LIGO-T1100497

Degrees-of-freedom estimation in the Student-t noise model.

Christian Röver

September 22, 2011

1 Introduction

The Student-t noise model was introduced in [1] as a robust alternative to the commonly used stationary, Gaussian noise model that constitutes the basis for many signal processing procedures. This model may also be used to generalize the matched filter in order to make it less outlier-sensitive [2]. The generalized Student-t model includes an additional set of degrees-of-freedom parameters ν_j along with the power spectral density $S(f_j)$. These d.o.f. parameters effectively specify the degree of heavy-tailedness of the noise that is accounted for or expected through the model. While there are many off-the-shelf procedures available for estimating the PSD $S(f_j)$, it is not so obvious how one would go about estimating the d.o.f. parameters. In the following, we will decribe a first attempt at eliciting these parameters from empirical measurements using some LIGO data. We will be using the notation explicated in [1, 2].

2 The empirical noise power distribution

2.1 The "Student-Rayleigh" distribution

In order to infer the heavy-tailedness behaviour of the (non-Gaussian) noise, it makes sense to reduce the data and only cosider the empirical *noise power* at the jth Fourier bin:

$$\frac{\left|\tilde{n}(f_j)\right|}{\sqrt{\frac{N}{4\Delta_t} S_1(f_j)}} \tag{1}$$

[2, Sec. IV.C]. If the Gaussian noise model was appropriate, then the noise power would be a random variable following a *Rayleigh distribution* (the squared expression would be χ^2-distributed with 2 degrees-of-freedom, or exponentially distributed with rate $\lambda = \frac{1}{2}$). Assuming instead the Student-t

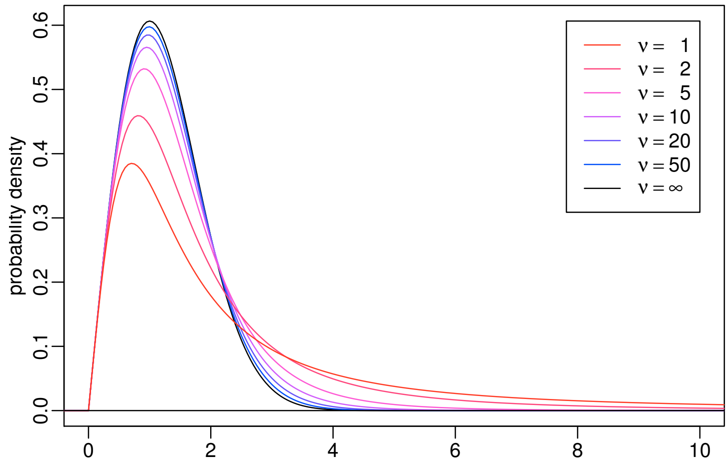

Figure 1: Probability density functions of "Student-Rayleigh" distributions for varying degrees of freedom ν and fixed scale $\sigma^2 = 1$. For $\nu = \infty$, the distribution corresponds to the usual ("Gaussian") Rayleigh distribution.

noise model, the same expression follows a more heavy-tailed probability distribution, whose functional form may be related to Snedecor's F-distribution, and which is explicitly given in [2, Appendix A.4]. We will refer to this distribution as the "*Student-Rayleigh*" distribution in the following.

Figure 1 illustrates the Rayleigh and Student-Rayleigh distributions in comparison. One can see how decreasing degrees-of-freedom values ν imply greater probabilities for extreme values, and how for $\nu \rightarrow \infty$ the distribution converges to the (usual, "Gaussian") Rayleigh distribution. Figure 2 shows how the Student-Rayleigh distribution's upper quantiles grow with decreasing degrees-of-freedom ν. While the median barely changes, the upper quantiles grow quite large for small values of ν.

2.2 Estimating the noise PSD

In most practical data analysis problems, the noise PSD $S(f)$ is a priori unknown and needs to be estimated from the data. A popular PSD estimation procedure is *Welch's method* [3, 4], which simply uses several noise samples to compute the empirical noise power and estimates the PSD via their average. A variety of this procedure commonly employed in LIGO data analysis uses the median instead of the mean in order to gain robustness against outliers.

An important point to note is that the estimation procedure introduces additional variation, or uncertainty, into the noise model. Even if the noise is perfectly Gaussian, with an uncertain, estimated PSD $S(f)$, the distribution of the normalized noise power (1) changes from a Rayleigh distribution to a

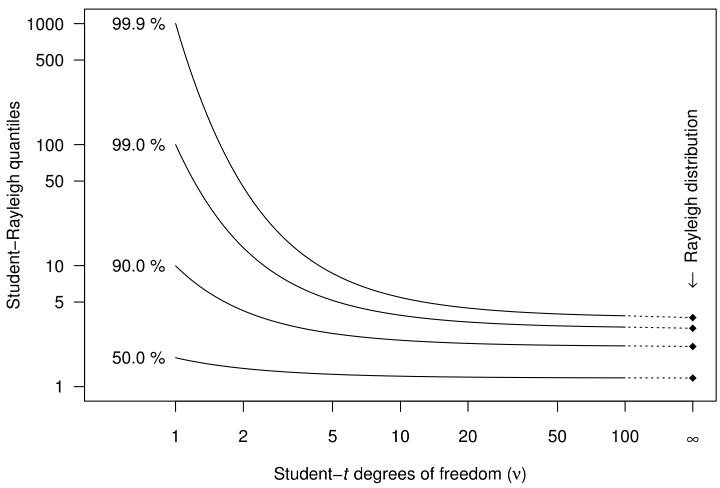

Figure 2: Selected quantiles of the Student-Rayleigh distribution as functions of the degrees-of-freedom parameter ν. The case labeled as $\nu = \infty$ indicates the limiting Rayleigh distribution.

heavier-tailed one.

If the noise PSD was estimated using Welch's method, using an average over k non-overlapping, independent noise samples, then the PSD estimator would follow a (scaled) χ^2_{2k}-distribution, which in turn would make the normalized noise residuals *exactly* Student-t distributed [5], and which would make the noise *power* (1) Student-Rayleigh distributed with $\nu = 2k$ degrees of freedom. In case the median method is used, this distributional form only holds approximately; the appropriate number of degrees-of-freedom is generally lower and should probably be $\approx \frac{2}{\pi} \times 2k$ instead.

2.3 Exploring the empirical noise residuals

Once we have sampled data, we can explore its empirical behaviour and contrast it with theoretical expectations based on the Gaussian or Student-t models. Since we are particularly interested in outliers and the distribution's tail behaviour, a *quantile-quantile plot (Q-Q-plot)* may be an appropriate exploration tool [see 2, Sec. IV.C]. Since an earlier investigation of the noise's (non-) Gaussian behaviour already suggested that the noise distribution changes with the corresponding frequency [6], we will in particular try to investigate the frequency dependence as well.

Figure 3 illustrates quantiles of the empirical normalized noise power (1), binned into 10-Hz-wide frequency bins. Normalisation of the noise power (1) is done in analogy to actual data analysis procedures, according to a PSD estimate based on a median from $k = 32$ preceding noise samples. (Note that this is different from the figures shown in [6], where the PSD was

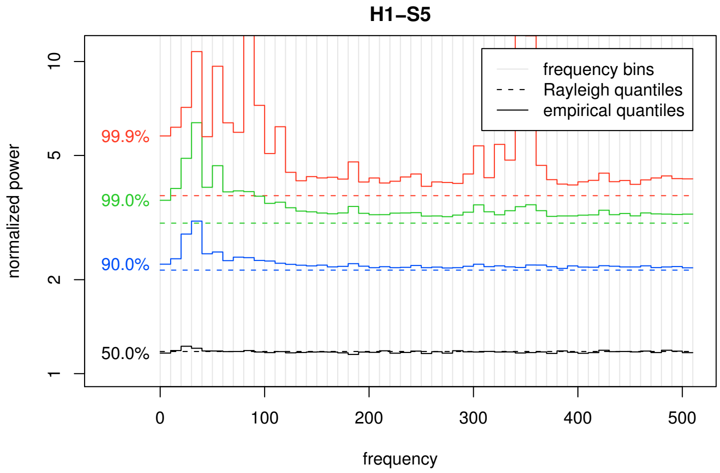

Figure 3: Empirical quantiles of the empirical, binned noise power. This figure is based on 200 noise samples, each 8 seconds long and sampled at sampling rate $\frac{1}{\Delta_t} = 1024\,\text{Hz}$. The noise PSD estimate used for normalisation was based on $k = 32$ equally-sized preceding noise segments. Quantiles are computed based on the distribution of the normalized noise power (1), binned across 10 Hz wide frequency ranges (so each of the 51 frequency bins here contained 16 000 noise samples). The dashed lines indicate the "theoretical" quantiles as implied by a Gaussian model, according to the Rayleigh distribution.

estimated via an *overall* median.) The dashed lines indicate the corresponding expected quantiles under the assumption of the Gaussian noise model (i.e., quantiles of the Rayleigh distribution). Note that while the median is matched well, i.e., the PSD estimation procedure actually does a good job at pinpointing the median power, the higher quantiles increasingly exceed the assumed values, and there also is a clear frequency dependence apparent.

It is important to note that the exact Fourier-domain noise distribution depends on the details of the (discrete) Fourier transform performed in the given analysis. This includes windowing of the data [7, 4], the sample size and sampling rate. Changing any of these will also affect the properties of the Fourier-domain noise $\tilde{n}_j$. When trying to match the noise model to the data, it is hence crucial to use samples that were of the same nature and subject to the same preconditioning as the data in the "actual" analysis.

2.4 Matching the degrees-of-freedom parameter(s)

The question now is how the Student-t model may be used to "fix" the discrepancies apparent in Fig. 3 and make the data analysis model reflect the observed noise properties. For now we will treat the problems of spectrum

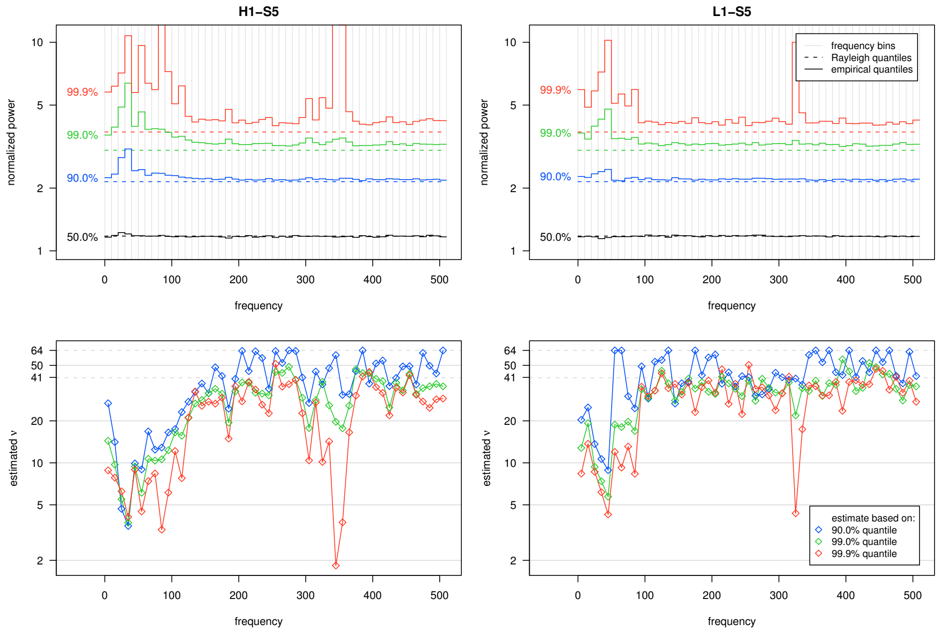

Figure 4: Empirical noise quantiles and estimated degrees-of-freedom parameters ν_j for some S5 data.

estimation and d.o.f. estimation separately. The noise spectrum, i.e., the Student-t-distribution's scale parameter, will be estimated "as usual" via a running median, and independent of the choice of ν_j. In order to fit the degrees-of-freedom parameter ν_j to the data, we will then (attempt to) match a particular Student-Rayleigh quantile to the empirically observed noise quantile. The idea is to take the empirical quantiles as in Fig. 3 and then invert the relationship shown in Fig. 2 to find the parameter ν for which the assumed distribution matches the empirically found one in the particular quantile. The mapping from quantiles to degrees of freedom may easily be derived numerically via a root-finding algorithm. Another sensible restriction to observe is that the fitted value ν in any case should probably not be greater than $2k$, which would be the appropriate number in the "optimal" case of Gaussian noise and a noise PSD derived via a mean estimator.

Figures 4–6 show more empirical quantiles for several pieces of instrumental data as well as simulated Gaussian noise. The corresponding estimated d.o.f. values ν_j, corresponding to the different quantiles, are shown in parallel. The d.o.f. estimates here were restricted to $\nu_j \leq 2k = 64$ (see the dashed line). For simulated Gaussian noise, the estimated values are indeed scattered around the anticipated value of $\nu_j = \frac{4}{\pi}k \approx 41$. The estimates based on different quantiles appear to roughly agree; one exception here is the right panel in Fig. 6, which (as a closer inspection reveals) appears to be

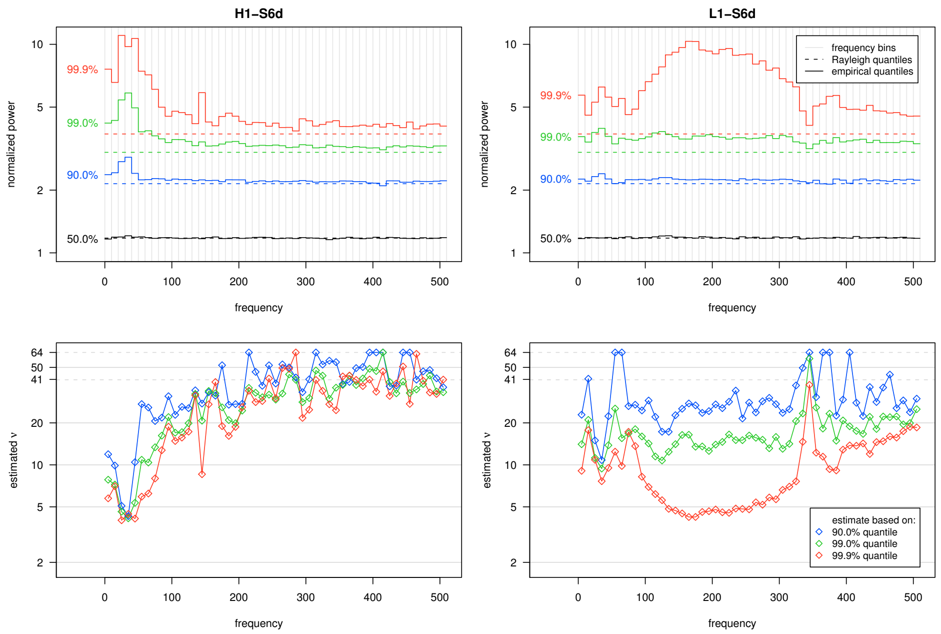

Figure 5: Empirical noise quantiles and estimated degrees-of-freedom parameters ν_j for some S6 data.

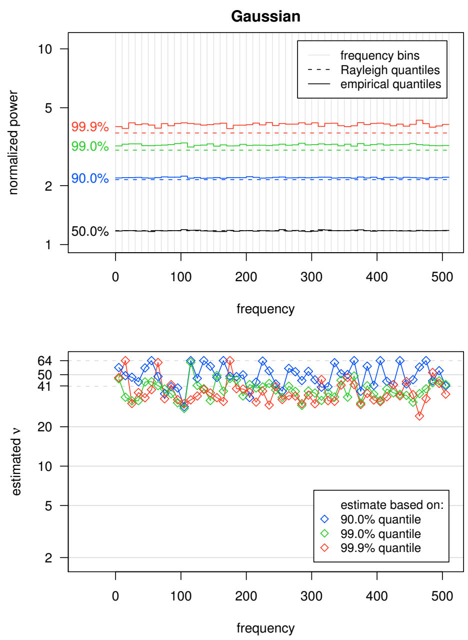

Figure 6: Empirical noise quantiles and estimated degrees-of-freedom parameters ν_j for simulated Gaussian data.

dominated by a transient noise event in this particular piece of data. The frequency-dependence of the noise distribution is also obvious.

3 Conclusions / to do

Estimation of the degrees-of-freedom parameters ν_j via empirical quantiles seems to work and and to give roughly consistent results even when usinh different quantiles. This also suggests that the Student-t distribution appears to match the actual noise properties to some degree.

For a practical application, there is a number of details yet to be determined. For example how (or whether) the binning should exactly be done, or what quantile the estimation should actually be based on. While a lower quantile will show less estimation variance, a higher quantile may be more suitable to reflect rare events. Another question is what time interval the estimates should be based on — one could modify the PSD estimation procedure to also return ν_j estimates, but on the other hand, these may be supposed to also reflect particularly rare events, and so one might want to base them on way longer time ranges. Eventually, the exploration of other ways of degrees-of-freedom estimation may also be worthwhile.

References

[1] C. Röver, R. Meyer, and N. Christensen. Modelling coloured residual noise in gravitational-wave signal processing. *Classical and Quantum Gravity*, 28(1):015010, January 2011.

[2] C. Röver. A Student-t based filter for robust signal detection. *Arxiv preprint 1109.0442v1*, September 2011.

[3] P. D. Welch. The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, AU-15(2):70–73, June 1967.

[4] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C: The art of scientific computing.* Cambridge University Press, Cambridge, 1992.

[5] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the theory of statistics.* McGraw-Hill, New York, 3rd edition, 1974.

[6] S. Waldman. Rayleigh distributions for H1, L1 for S6a. LIGO-Virgo collaboration internal report. URL: `https://www.lsc-group.phys.uwm.edu/ligovirgo/cbcnote/S6Plan/091117085501AnalysisRayleigh_distributions` (CBC Wiki), November 2009.

[7] F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978.