A population-based approach to background discrimination in particle physics

Federico Colecchia

Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UNITED KINGDOM

Now at Brunel University, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UNITED KINGDOM

E-mail: `federico.colecchia@brunel.ac.uk`

Abstract. Background properties in experimental particle physics are typically estimated from control samples corresponding to large numbers of events. This can provide precise knowledge of average background distributions, but typically does not take into account statistical fluctuations in a data set of interest. A novel approach based on mixture model decomposition is presented, as a way to extract additional information about statistical fluctuations from a given data set with a view to improving on knowledge of background distributions obtained from control samples. Events are treated as heterogeneous populations comprising particles originating from different processes, and individual particles are mapped to a process of interest on a probabilistic basis. The proposed approach makes it possible to estimate features of the background distributions from the data, and to extract information about statistical fluctuations that would otherwise be lost using traditional supervised classifiers trained on high-statistics control samples. A feasibility study on Monte Carlo is presented, together with a comparison with existing techniques. Finally, the prospects for the development of tools for intensive offline analysis of individual interesting events at the Large Hadron Collider are discussed.

1. Introduction

Background discrimination in particle physics is usually performed by identifing events that are more likely to contain a physics process of interest, the primary goal being rejection of contributions from uninteresting processes that mimic the signal and thus make its extraction and measurement more complicated. Traditional approaches achieve this goal by focussing on entire events, comparing kinematic and topological properties with reference distributions usually obtained from control samples.

This article presents a novel approach that builds on a population-based view of particle physics events, which are treated as mixtures of subpopulations comprising particles originating from different physics processes such as a hard scattering of interest as opposed to background. The main goal is to decompose an input data set by assigning individual particles a probability for them to originate from a given process based on particle-level information.

This is achieved by adapting and applying mixture decomposition techniques [1] that are well established in statistics and that have been used in other disciplines to solve formally-similar problems. In this formulation, events are treated as heterogeneous statistical populations comprising particles whose kinematics reflects the process they originated from.

This contribution describes an initial investigation of the possibility to use mixture model decomposition techniques for background discrimination at the Large Hadron Collider (LHC). The study is based on a sampling algorithm inspired by the Gibbs sampler [2] and by Expectation Maximization (EM) [3] whose goal is to decompose an input data set into collections of particles originating from a hard scattering of interest as opposed to background, mapping individual particles to signal or background on a probabilistic basis. A number of well-established methods and results set a context for this investigation in addition to the Gibbs sampler and to EM, namely (i) other simulation-based methods such as [4], (ii) a more general use of Markov Chain Monte Carlo (MCMC) techniques, recently applied to the study of the Cosmic Microwave Background radiation [5], (iii) a recent renewed interest in Bayesian numerical methods for data analysis in particle physics [6] [7] [8], in addition to (iv) the use of MCMC with reference to specific optimization problems in the field [9].

In this study, the proposed sampling algorithm was used to classify individual particles into signal and background. Results obtained on a collection of ~ 600 simulated particles from a hard scattering and from background are presented and discussed in this article, together with cross-checks on toy Monte Carlo as described in the appendix.

In general, different events in particle physics can look very different from one another even when the underlying physics processes are the same, and statistical fluctuations can be non-negligible in low-statistics data sets. When classification is performed using traditional supervised algorithms, fluctuations are usually not taken into account, since training typically relies on high-statistics control samples. On the other hand, the algorithm presented in this article can estimate properties of signal and background probability density functions (PDFs) from the data: in principle, this makes it possible to use information obtained from a data set of interest to improve on the description of background PDFs obtained from control samples, which do not normally take statistical fluctuations into account.

From a broader perspective, this contribution illustrates a new population-based approach that aims to improve on the description of background PDFs obtained from a high-statistics control sample by using information about statistical fluctuations extracted from a lower-statistics data set: this is done by assigning individual particles a probability for them to originate from signal or background, i.e. by decomposing an input collection of particles into signal and background-associated subpopulations.

2. The sampling algorithm

This approach to background discrimination is presented with reference to the general problem of decomposing a collection of particles from high-energy particle collisions into subpopulations associated with different underlying physics processes and described in terms of different PDFs.

The input data set consists of a mixture of particles, some of which originated from a hard scattering of interest, others from background. Provided that the corresponding subpopulations can be characterized sufficiently well in terms of their kinematic or topological properties, it is in principle possible to ask, for each particle, what the probability is for it to originate from signal as opposed to background. In particular, the proposed algorithm estimates such probabilities by iteratively sampling from subpopulation PDFs.

As opposed to classical mixture models, which typically rely on a parametric formulation requiring the shapes of the subpopulation PDFs to be known a priori, our formulation is based on a more general mixture of the form

$$\sum_{j=1}^{K} \alpha_j f_j(x) \tag{1}$$

where the PDFs f_j satisfy a set of constraints associated with a histogram regularization procedure as outlined in section 3. Subpopulation fractions α_j ("mixture weights") are required to sum to unity, i.e. $\sum_{j=1}^{K} \alpha_j = 1$.

The variable x can correspond to particle pseudorapidity η, a kinematic variable related to the particle polar angle θ in the laboratory frame in terms of $\eta = -\ln(\tan\theta/2)$, or p_T i.e. the transverse momentum of the particle with respect to the beam direction. The subpopulation PDFs f_j are defined in terms of regularized histograms of x, as described in section 3, where the associated constraints imposed on the PDFs are detailed. The symbol φ_j will be used to denote the estimate of the generic subpopulation PDF f_j throughout the text.

The choice of (1) was driven by our previous studies, where assuming a predefined PDF functional form led to significant bias on mixture weight estimates. That bias ultimately related to assuming that PDFs obtained from high-statistics control samples were also appropriate to describe the corresponding probability distributions in a lower-statistics data set. However, statistical fluctuations are sometimes appreciable, and for this reason it is necessary for the model to provide more flexibility if fluctuations in the data set of interest are to be described. While a rigorous treatment may call for the use of nonparametric Bayesian methods [10], which can be used to provide an additional dimension of flexibility to statistical models, it was decided to adopt a simplified intuition-driven approach for this study, in order to avoid introducing additional complications not related to the algorithm itself in this phase of the development.

The histogram regularization procedure described in section 3 can be seen as a simplified version of established methods such as Tikhonov regularization [11], which can be used to impose smoothness constraints to a likelihood maximization problem. From a conceptual point of view, an alternative way of interpreting the model used in this study is as a simplified version of established kernel or wavelet-based techniques, where regularized histograms effectively play the role of a set of basis functions. In the absence of any constraints to the PDFs in the mixture, the statistical model (1) would instead not be well defined, so this is an essential ingredient. Additional remarks about existence and uniqueness of a stationary distribution for the Markov Chain associated with the algorithm in the configuration used for this study will be provided in section 3 after the discussion of the Monte Carlo analysis.

Given the mixture of probability distributions (1) and a set of observations $\{x_i\}_{i=1,...,N}$, the problem of clustering the latter into K groups by probabilistically associating each of them with a distribution of origin has been solved numerically in a Bayesian framework using MCMC techniques. In particular, the Gibbs sampler [1], which directly inspired this work, has been used for this purpose in different disciplines.

The basic pseudocode of the proposed sampling algorithm is reported below. The value of variable v at iteration t is indicated with $v^{(t)}$ throughout.

```
(i) Initialization: Choose $\underline{\alpha}^{(0)} = \{\alpha_j^{(0)}\}_j$ and $f_j^{(0)}$=$\varphi_j^{(0)}$, $j = 1, ..., K$ as described in section 3.
(ii) Iteration t:
    (a) Generate the allocation variables $z_{ij}^{(t)}$, $i = 1, ..., N$, $j = 1, ..., K$ based on probabilities $P(z_{ij}^{(t)} = 1|\alpha_j^{(t-1)}, \varphi_j^{(t-1)}, x_i)$ proportional to $\alpha_j^{(t-1)} f(x_i|\varphi_j^{(t-1)})$. The quantity $z_{ij}^{(t)}$ equals 1 when observation $i$ is mapped to distribution $j$ at iteration $t$, and 0 otherwise. In general, the variables $z_{ij}^{(t)}$ depend both on the mixture weights $\alpha_j$ and on the estimates $\varphi_j$ of the subpopulation PDFs from the previous iteration.
    (b) Generate $\underline{\alpha}^{(t)}$ from the probability density function of $\underline{\alpha}$ given $\underline{z}^{(t-1)} = \{z_{ij}^{(t-1)}\}_{ij}$, $\rho(\underline{\alpha}|\underline{z}^{(t-1)})$. Knowledge of which particles are mapped to process $j$ at iteration $t-1$ makes it possible to generate the subpopulation fractions $\underline{\alpha}$ at iteration $t$.
    (c) Obtain an updated estimate of the subpopulation PDFs from the data $\underline{x}$ based on knowledge of which particles are mapped to subpopulation $j$ at iteration $t-1$. Details are provided in section 3.
```

A specific choice for the function ρ and a way to obtain updated estimates of the subpopulation PDFs f_j are described in section 3 with reference to the Monte Carlo study.

The central idea of the algorithm is the following: the better the observations $\{x_i\}_i$ are mapped to the subpopulations $j = 1, ..., K$, the more accurate the estimates of $\rho(\underline{\alpha}|\underline{z})$ and of the subpopulation PDFs φ_j. Once some correct values of z_{ij} are found, $\rho(\underline{\alpha}|\underline{z})$ and φ_j begin to roughly reflect the correct distributions, which in turn leads to additional correct mappings z_{ij} to be found at subsequent iterations.

The above pseudocode corresponds exactly to the Gibbs sampler, where updated estimates of subpopulation PDFs are obtained at each iteration, as indicated at step (c). On the other hand, when step (c) is removed from the pseudocode, particles are mapped to signal and background based on the subpopulation PDFs provided at initialization, and the algorithm is then more akin to EM. Throughout the paper we will refer to the two versions of the algorithm with step (c) included or not in the pseudocode as “unconstrained sampler" and “constrained sampler", respectively.

The primary objective of this article is to study the use of the proposed sampling technique in different configurations in order to:

(i) Obtain estimates φ_j of the subpopulation PDFs from the input data set.

(ii) Estimate the subpopulation weights α_j. In the context of this study, this corresponds to estimating the fractions of background and signal particles contained in the input data set.

(iii) Assign individual particles a probability for them to originate from a given process based on the subpopulation PDFs estimated at step (i) as opposed to relying exclusively on high-statistics templates. In the context of this study, this allows classification of individual particles into signal and background.

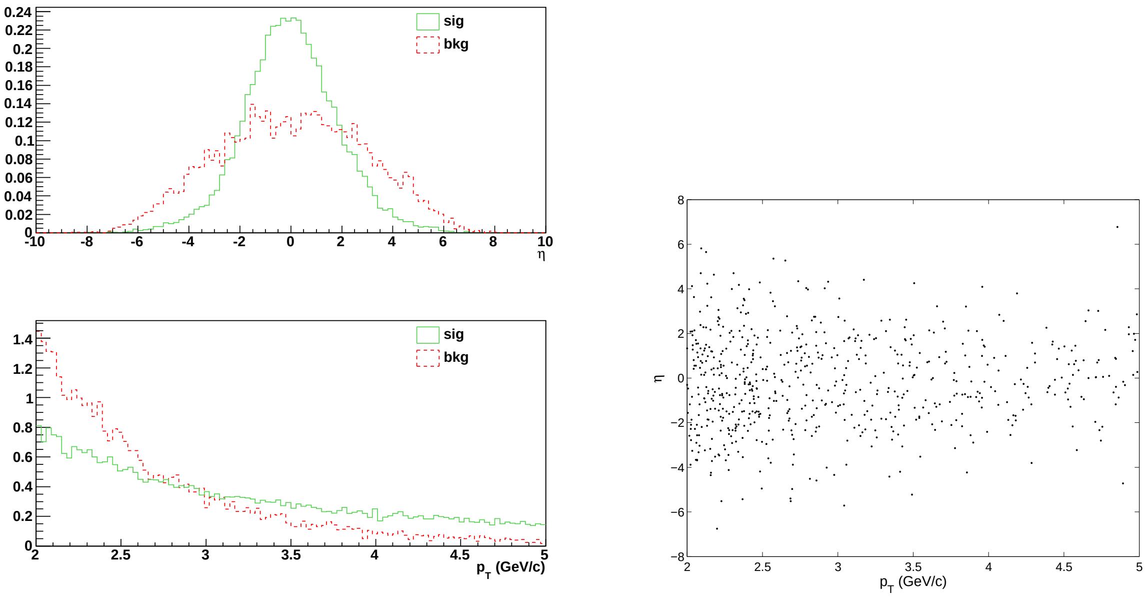
3. Monte Carlo study

The algorithm was applied to a Monte Carlo data set generated using Pythia 8.140 [12] [13], obtained superimposing $gg \rightarrow t\bar{t}$ signal events from pp interactions at $\sqrt{s} = 14$ TeV to soft QCD interactions, so called Minimum Bias events, in order to simulate background. The signal process was chosen in order to illustrate the use of the algorithm for background discrimination at the particle level. Further studies will be needed in order to extend these results beyond the initial investigation presented in this article, and to assess the potential of population-based techniques for background discrimination in the context of specific analyses at the LHC.

The sampler was run over a collection of charged particles with 2 GeV/c $< p_T <$ 5 GeV/c, and individual particles were assigned a probability for them to originate from signal as opposed to background based on their η and p_T values.

The pseudocode of the algorithm used for this application is shown below. Subscripts *sig* and *bkg* relate to signal and background, respectively.

```
(i) Initialization: Set $\alpha_{bkg} = \alpha_{bkg}^{(0)} = 0.5$, $f_j = \varphi_j^{(0)}$. Initial conditions for the estimates $\varphi_j^{(0)}$ of the subpopulation PDFs $f_j^{(0)}$ are given by regularized $\eta$ and $p_T$ distributions from the high-statistics control sample, as described in section 3.1.
(ii) Iteration $t$:
    (a) Generate $z_{ij}^{(t)}$ for all particles ($i = 1, ..., N$) and distributions ($j = 1, 2$ corresponding to background and signal, respectively) according to $P(z_{ij}^{(t)} = 1|\alpha_j^{(t-1)}, \varphi_j^{(t-1)}, x_i) \propto \alpha_j^{(t-1)} f_j(x_i|\varphi_j^{(t-1)})$, where $\alpha_1 = \alpha_{bkg}$, $\alpha_2 = 1 - \alpha_{bkg}$.
    (b) Set $\alpha_j^{(t)} = \sum_i z_{ij}^{(t-1)}/N$, $\forall j$. This corresponds to the simplest choice of setting $\rho(\alpha_j|\underline{z}^{(t-1)}) = \delta(\alpha_j - \sum_i z_{ij}^{(t-1)}/N)$ for the probability density function of $\underline{\alpha}$ given $\underline{z}$.
```



(a)

(b)

Figure 1. (a) Generator-level η and p_T distributions for signal (solid green histograms) and background particles (dashed red histograms) with 2 GeV/c $< p_T <$ 5 GeV/c from the high-statistics control sample. The distributions correspond to a total number of $\sim 33,000$ particles and are normalized to unit area. (b) The corresponding two-dimensional distribution.

(c) Obtain updated estimates of the subpopulation PDFs by regularizing the η and p_T distributions corresponding to particles mapped to the relevant subpopulation at iteration $t-1$, i.e. based on $z_{ij}^{(t-1)}$.

In general, the functions f_j are the joint PDFs for η and p_T corresponding to background ($j = 1$) and signal ($j = 2$) particles. This study is restricted to charged particles with 2 GeV/c $< p_T <$ 5 GeV/c, which makes it possible to neglect the correlation between η and p_T as a first approximation. For this reason, the joint PDFs take the form $f_{sig/bkg} = f_{sig/bkg}^{(\eta)} f_{sig/bkg}^{(p_T)}$, and obtaining updated estimates of the subpopulation PDFs reduces to regularization of one-dimensional histograms, as described in the following.

As for the number of iterations to be used with the algorithm, no rule is documented in the statistics literature with reference to related techniques, and the choice is generally problem-dependent. The number of iterations was set to 1,000 in this study, and probabilities were averaged over the last 100 iterations. Runs were also performed letting the sampler run for a longer time: the algorithm exhibited a relatively fast convergence on the data set analyzed, and no gain was found in choosing a higher number of iterations. Moreover, multiple runs were performed under different initial conditions in order to make sure that the algorithm converged. In particular, the initial conditions for the subpopulation PDFs were perturbed by using different initial conditions for the fits to the high-statistics distributions from the control sample. Similarly, the generation parameters in the toy Monte Carlo study were varied around their nominal values by $\pm 10\%$, with no appreciable difference in the results.

In order to obtain initial conditions $\varphi_j^{(0)}$ for the subpopulation PDFs, a Monte Carlo data set was used containing a total of about 33,000 charged particles in the kinematic range 2 GeV/c $< p_T <$ 5 GeV/c. In addition to estimation of $\varphi_j^{(0)}$, this high-statistics control sample was also used to guide the histogram regularization procedure as described in the following. Figure 1 (a) shows the η and p_T distributions for signal and background particles (solid green and dashed red histograms, respectively).

As anticipated, one of the goals of the sampler is to estimate the background PDFs from the input collection of particles. In other words, the algorithm classifies particles into signal and background without relying exclusively on predefined background templates: the background PDFs that are estimated by the algorithm are thus expected to reflect the specific background conditions in the input data set, which can be different from the average conditions obtained from a high-statistics control sample.

The algorithm basically tries to uncover a signal and a background subpopulation in the input collection of particles based on the data and on initial conditions on the subpopulation PDFs. The results presented in this article relate to an input data set comprising 636 charged particles in the kinematic region 2 GeV/c $< p_T <$ 5 GeV/c, out of which 481 originate from a signal hard process and 155 from background, corresponding to a fraction of background particles of $\sim 24\%$. The total number of particles in the input data set is in line with typical charged particle multiplicities at the LHC as of July 2011.

The signal and background η and p_T distributions corresponding to the Monte Carlo input data set used in this study are shown in figure 2. The solid green (dashed red) histograms in the upper panel display the signal (background) η distributions, normalized to unit area. The corresponding p_T distributions are given in the lower panel. It is worth noticing that some of these distributions are appreciably different from the corresponding ones obtained from the control sample due to statistical fluctuations, as expected. In particular, the background η distribution exhibits two modes that are shifted with respect to zero, while the corresponding distribution from the control sample is centered around zero.

In order to illustrate the histogram regularization procedure used in this study, figure 3 shows an example of the η distribution of particles mapped to the signal subpopulation at a given iteration of the algorithm. As opposed to assuming a functional form for the PDF and fitting a function to the histogram, the histogram is regularized i.e. the subpopulation PDF is obtained by means of spline interpolation of the histogram contents, as further discussed in the following. The superimposed curve on the figure corresponds to the regularized histogram, and is used by the algorithm as an estimate of the corresponding subpopulation PDF.

As anticipated, this approach gives the algorithm more flexibility in terms of estimating the subpopulation PDFs from the input data set with respect to our previous attempts relying on a predefined PDF functional form, while still leading to a well-defined target distribution for the associated Markov Chain.

3.1. Regularization

Step (c) in the pseudocode shown in section 3 requires iterative PDF updates based on current mapping of individual particles to different subpopulations. This operation is performed when the algorithm is operated in unconstrained mode, as discussed below.

As anticipated, it was decided to adopt a simplified statistical model for the purpose of this investigation, while at the same time providing enough flexibility for the algorithm to be able to describe fluctuations. In the context of this study, this was done by performing spline interpolation of one-dimensional η and p_T histograms. As previously mentioned, this can be seen as a simplified version of established regularization techniques, for instance as a way to use a priori information about the underlying distributions in order to get rid of spurious oscillatory components (such methods have been analyzed in detail in particle physics in order to develop unfolding procedures, with a view to “removing" detector effects from observed distributions, see e.g. [14]).

The complexity of the histogram regularization procedure that was used in this study was

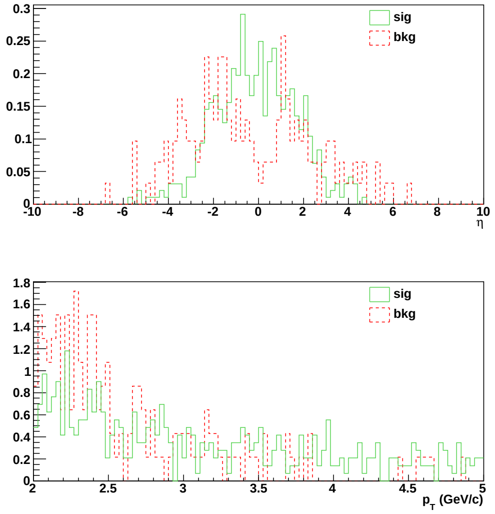


Figure 2. Particle η and p_T distributions from the Monte Carlo input data set used in this study. Solid green and dashed red histograms correspond to signal and background, respectively. Distributions are normalized to unit area.

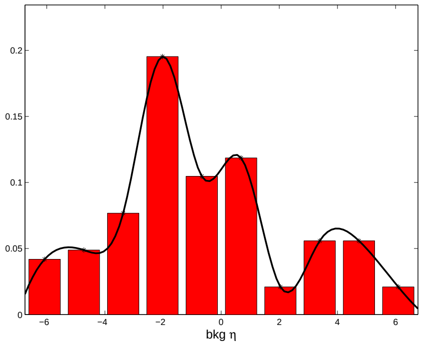


Figure 3. Example of the pseudorapidity η distribution of particles mapped to the background subpopulation at a given iteration of the algorithm. The superimposed curve is the result of the regularization procedure described in the text, and is used by the algorithm as an estimate of the corresponding subpopulation PDF.

intentionally kept minimal in order to avoid the introduction of additional complications that might obscure the response of the algorithm at this stage of the development. Further studies will be needed in order to understand in detail how results are affected by the regularization procedure.

In the context of this investigation, a priori information about signal and background PDFs was obtained from the high-statistics control sample. When subpopulation PDFs are updated iteratively during the execution of the sampler, i.e. when the algorithm is operated in unconstrained mode, a “regularization window" is applied to the η histograms in order to get rid of outliers: in other words, a spline interpolation of the histogram contents is obtained using only the part of the histogram that lies between a minimum and a maximum η value, which leads to extreme statistical fluctuations on the tails of the distribution to be excluded. It is worth noticing that assuming a functional form for the PDF and fitting it to the histogram would effectively produce a similar result, i.e. it would reduce the impact of outliers on the estimated PDF. However, as anticipated, that approach was observed to introduce significant bias in previous studies, and was thus abandoned in favor of the statistical model presented in (1), where subpopulation PDFs are defined as the output of a histogram regularization procedure without reference to any predefined functional form, but still subject to regularization constraints. Figure 4 shows signal (solid green) and background (dashed red) η distributions from the high-statistics control sample, with superimposed arrows indicating the regularization window. The maximum $|\eta|$ value was set to $|\eta| = 5$ ($|\eta| = 7$) for signal (background).

On top of this, again with a view to getting rid of extreme statistical fluctuations when regularizing histograms, boundary conditions were introduced on the η and p_T PDFs, constraining the value of f_j to points chosen based on control sample distributions: in particular, the signal (background) η PDF was constrained to 0 when $|\eta| > 5$ ($|\eta| > 7$), and signal (background) p_T PDFs were constrained at 2 GeV/c and 5 GeV/c to 0.7 (1.2) and 0.1 (0)

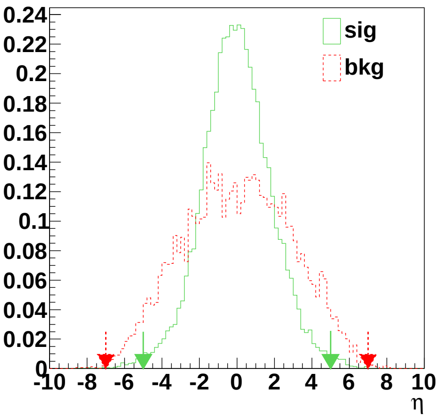


Figure 4. Illustration of the regularization window used in this study. The histograms correspond to signal (solid green) and background (dashed red) η distributions from the high-statistics control sample. The position of the solid green and dashed red arrows correspond to the regularization window: at each iteration of the algorithm, only the part of the distribution that lies between the arrows is used for spline interpolation, which makes the results robust against outliers. Additional details are given in the text.

(see figure 1(a)).

Results were found to be stable with respect to reasonable changes to the above regularization contraints.

3.2. Choice of configuration

As anticipated, the algorithm can be operated in unconstrained or constrained mode, depending on whether step (c) in the pseudocode given in section 3 is included or not.

As already pointed out in section 2, the algorithm can process an input collection of particles in order to obtain one or more of the following results:

(i) Estimate the subpopulation PDFs from the input data set.

(ii) Estimate the fraction of particles associated with a given process in the input data set, e.g. the fraction of background particles.

(iii) Assign individual particles a probability for them to originate from a given process, such as a hard scattering of interest as opposed to background, based on subpopulation PDFs estimated at step (i) as opposed to relying on predefined templates that reflect average background conditions.

Depending on the objective, it may be appropriate to run the algorithm in different modes.

For instance, the histogram regularization procedure that is used here to obtain iterative estimates of the subpopulation PDFs when the algorithm is operated in unconstrained mode inherently leads to bias on mixture weights, because imposing a regularization window changes the number of particles that are mapped to signal or background at a given iteration. For this reason, it may be more appropriate to use a different approach in order to estimate the fraction of background particles.

One option is described below:

(a) The constrained sampler is first used to estimate the mixture weights. In the two-subpopulation scenario described in this study, goal (ii) above corresponds to estimating the fraction of background particles contained in the input data set. The initial conditions for the mixture weights are $\alpha_1^{(0)} = \alpha_2^{(0)} = 0.5$, corresponding to no prior knowledge about the fraction of background particles in the input sample. The subpopulation PDFs are kept fixed at the estimates provided by the high-statistics control sample. The corresponding results are described in section 3.2.1.

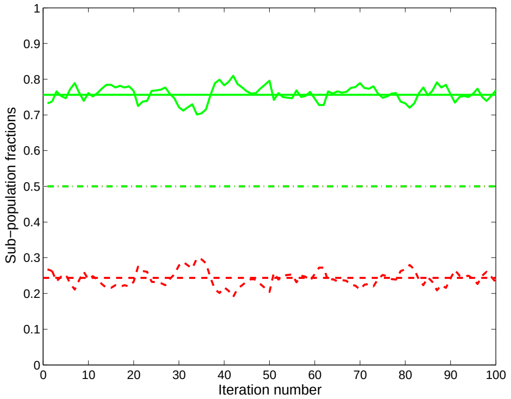


Figure 5. Mixture weights obtained running the constrained sampler on the Monte Carlo input data set. Results from the last 100 iterations are shown. The solid green (dashed red) curve denotes the estimated fraction of signal (background) particles. The solid green (dashed red) horizontal line indicates the true value for signal (background) from the simulation, and the dash-dot line corresponds to the initial conditions for the mixture weights.

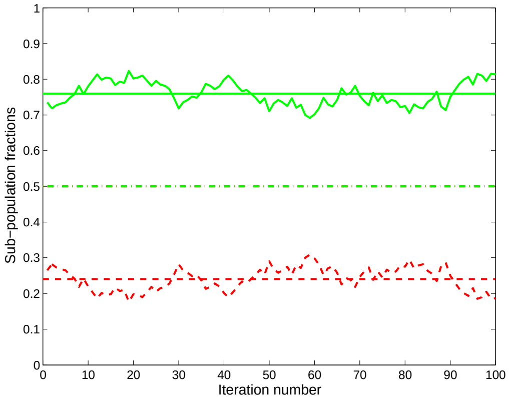


Figure 6. Mixture weights obtained running the constrained sampler on a toy Monte Carlo data set, as described in the text. Results from the last 100 iterations are shown. The solid green (dashed red) curve corresponds to the estimated fraction of signal (background) particles. The solid green (dashed red) horizontal line indicates the true value for signal (background) from the toy Monte Carlo, and the dash-dot line corresponds to the initial conditions for the mixture weights.

(b) The algorithm is then run again on the input data set in unconstrained mode, i.e. subpopulation PDFs are now updated at each iteration, starting from initial conditions corresponding to regularized distributions from the high-statistics control sample. However, mixture weights are kept fixed at the results from the previous step.

It is worth noticing that the algorithm differs from a proper Gibbs sampler in both cases.

As for assigning individual particles in the input data set a probability for them to originate from signal as opposed to background, the most appropriate approach may again depend on the specific application. In general, probabilities may be assigned directly using the unconstrained sampler at step (b) above, as done in this study, or an additional run of the algorithm in constrained mode may alternatively be added after the previous two, with fixed PDFs given by the estimates from step (b). Further studies will be necessary in order to better understand the classification performance of the algorithm in different configurations and to guide this choice.

Results obtained running the constrained an unconstrained sampler as described above on the Monte Carlo input data set used in this study are reported and discussed in the following sections.

3.2.1. Constrained sampler As anticipated, the algorithm in constrained mode was primarily used in this study in order to estimate mixture weights, i.e. the fraction of background particles in the input data set. Figure 5 shows the corresponding estimates over the last 100 iterations. The solid green and dashed red curves correspond to the estimated fractions of signal and background particles, respectively. The solid green (dashed red) horizontal line indicates the signal (background) true value from the simulation, while the dash-dot line corresponds to the initial conditions for the mixture weights.

Additional runs on toy Monte Carlo samples were performed as a cross-check, as described in the appendix. In particular, figure 6 displays the estimated mixture weights obtained by running the constrained sampler on a toy Monte Carlo data set with subpopulation PDFs kept fixed at truth information.

3.2.2. Unconstrained sampler The unconstrained sampler was used in this study in order to estimate the signal and background PDFs from the input data set, while keeping the mixture weights fixed at the results obtained from the previous run of the algorithm in constrained mode.

Figure 7 shows the subpopulation PDFs estimated by the algorithm on the Monte Carlo input data set. The curves correspond to the output of the histogram regularization procedure averaged over the last 100 iterations, superimposed to the true distributions (histograms). The η (p_T) distributions are displayed in the top (bottom) plots, figures on the left-hand (right-hand) side correponding to background (signal). All distributions are normalized to unit area. The bottom panel in each figure shows the corresponding ratio between the relevant subpopulation PDF estimated by the algorithm and truth information.

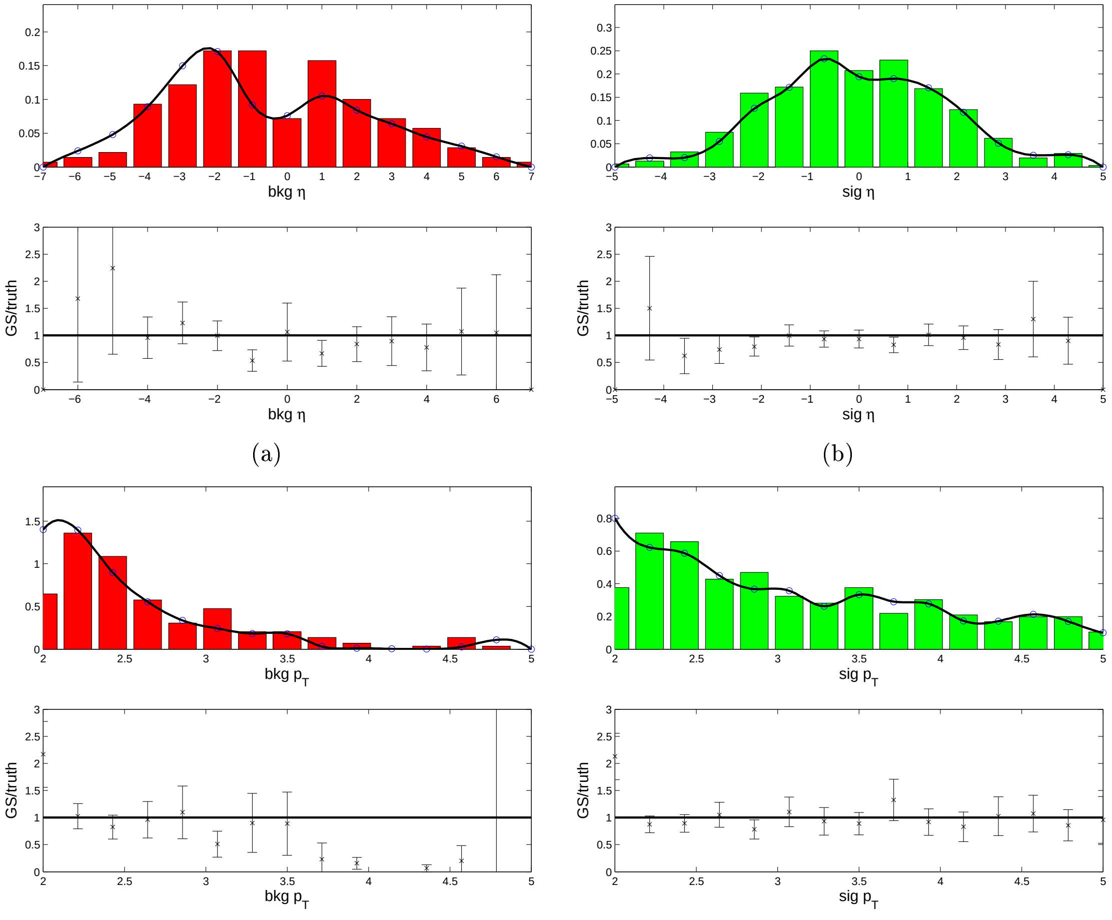
The figure illustrates a distinctive characteristic of the proposed algorithm as opposed to well-established techniques. As already pointed out, the background η distribution in the Monte Carlo data set used in this study differs appreciably from the corresponding distribution obtained from the control sample, as shown by the two modes around $\eta \simeq -2$ and $\eta \simeq 1$ in the figure, as opposed to the symmetric distribution centered around $\eta \simeq 0$ that is obtained from the high-statistics data set. As it can be seen, the sampler was able to identify with reasonable performance the presence of such deviations with respect to the control sample templates. Such properties of the background PDFs are specific to the data set under investigation, and could not have been extracted using traditional supervised classification techniques, since those would have relied on predefined background templates those features would have been absent from in the first place.

In conclusion, although this study shows that traditional techniques can in some cases outperform this algorithm in terms of classification performance as discussed in the following, the primary objective of the proposed approach is not to improve on existing methods in terms of classification performance, but rather to extract information about statistical fluctuations from a data set of interest.

In addition to obtaining data-driven estimates of the subpopulation PDFs in unconstrained mode, one of the goals of the algorithm in this application is to assign individual particles a probability for them to originate from a given process, such as a hard scattering of interest as opposed to background. In this study, the latter probabilities were obtained from the same unconstrained run of the algorithm that provided the PDF estimates shown in figure 7. In general, other choices are possible, such as performing an additional run of the algorithm in constrained mode with subpopulation PDFs kept fixed at the estimates shown in figure 7, as previously mentioned. Detailed studies will be necessary in order to understand the implications of different choices before population-based tools for background discrimination can be applied to physics analysis at the LHC.

An initial comparison of the classification performance of the algorithm in the configuration chosen for this study with the corresponding performance of existing techniques is described in section 3.3.

The probabilities returned by the algorithm were validated by comparing the true kinematic distributions with the corresponding ones for particles with $P_{sig} > 0.5$, P_{sig} being the estimated probability for a given particle to originate from the signal process, averaged over the last 100 iterations. Results are shown in figure 8, where histogram bars indicate the η distribution for particles with $P_{sig} > 0.5$ and stars correspond to true distributions.



(a)

(b)

(c)

(d)

Figure 7. Subpopulation PDFs estimated by the unconstrained sampler on the Monte Carlo input data set used in this study, averaged over the last 100 iterations. (a) Background η. (b) Signal η. (c) Background p_T. (d) Signal p_T. In each subfigure, the upper panel shows truth information (histogram bars) superimposed to the result of the regularization procedure averaged over the last 100 iterations (curve). The lower panels display the ratio between subpopulation PDFs estimated by the algorithm and the corresponding truth-level information.

Table 1. Average numbers of pile-up particles (second column) expected at different LHC instantaneous luminosities (first column) [17]. A 25 ns bunch crossing is assumed. The third column reports the corresponding ratios between the number of background and signal particles observed in the kinematic region considered in this study. These estimates were used to generate the curves shown in figure 10, as described in the text.

$L(\mathrm{cm}^{-2}\mathrm{s}^{-1})$	$\langle n_{PU} \rangle$	N_{bkg}/N_{sig}
10^{33}	2.3	0.1
10^{34}	23.0	0.9
10^{35}	230.0	4.4

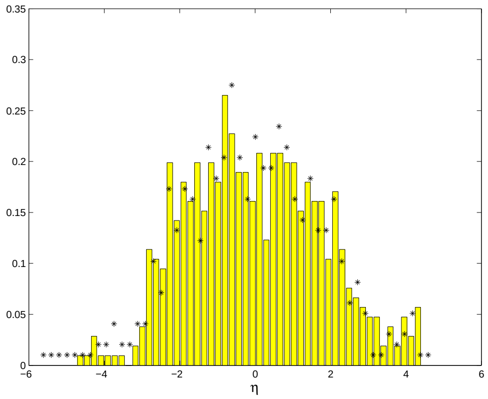


Figure 8. Comparison between the η distribution for particles with $P_{sig} > 0.5$ (histogram bars) and the corresponding distribution from truth (stars). Additional information is given in the text.

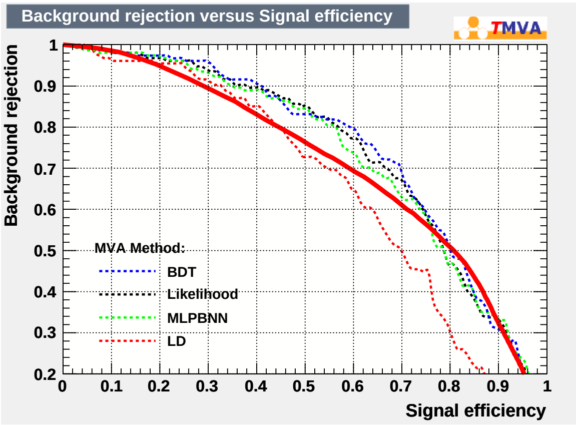


Figure 9. Comparison between the ROC curve obtained using the unconstrained sampler, shown here as background rejection rate as a function of signal efficiency, and the corresponding curves obtained using existing supervised classification techniques from TMVA, as described in the text. The solid red line is the curve from the unconstrained sampler corresponding to the last 100 iterations. The other curves correspond to TMVA algorithms, namely Boosted Decision Trees (dashed blue), Naive Bayes classification (dashed black), the Neural Network-based classifier MLPBNN (dashed green), and Linear Discriminant (dashed red). Additional information is given in the text.

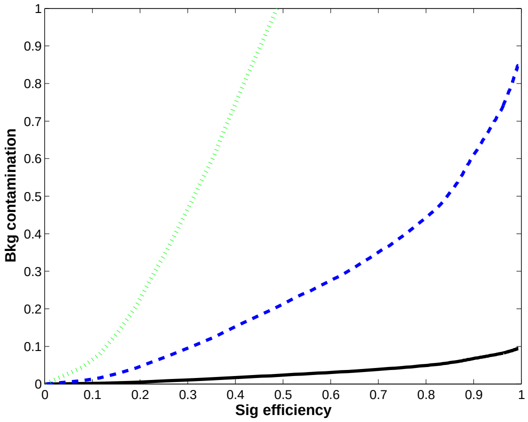


Figure 10. Background contamination as a function of signal efficiency at different LHC instantaneous luminosities (10^{33} cm^{-2}s^{-1} solid black, 10^{34} cm^{-2}s^{-1} dashed blue, 10^{35} cm^{-2}s^{-1} dotted green). Background contamination is defined as number of misclassified background particles normalized to number of signal particles. Misclassification probabilities correspond to the ROC curve from the sampler in figure 9. Additional information is given in the text.

3.3. Classification performance

Operating the sampler as presented in this article is equivalent to using it as a binary classifier. Its performance can thus be quantified using the Receiver Operating Characteristic (ROC) curve, which displays true-positive as a function of false-positive probability. The area under the curve is a number between 0 and 1: the higher its value, the better the classifier is able to discriminate between the two categories (signal and background in this case). The ROC curve of a random classifier would be a straight line along the main diagonal on the true-positive vs false-positive plane (“chance diagonal" [15]).

Figure 9 shows a comparison between the ROC curve obtained using the unconstrained sampler on the Monte Carlo input data set used for this study and the corresponding curves from different supervised multivariate classification methods using TMVA [16] V04-01-00. The curves are displayed using an equivalent representation in terms of background rejection rate as a function of signal efficiency[1] . The dashed lines refer to different TMVA methods[2], namely Boosted Decision Trees (dashed blue), Naive Bayes classification (dashed black), the Neural Network-based classifier MLPBNN (dashed green), and Linear Discriminant (dashed red).

The solid red line corresponds to the proposed algorithm. The figure suggests that classification performance of the sampler is similar to that of existing supervised methods, although other methods perform better in terms of ROC curve on the data set used in this study. However, the advantage of the proposed sampling algorithm with respect to existing methods is not in terms of improved classification performance, but instead relates to estimating features of the background distributions that reflect statistical fluctuations in the data, which is generally not possible using established supervised classifiers trained on control samples.

It may also be useful to provide a more precise idea of the background rejections and signal efficiencies that can be achieved using the proposed algorithm corresponding to different LHC instantaneous luminosities[3]. Figure 10 shows estimates of background contamination as a function of signal efficiency at three different LHC instantaneous luminosities. Background contamination is defined as the number of misclassified background particles normalized to the number of signal particles, and is calculated by rescaling the abscissa of the ROC curve by the ratio between the number of background and signal particles in the kinematic region considered in this study, as given in table 1[4]. The abscissa of the ROC curve in fact corresponds to false-positive rate, i.e. to the probability for a background particle to be misclassified as signal, and multiplying it by the ratio between the number of background and signal particles provides the desired result. The three curves in figure 10 correspond to instantaneous luminosities of 10^{33} cm^{-2}s^{-1} (solid black), 10^{34} cm^{-2}s^{-1} (dashed blue), and 10^{35} cm^{-2}s^{-1} (dotted green).

3.4. Convergence issues

A remark is necessary with regards to the convergence properties of the Markov Chain associated with the proposed sampling algorithm in the form presented in this article. The proposed technique is here justified primarily based on the results it provides, and based on its ability to extract additional information related to statistical fluctuations from a data set of interest. This is to be compared with the description of background distributions obtained using control samples, which, despite its level of precision, usually only reflects average background conditions and does not take statistical fluctuations into account.

Although the statistical model (1) may be questioned from a theoretical point of view and a more rigorous approach based on Bayesian nonparametric methods may be required, the model presented here in practice leads to a well-defined target distribution for the algorithm to sample

[1] Based on our previous terminology, “background rejection rate" is equivalent to $1 - P_{bkg \to sig}$, and “signal efficiency" corresponds to $P_{sig \to sig}$, where $P_{bkg \to sig}$ ($P_{sig \to sig}$) is the probability for a background (signal) particle to be mapped to the signal subpopulation.

[2] The algorithms were run using the high-statistics control sample for training and the same collection of particles the sampler was run on for testing.

[3] The expected average number of pile-up interactions, i.e. the expected average number of primary vertices in the events, is here taken as a measure of background activity for illustrative purposes.

[4] The average numbers of pile-up interactions at different LHC instantaneous luminosities are taken from [17], and correspond to a 25 ns bunch crossing.

from. As anticipated, this is primarily due to the constraints associated with the histogram regularization procedure adopted in this study, which effectively restricts the search space and leads to the existence of a well-defined stationary distribution for the Markov Chain. This was also verified explicitly by using flat distributions as initial conditions for the subpopulation PDFs, making sure that reasonable estimates of the PDFs were still obtained by the sampler.

3.5. Dependence on initial conditionus

One more issue that is worth discussing is dependence of results on initial conditions. The ability to reach the equilibrium distribution regardless of the starting point is a defining feature of Markov Chains. Throughout this study, it has been verified that the initial conditions on the subpopulation PDFs can be perturbed without altering the final resuls. Results are actually independent of the PDF initial conditions well beyond the deviations that are normally expected given the high level of precision with which initial PDF estimates are generally obtained from control samples.

The similarity of the PDFs estimated by the sampler with the PDF initial conditions obtained from the control sample should not be mistaken for a limitation of the proposed method, but should rather be seen as a defining feature. Although incorporating a more rigorous formulation of the statistical model will be important for the method to be developed further, it should be noticed that one of the goals of the algorithm is to improve on knowledge of background PDFs obtained from high-statistics control samples by extracting additional information about statistical fluctuations from a data set under study. For this reason, the PDFs estimated by the sampler will normally be similar to the initial PDFs, and the associated Markov Chain will generally exhibit a relatively-fast convergence by construction.

3.6. Concluding remarks

The possibility to estimate features of signal and background distributions from a data set under study is a distinctive characteristic of the proposed method as compared to existing multivariate approaches such as those available in ROOT [18] with TMVA. Although established techniques in some cases provide better classification performance on the data set analyzed in this study, as shown by the comparison in figure 9, existing methods are in general unable to describe features of the background distributions that are not already encoded in the training sample. And since the latter typically corresponds to a high-statistics control sample, this usually leads to statistical fluctuations in the input data set being neglected.

This technique has been investigated with the prospective goal of developing novel methods for intensive offline analysis of individual interesting events at the LHC, and more generally in particle physics. Data analysis in the field in fact often results in a number of candidate events that may contain a signal process of interest. Traditional methods perform background subtraction based on fixed templates that typically provide a precise description of average background properties. However, this approach normally leads to neglecting features of background distributions due to statistical fluctuations that may be present in the candidate events of interest even though those features cannot be spotted from background templates obtained from control samples. Developing dedicated tools for background subtraction based on event-level templates taking fluctuations into account may then lead to improved background subtraction and to lower systematic uncertainties. This aspect will be the subject of future studies, as will quantification of the impact of the algorithm in a realistic analysis environment.

It is also worth noticing that, from a conceptual point of view, the proposed population-based approach is in a sense based on a similar phylosophy as particle flow analysis, which has been increasingly used in particle physics [19], in that the focus is on individual particles inside events. However, the prospective objective of the proposed technique is different, and concentrates on extracting from the data event-level background templates that take statistical fluctuations into account.

Efforts to eliminate noise in event-by-event analysis of high-energy multiparticle production are reported in the literature, most notably with reference to the study of dynamical fluctuations in heavy-ion collisions, where the notion of “event-by-event fluctuations" was introduced [20], e.g. for mean transverse momentum or mean transverse energy measurement. In the context of such studies, the focus is e.g. on analytically obtaining moments that can be used to eliminate statistical fluctuations from the data with a view to extracting information about the underlying dynamics [21]. Although those studies are conceptually related to the prospective goal of the approach presented in this article in that they aim to subtract noise from individual events, they are fundamentally different. First of all, [21] requires fluctuations to be Poissonian, while this method works under more general conditions. Moreover, one of the novel aspects of this work is the idea of concentrating on individual particles inside events, reformulating background discrimination in terms of a classification problem at the particle level. In other words, the emphasis of this work on a new population-based view of particle physics events is an important aspect that distinguishes the proposed approach from previous efforts.

As a concluding remark, it should also be noted that the iterative nature of the algorithm leads to a disadvantage with respect to established multivariate algorithms in terms of execution time. However, the running time of the sampler corresponding to 1,000 iterations on the Monte Carlo input data set used in this study was $\sim$ 20 s on a 2 GHz Intel Processor with 1 GB RAM, so still reasonable for offline use. In any case, given the parallelization potential of the sampler, which is a consequence of a similar property of the Gibbs sampler as pointed out in [2], improvements may be possible in this respect, for example using commodity Graphics Processing Units (GPUs) that have been used extensively both in particle physics and in other disciplines for compute-intensive applications.

4. Conclusions and outlook

This contribution has presented an initial investigation of a novel approach to background discrimination in particle physics that builds on a population-based view of events from high-energy particle collisions. Collections of particles are treated as mixtures of subpopulations associated with different physics processes, and sampling techniques related to statistical mixture decomposition models are used to assign individual particles a probability for them to originate from a hard scattering of interest as opposed to background. This application of the proposed sampling algorithm to a classification problem at the particle level has been pursued with the prospective goal of developing a suite of tools for extraction of background properties from individual interesting events at the LHC, and more generally in particle physics. For instance, a major objective is to obtain estimates of PDF shapes from the data without relying exclusively on templates from high-statistics control samples and without assuming predefined functional forms.

This study has highlighted strengths and limitations of the algorithm operated in different configurations. In general, systematic uncertainties associated with the use of the algorithm will have to be evaluated in the context of a given analysis.

Detailed understanding of how classification performance in different configurations compares to existing techniques will also require further study, as will the possible development of subsequent versions optimized in terms of execution time, building on the inherent parallelizability of the algorithm.

As anticipated, the total number of particles in the Monte Carlo input data set used in this study is in line with typical charged particle multiplicities at the LHC corresponding to operating conditions as of July 2011. For this reason, the results presented in this article are a promising starting point for futher development, with a view to building dedicated software tools for offline analysis of individual interesting events at the LHC.

Acknowledgments
The author wishes to thank the High Energy Physics Group at the Department of Physics and Astronomy, University College London, for their kind hospitality that created the conditions for this study to be completed. Particular gratitude goes to Prof. Jonathan M. Butterworth for his precious comments on this project. The author also wishes to thank Prof. Trevor Sweeting at the Department of Statistical Science, University College London, for his feedback, and Dr. Alexandros Beskos at the same department for fruitful discussions. Particular gratitude also goes to the Department of Astronomy and Theoretical Physics at Lund University for their kind hospitality when this project was started, and particularly to Prof. Carsten Peterson and to Prof. Leif Lönnblad for their advice and for fruitful discussions.

Appendix: Toy Monte Carlo studies
Results from the Monte Carlo study described in section 3 were cross-checked on toy Monte Carlo data sets. Samples of ~ 600 signal and background particles were generated according to η and p_T distributions similar to those obtained using Pythia. Particle η and p_T were generated independently: Gaussian PDFs centered at zero with standard deviations comparable to those observed in Monte Carlo were used for η, and p_T values were generated based on polynomial PDFs in the range 2 GeV/c $< p_T <$ 5 GeV/c parametrizing the corresponding Monte Carlo distributions.

Additional cross-checks were performed by varying toy Monte Carlo generation parameters by $\pm 10\%$ with respect to their nominal values, in order to make sure that results did not depend on a specific parameter choice. The algorithm was also run on different numbers of particles in the input data set, with no appreciable changes to the results.

References

[1] Marin J M 2005 *Handbook of Statistics* eds Dey D and Rao C R (Elsevier North-Holland) chapter 16, pp 459-503

[2] Geman S and Geman D 1984 *IEEE T. Pattern Anal.* **6** 721

[3] Dempster A P, Laird N M and Rubin D B 1977 *J. Roy. Statist. Soc. Ser. B* **39**(1):1-38

[4] Tanner M A and Wong W H 1987 *J. Amer. Statistical Assoc.* **82** (398):528-540

[5] Groeneiboom N E 2009 A self-contained guide to the CMB Gibbs sampler (*Preprint* astro-ph.CO:0905.3823)

[6] D'Agostini G 1999 *Bayesian Reasoning in High Energy Physics - Principles and Applications* CERN Yellow Report 99-03

[7] D'Agostini G 2003 *Bayesian Reasoning in Data Analysis: A Critical Introduction* (Singapore:World Scientific Publishing)

[8] Ciuchini M *et al.* 2001 *J. High Energy Phys.* JHEP07(2001)013

[9] Buckley A *et al.* 2006 (Preprint hep-ph/0605048)

[10] Ferguson T 1973 *Ann Stat* **1**(2) 209-30

[11] Tikhonov A N 1943 *Doklady Akademii Nauk SSSR* **39**(5):195-8

[12] Sjöstrand T, Mrenna S and Skands P 2006 *J. High Energy Phys.* JHEP05(2006)026

[13] Sjöstrand T, Mrenna S and Skands P 2008 *Comput. Phys. Comm.* **178**

[14] Hoecker A and Kartvelishvili V 1996 *Nucl. Instrum. Methods* A **372**(3) 469-81 (*Preprint* hep-ph/9509307)

[15] Krzanowski W J and Hand D J 2009 *ROC Curves for Continuous Data* (Boca Raton:Chapman & Hall/CRC Monographs on Statistics & Applied Probability)

[16] Hoecker A *et al.* 2007 PoS(ACAT) **040**

[17] Lockman W 2009 SLUO/LHC Workshop

[18] Brun R and Rademakers F 1996 *Proc. Fifth International Workshop on New Computing Techniques in Physics Research (Lausanne)*
Brun R and Rademakers F 1997 *Nucl. Instrum. Methods* A **389** 81-86
http://root.cern.ch

[19] The ALEPH Collaboration 1995 *Nucl. Inst. Meth.* A **360** 481-506

[20] Voloshin S A, Koch V and Ritter H G 1999 *Phys Rev* C **60** 024901

[21] Jinghua F and Lianshou L 2003 *Phys Rev* C **68** 064904