Particle identification at LHCb

New calibration techniques and machine learning classification algorithms

Anton Poluektov

University of Warwick, UK

5 July 2018

On behalf of LHCb collaboration
One-arm spectrometer optimised for studies of beauty and charm decays at LHC

- Good vertexing: measure $B^0$ and $B_s^0$ oscillations, reject prompt background
- Particle identification: flavour tagging, misID background
- High-resolution tracking
- Calorimetry: reconstruct neutrals ($\pi^0, \gamma$) in the final state
- Efficient trigger, including fully hadronic modes
Excellent Particle identification performance is vital for LHCb physics

\[ B^0_s \rightarrow \mu^+\mu^- \quad B^0_s \rightarrow K^+K^- \quad B \rightarrow K^*\gamma \]

- Background rejection for rare decays
- Classification of final states with the same topology
- Reduction of bandwidth in the trigger

[PLR 118, 191801 (2017)]

[JHEP 10 (2013) 183]

Identify long-lived final state particles based on information from subdetectors:

- Charged: $\pi, K, p, e, \mu$
- Neutral: $\pi^0, \gamma$
Identify long-lived final state particles based on information from subdetectors:

- **Charged:** $\pi, K, p, e, \mu$
- **Neutral:** $\pi^0, \gamma$
Areas for machine learning in PID:

- Identification of final state particles: supervised learning, multiclass classification
- Evaluation of PID efficiency from calibration data samples: unsupervised learning, density estimation
- Simulation of PID response: generative models.

See [next talk by Fedor Ratnikov]
“Fast calorimeter simulation in LHCb”

PID strategy and performance in Run2: see [talk by Carla Marin Benito]
PID variables

- Low-level PID information: likelihoods obtained from info of individual detectors
  - Rings in RICH detectors
  - Clusters in calorimeter
  - Hits in muon system

- Higher-level variables (ProbNN):
  - ANN output combining the above (+auxiliary info from tracking etc.)
  - 6 models for each of charged PID hypotheses + “ghost” (tracks not representing real particles)
  - Trained on MC
  - Baseline approach: MLP implemented in TMVA, 1 hidden layer
Advanced classification techniques for charged PID

Trying new classification techniques

- XGBoost [arXiv:1603.02754]
- CatBoost [arXiv:1706.09516]
- Boosting to flatness [JINST 10 (2015) T03002]
- Deep Neural Networks (keras library)

Improvements are possible with advanced classifiers, but careful choice of training samples is needed (more sensitive to kinematic properties than baseline).
Typically, PID performance depends on track kinematics ($p, \eta$) and event multiplicity.

Systematics-limited measurements: having a classifier with efficiency independent of kinematics/multiplicity is an advantage.

Flat4d: classifier trained with flatness term in loss function

$$\mathcal{L} = \mathcal{L}_{\text{exp}} + \alpha \mathcal{L}_{\text{FL}}, \quad \mathcal{L}_{\text{FL}} = \sum_b \int |F_b(s) - F(s)|^2 ds$$
Radiative decays (e.g. $B \rightarrow K^* \gamma$): sensitive to New Physics, energetic photons in the final state

Large backgrounds from $\pi^0$: high-momentum $\pi^0$ do not form separated clusters in ECAL.

Pattern recognition to separate $\gamma$ from $\pi^0$
$\gamma/\pi^0$ separation: baseline classifier

- Input features based on $3 \times 3$ “image” around a center of the cluster:
  - Shape of the cluster (width, tails, eccentricity, orientation)
  - Energies of the most and 2nd-most energetic cells
  - Hit multiplicity and shape in the preshower cells
- Output: MLP with 2 hidden layers in TMVA

---

**LHCb simulation**

**LHCb data**

$B^0 \rightarrow K^*\gamma$

- No $\pi^0$ suppression
- With $\pi^0$ suppression
  - $\epsilon_{\text{sig}} = 98\%$, $\epsilon_{\text{bkg}} = 55\%$
$\gamma/\pi^0$ separation: new classifier

- Input features: energy deposition in $5 \times 5$ ECAL and PS cells ("raw images")
- Training samples: $B \rightarrow K\pi\gamma$ (signal) and $B \rightarrow K\pi\pi^0$ (background)
- Several classifiers tried:
  - ANNs with 1–2 hidden layers, different optimisers (Adamax, Adagrad, SGD)
  - BDTs (XGBoost, CatBoost, LightGBM)

BDT with XGBoost shows the best performance (AUC=0.95)
PID resampling and correction

PID response is widely used in physics selections ⇒ need to reproduce it precisely in simulation to evaluate selection efficiency, background contamination.

PID performance is a complicated function of track kinematics and event multiplicity ⇒ multivariate problem.

Two procedures developed at LHCb:

- **Resampling (PIDGen)**: Using the known 4D distribution of calibration sample in $PID$ variable, track kinematics ($p_T$ and $\eta$) and event multiplicity ($N_{tracks}$), generate PID variable that looks like in data for any given track kinematics and multiplicity.

- **Variable transformation (PIDCorr)**: Using the above 4D distributions for data and MC, construct a function that transforms simulated PID response such that it matches data. This approach preserves correlations between different PID responses for the same track (e.g. $\pi$ and $K$ probabilities).

[arXiv:1803.00824]
PIDGen and PIDCorr: input variables

sPlot technique applied to calibration samples to statistically subtract background

Describe PDFs of the sWeighted calibration sample in 4 variables:
- PID variable (transformed to avoid sharp peaks)
- $\log p_T$
- Pseudorapidity $\eta$
- Track multiplicity $\log N_{\text{tracks}}$

sWeighted $D^{\ast\pm} \rightarrow D^0 \pi^{\pm}$ calibration sample
Four-dimensional kernel density estimation of calibration data performed using Meerkat library [JINST 10 (2015) P02011] [HepForge]

- Provides kernel-based correction to the approximated density
- Efficient with multidimensional PDFs

Example: two-dimensional projections onto PID $\log p_T$:
**PIDGen: validation of resampled variables**

**PIDGen**: discard simulated PID response, resample from calibration density for a given track $p_T, \eta$ and track multiplicity

$$\text{PID}_{\text{corr}} = P_{\exp}^{-1}(\xi|p_T, \eta, N_{\text{tracks}})$$

Performance is validated on independent clean high-statistics data samples.

PIDGen resampled variables (ProbNNK and ProbNNpi) for a kaon track from sWeighted $\Lambda^0_b \to \Lambda^+_c \pi^-$, $\Lambda^+_c \to pK^-\pi^+$ sample.
**PIDCorr**: preserve correlations between different PID responses for the same track. Transformation of simulated PID instead of complete resampling.

\[
\text{PID}_{\text{corr}} = P_{\text{exp}}^{-1} \left( P_{\text{MC}}(\text{PID}_{\text{MC}} | \rho_T, \eta, N_{\text{tracks}}) | \rho_T, \eta, N_{\text{tracks}} \right)
\]

Reproduce not only individual PID responses (ProbNNpi, ProbNNK, etc.), but also their combinations

\[c\Lambda_K \text{ from } P_{\text{cor}}, \quad c\Lambda_K \text{ from } (1-\text{ProbNNpi}) \times \text{ProbNNK0}\]

sWeighted \( \Lambda^0_b \rightarrow \Lambda_c^+ \pi^- \), \( \Lambda_c^+ \rightarrow pK^- \pi^+ \) data,
uncorrected simulation,
corrected simulation (PIDGen or PIDCorr).
Summary

- Particle identification at LHCb: a broad area to apply advanced machine learning techniques
- Several new approaches tested on Run1/Run2 data:
  - Multivariate classifiers for charged and neutral particle classification
  - Density estimation of calibration data: resampling and correction of MC PID response
- PID will be even more important after LHCb upgrade: software trigger including PID information
Backup
Input variables for charged ANN classifiers

**Tracking**
- Total momentum
- Transverse momentum
- Quality of the track fit
- Number of clusters associated to the track
- ANN response trained to reject ghost tracks
- Quality of the fit matching track segments upstream and downstream of the magnet

**RICH detectors**
- Geometrical acceptance of the three radiators, depending on the direction of the track
- Kinematical acceptance due to Cherenkov threshold for muons and kaons
- Likelihood of the electron, muon, kaon, and proton hypotheses relative to the pion
- Likelihood ratio of the below-threshold and pion hypotheses

**Electromagnetic calorimeter**
- Likelihood ratio of the electron and hadron hypotheses
- Likelihood ratio of the muon and hadron hypotheses
- Matching of the track with the clusters in the preshower detector
- Likelihood ratio of the electron and pion hypotheses, after recovery of the Bremsstrahlung photons

**Hadronic calorimeter**
- Likelihood ratio of the electron and hadron hypotheses
- Likelihood ratio of the muon and hadron hypotheses

**Muon system**
- Geometrical acceptance
- Loose binary requirement already available in the hardware trigger
- Likelihood of the muon hypothesis
- Likelihood of the non-muon hypothesis
- Number of clusters associated to at least another tracks
PIDGen and PIDCorr approaches

Sampling

Data

Simulation

Random generator in [0, 1)

Transform

PID-corrected dataset

Simulated dataset
Meerkat approach to density estimation

Traditional kernel density: data points $x_i$, kernel $K(x)$

$$P_{\text{KDE}}(x) = \sum_i K(x - x_i)$$


$$P_{\text{corr}}(x) = \frac{\sum_{i=1}^{N} K(x - x_i)}{(P_{\text{appr}} \otimes K)(x)} \times P_{\text{appr}}(x).$$

In other words, we represent the PDF as a product of approximation PDF and kernel correction:

$$P_{\text{corr}}(x) = f(x)P_{\text{appr}}(x)$$

$P_{\text{appr}}(x)$ takes care of boundary effects and narrow structures.

In the practical implementation, use binning with multilinear interpolation:

$$P_{\text{interp}}(x) = \frac{\text{Bin} \left[ \sum_{i=1}^{N} K(x - x_i) \right]}{\text{Bin} \left[ (P_{\text{appr}} \otimes K)(x) \right]} \times P_{\text{appr}}(x).$$