Data-driven estimation of neutral pileup particle multiplicity in high-luminosity hadron collider environments

Federico Colecchia
Brunel University London, Kingston Lane, Uxbridge UB8 3PH, UNITED KINGDOM
E-mail: federico.colecchia@brunel.ac.uk

Abstract. The upcoming operation regimes of the Large Hadron Collider are going to place stronger requirements on the rejection of particles originating from pileup, i.e. from interactions between other protons. For this reason, particle weighting techniques have recently been proposed in order to subtract pileup at the level of individual particles. We describe a choice of weights that, unlike others that rely on particle proximity, exploits the particle-level kinematic signatures of the high-energy scattering and of the pileup interactions. We illustrate the use of the weights to estimate the number density of neutral pileup particles inside individual events, and we elaborate on the complementarity between ours and other methods. We conclude by suggesting the idea of combining different sets of weights with a view to exploiting different features of the underlying processes for improved pileup subtraction at higher luminosity.

1. Introduction

The contamination, or background, from low-energy processes described by Quantum Chromodynamics (QCD) is a major challenge at the Large Hadron Collider (LHC), and the potential impact on physics analysis is anticipated to become even more significant in the upcoming operation regimes of the accelerator. In fact, the average pileup rate, i.e. the rate of low-energy interactions between other protons, will increase with the instantaneous luminosity of the collider, and this is anticipated to place stronger requirements on the correction techniques employed at the LHC experiments.

The presence of multiple vertices inside collision events due to pileup can significantly complicate the extraction from the data of the physics quantities of interest, and calls for the use of dedicated subtraction techniques. Established methods that are part of the core reconstruction pipelines at the LHC experiments rely on the use of tracking information for charged particles, as well as on estimates of the pileup energy flow associated with neutral[1] particles [1], for which the task of assigning a vertex of origin is in general significantly more difficult.

In the light of the upcoming operation regimes of the LHC, algorithms of a different nature have recently been proposed and are currently being evaluated. Methods such as those presented in [2, 3, 4] assign individual particles inside collision events weights that reflect the probability of the particles originating from soft, i.e. low-energy, QCD interactions as opposed to the signal hard parton scattering. The use of the weights to rescale the particle four-momentum vectors [2]

[1] Whenever neutral particles are referred to in the text, neutrinos are not considered.

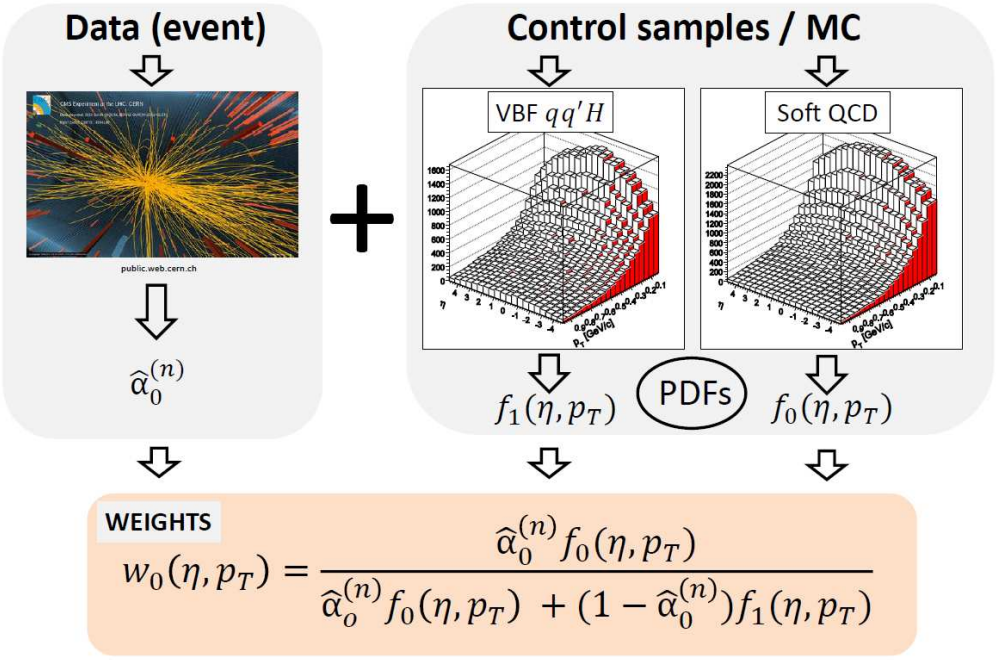


Figure 1. Schematic representation of the data processing involved in the calculation of the particle weighths, as described in the text. High-statistics control samples are used to estimate the shapes of the particle-level PDFs for signal (Standard Model Higgs boson production via vector boson fusion (VBF) from proton-proton interactions at $\sqrt{s} = 14$ TeV) and for background soft QCD interactions. The overall fraction of neutral particles originating from background is calculated event by event as described in section 3.2. This information is then combined into the particle weights, $w_0(\eta, p_T)$, which reflect the probability of individual neutral particles originating from soft QCD interactions as opposed to the hard parton scattering, based on the (η, p_T) bin they belong to. The results reported in this article correspond to 50 pileup vertices per event.

has been shown to result in improved performance in terms of pileup subtraction when compared to traditional methods. It is worth noting that the implications of assigning weights to individual particles were also explored from a different perspective in [5], where a connection was established with the idea of multiple interpretations of the data, and where the potential benefit to physics analysis was discussed.

2. Our approach

The contribution of this article is two-fold.

- *A different choice of particle weights.* We are proposing a different definition of the particle weights, which makes use of information about the underlying physics that, to our knowledge, is not employed by other methods. As opposed to exploiting the existence of collinear singularities in the physics that underlies the showering process [2], our technique relies on the particle-level kinematic signatures of the hard parton scattering and of the soft QCD interactions. Our aim is to estimate the variability in the shapes of the distributions inside individual collision events that is associated with statistical fluctuations in the data.

- *A different application of the weights.* We investigate a different use of the weights, whereby the particle-level kinematic distributions in the data are reshaped in order to estimate the number of neutral pileup particles in different kinematic regions inside individual events.

We treat each event as a heterogeneous statistical population of particles that have their origin either in the signal hard parton scattering or in background soft QCD interactions. Although it is generally not possible to map individual particles to a single physics process in a hadron collider environment due to colour connection, the use of weights that reflect the likelihood of particles originating from either process provides a way of addressing this conceptual issue.

Figure 1 shows a schematic representation of the procedure employed to calculate the weights. The shapes of the particle-level probability density functions (PDFs) corresponding to the signal high-energy parton scattering (in this article, Standard Model Higgs boson production via vector boson fusion) and to soft QCD interactions are obtained from control samples. Since the data sets in question are high-statistics, the effect of statistical fluctuations in the data is averaged out, and the shapes of the distributions reflect the expectation from the underlying physics processes. On the other hand, the overall fraction of neutral particles associated with soft QCD interactions is estimated event by event, which takes into account the variability of the neutral pileup particle fraction across collisions. This information is then combined into the particle weights, $w_0(\eta, p_T)$, which reflect the expected fraction of neutral soft QCD particles in each event as a function of particle η and p_T.

We discuss the results of an initial study on Monte Carlo data at the generator level, and show that our weights can be used to estimate the number of neutral pileup particles across the kinematic space inside individual events with reasonable accuracy.

The technique described in this article is based on a deterministic variant of a Markov Chain Monte Carlo algorithm that we proposed for particle-by-particle filtering of individual events at the reconstruction level [6, 7]. Our main goal is to contribute to improve on the subtraction of soft QCD background in high-luminosity hadron collider environments using algorithms that can be implemented at the reconstruction level. Specifically, we are targetting a processing stage upstream of jet reconstruction, i.e. before the particles are clustered together according to their likelihood of originating from the same scattered hard parton. The advantages of this algorithm over the previous stochastic version are its parallel nature and the simplicity of the calculations involved.

The results shown in this article complement those discussed in [8] with reference to a different signal process: as opposed to $t\bar{t}$ production via gluon fusion, the signal distributions reported in the following relate to Standard Model Higgs boson production via vector boson fusion, which does not involve colour exchange between the colliding protons.

3. The algorithm

This section describes the calculation of the particle weights, as well as the use of the weights to estimate the number of neutral pileup particles in different regions of the particle (η, p_T) space inside individual events. This study concentrates on the region $-5 < \eta < 5$, $0 < p_T < 1$ GeV/c, which contains most particles associated with soft QCD interactions. The (η, p_T) space was subdivided into bins of widths $\Delta\eta = 0.5$ and $\Delta p_T = 0.05$ GeV/c.

3.1. Control sample PDF templates

High-statistics control samples were used to obtain the shapes of the average particle-level (η, p_T) distributions corresponding to particles originating from the signal hard parton scattering and from background soft QCD interactions. Monte Carlo data sets were generated using Pythia 8.176 [9, 10], corresponding to the following:

- ***Signal***: ~300,000 final-state particles associated with Standard Model Higgs boson production via vector boson fusion, i.e. $qq' \rightarrow qq'WW(ZZ) \rightarrow qq'H$, from pp collisions at $\sqrt{s} = 14$ TeV.
- ***Background***: ~300,000 soft QCD particles corresponding to 50 pileup vertices per event.

Such high-statistics distributions reflect the expectation from the corresponding physics processes whereby the effect of statistical fluctuations in the data is averaged out, and can therefore be used to estimate the expected, or average, number of neutral pileup particles in each (η, p_T) bin. On the other hand, the corresponding unknown actual numbers generally deviate from the expected values, and, given the typical particle multiplicity inside LHC events, the discrepancy is often non-negligible. The calculation of the statistical uncertainty on the estimated number of neutral soft QCD particles in each (η, p_T) bin is discussed in section 3.4.

The high-statistics (η, p_T) distributions obtained in this study for neutral final-state particles associated with the hard parton scattering and with soft QCD interactions are shown as part of the schematic representation of the workflow of the analysis in figure 1. The plots were rotated around the vertical axis in order to make the distributions more clearly visible.

3.2. Event-by-event neutral particle fractions

In addition to the shapes of the probability distributions described in section 3.1, the calculation of the expected number of neutral soft QCD particles in each (η, p_T) bin, $\nu_b(\eta, p_T)$, also requires an event-by-event estimate of the overall fraction of neutral particles originating from soft QCD interactions as opposed to the hard parton scattering, $\hat{\alpha}_0^{(n)}$. For the purpose of calculating the particle weights, assigning a value to $\hat{\alpha}_0^{(n)}$ is essentially equivalent to specifying the relative normalisation of the signal and background distributions.

The quantity $\hat{\alpha}^{(n)}$ was estimated in each event based on the corresponding charged particle fraction, $\hat{\alpha}_0^{(c)}$, according to this formula:

$$\hat{\alpha}_0^{(n)} = \min(k\hat{\alpha}_0^{(c)}, \hat{\alpha}_0^{(c)}), \tag{1}$$

where the use of "min" ensures that $\hat{\alpha}_0^{(n)} \in [0, 1]$.

The correction factor k takes into account the difference between charged and neutral particle kinematics, including the effect of charged particles with $p_T \lesssim 0.5$ GeV/c not reaching the tracking detectors. The ratio between the fraction of neutral soft QCD particles and the corresponding charged fraction, calculated using Monte Carlo, is shown in figure 2. The distribution was obtained over the events analysed in this study, and the corresponding average value, $\langle k \rangle = 1.02$, was used as the value of k.

3.3. Particle weights

The above information was combined into the definition of the particle weights as follows:

$$w_0(\eta, p_T) = \frac{\hat{\alpha}_0^{(n)} f_0(\eta, p_T)}{\hat{\alpha}_0^{(n)} f_0(\eta, p_T) + \hat{\alpha}_1^{(n)} f_1(\eta, p_T)}, \tag{2}$$

where $\hat{\alpha}_1^{(n)} \equiv 1 - \hat{\alpha}_0^{(n)}$. Figure 3 (a) displays $w_0(\eta, p_T)$ in the region $-5 < \eta < 5$ and $0 < p_T < 1$ GeV/c, corresponding to one of the events analysed.

This choice of weights reflects the probability of individual particles originating from soft QCD interactions as opposed to the hard parton scattering, based on the shapes of the expected (η, p_T) PDFs as well as on the estimated overall fraction of neutral soft QCD particles in each event. This highlights the difference between our approach and other weighting methods that rely on

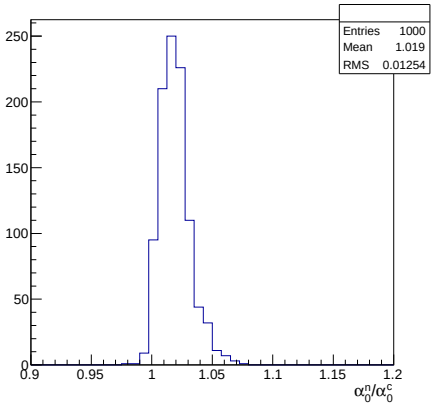
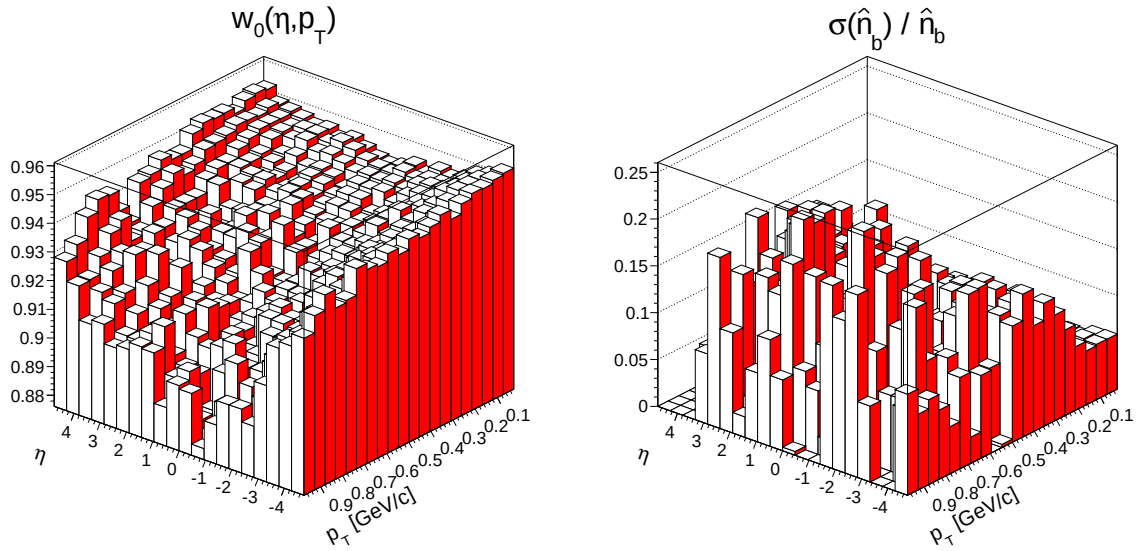


Figure 2. Ratio between the fraction of neutral particles associated with soft QCD interactions and the corresponding charged fraction, $\alpha_0^{(n)}/\alpha_0^{(c)}$, from Monte Carlo. The distribution was obtained over the events generated in this study.

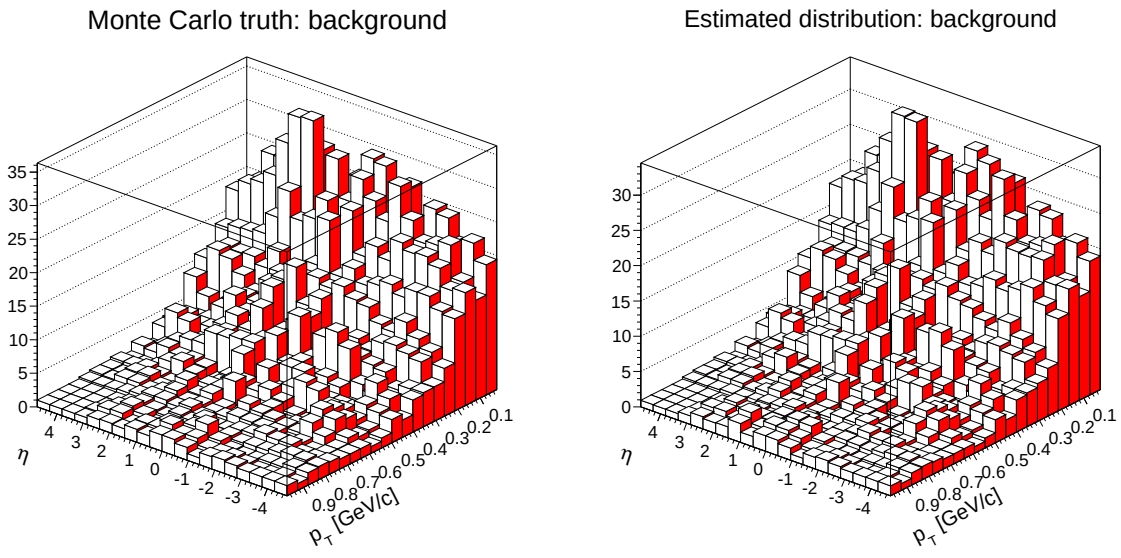


(a) (b)

Figure 3. (a) Particle weights, $w_0(\eta, p_T)$, across the (η, p_T) space in the region investigated, $-5 < \eta < 5$, $0 < p_T < 1$ GeV/c. The plot, which corresponds to one of the events analysed, was rotated around the vertical axis in order to make the distribution more clearly visible. (b) Relative uncertainty on the estimated number of neutral pileup particles, $\sigma_{\hat{n}_b}(\eta, p_T)/\hat{n}_b(\eta, p_T)$.

particle-to-particle proximity measures and that exploit different properties of the underlying physics, e.g. the existence of collinear singularities in the showering process [2].

It should be noted that our decomposition of the particle (η, p_T) space is relatively coarse-grained, particularly along the η axis. This is one of the reasons why we are not proposing that these weights be used in isolation, but rather in conjunction with those employed by other methods. It is our opinion that combining different sets of weights reflecting different properties of the underlying physics processes, e.g. using multivariate techniques, can result in improved rejection of neutral pileup particles. For instance, some of the results shown in [2] seem to suggest over-subtraction of soft QCD radiation, whereby particles originating from the hard parton scattering can be interpreted as pileup-related, and we expect the addition of a weighting



(a) (b)

Figure 4. (a) True particle-level (η, p_T) distribution of neutral soft QCD particles corresponding to one of the events analysed in this study. The effect of statistical fluctuations on the shape of the distribution is apparent. (b) The corresponding (η, p_T) distribution estimated using this algorithm.

algorithm not based on particle proximity to result in improved background rejection as the average pileup rate increases.

3.4. Reshaping the particle-level kinematic distributions

In this section, we illustrate a different use of the particle weights with reference to the quantity $w_0(\eta, p_T)$ discussed in section 3.3. As opposed to employing the weights to rescale the particle four-momentum vectors [2, 3, 4], we use them to reshape the particle-level (η, p_T) distribution in each event. We show that this approach allows the estimation of the number of neutral soft QCD particles in each (η, p_T) bin with reasonable accuracy regardless of whether or not signal particles are present.

Given an event and the definition of $w_0(\eta, p_T)$, the expected number of neutral soft QCD particles in each (η, p_T) bin is given by:

$$\nu_b(\eta, p_T) = w_0(\eta, p_T) n(\eta, p_T), \tag{3}$$

where $n(\eta, p_T)$ is the corresponding number of neutral particles in the data. Given $\nu_b(\eta, p_T)$, the unknown number of neutral soft QCD particles in each (η, p_T) bin can be treated as a binomial random variable with mean given by (3) and standard deviation:

$$\sigma_{\hat{n}_b}(\eta, p_T) = \{n(\eta, p_T) w_0(\eta, p_T) [1 - w_0(\eta, p_T)]\}^{\frac{1}{2}}. \tag{4}$$

In conclusion, the number of neutral soft QCD particles in each bin can be estimated in terms of:

$$\hat{n}_b(\eta, p_T) = w_0(\eta, p_T) n(\eta, p_T) \pm \sigma_{\hat{n}_b}(\eta, p_T). \tag{5}$$

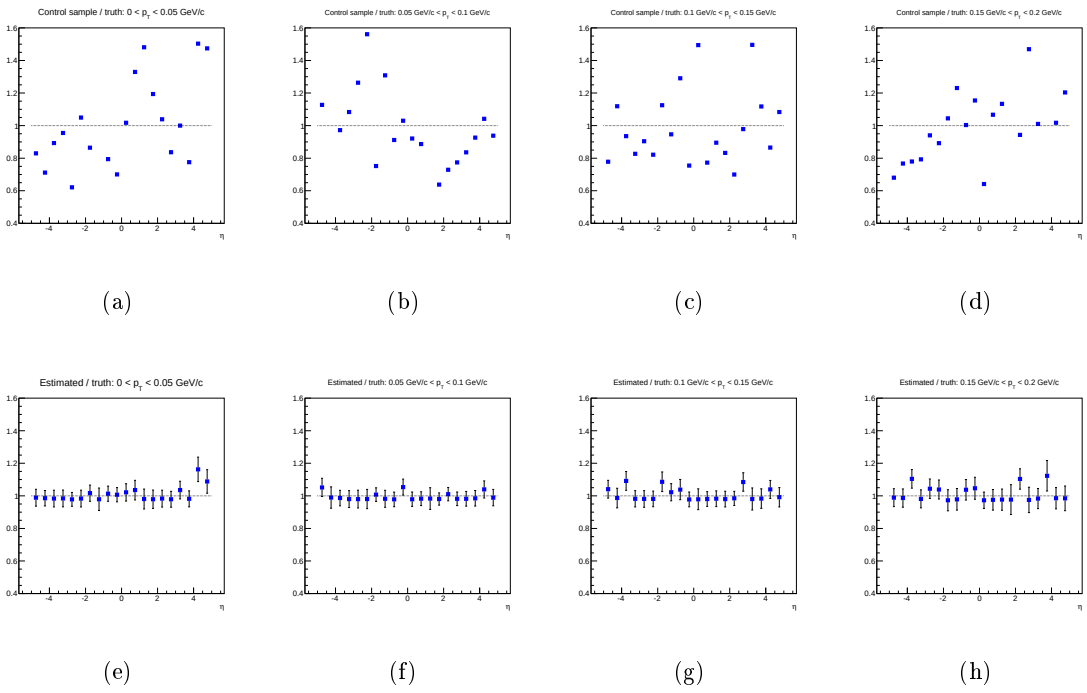


Figure 5. (a-d) Ratio between the control sample and the true (η, p_T) distribution of neutral soft QCD particles, both normalised to unit volume. The ratios are shown as a function of particle η in different p_T bins in the region $0 < p_T < 0.2$ GeV/c. The error bars (not visible on this scale) correspond to one Poisson standard deviation on the control sample bin contents. (a) $0 < p_T < 0.05$ GeV/c. (b) 0.05 GeV/c $< p_T <$ 0.1 GeV/c. (c) 0.1 GeV/c $< p_T <$ 0.15 GeV/c. (d) 0.15 GeV/c $< p_T <$ 0.2 GeV/c. (e-h) Ratio between the estimated (η, p_T) distribution of neutral soft QCD particles and the corresponding true distribution, both normalised to unit volume. The ratio is displayed as a function of particle η in different p_T bins, and the error bars correspond to one binomial standard deviation on the bin contents in $\hat{n}_b(\eta, p_T)$. (e) $0 < p_T < 0.05$ GeV/c. (f) 0.05 GeV/c $< p_T <$ 0.1 GeV/c. (g) 0.1 GeV/c $< p_T <$ 0.15 GeV/c. (h) 0.15 GeV/c $< p_T <$ 0.2 GeV/c.

The quantity $\sigma_{\hat{n}_b}(\eta, p_T)/\hat{n}_b(\eta, p_T)$ is shown in figure 3 (b) corresponding to one of the events analysed in this study. The performance of the algorithm is illustrated in figure 4, which provides a comparison between the true (a) and the estimated (b) number of neutral pileup particles across the (η, p_T) space in the same event. The accuracy of the estimated shape of the distribution of neutral pileup particles in the event is further illustrated in figure 5, where the ratio to the true distribution of the control sample (a-d) and of the estimated distribution (e-h) is displayed as a function of particle η in different p_T bins in the region $0 < p_T < 0.2$ GeV/c. The error bars in (e-h) correspond to one binomial standard deviation on the bin contents in $\hat{n}_b(\eta, p_T)$.

It should be noted that the idea of employing the weights to reshape the (η, p_T) distribution in the data can be used in conjunction with any definition of the weights, and that, in particular, it does not require all particles in the same (η, p_T) bin to have equal weights. In fact, if $S(\eta, p_T)$ denotes the set of particles i in the event inside a given (η, p_T) bin, the procedure outlined above is equivalent to estimating $\hat{n}_b(\eta, p_T)$ according to $\hat{n}_b(\eta, p_T) = \sum_{i \in S(\eta, p_T)} w_i$, where w_i is the weight assigned to particle i. The quantity $\sum_{i \in S(\eta, p_T)} w_i$ in fact reduces to $w_0(\eta, p_T) n(\eta, p_T)$ when all particles in the same (η, p_T) bin have the same weight, $w_i = w_0(\eta, p_T)$.

It is also worth noticing that the algorithm is inherently parallel, since different bins can be processed independently. We believe that the simplicity and parallelisation potential of this technique make it a suitable candidate for inclusion in future particle-level event filtering procedures upsteam of jet reconstruction at high-luminosity hadron collider experiments.

4. Conclusions

With reference to the upcoming higher-luminosity regimes of operation of the Large Hadron Collider, it is our opinion that the combination of different sets of particle weights encoding complementary information about the underlying physics processes has the potential to improve further on pileup subtraction, i.e. on the rejection of background particles originating from other proton-proton collisions.

We have discussed a choice of weights that, unlike that employed by other methods, is not based on particle-to-particle proximity, but rather on the particle-level kinematic signatures of the signal hard parton scattering and of the low-energy strong interactions. We have also shown that, when the weights are used to reshape the particle-level kinematic distributions inside individual collision events, they lead to reasonable estimates of the number density of background neutral particles across the kinematic space.

As more particle weighting methods become available, we envisage the possibility of developing algorithms based on multivariate combinations of different sets of weights with a view to exploiting all the particle-level information available in the data to reject neutral pileup particles. This study is based on a deterministic variant of a Markov Chain Monte Carlo algorithm that we previously discussed in conjunction with the idea of filtering individual collision events on a particle-by-particle basis at the reconstruction level in high-luminosity hadron collider environments. The main advantages of this approach, as compared to the previous stochastic version, are its parallelisation potential and the simplicity of the calculations involved.

Acknowledgments

The author wishes to thank the High Energy Physics Group at Brunel University London for a stimulating environment, and particularly Prof. Akram Khan, Prof. Peter Hobson and Dr. Paul Kyberd for fruitful conversations, as well as Dr. Ivan Reid for help with technical issues. Particular gratitude also goes to people the author had fruitful discussions with at an early stage of development of this research idea, namely Prof. Jonathan Butterworth, Prof. Trevor Sweeting and Dr. Alexandros Beskos at University College London, as well as Prof. Carsten Peterson and Prof. Leif Lönnblad at Lund University.

References

[1] The CMS Collaboration 2014 PAS JME-14-001
[2] Bertolini D, Harris P, Low M and Tran N 2014 *J. High Energy Phys.* 1410 059
[3] Cacciari M, Salam G P and Soyez G 2014 (Preprint arXiv:1407.0408 [hep-ph])
[4] Berta P, Spousta M, Miller D W and Leitner R 2014 *J. High Energy Phys.* 1406 092
[5] Kahawala D, Krohn D and Schwartz M D 2013 *J. High Energy Phys.* 1306 006
[6] Colecchia F 2012 *J. Phys.: Conf. Ser.* **368** 012031
[7] Colecchia F 2013 *J. Phys.: Conf. Ser.* **410** 012028
[8] Colecchia F 2014 (Preprint arXiv:1412.1989 [hep-ph])
[9] Sjöstrand T, Mrenna S and Skands P 2006 *J. High Energy Phys.* 0605 026
[10] Sjöstrand T, Mrenna S and Skands P 2008 *Comput. Phys. Comm.* **178**(11):852-67