



Aspects of HEP – Data Reconstruction (2) and Analysis

Thiago R. F. P. Tomei

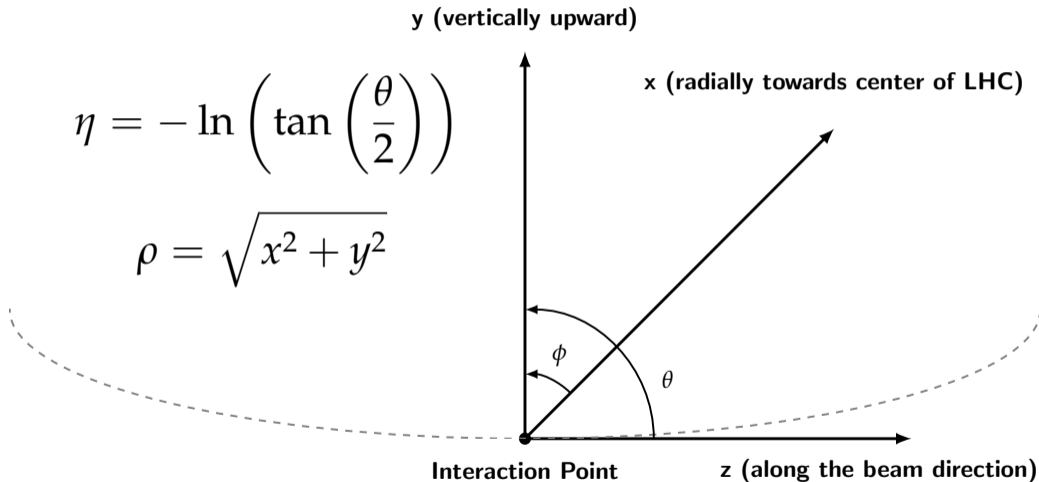
SPRACE-Unesp

Introduction to Reconstruction

After our particles have interacted with our detector, we have to “read out” that information, and format it in such a way that we can understand it. Generally, that entails some sort of transducer that converts the energy deposited in our detector elements and transforms it into electrical signals. It is also customary to digitise the signals for further data processing.

- **Local reconstruction** starts with the detector readout (e.g., channel #28 has 7 ADC counts), and transforms it into intelligible data – a “hit”. This is usually low-level information localised to one subsystem, e.g., for a tracker detector it represents not much more than “a charged particle passed through this sensor”.
- **Global reconstruction** starts with the hits and tries to group them at a higher-level. Connecting the hits in a silicon detector to reconstruct the trajectory of a charged particle, or clustering the hits in a scintillator to reconstruct the deposit of a photon, fall in this category.
- **Global event description** uses the high-level objects to describe the event as a whole, ready to be connected to a physics process interpretation. Connecting energy clusters to a track to try to reconstruct an electron or a charged hadron, or connecting tracks in two different tracker systems to reconstruct a muon, fall in this category.

Coordinate System



Global Reconstruction: Tracking as Example

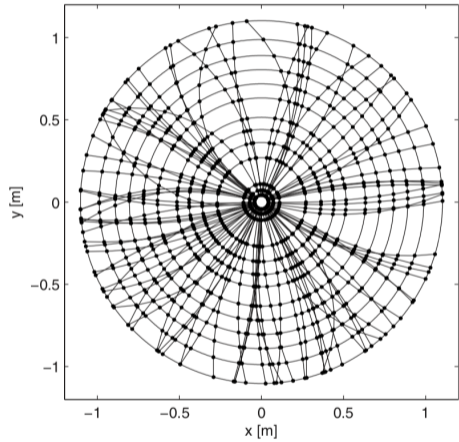
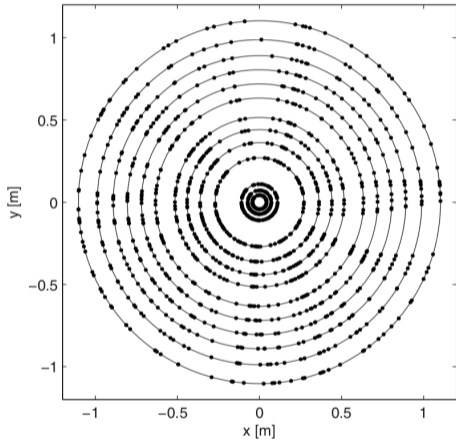
Track reconstruction refers to the process of using the hits, obtained from the local reconstruction, to obtain estimates for the momentum and position parameters of the charged particles responsible for the hits (tracks).

- Translate the hits' local coordinate system to the tracks' global coordinate system.
 - Misalignment of detector elements: additional uncertainty.

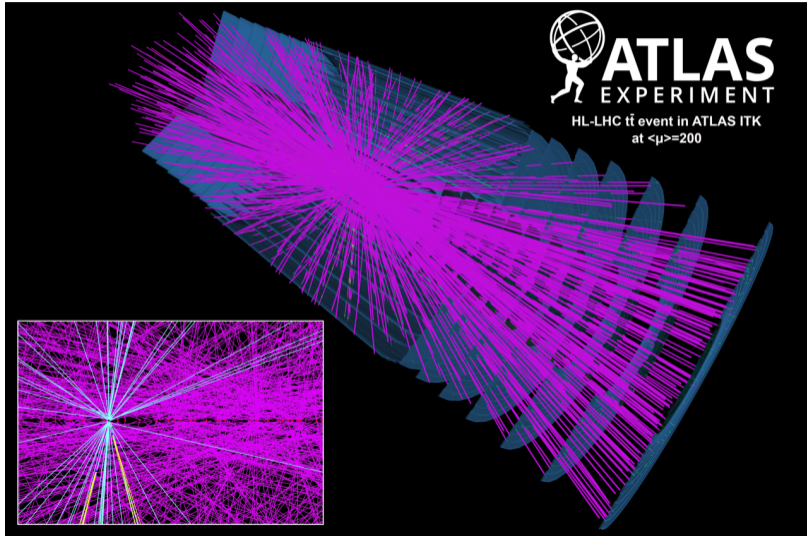
Traditionally divided in two steps:

- **Track finding:** decide, within the set of reconstructed hits $\{X_i\}$, which subset $\{X_i\}^k$ makes up a given track candidate k .
- **Track fitting:** from the subset of hits that make up a given track candidate, estimate the parameters – production vertex position \mathbf{x}_v and momentum at vertex \mathbf{p}_v – of the originating particle.

Tracking: A Hard Problem



Tracking: A **Very** Hard Problem



Motion in Magnetic Field

In a magnetic field the motion of a charged particle is determined by the Lorentz Force and finally

$$\frac{d\mathbf{p}}{dt} = e\mathbf{v} \times \mathbf{B}$$

Since magnetic forces do not change the energy of the particle

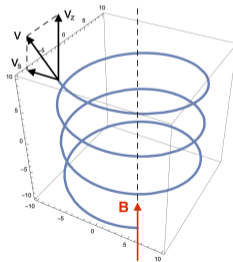
$$m_o\gamma \frac{d\mathbf{v}}{dt} = e\mathbf{v} \times \mathbf{B}$$
$$m_o\gamma \frac{d^2\mathbf{r}}{dt^2} = e \frac{d\mathbf{r}}{dt} \times \mathbf{B}$$

using the path length s along the track instead of the time t , $ds = vdt$ we have

$$m_o\gamma v \frac{d^2\mathbf{r}}{ds^2} = e \frac{d\mathbf{r}}{ds} \times \mathbf{B}$$

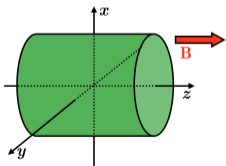
$$\frac{d^2\mathbf{r}}{ds^2} = \frac{e}{p} \frac{d\mathbf{r}}{ds} \times \mathbf{B}$$

- Inhomogeneous $\mathbf{B}(s)$: solve differential equation to find the trajectory $r(s)$.
- Homogeneous \mathbf{B} : helix trajectory – circle in plane transverse to \mathbf{B} .



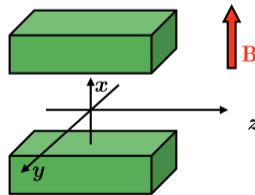
Solenoidal field

- cylindrical symmetry
- deflection in $x - y$ ($\rho - \phi$) plane
- tracking detectors arranged in cylindrical shells
- measurement of curved trajectories in $\rho - \phi$ planes at fixed ρ



Dipole field

- rectangular symmetry
- deflection in $y - z$ plane
- tracking detectors arranged in parallel planes
- measurement of curved trajectories in $y - z$ planes at fixed z



The Helix Equation

Parametric form:

$$x(s) = x_o + R \left[\cos \left(\Phi_o + \frac{hs \cos \lambda}{R} \right) - \cos \Phi_o \right]$$

$$y(s) = y_o + R \left[\sin \left(\Phi_o + \frac{hs \cos \lambda}{R} \right) - \sin \Phi_o \right]$$

$$z(s) = z_o + s \sin \lambda$$

λ is the dip angle. $h = \pm 1$ is the sense of rotation on the helix. The projection on the $x - y$ plane is a circle:

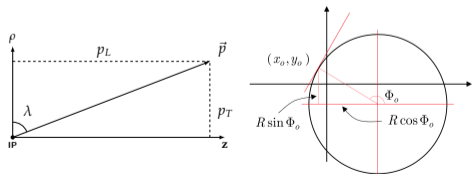
$$(x - x_o + R \cos \Phi_o)^2 + (y - y_o + R \sin \Phi_o)^2 = R^2$$

x_o and y_o are the coordinates at $s = 0$, while Φ_o is related to the slope of the tangent to the circle at $s = 0$.

In the transverse plane $x - y$, we can relate the curvature radius R to the transverse momentum p_T as:

$$R = \frac{p_T}{0.003 q B}$$

R in cm, p_T in GeV, q in elementary charge, B in T.



Conformal Mapping

Circle going through the origin: $(x - a)^2 + (y - b)^2 = R^2 = a^2 + b^2$. Define:

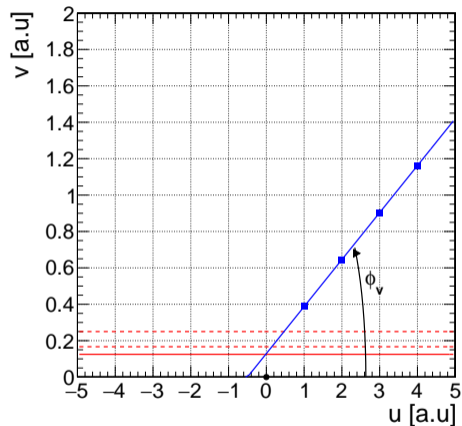
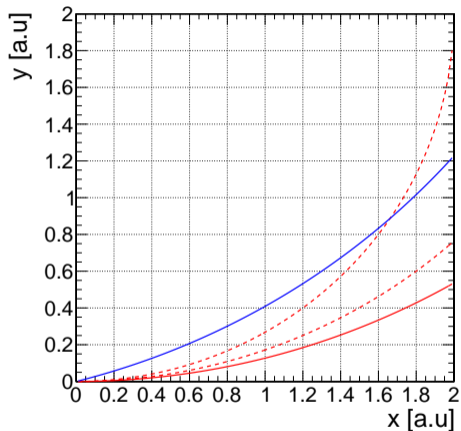
$$u = \frac{x}{x^2 + y^2}, v = \frac{y}{x^2 + y^2}, \text{ leading to } v = \frac{1}{2b} - u \frac{a}{b}$$

How to obtain track candidates:

- Convert hits from (x,y) to (u,v) .
- In the (u,v) coordinate system, the larger the momentum, the closer line is to origin.
- Approximate line as going through origin – hits of a track have constant ϕ_v .

In practice, we search for peaks in the histogram of ϕ_v variable of all hits.

Conformal Mapping Illustrated



All the blue points have, approximately, the same ϕ_v .

Hough Transform

Remember: the radius of curvature is $R = \frac{p_T}{0.003 qB}$. We are interested in particles originating at or close to the interaction point. The trajectory of such particles in the transverse plane is described by the following equation:

$$\frac{r}{2R} = \sin(\varphi - \phi) \approx \varphi - \phi$$

Here ϕ is the angle of the track in the transverse plane at the origin ($\phi = \Phi_o^{\text{helix}} - \pi/2$). The small angle approximation is valid for tracks with $p_T \gtrsim 2$ GeV. Combine the two equations:

$$\phi = \varphi - \frac{0.0015 qB}{p_T} \cdot r$$

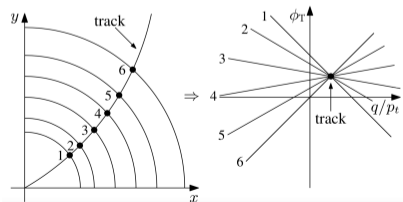
So, one hit in (r, φ) maps to a straight line in the track parameter space $(q/p_T, \phi)$.

If several hits are produced by the same particle, all the lines intersect at the same point.

- Locate the intersection to identify a track.
- Intersection coordinates locate provide track parameters $(q/p_T, \phi)$.
- **Caveat:** the gradient of each line is proportional to the hit radius r – always positive. Better use variables r_T, ϕ_T instead:

$$r_T = r - T, \quad \phi_T = \phi_{\text{track}}(r_T)$$

Now r_T can have both signs; this leads to a larger range of line gradients and more precision in the measurement of the intersection point.



Track Road and Track Following

These are local approaches, i.e., start already from a subset of the hits.

- **Track Road:** start from hits that could have been created by the same particle. Use the expected trajectory shape to interpolate amongst the hits, creating a "road" around the trajectory. All hits that fall in the road constitute the track candidate.
- **Track Following:** start from a "track seed" – a subset of hits that for some reason are deemed to be more accurate. From the seed, extrapolate until the next detector element containing a hit. The closest hit becomes part of the track candidate. Iterate until either 1) too many layers have missing hits or 2) the end of the detector system is reached.

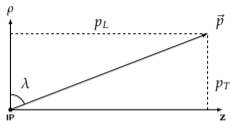


Elements of Track Fitting

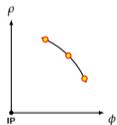
- Track parameterization: depends on detector geometry. For helix trajectory in solenoid, we could use the parameter set $\{x_o, y_o, \Phi_o, \lambda, R\}$ we discussed.
- Track model: how the track parameter or state vector \mathbf{q} at a given surface k depends on the state vector on a different surface i : $\mathbf{q}_k = \mathbf{G}_{k|i}(\mathbf{q}_i)$, where $\mathbf{G}_{k|i}$ is the track propagator from surface i to j .
- Error propagation: during the track parameter estimation procedure, propagation of the track parameter covariance matrix along with the track parameters themselves is often requested.
- Material effects: ionization energy loss, multiple Coulomb scattering, bremsstrahlung, all affect the particle propagation.
- Measurement model: \mathbf{h}_k describes the functional dependence of the measured quantities in layer k , \mathbf{m}_k , on the state vector at the same layer: $\mathbf{m}_k = \mathbf{h}_k(\mathbf{q}_k)$.

Track Fitting – Momentum Measurement (1)

Particle momentum projected along two directions:

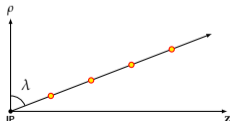


In the $\rho - \phi$ plane we measure the transverse momentum p_T through the curvature radius R :



$$p_T = p \cos \lambda = 0.003 BR$$

In the $\rho - z$ plane we measure the dip angle λ :

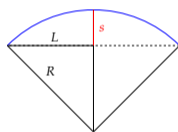


Orders of magnitude:

$$p_T = 1 \text{ GeV} \quad B = 2T \quad R = 1.67 \text{ m}$$

$$p_T = 10 \text{ GeV} \quad B = 2T \quad R = 16.7 \text{ m}$$

The **sagitta** s :



$$R^2 = L^2 + (R - s)^2$$

$$R = \frac{s}{2} + \frac{L^2}{2s}$$

$$s \ll R \Rightarrow R \approx \frac{L^2}{2s}$$

Assuming a track length of 1m:

$$p_T = 1 \text{ GeV} \quad s = 7.4 \text{ cm}$$

$$p_T = 10 \text{ GeV} \quad s = 0.74 \text{ cm}$$

Track Fitting – Momentum Measurement (2)

Once we have measured the transverse momentum and the dip angle the total momentum is

$$p = \frac{p_T}{\cos \lambda} = \frac{0.003 BR}{\cos \lambda}$$

The error in the momentum is easily calculated

$$\frac{\partial p}{\partial R} = \frac{p_T}{R}$$
$$\frac{\partial p}{\partial \lambda} = -p_T \tan \lambda$$

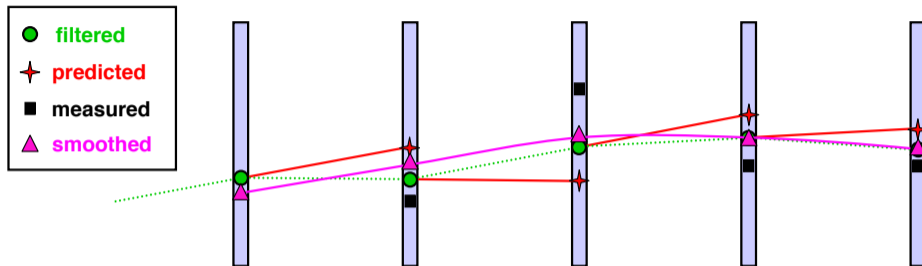
$$\left(\frac{\Delta p}{p}\right)^2 = \left(\frac{\Delta R}{R}\right)^2 + (\tan \lambda \Delta \lambda)^2$$

We need to study:

- the error on the radius measured in the bending plane $\rho - \phi$;
- the error on the dip angle in the $\rho - z$ plane;
- contribution of multiple scattering to momentum resolution.

In an hadronic collider the main emphasis is on the **transverse momentum** p_T . Since the elementary processes happen amongst partons that are not at rest in the lab frame, we employ momentum conservation only in the transverse plane.

Kalman Filter – The Ideas



- To start the Kalman Filter we need a “seed”;
- the position on the next plane is **predicted**;
- the **measurement** is considered;
- prediction and measurement are merged (**filtered**);
- iterate: **prediction** – **measurement** – **filtering** – ...

— **filtered trajectory**
— **predicted trajectory**
— **smoothed trajectory**

Filtering: weighted average of the new measurement y_n and the new prediction y_p .
Smoothing: refine each position estimate with information from all measurements.

Kalman Filter – The Equations

Dimensions

- P parameters for track model.
- D measurements \mathbf{m}_k at each point (usually $D < P$).
- K measurement points ($N = K \times D$).
- State of the system at index k is \mathbf{q}_k .

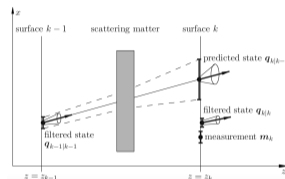
Iterative method

- Using system equation $\mathbf{q}_k = \mathbf{G} \cdot \mathbf{q}_{k-1} + \boldsymbol{\omega}_k$.
 - $\mathbf{G} = P \times P$ matrix, $\boldsymbol{\omega} =$ perturbation associated with covariance $P \times P$ matrix V_ω .
- Add new point, update parameters and cov., using meas. equation $\mathbf{m}_k = \mathbf{H} \cdot \mathbf{q}_k + \boldsymbol{\epsilon}_k$.
 - $\mathbf{H} = D \times P$ matrix, $\boldsymbol{\epsilon} =$ measurement error associated with diagonal cov. $D \times D$ matrix V_m .
 - Weighted means of prediction and measurement using variance $\Leftrightarrow \chi^2$ fit.

$$\mathbf{q}_k = \left(V_{k|k-1}^{-1} \mathbf{q}_{k|k-1} + \mathbf{H}^T V_{m_k}^{-1} \mathbf{m}_k \right) \cdot \left(V_{k|k-1}^{-1} + \mathbf{H}^T V_{m_k}^{-1} \mathbf{H} \right)^{-1}$$

Starting point

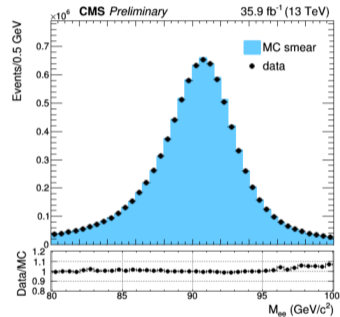
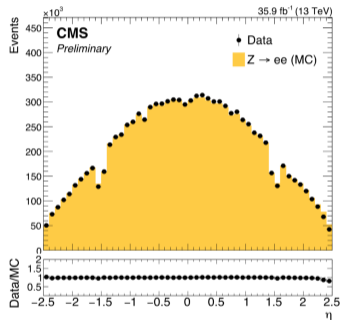
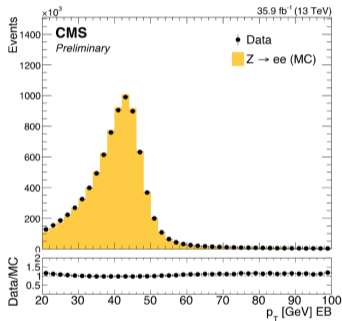
- Initial set of parameters (seeds): first measurements.



Global Event Description, a.k.a. “Particle Flow (PF)”

- All of the subsystems provide basic elements: silicon tracks, calorimeter hits (EM and hadronic), muon spectrometer tracks, ...
- Basic idea: correlate basic elements to identify **all final-state particles**. Combine corresponding measurements for ultimate accuracy in particle properties.
 - Initially developed and used by ALEPH (LEP).
 - Had not been used with much success in hadron colliders – detector granularity was not fine enough!
- Complicated algorithm that implements a set of links:
 - Electrons leave both tracks and EM calorimeter deposits; photons leave only the latter.
 - Muons leave tracks both in silicon trackers and in muon spectrometers.
 - Charged hadrons leave both tracks and hadronic calorimeter deposits; neutral hadrons leave only the latter.
 - Electrons emit photons during bremsstrahlung, all that energy must be reclustered.
 - A hadronic τ decay produces just a few hadrons and one neutrino.
- The output of the PF algorithm is a list of final-state reconstructed particles.

Example: Electron Reconstruction with ECAL + Tracker



Hadronic Jets and Missing p_T

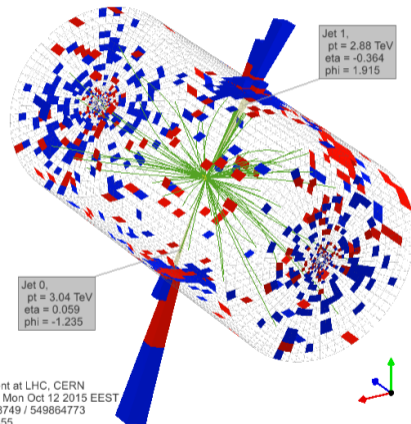
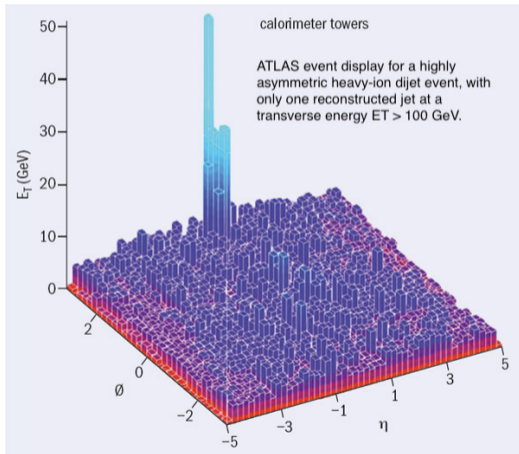
We saw in the the first lecture that quarks and gluons fragment and hadronize, producing a collimated, high-momentum spray of hadrons flying roughly in the same direction – a hadronic jet. A **jet algorithm** is an iterative, tuneable procedure that clusters close particles and creates the jet objects.

In the same vein, we can cluster all observed particles. This allows for an important cross-check: the total transverse momentum in a hadron collision should be zero. If, after the overall clustering, a sizeable transverse momentum remains, that would be evidence for an undetected particle being produced!

We define the **missing transverse momentum** \vec{p}_T^{miss} :

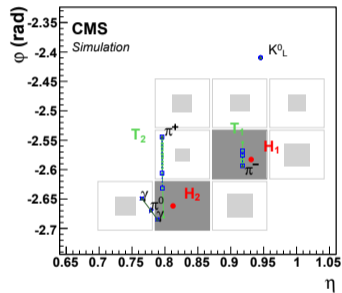
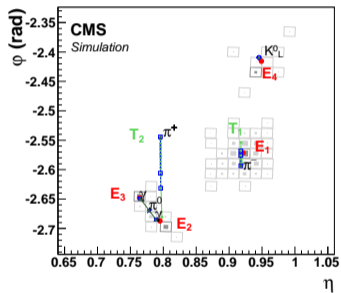
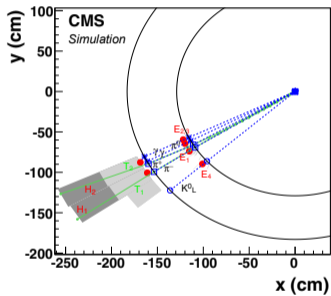
$$\vec{p}_T^{\text{miss}} = - \sum_i \vec{p}_T^i$$

Event Displays of Calorimeter Jets



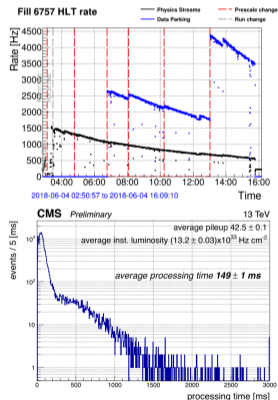
CMS Experiment at LHC, CERN
Data recorded: Mon Oct 12 2015 EEST
Run/Event: 258749 / 549864773
Lumi section: 355
Dijet Mass: 6.14 TeV

Event Display of a PF Hadronic Jet

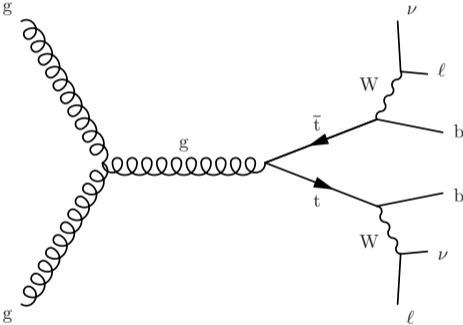


Online Reconstruction

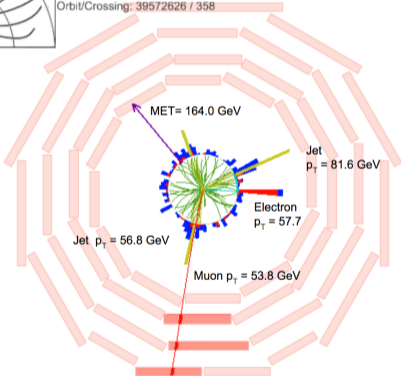
- With a collision rate close to 40 MHz, impossible to record and store all collisions happening at the LHC experiments.
- **Trigger system** implemented to keep only the more interesting collisions – quasi-real time reconstruction.
- Level 1 (L1) trigger in CMS:
 - Customizable hardware (ASICs and FPGAs).
 - Output rate: 100 kHz (detector readout constraint).
 - Timing: 4 μ s (available buffer).
 - Coarse granularity.
- High Level Trigger (HLT) in CMS:
 - PC farm with light version of the offline software.
 - Output rate: 1 kHz average over LHC fill.
 - Timing: \lesssim 380 ms.
 - Full detector granularity.



A Fully Reconstructed Event



CMS Experiment at LHC, CERN
Data recorded: Wed Jul 8 19:26:24 2015 CEST
Run/Event: 251244 / 83494441
Lumi section: 151
Orbit/Crossing: 39572626 / 358





Reconstruction is ready...

... onwards to Data Analysis!

Introduction to Data Analysis

In broad terms, we can divide the LHC data analyses in two camps:

- **Measurements**, in which one is trying to measure some known standard model quantity. Examples go from very simple quantities, like the Higgs boson mass M_H , to convoluted quantities like the fully inclusive top quark pair production at the LHC at 14 TeV, $\sigma(\text{pp} \rightarrow t\bar{t} + \text{anything})$.
- **Searches**, in which one tries to uncover evidence of discrepancies between the standard model predictions and the observed data. Examples include new searches for resonances, supersymmetry, dark matter. . .

But this distinction is a bit artificial! Consider:

- $\text{H} \rightarrow \mu\mu$ at the 13 TeV LHC has a well-defined SM prediction: $\sigma(\text{pp} \rightarrow \text{H} \rightarrow \mu\mu) \simeq 12.08 \text{ fb}$. This has not been observed yet, so we call it a search.
- When searching for a new resonance, $\text{pp} \rightarrow Z' \rightarrow ee$, we usually make (multiple) assumptions on the value of its mass, spin, etc. We then try to measure its production cross-section – usually coming up with a value statistically compatible with zero.

What do we Actually Measure?

Easy answer – **we count the number of events in a given configuration**. The observed number of events for a given physics process N_{obs} is:

$$N_{\text{obs}} = \sigma \times \mathcal{L} \times A \times \epsilon$$

where:

- σ is the production cross-section. This is a purely theoretical-driven quantity – it is the very σ we learned how to calculate on the first lecture.
- \mathcal{L} is the accelerator luminosity. This is a measure of how many particles we are able to fit through a given space in a given time. We discussed it on the second lecture.
- A is the acceptance. It measures, for that given process, the ratio of detectable particles that actually go into the detector volume.
 - Technically, this is also theoretically-driven, but it is customary to factor it out like this.
- ϵ is the selection efficiency. It can be subdivided into two parts, $\epsilon = \epsilon_o \times \epsilon_a$:

- ϵ_o is the efficiency of reconstructing a given set of objects in the detector. We discussed it on the third lecture.
- ϵ_a is the efficiency of any further requirements done in the analysis (a.k.a. fun ☺).

The “given configuration” we discussed above is determined by the acceptance \times efficiency product. The total number of events we observe in that configuration is:

$$N_{\text{obs}} = \sum_i N_{\text{obs}}^i + N_{\text{fakes}}$$

where the sum runs over **all** of the physics process that produce events in that configuration. There may be spurious contributions N_{fakes} from any kinds of non-collision effects. Those include:

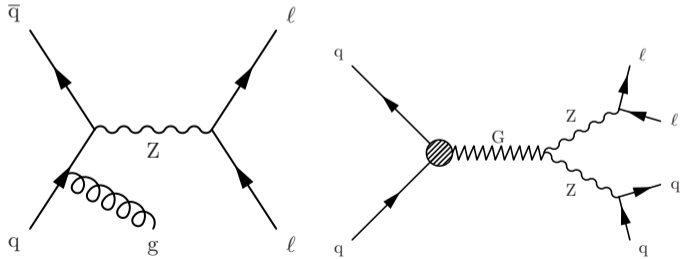
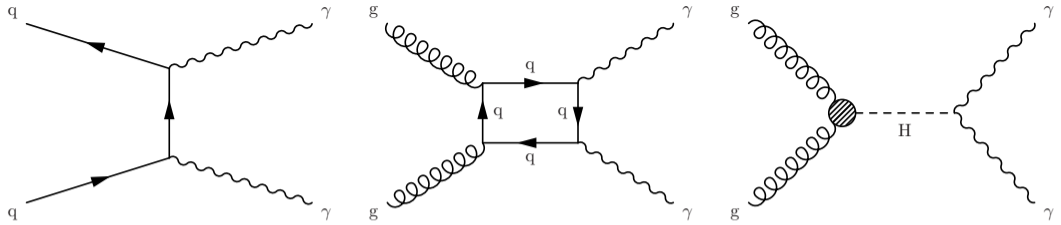
- Cosmic rays going through the detector (fake muons).
- Longitudinal particles from beam interactions with accelerator elements (beam halo).
- Hardware failures: “hot cells”, dead channels, high voltage spikes.
- Software failures: e.g., unusual configurations of hits leading to high numbers of fake tracks.
- The phase of the moon (?)
 - Nucl. Instrum. Methods Phys. Res., A 357 (1995) 249-252
- The seasonal variation of rainfall (???)
 - Proceedings of the 1999 Particle Accelerator Conference, New York, 1999
- The schedule of the French high-speed rail trains (?????)
 - Nucl. Instrum. Methods Phys. Res., A 417 (1998) 9-15

But let's go back to collision processes. . .

Signal and Background

The fact that the sum $\sum_i N_{\text{obs}}^i$ goes over all physics processes means that we cannot readily separate the process in which we are interested – the **signal**. All of the other processes constitute the **background** for that measurement. We can separate backgrounds in two sets:

- Irreducible backgrounds are those that share the exact same final state as the signal. For instance, the nonresonant diphoton production $pp \rightarrow \gamma\gamma$ is an irreducible background to Higgs boson production in that channel, $pp \rightarrow H \rightarrow \gamma\gamma$. The only option is to model them as well as possible.
- Reducible backgrounds are those where the final state differs from that of the signal, but due to various reasons end up being selected by our analysis. An example would be inclusive $Z \rightarrow \ell\ell$ production being a background for a $ZZ \rightarrow \ell\ell qq$ search: the former could appear as a “dilepton + jet” final state, and that jet could be mistaken to be the $Z \rightarrow qq$ leg of the latter.



Separating Signal and Background

Our first task is to search for observables that are differently distributed for signal and background.

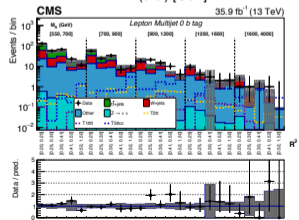
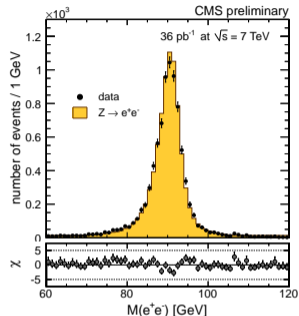
Some searches are easy, for $Z \rightarrow \ell\ell$:

- 2 high- p_T leptons.
- Same flavour, opposite charges.
- Invariant mass $M_{\ell\ell}$ in 70–110 GeV range.

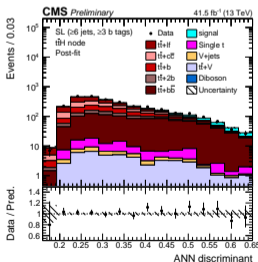
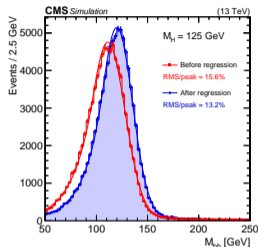
Other searches are harder. The Razor variables used for SUSY searches:

$$M_R \equiv \sqrt{(|\vec{p}^{j_1}| + |\vec{p}^{j_2}|)^2 - (p_z^{j_1} + p_z^{j_2})^2} \quad \text{and} \quad R^2 \equiv \left(\frac{M_R^R}{M_R} \right)^2$$

$$\text{with } M_T^R \equiv \sqrt{\frac{p_T^{\text{miss}} (p_T^{j_1} + p_T^{j_2}) - \vec{p}_T^{\text{miss}} \cdot (\vec{p}_T^{j_1} + \vec{p}_T^{j_2})}{2}}$$



Machine Learning in Analysis

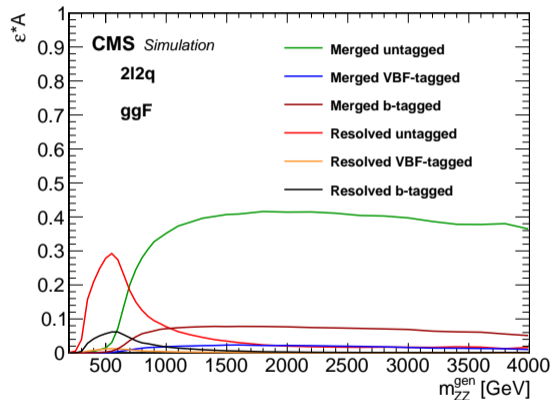


Some times, we don't need to search for variables. . . the computer can do it for us!
Machine Learning (ML) techniques can be applied to two kinds of problems:

- **Classification problems:** the output variable takes class labels. Useful for analysis, e.g. classify the event as signal or background.
 - Usually we unpack the classification and work with the output variable directly.
- **Regression problems:** the output variable takes continuous values. Useful for reconstruction, e.g. energy of a b-jet based on its kinematics and flavour content.

Cut-Based Analyses

- Design a set of cuts by as unbiased a procedure as possible.
 - Blind study: try to avoid observer bias and confirmation bias. Don't look at the data until you have frozen the analysis.
- Choose cuts that optimise the final overall accuracy of the result.
 - Difficult tradeoff between statistical and systematic uncertainties (see later).
- Always study the marginal effect of each of your cuts by tables and plots.
 - Cuts with no marginal effect (that is, they remove no events after all other cuts) are quite useless.

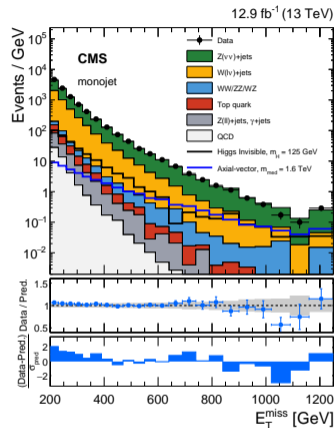
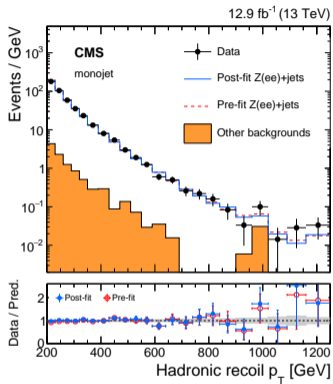
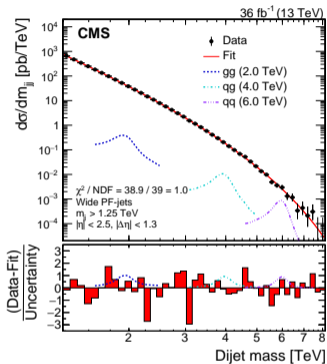


Background Estimation

After you optimise your analysis, some background always* remains. An array of options are available to estimate it:

- Fully trust the simulation: not recommended, except maybe for backgrounds so small that even an error by a large factor would make no difference.
- Trust a “data corrected” simulation: usually done by defining **control regions** (a.k.a “sidebands”), in which you expect similar behaviour of your background but a near-absence of your signal.
- Model your background “in situ”: the same, but your control regions act simultaneously as measurement regions for some other modality of your search.

* Even if it doesn't, you still have to find a way to put an uncertainty on that zero!



Basic Concepts in Probability

□ **Mathematical probability:** given the set of all possible exclusive, elementary events X_i , probability of occurrence of X_i is $P(X_i)$ and follows the Kolmogorov axioms:

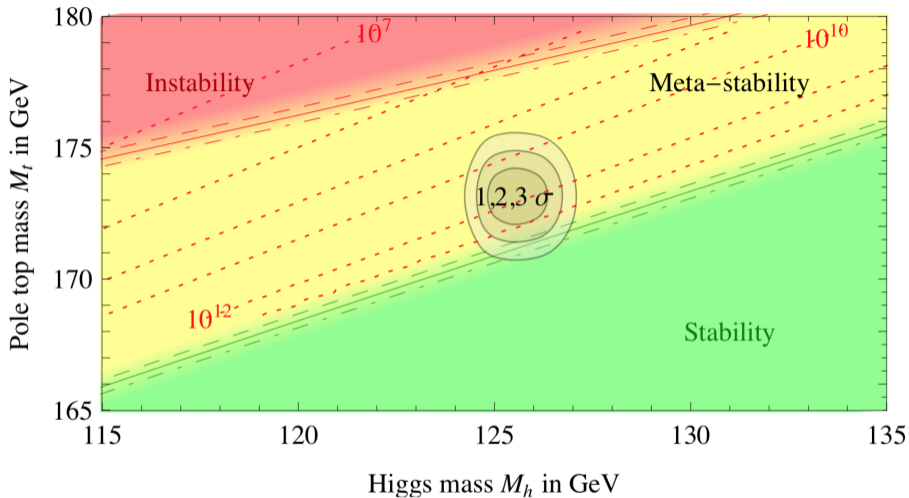
- $P(X_i) \geq 0$ for all i
- $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$
- $\sum_{\Omega} P(X_i) = 1$

□ **Frequentist probability:** if you observe N events, and n of them are of type X , the probability that any single event will be of type X is the “empirical” limit of the frequency ratio:

$$P(X) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

- Approximate the probability by making N large.
- Experiments have to be **repeatable** – but repeatable means that all the **relevant** conditions are the same. Good science should produce **reproducible results**.

- **Bayesian probability:** it is the degree of belief in X . Operational definition is based on the **coherent bet:** [Finetti1974]
- “The idea is to determine how strongly a person believes that X will occur by determining how much he would be willing to bet on it, assuming that he would be willing to bet on it, assuming that he wins a fixed amount if X does later occur and nothing if it fails to occur. Then $P(X)$ is defined as the largest amount he would be willing to bet, divided by the amount he stands to win.” [James2006].
 - This follows the Kolmogorov axioms.
 - However, it is a property of both the observer and the observed system – it will in general change if the observer obtains more knowledge. It is a **subjective** probability!
 - On the other hand, it helps addressing some questions that we want to try to answer:
 - “What is the probability that the universe is (cosmologically) flat?”
 - “What is the probability that the Higgs vacuum is stable?”
 - There is a lot of work in studying objective Bayesian statistics (H Jeffreys, E. T. Jaynes, S. James, J. Berger, The science is far from settled!



Bayes Theorem

For discrete events:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \xrightarrow{\text{Bayesian}} \quad P(\theta_i|\mathbf{X}^0) = \frac{P(\mathbf{X}^0|\theta_i) \cdot P(\theta_i)}{P(\mathbf{X}^0)}$$

For continuous random variables:

$$q(Y|X) = \frac{p(X|Y) h(y)}{g(X)} \quad \xrightarrow{\text{Bayesian}} \quad p(\theta|\mathbf{X}^0) = \frac{p(\mathbf{X}^0|\theta) p(\theta)}{\int p(\mathbf{X}^0|\theta) p(\theta) d\theta}$$

- $p(\theta|\mathbf{X}^0)$ is a p.d.f, the posterior probability density for θ .
- $p(\mathbf{X}^0|\theta)$ is the likelihood function $L(\theta)$. It is not a p.d.f
- $p(\theta)$ is the prior probability density for θ . **Here lies the major problem!**

Example (straight from Wikipedia):

- There are two subspecies of beetle – the “common one” C and the “rare one” R .
- An entomologist spots what might be a rare subspecies of beetle, due to the pattern X on its back.
- In the rare subspecies, 98% have the pattern, or $P(X|R) = 98\%$. In the common subspecies, 5% have the pattern, or $P(X|C)$.
- The rare subspecies accounts for only 0.1% of the population.
- How likely is the beetle having the pattern to be rare, or what is $P(R|X)$?

Example (straight from Wikipedia):

- There are two subspecies of beetle – the “common one” C and the “rare one” R .
- An entomologist spots what might be a rare subspecies of beetle, due to the pattern X on its back.
- In the rare subspecies, 98% have the pattern, or $P(X|R) = 98\%$. In the common subspecies, 5% have the pattern, or $P(X|C)$.
- The rare subspecies accounts for only 0.1% of the population.
- How likely is the beetle having the pattern to be rare, or what is $P(R|X)$?

$$P(R|X) = \frac{P(X|R)P(R)}{P(X)} = \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.05 \times 0.999} \simeq 1.9\% (!!!)$$

The Prior ProblemS – stolen from M. Pierini (CERN)

- The need of priors in Bayesian statistics is a problem for some physicist

The origin of the problem lies in the very first Bayesian assumption, namely that unknown model parameters are to be understood as mathematical objects distributed according to PDFs, which are assumed to be known: the priors. Obviously, the choice of the priors cannot be irrelevant; hence, the Bayesian treatment is doomed to lead to results which depend on the decisions made, necessarily on unscientific basis, by the authors of a given analysis, for the choice of these extraordinary PDFs.

J. Charles et al. hep-ph/0607246

- The lack of priors in nonBayesian statistics is a problem for some statistician

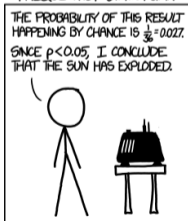
The frequentist approach to hypothesis testing does not permit researchers to place probabilities of being correct on the competing hypotheses. This is because of the limitations on mathematical probabilities used by frequentists. For the frequentists, probabilities can only be defined for random variables, and hypotheses are not variables (they are not observables)... This limitation for frequentists is a real drawback because the applied researcher would really like to be able to place a degree of belief on the hypothesis. He or she would like to see how the weight of evidence modifies his/her degree of belief (probability) on the hypothesis being true.

J. Press, *Subjective and Objective Bayesian Statistics*

DID THE SUN JUST EXPLODE? (IT'S NIGHT, SO WE'RE NOT SURE.)



FREQUENTIST STATISTICIAN:



BAYESIAN STATISTICIAN:



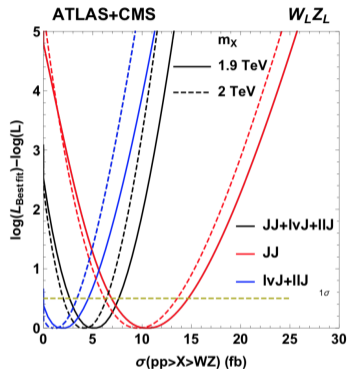
- Point estimation – find a single value $\hat{\theta}$ that is “as close as possible” to the true parameter θ we want to measure. We usually use the maximum likelihood estimator

$$\frac{\partial \ln L}{\partial \theta_i} = 0$$

which is optimal in the asymptotic limit of large N .

- But in general it is better to report the likelihood function itself, at least near its maxima.
- Interval estimation – find the range $\theta_a \leq \theta \leq \theta_b$ that contains the true value θ_0 with probability β .
 - 1D: trivial, use the Neyman construction with the Feldman-Cousins “unified approach”;
 - ND: use profile likelihood (MINOS). As a bonus, it allows for the removal of nuisance parameters μ by maximising the full likelihood, at each value of the parameter of interest θ .

Example of Profile Likelihood



The profiled combined likelihoods for ATLAS and CMS Run 1 diboson resonance searches. The best-fit cross-section for the $W' \rightarrow W_L Z_L$ with a W' mass of 1.9 TeV was $\sigma = 5.3^{+2.3}_{-2.0}$ fb.

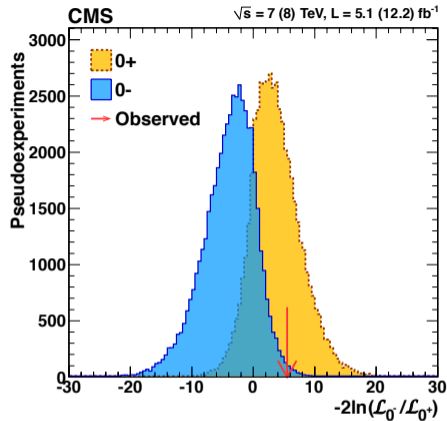
Hypothesis Testing – from Bob Cousins

Consider two hypotheses:

- H_0 : is the null hypothesis. For instance, “the Standard Model is a true description of nature at the scales probed by the LHC”.
- H_1 : is the alternative hypothesis. For instance, “⟨INSERT YOUR MODEL HERE⟩ is a true description of nature at the scales probed by the LHC”.

$L(X, \theta)$ is different for H_0 and H_1 . How do we test the two hypotheses against each other?

- For the null hypothesis H_0 , order possible observations x from least extreme to most extreme, using an ordering principle (which can depend on H_1 as well). Choose a cutoff α (smallish number).
- “Reject” H_0 if the observed x_0 is in the most extreme fraction α of observations x (generated under H_0). By construction:
 - α = probability (with x generated according to H_0) of rejecting H_0 when it is true;
 - β = probability (with x generated according to H_1) of not rejecting H_0 when it is false.



Expected distribution of $-2\ln(L(0^-)/L(0^+))$ under the pure pseudoscalar and pure scalar hypotheses (histograms) for the Higgs boson. The arrow indicates the value determined from the CMS observed data with the discovery dataset (7 TeV, 5.1 fb^{-1} and 8 TeV, 12.2 fb^{-1}).

A decision on whether or not to declare discovery (falsifying H_0) requires 2 more inputs, both of which can affect the choice of α :

- **Prior** belief in H_0 vs H_1 .
- Cost of Type I error (false discovery claim) vs cost of Type II error (missed discovery).

A one-size-fits-all criterion of α corresponding to 5σ is without foundation!

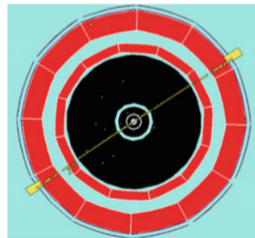
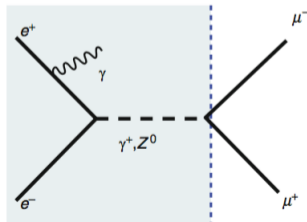
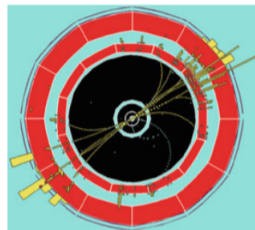
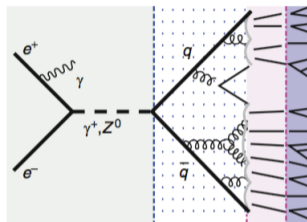
...and still, I (Thiago) am positive it will still be used for the rest of the days of the LHC.

Example: Ratio of Hadronic & Leptonic Cross Sections in e^+e^- Reactions

- For energy range where five quark flavours contribute and below the Z resonance (for lowest order in perturbation theory)

$$R_\gamma = \frac{\sigma(e^+e^- \rightarrow \text{hadrons})}{\sigma(e^+e^- \rightarrow \mu^+\mu^-)} = \frac{\sigma_{\text{had}}}{\sigma_{\text{lep}}}$$
$$= N_c \sum_q e_q^2 = N_c \frac{11}{9}$$

- Goal: determine or constrain the number of colour states (N_c)



- Count the number of events with hadronic (N_{had}) and leptonic (N_{lep}) final states

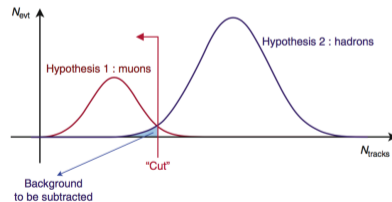
$$R_{\gamma} = \frac{\sigma_{\text{had}}}{\sigma_{\text{lep}}} = \frac{N_{\text{had}}/\mathcal{L}}{N_{\text{lep}}/\mathcal{L}} = \frac{N_{\text{had}}}{N_{\text{lep}}}$$

- Define event selection and estimate backgrounds, possibly with input from simulation to define discrimination variables.

- In this case, the number of charged particles (N_{tracks}) in each event helps to separate hadronic from leptonic events. Usually leptonic events have few tracks, whilst hadronic events have many more.
- But there is some overlap, so the selection has an efficiency ($\epsilon \leq 1$) to select a given type of event.

$$N_{\text{had/lep}}^{\text{meas}} = \epsilon_{\text{had/lep}}^{\text{meas}} \cdot N_{\text{had/lep}}^{\text{true}}$$

- $N_{\text{had/lep}}^{\text{meas}}$ should be corrected. In practice, what we do is estimate the efficiencies (using either simulation or data).



□ Analysis

- In general efficiency is not only due to the cut, but also includes: $\epsilon = A \cdot \epsilon_{\text{tri}} \cdot \epsilon_{\text{rec}} \cdot \epsilon_{\text{cut}}$
 - A : detector acceptance
 - $\epsilon_{\text{tri}}, \epsilon_{\text{rec}}, \epsilon_{\text{cut}}$: trigger, reconstruction, and cut efficiencies

- Background subtraction

$$N_{\text{lep}}^{\text{true}} = (N_{\text{lep}}^{\text{meas}} - N_{\text{bckg}}) / \epsilon_{\text{lep}}$$

- Statistical uncertainties

- Can be estimated considering the statistical distributions followed in each measurement.
- For counting experiment a Poisson distribution can be a proper choice.
- For efficiency measurement (pass or fail), an uncertainty following a binomial would be preferable.

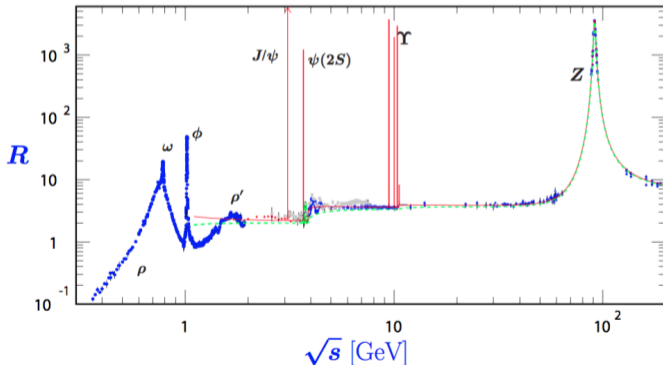
- Systematic uncertainties

- In general, it will depend on each analysis (no standard procedure).
- E.g. in the present case we could assign an uncertainty associated to the mismodelling of detector response by the simulation, which was used for efficiency correction.

Ratio of Hadronic & Leptonic Cross Sections in e^+e^- Reactions: Result

□ Measurements of R from different lepton colliders

- The relation for R_γ in previous slides applies to region $\sqrt{s} > 10$ GeV (where quarks $u, d, s, c,$ and b contribute) and far from the Z boson peak.
- In this case we have $R_\gamma = N_c \frac{11}{9} \approx \frac{11}{3} \Rightarrow N_c \approx 3$.





ROOT is a modular scientific software toolkit. It provides all the functionalities needed to deal with big data processing, statistical analysis, visualisation and storage. It is currently used by all the LHC experiments.



SciPy is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, some of the core packages are NumPy, SciPy library, Matplotlib, IPython, SymPy, pandas.