3B2v7.51c GML4.3.1 NIMA : 14205 Prod.Type:COM pp.1–7(col.fig.:NIL) ED:VijayaT PAGN: VS SCAN: padma

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Nuclear Instruments and Methods in Physics Research A ∎ (∎∎∎∎) ∎∎∎–∎∎∎

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

www.elsevier.com/locate/nima

Run II physics at the Fermilab Tevatron and advanced analysis methods

P.C. Bhat

Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA

Abstract

The Fermilab Tevatron has the unique opportunity to explore physics at the electroweak scale with the highest ever proton–antiproton collision energy of $\sqrt{s} = 1.96$ TeV and unprecedented luminosity. About 20 times more data are expected to be collected during the first phase of the collider Run II which is in its second year of data-taking. The second phase of Run II, expected to begin in 2005, will increase the integrated luminosity to about 10–15 fb^{-1}. Discovering a low-mass Higgs boson and evidence for Supersymmetry or for other new physics beyond the Standard Model are the main physics goals for Run II. It is widely recognized that the use of advanced analysis methods will be crucial to achieve these goals. I discuss the current status of Run II at the Tevatron, prospects and foreseen applications of advanced analysis methods.

PACS: ■; ■; ■

Keywords: ■; ■; ■

1. Introduction

The first phase of the second major proton–antiproton collider run (Run IIa) is well underway at Fermilab. Major upgrades to the accelerator complex include a brand new 150 GeV proton synchrotron called the Main Injector and a permanent magnet based antiproton recycler storage ring. The Main Injector enables 10 times more protons to be injected into the Tevatron, as compared to run I. The recycler helps recover the unused antiprotons from colliding beams of the Tevatron, store and reuse them in subsequent collisions. The collision energy is upgraded to $\sqrt{s} = 1.96$ TeV, up from $\sqrt{s} = 1.80$ TeV in Run I.

The CDF and DØ experiments also underwent major upgrades in preparation for Run II [1]. The CDF detector had the inner tracker replaced, a plug calorimeter added and muon detectors upgraded. The DØ detector acquired a new central tracker with silicon microstrip and scintillating fiber tracking layers inside a 2 T solenoidal magnetic field and new scintillating fiber pre-shower detectors. The muon system has been upgraded with a new layer of scintillators in the central region and an all new forward muon system. A forward proton spectrometer has been added to enhance capabilities for diffractive

E-mail address: pushpa@fnal.gov (P.C. Bhat).

0168-9002/03/$ - see front matter © 2003 Published by Elsevier Science B.V.
doi:10.1016/S0168-9002(03)00438-8

physics. Both experiments required new trigger and data acquisition systems.

A rich harvest of physics is expected from an order of magnitude more data that is expected to be collected in Run IIa alone. The broad physics program consists of the study of Quantum Chromodynamics via the study of jets, (particularly a high statistics study with the high transverse energy jets), electroweak physics with the W and Z bosons, beauty and charm quark physics, top quark physics including the possible evidence and study of the electroweak production of single top, and searches for the Higgs boson and for signals of new physics beyond the Standard Model, notably, the signatures for supersymmetry, lepto-quarks, technicolor or extra spatial dimensions.

In the following section, I give a short status report on the standard physics signals that are now being studied at the CDF and DØ experiments in an effort to understand the detectors and pursue studies on a wide range of physics topics. Then, I discuss, in the subsequent sections, the advanced analysis methods that are of importance in improving the various aspects of physics analysis and prospects for exciting physics applications.

2. Early physics results from run II

By the end of spring of 2002, after a year of running, the Tevatron had delivered about 55 pb^{-1} each to the CDF and DØ experiments. Most of these data have been utilized to commission and tune the detectors. Some data have been used to look at some standard physics signals. The best way to evaluate the performance of the detectors and tune the event reconstruction algorithms and assess calibration errors in these early days of the run is to look at known signals in accessible mass regions. I present here some preliminary results [2] from such studies using Run II data.

Fig. 1 shows the invariant mass spectrum from two unlike sign tracks measured in the DØ Silicon detector only, which reveals the $K_s^0 \to \pi^+\pi^-$ decays. Including measurements from the central fiber tracker improves the mass resolution to 5 MeV.

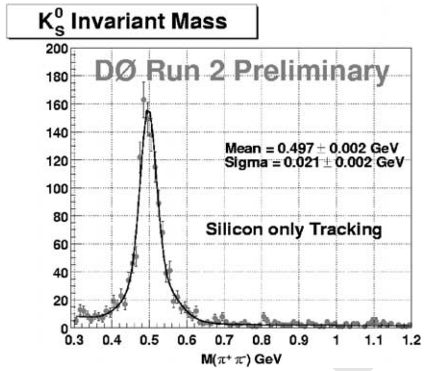

Fig. 1. The $K_s^0 \to \pi^+\pi^-$ invariant mass spectrum using tracks from the Silicon detector.

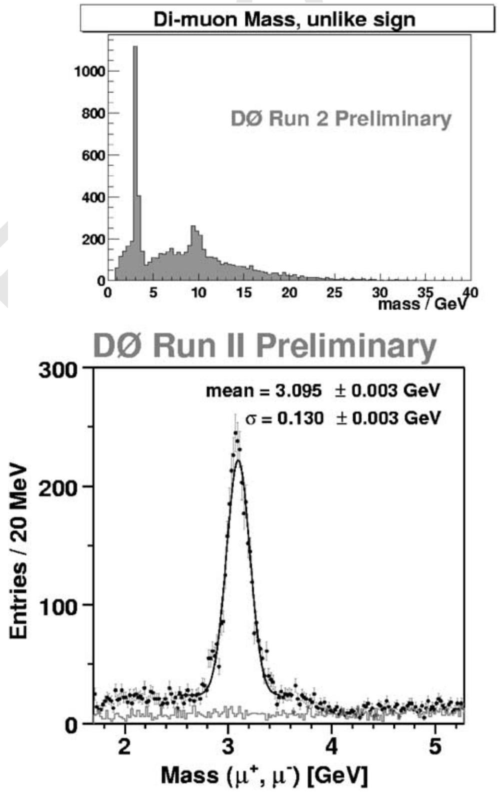

Fig. 2. The di-muon invariant mass spectrum showing the J/Ψ and the Υ peaks (top). The di-muon invariant mass spectrum after matching muons found in the muon system with tracks from the central detector shows a clean signal of J/Ψ (*bottom*).

Better alignment of the sub-detectors using data is expected to provide further improvement in track parameter measurements.

The dimuon invariant mass spectrum showing the J/Ψ and the Υ peaks using muons tracked and measured in the DØ muon system alone and the cleaner J/Ψ peak resulting after matching the muons with the central detector tracks are displayed in Fig. 2.

The electroweak gauge bosons (W, Z) are of paramount importance both in their own right and in inclusive signal and/or background channels in many interesting physics processes. The $Z \to l^+l^-$ event samples serve as good calibration tools for lepton measurements. The invariant mass distributions of $Z \to ee$ from DØ and $Z \to \mu\mu$ from CDF are shown in Figs. 3 and 4, respectively.

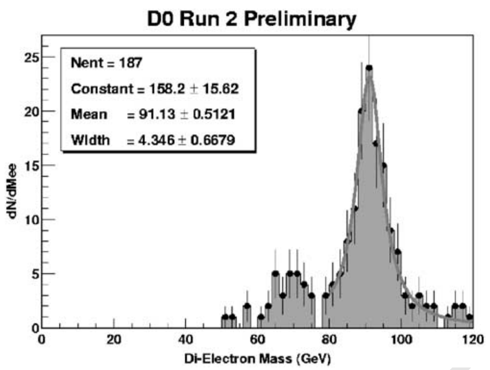

Fig. 3. The di-electron invariant mass spectrum (from DØ) showing the Z boson peak.

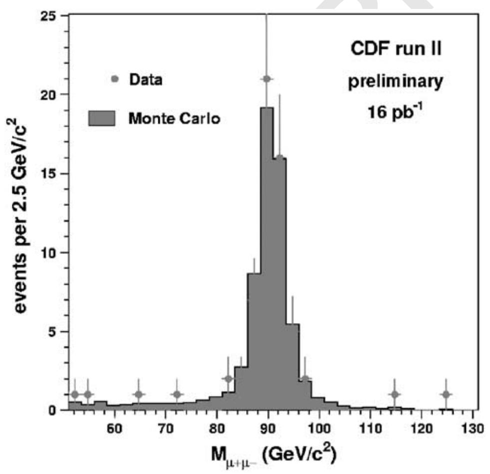

Fig. 4. The $Z \to \mu\mu$ signal from CDF Run II data [3].

The inclusive p_T spectrum of jets and the dijet mass spectrum measured in the DØ calorimeter in the central rapidity region of $|\eta| < 0.5$ are shown in Fig. 5. The jet energy corrections used are preliminary. With the full Run II data-set the transverse energy distribution of the jets should extend to beyond 600 GeV.

Enormous progress is being made at both experiments in understanding calibration and corrections, and in tuning event reconstruction and particle identification algorithms. A large number of physics analyses are in progress.

3. Advanced analysis methods

Uncovering the signals of new physics in a hadron collider environment is extremely challen-

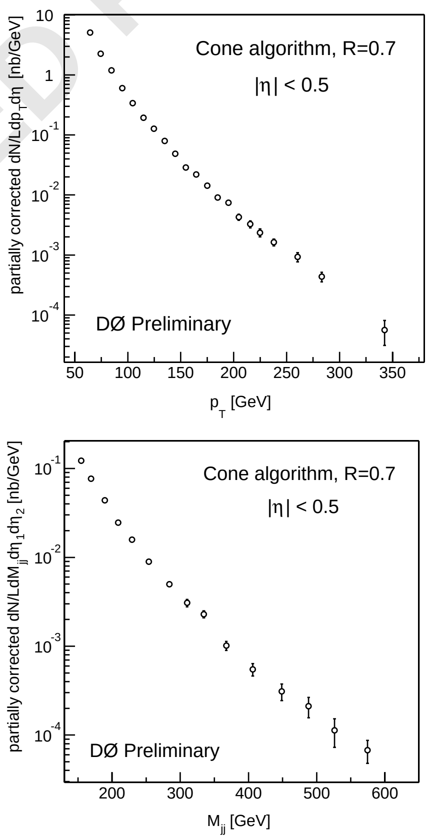

Fig. 5. Inclusive jet p_T spectrum (top) and dijet mass spectrum (bottom) from DØ ($\int L\,dt = 1.9\ \mathrm{pb}^{-1}$).

ging because of a wide variety of processes that can mimic a given signature. Therefore, the use of advanced data analysis techniques are absolutely necessary for optimal separation of signal and background. The main data analysis tasks performed in HEP are particle identification, signal/background event classification, parameter estimation (precision measurements), functional approximation (fitting) and data-driven feature extraction or exploration. The best use of data is ensured only with multivariate treatment.

Suitable choice and representation of multivariate data are important first steps for a successful application. These could be labelled simply as intelligent pre-processing of data. In some applications this pre-processing might be the only necessary multivariate treatment of the data. The selection of variables for a given analysis application can be performed using the characteristic physics information or with algorithmic approach, employing, e.g., grid searches. Having selected a set of variables, one might like to apply a suitable transformation to the variables that would yield a representation of the data that most clearly exhibits certain desirable or "interesting" properties. That is, if $\mathbf{x}$ is the original multidimensional datum, then we seek $\mathbf{s} = \mathbf{f}(\mathbf{x})$ which has the desirable properties. If a linear transformation is employed, then, $\mathbf{s} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is the transfer matrix.

The transformation of variables is equivalent to extracting a map $f: R^d \rightarrow R^N$. If $N < d$, then we would have effected a dimensionality reduction. There are nonlinear algorithms that use probability density estimation—histogramming, kernel-based methods and the methods of adaptive mixtures, and those that use stochastic optimization such as neural networks (NN). For a review of these methods, see Ref. [4]. In the following, I have chosen to discuss a few potentially interesting methods for HEP applications.

3.1. Grid search, principal and independent component analysis

Grid searches provide a systematic way of finding good variables and optimal cuts in multidimensional space, albeit not taking into account the correlations between variables. We developed a simple random grid search method [5] that can be used to compare the efficacy of variables as well as for a rapid search for optimal cuts for univariate classification. The results of such a grid search can be used as a benchmark to compare more sophisticated multivariate analyses.

To do a more advanced analysis, one would like suitably transformed variables as discussed earlier. Geometrically speaking, then, in a grid search one finds optimal cuts along the given coordinates while in the Principal and Independent Component Analyses (PCA & ICA) one finds interesting new directions (coordinates) in the multivariate space.

In the PCA algorithm (also known as Karhunen–Loeve transform or Hotelling transform), the variance along the axes is used as the interesting feature while finding transformed variables. The new set of orthogonal basis is obtained by finding eigenvectors $\mathbf{u_i}$ and eigenvalues $\boldsymbol{\lambda_i}$ as solutions of the equation,

$$\mathbf{C}\mathbf{u_i} = \boldsymbol{\lambda_i}\mathbf{u_i}$$

where $\mathbf{C} = \mathbf{E}(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^{\mathbf{T}}$ is the covariance matrix of the data set $\mathbf{x}$. The transformation matrix, $\mathbf{W}$ has as its columns the eigenvectors and the transformed variables are $\mathbf{s} = \mathbf{W}\mathbf{x}$.

The ICA [6] is a relatively new technique, invented only in the past decade. It can be seen as a powerful extension of statistical factor analysis and PCA. The observed variables are assumed to be linear or nonlinear mixtures of unknown latent variables. The ICA technique enables transformation of data variables to extract these underlying statistically independent factors.

The ICA technique has been used in analyzing medical images, signal processing, and in the field of economics. Application to HEP would be a completely new exercise, which we have recently undertaken.

3.2. NN and their ensembles

Artificial NN, though inspired from biology conceptually, are rigorous mathematical models for developing the map - $f: R^d \rightarrow R^N$ where $N \ll d$ and generally $N \sim 1$, without requiring a mathematical description of how the output(s) depend on the inputs.

Good generalization, that is good predictions for new inputs, is extremely important to minimize classification errors or to avoid over-fitting of data. The conventional methods for achieving good generalization have been (a) optimizing the size of the network given the training sample and (b) regularizing the training by penalizing model complexity. But there are new and sophisticated approaches to achieve better generalization. (See, e.g., Ref. [7].) And these involve using of ensembles of networks such as "committees" or "stacks." The basic concept in the usage of ensembles is to use many "nearly the best" networks with varying models (i.e., in architecture and input variables) rather than the "best" network, in ways that would help reduce the generalization error. It would be useful to arrive at rigorous or heuristic approaches to efficiently arrive at such ensembles.

3.3. Self organizing maps (SOM)

The SOM algorithm is an unsupervised technique which can be used for model-independent exploration of data by finding cluster patterns in data. The idea of SOM was first introduced and developed by Kohonen in early 1980s (hence called Kohonen map, as well). The algorithm maps multidimensional feature space onto, usually, a two-dimensional space with a lattice of nodes. Each node is associated with a vector of weight $\mathbf{w}$ of dimensionality $\mathbf{d}$ of the original space. An input vector is compared to all the weight vectors and assigned to the node that best matches the input.

3.4. Support vector machines (SVM)

The SVM algorithm is a fairly new one. The main idea is to map the feature space into a space of sufficiently high dimensions ($f: R^d \rightarrow R^N;\ N \gg d$) so that the optimal discriminating boundary between classes is a hyperplane and hence, can be found using linear methods. Given the feature vector $\mathbf{x}$, the optimal hyperplane is $\mathbf{w} \cdot \mathbf{b} + \mathbf{x}$ where $\mathbf{w}$ is the unit vector normal to the hyperplane and $|b|$ is the distance of the plane from the origin.

For more details on the SVM method see contribution from Vaiciulis [8].

3.5. Multivariate analysis issues

The important issues to pay attention to in performing a multivariate analysis are the following:

- choosing a set of variables without losing information,
- choosing the right method for the problem, which in many cases, has to be done by trying out a few methods,
- controlling model complexity, i.e., keeping the number of free parameters in the multivariate model small compared to the sample size,
- testing convergence of training in stochastic optimization algorithms, i.e., to have a good criteria to know when the training is optimal and cannot be improved further,
- validating the learning or modeling, i.e., quantifying the correctness of modeling or goodness of learning, especially given a limited sample,
- computational efficiency of the method and/or algorithm—it is important that the algorithm is computationally efficient so that the analysis can be repeated for many scenarios to ensure the robustness of the results.

4. Applications and prospects

The key factors responsible for the sweeping success of NN algorithms for multivariate analysis are their power, ease of use and many successful applications in HEP. To cite a few examples from Run I Tevatron physics—(1) the top quark discovery at DØ benefited from comparisons of conventional analysis with results from NN analysis [9], (2) precision measurements of the top quark mass at DØ in lepton + jets and dilepton channels where the advanced methods helped reduce the statistical uncertainities by a factor of two, (3) top quark study in all-jets decay mode and searches for single top production at DØ and CDF, (4) world's best limit on first generation scalar leptoquark mass obtained by DØ. There are many spectacular applications of multivariate methods at LEP and HERA experiments. Some example applications and prospects have been presented in other talks at this workshop [10].

In 1990, I believed that multivariate methods would provide huge gains in top quark searches. We employed multivariate methods, at DØ, particularly NN, to optimize signal selection cuts that helped in top quark physics studies from discovery to precision measurements [11]. Using advanced multivariate and Bayesian methods, the DØ collaboration measured the top quark mass to be 173.6 ± 5.6 ± 6.0 GeV in the lepton + jets channel with a better than expected statistical precision of 5.6 GeV. This extraordinary feat in Run I of precision top quark mass measurement has now been surpassed by exploiting probabilistic information for each event in the matrix element method [12] using the same Run I sample. The comparisons of discrimination between signal and background are shown in Fig. 6. The new measurement yields a top mass of 179.9 GeV with a statistical uncertainty of 3.6 GeV. Run II, with an expected yield of the order of 500 b-tagged $t\bar{t}$ events in lepton + jets final state alone per fb^{-1} recorded, ushers in an era of a variety of precision measurements in top quark physics.

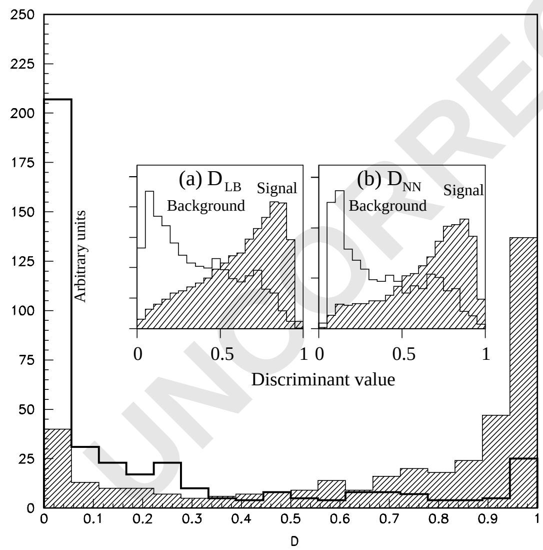

Fig. 6. Likelihood discriminant distributions for signal (hatched histograms) and background events in $t\bar{t} \rightarrow$ lepton + -jets channel using the new matrix element method [12]. The inset shows results from earlier analyses using a multivariate likelihood method (left) and NN (right).

Topping the list of interesting searches in Run II is that for the Higgs boson. The Higgs mechanism is one vital piece of the standard model that still awaits experimental evidence. Therefore, the Higgs boson would be the most sought after particle in Run II. The discovery of an SM-like Higgs boson will lend credence to the popular theories of the origin of mass. The Tevatron Run II Higgs working group explored the discovery reach for the Higgs boson and the results are shown in Fig. 7. The details of the analysis are described in published papers and the working group report [13]. There are valid and intriguing reasons for the prevailing optimism that the Higgs boson and Supersymmetry may be around the corner. The most favored Higgs boson mass from constraints from precision measurements is in the neighborhood of 100 GeV. In most SUSY models, the Higgs mass is below 150 GeV. The Tevatron, although, has good prospects for discovering a low-mass Higgs boson, it is not going to be easy. It is important to emphasize, however, that the discovery reach at a given mass requires half the integrated luminosity if multivariate methods are adopted instead of conventional univariate methods.

Searches for signatures from new physics beyond the Standard Model such as leptoquark production, supersymmetry or technicolor, are also employing multivariate methods in various stages of Run II data analysis. Advanced multi-

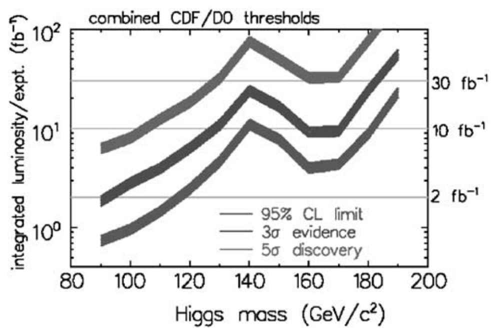

Fig. 7. The required integrated luminosity for 5σ, 3σ observation or 95% C.L. exclusion of the SM Higgs boson in Run II. For details of the analysis see Ref. [13].

variate and statistical techniques [14] form a powerful combination that enable optimal use of data, consistent treatment of uncertainties and meaningful model comparisons.

5. Summary

Run II is well underway at Fermilab. Early physics results from the upgraded CDF and DØ experiments promise an exciting physics program in the years ahead. A new era of precision measurements in standard model physics and of exciting opportunities to discover the agent(s) of electroweak symmetry breaking and new physics beyond the standard model has commenced. There are strong theoretical motivations for new discoveries and valid reasons for the prevailing optimism. The multivariate methods will provide sensitivity to new particles with masses beyond the reach of conventional methods of analysis based on univariate cuts. Advanced statistical methods adopting a fully probabilistic approach will enable better precision measurements and better explorations of model parameters. In short, the use of advanced multivariate and statistical techniques will enable new discoveries and produce results with better precision, robustness and clarity.

References

[1] CDF Collaboration, The CDF II Detector Technical Design Report, FERMILAB-PUB-96/390-E, 1996; DØ Collaboration, The D0 upgrade: The Detector and its Physics, FERMILAB-PUB-96-357-E, 1996.
[2] M. Verzocchi, in: Proceedings of the 37th Rencontres de Moriond on Electroweak Interactions and Unified Theories, Les Arcs, March 2002; hep-ex/0205049;
M. Rescigno, in: Proceedings of the 37th Rencontres de Moriond on Electroweak Interactions and Unified Theories, Les Arcs, March 2002; hep-ex/0205092.
[3] F. Bedeschi, in: Proceedings of ICHEP'02, Amsterdam, July 2002; http://www.ichep02.nl.
[4] P.C. Bhat, in: Proceedings of ACAT2000, AIP, Batavia, October 2000, p. 22.
[5] H. Prosper, et al., Proceedings of ICHEP'95, World Scientific, Rio de Janeiro, 1995, 1996.
[6] A. Hyvärinen, K. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001; http://www.cis.hut.fi/projects/ics/.
[7] C. Bishop, Neural Networks and Pattern Recognition, Oxford University Press, Oxford, 1995.
[8] A. Vaiciulis, Nucl. Instr. and Meth., these proceedings.
[9] P.C. Bhat, in: Proceedings of the Eighth meeting of the Division of Particles & Fields of the American Physical Society, World Scientific, Albuquerque, 1994, p. 705;
P.C. Bhat, in: Proceedings of the 10th Topical Workshop on Proton–Antiproton Collider Physics, AIP, Batavia, 1995, p. 308.
[10] E. Boos, L. Dudko, D. Smirnov, Nucl. Instr. and Meth., these proceedings;
L. Litov, Nucl. Instr. and Meth., these proceedings;
J.-C. Prévotet, et al., Nucl. Instr. and Meth., these proceedings;
A. Badalà, et al., Nucl. Instr. and Meth., these proceedings;
J. Zimmermann, C. Kiesling, P. Holl, Nucl. Instr. and Meth., these proceedings.
[11] P.C. Bhat, H.B. Prosper, S.S. Snyder, Int. J. Mod. Phys. 13 (1998) 5113 and references therein.
[12] J. Estrada, in: Proceedings of the DPF Meeting of the American Physics Society, Williamsburg, May 2002.
[13] P.C. Bhat, R. Gilmartin, H.B. Prosper, Phys. Rev. D 62 (2000) 074022;
T. Han, et al., Phys. Rev. D 59 (1999) 093001;
Tevatron Run II Higgs Group Report hep-ph/0010338.
[14] Fermilab Advanced Analysis Methods Group, http://projects.fnal.gov/run2aag/.