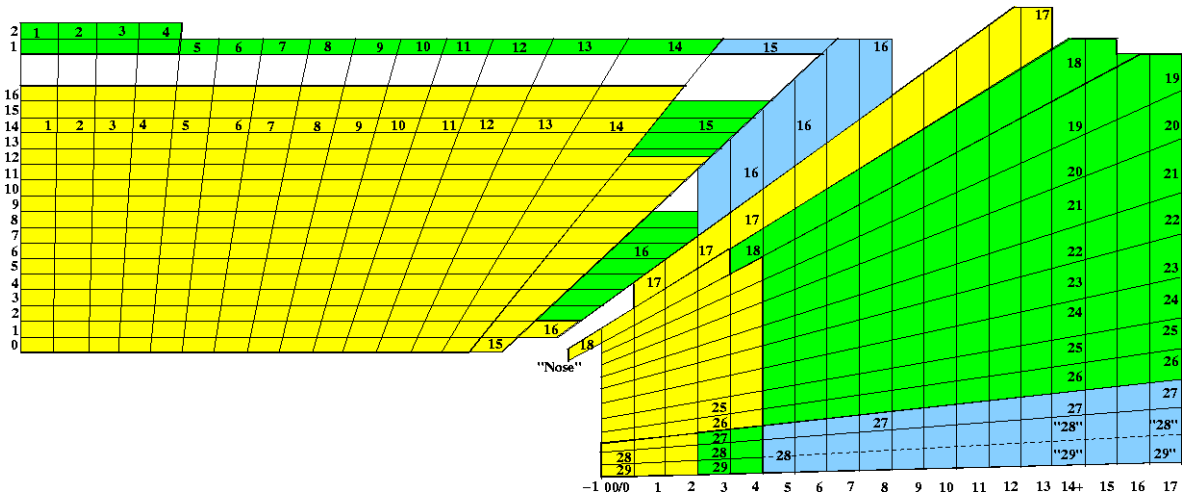


Figure 2.5: A schematic view of the tower mapping in r-z of the HCAL barrel and endcap regions.

the Barrel calorimeter Endcaps.

The HB consists of 36 azimuthal wedges which forms two half barrels HB+ and HB-. Each Wedge is composed of absorber plates (one 40-mm-thick front steel plate, eight 50.5-mm-thick brass plates and one 75-mm-thick back steel plate) and active scintillating medium. The total absorber thickness is 5.82 interaction lengths (λ_I) at 90° and increases with polar angle θ as $\frac{1}{\sin\theta}$ and reaches a value of $10.6\lambda_I$ at $|\eta| = 1.3$.

The plastic scintillator is divided into 16 η sectors, resulting in a segmentation $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$.

The energy resolution of HCAL is $\frac{\sigma_E}{E} \approx (65\%/\sqrt{E} \oplus 5\%)$ in the barrel and $\frac{\sigma_E}{E} \approx (83\%/\sqrt{E} \oplus 5\%)$ in the endcaps.

2.3.5 Muon System

The CMS Muon System is designed to reconstruct the momentum and charge of muons over the entire kinematic range of LHC.

In the barrel region (Fig. 2.7)we can use drift chambers with standard rectangular drift cells because the magnetic field is mostly contained in the return yoke and

the neutron-induced background is small. The Barrel Drift Tube (DT) chambers cover the pseudorapidity range $|\eta| < 1.2$ and are organized into 4 stations.

The first 3 stations contain 8 chambers divided in 2 groups which measure the muon coordinate in the $r - \phi$ plane and 4 chambers which provide a measurement in the z direction. The fourth station does not contain the z-measuring planes.

The two set of chambers in each station are separated as much as possible to achieve the best angular resolution.

The drift cell of each layer are offset by a half-cell width with respect to their neighbor. In this way we eliminate dead spots and we can measure muon time with sufficient resolution to obtain standalone bunch crossing identification.

In the endcap regions the muon system uses Cathod Strip Chambers (CSC), which with their fast response time, fine segmentation, and radiation resistance can identify muons between $0.9 < |\eta| < 2.4$ in a high field region and with a high rate of muons and background.

Each endcap is made of 4 stations of CSC. The chambers are perpendicular to the beam line and the cathod strips provide a precision measure of the $r-\phi$ bending plane. The anode wires are read to obtain a beam-crossing time measure of muons as well as a η measurement.

In the muon system there is also a Resistive Plate Chambers system (RPC) which can combine adequate spatial resolution with an excellent time resolution comparable to that of scintillators.

With these characteristics the RPC can improve the muon trigger efficiency and can provide an unambiguous assignment of the bunch crossing.

Drift Tubes

A drift tube chamber (Fig. 2.8) is made of 3 or 2 superlayers (SL).

Each superlayer is made of 4 layers of rectangular drift cells staggered by half a cell. The SL is the smallest independent unit of the design.

Each cell is 42×13 mm^2 and has a stainless anode wire with diameter 50 μm and a length varying from 2 to 4 meters depending on the station. Cells are filled with an $Ar - CO_2$ mixture.

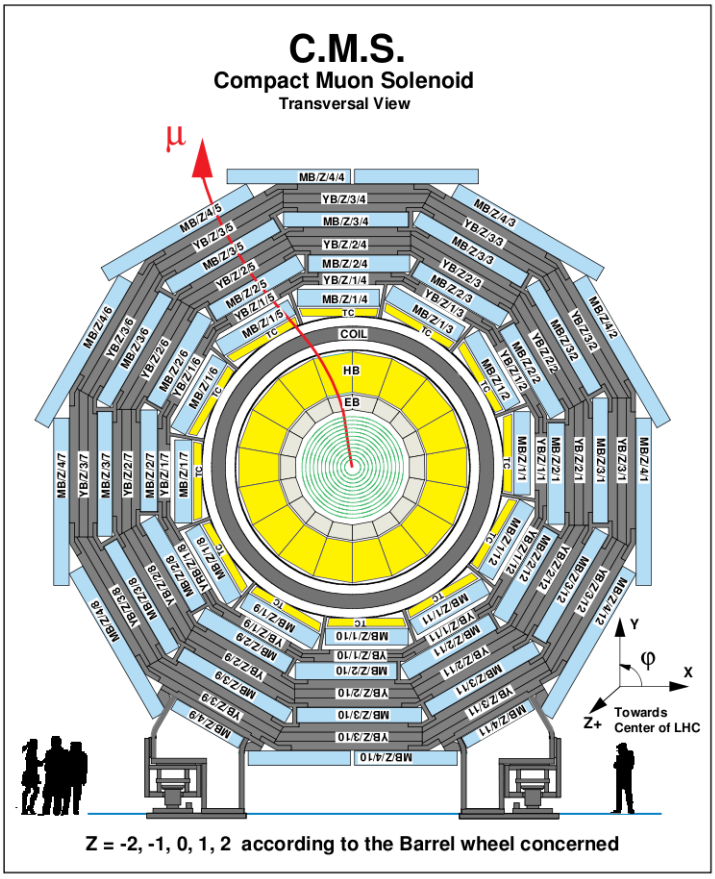


Figure 2.6: Transverse view of the CMS Detector showing the Muon Stations in the barrel Region.

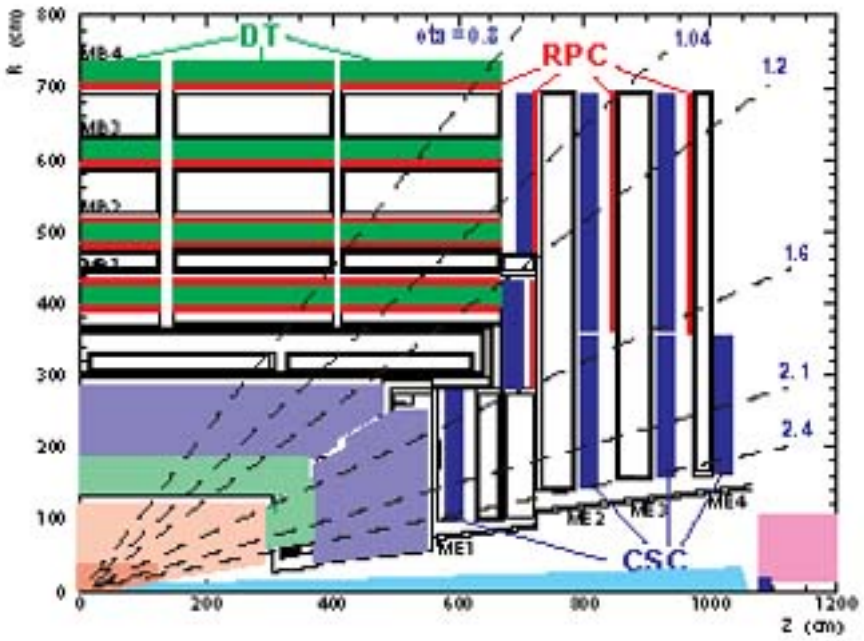


Figure 2.7: Transversely view of the CMS Detector showing the Muon Stations in the barrel Region.

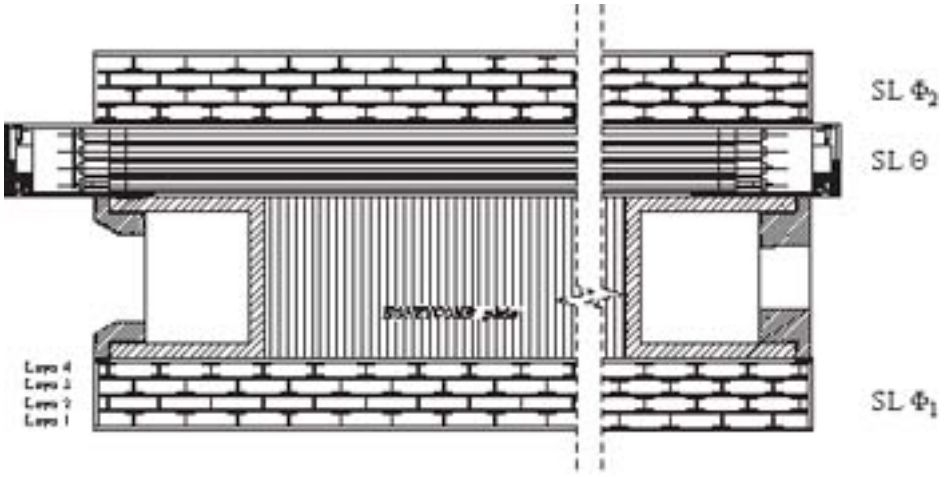


Figure 2.8: The layout of a DT chamber inside a muon barrel station

Cathode Strip Chambers

CSC are multiwire proportional chambers with segmented cathod readout.

Chambers are trapezoidally shaped (Fig. 2.9) with a maximum dimension of 3.5×1.5 m^2 and have 6 anode wire planes and 7 cathode panels. The gas mixture filling CSC is made of $40\%Ar + 50\%CO_2 + 10\%CF_4$

Resistive Plate Chambers

RPC are gaseous parallel plate detectors.

The CMS RPC consists of 2 gaps (up and down gaps) operated in avalanche mode with common readout strips.

This double gap configuration allows the single gap to operate at low gas gain with an effective detector efficiency higher than for a single gap. The gas filling RPC is a mixture of 96.2% $C_2H_2F_4$, 3.5% iC_4H_{10} and 0.3% SF_6.

2.3.6 Trigger and Data Acquisition System

LHC provide high rate proton-proton collisions. There will be one bunch crossing every 25 ns (40 MHz).

At a luminosity of $\mathcal{L} = 10^{34}$ cm^{-2} s^{-1} there will be an average of 20 interactions per bunch crossing.

Since it is impossible to store such an amount of data (100 TByte/s), a drastic rate reduction must be achieved. The CMS trigger system is responsible for this rate reduction.

The rate is reduced in two steps: Level-1 trigger (L1T) and High Level Trigger(HLT) (Fig. 2.10).

Level-1 trigger

The L1 trigger is a custom designed, highly programmable electronic system.

The output rate of L1 trigger has to be of about 100 kHz.

A pipelined processor architecture permits to store data on a temporary memory for a time of 3.2 μs corresponding to 128 bunch crossings before rejecting or accepting an event.

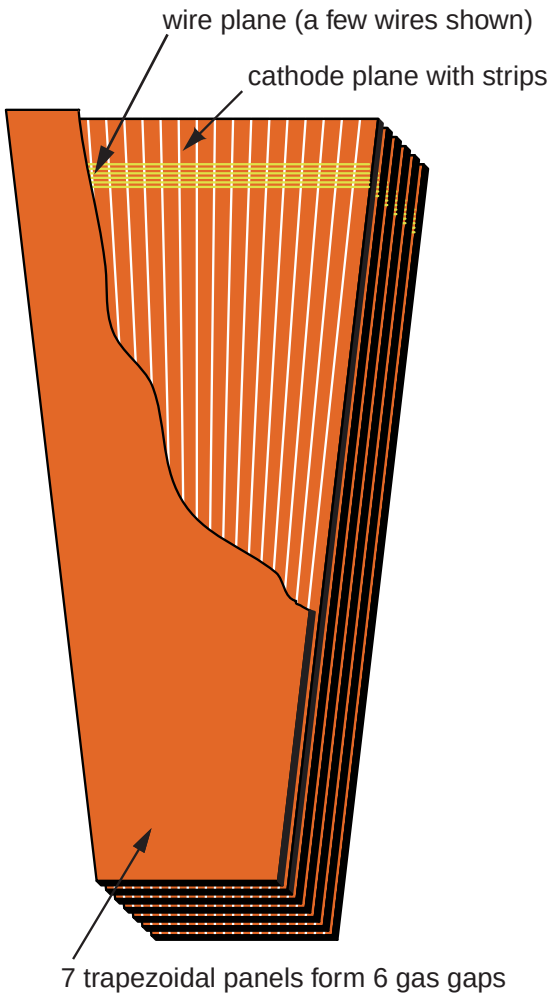


Figure 2.9: Schematic View of a CSC chamber

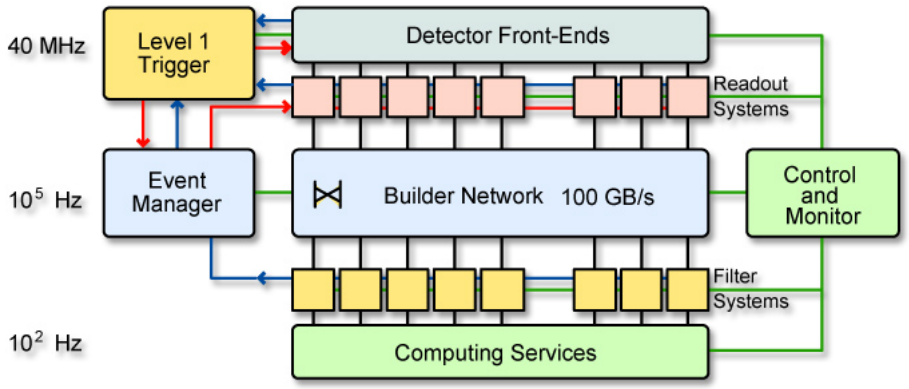


Figure 2.10: Architecture of the CMS DAQ System

High Level trigger

HLT is a Software system hosted in a processing farm of about 1000 processors and has access to the full event information.

The reduction factor of event rate for HLT should be above 10^3 to obtain the final 100 Hz rate corresponding to a data flux of 100 MByte/s.

Chapter 3

CMS Computing and Software

3.1 Computing Model

CMS has adopted a distributed computing model [5] in order to cope with the requirements for storage, processing and analysis of the huge amount of data the experiment will collect. In the CMS computing model, resources are geographically distributed and operated by means of Grid Software.

3.1.1 Tier Architecture

The CMS offline computing system has a tiered structure.

A single Tier-0 center, located at CERN, accepts data from the CMS On line Data Acquisition System, saves them on the permanent mass storage and performs first pass reconstruction. After that, data are distributed to a set of Tier-1 centers.

Tier-1 centers are located in CMS collaborating countries. They provide large CPU facilities and mass storage. Tasks carried out at Tier-1 are custodial of fractions of the experiment data and organized sequential processing of data and extraction of dataset to be sent to Tier-2 centers.

Tier-2 centers are hosted at CMS institutes and they provide computing resources for the activities of the physics groups. Tier-2 centers rely upon Tier-1 for access to large dataset and secure storage. Typical tasks performed at Tier-2 centers are Monte Carlo Simulation, calibration activities and final stage analysis.

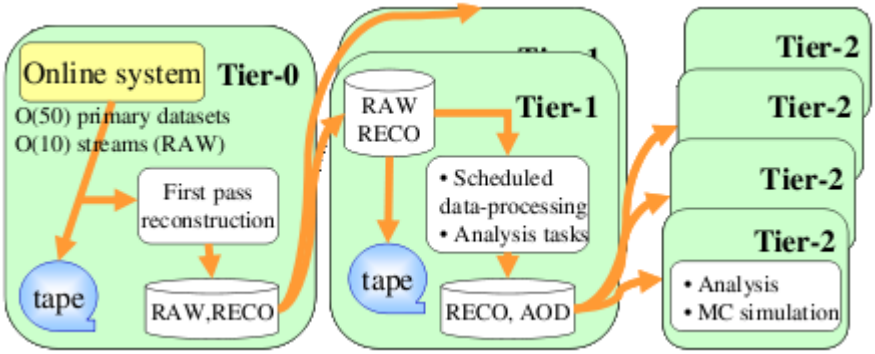


Figure 3.1: CMS Data Flow

3.1.2 Data Organization and Flow

CMS will use several data formats varying in size, detail and refinement.

RAW data will contain all detector data after on line formatting, the Level-1 trigger result, HLT selections. This data format is designed to have a size of 1.5 MByte/event (2 MByte/event for simulated data).

RECO data will contain all reconstructed objects (jets, tracks, muons, etc.) and reconstructed hits and clusters. This data format is designed to have a size of 0.25 MByte/event (0.4 MByte/event for simulated data).

AOD will contain Analysis Object Data. They are derived from RECO data to provide a convenient and compact format for physics analysis. This data format is designed to have a size of 0.05 MByte/event. They will contain all the high level physics objects and just a summary of the RECO information.

The Tier-0 will be accepting 225 MB/s of RAW data from the CMS Detector.

At the Tier-0 RAW data will be reconstructed and archived on tape at CERN and distributed to Tier-1 centers. At Tier1-centers will be produced AOD data and will be run additional processing of RAW, RECO and AOD data triggered by Physics Group requests. A fraction of the data produced at Tier-1 centers will be transferred to Tier2-centers, which support iterative analysis.

If we assume a rate of approximately 150 Hz as HLT output, the amount of RAW data produced will be of 4.5 PB/year (there will be two copies of raw data, one stored at the Tier-0 and another stored in a Tier-1 center). The total volume of

reconstructed data (RECO), including 3 reprocessing, will be of about 2 PB/year and the total volume of AOD data will be of about 2.6 PB/year (each Tier-1 center will store a whole copy of the AOD data).

3.1.3 Distributed environment: GRID

Since the CMS experiment has adopted a distributed computing model, CMS experiment needs GRID technologies.

The computational Grid concept refers to a reliable, flexible, secure resource sharing among Virtual Organizations.

A virtual organization is a group of Grid users who are able to work collaboratively with other members of the group and/or share resources (data, software, CPU, storage space, etc) regardless of geographical location.

CERN and LHC experiments decided to develop, build and maintain a distributed computing infrastructure, Worldwide LHC Computing Grid Project (WLCG)[6].

The WLCG architecture consists of a set of services and applications running on the Grid infrastructures provided by the WLCG partners. These infrastructures at the present consist of those provided by the Enabling Grids for E-sciencE (EGEE) project in Europe, the Open Science Grid (OSG) project in the U.S.A. and the Nordic Data Grid Facility in the Nordic countries.

Grid services can be summarized in

- Storage Element Services (SE) are Mass Storage systems providing predefined interfaces for data access.
- File Transfer Services are services that should provide a reliable data transfer.
- Compute Resource Services (CE) are a set of services to provide access to a local batch system.
- User Interface is the access point to GRID services. A user with an account and a certificate installed on the UI can be authenticated and authorized to access GRID Services through a command line interface and API libraries.
- Workload Management System is a pool of services responsible for the acceptance of job submits and the dispatching of those jobs to the appropriate CE depending on the job requirements and the available resources.

3.1.4 CMS data and workload management tools

CMS has built several computing services for data management and workload management that run on top of the Grid infrastructure. CMS has developed also tools to submit and monitor analysis jobs in a Grid environment.

Data Management

The CMS Data Management System [8] is based on a set of components that collaborate to provide the necessary functionality.

- Data Bookkeeping System (DBS) keeps data description and permits to execute queries to find available data and their logical organization in terms of files and file-blocks.
- Data Location Service (DLS) keeps the mapping between file blocks and Storage Elements where they are located.
- Local Catalog is responsible for translating logical file names to physical file path within the site itself.
- Data placement and transfer systems are implemented by the PhEDEx project [9]. PhEDEx manages allocations and releases of storage resources and the replication and movement from multiple sources to multiple destinations.

BOSS

BOSS (Batch Object Submission System)[10] has been developed in the context of CMS experiment to provide a common interface for submission, monitoring and output retrieval of jobs on the Grid or on a local farm. The key features of BOSS are the use of a local relational database (MySQL or SQLite at the moment) to store the job related information and the use of registrable scripts to perform typical batch system operations. Using registrable scripts BOSS can be adapted by a site administrator, to almost every batch system, local or not, implementing the specific scripts for job submission, deletion etc. Furthermore BOSS permits to the user the definition of job specific information to be retrieved from the standard output of

the job. Once these information are defined, BOSS retrieves and stores them in a specific table on the database when the output of the job is retrieved.

A BOSS task is a tree of jobs and each job can be a chain of several programs to be executed in a defined order (at the moment just a linear order is implemented).

BOSS can also be used to real-time monitor running jobs, using a Real Time Server, which is a service (currently a bare MySQL database) which has to be reachable from the worker node and from the client, via a registrable plug in.

BOSS is distributed with a working set of scripts for the more common batch systems and Grid interfaces and with an implementation of a real time client.

BOSS has been used as the interface toward the grid and the local batch system by several other tools for analysis and production (CRAB, PRODAGENT) during CMS official testing activities and Monte Carlo Productions.

CRAB

CRAB (CMS Remote Analysis Builder) is a tool for the physics analysis user which permits to develop locally the analysis code and then submit it to the GRID to reach data on distributed storage systems.

CRAB needs a configuration file where the user writes where the code is, which dataset he wants to analyze and other parameters as the number of events per job. With these information CRAB finds where the data files are located using the CMS Data Management infrastructure and creates the number of jobs required to analyze the whole dataset. At this point the user can submit jobs, check their status and retrieve the output.

Of course when the number of jobs becomes huge, submitting, checking, resubmitting can be a hard work. In order to automate as much as possible all the analysis operations, it is under development an Analysis Server that should be as much transparent as possible to the end user and that has the same CRAB interface. This Analysis Server takes care of jobs from the submission to the output retrieval, leaving to the end user just the preparation of the job and the output evaluation.

ProdAgent

The Monte Carlo production [12] is the activity that has most advantages from automation. Handling user requests, tracking jobs execution and monitoring the

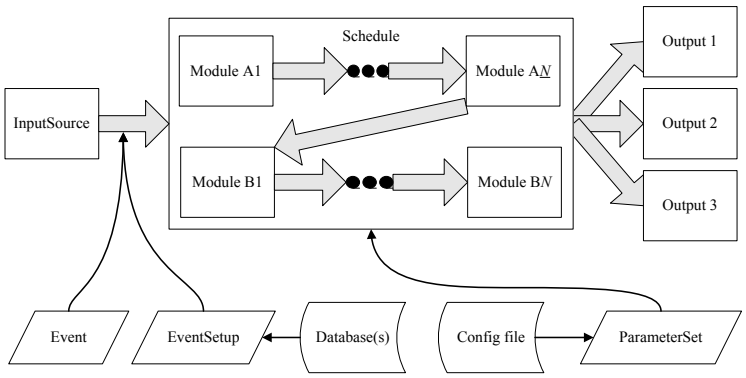


Figure 3.2: CMSSW Framework and EDM

production are the main task of a Monte Carlo Production tool.

CMS has developed a Monte Carlo Production system based on a tool for handling Physics Groups Requests and on a tool for distributing requests (Production Manager) among several instances of Production Agents.

The Production Agent is an automated system which takes requests from the Production Manager and prepares and submits jobs, resubmits them in case of failure, merges the output and registers completed productions in the CMS data management infrastructure (DBS-DLS).

3.2 Experiment Software

The CMS experiment software (CMSSW) is based on a framework and on an Event Data Model, physics software modules, services and utility tool kits.

3.2.1 CMSSW Architecture

The framework defines the top level abstractions, their behavior and collaboration patterns.

It comprises two components:

- a set of classes that capture CMS specific concepts and event features and a control policy that takes care of the flow of control, module scheduling,

input/output, etc.

- The physics and utility modules that are written by detector groups.

The physics modules communicate with each other only through the data access protocols that are part of the framework itself.

The service and utility toolkit consists of physics type services and computer services.

Both the application framework and the service and utility toolkit shield the physics software modules from the underlying technologies. This will ensure a smooth transition to new technologies with changes localized in the framework and in specific components of the service toolkit.

3.2.2 Event Data Model

The Event Data Model is built around the *Event* which contains all data taken during a triggered physics event and all data derived from them.

Events are processed by a user specified sequence of modules and data can be read and stored in the Event, with all provenance information. Modules can be of several types:

- event data producers (EDProducers) used in triggering, reconstruction and simulation. These are the modules that put data products into the Event.

- output: each of these modules write the event data to one of several persistent forms.

- filter (EDFilters) used in triggering. These modules control the flow of processing for the trigger lists.

- analyzers (EDAnalyzers). These modules do not modify the event data, but can use it to create histograms or other event summaries.

The CMS Event Data Model implements an event persistence that allows ROOT [13] to be used directly for interactive analysis of standard CMS event data.

3.2.3 ROOT

ROOT is an object-oriented framework aimed at solving the data analysis challenges of high-energy physics.

ROOT has a command line interface and a C++ interpreter (CINT) which can be used to run scripts for data analysis.

The CMSSW Framework uses many classes from the ROOT Framework and in particular the output of files from CMSSW is a ROOT format.

3.3 How can this all work together

All these tools have been used more or less directly to write this thesis.

Monte Carlo samples described in chapter 5 were produced with ProdAgent, registered in DBS-DLS and transferred with PhEDEx to some storage element.

To analyze them we have developed an EDAnalyzer with CMSSW Framework and we have used a GRID user interface with CRAB/BOSS to find data location and to pack and prepare our code for the submission to the WMS and to execute it over the GRID, where data files are stored.

Once we have got back our output ROOT files with relevant data for this analysis, we have used again ROOT to run the *Toolkit for Multivariate Analysis*[14] and to produce histograms and plots.

The CMS computing and software realize a big complex system, facing complex problems, while trying to hide all this complexity to the users.

Chapter 4

Multivariate Analysis Techniques

4.1 Introduction

Pattern recognition consists in assigning an event to a predetermined category or class.

A pattern recognition system implies a sensor (CMS detector), a preprocessing mechanism (CMSSW) to extract event features, a feature extraction mechanism, a classification algorithm and some examples already classified (Monte Carlo samples).

In our case features are variables built on reconstructed event objects.

To classify an event we will use a vector of features.

The quality of a feature vector is given by its ability to discriminate signal from background.

With our feature vector and Monte Carlo Samples, we can train some classifiers to distinguish between signal and background.

The last step is to evaluate the performance of our classifier.

4.2 Classifiers

The task of a classifier is to divide the feature space in regions and to assign each region to a class.

The classification of a feature vector consists of determining which region it belongs to.

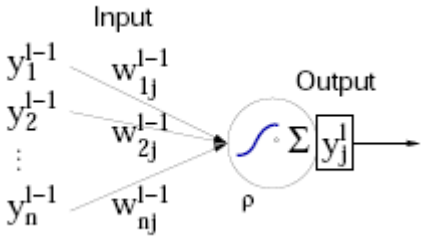


Figure 4.1: Multi Layer Perceptron with one hidden layer.

4.2.1 Artificial Neural Network

An artificial Neural Network (ANN) is a simulated collection of interconnected neurons.

A neuron is an entity which has a defined output response to a given set of input signals.

In the collection of neurons of the ANN we identify a set of n input neurons (the number of variables we calculate from data) and a set of m output neurons (typically 1) realizing a mapping of the n dimensional variables space to R.

If the response of neurons to the input signal is not linear, this mapping is also non linear.

Multi Layer Perceptron

If we organize neutrons in layers, allowing connections only in one direction from one layer to the next one, what we realize is a Multi Layer Perceptron. We call the first layer the input layer, the last layer the output layer and the layers between the hidden layers (Fig. 4.1.

We can often separate the neural network response function in a synapse function and a neuron activating function.

The synapse function is typically a sum (sum of squares or sum of absolutes) of neuron output $\times$ neuron weight.

The neuron activation function typically is linear or sigmoid or radial or Tanh.

To optimize the classification performance we have to calculate the neuron weights.

The most common algorithm used to fulfill this task is the BackPropagation algorithm.

We can define an Error function E which depends on ANN weights and input neurons, and then, at each step ρ of the training, move each weight $\mathbf{w}$ of a small amount η in the direction where E decreases more rapidly.

$$\mathbf{w}^{\rho+1} = \mathbf{w}^{\rho} - \eta \nabla_{\mathbf{w}} E \tag{4.1}$$

The MLP neural network implements a variable ranking which is based on the sum of the weights squared of the connections of the input neuron.

4.2.2 Boosted decision trees

A decision tree is a binary tree which splits the variables space in a collection of hypercubes where each hypercube is classified as signal or background according to the hypercube dominant class (Fig. 4.2).

At each step of the training, the algorithm finds the variable and the cut for that variable that gives the best separation between signal and background. Then this procedure is iterated over each subset until the events in the final hypercubes are under a predefined number.

The main problem with a single decision tree is its instability with respect to statistical fluctuations of the training sample.

The solution to this problem is the construction of an ensemble of decision trees (a forest) which are built from the same training set. The elements of the training set are subject to a boosting procedure which modifies their weights in the sample.

Boosting Algorithm

The most popular boosting algorithm is called AdaBoost (Adaptive Boost).

In this algorithm events misclassified while training a decision tree are given a higher weight in the training of the next tree, forcing the classifier to concentrate on them.

The common boosting weight α is derived from the misclassification rate *err* of the previous tree,

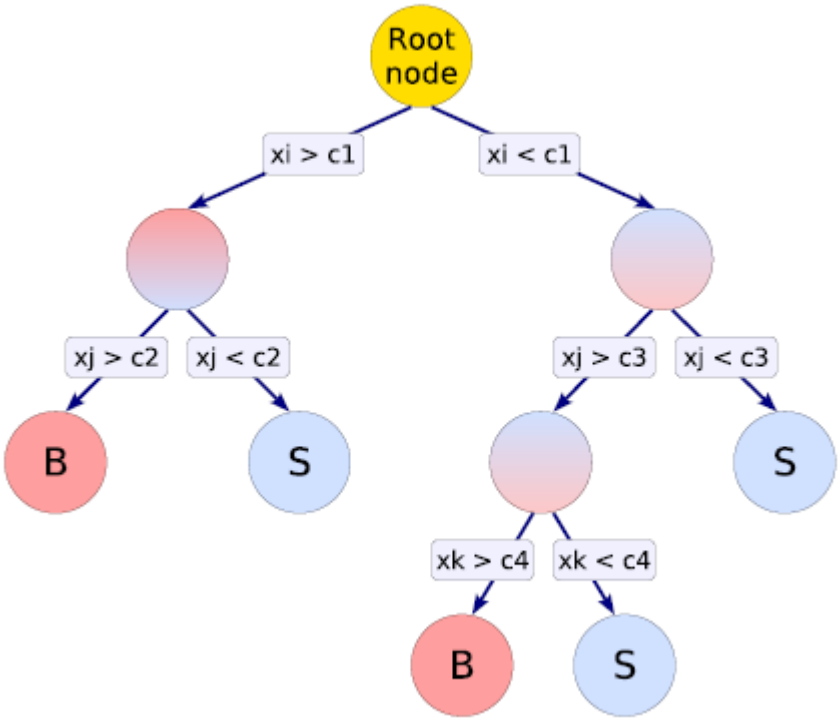


Figure 4.2: Decision Tree structure.

$$\alpha = \frac{1 - err}{err} \tag{4.2}$$

and its meaning is to give more importance to the more accurate tree.

The output of the forest is

$$y_{BDT} = \sum_{i \in forest} \ln(\alpha_i) \cdot h_i(x) \tag{4.3}$$

where $h_i(x)$ is the result of the individual tree for the x input variables.

Small values of y_{BDT} indicate background events while large values indicate signal events.

To avoid the classifier over training, a pruning process is applied after the boosting process. Pruning consists in the cutting of a tree from the bottom up to eliminate statistically insignificant nodes.

4.2.3 Support Vector Machine

The Support Vector Machine is a linear classifier which minimize the empirical classification error and maximize the geometric margin.

This means that the Support Vector Machine maps input vectors to a higher dimensional space (also with non linear transformations) where signal and background are linearly separable, then it builds two parallel hyperplanes separating the classes.

The separating Hyperplane is the plane that maximize the distance from the two parallel hyperplanes (Fig. 4.3).

The assumption is that the larger the distance, the better the generalization error of the classifier will be.

Since the separating hyperplane depends only on the boundaries of the classes to discriminate, the most relevant vectors are the ones near the border (support vectors).

SVM performances can be tuned modifying kernel parameters and a Cost parameter.

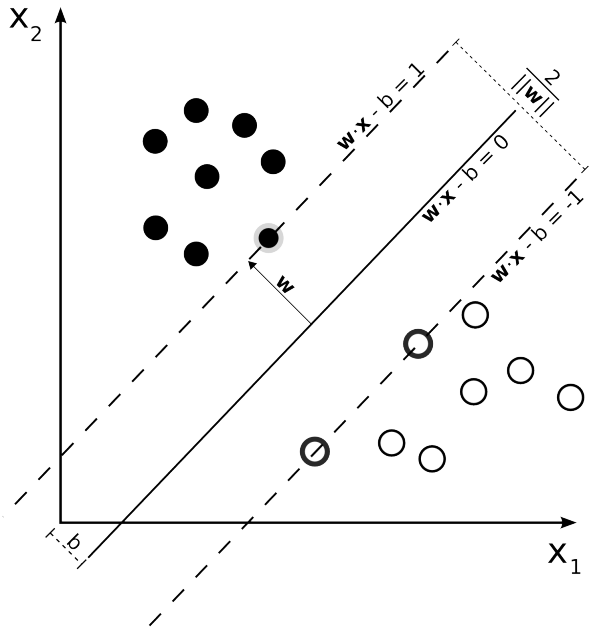


Figure 4.3: Separating hyperplane.

Chapter 5

Event Samples and High Level Objects

5.1 Monte Carlo Samples

The data used for the present analysis are from the Spring07 official production for signal events and from CSA07 official production for background events.

Signal samples for top pair production were generated with PYTHIA/TopRex generator and reconstructed with CMSSW version 1_3_1.

Background samples were generated with Pythia version 6_409 and reconstructed with CMSSW version 1_5_2.

Table 5.1 lists the datasets used for this analysis

5.2 Jet Reconstruction

In this analysis we will use jets reconstructed with the iterative cone algorithm.

This algorithm takes two parameters, a cone size R and a seed threshold, and creates a list ordered in E_{T} from input objects (calorimeter towers for reconstructed jets); starting from the first object it builds a cone of size R in $\eta - \phi$ space.

All the objects falling inside this cone are used to calculate a proto-jet direction and energy. The resulting proto-jet direction is used to iterate the process considering a new cone of size R and adding to the proto-jet all the objects falling in this

$t\bar{t}$ signal			
Channel	dataset	Number Of Events	Cross Section (nb)
$t\bar{t}$ inclusive	ttbar_inclusive_toprex	2762326	0.833
$t\bar{t}$ all hadronic	ttbar_inclusive_toprex	1261378	0.38

QCD background			
$\hat{p_t}$ range (GeV)	dataset	Number Of Events	Cross Section (nb)
$50 < \hat{p_t} < 80$	QCD_Pt_50_80	893240	2.16×10^4
$80 < \hat{p_t} < 120$	QCD_Pt_80_120	1243257	3.08×10^3
$120 < \hat{p_t} < 170$	QCD_Pt_120_170	1260951	4.94×10^2
$170 < \hat{p_t} < 230$	QCD_Pt_170_2300	934870	1.01×10^2
$230 < \hat{p_t} < 300$	QCD_Pt_230_300	800840	2.45×10^1
$300 < \hat{p_t} < 380$	QCD_Pt_300_380	1272037	6.24
$380 < \hat{p_t} < 470$	QCD_Pt_380_470	781003	1.78
$470 < \hat{p_t} < 600$	QCD_Pt_470_600	1317613	6.83×10^{-1}
$600 < \hat{p_t} < 800$	QCD_Pt_600_800	592580	2.04×10^{-1}
$800 < \hat{p_t} < 1000$	QCD_Pt_800_1000	718458	3.51×10^{-2}
$1000 < \hat{p_t} < 1400$	QCD_Pt_1000_1400	615085	1.09×10^{-2}
$1400 < \hat{p_t} < 1800$	QCD_Pt_1400_1800	298782	1.06×10^{-3}

Table 5.1: Monte Carlo samples.

new cone. The process goes on until proto-jet energy and direction variation are under a fixed threshold (1% in energy and $\Delta R < 0.01$ in direction).

Once a stable proto-jet is found, all objects in the proto-jet are removed from the list of objects and the proto-jet itself is added to the list of jets.

The algorithm ends when in the object list there are no objects with an E_{T} over the seed threshold.

5.3 Jet Calibration

The jet energy resolution depends on a lot of factors which are related to the jets physics and to the detector response.

From the physics point of view, initial and final gluon radiation, jet fragmentation and pile-up are relevant for the jet energy resolution.

From the detector point of view, electronic noise, dead materials, cracks, low resolution for low E_{T}, jet containment and separation have a direct effect on jets energy resolution.

The goal of jet calibration is to obtain a calibration function depending on direction, energy and flavor of jet.

To achieve this result, minimum bias events will be used to test the uniformity of the energy scale in ϕ; isolated energetic particles identified by the tracker will be used to calibrate the calorimeter up to $|\eta| < 2.4$ and E_{T}-balance in di-jets events or $\gamma/Z + jet$ will be used for $|\eta| > 2.4$.

Jets corrections used in Monte Carlo production are built on studies of the jet response over fully simulated QCD di-jets events in the full p_{T} range $0 < p_{\mathrm{T}} < 4000$ GeV/c.

5.4 QCD Rates and $\mathrm{t\bar{t}}$ efficiency

The number of events from QCD production is huge.

To reduce the QCD rate to an acceptable value without throwing away too many signal events, we have to set up a trigger based on the E_{T} of the four leading jets.

Defined ϵ the selection efficiency, the effective cross section becomes

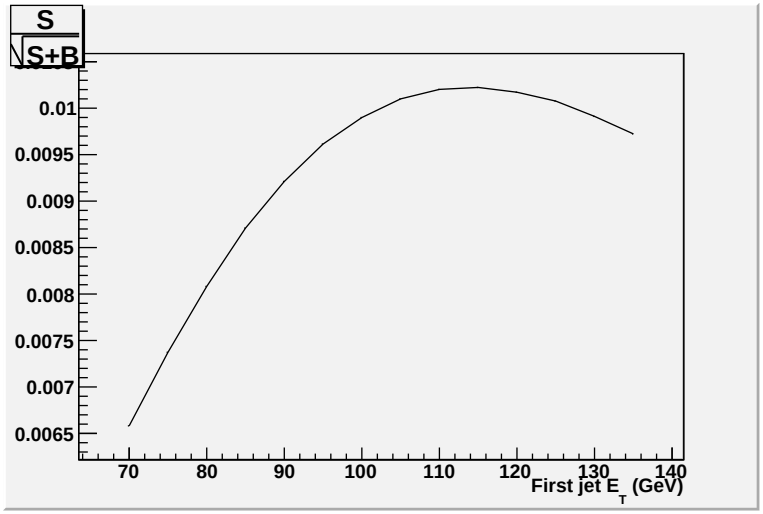


Figure 5.1: Significance of the first Jet Et.

$\hat{\sigma} = \sigma \times \epsilon$

the rate is defined in terms of the instantaneous luminosity $\mathcal{L}$ as

$\text{Rate} = \hat{\sigma} \times \mathcal{L} = \sigma \times \epsilon \times \mathcal{L}$

if we express the cross section in nb, the rate for an instantaneous luminosity of $\mathcal{L} = 2 \times 10^{33}\ \text{cm}^{-2}\,\text{s}^{-1}$ is:

Rate(Hz)=$2 \times \hat{\sigma}$(nb)

To guide our cuts in E_{T} we looked at the significance defined as

$$\frac{S}{\sqrt{S+B}} \tag{5.1}$$

trying to stay as near as possible to the maximum for each cut (Fig 5.1). The selected cuts are:
$E_{\text{T}} > 115$ GeV on the first jet
$E_{\text{T}} > 70$ GeV on the second jet

$E_T > 60$ GeV on the third jet
$E_T > 40$ GeV on the fourth jet

With these cuts we reach a QCD rate of about 211 Hz and a $t\bar{t}$ efficiency of about 24% (33% for the all hadronic channel).

Of course this QCD rate is much too high for CMS trigger requirements and should be strongly diminished applying also a b-tag selection. Due to the low statistics of our samples we will not apply a stronger reduction of the QCD rate for this analysis, so we leave to future developments the task of reducing the trigger rate down by a factor of ≈ 10.

After this selection we decided to choose an event topology where the number of jets N_{jets} is $6 \leq N_{jets} \leq 8$.

This selection reduced the $t\bar{t}$ efficiency to 7.9% (13% for the all hadronic channel).

To calculate the number of jets, we considered only jets with a $E_T > 30$ GeV and $|\eta| < 2.4$.

In table 5.2 we see the $t\bar{t}$ efficiency and the resulting QCD rates after the trigger application and the requirements on the number of jets.

The $t\bar{t}$ effective cross section resulting after these selections is of about 66 pb (50 pb for the all hadronic channel).

The signal to background (S/B) ratio after the multi-jet trigger amounts to about $\frac{1}{517}$ ($\frac{1}{862}$ for the all hadronic channel).

Applying the topological request we get a S/B ratio of about $\frac{1}{106}$ ($\frac{1}{140}$ for the all hadronic channel).

$t\bar{t}$	Rate [Hz]		
	Production	multi-jet trigger	$6 \leq N_{jets} \leq 8$
all hadronic	.76	0.251	0.099
$\hat{\sigma}$ [nb]	0.38	0.123	0.050

QCD	Rate [Hz]		
$\hat{p_T}$ range [GeV]	Production	multi-jet trigger	$6 \leq N_{jets} \leq 8$
50 - 80	43 k	10	0.15
80 - 120	6.2 k	66	2.2
120 - 170	1 k	75	4.7
170 - 230	202	39	3.8
230 - 300	49	14	1.95
300 - 380	12.5	4.5	0.79
380 - 470	3.6	1.5	0.3
470 - 600	1.4	0.62	0.15
600 - 800	0.4	0.2	0.01
Total	50.5 k	211	14
Total $\hat{\sigma}$ [nb]	25250	106	7

Table 5.2: $t\bar{t}$ and QCD rates at production level and after the trigger application. As we can see from table 5.1 the contribution to the QCD rate of samples with $\hat{p_T} > 800$ GeV is less than 1 Hz at $\mathcal{L} = 2{\times}10^{33}$ cm^{-2} s^{-1}.

Chapter 6

Application to top quark cross section measurement

The S/B value obtained after the multi-jet trigger and the topological request $6 \leq N_j \leq 8$ is very small and we need an additional selection.

6.1 Event Variables

To discriminate signal events from background events, we use a set of kinematical variables which try to characterize the $t\bar{t}$ events from the QCD background.

These variables are:

- Total Transverse Energy of the jets (SumEt);
- non leading jet total transverse energy obtained removing the two most energetic jets (SumEt3);
- sphericity, defined as $\frac{3}{2}(Q_1 + Q_2)$ where Q_1 and Q_2 are the two smallest eigenvalues of the Sphericity Tensor S [1]
- aplanarity, defined as $\frac{3}{2}Q_1$ where Q_1 is the smallest of the three normalized eigenvalues of the Sphericity Tensor;

[1] The Sphericity Tensor is defined as: $S^{\alpha\beta} = \sum_j p_j^\alpha p_j^\beta$ where $\alpha, \beta = 1, 2, 3$ corresponds to the x,y and z components. By standard diagonalization of $S^{\alpha\beta}$ one may find three eigenvalues $Q_3 \geq Q_2 \geq Q_1$ with $Q_3 + Q_2 + Q_1 = 1$

- centrality, which is the fraction of the hard scattering energy going in the transverse plane;

- E^*_{T1} and E^*_{T2}, for the leading jet and the next to leading jet, where $E^*_T = E_{\mathrm{T}} \sin^2\theta$ and where θ is the angle with respect to the proton beam, as measured in the all-jets center of mass frame.

- the minimum di-jet invariant mass (DijetMinMass);

- the maximum di-jet invariant mass (DijetMaxMass);

- the minimum tri-jet invariant mass (TrijetMinMass);

- the maximum tri-jet invariant mass (TrijetMaxMass);

figures from 6.6 to 6.16 show the distribution of these variable normalized to the same area and normalized to a common luminosity of 1 fb^{-1}, after requiring $6 \leq N_{jet} \leq 8$, with $E_{\mathrm{T}} > 30$ GeV and $|\eta| < 2.4$.

6.2 Classifiers performance

The training of the classifiers has been done using the $t\bar{t}$ fully hadronic sample and the QCD background in the range $80 < \hat{P}_t < 470$ GeV where signal and background have similar behavior.

Each sample has been weighted with the effective cross section to give it the right relevance.

The trained classifiers have been then applied to the whole background sample and on the whole signal sample.

In Fig. 6.1 and Fig. 6.2 we can see the output of the three classifiers on the signal events and background events not normalized to a common luminosity. Looking at these graphs one can think of a good separation of signal from background, but we have to remember that our samples will be normalized to a luminosity of 1 fb^{-1} and then we will see that the rejection of background is not so good in terms of absolute background ratio (Fig. 6.3).

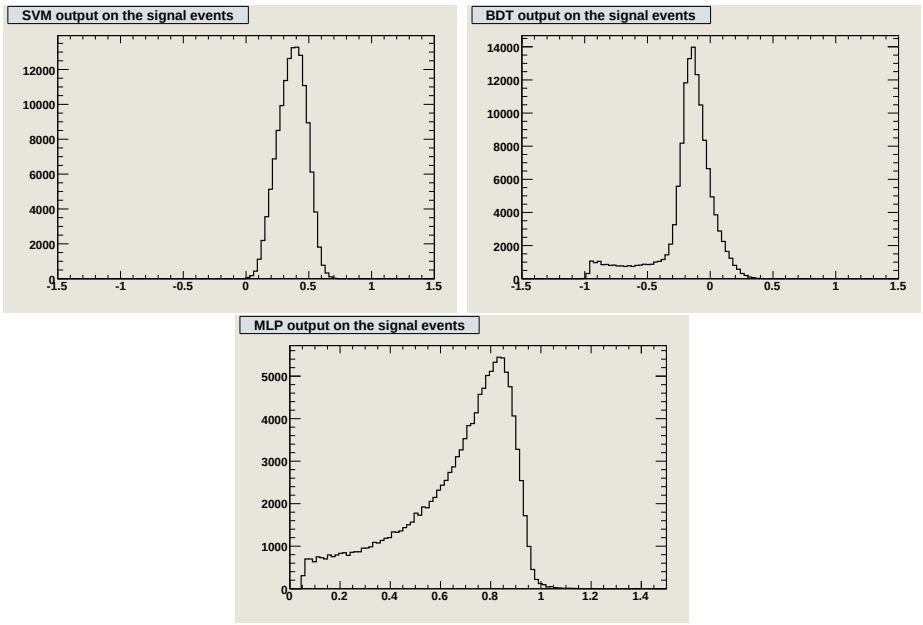


Figure 6.1: Output of the SVM,BDT and MLP classifiers on the signal events.

Fig. 6.4 shows the rejection of background versus the signal efficiency for the SVM, BDT and MLP classifiers. From these graphs we can see that the three classifiers have comparable performance.

Even if the performance is similar, anyway, applying the trained classifier to the data is much faster for the Multi Layer Perceptron than for the other two classifiers, and so from now on we will use MLP.

In Fig. 6.5 we can see the significance values and $\frac{S}{B}$ values obtained with MLP classifier. These values will be used to estimate a statistical significance for the Cross Section Measurement.

6.3 Cross section measurement

The cross section measured would be

$$\sigma = \frac{n - b}{\epsilon L} \tag{6.1}$$

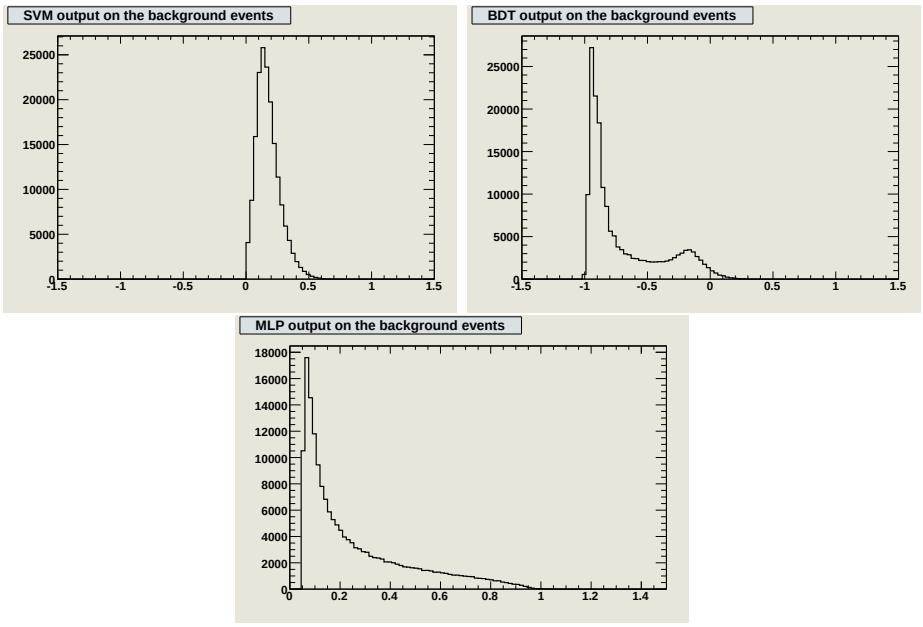


Figure 6.2: Output of the SVM,BDT and MLP classifiers on the background events.

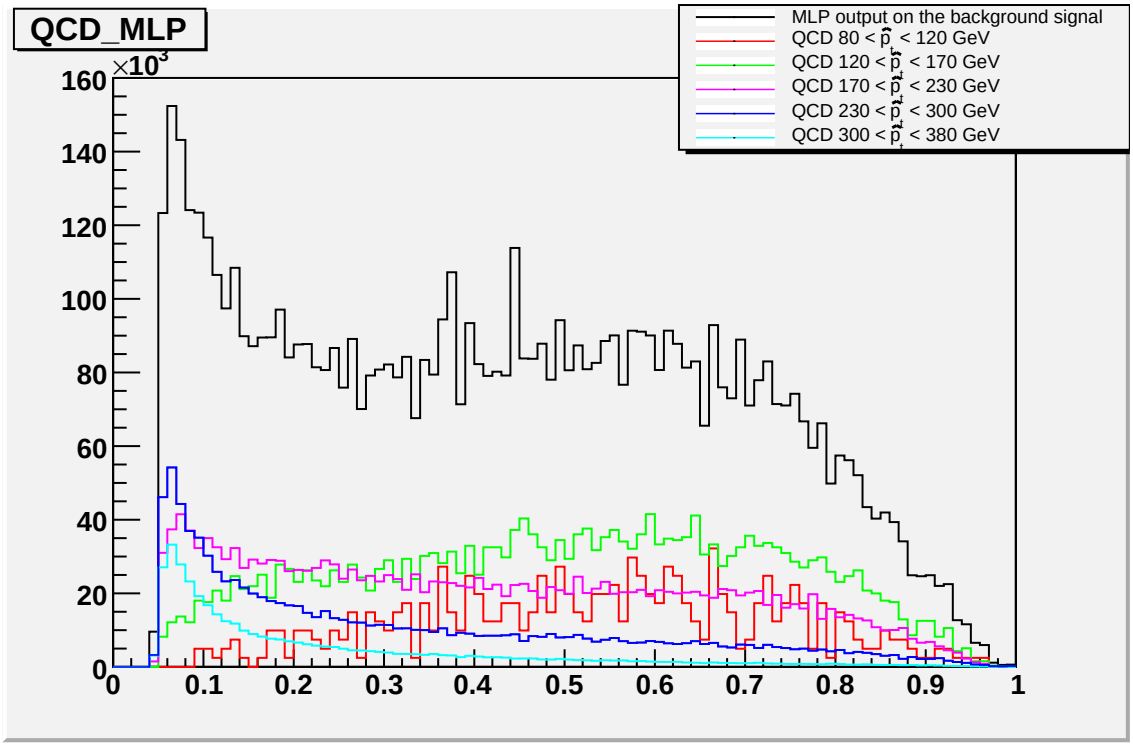


Figure 6.3: MLP output on the background sample. Contribution of the various samples of QCD background is shown.

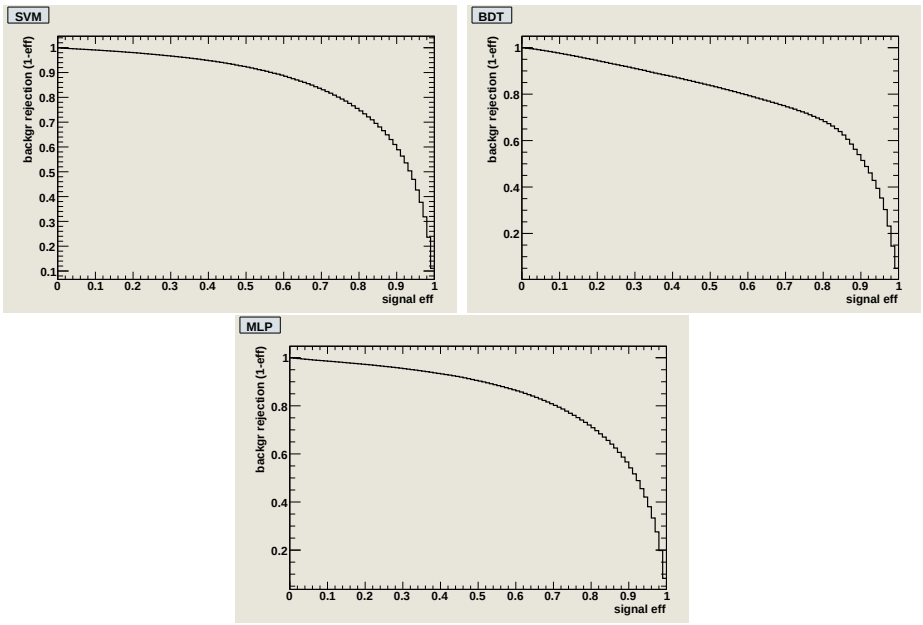


Figure 6.4: Background rejection versus signal efficiency for the SVM, BDT and MLP classifiers.

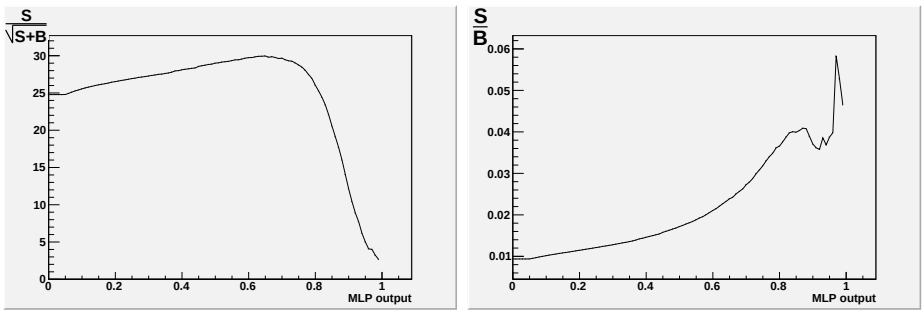


Figure 6.5: Significance and signal to background ratio for Multi Layer Perceptron.

where n is the number of candidate events, b is the expected background, ϵ is the total signal efficiency and L is the integrated luminosity.

To apply this formula to our samples we have to set n = N($t\bar{t}$)+N(QCD) and being b=N(QCD) leaves $n - b$=N($t\bar{t}$).

Assuming we are able to estimate quite accurately the background b, then the expected statistical uncertainty of the cross section would be :

$$(\Delta\sigma)_{stat} \approx \frac{(\Delta n)_{stat}}{\epsilon L} = \frac{\sqrt{n}}{\epsilon L} = \sigma\frac{\sqrt{n}}{n-b} \tag{6.2}$$

so

$$(\frac{\Delta\sigma}{\sigma})_{stat} = \frac{\sqrt{n}}{n-b} \tag{6.3}$$

If we cut on the MLP output at 0.7, the point which gives the maximum of $\frac{S}{\sqrt{S+B}}$, we expect to collect 33 000 signal events and 1 215 000 background events in the first fb^{-1} of collected data; this correspond to $\left(\frac{\Delta\sigma}{\sigma}\right)_{stat} = 3.4\%$

After the selection based on the MLP output, we tried to require also that in the selected events at least one jet is a jet generated by a b quark.

The 7% (0.3%) of the selected signal (background) events have a jet tagged as a b-jet. This means a $\frac{S}{B} \approx \frac{1}{1.5}$ but the statistical uncertainty on the cross section remains of about 3.3%.

The expected background, the $t\bar{t}$ efficiency and the integrated luminosity are all subject to systematic uncertainties which are not possible to discuss at present time because we are missing the Monte Carlo samples generated with different assumption of jet energy scale, initial/final state radiation, parton distribution functions, etc.

We agree, however, that the dominant sources of systematic uncertainty will come from the uncertainty on the knowledge of the Jet Energy Scale(JES). An educated guess estimates that in the first years of running the jet energy scale would be known at a 5% level.

Such 5% uncertainty would translate into a 21.7% uncertainty on the efficiency and in a 22% uncertainty on the cross section.

This is a large uncertainty. In order to obtain a more accurate σ measurement we clearly need to improve our knowledge of JES.

A possible solution could be to measure simultaneously the mass and the cross section, where the mass is the outcome of a two dimensional fit including JES, and the W masses are used as a constraint on JES.

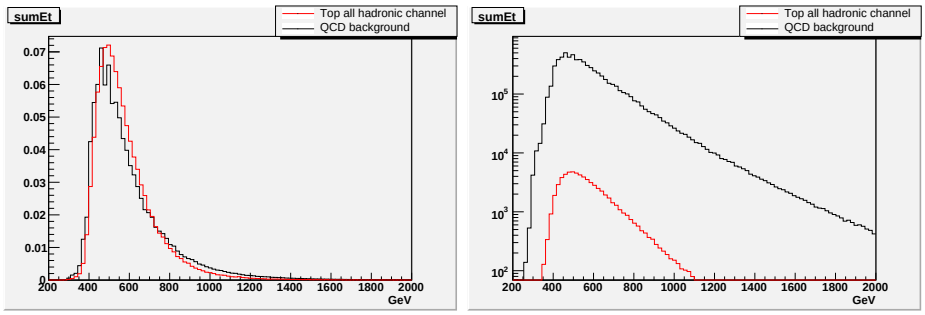


Figure 6.6: SumEt for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

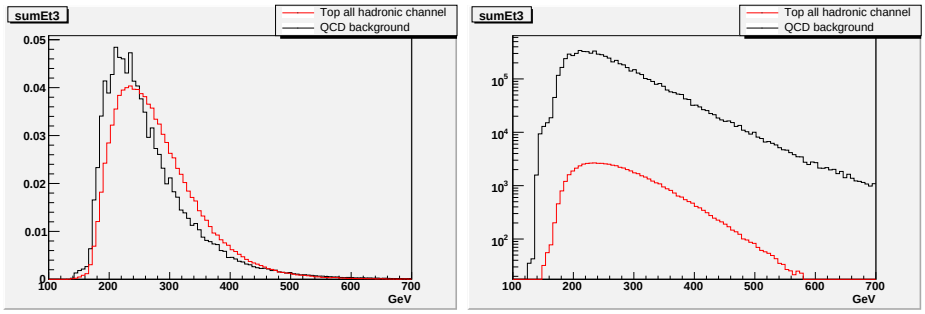


Figure 6.7: SumEt3 for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

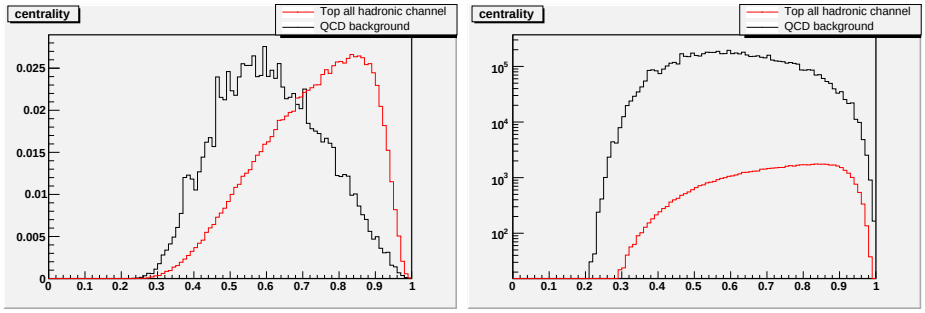


Figure 6.8: Centrality for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

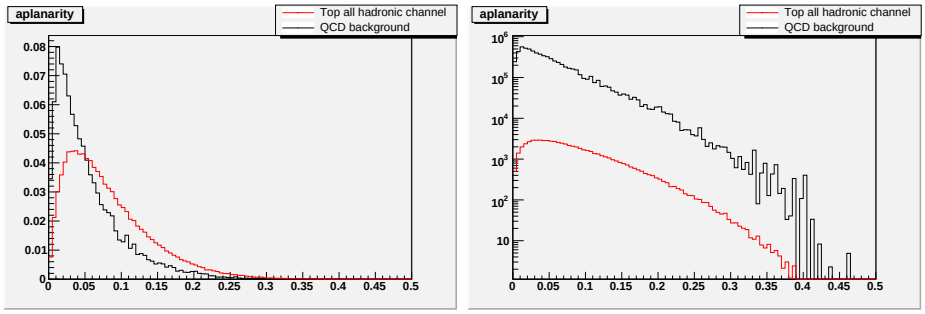


Figure 6.9: Aplanarity for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

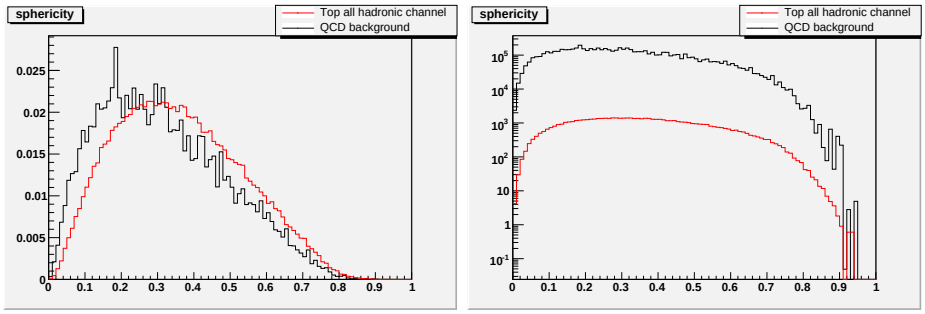


Figure 6.10: Sphericity for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

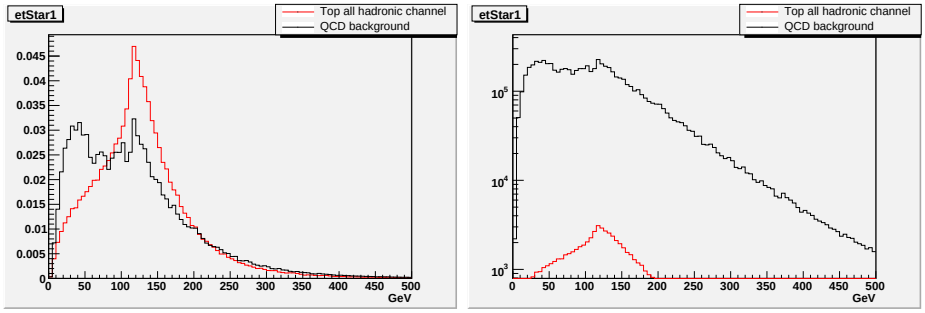


Figure 6.11: E^*_{T1} for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

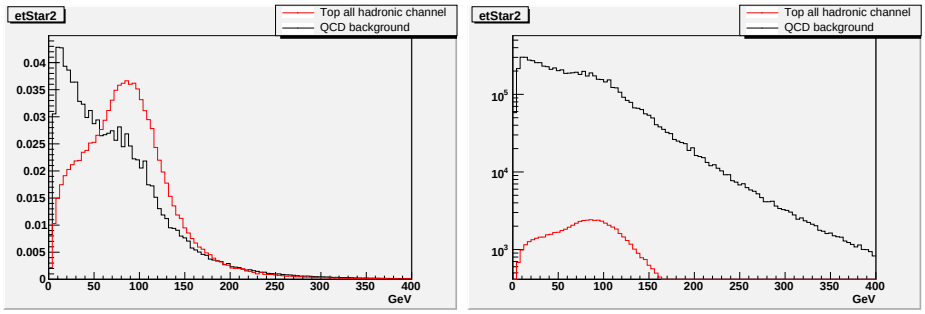


Figure 6.12: $E^*_{T_2}$ for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

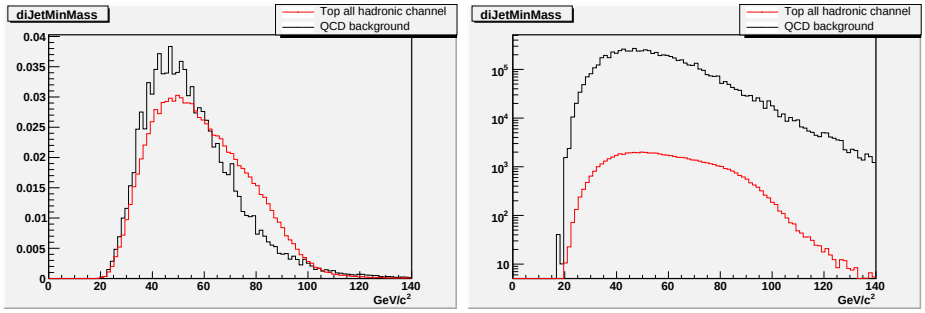


Figure 6.13: DiJetMinMass for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

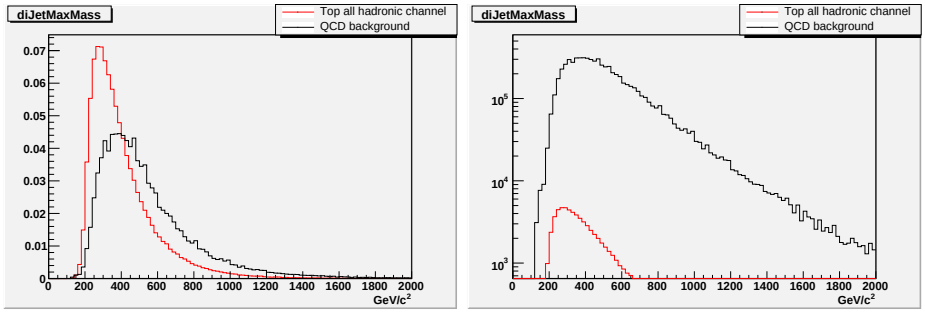


Figure 6.14: DiJetMaxMass for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

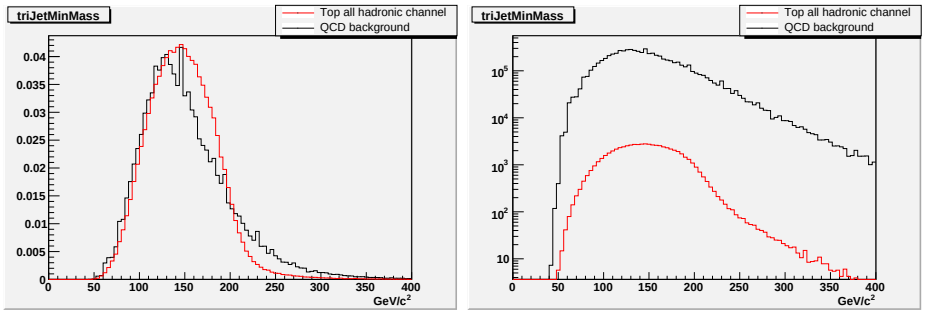


Figure 6.15: TriJetMinMass for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

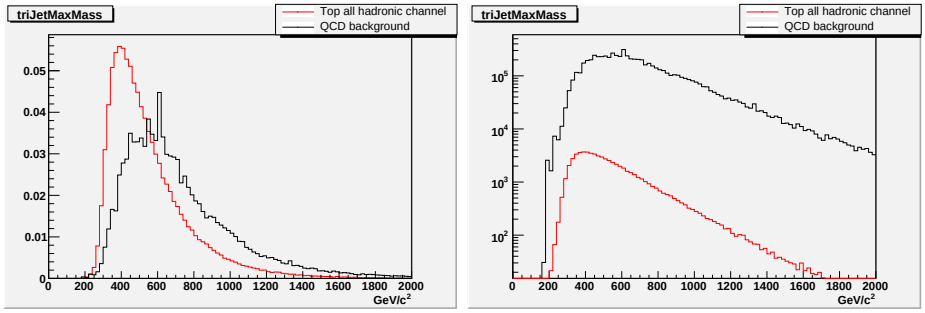


Figure 6.16: TriJetMaxMass for signal and background events. Distribution normalized to the same area (left) or the expected amount in 1 fb^{-1} (right)

Conclusion

In this thesis we set up a procedure to apply multivariate analysis techniques to the study of the Top quark fully hadronic decay channel.

We used tools which are available within the software of the CMS Experiment as building blocks for the whole processing.

In particular we used a package integrate with ROOT, TMVA (Tools for MultiVariated Analysis) which implements the most common classifiers.

We started with data samples produced in official CMS activities for the Top quark signal and QCD background producing a set of jet based variables which should characterize $t\bar{t}$ events.

We used the Multi Layer Perceptron classifier, the Boosted Decision Trees classifier and the Support Vector Machine classifier to determine a kinematical selection of the signal events from the background events using the defined variables.

The performance of the three classifiers seems equivalent and a selection based only on kinematical variables seems not sufficient to achieve a useful signal background ratio.

In particular, we see that after the multi-jet trigger selection the number of events of QCD background is not sufficient to permit an adequate training, even weighting each sample with a specific weight.

Experience from CDF teaches that these classifiers perform better if the background signal used for the training is well characterized or even better if the background signal comes from real data. So we expect that the procedure we set up using Monte Carlo background simulation will give better results when we will be able to use real data for the training.

Bibliography

[1] S. Fukuda et al. Tau neutrinos favored over sterile neutrinos in atmospheric muon neutrino oscillations. *Phys. Rev. Lett.*, 85:3999-4003, 2000.

[2] M. Ambrosio et al. Matter effects in upward-going muons and sterile neutrino oscillations. *Phys. Lett.*, B517:59-66, 2001.

[3] M. Turner E. Kolb. *The Early Universe*. 1989.

[4] M. Beneke *et al.*, "Top quark physics," arXiv:hep-ph/0003033.

[5] *CMS, The Computing Project - Technical Design Report*, CERN/LHCC 2005-023, (2005)

[6] LHC Computing Grid (LCG), Web Page, http://lcg.web.cern.ch/LCG/ and *LCG Computing Grid - Technical Design Report*, LCG-TDR-001 CERN/LHCC 2005-024, (2005)

[7] EGEE Homepage, http://www.cern.egee

[8] A. Fanfani,*Distributed Data Management in CMS*, CHEP06, Mumbay 2006.

[9] PhEDEx - Physics Experiment Data Export - Project Homepage, http://cms-project-phedex.web.cern.ch/cms-project-phedex/

[10] BOSS, Batch Object Submission system - Project Homepage, http:/boss.bo.infn.it/ and C. Grandi, A. Renzi, *Object Based system for Batch Job Submission and Monitoring (BOSS)* CMS NOTE 2003-005 (2003) and G. Codispoti et al., *BOSS: the CMS interface for job summission,*

monitoring and bookkeeping, Workload Management and Workflows at the EGEE User Forum, CERN March 1st-3rd 2006

[11] CRAB homepage, http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/ and D. Spiga et al., *The CMS Remote Analysis Builder (CRAB)* 14th Int. Conf. on High Performance Computing (HiPC 2007). Goa, India. Dec 18-21 2007. (vol. 4873, pp. 580-586). ISBN/ISSN: 978-3-540-77219-4.

[12] D. Evans et al. *CMS MC Production System Development & Design*, CHEP'07 International Conference on Computing in High Energy and Nuclear Physics , Victoria BC , Canada, 2-7 Sept 2007

[13] ROOT - An Object Oriented Data Analysis Framework in AIHENP 96 *Workshop*, volume Phys. Res. A 389, pp. 81-86. Lausanne, Switzerland, September, 1996, 1997.

[14] http://tmva.sourceforge.net/

[15] T. Sjstrand, P. Edn, C. Friberg, L. Lnnblad, G. Miu, S. Mrenna and E. Norrbin, Computer Physics Commun. 135 (2001) 238

[16] Comput.Phys.Commun. 148 (2002) 87-102 arXiv:hep-ph/0201292v1

[17] R. Demina et al., Calorimeter Cell Energy Thresholds for Jet Reconstruction in CMS, CMS Note 2006/020 (2006).

[18] A. Heister et al., Measurement of Jets with the CMS Detector at the LHC, CMS Note 2006/036 (2006).

[19] O. Kodolova, Jet Energy Measurements in CMS, CMS CR 2005-019 (2005). Presented at HCP 2005, Hadron Collider Physics Symposium, Les Diablerets, Switzerland, 46 July