# PhyStat2007

# Report of Contributions

Contribution ID: **1**                                                    Type: **not specified**

# TMVA - Toolkit for Multivariate Data Analysis with ROOT

In high-energy physics, with the search for ever smaller signals in ever larger data sets, it has become essential to extract a maximum of the available information from the data. Multivariate classification methods based on machine learning techniques have become a fundamental ingredient to most analyses. Also the multivariate classifiers themselves have significantly evolved in recent years. Statisticians have found new ways to tune and to combine classifiers to further gain in performance. Integrated into the analysis framework ROOT, TMVA is a toolkit which holds a large variety of multivariate classification algorithms. They range from rectangular cut optimization using a genetic algorithm and from likelihood estimators, over linear discriminants and non-linear neural networks, to sophisticated more recent classifiers such as boosted decision trees, rule ensemble fitting and a support vector machine. TMVA manages the simultaneous training, testing, and performance evaluation of all these classifiers with a user-friendly interface, and expedites the application of the trained classifiers to data.

**Authors:**   Dr HOECKER, Andreas (CERN);  Dr TEGENFELDT, Fredrik (Iowa State University);  Dr VOSS, Helge (MPI fur Kernphysik Heidelberg, Germany);  Dr STELZER, Jörg (CERN);  Dr VOSS, Kai (University of Victoria, Canada)

**Presenter:**   Dr TEGENFELDT, Fredrik (Iowa State University)

Contribution ID: **2**                                        Type: **not specified**

# Statistical bias of fit parameters

In the presentation we consider systematic biases of fit parameters arising from a chiˆ2 minimization procedure, which we call "statistical biases of fit
parameters". We discuss several possible techniques, which may reduce those
statistical biases.

That may be extremely useful, in paticular for precision experiments with high statistics, if for technical reasons, for systematic studies, etc., the whole
data set has to be divided into many equal parts and then fit separately,
and the weighted average being used as a final result.

Some numerical estimates for the muon g-2 experiment are presented.

**Author:**   REDIN, Sergei (Budker Institute)

**Presenter:**   REDIN, Sergei (Budker Institute)

Contribution ID: **3**                                        Type: **not specified**

# Confidence distributions in statistical inference

This report reviews new methodology for statistical inferences.
Point estimators, confidence intervals and $p-$values have been
fundamental tools for frequentist statisticians.
Confidence distributions, which can be viewed as "distribution
estimators", are often convenient devices for constructing all the
above statistical procedures plus more.

**Author:**   Dr BITYUKOV, Sergey (INSTITUTE FOR HIGH ENERGY PHYSICS, PROTVINO)

**Co-author:**   Prof. KRASNIKOV, Nikolai (Institute for nuclear research RAS, Moscow, Russia)

**Presenter:**   Dr BITYUKOV, Sergey (INSTITUTE FOR HIGH ENERGY PHYSICS, PROTVINO)

Contribution ID: **4**                                              Type: **not specified**

# A Bayesian approach to the Constrained MSSM

We employ a Markov Chain Monte Carlo scanning technique and a Bayesian analysis to perform efficient parameter inference of the CMSSM (Constrained Minimal Supersymmetric Standard Model). The approach allows us to vary simultaneously all the CMSSM parameters and relevant Standard Model (nuisance) parameters, and to properly treat experimental constraints from collider physics and cosmology, fully incorporating all relevant sources of uncertainty. This novel application of Bayesian technology to analysing 'new physics' models leads to much more informative results than the usual fixed-grid scanning method. We delineate probability distributions of the CMSSM parameters, of collider and cosmological observables as well as a dark matter direct detection cross section. The method allows us to derive global properties of the model. In particular, by taking very broad flat priors on the CMSSM parameters, we find that the 68\% posterior probability range for the light, SM-like Higgs mass is between 115.4 GeV and 120.4 GeV. Implications for Higgs searches at the Tevatron, Higgs and superpartner searches at the LHC are also explored. An extension to other SUSY models and multi-parameter frameworks is straightforward.

## Summary

This talk will be based on our recent papers:

1) Implications for the Constrained MSSM from a new prediction for b to s gamma.
Leszek Roszkowski, Roberto Ruiz de Austri, Roberto Trotta . May 2007.
e-Print: arXiv:0705.2012

2) On the detectability of the CMSSM light Higgs boson at the Tevatron.
Leszek Roszkowski (Sheffield U. & CERN) , Roberto Ruiz de Austri (Madrid, Autonoma U.) , Roberto Trotta (Oxford U.) . Nov 2006. Published in JHEP 0704:084,2007.
e-Print: hep-ph/0611173

3) A Markov chain Monte Carlo analysis of the CMSSM.
Roberto Ruiz de Austri (Madrid, Autonoma U.) , Roberto Trotta (Oxford U.) , Leszek Roszkowski (Sheffield U.) . Feb 2006. Published in JHEP 0605:002,2006.
e-Print: hep-ph/0602028

and on some yet unpublished results

**Author:**   Prof. ROSZKOWSKI, Leszek (CERN and University of Sheffield)

**Co-authors:**   Dr RUIZ DE AUSTRI, Roberto (Univ. Autonoma Madrid); Dr TROTTA, Roberto (Oxford University)

**Presenter:**   Prof. ROSZKOWSKI, Leszek (CERN and University of Sheffield)

Contribution ID: **5**                                              Type: **not specified**

# Probability Matching Priors in LHC Physics: A Pragmatic Approach

Probability matching priors (PMP's) provide a bridge between Bayesian and Frequentist inference by yielding Bayesian posterior intervals with Frequentist validity. PMP's also allow the Frequentist access to the powerful computational tools of Bayesian methodology. Unfortunately, such priors are, in general, extremely challenging to implement as they are defined as the solution to a potentially high-dimensional and highly non-linear PDE. Outside the orthogonal case, no general framework exists for the implementation of PMP's. Recent work by Levine & Casella (2003), and Sweeting (2005) has made progress in this area, although neither approach can be applied in full generality. We consider implementation of a PMP for the three Poisson system arising in LHC experiments (as per 'The Banff Challenge', to be presented by Joel Heinrich). In this example, 'approximate probability matching' priors are sought by applying the class of priors from Tibshirani (1989). While derived as first order matching priors for orthogonal parameterizations, we propose applying prior distributions of this form to the non-orthogonal setting. An orthogonality metric is introduced to determine suitability, and relative information surfaces are used to understand the underlying structure of the problem. Simulation results and coverage properties are presented, together with comparison to a variety of alternative Bayesian techniques.

**Author:**   Mr BAINES, Paul (Harvard University)

**Co-author:**   Prof. MENG, Xiao-Li (Harvard University)

**Presenter:**   Mr BAINES, Paul (Harvard University)

Contribution ID: **7**                                          Type: **not specified**

# Evaluation of two methods for incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process

Hypothesis tests for the presence of new sources of Poisson counts amidst background processes are frequently performed in high energy physics, gamma ray astronomy, and other branches of science. While there are conceptual issues already when the mean rate of background is precisely known, the issues are even more difficult when the mean background rate has non-negligible uncertainty, as some commonly used techniques are not on a sound foundation. In this paper, we evaluate two classes of algorithms by the criterion of how close the ensemble-average Type I error rate (rejection of the background-only hypothesis when it is true) compares with the nominal significance level given by the algorithm. Following J. Linnemann, we recommend wider use of an algorithm firmly grounded in frequentist tests of the ratio of Poisson means.

**Authors:**   TUCKER, Jordan (UCLA);   COUSINS, Robert (UCLA)

**Presenter:**   TUCKER, Jordan (UCLA)

Contribution ID: **10**                                         Type: **not specified**

# Statistical Software in ROOT

Advanced mathematical and statistical computational methods are required by the LHC experiments for analyzing their data. Some of these methods are provided by the ROOT project, a C++ Object Oriented framework for large scale data handling applications. Various statistical classes and methods currently exist in ROOT spread in various libraries. Examples include methods for regression analysis, for estimating confidence levels or for classification using multi-variate analysis techniques.

An effort is on-going to re-organize these classes and integrate them together with new tools, which are currently developed by the high energy physics community, in a set of coherent and modular statistical libraries. Emphasis will be on the quality and the performance of the methods but also on the easiness of use in order to allow the comparison of similar methods without difficulty. Furthermore, a coherent design of the libraries will allow extensions and easy integration of new statistical methods developed by the physics or statistical community. The final goal is to provide common standard implementations of statistical methods required for the analysis of the LHC data.

We will briefly review the current statistical classes present in ROOT and we will present in greater detail our plans for developing these new common statistical libraries in collaboration with the LHC experiments.

**Author:**   Dr MONETA, Lorenzo (CERN)

**Co-author:**   Dr BRUN, Rene (CERN)

**Presenter:**   Dr MONETA, Lorenzo (CERN)

Contribution ID: **11**     Type: **not specified**

# Event Weighting with near-optimal variance for background-subtracted data

It is possible to find event weighting schemes which produce parameter estimates with variance nearly the same as a ML estimate. But there are situations in which a full ML estimate is inconvenient, usually for computational reasons (iteration over large data sets for example). If an variable x associated with the events is a candidate discriminating variable (that is, its distribution for signal and background differ, so that s(x) is not equal to b(x)), a weight function can be defined using s(x) and b(x) which allows estimation of a signal fraction or a number of signal events in a sample with a variance approaching that of a maximum likelihood estimate of the same quantity. In the case in which there is an external estimate of the amount of background in the sample, this is also possible, with improved variance. We derive a formula for this case and discuss it in the context of more general results on event weighting from earlier papers by Barlow and by Tkachov. The specific context came from gamma ray astronomy but the techniques may also be of interest at the LHC.

**Author:** LINNEMANN, James (Michigan State University)

**Co-author:** Dr SMITH, Andrew (University of Maryland)

**Presenter:** LINNEMANN, James (Michigan State University)

Contribution ID: **12**                                    Type: **not specified**

# A pitfall in estimating contributions of systematic errors

A common practice in evaluating the contribution of various systematic uncertainties to a result is to evaluate the contribution of each uncertainty separately (typically changing the parameter by what are felt to be + and - one sigma of systematic uncertainty), then to add the induced changes in the result in quadrature. Statisticians would refer to this as "One Factor at a Time" experimental design (see the plenary talk by Nancy Reid). This practice is in general not recommended by statisticians, despite its popularity with working scientists. Typically if pressed for a justification, the first order error propogation formula is invoked: the individual systematic effects are statistically independent, so that their effects should also add in quadrature with no cross terms because of their independence. However, if the response to the uncorrelated parameters being varied contains nonlinearities, in particular the response of one parameter depends on the value of another parameter ("interaction" in the statistical Design of Experiments jargon), a significant portion of the actual uncertainty may be missed by applying the usual formula.

**Author:**   LINNEMANN, James (Michigan State University)

**Presenter:**   LINNEMANN, James (Michigan State University)

Contribution ID: **14**                                                         Type: **not specified**

# Dilution of a statistical significance of a signal in the Higgs boson searches in the H->ZZ channel at LHC

Should an event excess compatible with the H->ZZ->4l decay channel be observed at LHC, the statistical significance of the access must be properly scaled down to account for the systematic errors and the fact that the search is performed in a wide-open range of possible Higgs boson masses. In this talk, we present results of studies addressing both of the two contributions and show that the required corrections in Higgs boson search in this particular channel are by far not negligible.

**Author:** Dr DROZDETSKIY, Alexey (University of Florida, Gainesville, FL, USA)

**Co-authors:** Prof. KORYTOV, Andrey (University of Florida, Gainesville, FL, USA); Prof. MITSEL-MAKHER, Guenakh (University of Florida, Gainesville, FL, USA)

**Presenter:** Dr DROZDETSKIY, Alexey (University of Florida, Gainesville, FL, USA)

Contribution ID: **16** Type: **not specified**

# SFitter: Determining supersymmetric parameters

If supersymmetry (or similar complex models) is found at the LHC,
the goal for all colliders over the coming decades will be to extract
its fundamental parameters from the measurements. Dedicated
state-of-the-art tools will be necessary to link a wealth of
measurements to an e.g. 20-dimensional MSSM parameter space.
Starting from a general log-likelihood function of this high-dimensional
parameter space we show how we can find the
best-fit parameter values and determine their errors. Beyond a single
best-fit point we illustrate how distinct secondary minima
occur in complex parameter spaces and how unnatural solutions can be
subsequently distinguished. In cases where there are flat dimensions in
the likelihood we comment on the benefits and limitations of
marginalizing over additional dimensions.

**Authors:** Dr ZERWAS, Dirk (LAL Orsay); Dr RAUCH, Michael (University of Edinburgh); Dr LAFAYE, Remi (LAPP Annecy); Dr PLEHN, Tilman (University of Edinburgh)

**Presenter:** Dr LAFAYE, Remi (LAPP Annecy)

Contribution ID: **17** Type: **not specified**

# Subtracting histograms using profile likelihood

It is known that many interesting signals expected at LHC are of unknown shape and strongly contaminated by background events. These signals will be difficult to
detect during the first years of LHC operation due to the initial low luminosity.
In this work, one proposes a method on how to obtain signal information of unknow shape from data even when there are very low signal and large background statistics. We present a method of subtracting histograms based on the profile likelihood when the background is previously estimated by Monte Carlo events. Estimators for each bin of the histogram difference are calculted so as limits for the signals with 68,3% Confidence Level for a low statistics case when one has a exponential background and a gaussian signal. Our results show a good performance and avoid the problem of negative values when subtracting histograms. This approach can be used to look for the Higgs particle.

**Authors:** Mr NEPOMUCENO, Andre (Federal University of Rio de Janeiro); Dr ALMEIDA, Fermando (Federal University of Rio de Janeiro)

**Presenter:** Mr NEPOMUCENO, Andre (Federal University of Rio de Janeiro)

Contribution ID: **18**                                        Type: **not specified**

# Combining Channels with Fits

We study a combined fit of several channels in a hypothetical search for new physics. In particular, we use toy Monte Carlo to investigate the potential advantages and disadvantages of imposing consistency requirements, such as demanding that the standalone fits in the individual channels yield masses that are compatible with each other before a combined fit is performed.

**Author:**   QUAYLE, William (Wisconsin)

**Presenter:**   QUAYLE, William (Wisconsin)