# Why and When to Optimize Efficiency Times Purity

Benno List
ETH Zürich, Institute for Particle Physics

31.7.2002

The quantity efficiency times purity $\epsilon \cdot \pi$ is in many situations the quantity that should be optimized by a selection. This note shows why this is the case, and when.

Let us consider the case where we want to measure a certain signal cross section $\sigma_{\mathrm{s}}$. The number of signal events that we expect to see is given by $\langle S \rangle = \epsilon \cdot \int \mathcal{L} \mathrm{d}t \cdot \sigma_{\mathrm{s}}$.

In the presence of background, we expect also a number of background events $\langle B \rangle$ in the signal region.

Let us assume that this *expected* number of background events in the signal region is known with a small uncertainty (small compared to the statistical error on the number of signal events). This can be the case if we have a background Monte Carlo sample which we trust very much, or if we do a fit of the background shape (e.g. around a mass peak) that we can extrapolate into the signal region with high accuracy.

Then the expected number of total observed events is

$$\langle N \rangle = \langle S \rangle + \langle B \rangle = \epsilon \cdot \int \mathcal{L} \mathrm{d}t \cdot \sigma_{\mathrm{s}} + \langle B \rangle.$$

Now we do the measurement, and observe $N$ events. From this the signal cross section $\sigma_{\mathrm{s}}$ is calculated as

$$\sigma_{\mathrm{meas}} = \frac{N - \langle B \rangle}{\epsilon \cdot \int \mathcal{L} \mathrm{d}t}. \tag{1}$$

Now we evaluate the statistical error of the cross section measurement; this is simply

$$\delta\sigma_{\mathrm{meas}} = \frac{\delta N}{\epsilon \cdot \int \mathcal{L} \mathrm{d}t} = \frac{\sqrt{N}}{\epsilon \cdot \int \mathcal{L} \mathrm{d}t},$$

so the expected statistical error is

$$\langle \delta\sigma_{\mathrm{meas}} \rangle = \frac{\sqrt{\langle N \rangle}}{\epsilon \cdot \int \mathcal{L} \mathrm{d}t},$$

Note that we do *not* add a statistical error for $\langle B \rangle$, because we assume that we know the *expected* number of background events very well.

Let us now define the purity

$$\pi = \frac{\langle S \rangle}{\langle S \rangle + \langle B \rangle} = \frac{\langle S \rangle}{\langle N \rangle}. \tag{2}$$

Then we can write

$$\langle N \rangle = \frac{\langle S \rangle}{\pi} = \frac{\epsilon \cdot \int \mathcal{L}\mathrm{d}t \cdot \sigma_\mathrm{s}}{\pi},$$

and we get

$$\langle \delta\sigma_\mathrm{meas} \rangle = \frac{\sqrt{\frac{\epsilon \cdot \int \mathcal{L}\mathrm{d}t \cdot \sigma_\mathrm{s}}{\pi}}}{\epsilon \cdot \int \mathcal{L}\mathrm{d}t}, = \sqrt{\frac{\sigma_\mathrm{s}}{\epsilon \cdot \pi \cdot \int \mathcal{L}\mathrm{d}t}}$$

or equivalently

$$\frac{\langle \delta\sigma_\mathrm{meas} \rangle}{\langle \sigma_\mathrm{meas} \rangle} = \frac{1}{\sqrt{\epsilon \cdot \pi \cdot \int \mathcal{L}\mathrm{d}t \cdot \sigma_\mathrm{s}}}.$$

**This shows that the expected statistical error $\langle \delta\sigma_\mathrm{meas} \rangle$ is minimal if the event selection minimizes the quantity "efficieny times purity" $\epsilon \cdot \pi$.**

This answers the "why".

To answer the "when," we must look into the assumptions that we made along the way:

The most important assumption is that we can neglect the statistical uncertainty of the expected number of background events $\langle B \rangle$, for all variations of the event selection that we consider.

This is especially the case if we make a fit to a mass peak of the form signal+background, and the get $\langle B \rangle$ by integrating the background under the signal region.

This is also the case if we have a very good background Monte Carlo that predicts the absolute normalization of the background with a small uncertainty.

There are, however, situations where other factors have to be taken into account:

In the case of a search, we may have a situation where a standard model Monte Carlo generator predicts the bulk of the data very well, but the description of the tails of distributions is not so certain. If then $\epsilon \cdot \pi$ may be optimal in a region where $\epsilon$ is small and $\pi$ is high, i.e. in a region where only a small fraction of the standard model background remains. In this region the systematic uncertainty on the standard model background may be high, and become the dominant systematic error. In that case, a selection with higher signal efficiency and lower pority, but also a lower relative systematic uncertainty on the background expectation may have a smaller overall uncertainty.

In the case of searches there is a second reason why one would *not* want to optimize $\epsilon \cdot \pi$: Whereas $\epsilon$ is independent of the signal cross section, $\pi$ does depend on the assumed signal cross section. This is not very elegant.

The way out of the dilemma is to say "we want to be sensitive to the smallest possible cross section," and thereby letting $\sigma_\mathrm{s}$ go to 0. Then $\pi$ becomes $\pi \approx \langle S \rangle / nbexp$, and $\epsilon \cdot \pi \approx \langle S \rangle^2 / \langle B \rangle$. So one has to optimize this quantity, or equivalently the more familiar $\langle S \rangle / \sqrt{\langle B \rangle}$, which is maximal if $\epsilon / \sqrt{\langle B \rangle}$ is maximal. Now the optimal working point does not depend on the assumed signal cross section anymore.

One might add that this formula should be modifie when $\langle B \rangle$ becomes small.

In another situation, the expected number of background events itself may have a significant statistical uncertainty, which may be coupled to the statistical unceratinty of the total number of events. This is the case if only very small sidebands around a mass peak can be fitted, or if the background is taken from a control sample (like–sign instead of unlike–sign pairs, negative instead of positive decay length) that represents only the $\langle B \rangle$ part.

This situation is discussed in the following section.

### Background is Taken from Control Sample

Let us consider the case where we have a signal sample, where we observe $N$ events, plus a control sample with $B$ events. Let us assume that the control sample is prepared such that the average number of events in the control sample is just the expected number of background events in the signal region $\langle B \rangle$.

Now our master formula Eq. (1) becomes

$$\sigma_{\mathrm{meas}} = \frac{N - B}{\epsilon \cdot \int \mathcal{L} \mathrm{d}t},$$

with the only (but important) difference that $\langle B \rangle$ has to be replaced by $B$.

Now we also have to take the statistical error on $B$ into account, and we get

$$\delta\sigma_{\mathrm{meas}} = \frac{\sqrt{N + B}}{\epsilon \cdot \int \mathcal{L} \mathrm{d}t},$$

or for the expected error

$$\langle \delta\sigma_{\mathrm{meas}} \rangle = \frac{\sqrt{\langle N \rangle + \langle B \rangle}}{\epsilon \cdot \int \mathcal{L} dt}.$$

From the definition of the purity Eq. (2) it follows that

$$\langle N \rangle + \langle B \rangle = \langle S \rangle \cdot \left( \frac{2}{\pi} - 1 \right),$$

which leads to

$$\frac{\langle \delta\sigma_{\mathrm{meas}} \rangle}{\langle \sigma_{\mathrm{meas}} \rangle} = \sqrt{\frac{2 - \pi}{\epsilon \cdot \pi \cdot \int \mathcal{L} \mathrm{d}t \cdot \sigma_{\mathrm{s}}}}.$$

So in this case not $\epsilon \cdot \pi$ has to be maximized, but rather $\frac{\epsilon \cdot \pi}{2 - \pi}$!

The denominator $2 - \pi$ shifts the optimal working point ot the analysis towards higher purities, as one would expect when the background error becomes important.

### Background is Taken from Control Sample, Signal is Diluted

A more complex situation arises if a fraction of the signal "leaks" into the control sample.

Let us define $C$ as the number of events in the control sample. We now have

$$\langle C\rangle = d \cdot \langle S\rangle + \langle B\rangle,$$

i.e. the control sample contains on average $\langle B\rangle$ background events and $d \cdot \langle S\rangle$ signal events. Let us call $d$ the dilution factor, and assume that it is known from Monte carlo simulations.

Now we have $\langle B\rangle = \langle C\rangle - d \cdot \langle S\rangle$, and therefore for the expected number of events $\langle N\rangle$ in the signal region

$$\langle N\rangle = \langle S\rangle + \langle B\rangle = \langle S\rangle + \langle C\rangle - d \cdot \langle S\rangle = (1-d) \cdot \langle S\rangle + \langle C\rangle.$$

Now the master formula becomes

$$\sigma_{\mathrm{meas}} = \frac{N - C}{(1-d) \cdot \epsilon \cdot \int \mathcal{L}\mathrm{d}t},$$

the statistical error is

$$\delta\sigma_{\mathrm{meas}} = \frac{\sqrt{N+C}}{(1-d) \cdot \epsilon \cdot \int \mathcal{L}\mathrm{d}t},$$

and the expected statistical error is

$$\langle \delta\sigma_{\mathrm{meas}}\rangle = \frac{\sqrt{\langle N\rangle + \langle C\rangle}}{(1-d) \cdot \epsilon \cdot \int \mathcal{L}\mathrm{d}t}.$$

Now $\langle N\rangle = \langle S\rangle/\pi$, $\langle B\rangle = \langle S\rangle(1/\pi - 1)$, and $\langle C\rangle = \langle S\rangle(1/\pi - 1)(1+d)$, so we get

$$\langle N\rangle + \langle C\rangle = \langle S\rangle \left(\frac{1}{\pi} + \left(\frac{1}{\pi} - 1\right)(1+d)\right) = \langle S\rangle \frac{(2+d) - (1+d)\pi}{\pi},$$

and finally

$$\frac{\langle \delta\sigma_{\mathrm{meas}}\rangle}{\langle \sigma_{\mathrm{meas}}\rangle} = \sqrt{\frac{(2+d) - (1+d)\pi}{(1-d)^2 \cdot \epsilon \cdot \pi \cdot \int \mathcal{L}\mathrm{d}t \cdot \sigma_{\mathrm{s}}}}.$$

So this time one has to maximize $\frac{(1-d)^2 \cdot \epsilon \cdot \pi}{(2+d) - (1+d)\pi} = \frac{(1-d)^2 \cdot \epsilon \cdot \pi}{(2-\pi) + d(1-\pi)}$.

## Special Case: Gaussian Signal over Linear Background

Finally, let us consider one important special case: A Gaussian signal over a linear background.

Let us assume that we have a Gaussian signal which contains (on average) $s_0$ events, over a linearly varying background that contains $b_0$ events within $\pm 2\sigma$ of the signal's mean value.

We define now $x$ as the number of sigmas around the signal's mean where we accept events, i.e. we accept all events where $m - \mu \leq x\sigma$, where $m$ is e.g. the event's reconstructed mass, $\mu$ is the mean mass, and $\sigma$ is the signal width.

Then the number of signal events that pass the cut is given by

$$\langle S \rangle (x) = s_0 \mathrm{erf}(x/\sqrt{2}),$$

where erf is the error function defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int\limits_0^x \exp(-t^2)\mathrm{d}t.$$

(Here, $\pi$ is the usual number $\pi$, not the purity!) Consequently, the efficiency is given by

$$\epsilon(x) = \mathrm{erf}\,(x).$$

The number of background events that pass the cut is

$$\langle B \rangle (x) = \frac{1}{2} b_0 x,$$

so one has

$$\pi(x) = \frac{s_0 \mathrm{erf}(x)}{s_0 \mathrm{erf}(x) + \frac{1}{2} b_0 x} = \frac{\mathrm{erf}(x)}{\mathrm{erf}(x) + \frac{b_0}{2 s_0} x}.$$

Using Mathematica, one can plot the resulting values of $\epsilon \cdot \pi$, see Fig. 1. One can also use Mathematica to evaluate the cut value that maximizes $\epsilon \cdot \pi$, as shown in Fig. 2. It is interesting to note that the optimal value is generally much smaller than the value 2 that is often used, typically around 1.3.

Fig. 3 (left) shows the relative improvement of $\epsilon \cdot \pi$ that is achieved by using the optimal cut value instead to cut at $2\sigma$. One can see that one cannnot gain much for values of $0.1 < s_0/b_0 < 10$.

However, it should be noted that for high values of $s_0/b_0 > 5$, a cut at e.g. $3\,\sigma$ is better, also because then 99.5 % imnstead of "only" 95.5 % of the events are included in the signal region, which will decrease the systematic uncertainty due to an uncertainty in the signal shape.

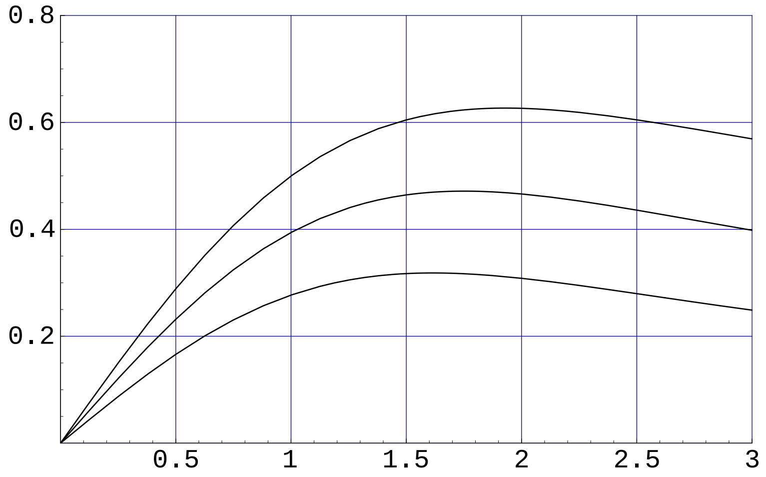


Figure 1: Efficiency times purity for a relation $s_0/b_0 = 0.5, 1, 2$ as function of the cut value $x$.

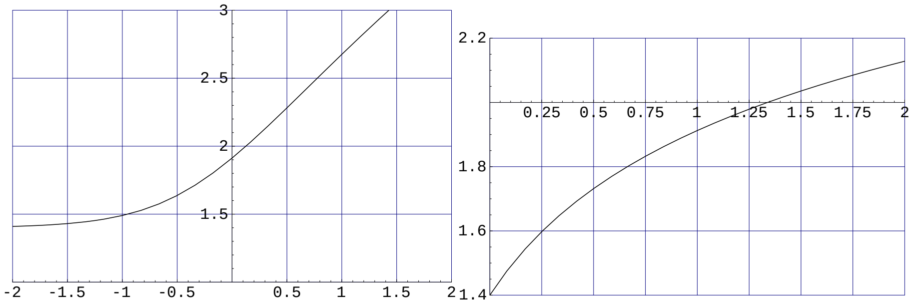


Figure 2: The optimal cut value $x$ as function of $\log_{10}(s_0/b_0)$ (left) and $s_0/b_0$ (right).

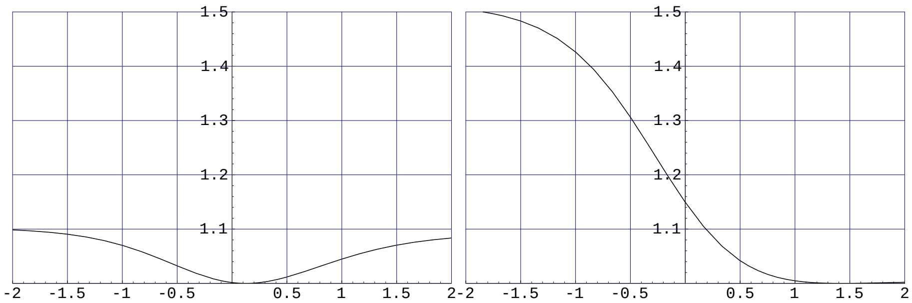


Figure 3: The Improvement of $\epsilon \cdot \pi$ when one uses the optimal cut instead of a cut at $x = 2$ (left) or $x = 3$ (right) as function of $\log_{10}(s_0/b_0)$.