Thomas Reitenspiess (ETHZ) on behalf of the CMS collaboration

CMS detector simulation tuning through machine learning
Detector Simulation Tuning

Motivation

- The $H \rightarrow \gamma\gamma$ analysis takes fully differential input variables to e.g. measure differential cross-section

Differential detector simulation correction using $Z \rightarrow e^+e^-$ tag and probe sample

- Using electrons from $Z \rightarrow e^+e^-$ to measure and fine-tune the ECAL response to isolated electromagnetic objects
  - Electrons leave very similar trace as photons in ECAL and can be reconstructed as such

- Measuring $H \rightarrow \gamma\gamma$, this requires a good understanding of the detector response to photons
  - Photons produce showers in CMS ECAL
    - Shower shape variables
  - Account for additional object in the detector
    - Isolation variables

The $H \rightarrow \gamma\gamma$ analysis takes fully differential input variables to e.g. measure differential cross-section
Detector Simulation Tuning
General Strategy

- Using the decay $Z \rightarrow e^+ e^-$ as a clean standard candle, since its signature in the detector is very well understood
- Using detector related variables
- Therefore: remaining differences come from not simulating detector perfectly
- This strategy allows effective tuning of simulation
Detector Simulation Tuning
Tag and Probe

Tag and Probe
- One electron is identified very tightly and used as tag
- Other electron is used as unbiased probe to test detector response and derive corrections
- Electrons leave very similar trace in ECAL as photons, can therefore be used to derive correction for photons
- Very large sample – $O(10^7)$ – gives the ability to fine-tune detector response

Tag
- Pass tight electron ID
- Match trigger candidate

Probe
- Unbiased
- Reconstructed as photon
Detector Simulation Tuning

Methods

Quantile Morphing

- Shift simulation to match data according to cumulative distribution functions
- Cdf can be obtained differentially in kinematics and event-energy-density using quantile regression
- Method works only on continuous distributions
- Generally applicable effective correction method

Stochastic Correction

- Works for discontinuous variables
- Developed for isolation variables
- Stochastic element needed to account for discontinuity caused by detector properties
- Effective correction method for effects caused by mismatch of number of objects in isolation cone
The higher the event-energy-density $\rho$ the more objects will populate the isolation cone.

The low pile-up behaviour is simulated well.

Therefore the simulated isolation distribution for low pile-up is taken as starting point.

The low pile-up isolation distribution is resampled $\mu$ times.

Between distributions for integer numbers of $\mu$ an interpolation is performed to get distribution continuous in $\mu$.

Likelihood fit to data and simulation of resampled distribution is performed with $\mu$ as free parameter, binned in $\eta$ and $\rho$.
Detector Simulation Tuning
Stochastic Correction

- In every $\eta$-bin, get $\mu$ for data and simulation for different values of $\rho$
- Perform linear interpolation between them
- Add transverse momentum to the cone in simulation according to the difference in $\mu$, depending on $\eta$ and $\rho$ for every event

Result shows good agreement between data and corrected simulation
- First bin ($\text{Iso}_\gamma = 0$) is corrected very well
- Method catches also non-trivial detector effects
Detector Simulation Tuning
Quantile Morphing

- Requirements:
  - Continuous distribution
  - Availability of pure control sample

1. Integrate
2. Match quantiles
3. Transform
Detector Simulation Tuning
Quantile Regression

Goal: Differential Corrections

Measuring differential fiducial cross-sections, therefore it is crucial to have a well corrected simulation in every region of the phase space

1. Train $n_q$ BDTs per variable to predict conditional shape of cdf depending on $p_t$, $\eta$, $\phi$, $\rho$ for data and simulation
2. Find two $q_t$ around $y$ w.r.t $x_i = [p_t, \eta, \phi, \rho]$ for data and simulation
3. Use linear interpolation between the two points $(\tau_i, \tau_i)$ to find $cdf(y|X)$ for data and simulation
4. Correct simulation by matching $cdf_{data}$ and $cdf_{mc}$
Detector Simulation Tuning

Boosted Decision Trees

Training Data Set \((x_i, y_i), i = 1, \ldots, n\)

Decision Tree splits dataset into regions in \(x = [p_t, \eta, \phi, \rho]\) in a binary fashion, to minimize Loss function

\[
\begin{align*}
\text{if } x > 0.5 & : x > 0.5 \\
\text{if } x > 0.25 & : x > 0.7 \\
\text{if } x > 0.1 & : 0.6 \\
\text{if } x > 0.9 & : 1.4
\end{align*}
\]

\[
L(y_i, q_t(x_i)) = \sum_i \rho_\tau(y_i - q_t(x_i))
\]

Quantile Loss function

\[
\rho_\tau = \begin{cases}
(\tau - 1)u, u < 0 \\
\tau u, u > 0
\end{cases}
\]

In practice

- Using gradient boosting
- Minimizing quantile loss function
- Catch dependence of corrected variables to \(x = [p_t, \eta, \phi, \rho]\)
- Input variables not limited to these, could add e.g. time
Detector Simulation Tuning
Quantile Regression

Some Results

$\sigma_\phi$

$\sigma_\eta$
Detector Simulation Tuning
Quantile Regression

Some Results

$R_9$

$S_4$

![Graphs showing the distribution of data and simulated events for $R_9$ and $S_4$.](image-url)
Detector Simulation Tuning
Quantile Regression
Detector Simulation Tuning
Quantile Regression

$R_9$
Conclusion

Introduced two methods to tune detector simulation

Differentially tune simulation using machine learning techniques
- Based on quantile regression applying boosted decision trees
- Ability to fine-tune simulation dependent on kinematics and event-energy-density
- Method is scalable to more input variables
- Method can be easily applied elsewhere

Tune discontinuous isolation variables
- Applying stochastic techniques to account for discontinuity caused by detector properties
- Differential in pseudorapidity and event-energy-density
- Method can track non-trivial detector effects very well
Backup
CMS ECAL

- CMS ECAL is a homogeneous calorimeter
- Barrel consists of 36 supermodules with 1700 crystals each, 61200 in total
- Edcaps consist of 4 “Dee’s” with 3662 scintillating PbWO4 crystals each, 14648 in total
- Mounted inside the 3.8T Magnet
- The crystals are alligned quasi-projectively
ECAL SuperCluster

- For unconverted photons, a supercluster results to be formed of the 5x5 crystals centered around the crystal with the highest transverse energy deposit.
- More complicated for converted photons, including more 5x5 blocks in $\phi$ direction.
- In endcaps, 5x5 blocks can overlap.
Six Variables defined in the SuperCluster are used in photon identification:

- \( R_9 = \frac{\sum_{3 \times 3} E_i}{\sum_{5 \times 5} E_i} \)
- \( S_4 = \frac{\sum_{2 \times 2} E_i}{\sum_{5 \times 5} E_i} \)

- \( \sigma_\eta = \sqrt{\sum_{SC} (\bar{\eta} - \eta_i)^2 E_i} \)
- \( \sigma_\phi = \sqrt{\sum_{SC} (\bar{\phi} - \phi_i)^2 E_i} \)

- \( \sigma_{i \eta \eta} = \sqrt{\sum_{5 \times 5} (\bar{\eta} - \eta_i)^2 w_i} \)
- \( \sigma_{i \eta \phi} = \sqrt{\sum_{5 \times 5} (\bar{\eta} - \eta_i)(\bar{\phi} - \phi_i) E_i} \)
Isolation Variables

- Isolation defined as amount of transverse momentum in cone with $\Delta R$ around the reconstructed photon coming from other photons, charged or neutral objects.
- Isolation variables are independent of SuperCluster.
- Photon isolation and charged isolation are used in photon identification.
- Isolation variables are discontinuous since only hits over energy threshold are recorded.