Technical Report

Application of the Dortmund Spectrum Estimation Algorithm to LHCb Monte Carlo Simulations

Tim Ruhe, Margarete Schellenberg, Bernhard Spaan

04/2018

technische universität dortmund

SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

Part of the work on this technical report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C3,C5.

Speaker: Prof. Dr. Katharina Morik
Address: TU Dortmund University
Otto-Hahn-Strasse 12
D-44227 Dortmund
Web: http://sfb876.tu-dortmund.de

1 Introduction

The reconstruction of experimentally inaccessible quantities, e.g. a particle's energy, from other observables, is a common challenge for experiments in particle- and astroparticle physics, as well as in other research areas. For the decay and interaction of particles, for example, where the underlying physics is governed by stochastical processes, this corresponds to solving an inverse problem, described by the Fredholm integral equation of the first kind.

$$g(y) = \int_a^b A(x,y) f(x) dx \,, \tag{1}$$

where $f(x)$ describes the distribution of the experimentally inaccessible quantity x, whereas $g(y)$ is the distribution of an observable y, obtained experimentally. $A(x,y)$ is generally referred to as the response function and includes the physics of a particle decay or interaction and all detector effects, such as a limited acceptance or additional smearing. In most cases the response function has to be obtained from Monte Carlo simulations. Obtaining a solution to Eq. 1 is often referred to unfolding or deconvolution.

Several algorithms for solving Eq. (1) exist. The most common ones are singular value decomposition (SVD) [1], iterative Bayesian unfolding [2, 3] and *TRUEE* [4], which is based on the popular *RUN*-algorithm [5] and uses Tikhonov-regularization [6]. Iterative Bayesian Unfolding and the SVD approach are also included in the RooUnfold package [7] for the ROOT analysis toolkit [8]. All of these algorithms have in common that – although the distribution $f(x)$ is reliably reconstructed – the information on the individual events is lost in the unfolding process. This lost information might, however, be valuable for the extraction of physically relevant parameters. Furthermore, the number of variables used in the unfolding is limited in most cases, which implies that the information available for unfolding is constrained as well.

The Dortmund Spectrum Estimation Algorithm (*DSEA*) aims at overcoming the aforementioned challenges, by treating a discretized version of Eq. (1)

$$\vec{g}(y) = A(x,y)\vec{f}(x) \tag{2}$$

as a multinominal classification task. In that case $\vec{g}$ and $\vec{f}$ are histogrammed versions of the respective distributions and $A(x,y)$ is the so-called response matrix. Approximating $f(x)$ by $\vec{f}$ is suitable for most practical applications, in case a sufficient number of bins is used. In *DSEA* every bin in $\vec{f}$ is interpreted as a specific class of events. The corresponding classification task can then be solved by an – at least in principle – arbitrary classifier, which returns the probability c_{jk} of an event k to belong to bin j in $\vec{f}$. The final bin content of f_j is then obtained by adding all c_{jk}. In order to avoid a potential bias on the input distribution *DSEA* can be used iteratively, using a uniform distribution as input. For algorithmic details on *DSEA* we refer to [9].

The versatility of the *DSEA* algorithm is tested using data recorded at the LHCb experiment. The LHCb experiment is one of the four large experiments operated at the Large Hadron Collider near Geneva, Switzerland. One of its main goals is the research on the asymmetry of matter and anti-matter in our universe. According to Big Bang theories matter and anti-matter have been produced in equal amounts 13.8 billion years ago.

However, in today's universe there is no evidence of large quantities of antimatter. Hence, it is assumed that physical laws influence matter and anti-matter in different ways. In order to find answers to this and other open questions, physicists are investigating decays of hadrons containing b and c quarks at the LHCb experiment. Precision measurements of observables from the flavour sector of the Standard Modal are performed to carry out an indirect search for New Physics. In order to maximise the sensitivity with respect to these goals, the LHCb detector is designed as a single-arm forward spectrometer (see Figure 1). At the interaction point of the proton beams, which lies within the Vertex Locator, a large amount of particles is created by many different physical processes. These particles decompose into new particles that fly through the detector and interact with the detector material. Various subdetectors are responsible for the reconstruction and identification of these particles. Entire decay chains are reconstructed by combining tracks and tracing them back to their heavier parent particles. Tracks of charged particles are bent by the magnet. The curvatures of the trajectories allow to determine the momenta of the particles. However, the detector can locate tracks with a limited precision. This leads to a limited momentum resolution, which is about $\Delta p/p = 0.5\%$ at low momenta and up to 0.8% at momenta around 100 GeV/c [10]. Experimental limitations like the momentum resolution have the consequence that distributions do not reflect the truth. For example, measured invariant masses have a resolution in the range of 10–20 MeV/c^2 [10]. As a test of the *DSEA* algorithm, the invariant mass of a K^{*0} meson, which is reconstructed through its decay into a charged kaon and pion, is considered. The nominal width of the K^{*0} resonance is 47.3 ± 0.5 MeV/c^2 [11], which is large compared to the mass resolution at LHCb.

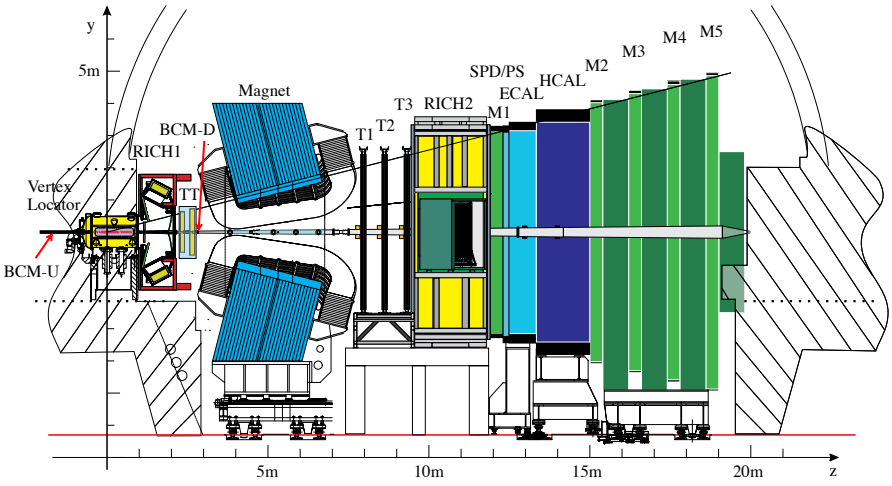


Figure 1: Scheme of the LHCb detector, illustrating the various subdetectors for identification and reconstruction of particles and their tracks.[12]

The paper is organized as follows. Section 2 discusses the convergence of *DSEA*, whereas the performance of the algorithm with respect to the agreement with the underlying distribution is presented in Sec. 3. The paper is concluded with a summary and an outlook in Sec. 4.

2 Convergence

Using *DSEA*, the reconstructed spectrum is obtained iteratively, which implies that the overall convergence of the algorithm is crucial for the success of the unfolding procedure. Following an approach presented in [13], the convergence of *DSEA* is quantified by comparing spectra obtained in two succeeding iterations via a χ^2-test. The χ^2-value for the k-th iteration is obtained using the following equation:

$$\frac{\chi^2}{n_{\text{bins}}} = \frac{1}{n_{\text{bins}}} \sum_{i=1}^{n_{\text{bins}}} \frac{(\hat{f}_{k,i} - \hat{f}_{k-1,i})^2}{\sigma_{k,i}^2}\,, \tag{3}$$

where $\hat{f}_{k,i}$ and $\hat{f}_{k-1,i}$ represent estimates for the i-th bin, obtained in the k-th and $(k-1)$-th iteration, respectively. The 1-σ uncertainties on the content of bin i are given as $\sigma_{k,i}$. For the systematic studies presented here, the convergence of *DSEA* was investigated with respect to the number of bins, the number of input examples and the confidence threshold selected to suppress contributions of small confidence values. The outcome of the studies are presented in Figs. 2 to 4 and discussed in the text.

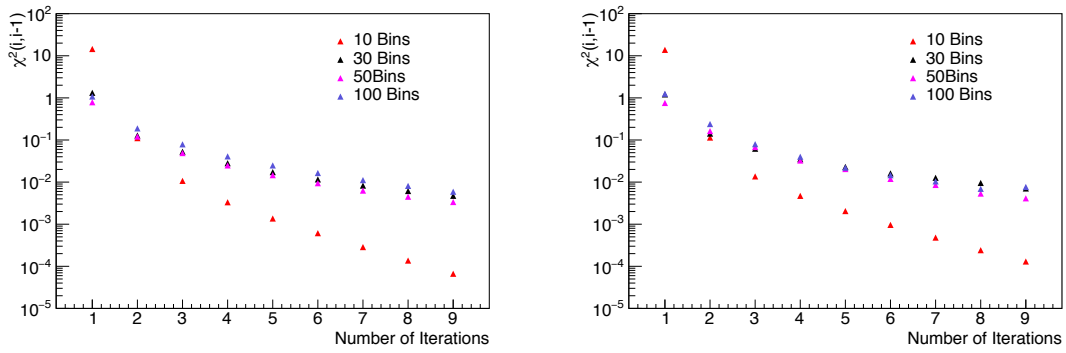
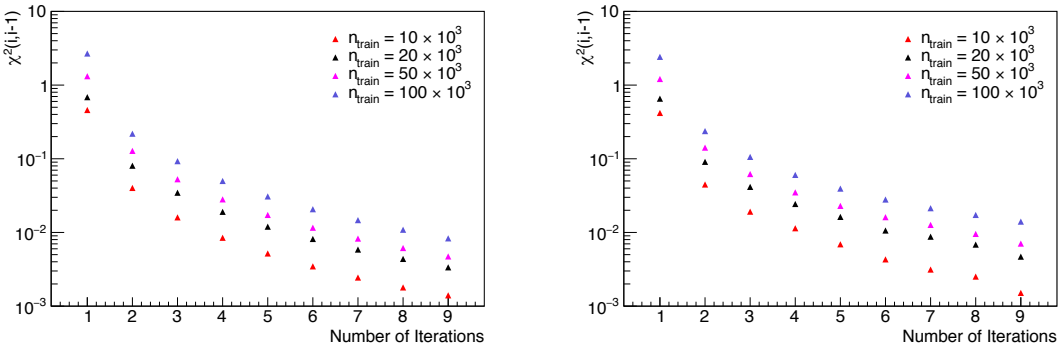


Figure 2: Convergence of *DSEA* for different numbers of bins, using 50,000 input examples and a confidence threshold of 0.01 (left) and 0.05 (right).

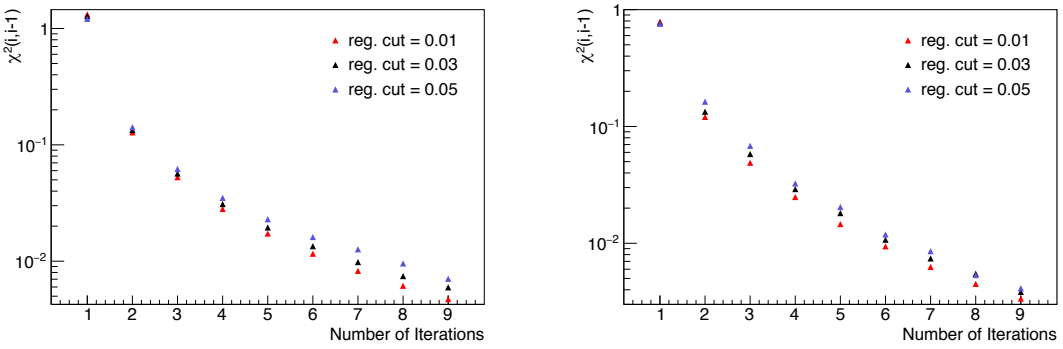
Figure 2 depicts the convergence of *DSEA* for different numbers of bins in the reconstructed spectrum, using 50,000 input examples. The results obtained with a confidence threshold of 0.01 are shown in Fig. 2a, whereas the convergence for a confidence threshold of 0.05 are presented in Fig. 2b. One finds that the algorithm converges reliably for both confidence thresholds, as the convergence criterion is met after at most two iterations, independent of the number of bins. It is further found that *DSEA* converges the fastest, when only 10 bins are utilized, whereas the convergence behaviour for 30, 50 and 100 bins is found to be comparable. Comparing the χ^2-values the algorithm converges to, with respect to the confidence thresholds investigated here one finds that this value is slightly smaller for a confidence threshold of 0.01.

Figure 3 presents the convergence of *DSEA* for different numbers of input examples n_{train}, studied for confidence thresholds of 0.01 (Fig. 3a) and 0.05 (Fig. 3b). It is found that



(a) (b)

Figure 3: Convergence of *DSEA* for different numbers of input examples, using 30 bins and a confidence threshold of 0.01 (left) and 0.05 (right).



(a) (b)

Figure 4: Convergence of *DSEA* for different confidence thresholds, using 50,000 input examples and 30 bins (left) and 50 bins (right).

the convergence criterion is met after at most two iterations, independent of the number of input examples, and therefore concluded that the algorithm converges reliably. One further observes that the χ^2 value *DSEA* converges to, is smaller for a smaller number of input examples.

Figure 4 shows the convergence of *DSEA* for different confidence thresholds, evaluated for 30 (Fig. 4a) and 50 bins (Fig. 4b). The convergence for a confidence threshold of 0.01 is shown in red, whereas the behaviour for confidence thresholds of 0.03 and 0.05 are depicted in black and blue, respectively. One finds that the convergence criterion is met after at most two iterations, independent of the selected confidence threshold and therefore it is concluded that the algorithm converges reliably. It is further found that *DSEA* converges slightly faster for smaller confidence thresholds. The χ^2-value the algorithm converges to, however, is found to differ only marginally.

3 Performance

Within this section the performance of *DSEA* with respect to the agreement with the underlying distribution is investigated. This agreement is quantified by means of the Hellinger distance, which, for the discrete case, is defined as:

$$H(f, \hat{f}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n_{\text{bins}}} (\sqrt{f_i} - \sqrt{\hat{f}_i})^2}\,, \tag{4}$$

where f and $\hat{f}$, represent the underlying and the reconstructed distribution, respectively. The smaller $H(f, \hat{f})$, the better the agreement between f and $\hat{f}$. Unlike other distance measures, e.g. a χ^2 distribution, the Hellinger distance does not take into account the estimated uncertainties. Using a χ^2 as a distance measure, was studied, but found to be rather unsatisfying with respect to finding an optimal setting for *DSEA*, as reconstructed spectra with large uncertainties were found to be favoured, due to the definition of the χ^2-distribution. Section 3.1 discusses the selection of optimal input parameters via an L-Curve, whereas exemplary spectra are discussed and compared to the underlying distribution in Sec. 3.2.

3.1 L-Curve

Unsurprisingly, the unfolding results depend on the input settings of the algorithm. Thus, finding a set of parameters, yielding an optimal unfolding result, is crucial for the success of the entire unfolding procedure. In addition to an optimal agreement with the underlying distribution, the algorithm should also have converged, which means that changes in the reconstructed spectrum are small between succeeding iterations. The overall performance of the algorithm is studied using so-called L-curves, where the agreement in terms of a distance measure (y-axis) is plotted versus the convergence criterion of the algorithm (x-axis). For the study at hand the Hellinger distance (see Eq. (4)) was used as a distance measure and the χ^2 between succeeding iterations was used as a convergence criterion (see Sec. 2). Selecting an optimal set of parameters generally implies finding a trade-off between agreement and convergence, as both criteria need to be simultaneously fulfilled. This includes the specification of a fixed number of iterations, which provides a stopping criterion.

As *DSEA* was found to converge reliably (see Sec. 2) the χ^2 between succeeding iterations generally decrease with an increasing number of iterations. It should therefore be noted that the number of iterations decreases from left to right in the L-Curve plots.

Figure 5 shows the Hellinger distance between the reconstructed spectrum and the underlying true distribution vs. the convergence, quantified as in Sec. 2. Different numbers of bins are compared using 50,000 input examples. Values obtained for a confidence thresholds of 0.01 and 0.05 are depicted in Fig. 5a and Fig. 5b. One finds that except for the 10 bins case, the optimal performance, indicated by small values of the Hellinger distance, is reached after the second (50- and 100 bins) or third (30 bins) iteration. In

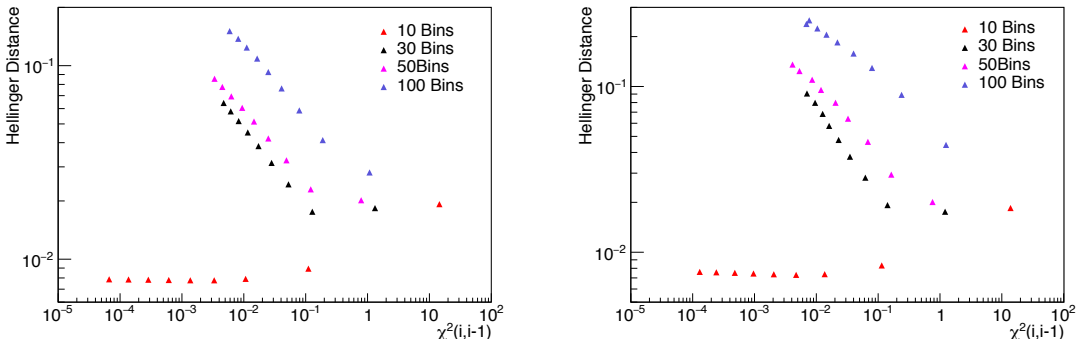


Figure 5: Hellinger distance vs. performance for different number of bins using 50,000 input examples and confidence thresholds of 0.01 (left) and 0.05 (right).

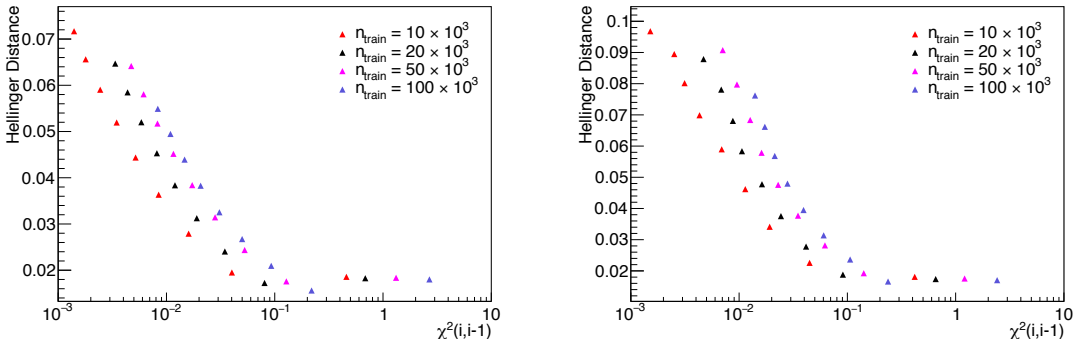
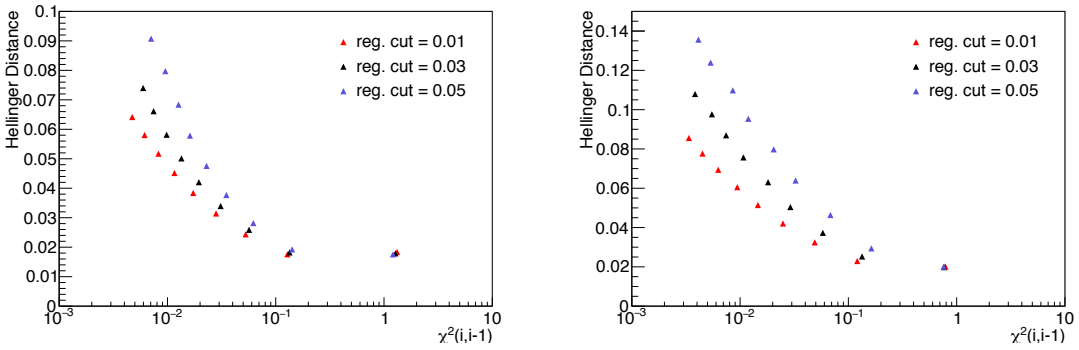


Figure 6: Hellinger distance vs. convergence for different numbers of input examples, using 30 bins and confidence thresholds of 0.01 (left) and 0.05 (right).

those cases an increase in the Hellinger distance, which indicates a decrease in the agreement of the compared distributions, is observed. For the 10 bins case, however, one finds that the algorithm reliably converges towards an optimal agreement with the underlying distribution. This can be understood from the fact, that a fixed number of 50,000 input examples was used, independent of the number of bins. When using for example 100 bins instead of 10, the number of examples available per bin is decreased by a factor of 10. Comparing Figs. 5a and 5b one further finds that the agreement is slightly better for a confidence threshold of 0.01.

Figure 6 depicts the Hellinger distance between the reconstructed spectrum and the underlying distribution vs. the convergence, for different numbers of input examples, obtained using 30 bins. Values obtained using confidence threshold of 0.01 and 0.05 are shown in Figs. 6a and 6b, respectively. One finds that an optimal performance is reached after two ($n_{\text{train}} = 10,000$ and $n_{\text{train}} = 20,000$) or three ($n_{\text{train}} = 50,000$ and $n_{\text{train}} = 100,000$) iterations. An increase of the Hellinger distance is observed with an increasing number of iterations. Therefore, a better agreement of the reconstructed spectrum with an increasing number of input examples available per class is observed



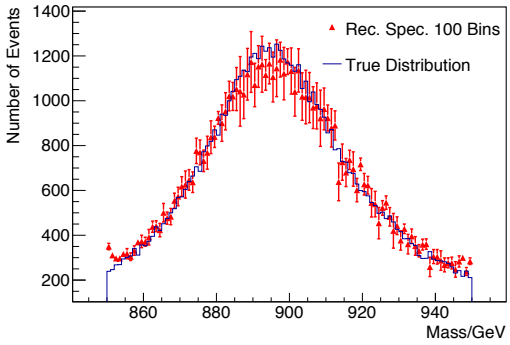
(a) (b)

Figure 7: Hellinger distance vs. convergence for different confidence thresholds using 50,000 input examples and 30 (left) and 50 bins (right).

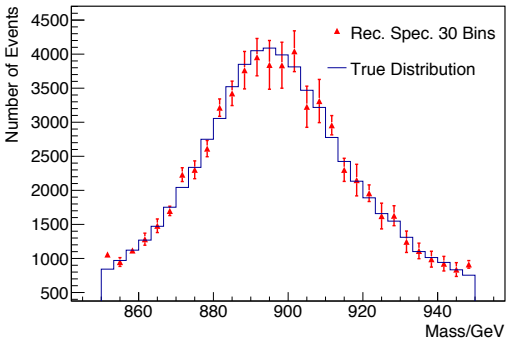
again. Comparing Figs. 6a and 6b one further finds that a slightly better agreement is achieved by choosing a confidence threshold of 0.01 (please note the different scales of the y-axes).

Figure 7 presents the Hellinger distance between the reconstructed spectrum and the underlying distribution vs. the convergence of DSEA for different confidence thresholds, using 50,000 input examples. The results obtained using 30- and 50 bins are shown in Figs. 7a and 7b. Again, one finds that an optimal performance, is reached after two (50 bins, Fig. 7b) or three (30 bins, Fig. 7a) iterations. An increase of the Hellinger distance is observed with an increasing number of iterations. One further finds that the values obtained for the Hellinger distance are rather similar, after two and three iterations, but deviate further and further as the number of iterations increases. Comparing Figs. 7a and 7b, one finds that a slightly better agreement is reached, when selecting 30 bins. This can be understood from the fact that a fixed number of 50,000 events was used in both cases, which leads to a 66% increase for the examples available per class, when 30 instead of 50 bins are selected.

Investigating the performance of *DSEA* with respect to different input settings, it was found that the algorithm generally reaches its optimal performance – indicated by a minimum in the Hellinger distance – after only two or three iterations. Except for the case, where only 10 bins were utilized, the Hellinger distance was found to increase, with an increasing number of iterations, which indicates a decrease in agreement. This behaviour is somewhat undesirable from an algorithmic point of view and will be discussed in greater detail in Sec. 4. As expected, the performance of the algorithm was found to increase with an increasing number of input examples available per bin. With respect to different confidence thresholds applied, it was found that the performance changes only marginally around the minimum of the Hellinger distance. Larger deviations were observed, as the overall performance decreases.

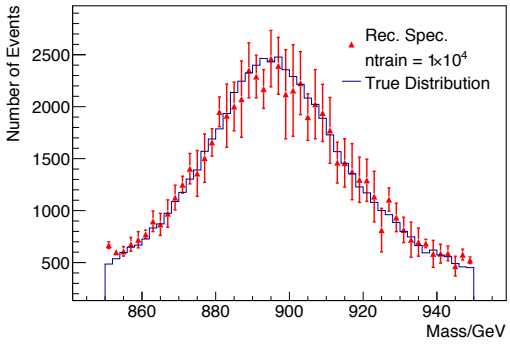


(a)

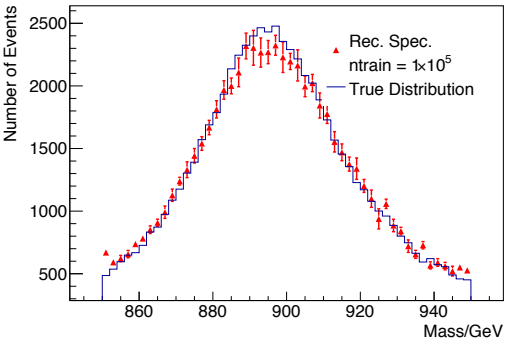


(b)

Figure 8: Comparison of the reconstructed spectrum (red) with the underlying distribution (blue) for 100 bins (left) and 30 bins (right), obtained using 50,000 input examples and confidence threshold of 0.01.



(a)

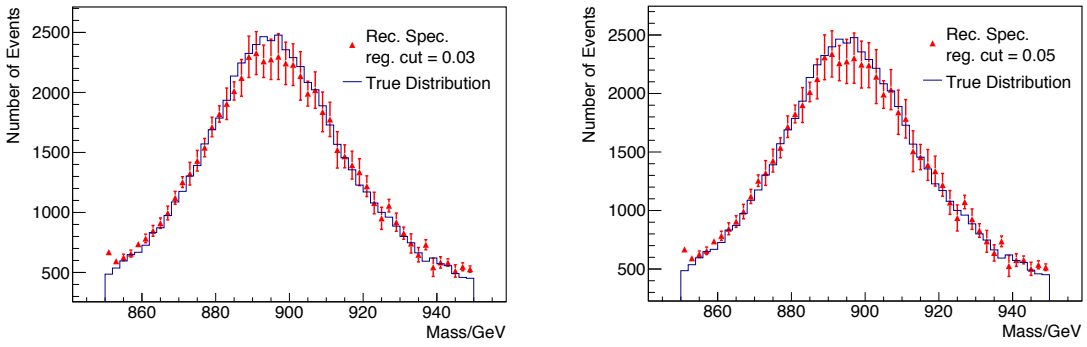


(b)

Figure 9: Comparison of the reconstructed spectrum (red) and the underlying distribution (blue), obtained for 50 bins and a confidence threshold of 0.03, using 10,000 (left) and 100,000 (right) input examples, respectively.

3.2 Exemplary Spectra

Figure 8 shows the comparison of the reconstructed spectrum (red) with the underlying distribution (blue). The depicted results were obtained using 50,000 input examples and a confidence threshold of 0.01. The outcome for 100 bins is shown in Fig. 8a, whereas the result obtained for 30 bins is presented in Fig. 8b. One finds that the distributions agree well in both cases. It is further observed that the unfolding result overestimates the bin content of the true distribution in the first and the last bin(s). Furthermore, the reconstructed spectrum has a tendency to underestimate the true bin content near the maximum of the distribution. Unlike for the first and last bins, however, this underestimation is generally well within the estimated uncertainties. In general, the reconstructed uncertainties were found to be smaller in the 30 bins example. As already discussed above, this can be understood from the smaller number of input events available per class in the 100 bin scenario.



(a) (b)

Figure 10: Comparison of the reconstructed spectrum (red) and the underlying distribution (blue) using 50 bins and 50,000 input examples, as well as confidence thresholds of 0.03 (left) and 0.05 (right), respectively.

Figure 9 depicts the comparison of the unfolding result and the underlying distribution, obtained using 50 bins and a confidence threshold of 0.03. The results obtained using 10,000 and 100,000 input examples are shown in Figs. 9a and 9b, respectively. The rising and falling edges of the distribution were found to be well reconstructed in both cases. Again, the first and last bins tend to be overestimated. When using 100,000 input examples, the reconstructed spectrum has a tendency to underestimate the true bin content near the maximum of the distribution. In addition to the generally good agreement, both spectra were found to agree well within the estimated uncertainties around the maximum, when using 10,000 input examples. As expected, the estimated uncertainties were found to be larger for the result obtained using 10,000 input examples.

Figure 10 shows the comparison of the underlying distribution and the unfolding result, obtained using 50,000 input examples and 50 bins. The outcome for a confidence threshold of 0.03 is presented in Fig. 10a, whereas the result for a confidence threshold of 0.05 is depicted in Fig. 10b. In both cases, the reconstructed spectrum was found to agree well with the true mass distribution. Again, the bin content appears to be underestimated for bins located near the upper- and the lower boundary of the histogram. The bins near the maximum of the distribution, as well as the rising and falling edges, however, were found to be reliably reconstructed. Furthermore, both spectra show similar features and the estimated uncertainties deviate only marginally. It is therefore concluded that for the study at hand the specific choice of the confidence threshold does not impact the reconstructed spectrum, as long as the other settings are selected close to the optimum. This conclusion is further supported by Figs. 7a and 7b, where the minimum of the Hellinger distance was found to deviate only marginally for different choices of the confidence threshold.

All in all it was found that the mass distribution of the decay of K^{*0}-mesons can be reliably reconstructed using Monte Carlo simulations from the LHCb experiment and the Dortmund Spectrum Estimation Algorithm *DSEA*. Spectra were successfully reconstructed using different settings and found to agree with the underlying distribution within the estimated uncertainties. As expected, the overall size of the uncertainties

was found to increase with a decreasing number of training events available per class. Furthermore, it was found that bins near the upper and lower edge of the spectrum are generally underestimated. This behaviour is of course undesirable in an unfolding, but can be circumvented by treating these bins as over- and underflow bins, respectively.

4 Conclusion and Outlook

This paper discussed the application of the Dortmund Spectrum Estimation Algorithm (*DSEA*) on Monte Carlo simulations of the LHCb experiment at CERN. The application focused on the reconstruction of mass spectra from the decay of K^{*0}-mesons in a mass regime between 850 and 950 GeV/c^2. The detailed investigations involved studies on the convergence behaviour of *DSEA*, as well as analyses on the overall agreement with the underlying distribution. Unlike studies on artificial data, so-called toy Monte Carlo, investigations on LHCb simulations provided insight into the applicability of the algorithm in a realistic setting, while at the same time allowing for detailed comparisons with the underlying distributions.

The Dortmund Spectrum Estimation Algorithm (*DSEA*) was investigated on its convergence behaviour with respect to the number of bins, the number of input examples and the selected confidence threshold. It was found that the convergence criterion is met after two to three iterations, independent of the selected settings. We therefore concluded that *DSEA* converges reliably, for analyses aiming at the reconstruction of the mass distribution in K^{*0}-decays.

It was found that mass spectra from the decay of K^{*0}-mesons can be reliably reconstructed in a mass regime between 850 and 950 GeV/c^2, using up to 100 bins. *DSEA* was found to converge reliably for all settings used in the study discussed here. Further observations showed that an optimal agreement with the underlying distribution was reached after only two to three iterations. Except for cases where only 10 bins were used for the reconstructed spectrum, the agreement was found to decrease with an increasing number of iterations. This behaviour is somewhat undesirable, as the optimal number of iterations needs to be select manually. Using a variable step width between the individual operations or combining *all* iterations in an ensemble are possible extensions of the algorithm, in order to avoid a manual selection of the optimal number of iterations. These possible extensions, however, will be investigated in future studies.

With respect to the number of bins and the number of examples used for the training of the classifier, it was found that these two parameters cannot be easily separated. The performance of *DSEA* was found to depend on the number of examples available for training per bin. This is understandable from a machine learning point of view, as a larger number of training examples per class generally leads to a better statistical description and a better classification performance.

In summary, it was found that *DSEA* is fully applicable to real-life problems, such as the reconstruction of spectra from the LHCb experiment and that the LHCb provides an excellent test bed for systematic studies on unfolding algorithms in general. Additional examinations on the algorithmic behaviour of *DSEA* are planned. These studies will

involve, for example, the reconstruction of mass spectra with decay widths below the experimental resolution of the LHCb detector. Such spectra appear as δ-peaks, which are generally hard to reconstruct in an unfolding, due to the applied regularizations, which assume a smooth behaviour of the sought after function $f(x)$. A successful application of *DSEA* to this type of distribution is expected to increase the energy resolution of the LHCb detector.

References

[1] A. Höcker and V. Kartvelishvili, "SVD approach to data unfolding," *Nuclear Instruments and Methods in Physics Research A*, vol. 372, pp. 469–481, Feb. 1996.

[2] G. D'Agostini, "A multidimensional unfolding method based on Bayes' theorem," *Nuclear Instruments and Methods in Physics Research A*, vol. 362, pp. 487–498, Feb. 1995.

[3] G. D'Agostini, "Improved iterative Bayesian unfolding," *ArXiv e-prints*, Oct. 2010.

[4] N. Milke and et al., "Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics," *Nuclear Instruments and Methods in Physics Research A*, vol. 697, pp. 133–147, Jan. 2013.

[5] V. Blobel, "The RUN manual," *Regularized unfolding for high-energy physics experiments. Technical Note TN361, OPAL*, 1996.

[6] A. Tikhonov, "On the solution of improperly posed problems and the method of regularization," *Sov. Math*, vol. 5, p. 1035, 1963.

[7] T. Adye, "Unfolding algorithms and tests using RooUnfold," *ArXiv e-prints*, May 2011.

[8] I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic, P. Canal, D. Casadei, O. Couet, V. Fine, L. Franco, G. Ganis, A. Gheata, D. G. Maline, M. Goto, J. Iwaszkiewicz, A. Kreshuk, D. M. Segura, R. Maunder, L. Moneta, A. Naumann, E. Offermann, V. Onuchin, S. Panacek, F. Rademakers, P. Russo, and M. Tadel, "ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization," *Computer Physics Communications*, vol. 180, pp. 2499–2512, Dec. 2009.

[9] T. Ruhe, M. Börner, M. Wornowizki, T. Voigt, W. Rhode, and K. Morik, "Mining for spectra – the Dortmund Spectrum Estimation Algorithm," in *Astronomical Data Analysis Software and Systems (ADASS XXVI)*, 2016. Accepted for publication.

[10] R. Aaij *et al.*, "LHCb detector performance," *Int. J. Mod. Phys.*, vol. A30, p. 1530022, 2015.

[11] M. Tanabashi *et al.*, "Review of particle physics," *Phys. Rev. D*, vol. 98, p. 030001, Aug 2018.

[12] R. Aaij *et al.*, "Letter of Intent for the LHCb Upgrade," Tech. Rep. CERN-LHCC-2011-001. LHCC-I-018, CERN, Geneva, Mar 2011.

[13] G. Choudalakis, "Unfolding in ATLAS," *ArXiv e-prints*, Apr. 2011.