Pileup Subtraction and Jet Energy Prediction Using Machine Learning

Jiakun Li, Vein Kong, Yujia Zhang
Department of Electrical Engineering, Stanford University

I. Overview

I (a). Motivation

In the Large Hadron Collider (LHC), a critical task is to reconstruct information for a single primary collision (jet) from multiple proton-proton collisions. These additional collisions are known as pileup.

This project focuses on: 1) to create a regression algorithm that predicts the jet energy without noise from a list of features of the collisions. 2) to select the most important features from ~30 candidates.

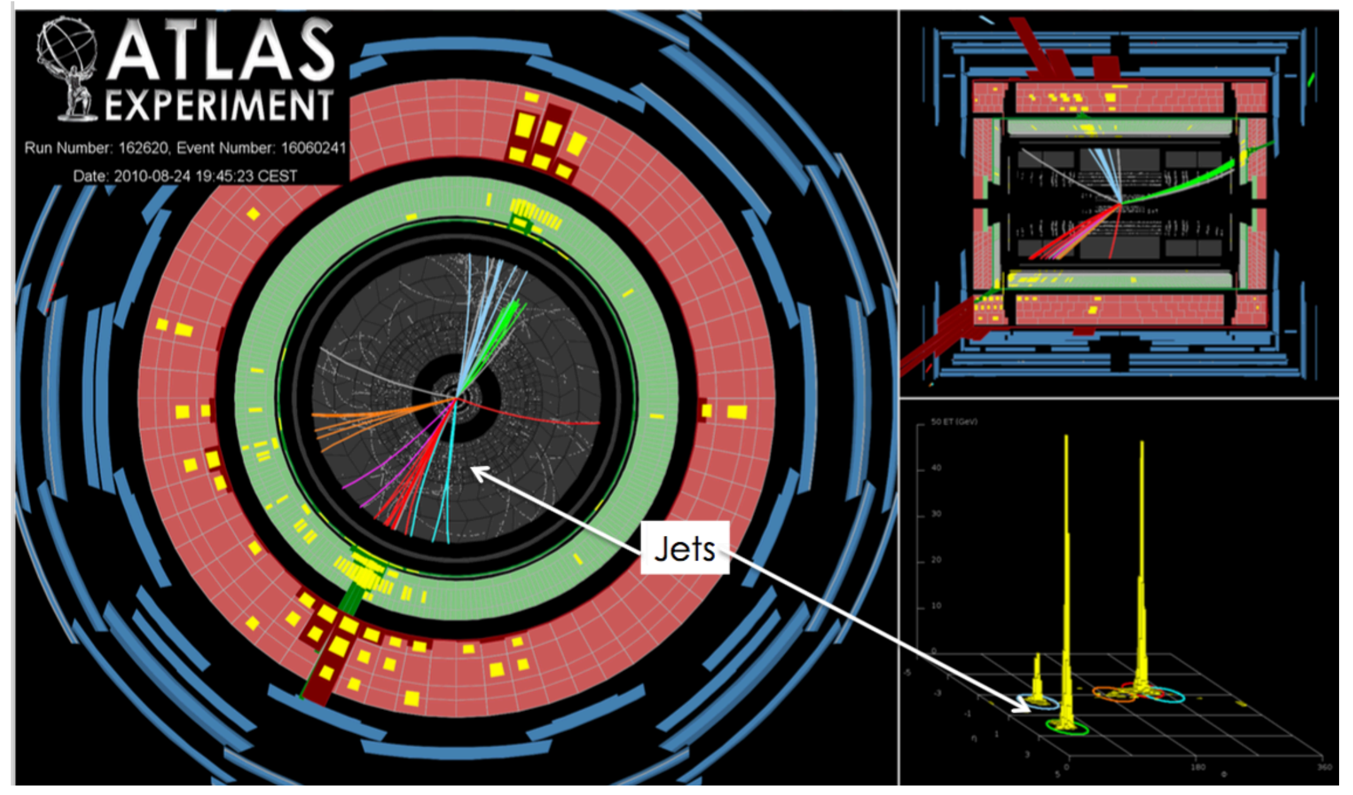
ATLAS EXPERIMENT
Run Number: 162620, Event Number: 16060241
Date: 2010-08-24 19:45:23 CEST
Jets

Fig 1. Conceptual representation of collisions and jets[1]

I (b). Features

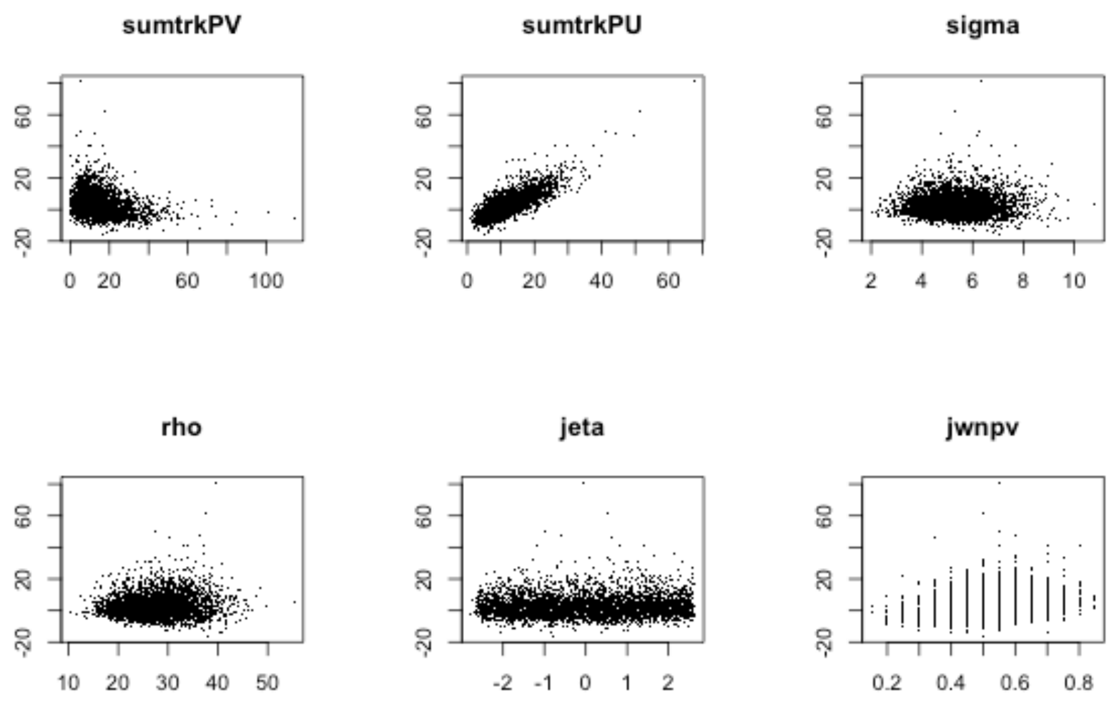

Fig 2. Correlation between (jpt-jptnoarea) and features

Feature	Physical Meaning
pt	Transverse momentum, the momentum that is perpendicular to the beamline of a particle detector
jpt	jet pt corrected by area based method: jpt = jptnoarea – rho∗area
jptnoarea	The raw observed jet transverse momentum including pileup energy contamination
sumtrkPV	Total pt of charged particles in the jet from the hard-scatter (signal vertex)
sumtrkPU	Total pt of charged particles in the jet from additional pileup vertices
rho	Event pt density
sigma	RMS of rho (measure of how large the point-to-point fluctuations of pileup are in the event)
jeta	Jet psedorapidity, spatial coordinate describing the angle of a particle relative to the beam axis
ptX	Jpt computed in circles of radius X/100, corrected by rho ∗ area
jwnpv	Fraction of vertices contributing to the jet with at least 1 charged particle
NPV	Number of pileup collisions in the event

I (c). Current method and Baseline

Matteo Cacciari and Gavin P. Salam[2] proposed an area-based approach to subtract pileup effects:

jpt = jptnoarea – rho∗ area

This approach suffers from inaccuracy by assuming pileup particles are uniformly distributed within each event and the pileup effect is only dependent on the jet area.

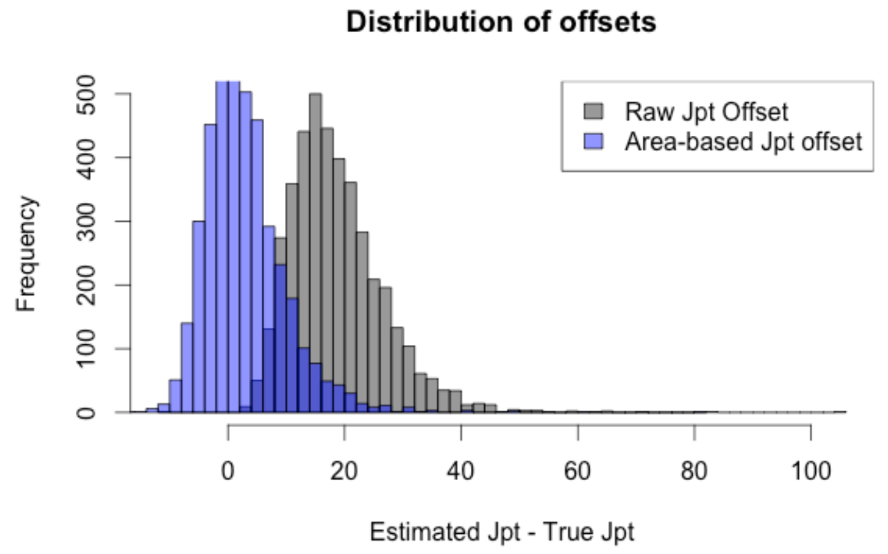

Fig 3. Distribution of offsets (raw and area-based correction)

II. Models

II (a). Linear Regression

- 80% training, 20% testing across samples of NPV from 20 - 50
- Tested with combination of features and targeted for
    - Min(mean(|tjpt – $\mathrm{jpt}_{\mathrm{predicted}}$|))
    - Min(var(|tjpt – $\mathrm{jpt}_{\mathrm{predicted}}$|))

II (b). Support Vector Regression

- SVR was used to check if high order features and interactions between features can improve regression models
- Results: training error decreased, test error increased, indication of over-fitting

II (c). Decision Tree

- Gradient boosting tree was applied
- No performance improvement from linear regression. Possible reason: Decision tree works best with piece-wise functions. The inherent model of collision data is not piece-wise.

III. Results

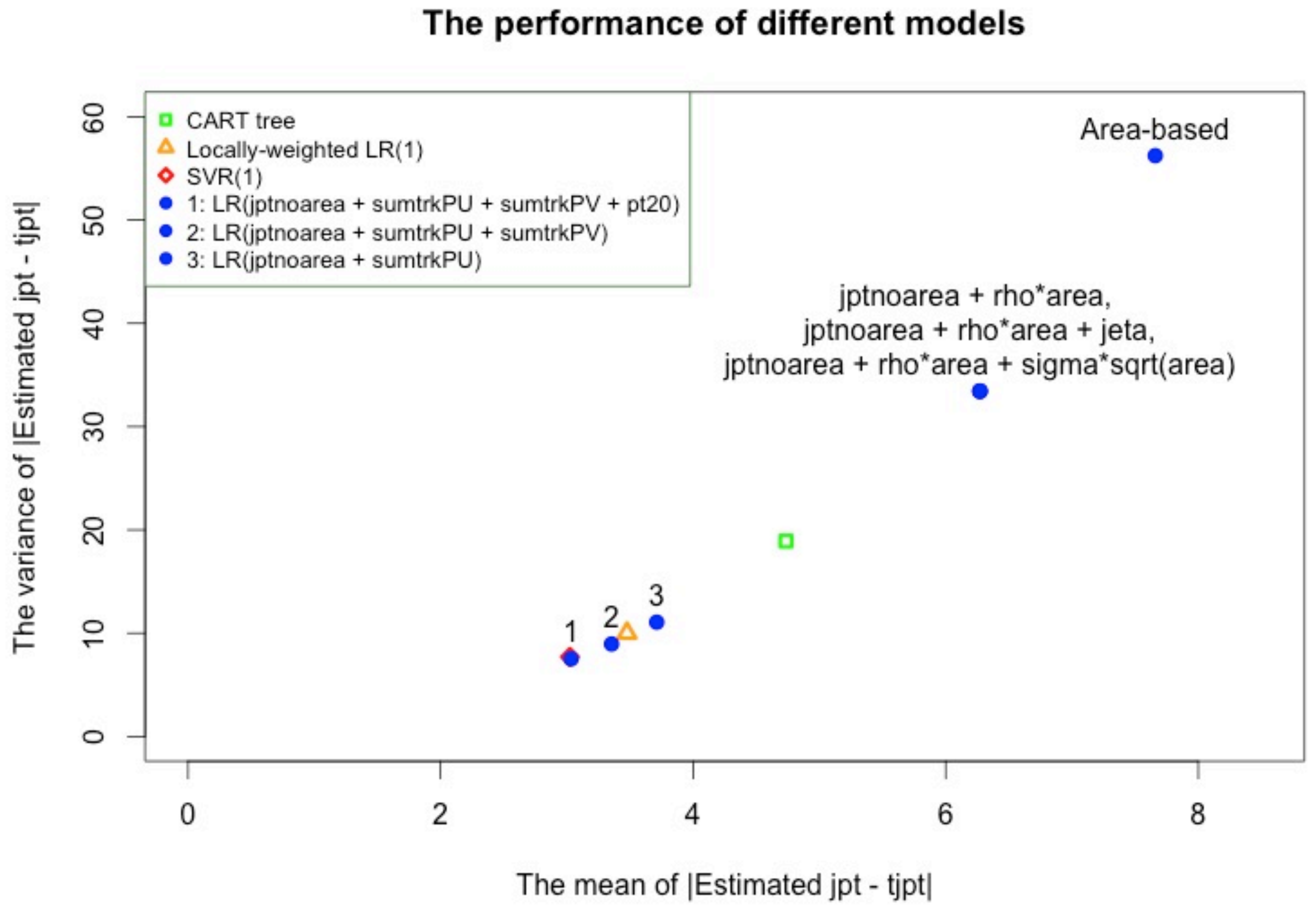

Fig 4. Performance comparison between different regression feature sets

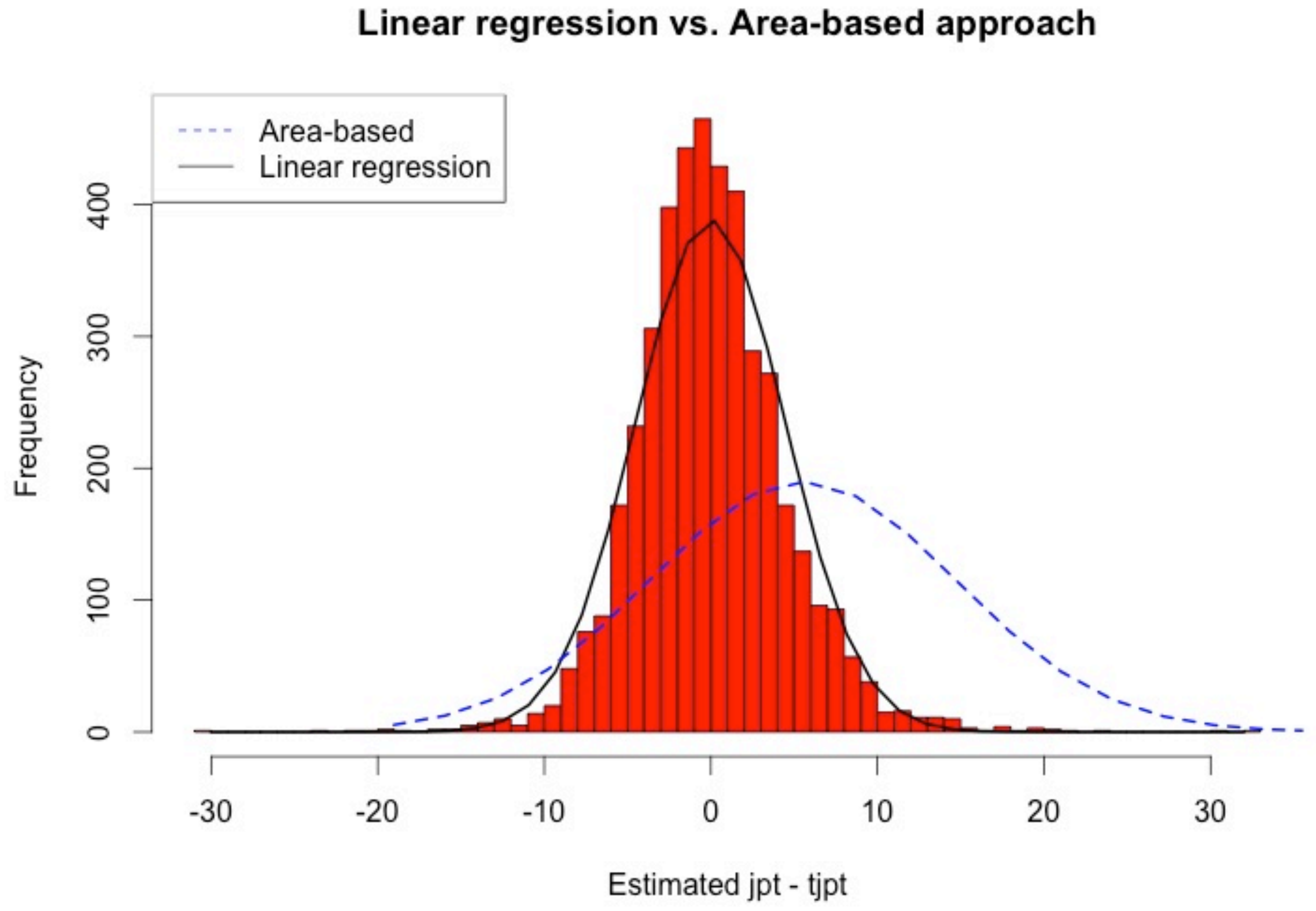

Fig 5. Performance comparison: best linear regression vs. Area-based approach

IV. Reference

[1] Lester Mackey, Ariel Schwartzman. Physics event reconstruction at the Large Hadron Collider, *Stanford Data Science Workshop*, 2015

[2] Matteo Cacciari, Gavin P. Salam. Pileup subtraction using jet areas. *Phys.Lett.B659*, pp. 119–126, 2008.