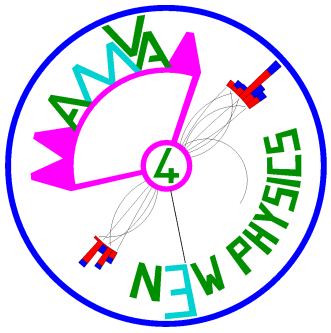
This Report is part of a project that has received funding from the **European Union's Horizon 2020 research and innovation programme under grant agreement Nº675440**



AMVA4NewPhysics ITN

Work Package 4 - Deliverable 4.1

Report of the Performance of Algorithms for Data-Driven Background Shape Modeling

AMVA4NewPhysics authors

February 27, 2017

Abstract

This document describes studies performed within the AMVA4NewPhysics network to identify performant methods for the modeling of background processes to new physics signals of interest. We focus our attention on the background arising from QCD multijet production, and the specific use case of modeling that process in the search for a small signal from Higgs boson pair-production and decay to two b-quark pairs. We consider three different methods for the task at hand, and discuss in detail the most promising one: a novel technique called "hemisphere mixing", which is shown to allow a precise modeling and offers several benefits with respect to competing methods. A set of detailed statistical tests proves the value of the algorithm, which is shown to provide a suitable background modeling for the $hh \to b\bar{b}b\bar{b}$ signal search.

Contents

1 Introduction

Processes mediated by Quantum Chromodynamics (QCD), which often yield multijet final state topologies, constitute a problematic background to searches for rare phenomena in hadron-hadron collisions. Even when Monte Carlo generators can be trusted to produce a reliable modeling of the final state under investigation, the issue is usually the huge cross section of the involved processes, which calls for very large simulations. Computational limitations typically end up preventing analysts from using QCD Monte Carlo samples for modeling purposes, or limit the statistical accuracy of the measurements. One example of the above situation is the search for non-resonant pair production of Higgs bosons in the 4-b-quark final state, $hh \to b\bar{b}b\bar{b}$, of interest to Work Package 1 of the AMVA4NewPhysics network. The data sample on which that search is based is collected by hadronic triggers, and is dominated by QCD processes. Monte Carlo simulations are usually unable to cope with the large required cross sections of the involved processes, such that when modeling multi-inverse-femtobarn datasets like those available in Run 2 at the LHC, they are not useful tools for anything other than coarse checks.

QCD multijet final states are also very complex. Not only is the QCD matrix element of the hard subprocess complicated by itself: to that one must add the presence of multiple parton scattering reactions, pileup effects, and multiple emission of initial and final state radiation following non-trivial color coherence patterns. Modeling in detail all the above features is a quite demanding task. In fact, Monte Carlo simulations can be trusted to produce a reasonable representation of real data only in the bulk of the phase space, and much less so in the unexplored corners where new physics signals are most frequently sought after.

In this document we describe a novel technique which we specifically designed to address the above problem. We also study its performance and its applicability to the search for pair production of Higgs bosons with a decay to two pairs of b-quark jets.

1.1 Plan of this document

Section 2 discusses methods we developed in the attempt of modeling the QCD background using estimates of the rate at which jets are found to contain a signal of b-quark hadronization. These methods are assessed as not sufficient for the complex task of identifying a small signal of hh pairs in LHC collider data; the following sections therefore focus on a different, more promising method, called *hemisphere mixing*. In Sec. 3 we describe in detail how the hemisphere mixing technique works. The data used for studies of the proposed algorithm are described in Sec. 4. Section 5 describes a number of statistical tests that we performed to verify the soundness of the method and its applicability to our considered use case. We draw some conclusions in Sec. 6.

2 Methods based on data-driven estimates of the b-tagging probability

2.1 Matrix-based methods

Historically, hadron collider experiments faced with the task of modeling the difficult QCD background in multijet final states have tried with mixed success both data-driven and simulation-based techniques. The seminal example –the first one where a parametrization of the b-quark tagging probability was studied– is the analysis which led to the discovery of the top quark in data collected by the CDF experiment at the Fermilab Tevatron in the nineties [3]. There, two methods were devised to estimate the rate of "W plus jets" events (which could hide the signature of single-lepton top-pair decays) due to QCD radiation off electroweak-produced W bosons. A "method 1" relied on a data-driven estimate of the rate at which jets were identified to contain a signal of b-quark hadronization (called b-tagging), carried out on independent samples of inclusive jets. An alternative "method 2" relied on Monte Carlo simulations of the background processes to estimate the same quantity.

While hadron collider searches studying final states that include real leptons (which originate from electroweak processes) benefit from the fact that the contributing background processes have a relatively small cross section, searches in fully-hadronic final states have to cope with QCD-driven processes, whose huge cross section makes a full simulation impractical. There, the technique most frequently deployed to estimate the rate of final states rich of b-quark jets (such as the ones we are focusing on in our activities within Work Package 1 of AMVA4NewPhysics) is a method-1-like one, in the jargon above; it has come to be called more frequently "matrix method". In it, the probability that a jet contains a b-quark tag is parametrised, in control samples of data designed to be poor in terms of signal contribution, as a function of the most relevant observable characteristics of the jet and of the event containing it. This probability can then be used to work out estimates of the number of b-tagged jets in an independent signal sample. Similar methods have been extensively used, e.g. in the publication that provided evidence for the first time an all-hadronic top-antitop signal in CDF data [2].

We studied the b-tag matrix approach as a preliminary investigation of the analysis strategy for the $hh \to b\bar{b}b\bar{b}$ search in CMS. The starting point consists in pre-selecting data which have a signal-like topology, i.e. events with at least four energetic jets; a requirement is made that at least three of the jets be tagged by the b-tagging algorithm, mimicking a pre-selection already applied by the CMS trigger used for the $hh \to b\bar{b}b\bar{b}$ search. Starting with this selected dataset B, which due to the huge rate of QCD processes is still dominated by background processes, one proceeds by identifying the variables characterizing the fourth jet (in order of the b-tag discriminator value) whose value is most strongly correlated with the relative rate at which the jet is labeled as b-tagged. The method does not attempt a distinction of jets that contain real b-quark decay products from jets that are incorrectly b-tagged by the secondary-vertex finding algorithm; rather, the *inclusive* rate of real and spurious b-tags is considered. The set of variables most likely connected with the relative b-tagging rate (which is effectively a b-tag probability) depends only mildly on the details of the data selection, and usually includes the jet transverse momentum and the jet pseudorapidity.

Once the variables most strikingly affecting the b-tag probability are sorted out, one may proceed to bin them in intervals small enough that within the bin support the probability can be considered costant. If one is, for example, considering three jet-related variables together, then depending on the size of the background-rich dataset B available one may decide what is the total number of bins that produces a detailed enough mapping of the b-tag probability, while keeping the statistical uncertainty in the individual estimates small enough. For example,

a one-million-jet sample divided into a three-dimensional feature space produces an average of 1000 events per bin if each variable has been divided in 10 bins. This may be appropriate if the shape of the b-tag probability as a function of the considered variables is smooth, while a finer division is required for more rapidly varying functions. One sees that already a 10x10x10 division of the data causes the typical relative statistical uncertainty in the binomial ratio of tagged over total jets to surpass the value of 3%, which seriously limits the usefulness of the approach; the typical variance-bias tradeoff is at work here, as a coarser binning reduces the variance at the expense of a potential systematic bias due to the approximation of a varying function with its mean value within each bin.

In our application we found it possible to attempt a reduction of the variance, by operating with only two variables: the jet p_T and pseudorapidity. Hence let us consider below the mathematical formulation of the matrix in the two-dimensional case. Under the assumption that we operate on a signal-free dataset (where it is understood that the signal would affect the probability estimates, as the b-tag probability of its jets would be higher), we may construct the matrix as follows:

$$P(i,j) = \frac{N^{b-tag}(i,j)_B}{N^{all}(i,j)_B} \tag{1}$$

where $N^{b-tag}(i,j)_B$ is the number of b-tagged jets contained in bin i,j in dataset B, and $N^{all}(i,j)_B$ is the total number of jets in the same bin of B; i,j are identifiers of the bins for the two considered variables over which the b-tag probability is parametrized by the matrix P. Once the matrix is constructed using the background-rich sample B, one may turn one's attention to a subsample of data S (independent from the previous one) selected in such a way that the signal component might be visible there. A prediction of the number of background-contributed b-tags in the signal sample S can be obtained as follows:

$$N_S{}^{bgr\,b-tags,exp} = \sum_i \sum_j P(i,j) N(i,j)_S \tag{2}$$

where $N(i,j)_S$ is the number of jets in bin (i,j) found in the data sample S. Since the above procedure can be applied to any subset of the signal-enriched dataset S, one may obtain with it a quantitative estimate of the number of background-produced b-tags integrated over all space or, which is even more proficuous, differentially as a function of whatever interesting variable one wishes to study. One common choice for the latter is a signal-discriminating variable. In the case of our interest, we have two discriminating variables of interest: the invariant mass of the first and the second dijet pair, once the four selected jets have been effectively paired to reconstruct the hh decay topology (the pairing algorithm is described below). In that case, one may verify whether the background prediction agrees with the observed data in regions where the signal is not expected to contribute; this is useful in order to confirm the soundness of the procedure and to provide an estimate of the systematic uncertainty on the background prediction.

In Fig. 2 we show how the b-tag parametrization, obtained using data in a control region devoid of signal, can faithfully reproduce the density of one-dimensional distributions of relevant kinematical variables. In Fig. 3 (left) we show how the number of events with four b-tags in a sample of pure background events, divided in the two-dimensional plane of the two reconstructed dijet masses, is well predicted by the matrix method. The same procedure evidences instead an excess of events over background predictions when the procedure is run on a mixed sample of background and signal (Fig. 3, right).

The overall evaluation of the matrix method for the search of non-resonant Higgs pair production is that it reliable, as it produces the expected results. However, there are two issues with the studied application. The first one is the need to rely on an insufficient base

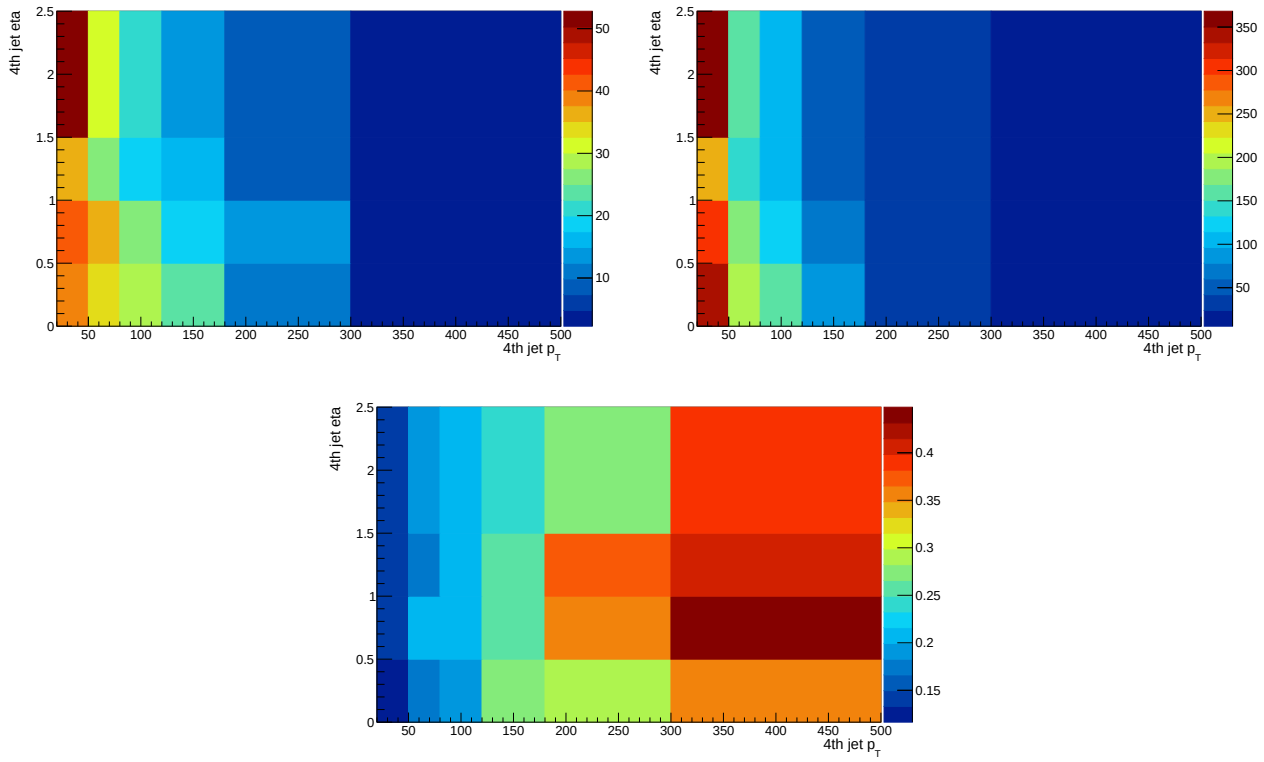


Figure 1: Example of parametrization of the b-tagging probability as a function of the p_T (horizontal axis) and absolute value of jet pseudorapidity (vertical axis) of the fourth jet (ordered by the b-tag discriminator value). Here is shown the result of the exercise on a Monte Carlo simulation of $t\bar{t}$ decays. The binning in the two variables has been chosen as a compromise of two different requirements: to try and optimize the relative population, obtaining a roughly constant variance of the resulting probability estimates, and to prevent the b-tag probability from varying too much within single bins. The left graph on the top row shows the distribution of b-tagged jets, the right graph shows the distribution of all considered jets in the B sample. The graph in the second row shows instead the estimated b-tag probability, obtained as the ratio of the two previous distributions.

of "control region" data, which results in the matrix-based probability estimate being affected by a significant statistical uncertainty, even when only a two-dimensional parametrization is used; the second is the fact that, at least for what concerns the search performed with the CMS experiment, the data used for the parametrization (named B sample) already contains three b-tagged jets because of the trigger requirements applied by the experiment's online data acquisition system, and thus is not completely signal-free; the signal contamination is small, but it is not negligible with respect to the signal contamination of whatever signal-rich region one may define once a selection of four-b-tagged events is applied. This *relative* non-negligible nature of the signal contamination in the control region affects the estimated b-tagging probability with a positive bias, which results in a negative bias on the extractable signal component in the signal region. This is a common issue of matrix-based methods, which is indeed usually corrected by an iterative procedure. However, in our case the contamination cannot be reduced to a level small enough that the resulting systematic uncertainties on the strength of the extractable signal be acceptably small. For these reasons, we conclude that we need to try and devise an algorithm which offers greater sensitivity to the modeling of the QCD backgrounds for the $hh \to b\bar{b}b\bar{b}$ search.

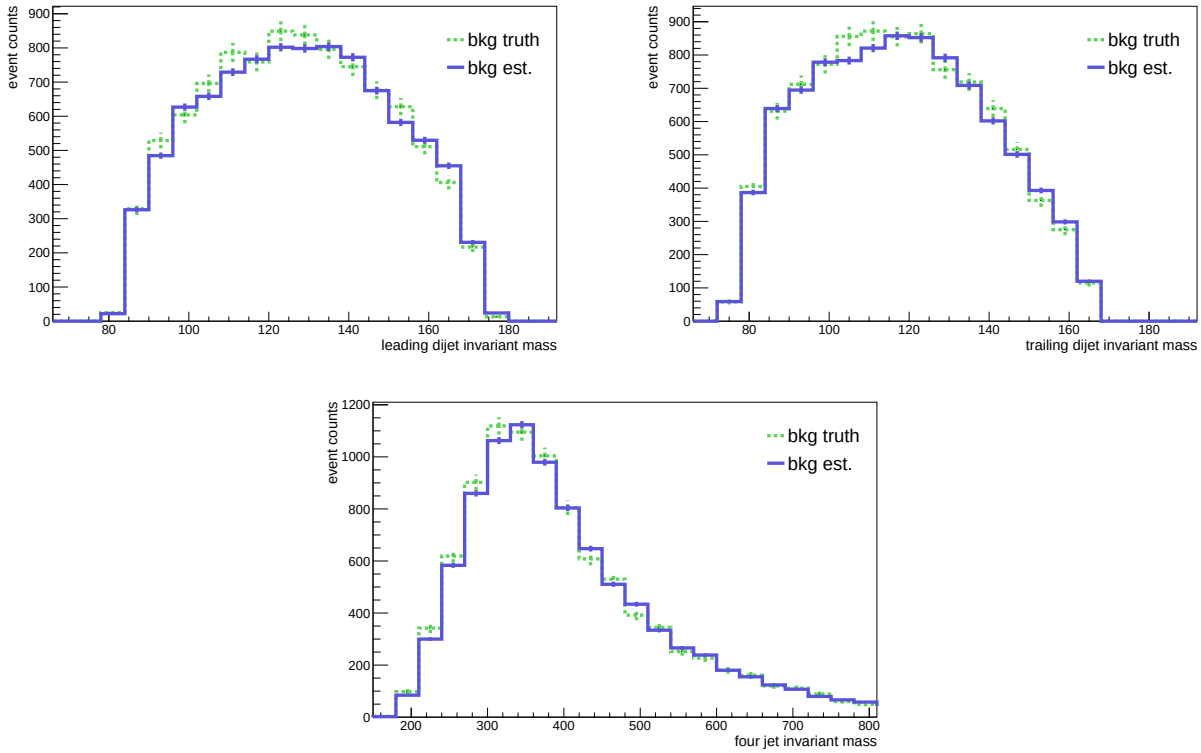


Figure 2: Top row, left: Distribution of the leading dijet mass in background-pure data compared to the background prediction; right: distribution of the trailing dijet mass. Bottom row: distribution of the di-Higgs mass. In all graphs, the original simulated data is in blue, and the matrix-based background prediction is in green.

2.2 The nearest-neighbor approach

A variant, and a possible improvement, of the matrix method described in the previous subsection is constituted by nearest-neighbor-based parametrizations of the b-tagging probability, in the feature space of kinematical observables correlated to the b-tag rate from spurious as well as real b-quarks. Rather than estimating the b-tagging rate in fixed intervals of the variables, as the matrix method does, nearest-neighbor estimates operate a continuous interpolation of the same quantity. Although the "curse of dimensionality" is not really reduced by the continuous estimate of the rate, the method does improve over matrix approaches, especially since it can easily handle situations where the b-tagging rate depends non trivially on more than three or four variables at the same time. A demonstration of the viability of this approach has been obtained by members of the network in the search for Higgs decays to b-quark pairs associated to additional, QCD-produced b-quark pairs [16], a signature of relevance for Supersymmetric models. This is exactly the same final state as the one considered in the $hh \to b\bar{b}b\bar{b}$ search we are focusing on here, so the conclusions one may draw from those studies apply here as well.

In the article cited above an adaptive definition of the distance metric was implemented, which optimized it as a function of the parameter space point where the probability estimate was to be evaluated. More details of the procedure are provided in Appendix A A.

Although the nearest-neighbor method remains viable and promising, we decided after some preliminary studies to not implement it for our Higgs pair production searches. The reason for our decision is again that its application requires the availability of large samples of data with a small number of b-tags (e.g. two, with respect to the four b-tags required to the data in the signal region). The trigger selections have become more strict as the LHC instantaneous luminosity grew from Run 1 to Run 2, and the data on which we are performing our searches

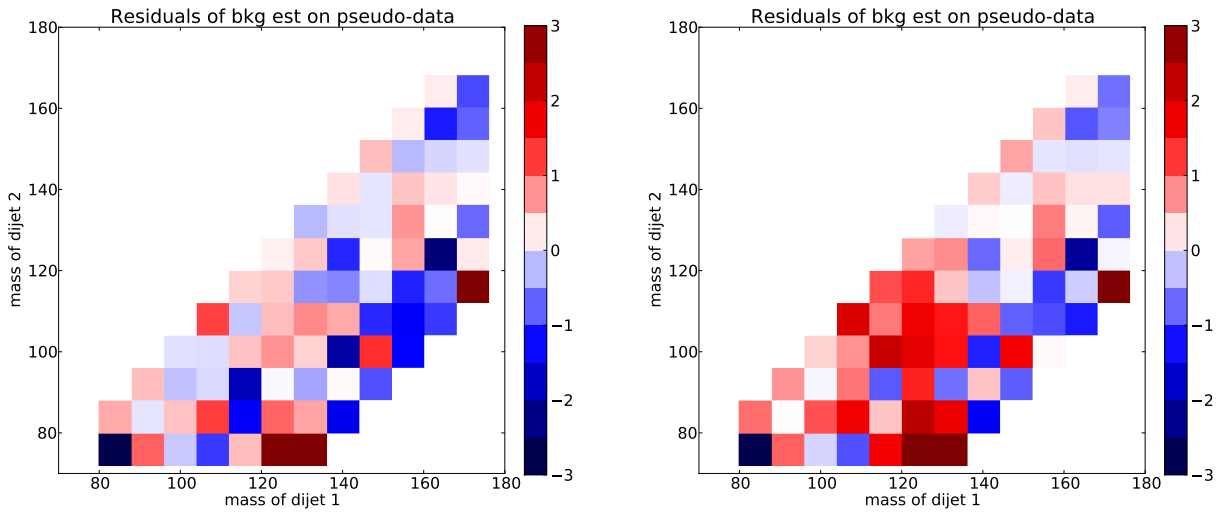


Figure 3: Left: Distribution of the residuals (see text) in the plane constructed with the reconstructed dijet invariant masses, for a background-pure sample. Right: residuals in the same plane, when the used sample includes a significant signal fraction. The accumulation of red and brown bins in the region between 100 and 150 GeV of the two dijet masses indicates that an excess of b-tagged jets is present there, betraying the presence of a hh signal component in the tested dataset.

today must pass tighter level of selection with respect to those collected in 2011 (which were the basis of the above cited article). This caveat applies in particular to the CMS experiment, which at the time of writing has not deployed an improved pixel detector in its tracking, and has therefore a harder task than ATLAS at rejecting events with light-flavoured jets at trigger level. The tighter selection means that the statistical precision of nearest-neighbor estimates of the b-tag probability would be seriously degraded, at least in the case of CMS. We therefore decided to move to an entirely different concept for a background modeling: one not connected with the quantitative estimate of the number of b-tagged jets as a function of the event kinematics, but rather aiming at a modeling of the background density by exploiting the specificities of QCD multijet production. Such a method can offer benefits to both LHC experiments in their searches for Higgs pair production. It is called "hemisphere mixing" and is introduced in the next section.

3 Hemisphere mixing

Event mixing techniques are not a novelty in experimental particle physics: they were used with success in many instances in the past. The setting was in almost all instances that of electron-positron collisions, where both the initial and the final state of the collision is relatively clean, and when the physics of the interaction makes the event more manageable from a purely topological point of view. Event mixing techniques were also used in a recent CMS publication based on work of AMVA4NewPhysics members; there, the context was that of low-energy collisions and the study targeted the "Bose-Einstein correlations" arising between pairs of pions or kaons of low momentum [21, 20]. To our knowledge, no other use of similar ideas have been made in the search for rare phenomena at the LHC; the same studies cited above used the idea of event mixing in a restricted way with respect to what we are going to describe here; the target of the mixing there was the property of charged tracks, which are considerably more simple than hadronic jets as a whole.

The complexity of the physics of high-energy multijet production makes the idea of employing event mixing techniques for the modeling of those processes weird enough that one could be tempted to discard it without a second thought. However, that would be a mistake. One can in fact think of multijet events as the result of a very simple, tree-level $2 \to 2$ parton-parton scattering, made complex by "second order" radiation effects. Taking such a standpoint one may more clearly accept that the kinematics of those two tree-level partons can be made the key to an effective mixing model. If those two partons can be considered separately, then they can also be substituted with others of similar characteristics. The details of the final state produced by each leading parton can thus be modeled independently. Working along this line of thought we have developed a very effective and entirely new modeling algorithm, which we called "hemisphere mixing". This technique, developed within Work Package 4 of the AMVA4NewPhysics network, is the main subject of the remainder of the present document.

3.1 The mixing idea

In the days of Z-pole electron-positron machines, the simplicity of the collisions allowed one to study events by defining a "thrust" direction using particles produced in the final state. The thrust axis is defined as the direction of space which maximizes the sum of particle momenta projections along itself, T:

$$T = \sum_j p_j |\cos \Delta\theta_{Tj}| \tag{3}$$

In passing, we also define a related variable we will use later, T_a:

$$T_a = \sum_j p_j |\sin \Delta\theta_{Tj}| \tag{4}$$

Above, $\Delta\theta_{Tj}$ is the angle that observed particle j makes with the thrust axis. If one studies Drell-Yan production, the thrust axis is a useful seed for a reference system describing the final state. Since the center of mass is at rest in the laboratory system, the event can be divided in two hemispheres, constructed such that the plane dividing them is orthogonal to the thrust axis. For a hadronic event this proves advantageous. Soft colour-reconnection processes do make the two hemispheres "talk" to each other, but this is a low-energy effect and it does not appreciably spoil the above picture.

In hadron-hadron collisions the center of mass is not at rest in the laboratory system, so even a two-body final state does not lend itself to an interpretation in terms of unrelated

hemispheres. A boost to the center-of-mass system would allow to employ that schematization, but that approach fails due to the limited acceptance in rapidity of relevant detector components (in particular the tracker, which we rely upon for b-tagging, one of the most powerful weapon at our disposal in the selection of a signal-rich sample of data). In presence of a reduction of the phase space, the boost creates non linearities in the resulting distributions which make any modeling arduous. What one can do is to abandon the idea of a three-dimensional thrust axis, and use instead a two-dimensional one, constructed in the transverse plane. Then, one can still talk of "hemispheres", provided that it is understood that these are in reality hemi-cylinders, with the beam line as the axis of the cylinder.

The question is whether the schematization of a multijet event as the incoherent sum of two hemispheres as defined above can be of any help in modeling hadron-hadron collisions. Let us push this idea a bit further. We may consider an energetic QCD proton-proton collision as it develops. At tree level, two quarks or gluons of high transverse momentum are emitted in the final state, almost perfectly back-to-back in azimuth. Forward and backward evolution add complexity to the picture, by including final-state radiation (FSR) and initial-state radiation (ISR) effects of progressively lower Q^2. What eventually remains of the two final state objects is a pair of hadronic jets, accompanied by many others which may compete with or even surpass the transverse momenta of the original pair. Additional parton-parton interactions in the same proton-proton reaction provide further confusion, but have usually a smaller Q^2; the same happens with pile-up from other proton-proton collisions. It remains to be seen to what extent do sub-leading processes modify the tree-level picture, spoiling the usefulness of the two-hemisphere idealization.

Rather than trying to answer that question directly, in this document we take a different approach: we assume a total absence of correlations in the kinematics of jet emissions in the two hemispheres, and proceed to exploit that assumption to model QCD multijet events kinematics by sewing together hemispheres observed in different events. A comparison of the model with the original data can then highlight the extent to which the method can be relied upon for specific applications.

3.2 The library of hemispheres

We take a dataset of QCD multijet events, and for each event we compute the thrust axis in the transverse plane, using the identified hadronic jets of the event. Once we know the thrust axis, we take at random one verse as a reference: this defines a two-dimensional unit vector $\vec{t}$ in the transverse plane. At this point we can divide the jet list in two, depending on the sign of the cosine of the angle between the jet azimuth and the thrust versor,

$$S = sgn(\cos(\phi_j - \phi_t)). \tag{5}$$

We store in a library separately the two collections: each determines one hemisphere; hence from a dataset of N events we obtain a library of $2N$ hemispheres. Each hemisphere can be characterized by the number of jets it contains, the number of b-tagged jets, the magnitude of the thrust contributed by the jets. Other possible kinematical variables can also be constructed, like the total invariant mass of the jets, etcetera. A list of some important kinematical variables for this study is provided in Sec. 4.5. A graphical description of the working of the algorithm is given in Fig. 4.

3.3 Mixing hemispheres

Let us go back to the original dataset of N events. Each event can be described by the variables related to the two original hemispheres that make it up: thrust of the jets contained in the

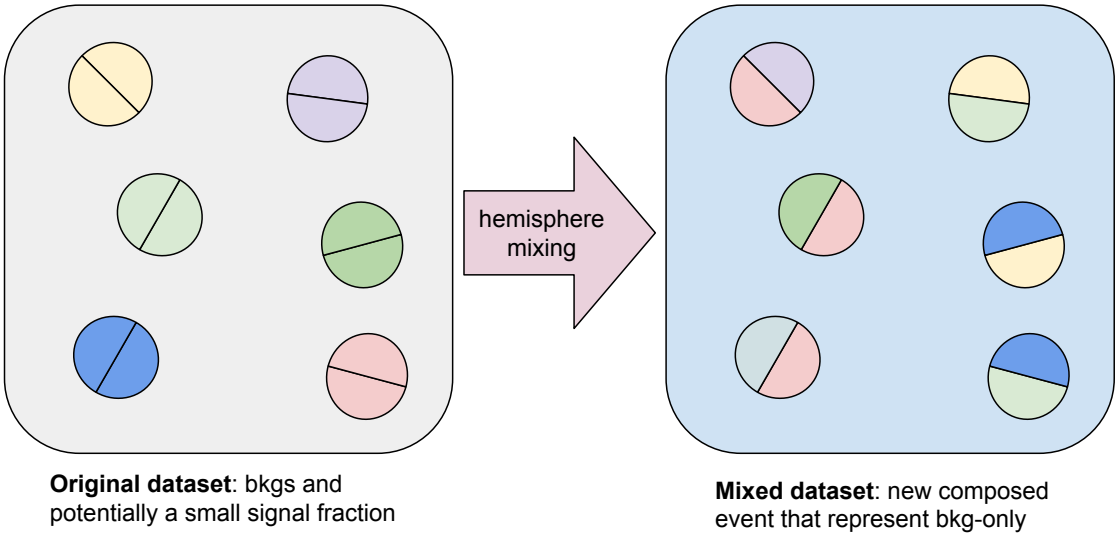


Figure 4: Schematic description of the algorithm concept.

hemisphere (T), number of jets (N_j), number of b-tags (N_t), combined mass of the jets (M), the variable called T_a (see above), and the sum of the jets p_z components. We may label them as $h_1(N_j, N_t, T, M, T_a, P_z)$ and $h_2(N_j, N_t, T, M, T_a, P_z)$. We now look in the library for two hemispheres h_p and h_q that are *similar* to h_1 and h_2, in the sense that they have the same value of N_j and N_t, and have small distance $D(1,p)$, $D(2,q)$ in the feature space spanned by the additional continuous variables T, M, T_a, P_z:

$$D(1,p)^2 = \frac{(T(h_1) - T(h_p))^2}{V_T} + \frac{(M(h_1) - M(h_p))^2}{V_M} + \frac{(T_a(h_1) - T_a(h_p))^2}{V_{T_a}} + \frac{(|P_z(h_1)| - |P_z(h_p)|)^2}{V_{P_z}} \tag{6}$$

$$D(2,q)^2 = \frac{(T(h_2) - T(h_q))^2}{V_T} + \frac{(M(h_2) - M(h_q))^2}{V_M} + \frac{(T_a(h_2) - T_a(h_q))^2}{V_{T_a}} + \frac{(|P_z(h_2)| - |P_z(h_q)|)^2}{V_{P_z}} \tag{7}$$

Above, V_T, V_M, V_{T_a}, and V_{P_z} are the variances of the four considered variables. Of course, we prevent h_p and h_q from being respectively equal to h_1 and h_2, in order to avoid reconstituting the same event from its two original hemispheres. Also, we match the sum of longitudinal jet momenta P_z by considering its absolute value (thus assuming, as it is safe to do in the case of ATLAS and CMS, that the detector acceptance is forward-backward symmetric), and invert jet pseudo-rapidities as needed in the matched hemisphere when we recombine two hemispheres to form the modeled event. The hemispheres are rotated such that they match the original thrust axis of the event to be modeled. Figure 5 gives a graphical description of the technique.

A naive approach to implement the previous procedure might be to loop over all hemispheres in the library for each event in the original dataset and find the hemispheres that better match those in the event. However, the computing time to apply the procedure to the whole dataset of N events this way would be $O(N^2)$. That means it is not well suited to large datasets. The matching of hemispheres is a multi-dimensional nearest-neighbor search, so it can be sped up by using clever data structures. Since we only want to match hemispheres having the same number of jets N_j and the same number of b-tags N_t, we start by partitioning hemispheres into independent subsets based on their value of N_j and N_t. For each subset, a k-d tree [10] is created, such that it partitions the multi-dimensional space in such a way that the best

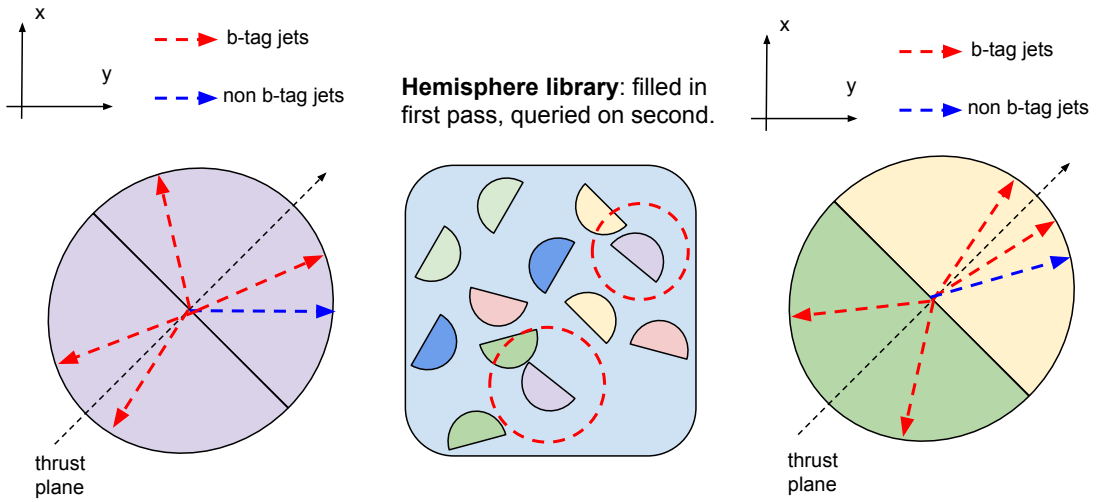


Figure 5: Graphical description of the creation of a mixed event from two hemispheres picked from the library.

matching hemispheres (i.e. the second-nearest neighbors) for each given hemisphere can be found in $O(\log N)$ comparison steps.

Through the above procedure we may therefore create an entirely new dataset, also constituted by N "events". These pseudo-events –we shall call them "hemisphere-mixed events" in the following– are constructed with hemispheres of real events, but not necessarily all of them, as individual hemispheres could be used multiple times, and others could not be used: if we were to create a new library of hemispheres from the hemisphere-mixed dataset we would not in general fully reproduce the original hemisphere library.

The alert reader might have noticed that the mixing procedure "breaks" any correlations that may exist between two halves of every event, beyond the rough agreement between the magnitude of the transverse thrust (which is indeed approximately preserved by the matching procedure). This is no doubt going to have some repercussions on the kinematic distributions we can construct with the events under study. In Sec. 5 below we study whether there are kinematic variables capable of distinguishing original and hemisphere-mixed events; yet for now, we work under the hypothesis that the modeling does indeed not wreak havoc in the kinematical distributions of QCD events.

One thing should already be pointed out here. If the original dataset contains a small fraction (say, a few percents) of a physical process different from the $2 \to 2$ QCD events we have been considering, then the mixing procedure is expected to smear out the difference between that contamination and the background, such that the final result is still unaffected by it. The reason of that effect is that the probability that a rare signal event gets replaced, by the mixing procedure, by two hemispheres taken from the same process scales with the square of the sample fraction. The actual probability, of course, depends not only on the signal fraction, but also on how "recognizable" are the hemispheres of the contaminating signal, according to the metric defined by the multi-variate distance D defined above. As long as D is constructed with variables that do not discriminate the signal from the QCD too much, any small signal will not affect the results of the mixing procedure, which will end up producing a hemisphere-mixed sample showing characteristics much more similar to the dominant process.

4 Datasets

The simulated samples used for this study were produced within the AMVA4NewPhysics ITN in order to carry out several studies on the applicability of multi-variate statistical tools to hh data analyses. In particular, in this analysis we will deal with signal sample corresponding to the $hh \to b\bar{b}\,b\bar{b}$ final-state and its main background component which is QCD multi-jet (mainly $b\bar{b}\,b\bar{b}$). In this section, details regarding the generation and detector-simulation of the datasets for the mentioned processes will be provided together with information about the reconstruction and event selection, which are preliminary steps to the application of the hemisphere mixing method.

4.1 $hh \to b\bar{b}\,b\bar{b}$ signal dataset generation

A total of 10 million events were generated at leading-order (LO), using matrix elements (ME) from MadGraph 5 [6], for the SM process $p\,p \to hh$ at a center-of-mass energy $\sqrt{s} = 13\,\text{TeV}$. The four-flavour scheme was used, with incoming partons being sampled from the `nn23lo1` [9] parton density function (PDF). The parton showering, hadronisation, and decays were handled by Pythia 8 [28]. The decay channels for the Higgs bosons are restricted to $h \to b\bar{b}$, so at the end all events are consistent with the final state of interest.

The following ME requirements were modified from the MadGraph 5.2.4.2 default values (at parton level):

- b-jets: $p_T \geq 20\,\text{GeV}$
- jets and b-jets: $|\eta| \leq 3$
- distance between two jets: $\Delta R_{jj} \geq 0.1$
- distance between two b-jets: $\Delta R_{bb} \geq 0.1$

while the rest where kept according to the default configuration. The leading-order production cross section reported by MadGraph 5 is $14.518 \pm 0.001\,\text{fb}$. Using next-to-next-to-leading-log (NNLL) matched to next-to-next-to-leading-order (NNLO) calculations, and accounting for top-quark mass effects to next-to-leading order (NLO), the LHC Higgs Cross-Section Working Group [18] calculates that the production cross section for $g\,g \to hh$ at $\sqrt{s} = 13\,\text{TeV}$ is $33.5^{+1.8}_{-2.3}\text{fb}$, for $m_h = 125\,\text{GeV}$. Ref. [18] also calculates that the branching ratios for $h \to b\bar{b}$ is $0.5824^{+0.0072}_{-0.0074}$. The total theoretical production cross section is therefore $11.37^{+0.64}_{-0.80}\text{fb}$ for the $hh \to b\bar{b}\,b\bar{b}$ process.

4.2 $pp \to b\bar{b}\,b\bar{b}$ background dataset generation

A total of 10 million events were generater at LO, using MEs from MadGraph 5, for the SM process $p\,p \to b\bar{b}\,b\bar{b}$. The four-flavour scheme was used, with incoming partons being sampled from the `nn23lo1` PDF. The parton showering, hadronisation, and decays were handled by Pythia 8.

The following ME requirements were modified from the MadGraph 5.2.4.2 default values (at parton level):

- b-jets: $p_T \geq 20\,\text{GeV}$
- jets and b-jets: $|\eta| \leq 3$

- distance between two jets: $\Delta R_{jj} \geq 0.1$
- distance between two b-jets: $\Delta R_{bb} \geq 0.1$

while the rest where kept according to the default configuration. The production cross section reported by MadGraph 5 is 1.7471 ± 0.0001 nb. Applying a LO→NLO rescaling factor calculated in Ref. [6] of $1.73^{+1.13}_{-0.73}$, the theoretical cross section for $pp \to b\bar{b}b\bar{b}$ is $3.0^{+2.0}_{-1.3}$nb.

4.3 Detector simulation

The simulation applied to both signal and background samples using Delphes [17, 25, 23] was configured to produce a response in between the ATLAS [1] and CMS [15] detectors. This choice was dictated by the need to allow researchers that belong to the two collaborations to profit equally from these studies. A middle-ground between CMS and ATLAS also allows results to be obtained which are not too dependent on experimental detail.

Delphes uses parametrised responses to allow the quick simulation of a real detector-environment by reconstructing final-state objects with given efficiencies, applying resolution effects, and simulating pile-up contributions. Whilst it is not expected to provide a simulation as accurate as that of Geant 4 [5, 4], it is expected to be sufficiently accurate to validate the proofs-of-concepts addressed, and the comparisons made, in this document.

4.4 Reconstruction and event selection

After simulation, events are reconstructed using the default algorithms provided by Delphes, resulting in a set of reconstructed objects (i.e. data structures) of different types: electrons, muons, taus, jets, photons and MET. However, only hadronic jets are relevant for the final-state considered in this analysis, which are clusters of detector signals produced by the hadronization products of energetic final-state partons. The reconstruction software used to reconstruct jets is the anti-k_T [12] algorithm with a radius parameter $R = 0.5$; the Delphes package includes the Fastjet [13] implementation of the algorithm.

Another relevant step in event reconstruction is b-tagging, which is the identification of jets from the decay of B-hadrons. In an actual proton collider detector, complex algorithms are used to take advantage of the fact that the lifetime of B-hadrons is long enough to produce displaced vertexes and tracks and achieve high identification efficiencies. Fast simulators such as Delphes cannot model this phenomenon in detail and therefore a simple b-tagging identification efficiency parametrization is used as a function of jet flavour, which should be of the same order of those produced by the CMS and ATLAS detectors.

The ultimate goal of the hemisphere mixing algorithm is to accurately modeling the background of the $hh \to b\bar{b}b\bar{b}$ and similar analyses, using real samples of data. Therefore, we have to impose realistic analysis cuts over the simulated dataset such that the results are comparable with those which could be obtained when applying this technique to data. Only the events that pass all the following requirements are selected and used in this study:

- At least four with $p_T > 30$ GeV and $|\eta| < 2.5$ have to be present in the event
- Of those jets, at least four have to be b-tagged

the rest of the events are filtered out.

4.5 Kinematic variables

Here we give a description of the kinematic variables that are used in this study to characterize the background and signal datasets. Events with a multi-jet topology can be univocally described by specifying the number of jets and the four-momentum of each of them; however, higher-level variables constructed with them can better discriminate the $hh \to b\bar{b}b\bar{b}$ decay process from QCD backgrounds. Below is given a list of the variables explicitly cited in the remainder of this document:

- H_T is the sum of the transverse momenta p_T of all observed jets in the event. It is a variable that discriminates the high-energy production of central jets, which can be originated by a massive particle decay, from the production of jets at smaller angles from the beams direction, which is more characteristic of QCD processes.
- M_{jj}^{lead} is the leading dijet invariant mass. This variable is computed by a procedure described in detail below.
- M_{jj}^{trail} is the trailing dijet invariant mass. This variable is also described below.
- p_T^1 is the transverse momentum of the highest-p_T jet of the event.
- $\Delta\phi_{12}$ is the azimuthal angle (computed, i.e., in the plane transverse to tbe beams direction) between the two highest-p_T jets in the event.
- $\Delta\phi_{23}$ is the azimuthal angle between the second and third jet, ordered by p_T.
- $\Delta\phi_{34}$ is the azimuthal angle between the third and the fourth jet, ordered by p_T.
- $\Delta\eta_{23}$ is the pseudorapidity difference between the second and third jet, ordered in p_T.

The procedure which selects the two pairs of jets that should be identified with possible decay products of Higgs bosons is the following. Unlike in the list above, where angular variables are defined, jets are now ordered by decreasing value of the b-tagging discriminant value, such that the first jets are those most likely originating from true b-quark decays. In hh signal events the first four jets in such a list are those corresponding to the decay of the two Higgs bosons, although mismatches and spurious b-tags do make the assignment imperfect. These four jets are then to be paired in order to reconstruct the mass of a hypothetical Higgs boson which produced them. The procedure by means of which the two dijet pairs are identified is the following. We first compute the dijet invariant mass of each of the possible pairs of jets constructed with the four b-tagged jets, obtaining three possible pairings: (M_{12}, M_{34}), (M_{13}, M_{24}), and (M_{14}, M_{23}). The chosen pairing among these is the one which results in the smallest value of the absolute dijet mass difference between the two values in the pair. This allows us to pick a pairing which is likely to be the correct one for jet pairs in hh decay events (for which the two dijet masses should be equal within experimental resolutions) without biasing ourselves toward accepting pairs whose dijet masses are closest to the true Higgs boson mass. The benefit is that we thus retain, in the resulting dijet mass variables (leading and trailing, as defined by their value - leading being the highest-mass one), a high discrimination power between signal and backgrounds. This can be confirmed by observing the distributions shown in the left panels of the middle rows in Fig. 8.

The eight variables listed above are only a subset of the many possible high-level feature variables one may construct with the jets of an event. In fact, since for each jet there are four momentum components to consider, a full characterization of a multi-jet event involves a total of $4N_j$ independent variables, where N_j is the number of jets. Many of the jets in the events

resulting from the selection detailed in Sec. 4.4 have 6, 7 or even more jets, making the number of possible independent features very large. We are thus facing the common problem of a high-dimensional feature space when we wish to perform statistical compatibility tests using these data samples. The solution, discussed in the next section, is to apply permutation techniques to tests that consider subsets of 20 kinematic variables originally identified to describe the most relevant kinematic characteristics of the selected events. Those 20 variables include, in addition to the eight described above, the transverse momenta and pseudorapidities of leading jets, as well as angles between them.

5 Tests of the hemisphere mixing procedure

5.1 Overview of the statistical framework

The main goal of the statistical studies is to check if the hemisphere mixing algorithm performs according to our expectations, i.e. whether it produces artificial observations having the same distribution as the original observed data. To this aim, we have performed both an exploratory analysis and a formal test to evaluate a potentially significant difference between the two distributions.

As a first exploratory step, we compare graphically the empirical distribution of the background data with the empirical distribution of hemisphere-mixed events. Since the two distributions are multidimensional, we consider univariate representations of the eight variables considered as the most relevant for the analysis, i.e. four kinematic features (H_T, transverse momentum, leading and trailing dijet invariant mass) and four angular features (azimuthal angle between jets and pseudo-rapidity difference - see Sec. 4.5).

The univariate visualization of the data is often performed by binning a given variable and drawing its histogram [26]. However, the comparison of two data sets sampled from a similar parent distribution using histograms is problematic due to the usually large variance of the per-bin uncertainty bars. Kernel density estimation [26, 27] is a better suited method for estimation and visualization of variable density. It mimics histograms while allowing for a greater flexibility and smoothness, and can be regarded as a generalization of the histograms. Given a univariate sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, a kernel density estimate is defined as

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i) \tag{8}$$

where K is a probability density function known as a "kernel" and h is a smoothing parameter selected according to some optimality criterion. In the following, Silverman's "rule of thumb" criterion is adopted [27, p.48].

The distribution estimates are presented for the original background observations and the hemisphere-mixed background events in Figs. 6 and 7. While not exactly identical, the distributions exhibit very similar behaviour, thus indicating a first descriptive evidence of the effectiveness of the hemisphere mixing algorithm. Since by eye the presented densities look quite alike we choose to also draw their ratio as a function of their domains. If the two distributions were equal, these densities ratios should oscillate around 0.5 without showing any systematic peaks. This is roughly what we observe in the graphs; the observed disproportion of background versus mixed events observed for the smallest values of the domain of some of the displayed variables is likely to be caused by the known issues of kernel density estimation on the boundaries.

For the sake of comparison of the original distributions, it is useful to also illustrate the difference between background and signal distributions. The plots of marginal distributions of the two different samples are presented in Figs. 8 and 9.

At this point it is important to check the effect of a signal contamination in the base of data used in constructing the library of hemispheres. For this purpose, a different set of hemisphere-mixed events are obtained from a dataset consisting in background observations contaminated with a 10% fraction of signal events. These mixed events are compared with the pure background sample (devoid of any signal contamination) in order to evaluate whether the hemisphere mixing algorithm enjoys the further property of smearing out a possible signal, as discussed in Sec. 3.3. In Figs. 10 and 11 we present the impact of a 10% signal contamination on the distributions and the smearing of its characteristics caused by the hemisphere mixing method. These figures indicate that the algorithm works according to its expectations. While

some of the distributions do show some very slight differences, the dijet mass distributions appear very well modeled. This is relevant as the dijet masses are the variables for which a precise modeling is most important, as in the experimental data analysis they are usually taken as the basis of inference on the presence of signal in the data. While in the hh search we will be dealing with possible signal contaminations well below one percent in selected data, here we voluntarily increase the signal fraction in order to see by eye what effect the mixing procedure has on the signal distributions. A 10% signal contamination would definitely be detectable by eye in the dijet mass region 100-150 GeV due to the distinctive peak in the leading and trailing dijet masses (see Fig. 8, graphs on the right in the second and third row). Figure 10 shows that this is not the case in hemisphere-mixed data, thanks to the smearing of the dijet mass features of the small signal component; this is also demonstrated by the ratio graphs.

The qualitative comparisons discussed above indicate that it is worth exploring the performance of the hemisphere mixing algorithm in more detail. We present below a proper multivariate statistical analysis and a sensitive hypothesis test with that aim.

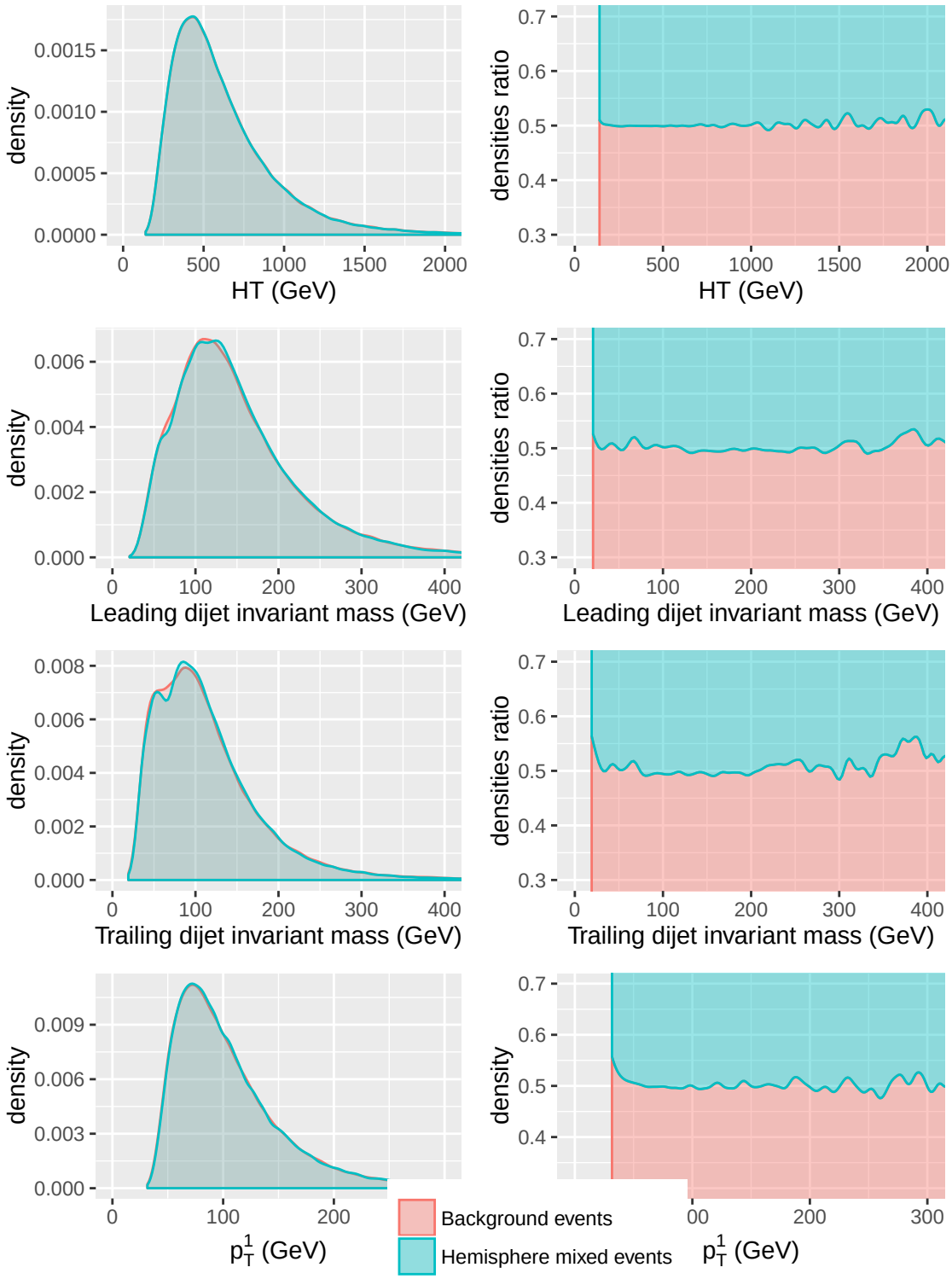


Figure 6: On the left side is shown the kernel density estimate of marginal distributions of pure background data (red) and of the hemisphere-mixed dataset obtained from it (blue); on the right side of each row is shown the ratio of the two densities as a function of their domain. The marginal distributions show four kinematic variables describing the events; see Sec. 4 for their definition. A Gaussian kernel and Silverman's "rule of thumb" method for bandwidth selection [27, p.48] are used.

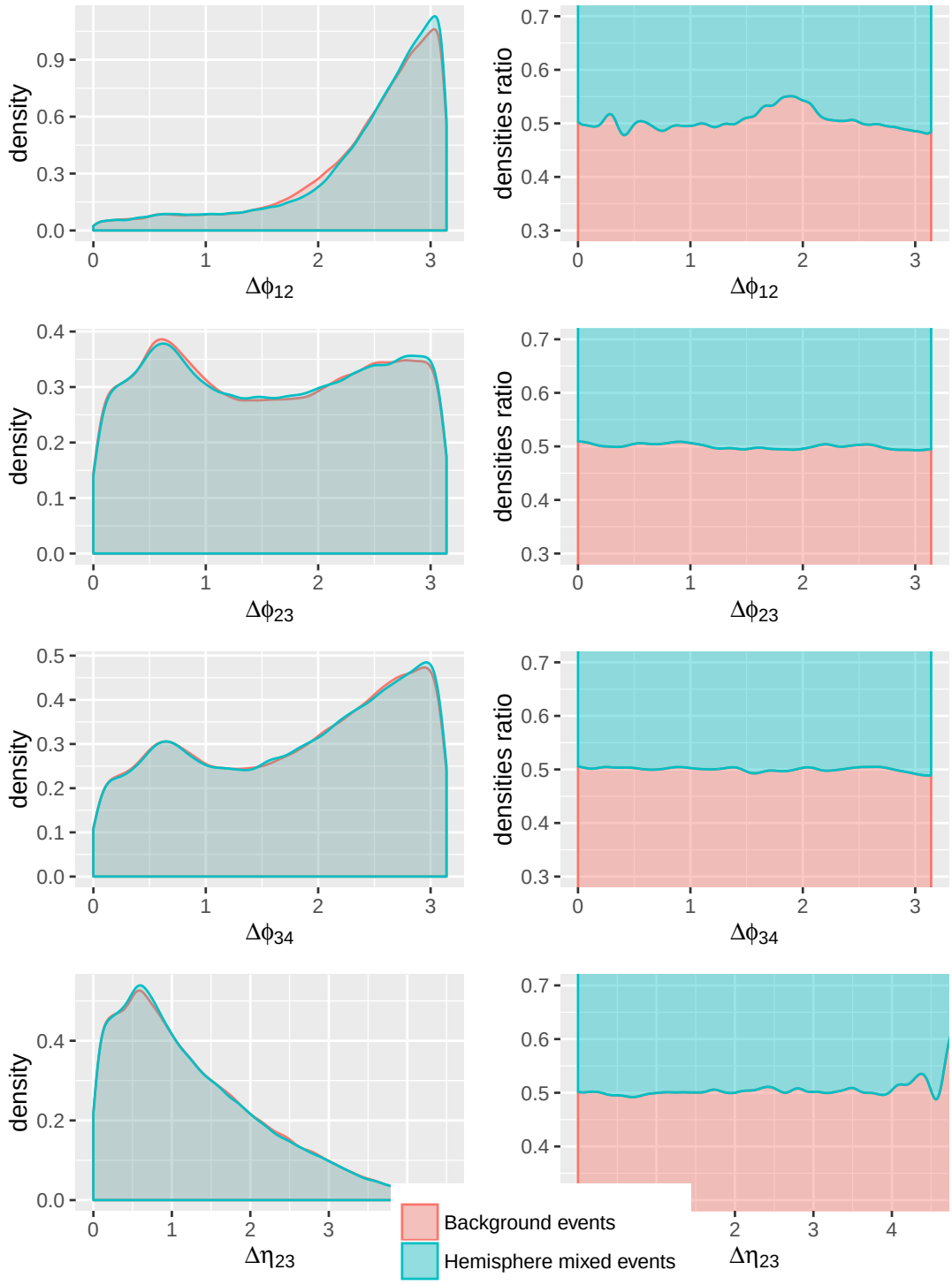


Figure 7: On the left side is shown the kernel density estimate of marginal distributions of pure background data (red) and of the hemisphere-mixed dataset obtained from it (blue); on the right side of each row is shown the ratio of the two densities as a function of their domain. The marginal distributions show four kinematic variables describing the events; see Sec. 4 for their definition.

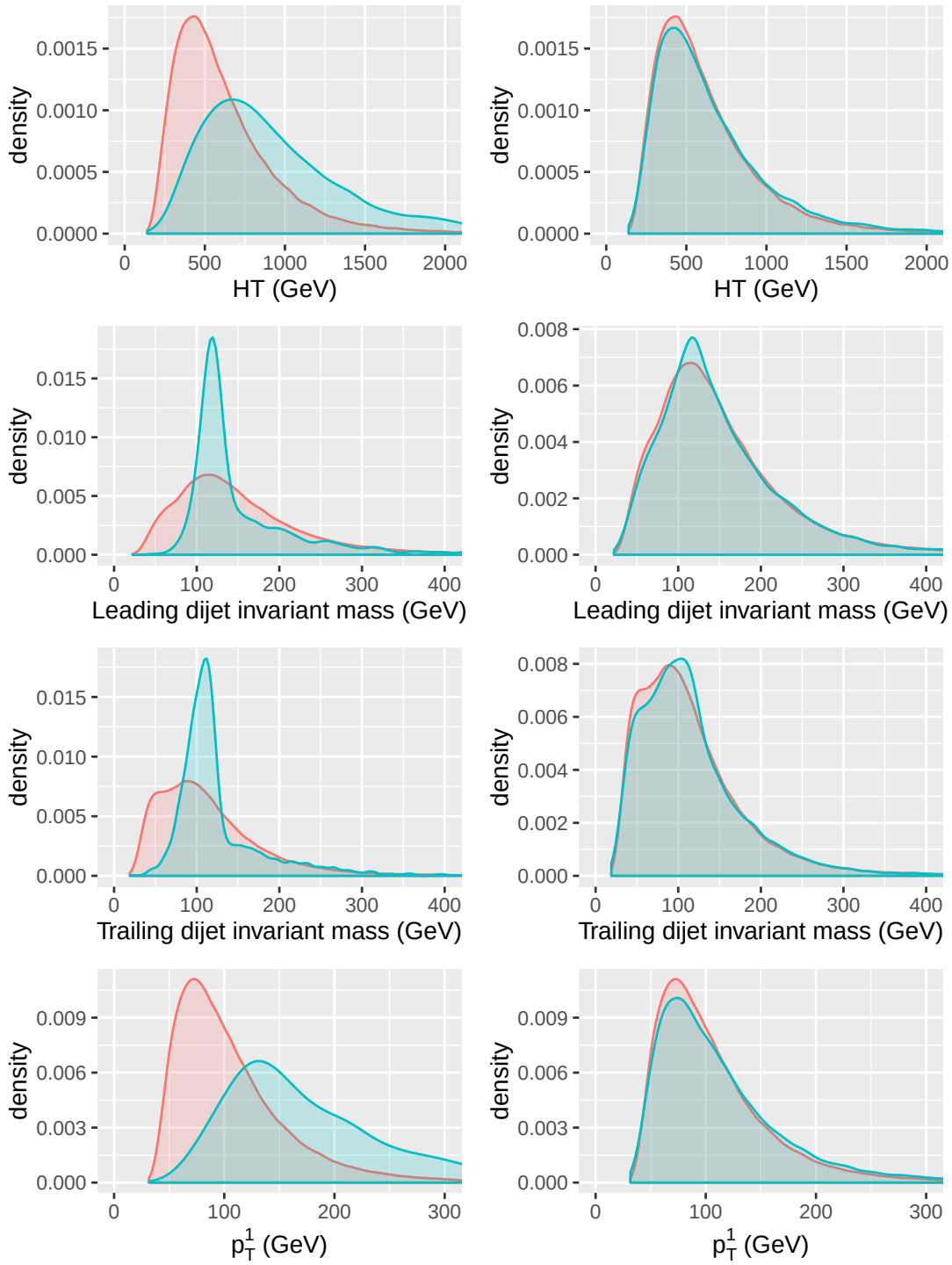


Figure 8: On the left side are compared the distributions of four kinematic variables for background (red) and signal (green). On the right a mixture of 90% *background and* 10% *signal (green) is compared to the background alone (red). From the graphs it is evident that a* 10% *contamination is very well distinguishable in these distributions, in particular those of the two-body invariant masses.*

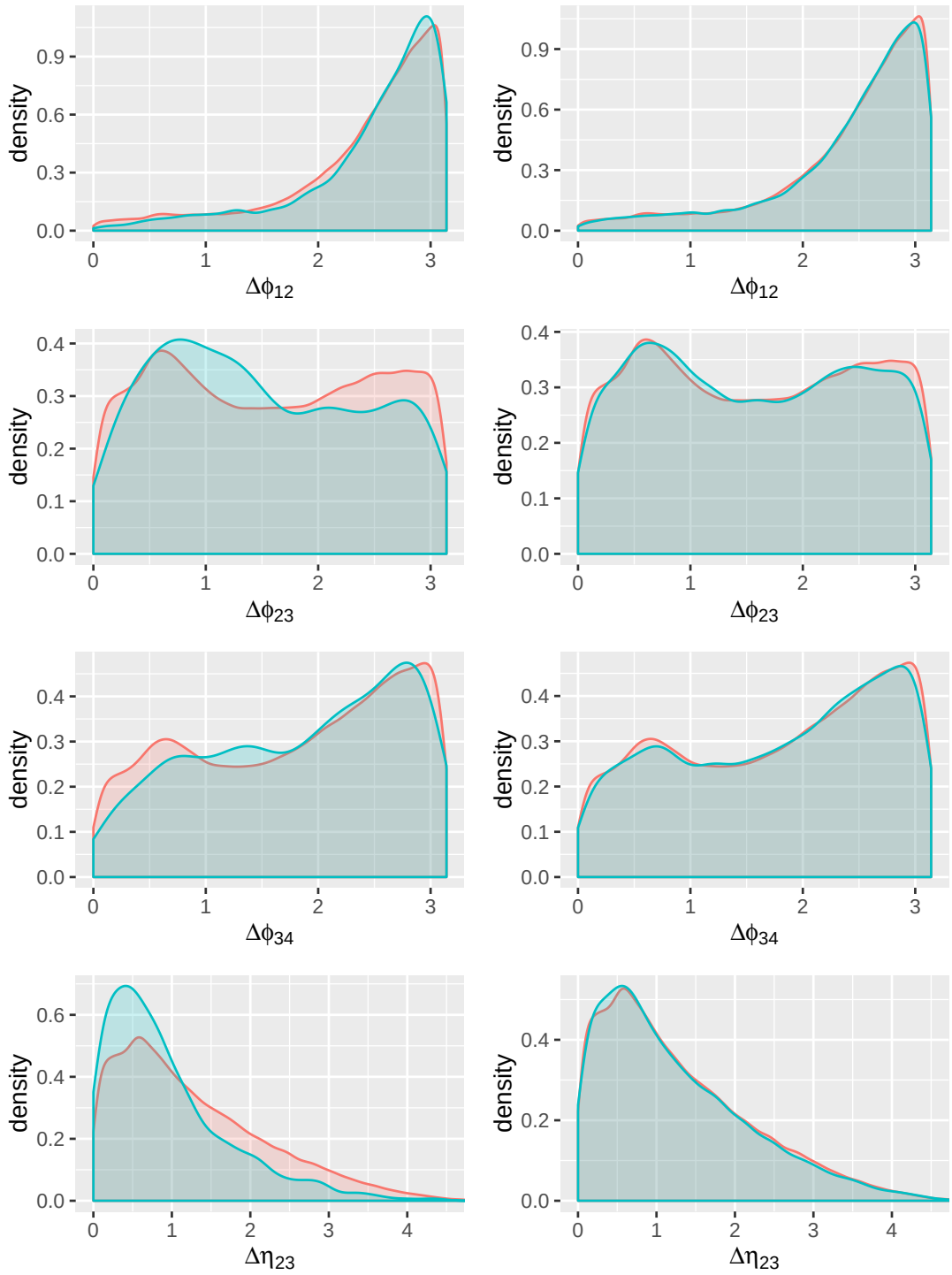


Figure 9: On the left side are compared the marginal distributions of four angular variables for background (red) and signal (green). On the right is shown the comparison of the marginals for a mixture of 90% *background and* 10% *signal (green) with those of background alone (red).*

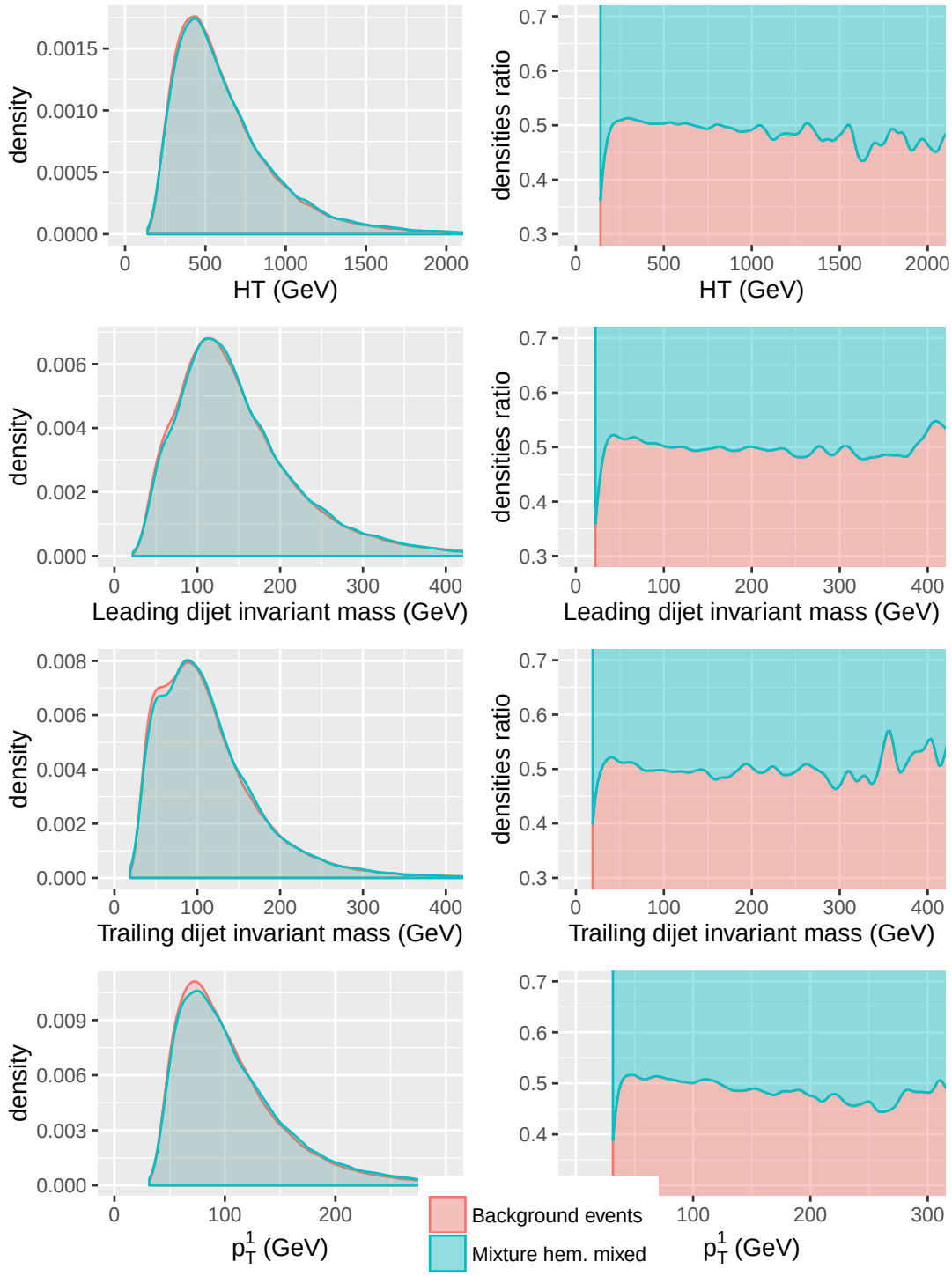


Figure 10: Left: comparison of the distributions of four kinematical variables for background alone and the hemisphere-mixed result of a sample constituted by 10% *signal and* 90% *background. The hemisphere mixing procedure makes the signal contamination almost invisible even when this is quite large; compare with the graphs on the right panels of Fig. 8. Right: ratio graphs of the distributions shown on the left.*

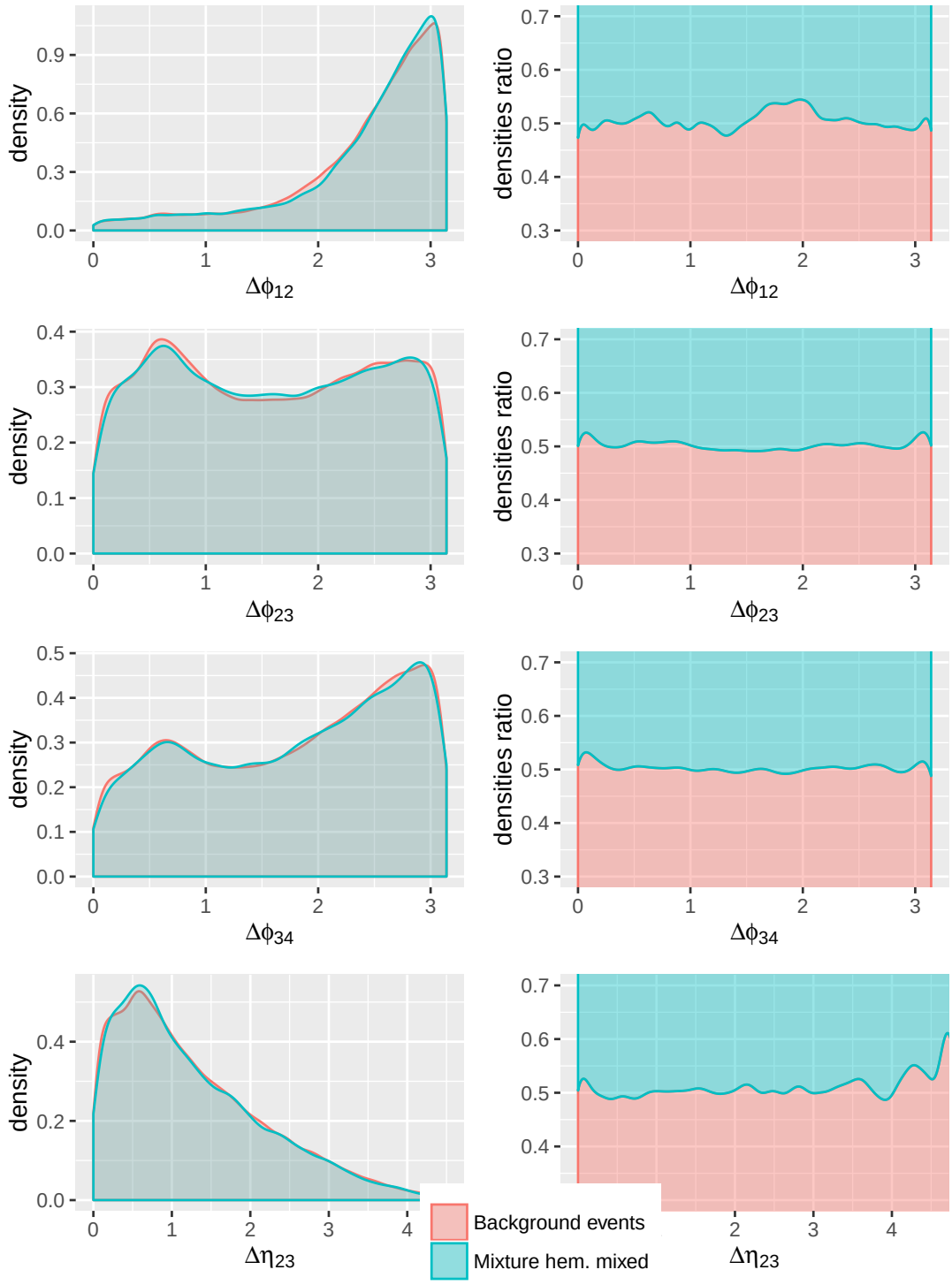


Figure 11: Left: comparison of the distributions of four angular kinematical variables for background alone and hemisphere-mixed result of a sample constituted by 10% *signal and* 90% *background. Compare with the graphs on the right panels of Fig. 9. Right: ratio graphs of the distributions shown on the left.*

5.1.1 Inferential analysis: a permutation-based approach

In order to confirm the empirical evidence supported by the visual exploration of the figures shown above, a formal statistical test should be performed. More specifically, let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{N_2}$ be two random samples from densities f_X and f_Y respectively, $\mathbf{X}_i, \mathbf{Y}_i \in \mathbb{R}^d$. Here the two samples refer to the background data and the hemisphere mixed data produced by the algorithm. The aim is to test the null hypothesis

$$H_0 : \quad f_X(\mathbf{x}) = f_Y(\mathbf{x})$$

for all $\mathbf{x}$ that are in the domain of variables, against the general alternative

$$H_1 : \quad f_X(\mathbf{x}) \neq f_Y(\mathbf{x}).$$

The issue of testing two samples for equal distributions is quite common in statistical inference and many solutions have been proposed [26, 30]. In the following we present the standard statistical tools that could be a common choice for the given problem.

The Kolmogorov-Smirnov test [26] can be used for the considered issue when data are unidimensional. The statistic is computed based on a distance between the empirical cumulative distribution functions of the samples and for this reason it is not restricted to location or scale changes only. This test has several attractive features. Among them is the robustness to outliers, as the statistic is only sensitive to the bulk of density function. On the other hand, this test has usually small power in comparison to others [26].

A more powerful alternative is the Wilcoxon rank sum test [26]. This is a common non-parametric univariate two-sample test, for which the alternative hypothesis is that the two distributions differ by some location shift $\mu \neq 0$ (for the two-sided case). For the considered data this test is not an optimal choice as it is also univariate and tests a different hypothesis (the same location in general does not mean the equality of distributions).

The Multivariate Analysis of Variance (MANOVA) [30] could be a better alternative as it is a multivariate test. However, the test is designed to spot the difference in means and therefore it also does not satisfy the hypothesis that are meant to be tested. Additionally, the assumptions for this test are that the variables have normal marginal distributions. However for a large number of observations the distribution of means is approximately normal (as it follows from the Cental Limit Theorem). For this reason the MANOVA test is robust to non-normal datasets [22].

As described above these standard statistical tests are not proper for our purpose. For this reason we have to identify a more sophisticated method. Duong et al. [19] recently proposed a kernel density-based two-sample test. The test makes no assumptions on the data distribution, it is multivariate and tests the required hypothesis. It relies on a kernel density estimation (KDE) of f_X and f_Y. The densities of the two samples are estimated as a generalization of Eq. 8,

$$\hat{f}_X(\mathbf{x}; H_X) = \frac{1}{N_1} \sum_{i=1}^{N_1} K_{H_X}(\mathbf{x} - \mathbf{X}_i) \quad \text{and} \quad \hat{f}_Y(\mathbf{x}; H_Y) = \frac{1}{N_2} \sum_{i=1}^{N_2} K_{H_Y}(\mathbf{x} - \mathbf{Y}_i)$$

where K is a kernel function and H_X, H_Y are the chosen bandwidth matrices. for $K \in \{X, Y\}$.

The integrated squared error is a measure of discrepancy between the density functions

$$T = \int \left[f_X(\mathbf{x}) - f_Y(\mathbf{x}) \right]^2 \mathrm{d}\mathbf{x}$$

where integration is taken over the appropriate Euclidean space and has been well studied for the optimal selection of smoothing parameters. Note that T could be also written in the following form:

$$T = \psi_{X,X} + \psi_{Y,Y} - \psi_{X,Y} - \psi_{Y,X}$$

where $\psi_{K,L} = \int f_K(\mathbf{x}) f_L(\mathbf{x}) d\mathbf{x}$. Therefore the discrepancy T could be estimated as

$$Z = \hat{\psi}_{X,X} + \hat{\psi}_{Y,Y} - \hat{\psi}_{X,Y} - \hat{\psi}_{Y,X}$$

where

$$\hat{\psi}_{X,X} = \frac{1}{N_1^2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_1} K_{H_X} \left(\mathbf{X}_i - \mathbf{X}_j \right), \quad \hat{\psi}_{Y,Y} = \frac{1}{N_2^2} \sum_{i=1}^{N_2} \sum_{j=1}^{N_2} K_{H_Y} \left(\mathbf{Y}_i - \mathbf{Y}_j \right),$$

$$\hat{\psi}_{X,Y} = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} K_{H_X} \left(\mathbf{X}_i - \mathbf{Y}_j \right), \quad \hat{\psi}_{Y,X} = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} K_{H_Y} \left(\mathbf{Y}_i - \mathbf{X}_j \right).$$

It has been shown that the Z statistic is asymptotically Normal. This property gives the KDE test a significant advantage over other multivariate tests in terms of computational speed. However, a drawback of the test is that the kernel density estimation is highly affected by the curse of dimensionality [7], hence its use is not recommended in dimensions higher than 6 [14].

Within the problem under consideration, the initial number of observed variables one may study is much higher than any dimension which could guarantee accurate results (as mentioned in Sec. 4, 20 relevant variables have been considered on the selected datasets; the eight most interesting among them have been presented in the figures shown *supra*). Therefore the idea is to perform the test on subsets of variables. Let $\mathbb{T}$ be the set of the variables from the data. We take P subsets of $\mathbb{T}$ and denote them as $\mathbb{T}_1, ..., \mathbb{T}_P$. For each $\mathbb{T}_k$ the statistical test is performed, a test statistic Z_k obtained and its respective p-value p_k, for $k = 1, \ldots, P$. Given the vector of p-values, a solution for combining them is required to evaluate the significance of the test.

Methods of inference for the combination of multiple p-values have been well described in the statistical literature [11]. A function known as a combinant is computed based on the obtained vector of p-values. The combinant is often chosen such that its distribution is known provided that some assumptions are met. Based on the known combinant distribution and the obtained combinant value, a single combined p-value is obtained, which allows us to decide whether the null hypothesis should or should not be rejected. In this report we consider two combinants: the Fisher combinant, which is defined as $p^F = -\sum_{k=1}^{P} \log(p_k)$; and the min-p, which is defined as $p^M = -\min_{k=1,2,...,P} p_k$.

One important point to make is that the distributions for the combinants are only known if the p-values obtained in the multiple tests are independent. Unfortunately, for our case this assumption is not met as the subsets $\mathbb{T}_k$ could have non-null intersections; in addition, the mutual dependence among the variables might cause the sets $\mathbb{T}_k$ to be dependent. One way to overcome this problem and turn out with a distribution of the combinants under the null hypothesis is to resort to a permutation framework [24]. Given $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{N_2}$, a new pair of samples is obtained by randomly exchanging observations from X to Y and viceversa. This would guarantee that the distribution of the permuted samples, whatever they are, are identical, i.e. that we are under the null hypothesis H_0. Then, a vector of test statistics can be computed from the permuted data. This operation is replicated B times and the following table is obtained.

$$\begin{array}{lccccc} \text{Variables} & \mathbb{T}_1 & \dots & \mathbb{T}_k & \dots & \mathbb{T}_P \\ \text{Original samples} & Z_{11} & \dots & Z_{1k} & \dots & Z_{1P} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{Permuted samples b} & Z_{b1} & \dots & Z_{bk} & \dots & Z_{bP} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{Permuted samples B} & Z_{(B+1)1} & \dots & Z_{(B+1)k} & \dots & Z_{(B+1)P} \end{array}$$

To combine the results, for each test statistic Z_{bk} the p-value is computed by columns as

$$p_{bk} = \frac{\sum_{l=1}^{B+1} \mathbb{1}\{Z_{bk} \leq Z_{lk}\}}{B+1}, \tag{9}$$

where $\mathbb{1}\{\cdot\}$ is the identity function. In this way the analogous matrix of p-values is constructed. In this matrix the particular combinants are computed by rows. Note that we obtain $B+1$ combined p-values each for a single permutation of samples as presented below.

$$\begin{array}{ccccccc} \mathbb{T}_1 & \dots & \mathbb{T}_k & \dots & \mathbb{T}_P & & \\ p_{11} & \dots & p_{1k} & \dots & p_{1P} & \to & p_1^F \\ \vdots & \ddots & \vdots & \ddots & \vdots & & \vdots \\ p_{b1} & \dots & p_{bk} & \dots & p_{bP} & \to & p_b^F \\ \vdots & \ddots & \vdots & \ddots & \vdots & & \vdots \\ p_{(B+1)1} & \dots & p_{(B+1)k} & \dots & p_{(B+1)P} & \to & p_{B+1}^F \end{array}$$

Given B combined p-values from tests on permuted samples, we obtain the empirical distribution of the combinant values under the null hypothesis. This distribution is used to obtain the final p-value for the considered permutation framework for the original sample combined p-value. The final p-value is given as a percentile rank of the original combinant values across the all obtained. In other words, the final p-value for the Fisher combinant in permutation framework is

$$p^F = \frac{\sum_{b=1}^{B+1} \mathbb{1}\{p_b^F \geq p_1^F\}}{B+1}. \tag{10}$$

For the min-p the final p-value is respectively given by

$$p^M = \frac{\sum_{b=1}^{B+1} \mathbb{1}\{p_b^M \geq p_1^M\}}{B+1}. \tag{11}$$

5.2 Performance of the statistical test

The aim of this section is to validate the KDE-permutation test described in Sec. 5.1.1 in terms of first type error and its power. Beyond the KDE-permutation test, for the sake of comparison we consider also the MANOVA test, as well as a Wilcoxon test included in a permutation setting based on the combination of univariate results to allow its use with multidimensional data.

The permutation framework of the KDE test has been adjusted to strike a compromise between power and computation time. The tested samples have size 2000 each and the test is performed in the three-dimensional space of the three selected variables. We restrict ourselves to three-dimensional subspaces because for higher dimensions the density estimation would not be accurate enough given the size of the samples; on the other hand, an increase of the size increases the computation time quadratically. We therefore take $P = 40$ sets of three variables

that span the space in which the consecutive tests are performed. The selection of sets is taken at random from all the possible choices of picking 3 out of 20 considered variables. The number of sample permutations B is set to be 400. The resulting CDF is presented in Fig. 12.

5.2.1 Type-1-error analysis

The accuracy of a statistical test can be defined as its resistance to incorrectly rejecting the null hypothesis at the nominal level α - the significance level or type-I error rate. Since the null hypothesis is rejected when p-value is lower than α, an accurate test produces under H_0 p-values that are uniformly distributed.

We are interested in verifying the accuracy of the KDE test in the permutation framework, i.e. checking if its p-values for random subsamples under the null are uniformly distributed. To do so we sample, without replacement, under the null hypothesis observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{N_2}$ from the sample of the background data, so that both sets are sampled from the common density. Given the two generated samples we perform the test and obtain its respective p-value for the tested hypothesis. This procedure is repeated R times in order to obtain an empirical cumulative distribution function (CDF) of p-values. This empirical CDF is compared with the uniform CDF to validate the accuracy (Fig. 12).

The accuracy is also computed for the Wilcoxon test in the described permutation framework and for the MANOVA test. Within the permutation framework, Wilcoxon tests are performed consequently on all $P = 20$ variables taken one at a time because the test is uniform. The samples are permuted also $B = 400$ times. For the MANOVA test the size of samples was increased to $n = 20000$ so that the asymptotic distribution of the test statistic would be reached.

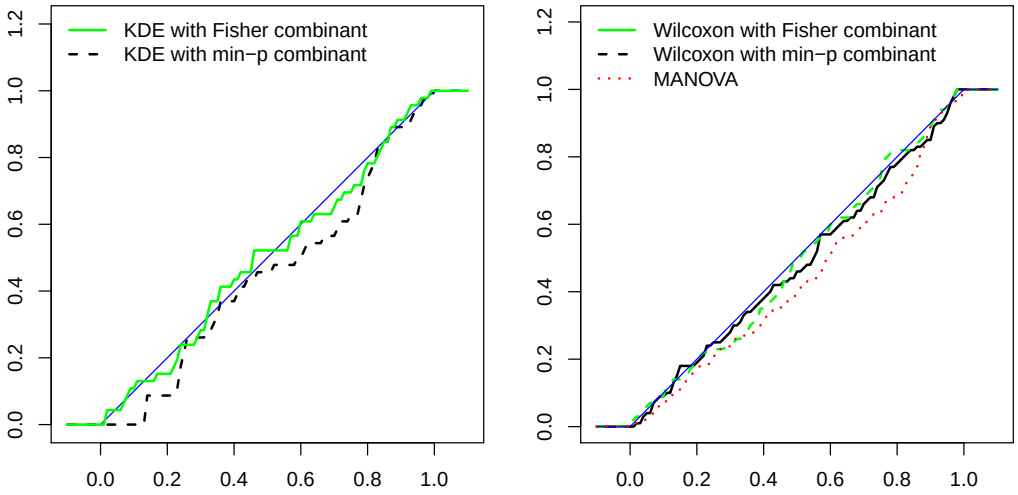


Figure 12: The empirical cumulative distribution function of p-values for the three considered tests under H_0 (on the left for KDE test and on the right for Wilcoxon and the MANOVA). The number of sampling is R equal to 46 for KDE test, 100 for Wilcoxon tests and 1000 for MANOVA. For the tests in permutation framework p-vales are computed for Fisher (green solid line) and min-p (black dashed) combinants. The CDF for MANOVA is shown by a red dotted line. The blue line on both plots is the uniform CDF.

The empirical CDF for the accurate test should be close to the uniform CDF. As the significance level α is usually a small value, both CDFs should overlap particularly well for the range of values in (0, 0.1). In the studied case the min-p combinant for the KDE test is too conservative, as seen by its CDF lying below the diagonal. The MANOVA test turns out to be conservative as well but rather for higher significance levels α. The other CDFs reflect the expected behaviour.

Given the chosen significance level $\alpha = 0.05$, the KDE test with the permutation framework for the Fisher combination would reject the null hypothesis for the tested 46 subsamples in 4.3% of cases; the MANOVA would reject the null in 4.5% of cases; and the Wilcoxon test with Fisher combinant in 6% of cases. Given the above, the KDE test with the Fisher combinant can be considered an accurate test. The same could be said about the Wilcoxon test in the permutation framework. The MANOVA test turned out to be conservative for the given data. In conclusion, the selected tests are accurate and proper for testing the stated hypothesis.

5.2.2 Power analysis

In order to analyze the performance of the studied algorithm we need to employ a statistical test that not only controls the first-type error but also offers a small second-type error probability, i.e. one that with high probability correctly rejects the null hypothesis when it is indeed false. The rate of type-II error (β) is equivalently determined by the *power* of the statistical test, which is defined as $1 - \beta$. Hence, a powerful test is one that with a high probability accepts the alternative hypothesis when a departure from the null hypothesis is present. This should be studied for the KDE-permutation test: in other words, we need to evaluate how often we are capable of rejecting the null hypothesis when the tested samples are drawn from different distributions. This is in general an ill-posed question, as we are not specifying the alternative hypothesis; however, we can take a simplified ansatz in the following.

In the considered framework, the more "separated" are the tested distributions the easier is to reject the null, hence the power of our KDE test can be measured as a function of the signal contamination in the samples. We compute it for a sequence of different signal fractions in a background sample.

To be more specific, consider two d-variate samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N_1}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{N_2}$, in which the first one is taken purely from the background dataset while the second one consists in $s\%$ of signal events and $100 - s\%$ of background events. In contrast to the previously described null distribution analysis, these two samples are indeed taken from different distributions. The difference between them increases as the signal fraction s increases. The p-value of such a test for each studied signal fraction s is shown in Table 1. The numbers indicate that the test has power for signal fractions of 5% and above. A larger size of data tested would cause the power to increase; unfortunately, CPU time constraints prevented us from studying those cases in more detail at the time of writing.

Significance level α	$s = 1\%$	$s = 5\%$	$s = 10\%$
0.01	0.013	0.018	0.038
0.05	0.050	0.118	0.175
0.10	0.138	0.200	0.275

Table 1: The fraction of cases for which the null hypothesis was rejected for significance levels α equal to 0.01, 0.05 *and* 0.10. *80 pairs of samples were generated under the alternative hypothesis for each background contaminated data with values of signal fraction s equal in turn to* 1%, 5% *and* 10%.

5.3 Tests of the hemisphere mixing performance

Since the accuracy and, to some extent, the power of the KDE test in the permutation framework with the Fisher combination function have been found to be appropriate, the test can be applied to the dataset with hemisphere mixed events. In contrast to previous approaches we do not have to perform many tests on different subsamples in order to analyze the distribution of its p-values. Rather a single test on the bigger samples would allow us to draw inference on the stated hypothesis, given the significance level $\alpha = 0.05$.

The hemisphere mixing algorithm is expected to produce artificial data with the same distribution as the original data. In order to validate this expectation, a test comparing a background-only sample with a hemisphere-mixed background sample is performed. The resulting p-value for this test is presented in the first row of Table 2. Its value shows that there is no evidence against the null hypothesis at any reasonable significance level α. In other words, based on the KDE-permutation test results, the pure background and hemisphere-mixed background are consistent with being sampled from the same distributions.

Sample tested against pure background	Obtained p-value
Hemisphere-mixed background events	0.224
Hemisphere-mixed data from mixture of 95% background and 5% signal	0.284
Hemisphere-mixed data from mixture of 90% background and 10% signal	0.005

Table 2: p-values for the KDE tests in the permutation framework describing the distribution equality of two samples, where the first sample is pure background and the second is a hemisphere-mixed sample of pure background (top row) and of background contaminated with a 5% *or* 10% *signal fraction (middle and bottom row). The tests are performed on samples of size 15000.*

A further desirable feature of the artificial data produced by the hemisphere mixing algorithm is that it is capable of smearing out the kinematical features of a possible signal component present in the data. This would be achieved if the hemisphere mixing of a data sample containing background events contaminated by a small signal component produced a dataset whose features distribute as a pure background sample. To test this property, we produce a hemisphere-mixed sample starting from a mixture of 95% background events and 5% signal events. A 5% contamination is absolutely off-scale in the case of the search for the tiny $hh \to b\bar{b}b\bar{b}$ signal predicted by the Standard Model in LHC data, so this test is meant to try and see where the background modeling "breaks down". The new sample distribution is tested against the pure background distribution; the resulting p-value of the test is given in the second row of Table 2. This also shows no evidence against the null hypothesis at the considered significance level α. The hemisphere mixing algorithm is thus seen to perform according to the expectations, for sample sizes typical of the LHC searches we aim at. We then test a 10% signal contamination, and verify that in that case the test does reject the null hypothesis at the chosen α level. We thus verify that a breaking point of the method is reached for very large signal contaminations. This further highlights the fact that for the signal contaminations of background-rich control samples typically used in the considered searches, which are usually smaller than a percent, the algorithm produces a very good modeling.

In conclusion, at a significance level $\alpha = 0.05$ we do not reject the null hypothesis of distributions equality in both tests discussed above. In other words, we see no evidence that the hemisphere mixed events have different distributions from those of a background-pure sample, both in the case of an application to a background-pure dataset, and in the case of a dataset which contain an up to 5% signal contamination. For this reason the algorithm may be successfully applied to the problem for which it was designed.

6 Conclusions

In this document we describe a technique which proves capable of modeling a data set predominantly constituted by high-p_T QCD production processes in a fully data-driven way. We show that the main kinematic features of the data are preserved by the proposed modeling algorithm, and that specific characteristics such as dijet masses and angles can be relied upon in the mixed model. The method is especially useful for the search of a resonant signal such as the one arising from $hh \to b\bar{b}b\bar{b}$ events because it is shown that the presence of such a minority component of the data used in the model does not affect the modeling of the background, which is driven entirely by the majority component. In particular, the Higgs boson "peak" in the dijet mass distribution of the signal component contaminating the data used in the model gets washed away by the recombination procedure. This is especially useful as one may then use the model as background-only template in a fit to data selected in the corner of phase space where the signal component is not negligible, as it remains insensitive of the contamination.

A Details of the Nearest Neighbor implementation

In this appendix we provide some detail on the algorithm mentioned in Sec. 2.2, which originates from the method called "k-nn" in professional statistics literature. Its previous use in high-energy physics research is very limited to our knowledge; one example is its adoption in a 2003 Tevatron study [8] where the problem was the one of regressing the dijet mass resolution of b-jet pairs for the Higgs boson search.

In the 2011 $b\bar{b}H$ search the scalar field to be estimated was, like in the matrix method discussed in the Introduction, the b-tag probability of hadronic jets. Differently from regression problems such as the one of the 2003 study, this quantity cannot be estimated on an event-by-event basis: one may only obtain it as a fraction of successes over trials by considering *samples* of events. This introduces a complication in the k-nn problem, which fully motivates the generalization of the usual "k-averaging" performed in such algorithms.

A.1 Details of the implementation

Data events were divided into two orthogonal sub-samples based on the value of a kinematic discriminator capable of effectively separating signal and background events; with it, a signal-depleted control region was defined. The data contained therein was used to estimate the b-tag probability P as a function of n relevant kinematical features of the jet and the event containing it. This was derived by selecting a sample of N_H events found to be the closest -i.e. the most *similar*- (in the multi-dimensional sense discussed below) to the point $X = x_1, x_2, ..., x_n$ where the function was to be evaluated. The function value $P(X)$ was then simply determined as the fraction of b-tagged events in the sample of N_H.

The generalized multi-dimensional distance at the basis of the selection of N_H neighbors was computed using the n variables describing the multi-dimensional feature space as follows:

$$D = \sum_{i=1}^{n_V} w_i^2 \left(x_i - y_{\alpha i}\right)^2 \tag{12}$$

where x_i are the variables defining the evaluation point X, and $y_{\alpha i}$ are the corresponding variables in events belonging to the control region α. The weights w_i account for the different importance of the features i in determining the value of P, and they are at the heart of the k-nn implementation described here.

The full range of variability of each variable x_i was divided in $n_B = 10$ bins of varying size, chosen such that have all the bins were equally populated by control region data; the b-tagging fraction was then computed in each bin. For any value of the variable x_i a single-dimensional weight w_i equating to the local gradient of the univariate b-tag probability was approximated by

$$w_i = \frac{f_{i_R} - f_{i_L}}{\bar{x}_{i_R} - \bar{x}_{i_L}}. \tag{13}$$

The above formula provides a crude estimate of the gradient of the function in each of the directions of space, as it assumes it to be independent on the particular location in the space where we need to estimate it. Indeed, it is only the starting point of a more refined estimate, detailed below.

A.2 The point-by-point weight estimation

One starts by collecting a large set of N_A events (with $N_A >> N_B$), selected from the control sample as the closest ones to the evaluation point X of the b-tag probability $P(X)$, using

the above simplified definition of weights in the distance calculation. The selected N_A events can then be used to compute a bias in the b-tag probability estimate. For each direction of space, spanned by x_i, a numeric estimate of the possible quadratic dependence of the b-tagging probability around the evaluation point (in the sense specified by the unrefined distance definition described above) is performed within the large hyperellipsoid spanned by the N_A events, using the calculation detailed below.

The difference in the value of the quadratic and constant function for $x'_i = x_i$ (the coordinate of the "center" of the hyperellipsoid containing the N_A events) may define the "bias" which one is subjected to if, by using evenly-distributed training events along x_i, one averages their b-tag probability neglecting the fact that the latter has a more complex variation than a constant or linear one. The magnitude of the bias along the direction x_i is used as an additional weight $W(X)$, this time fully varying throughout the feature space, which multiplies the relative distance component of control region events from the evaluation point:

$$D(X) = \sum_{i=1}^{n_V} W(X)^2 w_i^2 \left(x_i - y_{\alpha i}\right)^2 . \tag{14}$$

In such a way the metric D of the space around the evaluation point X is adaptively improved. In other words, the resulting set of N_H events which will be found the closest to X is distributed within a hyperellipsoid whose shape is characteristic of that particular point of the multi-dimensional space.

The approximation of the b-tagging probability in each point of space X with a quadratic function is the "lowest-order" improvement to a linear function, but it provides already a very effective estimate of the bias due to evaluating the b-tagging probability away from X by moving along each coordinate. The method provides a significant improvement in the choice of the most "similar" events to the testing one, as far as the b-tag probability is concerned.

A.2.1 Details of the weight calculation

The b-tag probability cannot be easily fit to a quadratic shape within the hyperellipsoid, because it is not a function defined on a per-event basis: events are either tagged or untagged, so the individual values are zeros or ones. However, we can estimate the parameters of the quadratic function as follows.

We write the b-tag probability as a function $P(X)$ defined in the surrounding of the evaluation point X, for which we take $x = 0$; here the variable x is the direction in the hyperspace along which we wish to estimate the bias in our estimate of b-tagging which we are subject to if we compute a linear average within the hyperellipsoid. So in general we have

$$P(x) = a_0 + a_1 x + a_2 x^2. \tag{15}$$

If we cannot easily fit for $P(x)$ within the hyperellipsoid, we can at least compute the moments of the distribution of b-tagged and untagged events in the surroundings of $x = 0$: these are defined as

$$\lambda_0 = \int_{-d}^{d} P(x)dx = 2da_0 + \frac{2}{3}d^3 a_2 \tag{16}$$

$$\lambda_2 = \int_{-d}^{d} x^2 P(x)dx = \frac{2}{3}d^3 a_0 + \frac{2}{5}d^5 a_2 \tag{17}$$

while of course odd moments are null. The two equations above allow to easily solve for a_0. We obtain

$$a_0 = \frac{9d^2\lambda_0 - 15\lambda_2}{8d^3}. \tag{18}$$

Now, if instead of assuming a quadratic form (the simplest form which produces a bias in the b-tagging rate if averaged over the hyperellipsoid) we decide, as we do, to simply estimate the b-tagging rate by averaging it in the $[-d, d]$ interval, we obtain

$$P(x) = a_0 = \frac{\lambda_0}{2d}. \tag{19}$$

The bias we are subject to because of linear averaging instead than considering the quadratic dependence is then

$$\Delta a_0 = \frac{\lambda_0}{2d} - \frac{9d^2\lambda_0 - 15\lambda_2}{8d^3} = \frac{15\lambda_2 - 5d^2\lambda_0}{8d^3}. \tag{20}$$

We can take the bias computed as above in each direction of space around the test point as our wanted estimate of the relative weight $W(X)$ in the calculation of the generalized distance in the space: the metric D is now therefore susceptible of how the b-tagging rate varies in a way directly proportional to the bias produced by our averaging procedure.

The λ coefficients are estimated by the formulas

$$\lambda_0 = 2d\frac{N_{tag}}{N_{jet}} \tag{21}$$

and

$$\lambda_2 = d^3\frac{2\Sigma_{tag}x^2}{3\Sigma_{jet}x^2} \tag{22}$$

where we have labeled N_{tag} the number of b-tagged events in the hyperellipsoid, N_{jet} the number of events not containing a b-tag in the considered jet, and the sums run on the respective sets of events.

B Software details

Software	Version	References	Use/Notes
NumPy	various	[31]	Data analysis and computation
R	version 3.3.2	[29]	Exploratory analysis and testing

References

[1] G Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *J. Instrum.*, 3:S08003. 437, 2008. Also published by CERN Geneva in 2010.

[2] F Abe, H Akimoto, A Akopian, MG Albrow, SR Amendolia, D Amidei, J Antos, S Aota, G Apollinari, T Asakawa, et al. First observation of the all-hadronic decay of t t pairs. *Physical Review Letters*, 79(11):1992, 1997.

[3] F Abe, MG Albrow, SR Amendolia, D Amidei, J Antos, C Anway-Wiese, G Apollinari, H Areti, P Auchincloss, M Austern, et al. Evidence for top quark production in p^- p collisions atv s= 1.8 tev. *Physical Review D*, 50(5):2966, 1994.

[4] S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003.

[5] J. Allison et al. Geant4 developments and applications. *IEEE Transactions on Nuclear Science*, 53(1):270–278, Feb 2006.

[6] J. Alwall et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[7] Adelchi Azzalini and Bruno Scarpa. *Data analysis and data mining: An introduction*. OUP USA, 2012.

[8] Levan Babukhadia, Fumi Ukegawa, Wade Fisher, Anna Goussiou, Richard Partridge, Martin Hennecke, Tom Junk, Weiming Yao, Boaz Klima, Luca Scodellaro, et al. Results of the tevatron higgs sensitivity study. Technical report, 2003.

[9] R.D. Ball et al. Parton distributions with LHC data. *Nucl. Phys.*, B867:244–289, 2013.

[10] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[11] JM Bibby, JT Kent, and KV Mardia. Multivariate analysis, 1979.

[12] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[13] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. Fastjet user manual. *arXiv preprint arXiv:1111.6097*, 2011.

[14] J. E. Chacón and T. Duong. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19(2):375–398, 2010.

[15] S. Chatrchyan et al. The CMS experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[16] CMS collaboration et al. Search for a higgs boson decaying into a b-quark pair and produced in association with b quarks in proton–proton collisions at 7 tev. *Physics Letters B*, 722(4):207–232, 2013.

[17] J. de Favereau et al. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014.

[18] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016.

[19] B. Duong, T. Goud and K. Schauer. Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences*, 109(22):8382–8387, 2012.

[20] V. Khachatryan et al. First measurement of bose-einstein correlations in proton-proton collisions at s= 0.9 and 2.36 tev at the lhc. *Physical review letters*, 105(3):032001, 2010.

[21] V. Khachatryan et al. Measurement of bose-einstein correlations in pp collisions at\ sqrt {s}= 0.9 and 7 tev. *Journal of High Energy Physics*, 2011(5):1–29, 2011.

[22] A. Khan and G. D. Rayner. Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics & Decision Sciences*, 7(4):187–206, 2003.

[23] A. Mertens. New features in Delphes 3. *J. Phys. Conf. Ser.*, 608(1):012045, 2015.

[24] F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

[25] M. Selvaggi. DELPHES 3: A modular framework for fast-simulation of generic collider experiments. *J. Phys. Conf. Ser.*, 523:012033, 2014.

[26] D. J. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[27] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[28] T. Sjöstrand et al. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

[29] R Core Team. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. 2013, 2014.

[30] H.E.A. Tinsley and S.D. Brown. *Handbook of applied multivariate statistics and mathematical modeling*. Academic Press, 2000.

[31] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.