Draft version April 5, 2025
Typeset using LaTeX **modern** style in AASTeX631

A Practical Line Cleaner for Narrow Lines in Gravitational Wave Data

Will M. Farr[1, 2]

[1] *Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA*
[2] *Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA*

ABSTRACT

I present a tool, `line_cleaner`, for fitting and regressing narrow-band lines from timeseries data such as appear frequently in gravitational wave detectors. Residuals produced by this tool can be passed onward to analyses that will benefit from removal of the (semi)coherent lines, as opposed to their suppression by an enhanced spectral density estimate around the line (a form of "notch filter").

1. THEORY

Consider a simple harmonic oscillator driven by a stochastically-fluctuating signal, $n(t)$,

$$\ddot{x}(t) + 2\gamma \dot{x}(t) + \left(2\pi f_0\right)^2 x(t) = n(t), \tag{1}$$

with damping rate γ and natural frequency ω_0. In the Fourier domain, the solution satisfies the algebraic equation

$$-\left(2\pi f\right)^2 X(f) + 4\pi i \gamma f X(f) + \left(2\pi f_0\right)^2 X(f) = N(f), \tag{2}$$

or

$$X(f) = \frac{N(f)}{\left(2\pi f_0\right)^2 - \left(2\pi f\right)^2 + 4\pi i \gamma f}, \tag{3}$$

if we ignore the free, homogeneous solutions to the differential equation (which anyway must damp out over timescales $\tau \gg \gamma^{-1}$). If n is wide-sense stationary and zero-mean, then

$$\left\langle N \right\rangle = 0, \tag{4}$$

$$\left\langle N^*(f) N(f') \right\rangle = S_n(f) \delta(f - f'), \tag{5}$$

will.farr@stonybrook.edu
wfarr@flatironinstitute.org

for some (positive) function S_n, called the power spectral density of the noise n. This implies

$$\langle X \rangle = 0, \tag{6}$$

$$\langle X^*(f) X(f') \rangle = \frac{S_n(f)}{\left((2\pi f_0)^2 - (2\pi f)^2\right)^2 + 16\pi^2 \gamma^2 f^2} \delta(f - f') \equiv S_x(f) \delta\left(f - f'\right). \tag{7}$$

If, furthermore, n is Gaussian distributed, then x will be as well, following an independent Gaussian distribution at each frequency f with mean and variance indicated above. Many narrow-band features (lines) in gravitational wave data can be adequately described by this model.

Assuming $\gamma / f_0 \ll 1$ (i.e. that the oscillator is severely under-damped), over short timescales τ that are longer than the natural period associated to the solution, but short compared to the damping time, i.e. $f_0^{-1} \ll \tau \ll \gamma^{-1}$, the solution is sinusoidal with frequency f_0. Over longer timescales the solution oscillates at the natural frequency, but with randomizing *phase*. For some lines in present-day gravitational wave detectors, the damping time is much longer than the analysis segments of typical signals; in such cases, it is possible to use data from before and after (and during) the analysis segment to *infer* the phase of the line through the analysis segment and *regress* it out of the data, improving the signal-to-noise ratio of the signal. This is the basic idea behind the line cleaner described in this note.

Given a stretch of data in the (discrete) Fourier domain from a total observation time T,

$$D_k \simeq D\left(f = \frac{k}{T}\right) = \frac{T}{N} \sum_{n=0}^{N-1} d_n e^{-2\pi i k n / N}, \tag{8}$$

and restricting our attention to a narrow range of frequencies $f_l < f < f_h$ around a line that we want to regress, it is reasonable to assume that the data are a sum of the line and some background, continuum noise; we can assume that the background noise power spectrum, S_0, is constant because the bandwidth is narrow. Then the likelihood for the data D given the line contribution X is

$$p\left(D_k \mid X_k, S_0\right) = N\left(D_k \mid X_k, \sqrt{T S_0}\right). \tag{9}$$

(In words: the data are normally distributed with mean X and standard deviation $\sqrt{T S_0}$.) It is also reasonable to assume that the line-driving noise spectrum S_n is constant over the narrow bandwith; we have seen above that the line itself is normally distributed with mean zero and variance $S_x(f) T$, so

$$p\left(X_k \mid S_n, f_0, \gamma\right) = N\left(X_k \mid 0, \sqrt{S_x\left(f_k \mid S_n, f_0, \gamma\right) T}\right). \tag{10}$$

The product of these two normal distributions is, itself a normal distribution; and it can be factorized in various ways (Hogg et al. 2020). One factorization writes it as a product of the marginal likelihood for the data given line parameters and continuum noise,

$$p\left(D_k \mid S_n, f_0, \gamma, S_0\right) = N\left(D_k \mid 0, \sqrt{S_x\left(f_k \mid S_n, f_0, \gamma\right) T + S_0 T}\right); \tag{11}$$

and the conditional posterior for the line X_k given the data and parameters

$$p\left(X_k \mid D_k, S_n, f_0, \gamma, S_0\right) \\ = N\left(X_k \mid \frac{S_x\left(f_k \mid S_n, f_0, \gamma\right) D_k}{S_x\left(f_k \mid S_n, f_0, \gamma\right) + S_0}, \sqrt{\frac{S_x\left(f_k \mid S_n, f_0, \gamma\right) S_0 T}{S_x\left(f_k \mid S_n, f_0, \gamma\right) + S_0}}\right). \tag{12}$$

Applying a prior on the line and continuum parameters, and using standard stochastic sampling techniques (MCMC, HMC, etc) allows to explore the posterior over S_n, f_0, γ, and S_0 given data; for each sample, we can then draw the line X_k from the conditional posterior above.

Subtracting a fair draw of the line from the data generates a residual sample,

$$R_k \equiv D_k - X_k, \tag{13}$$

without any line content which can then be passed to downstream data analysis pipelines that may benefit from the reduced noise level. This is in contrast to more standard analyses that include the line in an estimated PSD, effectively "notch filtering" out the data dominated by the line (Veitch et al. 2015; Littenberg & Cornish 2015). We choose to pass onward a fair sample of the residuals instead of, say, the maximum likelihood residuals so that the noise properties of the residuals are un-biasedly described by the spectral density S_0; see Appendix A.

2. IMPLEMENTATION

We have implemented[1] a model for line cleaning as described above using Hamiltonian Monte Carlo (Neal 2011) as implemented in `numpyro` Bingham et al. (2019); Phan et al. (2019) to sample over the line parameters and continuum noise level; the same model draws the line from the conditional distribution described above at each sample of the line parameters. The model fitting and line subtraction for an arbitrary number of lines in narrow bandwidth ranges supplied by the user is encapsulated in a single function, `clean_lines`, which returns time-domain residuals and optionally the output of the MCMC sampling for each line.

Time domain data are tapered with an apodizing Tukey window before being discrete Fourier transformed for the fitting in the frequency domain. Residuals are returned for all data *outside* the tapered region.

2.1. *Fitting Considerations*

[1] https://github.com/farr/LineCleaner

One consideration in this sort of analysis is the length of data to fit. The user should ensure (after tapering) there are at least several correlation times $\tau = 1/\gamma$ of the line avaliable to fit; longer data segments will not substantially improve the subtraction of the line (even if they will improve knowledge of the line's parameters), because the phase of the line "resets" after each correlation time. Long data segments can hurt if the data are significantly non-stationary, since the model assumes stationarity; this can be particularly bad if the actual physical line parameters are changing throughout the longer data segment. In practice, in LIGO data, the cleaner seems to work well with between several tens and several thousands of seconds of data, depending on the intrinsic width of the line.

The bandwidth should be chosen as narrow as possible, but wide enough to ensure that the wings of the line have fallen well below the continuum noise level. This is for two reasons: first, the line will only be subtracted from the data within the chosen bandwidth, so you want to include all data where the line could be relevant; second, you need to ensure that the model has a good estimate of the continuum noise level.

2.2. *Example*

Here we show an example of cleaning lines at 60 Hz, 120 Hz, and 180 Hz out of data from the Hanford LIGO observatory around the time of GW150914. The data are available from the Gravitational Wave Open Science Center (GWOSC) Abbott et al. (2021, 2023); GWOSC (2024). We use 4096 s of data sampled at 4096 Hz. The code is available in a Jupyter notebook[2]. The full original and cleaned power spectral densities are shown in Figure 1, and in a narrow band around each cleaned line in Figure 2. The line cleaner is able to remove the lines and leave a residual consistent with the continuum noise level.

Software: This work made use of the following software packages: `Jupyter` (Perez & Granger 2007; Kluyver et al. 2016), `matplotlib` (Hunter 2007), `numpy` (Harris et al. 2020), `pandas` (Wes McKinney 2010; pandas development team 2024), `python` (Van Rossum & Drake 2009), `scipy` (Virtanen et al. 2020; Gommers et al. 2025), `ArviZ` (Kumar et al. 2019; Martin et al. 2024), `JAX` (Bradbury et al. 2018), `numpyro` (Phan et al. 2019; Bingham et al. 2019), `tqdm` (da Costa-Luis et al. 2024), and `xarray` (Hoyer & Hamman 2017; Hoyer et al. 2025). Software citation information aggregated using `The Software Citation Station` (Wagg & Broekgaarden 2024; Wagg et al. 2024).

APPENDIX

[2] https://github.com/farr/LineCleaner/blob/main/LineCleaner.ipynb

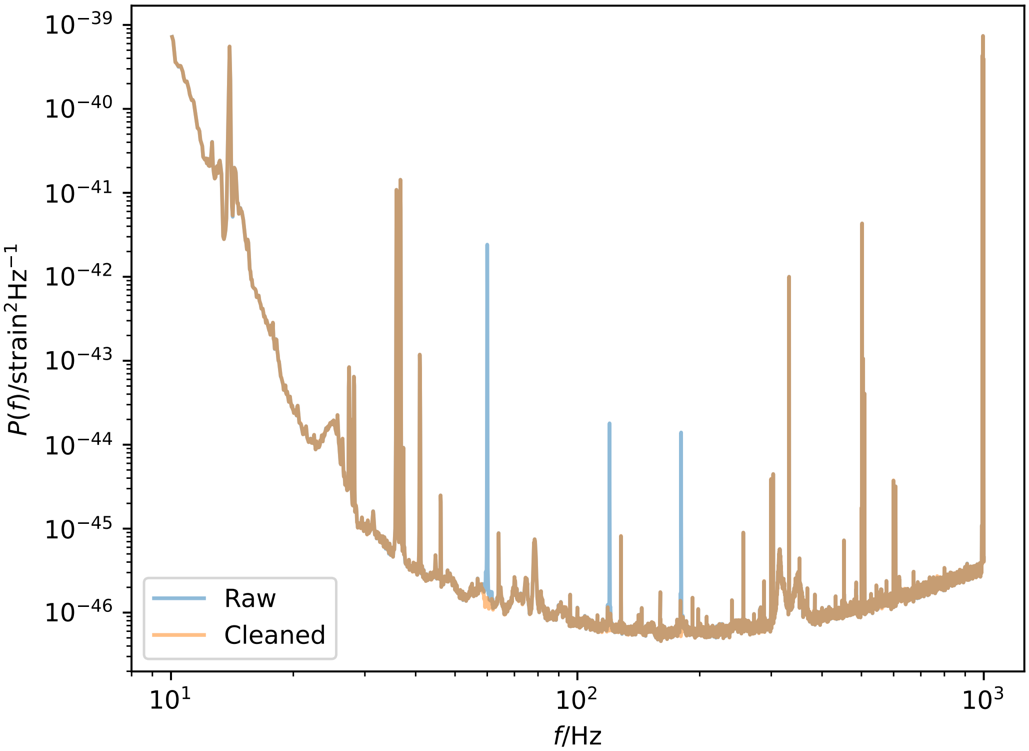

Figure 1. An example of a cleaning of data from the Hanford LIGO observatory around the GW150914 event Abbott et al. (2016, 2019); Abbott et al. (2021, 2023); GWOSC (2024). The raw and cleaned data power spectral density is shown from 10 Hz to 1 kHz. The lines at 60 Hz, 120 Hz, and 180 Hz are clearly visible in the raw data, but are removed by the line cleaner; the cleaning used 4 Hz bandwidth around each line. Figure 2 shows a zoomed view of the PSDs in the vicinity of each line.

A. AVOIDING OVER-SUBTRACTION

As discussed in Section 1, the line cleaner algorithm subtracts a fair draw from the posterior over the line parameters and line values, instead of using the maximum-likelihood values for either of these quantities. This procedure avoids oversubtraction of the line, which is particularly important if downstream analyses are producing, for example, a noise estimate from the residuals. As can be seen in Figure 2, the residuals are consistent with the continuum noise level when a fair draw is subtracted; this would not be the case were the maximum-likelihood line used to produce the residuals.

To motivate this choice, consider the following very simple example which nevertheless retains crucial features of the line cleaning problem. Let us suppose that we are trying to estimate a single, scalar quantity, x, from data d that is a sum of x and some normally distributed noise, $n \sim N(0, \sigma)$,

$$d = x + n. \tag{A1}$$

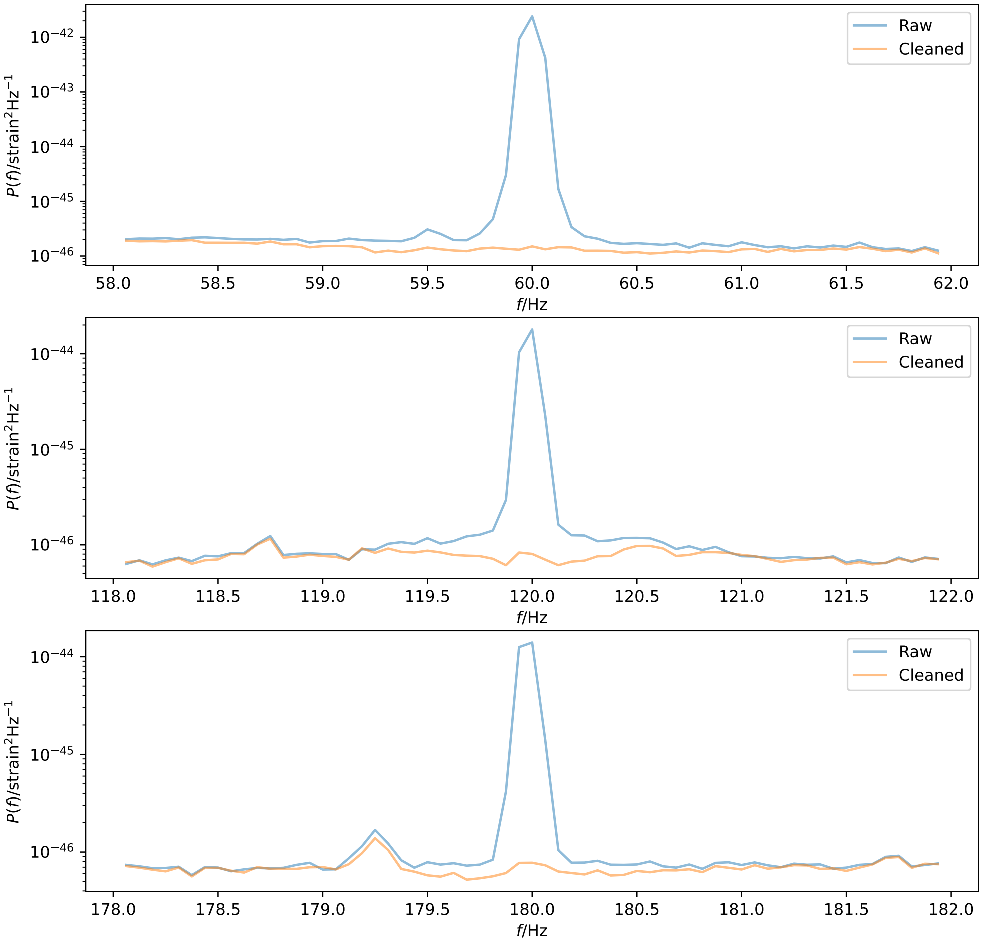

Figure 2. A zoom-in of the power spectral densities of the raw and cleaned data shown in Figure 1 around the lines cleaned at .

The maximum likelihood estimate for x, denoted $\hat{x}$, is

$$\hat{x} = d, \tag{A2}$$

which leaves zero residual when subtracted from the data:

$$r = d - \hat{x} = 0. \tag{A3}$$

On the other hand, the posterior for x given d is normally distributed with mean d and standard deviation σ,

$$p(x \mid d) = N(x \mid d, \sigma). \tag{A4}$$

A fair draw from this posterior, subtracted from d, leaves a residual that is also normally distributed with mean zero and standard deviation σ, which is exactly the correct distribution for residual noise to pass onward to downstream analyses. The maximum-likelihood model *over-subtracts* relative to the fair draw, producing residuals that are biased low compared to the actual noise process in the data. For this reason, we choose to subtract a line estimated by a fair draw from our posterior to produce the residuals that we pass onward.

REFERENCES

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2016, PhRvL, 116, 061102, doi: 10.1103/PhysRevLett.116.061102

—. 2019, Physical Review X, 9, 031040, doi: 10.1103/PhysRevX.9.031040

Abbott, R., et al. 2021, SoftwareX, 13, 100658, doi: 10.1016/j.softx.2021.100658

—. 2023, Astrophys. J. Suppl., 267, 29, doi: 10.3847/1538-4365/acdc9f

Bingham, E., Chen, J. P., Jankowiak, M., et al. 2019, J. Mach. Learn. Res., 20, 28:1. http://jmlr.org/papers/v20/18-403.html

Bradbury, J., Frostig, R., Hawkins, P., et al. 2018. http://github.com/google/jax

da Costa-Luis, C., Larroque, S. K., Altendorf, K., et al. 2024, doi: 10.5281/zenodo.14231923

Gommers, R., Virtanen, P., Haberland, M., et al. 2025, doi: 10.5281/zenodo.14630489

GWOSC. 2024, GW150914, https://gwosc.org/eventapi/html/GWTC-1-confident/GW150914/

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357, doi: 10.1038/s41586-020-2649-2

Hogg, D. W., Price-Whelan, A. M., & Leistedt, B. 2020, arXiv e-prints, arXiv:2005.14199, doi: 10.48550/arXiv.2005.14199

Hoyer, S., & Hamman, J. 2017, Journal of Open Research Software, 5, doi: 10.5334/jors.148

Hoyer, S., Roos, M., Joseph, H., et al. 2025, xarray, v2025.01.2, Zenodo, doi: 10.5281/zenodo.14777280

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90, doi: 10.1109/MCSE.2007.55

Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Schmidt, IOS Press, 87 – 90

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. 2019, Journal of Open Source Software, 4, 1143, doi: 10.21105/joss.01143

Littenberg, T. B., & Cornish, N. J. 2015, PhRvD, 91, 084034, doi: 10.1103/PhysRevD.91.084034

Martin, O. A., Hartikainen, A., Abril-Pla, O., et al. 2024, ArviZ, v0.20.0, Zenodo, doi: 10.5281/zenodo.13854799

Neal, R. 2011, in Handbook of Markov Chain Monte Carlo, 113–162, doi: 10.1201/b10905

pandas development team, T. 2024, pandas-dev/pandas: Pandas, v2.2.3, Zenodo, doi: 10.5281/zenodo.13819579

Perez, F., & Granger, B. E. 2007, Computing in Science and Engineering, 9, 21, doi: 10.1109/MCSE.2007.53

Phan, D., Pradhan, N., & Jankowiak, M. 2019, arXiv preprint arXiv:1912.11554

Van Rossum, G., & Drake, F. L. 2009, Python 3 Reference Manual (Scotts Valley, CA: CreateSpace)

Veitch, J., Raymond, V., Farr, B., et al. 2015, PhRvD, 91, 042003, doi: 10.1103/PhysRevD.91.042003

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: 10.1038/s41592-019-0686-2

Wagg, T., Broekgaarden, F., & Gültekin, K. 2024, TomWagg/software-citation-station: v1.2, v1.2, Zenodo, doi: 10.5281/zenodo.13225824

Wagg, T., & Broekgaarden, F. S. 2024, arXiv e-prints, arXiv:2406.04405. https://arxiv.org/abs/2406.04405

Wes McKinney. 2010, in Proceedings of the 9th Python in Science Conference, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61, doi: 10.25080/Majora-92bf1922-00a