Data analysis

Peter H. Hansen

University of Copenhagen

Content

preface

- Data analysis depends on the experimental setup and the scientific objective. So it is not ideal to use only one example:
  The direct search of the Higgs boson using the ALEPH detector at LEP.
- However, this example illustrates the general need to construct a hierarchy of estimators from the raw data:
  - The hit coordinates of the seen particles
  - The momenta at their birth
  - The identity and origin of the particles.
  - The entire final state (is it signal or background?).

From hits to coordinates

- Most detectors samples a charge cloud caused by a high-energy particle on finite size electrodes.
- To determine the coordinate of the particle, we need the geometrical convolution of the cloud and the electrode as a function of the particle coordinate.
- Consider a charge cloud depositing charge on only one electrode of size Δ. The coordinate estimate is the electrode center with a resolution of $\Delta/\sqrt{12}$.

From hits to coordinates

Consider a charge shared between two electrodes of size Δ. With a typical Gaussian charge cloud and measured pulseheights P_1 and P_2, this leads to a δx between the coordinate and the midpoint between the two electrodes:

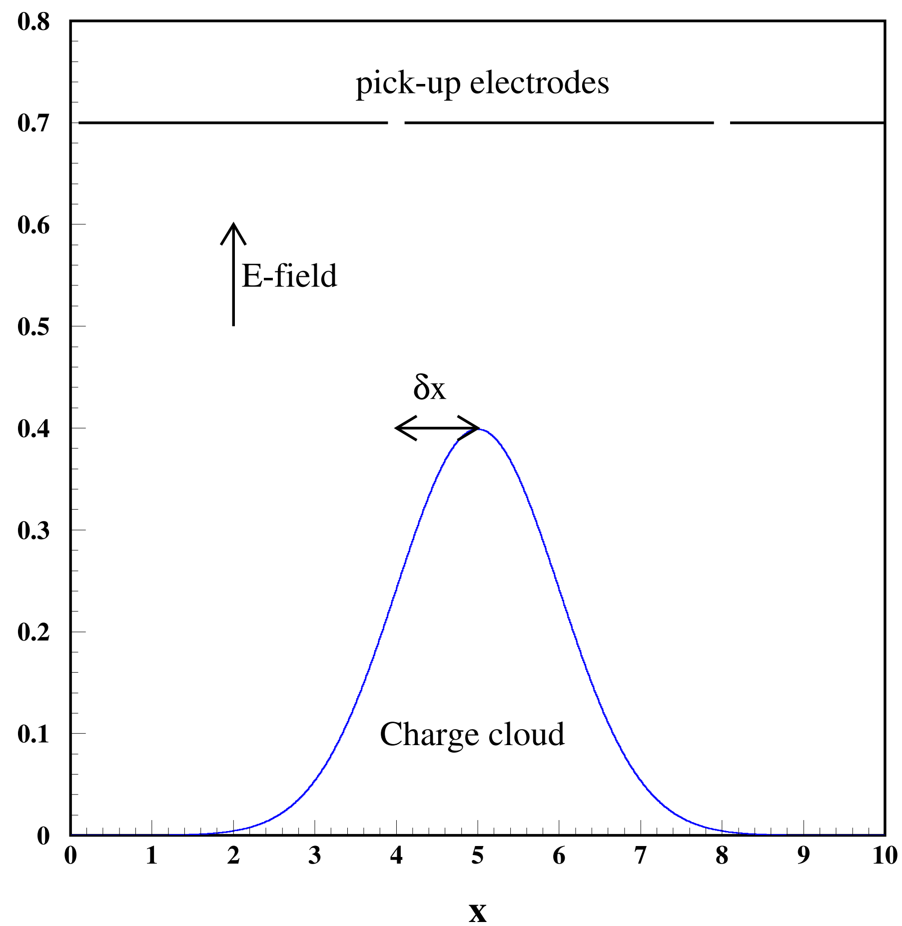
$$\frac{\delta x}{\Delta} = \sigma^2 \log \frac{P_2}{P_1}$$

where σ is the (known) width of the charge cloud. With an exponential cloud, the δx estimate would be slightly different.

In any case, the spatial resolution is much better when the reponse is shared between two electrodes. Therefore micro-strip detectors often have more capacitively coupled strips than it is able to read out.

From hits to coordinates

Pulse sharing between two pick-up electrodes.

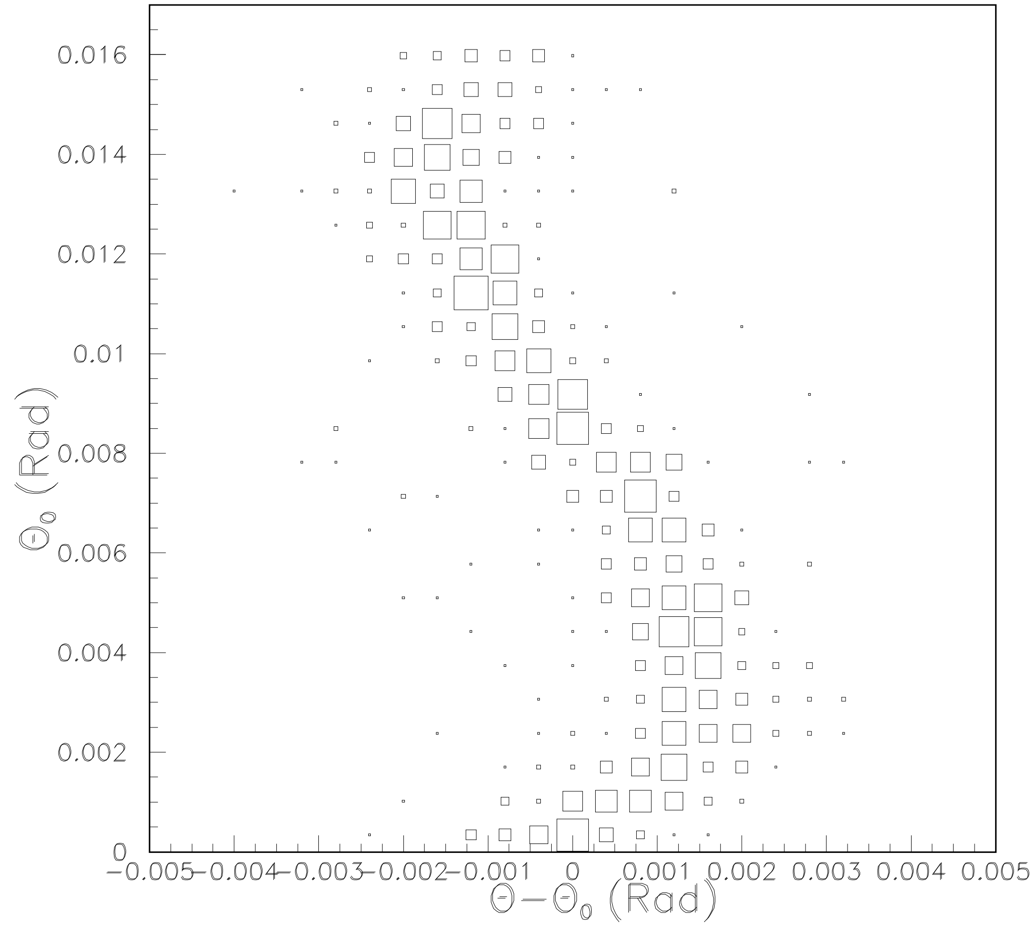


From hits to coordinates

Consider a charge shared among three or more electrodes. Here, a two parameter fit is possible, determining both the position and the width of the "charge cloud". The barycenter (a pulseheight-weighted average of electrode centers) is a widely used estimator of the coordinate – although not a perfect one.

Finite size effect

Barycenter versus true position for em-showers in a lead-gas calorimeter read out with 3×3 cm electrodes.



The response function

For reasons of noise and cost, the detector is often designed so that a two-electrode hit is the most probable. Therefore, the width of the "charge cloud" needs to be known for each hit to obtain an accurate response function, i.e. the coordinate as a function of the measured pulseheights. This function may, however, depend on local parameters:

- The relative alignment of $\bar{E}$ and $\bar{B}$ fields.
- The drift-distance of the drifting charge cloud and the the angle of the track to the plane of the electrodes.

Track coordinates in calorimeters

- The readout often subdivides a calorimeter into cells so that a high-energy particle will deposit its energy over a cluster of hit cells. The impact coordinate is estimated by the barycenter of energy-weighted cell-centers.
- Local corrections come from shower losses near cracks or edges, and from the finite cell size.
- The energy of the particle is taken as the sum of cell energies in the cluster, corrected for leakage out of the back and for ionization-losses in the preceeding material.

Pattern recognition and track parameters

- Consider a tracker with N planes normal to a local z axis. Each plane measures a x or a y coordinate, or both, of a track passing through.
  The task is now to determine which coordinates belong to which track and to fit the parameters of each track.
  For a cylindrical detector with an axial magnetic field, the tracks form ideally a helix with parameters $\bar{\alpha}$, e.g.:
  - $\pm 1/R$ (signed inverse radius of curvature)
  - d_0, ϕ_0, z_0 (point of closest approach to the beam-axis)
  - θ (polar angle wrt the beam-axis)

The Kalman filter

- The task of assigning the hits to the tracks is an art for which it is hard to give a general recipy. However, the fitting of the track parameters is an exact science. A popular algorithm performing both tasks at once is the Kalman filter.
- This algoritm can incorporate not only the random measurement errors in each plane but also the effect of multiple scattering and energy loss in the detector material.

The Kalman filter

- The Kalman filter consists of two separate algorithms: The filter and the smoother. The filter starts from the first hit closest to the interaction point, then predicts the hit in the next plane and so on, using a specific linear propagator $\mathbf{F}$:

$$\begin{aligned} \bar{\alpha}_k &= \mathbf{F}_{k-1}\bar{\alpha}_{k-1} \\ \mathbf{C}_k &= \mathbf{F}_{k-1}^T\mathbf{C}_{k-1}\mathbf{F}_{k-1} \end{aligned}$$

where $\mathbf{C}_{k-1}$ is the covariance matrix (here, we ignore multiple scattering). If a hit is found within acceptable distance from the prediction, it is used to update the track parameters and refine the propagator into the following plane. The algorithm is repeated until as many as possible of the measured coordinates are assigned to tracks.

The Kalman smoother

- The smoother algorithm then steps backwards, and thus use all the available information from the other hits to predict a particular hit. Again the track parameters and covariance matrix is updated at each step. For Gaussian measurement errors, it corresponds to a linear least squares fit – but without the need to invert large matrices.
  More info about the Kalman filter can be found at the site `egret0.stanford.edu/bbjones/explainkalman.ps`

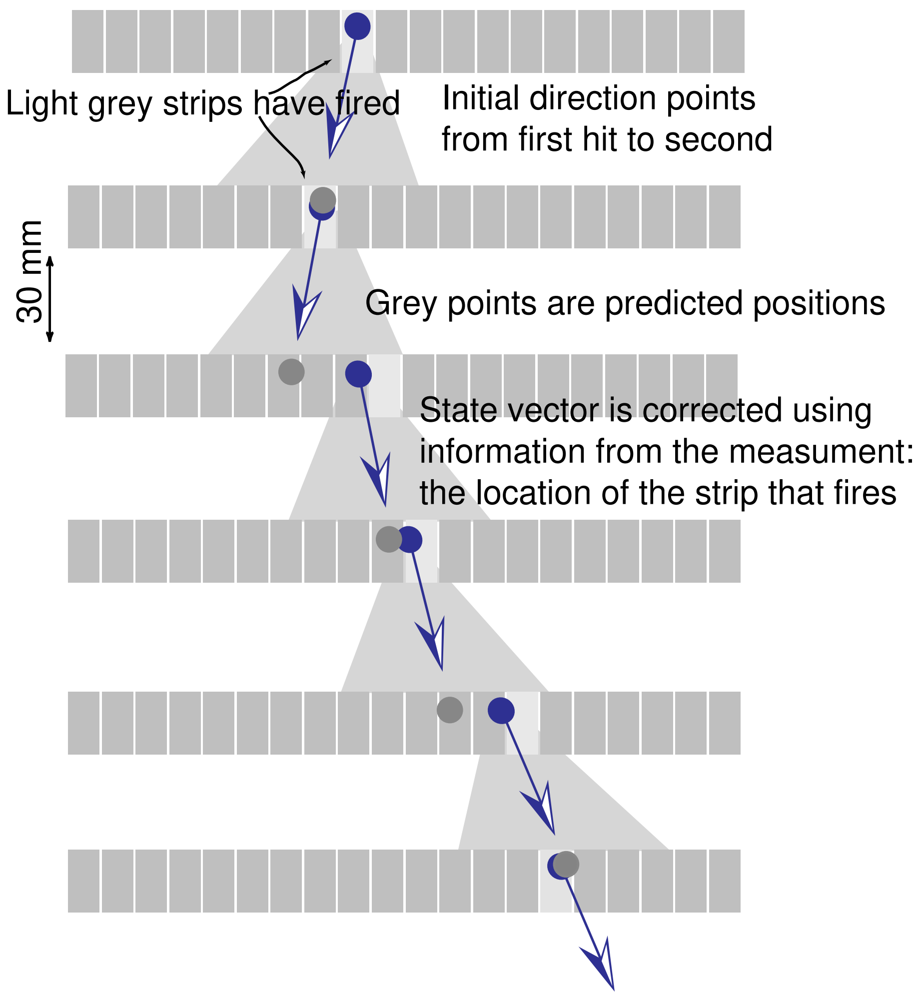


Fig. 1.— The Kalman filtering process as implemented for the *GLAST* science prototype beam test in October 1997.

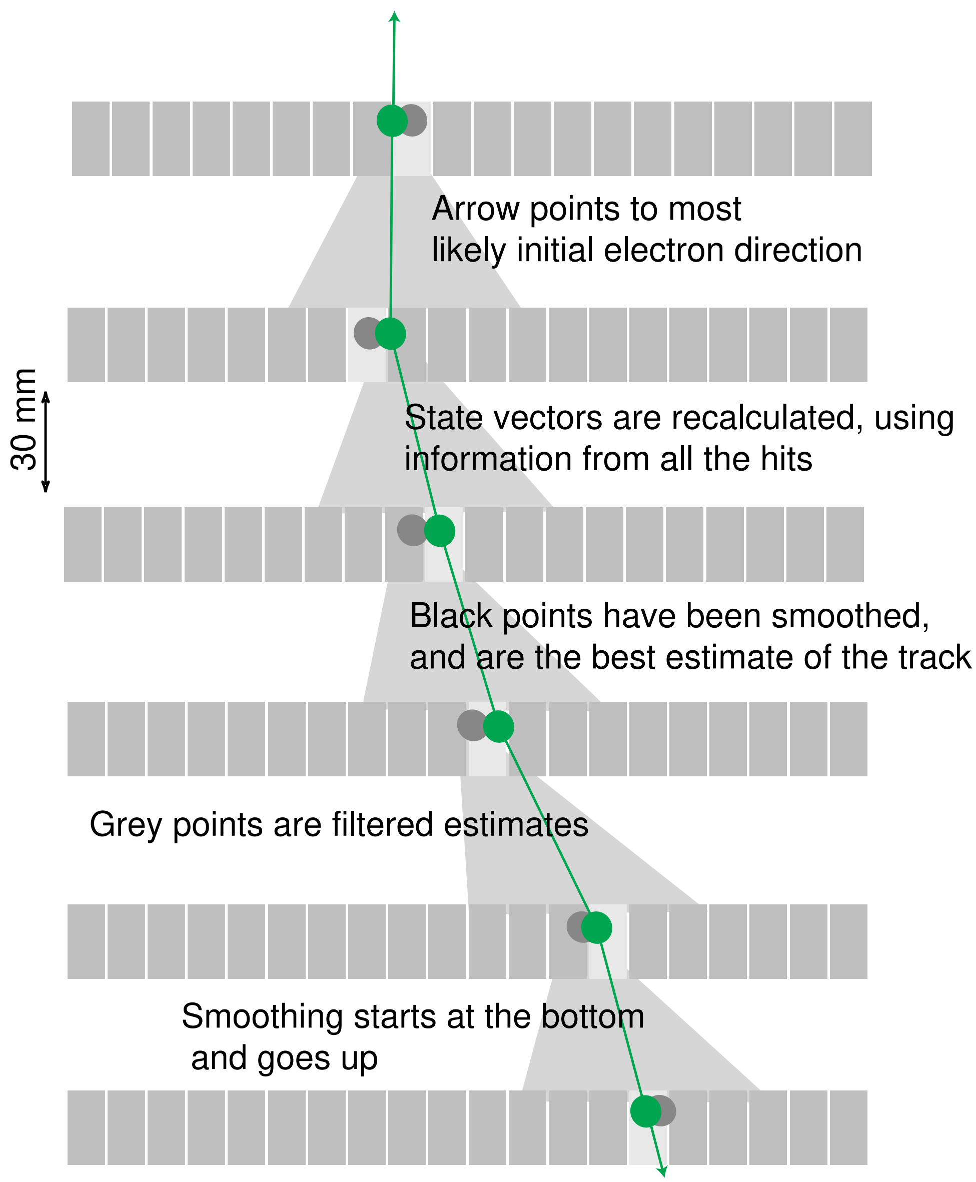


Fig. 2.— The Kalman smoothing process.

Calibration of drift-chambers

- For drift-chambers, the coordinate along the $\bar{E}$-field is (to first order) given by:

$$z = v(t - t_0)$$

  where v is the drift-velocity and t_0 the time offset. These can be calibrated e.g. using laser-pulses or a large set of tracks illuminating uniformly the drift-cell.

Calibration of calorimeters

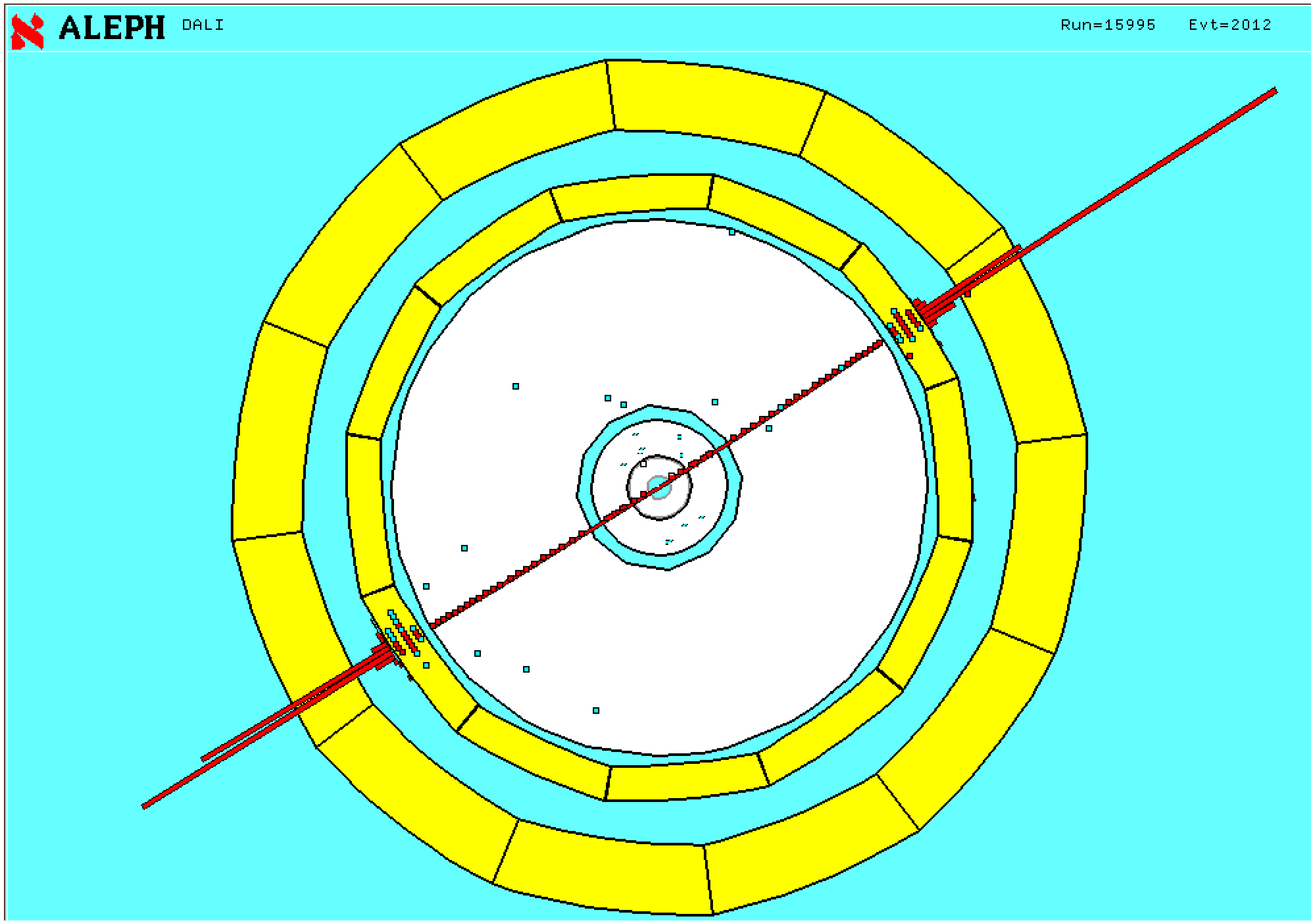
For calorimeters, a pulse-height P in cell number i translates to the energy:

$$E_i = \alpha_i(P - P_0)$$

where P_0 is the "pedestal" and α_i the calibration constant (when ignoring leakage and other nonlinearities). These constants are determined by successive steps:

- Test beam calibration of at least some modules before assembly.
- Pulsing of the electronics with known pulses.
- P_0 is measured and subtracted on-line using empty triggers.
- Radioactive or light sources may be used during data-taking to follow the calibration "constant" for each module closely in time.
- The energy of tracks from known processes are used for the final off-line calibration. This could typically be $Z \to e^+e^-$.

A $Z \to e^+e^-$ event



Alignment

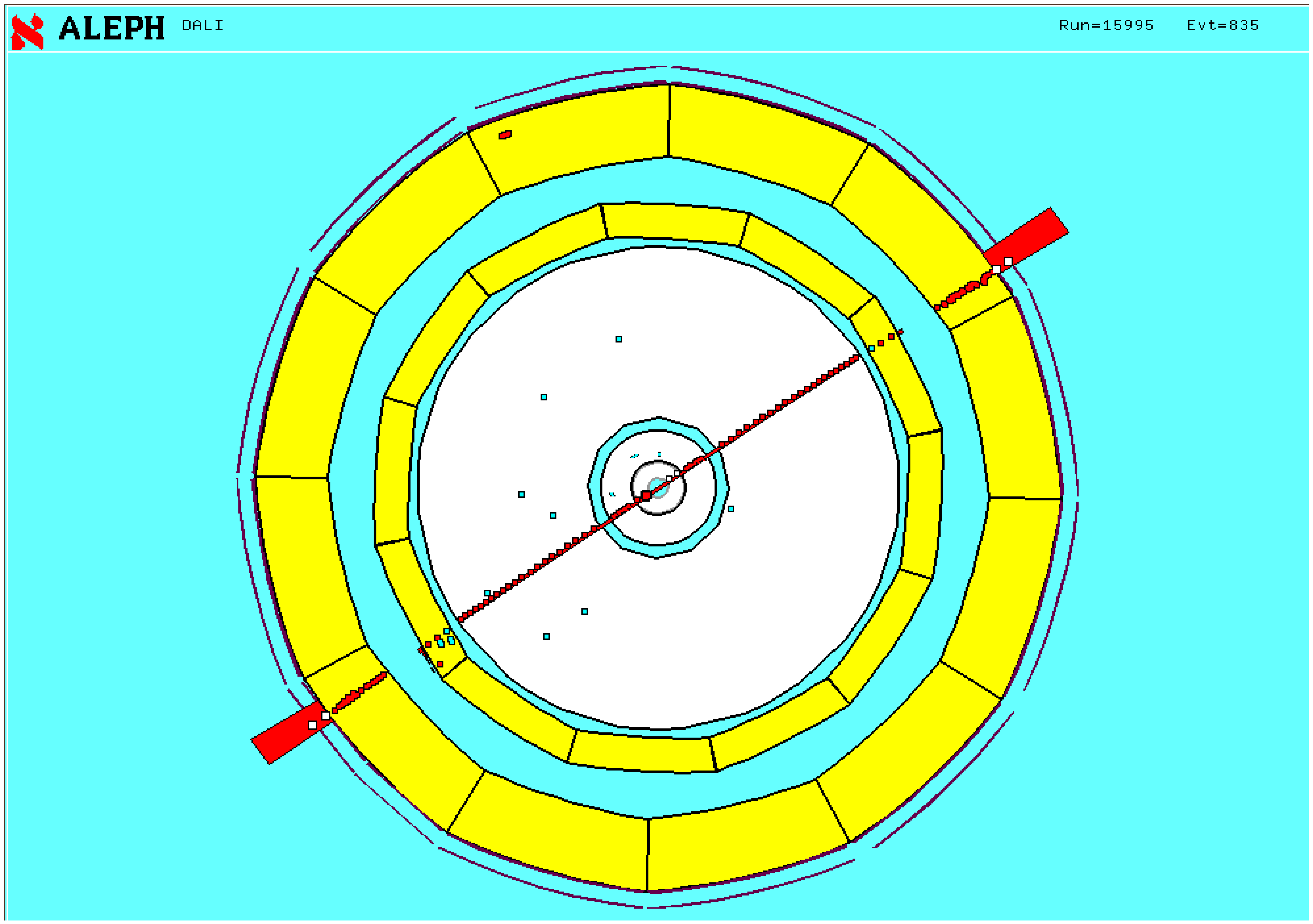
Alignment is important for tracking performance. The location in space of each elementary detector-element is established in many steps:

- Optical surveys of relative wire/strip positions during construction.
- Optical surveys of visible reference points after assembly.
- Before beam is turned on use cosmic ray alignment.
- Between bunches or fills use Laser alignment.
- Alignment using reconstructed tracks, such as muons from $Z \to \mu^+\mu^-$.

Alignment

The final adjustment uses a large sample of tracks, illuminating each elementary cell. The ideal algorithm for determining the alignment constants performs a simultaneous minimization of all differences between measured and predicted coordinates, taking constraints by vertex and kinematics of the alignment tracks into account, as well as the mechanical constraints of the detector. This requires, however, the inversion of a huge matrix (a way to do this is described in e.g. `http://www.desy.dk/ blobel/milleped.html`).

A $Z \to \mu^+\mu^-$ event



Particle id – electrons

- An "electromagnetic particle" deposits its energy in a limited region of the electromagnetic calorimeter. This goes for the longitudinal direction, where it is much less penetrating than a hadron, and for the transverse size which is confined inside a "Moliere radius" (see Leo).
- If a track is associated with the compact em-cluster, it is identified as an electron or positron. Often, the dE/dx, measured in the tracker, provides an additional discriminant against hadrons, due to the huge γ factor of electrons. At extremely high γ, some detectors exploit the transition radiation from thin foils with an index of refraction different from air to identify electrons.

Particle id – photons

- If there is no track, the compact calorimeter object is deemed a photon.

In ALEPH the hadron impurity among identified electrons in e.g. b-quark jets (from the weak b decay) is typically 10^{-3} with an efficiency for identifying a true electron of typically 0.7.

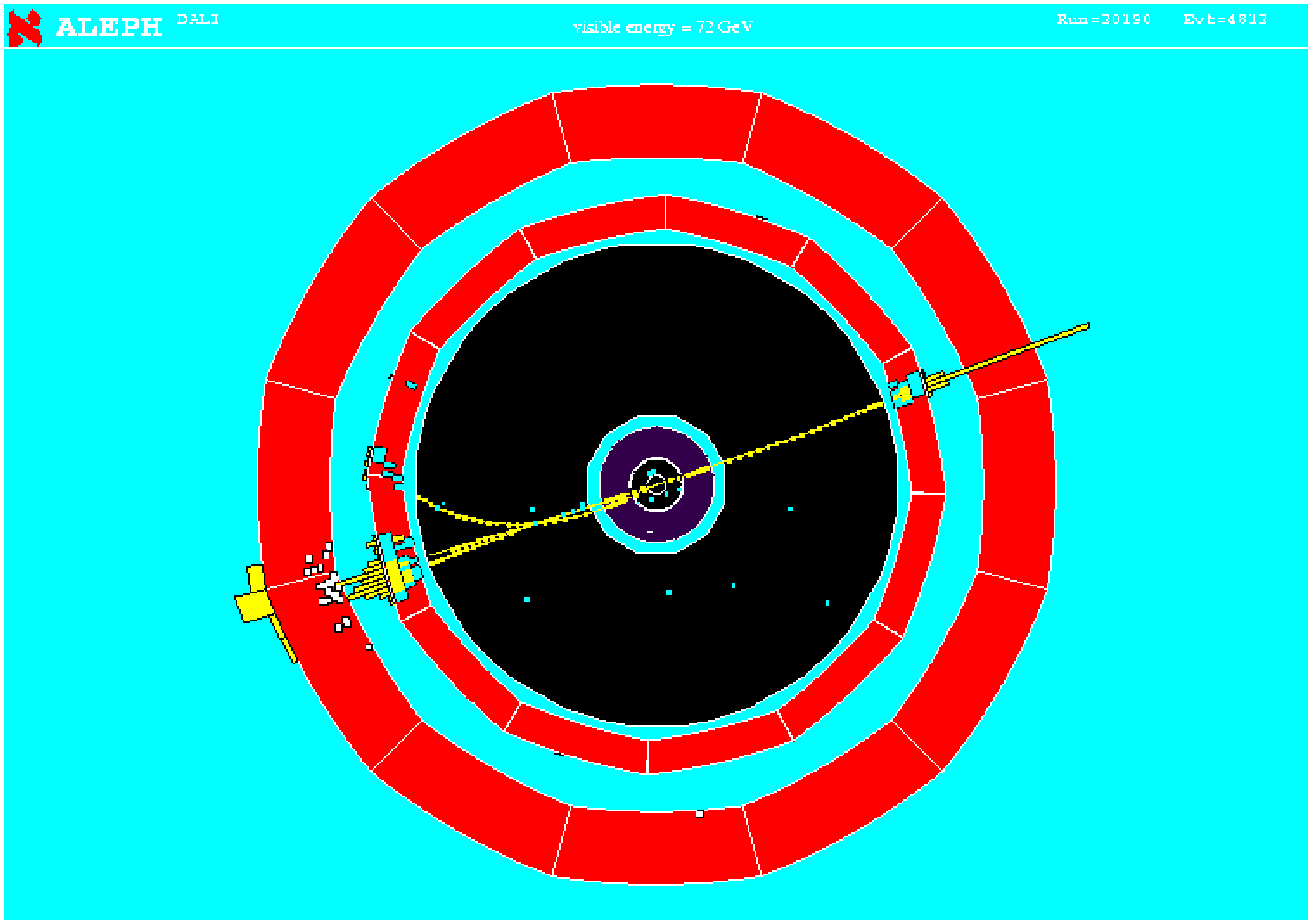
Particle id – muons

- A muon has a mass 200 times greater than an electron and suffers thus 40000 times less bremsstrahlung (see Leo). It has no strong interactions. In most cases, it sails through the calorimeters leaving only minimum ionising energy behind. A muon is identified by hits in muon chambers outside the calorimeter, and only minimum ionising energy around the track extrapolation in the calorimeter.

Particle id – taus

- A tau-lepton decays to either an electron, a muon or a small number of hadrons, in all cases accompagnied by one or two neutrinos. High-energy $\tau^{\pm}$ leptons are identified from
  - A narrow jet with a maximal mass of 1.7 GeV
  - An uneven small number of charged tracks, and a small number of neutral particles
  - Undetected neutrino's causing missing energy.

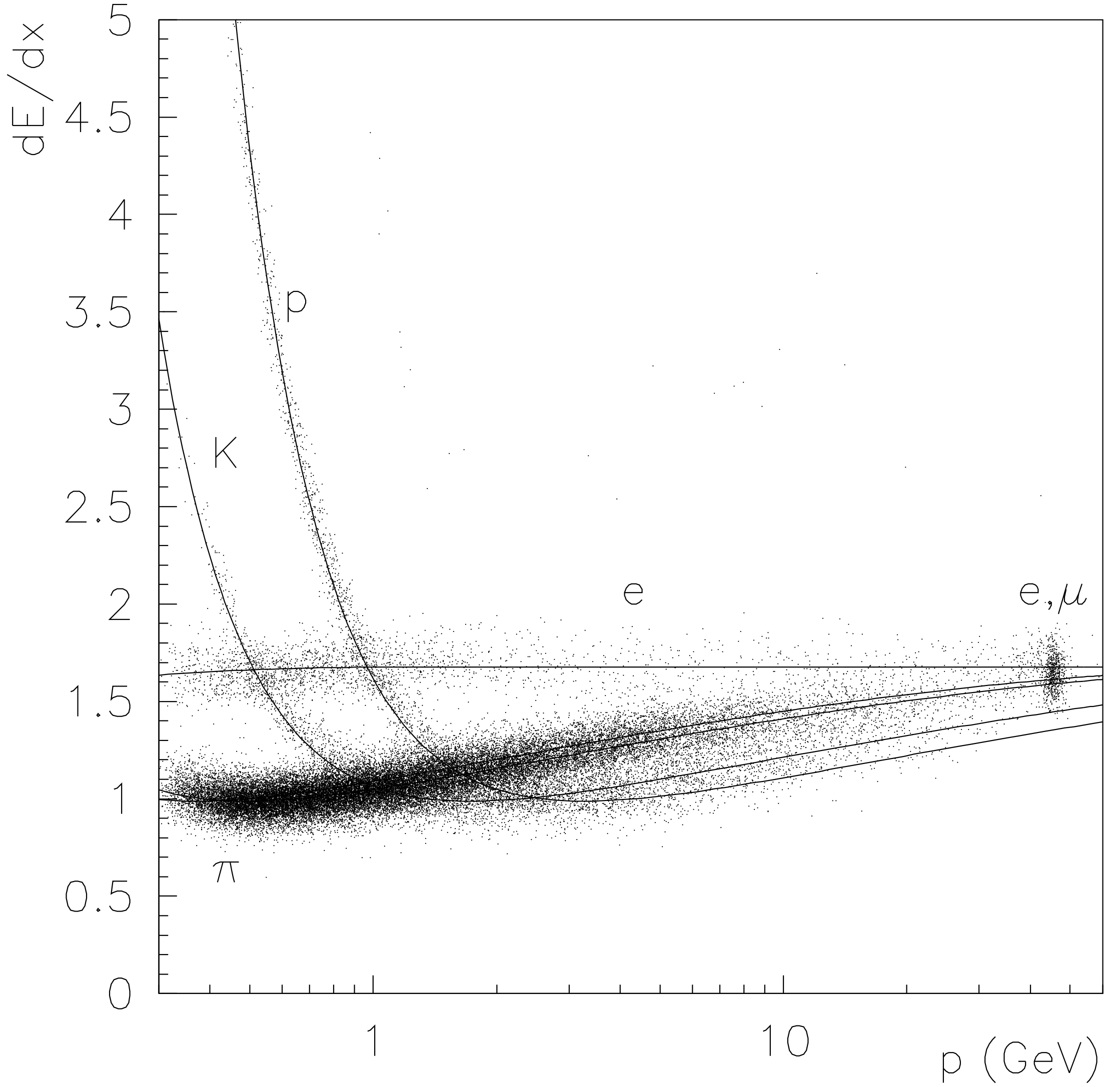
A $Z \to \tau^+\tau^-$ event



Particle id – by dE/dx, Cherenkov and TOF

- In a driftchamber, measurements of the pulseheight from a track on many sense wires provide an estimate of the mean dE/dx. Due to Landau tails, a truncated mean over the pulseheights on the wires gives a better estimator. A likelyhood fit to the Landau distribution would be even better, but is expensive in processing power. ALEPH got a 2σ separation between pions and kaons for $p > 2$ GeV (in the cross-over region around 1 GeV there is no information).
- For high energy particles Ring-Imaging-Cherenkov counters were used in e.g. DELPHI and HERA-B to determine the particles velocity and thereby its mass.
- At lower energies, the Time-Of-Flight measured at scintillator stations can provide information on the particle velocity.

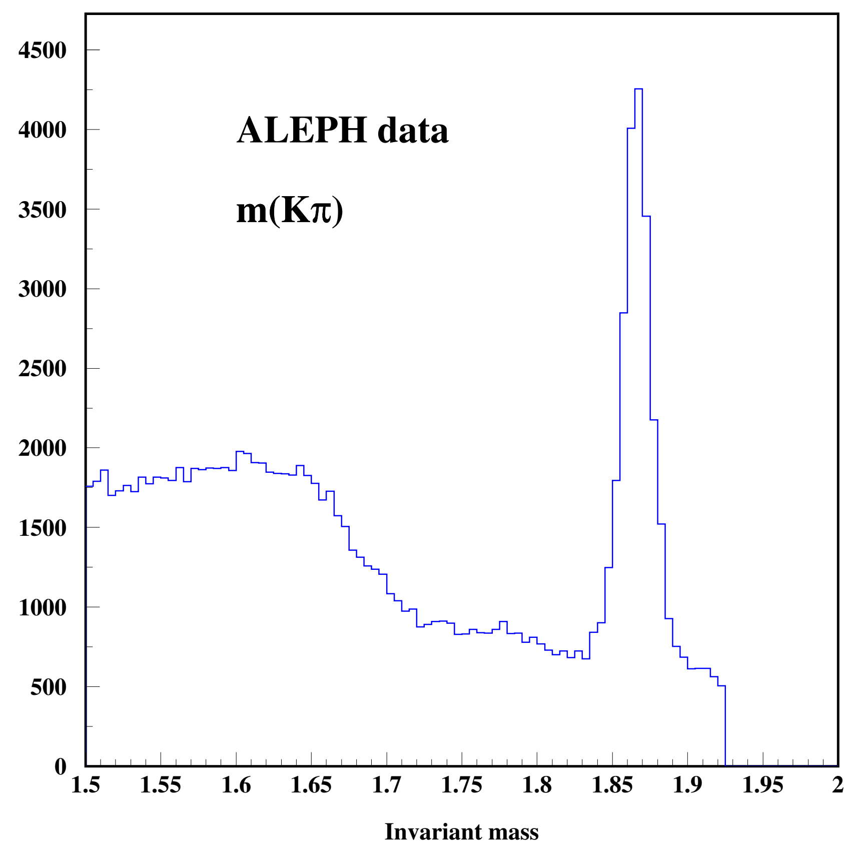
Using dE/dx for particle id



Particle id – by invariant mass

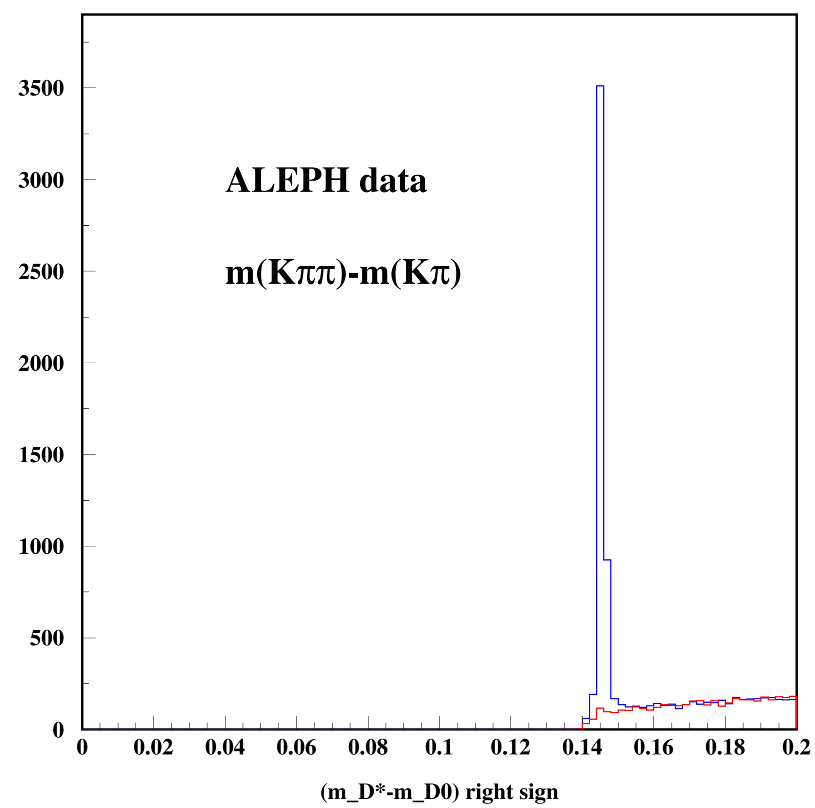
- Neutral strange particles, such as K^0 and Λ, are identified by their decay kinematics. They fly undetected a long way before decaying into a pair of charged particles, sometimes in the middle of the track detector.
- Some unstable particles decaying at the primary vertex can also be identified by the invariant mass of their decay products. For example, the tiny mass difference between the D and its exited state D^* ensures that the decay $D^* \to D + \pi$ can be identified almost without background.

Identifying particles by their decay products



Mass combinations of oppositely charged pion and kaon tracks (identified by dE/dx). The peak is due to $D^0 \to K\pi$.

Identifying particles by their decay products



Mass difference between $D^0 \to K^- \pi^+$ and $D^0 \pi^+$ combinations. In red, the result of using a π^- instead. The peak is due to $D^{*+} \to D^0 \pi^+$.

Particle id – by missing mass

- In exclusive reactions even neutrons can be identified, even if this particle would not be measured at all.
- If all other final state particles have their four-momentum accurately measured, the total p_{meas} is subtracted from the initial state p_{init}. The square of the remainder, the missing mass, should then be the neutron mass.

Energy flow

- The simplest way to determine the total energy in an event is by summing up the contents of all the calorimeter cells. In ALEPH the energy in an e^+e^- collision is determined this way with an accuracy of $\sigma(E) = 1.2\sqrt{E}$ (with E in GeV).
- Much better results can be obtained from an energy flow algorithm such as the following, resulting in a resolution of $\sigma(E) = 0.7\sqrt{E}$.

An energy flow algorithm

- Remove noise objects as well as possible.
- The energy of charged tracks (except special identified particles) are taken as $\sqrt{p^2 + m_\pi^2}$ and the associated calorimetric energy is removed.
- For electrons recover bremsstrahlung energy in a e.m. calorimeter cone around the track.
- The energy of photons in taken from the e.m. calorimeter.
- The remaining unassociated calorimeter energy clusters are attributed to neutral hadrons. The longitudinal readout compartments are here given new weights, optimizing the calorimeter response to hadrons.

Partons and jets

- Hadronic events involving large momentum transfers imply scattering of partons, i.e. quarks and gluons. The scattered partons emit gluon bremsstrahlung and form a parton shower. When all the created partons have separated about 1 fm, they fragment into hadrons, some of which are unstable and decay into lighter hadrons.
- In spite of all this complexity, the momenta of the final hadrons are still aligned with the original high-energy partons (a typical transverse momentum is $p_T \approx 0.35$ GeV). Therefore, the original partons may be reconstructed by a jet-algorithm using some metric, y, a measure of the momentum distance between two jets or particles.

A jet algorithm

- Start the first jet with a seed particle (e.g. the Energy Flow object with highest energy).
- Find the closest particle in y. If y is smaller than some cut-off y_{cut}, the particle is added to the jet. Then consider the nearest particle to the new jet for membership.
- Proceed adding particles until y exceeds y_{cut}, and then start a new jet. Repeat until all measured hadrons are assigned to jets.

Metrics and schemes

- A popular metric is the JADE metric:

$$y_{ij} = 2E_iE_j\sqrt{1 - \cos\theta_{ij}}$$

and the Durham metric:

$$y_{ij} = 2\max(E_i^2, E_j^2)\sqrt{1 - \cos\theta_{ij}}$$

- The “addition” of a new particle to a jet may proceed either by adding the two four-momenta (the E-scheme), or by adding the two three-momenta, while keeping the jets always mass-less (the P-scheme).

Metrics and schemes

- The best choice of metric and scheme depends on the application. So does the choice of y_{cut}.
- In some applications the event is always clustered into, say, four jets. In that case, the minimal metric distance y_{34} can be used to characterize the event.

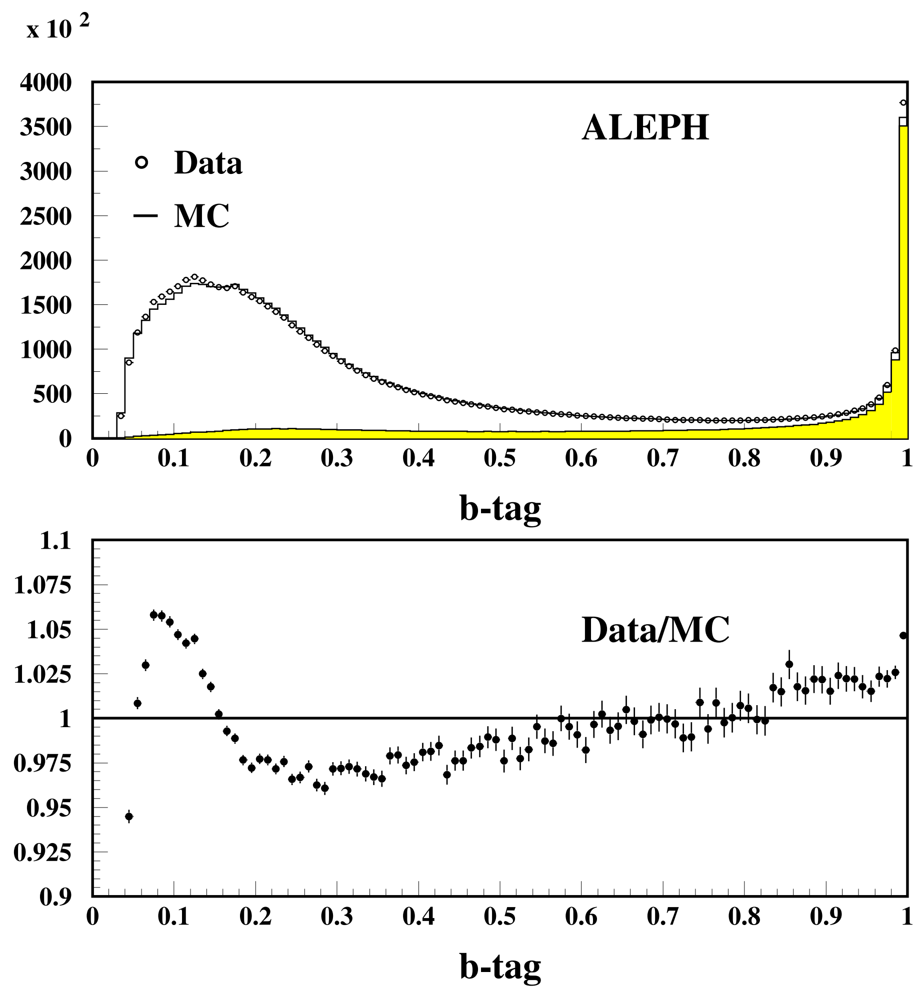
b-quark tagging

- The primary vertex is determined as the point in the interaction region where most measured tracks intersect within their errors.
- A b-quark ends typically up in a B-meson carrying a large fraction of the original b momentum. The lifetime of the B-meson averages 1 ps, and it decays into typically six charged particles (and a similar number of neutrals). Its decay-length is of the order of mm. This is used to tag b-jets.
- In ALEPH, this and five other variables for each jet were presented as inputs to a four-layer neural network. The output b-tag were trained on MC events to recognize b-jets.

Inputs for a neural-net b-tag

- The Sum of distances from the charged tracks closest approach to the jet-axis to the primary vertex (divided by the errors).
- The Lepton p_T w.r.t. the jet-axis. Leptonic b decays are relatively frequent (20%), and the p_T is high because of the large B mass (5 GeV).
- The Jet "fattiness" which is larger than is usual for lighter quarks.

The neural net b-tag in $Z \to q\bar{q}$ events



Kinematic fit

- If we have reconstructed an e^+e^- collision into a few jets, we can greatly enhance the energy resolution by performing a kinematic fit. Both the angles and the energies of the jets are then varied within their errors, until a best fit is obtained under the boundary conditions of total energy and momentum conservation.

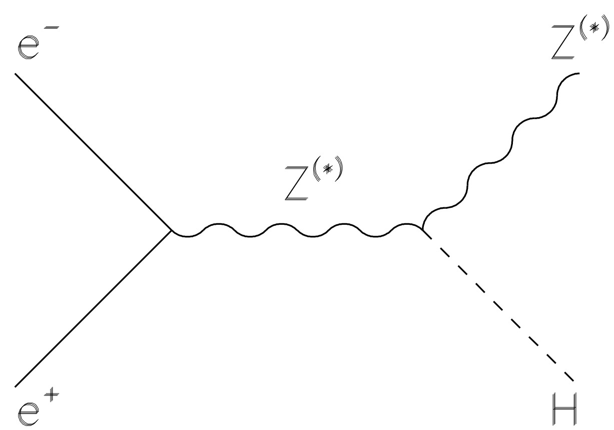
$$\sum p_{\text{jets}} = (2E_{\text{beam}}, 0, 0, 0)$$

If there is a lepton and a neutrino in the hypothesis, we have:

$$\begin{aligned} \bar{p}_{\text{jets}} + \bar{p}_l &= -\bar{p}_\nu \\ E_{\text{jets}} + E_l + p_\nu &= 2E_{\text{beam}} \end{aligned}$$

The Higgs search at LEPII

- at LEPII, the energy of the beams were gradually raised to a sustained maximum of 103.5 GeV each. The main aim with this was to find the Higgs boson. The most probable production mechanism was "Higgs-Strahlung": $e^+e^- \to Z^* \to Z + H$. Since the Z has a mass of 91.2 GeV, there might be a chance to find the Higgs if it is lighter than $207 - 91.2 = 115.8$ GeV, given enough luminosity.



The Higgs search at ALEPH

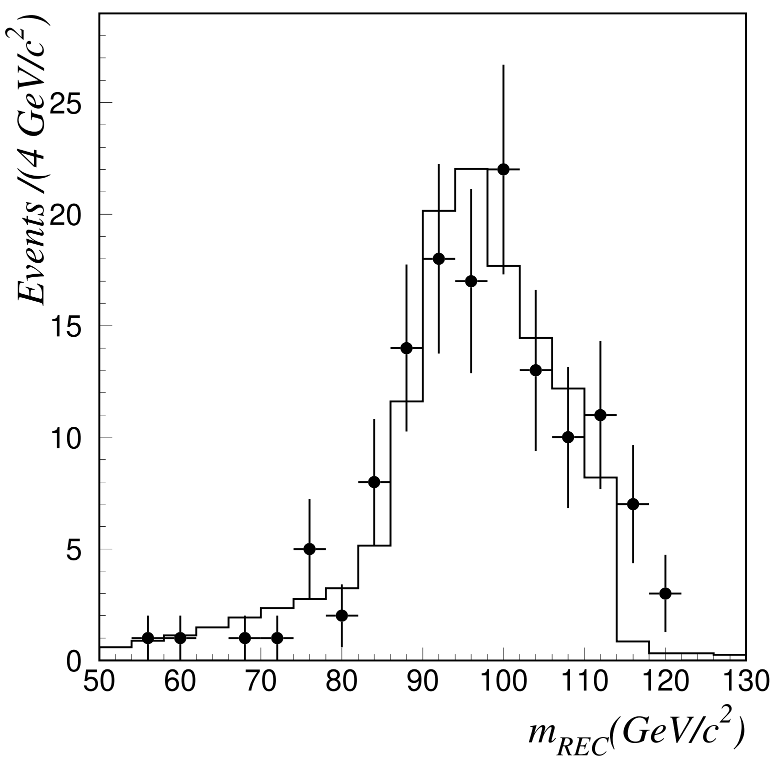
- All decay channels of the Z and the Higgs must be looked for to optimize the chance of discovery. To find the most probable decay channels, $Z \to q\bar{q}$ and $H \to b\bar{b}$, it is needed to identify b-jets and measure masses of di-jet systems.
- Two different analysis streams were used in ALEPH:
  - The cut-based stream using cuts on discriminating variables, such as the b-tags in four-jet events and their inter-jet angles and inter-jet masses, to select Higgs candidates.
  - The NN-based stream presenting 19 such variables as inputs to a neural net, trained by simulated Higgs events and background events to distinguish between the two. This is the standard analysis.

Background processes

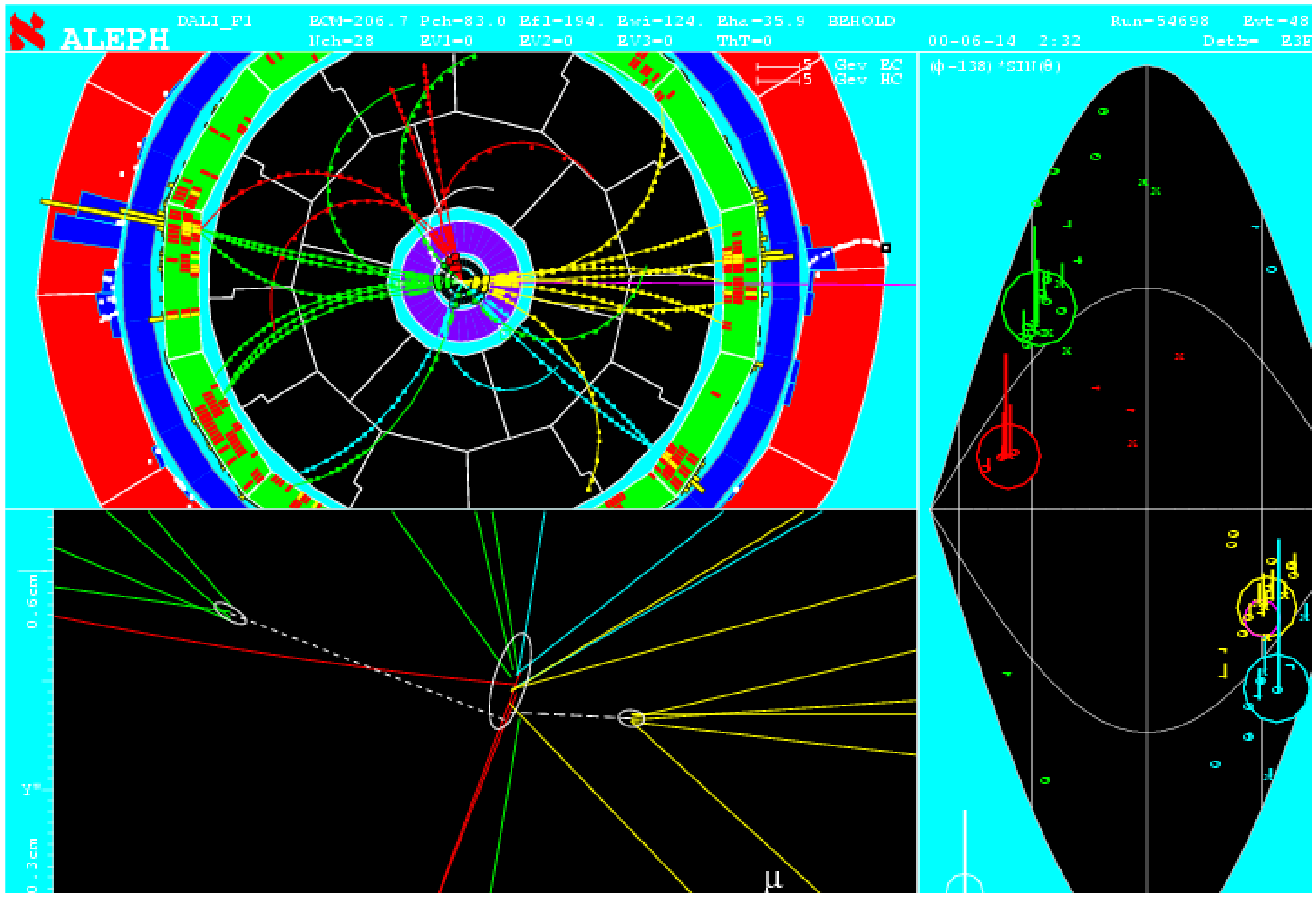
- The background processes are
  - $e^+e^- \to Z/\gamma \to W^+ + W^-$
  - $e^+e^- \to Z/\gamma \to Z + Z$
  - $e^+e^- \to Z/\gamma \to q + \bar{q} + gluons$

The observed signal

About 7 Higgs bosons was expected to be selected by the analysis, for a Higgs mass of 114 GeV$/c^2$. Even more excess events were actually observed, of which four were especially "Higgs like" (large NN output):



A Higgs candidate



The test statistic

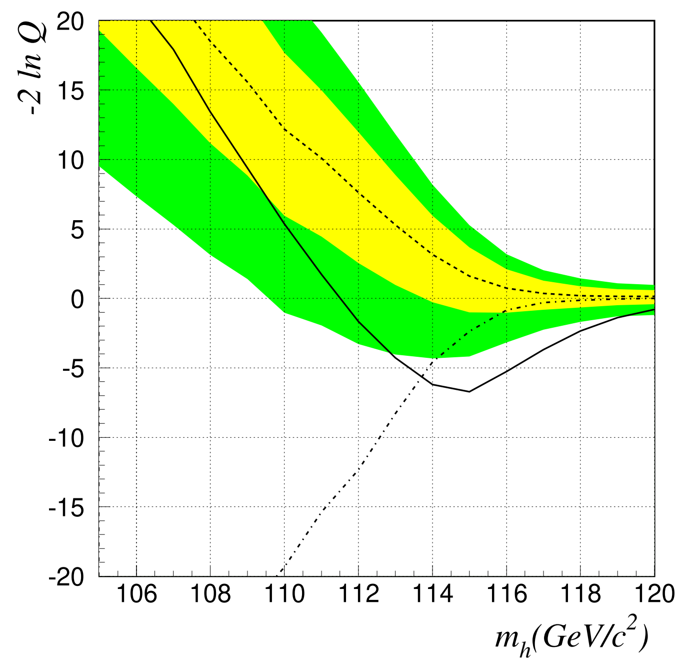
- From Monte-Carlo simulation, a joint p.d.f. is constructed for M_{rec} and NN, both for the background only hypothesis and for the background with a Higgs signal added.
- The likelihood ratio is in principle the best discriminator between the two:

$$Q = \frac{L_{s+b}}{L_b} = \frac{e^{-(s+b)}}{e^{-b}} \prod_{i=1}^{n_{obs}} \frac{s f_s(M_{rec}, NN) + b f_b(M_{rec}, NN)}{b f_b(M_{rec}, NN)}$$

Neglecting the f-terms, this is simply the ratio of Poisson probabilities for observing n_{obs} events. The f-terms give the likelihoods for the events to be observed with the particular configuration of M_{rec} and NN.

The signal

A signal would produce a lower than expected $-2\log Q$. The most likely Higgs mass is thus the one for which $-2\log Q$ of the observed data is minimal. It is shown below both for the observed data (black line) and for simulated background experiments. The bands are the contours containing 68% and 95% of the background experiments. The dashed-dotted line is the expectation for the signal plus background hypothesis.

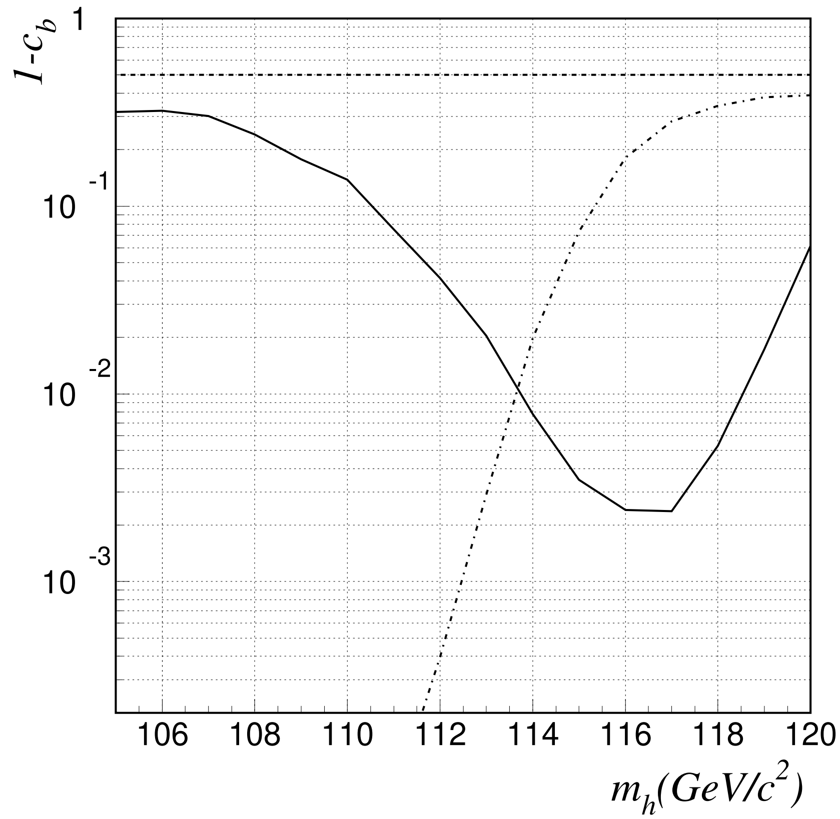


The Confidence Level

For each hypothesis, the probability of obtaining a smaller Q than observed is called the confidence level (CL). For the true hypothesis, the CL is uniformly distributed between 0 and 1 with a median value of 0.5 (as for any cumulative distribution). A signal would produce a drop in $1 - \mathrm{CL}_b$, where CL_b is the CL for the background only hypothesis.

The Confidence Level

The expected 1-CL is shown for for background only toy-experiments (dashed), signal plus background toy-experiments (dashed-dotted) hypothesis and the observed $1 - \mathrm{CL}_b$ (black).



Systematic errors

The systematic errors are already included in $1 - \mathrm{CL}_b$ by increasing the number of expected background events according to 1σ systematic uncertainties in the background simulation. The sources of these uncertainties, and their effect on the observed significance of the possible signal are:

Source	Effect on significance
MC Statistics	$\pm 0.07\sigma$
b-jet tagging	$\pm 0.08\sigma$
Gluon splitting probability	$\pm 0.04\sigma$
Jet momentum reolution	$\pm 0.05\sigma$
Selection variables	$\pm 0.06\sigma$
α_S	$\pm 0.06\sigma$
all	$\pm 0.15\sigma$

Summary of the Higgs search at ALEPH

The number of preselected events compared with expectations from background and from a Higgs boson of mass 115 GeV/c^2:

Channel	exp. Signal	exp. Background	Observed
$hq\bar{q}$	3.0	47.7	53
$h\nu\bar{\nu}$	1.0	37.7	39
hl^+l^-	0.4	30.8	30
$\tau^+\tau^-q\bar{q}$	0.3	13.7	15

Summary of the Higgs search at ALEPH

While this does not look earth-shattering, $1 - \mathrm{CL}_b$ nevertheless reaches a minimum of 2.4×10^{-3}, corresponding to an excess of 2.8σ, due to four especially “Higgs-like” events in the data. The data are, on the other hand, compatible with a $m_{\mathrm{Higgs}} = 115\,\mathrm{GeV}/c^2$ signal plus background hypothesis at the 1.1σ level.

Due to the excess, the lower limit on the Higgs mass obtained from the $1 - \mathrm{CL}_b$ curve is lower than expected:
$m_{\mathrm{Higgs}} > 111.5\,\mathrm{GeV}/c^2$ at the 95% confidence level

Summary of the Higgs search at LEP

However, the signal observed in ALEPH was not confirmed by observations in the other three LEP experiments, DELPHI, OPAL and L3. On the other hand, the three experiments could not exclude the signal either:

The combined LEP significance of a 115-116 GeV/c^2 signal is 1.7σ.

The combined 95% CL lower limit on the Higgs mass is 114.4 GeV/c^2.