$\mathcal{F}$-statistic bias due to noise-estimator

Reinhard Prix

LIGO-T1100551-v1 July 2006

1 Overview

- As reported by Iraj (see here) there appears to be a positive bias in the expectation-value of the $\mathcal{F}$-statistic in the pure-noise case, of about $1-2\%$ with a running-median size of 50. A similar effect has been reported also by Greg in the context of StackSlide (see here)
- Here we investigate the quantitative bias expected from the finite-sample estimator of the power-spectral density, namely the bias due to the fact that in general

$$E\left[\frac{1}{x}\right] \neq \frac{1}{E[x]}\,. \tag{1}$$

- Note: the effect investigated here has *nothing* to do with using the median versus the mean. In fact we'll be asssuming that we use the *mean* to estimate the power.

2 $\mathcal{F}$-statistic of pure noise

Schematically, the $\mathcal{F}$-statistic in the noise-only case can be written as

$$2\mathcal{F} = \sum_{\mu=1}^{4} \frac{n_\mu^2}{E[n_\mu^2]}\,, \tag{2}$$

where n_μ are 4 Gaussian random variables with zero-mean and variance σ^2, i.e. $E[n_\mu] = 0$ and $E[n_\mu^2] = \sigma^2$. It is obvious that in this case we have $E[2\mathcal{F}] = 4$.

In order to compute the $\mathcal{F}$-statistic, we therefore need to compute quantities the form

$$\zeta \equiv \frac{n_\mu^2}{E[n^2]}\,. \tag{3}$$

In practice, we use a finite-sample estimator for the variance $E[n^2]$, and for the sake of simplicity here we use the mean (instead of the median), namely

$$P_N \equiv \left\langle n^2 \right\rangle_N = \frac{1}{N}\sum_{i=1}^{N} n_i^2\,, \tag{4}$$

where the n_i are different (uncorreleated) noise-realizations. It is obvious that P_N is an *unbiased* estimator for σ^2, namely

$$E[P_N] = \sigma^2 = E[n^2]\,. \tag{5}$$

It is straightforward to show for the Gaussian variables n that

$$E[n^4] = 3\sigma^4\,, \tag{6}$$

and therefore we find

$$\begin{aligned} E[P_N^2] &= \frac{1}{N^2}\sum_{i,j=1}^{N} E[n_i^2\, n_j^2] \\ &= \frac{1}{N^2}\left(\sum_{i=j}^{N} E[n^4] + \sum_{i\neq j}^{N} E[n^2]^2\right) \\ &= \frac{1}{N^2}\left(3N\,\sigma^4 + N(N-1)\sigma^4\right) \\ &= \sigma^4\left(1 + \frac{2}{N}\right)\,. \end{aligned} \tag{7}$$

For $N \gg 1$ the noise-floor estimator P_N will be well approximated (central-limit) by a Gaussian distribution with mean μ_N and variance σ_N^2 given by

$$\mu_N = \sigma^2, \quad \text{and} \quad \sigma_N^2 = \frac{2\sigma^4}{N},\,. \tag{8}$$

3 The bias in the inverse noise-floor estimator

Although P_N is an unbiased estimator for σ^2, the expression for the $\mathcal{F}$-statistic (2) involves (four) terms of the form

$$\zeta_N \equiv \frac{n_\mu^2}{P_N}\,. \tag{9}$$

So the question is: given a Gaussian variable $y \equiv P_N$ with mean μ and variance σ^2, what is the expectation-value of $z = 1/y$? (Note that in practice there are most likely also correlations between n^2 and P_N, as we will be using 32 frequency-bins to compute the nominator, and 50 frequency-bins to estimate the noise-floor in the denominator. However, for very large N this effect should also vanish, and we neglect it here for "first-order simplicity" of the argument).
The Gaussian distribution of y is

$$\mathrm{pdf}[y] = f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \,, \tag{10}$$

and the distribution of $z = 1/y$ can be found as

$$\mathrm{pdf}[z] = g(z) = \frac{1}{z^2}\, f\left(\frac{1}{z}\right) \,. \tag{11}$$

What we want to estimate is $\bar{z} \equiv E[1/y]$, which is given by

$$\begin{aligned} \bar{z} &= \int_{-\infty}^{\infty} z\, g(z)\, dz = \int_{-\infty}^{\infty} \frac{1}{z}\, f\left(\frac{1}{z}\right) dz \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{1}{z} e^{-\left(\frac{1}{z}-\mu\right)^2/2\sigma^2}\, dz \,. \end{aligned} \tag{12}$$

Introducing a new dimensionless integration-variable ε as

$$\frac{1}{z} = \mu(1-\varepsilon) \,, \tag{13}$$

which gives

$$\frac{dz}{z} = \frac{d\varepsilon}{1-\varepsilon} \,, \tag{14}$$

the integral (12) now reads as

$$\bar{z} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{1}{1-\varepsilon}\, \exp\left[-\frac{\varepsilon^2}{2(\sigma/\mu)^2}\right] d\varepsilon \,. \tag{15}$$

If the Gaussian distribution of y has a maximum well-separated from zero, i.e. $\mu \gg \sigma$, then the exponent will vanish rapidly for $\varepsilon \gtrsim (\sigma/\mu)^2$, and we can therefore assume $\varepsilon \ll 1$ and expand

$$\begin{aligned} \bar{z} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \left(1+\varepsilon+\varepsilon^2+...\right) \exp\left[-\frac{\varepsilon^2}{2(\sigma/\mu)^2}\right] d\varepsilon \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \left(\sqrt{2\pi}\frac{\sigma}{\mu} + 0 + \sqrt{2\pi}\left(\frac{\sigma}{\mu}\right)^3 + ..\right) \\ &= \frac{1}{\mu}\left(1 + \left(\frac{\sigma}{\mu}\right)^2 + ...\right) \,. \end{aligned} \tag{16}$$

This shows there is a *positive* bias of $E[1/y]$ with respect to $1/E[y]$, which to first order is given by $(\sigma/\mu)^2$ in terms of the mean and variance of the underlying Gaussian distribution of y.
Applying this on (9) and assuming (for simplicity) the nominator and denominator to be independent, one gets

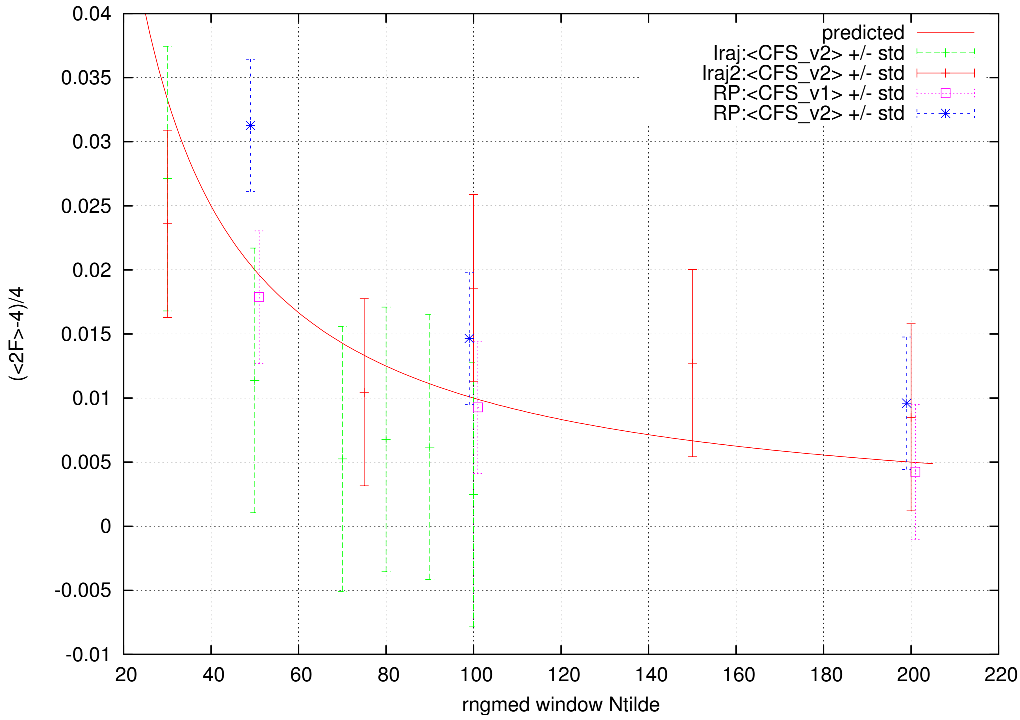
$$E[\zeta_{\mathcal{N}}] = E[n^2]\, E\left[\frac{1}{P_N}\right] = \sigma^2 \frac{1}{\mu_N}\left(1 + \left(\frac{\sigma_N}{\mu_N}\right)^2 + ...\right) = 1 + \frac{2}{N} + ... \quad (17)$$

There is a bias of $+2/N$, where N is the number of independent samples used to compute the mean of n^2. When using a window $\tilde{N}$ in the frequency-domain, this corresponds to $N = 2\tilde{N}$, as each frequency-bin contains two independent samples.
Using the default windowsize of $\tilde{N} = 50$, one would find a bias of

$$\frac{2}{N} = \frac{2}{2 \times 50} \approx 0.02\,. \quad (18)$$

4 Comparison to measurement

Taking measured results for $\langle 2F \rangle$ as a function of running-median window-size $\tilde{N}$ and comparing this to the prediction (17):



The error-bars on the measurements are $\pm\sigma_{\langle \mathcal{F} \rangle}$ of the F-average, given by

$$\sigma_{\langle \mathcal{F} \rangle} = \sqrt{\frac{8}{n_0}}\,, \tag{19}$$

where n_0 is the number of trials used to compute $\langle F \rangle$. Iraj's measurements used $T = 40$ hours (and L1), my measurements were for $T = 50$ hours and H1 ($T_{\mathrm{SFT}} = 30$ mins).

5 Discussion

- My measurements using CFSv1 (300,000 *independent* trials) seem to agree well with the prediction, but the CFSv2 results appear to have a slight additional bias.
  Could that come from the additional noise-weighing in the antenna-pattern functions??
- Why was the CFSv1-bias not seen in an earlier check discussed link?