Review of HEP Analysis Strategies

Markus Klute
Massachusetts Institute of Technology
May 22nd, 2012
HEP Data Analysis

• **Analysis**
  - statistical interpretation of an *ensemble of events* collected in a HEP experiment
  - typically, we have a model for *signal* and *background*
  - extract properties from *statistical analysis*

• **Examples**
  - OPAL, AMS, LHCb and CMS
  - search for the Higgs boson in CMS
HEP Event

- Event
  - readout cycle of the experiment
    - bunch crossing in accelerator structure
    - interaction of particle with detector
  - event consists of sub-detector measurements
  - particles are identified in a reconstruction step

pile-up

event reconstruction

event and event sample
Event Samples and Analysis Infrastructure

OPAL at LEP

LHCb at LHC

AMS at ISS

CMS at LHC
Example: OPAL at LEP

- OPAL (and the other LEP exp.) used analysis frameworks based on PAW
- Analysis software mainly in FORTRAN
- GEANT detector simulation
- HEP community transitioned to ROOT/C++ and other object-oriented programming at the end of the experiment lifetime

- OPAL in 1991
  - trigger rate 4Hz
  - event size 12-40 kB

- OPAL total data volume
  - reconstructable dataset ~2 TB
  - analysis dataset ~300 GB

- Storing and reconstructing OPAL data in 1991 was a challenge

- Parallel event processing on “computing farm”

- Trivial with today’s technology

- Event reconstruction took 30-60s at the time (on a 17SPECMark CPU)
Example: AMS at the ISS

- Alpha Magnetic Spectrometer Experiment on the International Space Station
- Computing **challenge** is the bandwidth limitation of 10Mb/s
- Detector output of ~7Gb/s is reduced using dedicated algorithms (zero suppression)
- Detector can buffer ~one week of data
- Primary data archive with 2 month buffer using laptop on ISS
- Data is transmitted using satellites to White Sands Ground terminal and via Huntsville to the AMS center at CERN
- AMS produces ~36TB of raw data per year
- Analysis framework is based on ROOT

special thanks to Paolo Zucccon for the content of this slide
Example: LHCb at the LHC

- LHC used luminosity leveling to control pile-up conditions
- Nominal interaction rate in ~20 MHz
- Reduced to 1MHz by L0 hardware trigger
- High level trigger farm (26k procs) reduced the rate to ~3.5kHz
- RAW event size ~50kB, resulting in ~1PB/y
- USER defined pre-selections reduce event sample to 10%
- Pre-selections are updated 2-3 times per year
- 150kB per reconstructed event. A micro format of 10kB per events used for large selections
- Software framework Gaudi also used by other experiments
- GEANT4 detector simulation
- Analysis framework based on ROOT

special thanks to Mike Williams for the content of this slide
Example: CMS at the LHC

- Reduce events rates using a hardware trigger (100kHz) and high level trigger farm (300Hz)
- Event samples are split in streams of ~50 Hz or less based of trigger signatures (e.g. DoubleMu)
- Event size for data
  - RAW (460kB), RECO (680kB), AOD (230kB)
- Total volume of data and MC stored on tape ~30PB
- 2011 AOD samples ~700TB. Multiple copies needed to support analysis
- USER access data (AOD) via grid submission
- GEANT4 detector simulation
- Analysis framework based on ROOT
Event Samples and Analysis Infrastructure

- Prerequisite for an effective analysis is the availability of datasets of manageable size
- Data reduction by data format
  - RAW ➜ RECO format ➜ Analysis format ➜ USER defined ➜ more USER defined
- Data reduction by event selection
  - reduce volume for analysis use
  - selections are applied at all data format levels
- Both methods have pros and cons
  - typically trade CPU and flexibility with storage
- Working point for an experiment (collaboration) needs optimization
  - Guideline (stating the obvious)
    - samples needed many times have to be small and always available
    - trade between event count and size / information
    - samples needed rarely can be large (RECO samples)
    - calculate derived quantities to minimize storage requirements, e.g. jets, electrons, missing ET
Event Samples and Analysis Infrastructure

- Data reduction by event selection
- Classify events in stream/skims which use a well defined event selection

- Subset of events contains all events need to perform an analysis
- Volume of events sample of “type D” can be much reduced wrt the total volume

Type D e.g. /DoubleMu
Event Samples and Analysis Infrastructure

- Experiments adopted computing / analysis strategy according to given challenges
- Main concepts a very similar across experiments with strong links between computing and analysis models
- Large scale experiments profit from improvements in computing technologies (grid)
- Standardized and collaborative tools
  - modeling of detectors with GEANT
  - physics using Monte Carlo generators with standardized interfaces
  - ROOT is main analysis tool and includes classes for statistical analysis
Example Analysis: H → WW in CMS

- Study the origin of electroweak symmetry breaking
  - how do W and Z bosons acquire mass?
  - can we explain fermion masses?
  - Higgs mechanism give answers
  - new particle is proposed in the SM: the Higgs Boson
    - mass is free parameter
    - all other properties are predicted in the SM
- Search for the Higgs Boson at the LHC
Example Analysis: $H \rightarrow WW$ in CMS

- Clean experimental signature
  - 2 leptons + 2 neutrinos
  - No mass peak
  - analysis is all about the background estimation
  - backgrounds have large cross sections

- Higgs: $16\text{pb}$
- $W \rightarrow lv$: $10438\text{pb}$
- $Z \rightarrow ll$: $1666\text{pb}$
- TTbar: $157\text{pb}$
- WW: $43\text{pb}$
- Higgs: $16\text{pb}$
Example Analysis: H → WW in CMS

- Analysis strategy
  - WW pre-selection
    - establish WW signature
    - data driven estimates of main backgrounds
  - Higgs selection
    - discriminate Higgs against WW background
    - cut-based selection or multivariate analysis discriminator

![Diagram showing analysis strategy and control regions](image-url)
Example Analysis: $H \rightarrow WW$ in CMS

- **Data streams and reduction in analysis flow**
  - /DoubleMu 63M events
  - /DoubleElectron 21M events
  - /MuEG 25M events
  - ~3.5M after pre-selection
  - ~100 events after WW selection

- **Monte Carlo needs**
  - 20M signal events
    - produced for 25 mass points from $110 \text{ GeV} < m_H < 600 \text{ GeV}$
    - POWHEG and PYTHIA
  - 110M background events
    - MADGRAPH, PYTHIA and POWHEG
  - MC generators are interfaced in software framework or used via standard LHE format

- **Pile-up (in-time and out-of-time)**
  - large effects on measurements of missing transverse energy, isolation of leptons, jet measurements, etc.
  - distribution unknown a priori
  - re-weighting of MC is required to match collision data
    - MC samples have been (digitized and) reconstructed using two different PU scenarios

- **Beam spot**
- **Center-of-mass energy**
Example Analysis: H → WW in CMS

- Boosted decision tree explores event kinematics
  - Multivariate classifier are also used in lepton selection
- Classification by # of jets and lepton flavor explore regions with varying S/B

Higgs is scalar boson
Statistical Interpretation

- **General framework**
  - signal strength modifier
  - nuisance parameter
  - likelihood
  - construct test statistic

- **Quantify an excess**
  \[ q_0 = -2 \ln \left( \frac{\mathcal{L}(\text{data} \,|\, b(\hat{\theta}_0))}{\mathcal{L}(\text{data} \,|\, \hat{\mu} \cdot s(\hat{\theta}) + b(\hat{\theta}))} \right), \quad \hat{\mu} \geq 0, \]

- **Quantify the absence of a signal**
  \[ q_{\mu} = -2 \ln \left( \frac{\mathcal{L}(\text{data} \,|\, \mu \cdot s(\hat{\theta}) + b(\hat{\theta}_\mu))}{\mathcal{L}(\text{data} \,|\, \hat{\mu} \cdot s(\hat{\theta}) + b(\hat{\theta}))} \right), \quad 0 \leq \hat{\mu} < \mu, \]

- modified frequentist CLs (similar to method used at LEP)
  \[ \text{CL}_{s+b} = P \left( q_{\mu} \geq q_{\mu}^{\text{obs}} \mid \mu \cdot s + b \right), \]
  \[ \text{CL}_b = P \left( q_{\mu} \geq q_{\mu}^{\text{obs}} \mid b \right), \]
  \[ \text{CL}_s = \frac{\text{CL}_{s+b}}{\text{CL}_b} \]
Statistical Interpretation

- If $\text{CLs} \leq \alpha$ for $\mu = 1$, conclude that signal is excluded at $(1-\alpha)$ CL
  - conservative approach given that proper CLs can not be determined w/o knowing the signal beforehand
- $\text{CL}_{s+b}$ and $\text{CL}_b$ are determined from independent toys
  - $O(1000)$ needed to have a statistical precision of $O(1\%)$
  - toys are numerical integrations of the likelihood functions
- Limit calculation procedure
  - data considered binned or unbinned
  - models can be number (cut & count), shapes or parametrized models that make predictions on presence of data
  - determine $q_{\mu}$ for an ensemble of $\mu$ values from a maximum likelihood fit of the model of data and calculate $\text{CL}_{s+b}$ and $\text{CL}_b$
  - systematic uncertainties incorporated as nuisance parameters to the fitted likelihood functions $L$ and integrated out
  - 95% CL upper limit
    - determine the value of $\mu$ where $\text{CLs} \leq 5\%$
Example: CMS Higgs Combination

- Limit calculation in CMS Higgs search
  - tools based on ROOTs statistic classes, alternatives available and used to cross check results
  - 43 channels (and growing)
  - 156-222 nuisance parameters (depending on tested mass hypothesis)
  - 183 mass hypotheses (from $110 \text{ GeV} \leq m_H \leq 600 \text{ GeV}$)

<table>
<thead>
<tr>
<th>Channel</th>
<th>$m_H$ range (GeV)</th>
<th>Luminosity (fb$^{-1}$)</th>
<th>Sub-channels</th>
<th>$m_H$ resolution</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H \rightarrow \gamma\gamma$</td>
<td>110–150</td>
<td>4.8</td>
<td>5</td>
<td>1–3%</td>
<td>[60]</td>
</tr>
<tr>
<td>$H \rightarrow \tau\tau$</td>
<td>110–145</td>
<td>4.6</td>
<td>9</td>
<td>20%</td>
<td>[61]</td>
</tr>
<tr>
<td>$H \rightarrow bb$</td>
<td>110–135</td>
<td>4.7</td>
<td>5</td>
<td>10%</td>
<td>[62]</td>
</tr>
<tr>
<td>$H \rightarrow WW^* \rightarrow 2\ell 2\nu$</td>
<td>110–600</td>
<td>4.6</td>
<td>5</td>
<td>20%</td>
<td>[63]</td>
</tr>
<tr>
<td>$H \rightarrow ZZ^{(*)} \rightarrow 4\ell$</td>
<td>110–600</td>
<td>4.7</td>
<td>3</td>
<td>1–2%</td>
<td>[64]</td>
</tr>
<tr>
<td>$H \rightarrow ZZ \rightarrow 2\ell 2\nu$</td>
<td>250–600</td>
<td>4.6</td>
<td>2</td>
<td>7%</td>
<td>[65]</td>
</tr>
<tr>
<td>$H \rightarrow ZZ^{(*)} \rightarrow 2\ell 2q$</td>
<td>130–164</td>
<td>4.6</td>
<td>6</td>
<td>3%</td>
<td>[66]</td>
</tr>
<tr>
<td></td>
<td>200–600</td>
<td>4.6</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H \rightarrow ZZ \rightarrow 2\ell 2\tau$</td>
<td>190–600</td>
<td>4.7</td>
<td>8</td>
<td>10–15%</td>
<td>[67]</td>
</tr>
</tbody>
</table>
Example: CMS Higgs Combination

- CPU requirements
  - toy evaluation for CL_{s+b} and CL_b typically ~1min
  - \( \mu \) (95% CL) typically determined from O(1000) toys per mass hypothesis and O(10) discrete values of \( \mu \)
  - results in ~3 CPU years
  - calculation can be parallelized. Submission of O(10000) grid jobs, i.e. results can be obtained within a few days
  - alternative statistical method (asymptotic calculation) can be used to speed up the process

![Graphs showing CMS Higgs boson mass limits and local p-values.](image)
What works, what can be improved

• CMS (and other experiments) are able to perform complex analyses
  • task becomes more complex as we search for rare processes

• Access to large data and MC samples is challenging
  • changing conditions require new productions and reprocessing
  • latencies in production and distribution
  • tails in availability limit the final result
  • data driven techniques limit the dependency on MC

• Computing resources
  • binding of CPU with data location has large overhead and causes inefficiency

• Advanced software frameworks based on common tools
  • GEANT for simulation
  • standard MC generator or interfaces
  • ROOT is main analysis framework
    • includes classes for statistical analysis
Summary

- General characteristics of an analysis in high energy physics
- Event samples in various high energy physics experiments
- Example analysis: $H \rightarrow WW$ in CMS
- Statistical interpretation: Combination of Higgs searches in CMS
- What works, what can be improved

- LHC Higgs results will be updated for ICHEP this summer