Fermilab Library
0 1160 0013934 9
LAL-92-37

LAL 92–37
June 1992

Laboratoire de l'Accélérateur Linéaire

COMPLEMENTARY ANALYSES AND ACCEPTANCE OPTIMIZATION IN NEW PARTICLE SEARCHES

ROUTE TO
NAME | LOCATION
RETURN TO

J.–F. Grivaz and F. Le Diberder

U.E.R
de
l'Université Paris-Sud

Institut National
de Physique Nucléaire
et
de Physique des Particules

Bâtiment 200 - 91405 ORSAY Cedex

JUL 2 1992

LAL 92–37
June 1992

Complementary analyses and acceptance optimization in new particle searches

J.-F. Grivaz and F. Le Diberder

Laboratoire de l'Accélérateur Linéaire, IN2P3-CNRS
et Université de Paris-Sud, F-91405 Orsay Cedex, France

Abstract

Two simple statistical methods are presented, which both allow a given exclusion confidence level to be set on the production of new particles, valid independently of any assumption on the potential backgrounds. Fistly, the use of analyses contributing little to the overall efficiency of a new particle search is justified, even if they are liable to contribute significantly to the expected background level. Secondly, a prescription to decide on the optimal location of the acceptance cuts in such a search is proposed.

1. Introduction.

The purpose of this report is to present two simple statistical methods which allow an exotic process, assumed not to take place, to be excluded at a given confidence level, with no assumptions on the potential sources of background being made, neither regarding their rates nor their distributions. In this respect, the methods advocated here fundamentally depart from the usual likelihood analysis which is to be applied if the background and the signal distributions are both known.

A convenient tool to calibrate the selectivity of a search analysis is provided by $< \sigma >$, the average over a large number of experiments of the value of the signal level σ which can be excluded at some given confidence level, say 95%. The lower $< \sigma >$ is, the more selective the analysis can be considered. The value of $< \sigma >$ depends on ε, the efficiency of the analysis for the signal, and on b, the average number of background events expected in that same analysis. In the case of simple event counting, it can be calculated as follows.

The number n of background events expected to be selected follows a Poisson distribution with average value b:

$$\mathcal{P}_b(n) = e^{-b}\frac{b^n}{n!} \ .$$

Let κ_n be the 95% CL upper limit set on the expectation value of a signal when n events are observed and no background subtracted:

$\kappa_n = 3.00,\ 4.74,\ 6.30,\ 7.75\ldots$, for $n = 0,\ 1,\ 2,\ 3\ldots$, respectively.

Then:

- in a fraction $\mathcal{P}_b(0)$ of the experiments, no events will be found and the 95% CL upper limit κ on the number of signal events to be expected within the acceptance of the analysis will be set at κ_0;
- in a fraction $\mathcal{P}_b(1)$ of the experiments, one event will be observed, from which κ will be set at κ_1;
- ...

The average value of κ is

$$\begin{aligned} < \kappa > &= \sum_{n=0}^{\infty} \kappa_n \mathcal{P}_b(n) \\ &\simeq 3(1 + 0.581\ b - 0.032\ b^2 + 0.005\ b^3 - \ldots) \end{aligned}$$

where the approximate expression holds for $b \lesssim \mathcal{O}(1)$.

And finally, the desired quantity is obtained:

$$< \sigma > = \frac{< \kappa >}{\varepsilon} \ .$$

In the following, two examples are presented of how this calibration tool can be used, and these applications are combined, in a somewhat unexpected way, in the last section.

2. On the use of complementary analyses.

A question which is frequently raised in the context of new particle searches is whether it is justified to make use of analyses which add little to the overall detection efficiency but may introduce some background events, therefore reducing the selectivity.

Consider for instance the case in which

- a main analysis A_1 yields a detection efficiency $\varepsilon_1 = 90\%$ for the process under study, with an expected background level of $b_1 = 0.15$ events,
- an additional analysis A_2 provides an efficiency $\varepsilon_2 = 10\%$ while adding $b_2 = 0.5$ to the expected background level.

If no signal is present, the original analysis A_1 leads to the observation of no events in a fraction $e^{-b_1} = 86\%$ of the experiments, in which case a 95% CL upper limit of $\kappa_0/\varepsilon_1 = 3.33$ will be set on the signal level. The upper limit becomes at least $\kappa_1/\varepsilon_1 = 5.27$ if one or more events are observed. The average value of the upper limit is $< \sigma_1 >= 3.62$.

If now analysis A_2 is included, the upper limit can be decreased to $\kappa_0/(\varepsilon_1+\varepsilon_2) = 3.00$, but only in a fraction $e^{-(b_1+b_2)} = 52\%$ of the experiments, while it will become greater than or equal to 4.74 otherwise. The average value of the upper limit is increased to $< \sigma_{1+2} >= 4.10$. It therefore seems that it would not be wise to take analysis A_2 into account since the result obtained using A_1 alone would probably be spoilt.

This conclusion does not necessarily hold, however, if not only the total number of events observed is considered, but also the information regarding the analyses by which they were selected. This can be achieved using as an estimator the product $\mathcal{E}_K$ of the probabilities associated with the observations in each of the K analyses. In the case of simple event counting,

$$\mathcal{E}_K(n_1,\ldots,n_K) = \prod_{k=1}^{K}\sum_{i=0}^{n_k}\mathcal{P}_{\nu_k}(i)$$

with

$$\mathcal{P}_{\nu_k}(i) = e^{-\nu_k}\frac{\nu_k^i}{i!} ,$$

instead of the usual

$$\sum_{i=0}^{n}\mathcal{P}_{\nu}(i) .$$

Here, k is the analysis label, n_k is the number of events observed in analysis k and ν_k is the number of events expected from the signal in that same analysis, while $n = \sum_k n_k$ and $\nu = \sum_k \nu_k$ are the total numbers of events observed and expected, respectively. Some attention must be paid to the fact that the value $\eta_{\rm CL}$ of $\mathcal{E}_K$ which corresponds to a given

exclusion confidence level CL — that is the value such that $\mathrm{Prob}(\mathcal{E}_K < \eta_{\mathrm{CL}}) = 1\text{-CL}$ — is no longer simply 1-CL, but this is only a minor technical complication.

Returning to the previous example, it turns out that the inclusion of analysis A_2 allows the 95% CL upper limit σ to become as low as $\kappa_0/(\varepsilon_1+\varepsilon_2) = 3.00$ in 52% of the experiments, while remaining $\leq \kappa_0/\varepsilon_1 = 3.33$ in another 34%, becoming 4.77 in another 8%, remaining $\leq \kappa_1/\varepsilon_1 = 5.27$ in another 5%, and greater than or equal to 6.37 in the remaining 1% of the experiments. The average value (which cannot be here simply obtained with the analytical expression given in Section 1) is $< \sigma >= 3.38$.

Not only is this average value smaller than $< \sigma_1 >$, but the upper limit obtained in any single experiment never becomes larger than if only A_1 were used. It is therefore clear that, in this example, analysis A_2 should be taken into account in spite of its comparatively large background level.

More generally, starting with a "main" analysis A_1, it can be considered that an additional analysis A_2 should be taken into account, irrespective of the background level it may introduce, as soon as it can never decrease the level of significance attached to the non-observation of events in the main analysis. In other words, the relation

$$\mathrm{Prob}(\mathcal{E}_2 < \mathcal{E}_2(0, n_2')) \leq \mathrm{Prob}(\mathcal{E}_1 < \mathcal{E}_1(0)) \qquad \forall n_2' \geq 0$$

should hold, which is achieved if

$$\mathcal{E}_2(0, n_2') < \mathcal{E}_2(n_1, n_2) \qquad \forall n_2' \geq 0, \forall n_1 \geq 1, \forall n_2 \geq 0.$$

This translates into

$$\mathcal{P}_{\nu_1}(0) < (\sum_{i_1=0}^{n_1} \mathcal{P}_{\nu_1}(i_1)) \times (\sum_{i_2=0}^{n_2} \mathcal{P}_{\nu_2}(i_2)) \qquad \forall n_1 \geq 1, \forall n_2 \geq 0 \ ,$$

which is automatically satisfied if

$$\mathcal{P}_{\nu_1}(0) < (\mathcal{P}_{\nu_1}(0) + \mathcal{P}_{\nu_1}(1)) \times \mathcal{P}_{\nu_2}(0) \ ,$$

i.e. if $\nu_2 < \log(1+\nu_1)$. If a 95% CL exclusion is aimed at, $\nu_1 = \kappa_0 = 3.00$ and this leads to $\nu_2 < 1.39 = 0.46\ \nu_1$. In other words, combining *via* the estimator $\mathcal{E}_2$ a complementary analysis A_2, with a detection efficiency ε_2, with a main analysis A_1, with an efficiency ε_1, is a no-lose strategy, irrespective of the background levels b_1 and b_2, as soon as $\varepsilon_2 < 0.46\varepsilon_1$. For larger efficiency increases, it is necessary to evaluate the amount of background associated with the complementary analysis in order to assess, using for instance the calibration tool $< \sigma >$, its potential usefulness.

An interesting consequence, which also goes against naive intuition, is that, if an analysis has been designed in some unbiased way in order to lead to no events in a large fraction of the experiments, its acceptance can be increased at no cost by up to 46%, even if some background may be introduced this way, provided the initial and additional

acceptance regions are treated separately as explained above, just as if different analyses applied in the two regions. Of course, the exact acceptance of the "additional analysis" must also be designed in an unbiased way.

3. Acceptance optimization.

The question of the justification of the location of acceptance cuts is also often raised in the context of new particle searches, but rarely answered in a quantitative way. The goal is to keep ε, the efficiency for the signal from the process being searched for, as high as possible, while at the same time reducing the background to a level b where it becomes harmless. Since these are conflicting requirements, a compromise must be found. Here, a simple prescription is provided, suited to the case where no background subtraction is performed, and in which this compromise is achieved in a well controlled way.

The optimal analysis is defined as the one which minimizes $< \sigma >$.

As defined in Section 1, $< \sigma >=< \kappa > /\varepsilon$ is the average value of the 95% CL upper limit set on the acceptance-corrected number of signal events when a large number of experiments is performed. The analysis acceptance is defined by the values of the cuts applied to a set of discriminating variables. Normally, when the cuts are tightened, ε and b decrease together. However, if a smaller b always leads to a lower $< \kappa >$, it does not necessarily lead to a smaller $< \sigma >$ because of the lower ε.

An example is shown in Figs. 1 and 2 in the simple case of a single variable v. The distributions, as a function of v, of the signal and of the expected background are displayed in Figs. 1a and 1c, respectively. The resulting distributions for ε and for b are presented, as a function of the cut value c, in Figs. 1b and 1d , respectively. Finally $< \sigma >$, the quantity to be minimized, is shown as a function of c in Fig. 2a. The optimal cut value is indicated on the various figures as a vertical line labelled (1).

A consequence of the advocated prescription is that, when the accumulated statistics increase, the cut locations should be modified since the expected background level would otherwise increase accordingly while the analysis efficiency would remain unchanged. Such an analysis update is shown in Fig. 2b in the case of the above example, now with the expected level of background b multiplied by a factor of five: the cut c should be tightened to the location indicated by the vertical line.

It is worth emphasizing that the minimization of $< \sigma >$ merely serves as a prescription to specify the analysis to be used. The actual exclusion confidence level obtained relies in no way on the assumed value for b. This confidence level is indeed conservative, and any additional knowledge on the backgrounds could only render it more stringent. In particular, if the background turns out to be very different from the one used to optimize the analysis, the exclusion power will be decreased, but the limit provided will remain valid.

4. And now...

In Section 2, it has been shown how a complementary analysis A_2 could be taken into account without ever running the risk of harming the main one, A_1, provided A_2 adds sufficiently little acceptance to that of A_1. This can now be used as follows.

Let the main analysis, A_1, be the one in which the acceptance cuts have been optimized as prescribed in Section 3, and let the search domain for A_2 be restricted to the region lying beyond the cuts set for A_1, thus ensuring that no overlap occurs between the acceptances of A_1 and A_2. A new set of cuts can now be optimized in this restricted search domain, with both the background level and the efficiency for the signal calculated within that same restricted domain. In the case of the example discussed above, the location of the new cut is shown in Fig. 2c, and indicated on the various figures as a vertical line labelled (2). Even if, in most of the situations occurring in practice such as the one of this example, A_2 adds much less to A_1 than the maximum allowed, there is no reason not to profit from A_2 since it is for free...

This procedure can of course be iterated *ad libitum*, supplementing a given analysis with a set of less and less powerful complementary analyses. In principle, the acceptance cut optimization should then be performed in a global way, with all the complementary analyses simultaneously taken into account *ab initio*.

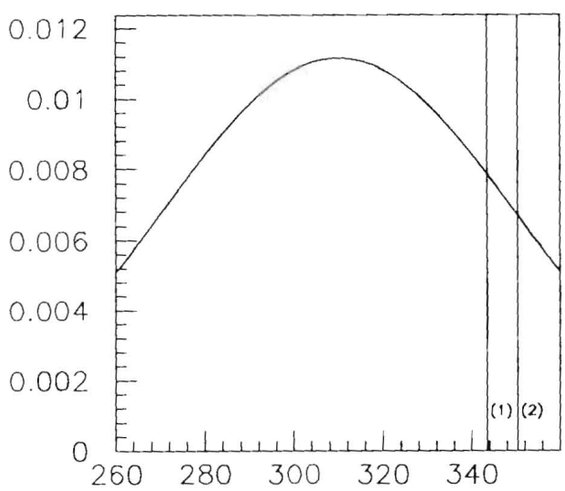

Figure 1a

V

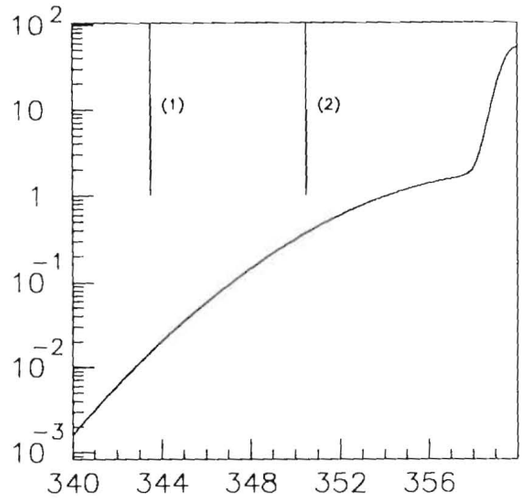

Figure 1c

V

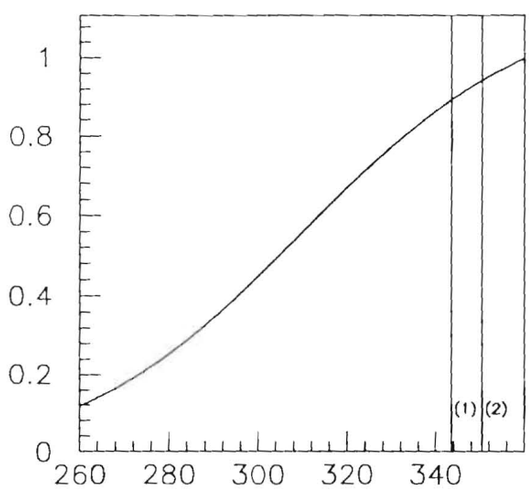

Figure 1b

C

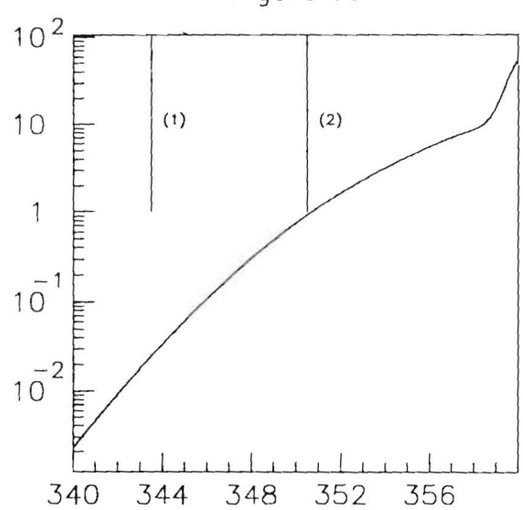

Figure 1d

C

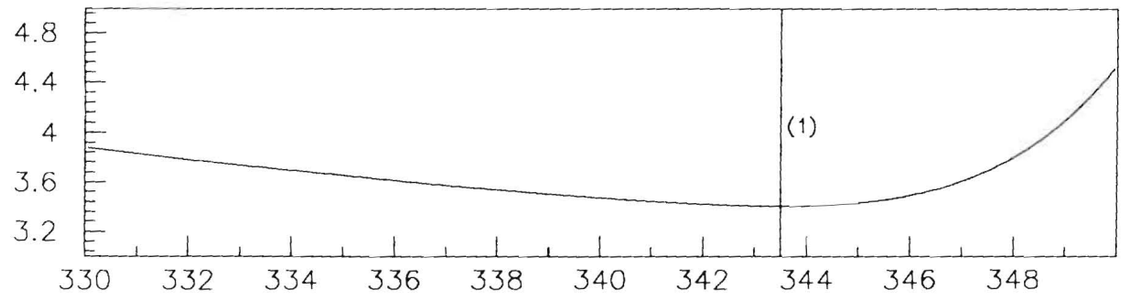

Figure 2a

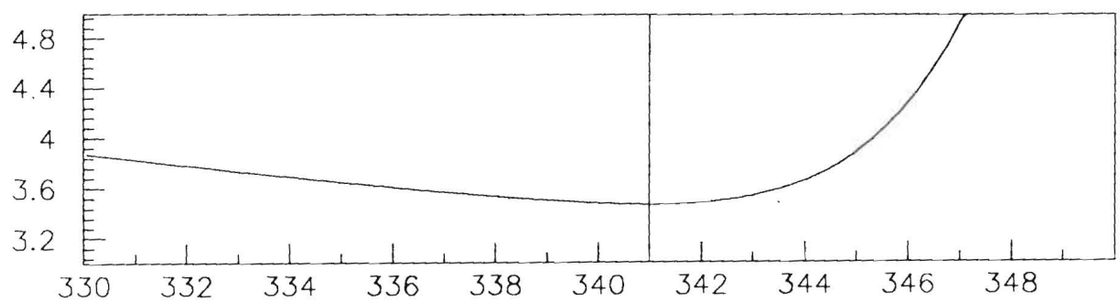

Figure 2b

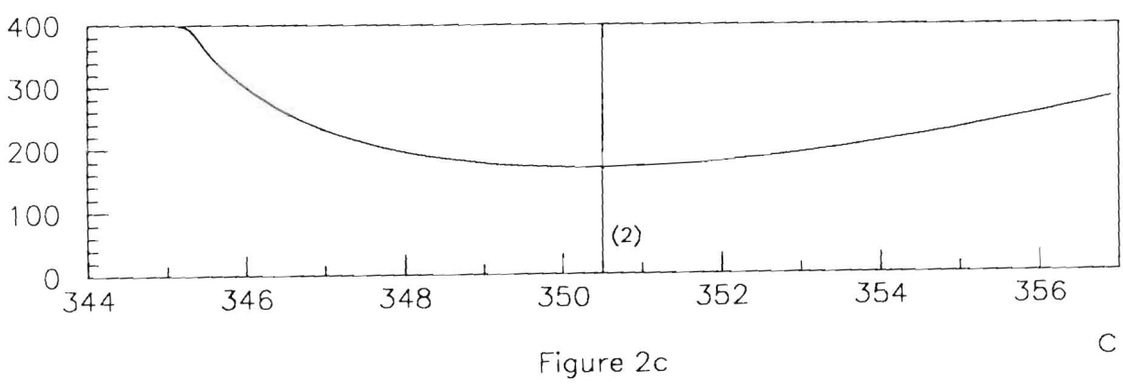

Figure 2c