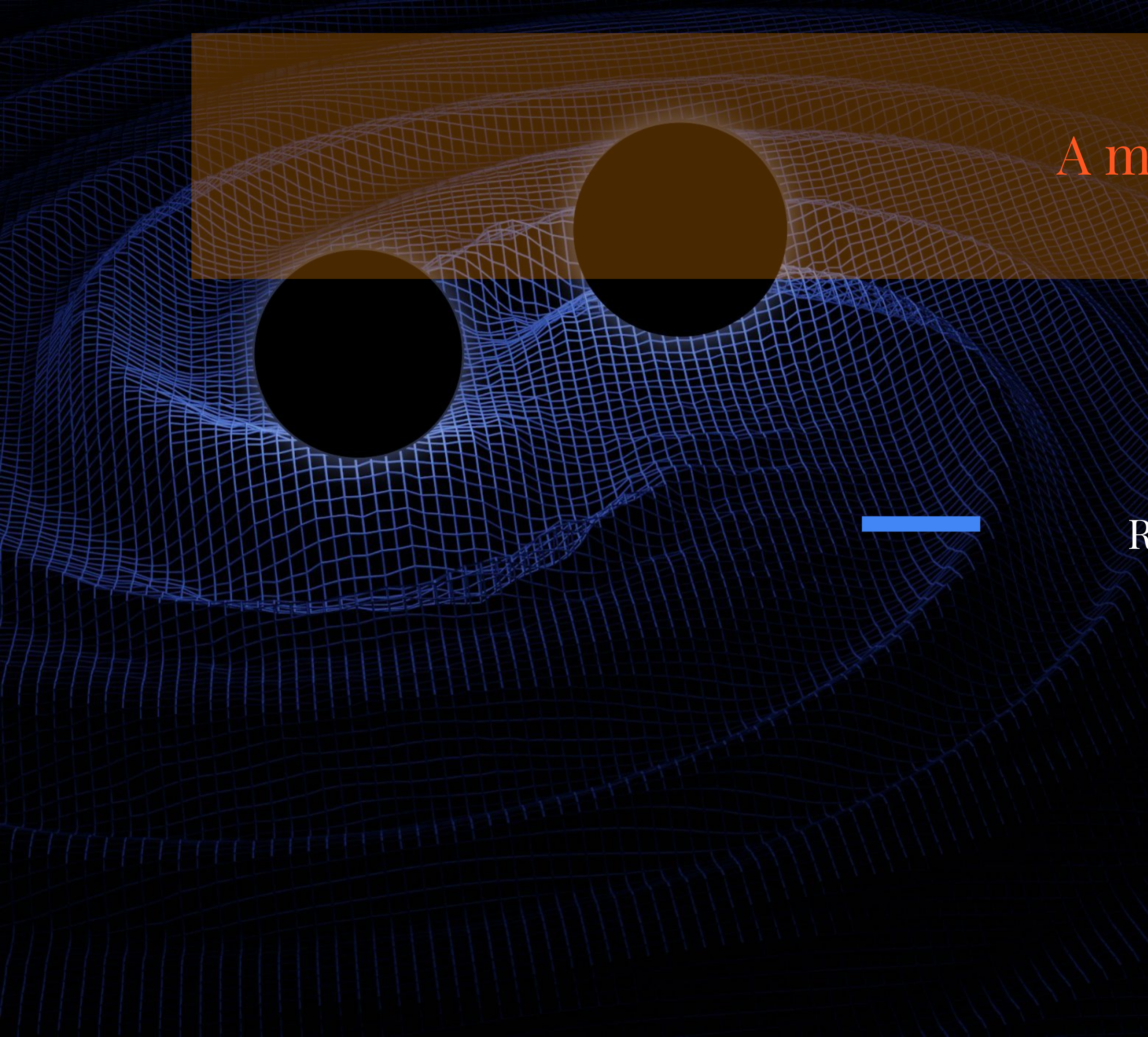
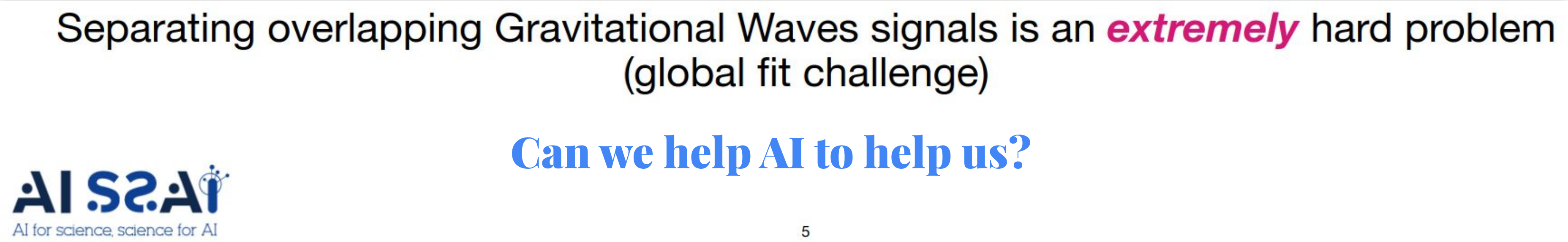


Beyond Gauss?
A more accurate model for LISA astrophysical noise sources

Heterogeneous Data and Large Representation Models in Science
Toulouse, FR
2024/10/01

R. Buscicchio
Università di Milano-Bicocca
riccardo.buscicchio@unimib.it

The global fit



Separating overlapping Gravitational Waves signals is an ***extremely*** hard problem (global fit challenge)

Can we help AI to help us?

AI SSAI
AI for science, science for AI

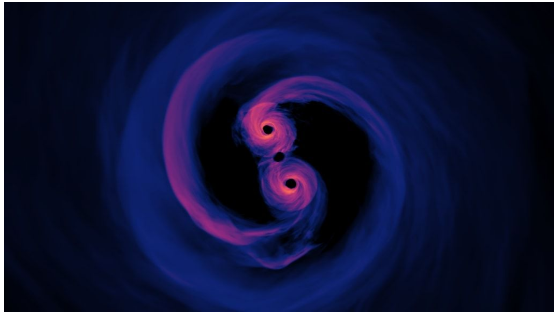
The global fit

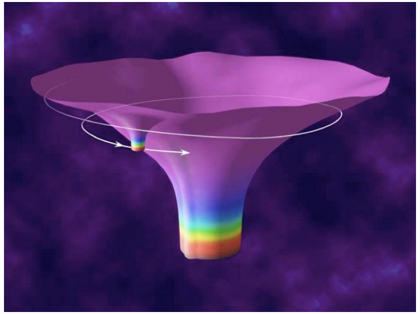
Scores from a penguin cacophony

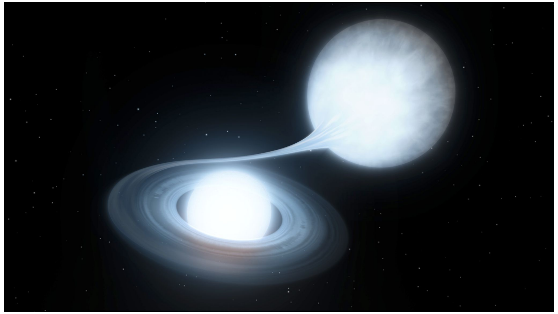
Hundreds

Tens to thousands

Millions

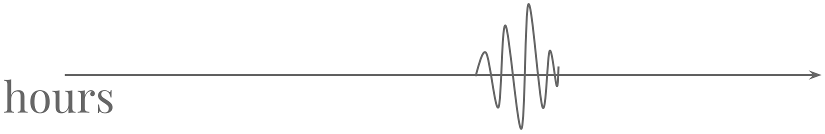






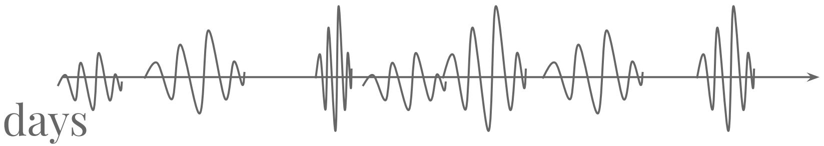
The data

Transients signals (i.e. MBHB)



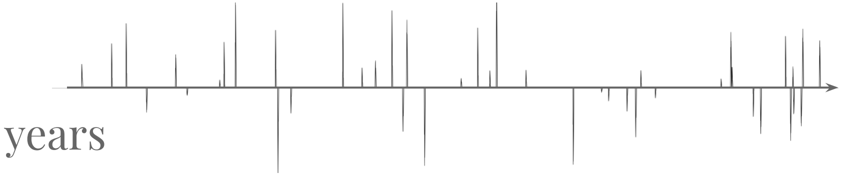
Is it stochastic?

Individual event search & parameter estimation



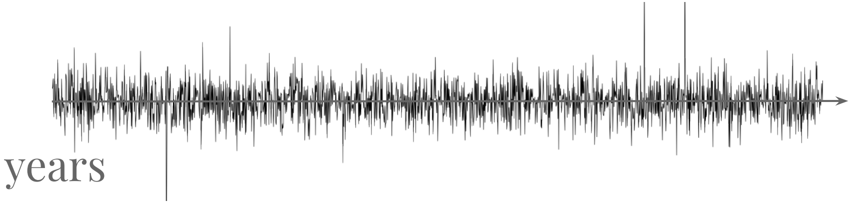
Is it stochastic?

Overlapping events S&PE



Is it deterministic?

Non-Gaussian stochastic signal S&PE

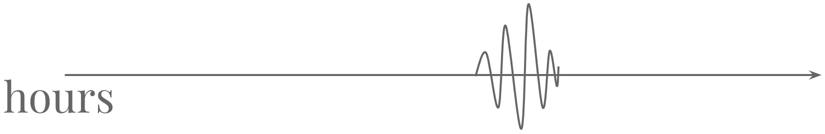


Is it deterministic?

Gaussian stochastic signal S&PE

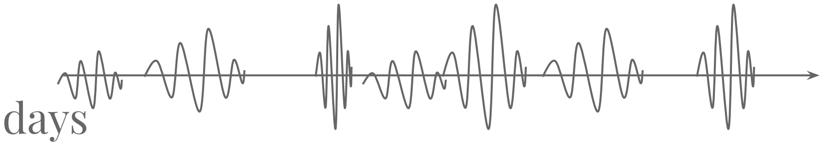
The data

Transients signals (i.e. MBHB)



Is it stochastic?

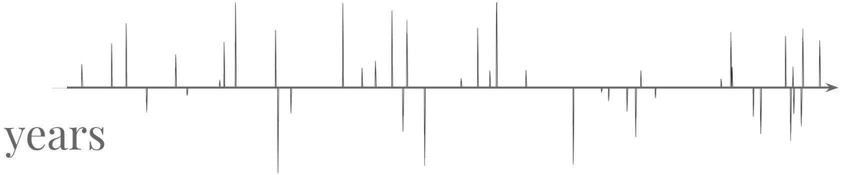
Individual event search & parameter estimation



Is it stochastic?

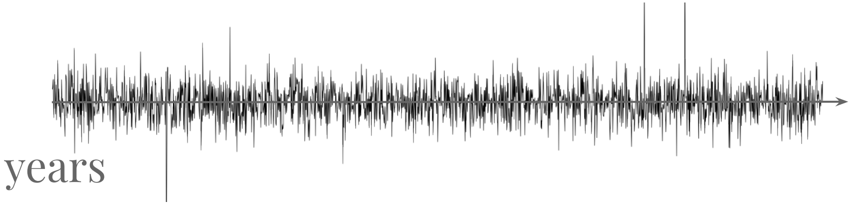
Overlapping events S&PE

$$\mathrm{SNR} > \rho_{\mathrm{thres}}(\mathrm{P}_{\mathrm{FA}})$$



Is it deterministic?

Non-Gaussian stochastic signal S&PE

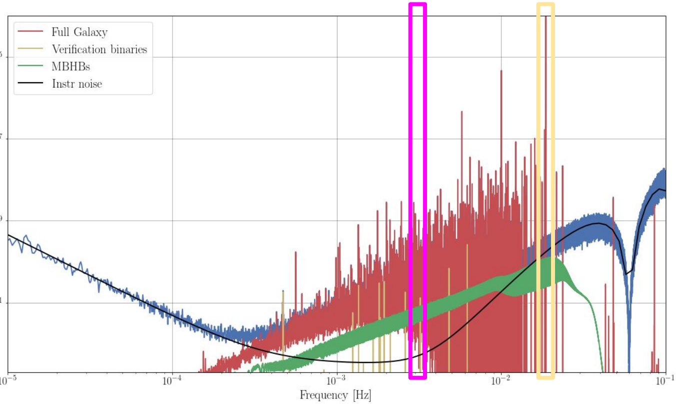


Is it deterministic?

Gaussian stochastic signal S&PE

The data

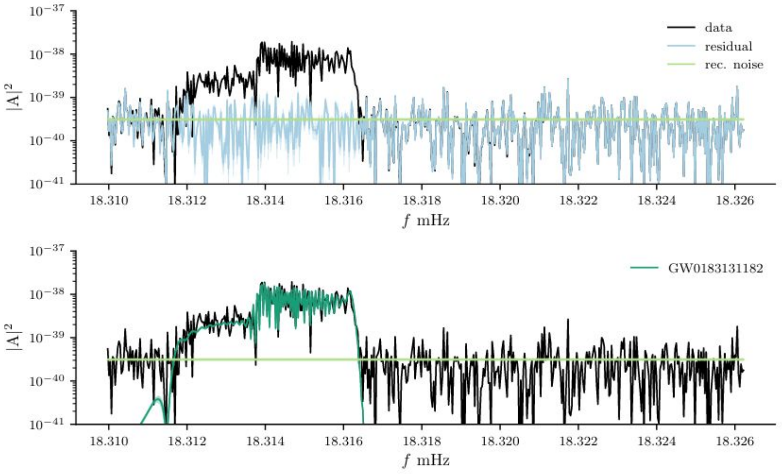
Persistent signals (i.e. GBs)



LISA Data Challenge

Is it stochastic?

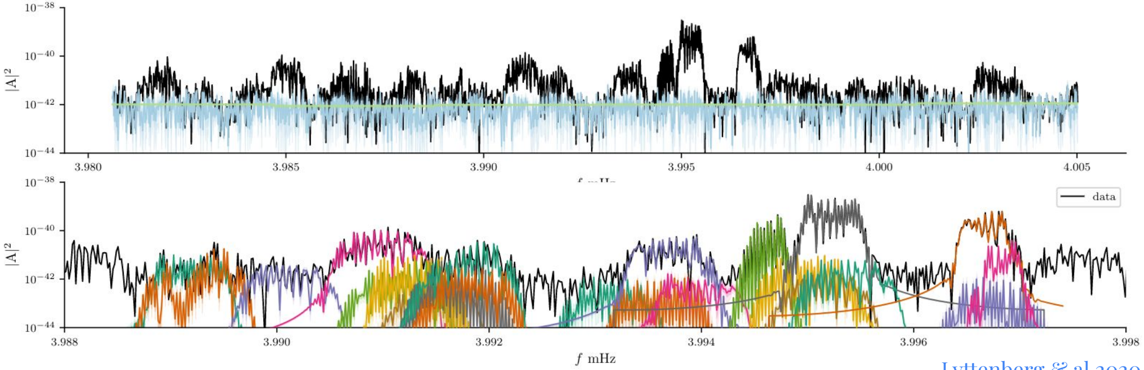
Is it stochastic?



Lyttenberg & al. 2020

Is it deterministic?

Is it deterministic?



Lyttenberg & al 2020

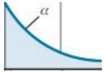
Stochastic or Deterministic

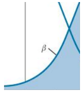
It is required by the data!

It is a modelling choice!

Detection statistics

Or “How to construct a frequentist detector”

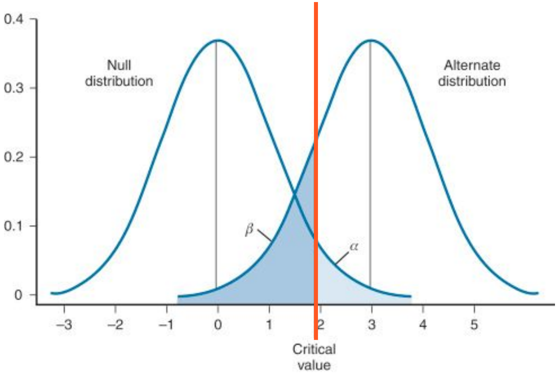
- Model the data under both hypotheses (noise, noise+signal).
- Fix the probability of false alarm P_{FA} : i.e. 

- Maximize the probability of detection P_D at fixed P_{FA} : i.e. (1- )

- Isolate the dependence on data in “sufficient statistics” **Y(s)** (SNR is just an example)
- Obtain a threshold as a function of P_{FA}: i.e.
- **Bonus**: the likelihood is $P(s|H_1)$

AI: classifiers

AI: regressors

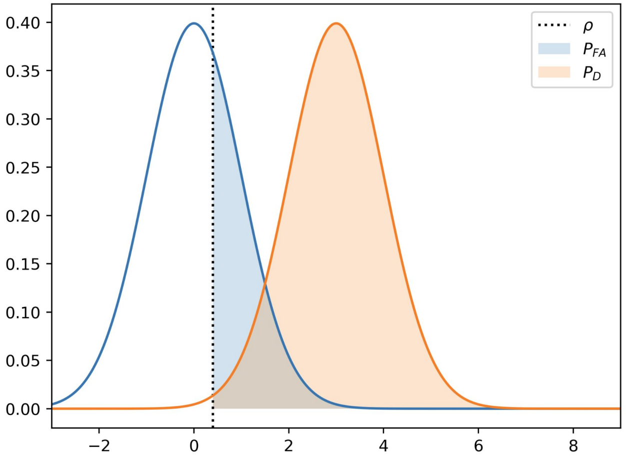


Detection

The simplest: a fixed number in Gaussian noise

$$p(d \mid \mathcal{H}_0) \qquad n = d \sim \mathcal{N}(0, \sigma)$$

$$p(d \mid \mathcal{H}_1) \qquad \mu + n = d \sim \mathcal{N}(\mu, \sigma)$$

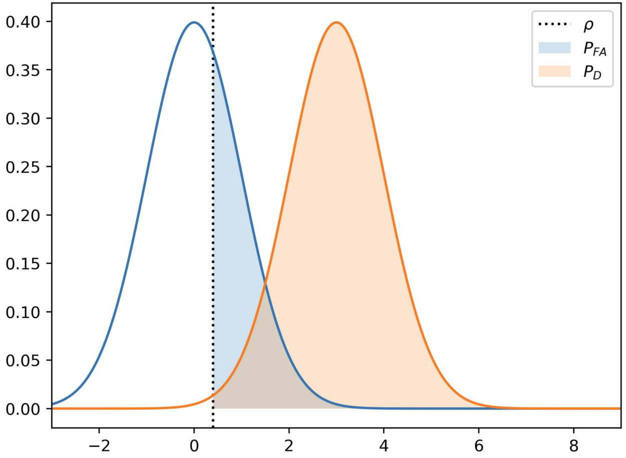


Detection

The simplest: a fixed number in Gaussian noise

$$p(d \mid \mathcal{H}_0) \qquad n = d \sim \mathcal{N}(0, \sigma)$$

$$p(d \mid \mathcal{H}_1) \qquad \mu + n = d \sim \mathcal{N}(\mu, \sigma)$$



Collect d, how do you decide?

- Better to minimize the probability of false alarm? →
- Better to maximize the probability of detection?
- Can you do both?

Optimal: Neyman Pearson detector

Threshold that maximizes the probability of detection at a fixed probability of false alarm.

$$\hat{s}(d) > \rho(P_{FA})$$

In GW context:

$$\mathrm{SNR}^2 = 4 \int_0^\infty df \frac{|\tilde{h}(f)|^2}{S_h(f)}$$

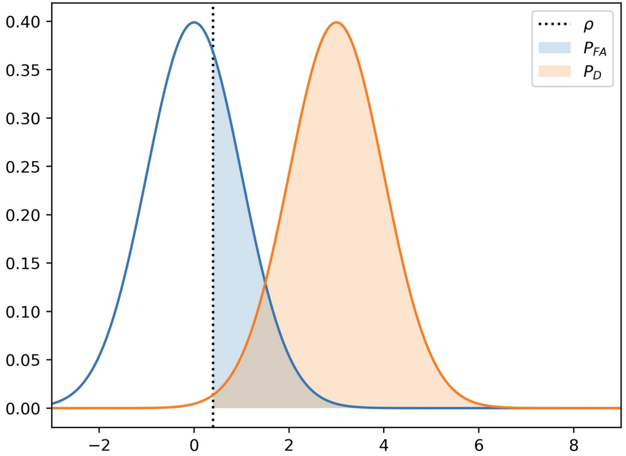
$$\mathrm{SNR} > \rho_{\mathrm{thres}}(\mathrm{P_{FA}})$$

Detection

The simplest: a fixed number in Gaussian noise

$$p(d \mid \mathcal{H}_0) \qquad n = d \sim \mathcal{N}(0, \sigma)$$

$$p(d \mid \mathcal{H}_1) \qquad \mu + n = d \sim \mathcal{N}(\mu, \sigma)$$



Collect d, how do you decide?

- Better to minimize the probability of false alarm?
- Better to maximize the probability of detection?
- Can you do both?

→

Optimal: Neyman Pearson detector

Threshold that maximizes the probability of detection at a fixed probability of false alarm.

$$\hat{s}(d) > \rho(P_{FA})$$

In GW context:

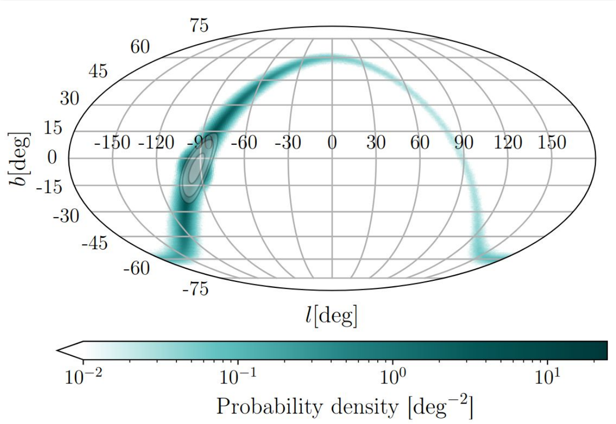
$$\mathrm{SNR}^2 = 4 \int_0^\infty df \frac{|\tilde{h}(f)|^2}{S_h(f)}$$

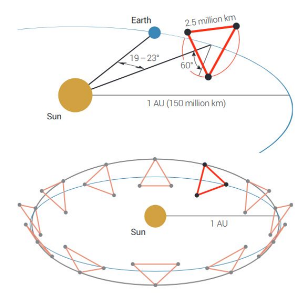
$$\mathrm{SNR} > \rho_{\mathrm{thres}}(\mathrm{P}_{\mathrm{FA}})$$

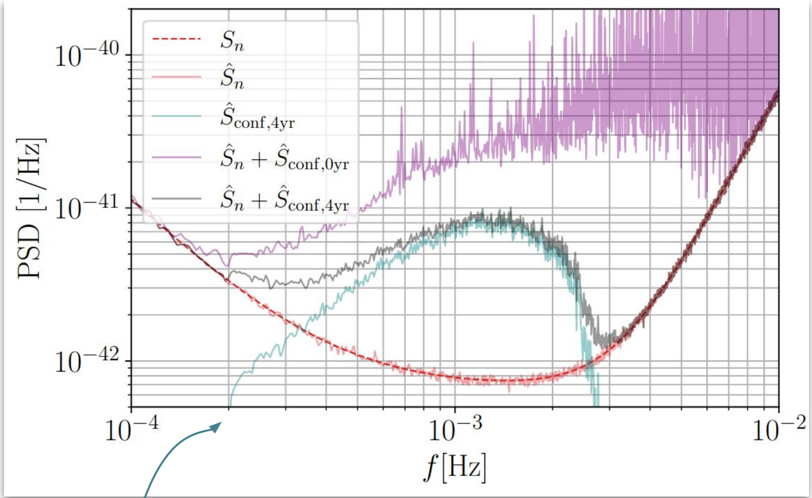
It is optimal to treat it as deterministic.

Galactic binaries

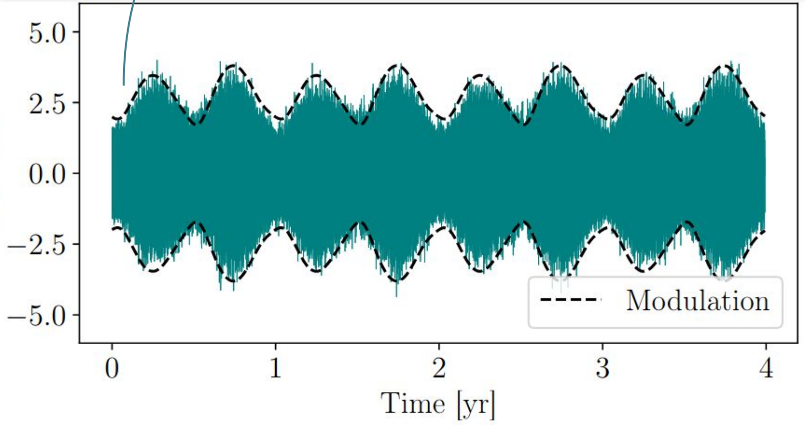
10^4 needles in 10^6 hay straws







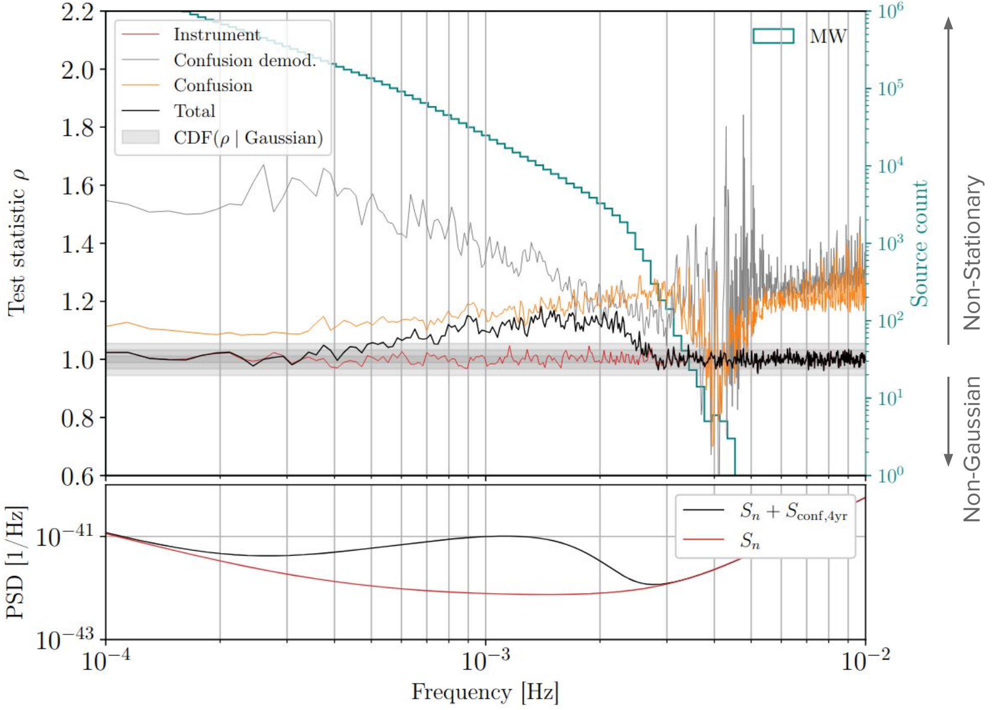
Is it ergodic?
Is it stationary?
Is it Gaussian?



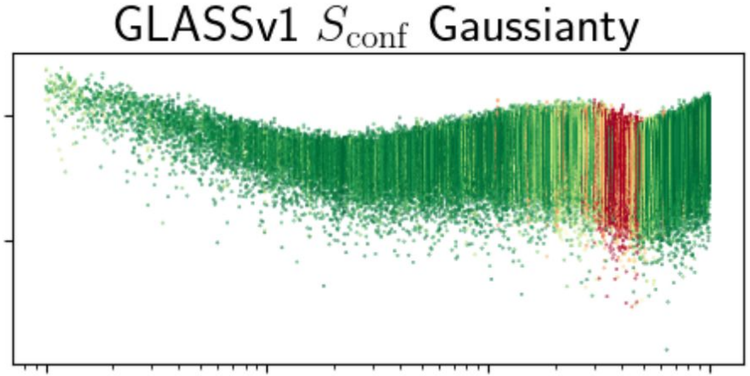
Galactic binaries

10^4 needles in 10^{6} hay straws

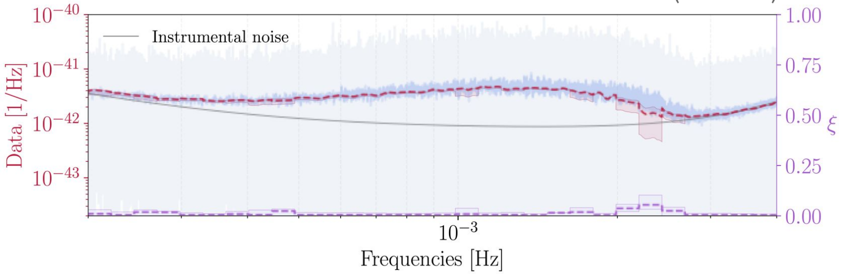
Buscicchio+ (out soon)



Rosati+ (out soon)



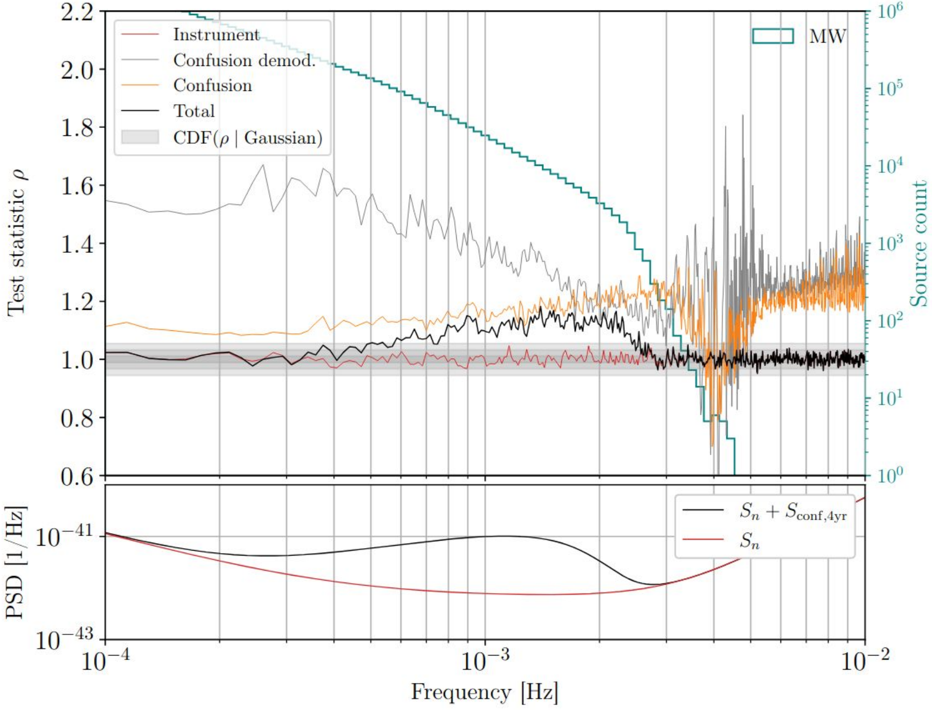
Karnesis+ (out soon)



Galactic binaries

10^4 needles in 10^{6} hay straws

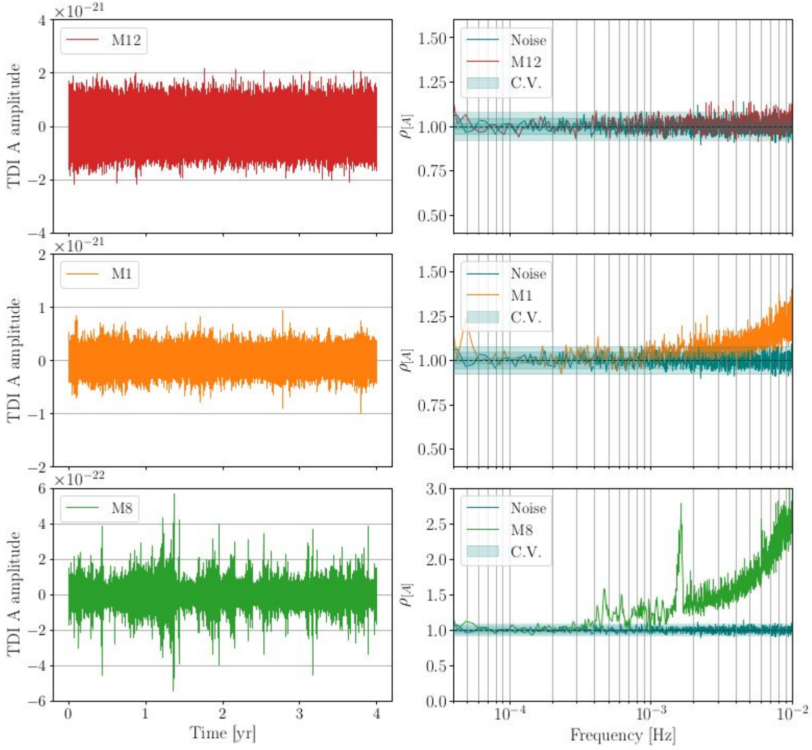
Buscicchio+ (out next week)



Bonus: EMRIs

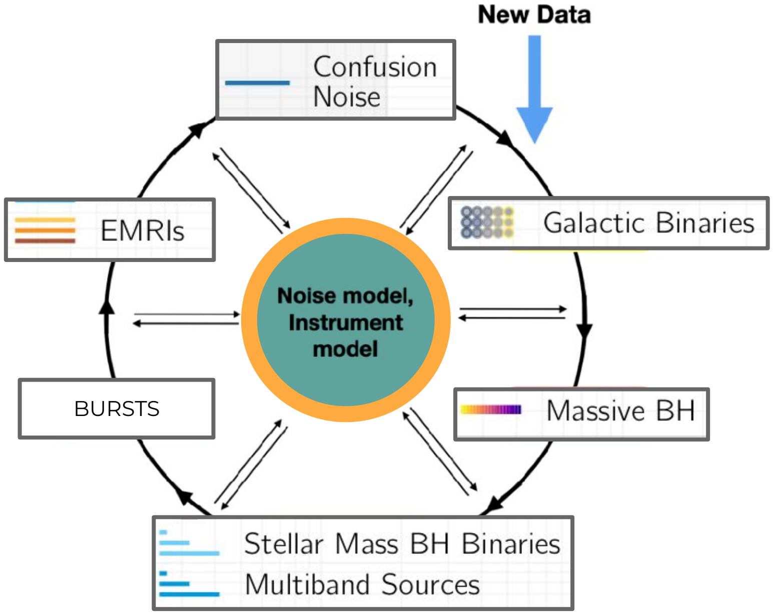
? needles in ? hay straws

Piarulli, Buscicchio, Burke+ (out next week)



Blocked Gibbs

A viable approach



New Data

Confusion Noise

Galactic Binaries

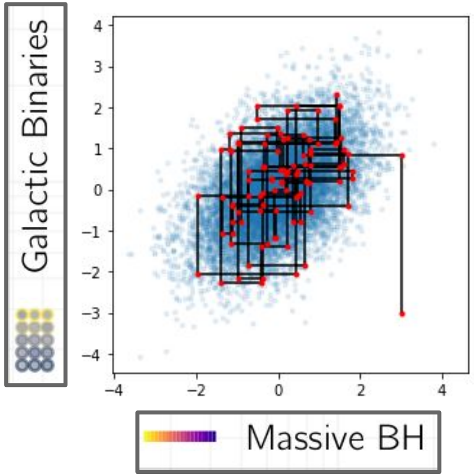
Massive BH

Stellar Mass BH Binaries
Multiband Sources

BURSTS

EMRIs

Noise model, Instrument model



Gibbs, “the sampler”

Sommese & al
Animal Cognition (2022)
25:701–705

$$X_1^{(t+1)} \big| \cdot \sim f\left(x_1 \big| x_2^{(t)}, \ldots, x_p^{(t)}\right),$$

$$X_2^{(t+1)} \big| \cdot \sim f\left(x_2 \big| x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}\right),$$

$$\vdots$$

$$X_{p-1}^{(t+1)} \big| \cdot \sim f\left(x_{p-1} \big| x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-2}^{(t+1)}, x_p^{(t)}\right),$$

$$X_p^{(t+1)} \big| \cdot \sim f\left(x_p \big| x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)}\right),$$

Likelihoods

for parameter estimation

(1) individual events: a model for the noise

$$\mathcal{L}(d \mid \mathcal{M}, q, \ldots) = p(n = d - h(\mathcal{M}, q) \mid \mathcal{M}, q, \ldots)$$

(2) multiple events: a model for the noise

$$\mathcal{L}(d \mid \mathcal{M}_1, q_1, \ldots) = p(n = d - \sum_i^N h_i(\mathcal{M}_i, q_i) \mid \mathcal{M}_1, q_1, \ldots)$$

(3) stochastic background: a model for two noises (or more)

$$\mathcal{L}(d \mid S_h, S_n) = p(d \mid S_h, S_n)$$

spectra

LISA $\mathcal{L}(d \mid \mathcal{M}_1, q_1, \ldots, N, S_h, S_n) = p(n = d - \sum_i^N h_i(\mathcal{M}_i, q_i) \mid \mathcal{M}_1, q_1, \ldots, \boxed{N}, S_h, S_n)$

individual sources

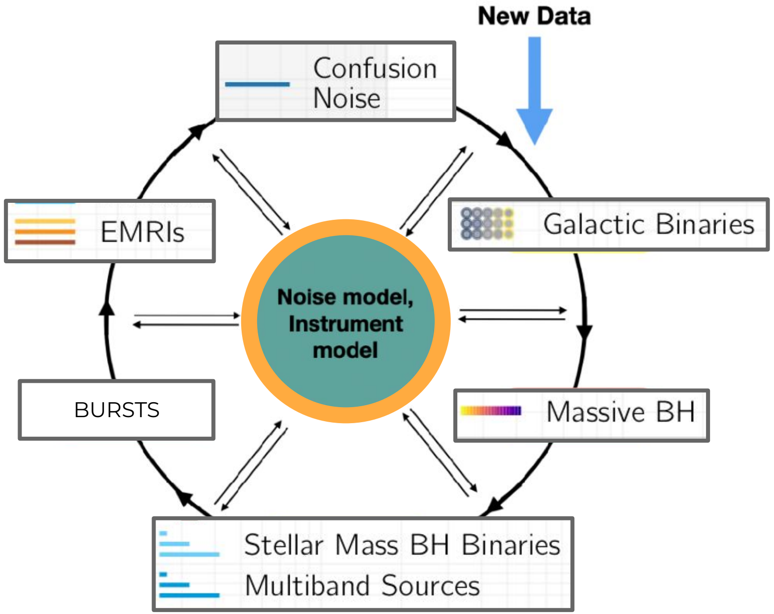
$$d = (M_X, M_Y, M_Z)$$

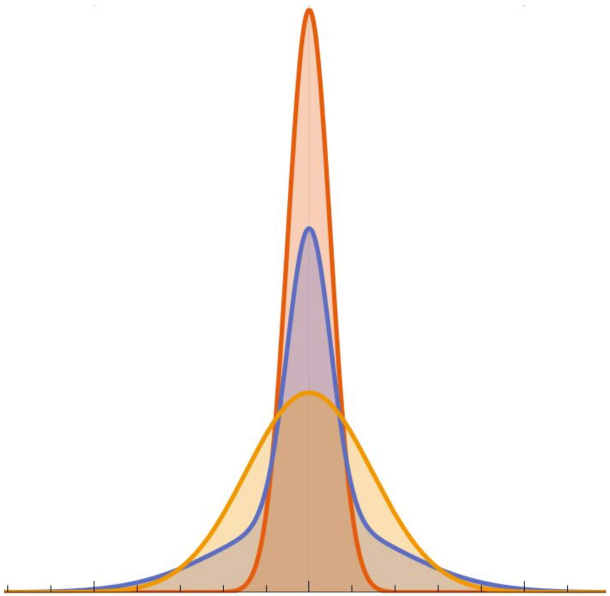
Why does it matter?

Source misidentification

Data model

$$s_i^{\mathcal{A}} = g_i^{\mathcal{A}} + h_i^{\mathcal{A}} + n_i^{\mathcal{A}}$$





Instrumental noise

Unresolved sources (True)

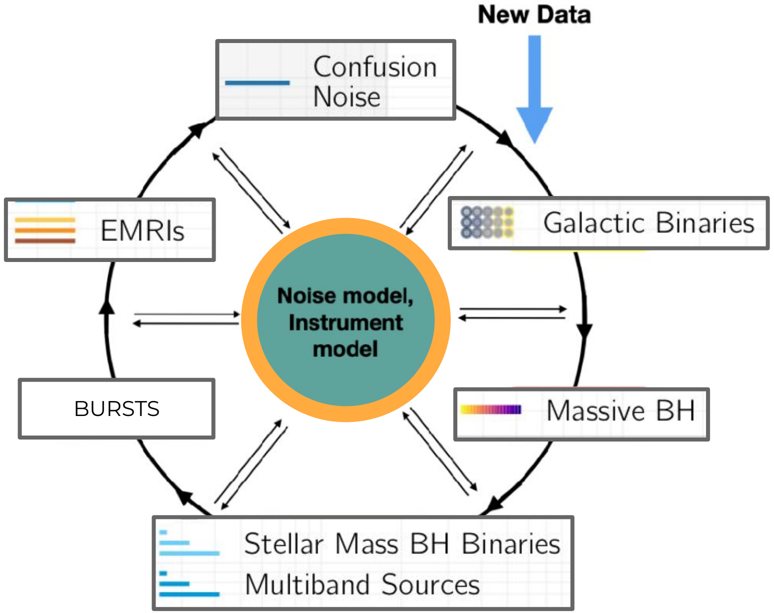
Unresolved sources (Gaussian approx.)

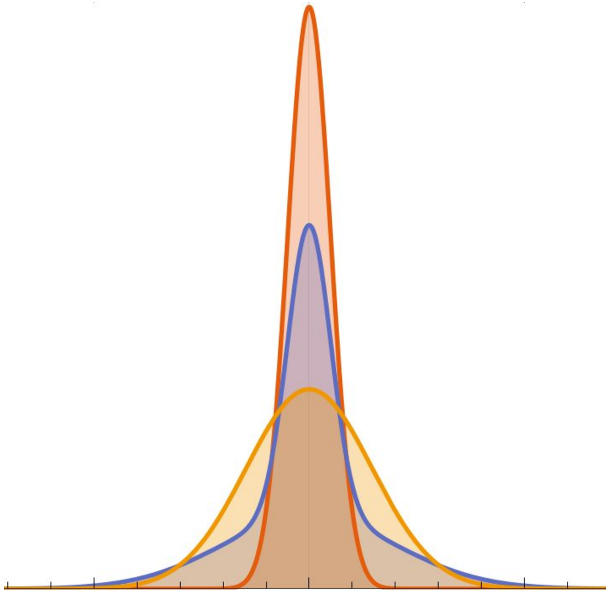
$\mathcal{L}(d \mid \theta)$ More difficult to absorb in SGWB

Easier to absorb in SGWB $\mathcal{L}(d \mid \theta)$

Why does it matter?

Source misidentification





Instrumental noise

Unresolved sources (True)

Unresolved sources (Gaussian approx.)

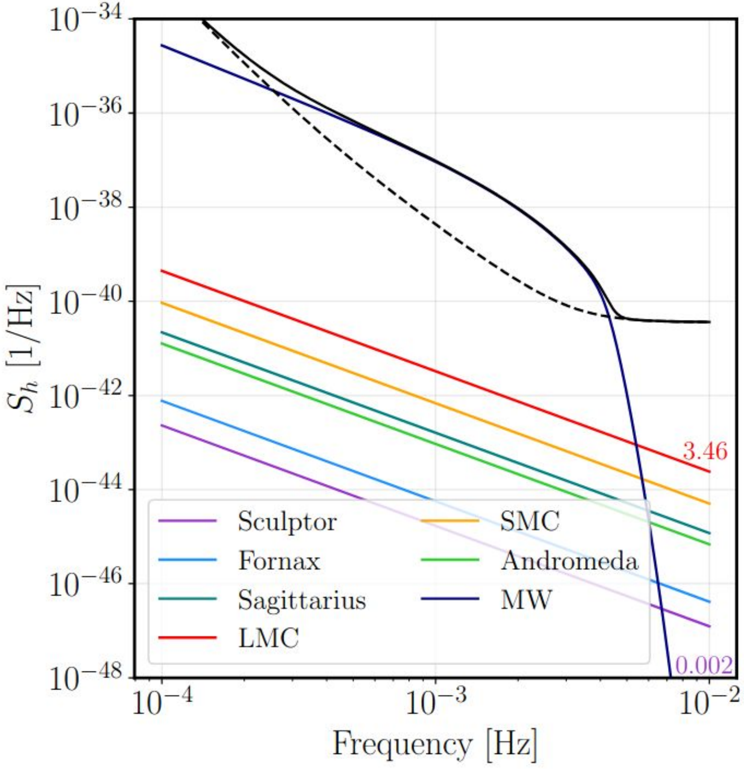
$\mathcal{L}(d \mid \theta)$

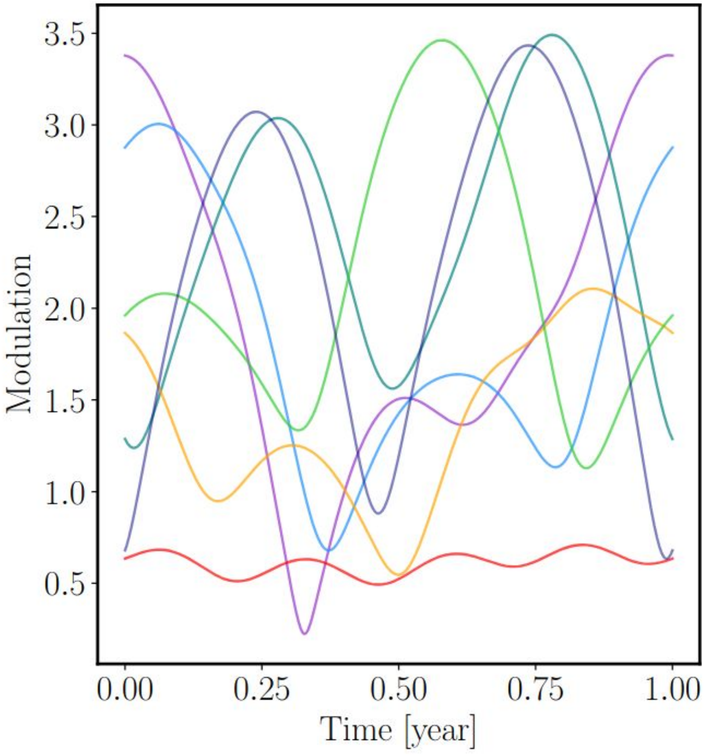
Can we help AI to help us? Likelihood-free methods?

$\mathcal{L}(d \mid \theta)$

How to help AI

Another example: quasi-stationarity





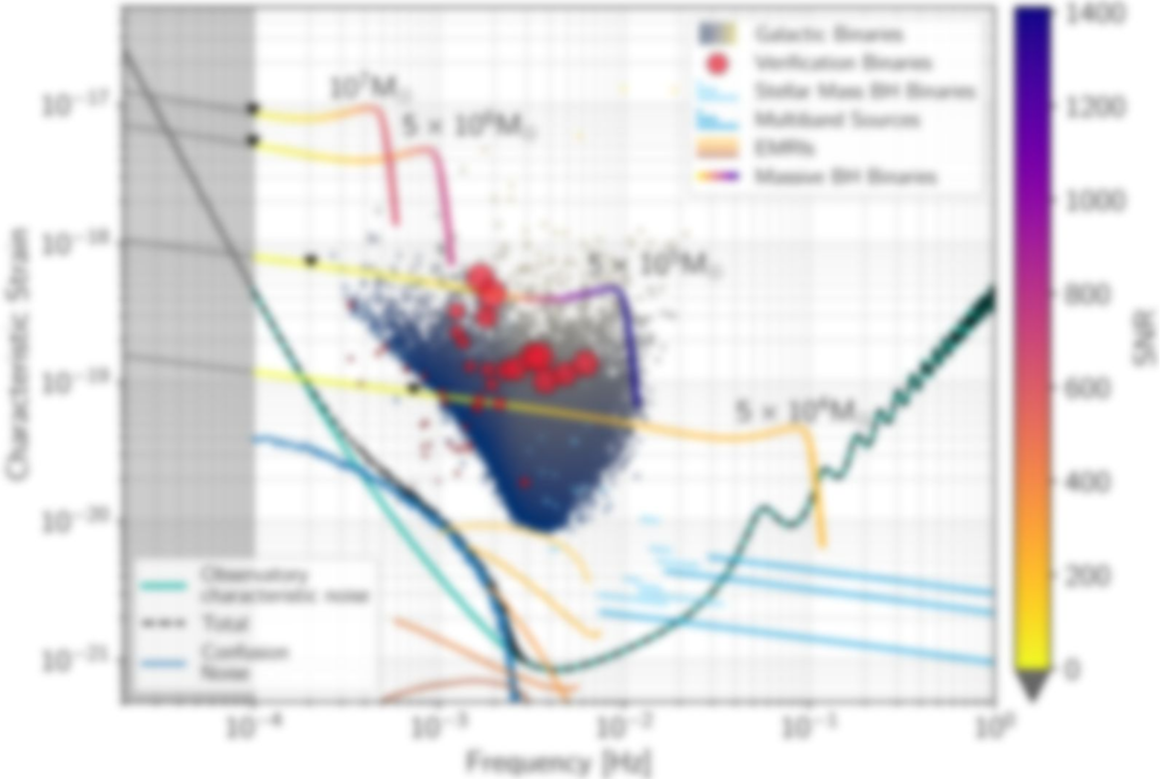
Pozzoli, Buscicchio+ (out next week)

Why does it matter?

10^4 needles in 10^{6} hay straws

Mismodelling biases:

- Biases in Ω_{GW}
- Misinterpret multiple $\{\Omega_{GW,i}\}$ as a single Ω_{GW}
- Individual source PE biases
- Population



Thanks

Questions?

Noise and signal

A new degree of freedom

Data model

$$s_i^{\mathcal{A}} = g_i^{\mathcal{A}} + h_i^{\mathcal{A}} + n_i^{\mathcal{A}}$$

$$p_n\left[n_i^{\mathcal{A}}\right] = \mathcal{N}_n \exp\left(-\tfrac{1}{2}\mathcal{W}_n(n,n)\right)$$

Noise model

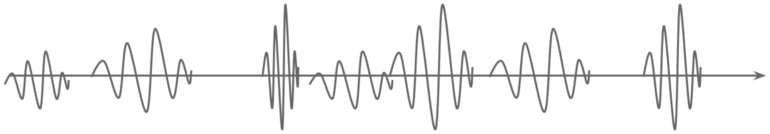
$$p_h\left[h_i\right] = \gamma_+\mathcal{N}(h_i;\sigma_+) + \mathcal{N}(h_i;\sigma_-)$$

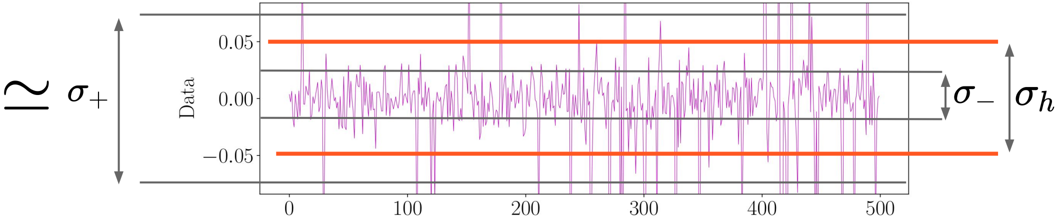
Signal model

$$\gamma_+ = \frac{\sigma_h^2 - \sigma_-^2}{\sigma_+^2 - \sigma_-^2} \qquad \gamma_- = \frac{\sigma_+^2 - \sigma_h^2}{\sigma_+^2 - \sigma_-^2}$$

$$\sigma_+ > \sigma_h > \sigma_-$$

Gaussian searches see only σ_h





Data

A new degree of freedom

$$p_s\left[s_i^{\mathcal{A}}\right] = \mathcal{N}_n\mathcal{N}_g \int_h \int_g p_h[h] \exp\left(-\tfrac{1}{2}\mathcal{W}_n(s-h-g, s-h-g) - \tfrac{1}{2}\mathcal{W}_g(g,g)\right)$$

$$= \mathcal{N}_{n+g} \int_h p_h[h] \exp\left(-\tfrac{1}{2}\mathcal{W}_{n+g}(s-h, s-h)\right)$$

Hard to model (Wick's theorem), easy to sample (see Buscicchio 2209.01400)

Frequentist approach

Bayesian approach

Improved statistics

A careful subtraction

Task: remove noise dominated non-zero terms under null-hypothesis

Most general solution:

$$\hat{y}_s = \hat{y}\left[\left(\sum_{\mathcal{A}} u^{\mathcal{A}}\right)^2\right] - \frac{1}{2^{N_D}} \underbrace{\sum_{\varepsilon_1=-1,1} \cdots \sum_{\varepsilon_{N_D}=-1,1}}_{\text{Combinatorics}} \hat{y}\left[\left(\sum_{\mathcal{A}} \varepsilon_{\mathcal{A}} u^{\mathcal{A}}\right)^2\right]$$

Exact, single-data statistics

Detector index

Noise-whitened data

Exact, single-data statistics

Remark 1: result does not depend on the specific choice of statistics.

Remark 2: result does not depend on GW model. It lives in “detector” indices.

Remark 3: result **is neither** perturbative in non-Gaussianity, **nor** in # of overlapping events (i.e. **neither** Regimbau, Mandic, **nor** Smith&Thrane)