Branching structure of QCD jets: new jet observables for quark-gluon discrimination

Joseph Davighi, University of Cambridge

Abstract

I have explored the fractal nature of hadronic jets and the potential use of fractal dimension in jet substructure physics. A more sophisticated set of parameters, named Branching Logarithmic Fit (BLF) parameters, has subsequently been developed to describe the fractal and corrections-to-fractal behaviour, the latter arising from QCD running of the branching probability. Theoretical motivation is given for these parameters, which have then been applied to the problem of quark-gluon discrimination. The BLF parameters are individually discriminating and only weakly correlated to variables currently used in quark-gluon discrimination. Consequently, their inclusion is expected to improve discrimination, and evidence is presented for this at the generator level.

1 Introduction

Quarks and gluons produced in collider experiments shower and hadronize to produce collimated jets of hadrons. The parton shower proceeds via repeated branching interactions, generating successive "jets within jets". The final distribution of hadrons in the jet should therefore exhibit some fractal character. In this report I explore new ways of describing the branching structure of jets, motivating corrections to a perfect fractal structure due to the running of the QCD branching probability.

The motivation for this study is two-fold. Firstly, we seek to develop our description of the fractal properties of jets. Jets, being the observable manifestation of quarks and gluons, are a fundamental component of high energy physics interactions. Thus, to better understand their complex branching structure is an inherently important problem in particle physics.

Secondly, we investigate a direct application of these tools to quark-gluon discrimination [1]. Whether a jet originates from a quark or gluon is determined by the process producing the jet, and thus quark-gluon discrimination can be very useful in selecting certain physics events. For example, vector boson fusion (VBF) always produces two hard quark jets, which tend to be in the gluon-dominated forward region of the detector. Tagging these quark jets can enhance selection of VBF events, which would in turn facilitate high precision Standard Model tests concerning Higgs boson searches and measuring electroweak couplings. Quark tagging is also expected to be very useful for physics searches beyond the Standard Model, including SUSY searches.

Quarks and gluons have different colour charges, which gives rise to different showering patterns and thus differences in observable properties of the resulting jets. For example, existing quark-gluon probability taggers [3, 2] exploit the higher multiplicity (number of particles in the jet) and wider jet cone of gluon jets to statistically discriminate between quark-initiated jets and gluon-initiated jets. Here, I have extracted new jet observables from jet branching structure, named Branching Logarithmic Fit (BLF) parameters, and explored their potential for improving quark-gluon taggers.

In section 2, a simple toy model for the branching structure of jets is developed. This motivates the definition of the BLF parameters in section 3, in which their value in quark-gluon discrimination is then investigated. Finally, discussion and conclusions are presented in sections 4 and 5 respectively.

2 Theory of jet branching structure

2.1 Minkowski dimension

The Minkowski dimension (or, more transparently, the "box-counting" dimension) serves as our starting point for describing jet branching structure[1]. To calculate the appropriately named box-counting dimension of an object, we consider how many boxes $N(\epsilon)$ of a given side-length ϵ are required to cover the set of points in an object. For perfect Euclidean shapes, as we vary the box scale we find that

$$N(\epsilon) \propto 1/\epsilon^D, \tag{1}$$

where D is the topological dimension of the shape (e.g. $D = 1$ for lines, $D = 2$ for squares, $D = 3$ for cubes). Thus, we can define a geometrical "box counting" dimension for Euclidean shapes by the relation (1), which is equal to the topological dimension.

We then generalise from these traditional Euclidean elements to exactly self-similar shapes to define the Minkowski dimension of a fractal, D, as the exponent such that (1) holds. The Minkowski dimension is formally defined:

$$D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}. \tag{2}$$

The Minkowski dimension of a fractal can take non-integer values, and always exceeds its topological dimension.

2.2 Fixed strong coupling

We now wish to calculate the box-counting dimension of a jet. In the subsequent study (section (3)) jets have been clustered using the anti-k_t algorithm, with a jet cone radius $R = 0.4$. The jet cluster cone is divided in (η, ϕ) coordinates into a square grid of cells, each cell having dimension ϵ. For a given scale ϵ, we calculate the number of cells $N(\epsilon)$ which register particle hits. This is the number of "boxes" required to pave our jet at the scale ϵ.

In the simplified case of a fixed branching probability, we anticipate perfect fractal branching as a QCD jet cascades to lower energies. Thus we expect $N(\epsilon) \propto 1/\epsilon^D$ to hold and the jet to possess an exact Minkowski dimension which we can calculate using (2).

In reality however we do not have infinitely self-repeating structures. Our hadronic jets have a finite multiplicity. Thus, as we increase the granularity of the detector (approaching $\epsilon \to 0$) the number of hits will saturate at the jet multiplicity. Hence (2) is not a useful definition. We can instead, for self-similar branching, identify D with the average slope

$$D = \left\langle \frac{\log N(\epsilon)}{\log(1/\epsilon)} \right\rangle, \tag{3}$$

[1] The terms Minkowski dimension, "box-counting" dimension, and fractal dimension shall be used interchangeably. The term "Hausdorff dimension" is often used elsewhere in this context, however this term has a stricter mathematical definition. Nonetheless in all naturally-occurring fractals, Hausdorff dimension and box-counting dimension are equivalent.

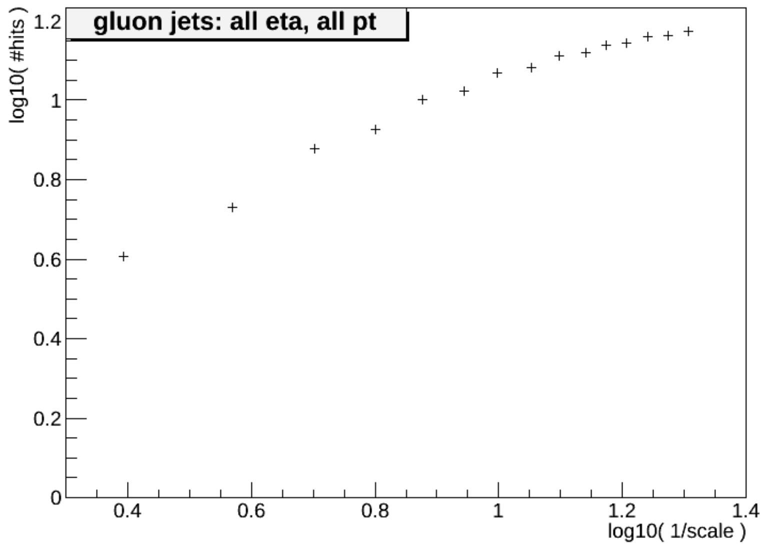

Fig. 1: Searching for suitable scales. For typical gluon jets, we do not observe an extended linear regime over which to calculate the Minkowski dimension. The running branching probability ensures the branching structure of jets is very much scale-dependent. The error bars are too small to see.

evaluated over a suitable range of scales where $\log N(\epsilon)$ is observed to increase linearly with $\log(1/\epsilon)$ (a "linear regime", indicative of fractal scaling) [4]. This provides a recipe for calculating the fractal dimension D of a jet in the limit of scale-invariant QCD.

QCD is, however, not scale-invariant. In Figure 1 we do not observe a distinct linear regime, but rather $\log N(\epsilon)$ levels off smoothly from large to small scales. The continuously varying gradient reflects the continuously varying branching probability. This observation implies the above recipe for extracting the Minkowski dimension may not be appropriate for characterising this scale-dependent branching structure.

I now present a simple argument to motivate an appropriate smooth function which describes the variation of $\log N(\epsilon)$ with $\log(1/\epsilon)$, incorporating how the probability of branching varies with scale.

2.3 Running strong coupling

A parton can branch via three process: $q \to qg$, $g \to gg$ and $g \to q\bar{q}$ (Figure 2). In all of these cases, the probability of a single branching interaction decreases at small energies, thus suppressing the number of hits $N(\epsilon)$ at small scales. Energy scales are proportional to angular scales, $Q \propto \epsilon$, as is seen by considering that two final state hadrons separated by a large angle will have branched from eachother early in the shower, thus from a highly energetic particle.

Let $P(\epsilon)$ be the probability of a single branching at a given angular scale ϵ. To relate the number of hits at a given scale of probing to the single-branching probability, we consider a small succession of branches at a particular scale ϵ ("small" in the sense that we can take $P(\epsilon)$ to be approximately constant over this set of branches).

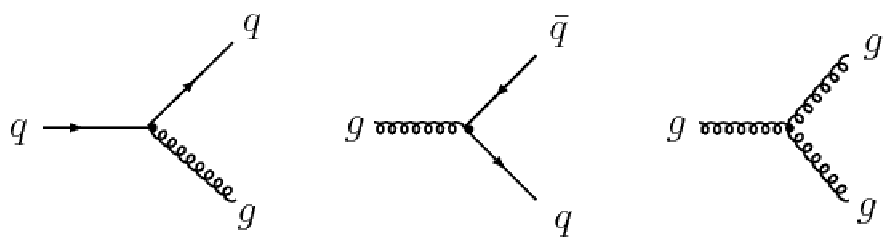

Fig. 2: QCD branching processes

- For our fixed scale ϵ (i.e. a fixed grid in (η, ϕ) space), the number of hits $N(\epsilon)$ will be proportional to the number of particles $N_{particles}(\epsilon)$ produced by the branches.
- The number of particles is one greater than the number of branching vertices, since each branching vertex produces an additional particle.
- The number of branching vertices is, on average, proportional to the *probability* of that number of branching vertices. Each vertex contributes a *multiplicative* factor $P(\epsilon)$ to the probability of this set of branches.
- Hence we expect the number of hits at a given scale to be approximately related to the branching probability at that scale by a power law:

$$N(\epsilon) \propto [P(\epsilon)]^{\delta}, \tag{4}$$

where δ is some characteristic (positive) exponent.

I then suggest a simple toy model for $N(\epsilon)$ in which we combine the scale-dependence of the branching probability only with the fractal increase in N with $1/\epsilon$ described in section 2.2 (that is, we naturally resolve more particles if the sampling grid is more finely divided):

$$N(\epsilon) = A\frac{[P(\epsilon)]^{\delta}}{\epsilon^{D}}, \tag{5}$$

where A, δ and D are all positive constants.

We must now derive the variation of $P(\epsilon)$ from perturbative quantum chromodynamics (pQCD). We remain in the perturbative regime because we are considering final state hadrons with an energy ~ 1 GeV, greater than the perturbative cut-off energy $\Lambda \sim 200$MeV. From pQCD, the dependence of the branching probability on energy scale Q is given by

$$P\left(Q^2\right) \sim \alpha_s(Q)\int^{Q}\frac{\log k^2 dk^2}{k^2} \sim \alpha_s(Q)\left(\log Q^2\right)^2, \tag{6}$$

which includes all terms from both the matrix element of the interaction (Figure 2), and also phase space factors, which depend on the energy scale. At the GeV scale of interest, α_s decreases approximately as

$$\alpha_s\left(Q^2\right) \approx \frac{1}{b_0\log\left(Q^2/\Lambda^2\right)}, \tag{7}$$

where the constant $b_0 > 0$. Assuming $Q \gg \Lambda$, we combine (6) and (7) to deduce:

$$P\left(Q^2\right) \sim \log Q^2.$$

Hence we see that the single branching probability decreases at small scales. This may at first seem counter-intuitive, since the strong coupling constant increases at small scales (small energies). However this running of α_s is more than offset by the restricting of phase space into which particles can branch, which also accompanies decreasing Q. Exploiting the energy-angular scale equivalence, we have $P(\epsilon) \sim \log \epsilon$.

Motivated by the definition of fractal dimension (2), we seek a relation between $\log N(\epsilon)$ and $\log(1/\epsilon)$. Inserting $P(\epsilon) \sim \log \epsilon$ into equation (5) and taking logarithms, we derive

$$\log N(\epsilon) = p_0 + p_1 \log(1/\epsilon) + p_2 \log(\log(1/\epsilon)), \tag{8}$$

where we have combined constants into the coefficients p_0, p_1 and p_2. Defining dependent and independent variables $y \equiv \log N(\epsilon)$ and $x \equiv \log(1/\epsilon)$ respectively, we write $y = p_0 + p_1 x + p_2 \log x$. The constant and linear terms are as for perfect fractal branching, in which case p_1 would traditionally be extracted as the Minkowski dimension of the jet. However, the running branching probability has produced the non-linear logarithmic term with coefficient p_2. We have thus derived the functional form of the pQCD correction to fractal branching in this toy model.

Using the formal definition (2), we see that the Minkowski dimension remains p_1, since

$$\lim_{x\to\infty} \frac{y}{x} = \lim_{x\to\infty} \left(\frac{p_0}{x} + p_1 + \frac{p_2 \log x}{x}\right) = p_1.$$

The fractal dimension is *formally* equal to the coefficient of the linear term. However the fractal dimension is not a meaningful way to describe our scale-dependent jet; the *set* of parameters p_0, p_1 and p_2 is a more appropriate characterization of the branching structure.

3 BLF parameters in Q/G discrimination

3.1 BLF and BQF parameter definition

Two classes of fitting functions have been used to model the variation of $\log N(\epsilon)$ with $\log(1/\epsilon)$ over a wide range of scales:

1. logarithmic fits of the form $y = p_0 + p_1 x + p_2 \log x$.
2. quadratic fits: $y = p_0 + p_1 x + p_2 x^2$.

In each case, the range of scales has been set by paving the jet cone with a square grid of {2,3,4,...,16} divisions in both η and ϕ coordinates[2].

A power series expansion of the logarithm term (about some non-divergent scale x_0) is here instructive to motivate the quadratic fit:

$$f(x) \equiv \log x \approx \log x_0 + \frac{1}{x_0}(x - x_0) - \frac{1}{2x_0^2}(x - x_0)^2 + ...$$

The first two terms simply apply corrections to the constant and linear terms in the logarithmic fitting function. The sign of the quadratic correction is negative for all choices of x_0, as required to produce small-scale suppression of $\log N(\epsilon)$. These fits are displayed, for generator level jets (defined in subsection 3.2), in Figure 3.

[2] For a given number of divisions, the scale ϵ is equal to twice the jet radius divided by the number of divisions.

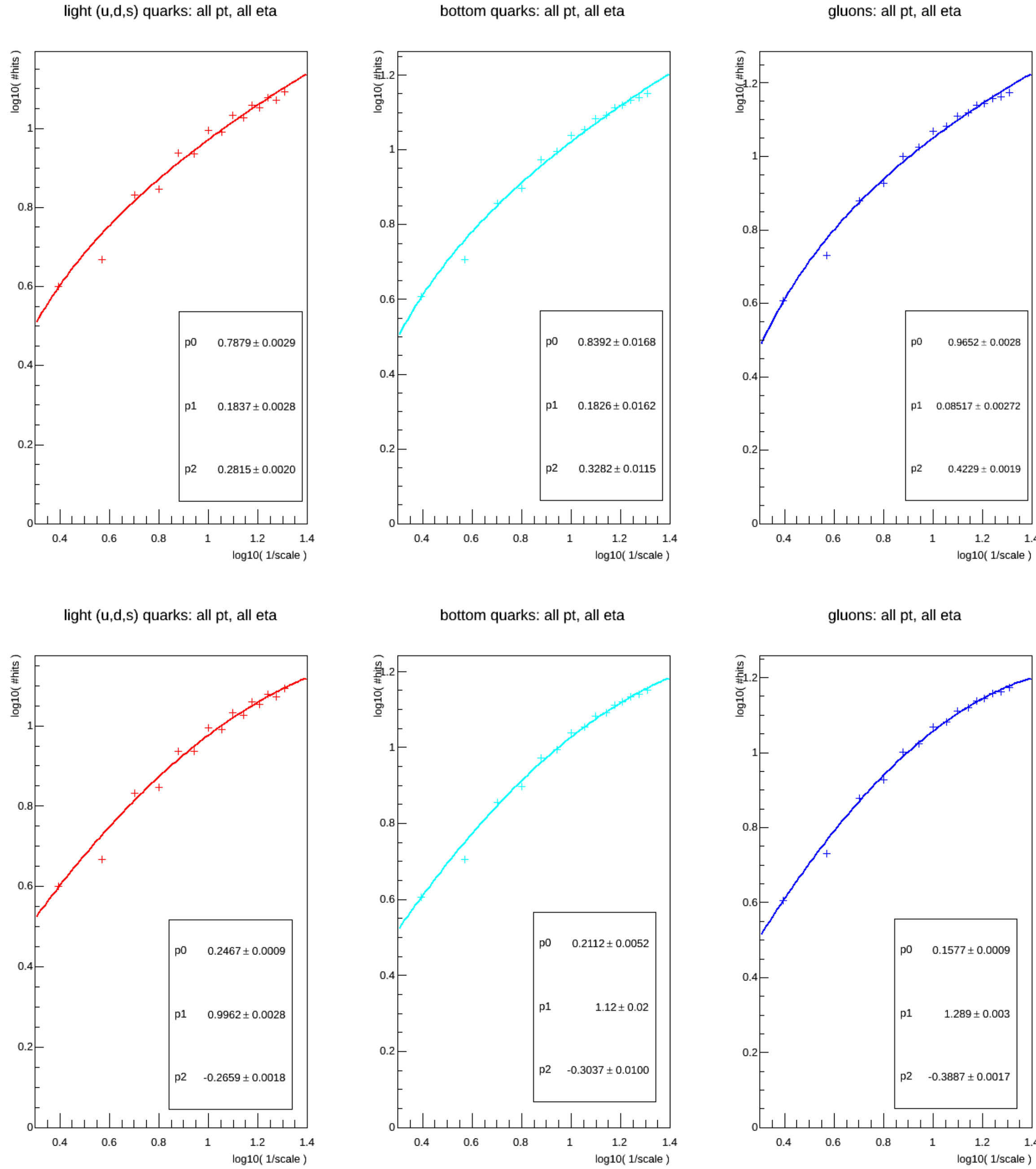

Fig. 3: Top: logarithmic fits to $\log N(\epsilon)$ against $\log(1/\epsilon)$ for light quarks, bottom quarks, and gluons, of the form $y = p_0 + p_1 x + p_2 \log x$. At small scales, $\log N(\epsilon)$ levels off as motivated in section 2.3. The error bars on each point are too small to see. Bottom: quadratic fits to $\log N(\epsilon)$ against $\log(1/\epsilon)$ of the form $y = p_0 + p_1 x + p_2 x^2$. We observe both the BLF and BQF parameters for the b-jets are midway between those of light jets and gluons.

The fitted coefficients, which are seen in Figure 3 to discriminate between light quarks, bottom quarks, and gluons, are extracted as new jet observables named the Branching Logarithmic Fit (BLF) and Branching Quadratic Fit (BQF) parameters. These parameters are intended to provide a simple characterisation of the complex branching structure.

3.2 Application to Quark/Gluon tagging: generator level study

A generator level study has been conducted to investigate the value of these parameters in Quark-Gluon discrimination. QCD dijet event samples generated using HERWIG at a centre-of-mass energy of 8 TeV are used throughout. All final state hadrons are required to have $p_T > 0.1\text{GeV}$. The flavour of a jet is determined by the standard method of searching through the generating partons, and matching each jet to the highest-p_T parton within $R < 0.3$ of the jet axis. The flavour of the jet is then identified with that of the matched parton. An event is classified as signal (background) if the leading jet is matched to a quark (gluon). All jet observables have then been evaluated for the leading jet in each event.

The BLF parameters were found to add more discriminatory power than the BQF parameters. Given also that the BLF parameters have a more direct theoretical motivation, in this report we shall consider only the use of the BLF parameters in Quark-Gluon tagging.

The variables currently used by the Compact Muon Solenoid (CMS) Quark-Gluon tagger are [2]:

- The number of particles in the jet (multiplicity) [5];
- The $p_T D$ variable, which describes the distribution of transverse momentum between the particles in the jet, with the definition $p_T D = \frac{\sqrt{\Sigma_i p_{T,i}^2}}{\Sigma_i p_{T,i}}$;
- The semi-minor axis of the jet in the $\eta - \phi$ plane, denoted σ_2.

There is no *a priori* reason to suppose any strong correlation between these three variables and the BLF parameters defined above, and indeed this is borne out at generator level. From Figure 4, the correlation coefficients between the BLF parameters and multiplicity, $p_T D$ and σ_2 are all suitably low, compared, for example, to the correlation between multiplicity and $p_T D$. We see very strong correlations *between* the BLF parameters, as is natural given they are parameters derived from the same fit. While the combination of all three BLF parameters adds little discrimination beyond that of a single parameter, their combination may nonetheless be used in constructing a Boosted Decision Tree (BDT) discriminator, which is designed to deal with correlated variables.

Figure 5 shows that the BLF parameters are also individually discriminating, particularly if we seek high signal efficiency. Thus, we have reason to hope that the addition of these parameters to a Quark-Gluon discriminator should improve performance. Figure 6 shows that the gain in discrimination given by replacing multiplicity with BLF parameters in the existing QG discriminator is greater than the gain achieved by including multiplicity in the existing discriminator in the first place. For example, for a quark efficiency of 60%, we have gluon rejection rates of 67.5% (for a BDT discriminator constructed using only $p_T D$ and σ_2), 68.4% ($p_T D$, σ_2 and multiplicity - the existing discriminator), and 69.6% ($p_T D$, σ_2 and the BLF parameters - our new discriminator). We see the improvements are at the $\sim 1\%$ level. Combining $p_T D$, σ_2, multiplicity and the BLF parameters was seen to give negligible further improvement.

Finally, a study of the dependence of these BDT discriminators on transverse momentum was undertaken, and it was found that discrimination is superior for high p_T jets, with the new discriminator performing slightly better in all p_T bins. An example is plotted in Figure 7.

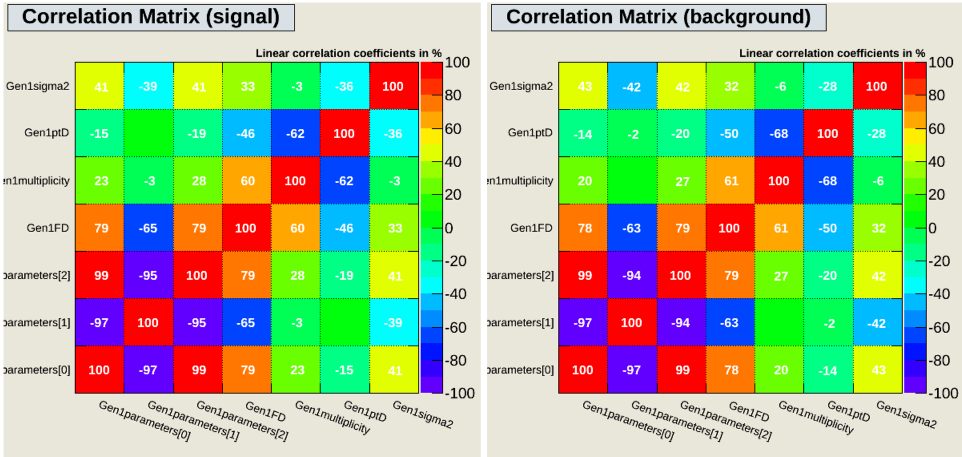

Fig. 4: Correlation matrices. We here display the linear correlation coefficients between pairs of variables. We include the existing QGD variables (multiplicity, $p_T D$ and minor axis (sigma2)), the BLF parameters, and also the Minkowski fractal dimension (FD), calculated from a linear fit over a small range of scales. The left matrix is for quark jets (signal) and the right for gluon jets (background). We see that the BLF parameters are almost perfectly correlated to eachother as we expect. More importantly, we see only weak correlations between the BLF parameters and the three existing QGD variables. We also see that the traditional fractal dimension is more strongly correlated to existing QGD variables, particularly multiplicity, as is expected from its definition.

4 Discussion

From this study, I suggest that small improvements to the CMS Quark/Gluon tagger should be possible if we include the BLF parameters - new jet observables motivated and defined in this report, which characterise the branching structure of jets. A more comprehensive generator level study using larger and/or different samples is desired. Perhaps more urgently, we must now study the performance of these new variables in reconstruction using, for example, the Particle Flow algorithm [6] or the PUPPI algorithm [7] to reconstruct particles in the presence of high pile-up.

Other ideas to extend this work further include:

- Investigating the p_T-dependence of the BLF parameters. Preliminary studies have suggested a strong p_T-dependence. It is therefore possible that the full discriminatory power of the BLF parameters is somewhat masked by their variations with p_T, in which case adding the p_T of the jet to the TMVA could improve discrimination.

- Exploring further related ways of characterising the branching structure. A development of the BLF parameters might involve exploring particle hits in a 3-dimensional space, perhaps spanned by η, ϕ and $1/z$, where z is the fractional transverse momentum of a particle in the jet. Further theoretical work will also be needed to motivate the extraction of new jet observables from such an approach.

- Applying the BLF parameters not to Quark-Gluon discrimination, but to the identification of pile-up jets.

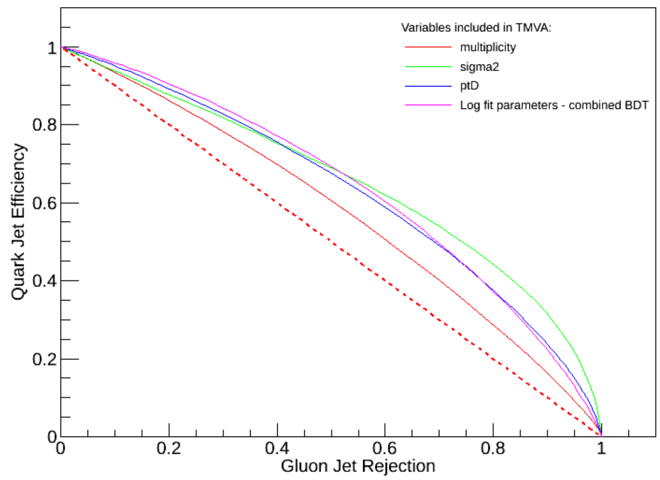

Fig. 5: Single variable performance comparison, in terms of quark efficiency and gluon rejection. ROC curves are plotted for each of multiplicity, $p_T D$, σ_2 and a BDT discriminator built from the BLF parameters alone (a single BLF parameter could have been used to very similar effect, due to their near-perfect correlation). BLF parameters, minor axis, and $p_T D$ are most discriminating. The BLF parameters are most discriminating for high signal efficiencies ($\gtrsim 70\%$), below which jet minor axis becomes most discriminating. Multiplicity is by far the least discriminating variable studied here.

This method of studying jet branching structure is a new approach, and as such there are many exciting ways in which we might proceed further with refining, re-defining, and applying these techniques to new problems in jet physics.

5 Conclusion

A simple theoretical model has been derived to describe the branching structure of jets, incorporating a correction to fractal branching due to running of the branching probability. This model produces good fits to the results of generator-level jet simulations, and the parameters fitted have then been extracted as new jet observables; BLF (Branching Logarithmic Fit) and BQF (Branching Quadratic Fit) parameters.

I have then sought to apply the BLF parameters to the problem of Quark/Gluon discrimination, because they are both individually discriminating and weakly correlated to the variables currently used by the CMS Q/G tagger. At the generator level, evidence has been presented for some modest improvement in discrimination by $\sim 1\%$ gluon rejection (for a given quark efficiency) when we replace multiplicity with the BLF parameters in the exisiting tagger.

Acknowledgements

I thank Philip Harris for supervising me in this work; for his continued generous support, guidance and discussion. I thank also the CERN Summer Student Program for facilitating this research.

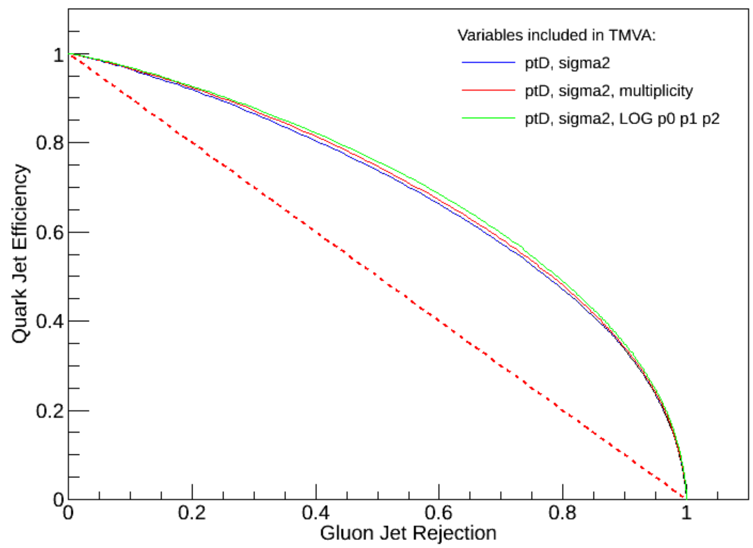

Fig. 6: Discriminator performance for combinations of variables. TMVA is used to construct a BDT discriminator from the given combinations of variables in each case. The red curve corresponds to the existing QG discriminant, and the green line the proposed discriminant in which multiplicity is replaced by the BLF parameters. This yields a slight improvement in gluon rejection for a given quark efficiency, at the $\sim 1\%$ level.

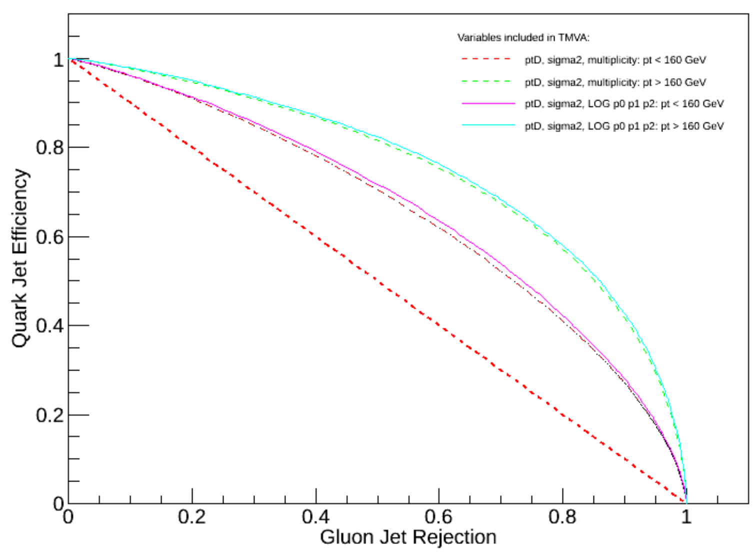

Fig. 7: p_T dependence of Quark-Gluon tagging performance. We see that discrimination improves at high p_T. In each p_T bin, performance is better for the new QG tagger.

References

[1] J. Gallicchio and M. D. Schwartz, *"Quark and gluon tagging at the LHC"*, Phys.Rev.Lett. 107, 172001 (2011), arXiv:1106.3076 [hep-ph]

[2] CMS Collaboration, *"Performance of quark-gluon discrimination using pp collision data at $\sqrt{s} = 8$ TeV"*, CMS Physics Analysis Summary CMS-PAS-JME-13-002 (2013)

[3] The ATLAS Collaboration, *"Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector"*, submitted to Eur. Phys. J. C (2014), arXiv:1405.6583v3 [hep-ex]

[4] M. Ruan, D. Jeans, V. Boudry, J. Brient and H. Videau, *"Fractal Dimension of Particle Showers Measured in a Highly Granular Calorimeter"*, Phys.Rev.Lett. 112 (2014) 1, 012001 arXiv:1312.7662 [physics.ins-det]

[5] E. D. Malaza, *"Multiplicity distributions in quark and gluon jets"*, Zeitschrift fur Physik C Particles and Fields 31 (1986) 143 – 150, doi:10.1007/BF01559605

[6] CMS Collaboration, *"Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus and E_T^{miss}"*, CMS Physics Analysis Summary CMS-PAS-PFT-09-001 (2009)

[7] D. Bertolini, P. Harris, M. Low and N. Tran, *"Pileup Per Particle Identification"*, submitted to JHEP (2014), arXiv:1407.6013 [hep-ph]