Report for Workshop 06w5054: Statistical inference Problems in High Energy Physics and Astronomy

Jim Linnemann (Michigan State University
Louis Lyons (Oxford University)
Nancy Reid (University of Toronto)

February 25, 2007

1 Introduction

In the analysis of data collected in Particle Physics experiments, the use of the best statistical techniques can produce a better quality result. Given that statistical computations are not expensive while accelerators and detectors are, it is clearly worthwhile to invest some effort on the former. The PHYSTAT series of Conferences and Workshops has been devoted to just this topic. It started at CERN in January 2000 with a Workshop on Confidence Limits - what one can say about the maximum possible strength of a hypothesised signal when no effect is seen in the data. The latest Workshop was held at Banff in the Canadian Rockies in July this year. This was also the culmination of the Workshop which had taken place earlier in the year at SAMSI (Statistical and Applied Mathematical Sciences Institute) in the North Carolina Research Triangle Park.

The Workshop was attended by 33 people, of whom 13 were statisticians, the remainder being mostly experimental Particle Physicists, with Astrophysicists making up the total. There were 3 graduate students. The Workshop concentrated on 3 specific topics: a) Upper Limits, in situations where there are systematic effects ("nuisance parameters"). b) Assessing the significance of possible interesting effects, in the presence of nuisance parameters. The subject of significance will hopefully be relevant for experiments

at the Large Hadron Collider (LHC) at CERN, where the exciting discoveries that may be made include the Higgs boson, supersymmetric particles, leptoquarks, pentaquarks, free quarks or magnetic monopoles, extra spatial dimensions, technicolour, the substructure of quarks and/or leptons, mini-black holes, etc. In all cases it will be necessary to distinguish among peaks which are merely statistical fluctuations, goofs and genuine signals of new Physics. c) The separation of events which are interesting signal from those due to boring background. This classification process is required in almost every statistical analysis performed in High Energy Physics. For each of these topics there was a Physics Co-ordinator and a Statistics one.

Of course, the 3 topics do interact with each other. Searches for new physics will result in an upper limit when no or little effect is seen, but will need a significance calculation when a discovery is claimed. The multivariate techniques are generally used to provide the enriched subsample of data on which these searches are performed. Just as for limits or significance, nuisance parameters can be important in multivariate separation methods too.

As this was a Workshop, participants were encouraged to be active in the weeks before the meeting. Reading material was circulated as well as some simulated data, on which participants could run computer programmes incorporating their favourite algorithms. This enabled all participants to become familiar with the basic issues before the start of the meeting. The Workshop started with two introductory talks (on "Brief Introduction to Particle Physics and typical statistical analyses" and "Monte Carlo Experiments in High Energy Physics"). These were primarily to describe for Statisticians the terminology used, the sort of physics issues that experimentalists try to investigate, what our statistical problems are and how we currently cope with them, etc. Jim Linnemann took the opportunity to publicise a new web site, `www.phystat.org`, which provides a repository for software useful in statistical calculations for physics. Everyone was encouraged to contribute suitable software, which can range from packages suitable from general use, to the code specifically used in preparing a physics publication.

2 General Summary

This section gives a brief summary of the main ideas touched on in the subsequent talks and discussion meetings. Summaries of individual talks

are provided in the Appendices, and for most talks there is a link on the conference web page to the slides or a relevant paper.

The discussion about limits ranged from a variety of Bayesian techniques, via profile likelihood to pure frequentist methods. An interesting suggestion from statisticians was that hierarchical Bayes might be a good approach for a search for new physics in several related physics channels. There was a lively discussion about the relative merits of the possible approaches, and even of what were the relevant criteria for the comparison. After a late evening session, it was decided that data would be made available by the limits convenor Joel Heinrich, for participants to try out their favourite methods; and Heinrich would compare the results. This work is expected to continue until November.

The significance issue was discussed in the context of Particle Physics and several Astrophysics ones too. Indeed it arises in a wide range of subjects where anomalous effects are sought. The Banff Physics convenor on significance, Luc Demortier, detailed 8 separate ways in which nuisance parameters can be incorporated in these calculations, and discussed their performance. This is going to be a crucial issue for new particle searches at the LHC, where some of the backgrounds will be known only approximately. Demortier also addressed the issues of whether it is possible to assess the significance of an interesting effect, which is obtained by physicists adjusting selection procedures while looking at the data; and why Particle Physics usually demands the equivalent of a 5 standard deviation fluctuation of the background before claiming a new discovery (The probability of obtaining such a large fluctuation by chance is below 1 part in a million).

The multivariate signal- background separation sessions resulted in very positive discussions between physicists and statisticians. Byron Roe explained the various techniques used for separating signal from background. For the MiniBooNE experiment, Monte Carlo studies showed that the 'boosted decision trees' approach yielded good separation, and was capable of coping with over 100 input variables. An important issue was assessing the effect on the physical result, in this case neutrino oscillation parameters, of possible systematic effects. One of the conventional methods for doing this is to vary each possible systematic effect by one standard deviation, and to see how much this affects the result; and then the different sources are combined. Roe pointed out that there is much to recommend an alternative procedure where the effect on the result is investigated of varying all possible systematic sources at random simultaneously.

This was a theme that was taken up in more detail by Toronto statistician Radford Neal, who also emphasised the need for any statistical procedure to be robust against possible uncertainties on its input assumptions. One of Neal's favourite methods uses Bayesian Neural Nets. He also described graphical methods for showing which of the input variables were most useful in providing the separation of signal and background.

Ilya Narsky gave a survey of the various packages that existed for performing signal-background separation. These included R, WEKA, MATLAB, SAS, S+ and his own StatPatternRecognition. Narsky suggested that the criteria for judging the usefulness of such packages should include their versatility, ease of implementation, documentation, speed, size and graphics capabilities. Berkeley Statistician Nicolai Meinshausen gave a useful demonstration of the statistical possiblities within R.

The general discussion in this sub-group covered topics such as the identification of variables that were not too useful, and whether to remove them by hand or in the programme; the optimal approach when there are several different sources of background; the treatment of categorical variables; and how to compare the different techniques. This last issue was addressed by a small group of participants working one evening using several different classifiers on a common simulated data set. Clearly there was not the time to optimise the adjustable parameters for each classification method, but it was illuminating to see how quickly it was possible to be able to use a new approach, and also to produce comparative performance figures. The results were presented by Reinhart Schweinhorst.

3 Conclusion

As far as the Workshop as a whole was concerned, it was widely agreed that it was extremely useful having Statisticians present to discuss new techniques, to explain old ones, and to point out where improvements could be made in analyses. It was noted, however, that while Astrophysics has been successful in involving statisticians in their analyses to the extent where their names appear on experimental papers, this is usually not the case in particle physics. Several reasons have been put forward to explain this. One is that statisticians like analysing real data, with all its interesting problems. But particle physics experimental collaborations tend to be very jealous about their data, and are unwilling to share it with anyone outside the collabora-

tion until it is too old to be interesting. This results in particle physicists asking statisticians only very general questions, which the statisticians regard as unchallenging and boring. If we really do want better help from statisticans, we have to be prepared to be far more generous in what we are ready to share with them. A second issue might be that in other fields scientists are prepared to provide financial support to a statistics post-doc, to devote his/her time and special skills to helping with the analysis of the data. In particle physics this is at present very unusual.

There was unanimous agreement among those there that the Banff meeting had been both stimulating and useful. The inspiring location and environment undoubtedly contributed to the dynamic interaction of participants. Not only were the sessions the scene of vigorous and enlightening discussion, but the work continued late into the evenings, with many participants learning new techniques, which they would be taking back with them to their analyses. There was real progress in understanding practical issues involved in the three topics discussed, and everyone agreed that it would be very useful and enjoyable to return to Banff for another Workshop in the future.

4 Appendix: Summaries of individual talks

4.1 Plenary Introductory Talks

Louis Lyons: Brief Introduction to Particle Physics and typical statistical Analyses

In this talk I gave an overview of the workshop organization and goals, and a brief introduction for statisticians of some of the main HEP experiments of interest, including the Tevatron, LHC and K2K experiments. I described three typical analyses; the first emphasizing estimation of unknown parameters, the second looking for an interesting signal, which of course must include a discussion of how to separate 'real peaks' from statistical fluctuations, and the third concerning the use of event variables and training data to determine how to separate background events from events of interest.

Jim Linnemann: Monte Carlo Experiments in High Energy Physics

My talk was an introduction to the use of Monte Carlo by particle physicists and a couple of other pieces of terminology. Monte Carlo simulations are used by particle physicists to estimate backgrounds and to calculate efficiencies for proposed new physics processes. The simulations consist of event generators with specific input physics and detector simulators describing how particles are observed in the apparatus.

The latter are often particularly slow, up to a minute per event. Both event generators and detector simulators have settings whose uncertainties generate systematic errors in the simulation results. I explained that cross sections are proportional to interaction probabilities, gave simple examples of kinematic quantities we cut on, and gave examples of cuts. Cuts are typically used to reduce data samples to a more manageable size, or remove regions which are difficult to simulate so as to reduce systematic errors. Physicists also use Monte Carlo methods for measuring performance of statistical methods, as do statisticians. Finally, the site phystat.org is now available for contributions to its code repository, and will link to the Banff workshop permanent web site.

Byron Roe: Setting the scene for multivariate signal/background separation

A number of modern multivariate classification are briefly described, with some emphasis on boosted decision trees and related methods. It is difficult to make comparisons in general. For limited tests with MiniBooNE Monte Carlo data, boosted decision trees performed as well or better than an other

method tested. It is noted that some hundreds of feature variables can be handled by the methods. Methods of reducing the number of variables are described. The use of simulations to estimate systematic errors varying parameters one at a time, or all together, is briefly noted. For some experiments, it is possible to estimate systematic errors while doing a fit for physics parameters. However, this method can have a problem if there are more systematic errors than data bins. A problem is noted in evaluating errors when doing log-likelihood fits in a region in which the usual analogy to chi-squared cannot be used.

Radford Neal: Statistian's view of the above

My talk tried to focus in on a series of questions, with some tentative thoughts on potential answers. First, it is helpful to be as specific as possible about the questions:

- What is the problem?
- What data is available?
- How can the data answer the questions?

Since Monte Carlo simulations play a large role, we need to consider several important issues related to this, including:

- How do we do inference with this?
- What form does the result of this take?
- How do we create PID variables?
- How can we detect and handle flaws in the models?
- How to run the MC simulation?

My view is that inference would normally be based on the likelihood function, and that a plot of the likelihood function is a very useful summary. I showed how to convert the original likelihood for a classification problem into something that depends only on the properties of the classifier, not on extraneous parameters. This explains why multivariate classifiers seem to be the right thing to do.

Now can we multiply these together to form an likelihood for N events. Is this wise? Possibly not, because our trained classifier is not perfect. A

robustness problem may enter at this point. It is also possible that this is less robust to problems in the original formulation. This motivates a kind of thresholding; which leads then to the simplified version of the problem of Poisson background and Poisson signal events.

Note that although this is not a statistical classification problemstatistically motivated classifiers may be very useful here; boosting is an example of this.

Joel Heinrich Setting the scene for limits and nuisance parameters

In this talk I sketched out what is needed in order to compare various proposed methods for computing limits, described some proposed methods for computing limits, and set out some parameter values that seem reasonable for initiating a systematic comparison. During the workshop this developed into a project for a definitive comparison of limits that will continue through the fall of 2006.

Luc Demortier: Setting the scene for p-values, including nuisance parameters

This talk gave an overview of the methods proposed for accommodating nuisance parameters the calculation of p-values. An introduction to the statistical theory of p-values was provided, and a summarization of their properties and their role as a measure of evidence. A large number of methods are available in the literature for incorporating the effects of nuisance parameters, and these were reviewed.

David van Dyk: Statistician's view of the above

I outlined the definition and interpretation of confidence intervals, and gave illustrations where the summary of the experiment by a confidence limit was not very informative. In many cases it is preferable to plot the likelihood function or the posterior distribution. I described in some detail my work with colleagues in high-energy astro-statistics.

Xiaoli Meng: Dealing with Nuisances: Principled and Ad Hoc Methods

My talk had three parts, all on dealing with nuisance parameters in hypothesis testing, particularly in testing the existence of emission lines in high energy astrophysics. The first part was a quick review of the posterior predictive approach, and the second part about how to create a useful pivotal quantity by introducing a "working" alternative model. The third part was on the idea of using moment methods, instead of maximization, to construct an approximation to profiled likelihood, a method that could be potentially useful when the usual approach of maximizing a likelihood provides unstable

results.

4.2 Parallel Session Specialized Talks

4.2.1 Limits and Significance

Roger Barlow: Significance and Likelihood ratio and confidence limits

I discussed whether Delta chi squared could be used as a measure of significance when comparing models with data, specifically for adding bumps to histograms. The conclusion is that you can't. Luc talked about this too the next day. Turns out there are papers in Biometrika that made all this clear long ago. I also asked whether my procedure to rank multichannel results for p-value purposes was sensible. There was no direct response, but the multichannel part of Joel's challenge will show the answer.

Anthony Davison: p-value functions

I was asked to talk briefly about significance functions. The starting-point is the observation that if Y is a continuous scalar random variable whose distribution function $F(y;\theta)$ depends upon a scalar parameter θ, then $U = F(Y;\theta)$ has the $U(0,1)$ distribution and is therefore a pivot: a function that depends on both data and parameter and whose distribution is known and does not depend on the parameter value. Confidence limits for θ based on an observed value y of Y may therefore be read off as the solutions in θ to the equations $F(y;\theta) = \alpha, 1-\alpha$; the resulting $1-2\alpha$ confidence interval has limits (θ_-, θ_+). There are obvious changes for upper and lower $1-\alpha$ intervals.

In practice we need to deal with three issues: we must replace Y with some general function of a data set; we must deal with nuisance parameters; and the data may be discrete. I discuss these in turn.

When Y represents a set of continuous data with log likelihood $\ell(\theta)$ and maximum likelihood estimator $\hat{\theta}$, then under mild regularity conditions on the underlying distribution we find that the likelihood root $r(\theta) = sign(\hat{\theta} - \theta)\left[2\left\{\ell(\hat{\theta}) - \ell(\theta)\right\}\right]^{1/2}$ has an approximate standard normal distribution, at least to first order; this means that if θ is the true parameter value, then

$$\{r(\theta) \le z\} = \Phi(z) + O(n^{-1/2}), \quad z \in Reals,$$

where n is an index of sample size, and Φ is the $N(0,1)$ distribution function. This implies that $r(\theta)$ is an approximate pivot, and that $\Phi\{r(\theta)\}$ may be

treated as a significance function from which confidence limits for θ may be obtained as solutions to $\Phi\ r(\theta)\ = \alpha, 1-\alpha$, as above. The resulting two-sided confidence interval typically has error of order $1/n$, while the one-sided intervals have error of order $1/\sqrt{n}$; an asymmetry term cancels from the expansions when the two-sided interval is used. Typically such intervals have better properties than those based on the score or Wald statistics. A third-order correct interval is obtained by replacing $r(\theta)$ in the above discussion with the modified likelihood root

$$r^*(\theta) = r(\theta) + r(\theta)^{-1} \log \left\{ \frac{q(\theta)}{r(\theta)} \right\},$$

where $q(\theta)$ depends on the problem; often it is either a score or a Wald statistic, but a fairly simple general form is available from work by Fraser and Reid. In this case the error for one-sided intervals drops to $O(n^{-3/2})$, and in many cases where exact computations are available or where simulations have been performed the error seems in fact to be numerically negligible.

When $\theta = (\psi, \lambda)$, where ψ is a scalar interest parameter and λ a possibly vector nuisance parameter, the log likelihood in the computation of the likelihood root is replaced by the profile log likelihood

$$\ell_{\rm p}(\psi) = \max_{\lambda} \ell(\psi, \lambda),$$

giving

$$r(\psi) = sign(\hat{\psi} - \psi) \left[2 \left\{ \ell_{\rm p}(\hat{\psi}) - \ell_{\rm p}(\psi) \right\} \right]^{1/2},$$

where $\hat{\psi}$ is the overall maximum likelihood estimator of ψ. Likewise $q(\theta)$ is replaced by a similar quantity $q(\psi)$ readily computed in many cases. Confidence intervals based on the resulting modified likelihood root $r^*(\psi)$ have again been found to be extremely close to exact ones, where these are available.

If the underlying data are discrete, then the discussion above applies with small modifications: Davison, Fraser, and Reid (2006, Journal of the Royal Statistical Society, series B) show that the error committed in taking the appropriate $r^*(\psi)$ will be of order $1/n$ rather than the $1/n^{3/2}$ seen in the continuous case, but that the numerical error is typically very small.

The overall implication is thus that excellent frequentist confidence limits can be obtained by using $\Phi\{r^*(\psi)\}$ as a significance function in both continuous and discrete cases. It turns out also that minor modifications produce

also Bayesian confidence intervals. More details and numerous examples are given in Brazzale, Davison and Reid (2006, Applied Asymptotics: Case Studies in Small-Sample Statistics, to be published by Cambridge University Press). Other accounts of this theory can be found in the books by Barndorff-Nielsen and Cox (1994, Inference and Asymptotics, Chapman & Hall), and Severini (2000, Likelihood Methods in Statistics, Oxford University Press).

Joel Heinrich: Stopping times and likelihood, a query

In certain situations, a physicist may be induced by observing events to publish right away, rather than wait for a pre-specified time. This is a definite change in procedure, the physicist's behavior is different than in the usual case. This leads to an altered frequentist ensemble, where the observed quantity is the waiting time t to the nth event, where n is a pre-specified constant. The parameter of interest s then becomes a scale parameter for the continuous (gamma) distribution of t. This is an easier problem from the frequentist perspective, and (in the b=0 case) there is an exact probability matching prior 1/s which yields perfect agreement between frequentist coverage and Bayesian credibility. From the subjective Bayesian point of view, however, the likelihood remains the same despite the change in the stopping rule, so nothing needs to be modified. More work will be done to fully explore the implications.

Tom Junk: Hypothesis Testing in HEP with Uncertain Nuisance Parameters, and an observation on Odd p-value Behavior

The multi-channel problem of testing for the presence or absence of a new particle is discussed. Typically searches produce histograms of data passing some optimized selection requirements, arranged in bins of some variable for which the signal to background ratio varies from bin to bin of the histogram. Sophisticated discriminant variables are often used, and there are many sources of uncertainty in the rates and shapes of these histograms. Often the bins of the histogram where little signal is expected serve as a calibration of one or more uncertain backgrounds. If the signal distribution shape is similar enough to the background shape, the effectiveness of the technique of fitting the sidebands weakens. Often the predictions in each bin of the signal and the background suffer from limited Monte Carlo samples, which introduces another source of uncertainty and a set of nuisance parameters in each bin.

One can address the problem very similarly to the approach Kyle Cranmer proposed at PHYSTAT03 – to maximize the likelihood separately with respect to the nuisance parameters for the null hypothesis and the test hy-

pothesis, and then apply known techniques for computing limits or confidence belts. I have a bias towards testing two hypotheses at a time and not many more, since the acceptance or exclusion of a test hypothesis should not depend on other test hypotheses considered or not considered.

One pitfall to avoid in doing a Bayesian limit with marginalization over the unknown nuisance parameters is double-use of the data sidebands to constrain backgrounds. The integration over nuisance parameters of the likelihood times the prior is effectively fitting the background shape to the data, as only those values of the background normalization nuisance parameters which best reproduce the data will be represented with significant weight in the integral. To use the data sidebands to construct the prior for the background (a Gaussian constraint), and then to use the same data in the likelihood function would double-count the effect of this data on the background uncertainty.

Software is available at

www.hep.uiuc.edu/home/trj/cdfstats/mclimit_csm1/index.html

It calculates both Bayesian upper limits and CLs ones. P-values are also computed for comparing the data to the null and test hypotheses.

An odd feature of p-values in low-statistics single-channel analyses is a manifestation of the discontinuous coverage curves. Over coverage is unavoidable, and particularly noticeable for channels with few expected events. If a second, much weaker channel is combined with a strong channel with little expected background (say s1=3, b1=1, s2=0.1, b2=2), then the observed limit can jump sharply when the second channel is added. Part of the over coverage of the single-channel case is now recovered by dividing the probability of each strong-channel's outcome into sub-outcomes indexed by the weak channel's outcome. If a continuous spectrum of signal and background can be constructed instead of a single counting experiment, then the distributions of the expected outcomes will not only be more optimal because of the extraction of more information from the data, but also will suffer less from the Poisson over coverage problem.

Toby Burnett: Detection of gamma rays "Finding point sources in the gamma-ray sky"

The main point here is that it is conventional in astrophysics to quote source discoveries with a "5 sigma" threshold, but without an analysis to demonstrate that the probability of a false positive is really at the level of a 5-sigma Gaussian. This is probably evident in the number of unidentified sources "found" by EGRET, and certainly demonstrated by the recent

GLAST data challenge.

I see it as a clear message to those performing such analyses that the null hypothesis does not depend on the position, so that the p-value must be determined empirically.

James Bueno: Bayesian upper limits I described the software I have written in Root C++ to calculate Bayesian upper limits for the Poisson problem with a choice of informative priors for background and efficiency via numerical Gaussian quadrature, and illustrated it on some examples.

Luc Demortier: Reference analysis I described the use of reference posterior distributions as a means of making inference about a parameter of interest in the presence of nuisance parameters and summarized their properties.

Eric Marchand: On the behaviour of Bayesian credible intervals for some restricted parameter space problems. This is recent work with Bill Strawderman. For estimating a positive normal mean, Zhang and Woodroofe (2003) as well as Roe and Woodroofe (2000) investigated HPD credible sets associated with priors obtained as the truncation of noninformative priors onto the restricted parameter space. They established the attractive lower bound of $(1-\alpha)/(1+\alpha)$ for the frequentist coverage probability of these procedures. W. Strawderman and I established that their lower bound is applicable for a substantially more general setting with underlying distributional symmetry. We showed that the lower bound still applies for certain types of asymmetry (or skewness), and we extended results obtained by Zhang and Woodroofe (2002) for estimating the scale parameter of a Fisher distribution. There is a wide scope of applications, including estimating parameters in location models and location-scale models, estimating scale parameters in scale models, estimating linear combinations of location parameters such as differences, estimating ratios of scale parameters, and problems with non-independent observations.

Gunther Zech: Likelihood vs coverage Coverage intervals versus Likelihood ratio intervals An example was constructed such that a coverage interval and a likelihood interval disagree by a large extent. It was demonstrated that the likelihood interval is at least intuitively much more attractive. The reason for this behavior is the fact that coverage intervals accept all parameter values which are compatible with the measurement whereas likelihood ratio intervals take into account the fact that only one and not several parameters can be true. As a consequence of this caveat of the coverage intervals, which is related to a violation of the likelihood princi-

ple, one should not require coverage for likelihood ratio or Bayesian intervals while coverage intervals which exclude relatively high likelihood ratios are very problematic.

Giovanni Punzi: Frequentist limits

I briefly described a fully frequentist method for incorporating nuisance parameters (systematic uncertainties) in the evaluation of confidence intervals, by means of a direct Neyman construction in multi-dimensional space. Thanks to an appropriate choice of ordering algorithm, results were obtained with good general properties: strict coverage, small overcoverage, and continuous behavior in the limit of small systematic uncertainty. The algorithm allows for both 1-sided and 2-sided limits (F-C or central) to be obtained, and puts no requirements on the distribution of the subsidiary measurements (related to the systematics), which may even be unavailable. Some results and comparisons for the benchmark problem of group A were presented.

Conrad, Jan and Cranmer, Kyle: Profile likelihood for marked Poisson processes

Bodhisattva Sen: Confidence intervals with nuisance parameters We discussed the Hybrid Resampling Method for dealing with nuisance parameters. We also proposed an extension of the Feldman and Cousins Unified method to include nuisance parameters. The Expectation-Maximization (EM) algorithm in relation to the signal plus noise model with marks (auxiliary variable associated with each event) is proposed. We also sketch a possible Bayesian hypothesis testing procedure for testing discovery in this scenario.

4.2.2 Multivariate

Stephen Bailey: Signal/background separation of supernova events in Supernova Factory images The Supernova Factory is developing a Support Vector Machine (SVM) approach to replace its current cut-based approach. The SVM works considerably better than cuts at 25% signal efficiency but performs about the same at 50% signal efficiency. The primary difficulty is a time varying background; participants provided several useful suggestions for training classifiers under under such conditions by including background parameters in the classification variables. At the suggestion of the participants, the Supernova Factory will also try Random Forest and Boosted Decision Tree classifiers which might work better with the noise and outliers in the dataset.

Radford Neal: Bayesian neural nets + robust classifiers

I described a neural network analysis of Byron Roe's data, which has 50 PID variables, to be used to classify events into one of two types. The NN has two hidden layers which connections from all inputs to all hidden units. Each connection has a weight attached to it, a 'bias' constant, and a tanh activation function. The output for each hidden unit is

$$tanh(b + \Sigma_i w_i V_i)$$

and this is converted at the end to a binary classifier by the logistic function.

This can be parameterized, by treating the weights and biases as parameters. It can be shown that even with just one hidden layer any function can be approximated with enough units.

In principle the parameters could be estimated by maximum likelihood, but this is not usually a good idea, since the data can be fitted exactly with enough units. NonBayesian methods incorporate either regularization or a naive method called 'early stopping', or an ensemble version of that analogous to cross-validation. This version of early stopping actually does quite well.

It turns out that he number of hidden units is not so crucial; early stopping basically corrects for this.

A Bayesian version works better, by integrating over the space of weights in the network, with a combination of prior and likelihood. This integration is carried out by MCMC. The Metropolis algorithm is very slow, but a hybrid MC version works well. The result is a probability for classification, but this is done on a large number of networks, and averaged (at the last step). The Bayesian version can incorporate a hyper-prior for the parameters on some of the parameters in the network. This allows some improved classification if the data turns out to be very predictable.

You can show that there is no statistical necessity to constrain the number of hidden units. However the computations do get slow as the number gets too large.

The next step is adding some boosting to this, to concentrate on the items that are hard to classify, but of course apply larger weights to the easy-to-classify items so that the output remains unbiased. This hopefully will reduce the computation time.

Toby Burnett: Decision trees

I made two points:

- Classification analysis can be used, not only to distinguish rather different entities, but to improve resolution. GLAST uses classification

trees to help characterize gammas that have well-measured energy and direction.

- It is productive to study misclassified background events, in order to discover new variables that can be used in a new classification. Many of the GLAST variables had such an origin.

Ilya Narsky: Multivariate classification Software for multivariate classification

Various software packages for multivariate classification have been used in HEP and elsewhere.

R, a popular tool among statisticians, implements many methods for classification and exploratory analysis of data. It is easy to install and use, has built-in graphics for display of data, and is generally well documented. However, R tends to be slow, especially on large data sets. R is a high-level interpreter suited for interactive analysis but hardly a reasonable choice for analyzing large amounts of data through batch jobs. WEKA, an object-oriented Java package, and MATLAB extensions implement many classification methods and are used by researchers in academia. SAS and S-plus offer software suites for industry but are less likely to be used in academia due to the high cost of their licenses.

HEP researchers have been using various implementations of neural networks for the last two decades. For example, Stuttgart Neural Network Simulator (SNNS) and JETNET have become popular at BaBar. Plenty of other implementations are available.

Several packages have been developed recently within the HEP community to implement advanced classifiers such as boosted decision trees, random forest and others. These include Byron Roe's m-boost, Narsky's StatPatternRecognition and TMVA (now available in Root). StatPatternRecognition, for instance, implements decision trees, boosting, arc-x4, random forest, a bump hunting algorithm (PRIM), linear and quadratic discriminant analysis, and interfaces to two SNNS networks.

At present, HEP analysts can choose software from a great number of available packages. Unfortunately, a HEP researcher working on specifics of a physics analysis, typically a grad student or a post-doc, often knows little about multivariate classification in general and even less about available software to make an informed choice. It would be very useful for the community to survey existing software packages and publish results of this survey online

or in any other form available to HEP researchers. The proposed categories for software comparison are listed here: " versatility and the scope of implemented methods " ease of installation and use " quality of manuals and documentation " CPU speed and memory consumption; how these quantities scale versus data size and dimensionality, both for the training cycle and for post-training classification; also maximal sample size and dimensionality that can be handled by the package " types of inputs that can be handled by the package (real, integer, categorical, mixed etc) " quality and convenience of the graphics interface, both for input and output " suitability for interactive analysis and/or batch jobs " ease of integration in the C++ framework Although it would be interesting to compare the predictive power of various implementations of the same method, this task would require a non-trivial amount of manpower. Because implementations of the same method vary among packages, such a comparison would not be possible without careful adjustment of input parameters for each implementation of the classifier, which is a time-consuming effort.

The proposed comparison would be a nice project for one or two undergrads or graduate students and a useful service to the community.

Blobel, Volker: Systematics for goodness of fit Dealing with systematics for chi-square and for log-likelihood goodness of fit

Systematic errors are at the origin of the unsatisfactory situation when data from many experiments are used in a global analysis and parameter estimation, and when attempts are made to determine uncertainties of predictions from parton distributions. Often the profile chi-square for single parameters and functions of parameters appear to be much to narrow.

The different error contributions in HEP experiments and the methods, to incorporate these error contributions in the log-likelihood expression are discussed. Often this is not done in an optimal way.

The normalization factor for all data of a single experiment is a product of many factors and its distribution can often be described by a log-normal distribution. In the log-likelihood expression the factor should be applied to the theoretical expectation, not to the experimental value, to avoid a bias.

Recent experiments publish a lot of data about the contributions to the overall covariance matrix from various systematic effects, and this information has to be used in parameter estimation. The contribution to the covariance matrix describing the statistical errors is usually given as a diagonal matrix. However due to finite experimental resolution. corrections for the bin-to-bin fluctuations have to be applied; correlations between data points

are often neglected and the given statistical errors are often too optimistic.

Nicolai Meinshausen: Using `R` for classification problems

The R-Project web page (www.r-project.org) has the code, all the various packages, a set of manuals, and so on. An advantage of R is that it is easy to play around with the data, fairly quickly, and there are packages written for many many statistical routines. A disadvantage is that it is quite as 'portable' as C++, and it is maybe not so easy to translate R fitted objects to another language. Also it can be slow as it is an interpreted program.

Variable Selection: With approximately 200 variables, one suggestion is that they all be kept in the classifier, and only deleted if they demonstrably don't have any impact on the classifier. Of course the variables in this category may be different between classifiers, but in general the set of really important ones "should" be consistent from one classifier to the next. Nicolai showed some plots available from random forests that identify variable importance according to a few measures related to classification.

Raoul LePage: Comments on Classification

I described how multivariate Gaussian based confidence regions may be obtained directly from scaled bootstrap plots with sufficient blocking to improve normality. This will apply to plots of a kind often seen in talks at this meeting, although not to all of them. The other part of my talk outlined a possible approach to the problem of developing useful classifiers for signal vs noise. The idea is to exploit the capabilities of modern methods of solving linear inverse problems by using them to directly obtain a density over the events "e" space that is interpreted at p(signal — e) for a particular choice of the probability p of signal (i.e. blending signal and noise events in the proportions p, 1-p). The idea also turns on telling the solver that integrals of some specified functions X(i) ¿ 0 (these are used by the inverse method to build the density) are identically one. Such functions can be chosen from any convenient class thought to be capable of building a good density for the purpose and would be normalized according to their integrals on the training set. In the present context favorite choices might include decision trees with specified parameters, logistic models, or any combination of these.

Neal, Radford: Handling Systematic Errors in Simulations

Reinhard Schwienhorst: Multivariate classification Comparisons

I gave a talk summarizing the work of several people, comparing different classifiers for the data sets that had been sent out before or during the meeting. We used the statistical data analysis package R and software written by Ilya Narsky (one of the workshop attendees) and data sets from MiniBoone,

Glast, Babar, and the D0 single top quark search. We found that Bayesian Neural networks, boosted decision trees and random forests did about as well as classifiers tuned within each experiment, even when using these classifiers out-of-the-box, without any special tuning. We also learned that installing and running R is straightforward and results can be obtained in a few hours, at least with the help of an expert

5 Summary of discussions

5.1 Limits with Nuisance Parameters

convenors David van Dyk (statistics) Joel Heinrich (intervals),

After listing the main methods that have been proposed to solve the upper limits problem, an attempt was made to collectively construct a matrix that listed their properties. This resulted in considerable discussion, and the addition of a few more methods. After the break, it was clear that the matrix was only reasonably complete in the column marked coverage, and that only for a single channel. As all the methods had reasonable coverage properties, more information was needed for a prospective user to decide which to use.

A collaborative project to supply the necessary information about each method was therefore initiated. Proponents of each method agreed to supply their resulting intervals for a common set of test cases. This would permit direct comparison of frequentist coverage, interval lengths, and Bayesian credibility. It was decided that 1 channel and 10 channel cases would be investigated. Those doing the work would have until the end of October to complete the task. Joel Heinrich agreed to generate the test cases and process the results.

The statisticians proposed adding a hierarchical Bayesian method to the list, and described in detail how this worked. This was accepted, after some discussion—physicists had been reluctant in the past to employ this strategy, but were persuaded by the statisticians that it would be worth trying.

5.2 Significance

convenors David van Dyk, Luc Demortier

The discussions on significance revolved around a set of questions that arise in various ways in searches for new physics. Statisticians were asked to

comment.

(1) Why a 5 sigma discovery threshold? Do we really believe in such small probabilities?

As previously explained by Louis Lyons, the motivation for the 5-sigma threshold in high energy physics is that many observed 2- and 3-sigma effects have been known to disappear, and that this is understood to be due to the "look-elsewhere" effect, and to unidentified or improperly modeled systematic effects. There is also a subjective aspect to the choice of threshold, in that physicists will require a very high standard of evidence for invalidating laws in which they believe strongly, such as energy conservation for example.

Although statisticians do not object to the 5-sigma threshold, they warn against using it as an excuse for ignoring the look-elsewhere and systematic effects. The latter should always be carefully investigated and taken into account in significance calculations. This warning becomes even more important as we move towards experiments of increasing size and complexity.

(2) Should the same significance threshold be used in all situations, regardless of the type of hypothesis being tested and regardless of sample size?

Although the evidence provided by p values against a given hypothesis has been shown numerous times (by professional statisticians) to depend on the type of hypothesis being tested as well as on sample size, the statisticians' answer to the above question was a measured "Yes, but be aware that the evidence provided by p values depends on these conditions."

(3) What methods are there to incorporate systematic uncertainties in p values? Which one(s) should we recommend?

Several methods were described at the meeting. High energy physicists seem to favor the "prior-predictive" method, which consists in averaging the p value over some prior density for the nuisance parameters. A concern, voiced by Kyle Cramner, is that a different way of setting up the prior in the calculation could lead two competing experiments to large differences in significance. A simple example is a Poisson p value with a Gaussian uncertainty on the mean. If the width of the Gaussian is independent of the mean, the resulting prior-predictive p value will tend to be conservative. On the other hand, if the width is proportional to the mean, the p value will be liberal.

(4) Are there general rules for choosing an optimal test statistic? What about in multiple dimensions, and with sparse data?

According to the statisticians present, the likelihood ratio statistic is the best bet in the vast majority of situations.

(5) What can we say about the likelihood ratio in non-standard situations, e.g. when a parameter is on the boundary of the maintained hypothesis, or when nuisance parameters appear under the alternative but not under the null?

This problem was discussed by Luc Demortier in his plenary talk and by Roger Barlow in a parallel talk. Non-standard likelihood ratio tests are quite common in high energy physics, for example when scanning a spectrum for a resonance with unknown mean and/or width. There have been several recent developments in this area by statisticians, in particular by econometricians. Among the statisticians present, Anthony Davison suggested some techniques based on moving average searches or on wavelet transforms. He also pointed out the usefulness of importance sampling (using the likelihood ratio as weight) in computations. Richard Lockhart pointed to recent work by Taylor and Worsley. Jim Linnemann provided two useful papers, one advocating the use of posterior-predictive p values, and the other a test statistic based on a score process.

(6) How should we handle a significant-looking discrepancy in one distribution out of many?

One possibility is to do a multiple test (Bonferroni or improvements thereon). Another possibility is to split the data sample in two, use the first half to "fish" and the second half to calculate significances. FDR techniques may be useful in well-planned fishing expeditions, as well as in particle ID algorithms.

(7) Should we seriously consider alternatives to p values?

Alternatives to p values do exist: Mike Evans' observed relative surprise, Jose Bernardo's Bayes reference criterion, Jim Berger's relative likelihoods. However, the statisticians present thought that these alternatives do not have enough "experience" yet to support their use, and they are not as well understood as p values. Nancy Reid mentioned that posterior hypothesis probabilities can sometimes be a useful alternative to p values.

A general comment is that a significance calculation will always be more persuasive when there exists a plausible alternative hypothesis that is shown to describe the data better than the null.

In addition to the discussion we had on the above questions, there were a couple of interesting talks. One was by Anthony Davison, on significance functions, where he showed how to eliminate nuisance parameters from these functions, how to handle discrete problems, and the relation between significance functions and reference posteriors. The other talk was by Tom Junk,

who discussed some interesting examples of p value calculations in high energy physics.

5.3 Multivariate Problems

convenors David van Dyk, Joel Heinrich (intervals), Luc Demortier (p-values)
Group C: Nancy Reid and Byron Roe

The following list of questions was compiled, and to some extent discussed, in the classification group:

1. go over classification methods: what can be said about applicability in different situations
2. Models not perfect; what can we say about robustness or other flaws in the model.
3. How do we get an intuitive feeling for the classification methods? Graphical methods?
4. Methods for variable selection? How to find the 'best' set of variables? Why do we need to reduce the number of variables?
5. How many data-sets have been looked at by different people.
6. Unisims, multisims: estimating systematic uncertainties.
7. Uniform framework? (Should we all be re-inventing the wheel?)
8. Question: is subjectivity adding to the quality of the analysis or not?
9. Can we learn about statisticians' methodology: Raoul, Nicolai

A graphical display was suggested by Radford Neal: Plot $\Delta\ log(p/(1-p))$ which is the change in the classifier output when variable 27 is changed by ϵ in the training case, vs variable 27. If the plot is straight, variable 27 affects the logit linearly; if the plot is curved then the variable isn't linear, but it is additive (no interaction). If the plot is scattered, then there are interactions going on with the other variables. (The computation of Δ"*prediction*" adjusts for the presence of the other 49 variables.)

Note that one plot needed for each PID, and that each plot has as many points as there are events in the training sample. There could be an 'ensemble' of such plots, one for each of the networks that go into the bagging, OR, it could be based on the averaged or bagged prediction.

Jim Linneman had the following thoughts

Variable preparation: If you have multiple backgrounds, you could consider training either separate classifiers for each; or training a multiple class classifier (even though all background classes would eventually be lumped into "non-signal"). Might be easier to diagnose that way.

You could consider removing from the training on a particularly resistant background which is truly nearly indistinguishable from your signal, rather than confusing the classifier by calling very similar events background in one case and signal in another. But as above, a multi-classifier might also address that issue differently.

For trees, no need to modify variables since they are invariant under uniform monotone transformations (as are cuts).

For nets, they like to have mean zero std deviation = 1 if you have no better idea. If you do so, then for the Bayes nets, you can use the same hyper-priors (the weight scales start out the same). If you have widely varying variables, say ones covering a wide range with a wide range of frequency, you might for example log-transform them. Ideally, if you have real reason to know some variable should be a good separator, you would ideally want it such that one unit of input change would correspond to one unit of output change, so an important variable might be given a larger initial standard deviation than one; that way similar weights would start out be causing bigger responses to this variable. But: such selection is not always obvious: a variable which by itself (1-d) shows little separation between signal and background can nonetheless be important in classification if it interacts strongly with other variables.

Terminology: physicists confuse the term "correlated or non-independent variables" with "interacting variables": I believe the issue is that correlation or independence would be a property of simply the signal or background distributions individually, while interaction has to do with the pair of variable's needing to be considered together to classify (predict a third variable, such as class membership); so it would be more a property of the ratio of the signal and background pdf's]

Non-ordered categorical variables. Say you have inputs which fall naturally in 3 classes, but which don't have any inherent ordering. For example,

you have observations with 3 different kind of electronics, but don't really want to claim they like in a "good better best" relationship. It's better to classify them into the 3 groups with similar characteristics if you can, than asking the classifier to figure it out from, say, the 100 different serial numbers. And the good way to encode those is to have 3 categorical input variables each with 0 or 1 values, so that only one of them is on for a given input. On the other hand, if there is a natural ordering relationship like "good better best" it may not be so bad to encode them as 1, 2, or 3 (though the response might be nonlinear–little difference between good and better, but big improvement for best). Or if you really have a quantitative number, it's better to use instantaneous beam intensity as a continuous variable than classifying them as low, medium, and high intensity.