Data analysis techniques

P. Govoni
CERN, Milano-Bicocca
introduction
Introduction

- my interpretation of “data analysis techniques” is here “doing a data analysis”

- follow the steps from the beginning (data taking) to the end (the result)
  - the luminosity
  - the trigger, from the point of view of the analysis
  - the reconstruction and detector response
  - the simulation
  - differential cross-section measurement: a di-jet correction
  - searches: the $H \to WW \to \ell\ell\ell$
  - multivariate techniques

Thanks to the following people, for interesting discussions, for liberally “borrowing” slides, or both: D. Benedetti, C. Bernet, T. Camporesi, G. Cowan, K. Cranmer, K. Ellis, S. Gennai, A. Ghezzi, A. Hoecker, R. Van Kooten, M. Nguyen, M. Paganoni, M. Pelliccioni, E. Rizvi...
access to data

RAW data

reconstruction

centralized analysis to identify final state objects (leptons, photons, jets...)

skimming

(centralized) copy of sub-set of data, dedicated to analyses (different final states)

analysis

(can be done more than once, to profit of better features derived from previous analyses)

profit of the latest reconstruction

well suited to the analysis needs (speed, bkg measurements)

do not decouple from the approved definitions

(personal) selections on the skims, to have a sample which is as small as possible
the cross-section

cross section: \( \sigma = \frac{N_{obs} - N_{bkg}}{\varepsilon \cdot \int L dt} \)

number of observed events

analysis efficiency \( \varepsilon = \varepsilon_{tr} \cdot \varepsilon_{reco} \cdot \varepsilon_{ID} \cdot \varepsilon_{sel} \)

background contamination in the sample

luminosity delivered by LHC

1 barn = \( 10^{-28} \) m\(^2\) = \( 10^{-24} \) cm\(^2\)
luminosity
\[
\sigma = \frac{N_{obs} - N_{bkg}}{\varepsilon \cdot \int \mathcal{L} dt}
\]

\[
\mathcal{L} = n_b f \frac{n_1 n_2}{4\pi \sigma_x \sigma_y}
\]

- number of colliding bunches
- revolution frequency
- beam transverse size

\[1 \text{ barn} = 10^{-28} \text{ m}^2 = 10^{-24} \text{ cm}^2\]
luminosity

Instantaneous Luminosity

Updated: 19:38:52

5 may 2011
delivered luminosity

- the **delivered luminosity** is what the LHC gives to an experiment
- the **recorded luminosity** is different from the delivered one, because of data taking inefficiencies
- the **certified luminosity** is different from the recorded one, because of detector problems
- not necessarily all studies need the same level of certification!
The number of interactions per bunch-crossing is poisson-distributed with mean $\mu$.

\[ \mu = \frac{\sigma L}{\int n_B} \]

\[ P(k = 0) = \frac{\lambda^0}{0!} e^{-\lambda} \]

- Hard to distinguish positive countings $\Rightarrow$ count the zeros and invert the poisson.
- Already with 10 interactions per bunch-crossing, the poisson is hard to invert (zero starvation).

Find another process which is linear in the luminosity and calibrate it.

\[ \frac{R_0}{\mathcal{L}_0} = \sigma_{vis} \]

Define $\sigma_{vis}$

\[ \mathcal{L}(t) = \frac{R(t)}{\sigma_{vis}} \]

Calculate the luminosity as a function of a rate.
the trigger
the trigger

- the vast majority of events are not interesting
- interesting physics happens at low rates (< 10 Hz)
- the final bandwidth is limited: can store up to O(100 Hz) of events (1 event ~ 1 MB)
- the decision has to be taken fast enough (bunch crossing rate = 1/25 ns)
trigger: the CMS example

- L1 based on regional information, dedicated electronics
- HLT is software-based, runs on commercial computers farm - can be implemented by std::physicist
- performs a first physics reconstruction of the event, with algorithms (very) similar to the ones used in the final analysis
- exploits the expected signatures of the event
what to trigger

• HLT searches for **interesting physics objects**:  
  • high pT leptons  
  • leptons with a certain degree of identification (isolation)  
  • presence of many leptons  
  • large missing energy  
  • presence of many jets (+ other requirements)

• HLT is based on the **topology of the analysis it aims for**

• make sure that the events one is interested in are actually triggered. If not, need to **implement a new one** and get it deployed

• low $p_T$, loose ID, few leptons are difficult to trigger
when the instantaneous luminosity increases, the triggers need to change, since the available bandwidth does not increase

- increase thresholds
- build sophisticated triggers
- prescale the trigger = take only a fraction \((1/p_i)\) of the events that would fire a given trigger

\[
N_{prod} = \frac{N_{obs}}{\varepsilon_{tr}} \quad \Rightarrow \quad N_{prod} = \frac{p_i \cdot N_{obs}}{\varepsilon_{tr}}
\]
**prescaling: example at CMS**

- example of a trigger table

<table>
<thead>
<tr>
<th>HLT path</th>
<th>inst. lumi (cm$^{-2}$s$^{-1}$)</th>
<th>Prescaler</th>
<th>L1 seed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2e31</td>
<td>1.4e33</td>
</tr>
<tr>
<td>SingleElectron</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLT_Ele25_CaloIdL_CaloIsoVL_TrkIdVL_TrkIsoVL_v1</td>
<td>200</td>
<td>150</td>
<td>100</td>
</tr>
<tr>
<td>HLT_Ele42_CaloIdVT_CaloIsoVL_TrkIdVL_TrkIsoVL_v1</td>
<td>75</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>DoubleElectron</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLT_DoubleEle10_CaloIdL_TrkIdVL_Ele10_v6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_DoubleEle45_CaloIdL_v1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_Ele17_CaloIdL_CaloIsoVL_Ele15_HFL_v6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_Ele17_CaloIdL_CaloIsoVL_Ele15_HFT_v1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_Ele17_CaloIdL_CaloIsoVL_Ele8_CaloIdL_CaloIsoVL</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_Ele17_CaloIdL_CaloIsoVL_Ele8_CaloIdL_CaloIsoVL</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_Ele17_CaloIdL_CaloIsoVL_Ele8_CaloIdL_CaloIsoVL</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HLT_CaloIdL_CaloIsoVL_v5</td>
<td>2000</td>
<td>1400</td>
<td>1000</td>
</tr>
<tr>
<td>HLT_Ele8_CaloIdL_CaloIsoVL_Jet40_v5</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>HLT_Ele8_CaloIdL_CaloIsoVL_v5</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>HLT_Ele8_CaloIdL_CaloIsoVL_v5</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>HLT_Ele8_CaloIdL_CaloIsoVL_v5</td>
<td>240</td>
<td>240</td>
<td>240</td>
</tr>
</tbody>
</table>
the trigger and the analysis

• events I am interested in (1) have to be triggered, (2) if not prescaled, it’s better

• the trigger is (usually) not 100% efficient on the analysis sample
  -> measure the efficiency (from data) of the trigger for the analysis

\[
\sigma = \frac{N_{\text{obs}} - N_{\text{bkg}}}{\varepsilon \cdot \int \mathcal{L} dt}
\]

\[\varepsilon = \varepsilon_{\text{tr}} \cdot \varepsilon_{\text{reco}} \cdot \varepsilon_{\text{ID}} \cdot \varepsilon_{\text{sel}}\]

• the **turn-on curve** is the trigger efficiency trend as a function of an offline selection

• the **objects reconstruction** at trigger level is different from the one used in the final analysis

• this produces an **efficiency curve** and a plateau that can be less than 1
trigger efficiency measurements

• different methods available
  • (by means of a software trigger emulator)
  • with tag & probe methods
  • compare to the efficiency of looser triggers (bootstrapping)
  • from a sample defined by an orthogonal trigger
• it changes with respect to the kinematics
  • perform measurements as a function of $p_T$, $\eta$
an example: the tag & probe

- select the **object that would fire the trigger** in a way independent of the trigger itself

- count **how many times it fires** the trigger

- **under the Z peak** basically only the Z production is expected

- given one good lepton, use the $M_{ll}$ constraint to identify it

- the result has to be corrected for combinatorial background under the Z peak (or the counting done by fitting the shapes)

- With sufficient statistics the efficiency can be evaluated in bins of $p_T$, $\eta$, $\phi$
an example: the tag & probe

- basic object of the muon reconstruction (track)
- minimum $p_T$ threshold applied
- $M(\text{tag,probe}) \sim M_Z$

• triggered by single muon trigger
• minimum $p_T$ threshold applied

$$\varepsilon_{\text{muon tr}} = \frac{\text{nb. of probes that fire the trigger}}{\text{nb. of probes}}$$

• both muons might fire the trigger

$$\varepsilon_{\text{muon tr}} = \frac{2TT + TP}{2TT + TP + TF} \quad \text{T} = \text{Tag}$$
$$\text{P} = \text{Probe that fires a trigger}$$
$$\text{F} = \text{probe that fails a trigger}$$
bootstrapping

• ask a utility trigger with **loose requirements**, to check a tight one

• **prescale** it (it will be needed, as requirements are loose)

• within the events triggered, search the ones that **survive the offline analysis selections** and match to the trigger object

• check whether these events **would pass also the tight trigger** and get an efficiency

• if the utility trigger is loose enough (es. a calorimetric deposit for electrons), it can be considered of efficiency 1 and the efficiency obtained is the one of the tight trigger

**keep an eye on the statistics:** a utility trigger is given lower rate + prescaling => not many events will survive the offline selections

build many utility triggers for different variables, rather than a single one with everything loose
other techniques

• use a trigger defined on information independent of the trigger with unknown efficiency (orthogonal)
  • muon triggers to test calorimetry triggers, or vice-versa

• when implementing a trigger for an analysis, need to be sure that also utility triggers are present, to measure the efficiency of the main one
  • they will probably be prescaled
combining triggers

• to increase the number of signal events, or increase the phase space covered:
  • different energies (with different prescales!)
  • different sub-detectors (2 muons in different regions)
  • different signals (electrons OR muons)
• different ways to do it
  • division: one trigger per phase space region
    the simplest, measure the efficiencies separately
  • exclusion: one analysis per trigger, according to the one that has the lowest prescale
    better performing
  • inclusion: the “OR” of all the triggers is considered
    the best one, can become complicated

Combining Triggers in HEP Data Analysis, arXiv:0901.4118
different energies

- choose the trigger as a function of jet energy (division method)
- choose the trigger with lowest prescale (exclusion method)
- select events if they fire any triggers (inclusion method)
the inclusion method

• the “OR” of all the triggers is considered:

\[
P_{tot}(evt) = 1 - \prod_{i=1}^{\text{triggers}} (1 - P_i(evt))
\]

• for two triggers:

\[
P_{tot}(evt) = P_1(evt) + P_1(evt) - P_1|2(evt)P_2(evt)
\]

• for the uncorrelated case:

\[
P_{tot}(evt) = P_1(evt) + P_1(evt) - P_1(evt)P_2(evt)
\]

• in general, correlations need to be considered
  • instrumental (common inefficient elements, common electronics, same level 1 trigger)
  • physical (jets and track triggers might be correlated)
a toy comparison

- keep the trigger simple

- the price payed in systematics might not be worth the effort of combining in the most sophisticated way, or sitting on the turn-on part of the efficiency curve
physics objects reconstruction
physics objects reconstruction

• obtain physics objects from the detector response
  • hits in the tracker and muon detectors
  • energy deposits in the calorimeters
• two ways are available in CMS
  • single objects reconstruction: build final objects (e.g. muons, electrons, jets) from the detector response
  • particle-flow reconstruction: build a coherent list of stable particles and produce the analysis objects on top of them
the CMS detector

- Hadronic calorimeter (HCAL)
- Electromagnetic calorimeter (ECAL)
- MUONS detectors
- Tracker
- Pixel
the particle flow

<table>
<thead>
<tr>
<th>Hits in the tracker</th>
<th>Cells in the calorimeter</th>
<th>Hits in muon detectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracker tracks</td>
<td>Calorimetric clusters</td>
<td>Muon tracks</td>
</tr>
</tbody>
</table>

Link the single objects with geometrical requirements on the extrapolated trajectories and create **blocks**
the particle flow

blocks

identify muons, promote the block
identify electrons, promote the block
match the remaining tracks to clusters, define charged hadrons and neutrals from calorimetric excess
the remaining calorimetric deposits define neutrals

the charged energy contribution is measured well from the tracker

the list of particles obtained (candidates) is used for high level objects classification and reconstruction, to be used in the analysis

\[
\text{MET} = - \sum_{i=0}^{N_{\text{particles}}} \vec{E}_T^i
\]
muons reconstruction

- **high purity** = fit with hits in both tracker and muon
- **high efficiency** = fit in the tracker + confirmation in the muon detector
- **momentum determination** from both tracker and muons information: best resolution from the tracker for $p_T < 200$ GeV, from the muons above (effect of multiple scattering)
- above 1 TeV, the **bremsstrahlung** is significant
electron reconstruction

from ECAL clusters or tracks:

- Seeding
- electron tracking: GSF
- cleaning of GSF track duplicates
- Identification of all the electron energy deposits in the ECAL

from ECAL footprint or tracks extrapolation:

- Identification against charged-particles interacting in the ECAL

use ECAL at high $p_T$, tracker at low $p_T$
electron reconstruction

search for the decay:

\[ J/\Psi \rightarrow e^+ e^- \]

contamination sources:
- real electrons, either from photon conversions or from semi-leptonic b-hadron decays,
- mis-identified charged hadrons.

- at most one hit missing in the pixel detector (reduce conversions)
- electrons originate from the same vertex (reduce the b-decay background)
- quality cuts to reject charged hadrons contamination
- opposite charge
photons reconstruction

- **ECAL clusters** not associated to a track, nor a deposit in the hadronic calorimeter

- ECAL detector response is **calibrated**, to account for the effect of the noise cut on the single crystals readout

- **check the photons energy scale** calibration with 2010 data, by looking at the $\pi^0$ peak position

- pair all photons with at least 400 MeV energy

- determine the peak position with a combined fit of signal + background
jet reconstruction

- jets are reconstructed with the AKT5 algorithm
- for the single object reconstruction: with calorimetric deposits
- for the particle-flow: with particle flow candidates

the jet energy resolution measured from 2010 data
tau reconstruction

- reconstructed as narrow jets in the standard case, as the sum of the particles compatible with the tau decay in a narrow cone in the particle flow case

![Diagram](image)

reconstructed taus $E_T$ compared to the expected one, test performed on a simulated $Z > \tau\tau$ sample
missing energy reconstruction

- Derived from (minus) the sum of “all the rest”
- Sensitive to uncertainties in all the other physics objects
- Noise effects, mis-calibrations, etc. generate fake missing energy in events without missing energy
- Perform a test on a di-jet sample

Data and expectations well agree

The scalar sum of energies used as a reference scale
reconstruction: in summary

• the reconstruction obtains from the detector measurements the physics objects in the final state

• in a coherent way, to close the kinematics (as much as possible)

• making use of the most precise sub-detector

• reconstruction and identification are not (always) disentangled, for example electrons need to be separated from jets

• data-driven techniques necessary to assess the performances

jet composition: only for neutral hadrons one cannot profit of tracker measurements
detector response

• the detector response is not perfect

• the output of the reconstruction needs to be calibrated for the detector response

• use known physics processes to get the calibrations and the relative uncertainty

• for example
  
  • resonances for leptons (energy scale, tag&probe)
  
  • cosmic rays (alignments)
  
  • transverse momentum balances
  
  • ....
ECAL calibration

- each ECAL channel needs a calibration factor to equalize the response of all detector elements
- for electrons, the energy is measured in the tracker and in the ECAL
- find the calibration coefficients by minimizing a $\chi^2$ of:

$$|\mathcal{E} \times \bar{c} - \mathcal{P}|$$

energy in single elements  electrons momenta  unknown coefficients

select good isolated electrons  statistical trend with luminosity  effect on $H \rightarrow \Upsilon \Upsilon$ invariant mass
jet energy corrections

- the jet energy scale needs to be calibrated, as a function of various variables

**Mandatory**

- L1: Offset
- L2: $\eta$ Relative
- L3: $p_T$ Absolute

**Optional**

- L4: “EMF”
- L5: Flavor
- L6: UE
- L7: Parton

Y+jet balance in the transverse plane

detector noise effects, pile-up

tag&probe like: di-jets events assumed to be balanced, get a relative correction

**Barrel Jet**

\[
p_{T}^{dijet} = \frac{p_{T}^{probes} + p_{T}^{barrel}}{2}
\]

**Probe Jet**

\[
B = \frac{p_{T}^{probes} - p_{T}^{barrel}}{p_{T}^{dijet}}
\]

\[
r = \frac{2+ < B >}{2- < B >}
\]
the simulation
**the simulation**

\[
\sigma = \frac{N_{obs} - N_{bkg}}{\varepsilon \cdot \int \mathcal{L} dt}
\]

\[\varepsilon = \varepsilon_{tr} \cdot \varepsilon_{reco} \cdot \varepsilon_{ID} \cdot \varepsilon_{sel}\]

- calculate what fraction of events from a given decay **falls within the detector acceptance and the selections** of the analysis

- need a **forecast of how the event develops** in space, after the interaction

- the **simulations** are necessary both for known physics objects (Z, W production) and, of course, to build searches for new physics

- the **uncertainty** in the input parameters is source of systematics
the simulation

• calculate **inclusive cross-sections**

• calculate **differential cross sections** as a function of variables of interest in the analysis

• provide **simulated events**, that mimic Physics, and have on average the behaviour foreseen by the theoretical model

---

**LHC**

- parton probability distribution in the proton
- radiation in the process

**detector**

- activity due to the proton remnants
the physics event generation

- add the underlying event
- let hadronic decay
- hadronize partons
- add the parton showers
- add initial and final state radiations
- generate hard process

Diagram:
- LHC
- detector
- Hadronization
- Parton Shower
- Hard SubProcess
- Parton Distributions
- Decay
- Minimum Bias + Collisions
- f(x, Q^2)
the simulation of the detector

- each experiment creates a **simulation of the detector**

- the GEANT program uses generator output (4-vectors) and simulates the **interaction of particles within the detector volume** (need a good description of the geometry):
  - particle ionization in trackers
  - energy deposition in calorimeters
  - intermediate particle decays/radiation

- the GEANT code is merged with (experiment specific) **detector simulation**

- final output: the response of the electronics readout

- MC events are in the **same format as real raw data**
the samples processing

Reality

Events (beam)

Data Acquisition

MC (Virtual Reality)

Event Generator

Detector Simulation

+MC "truth"

Reconstruction, Event Selection

+MC "truth"

Physics Analysis

+MC "truth"

Result

Generates 4-vectors for the particles, resonances, ang. dist., decays, etc. (PYTHIA, HERWIG, ALPGEN, Sherpa...)

Generates detector relevant quantities (GEANT 4)

Apply boundary conditions Acceptance

Inv. mass, efficiency, purity backgrounds, any dist.

Precision $\approx \frac{1}{\sqrt{N}}$

usually:

$N_{\text{MC signal}} \gg N_{\text{Data signal}}$

$N_{\text{MC backg}} \ll N_{\text{Data backg}}$ (QCD)
levels of simulation

Three typical levels of MC simulation:

• Full
  Particle → Energy Deposit → Detector Response → Electronics → Analog Signal → Digitization
  Time consuming, smaller samples

• "Fast" or parameterized
  Intelligently smeared 4-vectors, efficiencies, noise (from data and full MC)
  And/or calorimeter shower libraries
  Larger samples

• Toy
  Only throw from the handful of prob. dist. functions that you care about (with correlations)

  "Roll your own", usually write (easy in root!) and run yourself
  Crazy-large samples, quickly
  To determine probability of fluctuations, checks for systematic effects, etc..

R. Van Kooten, Experimental Techniques
comparison with data

• the simulation is a multi-dimensional parametrization of the knowledge of the detector and standard model predictions

  • is the theoretical simulation correct for the analysis?

    • additional jets production is crucial for analyses that apply a jet veto

    • spin correlations in the Higgs decay need to be treated correctly

  • is the behaviour of the simulation in agreement with data, in the phase space of interest for the analysis?
the pile-up
the pile-up

- At LHC, the interaction rate is higher than the bunch crossing rate.
- Within a bunch crossing in LHC, more interactions happen.
- An event of interesting physics will be recorded together with other events overlapped, that are proton-proton interactions with low physics interest.
- They are equivalent to a non-interesting event (minimum bias).

- Given an average number of interactions, the number of PU events per bunch-crossing is expected to have roughly a poissonian distribution.
measure the pile-up

- multiply the luminosity (per bunch) by the minimum bias cross-section (71.3 mb) gets the expected rate per bunch:

\[ \text{Rate}_{\text{pileup}, \text{xing}, \text{ls}} = \mathcal{L}_{\text{xing}, \text{ls}} \cdot \sigma_{\text{minimum bias}} \]

- divide by the revolution frequency of a bunch to get the number of PU events:

\[ N_{\text{pileup}, \text{xing}, \text{ls}} = \frac{\mathcal{L}_{\text{xing}, \text{ls}} \cdot \sigma_{\text{minimum bias}}}{\text{circulation rate}} \]

- calculate average distributions over longer periods, weighting by the luminosities
effects of pile-up

• fill in the detector with deposits:
  • jet reconstruction algorithms incorporate pile-up deposits
  • lepton isolation cones are filled in with pile-up deposits
  • new jets might appear in the event
  • more hits in the tracker appear
  • the trigger is affected
  • MET resolution worsens
  • ....
how to deal with it

• apply strict **requirements on the vertexing of tracks** - need a precise vertex reconstruction algorithm

• measure the **pile-up density** event by event, and use it to subtract from the jets energy a pile-up term (FastJet)

• do the same with isolation cones

• subtract in the isolation cone the contribution of tracks that do not aim at the same vertex of the lepton

• reconstruct the MET only with particles that aim at a given vertex