Optimizing Parameters of Information-Theoretic Correlation Measurement for Multi-Channel Time Series Datasets in Gravitational-Wave Detectors

Piljong Jung[1], Sang Hoon Oh[1], Young-Min Kim[2], Edwin J. Son[1], John J. Oh[1]

Motivation

The gravitational wave detectors are one of the complex and elaborate systems in which there are easily influenced by various electronics and devices surrounding instruments and environments. Thus, it is crucial to understand the effects of non-linear couplings between multi-channels in the complex devices and the origin and propagation of, for they act as an obstacle to GW observation.

Among various methods to identify non-linear correlations, the Maximal Information Coefficient (MIC) [1-2] provides a remarkable performance to capture the complex associations since it can estimate and characterize the strength of dependence between data sets. However, despite its capacity, MIC has difficulty in interpreting the results for the following reasons:

1) the association strength of estimators varies by choosing parameters,
2) if the data size of two data sets is not enough to estimate their coupling, it would be unable to calculate any correlations,
3) in case of using time-series data of multi-channel, it is required to match sample-rate between two data.

Therefore, we investigated the optimized configuration based on the statistical power method when we utilize MIC for estimating the non-linear correlation between time-series data of multi-channel of GW detectors.

Methods

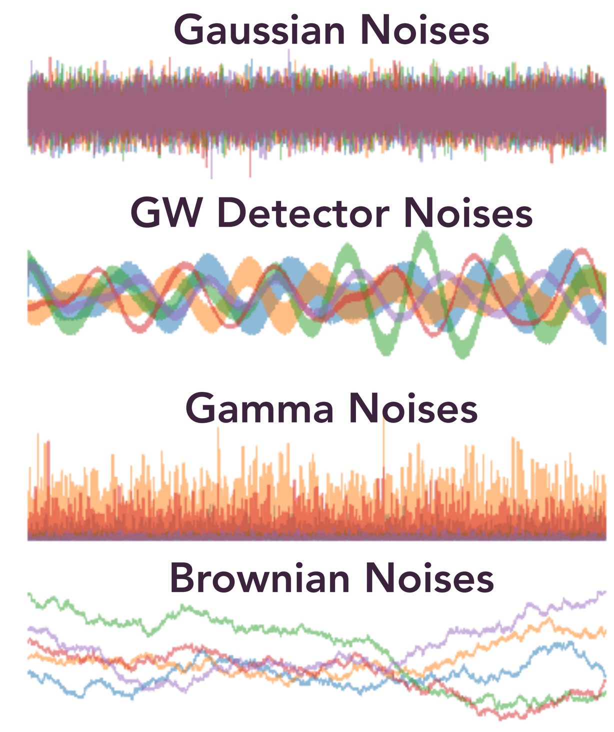
Alternative Hypothesis
$$X_1(t) = x(t) + \xi(t)$$
$$Y_1(t) = \sigma y(t) + \mathcal{N}\xi'(t)$$

Null Hypothesis
$$X_0(t) = x(s) + \xi(t)$$
$$Y_0(t) = \sigma y(s) + \mathcal{N}\xi(t)$$

$t \in \mathbb{R}$ and $\forall s \in \mathbb{R}$ is chosen arbitrarily

$$NSR_{Y/X} = \frac{E(\xi^2)E(x^2)}{E(\xi^2)E(y^2)} = \frac{\mathcal{N}^2}{\sigma^2}$$

Noise added



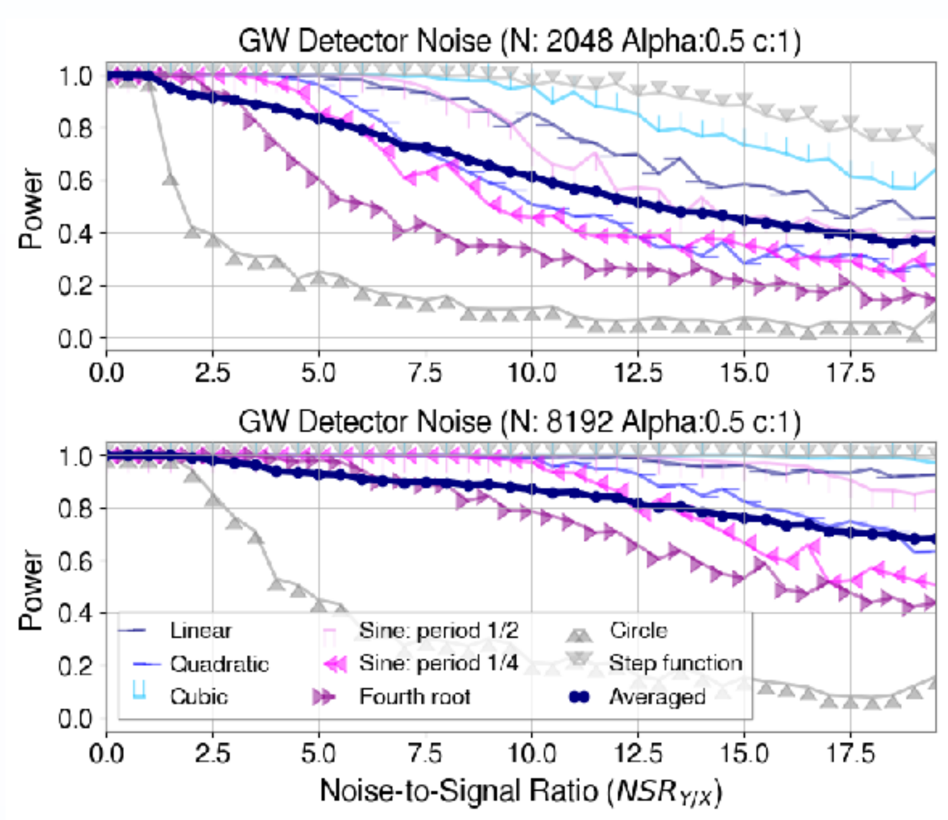
$$MICe(X, Y, \alpha, c) = \max_{ab<B(N)} \left\{ \frac{\max I^{[*]}(D, a, b)}{\log_2 \min\{a, b\}} \right\}$$

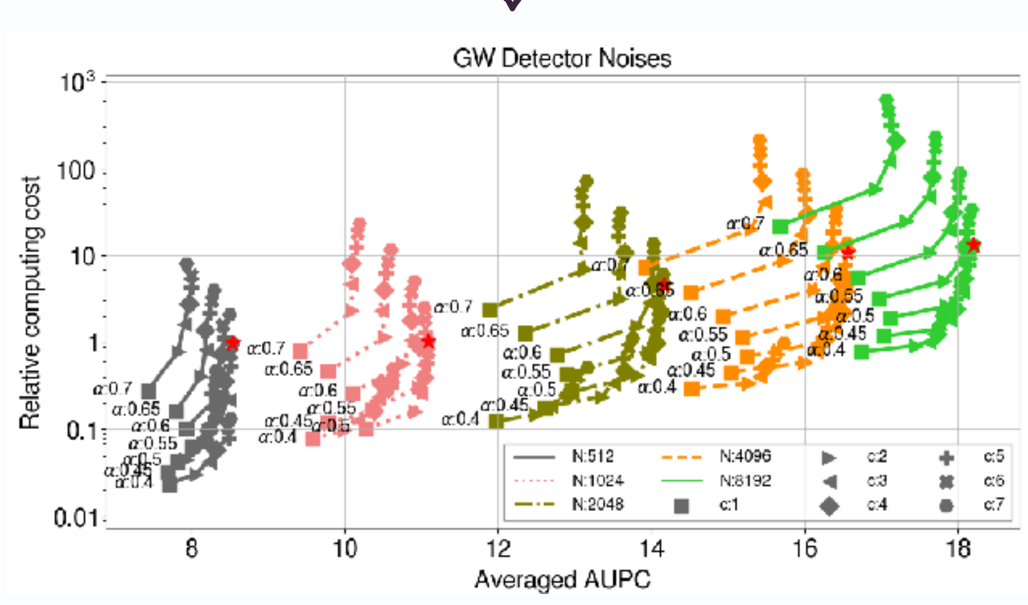
Sample number: 500
Statistical significance: 5 percent

$$D_0^{95\%}(\alpha, c) \in S_0^{95\%}(\alpha, c) \equiv \left\{ MICe\left(X_0(t), Y_0(t), \alpha, c\right) \right\}_{>95\%}$$
$$S_1^{5\%}(\alpha, c) \equiv \left\{ MICe\left(X(t), Y(t), \alpha, c\right) > D_0^{95\%}(\alpha, c) \right\}$$
$$\longrightarrow \quad Power \equiv N[D_1^{5\%}(\alpha, c)]/N_1$$

Results

A. Data samples and Parameter selection





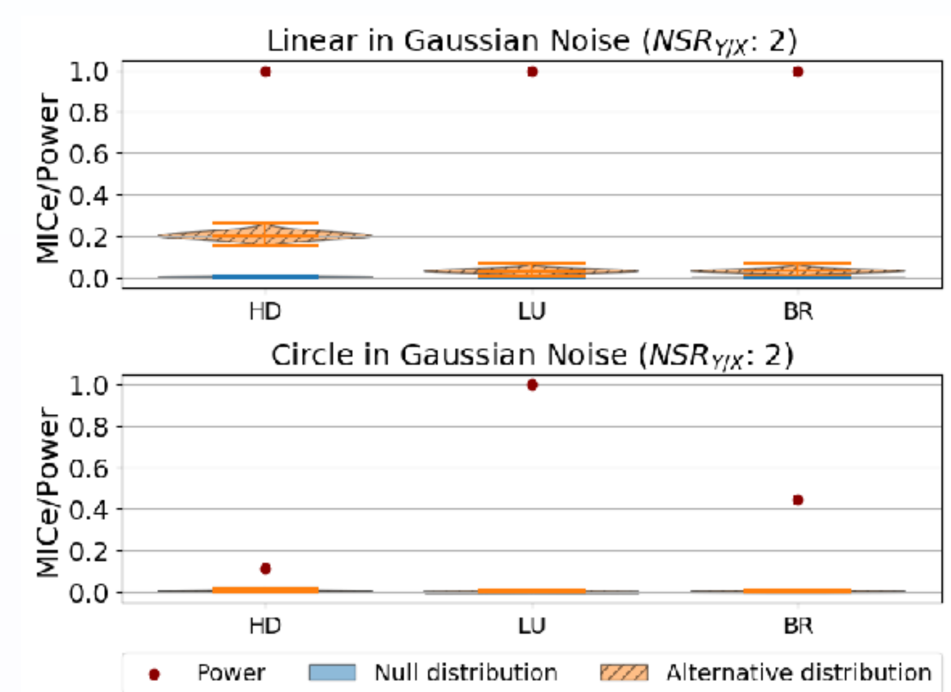
Optimized parameters

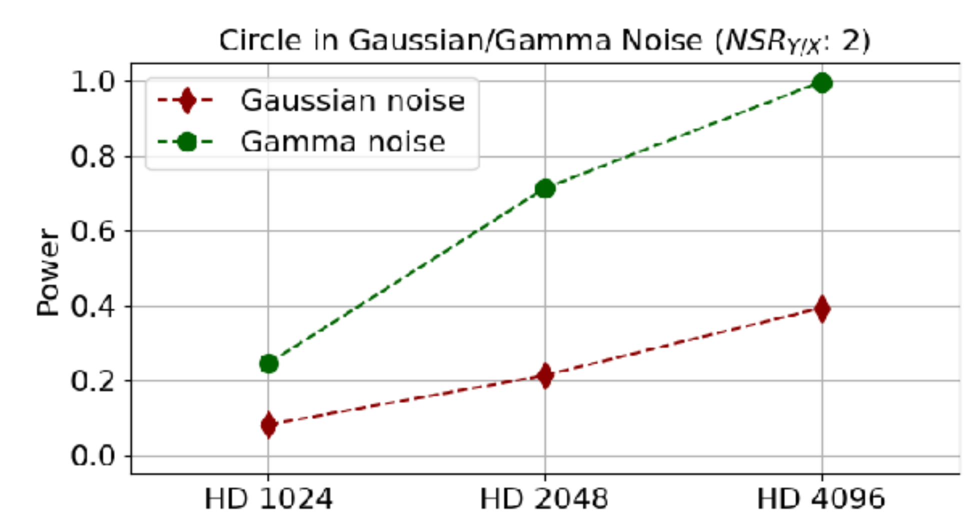
Noise	N	Alpha	c	Avg. AUPG	Cost
Gaussian Noise	512	0.35	7.0	5.434	1.000
	1024	0.35	2.0	6.899	1.286
	2048	0.30	5.0	9.166	1.625
	4096	0.25	7.0	11.465	3.069
	8192	0.25	7.0	13.742	5.694
GW Detector Noise	512	0.55	7.0	8.535	1.000
	1024	0.50	7.0	11.092	1.040
	2048	0.55	6.0	14.164	4.561
	4096	0.55	6.0	16.566	10.781
	8192	0.50	7.0	18.199	13.330
Gamma Noise	512	0.60	7.0	16.752	1.000
	1024	0.50	7.0	18.234	0.493
	2048	0.45	7.0	18.955	0.531
	4096	0.40	7.0	19.346	0.466
	8192	0.40	7.0	19.614	1.069
Brownian Noise	512	0.60	6.0	13.320	1.000
	1024	0.55	7.0	15.736	1.613
	2048	0.50	6.0	17.495	1.252
	4096	0.50	5.0	18.652	2.014
	8192	0.50	5.0	19.367	4.886

We estimate the statistical power of MICe for every functionally associated dataset under different background noises. To figure out the effect of parameters on the power of MICe, we investigate two factors for optimizing parameters of MICe- an area under the power curve(AUPC) and a computational cost. The AUPC is defined as an area under the statistical power curve for a given parameter, and the Cost is the relative running time.

When the sample size becomes large, the statistical power also remains efficient as the NSR level increases. Based on these results, we selected suitable parameter sets for producing the highest statistical power of MICe. Optimized parameters for each noise type are organized in the above table. The full results are shown in reference [3].

B. Resampling for multi-channel datasets





We considered three scenarios to match the different sampling rates: both down-sampling from high to low (HD), 2) both up-sampling from low to high (LU), 3) both up/down-sampling in middle frequency (BR).

Thus, except for a specific case of circular association for Gaussian and Gamma noises, the resampling effect does not affect the statistical power of MICe. This is because the down-sampled data size is not enough to distinguish null/alternative distributions. If the sufficient size of data samples is guaranteed, the aspect of data resampling does not affect the statistical power of computing MICe. The full results are shown in reference [3].

Discussion

With this study, we investigated that the statistical power of MIC depends upon the choice of parameter sets, the noise level of data, and the data sample size. Also, the value of parameters relies on the type of background noise and data sample size. For computing gauges of NSR and Power, we can choose the set of parameters, alpha and c, yielding the most optimal result. In addition, for dealing with data of different sampling rates, it is essential to have a sufficient data sample size regardless of choosing the resampling scenarios.

Even if we may improve the MIC algorithm by suggesting other methods, it is adequate to identify the non-linear coupling between two variables from different channels. To make a more reliable decision, it is of importance to have a consistent standard for interpretations.

References

[1] D. N. Reshef, et al., Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524. doi:10.1126/science.1205438
[2] Y. A. Reshef, et al., Measuring dependence powerfully and equitably, Journal of Machine Learning Research 17 (211) (2016) 1–63
[3] doi:10.5281/zenodo.4964870

THE 8TH KAGRA INTERNATIONAL WORKSHOP

[1]National Institute for Mathematical Sciences
[2]Department of Physics, Ulsan National Institute of Science and Technology