PROCEEDINGS OF SCIENCE

GARCON - Genetic Algorithm for Rectangular Cuts OptimizatioN

A. Drozdetskiy[*]
University of Florida, Gainesville, USA
E-mail: Alexey.Drozdetskiy@cern.ch

S. Abdullin[†]
Fermilab, IL, USA
E-mail: Salavat.Abdoulline@cern.ch

We will present Genetic Algorithm for Rectangular Cuts OptimizatioN (GARCON) program and demonstrate its functionality on a simple HEP analysis example. The program automatically performs rectangular cuts optimization and verification for stability in a multi-dimensional cuts phase space. The program has been successfully used by a number of different analyses presented in the Compact Muon Solenoid (CMS collaboration) Physics Technical Design Report (Large Hadron Collider (LHC), CERN, Geneva, Switzerland), corresponding results are also published in a number of papers in 2006.

XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research
April 23-27 2007
Amsterdam, the Netherlands

[*]Speaker.
[†]On leave of absence from IHEP, Moscow, Russia

PoS(ACAT)052

1. Introduction

Typically HEP analysis has quite a few selection criteria (cuts) to optimize for example a significance of the "signal" over "background" events: transverse energy/momenta cuts, missing transverse energy, angular correlations, isolation and impact parameters, etc. In such cases simple scan over multi-dimensional cuts space (especially when done on top of a scan over theoretical predictions parameters space like for SUSY e.g.) leads to CPU time demand varying from days to many years... One of the alternative methods, which solves the issue is to employ a Genetic Algorithm (GA), see e.g. [1, 2, 3].

We wrote a code, GARCON [4], which automatically performs an optimization and results stability verification effectively trying $\sim 10^{50}$ cut set parameters/values permutations for millions of input events in hours time. Examples of analyses are presented in the CMS Physics TDR [5] and recent papers [6, 7, 8, 9].

The GARCON program among many other features allows user:

- to select an optimization function among known significance estimators, as well as to define user's own formula, which may be as simple as signal to background ratio, or a complicated one including different systematic uncertainties separately on different signal and background processes, different weights per event and so on;
- to define a precision of the optimization;
- to restrict the optimization using different kind of requirements, such us minimum number of signal/background events to survive after final cuts, variables/processes to be used for a particular optimization run, number of optimizations inside one run to ensure that optimization converges/finds not just a local maximum(s), but a global one as well (in case of a complicated phase space);
- to automatically verify results stability.

GARCON, GA-based programs in general exploit evolution-kind algorithms and uses evolution-like terms:

- Individual is a set of qualities, which are to be optimized in a particular environment or set of requirements. In HEP analysis case an individual is usually a set of lower and upper rectangular cut values for each of variables under study/optimization.
- Environment or set of requirements of evolutionary process in HEP analysis case is a Quality Function (QF) used for optimization of individuals. The better QF value the better is an individual. Quality Function may be as simple as $S/\sqrt{B}$, where S is a number of signal events and B is a total number of background events after cuts, or almost of any degree of complexity, including systematic uncertainties on different backgrounds, etc.
- A given number of individuals constitute a Community, which is involved in evolution process.

- Each individual involved in the evolution: breeding with possibility of mutation of new individuals, death, etc. The higher is the QF of a particular individual, the more chances this individual has to participate in breeding of new individuals and the longer it lives (participates in more breeding cycles, etc.), thus improving community as a whole.

- Breeding in HEP analysis example is a producing of a new individual with qualities (set of min/max cut values) taken in a defined way from two "parent" individuals.

- Death of an individual happens, when it passes over an age limit for it's quality: the bigger it's quality, the more it lives.

- Cataclysmic Updates may happen in evolution after a long period of stagnation in evolution, at this time the whole community gets renewed and gets another chance to evolve to even better quality level. In HEP analysis case it corresponds to a chance to find another local and ultimately a global maximum in terms of quality function. Obviously, the more complicated phase space of cut variables is used the more chances exist that there are several local maximums in quality function optimization.

- There are some other algorithms involved into GAs. For example mutation of a new individual. In this case newly "born" individual has not just qualities of its "parents", but also some variations, which in terms of HEP analysis example helps evolution to find a global maximum, with less chances to fall into a local one. There are also random creation mechanisms serving the same purpose.

There is nothing special involved in GARCON input preparation. One would need to prepare a set of arrays for each background and a signal process of cut variable values for optimization. Similar to what is needed to have to perform a classical eye-balling cut optimization.

In comparison to other automatized optimization methods GARCON output is transparent to user: it just says what rectangular cut values are optimal and recommended in an analysis. Interpretation of these cut values is absolutely the same as with eye-balling cuts when one selects a set of rectangular cut values for each variable in a "classical" way by eye.

2. LM6 with PYTHIA: a Toy Study

We are working in the framework of mSUGRA model [10] which is derived from more general MSSM [11] model using constrains inspired by the super-gravity unification. In case of mSUGRA, the number of independent MSSM parameters is reduced to just five. For our illustration we selected a point in mSUGRA parameter space with the following values of mSUGRA parameters:

- the universal gaugino mass $m_{1/2} = 400$ GeV,

- the scalar mass $m_0 = 85$ GeV,

- the trilinear soft supersymmetry-breaking parameter $A_0 = 0$,

- the ratio of Higgs vacuum expectation values, $\tan\beta = 10$,

- sign of Higgsino mixing parameter, $sign(\mu) > 0$.

Characteristic qualities of SUSY events, following from a consideration of signal Feynman diagrams are: large MET (mainly due to massive stable SUSY particles, LSP) and large jet E_Ts (due to heavy SUSY particles cascade decays).

Background processes considered in this study are QCD, W/Z+jets, double weak-boson production and $t\bar{t}$.

The main generation tool is PYTHIA 6.227 [12]. In addition, ISASUGRA, part of ISAJET 7.69 [13] is linked to PYTHIA to provide mSUGRA masses, couplings and branchings for the signal simulation.

All simulations and analysis is done for an integrated luminosity of $10\,fb^{-1}$. More details on the analysis can be found elsewhere [4].

2.1 Variables and preselection

Several variables characterizing the event were stored in the GARCON input files:

- number of muons (N_μ),
- the highest muon p_T (p_T^1),
- isolation parameter for the highest p_T muon[1] ($ISOL_\mu^1$),
- number of jets with $p_T > 40$ GeV (N_j),
- E_T of the highest jet E_T (E_T^1),
- E_T of the third highest jet (E_T^3),
- missing transverse energy (E_T^{miss}),
- azimuthal angle between the highest-p_T muon and E_T^{miss} (if any) ($\Delta\phi(\mu^1, E_T^{miss})$),
- azimuthal angle between the highest-E_T jet and E_T^{miss} ($\Delta\phi(jet^1, E_T^{miss})$),
- circularity - $Circ = 2 \cdot \min(\lambda_1, \lambda_2) \,/\, (\lambda_1 + \lambda_2)$, where λ_1, λ_2 are eigenvalues of a simple matrix $C^{\alpha,\beta} = \Sigma E_i^\alpha E_i^\beta$, where Σ means sum over energies of all objects (leptons, jets, missing energy) and α, β = 1,2 correspond to x and y components. In case of back-to-back di-jets *Circ* is close to 0, while in case of multi-jet topology *Circ* tend to be closer to 1.

Jets are reconstructed using a cone algorithm with merging-splitting of overlapping clusters. In order to reduce the number of events in the data files, a minimal E_T^{miss} cut of 50 GeV is applied at generator level. Another pre-selections include the requirement to have at least two jets above 40 GeV in every event and a cut on the leading jet E_T in the event to be above 200 GeV. The latter results from the fact that it does not look possible to simulate an appropriate number of QCD events with $\hat{p}_T < 200$ GeV/c

[1] ISOL = $\sum p_T^i$ (p_T with respect to the beam direction) should be less or equal to 0, 0, 1, 2 GeV for the four muons when the muons are sorted by the ISOL parameter. The sum runs over only charged particle tracks with p_T greater then 0.8 GeV and inside a cone of radius $R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} = 0.3$ in the azimuth-pseudorapidity space. A p_T threshold of 0.8 GeV roughly corresponds to the p_T for which tracks start looping inside the CMS Tracker. Muon tracks are not included in the calculation of the ISOL parameter

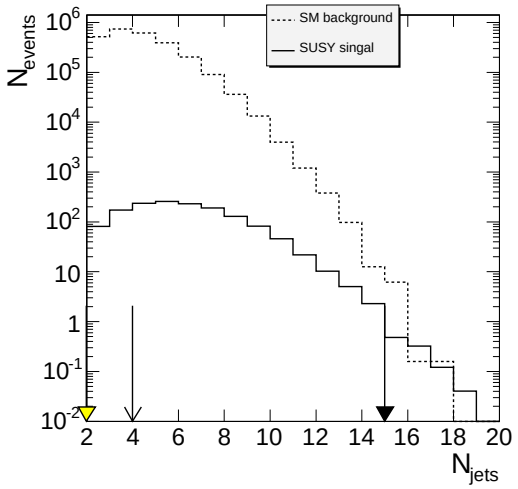

Figure 1: Number of jets with $E_T > 40$ GeV. Solid lines denote the SUSY signal, while dashed lines - the sum of the SM background distributions. Empty arrow is for classical analysis cut choice, filled colored arrows (black and gray/yellow) are GARCON optimized cuts (values for verification step).

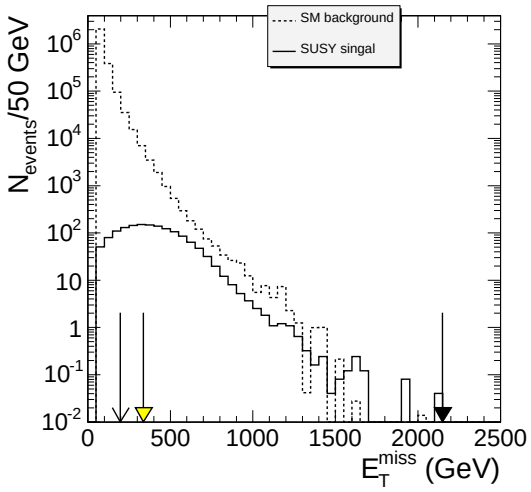

Figure 2: Distribution of missing transverse energy. The same notations as for Fig.1.

2.2 Significance estimator

The S_{c12} significance estimator [14] was used for optimization: $S_{c12} = 2 \cdot (\sqrt{B+S} - \sqrt{B})$, where B - is a number of all the background events after cuts, and S - is a number of signal events after cuts. Results are presented also in terms of $S_{cL} = \sqrt{2 \cdot (S+B) \cdot log(1.0+S/B) - 2 \cdot S}$ which follows true Poisson probability for small number of events better than S_{c12}, is shown in Ref. [6].

2.3 Splitting statistics in two parts

We divided statistics in two parts: to perform cuts optimization on one of them and then to verify stability of results on the other. It's especially important for the analyses with limited statistics: in such cases one risks to optimize cuts around a statistical fluke of a signal over backgrounds significance. "Blind experiment" verification approach allows to exclude such unstable cases.

3. Classical Search

3.1 Distributions and eye-balling search for cuts

Figures 1 – 6 show some of the simulated data distributions which are used in the current analysis. Solid lines denote the SUSY signal, while dashed lines - the sum of the SM background distributions.

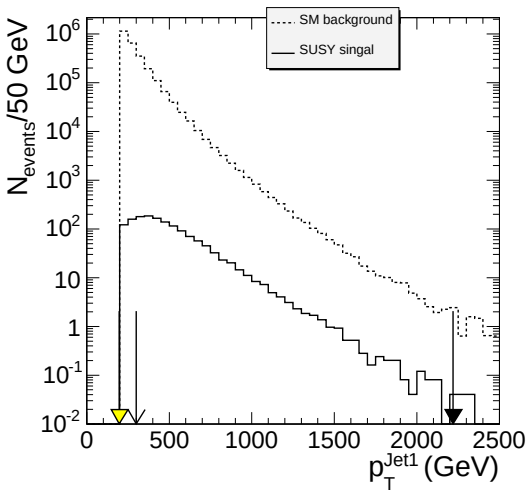

Figure 3: Transverse energy of the hardest-E_T jet. The same notations as for Fig.1.

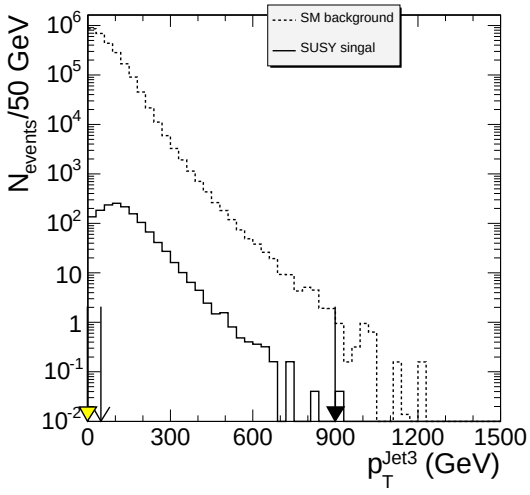

Figure 4: Transverse energy of the third-hardest-E_T jet. The same notations as for Fig.1.

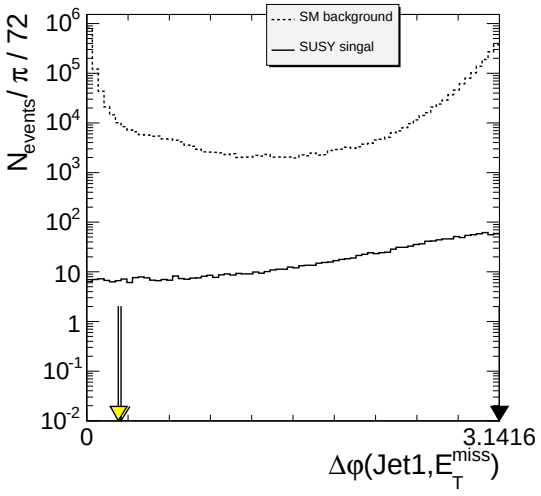

Figure 5: Azimuthal angle distance between leading jet and transverse missing energy. The same notations as for Fig.1.

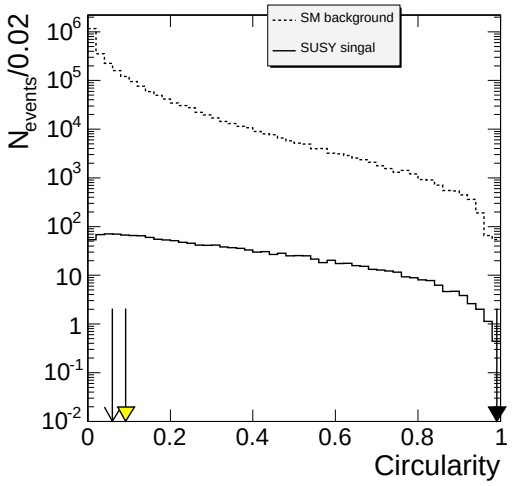

Figure 6: Distribution of the circularity. The same notations as for Fig.1.

4. GARCON Analysis

GARCON uses the same input information as a classical analysis: arrays of variable values, the same what is needed to perform a classical eye-balling cut optimization.

Details on chosen variables are given in Sec. 2.1.

4.1 Optimization

Each cycle/"year" of evolution includes a community update, that is breeding process, possible mutation of new individuals, quality and age calculations for each individuals, death of worst and

too old individuals, etc.

As described above the better is an individual QF, the longer it lives and hence the more chances it has to produce new individs, improving quality of a community as a whole and the very best individ quality as a final goal. This very best individ or the very best set of min/max cut variable values, which corresponds to the best achievable quality function (significance of signal over background) is a final goal and final output of the GARCON optimization step: rectangular cut values recommended by the optimization procedure.

Figures 7, 8, 9 and 10 show dynamics/evolution of the S_{c12} quality function, dynamics on MET and circularity cut variable values and amount of time used for optimization.

Typical optimization procedure with GARCON takes from a few seconds to several hours depending on the amount of statistics and additional requirements like minimal number of events to survive after all cuts, etc. As one can see from Fig. 10 results close to the best are already achieved before the first cataclysmic update, which happened at year < 50 and required less than 3.5 hours of CPU time for 10 variables (Sec. 2.1) or 20 optimized parameters with precision on each 2.5% and about $4 \cdot 10^5$ generated events on input (after pre-selection, see Sec. 2.1).

Optimized values for all the cut parameters are listed in Table 1. Results in terms of chosen significance estimator as well as signal to background number of events ratio, final event numbers are listed in Tab. 2. Cuts are also illustrated on cut parameter distribution in Figs. 1-6.

Table 1: Min and max values for cut parameters. Cut values for the classical analysis are the same. Cut values for GARCON verification are rounded off in comparison to those we have from optimization to reflect resolution effects and possible lower/upper limits.

cut parameter	classical	GARCON optimization	GARCON verification
N_{mu}	0-inf	0-5	0-5
p_{T}^1, GeV	0-inf	0-1020	0-inf
$ISOL_{mu}^1$, GeV	0-inf	0-1080	0-inf
N_j	4-16	2-16	2-16
E_{T}^1, GeV	300-inf	200-2220	200-inf
E_{T}^3, GeV	50-inf	0-901	0-inf
$\mathrm{E}_{\mathrm{T}}^{miss}$, GeV	200-inf	342-2150	340-inf
$\Delta\phi(\mu^1, E_{\mathrm{T}}^{miss})$, rad	0- π	0.297-π	0.297-π
$\Delta\phi(jet^1, E_{\mathrm{T}}^{miss})$, rad	0.262-π	0.245-π	0.245-π
circularity	0.06-1	0.0924-0.993	0.0924-1

Analyzing cut values and their distributions (Figs. 1-6) one can see that some variables after GARCON optimization converge to the limits of a particular distribution. From the technical point of view the reason for this is because GARCON works only with input values and doesn't have plus or minus infinity e.g. at its disposal. From the practical point of view, it means that min or max cut on a particular variable or the whole variable is not useful in comparison to other variables in terms of improving signal to background significance and GARCON shows it. As an example we can consider E_{T}^1 and E_{T}^3 before and after all cuts (except the cut on E_{T}^1 or E_{T}^3 correspondingly), the examples of variables for which GARCON and classical cut values are different: compare Figs. 3

and 4 for distributions before and Figs. 11 and 12 - after the cuts applied.

4.2 Verification

As mentioned earlier, the available MC statistics was divided in two parts. The second part is used for a "blind" analysis or results stability verification.

After we got the cut values from optimization step, we round them off to the level of expected precision [2] for each parameter (see Tab. 1) and apply them to the second half of the statistics.

Results are shown in Tab. 2. One can see that results are stable[3].

Table 2: Final results comparison for classical and GARCON (for optimization and verification steps) approaches in terms of S_{c12} and S_{cL} significance estimators as well as ratio of final number of signal to total background events and those numbers of events with MC statistical error included.

parameter	classical optimization	classical verification	GARCON optimization	GARCON verification
S_{c12}	8.1	8.0	15.3	14.7
S_{cL}	8.1	8.1	15.8	15.2
S/B	0.102	0.102	0.506	0.469
N_{signal}	665 ± 7	663 ± 7	574 ± 7	567 ± 7
$N_{background}$	6496 ± 160	6503 ± 160	1130 ± 121	1210 ± 121

4.3 Comparison between classical and GARCON approaches

Difference in performance in terms of significance (8 vs. 15) and signal to background events number ratio (0.1 vs. 0.5) may not be a typical gain when GARCON is used vs. a classical approach: classical approach may be pretty sophisticated (as well as time dedicated to it may be large). What is important to emphasize is that GARCON does optimization and verification of results stability in an automatic manner, not requiring any special treatment of either input data or output results and does converge to virtually the best set of cuts in typically hours time.

Different cases which show GARCON usage in much more complicated analyses cases can be found elsewhere [6, 7, 8, 9].

5. Summary

All-in-all GARCON is a simple yet powerful ready-to-use tool with flexible and transparent optimization and verification parameters setup. It is publicly available along with a paper on it [4] consisting of an example case study and user's manual.

[2] Expected precision, which includes detector resolution, of course is different for different parameters (muon p_{T}, jet E_{T}) and different HEP experiments.

[3] NOTE: in case there are zero generated events left after final cuts we use 0 ± 1 generated events, taking slightly pessimistic estimation for MC statistical error and hence corresponding number of expected events: $0 \pm one - generated - event - weight$.

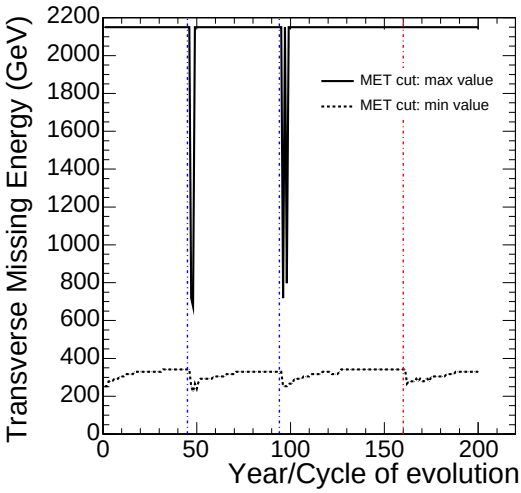

Figure 7: Evolution of the cuts on MET. Upper and lower curves are for min and max cut values on the variable. Vertical dotted-dashed lines show cataclysmic update times, the right one corresponds to a cataclysmic update after the best result achieved.

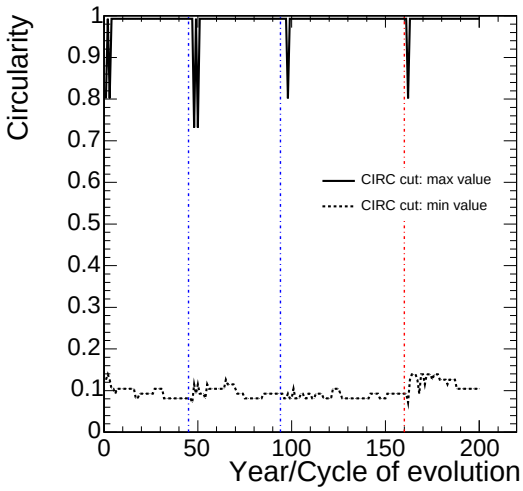

Figure 8: Evolution of the cuts on CIRC. Notations are the same as for Fig.7.

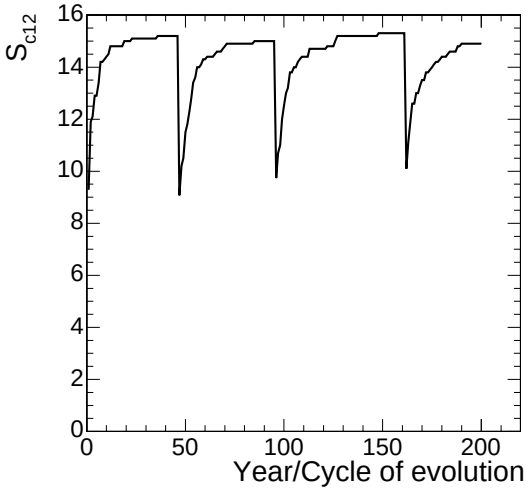

Figure 9: Significance (S_{c12}) estimator value dynamics. Notations are the same as for Fig.7.

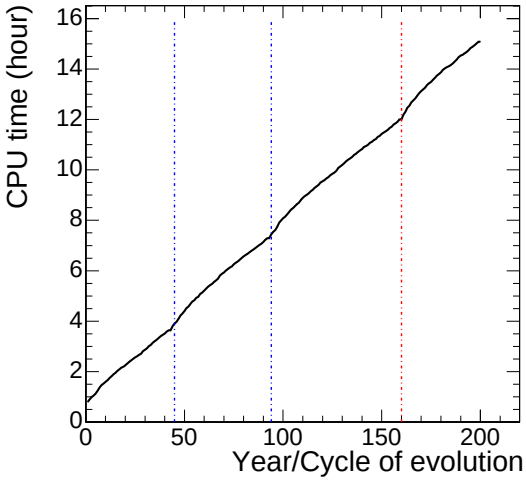

Figure 10: Amount of time spent for evolution. Notations are the same as for Fig7.

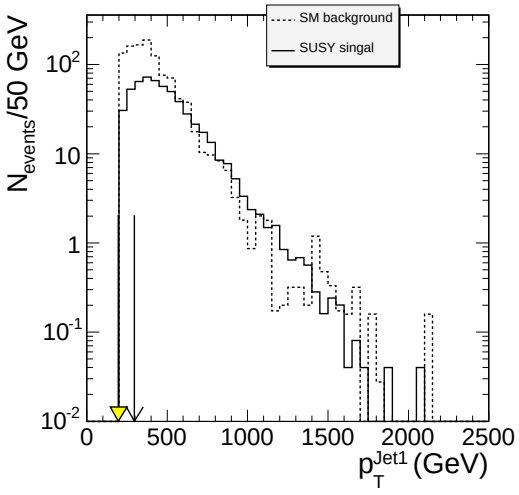

Figure 11: Transverse energy of the hardest-E_T jet. The same notations as for Fig.7.

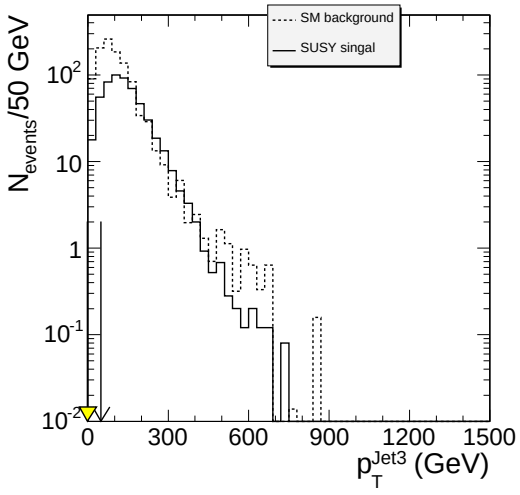

Figure 12: Transverse energy of the third-hardest-E_T jet. The same notations as for Fig.7.

References

[1] John H. Holland. *Adaptation in natural and artificial systems*. The University of Michigan Press, Ann Arbor, 1975.

[2] David E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison Wesley, 1989.

[3] S.Abdullin, *Genetic Algorithm for SUSY Trigger Optimization In CMS Detector At LHC*, NIM A 502 (2003) 693-695.

S.Abdullin,S. Kunori, *Revisited Trigger Optimization for the Low-Mass Scale SUSY*, CMS IN 2003/006

S.Abdullin, *On Optimization Of The Trigger Selection For The Low-Mass Scale SUSY*, CMS IN 2002/036.

S.Abdullin, *Genetic Algorithm for SUSY Trigger Optimization*, talk given at the IV Conference " LHC Days in Split", October 8-12, 2002 http://cmsdoc.cern.ch/∼abdullin/events/talks/Split2002.pdf

[4] S. Abdullin, A. Drozdetskiy et al., *GARCON: Genetic Algorithm for Rectangular Cuts OptimizatioN. User's manual for version 2.0*, hep-ph/0605143, http://drozdets.home.cern.ch/drozdets/home/genetic/

[5] M.Della Negra, A. Petrilli, A. Ball, L. Foa et al., *CMS Physics Technical Design Report, volume II*, J. Phys. G: Nucl. Part. Phys., vol34 (2007), 995-1579.

[6] S. Abdullin et al., *Search Strategy for the Standard Model Higgs Boson in the* $H \to ZZ^{(*)} \to 4\mu$ *Decay Channel using* $M(4\mu)$*-Dependent Cuts*, CMS Note 2006/122.

[7] V. Abramov et al., *Selection of single top events with the CMS detector at LHC*, CMS Note 2006/084.

[8] W.de Boer et al., *Trilepton final state from neutralino-chargino production in mSUGRA*, CMS Note 2006/113.

[9] D. Acosta et al., *Potential to Discover Supersymmetry in Events with Muons, Jets and Missing Energy in pp Collisions at sqrt(s) = 14 TeV with the CMS Detector*, CMS Note 2006/134

[10] For mSUGRA model description - see for example Report of SUGRA Working Group for Run II of the Tevatron and references therein : S.Abel*et al.*, hep-ph/0003154.

[11] For MSSM see, for instance : J. Ellis. S. Kennedy and D. V. Nanopoulos, *Phys. Lett.* **B260** (1991) 131; P. Langacker and M. X. Luo, *Phys. Rev.* **D44**(1991) 817; U. Amaldi, W. De Boer and H. Furstenau, *Phys. Lett.* **B260** (1991) 447; F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, *Nuovo Cimento* **104 A** (1991) 1817.

[12] T. Sjostrand, L. Lonnblad and S. Mrenna, *PYTHIA 6.2 Physics and Manual, report LU-TP-01-21*, Aug 2001, arXiv:hep-ph/0108264.

[13] F. Paige and S. Protopopescu, in *Supercollider Physics*, p. 41, ed. D. Soper (World Scientific, 1986); H. Baer, F. Paige, S. Protopopescu and X. Tata, in *Proceedings of the Workshop on Physics at Current Accelerators and Supercolliders*, ed. J. Hewett, A. White and D. Zeppenfeld (Argonne National Laboratory, 1993).

[14] S.I.Bityukov *et al.*: *"Uncertainties and Discovery Potential in Planned Experiments"*, hep-ph/0204326.