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

Technical Note	LIGO-T030111-00-D	04/22/2003
Notes on the sampling distribution of the correlation coefficient and normal density fits		
S. Mohanty, Sz. Márka, R. Frey, R. Rahkola, S. Mukherjee		

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 Definitions

The primary reference for Section 1 and 2 is [1].

This note is concerned with the *correlation coefficient* estimator 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 $[0, N-1]$.

2 Sampling distribution of r

Consider the case where the samples within each time series are independent and identically distributed. Let x_i and y_j have a bivariate normal[1] distribution,

$$N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \delta_{ij}\rho\sigma_x\sigma_y \\ \delta_{ij}\rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right] ,$$

where δ_{ij} is the Kronecker delta, $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. The above means that each sample also has a normal distribution and that samples x_i and y_i have a correlation coefficient ρ and x_i is not correlated with y_j if $i \neq j$. The quantity r defined in Eq. 1 is an *estimator* of ρ.

The probability density function (pdf) $p(r)$ of r under the above conditions is given in [1]. The information that is relevant to us is,

1. $p(r)$ should be independent of μ_x, μ_y, σ_x and σ_y. This follows from the invariance of r w.r.t an overall change in location and scale. Thus, this statistic makes us robust against non-stationarity in the mean and variance of each time series provided the non-stationarity is not significant within the on-source or off-source segments themselves. On the other hand, the distribution of the cross-correlation statistic alone which is the numerator of Eq. 1 does depend on σ_x and σ_y.

2. In the case of *off-source* data when there is no GW signal and assuming that the noise in the two time series is uncorrelated ($\rho = 0$), we can obtain a closed form expression for $p(r)$,

$$p(r) = \frac{\Gamma\left[\frac{1}{2}(N-1)\right]}{\Gamma\left[\frac{1}{2}(N-2)\right]\sqrt{\pi}}(1-r^2)^{\frac{1}{2}(N-4)} . \quad (3)$$

   There is no closed form expression for $p(r)$ when $\rho \neq 0$ but analytical expressions exist in the form of infinite series and can be obtained from [1].

[1] Same as a Gaussian distribution. The notation $N(\mu, \mathbf{\Sigma})$ denotes a multivariate Gaussian distribution with mean μ and covariance matrix $\mathbf{\Sigma}$.

3. Although $p(r)$ is not a normal density, it can be approximated by one. See section 2.1 and 4.

4. For *on-source* data containing a GW signal $h(t)$, the mean value of x_k becomes $h_k + \mu_x$ which is is not a constant across sample index. (Similarly for y_k.) The density of $p(r)$ is not tabulated in [1] for this case. This is the subject of a separate note [2].

2.1 Limiting distribution of r

As one would suspect, the limiting form (as N becomes large) of $p(r)$ is a normal density. The limiting Normal density (for arbitrary ρ) is,

$$p(r) \to N(\rho, (1-\rho^2)^2/N) \ , \tag{4}$$

which corresponds to mean ρ and variance $(1-\rho^2)^2/N$. For $\rho = 0$, we get $N(0, 1/N)$ as the limiting density.

It can be shown that the quantity

$$z = \ln\left[\frac{1+r}{1-r}\right] \ , \tag{5}$$

knows as Fisher's z statistic, converges to a limiting normal distribution whose variance is *independent* of ρ unlike the limiting distribution of r (c.f., Eq. 4). This convergence is also faster than that of the distribution of r alone. We have not explored the use of Fisher's z further so far.

3 Validity of Normal density fit

We would like to quantify the range of N for which the exact form of $p(r)$ as given in Eq. 3 is approximated poorly by a normal density fit (which could be measured from a histogram of r values for instance). As stated earlier, the normal approximation becomes better as N becomes large. The question is how large is "large".

For a given signal, It turns out [3] that the optimum value of the integration length depends quadratically on the rms amplitude, $h_{\rm rms}$, of the signal and linearly on the signal duration. For signal rms ampltiude $h_{\rm rms} \sim 1$, the optimum integration length is close to the signal duration. Thus, for signals with short duration *and* $h_{\rm rms} \sim 1$, a sensible detection strategy would use short integration lengths. This is not a very interesting situation, however, as far as a single trigger analysis goes since such signals would be too weak to allow a high confidence detection. But in a multiple trigger analysis where individually weak signals are combined to improve the signal to noise ratio, we can imagine using short integration lengths.

3.1 Validity of asymptotic Normal form

We will do some numerical experiments for the $N = 10$ case. For comparison, we will also cite the $N = 20$ results. One way to quantify the deviation between the exact and the approximate Normal density is to look at the difference in significance [4] that one would compute using the two densities. This is also motivated by the fact that a mutiple trigger analysis might in some cases combine only that part of data from each on-source segment which has a "small" significance.

Consider the density functions given by Eq. 3 and the asymptotic form in Eq. 4. Fig. 1 shows the significance as a function of $|r|$ for the two densities. From Fig. 1, one sees that the significances calculated from the exact and approximate densities begin to diverge quite a bit for values of significance $\sim 10^{-3}$ and smaller. For $N = 10$, $|r| \simeq 0.85$ gives a significance of $\simeq 10^{-3}$ when the exact form of $p(r)$ is used while the significance is $\simeq 10^{-2}$ for the Normal density.

Moreover, if one fixed a threshold on $|r|$ corresponding to a given significance, the two densities would give substantially different values. Fortunately for us, the Normal density gives more conservative thresholds. More importantly, note that for sufficiently low values of significance, *there is no solution for a threshold on* $|r|$ *within the valid range* $(0, 1]$ *when the normal density is used.*

3.2 Validity of empirically determined Normal fit

What about fitting a normal density to $p(r)$ instead of using the asymptotic, $N(0, 1/N)$, normal density? Maybe the best fit normal density for a given N is not the asymptotic normal density. Fig. 2 shows the χ^2,

$$\chi^2 = \int_{-1}^{1} d\,r\ (p(r) - N(0, \sigma^2))^2 / N(0, \sigma^2)\ , \tag{6}$$

between $p(r)$ and a normal density for different value of σ (in units of $1/\sqrt{N}$). The use of a χ^2 measure is motivated by the fact that we would use it to find the best fit normal density if we did not know $p(r)$ but had access to a histogram of r values for an arbitrarily large number of trials.

From Fig. 2 one sees that (1) the best normal density fit to $p(r)$ is already close to $N(0, 1/N)$ (c.f., the minimum value of χ^2) and (2) the best fit approximation becomes closer to $N(0, 1/N)$ as N increases, as it should. Thus our conclusions made earlier with $N(0, 1/N)$ remain valid. Moreover, the actual minimum of χ^2 occurs at $\sigma > 1/\sqrt{N}$. This means that the problem of indeterminable thresholds actually becomes worse with a best fit.

4 Summary

From the above numerical results we conclude that a normal density fit to the distribution of r leads to large errors for significance or threshold calculations when $N \sim 10$ and significance $\geq 10^{-3}$. However, the situation improves quite

fast as N increases (c.f., the numerical results for $N = 20$). This discussion gives us an idea of the typical values of N and significance for which we may expect problems if we use normal density fits to the distribution of $|r|$.

References

[1] T. W. Anderson, *An introduction to Multivariate Statistical Analysis*, second edition (John Wiley, 1984).

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

[3] Soumya D. Mohanty *et al*, LIGO-T030113-00-D.

[4] Significance is defined as the probability of obtaining a value of the correlation coefficient $\geq$ the *observed* value of r due to noise alone. If we were to fix a threshold for detection on $|r|$, then the probability of the observed value exceeding this threshold due to noise alone is called the *false alarm probability*.

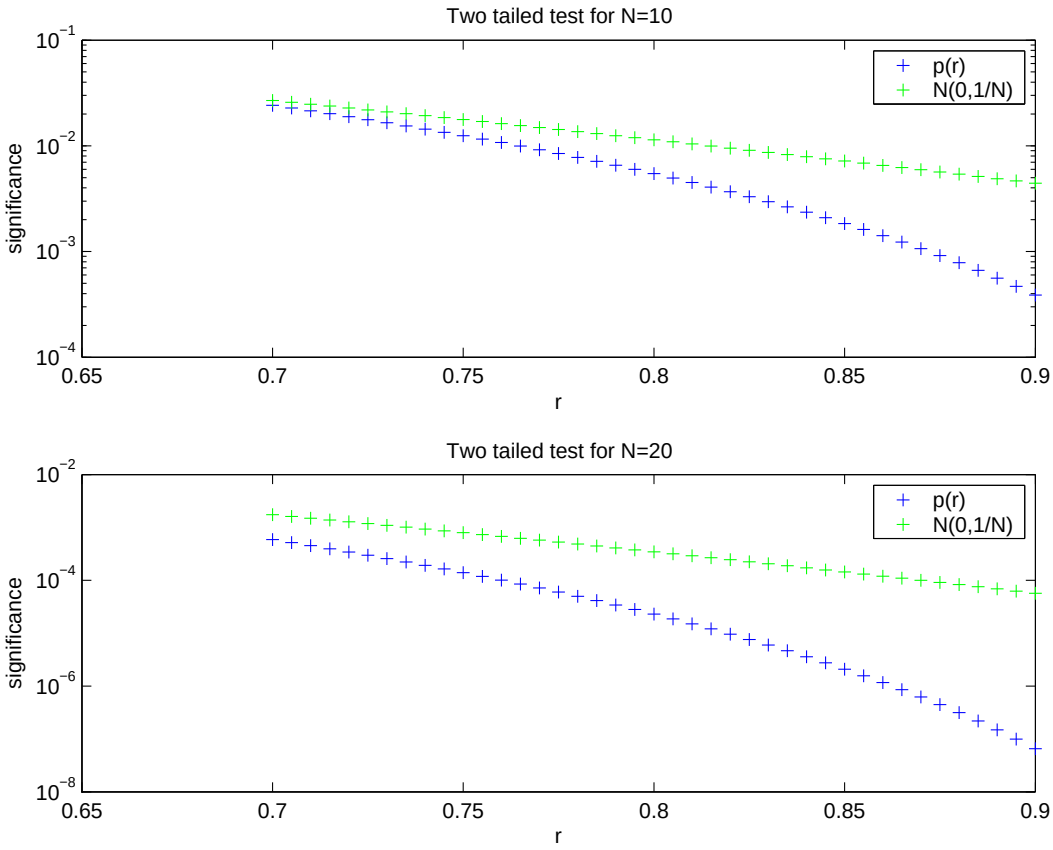
Two tailed test for N=10
p(r)
N(0,1/N)
significance
r
Two tailed test for N=20
p(r)
N(0,1/N)
significance
r

Figure 1:

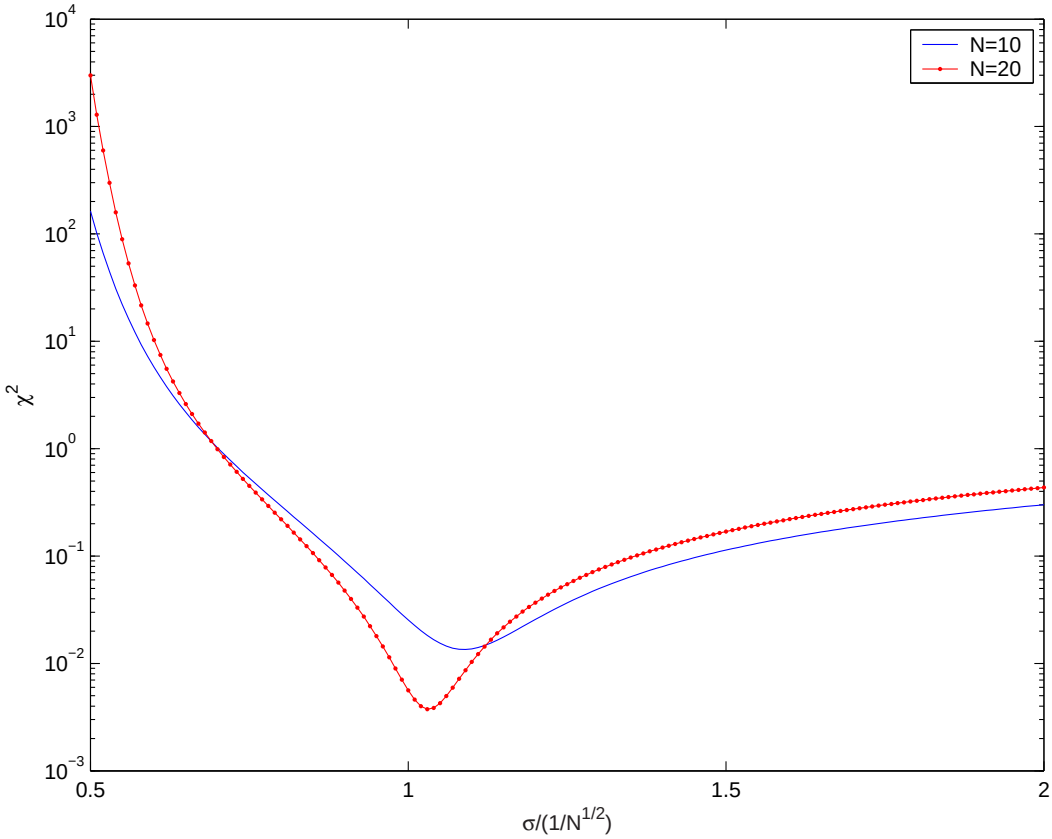

Figure 2: