CDF/PUB/STATISTICS/PUBLIC/6428

Calculation of Cross Section Upper Limits Combining Channels Incorporating Correlated and Uncorrelated Systematic Uncertainties

John Conway
Rutgers University

March 21, 2005

Abstract

A common problem arising in new particle search experiments is to determine upper limits on a signal cross section, combining more than one channel (for example e and μ) and incorporating the effects of systematic errors, which may or may not be correlated across channels. This note describes a Bayesian procedure for performing such a calculation.

1 Motivation

Quite often one searches for a new particle or phenomenon using a number of final states, or "channels." For example, in the search for $Z' \to \ell^+\ell^-$, one can look at final states with $Z' \to e^+e^-$ as well as $Z' \to \mu^+\mu^-$. Given a model for the production and decay branching ratios of the Z' to these final states, one may wish (in the absence of a discovery!) to set 95% CL upperlimits on the Z' production cross section.

In general, also, in setting limits one would like to incorporate the effects of systematic uncertainties. For example the expected background, the signal accceptance, and other factors may not be perfectly known; this lack of knowledge should weaken the limit one obtains.

Further complicating the picture is the fact that in many cases these systematic uncertainties can be correlated, which can weaken further the limit one might obtain. As the magnitude of such uncertainties grows the effects of correlated and uncorrelated uncertainties cannot be neglected.

This note outlines a procedure based on the commonly employed Bayesian integration technique for calculating such upper limits. For simplicity the case illustrated here is that of combining two or more single-channel counting experiments. The method has been employed in the recent stop → tau [1] and Z' [**?**]searches in CDF, which are examples of this. The method, however, is generalizable to much more complicated situations, including fits to kinematic spectra and unbinned likelihoods.

2 Bayesian limits with no systematics

Let the expected cross section of some signal process for which we search be σ_{sig}. Suppose there are N_{ch} search channels, in which one observes n_i events, with $i = 1, 2, ...N_{ch}$. in each channel, furthermore, assume the expected number of background events is b_i, and the acceptance is ϵ_i.

In the absence of any systematic errors one can form a joint likelihood from the product of the individual channel likelihoods as

$$\mathcal{L}(\bar{n}|\sigma_{sig}, \bar{b}, \bar{\epsilon}) = \prod_{i=1}^{N_{ch}} \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \tag{1}$$

where the expected number of events μ_i in each channel can be written

$$\mu_i = L_i \sigma_{sig} \epsilon_i + b_i \quad . \tag{2}$$

Here L_i is the integrated luminosity of the data sample used for each channel. (The overbars indicate that the variables are arrays carrying an i index.)

Given the likelihood, Bayes' Theorem can be used to derive a posterior probability density as a function of the signal cross section. We write this as

$$\mathcal{P}(\sigma_{sig}|\bar{n}, \bar{b}, \bar{\epsilon}) = \frac{\mathcal{L}(\bar{n}|\sigma_{sig}, \bar{b}, \bar{\epsilon}) P(\sigma_{sig})}{\int_0^\infty \mathcal{L}(\bar{n}|\sigma'_{sig}, \bar{b}, \bar{\epsilon}) P(\sigma'_{sig}) d\sigma'_{sig}} \quad . \tag{3}$$

The function $P(\sigma_{sig})$ is the prior probability density for the (unknown) signal cross section and expresses the prior knowledge of its value. For example, we know that $P(\sigma_{sig} < 0) = 0$ for real physics processes. Typically one takes $P(\sigma_{sig}$ to be uniform, at least up to some large value well beyond the eventual upper limit so that it remains normalizable. Of course the eventual answer one obtains can depend on the choice of prior, and this is a matter of some debate in the statistics community, but beyond the scope of this note. The method described here can be employed with whatever prior is chosen.

To obtain a 95% CL upper limit, then, one would find that value of σ_{95} such that

$$\int_0^{\sigma_{95}} \mathcal{P}(\sigma'_{sig}|\bar{n}, \bar{b}, \bar{\epsilon}) d\sigma'_{sig} = 0.95 \quad . \tag{4}$$

3 Bayesian limits with systematics

The usual way to incorporate the effects of systematic uncertainties (also called nuisance paramters) is to convolute the Poisson probability with a function representing the prior probability density in each paramter; usually one uses a Gaussian. Then one "marginalizes" the nuisance parameters by integrating over them. We might, for example, write the likelihood for a single channel, marginalized with respect to an uncertain acceptance, as

$$\mathcal{L}(n|\sigma_{sig}, b, \epsilon) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \int_0^\infty \frac{\mu'^n e^{-\mu'}}{n!} e^{-(\epsilon'-\epsilon)^2/2\sigma_\epsilon^2} d\epsilon' \quad . \tag{5}$$

The extension of the technique to additional systematic uncertainties is straightforward. Note that the value of the integrand ϵ' is used to calculate μ at any point.

One can evaluate such an integral by standard grid sampling techniques, or by a Monte Carlo technique. To comprehend the Monte Carlo approach, firstly note that mathematically, the integral is in essence a Gaussian-weighted average Poisson probability. One can readily calculate this average by repeatedly sampling the Poisson probability, drawing the value of ϵ' from a Gaussian distribution, and averaging the samples.

For the more complicated cases where there are multiple systematic uncertainties, and multiple channels, the Monte Carlo techniques converges very rapidly – much more rapidly than when standard grid sampling numerical techniques are applied. The accuracy of the intregration scales as the inverse of the square root of the number of Monte Carlo samples.

4 Correlated and uncorrelated systematics

Having described how to incorporate systematic uncertainties in general by marginalization, it remains to deal separately with correlated and uncorrelated uncertainties.

An example of a correlated systematic error might be the uncertainty on the integrated luminosity for each channel L_i. Though the L_i may be different, the relative systematic uncertainty may be the same for each, and completely correlated between the L_i. For the acceptance in each channel, there may be sources of uncertainty which are correlated (such as Q^2 scale effects) and those which are uncorrelated (such as Monte Carlo statistics or lepton ID efficiency).

Noting that the systematic effects all enter through the "smeared" expected number of events μ'_i. It is this value that varies from random point to random point during the Monte Carlo integration. It is convenient to express the shifts in the parameters affected by systematic uncertainty by multiplying the central value of the parameter by a term of the form $(1+f)$ where f is a random relative offset which takes into account the Gaussian variation of the parameter around its mean. In each channel, we can write the expected number of events during the Monte Carlo integration as

$$\mu'_i = (1+g_L)L\sigma_{sig}(1+f_{\epsilon i})(1+g_\epsilon)\epsilon_i + (1+f_{bi})(1+g_b)b_i \quad . \tag{6}$$

The key to the technique described here is that the f and g factors here encode the *relative* uncorrelated and correlated systematic uncertainties, respectively. (Note that the f's carry i indices and the g's do not.) To perform the Monte Carlo integration, then, it suffices to allow the f and g factors to vary within their Gaussian widths around a central value of zero, and average the resulting likelihoods.

Of course, this is not the only form that the expected number of events can take. For example, the expected background (or a portion of it) may depend on the integrated luminosity and would then be affected by that systematic error. However, as long as the expected number of events can be written in a form such that the systematic uncertainties can be expressed through f and g terms, the method can be readily adapted to calculate the marginalized likelihood.

Given the marginalized likelihood calculated in this way, one can again convert it to a posterior density in the signal cross section using Bayes' Theorem as before and obtain the desired confidence intervals..

5 Performance of the method

The method described above has been implemented in a program which steps through 10000 signal cross section points over a suitable range, and at each point performs a Monte Carlo

n_1	ϵ_1	$f_{\epsilon 1}$	b_1	f_{b1}	n_2	ϵ_2	$f_{\epsilon 2}$	b_2	f_{b2}	g_ϵ	g_b	σ_{95}
3	0.2	0.2	3.0	0.1	-	-	-	-	-	-	-	31.6
3	0.2	0.2	3.0	0.1	3	0.2	0.2	3.0	0.1	0.0	0.0	18.4
3	0.2	0.0	3.0	0.1	3	0.2	0.0	3.0	0.1	0.2	0.0	20.0
3	0.2	0.0	3.0	0.1	3	0.2	0.0	3.0	0.1	0.3	0.0	28.6

Table 1: Result of 95% CL upper limit on signal cross section, from a single channel, the combination of two channels with uncorreelated systematic errors, and combining two channels with correlated systematic errors.

integration utilizing thousands of random samples. The convergence of the integration can be studied by plotting the limit as a function of $1/\sqrt{N_{MC}}$ and fitting a line: the intercept corresponds to extrapolating to infinite statistics.

There are several features of the method described above that one desires:

- The method should give a smaller (more stringent) upper limit when channels are combined than those for the individual channels.
- The method should give less stringent limits as the systematic uncertainties increase.
- The method should give less stringent limits for correlated uncertainties than for equivalent uncorrelated uncertainties.

To test whether these properties hold, we can inspect the limits reported in Table 1 for various cases involving one or two channels, and a value $L = 1$ with no systematic uncertainty. (The diligent reader will find that they indeed do hold.)

While these are necessary properties for any channel combination method, they are perhaps not sufficient. Ultimately, frequentists need to study the frequentist coverage of the confidence intervals; Bayesians are unconcerned with coverage.

Appendix - corlim.f Program

There exists a Fortran program to calculate the upper limit on a signal process, based on combining the observation in one or more channels using the method described in this note. [4]

The program, corlim.f, is driven by a data file containing a table of the experimental results, the integrated luminosity, the acceptance and expected background, and all coorelated and uncorrelated uncertainties. Figure 1 shows the data file used to calculate the results shown in Table 1, with annotations indicating the various components of the table.

References

[1] J. Conway, *et al.*, Search for Pair-Production of Scalar Top Quarks in R-parity Violating Decay Modes, CDF Note 6316.

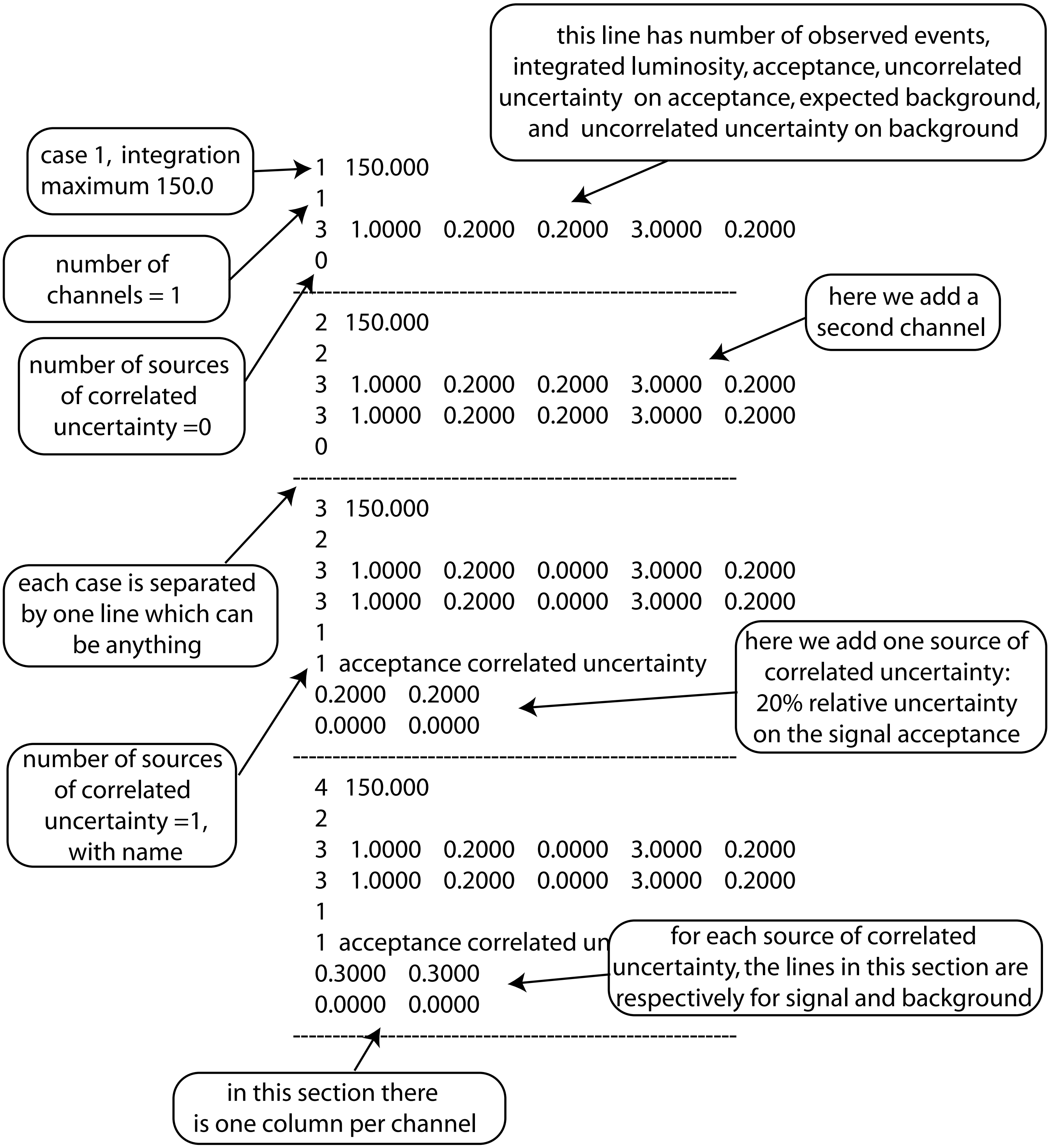

Figure 1: Example data file for the corlim.f program, annotated with the locations of various parameters.

[2] J. Conway, *et al.*, Combined Limits on Z' and RS Graviton in High Mass Dielectron and Dimuon Channels Using CDF Run 2 Data, CDF note 6386.

[3] Program available on request from the author.