MAGIC 99-4
30 April 1999

Some thoughts on gamma/hadron separation using composite probabilities

R.K.Bock

1. Problem definition

This note concentrates on optimizing the separation of hadrons and gammas, based on images seen in a telescope camera. It is assumed that Monte Carlo samples of gammas (signal) and hadrons (background) exist as training samples. As illustration existing Monte Carlo data for the 127-pixel CT1 of HEGRA are used, at 0 zenith angle, more specifically events in runs 08 (11715 gammas) and 11 (58532 hadrons). From these, a single test statistic is eventually derived which seems to give good, perhaps presently optimal separation for any chosen contamination. Ultimately, the intention is to use similar methods (hopefully cross-checked with other ideas) for the final data from the MAGIC telescope.

The term optimization needs definition, of course. Most abstract optimization procedures require the problem to be reduced to a single variable which is derived from the training samples using various parameters. The variable is minimized (or maximized) as a function of these parameters. It is suggested that we generalize this notion in the sense that we optimize an entire curve, which relates the coordinates in the Neyman- Pearson (or decision quality) diagram. The two coordinates of that diagram are the acceptance of background and the acceptance of signal, both normalized to cover the range 0 to 1. Obviously, two samples that can be fully separated, i.e. an ideal test situation, would allow us to reach the point (0;1), two meaningless tests that accept or reject everything, respectively, would give the points (1;1) and (0;0). A non- discriminating variable with a changing cut parameter (or a number of irrelevant tests varied somehow) gives us points lying on a 45-degree line, assuming linear coordinates. The goal of optimizing in this diagram is, therefore, to find a curve that lies above all points that can be found by any other selection procedure conceived; in other words, we want to have a procedure that gives maximal signal acceptance (also called efficiency, and complementary to loss or cost or significance, in statistics parlance) for a given background acceptance (also called contamination, the complement of power or purity). It seems important to me that we know how to optimize our efficiency for different values of background suppression, as it will often depend on the physics to be extracted, at which operational point we want to work, and the existence of different physics will best be shown by highlighting differences between samples that have been enriched in different and controlled ways.

2. Cuts in variables, and their optimization

An intuitively simple (and widely used) method of separating different samples consists of expressing each event (in the case of CT1 an image with 127 pixel values) in terms of several test statistics, variables that show statistical distributions different for signal and background samples. Statistical distributions are probability density functions visualized as histograms of observed variables. Any test statistic defines a projection of the image space onto an axis; the smaller the overlap of the signal and background events in such a projection, the more the variable contributes to overall separation. Moving a cut parameter across the overlap zone in a projection generates a curve in the Neyman-Pearson (NP) diagram which goes from the coordinates (0;0) to (1;1), passing more or less close to the ideal point (0;1). Multiple test statistics make it difficult to generate a single curve by cuts: any combination of cut parameters generates a point in the NP diagram, and what we are then out to find is the optimal envelope of these points. Usually, this becomes largely an operation of trial and error. I have tried to approach the optimization systematically in the space defined by the 7 variables found on the standard ntuples (the extended Hillas parameters: size, alpha, dist, length, width, mdist, conc); I call these the 'basic' variables. Optimizing, of course, needs a single variable, to be minimized as a function of the 7 cuts (or more, if high and low cuts are applied); even if a well-defined variable can be found, the optimizing programs require guidance (step sizes, range to be explored, etc) and judgement so as to avoid falling into local minima. I have used multiple starting points and explored the envelope curve by stepping along with the best gradient, or maximizing signal acceptance with the ratio of acceptance(sig)/acceptance(bgnd) constrained to remain the same, thus defining a straight line orthogonal to the presumed envelope. These results are shown in figure 1.

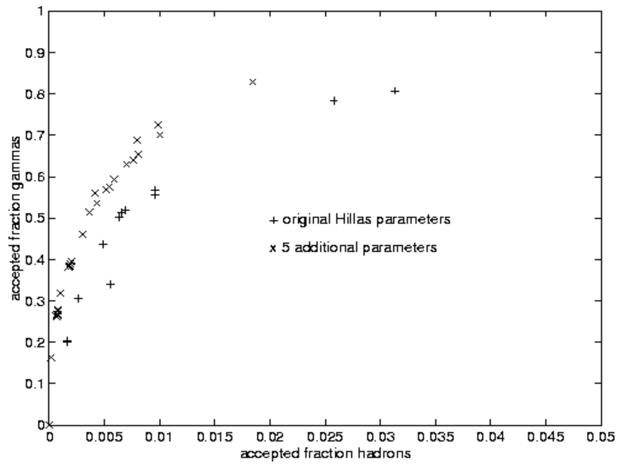

Figure 1: Various attempts to optimize single-variable cuts; the additional parameters are defined in the appendix (variables 9-13). In this optimization, a pre-cut sample of events (see section 4 below) has been used, and the acceptances corrected accordingly.

Following the idea of 'dynamical cuts' as advanced in Ref 1 and Ref 2, I also introduced variables that are combinations of the basic variables, essentially stretching the original variables' space such that the dynamic cuts can be approximated by simple cuts in the new projections. Also some of these results are shown in fig.1. Each point thus found has, of course, its own set of cuts, and interpolation between the points on the envelope curve is non-trivial.

3. Defining a single test statistic: gammaness or hadronicity, as a composite event probability

It is conceivable, to define an overall probability for a given event to belong to the signal or background distribution, still based on the statistical properties of a number of different test variables as above. One can take the 7 basic variables as above, histogram them separately for signal and background, and determine the fractional probability for each bin. Each event is then assigned as composite probability the product of the individual bin probabilities that correspond to the values of its specific variables. A signal event will, on average, have a higher composite probability for the signal histograms, and equally so for the background events which will be more likely to resemble the background histograms. The potential advantage lies in the fact that such probabilities (call them gammaness or hadronicityadr) constitute a single test statistic, with no hard cuts necessary in any projection. Using this variable for applying a cut, a single smooth curve in the NP diagram results.

There remains a problem of optimally choosing the variables and the histogram bin positions and widths; I dodge the first problem by using all variables (but see the section on pitfalls below), and have largely circumvented the second problem by determining variable bins such that each bin, for one sample (gammas), contains the same number of entries. In practice, this is achieved by sorting all values of each variable in increasing order, and taking the limits of quantiles as the bin limits. Example: if the sample size is 500 events and every 10th value (the 10-quantiles) is taken from the sorted list as bin limit, each bin will contain exactly 50 events for this sample. Fig. 2 shows, for illustration, the probabilities for hadrons in variable bins derived from gammas, for the variable alpha, a well-discriminating single test statistic.

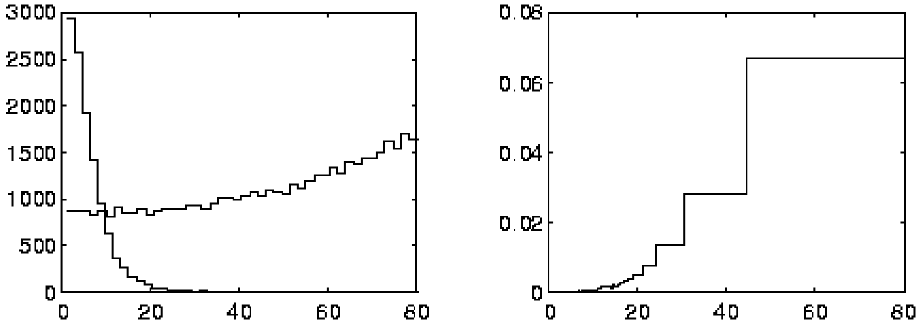

Figure 2: Left: the plain histograms for the variable alpha, for gammas (peaking at small values) and hadrons (covering the entire range from 0 to 90 degrees). Right: bin probabilities for hadrons, in bins determined to give a flat distribution for gammas.

In the following figure 3, the basic variables and 10 variables (the seven basic variables, and 3 additional ones) have been used to derive a composite probability: 300 variable-size bins were determined for each variable from the gamma sample, and the bin probabilities derived for hadrons. Each event was then assigned a hadronicity by forming the product of all bin probabilities corresponding to the events variables. The method is competitive with the above cut optimization as shown in figure 1, although not consistently so (the variables used were not exactly the same). The effect of the additional three variables is clearly visible at low acceptances.

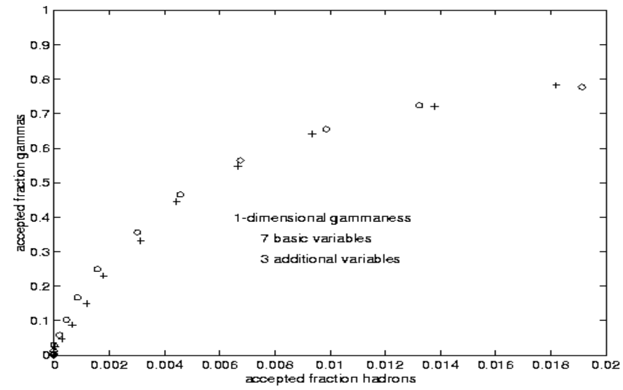

Figure 3: the NP diagram for a one-dimensional gammaness, a composite probability derived from two different sets of variables: + for the 7 basic variables, o for the composite probability derived from 10 variables.

4. Gammaness from 2-dimensional histograms

The above advantages of a single probability can be combined with the wish to capture also the interrelations between variables, which are at the origin of the dynamic cuts of Ref 1 and Ref 2. We can generalize the variable-bin histogram to two dimensions by first dividing the plane spanned by two characteristic variables along one axis using the quantile method, then subdividing, inside each slice thus defined, into two- dimensional histogram bins according to the distribution of the second variable in this slice. This is demonstrated in fig.4 which shows constant-probability bins for the two variables width and alpha, using gammas, and the probabilities derived for these bins from hadrons.

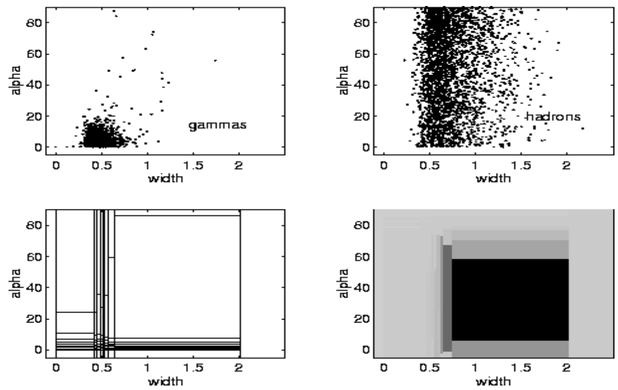

Figure 4: Top row: the two plots of alpha vs. width for gammas and hadrons. lower left: the constant-probability bins for gammas (16 are used along each axis, only 8 are shown for

clarity); lower right: the probability density for hadrons in the gamma bins shown as gray scales; overflows are possible.

Figure 5 below shows what can be obtained from the basic variables using composite probability from all possible combinations (7 variables, 21 two-dimensional plots). Also shown is the effect of adding some of the new test statistics. Both curves are mostly superior to the optimized cuts of figure 1, except for the smallest hadron acceptances.

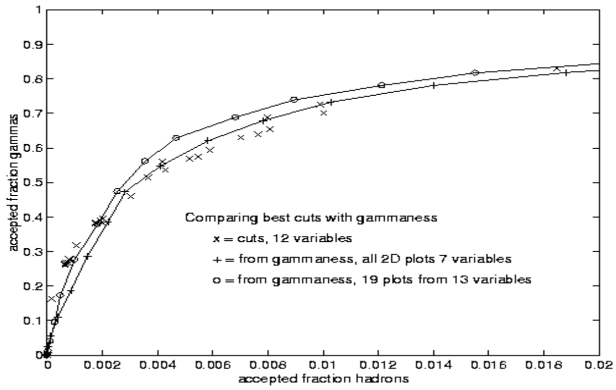

Figure 5: the NP diagram (+) for composite probabilities using all 2D plots for the 7 basic variables (21 2D plots). Using only the most significant 16 plots of these, and adding 3 plots (variables 11, 13, and 14 against energy (size), see appendix for variable definition), a somewhat better result is obtained (o). Two of the additional variables are derived from the basic variables, to approximate the non-linear oblique dynamic cuts, one is derived directly from the pixels. Both curves are mostly superior to the cut optimized in figure 1 (which use all additional variables 9-14, see appendix), except for the smallest hadron acceptances.

Next, I have used all 13 variables with a total of 57 2D-plots entering the composite probabilty: I used all plots relating the basic variables among themselves, and all plots relating basic and additional variables. I have then attempted to hand-optimize the use of plots, as it became apparent that using additional information does not systematically improve results. In fact, using only half the plots gives consistently higher gamma acceptances for constant acceptance of hadrons, if the selection of plots is done on some criteria describing the deviation of the probability distribution in a given plot from its average value, i.e. the difference between the gamma and hadron probabilities in this plot. Let us call this general effect 'smearing by irrelevant information'.

Along with this, I found out that a similar effect can be shown by pre-cutting the data samples: using simple cuts in the variables alpha and length, the original hadron sample can be reduced by a factor 6.5, cutting into the gammas by only 0.015. This pre-cutting has no effect when optimizing cuts (and has, in fact, been used in fig.1), but improves the composite probabilty discrimination, presumably again by an effect of smearing. This is demonstrated in fig.6.

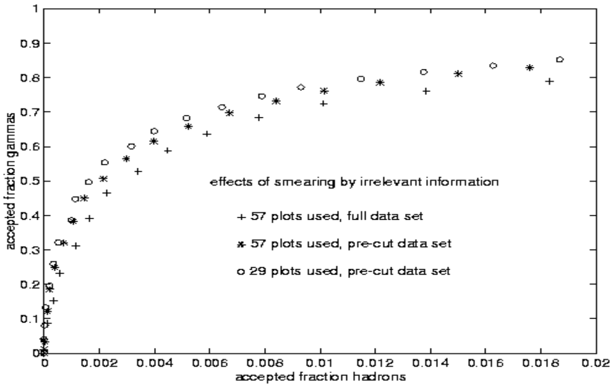

Figure 6: using more information can smear results unfavorably: the effect of selective use of 2D plots, and the effect of pre-cutting. All results in fig.6 are consistently superior to the best optimized cuts of fig.1, or to any results obtained with one-dimensional probability distributions, as in fig.3.

The method lends itself to computer optimisation. For instance, if one starts with composite probabilities derived from all possible 2-dimensional plots (78 for 13 variables), and subsequently starts removing those plots that show the worst smearing effect, one at a time, one can continually improve the discrimination until some optimum is reached. I have defined as relevant parameter for optimizing the average gamma efficiency for fixed hadron acceptance, using the (interpolated) points for the five hadron acceptances 0.001, 0.002...0.005. Some results, again showing the effect of information smearing, are shown in figure 7. It also appears that the pre-cutting of data has no effect on the final result, which is a reassuring result because giving weight to the optimizing procedure used.

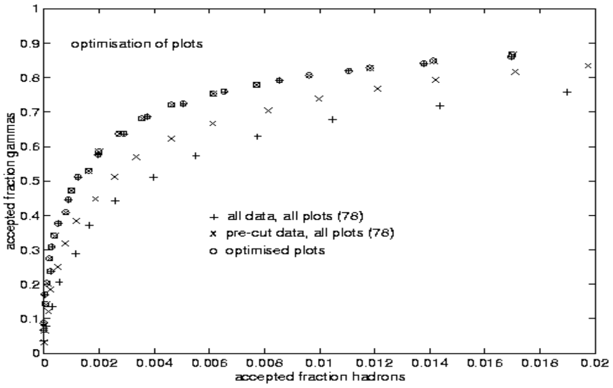

Figure 7: Systematic optimization of the 2-dimensional distributions entering composite probabilities. The optimized curves (encircled symbols) are much the same for full and pre-cut data sets, although the variables and distributions used are not: the unoptimized curves use 78 distributions, 17 of which are retained for the full, and 29 for the pre-cut data set.

The best obtained values are given in the following table. To compare with the best discrimination values obtained by D.Kranich (private communication), I have also used the same method on data samples selected for high ($\geq$5TeV) Monte-Carlo energy; a better gamma/hadron separation (compared to the full sample) is possible on this hadron-enriched sample (4692 gammas, 45187 hadrons) also with the method of composite probabilities. Note that all optimizations (except the first line) start with 2D distributions from all 13 variables. The first line, using only the basic variables, shows that the 6 additional variables introduced (see appendix) have a noticeable effect.

Hadron acceptance	plots	.001	.002	.005	.010	.020	.050	.100
Gamma acceptances								
7 variables, opt.	8	.352	.468	.656	.763	.859	.928	.966
pre-cut, hand-opt.	29	.386	.533	.676	.779	.859	.944	.978
all	78	.271	.398	.553	.669	.762	.854	.913
opt. from all	17	.465	.578	.722	.807	.873	.921	.948
pre-cut	78	.357	.461	.633	.739	.837	.941	.981
opt. from pre-cut	29	.471	.585	.729	.810	.881	.952	.979
opt. high-E	18	.639	.726	.827	.881	.923	.968	.986

5. Pitfalls, detours, remaining work

Zero probabilities: if bins are defined from one statistical distribution (I have used the gamma sample throughout), and the probabilities derived for the other (hadron) sample, then some bins will have zero content and hence zero probability. As the composite probability is obtained as product of individual bin probabilities, a single zero will suffice to set the composite probability to zero, resulting in ugly peaks and step functions. I circumvent the problem by counting in every bin at least half an event; this smoothes out the distributions in a very satisfactory way.

Data splitting: if all data are used for defining bins and probabilities, and subsequently for testing the discrimination quality, then the discrimination becomes uniformly better the finer the bin size is set. Convince yourself by considering infinitely fine bins: all events that have been used for defining probabilities (hadrons) will have non-zero probability, all others (gammas) will have most or all probabilities zero. An unrealistic ideal discrimination would result. I have, therefore, systematically split the hadron sample in two equal parts, using one half for defining the probabilities, and the second half for testing the quality of the method. The same data splitting used for defining the variable-size bins seems to have no effect, but no exhaustive tests have been done. The need for data splitting reminds of a very similar problem arising in the training and testing of artificial neural networks.

More than two dimensions: clearly, going from one-dimensional bin probabilities to two-dimensional ones much improves the result. It is a natural question, then, why higher-dimensional bin probabilities should not be used. I have not tried this, mostly because the bin contents (and hence the statistical error of the probabilities) will dwindle away quickly, and too many choices (which 3D plots?) have to be taken.

Further optimizing: the pre-cutting has not been systematically optimized, although the method lends itself to an exploration by computer methods beyond that of selecting the optimal combination of distributions to use. Also additional variables might be envisaged, although the demonstrated smearing effect must be kept in mind: they may distort the weights given to the underlying independent variables. As was also shown, the optimization described takes care of smearing.

Defining bins and probabilities: there is no reason why bins and probabilities should be defined from the gamma and hadron samples, respectively (as I have done); inverting their roles could give a different result, and using the two variables thus obtained together might improve the overall result. I have used for all plots 16 bins in each variable, also this may not be optimal.

Curves intersecting in the NP diagram: the examples shown above make it clear that curves in the NP diagram can intersect, so that in different parts of the diagram different optimal parameters will be found. A uniformly optimal procedure has not been found, so far; conceivably, multiple ways of calculating composite probabilities may have to be introduced, depending on the part of the curve one is interested in. Certainly the optimization used pushes up the low hadron-acceptance end of the curve, and ignores the acceptances larger than 0.005.

External parameters and real data: the validity of the above study is limited to the Monte Carlo runs 8 and 11, with no change of external parameters like zenith angle or impact point. Eckart also pointed out, rightly, that robustness with respect to misalignments, background light etc. need to be estimated, for this and other methods.

MAGIC telescope data: it is most important now to try these and competing ideas on MC-generated data for MAGIC. The composite probabilities described above may well be outdone by better ideas, but a metric of comparison similar to the above should be rigorously applied.

6. Conclusion

The method of deriving, for each event, a single composite probability from Monte-Carlo generated two-dimensional probability distributions, allows to reduce the multi-dimensional space to a single variable. This variable may be called gammaness, and different cut parameters in this variable permit to choose different points along a curve relating the accepted fraction of gammas and the accepted fraction of hadrons.

The method lends itself to computer optimization, and the obtained results seem in all aspects superior to 1-dimensional cuts in multiple variables. It was also shown that additional variables can improve the results, if care is taken to introduce a minimal set of two-dimensional probability distributions: redundant probability components or superfluous variables have a tendency of smearing the result.

Much more work is needed to adapt the method to data simulated for the MAGIC detector, and to study its robustness with respect to changing external parameters like zenith angle, background light, etc. This analysis also needs embedding in the future analysis chain; for the results presented, stand-alone programs in Matlab have been used on a PC, with ASCII files as input (variables were derived from images by a small C program, or converted from ntuples).

Appendix

The basic variables that were used in this study were:

2 log(size), 3 alpha, 4 dist, 5 length, 6 width, 7 mdist, 8 conc,

I added for the above analysis six further variables, without having studied in detail their individual value:

9 and 10: longitudinal and transverse pixel spread of the cluster: these are like length and width, but not weighted with energy;

11: longitudinal asymmetry, obtained by (signed) sum of square distances along the longer cluster axis, weighted with energy;

12: value of maximum pixel

13: derived from basic variables: width - sqrt(log(size)/3)

14: derived from basic variables: 10*width + log(size)

References

Ref 1: MPI-PhE/97-11, D.Kranich, Verbesserung der Gamma/Hadron-Trennung;

Ref 2: MPI-PhE/97-27, D.Petry, Beobachtung hochenergetischer Gamma-Strahlung.