LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO –
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Technical Note	LIGO-T030113-00-D	04/25/2003
Optimum integration length for computation of the cross-correlation coefficient		
S. Mohanty, Sz. Márka, S. Mukherjee, R. Rahkola, R. Frey		

Max Planck Institut für Gravitationsphysik
Am Mühlenberg 1, D14476, Germany
Phone +49-331-567-7220
Fax +49-331-567-7298
E-mail: office@aei.mpg.de

California Institute of Technology
LIGO Laboratory - MS 18-34
Pasadena CA 91125
Phone (626) 395-212
Fax (626) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Laboratory - MS 16NW-145
Cambridge, MA 01239
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

www: http://www.ligo.caltech.edu/

1 Statement of problem

The cross-correlation coefficient r is defined as,

$$r = \frac{\sum(x_k - \widehat{\mu}_x)(y_k - \widehat{\mu}_y)}{\sqrt{\sum(x_k - \widehat{\mu}_x)^2}\sqrt{\sum(y_k - \widehat{\mu}_y)^2}} , \quad (1)$$

$$\widehat{\mu}_x = \frac{1}{N}\sum x_k , \ \widehat{\mu}_y = \frac{1}{N}\sum y_k , \quad (2)$$

where $\{x_k\}$ and $\{y_k\}$, $k = 0, \ldots, N-1$ are two time series segments and all the summations range over $[0, N-1]$. Following Marka, we call N the *integration length*.

We would like to arrive at an "optimum" choice of integration length for a given signal duration.

2 Criterion for optimality

Marka adopted the following procedure for fixing an optimum integration length in a signal injection based upper limit estimation [1]. For a fixed signal waveform and a given integration length, the histogram of r values were plotted for the two cases (a) signal present and, (b) signal absent. One expects that the scatter in r will decrease with an increase in N since one is averaging over a larger number of samples. At the same time, the difference in the mean values of the two histograms will decrease with increase in N since one is bringing in more and more noise at the expense of the signal. The integration length was varied and a [visual inspection?] suggested an optimum value for the integration length where the difference in mean compared most favorably with the scatter in the r values.

We derive, from the approach outlined above, a more formal criterion for optimality. Imagine computing a *running* correlation coefficient from the two data streams, i.e., calculate r by sliding a rectangular window N samples long over each of the two data streams. When the input is just noise, the output from the running r will be smoothed noise. When a signal is present, there will be a bump in the output lasting $\sim N$ samples. The "visibility" of this bump can be measured by the ratio of the mean value of r, for a segment containing the full signal, to the standard deviation of the running r output in the absence of a signal. This is the *signal to noise ratio*, ρ_r, for a detection based on r. Based on this intuitive picture, the criterion we adopt is the following: For a given signal, the optimum integration length, $N_{\rm opt}$ is the one for which ρ_r is maximized.

3 Analytic result

This is a crude calculation! Needs to be improved.

Let the two data streams be uncorrelated, zero mean, unit variance, Gaussian white noise sequences in the absence of a signal. From the limiting normal

density [3] for the probability density $p(r)$ of r, we get $1/\sqrt{N}$ as the standard deviation for the running r output. One could use the *exact* probability density given in Eq. (3) of [3] but that leads to problems, in a numerical computation, for values of $N \geq 90$. For simplicity, we use the normal density approximation instead. From [2], we get the expression for r in the presence of identical signals $h[k]$,

$$\begin{aligned} r &= \frac{\sum_i (n_x[i]+h[i])(n_y[i]+h[i])}{||n_x+h||||n_y+h||} \,, \\ &= \frac{\langle n_x, n_y\rangle + ||h||\left(||h|| + \langle n_x, \widehat{h}\rangle + \langle n_y, \widehat{h}\rangle\right)}{\sqrt{||n_x||^2 + ||h||\left(||h|| + \langle n_x, \widehat{h}\rangle\right)}\sqrt{||n_y||^2 + ||h||\left(||h|| + \langle n_y, \widehat{h}\rangle\right)}} \,, \quad (3) \\ \widehat{h} &= \frac{h}{||h||} \,. \quad (4) \end{aligned}$$

Taking the ensemble average on both sides of Eq. 3, we get,

$$E[r] = \frac{\rho^2}{N+\rho^2} \,, \quad (5)$$

where ρ is the matched filtering signal to noise ratio. We can re-express ρ as $\rho^2 = M h_{\rm rms}^2$,

$$h_{\rm rms} = \sqrt{\frac{1}{M}\sum_{i=0}^{M-1} h_i^2} \,. \quad (6)$$

where M is the length of the signal [1].

Thus, ρ_r is,

$$\rho_r = \frac{E[r]}{1/\sqrt{N}} = \frac{\sqrt{N}}{Na+1} \,, \quad (7)$$

where $a = 1/(M h_{\rm rms}^2)$. The maximum of ρ_r occurs at,

$$\begin{aligned} N_{\rm opt} &= 1/a \,, \\ &= M h_{\rm rms}^2; . \quad (8) \end{aligned}$$

Thus, one sees that for weak signals $h_{\rm rms} \sim 1$, $N_{\rm opt} \sim M$ but increases quite rapidly with $h_{\rm rms}$ (or equivalently ρ).

4 Monte Carlo simulation

We perform Monte Carlo (MC) simulations to compute $N_{\rm opt}$. The simulations use the same noise model as used in Section 3.

[1] Note that this automatically restricts us to $N \geq M$ since we have assumed that the entire signal is contained within the segments. Also, we must assume the mean of the signal waveform to be zero. Otherwise, a correction factor [2] enters ρ which depends on the signal mean and the signal duty cycle.

We use a MATLAB script for the MC simulations [4]. The pseudo-code for the script follows.

1. Fix a signal waveform, $h(t)$. Fix the sampling frequency, f_s, for the data. The signal must have a finite duration of M samples. fix the number of trials, $N_{\rm trials}$.

2. Loop over ρ.

   (a) Loop over N. Constraint: $N \geq M$.

      i. Calculate $1/\sqrt{N}$, the standard deviation of the running r output.

      ii. Loop over $k = 1$ to $N_{\rm trials}$

         A. Inject $h(t)$ into pairs of independent white Gaussian noise sequences of length N with mean zero and unit variance.

         B. Compute r. Store.

      iii. goto 2(a)ii.

      iv. Compute the mean, $\widehat{\mu}_r$, of r.

   (b) goto 2a.

3. Plot $\widehat{\mu}_r\sqrt{N}$ versus N (c.f., Eq. 7).

4. goto 2.

The results are shown in Figures 2.

5 Results

Fig 2 was generated using the Gaussian signal shown in Fig 1. The duration of this signal was $M = 10$. The amplitude was scaled according to ρ. The other parameters were $N_{\rm trials} = 5000$ and $f_s = 800$ sec. (Thus, the signal was 12.5 msec in duration.) The value of f_s was chosen to reflect the bandwidth (200 to 600 Hz) that has been used thus far.

The MC results are in qualitative agreement with the analytic result (c.f., Sec. 3).

1. The value of $N_{\rm opt}$ (location of maximum for each curve in Fig. 2) increases with increase in ρ or equivalently $h_{\rm rms}$. (We have not yet verified whether $N_{\rm opt} \propto \rho^2$ or not but it certainly is a faster than linear dependence.)

2. For small $h_{\rm rms}$ or the weak signal case, $N_{\rm opt} \sim M$. For the lowest ρ of 2.0, $N_{\rm opt} \simeq 2.0 \times M$. (The analytic calculation in Sec 3 is certainly not correct. We just took the ensemble average of the numerator and denominator independently. See Fig 3 and the associated caption.)

(Note: the Gaussian signal violates the stipulation that the mean of the signal should be zero which underlies some of the expressions derived in Section 3. However, we are only seeking qualitative results here.)

6 Discussion

The main result in this note is the near quadratic dependence of the optimum value of integration length on the signal to noise ratio (equivalently $h_{\rm rms}$) of the signal. This result is obtained both analytically and via simulations. **This introduces a complication in detection strategies based on the use of r.** Unlike a *linear filter* such as a matched filter, the "filter bank" (i.e., values of N) for an r based strategy will also depend on the amplitude of a given signal type. This is expected, in general, for non-linear detectors.

One of the issues in using r in external trigger analysis has been the probability density of r and whether a normal density fit is acceptable or not. As shown in [3], the probability density function of r is not well approximated by a normal density for small sample numbers ($N \sim 10$). The question then was whether we would ever need to use integration lengths as small as this.

The conclusion seems to be in favour of just sticking to the normal density even though the optimum integration length is indeed $\sim M$ (the signal duration) for weak signals. A detection strategy will usually not be concerned with signal to noise ratios ≤ 4. For interval estimation, we do require the probability density for all values of signal to noise ratios and hence, the weak signal case is of importance. On the other hand the deviation between the probability densities is only important well into the tails, which may not matter much for interval estimation. It appears then that for a *single trigger* case, the use of a normal density for r is a good approximation.

Some more thought is needed to analyze the situation of *multiple triggers* because there we are indeed dealing with individually weak signals that lead to detection only when combined together.

References

[1] S. Marka, external triggers telecon minutes.

[2] Soumya D. Mohanty *et al*, LIGO-T030111-00-D.

[3] Soumya D. Mohanty *et al*, LIGO-T030112-00-D.

[4] Soumya D. Mohanty, *optintlength.m*, External Triggers CVS code/SDM_code/matlab.

[5] S. Marka *et al*, "Initial Scan of correlated behaviour of LIGO interferometers around the outstanding HETE trigger GRB03029 ", LIGO-T030069-00-D.

[6] Soumya D. Mohanty, *optintlength2.m*, External Triggers CVS code/SDM_code/matlab.

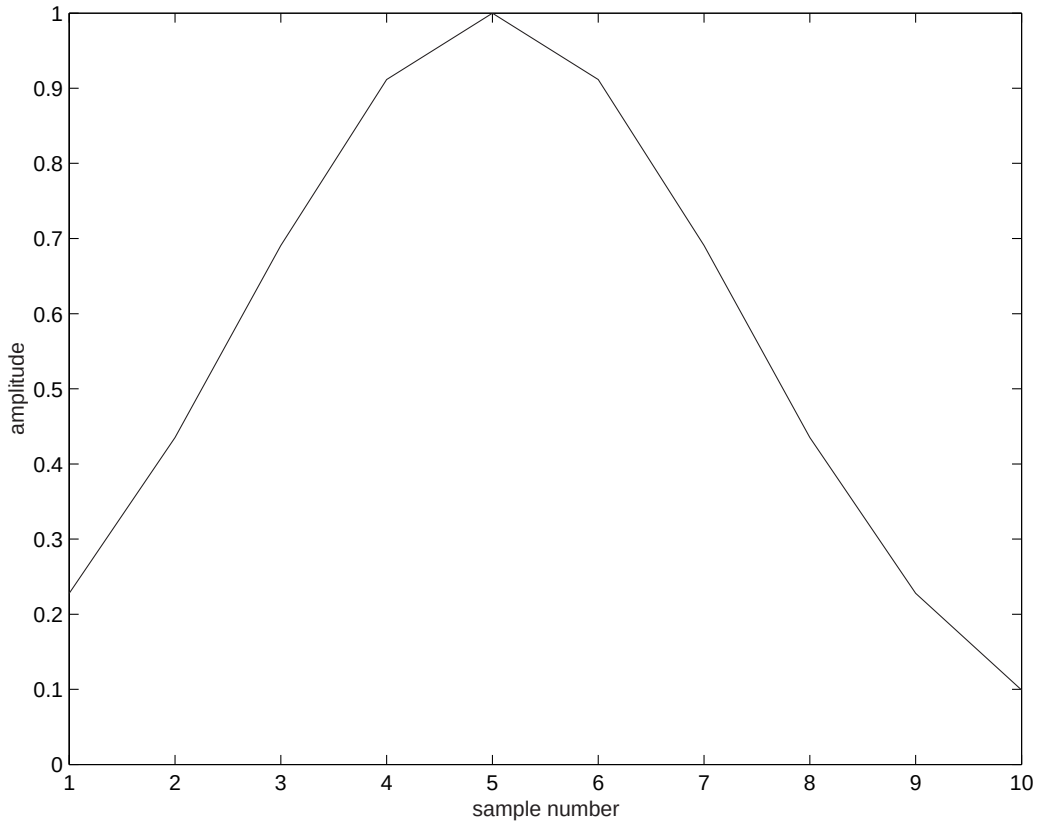

Figure 1: Signal Waveform.

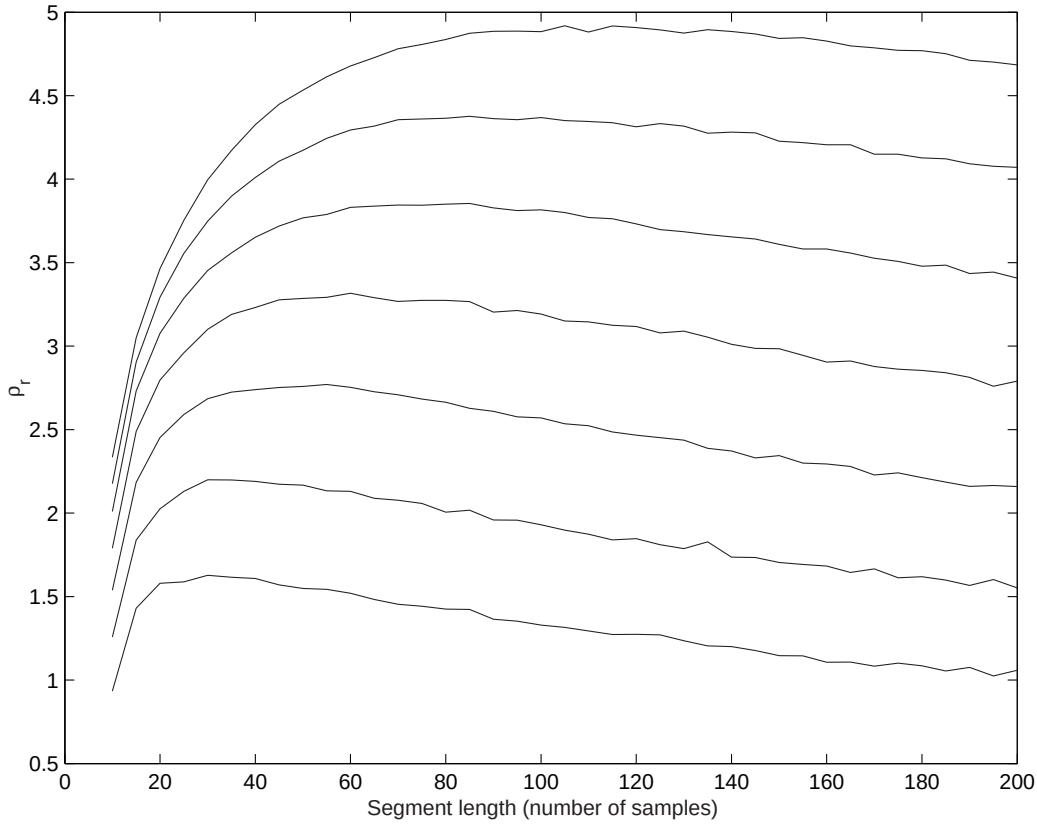

Figure 2: The value of $N_{\rm opt}$ is the location of the peak for each curve. The matched filtering signal to noise ratio increases from the bottom curve to top, $\rho = [2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]$. This figure was produced using the MATLAB code [4] (v1.2).

.

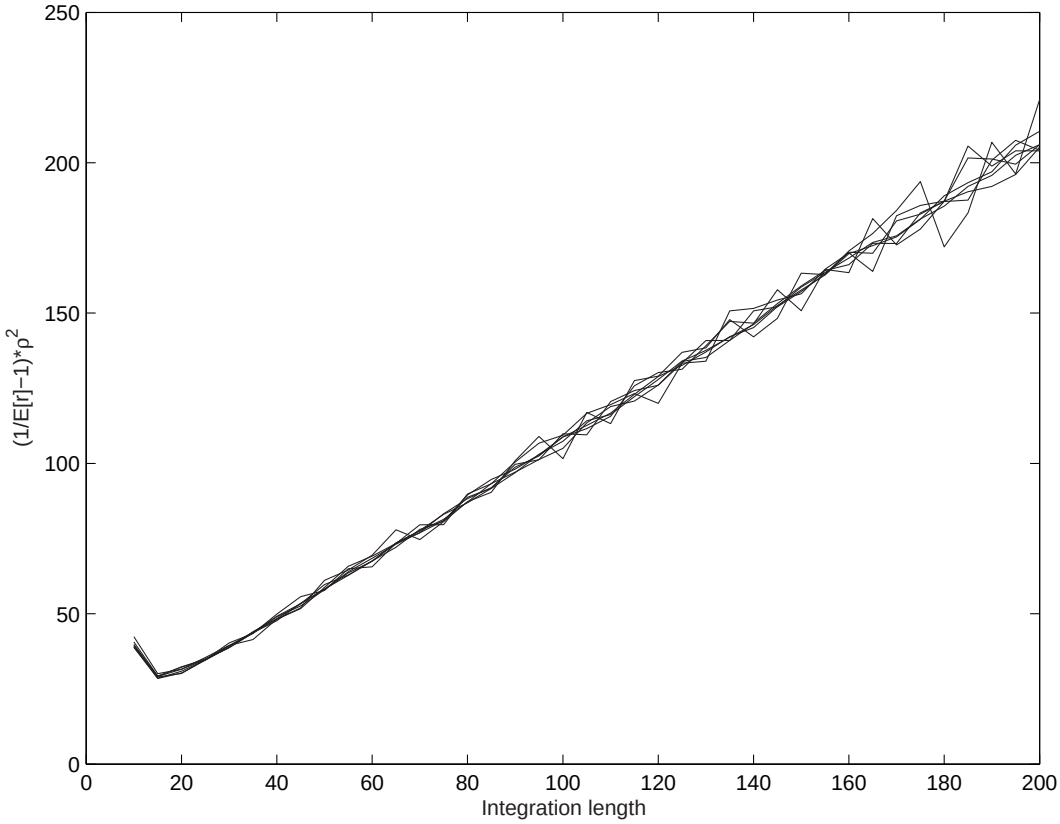

Figure 3: The variation of $\rho^2(1/E[r]-1)$ w.r.t N. According to Eq. 5, this quantity should be equal to N. The curves in this figure correspond to $\rho = [2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]$ and were produced using the Gaussian signal in Fig **??**. One does observe the predicted linear dependence though the slope is slightly different. More importantly there is a dip near $N \sim M$ which is neither predicted nor expected. This could be a simulation artifact but its origin needs to be resolved lest it affects results from other simulations. Apart from the dip the MC results seem to suggest that the calculations in Sec 3 are along the right track though off by, not very large, factors. This figure was produced using the MATLAB code [6] (v1.1).

.