ANI

Applied Nonparametric Inference

Reference Manual Version 19.1

Ashot Chilingarian

Cosmic Ray Division of Yerevan Physics Institute, 2 Alikhanian Brothers, Yerevan 36, Armenia

e-mail - chili@aragats.am
ANI home-page: http://www.crd.yerphi.am/ANI_User_Guide_Introduction

Science is in the business of making up stories called hypotheses and testing them, then trying its best to make up better ones.

Yerevan 2019

CONTENTS

CONTENTS **3**

LIST OF FIGURES **5**

LIST OF TABLES **5**

FOREWORD **7**

WHAT ANI IS INTENDED TO DO 7

WHAT ANI IS NOT INTENDED TO DO 7

NOTATION 7

REFERENCES 8

FURTHER REMARKS 8

ACKNOWLEDGMENTS 8

COPYRIGHT NOTICE 9

CHAPTER 1. INTRODUCTION **10**

1.1 PROBLEMS OF DATA ANALYSIS IN COSMIC RAY PHYSICS 10

1.2 SIMULATION FOR EXPERIMENTS IN THE ASTROPARTICLE PHYSICS 11

1.3 ANI STRATEGY 12

1.3.1 SELECTION OF THE BEST SUBSET OF VARIABLES 12

1.3.2 BAYESIAN ANALYSIS 13

1.3.3 NEURAL NET SOLUTIONS 13

1.3.4 ROBUSTNESS CONCEPT 14

1.3.5 VISUALIZATION 14

1.3.6 LIMITATIONS AND PERSPECTIVES OF DEVELOPMENT 14

CHAPTER 2. HOW TO USE ANI **15**

2.1 PROGRAM SUMMARY 15

2.2 KEY WORDS 15

2.3 SOURCE CODE 15

2.4 HISTORY OF ANI VERSIONS 16

2.5 RESTRICTIONS ON DATA SIZE 16

2.6 THE MAIN ANI PROCEDURES (MODES) 17

2.7 DATA FILES 17

2.8 ANI-SETUP 18

2.8.1 BOOKKEEPING SETUP 18

2.9 DATA TO DRIVE ANI DESCRIPTION AND EXAMPLE OF THE B.IN FILE: 19

CHAPTER 3. STATISTICAL INFERENCE IN COSMIC RAY PHYSICS **23**

3.1 NONPARAMETRIC INFERENCE 23

3.2 PARAMETRIC CLASSIFICATION 23

3.3 NONPARAMETRIC CLASSIFICATION - MONTE CARLO STATISTICAL INFERENCE 24

3.3.1 BAYESIAN PARADIGMA 26

3.3.2 BAYESIAN DECISION RULES 27

3.3.3 NONPARAMETRIC PROBABILITY DENSITY ESTIMATORS 28
3.3.4 NONPARAMETRIC REGRESSION 29
3.3.5 BAYES ERROR ESTIMATION 29
3.3.6 FRACTION ESTIMATION 31
3.3.7 THE BOOTSTRAP PROCEDURE 31
3.4 THE NEURAL CLASSIFICATION TECHNIQUE 33
3.4.1 NEURAL ESTIMATION 34
3.4.2 FF NEURAL NETWORK TRAINING 35
3.4.3 GENETIC ALGORITHMS 35
3.4.4 NET TOPOLOGY 36
3.4.5 STOPPING RULES 36
3.5 KNN ALGORITHM OF FRACTAL DIMENSION ESTIMATION 36
CHAPTER 4. ANI TESTING (GAUSSIAN DATA) 38
4.1 BAYESIAN ANALYSIS 38
4.1.1 PROBABILITY DENSITY ESTIMATES 38
4.1.2 THE BOOTSTRAP STATISTICAL MOMENTS 40
4.1.3 DISTRIBUTION MIXTURE SHARE ESTIMATION 41
4.1.4 BAYESIAN MAPPING 43
4.2 ARTIFICIAL NEURAL NETWORKS MODELS 45
4.2.1 NEURAL CLUSTERS 45
4.2.2 DETECTION OF THE CRAB NEBULAE BY THE WIPPLE COLLABORATION 45
CHAPTER 5. PRIMARY NUCLEI CLASSIFICATION IN 3 CATEGORIES (KASCADE DATA ANALYSIS) 47
5.1 THE EXTESIVE AIR SHOWER (EAS) SIMULATION 47
5.2 VALIDATION OF MODELS 49
5.2.1 COMPARISON OF THE SINGLE EAS VARIABLES 49
5.2.2 CORRELATION ANALYSIS 49
5.2.3 PROBABILISTIC DISTANCES 50
5.2.4 KASCADE EXPERIMENTAL DATA 51
5.2.5 QGS AND VENUS MODELS COMPARISON 51
5.2.6 THE KASCADE CLASSIFICATION MATRICES 52
5.2.7 COLORED NUCLEAR MAPS (MASCS) 54
5.2.8 FRACTION ESTIMATION 58
5.3 THE EXAMPLES OF ANI OUTPUTS 61
BIBLIOGRAPHY 64
APPENDIX A B.IN INPUT FILE EXAMPLE 69

LIST OF FIGURES

Figure 4.1: Parzen density estimates of standard normal distribution....39
Figure 4.2: The histogram of "probability integrals"....40
Figure 4.3 The Bayesian clusters for the samples from Gaussian populations with different means....44
Figure 4.4 The Bayesian clusters for the samples from Gaussian populations with same means....44
Figure 4.5 Deterministic Search....46
Figure 4.6 Genetic Search....46
Figure 5.1: Features distribution for proton and iron....47
Figure 5.2. Proton and iron events distribution in 3-dimensional space of features....48
Figure 5.3: QGS (red area) and VENUS (white area) clusters and experimental events distribution in N_e, N_μ^{tr}....52
Figure 5.4: 3-way map, calorimeterinformation. Green points represent oxygen MC data....54
Figure 5.5: 3-way map, array information. Green points represent oxygen MC data....55
Figure 5.6: QGS model: 3-way map, array information. $p_{p\equiv red,}\, p_{0\equiv green,}\, p_{Fe\equiv blue.}$ Black triangles represent oxygen MC data. $E_{MC} \in [1x10^{15}, 3x10^{15}]eV$....55
Figure 5.7: QGS model: 3-way map, array information. $p_{p\equiv red,}\, p_{0\equiv green,}\, p_{Fe\equiv blue.}$ Black triangles represent oxygen MC data. $E_{MC} \in [3x10^{15}, 3x10^{16}]eV$....56
Figure 5.8: VENUS model: 2-way map, array information. $p_{p\equiv red,}\, p_{Fe\equiv blue.}$ Black triangles represent oxygen MC data. $E_{MC} \in [1x10^{15}, 3x10^{15}]eV$....56
Figure 5.9: VENUS model: 2-way map, array information. $p_{p\equiv red,}\, p_{Fe\equiv blue.}$ Black triangles represent oxygen MC data. $E_{MC} \in [3x10^{15}, 3x10^{16}]eV$....57
Figure 5.10: VENUS model: 2-way map, array information. $p_{p\equiv red,}\, p_{0\equiv green,}\, p_{Fe\equiv blue.}$ Black triangles represent oxygen MC data. $E_{MC} \in [1x10^{15}, 3x10^{15}]eV$....57
Figure 5.11: VENUS model: 2-way map, array information. $p_{p\equiv red,}\, p_{0\equiv green,}\, p_{Fe\equiv blue.}$ Black triangles represent oxygen MC data. $E_{MC} \in [1x10^{15}, 3x10^{15}]eV$....58
Figure 5.12: VENUS model: Reconstructed classification results using two (p, Fe) (lower graphs) and three (p, 0, Fe) (upper graphs) classes for different sets of parameters....58
Figure 5.13 : QGS model: Reconstructed classification results using two (p, Fe) (lower graphs) and three (p, 0, Fe) (upper graphs) classes for different sets of parameters....59

LIST OF TABLES

Table 4.1: The quality check of the Parzen density estimator, samples from Gaussian populations N(0,1)-N(1,1), adaptive estimator....40
Table 4.2: Bootstrap expectations and bootstrap standard deviations of sampling statistics....41
Table 4.3: Fraction estimation, M=100; B=10....42
Table 4.4: Fraction estimation, M=1000; B=10....42
Table 4.5: Fraction estimation, N=4, B=1000; B=10....43
Table 4.6: WHIPPLE Crab detection, 1988-1989....45
Table 5.1: EAS features detected by KASCADE experiment....48
Table 5.2: P-values of statistical tests for proton and iron classes for different models: t - Student, D - Kolmogorov- Smirnov, U -Mann-Whitney....49
Table 5.3: Correlation matrix for QGS model....50
Table 5.4: Correlation matrix for Venus model....50
Table 5.5: The best feature subsets according to the Bhattacharya distance....51

Table 5.6: Experimental data homogeneity t est features used: N_μ^{CD}, E_h^{sum} 51
Table 5.7: Experimental data homogeneity test features used: N_e, N_μ^{tr},N_μ^{CD}, E_n^{sum} 51
Table 5.8: Experimental data homogeneity test features used N_e, N_μ^{tr} 51
Table 5.9: One dimensional tests : *t*-Student, *D*-Kolmogorov-Smirnov, U-Mann-Whitnay 51
Table 5.10. Comparison of exp. data with VENUS and QGS models.....52
Table 5.11. One dimensional tests for models andexperiment 52
Table 5.12: Calorimeter data, features used N_μ^{CD}, E_h^{sum} 53
Table 5.13: Array data, features used N_e, N_μ^{tr} 53
Table 5.14: KASCADE data, features used N_e, N_μ^{tr},N_μ^{CD}, E_h^{sum}.....53
Table 5.15: 3-way classification by N_μ^{CD}, E_h^{sum}.....53
Table 5.16: 3-way classification by N_e, N_μ^{tr} 53
Table 5.17: 3-way classification by N_e, N_μ^{tr},N_μ^{CD}, E_n^{sum} 53
Table 5.18:2-way classification by N_μ^{CD}, E_h^{sum} 53
Table 5.19:2-way classification by N_μ^{CD}, E_h^{sum} 53
Table 5.20:2-way classification by N_e, N_μ^{tr},N_μ^{CD}, E_h^{sum} 53
Table 5.21: Separability index forKASCADE.....54
Table 5.22: $4.1 \leq lg_{10} N_\mu^{tr} \leq 4.4, M_{TS} = 150, M_{exp} = 64$ 59
Table 5.23: $lg_{10} N_\mu^{tr} \geq 4.4, M_{TS} = 120, M_{exp} = 20$.....59
Table 5.24: $4.1 \leq lg_{10} N_\mu^{tr} \leq 4.4, M_{TS} = 150, M_{exp} = 64$.....59
Table 5.25: $lg_{10} N_\mu^{tr} \leq 4.4, M_{TS} = 120, M_{exp} = 20$.....59
Table 5.26: $3.39 \leq lg_{10} N_\mu^{tr} \leq 3.65, M_{TS} = 555, M_{exp} = 68420(array)$ 60
Table 5.27: $3.65 \leq lg_{10} N_\mu^{tr} \leq 3.85, M_{TS} = 215, M_{exp} = 56100(array)$ 60
Table 5.28: $3.85 \leq lg_{10} N_\mu^{tr} \leq 4.1, M_{TS} = 140, M_{exp} = 20400(array)$.....60
Table 5.29: $4.1 \leq lg_{10} N_\mu^{tr} \leq 4.4, M_{TS} = 135, M_{exp} = 7540(array)$ 60
Table 5.30: $lg_{10} N_\mu^{tr} \leq 4.4, M_{TS} = 110, M_{exp} = 2285$ 60

FOREWORD

WHAT ANI IS INTENDED TO DO.

ANI (Applied Nonparametric Inference) code represents a unified methodology of big data analysis intensively using Machine Learning algorithms. Ani was widely used in high-energy Astroparticle physics community, in genome analysis and other. ANI platform consisted of different approaches to draw scientific inference and provide users with whole system of coherent methods for:

- optimal utilization of information contained in experimental data and in the statistical models;
- best feature subsets selection and initial dimensionality reduction of big data;
- optimized methods of multivariate probability density estimation;
- scanning of multivariate spaces to reveal embedded nontrivial structures;
- nonparametric estimation of regression function;
- Neural and Bayesian classification and background rejection.

The main problems solved with ANI:

- event -by - event analysis of Extensive Air Shower (EAS) data;
- determination of the type and the energy of primary particles;
- hadronic background rejection in detection of very high energy gamma rays with imaging Cherenkov telescopes;
- genome analysis: finding a subset of genes responsible for the colon cancer;
- Lightning classification.

WHAT ANI IS NOT INTENDED TO DO.

- Simulation of nuclear - electromagnetic cascade in the atmosphere;
- estimation of the detector response;
- ANI is not intended also for the repeated solution of identically parameterized problems (such as shower size reconstruction) where a specialized program will be in general much more efficient.

NOTATION

d Dimensionality

L Number of classes

M Number of experimental events

v_i Vector of measured variables

u_i Vector of simulated events

(A, P) Stochastic mechanism which generates experimental data

$(A, \hat{P})$ "Controlled" stochastic mechanism, obtain with simulation code

V Event (measurement, feature) space

A Basic state space

P_A Prior measure

$C_{A\hat{A}}$ Losses (cost) measure

$p(v / A_k)$ Conditional probability density function (pdf)

$\hat{p}(v / A_k)$ Estimate of conditional pdf

$\hat{p}(A_i / v)$ Posterior density.

REFERENCES

The basic idea of applied Bayesian statistics, that the probability is orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information, was presented in [1, 2].

The posterior robustness of Bayesian decisions, as a key issue for applied statistical analysis was discussed and illustrated in [3, 4].

Among the current books on Bayesian statistics one of the best ones is [5].

Though simulation in data analysis in high energy physics is widely used, we can aware of a very few systematic investigations of theoretical aspects about how data may be compared with their simulated counterpart [6, 7, 9, 10].

The development of the Monte Carlo statistical inference for high-energy astroparticle physics experiments is presented in [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21].

The references on another key problem of applied Bayesian inference- nonparametric multivariate density estimation, could be found in tutorial section.

FURTHER REMARKS

This manual consists of four chapters:

- Introduction
- A reference guide explaining the concepts and ANI user interface;
- A tutorial about mathematical foundations of the Bayesian nonparametric statistical inference;
- Two tutorials on the ANI implementation and interpretation of results for the samples from Gaussian populations and for data from surface arrays registering Extensive Air Showers (EASs), namely MAKET-ANI and KASCADE experiments [22].

ACKNOWLEDGMENTS

I am grateful to Prof. Dr. G. Schatz and Prof. Dr. H. Rebel and all members of KASCADE collaboration for the stimulating scientific atmosphere and supporting the interest in the development of the ANI program for KASCADE data analysis.

Dipl. Phys. M.Roth prepared full simulations of KASCADE experimental data used as a priory information for ANI, and he was instrumental in implementing different modes of ANI for KASCADE data analysis. Dipl. Phys. A.Vardanyan achieved several improvements of the neural modes of ANI and participated in testing ANI in various regimes. He wrote the description of ANI graphical interface.

Both M.Roth and A.Vardanyan have been very helpful in preparation of this manual. The chapter 5 of this manual concerning KASCADE data analysis was prepared in collaboration with Prof. Dr. H.Rebel, Dipl. Phys. M.Roth and Dipl. Phys. A.Vardanyan.

I'm grateful to Dr. J.Knapp and Dr. M.Cowley for useful discussions and valuable remarks concerning new methods of data analysis. I thank the Wipple collaboration for the permission of the use of the Crab Nebula database.

The ANI graphical interface and genetic algorithm were designed and programmed by S.Chilingaryan. The calculations presented on figure (4.5) were prepared by E.Sevinian.

This work was a part of a WTZ project (X131.2) of scientific cooperation between Forschungszentrum Karlsruhe, Germany and Yerevan Physics Institute, Armenia. It has been partly supported by the research grant number 94694 of Armenian government and by A116 ISTC grant.

In 2019 the ANI code was moved to modern platform and revised.

COPYRIGHT NOTICE

CHAPTER 1.
INTRODUCTION

PROBLEMS OF DATA ANALYSIS IN COSMIC RAY PHYSICS

No model is true, only useful

The scientific method is characterized by data classification, the study of their interrelations and relations to past experience, accumulated in various theories and hypotheses. Usually, it is impossible either to prove or to refuse hypotheses by deductive method. The challenge is to draw sensible conclusions from noisy, discrepant information.

Modern arrays of particle detectors covering a large area are measuring different numerous secondary products of the primary cosmic ray (PCR, mostly protons and fully stripped nuclei) interactions in the atmosphere. Only a simultaneous measurement of a large number of independent parameters in each individual Extensive Air Shower (EAS) can yield reliable information to reconstruct the particle mass and its energy as well as the phenomenological characteristics of strong interaction with atmosphere nuclei.

The ambiguity of interpretation of the results of experiments with cosmic rays is connected with significant gaps in our knowledge of the characteristics of hadron - nuclear interaction at highest energies, and, with strong fluctuations of all shower parameters. The extra difficulties are due to the use of Monte - Carlo simulations of development and detection of different components of nuclear electromagnetic cascade and inherent reductionism of these models.

To make the conclusions about the investigated physical phenomenon more reliable and significant, it is necessary to develop a unified theory of statistical inference, based on non-parametric models, in which various nonparametric approaches (density estimation, Bayesian decision making, error rate estimation, feature extraction, sample control during handling, neural net models, etc…) would be incorporated.

The most important part of the presented approach is the quantitative comparison of multivariate distributions and use of a nonparametric technique to estimate the probability density in the multidimensional feature space. As compared to the earlier used methods of inverse problem solution, in ANI the object of analysis is each particular event (a point in the multivariate space of measured parameters - feature space) rather than alternative distributions of model and experimental data.

By considering all measured EAS parameters simultaneously, we are able to incorporate important information about their relationship and outline in multidimensional feature space nonlinear regions where events of definite type mostly grouped. That is why, along with the averaged characteristics, the belonging of each experimental event to a certain class (primary nuclei group, hadron or gamma ray image) is determined.

The advocated approach was used to estimate the upper limit of the iron nuclei fraction according to the gamma - family characteristics, registered by PAMIR collaboration [23, 24, 25, 26, 27]. It was the first attempt to make PCR studies on event-by-event basis.

A multidimensional analysis was applied for classification of the Cherenkov images of air showers registered by the Whipple observatory. It was shown that the use of several image parameters together with their correlations can lead to a reduction of the background rejection

down to a few tenths of a percent while retaining about 50% of useful (gamma-rays induced) events. The application of multivariate technique to the famous Crab detection data file (Whipple observatory - 1988-1989) [29], proves the advantage of the new background suppression technique and - achievement of considerable enhancement of source detection significance [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45].

ANI was intensively used for KASCADE and MAKET ANI experiment data analysis [46, 47, 48, 49, 50, 51, 52, 53]. Obtained results on the energy spectra of different nuclei groups of PCR represent a "mass spectroscopy" in "knee" region, and allows to obtain the basic results in the long standing PCR origin and acceleration problems.

1.2 SIMULATION FOR EXPERIMENTS IN THE ASTROPARTICLE PHYSICS

The most difficult and most important part data analysis in the high-energy physics is the comparison of competitive hypotheses and decision making on the nature of the investigated physical phenomenon.

In the cosmic ray physics the main technique of statistical inference, connected the with problem of determination of initial physical parameters (such as mass composition and energy spectrum of PCR, strong interaction characteristics, flux of very high energy gamma rays from point sources, etc…), - is the direct problem solution with detailed simulation of the cosmic ray traversal through the atmosphere and the detectors with a following comparison of the simulated and experimental data. Actually, an algorithm is constructed, which describes EAS development and registration of its different components on the observation level, which is based on a certain model of the process investigated, i.e. the set of the parameters that characterize the PCR flux and interaction of incident hadrons with the air nuclei.

By simulations with different models and comparing the experimental and model data, a class of models is selected, which describe the experimental data satisfactorily. Such an approach allows us to discard a certain class of non-satisfactory models, but the available experimental data usually do not allow one to select a unique model among the many proposed.

For almost all problems of inference, the crucial question is whether the used models are in fact consistent with data. Of course, any inference is conditioned on the model used, and, if the model is oversimplified, so that essential details are ever omitted, or improperly defined, at best only qualitative conclusions may be done. The actual need of a reliable M.C. code can be illustrated by contradicting results of physical inference on elemental composition, presented in the literature on basis of different simulation procedures. Based on the measured intensity of gamma-families detected by emulsion chamber experiments at high mountain altitudes [56], the Fuji-Kanbala group concluded that beyond the knee iron nuclei are dominating in the primary flux [57]. On the other hand, using the same data, the Chacaltaya and Pamir collaborations [58] insist that the origin of the knee is due to a change of the character of the hadronic interaction, while the elemental composition remains approximately unaltered.

Using an alternative observation technique, the Fly's Eye group claims [59] a significant

change of primary composition towards approximately pure proton content at energies larger than 10^{17}eV, in contrast the Akeno group could not find a significant change in composition at these energies [60]. These cases of conflicting results, reflect essentially the different theoretical "calibration" of the data and inadequate simplifications of the analysis techniques.

There is a general agreement [55] about the vital importance to develop a simulation program that invokes the actually best and most detailed treatments of all physical processes, relevant to EAS development, in order to be used, tested and cross checked for consistency by different groups, without adapting the model parameters for each actual case in different way. With the CORSIKA [54] program developed in context of the KASCADE experiment there is a modern code available, using efficient Monte-Carlo techniques. It includes various options of alternative interaction models, generally accepted to be valid up to 10^{16}-10^{17} eV.

Similarly, there is an established code: The CERN detector simulation package GEANT [61] for the simulation of detector response function of complex detector setups in presence of various different radiation sources. This widely used code is widely used to transform the theoretical EAS variables into the form “registered” by the actual apparatus and can be compared with the experimental data. Both programs CORSICA and GEANT are published, freely accessible and matter of continuous refinements by concerned study groups.

For comparing experimental and simulated data, we need to generate large samples of the observables, including realistically modeled all types of fluctuations for mimicking experimental detectors measuring with high precision many EAS parameters. Additionally, a "comparator" is necessary, based on coherent statistical methods for the analysis of nonparametric multivariate distributions with hidden nonlinear dependences. What we need is a well-defined technique, what one can call Monte - Carlo Inference. The presented approach to develop such a techniques considers the classification and hypothesis testing problems in the framework of Bayesian and Machine Learning paradigms and the main steps of the unified data analysis methodology are as following.

1.3 ANI STRATEGY

1.3.1 SELECTION OF THE BEST SUBSET OF VARIABLES

The data preprocessing is a first step in data analysis.

Both experimental and simulation event are checked for "outliers" - events with very big deviations according to "expected" values.

Then the "best subset of measurements" (in the sense of great discriminative power in classification to 2 categories) is selected from a variety of potentially useful variables. Proceeding from the initial dimensionality, a "worst" feature is selected, according to reduction of Bhattacharya distance, calculated for all variants of the reduced variable subsets.

The quantitative comparison of variables is done by means of the, so called, P-values of statistical tests, showing the relative discriminative power of the variable. The greater this value, the smaller the probability of the H_0 hypothesis to be correct. H_0 consists in the statement, that the two independent samples come from the one and the same population. The smaller this probability, one can reject this hypothesis with greater confidence and accept the alternative hypothesis: that two samples come from different populations. And the "distance" between populations is proportional to the P-value. Three different tests are used: the parametric Student test, the nonparametric Kolmogorov - Smirnov and Mahn - Whitney tests, based on ranks. The last two

does not require any assumptions about the shape of the underlying distributions.

Since interdependencies among the variables affect most of multivariate analyses procedures, it is worth examining the correlation matrix of EAS parameters.

The correlation analysis can help in the selection of the "best" feature pairs. The calculated Fisher matrix point on significant difference of pair-wise correlation in different classes.

1.3.2 BAYESIAN ANALYSIS

The Bayesian approach provides the general framework of incorporating of prior and experimental information. Bayesian decision rule, that assigns observable v to the class with the minimal posterior losses, takes into account estimates of conditional probability density and all possible losses due to any decision.

The posterior density is basis of statistical decisions on particle type and on the closeness of the simulated and experimental data samples. The term closeness refers to the degree of coincidence, similarity, correlation, overlapping or any such measure. Bayes classifier provides minimal losses (probability of error) among all classifiers for the same feature set. However, the Bayes classification meets several difficulties, as the analytic expressions of conditional densities and, hence, the posterior ones, are unknown. Therefore, we are obliged to use their nonparametric estimates. Nonparametric in the sense, that density function is not a particular member of a previously chosen parametric distribution family, but an estimate based only on sample information, and - on very mild conditions on the underlying density (usually only continuity). The well-known Parzen and K Nearest Neighbors (KNN) estimators are used in ANI.

The nonparametric regression is used for energy estimation. The peculiarity of solution of the regression problem in the cosmic - ray physics is the fact that neither the true spectrum $f(E)$ nor the conditional density $P(v/E)$ are known in the general case. The method is based on the obvious fact that the events close to each other in some metric in the feature space have similar energy - the geometrical consistence.

1.3.3 NEURAL NET SOLUTIONS

The alternative very powerful classification and estimation technique is connected with the development of mathematical models of Feed-Forward Neural Nets (FFNN). The input layer of the feed - forward network have one node for each feature, signal processing is performed layer by layer starting from the input. Neurons of successive layers receive input only from neurons of the previous layer and each neuron in a given layer sends its output to all nodes in the next layer. The neuron(s) of the output layer produce the discriminant function(s).

The training is performed with simulated data or/and calibration results (if available). The initial values of net parameters are chosen randomly from Gaussian population with zero mean and not very large variance. The training of FFNN consists in multiple processing of all training samples with iterative modifications of connection coefficients (weights).

The quality function minimization is usually done by the "so called" back propagation method, the gradient descend is performed on the quality function with respect to the weights in order to minimize the deviations of the network response from the desired "goal" response. The main drawback of such methods is their convergence to local minimum, in contrast, implemented in ANI different scenarios of random search allows to escape from the local minimum region and continue the search till the better solution will be found.

A common complaint of NN training techniques is the dependence of the final classification scheme on the purity and finiteness of training sets (small training samples effects). However, due

to the inherent robust characteristics of FFNN, neural classification is relatively insensitive to modest impurities in the training sets.

1.3.4 ROBUSTNESS CONCEPT

In general, referring to a statistical estimator, the robustness means the insensitiveness to small departures from the idealized assumptions for which the estimator is optimized [63, 64]. The word "small" refers to both: small departures for all data, or large departures for a small number of data.

For example, the particular nonparametric density estimator has to be tuned according to the unknown distribution function. Our modification of estimators – the, so-called, probability density L-estimator didn't require determination of the unique "best" parameter for the whole data set, rather a wide interval of parameters, one of which will be automatically chosen for the appropriate data point.

1.3.5 VISUALIZATION

For such abstract procedures, as multivariate mapping and classification in multidimensional spaces, the visualization is of crucial importance. After classifying experimental data according to training sample classes, a necessary analysis step is to examine the initial feature space for outlining the regions of acceptation of one or another hypothesis. Usually there are physical arguments about location of such "clusters".

For example, "heave-nuclei" initiated EASs in $N_e - N_\mu$ coordinates tend to occupy left- top quarter; the Cherenkov images, initiated from primary γ –quanta, have very specific shape in Hillas parameter space, etc.

A special "DENCURVE" key word of ANI provide possibility to store and visualize the multidimensional clusters in CERN PAW ntuples.

A special Bayesian scanning of multivariate space produces nonlinear multidimensional clusters corresponding to EAS, initiated from chosen primary nuclei group.One can easily examine nonlinear interdependencies between EAS variables using wide possibilities of PAW utility.

1.3.6 LIMITATIONS AND PERSPECTIVES OF DEVELOPMENT

The potential difficulties and limitations of the ANI package are connected with model dependence of statistical inference. The question of correctness of the model itself is always open and we need a more general procedure to check the model validity and obtain physical results not so crucially depending on the prechosen models.

One possibility of model - independent inference is connected with cluster analysis: to scan the multidimensional feature space to find singularities of probabilistic measure without incorporating any simulations. However, difficulties will encounter with physical interpretation of the embedded structures.

The second one is connected with the idea of integration over plausible models. Proceeding from a list of acceptable models – a model-integration procedure (committee method) can be defined for tuning both astrophysical parameters (composition, spectra) and strong interaction parameters.

CHAPTER 2.
HOW TO USE ANI

PROGRAM SUMMARY

Title of program - ANI - Applied Nonparametric Inference

Computers - DEC-ALPHA, SGI, PENTIUM based UNIX workstations.

Operating system - UNIX, LINUX

N of bits in a word - 32

Programming language used - FORTRAN 77

Number of code lines > 10,000

2.2 KEY WORDS

- Monte-Carlo Statistical Inference;
- Nonparametric Methods;
- Pattern Recognition;
- Multivariate Statistical Techniques;
- Bayes Risk estimation;
- Probability Density estimation;
- Classification;
- Artificial Neural Networks in Data Analysis;
- Sampling methods;
- Genetic Algorithms;
- Evolutionary Programming.

2.3 SOURCE CODE

The source code is written in standard FORTRAN77 including several routines from CERN program library. The "structure-creating" style of Fortran programming was used. Any function or procedure are represented by separate units (subroutines, or IF loops), provided with lines of explanations.

The CMZ source code management system is used for bookkeeping and version archiving [65]. The same source code is available for the all platforms mentioned. The automatically check of platform will activate the appropriate to this platform translators, linkers and program libraries. The modifications of code performed on one platform are fully available for others, of course, if there is no significant difference in translators.

For installation on a new platform the paths to system and CERN libraries have to be mentioned explicitly in CMZ KUIP files.

2.4 HISTORY OF ANI VERSIONS

There exist several versions of ANI package developed for different computers and operation systems.

YerPhI BESM-6 VERSION - 1985.5

CERN IBM 3090 (VM) VERSION - 1986.6

FIAN PDP 11/70 1987.2

FIAN VAX VERSION - 1989.1

YerPhI EC 1045 VERSION - 1989.2

PATCHY VERSION - 1990.6

DUBLIN VAX (UNIX) VERSION - 1990.7

KfK IBM 3090 (MVS) VERSION 1993.5

MPI VAX (VMS) Version 1993.6

KfK UNIX version 1994.5

YerPhI LINUX version 1994.8

CERN NOMAD version 1995.4

YerPhI Silicon Graphics version 1995.10

YerPhI Pentium version 1996.1

YerPhI CMZ alpha-97 version 1997.03

CRD: HP ProLiant ML370 G5 Server version 2019.01

2.5 RESTRICTIONS ON DATA SIZE

Depending on platform used and memory available different restrictions on possible sizes of executed data files and formats are made by following declarations of Fortran PARAMETER command:

```
parameter (in=8, imb=50000, il=5, imp=50000, ikcl=ll, ipr=5)
parameter (maxley=5, maxnod=13)
```

Current parameter settings available on FZK ALFA's are as follows:

IL	maximal number of classes	5
IN	maximal data dimensionality	10
IMB	maximal number of events in training sample	2000000
IMP	maximal number of experimental events	2000000
IKCL	maximal number of nuclei width variants	17
MAXLEY	maximal number of neural net layer	5
MAXNOD	maximal number of nodes in each layer	13
IPR	number of cost function variants	5

All array declarations in ANI are made implicitly using above mentioned restrictions. The

declaration of arrays in subroutines also is made implicitly via transferred list of the formal parameters. These restrictions help to avoid main obstacle of Fortran programming_absence of the utilities controlling the language structures. If any erroneous array dimension request encounters in input stream, detailed error report message is send and execution of program stopped.

2.6 THE MAIN ANI PROCEDURES (MODES)

The most important Key word is the analysis MODE, specifying the particular statistical procedure to be used. The selected operation *MODE* is printed in the first line of analysis passport, containing also description of all data subsets executed, and parameters of data analysis procedures used. Usually data analysis started with determination of intrinsic dimensionality of data - *DIMDIM* mode, the two figures used for dimensionality estimates: the average of local dimensionality, and the global correlation dimensionality.

Then a best data subset (in sense of discriminative value) is selected by the BHATA mode: proceeding from the initial dimensionality, on each step of dimensionality reduction a "worst" feature is selected and eliminated, according to the value of Bhatacharya distance, calculated for each variant of obtained subsamples.

The *ONE DIMEN.* mode is examining single variables and evaluated their discriminative power.

The COVCOR mode can help in the selection of the best pairs of the features.

The calculation of the minimal achievable Bayes risk is performed in *ONE-LEAVE-OUT-* mode implemented to the training samples with known category.

The *CLASSIFICATION* mode performs the attributing of the experimental events according to training sample classes (a prior knowledge) using Bayesian decision rules. The true fraction of different types of events in the specially constructed distribution mixture is calculated if RECONSTRUCT key word is selected.

The *REGRO* mode is used for energy and mass estimation.

The alternative very powerful classification and estimation techniques represent the Neural Networks models. The parameters *LEARNING and CLASSIFI* are used as an analogical to Baysian modes; the topology of net (number of layers and number of neurons in each layer) and parameters of random search of best for discrimination net parameters are specified in the b.in input file .

2.7 DATA FILES

ASCII format files with standard headers are used for the input of information. Also, format of PAW NTUPLES is supported. Data files are defined and referred by their names. The procedure of data reading can be checked line-by-line. Different selections, according to variables subsets, variables cuts, events numbers are possible.

Special data files with fixed names are created to provide possibility of information exchange between different ANI modes. For instance the results of classification can be stored in a file to be executed by another ANI mode for estimation task (energy of the selected light and heavy nuclei). Several ANI modes are using statistical parameters and estimates calculated in previous runs. The Neural Net training scenarios implemented in ANI provided possibilities of multi-step search with tuning parameters and changing particular algorithms. The Bayes error estimates, calculated in *ONE-LEAVE-OUT-* mode are used in *CLASSIFICATION* mode for fraction estimation.

Thus, following files with fixed names connected to particular mathematical numbers, provided possibility of data exchange between different ANI modes:

b.tem - for Bayes risk estimates;

learn.dat - current values of neural net weights;

For user interface following files are used: b.in - input stream; b.out - output stream; b.sys - error reporting.

2.8 ANI-SETUP

The *ANISETUP* graphic interface is designed for ***ANI*** *(Analysis and Nonparametric Inference)* statistical analysis package. It is written on *TCL.7.6* script language 7.6 and *TK.4.2* toolkit, which are available on most of ***UNIX*** platforms. The interface consists of two main parts:

- bookkeeping of input and output information (Figure 2.1).
- main input script setup for running the ANI program (Figure 2.2).

2.8.1 BOOKKEEPING SETUP

Select b.in - Click on the icon and select "b.in" (input file). If it is the first run, the default file, named "b" will be downloaded.

Save b.in to... - Specify the name for current "b.in" to be saved. Different input files corresponding to the various operating modes will be archived under different names. By the default the "b" is saving with the same name.

Delete button - One can select and delete the input file from the archive.

Run button - Runs the main input setup.

Next part is for viewing run results. Variants of red color message: *New running, Program terminated correctly, Error detected.* If the second message is printed, one can view the current results. In case of third message one have to view the *"b.sys"* file for error report.

Run PAW++ - Interface to CERN PAW++

View b.out - View the output ASCII file, which is available in each running and contains the resulting information on current run.

View b.sys - Detected errors during current running.

Run CMZ - Run CMZ, change the source code and recompile the program.

At each new run all output files, besides the b.tem file (*"b.hbook", "b.out", "learn.dat")* will be overwritten.

File to save b.out - Specify the name for *"b.out",* write comments if necessary (next icon) and press *Save.*

Save PAW output - Specify the name for *"b.hbook"* and press *Save.*

Select learn.dat - from first icon of this line one can select the archived *"learn, dat"* file, which contains the trained *neural network* parameters and *Restore* it for continuing net training from the point reached at the previous training cycle. After net training one can specify the file name to save the obtained "learn.dat" file.

PAW file: - Select saved hbook file and run paw++*(View PAW Arc.),* or delete it *(delete PAW Arc.).*

OUT file: - Select saved output file and and *View...,* or *Delete...* it. ***EXIT*** button - Exit from ANI-SETUP.

2.9 DATA TO DRIVE ANI DESCRIPTION AND EXAMPLE OF THE B.IN FILE:

1. PARAMETER CONTROLS THE OUTPUT STREAM, (by selecting numbers from zero to 8, one can include various additional output information, to be printed into output file b.out, also permanently attached under mathematical number NT=2). DEBUG=
   3
2. NUMBER OF DIFFERENT DATA CLASSES TO BE HANDLED AND NUMBER OF BUTSTRAP REPLICAS (L and NBUT numbers):
   2, 10
3. TRAINING SAMPLES NAMES:
   KASCADE1000
   KASCADE1000
4. CONTROL(EXPERIMENTAL) SAMPLE NAME:
   KASCADE1000S
5. DUMP (ARCHIVE) SAMPLE NAME (name of the file to store current training sample with applied cuts and selections.
   NN-EST
6. PAW HBOOK NAME:
   b.hbook
7. TOTAL NUMBER OF VARIABLES IN DATA FILES (dimensionality of training and control samples):
   5
8. THE RELATIVE COORDINATES (coordinate of the first event)FOR EACH DATA FILE:
   0,0
9. TOTAL NUMBER OF EVENTS TO BE READ FROM EACH DATA FILE:
   500,500
10. FIRST EVENT COORDINATE and SIZE OF CONTROL (EXPERIMENTAL) DATA FILE:
    0,100000
11. STATUS (DENCURVE - producing numerous PAW plots):
    DENCURVE
12. OPERATION MODE*:
    (JMODE = DIMDIM, BHATA, ONE DIMEN., COVCOR, FWRITE, BETEST, ONE-LEAVE-OUT-, CLASSIFICATION, REGRO, BUTSTRAP, LEARNING, CLASSIFI, FAST, SUPERCUT, SOBOL-CUT, MULTI-CUT)
    LEARNING
13. THE TYPE OF DATA FILES(ASCII with header and weights, ASCII without HEADER, OR PAW NTUPLES, (ACCESS = SEQUENTAL, SIMPLE, NOMAD):
    SEQUENTAL
14. DENSITY ESTIMATION MODE. Two general nonparametric modes are implemented, kernel density estimator and K nearest neghbours estimator.
    (JDEN = PARZ or KNN):
    PARZ
15. WEIGHTS IN REGRO MODE (JDIST = LINEAR, SQUARE, UNIFORM):
    UNIFORM
16. FORMAT OF SEQUENTAL INPUT (FORMAT of ASCII string):
    (10F10.5)

17. NUMBER OF DIFFERENT A PRIORY PROBABILITIES AND LIST OF PROBABILITIES:
    (IAP number and AP array):
    3
    0.5, 0.5
    0.1,0.9
    0.01,0.99
18. RECONSTRUCT FIRST TYPE EVENTS PORTION?:
    RECONSTRUCTT
19. NUMBER AND VALUE OF NUCLEI WIDTHS (OR LIST OF NEAREST NEIGHBORS):
    (KCL number and F array)
    5
    0.3,0.4,0.5,0.6,0.7 or 15,25,50,100,1502,5 пт
20. MAXIMAL EXPONENT IN PARZEN DENSITY ESTIMATION AND STRANGNESS CRITERIUM IN BAYESIAN DECISION RULE: (expmax and Strange):
    9000000.,0.0000000000000000001
21. NUMBER OF NEAREST NEGHBOURS FOR DIMDIM AND REGRO MODES, NUMBER OF PRINCIPAL COMPONENTS IF PCA MODE SELECTED (NEI=):
    1 7
22. VARIABLES TO BE PROCESSED (AMOUNT AND RELATIVE NUMBERS) (N number and NUMB array):
    2
    3,4
23. THE MINIMAL DIMENSION OF BEST VARIABLES SUBSET TO BE CHOOSEN BY BHATA SUBROUTINE (INTDIM=):
    1
24. LOWER BOUNDS OF VARIABLES (AMIN array):
    -9999999,-9999999
25. UPPER BOUND (AMAX array):
    9999999,9999999
26. Random generator used (genert = pseudo, or lp-tau - a uniform sieve in N-dimensions): In BETEST mode also RANNOR, NORRAN and NORMCO generators are used,
    Pseudo
27. DATA TRANSFORMATION TYPE (normalization to 0-1 - renorm, principal component transformation -pea), :
    norenorm
28. PARAMETERS OF PAW HBOOKS (FOR READ AND WRITE):
    (ntupw (r) - opening code; ntinw(r) - ntuple or histogram ID, MEMw(r) - memory size, IFOw(r) - hbook ID)
    1,10,100000,11
    0,20,1000000,22
29. ID of Ntuple from which input data is downloaded in the NOMAD mode
    1, 2
30. NEURAL NET CONFIGURATIONS OF LAYERS, N OF NODES IN EACH LEYER. (LEYERS number, NODES array):
    3,2,5,1
31. N OF ITERATIONS, STEP VALUE, SIGMA CRITERIUM,
    INITIAL SPREED, RANDOM GENERATOR SHIFT (NITER, cf, sim, spread, Ishift

numbers):
3000,100,0.02,88

32. SPEED (the power index for weighting the difference between actual and desired NN ootput).WSPEED (the power index for selecting event's weight variants):
    sss(jj)=sss(jj)+abs((o(jj)-b(its+ii)))**speed*u(ITS)**wspeed
    l.,l. 4 пт
33. QUALITY FUNCTION SYMMETRIZATION WEIGHTS (WIGHT array):
    0.5,0.5
34. SEARCH MODE (search = single - one dimensional search; MULLTI - all net parm. modificated simultaneously; neuron - all couplings and threshold of a randomly selected neuron)
    neuron
35. QUALITY FUNCTION TYPE (qualit = montec - training with M.C.; sigmaa - with ON/OFF pairs; estima - neural estimation:
    montec
36. QUALITY FUNCTION MODE, qtype = msd, (massa, kolm modes - now suspended)
    msd
37. MEMORY TYPE (memory = simple - no memorization of better point during search; memory - the best changes are accumulated)
    memory
38. BEGIN RANDOM SEARCH FROM (begin = random point; or - better point, found in previous search cycles):
    random
39. STOP ITERATIONS IF QUALITY FUNCTION IS LESS THAN (stiter number):
    0.00001
40. DECISION POINT (for 2 class case if qualit ne.montec):
    else if (qualit.eq.'sigmaa'.or.qualit.eq.'spectr'.or.qualit.eq.'pure') then
    if(o(1).le.dpoint) then…
    0.51
41. number of different partitioning of last neuron output (0-1) interval (analog of a priory probabilities for the neural decision making) and list of partitioning and goal functions for all classes.
    nipr number and part and goal arrays):
    3
    0.5, 1
    0.1,0.9
    0.1,1.
42. MULTIDIMENSIONAL "BINS" NUMBER FOR FEATURE SPACE SCANNING (revealing of multidimensional nonlinear cluster shape), ndel number:
    1000
43. ESTIMATION MODE: NUMBER OF REGRESSANDS (ONE OR TWO) AND IT'S NAMES (ny number and numreg array):
    1
    MUON

Operation modes - The main and most important key word in ANI. Numerous operating modes are implemented in ANI program, to perform a different statistical procedures:

- *DIMDIM, DIMFLAT* - intrinsic dimensionality analyses.
- *BHATA* - Bhattacharyya distance calculation.
- ***ONE-DIMEN*** - tests for comparing single variables.
- ***COVCOR*** - covariances analyses.
- *FWRITE* - data subsamples archiving.
- ***ONE-LEAVE-OUT-*** - Bayesian learning and Bayes error estimation.
- ***CLASSIFICATION*** - Bayesian classification of experimental data.
- ***BOOTSRAP*** - Bootstrapization of Bayesian learning and classification, fraction reconstruction.
- ***REGRO, REGRO-AD*** - Nonparametric regression for energy estimation (KNN and Parzen types).
- ***LEARNING*** - Neural network training for both, classification and estimation.
- ***CLASSIFI*** - Neural classification and estimation of control events.
- ***EXP*** - Neural classification and estimation of experimental data.
- *FAST, SUPERCUT, SOBOL-CUT, MULTI-CUT* - for on-line analyses of atmospheric Cherenkov telescope data.
- *SAMEPT* - abandoned.

CHAPTER 3
STATISTICAL INFERENCE IN COSMIC RAY PHYSICS

NONPARAMETRIC INFERENCE

The scientific method is characterized by data classification, the study of their interrelations and relations to past experience, accumulated in various theories and hypotheses. Usually, it is impossible either to prove or to refute hypotheses by deductive method. The challenge is to draw sensible conclusions from noisy, discrepant information.

The main aspect of statistics is collection and interpretation of data, the interpretative aspect being the one that is now regarded as the essence of the subject [66]. The fundamental idea of statistics is that useful information can be obtained from individual small bits of data. An inductive method leads to empirical statements, that may be connected with theoretical ones by means of rational inductive conclusion rules [67].

The most natural and most general framework in which to formulate solutions to the physical inference in cosmic ray physics is a statistical one, which recognized the probability nature both of the physical processes of propagation of cosmic radiation through the atmosphere and the detectors, and - of the form in which data analysis results should be expressed.

However, it is very important to provide the scientist with objective criterion by which to judge the claims of hypotheses (models) under investigation *(problem solving strategy).* By model we mean a complete probability statement of what currently supposed to be known a priori about the mode of generation of data and of uncertainty about the parameters [68].

If this statement consists in the existence of an analytic distribution family, (like Poisson or Gaussian), appropriate to the problem in hand, we have prescribed parametric model. For such parametric models a well-known concept of statistical inference consists in obtaining estimates of its parameters and verifying the validity of a chosen family [69].

3.2 PARAMETRIC CLASSIFICATION

The classification problem in parametric case (Newman-Pearson test) is traditionally described in terms of null and alternative hypothesis, critical and acceptance regions and level of significance [70]. The "best" critical region (the region of rejection of null hypothesis) is constructed by means of a Likelihood Ratio(LR):

$$LR(v) = \frac{p(v/\psi_1)}{p(v/\psi_2)}, \qquad (3.1)$$

each of two classes is defined by values of - the parameter of a prechosen analytic probability density function, $v$is a multivariate observation vector (point in multidimensional feature space) $p(v/\psi_1), (v/\psi_2)$- are conditional probability density functions describing distinct, mutually exclusive (non-overlapping) and full $p(v/\psi_1) + p(v/\psi_2) = 1$statements (null and alternate hypothesis).

The threshold value reflects the costs of consequences of statistical decision. Usually one select this value to keep on some constant minimal level error for one class, while maintaining to minimize the error of the other class.

For the Kclass case the $p(\psi)$ -will be chosen as a "true" class

$$\psi = argmax_{\psi_i} p\left(\frac{v}{\psi_i}\right), i = 1, \ldots K. \qquad (3.2)$$

Ifψtakes infinite number of values from some metric space Ψthen we deal with an estimation problem and the Maximal Likelihood Estimate (MLE) is assymptotically unbiased and effective

$$\psi_{mle} = argmax_{\psi} \sum_{i=1}^{M} lnf(v_i/\psi), \psi \subset \Psi. \qquad (3.3)$$

where $\{v_i\}, i = 1, M$are the experimental events. The parametric estimation uses whole experimental sample set, instead of only one event in the classification problem, with the benefit of solving regression problem (parameter estimation) for all possible experimental situations. The analytical function $f(v/\psi_{mle}) \equiv f(v)$ can be used for energy estimation, of course if the shape of particular functional family $f(\bullet)$ is known.

Although the results of analysis using parametric statistics usually are easy to present and understandable, it is very important to remember that any inferential conclusion based on parametric technique are not exactly valid unless every assumption is satisfied.

If these assumptions cannot be substantiated, or are discarded, or are not even known to the investigator, then the inference may be less reliable than a judicious opinion, or even arbitrary guess [4].

The parametric methods superimpose very restrictive assumptions on the nature of the population from which the sample is drawn. For example, the assumption of a normal distribution implies a continuous, symmetric, bell shaped distribution with infinite domain and a specific mathematical function. And statistical inference is exact for these sampling distributions only and may not even be close to the obtained one, if the population assumption comes to be incorrect.

3.3 NONPARAMETRIC CLASSIFICATION - MONTE CARLO STATISTICAL INFERENCE

Usually, for experimental physics data analysis, the Likelihood Function cannot be written explicitly, and we deal with implicit, nonparametric models, for which no parametric form of underlying distribution is known, or can be assumed.

Nonparametric methods use much less stringent assumptions about population than those made in parametric statistics. Usually the underlying population distribution is assumed to be continuous only. Of course, this assumption is rather mild comparing with the very specific assumptions made in parametric case.

Let us consider the stochastic mechanism $(\mathcal{A}, \mathcal{V})$ which generates the observations v in a multivariate feature space - $\mathcal{V}$, v is a *d*-dimensional vector of EAS parameters measured experimentally. We assume that observations are random and can be described by some conditional probability density function depending on the primary particle type. The feature space V covers possible acceptable values of EAS parameters including cuts on age and Ne parameter, etc...

The basic states space $\mathcal{A}$ consists of alternative models or classes (the alternative strong interaction models, or - different primary nuclei). The appropriate statistical model to describe this situation is the probability mixture model:

$$p(v) = \sum_{k=1}^{L} P_k p\left(\frac{v}{\mathcal{A}_k}\right). \qquad (3.4)$$

And the main problem in EAS physics is to determine the proportions (frequencies) of events in each category $\mathcal{A}_k$.

We don't know the full statistical description (conditional probability density functions $p(v/\mathcal{A}_k)$ of how nature produces EAS from incident particles, nor the possibility to use particle beams outside the atmosphere to calibrate the installations, that is why, to determine the mutual probability measure on the direct product of $\mathcal{A}$ and $\mathcal{V}$ spaces the total Monte - Carlo simulation of the EAS development in the atmosphere and in detectors is performed, including experimental data registration and handling for alternative primary particles and possible strong interaction models in a wide energy range.

The set of *d*-dimensional $\mathcal{U}$ vectors obtained in simulations is an analog of the experimentally measured values of $\mathcal{V}$. But, as opposed to experimental data, it is known to which of the alternative classes each of these events belongs. These "labeled" events include a priori information about dynamics of the EAS development and registration, which is given in a nonparametric form, in form of simulation trials.

The sequence $\{\mathcal{U}_i, t_j\}$, where $i = 1, M_j, j = 1, L$, t-is the class index, is generated by a detailed Monte Carlo simulation program like CORSIKA and consists of L classes each containing M_i, simulation trials. This "controlled" stochastic mechanism we denote by $(\mathcal{A}, \mathcal{P})$and name training sample (TS). The training sample is the basis of all statistical procedures in applied Bayesian and neural approaches. The corresponding distribution mixture model takes the form:

$$\hat{p}(v) = \sum_{k=1}^{L} \hat{P}_k \hat{p}(v/\mathcal{A}_k) \qquad (3.5)$$

Of course, this substitution of unknown conditional densities $p(\mathcal{V}/\mathcal{A}_k)$ by their " simulation" analog $\hat{p}(\mathcal{V}/\mathcal{A}_k)$ is only valid if used model is adequate. And validation of the model remains the most crucial and yet unsolved problem in EAS data analysis.

Of course, for reliable estimation of conditional densities we'll need significant amount of training trials to cover all intrinsic variations of measurable EAS parameters and completely represent all categories (primary nuclei).

Since both physical processes of particle production and those of registration are stochastic, only by careful measurement of probabilities we can gain an understanding of the EAS phenomena. We can't expect simple solutions, as multidimensional distributions of EAS parameters overlap significantly and any decision on primary particle type and its energy will contain uncertainty.

The only thing we can require when classifying a distribution mixture is to minimize the losses due to incorrect classification to some degree and to ensure use of a priori information completely. Such a procedure is the *Bayes decision rule with nonparametric estimation of the multivariate probability density function.*

3.3.1 BAYESIAN PARADIGMA

The Bayesian approach of the statistical inference is a modification of the opinions of consistent experts (a-priori knowledge) in the light of new evidence and the Bayes theorem specifies how such modification should be made.

Moreover, as we believe, Bayesian a posteriori measures are only trustworthy and sensible measures of how the uncertainty about the phenomenon under investigation should be modified after new experimental data are achieved [3].

The Bayesian approach formalizes the account of all the losses connected with probable misclassification and utilizes all the differences of alternative classes [71]. The decision problem in a Bayesian approach is simply described in terms of the following probability measures defined on metric spaces:

- The space of possible states of nature – $\mathcal{A} \equiv (p, \alpha, O, N, Fe)$-groups of primary nuclei;
- The space of possible statistical decisions-$\tilde{\mathcal{A}} \equiv (\tilde{p}, \tilde{\alpha}, \tilde{O}, \tilde{N}, \tilde{F})$ where $\tilde{p}, \ldots \tilde{F}$ are the decisions that the examined event is caused by a primary proton, or…, iron nuclei;
- Cost (loss) measure $c_{A\tilde{A}}$, or $c_{A_i A_j}$ or in simple notion c_{ij}. This measure is defined on the direct product of nature states and decision spaces($\mathcal{A} \otimes \tilde{\mathcal{A}}$). Losses, connected with definite statistical decision $\tilde{\mathcal{A}}$ are equal to

$$c_i = \sum_{j=1}^{L} c_{ij,}\, i, j = 1, L \tag{3.6}$$

At correct classification of primary particles into "proton" and "iron" classes the losses are equal to zero

$$c_{Fe\tilde{F}e} = c_{p\tilde{p}} = 0 \tag{3.7}$$

or for problem of background rejection in TeV gamma-ray astronomy

$$c_{\gamma\tilde{\gamma}} = c_{h\tilde{h}} = 0 \tag{3.8}$$

If we misclassify the signal event, we decrease the efficiency of the γ-event registration. If we attribute hadronic images to a γ-ray one, we increase the background contamination. As we expect a significant excess of background against signal, we are interested in a strong background suppression. Thus, it is reasonable to admit the non-symmetric loss function for this case

$$c_{\gamma\tilde{h}} = 0.9, c_{h\tilde{\gamma}} = 0.1. \tag{3.9}$$

For elemental composition studies one can take uniform a priori losses function

$$C_p = C_\alpha = C_O = C_N = C_{Fe} = 0.2 \tag{3.10}$$

- Event (measurement, feature) space $\mathcal{V}$- a set of measurable characteristics of EAS, Cherenkov image parameters etc.
- The prior measure $P_A \equiv \left(P_p, P_{Fe} \ldots\right)$.
- Conditional densities (Likelihood functions):

$$\{\hat{p}(v/p), \{\hat{p}(v/\alpha), \{\hat{p}(v/O), \dots\} \qquad (3.11)$$

These density functions are estimated by means of training samples obtained in simulation trials with different primaries.

Multivariate probability density estimation is a fundamental problem in data analysis, pattern recognition, and machine learning. The estimation of the conditional (on particle type) density on the basis of a collection of simulations is also a key problem in cosmic ray and high energy physics.

3.3.2 BAYESIAN DECISION RULES

The Nonparametric Bayesian decision rule takes the form

$$\tilde{\mathcal{A}} = \eta(v, \mathcal{A}, \tilde{\mathcal{P}}) = argmax_i\{C_i\hat{p}(\mathcal{A}_i/v)\}, i = 1, \dots, L \qquad (3.12)$$

Where c_i is the losses connected with $\tilde{\mathcal{A}}$ decision, $\hat{p}(\mathcal{A}_i/v)$ is the nonparametric estimates of the a posteriori density, connected with conditional ones by Bays theorem:

$$\tilde{p}(\mathcal{A}_i/v) = \frac{\hat{P}_i\hat{p}(v/\mathcal{A}_i)}{\hat{p}(v)} \qquad (3.13)$$

And finally, substituting a posteriori density by the conditional ones we get the Bayesian decision rule in the form

$$\tilde{\mathcal{A}} = argmax_i\{C_iP_i\hat{p}(v/\mathcal{A}_i)\}, i = 1, \dots, L \qquad (3.14)$$

As one can easily see from above equation, the Bayesian statistical decision is dependent on multiplicator C_iP_i; therefore, we cannot separate the influence of losses (cost) measure and prior measure on the decision rule. Changes in losses can be compensated in changes in the prior to keep constant the Bayesian decision. We think, that it is reasonable to treat C_jP_j as single entity and denote it as a priori losses.

The robust Bayesian inference claims that after considering repeated evidence, the initial used prior distribution can't influence a posteriori distribution heavily [3]. Thus, the choice of prior distribution isn't of critical importance for fraction estimation, because of the very big volumes of experimental data reshaping the initial prior knowledge.

For the investigation of the influence of the a priori losses on the classification results, the statistical decision is made simultaneously for different alternative variants of a priori losses. Examining the, so-called, "influence curves" obtained with different losses, one can select the preferable regime of the estimator operation. For example, it is possible to select the desired ratio of background rejection and signal detection efficiency.

In ANI package provision is made to avoid statistical decision if all classes are very far from experimental events (outliers' problem). If:

$$\hat{p}(v/\mathcal{A}_i) < ST \text{ for all } i = 1, \dots K \qquad (3.15)$$

then the outliers report is sent to output stream. *ST* is, so called, Strangeness criteria, usually set to very small number. Conditional densities are estimated by the $TS(A, \tilde{\mathcal{P}})$using one of many nonparametric methods available, L is the number of classes.

The Nonparametric Likelihood Ratio for classes A_1, A_2 and experimental event **v** can be represented as

$$LR(v) = \frac{\hat{p}(v/\mathcal{A}_1)}{\hat{p}(v/\mathcal{A}_2)} \qquad (3.16)$$

Usually for comparison purposes we will use the sampling mean of Log Likelihood ratio. The nonparametric Log-likelihood function for the *k-th* class has the form:

$$\mathcal{L}_k = \sum_{i=1}^{M} \ln \hat{p}(v_i/\mathcal{A}_k), \qquad k = 1, L \tag{3.17}$$

where *M* is number of experimental events. The negative of Log Likelihood function is calculated in ANI, therefore the smaller values will correspond to most probable model.

3.3.3 NONPARAMETRIC PROBABILITY DENSITY ESTIMATORS

To estimate conditional densities, we used Parzen and KNN methods [75, 77, 78, 79, 80, 81, 82, 83, 84] with automatic "best" method parameter adaptation (kernel width - for Parzen, and number of nearest neighbors - for KNN) [85].

Several probability density estimates are calculated simultaneously. Then the obtain sequence is ordered and the median of this sequence is chosen as final estimate (so called L-estimate). Depending on the intrinsic probability density in the vicinity of point v where the density is estimated, due to stabilizing properties of the median, each time the best estimate will be chosen [74]. The Parzen kernel probability density is estimated by:

$$\hat{p}\left(\frac{v}{\mathcal{A}_i}\right) = \frac{|\Sigma_i|}{2\pi^{\frac{d}{2}} h^d} \sum_{j=1}^{M_i} e^{-\frac{r_j^2}{2h^2}} \omega_j, \quad i = 1, \dots, L, \qquad \sum_{j=1}^{M_i} \omega_j = 1 \tag{3.18}$$

where *d* is the feature space dimensionality, M_i is the number of events in the *i*-th TS, ω_j are the event weights, *h* is the kernel width (parameter controlling the degree of the "smoothness" of an estimate), r_j is the distance from experimental event v to the *j*-th event of the TS in the Mahalanobis metric

$$r_j^2 = (v - u_j)^T \sum\nolimits_i^{-1} (v - u_j) \tag{3.19}$$

where Σ_i is the sampling covariance matrix of the class to which u_j, belongs.

The *KNN* estimate (if the equal weights are assumed) takes the form:

$$\hat{p}(v/A_i) = \frac{k-1}{M_i V_k(v)} \tag{3.20}$$

where $V_k(v)$ is the volume of a d-dimensional hypersphere containing K nearest neighbors to the experimental event v:

$$V_k(v) = V_d |\Sigma_i|^{1/2} r_k^d, V_d = \frac{\pi^{d/2}}{\Gamma(\mathrm{d}/2+1)}, \tag{3.21}$$

where r_k is the distance to the *k*-th nearest neighbor of v, Γ(.) is the gamma function. $|\Sigma_i|$ is the determinant of the covariance matrix of the class to which the K-th nearest neighbor belongs.

3.3.4 NONPARAMETRIC REGRESSION

As well as for density estimation, described in previous section, we use the KNN and Parzen window for nonparametric regression. The choice of nonparametric methods for energy estimation is obvious: the a priori information about the shape of energy spectra in the "knee" region (ascribed both from measurements and existent models of particle generation and acceleration in interstellar media) predicts rather complicated character of spectra and change to different modes in various energy domains. Therefore, we couldn't expect that any parametric family with not very large number of parameters will describe the data satisfactory.

So, also in the case of energy estimation, as for classification, the nonparametric methods only allow an event-by-event analysis of EAS data.

The method is based on the fact that the events close to each other in feature space v should have close to each other energies (geometric consistency hypothesis). The Parzen regression energy estimate takes form

$$\hat{E}(v_j) = \sum_{i}^{M_{TS}} C_i E_i \qquad (3.22)$$

where

$$C_i = \frac{|\Sigma_i|}{2\pi^{d/2} h^d} e^{-r_{ij}^2/h^2} \omega_i. \qquad (3.23)$$

Here, r_{ij} is the distance from the observable v_j to the u_i point of the TS, ω_i is the training event weight. The Parzen estimate is calculated for different prechosen values of kernel widths h. The median of the estimates sequence is used as final estimate.

The KNN regression (for equal weights) energy estimate takes slightly different form

$$\hat{E}(v_j) = \sum_{i}^{K} C_i E_{[i]} \qquad (3.24)$$

$$\sum_{i}^{M_{TS}} C_i = 1 \qquad (3.25)$$

where $E_{[i]}$ stands for the sequence of energies of K nearest to v_j neighbors of TS. C_j coefficientsare inverse proportional to distance (or square distance) between v_j and u_i.

The KNN estimate is also calculated for different prechosen values of K. The median of the estimates sequence is used as final estimate.

3.3.5 BAYES ERROR ESTIMATION

The classification methods, like all the statistical ones, include a quality check as an absolutely necessary procedure. The proposed procedure can be used as well for the determination of the frequencies of the probability mixture (3.5).

The most natural measure for quality check is the error probability (classification error) which

depends on both the degree of overlapping of alternative multivariate distributions and – on the decision rule being used (Bayes decision rules provides minimal error as compared to any other decision rule using the same features):

$$R^B = E\{\theta[\eta(v, \mathcal{A}, \mathcal{P})]\} \qquad (3.26)$$

where

$$\theta[\eta(v, A, P)] = \begin{cases} 0, \text{for correct classification} \\ 1, \text{otherwise} \end{cases} \qquad (3.27)$$

The mathematical expectation is taken over the whole $\boldsymbol{d}$ dimensional feature space $\boldsymbol{v}$. In other words, the Bayes error is a measure of the overlapping of alternative distributions in the feature space, e.g. the expected proportion of the "incorrect" classification. Since we do not know the analytical form of the distribution of measurements, we obtain an estimate of $\hat{R}^B$ via the TS:

$$\hat{R}^e = \frac{1}{M_{TS}} \sum_{i=1}^{M_{TS}} \theta\{\eta(u_i, \mathcal{A}, \tilde{P})\} \qquad (3.28)$$

i.e. we classify the $\{u_i\}$, i = 1, $M_{TS,}$ then check the correctness of the classification over the index of the class t_j, $j = 1, L$. The expectation is taken over all possible samples of volume M_{TS}.

However, as numerous investigations have shown (e.g. [72]), this estimate is systematically biased and hence, a one-leave-out-for-a-time estimate is preferable

$$\hat{R}^e = \frac{1}{M_{TS}} \sum_{i=1}^{M_{TS}} \theta\{\eta(u_i, \mathcal{A}, \tilde{P}_{(i)})\} \qquad (3.29)$$

where $(\mathcal{A}, \tilde{P}_{(i)})$is a TS with a removed *i*-th element, which is classified and then "returned" to the sample. This estimate is unbiased and has an essentially smaller m.s. deviation compared with other estimators [73]. The advantage of $\hat{R}^e$ is especially notable when the feature space has a high dimensionality. Note, that we have the possibility of repeated estimation of the error probability by classifying various TS classes - $\{u_i, t_j\}$, j = 1,L.

By R_{ij}^e (or simply R_{ij}) we denote the probability of classifying the *i*-th class events as belonging to the *j*-th class (misclassification). By R_{ii} the "true" classification probability is denoted. For EAS classification according to 5 primary groups, each element of the “classification matrix” can be determined by the Bayes risk estimator (3.29):

$$\begin{pmatrix} R_{p\to p} & R_{p\to\alpha} & R_{p\to 0} & R_{p\to si} & R_{p\to fe} \\ R_{\alpha\to p} & R_{\alpha\to\alpha} & R_{\alpha\to 0} & R_{\alpha\to si} & R_{\alpha\to fe} \\ R_{0\to p} & R_{0\to\alpha} & R_{0\to 0} & R_{0\to si} & R_{0\to fe} \\ R_{si\to p} & R_{si\to\alpha} & R_{si\to 0} & R_{si\to si} & R_{si\to fe} \\ R_{fe\to p} & R_{fe\to\alpha} & R_{fe\to 0} & R_{fe\to si} & R_{fe\to fe} \end{pmatrix}$$

This matrix accumulates a priori knowledge on the possibility of data classification into 5 categories. If all diagonal elements are greater than 0.6 (and therefore - the sum of all non-diagonal elements in each line is less than 0.4), you can expect unambitious results of fraction estimation after reconstruction procedures explained in the next section.

The overall index reflecting the "goodness" G of features used for classification (index of separability) takes the form:

$$G = \left(\prod_{i=1}^{L} R_{ii} \right)^{1/L} \tag{3.30}$$

This averaged product of diagonal elements represents the "mean" probability of true classification into one of L categories. This index, of course, is directly connected with Bayes error.

3.3.6 FRACTION ESTIMATION

Now let us estimate a posteriori fraction of various classes in the distribution mixture.

The best estimate of a posterior fraction [8] (in case of a uniform a priori information and absence of systematic errors) is the empirical fraction

$$P_i^e = \frac{M_i}{M_{tot}} \tag{3.31}$$

where, M_i is the number of events classified by the Bayesian decision rule (3.12) as belonging to the class A_i, M_{tot} is the total number of events. With account of classification the matrix of the classification errors, the corrected fraction (proportion) can be obtained as the solution of the following set of linear equations:

$$\sum_{k=1}^{L} \hat{P}_k R_{ki} = P_i^e, \quad i = ,1, \ldots L. \tag{3.32}$$

where $\hat{P}_k$ is the estimate of the proportions P_k of the distribution mixture (3.4).

The accuracy of the estimates is defined by the TS size and number of control data as well as by the value of the Bayes risk, which represents the "quality" of discrimination with the chosen feature subset (index of separability). Note, that the set (3.32) is a poorly defined system and at large values of classification errors the solution is unpredictable and hence, the choice of a feature combination providing a high percentage ($\geq 60\%$) of correct classification is a necessary preliminary stage.

For classification in two categories (for example "heavy" - *Fe*, and "light" - *p* nuclei) the system of two equations can be easily solved explicitly:

$$P_{Fe} = \frac{P_{Fe}^e - R_{p,Fe}}{1 - R_{p,Fe} - R_{Fe,p}}, \quad P_p = 1 - P_{Fe} \tag{3.33}$$

3.3.7 THE BOOTSTRAP PROCEDURE

As we have shown in the previous section, to estimate the proportion of various nuclei in the primary flux, besides the classification of the experimental data by a TS, it is also necessary to calculate the misclassification rates R_{ij}. The final accuracy of the obtained nuclear composition is a function of both accuracies of classification and the determination of R_{ij}.

The possibility to decrease the bias and variance of the estimates of misclassification rates (3.32) was discussed in [28], where it was demonstrated that it is possible to improve the accuracy of the estimates, if the TS size is large, and, therefore we can obtain the estimate of (3.29), dividing the training sample to independent subsamples. Unfortunately, time consumption per model event generation increases abruptly with primary particle energy and the TS size is rather small for high energy events. Thus, the problem of an efficient use of the information contained in training samples is very important for cosmic ray physics, since samples corresponding to the highest energies are very limited. Of the greatest importance is also the estimation of the statistical errors of obtained fractions of different primary nuclei. We propose to use the advanced resampling methods for fraction error estimation. The resampling methods of statistical error estimation were widely used since the last two decades. An efficient procedure actively developed in both applied and theoretical aspects is the bootstrap method [87] which lies in replication of the initial sample many times by means of random sampling with replacement. Thus, obtained in such a way conditionally independent bootstrap-replicas stand for independent samples from the general population (under the condition of sufficiently large size of the initial sample) and can be used for statistical error estimation [88]. In fact, the bootstrap method substitutes the unknown general population by a single sample. The theoretical basis of the bootstrap method is the analog of the Central Limit Theorem (CLT) proved in [89]:

$$P\{\sqrt{B}(\mu_* - \mu_M) < tS_M | v_1, \ldots, v_M\} \to N(t), \qquad (3.34)$$

when $M, B \to \infty$ and $v_1, \ldots, v_M$ are independent, identically distributed (IID) random quantities, $N(t)$ is a standard Gaussian distribution and t is it's quantile, μ_M and S_M are estimates of the first and the second statistical moments,

$$\mu_* = \frac{\sum_{j=1}^{B} \mu_j^b}{B}, \qquad \mu_j^b = \frac{\sum_{i=1}^{M} v_i^{(j)}}{M} \qquad (3.35)$$

μ_j^b is the j - th bootstrap replica's mean, μ_* is the bootstrap first moment. Moreover, analogies between sampling and the bootstrap procedures are valid also for many other statistics. Referring to [90], we shortly summarize the main idea of the new method: a new procedure - the bootstrap-moments (denoted by μ_* and σ_*) are introduced, which in many cases substitute the statistical moments calculated according to a distribution function (in most cases of interest - unknown). Of course, the analytical calculation of the bootstrap moments is usually impossible. However, and here lies much of the strength of the bootstrap approach, these quantities may be computed, to any desired level of accuracy, by a Monte Carlo simulation [90].

Returning to the problem of distribution mixture coefficient estimation we consider two ways of the procedure bootstrapisation:

- obtain the bootstrap estimate of the misclassification coefficients R_{*ij} and empirical ratio $P_{*i}^e, i = 1, L$, then reconstruct the fraction according to (3.32).
- carry out procedure of fraction estimation using each bootstrap replica, then obtain B bootstrap estimates of the fraction $\hat{P}_{*i}^e, (i = 1, L, j = 1, B)$.

The second way is preferable, due to the possibility of explicit calculation of the systematic errors and, therefore - to evaluate the m.s.d. for the obtained fraction estimates. By several

bootstrap replicas we calculate the bootstrap expectation $\hat{P}_{*\mathrm{i}}$ and bootstrap standard deviations of the mixture coefficients $\hat{P}$, which will be used as final estimates of the fraction of different nuclei groups in the primary flux.

3.4 THE NEURAL CLASSIFICATION TECHNIQUE

The basic computing element in a multi layered Feed-Forward Neural Network (FFNN) is a node (formal neuron). A general i-th node receives signals from the outputs of the all neurons of the previous layer:

$$IN_i^{l+1} = T_i + \sum_{j=1}^{NODES(l)} W_{ij}^l \times OUT_j^l, i = 1, NODES(l+1), l = 1, L-1. \tag{3.36}$$

where the threshold T_i, and connection strengths (weights) W_{ij}^l are parameters associated with the node *i,l* is the layer index, *L*is the total number of layers, NODES (/) is the number of neurons in the *l*-th layer and OUT_j^l - is the output of the *j*-th neuron in *l*-th layer. The index *i* corresponds to the next to j layer.

The output of the neuron is assumed to be a simple function of its input, usually it is formed by the, so called, nonlinear sigmoid function:

$$OUT_i^l = \frac{1}{(1+e^{-IN_i^l})}, i = 1, NODES(l), l = 2, L. \tag{3.37}$$

where IN_i^l- is the input of the i-th neuron in the l-th layer.

Thus, feature, entering the first layer are translated from input through hidden layers to the output nodes. Therefore, FFNN provides the mapping of N-dimensional feature space to the space of the lower dimensionality, in ideal case into class assignments. For classification purposes this mapping takes a special form with aim to "shift" different classes of TS from each other as much as possible. Therefore the "goal" output $OUT^{goal}(k)$ for events of the k-th category could be chosen as follows:

$$OUT^{goal}(k) = \frac{k-1}{K-1}, k = 1, K. \tag{3.38}$$

where *K* is total number of classes. Of course, it is possible to define another set of "goal" value.

In the case of two classes, i.e. signal and background events, the "goal" outputs, as one can easily see, are equal to zero and one. The actual events classification is performed by comparing the obtained output value with the "goal" one. We expect, that the data flow passing through the trained net will be divided in two clusters concentrated in the opposite regions of the (0, 1) interval. Choosing an appropriate point in this interval (the so-called decision point c), the classification procedure can be defined: an event with an output greater or equal than the decision point is attributed to the background class, while all the other events - to the signal class:

$$OUT(v) = \begin{cases} < c, & v \text{ is classified as signal} \\ \geq c, & v \text{ is classified as background} \end{cases} \tag{3.39}$$

where $OUT(v)$ is the output node response to a particular experimental measurement v. This

decision rule is a Bayesian decision rule; therefore, the output signal of a properly trained feed forward neural network is an estimate of the a posteriori probability density [91].

For the multi-way classification one can define a set of not overlapping intervals in (0 — 1). This set, along with the chosen "goal" values, will determine the mapping of net output into class labels. The figure of merit to be minimized is simply the discrepancy of apparent and target outputs over all training samples (so called classification score):

$$Q = \sum_{k=1}^{K} \sum_{j=1}^{M_k} (OUT(k) - OUT^{goal}(k))^2 \tag{3.40}$$

where $OUT_j(k)$ is the actual output value for the *j*-th training event, belonging to the k- th class, and the $OUT^{goal}(k)$ is the target value for the *k*-th class output, where *K* is number of categories and M_k is number of events in the k-th training set.

In many cases of interest, it is preferable (and possible) not to use the simulations at all. If the calibration of experiment with background cosmic radiation, as in case of atmospheric Cherenkov telescopes is available, a new model-independent quality function can be determined. Searches for discrete very high-energy gamma-ray sources consisted in the detection of an abundance ($N_{on} - N_{off}$) of events coming from the direction of a possible source(N_{on}) as compared with the control measurement, when pure background is registered(N_{off}). As the expected fluxes are rather weak and background is abundant (the signal to background ratio not exceeding 0.01 and usually is much smaller), one should always answer the following question: is the detected abundance a real signal or only a background fluctuation? The measure of statistical significance used in gamma-ray astronomy is some empirical analog of the P-value of the Student statistical test [92]:

$$\sigma = \frac{N_{on} - N_{off}}{\sqrt{N_{on} + N_{off}}} \tag{3.41}$$

The larger σ, the smaller the probability that the detected excess is due to background fluctuation. The telescope design and the development of new data handling methods have the purpose to enlarge the value of σ. After introducing some “cut” in the image parameter space and selection of the "gamma-like" events from raw data (both from the ON and OFF samples), the σ criteria value can significantly enlarge:

$$\sigma_{cut} = \frac{N^*_{on} - N^*_{off}}{\sqrt{N^*_{on} + N^*_{off}}} \tag{3.42}$$

where N^*_{on}, N^*_{off} are the numbers of events "surviving the cut. After executing all *ONN, OFF* samples, the σ value is calculated each time by means of "survived" (classified) events.

3.4.1 NEURAL ESTIMATION

The same FFNN with different quality function is used for primary energy estimation. The following ”quality” function have to be minimized

$$Q = \sum_{j=1}^{M} \omega_j \left(OUT(u_j) - OUT^{true}(u_j) \right)^2, \tag{3.43}$$

where, $OUT(u_j)$ is the output of the FFNN last layer and the $OUT^{true}(u_j)$ is the parameters used in simulation (energy of cosmic ray nuclei or very high-energy gamma ray), $\omega_j$is the event weight (usually the highest energy events get higher weight).

3.4.2 FF NEURAL NETWORK TRAINING

Two main scenarios of net training are implemented in ANI.

- The deterministic mode: the multidimensional quasi-random sieves are used for scanning of the net parameter space. Positioning the sieve center at the previously found best point, and subsequently decreasing sieve size, we'll arrive to the best NN parameters.
- The random search algorithm uses pseudo-random numbers to select the particular net node and randomly change all its weights. If new weights bring an improvement of quality function, then this change survives, and a new random search step is performed, in opposite case the changes are subtracted and another random step is made from the previous point.

The total number of searching NN parameters equals:

$$NTOT = \sum_{l=2}^{L} NODES(1) + \sum_{l=1}^{L-1} NODES(l)NODES(l+1)\,. \tag{3.45}$$

The random change (addition, or subtraction) Δ_i, is selected on the i—th iteration of search procedure in the following way

$$\Delta_i = STEP f(Q_{i-1})(RNDM - 0.5) \tag{3.46}$$

where RNDM - is randomly distributed in the (0 - 1) interval, $f(Q_{i-1})$ is the power function, controlling the random step size during reaching the global minimum and STEP is normalizing factor.

Also, complementary search modes can be used:

- Single mode (one of NTOT) - single parameter is randomly chosen and randomly changed;
- Multi-mode - all net parameters simultaneously are randomly changed.

It is possible to combine different search modes. Each will start from the best point reached in a previous search. Changing modes and search parameters helps to escape the local minima region and finally obtain desired solution.

3.4.3 GENETIC ALGORITHMS

As was described in the previous section, the NN training (search of the best parameters) is a very time-consuming procedure with no guarantee to find the unique best point in multidimensional (dimensionality may be 100 and more) NN parameters space.

To fasten this process, the genetic approach is proposed. Starting from a "pool" of "good" solutions, obtained with different searching scenarios starting from different initial points we determine the "genetic" operations on it.

Each solution is represented by a "chromosome" - collection of gene (each gene consists of a single neuron parameters). The main genetic operations are "crossing-over" - exchange of randomly chosen neurons in randomly chosen chromosome, and mutation - rare process of the random changing the single " allele".

Then the evolution started (of course the rules of forming of the next generations have to be defined) and "fitness survival" mechanism is triggered.

We'll demonstrate in the ANI testing chapter, that some of "offsprings" demonstrate better characteristics than "parents".

The genetic module is a separate program written in Cfor UNIX by S.Chilingarian.

3.4.4 NET TOPOLOGY

As for many other nonparametric techniques, for FFNN training it is very difficult to find an appropriate method for net parameter determination (e.g. the number of nodes and layers). Of course, one can form the initial "pool" from networks with different topologies, and leave evolving population to find the best one. But this approach seems to be too complicate, due to numerous variants and possibilities.

Instead, the "Occam" principle is used. Start from minimal configuration (single hidden layer with 3-5 neurons); then increase NN step-by-step; check for improvement and stop when no more improvement take place.

It is worth to mention, that one cannot use very complicated nets, if training samples are limited in size. The empirical rule requires as minimum 10 training events for each net parameter. Therefore, you can't use more than the simplest net (4::3::1) if your training sample consists of 100 events and you try to make 3-way classification of 4-dimension EAS parameters.

3.4.5 STOPPING RULES

Usually, the net training iterations canceled when the value of quality function is stabilized, and no more improvement takes place. Or, when the requested maximal number of iterations has been reached. Then the obtained set of net parameters can be used for experimental data classification.

But, there always is the danger of "overtraining" (especially for small training sets). The obtained network will very good describe the particular training set, but not the required overall dependence. So, the "training" error could be minimized, but the "generalization" error, obtained with the same net, but with an independent sample can give a rise of classification errors. Therefore, in each step of training it is important to check the results with an independent sample and stop training when the "generalization" error starts to increase.

3.5 KNN ALGORITHM OF FRACTAL DIMENSION ESTIMATION

The basic approach to dimensionality analysis lies in characterizing physical systems by the singularities of the invariant probability measure [93]. To do this, let us determine the scaling of moments of the random quantity $p_i(l)$ of order q at scale l:

$$C_q(l) \equiv < p_i(l)^q > \equiv \sum_{i=1}^{N(l)} p_i(l)^{q+1} \sim l^{\phi(q)}, \phi(q) = q d_{q-1} \qquad (3.47)$$

where d_q are the Renyi dimensions (generalized dimensions) determined for $-\infty < q < \infty$.

If q=*-1*, the relation (3.47) determines the capacity dimension $d_F = d_0$, if q= *0* - the information dimensionality d_1, and if q=*1* the correlation dimension d_2. The estimates of the Renyi dimensions are defined as a slope connecting some values of $\{l_i\}$ with the corresponding values of $\{C_q(l_i)$ in a double-logarithmic scale. While the formal definition of the generalized dimension is given in the limits of very small scales and infinite population, in practical applications only a limited number of events is accessible, so only a finite scale can be considered. In fact, if we cover the distribution with boxes that are too small, most of them will contain just one particle or nothing at all. Therefore, for the direct Renyi dimension calculation there are no instructions regarding the choice of the box-size$\{l_i\}$. Algorithms based on nearest neighbor information (KNN-algorithms) are much more efficient than the box- counting algorithms and they introduce a natural scale - the sample-averaged distance to NN:

$$\bar{R}_k, k = 1,2,\dots M-1 \qquad (3.48)$$

where M is the total number of events in the sample. Using the ergodic theorem one can make a replacement [94, 95]:

$$\sum_{i=1}^{N(l)} p_i(l)^{q+1} \sim \sum_{j=1}^{M} \tilde{p}_j(l)^q \simeq Q_1 \qquad (3.49)$$

where $\tilde{p}_j$ is the probability to find the point of the studied set not in the box of size l, but inside the hypersphere of radius l, centered at some other point of the studied set and Q_1 is the total number of q-tuples within this sphere. For the $\bar{R}_k$ sequence, the scaling relation takes the form:

$$Q_{\bar{R}_k} \sim \bar{R}_k^{\phi(q)} \qquad (3.50)$$

For q = *1*(correlation dimension) the number of q-tuples is simply equal to the number of the sample events within /-spheres, and the left-hand side of (3.50) is equivalent to the mean number of the sample points inside a hyper-sphere with radius equal to the average distance to the *K*-th neighbor, i.e. is equal to the number *k*, so:

$$k \sim \bar{R}_k^{d_2} \qquad (3.51)$$

Hence, the modified algorithm defines d_2 as a slope of the *k*-dependence of R_k in a double-logarithmic scale. Note that the obtained dimensionality is not in any way connected with the regions where singularities of the probability measure arise, i.e. it is impossible to recover the spatial structure of the multi-fractal support by the d_q spectrum. That is why we believe that the local dimensionality may be useful in separating the space regions where considerable fluctuations of the invariant probability measure are observed. Apart from sample averaging, there is also one more way to get a linear equation for dimensionality determination [96]. We can determine the same procedure as described above not for the sample averaged distances, but for the actual distances to the nearest neighbors of each point in the sample. For this, one must choose the series $\{k_j\}$ such, that the density estimates are very close to each other and hence, the dependence of $\hat{p}_k(v)$ on k can be ignored. Using chosen in such way $\{\overline{R}_{k_j}(v_i)\}$ values and corresponding $\{k_j\}$ values, one can estimate the local dimensionality at a point v_i. The mode of the histogram of the local dimensionalities usually approaches the global one, and local inhomogeneities of the sample can also be readily seen from the histogram as several local pikes.

CHAPTER 4
ANI TESTING (GAUSSIAN DATA)

4.1 BAYESIAN ANALYSIS

4.1.1 PROBABILITY DENSITY ESTIMATES

To check the ANI density estimation modes, we use samples from multivariate Gaussian populations with different means (0 - for the first class and 1 - for the second) and equal variances (1 for both classes). We compare the theoretical value of the Bayes error, which for Gaussian distributions is directly related to the Mahalonobis distance (3.19) between first statistical moments of two classes:

$$R^B = \Phi\left(-\frac{r_{mah}}{2}\right) \qquad (4.1)$$

where Φ is the cumulative standard Gaussian distribution function. For univariate Gaussian population many theoretical results exist on the closeness of estimated and true probability density function ([76]). The two main measures used to describe this closeness are the L_1 and -L_2 (the integrated mean square error) metrics.

$$L_1 = \int E(|\hat{p}_m(v) - p(v)|)\, dv \qquad (4.2)$$

$$L_2 = \int E((\hat{p}_m(v) - p(v))^2)\, dv \qquad (4.3)$$

where $\hat{p}_m$ is the nonparametric density estimate obtain by implementing one of nonparametric density estimators (3.18) with a sample consisting of *m* events. From the figure (4.1) one can see the probability density curves corresponding to the different smoothing parameters of the Parzen estimator overlaid onto the standard Gaussian density ([86]). For the small kernel width h = 0.1 the estimated is discreet, for the width - *h = 0.7 — oversmoothed.* There is a number of notions for "optimal" kernel width ([76]), for samples from Gaussian populations, for example

$$h = 1.66M^{-1/5} \qquad (4.4)$$

This equation is valid only if Mahalonobis metric is used. As one can easily calculate, the optimal kernel width increases from 0.41 for sample size *M*= 1000, to - 0.67 for sample size *M*=100. For multivariate Gaussian distributions, of course, one has to have taken larger values.

The "optimal kernel" and adaptive estimates (L-estimate) show better approximation, compared with estimates with fixed parameters. Note, that the L-estimate didn't use very specific a priory information on distribution function shape, as an "optimal kernel" estimate, therefore, it is robust and can be used for samples taken from distributions, which analytic shape is unknown (common case in analysis of big data). The sequence of fixed kernels, used for constructed of L-estimate, should cover some wide interval depending, of course, on the available sample size.

To check the density estimator, we calculate the probability integral for several independent samples from standard Gaussian population:

$$\int \hat{p}(v)\, dv \qquad (4.5)$$

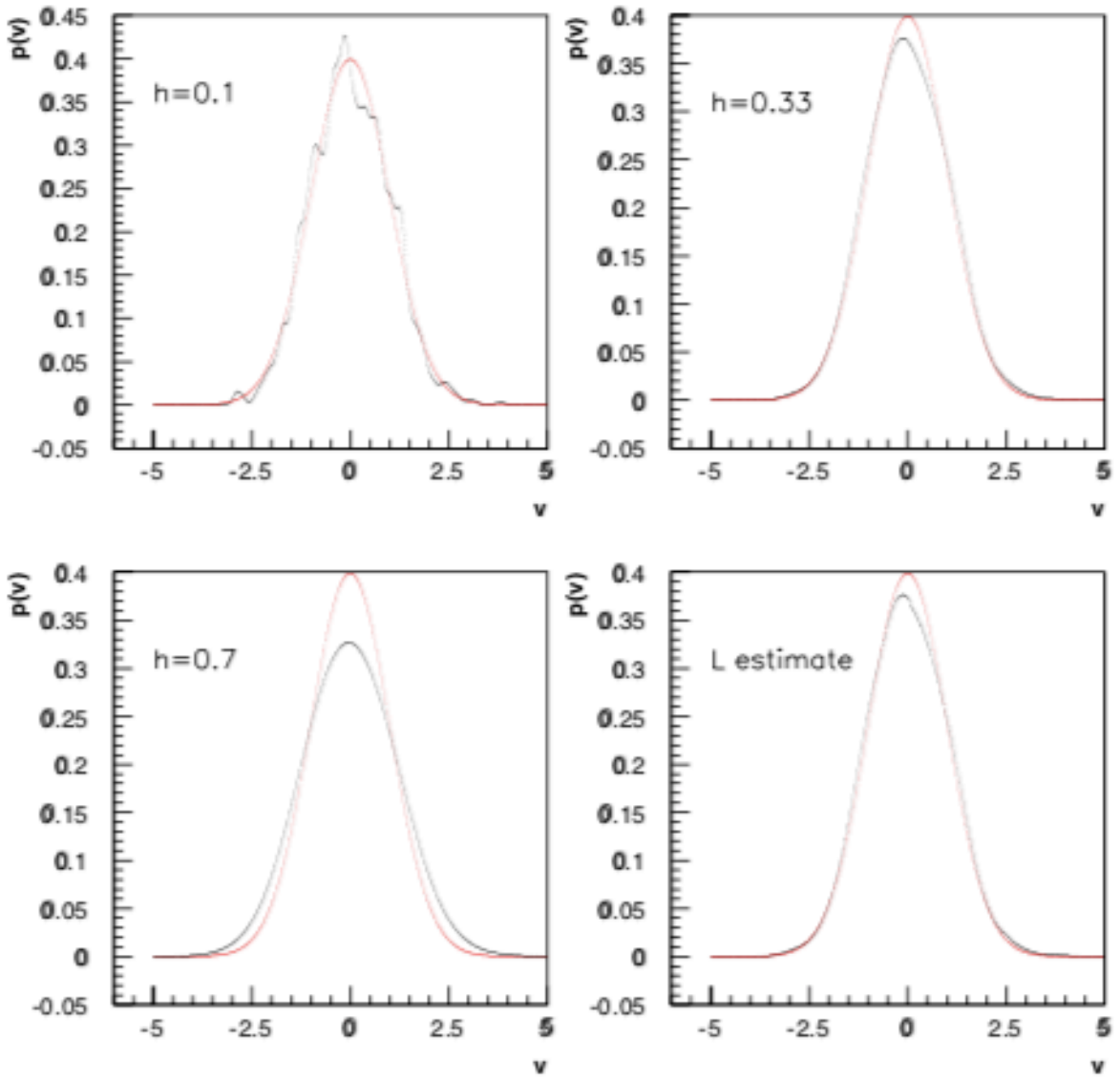

Figure 4.1: Parzen density estimates of standard normal distribution

As one can see from the figure 4.2, the distribution of probability integrals is rather smooth approaching 1 from the left (the bottom Darbu sum), and ensuring correct normalization of the estimated densities (3.18).

The results of the adaptive estimator are summaries in Table 4.1. The densities were estimated in 51 points uniformly distributed in (-5+5) interval. The L_1 and L_2 measures were calculated for samples from standard Gaussian population. The Bayes risk estimates were done for samples from standard Gaussian and Gaussian with mean 1 and variance 1, according to (3.29). The probability densities were calculated simultaneously for 7 kernel widths (from 0.2 to 0,8). Corresponding Bayes errors and L_1, L_2 measures as well as first and second statistical moments, were calculated for each from 1000 trials of independent samples of size 10, 25,100; 100 trials for - 400; and 10 trials for - 1000. Rather well agreement with other estimators reported in ([82, 15]), demonstrates the consistence and unbiasedness of density estimators used in ANI.

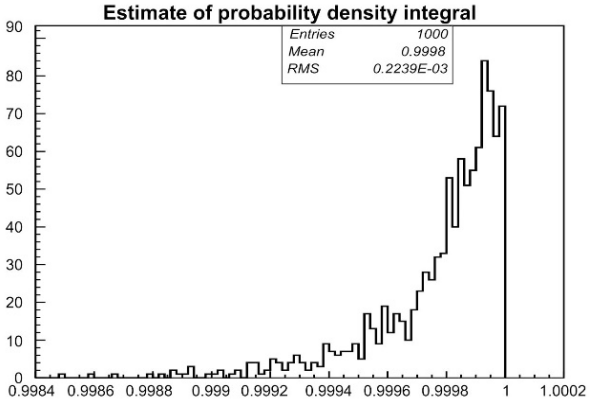

Figure 4.2: The histogram of "probability integrals"

Table 4.1: The quality check of the Parzen density estimator, samples from Gaussian populations N(0,1)-N(1,1), adaptive estimator

MxB	mean		variance		R^e	$L1$	$L2$
10 X 1000	0.002	0.996	1.437	1.463	0.332±0.210	0.345±0.150	0.0310±0.024
	-0.050	0.978	1.340	1.330	0.325±0.190	0.340±0.140	0.0310±0.024
	-0.004	0.998	1.400	1.430	0.332±0.200	0.340±0.140	0.0300±0.023
25 X 1000	-0.008	0.995	1.144	1.139	0.317±0.140	0.240±0.100	0.0154±0.013
	-0.033	0.978	1.080	1.080	0.320±0.140	0.250±0.100	0.0166±0.012
	-0.005	0.990	1.140	1.140	0.317±0.145	0.240±0.100	0.0155±0.012
100 X 1000	-0.002	0.994	1.031	1.033	0.317±0.074	0.141±0.050	0.0058±0.0041
	-0.003	0.966	0.980	0.980	0.314±0.080	0.153±0.048	0.0065±0.0042
	-0.006	0.948	1.030	1.030	0.312±0.071	0.141±0.050	0.0057±0.0040
400 X 100	-0.004	1.000	1.010	1.001	0.308±0.060	0.083±0.024	0.0021±0.0013
	-0.030	0.965	0.960	0.960	0.310±0.050	0.097±0.028	0.0026 ±0.0014
	-0.000	0.990	1.000	1.010	0.310±0.055	0.089±0.032	0.0023 ±0.0015
1000 X 10	-0.001	0.996	1.001	0.992	0.310±0.100	0.062±0.028	0.0012±0.0008
	-0.035	0.959	0.950	0.950	0.310±0.100	0.077±0.033	0.0016±0.0010
	-0.004	0.998	0.990	1.000	0.310±0.100	0.064±0.025	0.0011 ± 0.0005

4.1.2 THE BOOTSTRAP STATISTICAL MOMENTS

To check the validity of bootstrap approach we test the biasness of bootstrap moments using samples from the standard distribution *N(0,1);* the sample size varied from 25 to 1000, the number of bootstrap replicas in a series was from 10 to 1000. The mean was calculated for each bootstrap replica according to (3.35), and for each bootstrap series - the bootstrap estimate of the mean $\mu_*$and - mean standard deviation - $\delta_* = \sigma_*/M,$ was evaluated. The results of calculations, which present in table 4.2, illustrate the validity of "butstrap" CLT (3.34) and consistency of using of the bootstrap moments. Although the mathematical theorems were proved for the asymptotic cases $M, B \to \infty$ even with small sample sizes and rather small numbers of bootstrap replicas$(M, B = 50)$, the obtained bootstrap estimates coincide with sampling statistical moments with high precision. Of course, enlarging of sample size and bootstrap replicas number improves the accuracy significantly.

4.1.3 DISTRIBUTION MIXTURE SHARE ESTIMATION

The bootstrap fraction estimator (see section 3.3.7) was checked by a pilot Monte Carlo study of 2-way classification with samples from Gaussian population $N(0,1)$ and $N(1, 1)$. For comparing the "bootstrap errors" of "reconstructed" proportion with expected from usual sampling procedures we make aN "experimental" sample with mixture from both samples in 0.2-0.8 proportion. The empirical fraction estimate P_1^e and the error matrix R^e were calculated using the Bayesian non-para- metric procedures (3.31 and 3.29). Then the "true" fraction $\hat{P}_1$ was calculated according to (3.32 and 3.33).

The sample averaged values of all 3 estimates was obtained with independent random samples from the same Gaussian populations. They are denoted in the table by brackets <>. The bootstrap estimates are denoted by * symbol. The examples of particular run are presented in ANI outputs section (5.3). In table 4.3, 100 events were used for training and 1000 for classification, 10 bootstrap samples was simulated. The dimensionality was varied from 1 to 4.

Table 4.4 represents results for 1000 event training and the same 1000 for classification and 10 bootstrap replicas. The last table 4.5 represents 2 big bootstrap trials 1000 for 100 training and 100 for 1000. Only dimensionality 4 was used. As it is seen from the tables, especially for the large samples and bootstrap replicas number, the bootstrap fraction estimates along with m.s.d. estimates are in consistence with sampling estimates.

Table 4.2: Bootstrap expectations and bootstrap standard deviations of sampling statistics

	B	10	50	100	200
$M=25$ $\delta_{25}=0.2$	$E\{\mu_*-\mu_M\}$	-0.0152	0.0031	-0.0048	-0.0003
	$\sigma\{\mu_*-\mu_M\}$	0.0639	0.0251	0.0174	0.0160
	$E\{\delta_*\}$	0.1891	0.1974	0.1929	0.1977
	$\sigma\{\delta_*\}$	0.0560	0. 0.0227	0.0031	0.0028
$M=50$ $\delta_{50}=0.1414$	$E\{\mu_*-\mu_M\}$	-0.0024	-0.0023	0.0003	-0.0001
	$\sigma\{\mu_*-\mu_M\}$	0.0402	0.0227	0.0149	0.0097
	$E\{\delta_*\}$	0.1481	0.1398	0.1396	0.1395
	$\sigma\{\delta_*\}$	0.0286	0.0182	0.0167	0.0154
$M=100$ $\delta_{100}=0.1$	$E\{\mu_*-\mu_M\}$	-0.0171	-0.0010	-0.0004	-0.0008
	$\sigma\{\mu_*-\mu_M\}$	0.0323	0.0152	0.0101	0.0066
	$E\{\delta_*\}$	0.0897	0.0959	0.1000	0.0988
	$\sigma\{\delta_*\}$	0.0212	0.0107	0.0097	0.0086
$M=200$ $\delta_{200}=0.0707$	$E\{\mu_*-\mu_M\}$	0.0038	-0.0017	0.0001	0.0000
	$\sigma\{\mu_*-\mu_M\}$	0.0231	0.0107	0.0082	0.0048
	$E\{\delta_*\}$	0.0593	0.0692	0.0694	0.0700
	$\sigma\{\delta_*\}$	0.0154	0.0078	0.0063	0.0049

$M=500$ $\delta_{500}=0.0447$	$E\{\mu_*-\mu_M\}$ $\sigma\{\mu_*-\mu_M\}$ $E\{\delta_*\}$ $\sigma\{\delta_*\}$	-0.0018 0.0115 0.0430 0.0095	0.0007 0.0072 0.0452 0.0043	0.0004 0.0040 0.0442 0.0033	0.0003 0.0032 0.0446 0.0024
$M=1000$ $\delta_{1000}=0.032$	$E\{\mu_*-\mu_M\}$ $\sigma\{\mu_*-\mu_M\}$ $E\{\delta_*\}$ $\sigma\{\delta_*\}$	0.0038 0.0079 0.0322 0.0073	0.0001 0.0050 0.0317 0.0033	0.0002 0.0030 0.0316 0.0022	0.0003 0.0022 0.0315 0.0017

Table 4.3: Fraction estimation, M=100; B=10

N	1	2	3	4
R^B	0.309	0.240	0.193	0.159
Index	0.632	0.719	0.788	0.825
R^e	0.285	0.280	0.205	0.165
P_1^e	0.367	0.338	0.297	0.285
$\hat{P}_1$	0.162	0.183	0.206	0.219
R_*^e	0.295 ± 0.094	0.264 ± 0.075	0.221 ± 0.045	0.173 ± 0.072
P_{*1}^e	0.394 ± 0.096	0.340 ± 0.045	0.277 ±0.041	0.270 ± 0.038
$\hat{P}_{1*}$	0.132 ±0.165	0.193 ±0.078	0.263 ±0.022	0.203 ± 0.039
$< R^e >$	0.325 ± 0.098	0.252 ± 0.082	0.221 ± 0.082	0.177 ±0.029
$< P_1 >$	0.355 ±0.103	0.393 ± 0.077	0.324 ± 0.057	0.311 ±0.039
$< \hat{P}_1 >$	0.184 ±0.075	0.156 ±0.087	0.220 ±0.033	0.173 ± 0.063

Table 4.4: Fraction estimation, M=1000; B=10

N	1	2	3	4
R^B	0.309	0.240	0.193	0.159
Index	0.700	0.767	0.813	0.839
R^e	0.300	0.233	0.187	0.159
P_1^e	0.367	0.338	0.297	0.285
$\hat{P}_1$	0.162	0.183	0.206	0.219

R_*^e	0.300 ± 0.044	0.231 ± 0.029	0.187 ±0.029	0.159 ±0.023
P_{*1}^e	0.376 ± 0.043	0.348 ± 0.018	0.301 ± 0.020	0.291 ±0.018
$\hat{P}_{*1}$	0.163 ±0.029	0.189 ±0.020	0.205 ±0.014	0.217 ±0.008
$< R^e >$	0.310 ±0.029	0.250 ±0.019	0.200 ±0.025	0.164 ±0.018
$< P_{*1}^e >$	0.405 ± 0.030	0.342 ± 0.025	0.329 ±0.024	0.309 ± 0.020
$< \hat{P}_1 >$	0.194 ±0.040	0.191 ±0.040	0.186 ±0.035	0.199 ±0.019

Table 4.5: Fraction estimation, N=4, B=1000; B=10

N=4	100x1000	1000x100
R^B	0.159	0.159
Index	0.825	0.839
P_1^e	0.260	0.285
$\hat{P}_1$	0.156	0.219
R^e	0.165	0.159
R_*^e	0.171 ±0.073	0.160 ±0.019
P_{*1}^e	0.270 ±0.042	0.287±0.015
$\hat{P}_{1*}$	0.205 ±0.051	0.215 ±0.012
$< R^e >$	0.170 ±0.062	0.165 ±0.021
$< P_1^e >$	0.312 ± 0.052	0.304 ± 0.020
$< \hat{P}_1 >$	0.208 ± 0.051	0.202 ± 0.023

4.1.4 BAYESIAN MAPPING

Bayesian decision rule (3.12) is defined in each point of feature space $\mathcal{V}$ and for each point the definite decision is made (of course each point $\upsilon \subset \mathcal{V}$ have to have the "physical" meaning, visa-versa the outliers selection rule (3.15) will be triggered). υ points attributed for one and the same category usually form, so called, "clusters" - compact regions in $\mathcal{V}$. The shape of this regions can be

highly nonlinear and ever disconnected. On the figure 4.3 we can see distinct clusters corresponding to samples from 2-climensional Gaussian population with means 1, 5, 7. Variances of all classes are equal to 1, 1000 events for each class were used. As the sample means are rather far from each other, the clusters get definite shape containing events from particular class, simultaneously rejecting events from the other classes. The aspheric shape of clusters is explained by the aspherical symmetry of 2-dimensional uncorrelated Gaussian population, the middle class variables are correlated, and therefore second cluster points (superimposed on the figure) had an elliptic shape.On the figure 4.4 the clusters corresponding to the two samples from the one and the same Gaussian population are presented. Very complicated form of cluster is explained by the sampling random fluctuation.

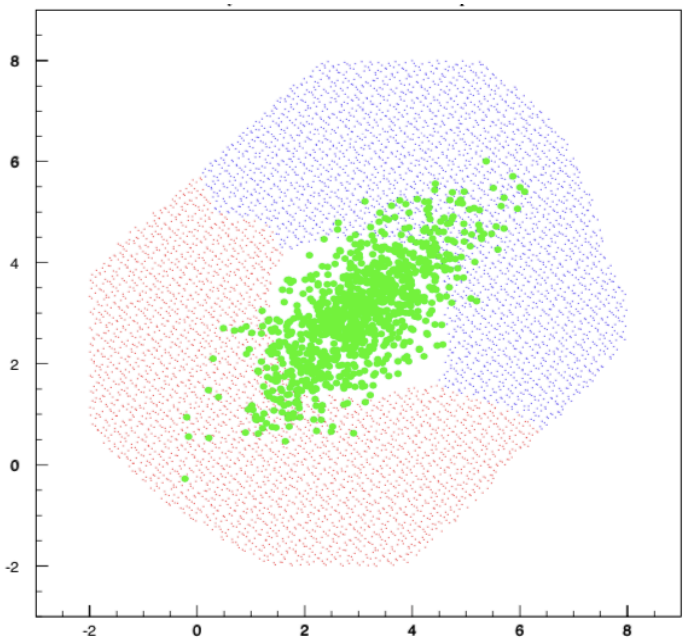

Figure 4.3 The Bayesian clusters for the samples from Gaussian populations with different means

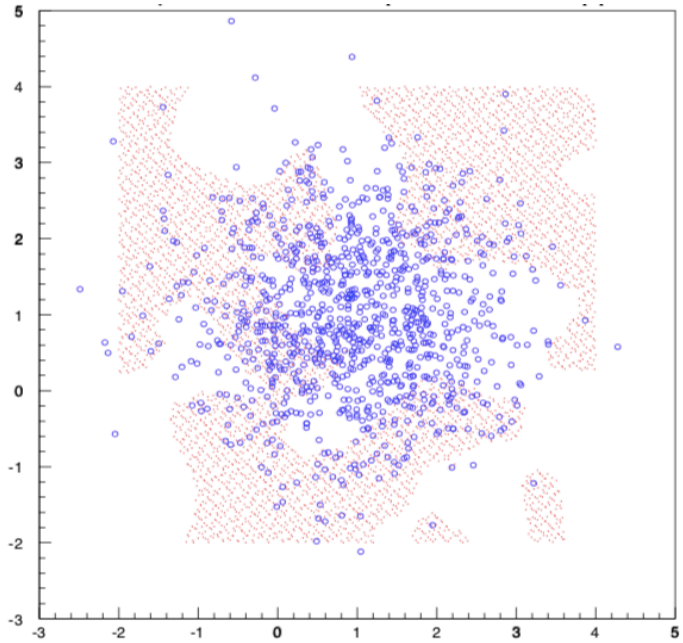

Figure 4.4 The Bayesian clusters for the samples from Gaussian populations with same means

4.2 ARTIFICIAL NEURAL NETWORKS MODELS

4.2.1 NEURAL CLUSTERS

We use the 2:4:2:1 feedforword network for detecting the 2-climensional clusters with radial symmetry. Training samples consists of two classes of 450 uniform distributed "background" events. The 50 "signal "events generated according to 2-climensional radial symmetric Gaussian population with mean - 0.5 and σ- 0.03 were added to one of background samples.

The goal of algorithm is to find a 2-climensional cluster maximizing the objective function (3.41), e.g. containing as much as possible "signal" events, and rejecting the background events.

Figure (4.5) preents the results of Sobol sieves search strategy (see section 3.4.2). A is size of pseudo-random sieve.

After obtaining several trained networks, the obtain "chromosome" where used as a "pool" for genetic algorithm. The "fittest descent" strategy after several generations of offsprings, as one can see from figure 4.6 improves situation. Both "mutations" and "crossovers" increase σ values.

4.2.2 DETECTION OF THE CRAB NEBULAE BY THE WIPPLE COLLABORATION

The best discrimination technique used in the WHIPPLE Observatory is the multidimensional cuts *(supercuts)* method proposed in ([37]) and then improved in ([30]) (four Cherenkov image parameters were used). The method consists of a posterior selection of the best gamma-cluster (multidimensional box), containing "gamma-like" events. The particular coordinates of the box were selected to maximize the σ value on the 1988-1989 Crab nebula observation data base (65 ON, OFF pairs ~ 10^6 events) ([29]). By implementing the supercuts method, the initial σ value was enlarged from 5(raw data) to 34. The parameters of the Cherenkov image, used for background rejection reflect the inherent differences in angular size and shape from two types of images (WIDTH, LENGTH) and differences in the image orientation (MISS, ALPHA), AS WELL AS the estimate of impact parameter of particle – DIST; the dispersion parameter (CONG) of the images have also been used. A single parameter can be defined which combines the shape and orientation criterion - AZWIDTH.

We use a simple 4::5::1 neural net to select the better nonlinear shape of the gamma- cluster. The net was trained on experimental ON&OFF events. For Neural analysis were used the same variables as for Supercut analysis: WIDTH, LENGTH, DIST, ALPHA. The comparison of different background suppression methods one can see in the table 4.6. DIFF $= N^*_{on} - N^*_{off}$ is the estimate of the signal, DIFF/N^*_{off}- is the estimate of the signal to noise ratio,$\frac{N^*_{off}}{N^*_{on}}$- is the estimate of background suppression by used technique.

Table 4.6: WHIPPLE Crab detection, 1988-1989

	N^*_{on}	N^*_{off}	σ	DIFF	DIFF/N^*_{off}	$\frac{N^*_{off}}{N^*_{on}}$
Raw	506255	501408	4.8	4847	0.01	
AZWIDTH	14622	11389	20.4	3233	0.28	0.0227
WEDGE cut ([37])	6017	3381	27.2	2636	0.78	0.0067
SUPERCUT ([30])	4452	1766	34.3	2686	X .52	0.0035
NEURAL 4::5::1	6278	2858	35.8	3420	1.20	0.0057

The neural nonlinear cluster is much less restrictive than supercut and AZWIDTH cuts. More "intelligent", smooth nonlinear shape of the γ-cluster ensures the significant enhancement of the signal detection efficiency along with higher rejection of the background.

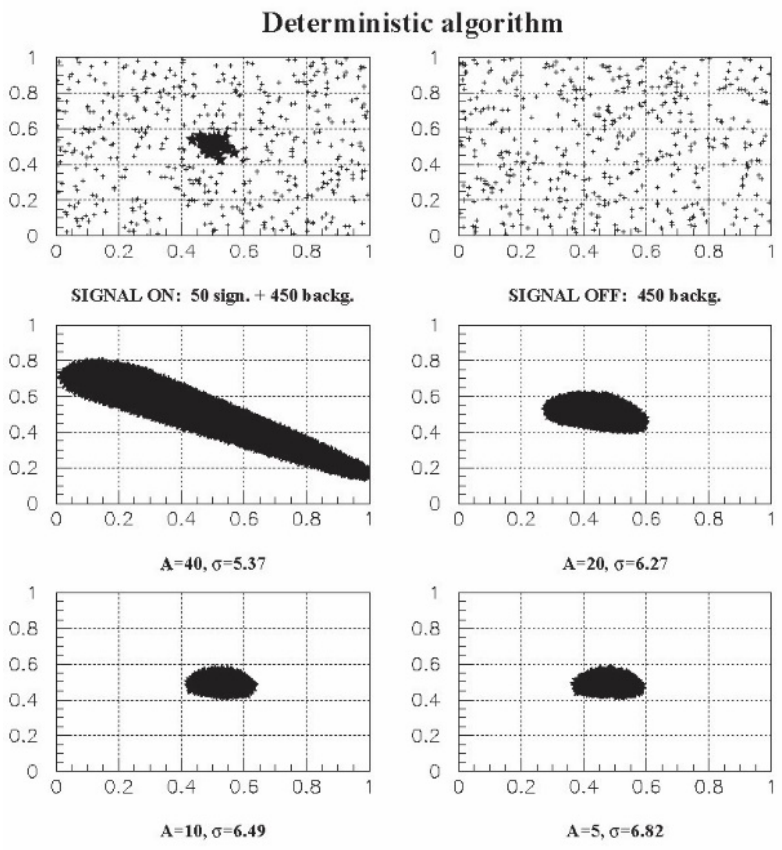

Figure 4.5: Deterministic Search

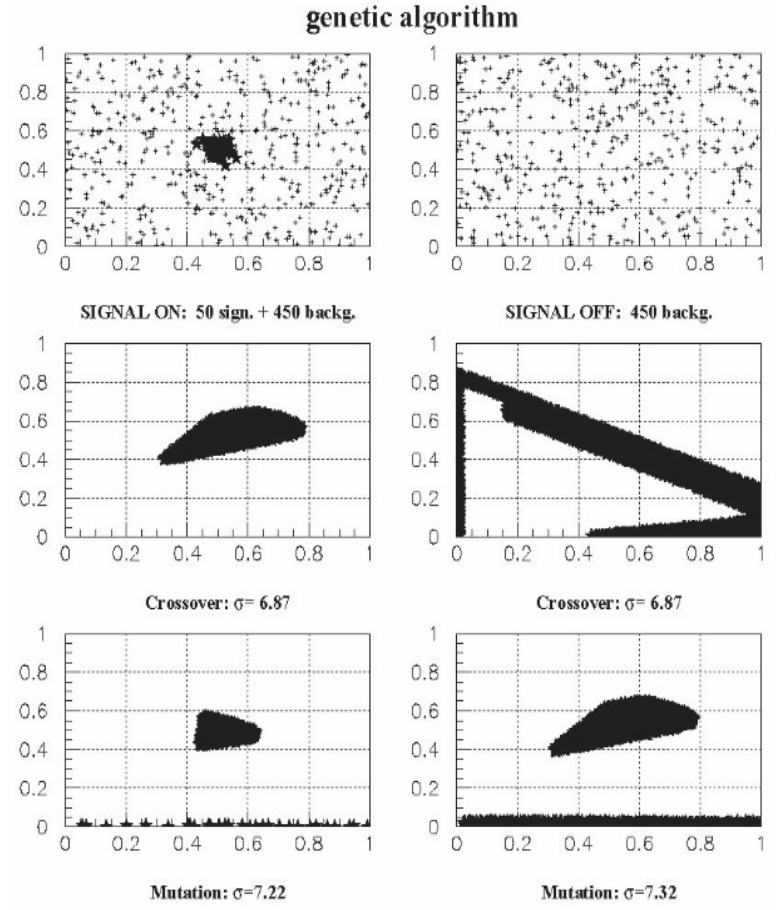

Figure 4.6: Genetic Search

CHAPTER 5
PRIMARY NUCLEI CLASSIFICATION IN 3 CATEGORIES (KASCADE DATA ANALYSIS)

5.1 THE EXTESIVE AIR SHOWER (EAS) SIMULATION

All statistical decisions and procedures are correct within the prechosen model. Thus, a realistic simulation is the key problem of any physical inference in indirect experiments. Research of Extensive air showers is a classic example of such a situation. Sure, adequate consideration of detector response and an identical reconstruction of experimental and simulated data are necessary steps of data analysis.

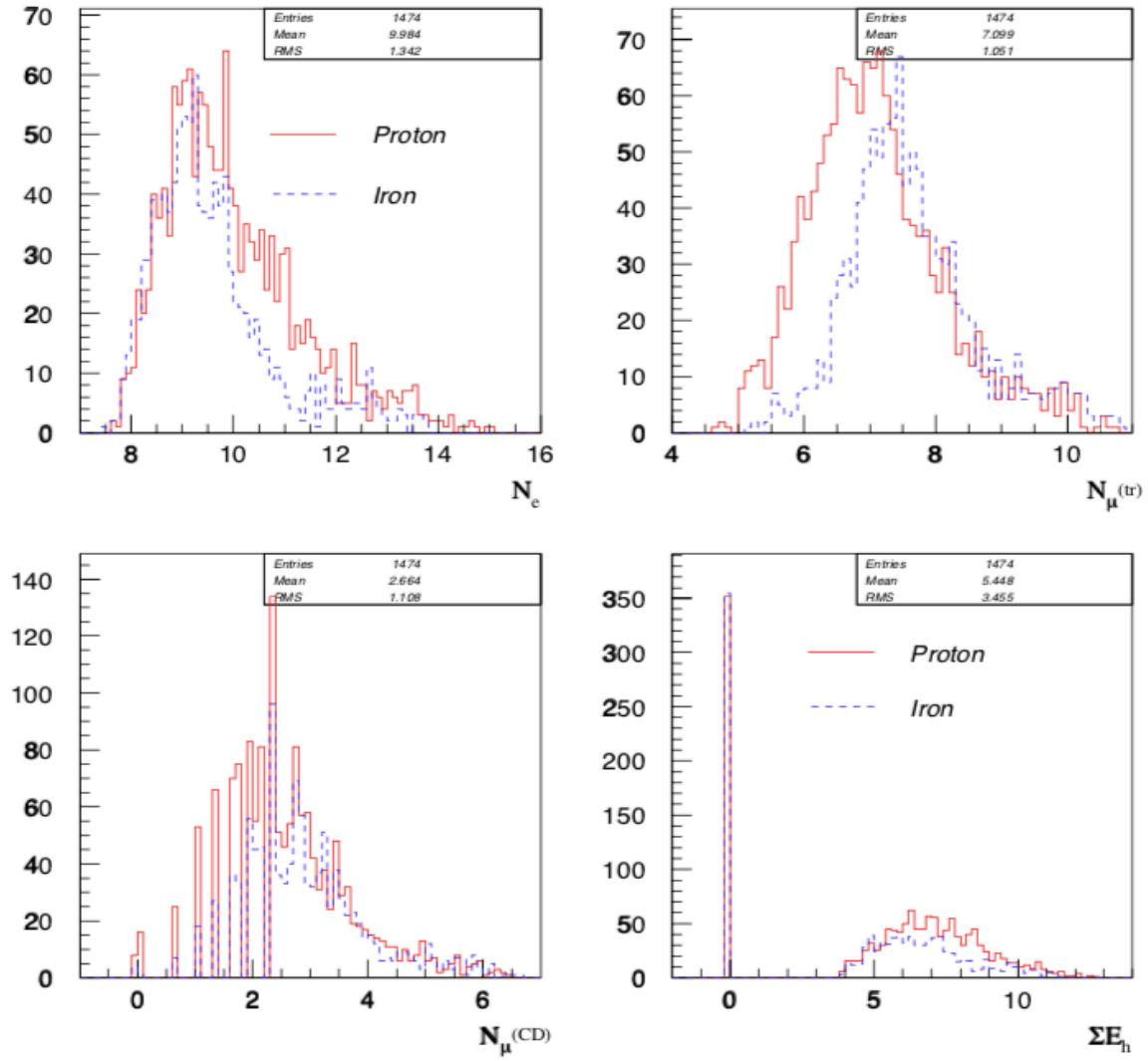

Figure 5.1: Features distribution for proton and iron

The first simulation data base of the KASCADE experiment fulfilling the above requirements is available since recently, and we use specific EAS parameters, like the numbers of electrons/photons (N_e) and truncated number of muons ($N_{\mu^{tr}}$)and the age (N_{30}) parameters as input for data analysis. The physical meaning of these variables will not be discussed; we only mention that the procedures of their treating are identical for experimental data and simulations. It is also very important to say that hypotheses about lateral distributions of muons and electrons at

very small and large distances are not of influence. The simulations of the EAS development in the atmosphere was done with the (CORSIKAcode (version5.2: VENUS and QGS models). For the calculation of the full detector response function the GEANT CERN package was used. The parameters of simulated showers were reconstructed with the same programs as experimental ones.

The measured EAS parameters by KASCADE detector are presented in Tab. 5.1 and Fig. 5.1. As one can see from the figure the overlapping of shower parameters corresponding ever to most distinct proton and iron classes is rather large and one can't expect reliable classification of primaries according to the single EAS features.

In the multidimensional features space, as one can see from Fig. 5.2 the differences between proton and iron samples is much larger. Therefore, the detailed examination of the discriminative power of all EAS characteristics and their correlations will allow to find a best subset of features to be used for experimental data classification.

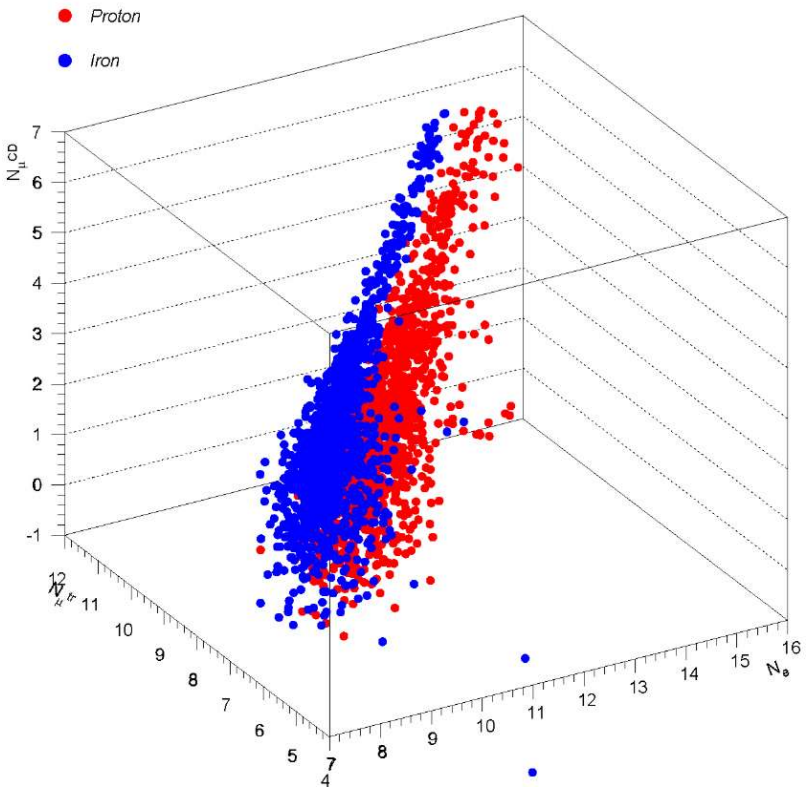

Figure 5.2. Proton and iron events distribution in 3-dimensional space of features

Table 5.1: EAS features detected by KASCADE experiment

N_e	Number of electrons in EAS
N_μ^{tr}	Truncated number of muons (number of muons in the range of 40 to 200m)
S_{30}	Shower age associated with a Molier radius 30m
N_μ^*	Number of muons in central detector
N_h	Number of hadrons
E_h^{max}	The sum of energy of most energetic hadrons
E_{sum}	Total energy of hadrons

5.2 VALIDATION OF MODELS

5.2.1 COMPARISON OF THE SINGLE EAS VARIABLES

First of all, we examine single variables to select primary mass discriminants and variables correlated with the primary energy. We use for these purposes simulated events, for which the "true" values of mass and energy are known. For all EAS variables we calculate the P-values of following statistical tests:

- Student's t-test

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

where the μ_1, σ_1,and μ_2, σ_2 are the mean values and the standard deviations of the first and second classes.

- Kolmogorov-Smirnov *D*-test

$$D = supremum_v |F_1(v) - F_2(v)|$$

where $F_1(v)$ and $F_2(v)$ are the cumulative probability function for first and second classes(models): $\left(F(v) = \frac{N(v_i<v)}{N(v)}\right)$

- Mann-Whitney U-test

where the T_1 and T_2 are the sum of ranks of events from first and second samples , and M_1, M_2- are the number of events in samples. The rank is the number of particular event in ordered sequence of events (so called variation sequence).

5.2.2 CORRELATION ANALYSIS

The correlation analysis was done to select the single variables and - best pairs of variables for distinguishing between classes.

Table 5.2: P-values of statistical tests for proton and iron classes for different models: t - Student, D - Kolmogorov- Smirnov, U -Mann-Whitney

QGS	t	D	U	VENUS	t	D	U
N_e	3.177	2.747	4.996	N_e	2.869	3.161	5.778
N_μ^{tr}	12.601	6.026	12.723	N_μ^{tr}	10.274	4.282	9.403
S_{89}	17.160	7.489	17.294	S_{89}	20.415	8.314	19.473
N_μ^{CD}	7.207	3.452	7.132	N_μ^{CD}	5.650	2.031	3.872
N_h	0.673	1.647	2.811	N_h	2.265	3.335	5.848
E_h^{max}	5.564	3.066	6.144	E_h^{max}	3.612	2.402	4.458
E_h^{sum}	2.478	3.126	3.985	E_h^{sum}	3.457	3.140	6.174

Table 5.3: Correlation matrix for QGS model

	Mass	E_0	N_e	N_μ^{tr}	S_{30}	N_μ^{CD}	N_h	$maxE_h$	$sumE_h$
Mass	1.00	0.21	-0.03	0.27	0.32	0.15	-0.03	-0.11	-0.07
E_0	0.21	1.00	0.92	0.95	-0.25	0.94	0.78	0.53	0.73
N_e	-0.03	0.92	1.00	0.90	-0.43	0.93	0.85	0.62	0.81
N_μ^{tr}	0.27	0.95	0.90	1.00	-0.23	0.93	0.78	0.52	0.72
S_{30}	0.32	-0.25	-0.43	-0.23	1.00	-0.33	-0.39	-0.33	-0.38
N_μ^{CD}	0.15	0.94	0.93	0.93	-0.33	1.00	0.86	0.60	0.82
N_h	-0.03	0.78	0.85	0.78	-0.39	0.86	1.00	0.70	0.95
$maxE_h$	-0.11	0.53	0.62	0.52	-0.33	0.60	0.70	1.00	0.73
$sumEh$	-0.07	0.73	0.81	0.72	-0.38	0.82	0.95	0.73	1.00

Table 5.4: Correlation matrix for Venus model

	Mass	E_0	N_e	N_μ^{tr}	S_{30}	N_μ^{*}	N_h	$maxE_h$	$sumE_h$
Mass	1.00	0.18	-0.06	0.18	0.33	0.09	-0.05	-0.09	-0.07
E_0	0.18	1.00	0.91	0.95	-0.24	0.94	0.80	0.52	0.76
N_e	-0.06	0.91	1.00	0.90	-0.40	0.93	0.89	0.60	0.85
N_μ^{tr}	0.18	0.95	0.90	1.00	-0.24	0.92	0.81	0.51	0.76
S_{30}	0.33	-0.24	-0.40	-0.24	1.00	-0.32	-0.41	-0.28	-0.40
N_μ^{CD}	0.09	0.94	0.93	0.92	-0.32	1.00	0.86	0.57	0.83
N_h	-0.05	0.80	0.89	0.81	-0.41	0.86	1.00	0.65	0.95
$maxE_h$	-0.09	0.52	0.60	0.51	-0.28	0.57	0.65	1.00	0.68
$sumE_h$	-0.07	0.76	0.85	0.76	-0.40	0.83	0.95	0.68	1.00

5.2.3 PROBABILISTIC DISTANCES

Another important measure of the separability of two samples is the *Bhattacharya distance,* which takes the form:

$$R_{BHATA} = \frac{1}{8}(\mu_2 - \mu_1)^T \left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{|\frac{\Sigma_1+\Sigma_2}{2}|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

where the μ_i and Σ_i are the first moments vector and covariance matrix of i-th class. The first term of this equation is the *Mahalanobis distance* and the second term is the correlation distance.

$$R_{Mahal} = (\mu_2 - \mu_1)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_2 - \mu_1)$$

$$R_{Corr} = \ln \frac{|\frac{\Sigma_1+\Sigma_2}{2}|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

We select the best subsets of EAS features according the Bhattacharya distance

5.2.4 KASCADE EXPERIMENTAL DATA

Table 5.5: The best feature subsets according to the Bhattacharya distance

	2 best	next best	worst
QGS	N_e N_μ^{tr}	N_μ^* S_{30} $sumE_h$	N_h $maxE_h$
VENUS	N_e N_μ^{tr}	S_{30} N_μ^* $sumE_h$	N_h $max\ E_h$

2371 events of central calorimeter and 450000 events of array data were r analyzed. The events were selected within 15 — 20° zenith angle range (the simulations were done for the same angles). To check the homogeneity of the experimental data we divide it to 3 subsamples and make multiple comparisons with techniques described above. Also, the negative Log likelihood function value $\pounds$ (3.17) and estimate of Bayesian error R^e (3.29) were calculated.

Table 5.6: Experimental data homogeneity t est features used: N_μ^{CD}, E_h^{sum}

	£	$R_{Mahal..}$	$R_{Bhata.}$	$R_{corr.}$	R^e
1 class	2.911	0.011	0.007	0.024	0.479
2 class	2.803	0.029	0.021	0.069	0.466

The homogeneity check for array and calorimeter data one can find in tables 5.6, 5.7, 5.8, 5.9. All tests demonstrate rather good agreement with the each other and prove the homogeneity of experimental data samples.

Table 5.7: Experimental data homogeneity test features used: N_e, N_μ^{tr},N_μ^{CD}, E_n^{sum}

	£	$R_{Mahal.}$	$R_{Bhata.}$	$R_{corr.}$	R^e
I class	4.434	0.129	0.038	0.086	0.429
2 class	4.289	0.056	0.041	0.135	0.453

Table 5.8: Experimental data homogeneity test features used N_e, N_μ^{tr}

	£	$R_{Mahal.}$	$R_{Bhata.}$	$R_{corr.}$	R^e
I class	1.428	0.002	0.000	0.000	0.490
2 class	1.425	0.001	0.000	0.000	0.494

Table 5.9: One dimensional tests :t-Student,D-Kolmogorov-Smirnov,U-Mann-Whitnay

	t	D	U
N_e	1.816	1.114	1.433
N_μ^{tr}	0.080	0.823	0.570
N_μ^{CD}	1.595	1.118	1.776
E_h^{sum}	1.418	0.407	1.358

5.2.5 QGS AND VENUS MODELS COMPARISON

To compare CORSIKA different strong interaction models, one have to have the same mass composition and energy distribution in experimental and simulated data, to avoid mass depended differences. The mass composition of primary cosmic radiation in low energy region (bellow $2*10^{15}$ eV) is measured by direct methods and the following proportion of different nucleus is assumed to be true [62]: *H- 24%; He- 31%; O - 21%>; Si - 12%>; Fe - 12%.*

To avoid energy spectrum based differences we choose rather narrow energy interval. The truncated muon interval (in logarithmic scale) $7.82 \leq N_\mu^{tr} \leq 9.21$ is corresponding to the $6 *$

$10^{14} \leq E_0 \leq 2 * 10^{15}$eV. Thus, we construct simulation samples from VENUS and QGS models with described above proportion of primaries in the mentioned energy range. The SAME selection was made in both, Monte Carlo and experimental data. Both models are very close to experimental data (see tables 5.11 and 5.10), but all tests give a small preference to the VENUS model. On the colored map 5.3 one can see the regions of preference with experimental data superimposed; VENUS model forms a more compact cluster compared with QGS.

5.2.6 THE KASCADE CLASSIFICATION MATRICES

The examination of classification matrix and it's index (3.30) gives clues for understanding the discriminative power of different EAS measurable for composition estimation.

The value greater than of 0.6 are still allow for solving the system of equation (3.32) . For the lower values the solutions didn't converge and fraction couldn't be reconstructed. Therefore we have to find appropriate variables, or reduce the number of classes. Only balance between expected classification errors and number of classes used, will allow to obtain reasonable and reliable estimates of the fraction.

Table 5.10. Comparison of exp. data with VENUS and QGS models

	L	R_{Bhata}	R^e
QGS	1.2036 ±0.01	0.023 ±0.001	0.456 ±0.01
VENUS	1.1818 ±0.01	0.014 ±0.001	0.469 ±0.02

Table 5.11. One dimensional tests for models andexperiment

VENUS	t	D	U	QGS	t	D	U
N_e	0.916	2.312	0.901	N_e	3.369	3.004	4.368
N_μ^{tr}	3.199	1.450	2.653	N_μ^{tr}	4.609	2.632	4.787

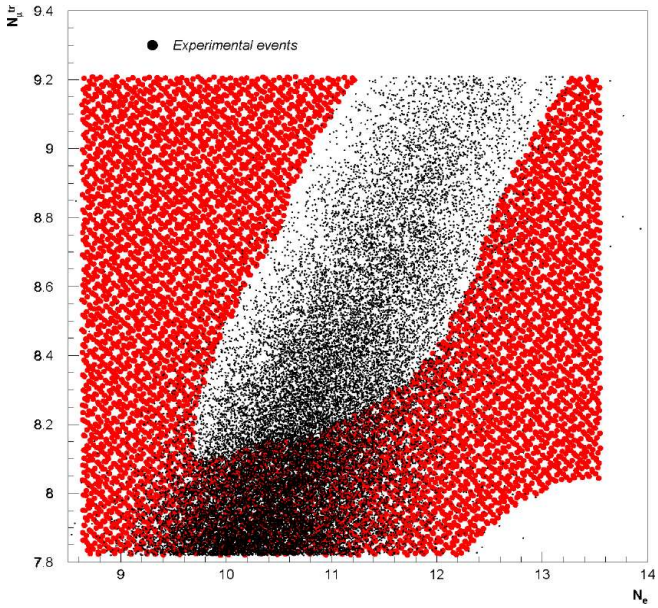

Figure 5.3: QGS (red area) and VENUS (white area) clusters and experimental events distribution in N_e, N_μ^{tr} space

As one can see in Tables (5.12), (5.13), (5.14), present status of a priori knowledge accumulated in M.C. models and represented in training samples, didn't support the attempts to make 5-way classification even for all available features.

Table 5.12: Calorimeter data, features used N_μ^{CD}, E_h^{sum}

0.5833	0.1172	0.0260	0.1777	0.0958
0.4211	0.1194	0.0244	0.2490	0.1861
0.3126	0.1138	0.0121	0.2701	0.2914
0.3362	0.1017	0.0276	0.2845	0.2500
0.3005	0.0863	0.0216	0.2311	0.3606

Table 5.13: Array data, features used N_e, N_μ^{tr}

0.5148	0.2620	0.1024	0.0565	0.0643
0.3240	0.2979	0.1846	0.0885	0.1050
0.1185	0.1803	0.2340	0.1721	0.2951
0.0707	0.1064	0.1942	0.1958	0.4330
0.0445	0.0659	0.1006	0.1362	0.6527

Table 5.14: KASCADE data, features used $N_e, N_\mu^{tr}, N_\mu^{CD}, E_h^{sum}$

0.4785	0.3085	0.1323	0.0438	0.0368
0.2678	0.3863	0.1943	0.0766	0.0750
0.0558	0.1932	0.2903	0.2200	0.2407
0.0367	0.1247	0.2311	0.2506	0.3570
0.0368	0.0706	0.1544	0.1471	0.5912

The situation with 3-way classification is much better, as we need much less a priori information, comparing with classification into 5 nuclei groups. As we can see from tables (5.15), (5.15), (5.17), even array information only allows to resolve the distribution mixture (3.4). The calorimeter information significantly increases the expected fraction reconstruction accuracy. The 2-way classification in "heavy" and "light" nuclei can be done with significant larger accuracy. See tables (5.19), (5.18), (5.20)

The information concern 5, 3 and 2 -way classifications for KASCADE different parts is **summarized in table (5.21), where the separability indexes are presented.**

Table 5.15: 3-way classification by N_μ^{CD},E_h^{sum}

0.5749	0.2712	0.1539
0.3443	0.3319	0.3239
0.2167	0.2976	0.4857

Table 5.16: 3-way classification by N_e, N_μ^{tr}

0.6831	0.2595	0.0574
0.2132	0.4849	0.3019
0.0919	0.3078	0.6003

Table 5.17: 3-way classification by N_e, N_μ^{tr},N_μ^{CD},E_h^{sum}

0.7108	0.2419	0.0473
0.1777	0.5176	0.3048
0.0779	0.2559	0.6662

Table 5.18:2-way classification by N_μ^{CD},E_h^{sum}

0.665	0.335
0.246	0.754

Table 5.19: 3-way classification by N_e, N_μ^{tr}

0.863	0.137
0.088	0.912

Table 5.20:2-way classification by N,N_μ^{tr},N_μ^{CD},E_h^{sum}

0.865	0.865
0.076	0.076

Table 5.21: Separability index forKASCADE

	Index-	Index-3	Index-2
CD	0.154	0.462	0.708
ARRAY	0.341	0.584	0.887
ARRAY+CD	0.38	0.626	0.894

5.2.7 COLORED NUCLEAR MAPS (MASCS)

It is of greatest importance to divide initial feature space according to different primaries. Each decision rule maps vi,(or ui) events to one of 3 nuclei groups. Visa-versa, each nuclei group is mapped by decision rule (3.12) to the definite region of feature space. By examining of such "nuclear maps" we can make insight to the possibility of defining the type of particular nuclei and about expected misclassification to the other nuclear groups. Overlaying the experimental data on the colored nonlinear "masks" we can visualize the Bayesian decision procedure. The different masks, for various variables, two energy regions and 2 strong interaction models are posted below. The colored maps are corresponded to 2 - and 3-way classifications, to Re estimates and tables from the previous section.

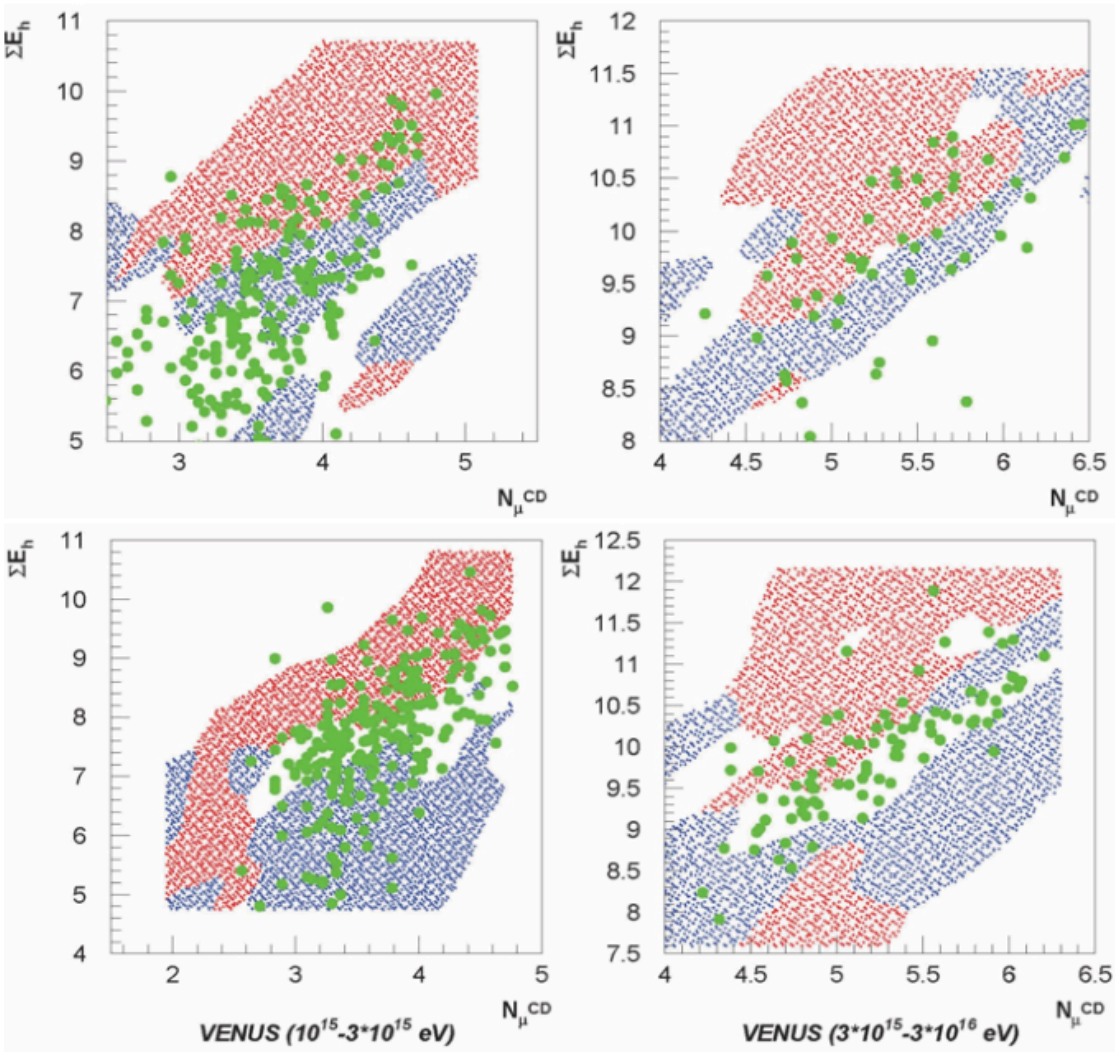

Figure 5.4: 3-way map, calorimeterinformation. Green points represent oxygen MC data.

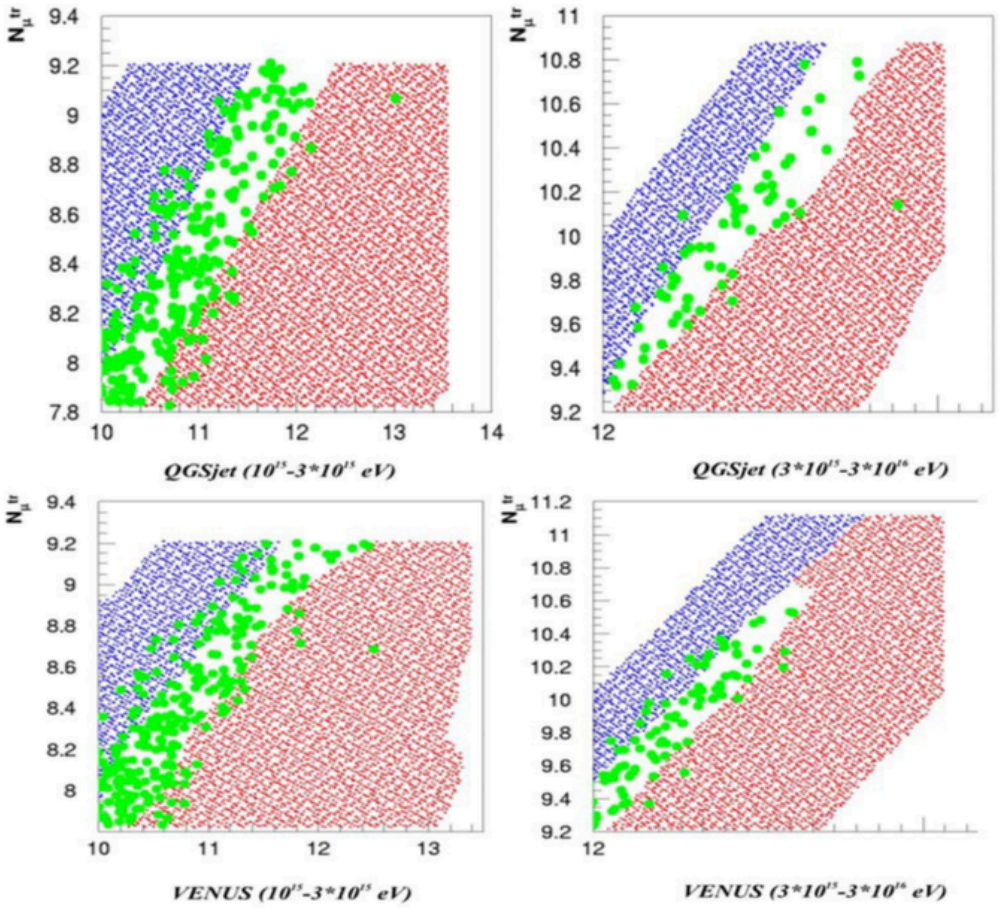

Figure 5.5: 3-way map, array information. Green points represent oxygen MC data.

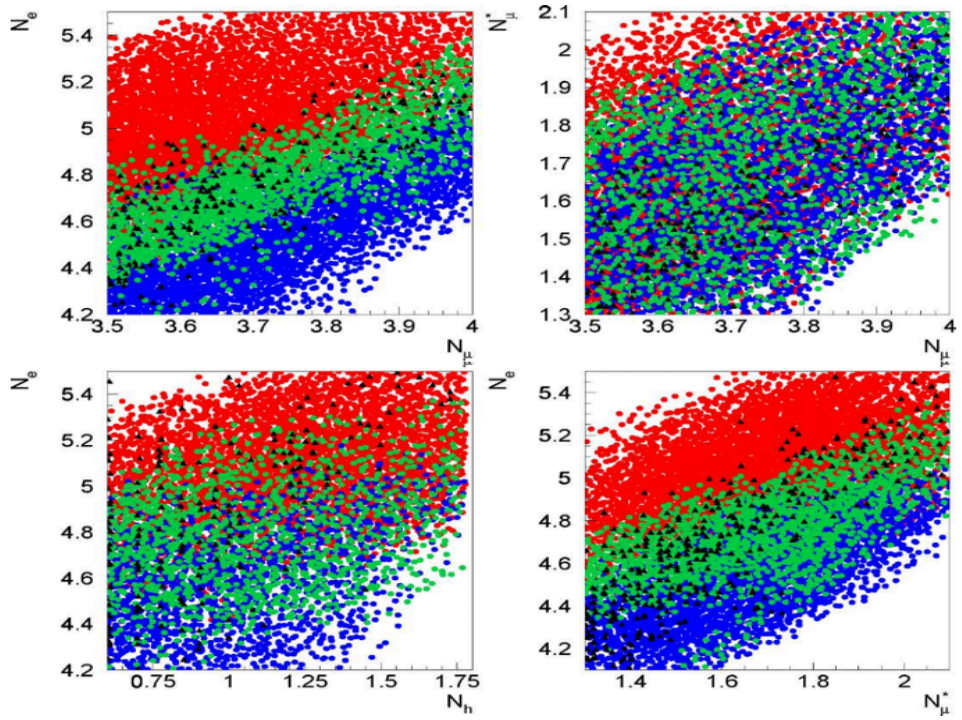

Figure 5.6: QGS model: 3-way map, array information $p_{p\equiv red,}\ p_{0\equiv green,}\ p_{Fe\equiv blue.}$ *Black triangles represent oxygen MC data.* $E_{MC} \in \left[1x10^{15}, 3x10^{15}\right] eV$

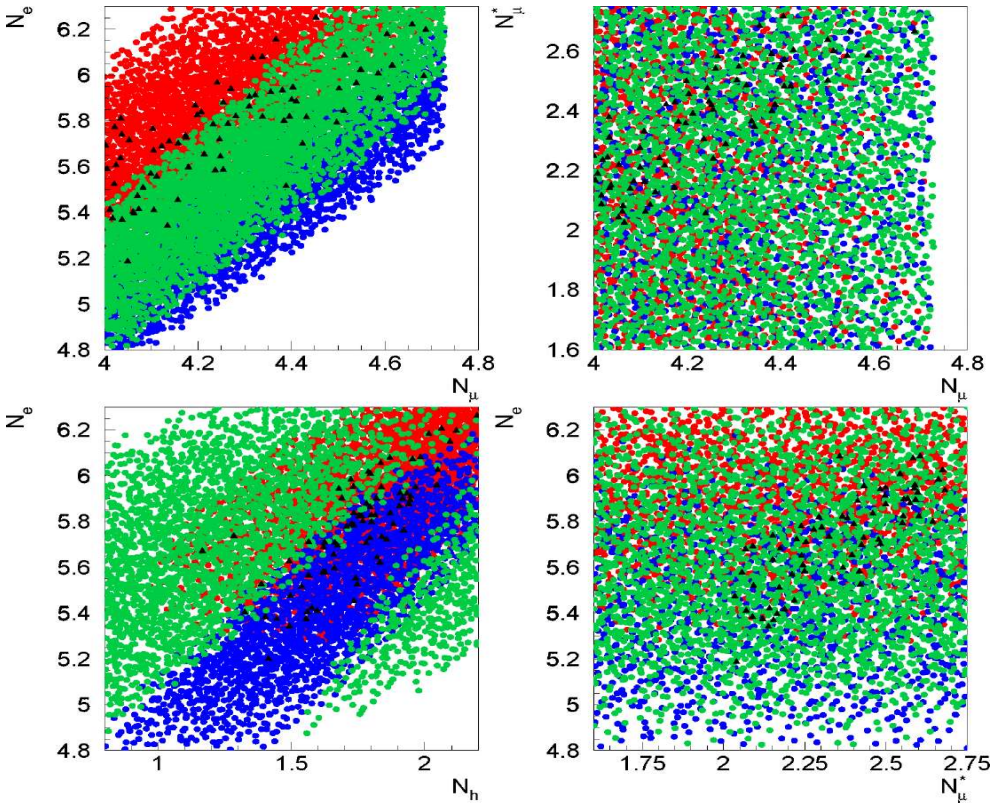

Figure 5.7: QGS model: 3-way map, array information $p_{p\equiv red,}$ $p_{0\equiv green,}$ $p_{Fe\equiv blue.}$ *Black triangles represent oxygen MC data .* $E_{MC} \in \left[3x10^{15}, 3x10^{16}\right] eV$

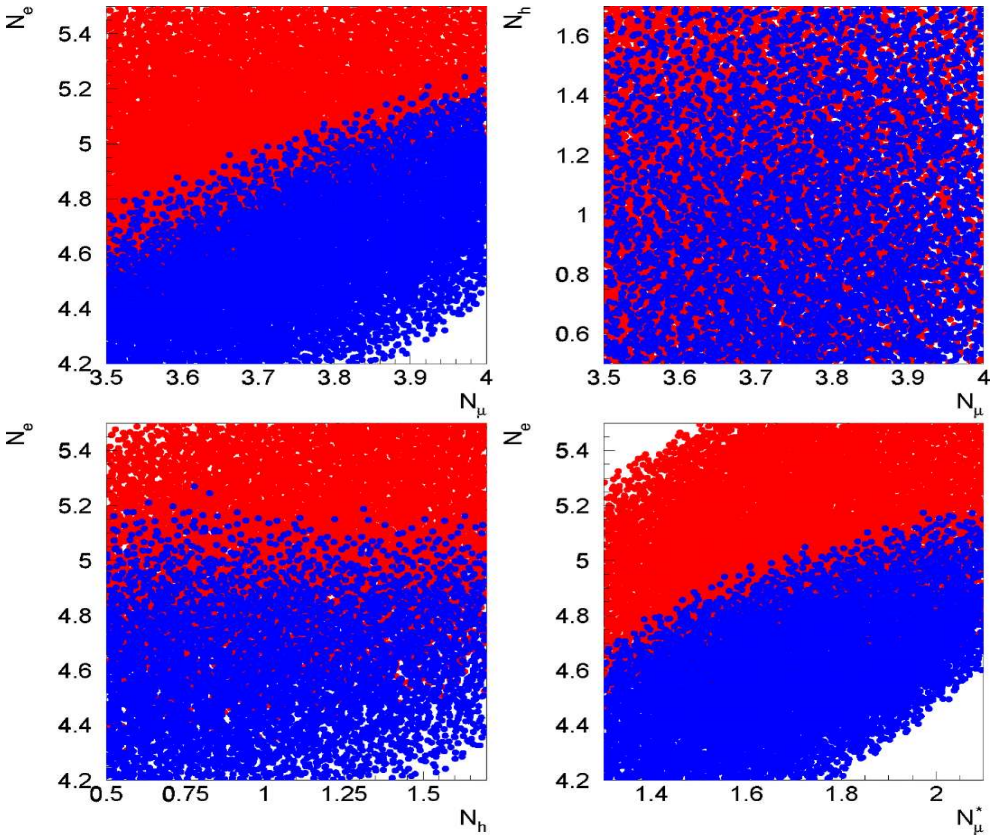

Figure 5.8: VENUS model: 2-way map, array information. $p_{p\equiv red,}p_{Fe\equiv blue.}$ $E_{MC} \in \left[1x10^{15}, 3x10^{15}\right] eV$

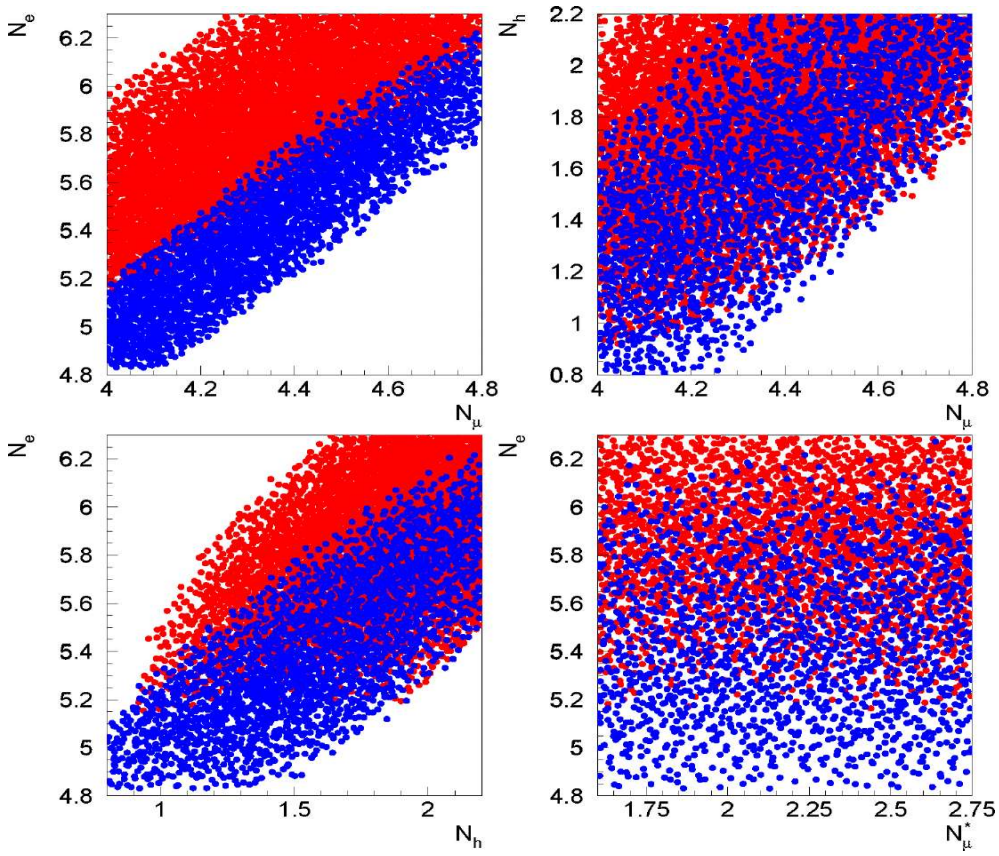

Figure 5.9: VENUS model: 2-way map, array information. $p_{p\equiv red}$, $p_{Fe\equiv blue}$.
$E_{MC} \in \left[3x10^{15}, 3x10^{16}\right] eV$

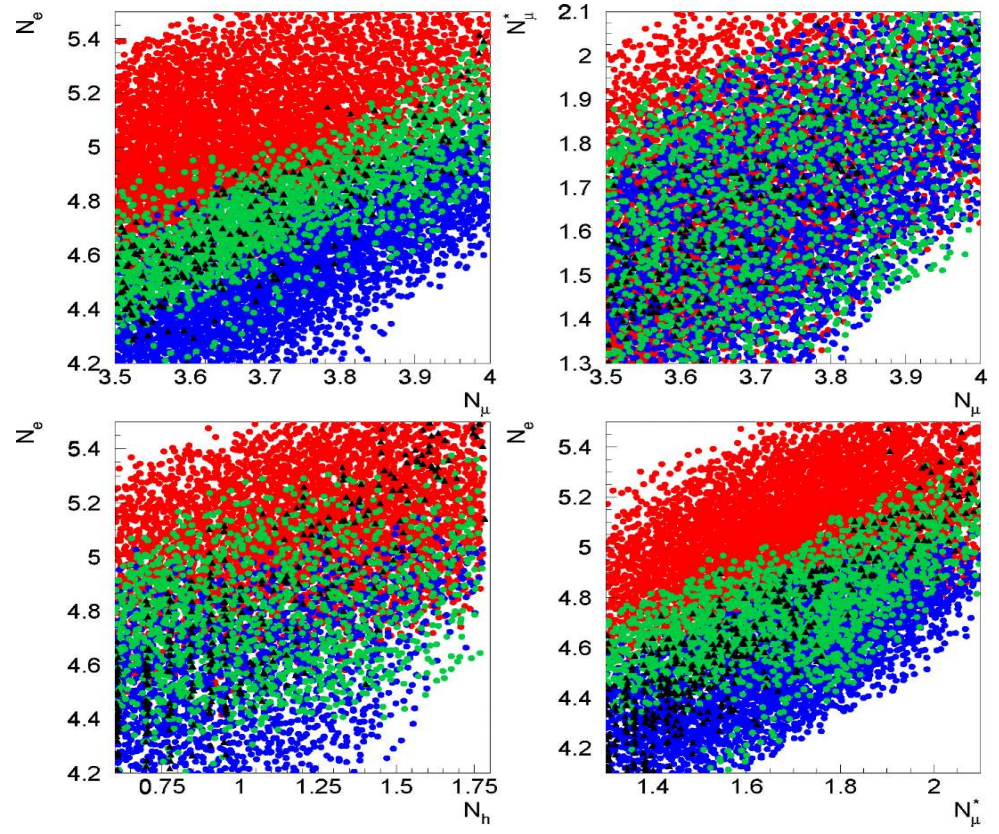

Figure 5.10: VENUS model: 3-way map, array information. $p_{p\equiv red}$, $p_{0\equiv green}$, $p_{Fe\equiv blue}$.
Black triangles represent oxygen MC data. $E_{MC} \in \left[1x10^{15}, 3x10^{15}\right] eV$

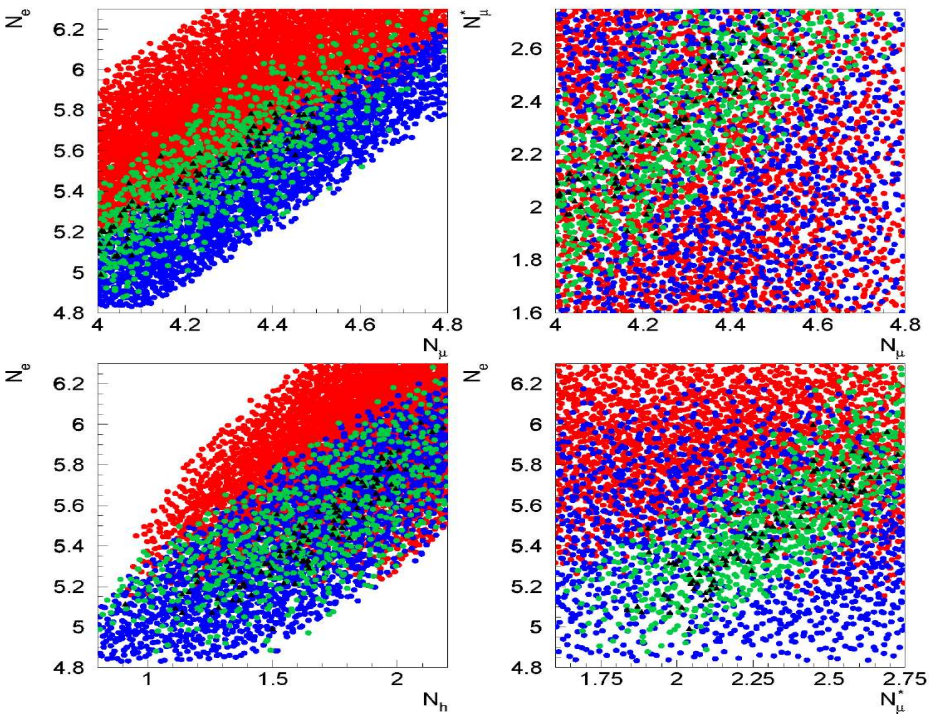

Figure 5.11: VENUS model: 3-way map, array information. $p_{p\equiv red}, p_{0\equiv green}, p_{Fe\equiv blue}$. Black triangles represent oxygen MC data. $E_{MC} \in \left[1x10^{15}, 3x10^{15}\right] eV$

5.2.8 FRACTION ESTIMATION

The KASCADE data fraction estimation was done in 5 energy bins using both calorimeter and array variables and both QGS and VENUS models. The bootstrapization procedure allows method error estimation and, as well, model error estimation. The steady tendency of heavier composition above the "knee" is detected for all variables used in analysis and for both models. Unfortunately, lack of simulations and experimental data for the highest energies didn't support yet more firm conclusions. Figures (5.12 and 5.13) and tables (5.22 -5.29) demonstrate the obtained results on the elemental composition energy dependence.

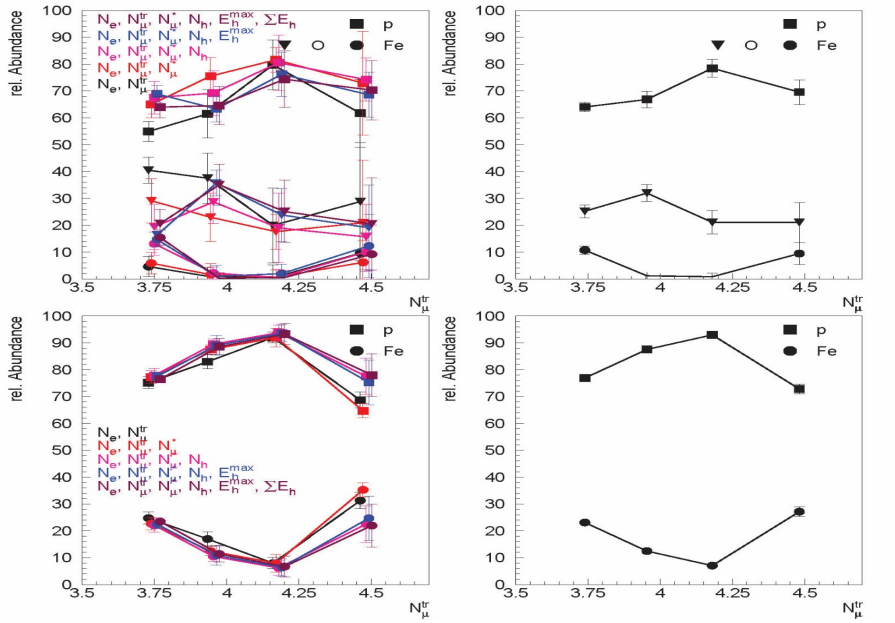

Figure 5.12: VENUS model: Reconstructed classification results using two (p, Fe) (lower graphs) and three (p, 0, Fe) (upper graphs) classes for different sets of parameters.

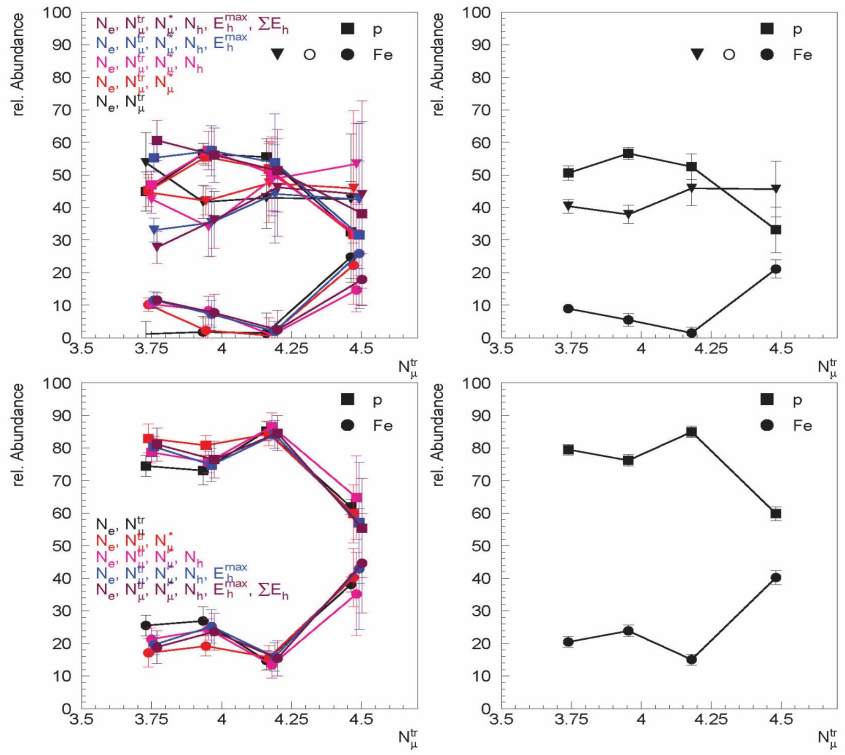

Figure 5.13 : QGS model: Reconstructed classification results using two (p, Fe) (lower graphs) and three (p, 0, Fe) (upper graphs) classes for different sets of parameters.

Table 5.22: $4.1 \leq lg_{10}N_{\mu}^{tr} \leq 4.4, M_{TS} = 150, M_{exp} = 64$

	%	stat. err.	meth. err.	model err.
p	52	±6	±2	±12
0	44	±3	±5	±12
Fe	4	±2	±2	±0.5

Table 5.23: $lg_{10}N_{\mu}^{tr} \geq 4.4, M_{TS} = 120, M_{exp} = 20$

	%	stat. err.	meth. err.	model err.
p	51	±11	±6	±18
0	33	±11	±10	±11
Fe	16	±8	±5	±6

Table 5.24: $4.1 \leq lg_{10}N_{\mu}^{tr} \leq 4.4, M_{TS} = 150, M_{exp} = 64$

		%	stat. err.	meth. err.	model err.
p		66	±11	±3	±7
Fe		34	±11	±3	±6

Table 5.25: $lg_{10}N_{\mu}^{tr} \leq 4.4, M_{TS} = 120, M_{exp} = 20$

	%	stat. err.	meth. err.	model err.
P	88	±5	±1	±3
Fe	12	±5	±1	±4

Table 5.26: $\mathbf{3.39 \leq lg_{10}N_{\mu}^{tr} \leq 3.65, M_{TS} = 555, M_{exp} = 68420(array)}$

	%	stat. err.	meth. err.	model err.
P	51	±0	±6	±6
0	42	±0	±11	±9
Fe	7	±0	±7	±1

Table 5.27: $\mathbf{3.65 \leq lg_{10}N_{\mu}^{tr} \leq 3.85, M_{TS} = 215, M_{exp} = 56100(array)}$

	%	stat. err.	meth. err.	model err.
P	51	±0	±6	±6
0	48	±0	±11	±7
Fe	1	±0	±7	±1

Table 5.28: $\mathbf{3.85 \leq lg_{10}N_{\mu}^{tr} \leq 4.1, M_{TS} = 140, M_{exp} = 20400(array)}$

	%	stat. err.	meth. err.	model err.
p	56	±0	±6	±6
0	44	±0	±11	±4
Fe	0	±0	±7	±2

Table 5.29: $\mathbf{4.1 \leq lg_{10}N_{\mu}^{tr} \leq 4.4, M_{TS} = 135, M_{exp} = 7540(array)}$

	%	stat. err.	meth. err.	model err.
P	63	±0	±8	±11
0	33	±0	±14	±12
Fe	4	±0	±8	±4

Table 5.30: $\mathbf{lg_{10}N_{\mu}^{tr} \leq 4.4, M_{TS} = 110, M_{exp} = 2285}$

	%	stat. err.	meth. err.	model err.
P	46	±1	±10	±12
0	46	±1	±14	±11
Fe	8	±0	±6	±8

5.3 THE EXAMPLES OF ANI OUTPUTS

RUN BEGINS AT 26/04/98 12.39.06

ANALYSIS AND NONPARAMETRIC INFERENCE(ANI- 98)

The platform is DEC Alppa workstation, 600 MHz CPU speed.
Revised version, May 1998, Karlsruhe.

JOB MODE - ONE-LEAVE-OUT-

Classification matrix calculation (3.29), 3.18.

JOB STATUS - noDENCURVE

DENSITY ESTIMATIORN MODE: PARZ

One of two available density estimators: PARZ or KNN

MAXIMAL EXPONENT 0.4000E+03

Maximal possible power index of exponent.

STRANGE EVENTS SELECTION, DENSITY<0.1000E -34

Bayes strengeness criterium (3.15).

Number of BOOTSTRAP replicas 10

Usually greater than 50, See (3.35)

Number of VARIABLES 3

Size of variables subset, **u**, **v** dimensionality.

CONTROLE (EXPERIMENTAL) SAMPLE:

n01 REL. COORDINATES: 9800 1000 Selected: 0

From file **n01** 1000 events are read starting from 9800, 200 - from Gaussian population N(0,1), and 800 from-N(1,1), those the fraction of "first type" events in "experimental" sample is 0.2.

TRAINING SAMPLE:
n01 REL. COORDINATES: 1000 1000 Selected: 1000

n01 REL. COORDINATES: 11000 1000 Selected: 1000

The "pure" cases - samples from Gaussian populatios N(0,1) and N(1,1), in file n01 there are 10000 five-dimensional events of both kinds, any events could be selected.

Minimal and Maximal values of Training Sample-

0.2899*E* + 01 <*X*1< 0.4073*E* + 01

-0.4220*E* + 01 < X2 < 0.3916*E* + 01

-0.4360*E* + 01 < X3 < 0.4432*E* + 01

Extremal values of each variable of selected samples.

SUPERIMPOSED BOUNDARIES:-0.5000E + 01 < XI < 0.5000E + 01

-0.5000*E* + 01 < X2 < 0.5000*E* + 01

-0.5000*E* + 01 < X3 < 0.5000*E* + 01

-0.5000*E* + 01 < X4 < 0.5000*E* + 01

-0.5000*E* + 01 < X5 < 0.5000*E* + 01

Selective "cuts" for each "measured" variable. Restrictions on the *V* feature space

ONE-LEAVE-OUT- TEST OVER TRAINING SAMPLE

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.799 0.798 0.796 0.794 0.791 0.788 0.769 0.796 0.794

1> 2 0.201 0.202 0.204 0.206 0.209 0.212 0.231 0.204 0.206

2> 1 0.181 0.177 0.175 0.168 0.166 0.164 0.145 0.175 0.168

2 > 2 0.819 0.823 0.825 0.832 0.834 0.836 0.855 0.825 0.832

ONE-LEAVE-OUT- TEST OVER TRAINING SAMPLE

The parameters of the Parzen density estimator (kernel widths). The summ is the mean estimate, **ada** - the L estimate - median of ordered sequence of estimates. Seven kernel widths are used.

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.799 0.798 0.796 0.794 0.791 0.788 0.769 0.796 0.794

1> 2 0.201 0.202 0.204 0.206 0.209 0.212 0.231 0.204 0.206

2> 1 0.181 0.177 0.175 0.168 0.166 0.164 0.145 0.175 0.168

2 > 2 0.819 0.823 0.825 0.832 0.834 0.836 0.855 0.825 0.832

Each row of this matrix represents the "classification" matrix obtained with Bayes dicisicion rule (3.12) with particularkernel width

RUN FINISHS AT 26/04/98 12.39.21 CPU TIME 0.1512E+02

RUN BEGINS AT 26/04/98 12.39.30

ANALYSIS AND NONPARAMETRIC INFERENCE(ANI- 98)

Experimental data classification.
Comparisons of exp.data with theoretical models.

JOB MODE – CLASSIFICATION

Minimal and Maximal values of Exp. Data

-0.316*E* + 01 < X1 < 0.4054*E* + 01
0.3583*E* + 01 <X2 < 0.3875*E* + 01
0.2320*E* + 01 < X3 < 0.4420*E* + 01

Calculated extremal values of experimental data.

MEAN OF LOG-LIKELIHOOD RATIO (first/second)

-0.501- 0.457-0.383-0.325-0.302-0.281-0.217- 0.360- 0.329

Calculated according (3.16), negative values corresponds to the 2 class preference, positive - to the 1.

LOG-LIKELIHOOD FUNC. - exp. according to theor. models

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 5.162 5.171 5.227 5.324 5.384 5.449 5.741 5.308 5.332
1 > 2 4.668 4.718 4.846 4.999 5.082 5.168 5.524 4.948 5.003

Negative of (3.17), the smallest values are correspond to the best model.

MEANVALUES OF BAYES ERROR AND PR. DISTANCES
R MAHALO R BHATA R CORR BAYES

1.548 0.204 0.041 0.267

Sampling estimates of probability distances between experimental and model data (3.19), (3.26).

Usually greater than 50, See (3.35)

LOG-LIKELIHOOD FUNC. - exp. according to theor. Models

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 5.162 5.171 5.227 5.324 5.384 5.449 5.741 5.308 5.332
1 > 2 4.668 4.718 4.846 4.999 5.082 5.168 5.524 4.948 5.003

BAYSIAN CLASSIFICATION OF CONTROL SAMPLE

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.309 0.309 0.306 0.298 0.296 0.292 0.280 0.305 0.2973
1 > 2 0.691 0.691 0.694 0.702 0.704 0.708 0.720 0.695 0.703

Empirical fraction estimation by implementing Bayes, decision rules (3.31), $\widehat{\boldsymbol{P}}$

RECONSRUCTED PROPORTION OF 1 TYPE EVENTS

0.207 0.213 0.211 0.208 0.208 0.205 0.216 0.209 0.206

Reconstructed proportion of first type events (3.33),

$\boldsymbol{P^e}$

SEPARABILITY MEASURE

0.654 0.657 0.657 0.661 0.660 0.659 0.657 0.657 0.661

THE ESTIMATED PROPORTIONS (CERM RQN PROGRAM)

***class* 0.90 1.00 1.20 1.40 1.50 1.60 2.00** *summ ada*

The solution of linear equations system (3.32).

1 > 1 0.207 0.213 0.211 0.208 0.208 0.205 0.216 0.2090.206
1 > 2 0.793 0.78 70.789 0.792 0.792 0.795 0.784 0.791 0.794

RUN FINISHS AT 26/04/98 12.39.38 CPU TIME 0.7967E+01

RUN BEGINS AT 26/04/98 12.39.49

ANALYSIS AND NONPARAMETRIC INFERENCE(ANI-98)

Bootstrapisation of Bayes risk, classification rates and fraction estimates. Obtaining of fraction errors.

JOB MODE - BUTSTRAP

MEAN OF ONE-LEAVE-OUT-TEST

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.804 0.804 0.801 0.797 0.792 0.789 0.773 0.801 0.796
1 > 2 0.196 0.196 0.199 0.203 0.208 0.211 0.227 0.199 0.204

2 > 1 0.183 0.182 0.180 0.174 0.171 0.167 0.157 0.179 0.173
2 > 2 0.817 0.818 0.820 0.826 0.829 0.833 0.843 0.821 0.827

The same as in "ONE-LEAVE-... mode, but averaged over B bootstrap replicas, see section (3.3.7).

$$R_*^e$$

BOOTSTRAPIZATION OF CLASSIFICATION

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.309 0.310 0.305 0.301 0.299 0.297 0.286 0.305 0.301
1 > 2 0.690 0.690 0.695 0.699 0.701 0.703 0.713 0.695 0.699

The same as in "CLASSIFICATION" mode, but averaged over B bootstrap replicas, see section (3.3.7).

$$\hat{P}_*^e$$

VARIANCE OF ONE-LEAVE-OUT-TEST

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 >1 0.019 0.020 0.025 0.029 0.033 0.036 0.053 0.025 0.029
1> 2 0.019 0.020 0.025 0.029 0.033 0.036 0.053 0.025 0.029

2 >1 0.013 0.015 0.019 0.025 0.028 0.032 0.047 0.020 0.025
2 >2 0.013 0.015 0.019 0.025 0.028 0.032 0.047 0.020 0.025

The m.s.d. of empirical risk estimates.

CLASSIFICATION VARIANCE

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.010 0.010 0.015 0.020 0.022 0.025 0.042 0.016 0.020
1 > 2 0.010 0.010 0.015 0.020 0.022 0.025 0.042 0.016 0.020

The m.s.e. of proportion estimates, see section

BOOTSTRAP AVERAGE OF RECONSTRUCTED PROPORTION

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.2040.2060.2010.2040.2060.2090.2100.2030.205
1 > 2 0.7960.7940.7980.7960.7940.7910.7900.7970.795

Reconstruction of fraction for each bootstrap replica then averaging see section (3.3.6).

$$\hat{P}_*$$

SDE OF RECONSTRACTED PROPORTIONS

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.0110.0120.0120.0130.0150.0180.0130.0120.014
1 > 2 0.0110.0120.0120.0130.0150.0180.0130.0120.014

The m.s.e. of reconstructed proportions

PROPORTIONS (AVERAGED RISKS AND CLASSIFICATIONS)

class 0.90 1.00 1.20 1.40 1.50 1.60 2.00 *summ ada*

1 > 1 0.2040.2060.2020.2040.2070.2100.2100.2030.205
1 > 2 0.7960.7940.7980.7960.7930.7900.7900.7970.795

Reconstructed fractions with bootstrap mean classification and risks.

RUN FINISHS AT 26/04/98 12.45.38 CPU TIME 0.7967E+04

BIBLIOGRAPHY

[1] R. Schlaifer, Probability and Statistics for Buseness Decisions, Mc.Graw-Hill, New York (1959).

[2] H. Haifa. H.Schlaifer. Applied Statistical Decision Theory, Harvard Univ., Boston (1961).

[3] J.O. Berger, The Robust Bayesian Viewpoint, in Robustness of Bayesian analyses, ed. Kadane J.B., Elsevier Science Publishers, (1984).

[4] W. Edwards, H.Lindman, L.J.Savage, Bayesian Statistical Inference, in Robustness of Bayesian analyses, ed. Kadane J.B., Elsevier Science Publishers, (1984).

[5] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Har- court Brace Jovanovich Publishers, (1990).

[6] J.N. Friedman, Data analysis techniques for high-energy physics, CERN Yellow Report, (1974).

[7] P.J. Diggle and R.J.Gratton, Monte Carlo methods for implicit statistical models, J.R.Statist. Soc.B, v. 46 (1984), 193.

[8] J.D. Hey, An Introduction to Bayesian Statistical Inference, Martin Robertson, (1983).

[9] P. Desesquelles Multivariate analysis in Nuclear Physics, Ann.Phys. FR., v. 20, (1995), 1.

[10] E.A.Thompson, Monte Carlo Likelihood in Genetic Mapping, Stat. Science., v.9, (1994), 355.

[11] A.A.Chilingarian, The review of data analysis methods for ANI experiment, Voprosi Atomnoj NAuki i Tekhniki, ser. tech. fiz. exp., v. 2 (8), Kharkov, (1981), 59.

[12] A.A.Chilingarian, The development of of Statistical methods in Cosmic Ray physics, Proc. 18 ICRC, v. 5, Bangalore, (1983), 524.

[13] N.Z.Akopov, Sv.Kh. Arutunian, A.A.Chilingarian, S.Kh.Galfayan, V.Kh.Matevosyan, M.Z.Zazyan, The design principle and structure of ANI data center, Preprint EPI 819(46), 1985.

[14] V.V. Avakyan, A.A.Chilingarian. et al, Bayesian identification of cosmic ray flux hadrons with TRD detector of PION installation, Preprint EPI 933(84), 1986.

[15] A.A.Chilingarian, Statistical decisions under nonparametric a prior information, Computer PhysicsCommunications, v. 54, (1989), 381.

[16] A.A.Chilingarian, H.Z.Zazyan, A bootstrep method of distribution mixture proportion determination Pattern Recognition Letters v.11, (1990), 781.

[17] A.A.Chilingarian, Development of data processing methods in HEP, from a Data Base to a Knowledge Base, Preprint EPI 1327 (22), 1991.

[18] A.A.Chilingarian, Statistical inference in cosmic ray physics, combined analysis of simulated and experimental data, Proc. 22 ICRC v.3, Dublin, (1991), 34.

[19] A.A.Chilingarian, Zazyan H.Z., On the possibility of investigation of the mass composition and energy spectra of PCR in the energy range 10^{15} — 10^{17} eV using EAS data, IL Nuovo Cimento v. 14C (6), (1991), 555.

[20] A.A.Chilingarian, H.Z. Zazyan, Experiments with particle bundels in cosmic ray physics. Determination of strong interaction parameters by a pattern recognition method, J. of Nuclear, phys. (russion) 54, (1991), 128.

[21] A.Chilingarian, S.Ter-Antonyan, A.Vardanyan, The comparison of Bayesian and Neural techniques in problem of classification to multiple categories, Xl.M. v. NIMA 1063, (1997) 230.

[22] G.Schatz , W.D.Apel, K.Bekk, E.Bollmann, H.Bozdog, I.M.Bancus, M.Brendle, J.N.Capdevielle, A.Chilingarian, K.Daumiller, P.Doll, J.Engler, M.Foeller, P.Gabriel, H.J.Gils, R.Glasstetter, A.Haungs, D.Heck, J.R.Hoerandel, K.H.Kampert, H.Keim, J.Kempa, H.O.Klages, J.Knapp, H.J.Mathes, H.J.Mayer, H.H.Mielke, D.Muehlenberg, J.Oehlschlaeger, M.Petcu, U.Raidt, H.Rebel, M.Roth, H.Schieler, G.Schmalz, H.J.Simonis, T.Thorny. J.Unger, B.Vulpescu, G.J.Wagner, J.Wdowczyk, J.H.Weber, J.Wentz, Y. Wetzel, T.Wibig, T.Wiegert, D.Wochele, J.Wochele, D.Wochele, J.Zabierowski, S.Zagromski, B.Zeitniz,
The KASCADE Experiment, Nucl.Phys.B (Proc.Suppl.) v. 60B, (1998) 151.

[23] A.A.Chilingarian, A.M.Dunaevski, et al, Multivariate analysis of Roentgen-Emulsion chamber data, 19 ICRC, v. 5, San-Diego, (1985), 392.

[24] S.Kh.Galfayan, A.A.Chilingarian, A.M.Dunaevski, M.Z.Zazyan, et al, Multiple comparisons of EAS and Roentgen-Emulsion chamber data with simulations, Izv. AN.SSSR., ser. fiz., 50(11) (1986), 2146.

[25] A.A.Chilingarian, S.Kh.Galfayan, M.Z.Zazyan, A.M.Dunaevski, et al, The upper boundary of of iron nuclei fraction in PCR obtained from PAMIR data, 20 ICRC, v.l, Moscow, (1987), 386.

[26] A.A.Chilingarian, S.Kh.Galfayan, M.Z.Zazyan, A.M.Dunaevski, et al, The new method of gamma-families analysis, 20 ICRC, v. 5, Moscow, (1987), 312.
[27] S.Kh.Galfayan, A.A.Chilingarian, A.M. Dunaevski, M.Z. Zazyan, et al, The iron nuclei in primary cosmic rays by gamma-families data, Izv. AN.SSSR., ser.(i/.. v. 53, (1989), 280.

[28] A.A.Chilingarian, S.Kh.Galfayan et al, Upper boundary of iron nucleiin primary cosmic rays at $E > 10^{16}$eF Lebedev Institute preprint 75, 1988.

[29] G.Vacanti, M.F.Cawley, et. al., Gamma-ray observations of the Crab Nebula at TeV energies, Astroph. J. 377, (1991), 467

[30] M.Punch, C.W.Akerlof, M.F.Cawley et.al., Supercuts: an improved method of selecting gamma-rays, Proc. 22 ICR C v.l, Dublin, (1991), 464.

[31] F.A.Aharonian, A.A.Chilingaryan, A.K.Plyasheshnikov, A.K.Konopelko, On the possibility for a higher efficiency of discrimination of gamma-rays from point sources by the pattern recognition method, Preprint YerPhI 1171(48), 1989.

[32] F.A.Aharonian, A.A.Chilingaryan, A.K.Plyasheshnikov, A.K.Konopelko, Analysis of the possibilities of suppression of the cosmic-ray background when detecting very high energy cosmic gamma-quanta by means of a system of Cherenkov gamma-telescopes with multichannel light receivers, Preprint YerPhI 1277(60), 1990.

[33] F.A.Aharonian, A.A.Chilingaryan, A.K.Plyasheshnikov, A.K.Konopelko, On the possibility of an improvement of background hadronic showers discrimination against gamma-ray coming from a discrete source by a multidimensional Cherenkov light analysis, 21 ICRC, v. 4, Adelaide, (1990), 246.

[34] A.A.Chilingarian, M.F.Cawley, Multivariate analysis of Crab Nebula Data, Wipple Collaboration internal report, Wipple Collaboration internal report, 5 July, 1990.

[35] F.A. Aharonyan, A.A.Chilingarian, et al, A multidimensional analysis of the Cherenkov images of air showers induced by very high energy gamma-rays and protons, NIM, v. A-302, (1991), 522.

[36] F.A.Aharonian, A.A.Chilingarian, A.K.Plyasheshnikov, A.K.Konopelko, Use of multi-dimensional analysis for classification of events registered by the system of Cherenkov gamma

telescopes with multichannel light receivers, Izv. Akademii Nauk, USSR, Phys.(russion) v.55, (1991), 734.

[37] A.A.Chilingarian, M.F.Cawley, Application of multivariate analysis to atmospheric Cherenkov imaging data from the Crab Nebula, Proc. 22 ICRC v.l, Dublin, (1991), 460.

[38] A.A Chilingarian, Neural Net Classification of the gamma and proton images registered with atmospheric Cherenkov technique, random search learning in feed-forward networks, Proc. 22 ICRC v.l, Dublin, (1991), 540.

[39] A.A.Chilingarian, A.K.Konopelko, A.V.Plyasheshnikov, New algorithms for gamma- quanta energy estimation by the telescopes with Cherenkov light imaging facilities, Proc. 22 ICRC v.l, Dublin, (1991), 480.

[40] A.A.Chilingarian, On the methods of the enhancement of the reliability of the signal detection with Cherenkov Atmospheric techniques, Izv. AN (ser. phys. (in russian) v.57, 186.

[41] F.A.Aharonian, A.A.Chilingaryan, R.G.Mirzoyan, A.K.Plyasheshnikov, A.K.Konopelko, The system of imaging atmospheric Cherenkov telescopes: the new prospects for VHE gamma ray astronomy, Exp. Astr. v. 2, (1993), 331.

[42] A.A.Chilingarian, M.FCawley, Optimizing the non-linear gamma-ray domain in VHE gamma-ray astronomy using neural-network classifier, Proc. 24 ICRC, v.3, Rome, (1994) , 742.

[43] A.A.Chilingarian, Neural classification technique for background rejection in high energy physics experiments, Neurocomputing, v. 6 , (1994), 497.

[44] A.A.Chilingarian, Detection of weak signals against background using neural network classifiers, Pattern Recognition Letters, v. 16, (1995), 333.

[45] A.A.Chilingarian. E.H.Sevinian, S.A.Chilingarian, The non-linear signal domain selection using a new quality function in neural net training, XI.M. v. A 389, (1997), 242.

[46] A.Chilingarian, M.,Halpaap, H.J.Gils, H.Rebel, A Comparative study of EAS energy estimation methods, Proc. 24 ICRC, v.l, Rome, (1995), 391.

[47] H.Rebel, G.Volker, M.Foller, A.A.Chilingarian, Arrival time distributions from extensive air showers as signature of the mass composition of cosmic rays, J.Phys. v. G: 21, (1995), 451.

[48] A.A.Chilingarian, S.Ter-Antonyan, A.Vardanyan, M.Roth. J.Knapp, H.J.Gils, H.Rebel, On the nonparametric classification and regression methods for the multivariate EAS data analysis, Nuclear. Phys. B, v. 52B, (1997), 237.

[49] H.Rebel,M.Roth, J.Knapp, H.J.Gils,A.Chilingarian, S.Ter-Antonyan, A.Vardanyan, On the accuracy of the elemental composition determination on the mountain altitudes and sea level, Nuclear Phys. B., v. 52B, (1997), 240.

[50] I.M.Brancus, B.Vulpesku, H.Rebel, M.Duma, A.A.Chilingarian, Correlated features of arrival time and angle-of-incidence distributions of EAS muons, Astroparticle Physics, v. 7, (1997), 343.

[51] A.A.Chilingarian for KASCADE Collaboration, S.Ter-Antonyan, A.Vardanyan, How to infer the mass composition from EAS observations demonstarated with KASCADE data, 25 ICRC, v. 4, Durban, (1997), 105.

[52] M.Roth for KASCADE collaboration, S.Ter-Antonyan, A.Vardanyan, How to infer the primary energy spectrum from EAS observations demonstarated with KASCADE data, 25 ICRC, v. 4, Durban, (1997), 157.

[53] A.A.Chilingarian, S.Ter-Antonyan, A.Vardanyan, M.Roth, H.J.Gils, J.Knapp, H.Rebel, Energy Spectra and Elemental Composition Determination on Mountain Altitudes and Sea Level, Nucl. Phys. B. v. 60B (1988) 117.

[54] Capdevielle J.N. et al., 1992, KfK Report 4998, Kernforschungszentrum Karlsruhe. J.Knapp and D.Heck, 1993, KfK Report 5196B, Kernforschungszentrum Karlsruhe. D.Heck, J.Knapp, J.N.Capdevielle, G.Schatz, T.Thouw, 1998, FZKA Report 6019, Forschungszentrum Karlsruhe.

[55] J.Knapp, 1997, Rapporteur talk, Proc. 25 th ICRC, Durban , 1997 (in publication).

[56] S.G,Bayburina et al, Nucl.Phys. B v. 391, (1981)1.

[57] J.R.Ren, Phys. Rev. D v. 38, (1988), 1404.

[58] L.T.Baradzei et al, Nucl.Phys. v. B 370, (1992)365.

[59] T.K.Gaiser et al, Phys.Rev. v.D 47, (1993), 1919.

[60] N.Hayashida et al, Proc. 25 th ICRC, v. 6, (1997), 241.

[61] GEANT, 1993, CERN program library, CERN.

[62] A.A. Watson, Charged Cosmic Rays above 1 TeV, Rapporteur talk: 25 ICRC, Durban (1997).

[63] Jaynes E.T., in Foundations of Probability Theory, Statistics Inference and Statiustical Theories of Science, ed. W.L.Harper, C.A. Hooker, Dordrecht, Rewidel.

[64] W.H.Press, S.A.Teukolsky, et al., in Numerical Recipes in Fortran, chapter 15.7. Cambridge Univ. Press.

[65] A Source Code Management System, user's guide and reference manual, version 1.46, CERN 1994.

[66] E.Lederman, Handbook of Applied Mathematics, Statistics, John Wiley and Sons, New-York, (1984).

[67] P.Hajek ,T.Havranek, Mechanizing Hypothesis Formation, Springer Verlag, Heidelberg, (1979).

[68] G.E.P.Box, The importance of practice in the development of statistics, Technometrics, 26 (1984), 1.

[69] E.A.Eadie, D.Drijard, F.E.James, M.Ross and B.Sadoulet, Statistical methods in experimental physics, North-Holland, Amsterdam, (1971).

[70] S.Zacks, The theory of statistical inference, John Wiley and Sons,New-York, (1977).

[71] D.V.Lindley, Bayesian statistics, Soc.for indust. and appl. math.,Philadelphia, (1978).

[72] G.T.Toussaint , Bibliography of misclassification, IEEE trans, on Information v.IT-20 (1974), 472.

[73] S.M.Snappin, J.D.Knoke, Classification error rate estimators evaluated by unconditional mean squared error, Technometrics, v.26 (1984), 371.

[74] B.Efron, Nonparametric standart errors and confidence intervals, Canadian J. Statist., v. 9 (1981), 139.

[75] L.Devroye, L.Gyorfi, Nonparametric density estimation. The LI view, Jown Wiley and Sons, New-York, (1985).

[76] L.Devroye, Universal smothing factor selection in density estimation: theory and practice, Technical report, McGill Univ., (1997).

[77] M.Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Stat., v. 27, (1957), 832.

[78] E.Parzen, On estimation of a probability density function and mode, Ann. Math. Stat., v. 33, (1962), 1065.

[79] E.Fix, J.L.Hodges, Discriminatory analysis. Nonparametric discrimination, Consistency Properties Project 21-49-004, Report l.l'SAF School of Aviation Medicine, Ran- dolf Field, Texes, (1951).

[80] D.O.Lofsgaarden and C.D.Quesenberry, A nonparametric estimate of a multivariate density function, Ann. Math. Stat., v. 36, (1965), 1049.

[81] P.C.Mahalonobis, On the generalized distance in statistics, National Inst.of India, v. 2, (1936), 49.

[82] R.A.Tapia,J.R.Thompson, Nonparametric probability density estimation, The John Hopkins University Press, Baltimore and London, (1978).

[83] K.Fukunaga, D.Himmels, Bayes error estimation using Parzen and KNN procedures, v. PAMI9, (1987), 634.

[84] L.R.Rabiner, E.Levinson, A.E.Rozenberg and J.G.Wilpon, Speaker - independent recognition of isolated words using clustering techniques, IEEE trans, on Acoustics, Speech, Signal Processing, v. ASSP-27, (1974), 336.

[85] A.A.Chilingarian and S.Kh.Galfayan, Calculation of Bayes risk by KNN method, Stat. Problems of Control,Vilnius, v. 66, (1984), 66.

[86] Deutsche Bundesbank, 10 DM banknote, 1949-2001.

[87] B.Efron, Bootstrap methods, another look at the jacknife, Ann. Statist., v. 7, (1979), 1.

[88] A.K.Jain, R.C.Dubes and C-C.Chen, Bootstrap Techniques for error estimation, IEEE Trans., v. PAMI-9, (1987), 628.

[89] P.J.Bickel, D.A.Freedman, Some Asymptotic Theory for the Bootstrap, Ann. Stat. v.9, (1981), 1198.

[90] G.A.Young, Bootstrap: more than a stab in the dark? Stat. Science, v. 9, (1994), 382.

[91] D.W.Ruck, K.S. Rogers et al, The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function, IEEE Trans, on Neural Networks v.l, (1990), 296.

[92] S.N. Zhang, D.Ramsden, Statistical data analysis for g-ray astronomy, Exp. Astronomy v.l,(1990), 158.

[93] C.Paladin, A.Vulpiani, Anomalous Scaling Laws in Multifractal Objects, Phys.Rep., v.156.No.4, (1987)

[94] K.Pawelzik, U.S.Sinister. Generalized dimensions an entropies from a measured time series, Phys.Rev.A, v.35, (1987) 481.

[95] J.G.Caputo, P.Atten, Metric entropy : an experimental means for characterizing and quantifying chaos, Phys.Rev.A, v. 35, (1987) 1311.

[96] K.W.Pettis, T.A.Baily, A.K.Tain, R.C.Dubes, An Dimensionality Estimator from Nearest Neighbour Information. IEEE Trans, on Pattern Anal, and Machine Intelligence, v. PAMI1, (1979), 25.

APPENDIX A *B.IN* INPUT FILE EXAMPLE

3
2,10
mhgg-light.dat
mhgg-heavy.dat
../../ARO/aroexp.txt
mhgg-heavy-e0.dat
mhgg-heavy-e0.hbook
8
0,0,0,0,0
400,400,10,10,10
0,2000000
non
LEARNING
SEQUENTAL
PARZ
SQUARE
(12f10.4)
1
0.5,0.5,0.2,0.2,0.2
RECONSTRUCT
7
0.01,0.03,0.05,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.9,1.
900000000.,0.000000000000001
7
2
3,4,5,6,8
2
-1000,-1000,-1000,-1000,-1000,-1000,-1000,-1000
1000,1000,1000,1000,1000,1000,1000,1000
pseudo
norenorm
1,1,1024,11
0,20,1024,22
1,2,3,4,5
3,2,5,1 - NN configuration, number of layers and number of neutrons in each layer : number of neutrons in the first layers equals number of variables
500000,10.,9.,1.,644 Number of iterations, radnom step, sigma criteria and shift of random number generator
1.,1.
0.53,0.47
neuron
montec
msd
memory
better
0.000001
0.51
5
0.01,1.,0.05,1.,0.1, 1.,0.2,1.,0.3,1.,0.5,1.
0,1.
200000
1
Ne
S