Threshold Optimization and Variable Construction for Classification in the MAGIC and FACT Experiments

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften

von

Tobias Voigt

Technische Universität Dortmund

Fakultät Statistik

Eingereicht am 08. Juli 2014

Mündliche Prüfung am 09. Oktober 2014

1. Gutachter: Prof. Dr. Roland Fried
2. Gutachter: Prof. Dr. Claus Weihs

Contents

1 Introduction

Binary classification problems are quite common in scientific research. In very high energy (VHE) gamma-ray astronomy for example, the interest is in separating the gamma-ray signal from a hadronic background. The separation has to be done as exactly as possible since the number of gamma-ray events detected is needed for the calculation of the energy-dependent gamma-ray flux (energy spectrum) and the time-dependent flux (light curve) as dedicated physical quantities of an observed source (see e.g. Mazin, 2007). From these quantities source internal features can be inferred, as e.g. the underlying emission processes (e.g. Aharonian, 2004) or the size of the region emitting gamma-rays (e.g. Schlickeiser, 2002). In this situation, there are some distinctive features characterising the data one has to handle.

The most important one is that there is a huge class imbalance in the data. It is known that hadron observations (negatives) are more than 100 to 1,000 times more common than gamma events (positives) (Weekes, 2003; Hinton, 2009). The exact ratio, however, is unknown. Thus the ratio with which classification algorithms are trained is in general not the one in actual data which has to be classified. A second feature is that one is not primarily interested in a good classification, but in the estimation of the number of signal events in a data set. Assessing the quality of the estimation is difficult, since the influence of individual misclassifications on the outcome of the analysis is unknown. If the individual influence was known it could be used as individual misclassification cost and there would be the possibility to make the used classifier cost sensitive. The classifiers commonly used in VHE gamma-ray astronomy like random forests (Albert et al, 2008), boosted decision trees (Ohm et al, 2009; Becherini et al, 2011), or neural networks (Boinee et al, 2006), are reviewed in Bock et al (2004) and Fegan (1997). Throughout this work we favor random forests (Breiman, 2001), as usually done in the MAGIC experiment (MAGIC Collaboration, 2014). One effective method of making these cost sensitive is the thresholding method (Sheng and Ling, 2006). The idea of this method is to minimize the misclassification costs with respect to the discrimination threshold applied in the outcome of a classifier. Obviously, this method is inapplicable as we do not have individual misclassification costs, but as stated above in VHE gamma-ray astronomy one is not primarily interested in the best possible classification of any single event, but instead one wants to know the total number of gamma observations (positives) as this is the starting point for astrophysical interpretations. Statistically speaking this means estimation of the true number of positives based on a training sample. As we will show in the first part of this work, the mean square error (MSE) of this estimation can be regarded as misclassification risk in the thresholding method. So the idea is to choose the discrimination threshold which minimizes the MSE of the estimated number of positives in a data set. Additionally, the unknown class imbalance is taken into consideration.

For the analyses in this work we use data from the MAGIC and FACT telescopes. The two telescopes on the Canary island of La Palma are imaging atmospheric Cherenkov telescopes (IACTs). Their purpose is to detect highly energetic gamma rays emitted by various astrophysical sources. The data collection process can be summarized as follows: A highly energetic particle interacts with molecules in the atmosphere and induces a so-called air shower of secondary particles. These air showers emit blue light flashes, so-called Cherenkov light, which is collected by Cherenkov telescopes like MAGIC or FACT. The camera of an imaging Cherenkov telescope uses the Cherenkov light to create very pixelized side views of the induced air shower. From the images, various information can be inferred for example about the energy and angle of the inducing particle.

As stated above, only gamma rays are of interest in the FACT experiment, but also many other particles, summarized as hadrons, induce air showers, which are also recorded by the FACT telescope. Because of this, a separation of gamma rays (called signal in the following) and hadrons (called background in the following) is needed. The separation is done using the classification methods stated above (in this work random forests), for which variables are extracted from the shower images.

In very high energy gamma-ray astronomy, so-called Hillas parameters are used as variables for classification. These variables are based on moment analysis parameters introduced by Hillas (1985). Stereoscopic features introduced by Kohnle et. al. (1996) cannot be used in the FACT experiment as the experiment only uses one telescope. The original parameters introduced by Hillas (1985) are based on fitting an ellipse to the shower image and using its parameters such as its length and width as features. Hillas parameters were first introduced in 1985 and were not especially made to uncover differences between signal and background. It is thus desirable to find new or additional variables to improve the classification. The approach we are following in this work is to extend the idea of fitting an ellipse to fitting bivariate distributions, as they allow the use of more information than ellipses. We use several distance measures for distributions to measure the distance between the fitted distribution and the underlying empirical distribution of the shower image. The idea of this is to fit a distribution to the shower image which fits well to signal, but not to background events. In the second half of this work we describe our approach of distance based variable generation. We introduce the distributions we fit to the showers as well as the distance measures used. We then use a FACT data set to investigate the quality of the classification into signal and background applying the newly generated variables along with the Hillas parameters to a FACT dataset.

In the following we describe the two telescopes in Chapter 2, introduce the method to optimize the discrimination threshold in the outcome of a random forest by minimizing the MSE in Chapter 3 and present newly constructed classification variables in Chapter 4. The threshold optimization is combined with the new variables in Chapter 5. We conclude in Chapter 6.

2 Motivation

In this work, we use data from two astrophysical experiments, MAGIC and FACT. The underlying idea and theory is the same for both experiments, but technical details are different. Differences and similarities between the two experiments as well as the data which is available from them is introduced below. Many processes which happen in astrophysical sources of cosmic rays are still not known to astrophysicists, for example the processes going on in jets of Active Galactic Nuclei (AGNs). The detection and analysis of cosmic rays sent out by such sources, especially the energy spectrum of the rays emitted are informative for astrophysicists and help understanding the processes going on in the sources of cosmic rays. Especially gamma rays are helpful here, because they do not have an electric charge and are not influenced by magnetic fields. Therefore, when they reach earth, their origin can be reconstructed. The detection process is described below.

2.1 Air Showers and Cherenkov Light

In this work, we differentiate mainly between two forms of cosmic rays. The first one is highly energetic gamma rays, which are basically light photons, but with a very high energy. We also call these gamma rays signal events or positives. The other form of cosmic rays we consider in this work is summarized as hadrons. Hadrons are for example protons and neutrons, but also further particles like muons. Those hadrons form the background, also called negatives in our application, as the gamma rays are the interesting events here. Hadrons are only observed because of the way the gammas are detected.

The detection of gamma rays is an indirect process, in which the earth's atmosphere is used as a calorimeter. When a gamma ray enters the atmosphere, it interacts with the air molecules and induces a chain reaction, a so-called air shower (Grieder, 2010) of secondary particles. Simulated air showers are displayed in Figure 1. Air showers can be seen by Cherenkov telescopes by making use of Cherenkov light (Cherenkov, 1934), which is blue light emitted by air showers. The blue light is emitted in very short flashes, lasting only some microseconds, and cannot be seen by the human eye. Cherenkov telescopes, however, collect the blue light flashes and direct them into a camera, which creates images of the air showers. The imaging is done in such a way that the shape of a shower as seen in Figure 1 remains in the image. The camera and thus the quality of the images depends on the telescope one is using. The cameras of the MAGIC and FACT telescopes and some sample images are introduced in the following chapters.

The problem with the data collection process is that not only the interesting gammas induce particle showers. Hadrons also induce them and they emit Cherenkov light just like the gamma-induced showers, which is also collected by Cherenkov telescopes. To analyse gamma events, it is thus necessary to separate gamma rays from the hadronic background. The separation is done using classification techniques described in later Chapters.

The whole data collection process with Cherenkov telescopes can be seen in Figure 2.

2.2 MAGIC

The two MAGIC telescopes (MAGIC Collaboration, 2014), which are two of the largest Cherenkov telescopes in the world with a mirror diameter of about 17

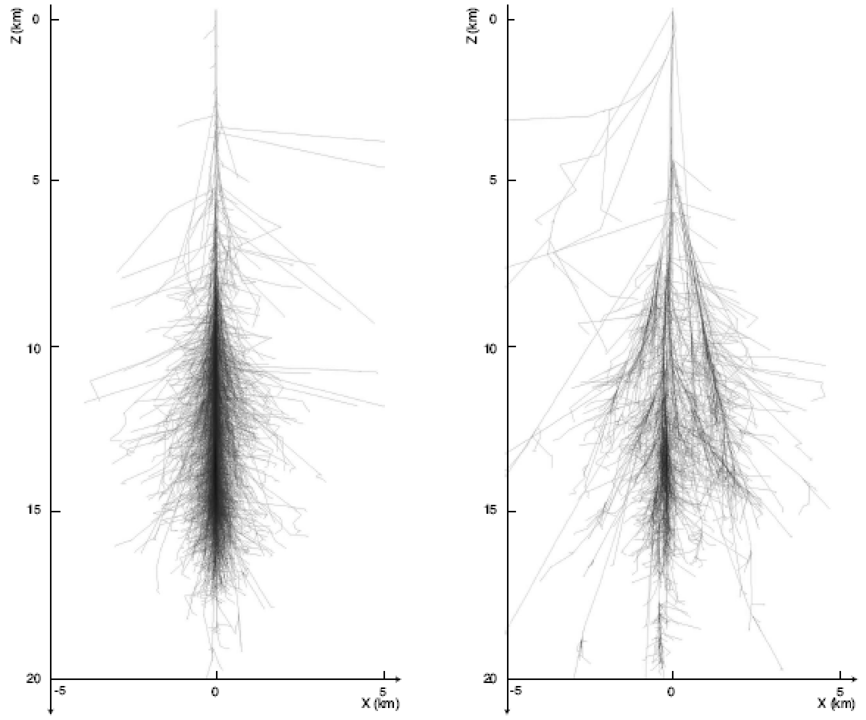


Figure 1: Simulated air showers of a gamma ray (left) and a proton (right)(Sidro Martin, 2008)

meters, are situated on the mountain Roque de los Muchachos on the Canary island of La Palma. In spite of being so large, the telescopes are highly mobile. They can be directed to every spot in the sky in about 25 seconds (Ferenc, 2005). Figure 3 shows the older MAGIC I telescope in the back and the newer MAGIC II telescope, which began operating in 2009 and features an improved camera (Hsu et al., 2007) and improved mirrors (Backes et al., 2007), in the front. As can be seen in the Figure, the telescopes mainly consist of parabolic mirrors, which collect and focus the Cherenkov light emitted from air showers, and a camera, which converts the collected light into photo electrons.

At the time of our analysis of MAGIC data, the MAGIC I camera consisted of 578 hexagonal photo multiplier tubes (PMTs, see e.g. Errando, 2006), but has been

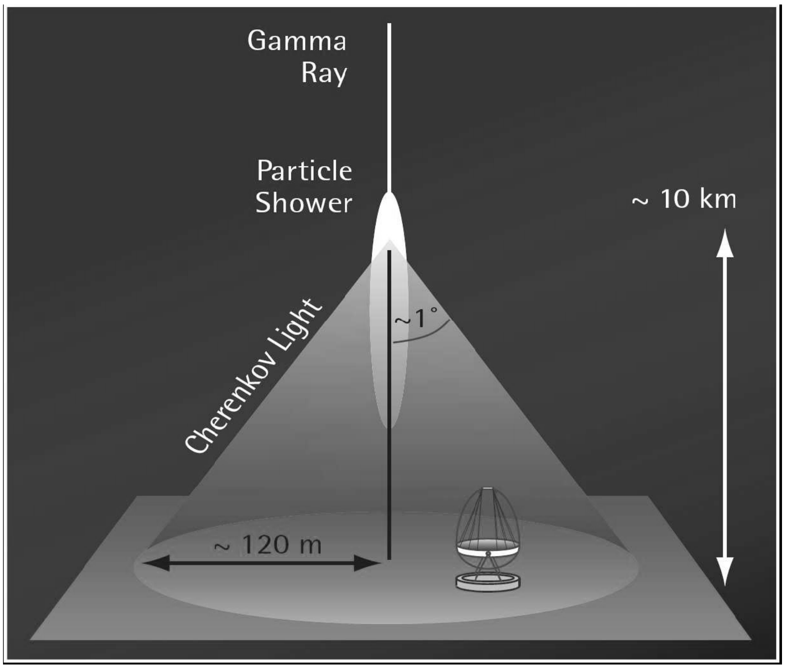


Figure 2: The data collection process with Cherenkov telescopes. A particle enters the atmosphere and induces a particle shower which sends out Cherenkov light. The light can be detected by a Cherenkov telescope. (Hadasch, 2008)

upgraded in 2012 to achieve full technical homogeneity between the two telescopes. Since then, both cameras consist of 1039 PMTs (Nakajima et al, 2013). We did not use the data from the new camera, as we used data from the FACT telescope at the time it was installed. This had several reasons, which we point out below.

PMTs convert light into photo electrons. The number of photo electrons triggered in a PMT depends on the intensity of the light registered in that PMT. To minimize the space between PMTs, they are complemented by hexagonal cones, so called Winston Cones. The combination of Winston Cone and PMT is called a pixel. A schematic view of the older MAGIC I camera and the new camera layout can be seen in Figure 4.

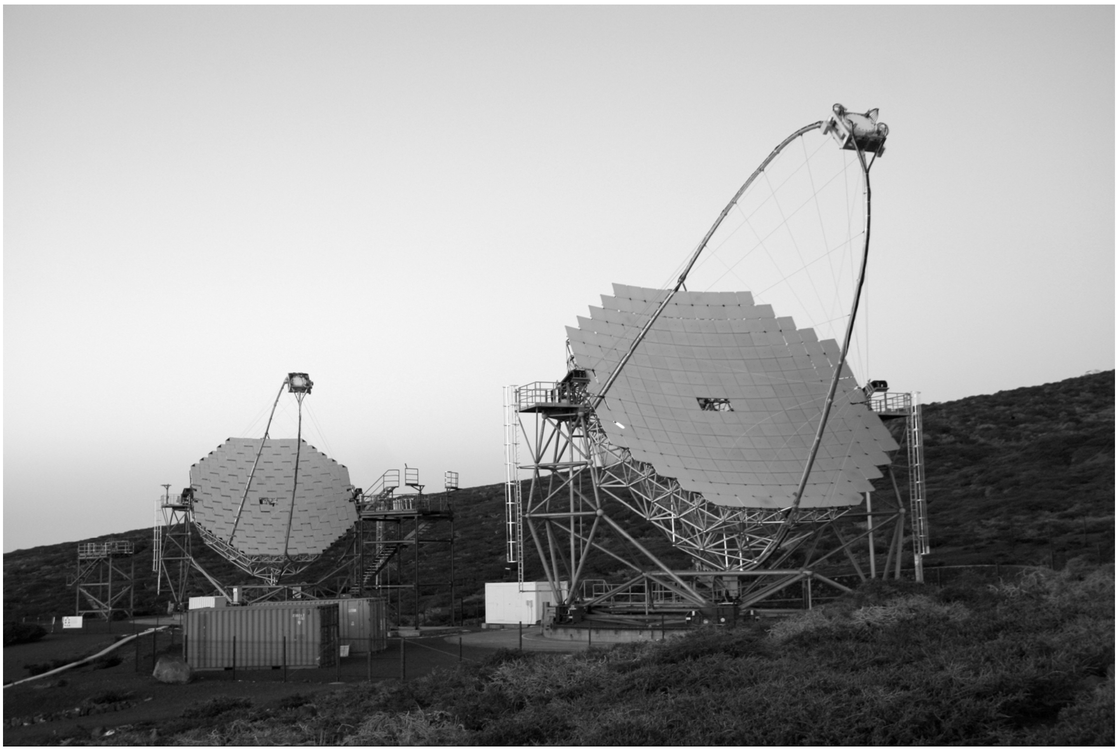


Figure 3: The two MAGIC telescopes. The older MAGIC I telescope (left) and the newer MAGIC II telescope (right) (MAGIC-Homepage, 2010).

2.3 FACT

The FACT telescope seen in Figure 5 is a single dish telescope also located on the MAGIC site. It is based on one of the former HEGRA telescopes and has a mirror plane of 9.5 sqm. The name FACT stands for **F**irst G-**A**PD **C**herenkov **T**elescope, which also describes its main characteristic: FACT is the first imaging Cherenkov telescope to use Geiger-mode avalanche photodiods (G-APDs) instead of photomultiplier tubes (PMTs), which need a lower voltage and have a higher photon detection efficiency (FACT Collaboration, 2014).

The FACT camera consists of 1440 G-APDs, that is, 1440 pixels. A schematic

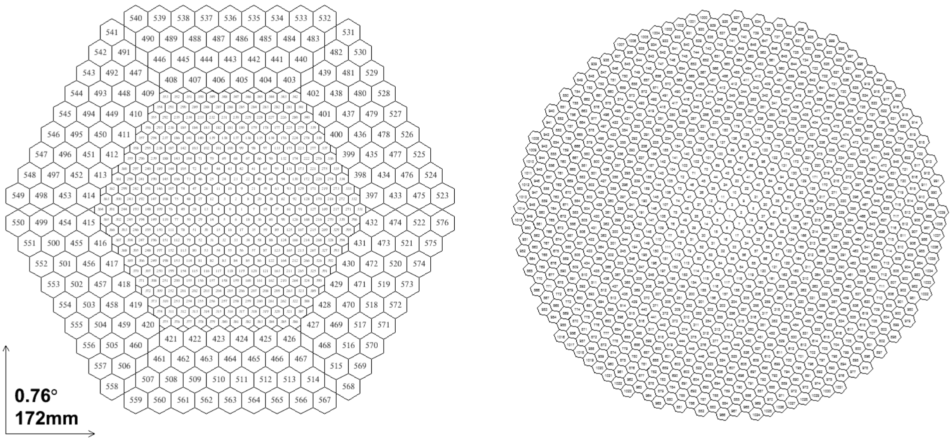


Figure 4: The cameras of the two MAGIC telescopes with numerized pixels. The older MAGIC I camera on the left and the newer MAGIC II camera on the right, which is also used in the MAGIC I telescope now. (MAGIC-Homepage, 2010).

view of the FACT camera with a simulated sample event can be seen in Figure 6.

Although technical details of the FACT telescope differ from the MAGIC telescopes, the aim and analysis of the experiments are basically the same: The detection of highly energetic gamma rays.

2.4 The Analysis Chain

The analysis chain used in the two expermiments MAGIC and FACT is basically the same. Besides the steps mentioned here, there are some more, which are mandatory and not interchangable and therefore not of interest here. For example the calibration of images.

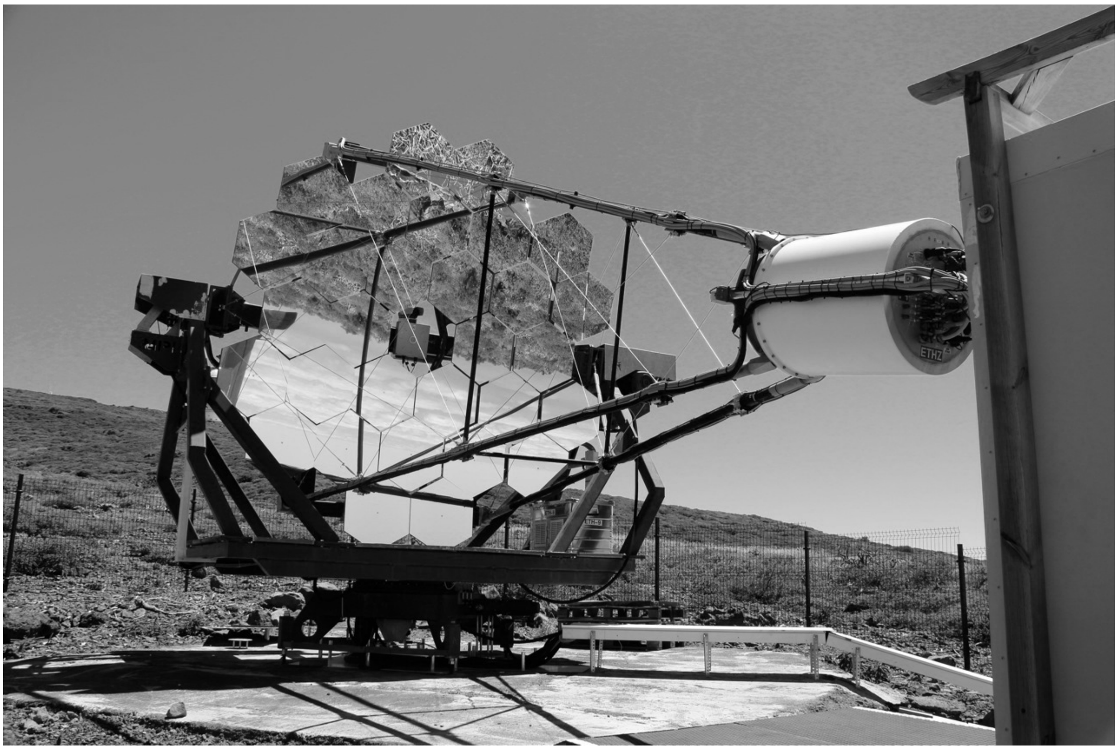


Figure 5: The FACT telescope. One can see the mirror plane and the camera embedded in the white metal cylinder. (MAGIC-Homepage, 2010).

2.4.1 Image Cleaning

When recording Cherenkov light with the MAGIC or FACT telescopes not only Cherenkov-photons are recorded. Background light emitted for example by nearby cities or the illuminated night sky (moon and stars) are also recorded. This background light is not to be confused with the hadronic background events. For better differentiation from background events we call the background light here noise. The noise is present in all camera images, regardless if it is a signal event or a background event. At full moon this noise can be so strong that taking data is almost impossible.

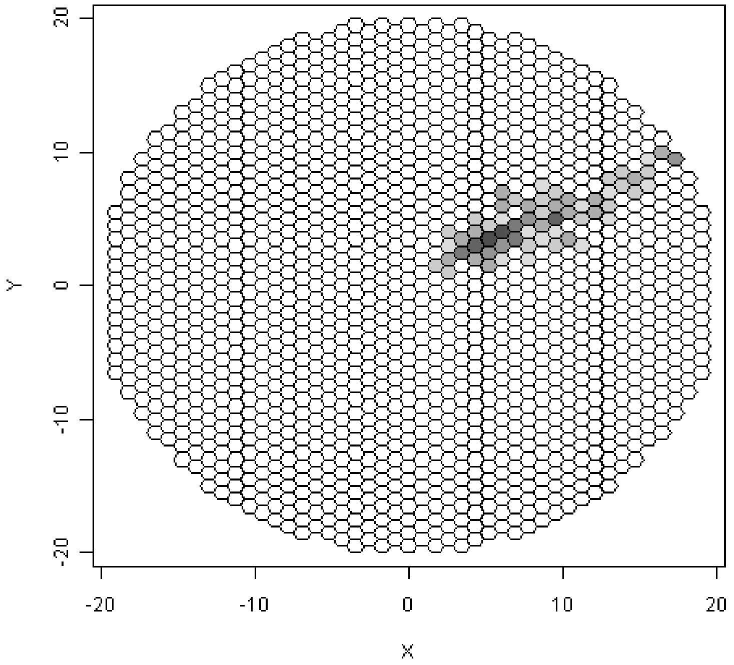


Figure 6: A sample image of an air shower induced by a simulated gamma ray. The greyscale shows the intensity of light in each pixel. The darker a pixel is, the more Cherenkov light photons have been collected.

Since the noise is always present, it is inherent in all images recorded. The actual shower is thereby overlapped by the noise and creates an image which has positive intensities in all pixels, with the intensity being largest where the actual shower is recorded. This can be seen on the left hand side of Figure 7. Would further steps of the analysis be used on this uncleaned picture, the results would be strongly biased.

For this reason a so-called image cleaning step is necessary. In this step, one tries to decide which of the pixels belong to the actual shower and which ones only contain noise. The intensity of pixels, which are thought to only contain noise, are set to 0. This can be seen on the right hand side of Figure 7. Here an image cleaning has been used to cut out the actual shower and suppress the noise.

Although the shower can usually be seen by the naked eye in most images, the image cleaning is not trivial. Especially on the edge of showers it is difficult to decide which pixels should be suppressed. Especially difficult is the image cleaning in images, which recorded events of smaller energy, because they emit less Cherenkov photons than events of higher energy, which leads to lower intensities in shower pixels.

The image cleaning in Figure 7 is the one typically applied in MAGIC, which is also used in FACT. The image cleaning done on the data we use consists of these steps (cf. Thom, 2009):

1. Determine Core Pixels: Pixels with an intensity larger than a threshold c_{high}
2. Determine Used Pixels: Pixels with an intensity larger than a threshold c_{low}, which are neighbours of Core Pixels.
3. Set intensity of all other pixels to 0.
4. Set intensity of isolated Core Pixels to 0, that is Core Pixels with neighbours all having intensity 0.

The thresholds c_{high} and c_{low} have to be chosen reasonably and are very different for the two telescopes. For MAGIC the standard values are $c_{high} = 8.5$ and $c_{low} = 4.5$. For FACT the values have to be much higher, because the G-APDs used are able to detect more light and thus also more background.

It is possible that there are some separate areas of intensity greater than 0 after the image cleaning. We call these areas islands. The biggest island in terms of used pixels is called Main Island, the other are called Sub Islands.

In an extended image cleaning also time information is considered.

We do not change the image cleaning in this work, but it should be kept in mind that the image cleaning is an important step in the analysis and shows large room for improvements. See Deiters (2013) for example for a comparison of filter techniques and their influence on the image cleaning and the further analysis.

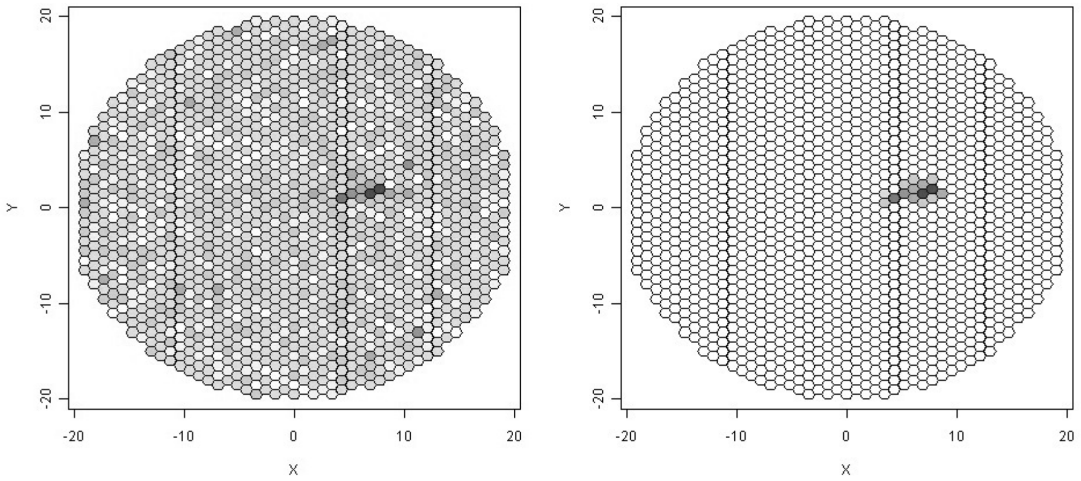


Figure 7: A sample image before (left) and after (right) the image cleaning. In this case the number of islands in the image is 1.

2.4.2 Quality Cuts

Quality Cuts (cf. Bretz, 2006) filter events, which are either impossible to classify (for example because they are too small) or which cannot possibly be signal events.

Only events which fulfill the following conditions are kept in the data set:

- Number of islands < 3. It is known that signal events cannot have more than two islands.

- Number of pixels > 5. The image must consist of more than five pixels with an intensity greater than zero after the image cleaning. Else it is too small to make any justifiable assertions about it.

- Leakage < 0.3. That means that more than 70% of the shower has to be within the camera plane. This is measured by the Hillas parameter Leakage, see below.

See Bretz (2006) for more, less intuitive and not interpretable quality cuts.

2.4.3 Variable Extraction

In order to classify the images into background and signal, proper classification variables have to be extracted from the raw images. The currently used variables are so-called Hillas parameters, which were first introduced by Hillas (1985) and extended for example by stereoscopic features introduced by Kohnle et. al. (1996) and variables describing the time evolution of a shower (e.g. Bastieri, 2005). The idea underlying the original Hillas variables is to fit an ellipse to a shower image and use the parameters of the ellipse along with some additional features as variables in the classification. This can be seen in Figure 8, where a few sample Hillas variables are displayed. The Hillas variables are explained in more detail in Chapter 4.2, where we also construct new variables to complement the currently used Hillas variables.

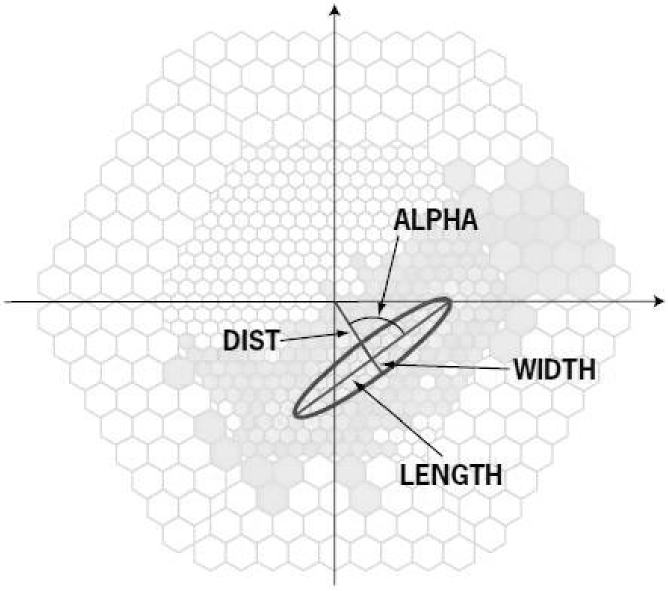


Figure 8: Some Hillas variables from fitting an ellipse.
(source: http://ihp-lx.ethz.ch/Stamet/magic/parameters.html)

2.4.4 Gamma-Hadron-Separation

With the Hillas variables extracted in the previous step, a classification is done to separate gamma ray events from hadronic background events. The separation is important, as a pure gamma sample is necessary to draw conclusions about the source one is observing.

Usually in the MAGIC and FACT experiments random forests (Breiman, 2001) with some minor changes described in Albert et al. (2008) are used for the classification.

The gamma-hadron-separation is the central aspect of this work. It is described in more detail in Chapter 3, where a method for optimizing the discrimination threshold of a random forest is introduced.

2.4.5 Unfolding of Energy Spectra

A central goal of the analysis of MAGIC and FACT data is to reconstruct energy spectra of the observed source, that is the distribution of the energy of gamma ray particles emitted by the source.

Energy spectra are usually estimated by histograms of the estimated energy of gamma particles. There are several uncertainties in estimating such a spectrum. For example the energy reconstuction of a particle is not exact, but only an estimation. Because of this the true energy of a particle is not always reconstructed accurately, so that a particle can be falsely sorted into a wrong energy range.

The most important uncertainty in this work comes from the difficult classification of signal and background. For the estimation of energy spectra, the data one estimates it from has to be a very pure gamma sample. Classification errors in the gamma-hadron-separation make it difficult to estimate the true energy spectrum and leads to a biased estimation. This means firstly that the classification has to be done as exactly as possible and secondly that the estimated spectrum has to be bias-corrected as long as the gamma-hadron-separation is not perfect.

The bias-correction is done using an unfolding procedure implemented in the unfolding program TRUEE (Milke et al., 2012). The idea of the unfolding is the following: The true energy distribution, or more exactly, its density $f(x)$, is assumed to be folded by a kernel function $A(y, x)$ into the observed distribution with density $g(y)$. This can be written using the Fredholm integral equation (Fredholm, 1903),

$$g(y) = \int A(y, x) f(x) dx + b(y)$$

where also a known background density $b(y)$ is added additively. Considering discrete or classed distributions this becomes a lot easier. Let $g_1, ..., g_k$ be the relative number of observations in the k observed energy classes and $f_1, ..., f_m$ the true probabilities of the m energy classes (the binning in the observed and true distributions can be different), then the above Fredholm integral can be rewritten using a $k \times m$-Matrix A and vectors $\mathbf{g} = (g_1, ..., g_k)^T$ and $\mathbf{f} = (f_1, ..., f_m)^T$:

$$\mathbf{g} = A\mathbf{f} + \mathbf{b}$$

This is a well-known linear model with the vector of observations **g**, design matrix A, parameter vector **f** and vector of errors **b**. Thus, the unfolding of the energy spectrum is an estimation of the vector **f**, where the vector **g** is observed and the transition matrix A is determined on simulated data and then used on real data. The estimation of **f** can be done using ordinary least squares, but experience shows that such estimations have large fluctuations. Therefore, more sophisticated approaches are used in TRUEE, such as the usage of B-Splines and a penalized minimization, which shows high resemblance to Ridge Regression.

The whole theoretical foundation of unfolding can be seen in Blobel (2010) and the exact procedure used in TRUEE can be seen in Milke (2012).

2.5 MAGIC and FACT Data

In this Chapter we describe the data we use in this work.

2.5.1 The FACT Data Set

The FACT data set we are using is a synthetic data set consisting of real background data and simulated signal. The real background is taken from the Crab Nebula and therefore consists also of gamma rays from this source. The background data is thus not only pure background, but also contains gamma events. However, the gamma count in this data is expected to be very small, as the signal to background ratio is usually about 1:1000 (Weekes, 2003). Additionally, real background also consists of some gamma rays, only that they are not from the source one is observing. Also, due to the relatively small number of events we use only a signal to background ratio of 1:100 in our analyses, so that the number of gammas in the background is so small that they should not make a difference. That is why using this data as background should not be much of a problem. The background data set consists of 12,768 events. The simulated data consists only of gamma events. The simulations are from an early stadium of simulations and represent data from the Crab nebula. The signal data set has 12,500 events.

The combined, synthetic data set thus consists of 25,283 events. However, this is before the quality cuts described in Chapter 2.4.2 are applied. After the application of quality cuts, 6,707 real background events and 6,472 simulated signal events remain in the data set, for a total of 13,179 events.

The FACT data set includes data from different analysis steps. We have uncleaned and cleaned camera pixel data available as well as the Hillas variables. That is why we can easily calculate new variables on this data set from the camera data and combine them with Hillas variables. However, the drawback of this data is that the data set is relatively small, so that it is difficult to make analyses with a real signal to background ratio. We also do not have information about the true or estimated energy of observations in this data.

2.5.2 The MAGIC Data Set

The MAGIC data we use here is synthetic data consisting of real background and simulated signal events. The data was taken by the two MAGIC telescopes operating in stereoscopic mode. The data set consists of 652,935 simulated gamma events and 357,850 hadronic background events, for a total of 1,010,785 events.

We only have Hillas parameters available in this data, along with a variable called estimated energy, which includes for each observation the energy estimated by a random forest regression (not to be confused with the random forest used for the gamma-hadron-separation).

2.6 High Class Imbalance and the Consequences

It is known that in real MAGIC and FACT data the ratio between signal and background is very unfortunate. Signal events are known to be about 100 to 1000 times less frequent than background (Weekes, 2003). This poses problems. First of all the classification becomes exceptional difficult and the search for gammas becomes looking for needles in a haystack. Especially often used measures of quality for classification like the accuracy, that is the ratio of correctly classified observations, are much less meaningful. With a ratio of signal to background of

1:1000, a classifier, which always classifies every event as background, will give an accuracy of 99.9%. The ratio also poses a problem for the reconstruction of energy spectra. The first step of the reconstruction is estimating how many signal events there are in the data. This becomes very difficult if too much background is classified as signal. ROC curves for example, which are often used in classification problems, are also not that meaningful here, because a larger area under the curve does not necessarily mean that the signal is estimated better.

We also have a problem with training data. As we do not know the exact signal to background ratio in real data, any number of signal and background events can or has to be used as training data, but the ratio has to be taken into account when assessing the quality of the classification. With a ratio of 1:1 in the training data for example, the number of falsely classified background events has to be very low on this data to achieve a pure sample of signal events on real data and keep the error of the estimation of the number of signal events as low as possible. That is why we want the falsely classified background to be in general very small, while we want to minimize the error we make on signal events.

2.7 Aim

The aim of this work is to improve upon the current MAGIC and FACT analysis chains. Some methods used in this analysis chain are either old for Cherenkov astronomy standards or heuristical or both.

The Hillas parameters for example were introduced by Hillas in 1985, which was at the very beginning of Cherenkov astronomy. They were a first attempt to describe shower images and were not specifically designed to find differences between the gamma ray signal and the hadronic background. In other words they were not constructed with the aim of the analysis in mind. As stated above, the aim of the analysis chain is the reconstruction of energy spectra. That means the whole analysis is only done for that aim to estimate the true energy spectrum of a source one is observing.

This has some consequences. In the gamma-hadron-separation step for example, classification methods (here usually random forests (Breiman, 2001; Albert et al., 2008)) are used to separate the gamma ray signal from the hadronic background. There are many quality measures for classification methods, for example simply the classification accuracy, which is the ratio of correctly classified observations. There is also the false positive rate (FPR) and true positive rate (TPR), also known as sensitivity and specificity, with the FPR being the ratio of falsely classified background observations and the TPR being the ratio of correctly classified signal observations. The two values are often considered together in ROC curves (Fawcett, 2006).

In our astronomical setting, however, where the overall aim of the analysis is to reconstruct energy spectra, considering only such quality measures can be misleading. For example increasing the accuracy of the classification even to very high values does not necessarily mean an improvement of the reconstruction of energy spectra. One reason for this is the very unfortunate signal to background ratio, introduced above.

The aim of this work is thus to improve upon the currently used analysis chain with respect to the objective of the analysis, that is the reconstruction of energy spectra. As we will see, the reconstruction can be considered as an estimation task. We can thus use quality measures for estimators, here especially the MSE, to measure the quality of our classification.

In a first step we use this MSE to optimize a discrimination threshold in the outcome of a random forest to improve the classification in such a way that the reconstruction of energy spectra is improved.

In a second step we construct new variables for the classification to improve upon the Hillas parameters. The new variables are constructed by fitting bivariate distributions to shower images and determining the distance between the observed and fitted distributions. As stated above, it is important in our context to maintain a very low FPR, for being able to properly estimate energy spectra. We investigate if the newly constructed variables improve upon the TPR while maintaining a very low FPR.

In a third step the threshold optimization and variable construction are used together. In this step we examine if the suggested methods in this work lead to improvements of the analysis chains of the two telescopes.

3 Threshold Optimization

In this chapter we optimize the threshold in the outcome of a random forest to improve upon the estimation of the number of signal events in a data set. We begin by stating the problem setup, then describe the method used to optimize the threshold and finally apply the method to synthetic MAGIC data. The contents of this chapter can also be seen in Voigt et al. (2014) and partly in Voigt & Fried (2013).

3.1 Problem setup

The problem we are facing is a binary classification problem. Some of the features described here are distinctive properties of our astrophysical setting. We have a random vector of input variables $\mathbf{X} = (X_1, ..., X_m)^T$ and a binary classification variable Y. We denote the joint distribution of $\mathbf{X}$ and Y with $P(\mathbf{X}, Y)$. We neither know this distribution nor can we make any justifiable assumptions about it. Additionally, in our application it is not possible to draw a training sample from the joint distribution. We are, however, able to draw samples from the conditional distributions $P(\mathbf{X}|Y = 0)$ and $P(\mathbf{X}|Y = 1)$. Thus, we have independent realizations $(\mathbf{x}_1, 0), ..., (\mathbf{x}_{m_{0\cdot}}, 0)$ and $(\mathbf{x}_{m_{0\cdot}+1}, 1), ..., (\mathbf{x}_{m_{1\cdot}+m_{0\cdot}}, 1)$ from the respective distributions with sizes $m_{0\cdot}$ and $m_{1\cdot}$, respectively, and $m = m_{0\cdot} + m_{1\cdot}$.

Based on these samples a classifier is trained. A classifier such as a random forest or logistic regression can be interpreted as a function $f : \mathbb{R}^m \rightarrow [0, 1]$. For a random forest for example the output of this function is the fraction of trees which voted for $Y = 1$. For a final classification into 0 and 1 we need a threshold c, so that

$$g(\mathbf{x}; c) = \begin{cases} 0, & \text{if } f(\mathbf{x}) \geq c \\ 1, & \text{if } f(\mathbf{x}) < c \end{cases}. \tag{1}$$

We consider f to be given and only vary c in this Chapter.

We call the above samples $(\mathbf{x}_1, 0), ..., (\mathbf{x}_{m_{0\cdot}}, 0), (\mathbf{x}_{m_{0\cdot}+1}, 1), ..., (\mathbf{x}_m, 1)$ the training data. As stated above, each of the m observations belongs to one of two classes, where $m_{1\cdot}$ is the number of observations in class 1 and $m_{0\cdot}$ is the number of observations in class 0. In the following, observations of class 1 are called positives and observations of class 0 are called negatives.
In addition to the training data, we have a sample of actual data to be classified, $\mathbf{x}_1^*, ..., \mathbf{x}_N^*$, for which the binary label is unknown. This data consists of N events with $N_{1\cdot}$ and $N_{0\cdot}$ being the unknown random numbers of signal and background events in this data. N, $N_{1\cdot}$ and $N_{0\cdot}$ are therefore random variables. This means our actual data, which we want to analyse, contains a random number of observations.

For a given threshold c, we denote the numbers of observations in the training data after the classification as $M_{ij}, i, j \in \{1, 0\}$, where the first index indicates the true class and the second index the class as which the event was classified by (1). We can display these numbers in a 2×2 table. With $N_{ij}, i, j \in \{1, 0\}$ defined analogously we get a similar 2×2 table for the actual data, see Table 1.

Table 1: True and classified numbers of positives and negatives in a training sample (left), in a sample of actual data (middle) and in Off-data (right).

		classified					classified					classified		
		1	0	$\sum$			1	0	$\sum$			1	0	$\sum$
true	1	M_{11}	M_{10}	$m_{1\cdot}$	true	1	N_{11}	N_{10}	$N_{1\cdot}$	true	1	0	0	0
	0	M_{01}	M_{00}	$m_{0\cdot}$		0	N_{01}	N_{00}	$N_{0\cdot}$		0	N_{01}^{Off}	N_{00}^{Off}	$N_{0\cdot}^{Off}$
	$\sum$	$M_{\cdot 1}$	$M_{\cdot 0}$	m		$\sum$	$N_{\cdot 1}$	$N_{\cdot 0}$	N		$\sum$	$N_{\cdot 1}^{Off}$	$N_{\cdot 0}^{Off}$	N^{Off}

It is obvious that we do not know the numbers in the first two rows of the table of the actual data as we do not know the true numbers of positives and negatives $N_{1\cdot}$ and $N_{0\cdot}$.

As we can see above, $N_{1\cdot}$ is considered to be a random variable and our goal is to estimate, or perhaps better predict, the unknown realization of $N_{1\cdot}$. The same applies for the number $N_{0\cdot}$. That is why we consider all the following distributions to be conditional on these values.

As explained above, we are only able to draw samples from conditional distributions. The difference between sampling from the joint distribution and sampling from the conditional distributions is that the marginal probability $P(Y = 1)$ can be estimated from samples from the joint distribution, but not from samples from the conditional distributions. This is because the samples from the conditional distributions are obtained from Monte Carlo simulations, where the sample sizes have to be set manually before the simulation.

To estimate the number of signal events $N_{1\cdot}$ we have additional information in the form of Off data available. In our astrophysical setting we collect Off data by observing a source-free position in the sky so that only background data is observed. This can be seen on the right hand side of Table 1. Technically speaking Off data represents another way to sample from the conditional distribution $P(\mathbf{X}|Y = 0)$. The sample size $N_{0\cdot}^{Off}$ in this data is assumed to be a random variable with the same distribution as the unobservable $N_{0\cdot}$. This is a reasonable assumption in our asrophysical setting, as background can be considered homogeneous, so that for equal observation time and collection area, the background should be the same, short of random effects. $N_{0.}^{Off}$ can therefore be used to assess the amount of background in our data, that is $N_{0.}$, and is used in estimators derived from a classification (see Chapter 3.2). Through classification of the $N_{0.}^{Off}$ Off-Events we get the number of correctly and falsely classified observations in this data, N_{00}^{Off} and N_{01}^{Off}, see Table 1. In our astronomical setting we get Off data by observing a source-free position in the sky, from which we know that no signal events are emitted. For more information about Off data see Chapter 3.4.

To summarize, we have three data sets:

- Training data, containing a manually fixed amount of positives, $m_{1.}$, and negatives, $m_{0.}$.
- Actual data, of which we want to estimate the number of positives $N_{1.}$.
- Off data, containing only negatives, where the number of negatives $N_{0.}^{Off}$ is random and has the same distribution as the number of negatives in the actual data.

Our main goal is to estimate the number of positives in the actual data, $N_{1.}$. This is not an easy task as $P(Y = 1)$ is known to be very small, of size around 1/100 or 1/1000. Because of this fact and because of the other characteristics of our application mentioned above, standard estimators of $N_{1.}$ or $P(Y = 1)$ will fail to estimate these values accurately.

We make some additional assumptions: We assume that the N_{i1}, the N_{01}^{Off} and the M_{i1}, $i \in \{1, 0\}$ are independent and (conditionally) follow binomial distributions. These assumptions also follow directly from an assumed multinomial distribution for the entries of Table 1:

$$N_{01}|_{N_{0\cdot}=n_{0\cdot}} \sim \text{Bin}\left(n_{0\cdot}, p_{01}\right), \tag{2}$$

$$N_{11}|_{N_{1\cdot}=n_{1\cdot}} \sim \text{Bin}\left(n_{1\cdot}, p_{11}\right), \tag{3}$$

$$N_{01}^{Off}|_{N_{0\cdot}^{Off}=n_{0\cdot}^{Off}} \sim \text{Bin}\left(n_{0\cdot}^{Off}, p_{01}\right), \tag{4}$$

$$M_{01} \sim \text{Bin}\left(m_{0\cdot}, p_{01}\right), \tag{5}$$

and

$$M_{11} \sim \text{Bin}\left(m_{1\cdot}, p_{11}\right), \tag{6}$$

with some probabilities p_{01} and p_{11}. As estimators for these two probabilities we use the True Positive Rate (TPR), which is also known as Recall or (signal) Efficiency, and the False Positive Rate (FPR)

$$TPR = \frac{M_{11}}{m_{1\cdot}} \tag{7}$$

and

$$FPR = \frac{M_{01}}{m_{0\cdot}}. \tag{8}$$

Obviously, both rates can only take values in the interval $[0, 1]$ and they depend on the discrimination threshold c, as M_{11} and M_{01} depend on it. When choosing

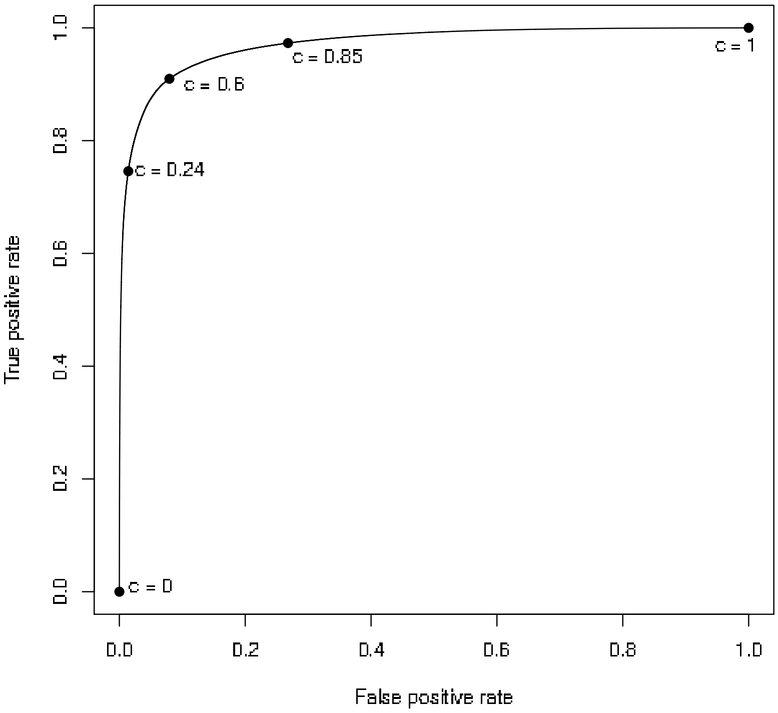


Figure 9: ROC curve for the MAGIC data. The higher in the top left corner the curve lies, the better is the classification. This curve shows that the classification in the MAGIC analysis chain is quite good. The problem is how to finally choose the threshold.

an appropriate c, a high value of TPR and a low value of FPR are desired. These two requirements, however, are contradictory, because a high value of the threshold leads to high values of TPR and FPR and vice versa. Figure 9 depicts a Receiver Operating Characteristic (ROC) curve (see Fawcett, 2006), which shows the TPR and FPR when altering the discrimination threshold. As we will see in the following Chapter, these two values are important in the estimation of the number of positives.

3.2 Estimation of the true number of positives

As stated above the estimation of the true number of positives $N_{1\cdot}$ is the main aim of our analysis. Depending on the individual problem such an estimator can look quite differently. For the given application of VHE gamma-ray astronomy, it is known that an irreducible portion of the recorded negatives will be misclassified (Sobczynska, 2007; Maier and Knapp, 2007). Thus, the number of positives is estimated via a difference estimator, as the number N_{01} of misclassified negatives in actual data can be approximated by the number N_{01}^{Off} of misclassified negatives in real Off data. According to Li and Ma (1983) the estimator

$$\tilde{N}_{1\cdot}^{\mathrm{LM}} = N - N_{0\cdot}^{Off} \tag{9}$$

is the Maximum Likelihood estimator in this situation if a Poisson distribution is assumed for $N_{1\cdot}$, $N_{0\cdot}$ and $N_{0\cdot}^{Off}$ and if we consider that the observation times of N and $N_{0\cdot}$ are the same and no classification is done. This estimator is unconditionally unbiased, $E\left(\tilde{N}_{1\cdot}^{LM} - N_{1\cdot}\right) = 0$, but a major problem of it is that its standard error is usually very high in case of a large class imbalance. The mean square error of this estimator is

$$\mathrm{MSE}\left(\tilde{N}_{1\cdot}^{\mathrm{LM}}\right) = \mathrm{E}\left(\left(\tilde{N}_{1\cdot}^{\mathrm{LM}} - N_{1\cdot}\right)^2\right) = \mathrm{Var}\left(N_{0\cdot}\right) + \mathrm{Var}\left(N_{0\cdot}^{Off}\right).$$

For example, if we assume independent Poisson distributions

$$N_{1\cdot} \sim Pois(\lambda_1),$$

$$N_{0\cdot} \sim Pois(\lambda_0),$$

and

$$N_{0\cdot}^{Off} \sim Pois(\lambda_0)$$

a signal to background ratio of 1:100 implies that the root mean square error of estimator (9) then is $\sqrt{100\lambda_1 + 100\lambda_1} = \sqrt{200\lambda_1}$. In case of $\lambda_1 = 1000$ the number of signal events would be estimated with a standard error of about 450, which is rather high. Additionally, the imbalance in our application is usually much worse than 1:100.

3.2.1 Estimation with known classification probabilities

The above estimation does not consider a classification before estimating the number of signal events. The idea of a classification in this situation is to suppress a large amount of the background events to receive a classified dataset with a more desirable signal to background ratio. If p_{11} was known, a corresponding estimator, which takes classification and the above mentioned binomial assumptions into account, would be

$$\tilde{N}_{1\cdot} = \frac{1}{p_{11}} \left(N_{\cdot 1} - N_{01}^{Off} \right). \tag{10}$$

Analogously to the estimator in equation (9), this quantity takes the difference between $N_{\cdot 1}$ and N_{01}^{Off} as an estimate for N_{11} and multiplies this with $\frac{1}{p_{11}}$ to compensate for the classification error in the signal events. It translates into the estimator in equation (9), when $p_{11} = p_{01} = 1$. Per construction this quantity can give negative results though $N_{1\cdot}$ is positive. Since we cannot achieve zero variance of the estimator, we cannot achieve an unbiased estimation when restricting its range to positive values. Because of the assumption that N_{01} and N_{01}^{Off} follow the same distribution we have $E(\tilde{N}_{1\cdot} - N_{1\cdot}) = 0$, justifying usage of $\tilde{N}_{1\cdot}$. Of course, when the results are to be interpreted in the application context, one might want to replace negative estimates by 0, but the corresponding estimator would be biased at least in case of zero or small values of $N_{1\cdot}$.

Obviously, the estimators in (9) and (10) cannot be used in problems where one does not have access to an Off data set. In this case, an alternative is for example the estimator $N_{\cdot 1}\frac{m_{1\cdot}}{M_{\cdot 1}}$. The following calculations have to be adapted to this estimator. In our application, however, this estimator cannot be used, as $m_{1\cdot}$ has to be set manually for the Monte Carlo simulations, so that using this estimator does not make sense.

Since we want to estimate the number of positives as exactly as possible we want to assess the quality of the estimator $\tilde{N}_{1\cdot}$. A standard measure of the quality of an estimator is the mean square error (MSE). As in applications we usually have fixed samples in which we want to estimate $N_{1\cdot}$, we calculate the MSE conditionally on $N_{1\cdot}$, $N_{0\cdot}$ and $N_{0\cdot}^{Off}$. The problem of high inaccuracy of the estimator $\tilde{N}_{1\cdot}^{\mathrm{LM}}$ remains in this situation, as its conditional bias is very high, although its variance is 0.

The MSE simplifies to the variance for unbiased estimators, so it is of interest if the estimator in (10) is unbiased for every fixed number of gamma events $N_{1\cdot}$, that is whether

$$\mathrm{Bias}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) = \mathrm{E}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) - N_{1\cdot}$$

equals 0, where $\mathrm{E}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right)$ is the conditional expectation of the estimator $\tilde{N}_{1\cdot}$ given the values of $N_{1\cdot}$, $N_{0\cdot}$ and $N_{0\cdot}^{Off}$.

Under the binomial assumptions made in the previous Chapter (equations (2) - (6)) we can easily calculate the conditional expectation of $\tilde{N}_{1\cdot}$:

$$\begin{aligned}
\mathrm{E}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) &= \mathrm{E}\left(\frac{1}{p_{11}}\left(N_{\cdot 1} - N_{01}^{Off}\right)|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) \\
&= \frac{1}{p_{11}}\mathrm{E}\left(N_{11} + N_{01} - N_{01}^{Off}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) \\
&= \frac{1}{p_{11}}\left(N_{1\cdot}p_{11} + N_{0\cdot}p_{01} - N_{0\cdot}^{Off}p_{01}\right) \\
&= N_{1\cdot} + \frac{p_{01}}{p_{11}}\left(N_{0\cdot} - N_{0\cdot}^{Off}\right)
\end{aligned}$$

The conditional bias of $\tilde{N}_{1\cdot}$ from equation (10) thus is

$$\mathrm{Bias}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) = \frac{p_{01}}{p_{11}}\left(N_{0\cdot} - N_{0\cdot}^{Off}\right). \tag{11}$$

This bias matches with what one would expect intuitively as it is small for p_{01} small and for p_{11} high and it reaches zero if no background is falsely classified or if the number of background events in the Off data is the same as in the actual data.

A standard measure for the quality of an estimator is the (conditional) MSE given by

$$\begin{aligned}
\mathrm{MSE}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) &= \mathrm{E}\left(\left(\tilde{N}_{1\cdot} - N_{1\cdot}\right)^2|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right) \\
&= \mathrm{Bias}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right)^2 + \mathrm{Var}\left(\tilde{N}_{1\cdot}|N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{Off}\right).
\end{aligned} \tag{12}$$

This variance term can easily be calculated by again using the assumption that

N_{11}, N_{01} and N_{01}^{Off} are independent and follow binomial distributions:

$$\begin{aligned}\operatorname{Var}\left(\tilde{N}_{1\cdot}|N_{1\cdot},N_{0\cdot},N_{0\cdot}^{Off}\right) &= \operatorname{Var}\left(\frac{1}{p_{11}}\left(N_{\cdot 1}-N_{01}^{Off}\right)|N_{1\cdot},N_{0\cdot},N_{0\cdot}^{Off}\right)\\ &= \frac{1}{p_{11}^2}\operatorname{Var}\left(N_{11}+N_{01}-N_{01}^{Off}|N_{1\cdot},N_{0\cdot},N_{0\cdot}^{Off}\right)\\ &= \frac{1}{p_{11}^2}\left(N_{1\cdot}p_{11}\left(1-p_{11}\right)+\left(N_{0\cdot}+N_{0\cdot}^{Off}\right)p_{01}\left(1-p_{01}\right)\right)\\ &= N_{1\cdot}\left(\frac{1}{p_{11}}-1\right)+\frac{p_{01}-p_{01}^2}{p_{11}^2}\left(N_{0\cdot}+N_{0\cdot}^{Off}\right)\end{aligned} \tag{13}$$

Using equations (13) and (11) the MSE in (12) becomes

$$\begin{aligned}\operatorname{MSE}\left(\tilde{N}_{1\cdot}|N_{1\cdot},N_{0\cdot},N_{0\cdot}^{Off}\right) &= \frac{p_{01}^2}{p_{11}^2}\left(N_{0\cdot}-N_{0\cdot}^{Off}\right)^2\\ &\quad + N_{1\cdot}\left(\frac{1}{p_{11}}-1\right)+\frac{p_{01}-p_{01}^2}{p_{11}^2}\left(N_{0\cdot}+N_{0\cdot}^{Off}\right).\end{aligned} \tag{14}$$

3.2.2 Estimation with Unknown probabilities

The above equation (14) depends on the values of p_{11} and p_{01}. As we do not know these values we have to estimate them. Consistent estimators for these values are TPR and FPR (equations (7) and (8)). Using TPR as an estimator for p_{11} in equation (10) we get the realistic estimator

$$\widehat{N}_{1\cdot}=\frac{m_{1\cdot}}{M_{11}}\left(N_{\cdot 1}-N_{01}^{Off}\right)=\frac{1}{TPR}\left(N_{\cdot 1}-N_{01}^{Off}\right). \tag{15}$$

As TPR and FPR are consistent estimators of p_{11} and p_{01}, and the sample sizes $m_{1\cdot}$ and $m_{0\cdot}$ are usually large ($> 10^5$), using the estimators instead of the true probabilities should only lead to a small difference. By estimating p_{11} with TPR and p_{01} with FPR in equation (14) we get the estimate

$$\begin{aligned}\widehat{\mathrm{MSE}}\left(\tilde{N}_{1\cdot}|N_{1\cdot},N_{0\cdot},N_{0\cdot}^{Off}\right) &= \frac{FPR^2}{TPR^2}\left(N_{0\cdot}-N_{0\cdot}^{Off}\right)^2 \\ &\quad + N_{1\cdot}\left(\frac{1}{TPR}-1\right)+\frac{FPR-FPR^2}{TPR^2}\left(N_{0\cdot}+N_{0\cdot}^{Off}\right).\end{aligned} \tag{16}$$

Note that $\tilde{N}_{1\cdot}$ is not a feasible estimator since p_{11} is unknown. Instead we propose $\widehat{N}_{1\cdot}$, but under our binomial assumptions neither the expectation nor the MSE of $\widehat{N}_{1\cdot}$ exist, since the term $E\left(\frac{1}{M_{11}}\right)$ is involved, which is infinite if $P(M_{11}=0)>0$. The binomial distribution, however, is only an approximation to reality as in practice we can simulate data until $M_{11}>0$. We find it hard to provide a realistic model for this process and use $\mathrm{MSE}\left(\tilde{N}_{1\cdot}|N_{1\cdot},N_{0\cdot},N_{0\cdot}^{Off}\right)$ as a measure of quality also for the estimator $\widehat{N}_{1\cdot}$ deduced from $\tilde{N}_{1\cdot}$. In our experience this gives good results, see Chapter 3.5.

As we see in the following Chapter, equations (15) and (16) can be used in an iterative manner to find an optimal discrimination threshold. Therein equation (16) is treated like the total misclassification costs in the thresholding method as value to be minimized over the threshold.

3.3 Minimizing the MSE to choose the threshold

As stated above, the values TPR and FPR directly depend on the discrimination threshold. If the number of positives $N_{1\cdot}$ was known we could minimize the MSE in equation (16) over all possible thresholds and thus find the one with which the

number of positives can be estimated best.
For simulations with a known ratio of positives and negatives, this is shown in Figure 10. It illustrates the MSE, calculated with equation (16) and depending on the threshold, for a fixed number of 5000 positives and different numbers of negatives. The solid curve, for a ratio of 1:50, that is 250000 negatives, is rather flat in comparison to the other curves, depicting ratios up to 1:1000. When looking at higher numbers of negatives one can see two major distinctive features. The first one is that the MSE is generally higher. This seems intuitive as it gets more difficult to separate signal and background when the background increases. The second feature is that the optimal value of the threshold approaches zero when the number of negatives is increased. An optimal threshold obviously depends on the class imbalance in the data, that is the ratio between positives and negatives, which is unknown in practice.

In the given astronomical application, this ratio can in general be different for different astrophysical sources the telescope is taking data from and even for different observation times of the same source. Thus using always the same threshold is not recommended. Instead it is desirable to adapt the threshold to the data and to the class imbalance in the data.
The problem here is that, as one knows the sum of positives and negatives N, knowing their ratio is equivalent to knowing the number of positives $N_{1\cdot}$, the value one wants to estimate. One needs to know $N_{1\cdot}$ to optimize the threshold to find the best estimate of $N_{1\cdot}$.
An intuitive idea to overcome this problem is to use a rough and easy to compute estimate of $N_{1\cdot}$ to optimize the threshold and then get a better estimate from a new classification. An easy to compute estimate can be obtained by setting the threshold to a fixed default value. According to Figure 10, $c = 0.1$ seems to be a reasonable initial choice. Using this value for the threshold we estimate $N_{1\cdot}$, then

calculate the optimal threshold for this estimate and estimate $N_{1\cdot}$ again for this new threshold. This procedure can be iterated until some convergence criterion is met. This approach leads us to the following algorithm:

1. Set an initial value c for the threshold.
2. Estimate $N_{1\cdot}$ using this threshold and equation (15).
3. Compute a new threshold through minimizing equation (16) over all thresholds using the estimates $\hat{N}_{1\cdot}$ for $N_{1\cdot}$ and $N - \hat{N}_{1\cdot}$ for $N_{0\cdot}$.
4. If a stopping criterion is fulfilled, compute a final estimate of $N_{1\cdot}$ and stop. Otherwise go back to step 2.

At this point we need a treatment for negative values of $\widehat{N}_{1\cdot}$, because negative estimates could lead to a negative estimate of the MSE. We then set negative values of (16) to 0. This happens very rarely, as the third term is positive and usually dominant. Please also note that minimization of the MSE takes place only over TPR and FPR, so the estimation of $N_{1.}$ does not diverge to $-\infty$, because the factors in the other two terms decrease faster in TPR than the second. As a stopping criterion, we require that the change in the threshold from one iteration to the next is below 10^{-6}.

In the following we refer to this algorithm as the MSEmin method. This method takes both into consideration: The class imbalance problem and the minimization of the MSE, that is, the overall misclassification cost. In the next Chapter we investigate the performance of this algorithm on simulated data and compare it to other possible approaches.

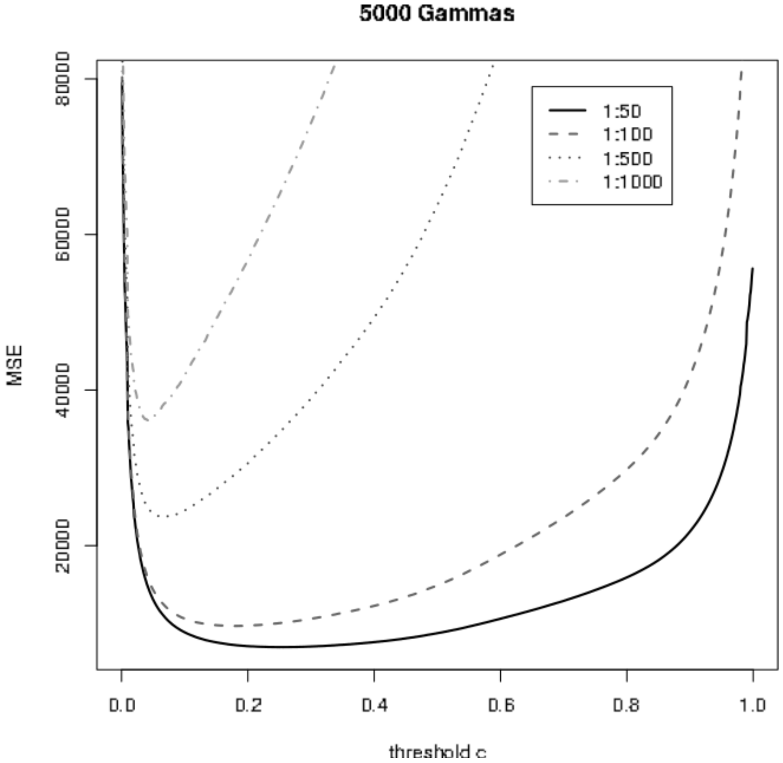


Figure 10: The MSE plotted against the threshold c for a fixed number of gamma events and different numbers of hadron-events according to several gamma-hadron-ratios

3.4 Application to astronomical data

In this Chapter we apply the MSEmin method to data from VHE gamma-ray astronomy collected with the MAGIC telescopes.

The MAGIC telescopes on the Canary island of La Palma are a stereoscopic imaging atmospheric Cherenkov telescope system. Its purpose is to detect highly energetic gamma-rays from astrophysical sources. This is done in an indirect way, employing the Earth's atmosphere as calorimeter: gamma-rays impinging the atmosphere induce cascades of highly energetic charged particles, emitting Cherenkov light (Cherenkov, 1934). For reviews of this research field see Hinton and Hofmann

(2009), Hinton (2009), and Chadwick et al (2008). One major difficulty is that not only gamma-rays induce such particle showers, but also many other particles summarized as hadrons, which are, for the strongest gamma-ray sources, 100 to 1000 times more common than the gamma-rays of interest (Weekes, 2003; Hinton, 2009). So the gammas, which are the positives in this application, have to be separated from the vast majority of hadrons, which represent the negatives. The separation was enhanced with a significant boost in sensitivity in 2009, when the second telescope was built and the telescope system began observation in stereoscopic mode (Aleksic et al, 2012).

The measurements are conducted in such a way that both the astrophysical sources of VHE gamma-rays and similar sky positions without (known) gamma-ray sources (Off data) are observed, which leads to N events in the data sample, and N^{Off} events in the Off data sample. For a time-efficient use of the telescopes these observations can even be conducted at once utilizing the false source method (Fomin et al, 1994).

In the MAGIC analysis chain (outlined in Chapter 2.4 or e.g. Firpo Curcoll et al, 2011) the classification is usually done by a random forest (Breiman, 2001; Albert et al, 2008). For the training of the random forest, Monte Carlo generated gamma-ray events are used as positive examples. The variables used for separation are based on the moment analysis parameters introduced by Hillas (1985), complemented by stereoscopic variables based on Kohnle et al (1996) and variables describing the time evolution of the shower images (Aliu et al, 2009). To obtain realistic distributions of these variables, the whole physical detection process is simulated, from the simulation of the particle showers and the emission of Cherenkov light (Heck and Knapp, 2010) up to the light detection and electronic digitalization process (Carmona et al, 2008; Majumdar et al, 2005). As the involved processes in hadronic shower development are much more complex and the telescopes' efficiency for recording hadron events is about a factor of 10 lower compared to gamma induced particle showers, the computation time for a sufficient sample of Monte Carlo simulated hadron events would require many CPU-hours of computation time for a similarly plentiful Monte Carlo sample of hadron events. Furthermore, the underlying particle physics processes are unaccessible for current precision measurements, so that predictions of different theoretical models differ by 20%–40% (Maier and Knapp, 2007). For these reasons, the negative examples for the training of the random forest are not Monte Carlo generated but taken from measured Off data.

The output of a random forest for each event is the fraction of trees which voted for this observation to be a hadron. A discrimination threshold c has to be applied in this fraction to finally classify each observation. From applying the trained random forest model f with a fixed discrimination threshold c on a test sample of n events, being comprised by $m_{1\cdot}$ Monte Carlo simulated gamma events and $m_{0\cdot}$ Off data hadron events, the M_{ij} and thus TPR and FPR can be inferred.

3.4.1 Energy-dependency

As for calculating energy spectra from the number of gamma events, the value of interest is not simply the total number of all gamma events by itself, but the number depending on the energy of the gamma-rays. Therefore we do not compute a single threshold for the whole data set. Instead the data is binned into several classes according to their energy and then the threshold is computed for every bin separately. This can lead to rather different values of the threshold in each energy bin. In reality, the energy of the gamma-rays is neither known nor directly measurable and thus has to be estimated, taking both the limited acceptance and limited energy resolution of the detector into account. This is done by applying an unfolding procedure (see Milke et al, 2011, 2012) and possibly also a random forest regression beforehand (Albert et al, 2007). Typically, this is conducted on a logarithmic-equidistant binning in energy, reflecting the detectors energy resolution and the power-law-like decline of event numbers depending on their energy. For this study, we neglect the unfolding procedure to not further complicate the issue and use the term energy for a random forest regression estimate of the true energy for simplicity.

3.4.2 Currently used Recall-methods

The currently used method to find the discrimination threshold, which is called hadronness cut in this application, is a rather simple one. Assuming that the Off data, introduced in Chapter 3.1, represent the real hadron-data exactly, that is $N_{0\cdot} = N_{0\cdot}^{Off}$, the estimate of $N_{1\cdot}$ in (15) becomes $\hat{N}_{1\cdot} = \frac{N_{11}}{TPR}$. This estimate is unbiased under the assumption of a binomial distribution in combination with $N_{0\cdot} = N_{0\cdot}^{Off}$, see (11), and its variance is, see equation (13),

$$\text{Var}\left(\frac{N_{11}}{TPR}|N_{1\cdot}\right) = \left(\frac{1}{TPR} - 1\right) N_{1\cdot}.$$

Minimizing this variance over all thresholds is equivalent to maximizing TPR. So under the assumption that the Off data perfectly represents the real hadron-data, the best discrimination threshold is the one which maximizes TPR. This results in a threshold where hadronness is equal to 0, classifying all events as gamma and thus receiving $TPR = 1$.

However, the assumption $N_{0\cdot} = N_{0\cdot}^{Off}$ is very strong, as only their underlying distributions are the same. Therefore, the approach which is used in practice is to set the threshold manually so that the TPR is “high, but not too high”. Often

TPR is set to values between 0.4 and 0.9 (e.g. Aleksic et al, 2010; Jogler, 2009; Hadasch, 2008) and the threshold is chosen accordingly.

3.4.3 Logistic Regression approaches

An approach to avoid energy binning is to fit a binary regression model to the random forest output, with energy as the covariate. The fitted curve can then be regarded as discrimination threshold. In this work we use a logistic regression model. Figure 11 illustrates this principle. In this Figure, the output of a random forest, here called Hadronness, is plotted against the energy of a particle. A logistic regression curve is fitted to the points and used as a variable, energy-dependend threshold. In this case each point above the curve is classified as background, whereas the points under the curve are classified as signal. We first fit the model to the training data using a standard Maximum Likelihood approach to estimate the model coefficients. In the following we refer to this method as Logreg.

Another approach is to combine the aforementioned logistic regression with our MSEmin method from Chapter 3.3. Instead of using a fixed threshold c in the algorithm, we optimize the parameters of the fitted logistic regression curve, β_0 and β_1, so that the MSE becomes minimal. The algorithm remains the same as before, with the only difference being that we do not minimize the MSE over one value, the threshold c, but over two, the model parameters. The minimization is done numerically via the Nelder Mead method (Nelder and Mead, 1965). We refer to this combination of the two methods as LogregMSE.

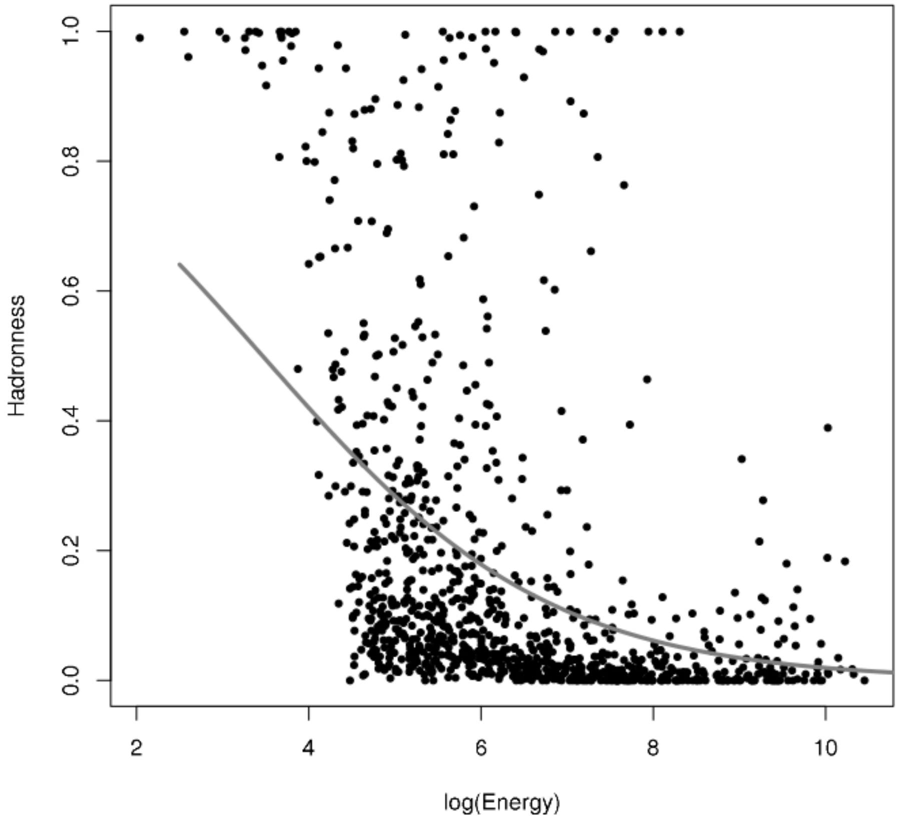


Figure 11: The random forest votes for a subset of 1000 sample observations, consisting of 500 gammas and 500 hadrons, and a fitted Logistic Regression curve. All observations under the curve are classified as Signal, the others as Background.

3.4.4 Comparison of the methods

To compare the MSEmin approach presented in Chapter 3.3 to the other methods, we apply all methods to several artificial test samples. For the method of manual choice we set TPR to 0.1, 0.2,..., 0.9 and refer to these methods as Recall01 to Recall09.

For the comparison we use the following data:

Test data: To represent actual data we simulate 500 samples for each gamma-hadron-ratio 1:100, 1:200,...,1:1000. The number of hadron-events in each sample is drawn from a Poisson distribution with mean 150,000. The number of gamma events is chosen to match the respective ratio.

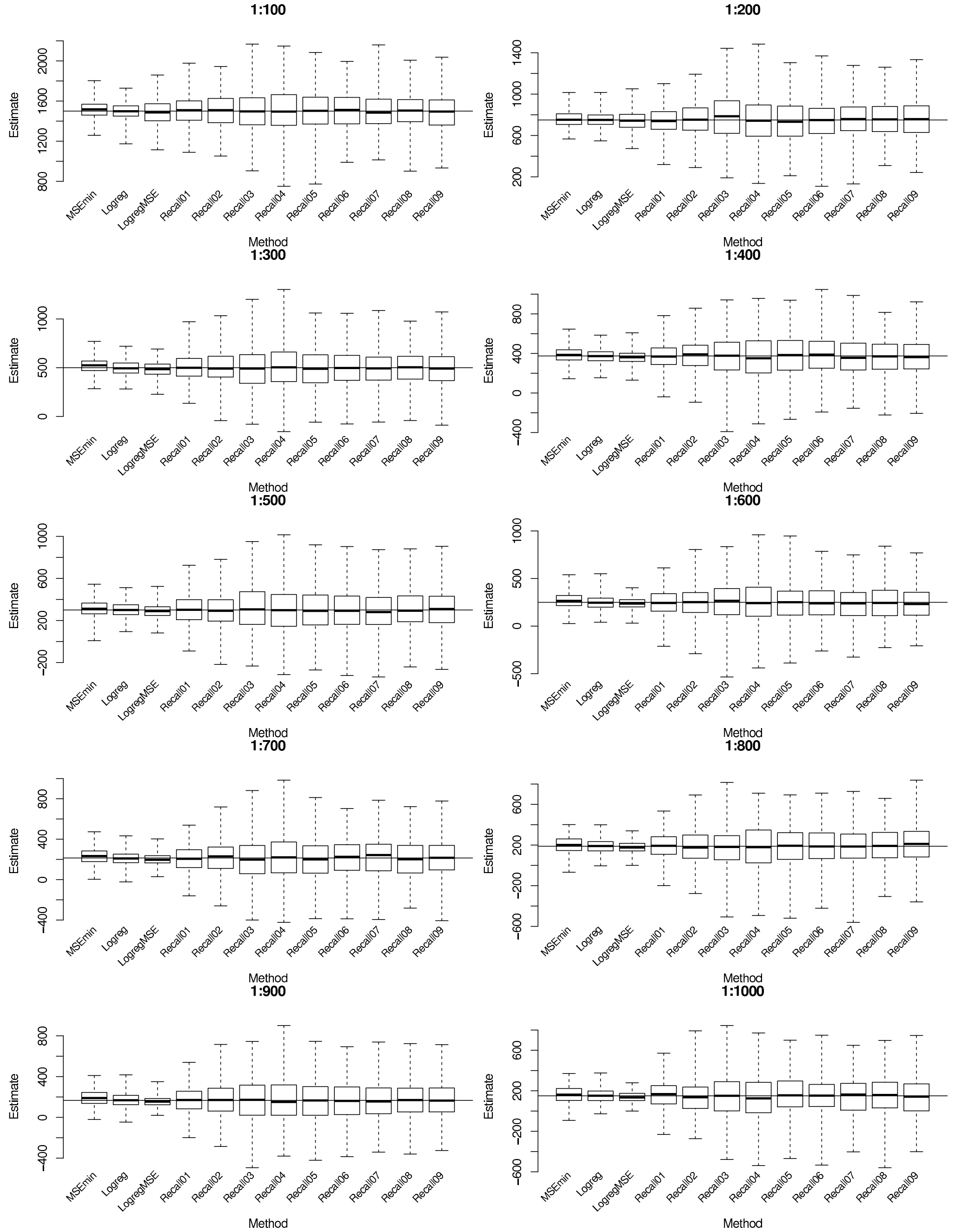
Training data: We use all gamma and hadron observations which are not used in the Test or Off data to represent the training data from which TPR and FPR are calculated. Note that the ratio of gammas and hadrons in this data has no influence on the outcome, as only TPR and FPR are calculated from this data.

Off data: For each test sample we draw a sample of hadron observations to represent the Off data. The number of hadrons in each sample is drawn from a Poisson distribution with mean 150,000.

As outlined above, all gamma events are simulated whilst the hadrons are real Off data observed from a sky-position with no gamma-ray source.

We split every sample into 12 energy-bins and apply the methods MSEmin and Recall01 to Recall09 to determine the optimal threshold in every energy bin. The classes are chosen to contain equal numbers of observations. This approach is in practice almost the same as choosing equidistant classwidths on a logarithmic scale, but it avoids (nearly) empty classes. The Logreg and LogregMSE methods are applied without binning the data.

For each method we additionally apply the θ^2-cut (Lessard et al, 2001; Domingo-Santamaria et al, 2005) as it is always applied in the MAGIC analysis chain. θ^2 is a variable, which is not included in the random forest for various reasons. Instead, an additional cut is applied in θ^2 after the classification. An event is classified as


1:100

1:200

1:300

1:400

1:500

1:600

1:700

1:800

1:900

1:1000

Figure 12: Boxplots of the estimates in the 500 samples with different gamma-hadron-ratios. The thick line in the middle of each box represents the median of the estimates. Between the upper and lower boundaries of each box lie 50% of the estimates. The whiskers range to the minimum and maximum of all the data. The true number is marked by the long horizontal line.

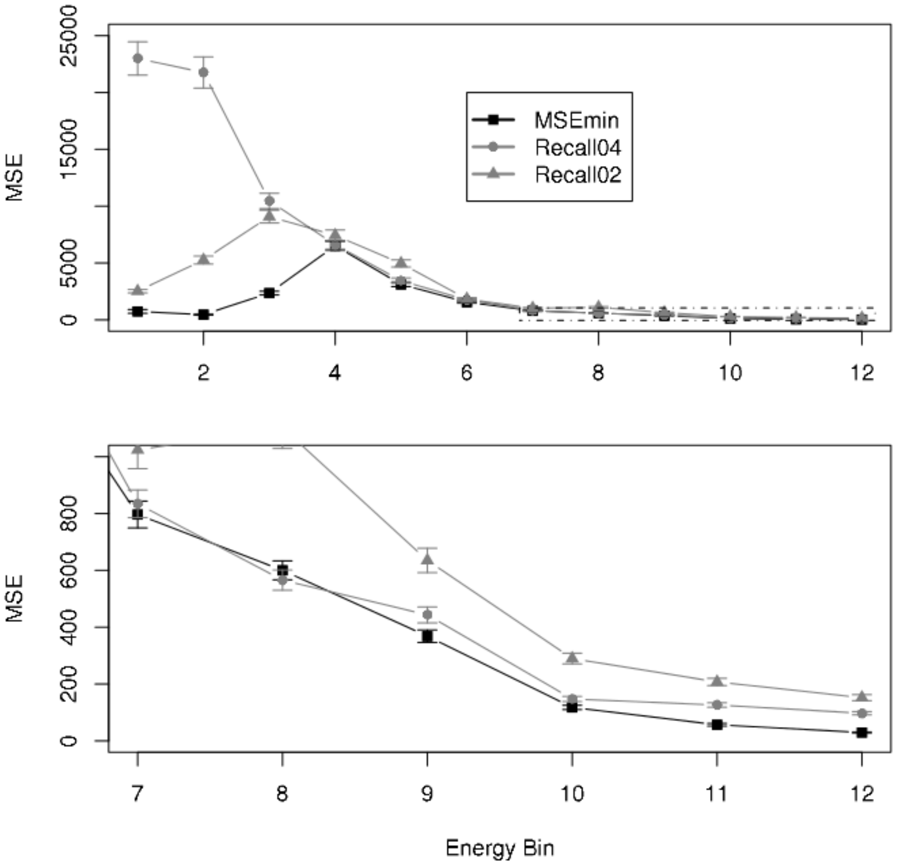


Figure 13: The empirical MSE of the estimation in each energy bin for a gamma to hadron ratio of 1:1000 and the three methods MSEmin, Recall04 and Recall02. The bottom panel shows the marked area of the top panel magnified. The lines are only drawn to guide the eye.

gamma if the used classification algorithm classifies it as gamma and its value of θ^2 is smaller than the respective cut. Here we choose the θ^2-cut to be 0.1. This cut is usually chosen energy dependent but for simplicity we choose it to be constant. Due to this cut it is sometimes not possible to reach the desired Recall in the method Recall09. We then choose the threshold so that the Recall is as high as possible.

Table 2: Empirical mean square error of the estimation of the number of gammas in each energy bin with standard deviations for a ratio of 1:1000.

bin	1	2	3	4	5	6
energy (GeV)	[0,25]	(25,35]	(35,49]	(49,66]	(66,88]	(88,123]
MSEmin	744 ± 164	459 ± 40	2352 ± 162	6513 ± 398	3117 ± 193	1559 ± 90
Recall04	22994±1465	21756±1376	10459 ± 678	6602 ± 388	3450 ± 230	1748 ± 106
Recall02	2522 ± 149	5261 ± 350	9081 ± 560	7429 ± 478	4965 ± 317	1831 ± 122
bin	7	8	9	10	11	12
energy (GeV)	(123,176]	(176,265]	(265,434]	(434,817]	(817,2023]	(2023,∞)
MSEmin	796 ± 47	600 ± 33	368 ± 22	117 ± 7	56 ± 5	30 ± 2
Recall04	834 ± 47	566 ± 36	443 ± 28	148 ± 9	126 ± 8	97 ± 6
Recall02	1024 ± 66	1102 ± 73	635 ± 43	289 ± 18	207 ± 13	152 ± 10

Figure 12 depicts boxplots of the estimated number of positives in the test samples. As we can clearly see for all considered ratios the overall number of gamma events is estimated with a smaller error for the MSEmin method than for any Recall method. The performance of MSEmin is quite similar to that of the standard logistic regression approach. The combination of these two methods give the best results with smaller errors in all ratios above 1:200. For 1:100 and 1:200 the performance of LogregMSE seems worse than that of Logreg. This is not intuitive as LogregMSE is made to produce smaller errors than Logreg. A possible explanation for this is that the Nelder Mead method may have ran into local optima.

Figure 13 and Table 2 show the MSE of the estimate in each energy bin for a ratio of 1:1000 and the three methods MSEmin, Recall02 and Recall04. The two Recall-methods were chosen because of their distinctive behavior in the first three energy bins. The other Recall-methods perform similar to one of them, but with higher MSEs.

As it is clearly seen in Figure 13, MSEmin drastically outperforms the other two in the first three energy bins, ranging from 0 to about 50 GeV. Although the performances of MSEmin and Recall04 at first glance look similar in the higher energy ranges, the magnification in Figure 13 and Table 2 show that MSEmin again per-forms significantly better than the other methods, with a 55% smaller error than Recall04 in the 11th energy bin and up to 70% in the 12th, covering energies higher than 2023 GeV.

Although Recall02 is better than Recall04 at low energies, it is the other way round at higher energies. In medium energy ranges between 123 and 265 GeV the difference between MSEmin and Recall04 is not significant (c.f. Table 2).

All three methods produce smaller MSEs in high energies than in lower ones (except for the first three bins for MSEmin and Recall02). This is because observations in higher energies are bigger and therefore easier to classify.

The different behavior at lower energy bins can also be explained. Monte-Carlo simulations for the gamma signal usually start at around an energy of 50 GeV. This corresponds to the fourth and higher bins. Nevertheless, as the energy of the observations is only estimated, there are also gamma ray observations in lower energy bins, even though not as many as in the other bins. The first three bins therefore represent situations, in which the gamma ray signal is very sparse and the classification is quite difficult due to the low energy range. In such situations the MSEmin method seems to work better than in other situations.

3.5 Investigation of the influence of the binomial assumption

In this Chapter the binomial assumptions from Chapter 3.2, $N_{ij}|_{N_{i\cdot}=n_{i\cdot}} \sim \text{Bin}\left(n_{i\cdot}, p_{ij}\right)$, $i, j \in \{1, 0\}$ and $N_{0j}^{Off}|_{N_{0\cdot}^{Off}=n_{0\cdot}^{Off}} \sim \text{Bin}\left(n_{0\cdot}^{Off}, p_{0j}\right)$, $j \in \{1, 0\}$, are investigated. For this we compare the MSE calculated using equation (16) with the empirical estimate

$$\widehat{MSE} = \frac{1}{m}\sum_{i=1}^{m}((\hat{N}_{1\cdot})_i - N_{1\cdot})^2$$

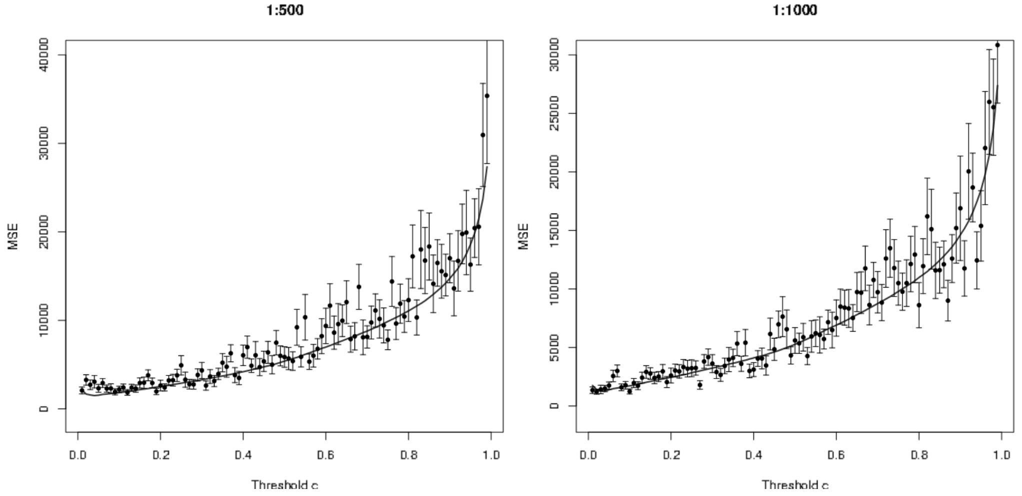


Figure 14: The MSE for ratios 1:500 (left) and 1:1000 (right) calculated via equation (16) which uses the assumption of a binomial distribution (solid line) and estimated from several samples (dots). Each dot is the mean of the MSEs estimated from 50 different samples. The errorbars indicate the standard error of each estimate

obtained from $m = 50$ estimates $(\hat{N}_{1\cdot})_i$ of $N_{1\cdot}$, based on different training samples, all with the same $N_{1\cdot}$ and $N_{0\cdot}$. Systematic differences between the two estimates of the MSE indicate a violation of the binomial assumption. Figure 14 shows estimates of the MSE with and without the use of the binomial assumption for different gamma-hadron-ratios with $m = 50$. Although there seem to be some systematic differences, the red line corresponding to formula (16) fits the empirical estimates quite well. The differences between them are small in comparison to the standard errors of the estimates. Thus, according to Figure 14 the assumption of the binomial distribution seems to be justifiable at least as a useful approximation.

4 New classification variables

In this Chapter, we construct new variables for the separation of FACT data into signal and background. The new variables are based on fitting bivariate distributions to shower images and calculating distance measures between the fitted and observed distributions. As we will see, the new variables lead to substantial improvement of the misclassification rates. The contents of this Chapter can also be seen in Voigt & Fried (2014a) and Voigt & Fried (2014b).

4.1 Distance based feature generation

The idea underlying the majority of the following chapter is to regard the images of gamma events as realizations of bivariate distributions. In other words as a shower image consists of a number of photons registered by the camera we assume a bivariate distribution for the random arrival coordinates (X, Y) of the photons on the camera plane. A camera image with its pixelized structure is thus seen as a kind of two-dimensional histogram with hexagonal classes. This way we get a natural embedding into the statistical context: The pixel-data can be seen as bivariate dataset, in which the data at hand is already binned.

In the following chapter we will see that the binning of the data can be problematic. Many statistical methods require that the data is available unbinned. If a method allows the incorporation of weights this problem can somehow be managed by assuming the value of each bin (here the intensity in a pixel) to be concentrated on one point, usually the centre of each pixel. As it is not known how the intensity is distributed in a pixel, the results can, however, be biased. This can be seen in the following.

When assuming that the shower images come from bivariate distributions a few questions arise. The first question is, obviously, which distribution underlies the events. Here it is important that the recorded Cherenkov light is emitted by showers induced by various particle types. The shape of the shower depends on the particle type which induced the shower. Proton induced showers for example have a different shape than gamma induced showers and the shape carries over to the shower image.

Here it is enough to model showers induced by gammas, as these are the important particles in our application. The approach here is thus an asymmetrical one similar to hypothesis testing. In fact, the resemblance to statistical testing is quite strong. We could hypothesize that a gamma shower comes from a bivariate distribution with cumulative distribution function (cdf) F_0 and see if the data contradicts this hypothesis.

Instead, we use a similar idea to construct variables for our classification. We assume a family of distributions for signal events and fit a distribution from this family to each shower. Then, we calculate distance measures between the fitted and observed distributions. If the assumed family of distributions is close to the true family of signal showers, the distance between the fitted and observed distributions should be small for signal events and higher for background events.

In the following, we try to find a distribution which fits well to signal images. We explain how we fit the distribution to a shower, which is not trivial due to the pixelized structure of the data. We will then show how with the fitted distributions and some distance measures new variables for the signal-background classification can be constructed. We will see that the new variables improve the classification on a sample of FACT data. We first explain, how the Hillas parameters are constructed and how our approach extends the idea underlying Hillas parameters.

4.2 Hillas variables

The idea of Hillas variables is to fit an ellipse to a shower image and use the parameters of the ellipse as variables for classification. The first two Hillas variables are the center coordinates of the ellipse, also known as the centre of gravity, `CoGX` and `CoGY`. Then there are the `Width` and `Length` of the fitted ellipse, that is the lengths of the two semiaxes, where the shorter one is always considered as `Width`. Another Hillas parameter directly calculated from the fitted ellipse is $\texttt{Area} = \pi \cdot \texttt{Width} \cdot \texttt{Length}$, the area of the fitted ellipse. The remaining two ellipse parameters are angles. The first is δ, the angle between the longer semiaxis of the ellipse and the x-axis of the camera. The second is α, which describes the angle between the longer semiaxis of the ellipse and the line between the centre of the camera and the centre of the ellipse. Besides, there are seven further Hillas variables mainly based on pixel brightnesses and ratios between them. Some Hillas variables and the underlying MAGIC telescope camera (MAGIC collaboration, 2014) can be seen in Figure 8 in Chapter 2.4.3. A summary of all Hillas variables available in our FACT data set can be seen in Table 3.

4.3 Distributions to fit

As stated above, the idea of the following chapter is to fit bivariate distributions to shower images. The distributions we try in this work are introduced below.

Table 3: The Hillas variables available in our data set.

Variable	Explanation
`Size`	Sum of intensities of all pixels in a shower
`Width`	Width of the fitted ellipse
`Length`	Length of the fitted ellipse
`Area`	$\pi \cdot Length \cdot Width$
`Delta`	Angle between the longer semiaxis of the ellipse and the x-axis
`Alpha`	Angle between reconstructed and expected direction of the shower
`Concentration`	Ratio of the brightest pixel to `Size`
`Concentration2`	Ratio of the two brightest pixels to `Size`
`NumberIslands`	The number of islands in a shower
`NumberShowerPixel`	The nubmer of pixels with intensity greater 0 after the image cleaning
`Leakage`	Ratio of the intensity of pixels on the edge of the camera to `Size`
`Leakage2`	Like Leakage, only additionally with neighbouring pixels

4.3.1 Gaussian Fit

Fitting an ellipse as done for the Hillas variables corresponds to fitting a contour line of a bivariate elliptic distribution. The most popular example of such a distribution is a bivariate Gaussian, which can be fitted for example by minimizing a χ^2-distance between the observed and the expected frequencies in the hexagonal telescope image. This allows to include information not only on the boundaries of the ellipse, but also on the values in the interior.

The relation between the parameters of the ellipse and the fitted Gaussian is as follows: Let $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}$ be the parameters of the fitted bivariate Gaussian distribution. The fitted distribution is thus $\mathcal{N}_2(\mu, \Sigma)$.

We look at the spectral decomposition of Σ:

$$\Sigma = U \Delta U^T$$

where

- $\Delta = \operatorname{diag}(\lambda_1, \lambda_2)$ is a diagonal matrix with the sorted eigenvalues of Σ: $\lambda_1 \geq \lambda_2$.
- $U = (u_1, u_2)$ is an orthogonal matrix containing the corresponding eigenvectors $u_1 \in \mathbb{R}^2$ and $u_2 \in \mathbb{R}^2$.
- U^T denotes the transposition of the matrix U.

With this notation, we can describe the relationship between the fitted Gaussian and the other Hillas variables:

- $\mu = \begin{pmatrix} \texttt{CoGX} \\ \texttt{CoGY} \end{pmatrix}$,
- $\lambda_1 = c \cdot \texttt{Length}$,
- $\lambda_2 = c \cdot \texttt{Width}$, where c is some constant,
- $\texttt{Area} = \pi \lambda_1 \lambda_2$,
- δ is the angle between the x-axis and u_1,
- α is the angle between μ and u_1.

Note that the Hillas variables contain the full information about the parameters of the fitted Gaussian distribution. However, it is known that the signal has a more regular shape than the background. For example, signal is known to be unimodal, while background can have several peaks. This information cannot be evaluated by fitting a single ellipse, but can be used by fitting a distribution, for example through the use of goodness of fit measures. Fitting a Gaussian distribution instead of an ellipse allows to incorporate information on the shape of signal images. The left hand side of Figure 15 shows a sample image on the underlying FACT telescope camera (The FACT collaboration, 2014) and a fitted bivariate Gaussian. It can be expected that a bivariate Gaussian fits better to signal images than to background images, because of the more regular shape of the former. We thus use distance measures between the observed image and the fitted Gaussian distribution in our classification, as these tend to take smaller values for signal events.

4.3.2 Skew-normal distribution

Fitting Gaussian distributions for constructing variables for classification can possibly be further improved. Signal images seem to be roughly elliptical, but not exactly. It is known that signal showers are skewed in one direction (de Naurois & Rolland, 2009). Because of this, an elliptical fit or fitting a Gaussian distribution can neither describe signal observations nor background observations very well. As the aim of our analysis is to find differences between signal and background, it makes sense to fit a distribution to the images which describes signal images well. Fitting a skewed distribution to the images, instead of a Gaussian, seems therefore to be mandatory.

A skewed extension of the bivariate normal is the bivariate skew-normal distribution introduced by Azzalini & Dalla Valle (1996). Azzalini & Capitanio (1999) extended the work by adding location and scale parameters. A random vector Y is said to have a multivariate skew-normal distribution with parameters $\xi \in \mathbb{R}^k$, $\Omega \in \mathbb{R}^{k \times k}$ and $\alpha \in \mathbb{R}^k$, short $Y \sim SN_k(\xi, \Omega, \alpha)$, if it has the density function

$$f_k(y) = 2\phi_k(y - \xi; \Omega)\Phi(\alpha^T \omega^{-1}(y - \xi))$$

where ϕ_k is the density function of a k-variate normal with zero mean, standardized marginals and correlation matrix Ω, Φ is the distribution function of a standard

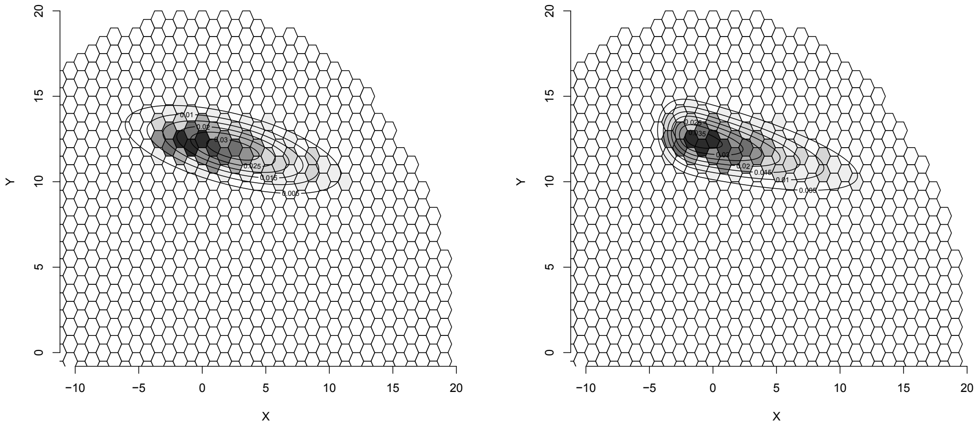


Figure 15: Contour lines of a bivariate normal (left) and bivariate skew-normal (right) distribution fitted to a random shower image. Only the relevant part of the camera is shown.

normal distribution and ω is the diagonal matrix of the square roots of the diagonal elements of Ω. The bivariate skew-normal distribution thus has seven parameters, whereas a bivariate normal has only five. A bivariate skew-normal distribution fitted to the same shower image as before can be seen on the right hand side of Figure 15.

4.3.3 Gaussian with fixed alignment

We can also use other information when fitting a distribution to a shower. We know for example that signal showers are always aligned to the center of the camera. This can be seen on the left hand side of Figure 16, where a simulated signal event is displayed. To use this information we force a fitted distribution to be aligned to the centre of the camera. As we know that signal events are aligned to the source (here the camera center), we force the fitted distribution to be aligned, too. In the case of the bivariate Gaussian, we fit the distribution under the constraint that at least one of the eigenvectors u_1 or u_2 has the same direction as μ, that is, one of the semiaxes of the elliptical contour lines is aligned to the camera center. For this we rotate the camera image so that the centre of gravity of the shower lies on the x-axis of the camera as seen in Figure 16. We then fit two one-dimensional distributions to the x- and y-directions of the shower. For now we fit two univariate normal distributions. These represent a bivariate normal distribution with zero covariance, so that we get a bivariate normal which is in one direction aligned to the center, when we rotate the image and the fitted distribution back.

After fitting a distribution in this way, we can apply the same distance measures as for the other distributions.

4.3.4 Skew-Normal with fixed alignment

While the skew-normal distribution takes the skewness and the normal with fixed alignment takes the alignment of the signal showers into account, both aspects together have not been considered, yet. We try to consider both by fitting a skew-normal distribution with fixed alignment. Like in the case of a normal, we rotate the shower image so that the centre of gravity of the shower lies on the x-axis. We then fit a univariate skew-normal to the x-direction and a univariate normal to the y-direction. We fit a skewed distribution to the x-axis and a symmetric one to the y-axis, because it is known that signal showers are skewed in the direction, which points to their source. The rotation of the images is done so that this direction is the x-axis in the rotated image for signal events. It is also known that signal showers are symmetric in the direction perpendicular to the one aligned to the source, which is the y-axis in the rotated image (de Naurois & Rolland, 1996).

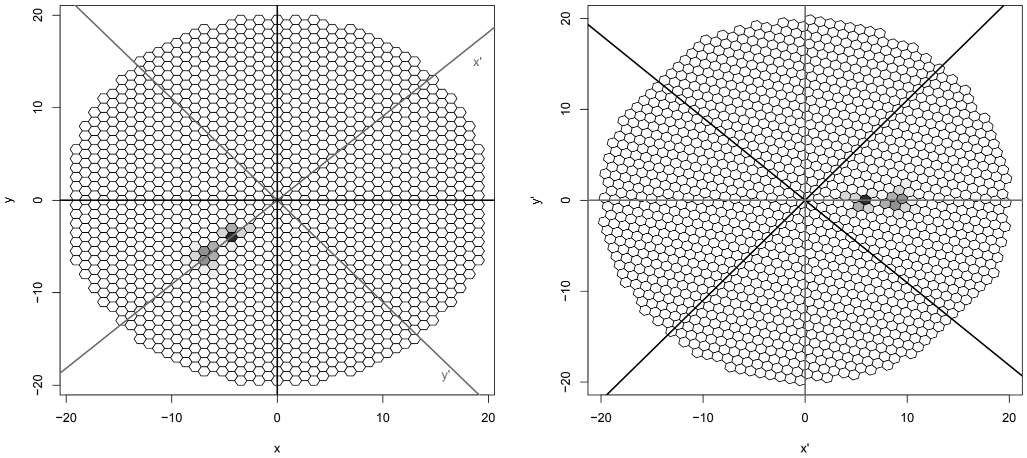


Figure 16: Rotating an image to fit two univariate normal distributions to the new x- and y-directions.

4.4 Maximum Likelihood Fitting

In this Chapter, we describe how we fit the distributions to the shower images. One possibility for fitting a density to the shower images is to use maximum likelihood estimators (MLEs) for the parameters of the distribution. For observations from normal distributions this is easy, as MLEs are known and simple to calculate, namely the arithmetic mean and the empirical covariance matrix. In our case of discretized data on a crude hexagonal pixel structure these estimators have to be adjusted for this discretization. A well-known possibility to adjust estimators for binned data are Sheppard's corrections (see e.g. Kendall & Stuart, 1963). These are, however, not applicable in our case, as they require equidistant (and therefore rectangular) class limits. As no corrections for hexagonal binning are known to us, we use intuitive, heuristical estimators. The principle underlying these is to use the center point of each bin weighted by the relative number of events in this bin.

The resulting estimators are weighted means, variances and covariance given by:

$$\hat{\mu}_x = \sum_i m_{xi} \frac{M_i}{M}$$

$$\hat{\mu}_y = \sum_i m_{yi} \frac{M_i}{M}$$

$$\hat{\sigma}_x^2 = \sum_i \frac{M_i}{M} \left(m_{xi} - \hat{\mu}_x\right)^2 ,$$

$$\hat{\sigma}_y^2 = \sum_i \frac{M_i}{M} \left(m_{yi} - \hat{\mu}_y\right)^2 ,$$

$$\hat{\sigma}_{xy}^2 = \sum_i \frac{M_i}{M} \left(m_{xi} - \hat{\mu}_x\right) \left(m_{yi} - \hat{\mu}_y\right) ,$$

where M_i is the intensity in the i-th pixel, M is the sum of all intensities in the shower and $m_i = \left(m_{xi}, m_{yi}\right)^T$ is the vector of the center coordinates of the i-th pixel.

We use these estimators to fit the normal and bivariate normal distributions. For the skew-normal distribution the fit is not that easy, as its parameters are not simply its mean and variance/covariance. In the case of the skew-normal distributions we use numerical ML-estimates of the parameters making use of the R package `sn` (Azzalini, 2014). As above, we consider the observations in each pixel to be concentrated on the center of the pixel. The calculated estimates are therefore no exact ML-estimates.

4.5 Discretization

The distributions described above are continuous, while the shower images are pixelized and thus discretized. In the following, we want to use distance measures to compare the observed and fitted distributions. These distance measures usually require that both distributions are either continuous or discrete. To use these distance measures, we discretize the fitted distribution.

The probability of a Cherenkov photon of shower i to fall in pixel j under the fitted distribution is

$$p_{ij} = \iint\limits_{\text{⬡}_j} f_0(x, y; \theta_i) dxdy$$

where f_0 is the probability density function (pdf) of the fitted distribution and θ_i is its parameter vector.

For calculating this integral, the hexagonal pixel can be split into two triangles and a rectangle:

$$p = I_1 + I_2 + I_3$$

with

$$I_1 = \int\limits_{m_x - r}^{m_x - \frac{r}{2}} \int\limits_{m_y - \sqrt{3}x + \sqrt{3}m_x - \sqrt{3}r}^{m_y + \sqrt{3}x - \sqrt{3}m_x + \sqrt{3}r} f(x, y) dxdy$$

$$I_2 = \int\limits_{m_x-\frac{r}{2}}^{m_x+\frac{r}{2}} \int\limits_{m_y-\frac{\sqrt{3}}{2}r}^{m_y+\frac{\sqrt{3}}{2}r} f(x,y)dxdy$$

$$I_3 = \int\limits_{m_x+\frac{r}{2}}^{m_x+r} \int\limits_{m_y+\sqrt{3}x-\sqrt{3}m_x-\sqrt{3}r}^{m_y-\sqrt{3}x+\sqrt{3}m_x+\sqrt{3}r} f(x,y)dxdy$$

Here, m_x and m_y are the coordinates of the centre of the pixel and r is the radius of the circumcircle of the pixel.

Even for the Gaussian distribution these integrals are not analytically solvable, so that they would have to be solved numerically for each pixel. We use the R package `cubature` (Narasimhan, 2013), which use algorithms for adaptive multidimensional integration, which are in detail described in Genz & Malik (1980) and Berntsen et al. (1991). The numerical solution poses a problem, however, because the FACT telescope camera consists of 1440 pixels. That means we have to numerically solve 4320 integrals per shower image. The compute cluster LiDo of the TU Dortmund University needs slightly more than half a second to compute one of these integrals numerically using `cubature`. This means that a single image needs more than half an hour of computation time. Considering that we have several million shower images every night this is far too much time needed for this step. Although the very slow computation times might be enhanced by using a different software than R, a faster method has to be used in R.

One possibility is to approximate the hexagonal pixels through rectangles, as seen in Figure 17. A rectangular pixel structure instead of a hexagonal one decreases the computation time by a factor of three. This is, however, still too long for the

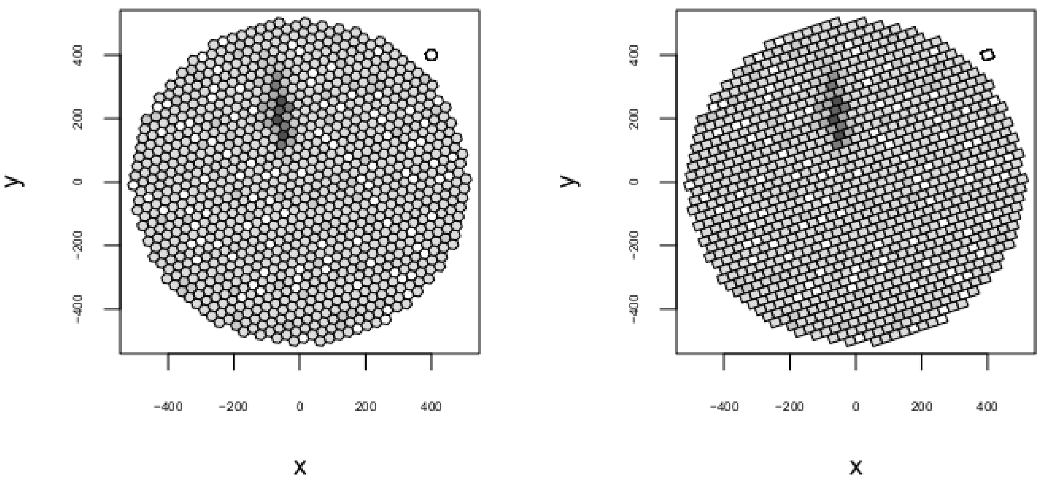


Figure 17: The original camera image (left) and an approximation with rectangles (right).

large data sets one has to handle in our astrophysical setting.

Another faster, but more inaccurate method is to just simulate from the fitted distribution and count the number of observations in each pixel. For a large enough number of observations the relative frequency in each pixel should be close to the true probability of an observation to fall into this pixel. This is the method used here. For each shower we draw 5000 observations from the fitted distributions and count the relative frequency of realizations in each pixel.

4.6 Distance measures

After we fit distributions to the showers we use the fitted distributions to construct new variables for classification. The idea here is that the fitted class of distributions describes the true distribution of signal events well, so that the fit should be better for signal events than for background events. We thus want to measure the distance between the fitted and the observed distribution of each of our observed events. Several distance measures between distributions exist. There are for example goodness of fit test statistics like the Chi-square (e.g. Greenwood & Nikulin, 1996), which can easily be extended to the bivariate case. There are also some distance measures for densities like the Kullback-Leibler divergence (Kullback & Leibler, 1951) or the Hellinger distance (e.g. Nikulin, 2001), described in the following. As we only need the discrete versions of the distances, we do not consider the continuous versions here.

4.6.1 Chi-square Statistic

An advantage of the Chi-square distance, underlying the corresponding test, in our situation is that it is based on binned data as given in our application. We can hence directly apply it to our data without the need of further modification.

The resulting Chi-square distance is

$$Q_n = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

where n_i is the observed intensity in pixel i, $n = \sum\limits_{i=1}^{m} n_i$, p_i is the probability that a gamma-induced photon arrives at pixel i and m is the number of pixels in the camera. Because we have discrete (binned) observed values we discretize the fitted distribution and directly calculate the p_i from it.

Large values of Q_n indicate that the observed event is likely to be a background event. That means we expect signal events to lead to smaller values of Q_n than background events.

4.6.2 Kullback-Leibler Divergence

One possibility to compare two distributions is to use distance measures for densities. The best known of these is the Kullback Leibler divergence (Kullback & Leibler, 1951).

Let P and Q be two discrete distributions and p and q their corresponding probability density functions (pdf). The asymmetric Kullback-Leibler divergence between p and q is defined as

$$D_k(P, Q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

for discrete distributions. The Kullback Leibler divergence is not symmetric, so in general $D_k(P, Q) \neq D_k(Q, P)$.

In our case we want to measure the distance between the empirical distribution with cdf F_n and pdf f_n and the fitted distribution with cdf F_0 and pdf f_0. As said in the previous paragraph we discretize the fitted distribution leading to the discrete empirical density given by $\frac{n_i}{n}$ with n and n_i being the same as above. The values of the discretized fitted distribution are given by $p_1, ..., p_m$, so that the Kullback Leibler divergence reduces in our case to

$$D_k(F_0, F_n) = \sum_{i=1}^{m} p_i \log \left(\frac{np_i}{n_i} \right)$$

and accordingly

$$D_k(F_n, F_0) = \sum_{i=1}^{m} \frac{n_i}{n} \log\left(\frac{n_i}{np_i}\right).$$

We use both versions in the following application and call the first one Kullback Leibler Divergence theor. and the other Kullback Leibler Divergence obs..

4.6.3 Hellinger Distance

The Hellinger distance is another distance measure between densities. For discrete distributions P and Q with pdfs p and q it is in general given by

$$D_h(P, Q) = 1 - \sum_x \sqrt{p(x)q(x)}.$$

The Hellinger distance is bounded by 0 and 1 and symmetric, $0 \leq D_h(P, Q) \leq 1$ and $D_h(P, Q) = D_h(Q, P)$.

With the same reasoning as above the Hellinger distance simplifies in our case to

$$D_h(F_0, F_n) = 1 - \sum_{i=1}^{m} \sqrt{p_i \frac{n_i}{n}}.$$

4.7 The new variables

The new variables we add to the Hillas parameters are the combinations of distributions and distance measures described above. A further variable added is the quotient of the variances in x- and y-direction in the rotated shower image we used to fit the aligned normal and skew-normal distributions. Like the aligned distributions it uses the information that signal events are aligned to the centre of the camera. It can be expected that for signal events this quotient is smaller than for background.

In one of the previous Chapters it could be seen how the camera is rotated in order to fit the aligned normal and skew-normal distributions. When fitting the two univariate Gaussians the two variances in the x- and y-directions need to be calculated. Those two variances are used here to construct one additional variable. Similar to the aligned distributions, the idea here is to use the alignment of signal showers opposing to the not aligned background. We additionally use, that signal showers are usually quite narrow, as can be seen in Figure 1 in Chapter 2.1. The information can be used in the ratio of the empirical variances S_y and S_x:

$$\frac{S_y}{S_x}$$

where S_y is the empirical variance of the rotated shower in y-direction and S_x is the empirical variance of the rotated shower in x-direction.

If this ratio is larger than one, the semiaxis of the fitted contour lines which is aligned to the center is smaller than the perpendicular one. This is a strong indicator that the event belongs to the background, because the longer semiaxis of signal events is usually aligned to the center, not the shorter one. If the ratio is 1, the fitted contour lines are circles. This is also an indicator that the event is background. For signal events we expect a ratio smaller than 1.

A summary of all new variables including the ratio can be seen in Table 4.

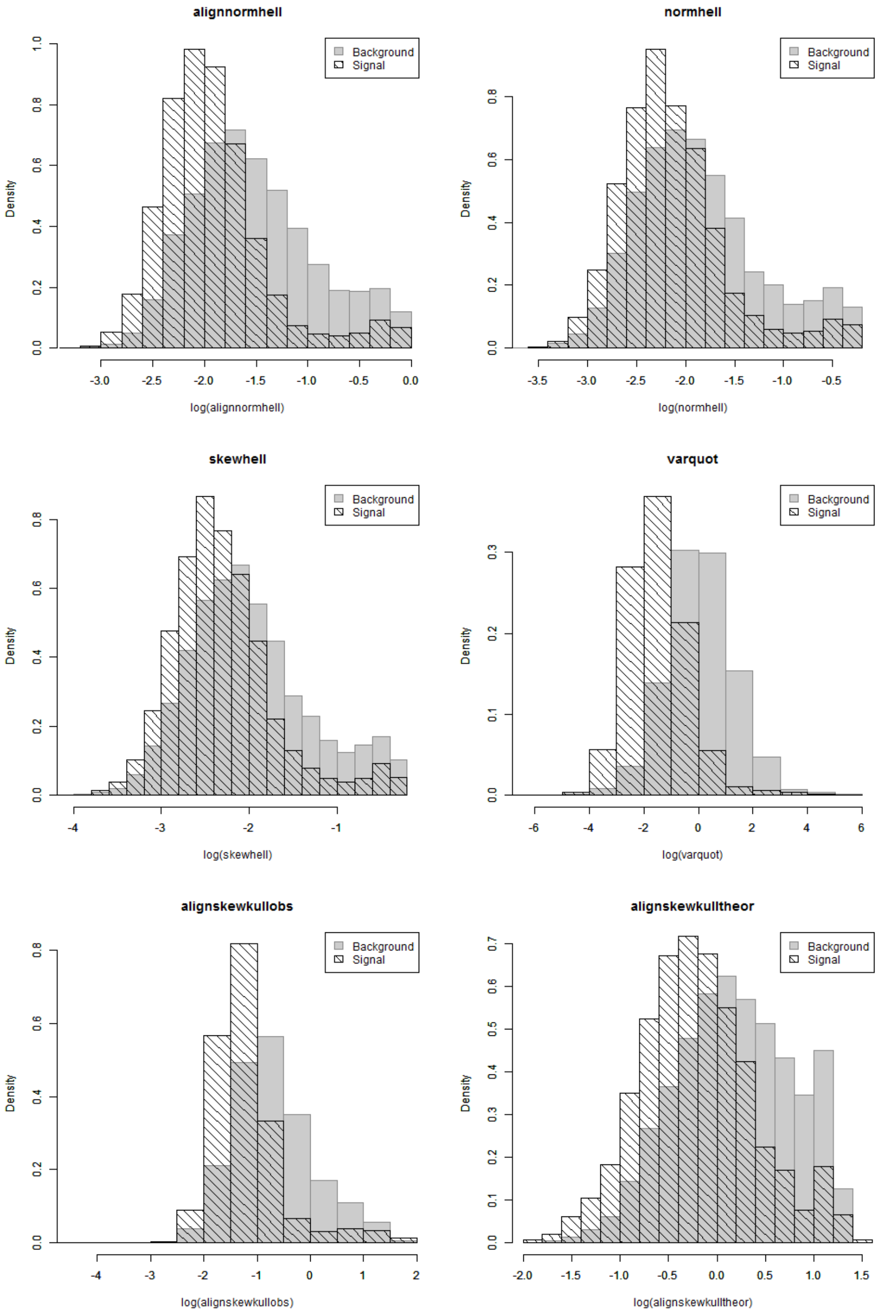


Figure 18: Some histograms of newly constructed variables. The x-axis is on a logarithmic scale to make signal and background better distinguishable, as the histograms of the original data are very skewed.

Table 4: Newly constructed features for classification.

Variable	Explanation
`normhell`	$D_h(F_n, F_0)$ with F_0: cdf of bivariate normal
`skewhell`	$D_h(F_n, F_0)$ with F_0: cdf of bivariate skew-normal
`alignnormhell`	$D_h(F_n, F_0)$ with F_0: cdf of bivariate normal aligned to source
`alignskewhell`	$D_h(F_n, F_0)$ with F_0: cdf of aligned normal/skew-normal
`normkullobs`	$D_k(F_n, F_0)$ with F_0: cdf of bivariate normal
`normkulltheor`	$D_k(F_0, F_n)$ with F_0: cdf of bivariate normal
`skewkullobs`	$D_k(F_n, F_0)$ with F_0: cdf of bivariate skew-normal
`skewkulltheor`	$D_k(F_0, F_n)$ with F_0: cdf of bivariate skew-normal
`alignnormkullobs`	$D_k(F_n, F_0)$ with F_0: cdf of bivariate normal aligned to source
`alignnormkulltheor`	$D_k(F_0, F_n)$ with F_0: cdf of bivariate normal aligned to source
`alignskewkullobs`	$D_k(F_n, F_0)$ with F_0: cdf of aligned normal/skew-normal
`alignskewkulltheor`	$D_k(F_0, F_n)$ with F_0: cdf of aligned normal/skew-normal
`normchi`	Q_n of χ^2-Test for bivariate normal
`skewchi`	Q_n of χ^2-Test for bivariate skew-normal
`alignnormchi`	Q_n of χ^2-Test for bivariate normal aligned to source
`alignskewchi`	Q_n of χ^2-Test for aligned normal/skew-normal
`varquot`	Ratio of variances in the rotated shower picture (S_y/S_x)

4.8 Application to FACT data

In this Chapter, we apply the above variable construction to FACT data. We measure the quality of a classification with and without the newly constructed variables with respect to the task of estimating the number of signal events as well as possible.

4.8.1 Distances on FACT data

In this Chapter, we have a descriptive look at the new variables, calculated on our FACT data set. Histograms of some of the new variables can be seen in Figure 18. We see that some of them show different shapes for signal and background.

Especially interesting are the new distances. The smaller these distances are the better fits the fitted distribution to the data. However, it can be assumed that there will be a lower bound on the distances, because of the pixelized camera, which makes it difficult to fit the distributions. So even if the underlying distribution is of the same family as the fitted one, the distance between observed and fitted distribution might be relatively far away from 0. To verify this we simulate data from bivariate normal distributions and discretize them on the camera surface. We choose the parameters of the bivariate normal so that the resulting camera images look reasonably similar to real showers.

The shower images we simulate consist of n_k realisations of $\mathcal{N}_k(\mu_k, \Sigma_k)$-distributions. The number of realizations n_k and the parameters of the distribution $\mu_k = \begin{pmatrix} \mu_{xk} \\ \mu_{yk} \end{pmatrix}$ and $\Sigma_k = \begin{pmatrix} \sigma_{xk}^2 & \sigma_{xyk} \\ \sigma_{xyk} & \sigma_{yk}^2 \end{pmatrix}$ are drawn randomly and independently from each other for each shower image:

- Because it is known that small showers are more common than large ones, we draw the number of photons n_k of a shower from an $Exp(0.002)$ distribution, rounding to the nearest whole number. The parameter value of 0.002 is chosen so that it resembles the shape of the distribution of `Size` in our data, which is the sum of the intensities of all pixels.

- μ_{xk} and μ_{yk} are drawn indepentently from a $Unif(-15, 15)$ distribution. The centre of gravity of the simulated shower thus lies randomly somewhere on the camera plane, but not too far on the sides.

- σ_{xk}^2 and σ_{yk}^2 are both drawn from an $Exp(\frac{1}{0.1 \cdot \sqrt{n_k}})$ distribution. This choice is arbitrary, but it can be seen that it creates realistic shower images. The idea is that a small number of photons n_k usually leads to smaller showers.

- For σ_{xyk} we draw a correlation randomly from a $Unif(-1,1)$ distribution. We multiply it with the square root of the product of σ_{xk}^2 and σ_{xk}^2 to get a random covariance.

Figure 19 shows exemplarily, that these arbitrary choices lead to simulated images, which look a lot like real shower images.

That way we simulate 2000 camera images where the underlying distribution is bivariate normal and fit bivariate normal distributions to them. We then calculate the distances described above and use the resulting distances as reference as for how close we can get to 0. If all of the distances calculated on FACT data are close to the reference, we can assume that we cannot get a better fit. In Figure 20 there are boxplots of the distances on only signal events in FACT data, including the distances on the simulated normal distributions, called Reference in the Figure. We only use signal events, because we want a fitted distribution to fit especially well to signal events. We see that the skew-normal distribution fitted to FACT data overall gives the smallest distances. That is not surprising, as it is the most flexible distribution we fit here. The aligned distributions on the other hand give higher distances. This seems unintuitive as with the aligned distributions we used the known alignment of signal events. However, in reality, the alignment to the camera center is often not perfect, so that a forced perfect alignment can lead to a worse fit than without forced alignment. We expect the background, however, to realize even greater distances, so that these variables may still work well in practice. We see later in this work that this is indeed the case.

We also see in the boxplots that the skew-normal with forced alignment gives high Kullback-Leibler Distances for our simulated background. We will see later that this variable has a high importance in a classification and is selected by variable

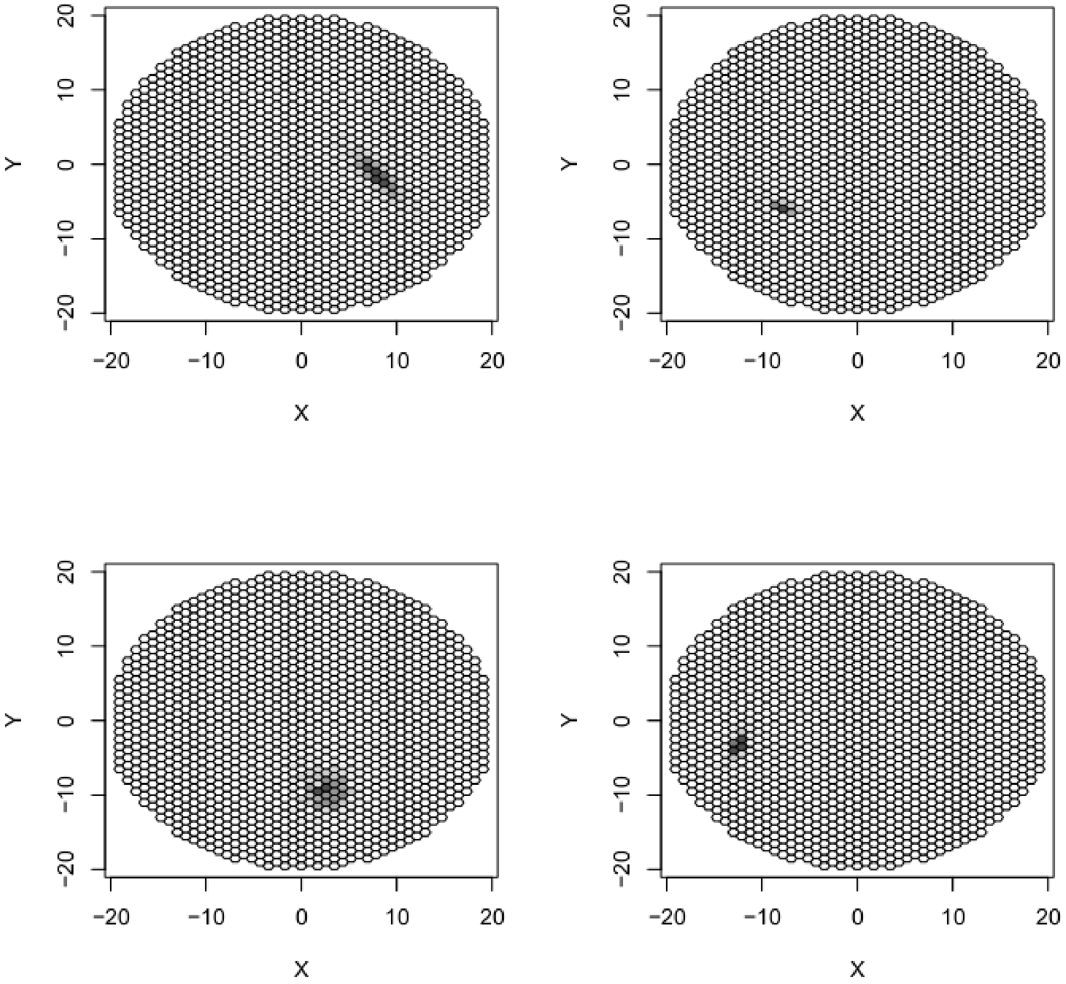


Figure 19: Exemplary four simulated shower images. 2000 of such images have been created.

selection methods.

Another observation in the boxplots is that the Chi-square distance is overall very small, in some cases the median is even slightly smaller than the reference. We conclude that the Chi-square distance cannot distinguish between different fitted distributions as well as the other distances.

4.8.2 Variable selection

As we have created quite a few variables in the previous Chapter, it is now of interest which of these variables improve the classification. As nearly all new variables are based on the same idea, it is possible that some of them are redundant. It is also possible that they are completely unimportant for the classification in which case it would be desirable to eliminate them from the dataset. In addition, it is of interest if some of the Hillas variables are not needed as well. If so, they can be left out, too.

This is why a variable selection is done here. After the selection, the results of a dataset with the selected variables is compared to a dataset containing all variables and one that contains only Hillas variables.

For the variable selection we use one wrapper and one filter variable selection (e.g. Guyon and Elisseeff, 2003). As a representative of the wrappers we use a variable selection based on the out of bag (OOB) error of a random forest (varSelRF). As a filter variable selection method we use a Minimum Redundancy Maximum Relevancy technique (MRMR). Both methods are used separately on the same dataset and the results are compared.

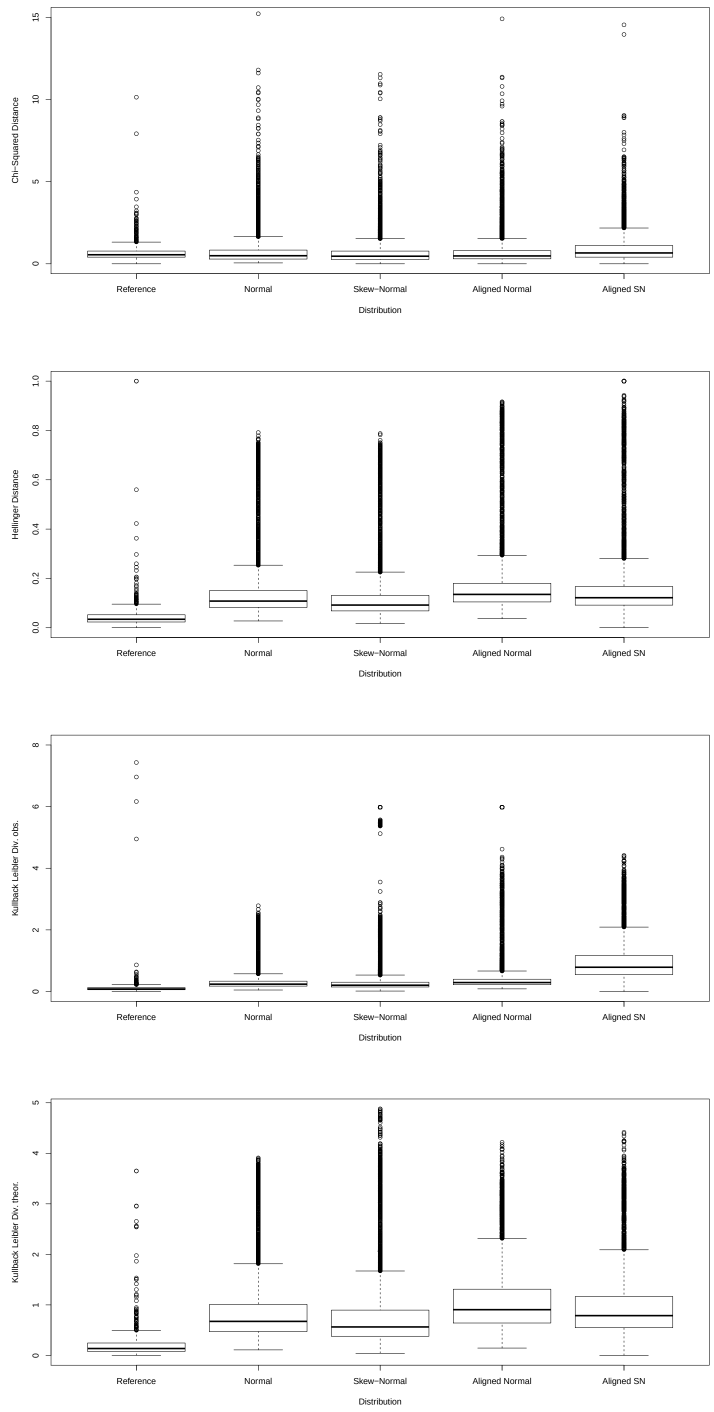


Figure 20: Boxplots of the newly constructed variables. Please note that in the case of the Chi-square distance there are a few outliers larger than 15 in the simulated data (the left boxplot), which have been omitted for the sake of better readability.

4.8.3 varSelRF

The varSelRF method is a backward selection algorithm which chooses variables for classification through random forests (Diaz-Uriarte & Alvarez de Andres, 2005). It begins with all variables and in each step it discards a number of variables based on the random forest OOB error estimate (Breiman, 2001). Table 5 shows which variables are chosen by this method. We see that of the newly constructed variables `varquot`, `alignnormhell`, `alignskewhell` and `alignskewkullobs` have been selected. We also see that the alignment seems to have a high influence on the classification, as the distance variables selected are all from aligned distributions.

4.8.4 MRMR

MRMR (Minimum Redundancy Maximum Relevancy; Peng et. al., 2005) deals with a problem of many other filter selection techniques. Many filter techniques only choose relevant variables. The problem is that such variables that work well for classification are often highly correlated, so that selection of only relevant variables often results in worse classification than discarding redundant variables in favor of less relevant ones. As the name suggests, MRMR tries to select variables, so that at the same time redundancy is minimized and relevancy is maximized. The technique uses a mutual information matrix. The number of variables has to be chosen manually for MRMR. We set it to 9 here, which is the number chosen automatically by the varSelRF method. Table 5 shows which variables are chosen by this method. In this selection only two newly constructed variables are selected. Both of them have been selected by varSelRF, too. The Kullback-Leibler divergence with aligned skew-normal distribution seems to be quite important in the classification. An investigation of the importance is reported later on.

Table 5: Variables selected by the two feature selection methods MRMR and VarSelRF.

MRMR		VarSelRF	
	Leakage		Alpha
	NumberIslands		Concentration2
	Leakage2		varquot
	varquot		alignnormhell
	alignskewkullbeob		alignskewhell
	Width		alignskewkullobs
	Delta		Leakage2
	Alpha		Size
	Concentration		Width

4.8.5 Quality of the classification and the estimation

In this Chapter we assess the quality of the classification with and without the newly constructed variables. Additionally, we look at classifications using only the variables chosen by the two variable selection methods, as seen in Table 5.

We draw 500 subsamples from the FACT data described above to use as test data. The number of background events in each of these subsamples is drawn from a Poisson distribution with mean 1000. We add signal events according to the ratio 1:100 and use the events left in the original dataset as training data. For each of the 1000 test samples of each ratio we train a random forest (Breiman, 2001) and classify the test data with the trained forest. We choose the ratio 1:100 because it is closest to reality we can get with our data. If we reduce the ratio more we do not have enough signal events in the test data. A random forest requires a threshold value as tuning parameter. A threshold value of α means here that $\alpha \cdot 100\%$ of the trees in the random forest have to vote for a background event to classify it as such. This threshold is the same we optimized in the previous chapter. For now we do not use this optimization here, but evaluate only the influence of the new variables on the classification using several different thresholds. We vary the threshold between 0.01 and 0.5. A combination of the previously introduced threshold optimization and the new variables is done in the next chapter.

One would usually use ROC curves (Fawcett, 2006) to compare the different classifications, but because of the high signal to background ratio, the ROC curve is not very meaningful, so that we look at the signal error ($1-$ true positive rate) and background error (false positive rate) separately.

Figure 21 reports the errors of the classification, depending on the cutoff in the test data sample. We see an overall very small classification error in the background, which is necessary at this signal to background ratio. The new features do not significantly improve the already very small background error. The two variable selection methods also do not lead to lower background errors. The signal error on the other hand is improved a lot by the new features and especially by the variable selection methods. We have an overall gain in the misclassification rate of about 3% for the new variables and about 10% for the variable selection methods.

Although the background error is not improved for the variable selection methods the signal error is improved very much. Because it can be expected that in real data there is a signal to background ratio of about 1:1000 (Weekes, 2003), a low signal error is desirable, while maintaining a very low background error, so that the new variables can be seen as an overall improvement.

4.8.6 Importance

We also have a look at the individual features to see, how important they are in the classification. A random forest has a built in importance measure for the features used: The mean Gini decrease (Breiman, 2001). This importance is provided in

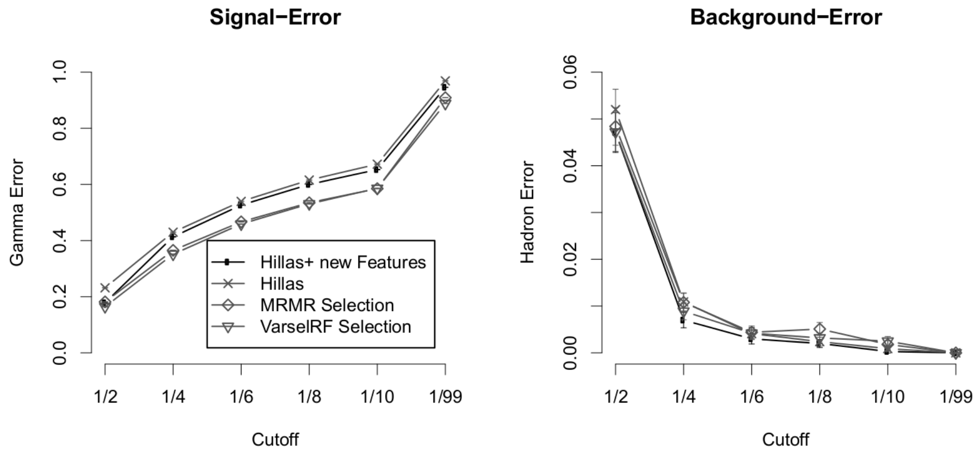


Figure 21: Error rates depending on the signal background ratio. The signal error is the ratio of falsely classified signal events to all signal events. Analogously the background error is the ratio of falsely classified background events to all events.

Table 6. Among the new features, varquot has overall a very high importance and alignskewkullobs has also a higher importance than most other variables, while most other new features have smaller importance.

Like in the variable selection methods we observe again the high importance of alignskewkullobs in comparison to other variables, especially the other Kullback Leibler Divergence, alignskewkulltheor, which is the same distance, but with the role of the fitted and observed distributions being switched. A possible explanation is that, as we see in Subsection 4.6.1, the Kullback Leibler divergence obs. has high contributions where the observed distribution has high values, whereas the Kullback Leibler divergence theor. has high contributions where the fitted distribution has high values. That means that KL divergence obs. detects mainly situations where the fitted distribution has smaller values than the observed distribution. In turn, the KL divergence theor. tends to have high values where the fitted distribution has larger values than the observed one. As background events

Table 6: The importance of each variable when using all variables in a random forest.

Variable	Importance	Variable	Importance
Hillas-Parameters:		**New Features:**	
Size	123.57	alignnormhell	129.02
Width	185.93	normhell	63.95
Length	126.63	skewhell	59.06
Area	103.01	normkullobs	69.23
Delta	56.56	normkulltheor	59.71
Alpha	440.92	normchi	100.48
Concentration1Pixel	137.56	alignnormchi	71.50
Concentration2Pixel	158.30	skewchi	75.71
NumberIslands	3.28	skewkullobs	63.92
NumberShowerPixel	48.88	skewkulltheor	60.93
LeakageBorder	98.12	alignnormkullobs	93.87
LeakageSecondBorder	266.03	alignnormkulltheor	63.88
		alignskewchisq	60.53
		alignskewhell	121.04
		alignskewkullobs	161.30
		alignskewkulltheor	76.19
		varquot	725.25

usually scatter more on the camera plane and tend to take extreme values more often than signal events, through the Kullback Leibler divergence obs. one is more likely to detect those events, because it puts more weight on such outliers.

`varquot` and `alignskewkullobs`, the two most important among the new variables, both use the information about alignment and the shape of signal events, which is why they might have such high importances in the classification. This shows that it is important to use all information available.

5 Combination of the two methods

In this last Chapter we combine the threshold optimization we did on MAGIC data with the variables constructed using FACT data, by using Hillas parameters and new variables together on a FACT data sample and optimizing the threshold of the random forest as outlined in Chapter 3.

Unlike MAGIC data, FACT data does not provide us with an estimate of the energy of each event, so that energy-bin-wise analysis cannot be done here. Instead, we use the threshold optimization on the whole data set.

From the FACT data we draw 500 times training, test and Off data. We draw the samples in the same way as in Chapter 4.8.5. The number of hadrons in the test and Off data is drawn from a Poisson distribution with mean 1000. A number of gammas is added according to the used signal to background ratio of 1:50 or 1:100. Each time we train a random forest on the training data and try to estimate the number of signal events in the training data. Instead of looking at varying thresholds like in Chapter 4.8.5, we here use different ways of setting the threshold and compare the quality of the estimation of signal events.

We compare the following settings:

- *Allvars_thropt*: All variables including the new variables constructed in Chapter 4.1 used with a random forest of which the threshold is optimized as presented in Chapter 3. This combines both methods introduced in this work.
- *Varsel_thropt*: Only the variables selected by the varSelRF selection method combined with the new threshold optimization.

- *Hillas_thropt*: Only Hillas variables combined with the new threshold optimization.

- *Allvars_rec*: Hillas variables and the new distance based variables used in a random forest where the threshold is chosen by setting a value for the recall (here 0.1, as this seems to be the best value as seen in Figure 12), with which we already compared the threshold optimization in Chapter 3.4.4.

- *Varsel_thropt*: The variables chosen by the varSelRF selection method combined with the recall threshold selection currently in use.

- *Hillas_rec*: Only Hillas parameters combined with the recall threshold selection currently in use. This is what is currently done in the MAGIC experiment.

The results can be seen in Figure 22 where boxplots of the 500 estimations of the number of signal events using the different approaches are plotted. It can be seen that for both ratios the new variables combined with the threshold optimization presented in this work (*Allvars_thropt*) lead to a smaller variance when estimating the number of signal events than when using the two methods currently in use (*Hillas_rec*). The current methods also seem to slightly underestimate the number of signal events. It can also be seen that the new method leads to a more symmetric estimation.

The variable selection, however, seems not to have a positive influence on the estimation of signal events, when combined with the new threshold optimization (*Varsel_thropt*). The boxplot is about as wide as for all variables in both signal to background ratios. It is a bit surprising that the estimation using the variable selection seems to be slightly worse than with all variables, as in Chapter 4.8 it could be seen that the variable selection improved the TPR of the classification significantly.

The variable selection combined with the currently used recall method (*Varsel_rec*) gives surprising results, especially for a ratio of 1:100. The variance is very small in this ratio, but there seems to be an undesirably large bias. The same effect, only at a smaller scale, can be seen in the ratio of 1:50. The Hillas and new variables combined with the recall method (*Allvars_rec*) seem to give the largest variance in the estimation.

Overall, using both, the Hillas variables and the newly constructed ones, together with the threshold optimization introduced in this work, seems to give the most desirable results, as their estimation of the number of signal events is unbiased and has smaller variance than most other methods.

The average threshold chosen by the methods is 0.0099 for the new method *Allvars_thropt* and 0.0002 for the currently used reference method *Hillas_rec*. The latter thus chooses smaller thresholds of about a factor of five. The standard deviation of the chosen thresholds is 0.0466 for the new methods and 0.0008 for the currently used methods, so the new method chooses much more varying thresholds. This indicates that the new method adapts more to the data at hand, while the reference chooses similar thresholds in all situations.

It would be desirable to do the threshold optimization binwise like we did in Chapter 3. That way the estimation could probably be improved further, as usually different thresholds are chosen for different energy bins by our threshold optimization. Furthermore, the logistic regression approaches cannot be used without information about the energy of a particle. However, unlike in the MAGIC experiment, in the FACT experiment the energy of the particles is not estimated before

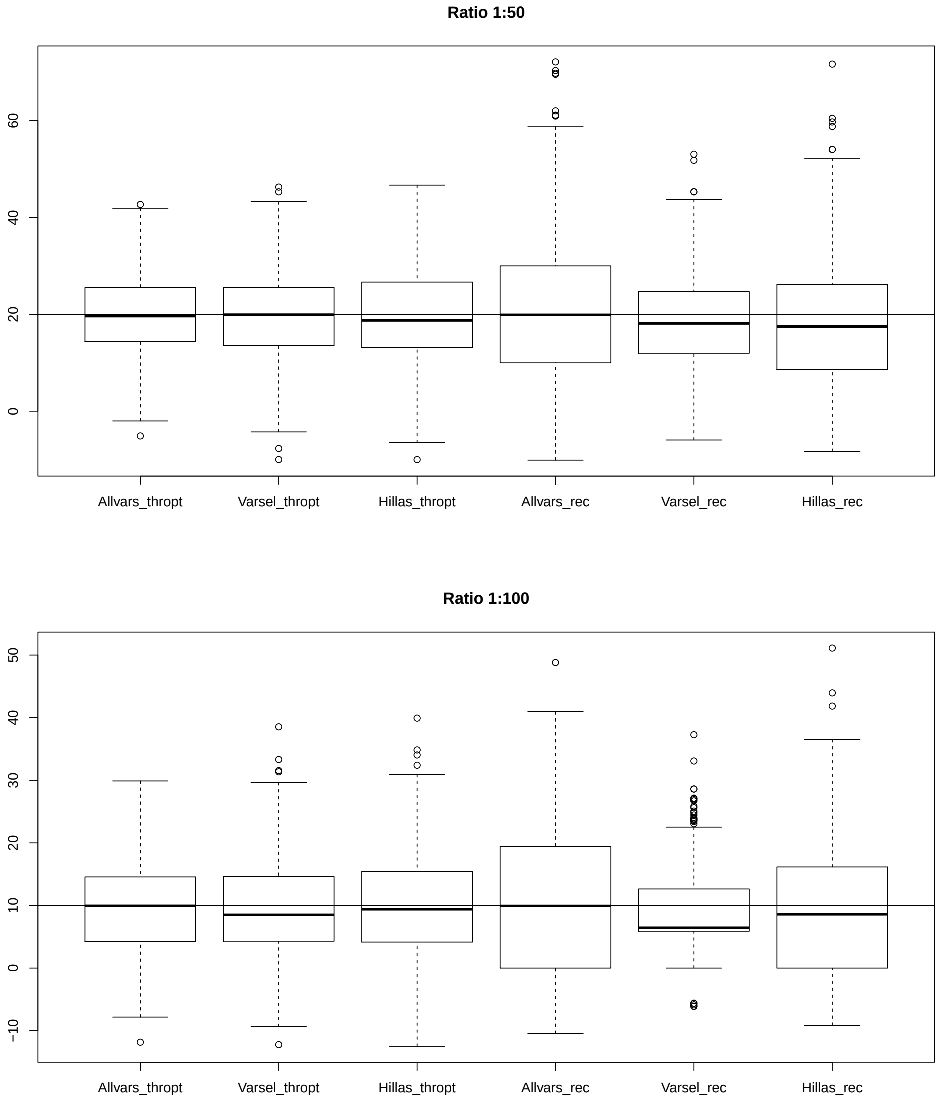


Figure 22: Boxplots of 500 estimations of signal events in some test data with two different signal to background ratios. Different approaches for the classification have been used. The estimation method itself is the same for all classifications.

the gamma hadron separation, so that in this experiment (at this time) the energy cannot be used. The results in Chapter 3 show that this information is very useful in the threshold optimization and should be made available in this experiment. But as seen above even without the energy binning the new methods estimate the number of signal events with a smaller error.

6 Conclusion

In this work we have introduced a method which chooses an optimal discrimination threshold in highly imbalanced data with unknown misclassification costs. To choose the threshold the method minimizes the MSE of an estimator of the number of positives in the data. Estimating this number well is important in the analysis chain of the MAGIC telescopes, used here to investigate the performance of the method. Minimizing the MSE is a general principle where the estimation of the number of positives is only one possible application. The method can be used whenever one is interested in estimations which are based on the outcome of a classification.

The MSEmin method can be used to optimize a threshold in the whole data or in subsets such as energy bins. In our application the best performance was reached by combining the principle of minimizing the MSE with a Logistic Regression, that is, a method where the threshold is not fixed, but depends on a covariate. Logistic regression is of course not the only method with which the MSEmin method can be combined. Other methods, which provide a fixed or variable threshold are equally possible.

We have applied some assumptions for developing the method introduced here. One is the assumption that the N_{ij} follow binomial distributions. Implicitly this means that all events are classified correctly or not with a constant probability, which is the same within positives and within negatives. In the MAGIC experiment, the probability of classifying an event correctly depends for example strongly on the energy of the particle which induced the event (e.g. Voigt, 2010). Nevertheless, the MSE based on this assumption fits the empirical MSE (as an estimator for the real MSE) quite well over the whole range of possible discrimination thresholds and for different class imbalances so that this assumption seems to provide good approximations. Another assumption we made implicitly is that there is no data set shift (Quionero-Candela et al, 2009), that is that there are no differences in the distributions between training data and actual data. As it is impossible to receive an actual dataset with 100% correct labels, it is likely that the probability distributions in our simulated data do not exactly match actual data, meaning it is likely that there is at least a covariate shift. However, this problem is inherent in any method in our application and thus a general examination of this problem is necessary, taking also any other steps of the MAGIC analysis chain into account. This is, however, beyond the scope of this work.

When looking at the training and Off data one may ask why the negatives in these data sets are treated separately. As the false positive rate (FPR) can also be inferred from Off data as well, the negatives in the training data seem to be obsolete. However, it is necessary to keep FPR and the number of falsely classified events in the Off data set, N_{01}^{Off}, independent from each other. Otherwise an estimation bias is introduced.

Keeping that in mind, the MSEmin method introduced here provides an estimation method for the number of positives in a data set with high class imbalance, which outperforms other methods currently in use. Further improvement can be achieved by combining the method with a Logistic Regression approach.

We have also discussed the usage of variables in a classification to separate the gamma-ray signal from a hadronic background. We have introduced Hillas parameters and reasoned why these parameters need improvement. We have established a connection between Hillas variables and bivariate Gaussian distributions. We have seen that Hillas variables alone cannot include all information available about signal events and have extended the idea of fitting an ellipse to fitting a bivariate distribution. Although most of the information on a Gaussian fit is given by the Hillas variables (except for a constant factor c), the Gaussian fit allows construction of some additional variables. The new variables are based on fitting a normal or skew-normal distribution to a shower and using distance measures for densities to evaluate the distance of an event to the fitted distribution. We have seen that the classification is improved by the new features. Especially a fitted skew-normal distribution with a fixed alignment combined with the Kullback-Leibler distance seems to be important in the classification along with an additional variable based on the quotient of variances. We have also seen that feature selection methods do not only decrease the number of variables, but also that the reduced number gives better results in the classification in terms of misclassification.

In a third step we combined the new variables with the threshold optimization. We have seen that the two methods combined lead to significant improvements of the estimation of the number of signal events, which is important for the reconstruction of energy spectra, one of the main goals of the analysis of this kind of data. It can be concluded that both new methods introduced here can improve the analysis chains currently in use by the MAGIC and FACT experiments.

7 References

Aharonian, F. A. (2004) *Very High Energy Cosmic Gamma Radiation A Crutial Window on the Extreme Universe.* World Scientific Publishing Co.Pte. Ltd.

Albert, J. et al. (2008) Implementation of the Random Forest Method for the Imaging Atmospheric Cherenkov Telescope MAGIC. *Nuclear Instruments and Methods in Physics Research A*, **588**, pp. 424 - 432.

Albert, J. et al. (2007) Unfolding of Differential Energy Spectra in the MAGIC Experiment. *Nuclear Instruments and Methods in Physics Research A*, **583**, pp. 494 - 506.

Aleksic, J. (2012) Performance of the MAGIC Stereo System Otained with Crab Nebula Data. *Astroparticle Physics*, **35**, pp. 435 - 448.

Aleksic, J. (2010) MAGIC TeV Gamma-ray Observations of Markarian 421 during Multiwavelength Campaigns in 2006. *Astronomy and Astrophysics*, **519**.

Aliu, E. et al. (2009) Improving the Performance of the Single-Dish Cherenkov Telescope MAGIC through the Use of Signal Timing. *Astroparticle Physics*, **30**, pp. 293 - 305.

Azzalini, A. (2014) *The R 'sn' Package: The skew-normal and skew-t Distributions (version 1.0-0).* URL http://azzalini.stat.unipd.it/SN

Azzalini, A. and Dalla Valle, A. (1996) The Multivariate skew-normal Distribution. *Biometrika*, **83**, p. 4

Azzalini, A. and Capitanio, A. (1999) Statistical Applications of the Multivariate skew-normal Distribution. *Journal of the Royal Statistical Society, Series B*, **61**, pp. 579 – 602.

Backes, M. et al. (2007) Long Term Monitoring of Bright TeV Blazars with the MAGIC Telescope, *Astronomische Nachrichten*, **328**, p. 677.

Becherini, Y. et al. (2011) A New Analysis Strategy for Detection of Faint Gamma-ray Sources with Imaging Atmospheric Cherenkov Telescopes. *Astroparticle Physics*, **34**, pp. 858 - 870.

Berntsen, J., Espelid, T.O., Genz, A. (1991) An Adaptive Algorithm for the Approximate Calculation of Multiple Integrals,*ACM Trans. Math. Soft.*,**17** (4), pp. 437 - 451

Blobel, V. (2010) Unfolding - Linear Inverse Problems, *Notes for the Terrascale workshop at DESY May 2010*

Bock, R.K. et. al. (2004) Methods for Multidimensional Event Classification: A Case Study using Images from a Cherenkov Gamma-ray Telescope. *Nuclear Instruments and Methods in Physics Research A*, **516**, pp. 511 - 528.

Boinee, P., Barbarino, F., de Angelis, A., Saggion, A. and Zacchello, M. (2006) Neural Networks for Gamma-Hadron Separation in MAGIC. In: Sidharth, B.G., Honsell, F. and de Angeles, A.: Frontiers of Fundamental and Computational Physics, p. 297.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, pp. 5 – 32.

Bretz, T. (2006) *Observations of the Active Galactic Nucleus 1ES 1218+304 with the MAGIC-Telescope*, PhD Thesis, Bayerische Julius-Maximilians-Universität, Würzburg

Carmona, E., et. al. (2008) Monte Carlo Simulation for the MAGIC-II System. In: *Proceedings of the 30th International Cosmic Ray Conference*, **3**, pp. 1373 - 1376.

Chadwick, P.M., Latham, I.J., Nolan, S.J. (2008) TOPICAL REVIEW: TeV Gamma-ray Astronomy. *Journal of Physics G Nuclear Physics*, **35**(3).

Cherenkov, P.A. (1934) Visible Emission of Clean Liquids by Action of Gamma Radiation. *Doklady Akademii Nauk SSSR*, **2**, pp. 451+.

Deiters, F. (2013) *Glättung und Image Cleaning im FACT Experiment*, Bachelor Thesis, TU Dortmund

Diaz-Uriarte, R. and Alvarez de Andres, S. (2005) *Variable Selection from Random Forests: Application to Gene Expression Data*, Technical Report.

Domingo-Santamaria, E., Flix, J., Rico, J., Scalzotto, V., Wittek, W. (2005) The DISP Analysis Method for Point-like or Extended Gamma Source Searches / Studies with the MAGIC Telescope. In: *Proceedings of the 29th International Cosmic Ray Conference*, **5**, pp. 363 - 366.

Errando, M. (2006) *Study of Optical Properties of Last Generation Photodetectors for Cherenkov Astronomy Applications*, Masters thesis, Universitat Autonoma de Barcelona

The FACT collaboration (2014) The FACT Telescope Web Pages,
URL: http://isdc.unige.ch/fact/

Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, pp. 861 - 874

Fegan, D.J. (1997) TOPICAL REVIEW: Gamma/Hadron Separation at TeV Energies. *Journal of Physics G Nuclear Physics*, **23**, pp. 1013 - 1060.

Firpo Curcoll, R., Delfino, M., Neissner, C., Reichardt, I., Rico, J., Tallada, P., Tonello, N. (2011) The MAGIC Data Processing Pipeline. *Journal of Physics Conference Series*, **331**(3), pp. 32 – 40.

Fomin, V.P., Stepanian, A.A., Lamb, R.C., Lewis, D.A., Punch, M., Weekes, T.C. (1994) New Methods of Atmospheric Cherenkov Imaging for Gamma-ray Astronomy. I. The False Source Method. *Astroparticle Physics*, **2**, pp. 137 - 150.

Fredholm, E. (1903) Sur une classe d´equations fonctionnelles, *Acta Math.*, **27** (832), pp. 365 - 390.

Genz, A. C., Malik, A. A. (1980) An Adaptive Algorithm for Numeric Integration over an N-dimensional Rectangular Region, *J. Comput. Appl. Math.*, **6** (4), pp. 295 - 302

Greenwood, P.E. and Nikulin, M.S. (1996) A Guide to Chi-Squared Testing. In: *Wiley Series in Probability and Statistics*, Wiley

Grieder, P.K.F. (2010) *Extensive Air Showers*, Springer, Heidelberg

Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, pp. 1157 – 1182.

Hadasch, D. (2008) *Study of the MAGIC Performance at High Zenith Angles and Application of the Results on a Very High Energy Gamma Ray Flare of the Blazar PKS 2155-304*, Diploma Thesis, Technische Universitaet Dortmund, Dortmund

Heck, D., Knapp, J. (2010) EAS Simulation with CORSIKA: A Users Manual. Forschungszentrum Karlsruhe, http://www-ik.fzk.de/corsika

Hillas, A.M. (1985) Cherenkov Light Images of EAS Produced by Primary Gamma. In: *Proceedings of the 19th International Cosmic Ray Conference ICRC*, San Diego, **3**, p. 445

Hinton, J. (2009) Ground-based Gamma-ray Astronomy with Cherenkov Telescopes. *New Journal of Physics*, **11**(5).

Hinton, J.A., Hofmann, W. (2009) Teraelectronvolt Astronomy. *Annual Review of Astronomy & Astrophysics*, **47**, pp. 523 - 565.

Hsu, C.C. et al. (2007) The Camera of the MAGIC-II Telescope, In: *Proceedings of the 30th International Cosmic Ray Conference ICRC*, Merida

Jogler, T. (2009) *Detailed Study of the Binary System LS I +61o303 in VHE Gamma-rays with the MAGIC telescope.* PhD thesis, Technische Universitaet Muenchen

Kohnle et. al. (1996) Stereoscopic Imaging of Air Showers with the First Two HEGRA Cherenkov Telescopes. *Astroparticle Physics*, **5**, pp. 119 – 131.

Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, pp. 79 – 86.

Lessard, R.W., Buckley, J.H., Connaughton, V., Le Bohec, S. (2001) A New Analysis Method for Reconstructing the Arrival Direction of TeV Gamma Rays Using a Single Imaging Atmospheric Cherenkov Telescope. *Astroparticle Physics*, **15**, pp. 1 - 18.

Li, T.P., Ma, Y.Q. (1983) Analysis Methods for Results in Gamma-ray Astronomy. *Astrophysical Journal*, **272**, pp. 317 - 324

The MAGIC collaboration (2014) The MAGIC Telescope Web Pages,
URL: https://magic.mpp.mpg.de/home/

Maier, G., Knapp, J. (2007) Cosmic-ray Events as Background in Imaging Atmospheric Cherenkov Telescopes. *Astroparticle Physics*, **28**, pp. 72 - 81.

Majumdar, P. et al. (2005) Monte Carlo Simulation for the MAGIC Telescope. In: *Proceedings of the 29th International Cosmic Ray Conference*, **5**, p. 203.

Mazin, D. (2007) *A Study of Very High Energy Gamma-ray Emission from AGNs and Constraints on the Extragalactic Background Light.* PhD thesis, Technische Universitaet Muenchen

Milke, N., Rhode, W., Ruhe, T. (2011) Studies on the Unfolding of the Atmospheric Neutrino Spectrum with IceCube 59 using the TRUEE Algorithm. In: *Proceedings of the 32nd International Cosmic Ray Conference*

Milke, N., Doert, M., Klepser, S., Mazin, D., Blobel, V., Rhode, W. (2012) Solving Inverse Problems with the Unfolding Program TRUEE: Examples in Astroparticle Physics *Nuclear Instruments and Methods in Physics Research A*, **697**, pp. 133 – 147.

Nakajima, D. et al. (2013) New Imaging Camera for the MAGIC-I Telescope, In: *Proceedings of the 33rd International Cosmic Ray Conference*, Rio de Janeiro

Narasimhan, B. (2013) *cubature: Adaptive Multivariate Integration over Hypercubes*, R package version 1.1-2., C code by Steven G. Johnson, http://CRAN.R-project.org/package=cubature

de Naurois, M. and Rolland, L. (1996) A High Performance Likelihood Reconstruction of γ-rays for Imaging Atmospheric Cherenkov Telescopes. *Astroparticle Physics.*

Nelder, J.A., Mead, R. (1965) A Simplex Method for Function Minimization. *The Computer Journal*, **7**(4), pp. 308 - 313.

Nikulin, M.S. (2001) Hellinger Distance. In: *Hazewinkel, M.: Encyclopedia of Mathematics*, Springer

Ohm, S., van Eldik, C., Egberts, K. (2009) Gamma/Hadron Separation in Very-high-energy Gamma-ray Astronomy Using a Multivariate Analysis Method. *Astroparticle Physics*, **31**, pp. 383 - 391.

Peng, H. and Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, pp. 1226 – 1238.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D. (2009) *Dataset Shift in Machine Learning.* The MIT Press

Schlickeiser, R. (2002) *Cosmic Ray Astrophysics.* Springer-Verlag, Berlin Heidelberg

Sheng, V., Ling, C. (2006) Thresholding for Making Classifiers Cost-sensitive. In: *Proceedings of the 21st National Conference on Artificial Intelligence*, AAAI Press, **1**, pp. 476 - 481

Sidro Martin, N. (2008) *Discovery and Characterization of the Binary System LSI +61303 in Very High Energy Gamma-Rays with MAGIC*, Dissertation, Universitat Autonoma, Barcelona

Sobczynska, D. (2007) Natural Limit on the Gamma/hadron Separation for a Stand Alone Air Cherenkov Telescope. *Journal of Physics G Nuclear Physics*, **34**, pp. 2279 - 2288.

Thom, M. (2009) *Analyse der Quelle 1ES 1959+650 mit MAGIC und die Implementierung eines Webinterfaces fuer die Automatische Monte-Carlo-Produktion*, Diploma Thesis, Technische Universitaet Dortmund, Germany

Voigt, T. (2010) *Exploration und Vorverarbeitung von MAGIC-Daten zur Gamma-Hadron-Separation.* Diploma Thesis, Technische Universitaet Dortmund, Germany

Voigt, T., Fried, R., Backes, M., Rhode, W. (2013) Gamma-Hadron-Separation in the MAGIC Experiment, In: Spiliopoulou, M., Schmidt-Thieme, L., Janning, R.: *Data Analysis, Machine Learning and Knowledge Discovery*, Springer

Voigt, T., Fried, R., Backes, M., Rhode, W. (2014) Threshold Optimization for Classification in Imbalanced Data in a Problem of Gamma-ray Astronomy, *Advances in Data Analysis and Classification*, Springer, DOI 10.1007/s11634-014-0167-5

Voigt, T., Fried, R. (2014a) Distance Based Feature Construction in a Setting of Astronomy, In: Lausen, B., Krolak-Schwerdt, S., Boehmer, M.: *Proceedings of the ECDA 2013*, Springer

Voigt, T., Fried, R. (2014b) Modeling Cherenkov Telescope Images for Variable Construction in Classification, In: *Proceedings of the IWSM 2014*, accepted

Weekes, T. (2003) *Very High Energy Gamma-Ray Astronomy.* Institute of Physics Publishing, Bristol/Philadelphia