Gravitational waves: data analysis

Stanislav Babak

Laboratoire AstroParticule & Cosmologie (APC), CNRS.
Univ. Paris Cite, France

Oct, 2025, Les Houches

Overview

Frequentist data analysis

Bayesian approach

Matched filtering

Matched filtering is a powerful technique for searching for a signal of known shape in noisy data.

- In practice the signal is parameterized by a vector of parameters θ (I also use $\theta_i = \vec{\theta}$).
- We will work in both time and frequency representations.
- Noise: $n(t)$. Assume zero-mean noise, stationary and Gaussian over the signal duration (non-stationarity: next lecture).
- Here are the properties of the noise: covariance kernel and PSD.

$$\mathbb{E}[\tilde{n}(t)] = 0, \quad C(\tau) = \mathbb{E}\Big[n(t)\, n(t+\tau)\Big]. \tag{1}$$

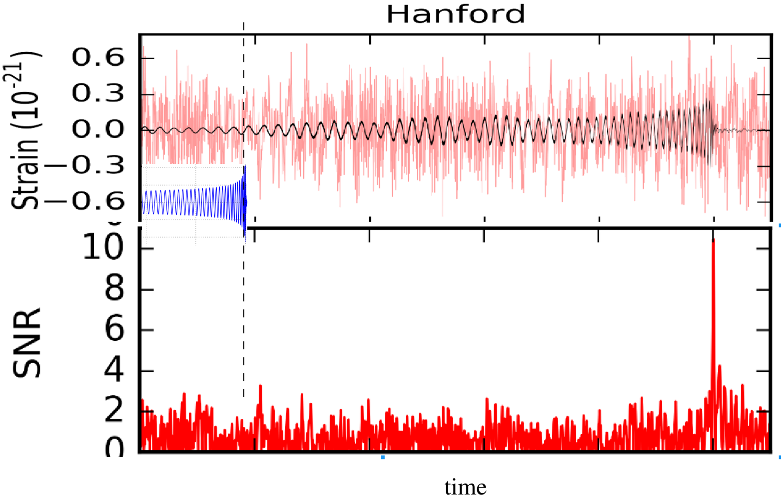
$$C(\tau) \;=\; \int_{-\infty}^{\infty} S^{(1)}(f)\, e^{2\pi i f \tau}\, \mathrm{d}f, \qquad S^{(1)}(f) \;=\; \int_{-\infty}^{\infty} C(\tau)\, e^{-2\pi i f \tau}\, \mathrm{d}\tau.$$

With one-sided PSD $S_n(f)$ for $f \geq 0$ (so that $\Phi(f) = \frac{1}{2} S_n(|f|)$),

$$C(\tau) \;=\; \int_0^{\infty} S_n(f)\, \cos(2\pi f \tau)\, \mathrm{d}f.$$

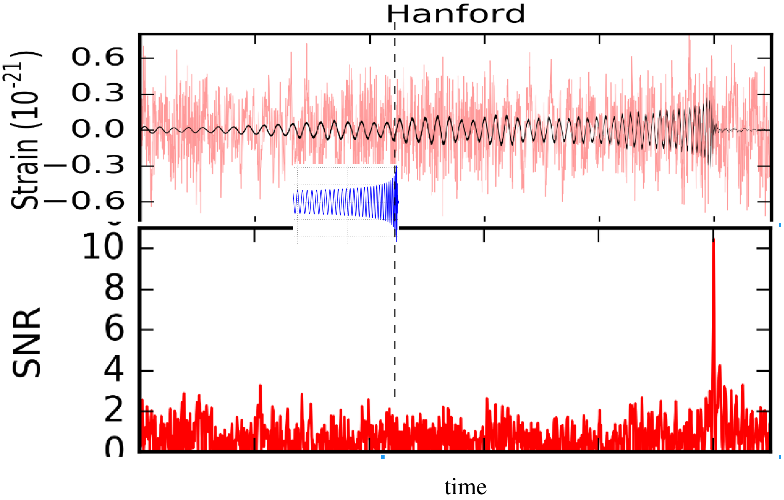
Matched filtering

We are searching for a signal of a specific shape buried in the noise:



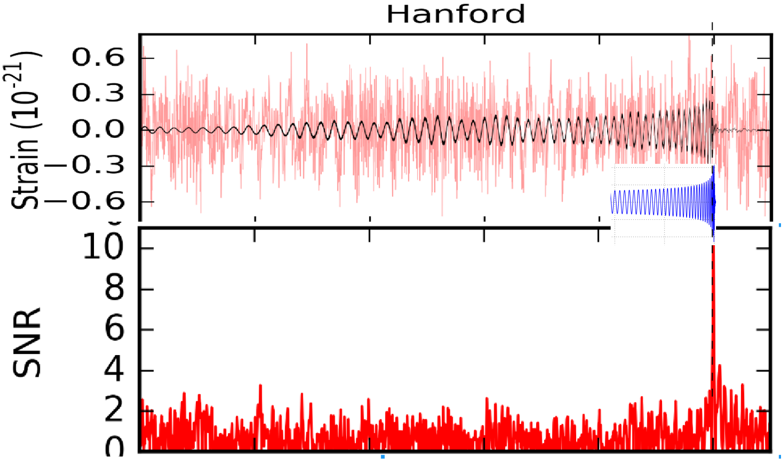
Matched filtering

We are searching for a signal of a specific shape buried in the noise:



Matched filtering

We are searching for a signal of a specific shape buried in the noise:



time

Matched filtering SNR

- Signal-to-noise ratio (SNR)

$$SNR = \sqrt{< h|h >}$$

where we have also introduced inner product

$$< d|h > = 4\Re \int_0^{f_{\max}} df \, \frac{\tilde{d}(f)\tilde{h}^*(f)}{S_n(f)} \tag{2}$$

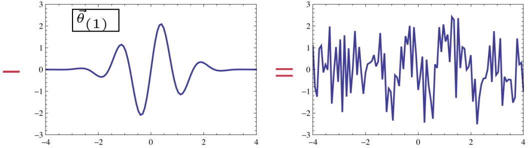
- Introduce matched filter SNR:

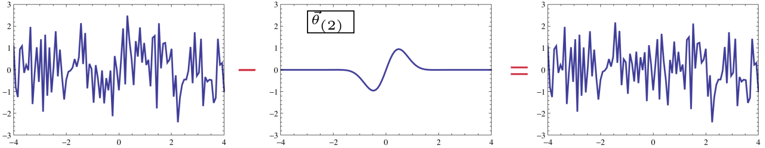
$$\rho = \frac{< d|h >}{\sqrt{< h|h >}} \tag{3}$$

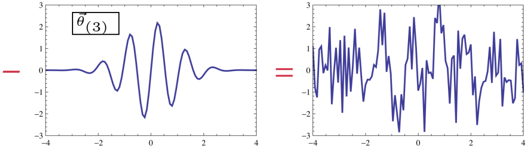
where $d(t)$ is data and tilde is a Fourier transform.
Note that $\bar{\rho} = SNR$, where bar means the average over noise realisations.

Matched filtering and parameter estimation

noise = data - signal



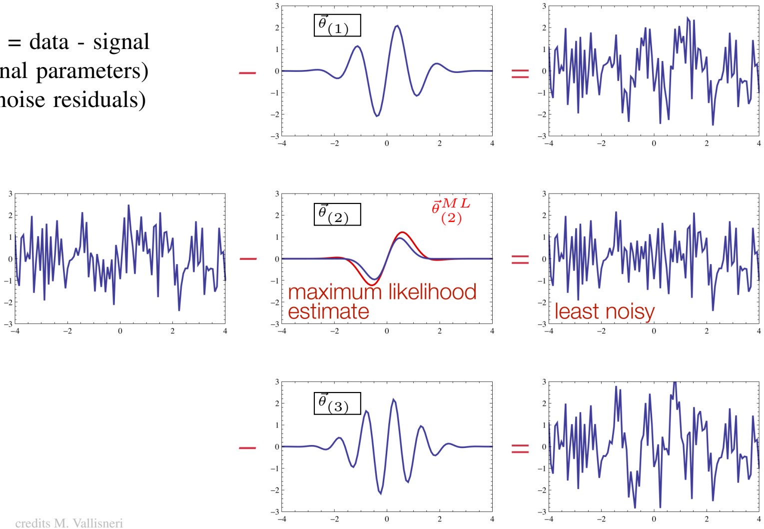




credits M. Vallisneri

Matched filtering and parameter estimation

noise = data - signal
p(signal parameters)
= p(noise residuals)



credits M. Vallisneri

Likelihood

Let us assume that the data contains the signal (signals are strong in LISA)

$$d(t) = n(t) + s(t) : \quad data = noise + signal.$$

The signal $s = s(\vec{\theta}, t)$ is a function of the parameters characterizing the GW source (signal). If the template (model of GW signal) represents *exactly*: $h(\vec{\theta}, t) = s(\vec{\theta}, t)$:

$$p(d(t)|s(\vec{\theta}, t)) = p(d(t) - s(\vec{\theta}, t)) = p_n$$

This is likelihood function: residuals after subtraction the correct signal should look like noise (with some statistical properties). For **gaussian noise**:

$$p(n) \propto e^{-\frac{1}{2} n(t_i) C^{-1}_{ij} n(t_j)} \tag{4}$$

or in frequency domain and taking into account $n = d - h(\vec{\theta})$

$$\mathcal{L}(d|h) \propto e^{-\frac{1}{2} < d - h(\vec{\theta}) | d - h(\vec{\theta}) >},$$

We will usually refer to $\mathcal{L}(d|\vec{\theta})$ (above) as likelihood. NOTE(!) I also use $p(d|\vec{\theta})$ as likelihood sometimes (when we consider Bayesian approach).

Frequentist approach

- Signal is deterministic; noise corrupts it.
- We want to detect the signal and estimate parameters.
- In Gaussian noise, the likelihood ratio is most powerful detection statistic (Neyman–Pearson).
- Competing models can be tested via their likelihoods.

Example: M_0 model (hypothesis) that data contains only noise $d(t) = n(t)$, M_1 model that data contains noise and a deterministic signal $s(t)$: $d(t) = s(t) + n(t)$. Then the log-likelihood ratio:

$$\Lambda = d^\top \boldsymbol{C}^{-1} h - \frac{1}{2} h^\top \boldsymbol{C}^{-1} h = < d|h(\vec{\theta}) > -\frac{1}{2} < h(\vec{\theta})|h(\vec{\theta}) > \quad (5)$$

Where we assumed that $s = h(\vec{\theta})$.
Note that $\bar{\Lambda} = \frac{1}{2}\mathrm{SNR}^2$

Frequentist approach: Detection

- Repeat the experiment conceptually to get sampling distributions of detection statistic (Λ) under model M_0 and M_1
- Search for maximum of Λ for each noise realisation under M_0 (backgroumd distribution) and M_1 (detection)
- Neyman-Pearson: Assume a false detection probability α:

$$\Pr_{\mathrm{M}_0}(\Lambda > \eta) = \int_{\eta}^{+\infty} \mathrm{p}(\Lambda, \mathrm{M}_0) d\Lambda = \alpha.$$

  this defines the threshold on the the detection η.
  We say that we have detection if $\Lambda(M_1) > \eta$ subject of false alarm α.
- Decreasing the f.d.p $\alpha \rightarrow$ the threshold increases $\rightarrow$ we need a higher Λ (SNR) to claim detection.
- only one experiment $\rightarrow$ rely on ergodicity and accurate noise modeling to calibrate thresholds.
- Requires estimation of distributions $p(\Lambda, M_{0,1})$: requires search for maximum likelihood for each simulation/experiment $\rightarrow$ computationally expensive

Frequentist approach: Parameter estimation

- We search for parameters that maximize the likelihood (or equivalently Λ) $\rightarrow$ we get *Maximum Likelihood Estimator (MLE)* of parameters of the signal $\vec{\theta}_{MLE}$
- For a given noise realisation: $\vec{\theta}_{MLE} \neq \vec{\theta}_{\mathrm{true}}$. (signal is corrupted by noise)
- Stronger the signal (higher SNR), closer the MLE to the true values (less influence of the noise):
$$|\vec{\theta}_{MLE} - \vec{\theta}_{\mathrm{true}}| \propto \frac{1}{\mathrm{SNR}}$$
- MLE is unbiased: $\overline{\vec{\theta}_{MLE}} = \vec{\theta}_{\mathrm{true}}$
- Maximization over amplitude: factor out the constant amplitude: $h \rightarrow Ah(t)$. Find MLE of the amplitude:
$$\frac{\partial \Lambda}{\partial A} = 0, \rightarrow A = \frac{< d|h >}{< h|h >}, \qquad \Lambda_{\max} = \frac{1}{2}\frac{< d|h >^2}{< h|h >} = \frac{1}{2} < d|\hat{h} >^2 = \frac{\rho^2}{2} \tag{6}$$
where we have introduced a normalised template $\hat{h} :< \hat{h}|\hat{h} >= 1$
- Unknown phase: Let $h(t;\phi) = h_c \cos\phi + h_s \sin\phi$ with $< h_c|h_s >= 0$, $< h_c|h_c >=< h_s|h_s >$. Then maximizing over ϕ: $\rho^2_{\max} =< d|\hat{h}_c >^2 + < d|\hat{h}_s >^2$.

Frequentist approach: parameter estimation

- **Bias.** Consider the case where $s \neq h(\vec{\theta}_{\rm true})$: our model is approximation of a true signal (usually the case)
- Minimising:
$$\min_{\vec{\theta}} < s - h(\vec{\theta})|s - h(\vec{\theta}) > \rightarrow \tilde{\theta}_i \quad (\vec{\theta}_{true} - \tilde{\theta}) = \delta\vec{\theta}$$
$\delta\vec{\theta}$ is a *bias* in parameter estimation.
- Bias does not depend on SNR. We should aim to keep bias below the statistical error $|\vec{\theta}_{true} - \vec{\theta}_{MLE}|$.
- **Overlap**. Overlap: $\mathcal{O}(s, h) =< \hat{s}|\hat{h} >= \cos\psi \in [0, 1]$. Ignores overall amplitudes: “angle between two signals"
- **Faithfulness**: $\mathcal{O}(s, h(\vec{\theta}_{\rm true}))$ – measure of similarity between the signal and the template (measure of goodness of approximation)
- The model might be not faithful but still fit well the signal on expense of the bias: **effectualness**. *fitting factor* is

$$\mathrm{FF} = \max_{\vec{\theta}} \mathcal{O}\Big(s, h(\vec{\theta})\Big), \tag{7}$$

$\mathcal{F}$-statistic

- Maximization of Λ over phase and amplitude is a partial case of $\mathcal{F}$-statistic
- Suppose $h(t) = \sum_{\mu=1}^{p} a^\mu h_\mu(t;\vartheta)$ with *linear* coefficients a^μ and *nonlinear* shape parameters ϑ. Define

$$X_\mu =< d|h_\mu >, \qquad M_{\mu\nu} =< h_\mu|h_\nu > . \tag{8}$$

Maximizing the likelihood over a^μ gives

$$\widehat{A}^\mu = (M^{-1})^{\mu\nu} X_\nu, \qquad 2\mathcal{F}(\vartheta) = X_\mu (M^{-1})^{\mu\nu} X_\nu. \tag{9}$$

In Gaussian noise: $2\mathcal{F} \sim \chi^2_p$ (central) under M_0; noncentral with parameter ρ^2 under M_1. Maximizing analytically over amplitudes reduces dimensionality (*fighting the curse of dimensionality*).

- Maximization over time of arrival (transient signals). Time shift τ corresponds to multiplication by $e^{2\pi i f\tau}$ in frequency. Use IFFT

$$\rho(\tau) = 4\Re \int df\, \frac{\tilde{d}(f)\hat{h}^*(f)}{S_n(f)} e^{2\pi i f\tau},$$

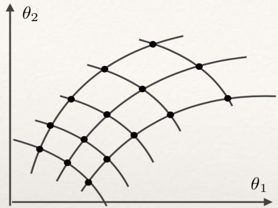
- Network inner product. If the noise is independent in each detector I

$$< h|h >_{\rm net} = \sum_I < h|h >_I \Rightarrow \rho^2_{\rm net} = \sum_I \rho^2_I.$$

Frequentist approach: parameter estimation

- For most parameters we need to do maximisation (search) numerically
- Problems: (i) likelihood surface is often multimodal: problem to find the global maximum; (ii) "Volume" of a signal (region of high likelihood) is small compared to to the total searched parameter space

Likelihood maximization: grid based search

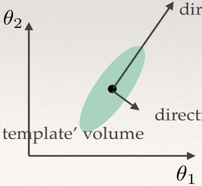


- We want to cover the parameter space (N-dim) by grid of points at equal distance from each other.
- Grid: not too coarse, not too fine
- The distance is determined **not** by a coordinate distance but by “proper” distance — correlation between nearby templates: introduce interval and metric

$$ds^2 = |\hat{h}(\theta_i + \delta\theta_i) - \hat{h}(\theta_i)| \approx (\hat{h}(\theta_i + \delta\theta_i) - \hat{h}(\theta_i)|\hat{h}(\theta_i + \delta\theta_i) - \hat{h}(\theta_i)) \approx \boxed{\left(\frac{\partial \hat{h}}{\partial \theta_i} \middle| \frac{\partial \hat{h}}{\partial \theta_j}\right)} \delta\theta_i \delta\theta_j$$

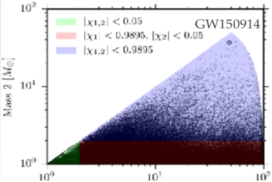
metric on the parameter maniold

Consider 2-D parameter space and fix ds = 0.01



the size and orientation of the ellipse is a function of a central pont

direction of weak correlation (chirp mass)



5

Grid-based search

- *pros:* Easy to parallelize, covers full parameter space, potentially can find all local maxima and several signals.
- *cons:* **VERY** hard to make a uniform grid in many dimentions (stochastic template bank). Computationally very expensive vbeyond 2-3-dim space.
- Hierarchical search: start with a coarse grid with low detectoin threshold and zoom in onto candidates.

Stochastic searches

We do a random walk in parameter space until we find the region(s) of high Λ. We use population approach $\rightarrow$ we choose a population of points scattered over parameter space. We "evolve population" towards high likelihood Λ.

- **PSO** (Particle Swarm Optimization). Maintain a population with positions $\boldsymbol{x}_i$ and velocities $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i \leftarrow w\,\boldsymbol{v}_i + c_1 r_1(\boldsymbol{x}_i^{\text{best}} - \boldsymbol{x}_i) + c_2 r_2(\boldsymbol{x}^{\text{gbest}} - \boldsymbol{x}_i), \tag{10}$$

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \boldsymbol{v}_i, \tag{11}$$

  with inertia w, accelerations $c_{1,2}$, and $r_{1,2} \sim U(0,1)$.

- **DE** (Differential Evolution). For target $\boldsymbol{x}_i$, pick distinct r_1, r_2, r_3 and form a donor

$$\boldsymbol{v}_i = \boldsymbol{x}_{r_1} + F\,(\boldsymbol{x}_{r_2} - \boldsymbol{x}_{r_3}), \quad F \in [0,2]. \tag{12}$$

  Generate a new point (crossover) for each particle j,

$$\boldsymbol{u}_j = \begin{cases} \boldsymbol{v}_i, & \text{if rand} < p_{\text{cross}}, \\ \boldsymbol{x}_j, & \text{otherwise}, \end{cases} \tag{13}$$

  with $p_{\text{cross}} \in [0,1]$. *Selection:* if $\mathcal{L}(\boldsymbol{u}_j) > \mathcal{L}(\boldsymbol{x}_j)$, set $\boldsymbol{x}_i \leftarrow \boldsymbol{u}_i$; else keep $\boldsymbol{x}_i$.

Bayesian approach

- As the name says $\rightarrow$ based on the Bayesian equality.
- Deeper: We treat the signal as deterministic (stochastic signals need a separate lecture) but parameters defining the model are *random variables*. We need to adopt *prior* distribution $\pi(\vec{\theta})$ before we start analysing the data
- Choice of the prior: based on the physical principles, on another expriment or assumed physical model (e.g. population of sources). Alternatively: non-informative prior (uniform, or log-uniform (scale invariant). see also Jeffreye prior.
- Likelihood updates our prior knowledge based on the observations (how likely to observe this data given a model)
- Updated results are *posteriors*

Bayesian approach

- Assume model M_is parametrized by a set of parameter $\vec{\theta}_i$:

Likelihood prior

$$\underbrace{P(\vec{\theta}_i|d, M_i)}_{\text{Posterior}} = \frac{\overbrace{P(d|\vec{\theta}_i, M_i)}^{\text{Likelihood}}\ \overbrace{\pi(\vec{\theta}_i)}^{\text{prior}}}{\underbrace{P(d|M_i)}_{\text{evidence}}} x$$

Posterior evidence

Evidence (or marginalised likelihood)

$$p(d \mid M_i) = \int p(d \mid \boldsymbol{\theta}, M_i)\pi(\boldsymbol{\theta} \mid M_i)\mathrm{d}\boldsymbol{\theta}.$$

Bayesian approach: Bayes factor

- We cna formulate Bayes theorem for model M_i:

$$P(M_i, d) = \frac{P(d|M_i)\pi(M_i)}{p(d)}$$

  Probability of modle M_i given observed data d. The problem is $p(d) \rightarrow$ requires considering a complete set of models $\rightarrow$ almost never possible

- Consider models M_i and M_j with priors $p(M_i)$ and $p(M_j)$, the *odds ratio* is

$$\mathcal{O}ij \equiv \frac{p(M_i \mid d)}{p(M_j \mid d)} = \underbrace{\frac{p(d \mid M_i)}{p(d \mid M_j)}}_{\mathcal{B}ij} \times \underbrace{\frac{\pi(M_i)}{\pi(M_j)}}_{\text{prior odds}}, \quad (14)$$

  where $\mathcal{B}ij$ is the *Bayes factor*.

- Often assume equal (uniform) prior on all considered models (very wrong for TGR)

Bayesian approach: detection

- Detection in Bayesian approach is based on computation of Bayes factor (or odd ratio)
- Examples of models (i) noise only or noise + signal (ii) model with 2 or with 3 sources...
- the big question is how to set threshold on the Bayes factor: c.d.c

Table: Evidence strength for Bayes factors in favor of M_1 over M_0 (all in B_{10}).

Strength label	Jeffreys (in B_{10})	Kass–Raftery (in B_{10})	Lee–Wagenmakers
Barely / Anecdotal	1–3.2	$1–e^1 \approx 2.72$	1–3
Substantial / Positive / Moderate	3.2–10	$e^1–e^3 \approx 2.72–20.1$	3–10
Strong	10–31.6	$e^3–e^5 \approx 20.1–148.4$	10–30
Very strong	31.6–100	$> e^5 \approx 148.4$	30–100
Decisive / Extreme	> 100	> 148.4	> 100

Notes: Jeffreys' original bands were in $\log_{10} B_{10}$; we converted via $B_{10} = 10^{(\log_{10} B_{10})}$. Kass–Raftery reported ranges in $2 \ln B_{10}$; we converted with $B_{10} = \exp\big((2 \ln B_{10})/2\big)$. Labels are heuristic; decisions should also state prior odds and costs.

Markov chain Monte-Carlo (MCMC)

- Come back to a single model. In Bayesian inference we target the posterior (parameter estimation)

$$p(\theta|d, M) \;\propto\; p(d|\theta, M)\,\pi(\theta|M), \tag{15}$$

- We construct Markov chain: stochastic process where the next point in the chain depends only on the previous. We use the transitional probability $\vec{\theta}_t \rightarrow \vec{\theta}_{t+1}$. The chain sampling the target distribution if is it stationary (time reversable) $\rightarrow$ satisfies detailed balance

$$p(\vec{\theta}_t)P(\vec{\theta}_{t+1}|\vec{\theta}_t) = p(\vec{\theta}_{t+1})P(\vec{\theta}_t|\vec{\theta}_{t+1})$$

Metropolis-Hastings transitional kernel

- Consider a particular way of building the tansitional probability (Metropolis-Hastings)
- introduce proposal density $q(\theta' \mid \theta)$.
- Given the current state θ, propose $\theta' \sim q(\cdot \mid \theta)$ and accept with probability

$$\alpha(\theta,\theta') = \min\left\{1,\ \frac{\mathcal{L}(d|\theta',M)\,\pi(\theta'|M)}{\mathcal{L}(d|\theta,M)\,\pi(\theta|M)}\cdot\frac{q(\theta\mid\theta')}{q(\theta'\mid\theta)}\right\}. \tag{16}$$

Log form (numerically stable):

$$\log\alpha(\theta,\theta') = \min\Bigg\{0,\ \underbrace{\log\mathcal{L}(d|\theta',M)-\log\mathcal{L}(d|\theta,M)}_{\text{likelihood change}} + \underbrace{\log\pi(\theta'|M)-\log\pi(\theta|M)}_{\text{prior change}} + \underbrace{\log q(\theta\mid\theta')-\log q(\theta'\mid\theta)}_{\text{proposal asymmetry}}\Bigg\}. \tag{17}$$

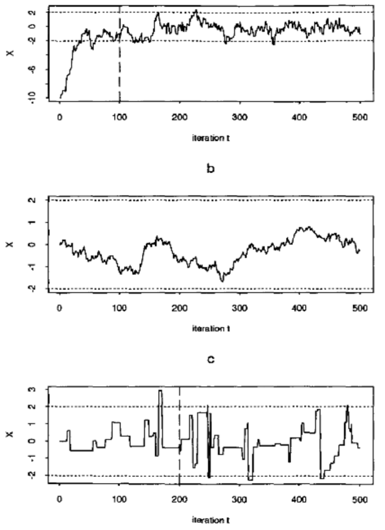
Metropolis-Hastings transitional kernel

- One can check that the detailed balance equation is satisfied:

$$p(\theta|d, M)\, q(\theta' \mid \theta)\, \alpha(\theta, \theta') \; = \; p(\theta'|d, M)\, q(\theta \mid \theta')\, \alpha(\theta', \theta),$$

- The final result does not depend on the proposal, but efficiency does.
- Efficiency: high acceptance rate and low autocorrelation length (see next slide)
- **Local geometry (random-walk)**: $q(\theta' \mid \theta) = \mathcal{N}(\theta, \Sigma)$ exploits local curvature; step scale Σ sets the acceptance/mixing balance. In high dimensions, scaling theory suggests an optimal acceptance near ~ 0.23 for isotropic random-walk MH.
- **Global jumps (independence)**: $q(\theta' \mid \theta) = g(\theta')$ can traverse modes if g approximates the posterior's bulk/tails; mis-matched g yields very low acceptance.
- **Heavy tails & robustness**: Student-t or mixture proposals improve mode-hopping and outlier robustness.
- **Preconditioning and blocking**: Use a covariance aligned with posterior correlations (e.g., empirical from pilot runs); update strongly correlated coordinates together; weakly coupled ones in separate blocks.
- **Mixtures and schedules**: Combine small/medium/large steps to balance local refinement and occasional long moves.

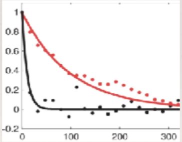
Metropolis-Hastings: proposal



- The theorem tells us that the chains will sample the posterior pdf (after some burn-in length) independent of the proposal distribution, BUT

- The efficiency of the sampling strongly depends on the proposal (proposal should resemble the posterior)

- Number of samples vs. number of independent samples (defined by autocorrelation length)



- Multimodal posterior require special treatment! (simulated annealing, parallel tempering)

10

Parallel tempering

- Parallel tempering runs C Markov chains in parallel at inverse temperatures $\beta_1 > \beta_2 > \cdots > \beta_C$ (with $\beta_1 = 1$ the *cold* chain). Chain c targets the tempered posterior

$$p_{\beta_c}(\theta) \;\propto\; \mathcal{L}(d|\theta, M)^{\beta_c}\,\pi(\theta|M), \qquad \beta_c \in (0, 1]. \tag{18}$$

Parallel tempering

- To share information across temperatures, occasionally propose a *swap* of states between chains c and c' (typically neighbors, $c' = c + 1$). Let $s(c, c')$ be the probability of proposing that pair (often symmetric and uniform). Given the current pair $(\theta_c, \theta_{c'})$, propose

$$(\theta_c, \theta_{c'}) \longrightarrow (\theta_{c'}, \theta_c)$$

and accept with probability

$$\alpha_{\mathrm{swap}} = \min\left\{1, \frac{p_{\beta_c}(\theta_{c'})\, p_{\beta_{c'}}(\theta_c)}{p_{\beta_c}(\theta_c)\, p_{\beta_{c'}}(\theta_{c'})} \cdot \frac{s(c', c)}{s(c, c')}\right\}. \tag{19}$$

With the tempered targets in (18) and a symmetric s, priors cancel and the ratio simplifies to

$$\alpha_{\mathrm{swap}} = \min\left\{1, \exp\left[(\beta_c - \beta_{c'})(\log \mathcal{L}(d|\theta_{c'}, M) - \log \mathcal{L}(d|\theta_c, M))\right]\right\}. \tag{20}$$

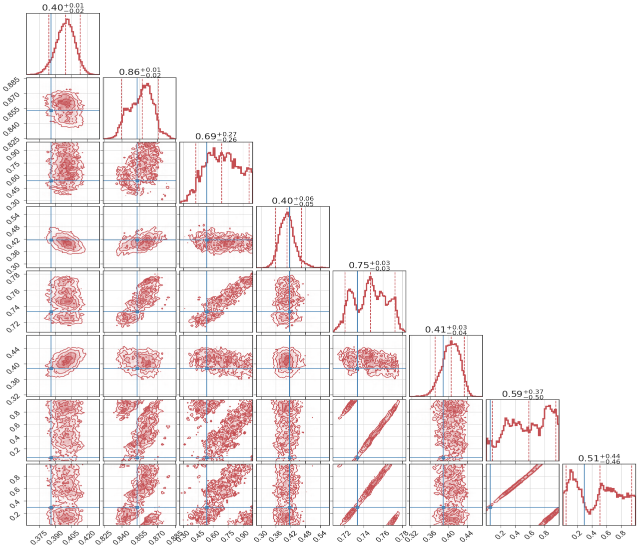
- Divide the chains in even/odd ad sawp parwise parallel ($0 \to 1$, $2 \to 3$, $4 \to 5$, ...)

Parallel tempering

Let $1 = \beta_1 > \beta_2 > \cdots > \beta_C > 0$ (equivalently $T_c = 1/\beta_c$). Guidelines:

- **Geometric spacing:** $\beta_c = \beta_1 \, r^{\,c-1}$ with $r \in (0, 1)$ is a common baseline.
- **Coverage:** choose $T_{\max}$ (smallest β_C) high enough that p_{β_C} is nearly prior-dominated, facilitating global moves.
- **Swap rates:** target neighbor swap acceptances in the range 0.2–0.6; refine spacing near $\beta \approx 1$ if needed (denser ladder where the posterior geometry is sharpest).
- **Adaptive (between runs/epochs):** estimate empirical log-likelihood variances to place $\{\beta_c\}$ so that adjacent energy (negative log-likelihood) distributions overlap sufficiently; adjust only between epochs to avoid breaking Markov stationarity.

Corner plot


$0.40^{+0.01}_{-0.02}$
$0.86^{+0.01}_{-0.02}$
$0.69^{+0.27}_{-0.26}$
$0.40^{+0.06}_{-0.05}$
$0.75^{+0.03}_{-0.03}$
$0.41^{+0.03}_{-0.04}$
$0.59^{+0.37}_{-0.50}$
$0.51^{+0.44}_{-0.46}$

Bayes factor: Nested sampling

- The Bayesian evidence (marginal likelihood) is $p(d \mid M) = p(d|M) = \int \mathcal{L}(d|\theta, M)\,\pi(\theta|M)\,\mathrm{d}\theta$. Nested Sampling rewrites this integral as a one-dimensional integral over the *prior-mass* variable $X \in [0, 1]$.
- So...

$$X(\lambda) \;=\; \int_{\{\mathcal{L}(d|\theta,M)>\lambda\}} \pi(\theta|M)\,\mathrm{d}\theta \;=\; \Pr_{\theta\sim\pi}\Big[\mathcal{L}(d|\theta, M) > \lambda\Big], \tag{21}$$

which is non-increasing from $X(0) = 1$ to $X(+\infty) = 0$.
By the layer-cake (Lebesgue) representation,

$$Z(M) \equiv p(d \mid M) \;=\; \int_0^{\infty} X(\lambda)\,\mathrm{d}\lambda \;=\; \int_0^{1} \mathcal{L}(X)\,\mathrm{d}X, \qquad \mathcal{L}(X) := \lambda \text{ such that } X(\lambda) = \tag{22}$$

Geometrically, X measures the remaining prior volume *inside* the iso-likelihood contour $\{\theta : \mathcal{L}(\theta) > \lambda\}$. As $X \downarrow 0$, one moves to higher likelihood levels.

-

Nested sampling: Algorithm sketch.

1. Initialize $N_{\rm live}$ points $\{\theta^{(n)}\}_{n=1}^{N_{\rm live}} \sim \pi(\theta|M)$; set $X_0 = 1$, $p(d \mid M) \leftarrow 0$.
2. For $i = 1, 2, \ldots$:
   - 2.1 Identify the worst live point θ_i with likelihood $\mathcal{L}_i$.
   - 2.2 Draw $t_i \sim \mathrm{Beta}(N_{\rm live}, 1)$ and set $X_i = t_i X_{i-1}$; weight $w_i = X_{i-1} - X_i$. (often replaced by $\mathbb{E}[t_i] = e^{-1/N_{\rm live}}$).
   - 2.3 Accumulate evidence: $p(d \mid M) \leftarrow p(d \mid M) + \mathcal{L}_i\, w_i$.
   - 2.4 Save θ_i with weight $\mathcal{L}_i w_i$ (for posterior).
   - 2.5 Replace θ_i by a new constrained-prior sample with $\mathcal{L} > \mathcal{L}_i$.
3. Terminate when $p(d \mid M)_{\rm rem} \;\approx\; X_i\ \max_{n \le N_{\rm live}} \mathcal{L}(\theta^{(n)}_{\rm live})$, is negligible; add the final live set contribution.

- Advice: map the prior to a uniform in unit hypercube
- Hard part: drawing a new θ from $\pi(\theta|M)$ *restricted* to $\{\mathcal{L} > \mathcal{L}_i\}$. Need to cover multimodality
- Existing tools: dynesty, multinest, polychord, cpnest, nessai, ...
- We need to compute evidence for each competing model to estimate Bayes factor

$$\mathcal{B}ij = \frac{Z(M_i)}{Z(M_j)}$$

Reversible-jump MCMC

- RJ MCMC - transdimensional extension of MH $\rightarrow$ can do model selection without explicitly computing evidence.
- Introduce model state m and corresponding parameter space $\vec{\theta}_m$ then the joined target distribution

$$p(m, \theta|d) \; \propto \; \mathcal{L}(d|\theta, m)\, \pi(\theta|m)\, p(m), \tag{23}$$

- Models could have different dimensionality: need to match dimensions during jumpe between model (changing the models state m)
- At a between-model update, a candidate change of model index from m to m' is *selected* with probability $r_{m\to m'} \in (0, 1]$ (e.g., "birth" vs. "death" move probabilities in a two-model setting). Conditional on selecting $m \to m'$, draw $u \sim q_{m\to m'}(u|\theta, m)$, and define a bijection

$$(\theta', u') \;=\; \mathcal{T}_{m\to m'}(\theta, u), \qquad \text{with} \quad \dim(\theta) + \dim(u) \;=\; \dim(\theta') + \dim(u').$$

Let $J = \left|\det \partial(\theta', u')/\partial(\theta, u)\right|$ be the Jacobian of $\mathcal{T}$.

RJ MCMC

- The Metropolis-Hastings acceptance for the between-model move is

$$\alpha\Big[(m,\theta)\to(m',\theta')\Big] = \min\left\{1,\ \frac{\mathcal{L}(d|\theta',m')\,\pi(\theta'|m')\,p(m')\,r_{m'\to m}\,q_{m'\to m}(u'|\theta',m')\,J}{\mathcal{L}(d|\theta,m)\,\pi(\theta|m)\,p(m)\,r_{m\to m'}\,q_{m\to m'}(u|\theta,m)}\right\}. \tag{24}$$

- Consider an example of two nested models $M_0(\theta)$ and $M_1(\theta,\psi)$
- Birth move $M_0 \to M_1$ (add ψ). Use identity embedding for θ and propose the new block via the auxiliary:

$$u \sim q_b(u|\theta,M_0), \qquad \mathcal{T}_{0\to1}:\ (\theta,u)\mapsto(\theta',\psi')=(\theta,u), \quad u'=\varnothing.$$

This mapping has $J=1$. Let $r_{0\to1}$ be the probability of proposing a birth (and $r_{1\to0}$ a death). The acceptance becomes

$$\alpha_{\text{birth}} = \min\left\{1,\ \frac{L(d|\theta,\psi,M_1)\ \pi_1(\theta,\psi)\ \pi(M_1)\ r_{1\to0}}{L(d|\theta,M_0)\ \pi_0(\theta)\ \pi(M_0)\ r_{0\to1}\ q_b(\psi|\theta,M_0)}\right\}. \tag{25}$$

RJ MCMC

- Death move $M_1 \to M_0$ (remove ψ). The reverse mapping drops the extra block:

$$\mathcal{T}_{1\to 0}: \ (\theta, \psi) \mapsto \theta' = \theta, \quad \text{and define } u' = \psi \text{ for the reverse density } q_b(u'|\theta', M_0).$$

Again $J = 1$. The acceptance is

$$\alpha_{\text{death}} = \min\left\{1, \ \frac{\mathcal{L}(d|\theta, M_0)\ \pi_0(\theta)\ p(M_0)\ r_{0\to 1}\ q_b(\psi|\theta, M_0)}{\mathcal{L}(d|\theta, \psi, M_1)\ \pi_1(\theta, \psi)\ p(M_1)\ r_{1\to 0}}\right\}. \tag{26}$$

- If the priors factorize as $\pi_1(\theta, \psi) = \pi(\theta)\pi(\psi)$ and $\pi_0(\theta) = \pi(\theta)$, then $\pi(\theta)$ cancels in (25) and

$$\alpha_{\text{birth}} = \min\left\{1, \ \frac{\mathcal{L}(d|\theta, \psi, M_1)\ p(M_1)\ r_{1\to 0}\ \pi(\psi)}{\mathcal{L}(d|\theta, M_0)\ p(M_0)\ r_{0\to 1}\ q_b(\psi|\theta, M_0)}\right\}. \tag{27}$$

- Bayes factor:

$$\frac{\text{time in } \mathrm{M}_1}{\text{time in } \mathrm{M}_0} \approx \frac{p(M_1|d)}{p(M_0|d)} = \mathcal{B}_{1,0}\frac{p(M_1)}{p(M_0)}$$

- **Prior-matching birth:** A simple, valid choice is $q_b(\psi|\theta, M_0) = \pi(\psi|\theta, M_1)$
- **Posterior-informed birth:** If feasible, center q_b near an estimate of the conditional posterior $p(\psi|\theta, d, M_1)$ (e.g., Laplace approximation)

RJ MCMC algorithm sketch

1. With probability $r_{0\to1}$ (if in M_0) propose birth: draw $\psi \sim q_b(\psi|\theta, M_0)$, set $(\theta', \psi') = (\theta, \psi)$, accept with α_{birth}.
2. With probability $r_{1\to0}$ (if in M_1) propose death: drop ψ and accept with α_{death}.
3. Otherwise, perform within-model updates of current parameters (e.g., MH on θ in M_0; MH on (θ, ψ) in M_1).

Gibbs sampling

- Let posterior target be

$$p(\theta \mid d, M) \;\propto\; L(d \mid \theta, M)\, \pi(\theta \mid M),$$

- partition the parameters into blocks $\theta = (\theta_1, \ldots, \theta_B)$. The full conditional of block b is

$$p(\theta_b \mid \theta_{-b}, d, M) \;\propto\; L(d \mid \theta_b, \theta_{-b}, M)\, \pi(\theta_b \mid \theta_{-b}, M),$$

with $\theta_{-b} = (\theta_1, \ldots, \theta_{b-1}, \theta_{b+1}, \ldots, \theta_B)$.

- Systematic scan. A Gibbs sweep replaces each block by an exact draw from its full conditional:

$$\theta_1^{(t+1)} \sim p(\theta_1 \mid \theta_{-1}^{(t)}, d, M), \quad \theta_2^{(t+1)} \sim p(\theta_2 \mid \theta_1^{(t+1)}, \theta_{-(1,2)}^{(t)}, d, M), \; \ldots,$$

$$\theta_B^{(t+1)} \sim p(\theta_B \mid \theta_{-B}^{(t+1)}, d, M).$$

- Each draw is accepted with probability 1 (if known analytically). Otherwise we can still use MH to draw from the conditional probabilities.

Product space

Estimation of Bayes factor using "product space" approach

- Idea is similar but avoids the dimensionality matching
- Static number of models for comparison
- Core idea: make a super (product) space of all parameters of all models
- Models are indexed by $j = 1, \dots, J$, with parameter blocks $\vec{\theta}_j \in \Theta_j$ and union $\theta = (\vec{\theta}_1, \dots, \vec{\theta}_J) \in \Theta = \prod_j \Theta_j$. Let the binary state vector be $m = (m_1, \dots, m_J) \in \{0,1\}^J$ (m_j=1 "on", m_j=0 "off").
- The *product-space* joint target is

$$p(m, \theta|d) \;\propto\; \mathcal{L}(d|\theta, m)\, \pi(m)\, \pi(\theta|m), \qquad \pi(\theta|m) \;=\; \prod_{j=1}^{J} \left[\pi_j(\vec{\theta}_j)\right]^{m_j} \left[\tilde{\pi}_j(\vec{\theta}_j)\right]^{1-m_j}, \tag{28}$$

- $\pi(m)$ is the prior on states, π_j is the *true prior* for block j when "on", and $\tilde{\pi}_j$ is a *pseudo-prior* supplying a proper density for the same block when "off".
- Likelihood: $\mathcal{L}(d|\theta, m) \to$

$$\log \mathcal{L} = -\frac{1}{2} \sum_{m_i} < d - m_i h(\vec{\theta}_i) | d - m_i h(\vec{\theta}_i) >$$

Product space

- Gibbs scheme: **state update** $\rightarrow$ **in-model update**
- State update (change m_j), θ are fixed

$$r_j = \frac{\pi(m^{(j=1)})}{\pi(m^{(j=0)})} \cdot \frac{\mathcal{L}(d \mid \theta,\, m^{(j=1)})}{\mathcal{L}(d \mid \theta,\, m^{(j=0)})} \cdot \frac{\pi_j(\theta_j)}{\tilde{\pi}_j(\theta_j)}, \tag{29}$$

  Probability of accepting $m_j \rightarrow 1$:

$$p(m_j{=}1 \mid \cdot) = 1/(1 + e^{-x}) \tag{30}$$

  where $x = \log r_j$
- in-model updates: we can use Metropolis-Hastings ratio

Product space: two nested models

- Consider $M_0(\theta)$ and $M_1(\theta, \psi)$ with shared θ and an extra block ψ in M_1. Introduce a model index $k \in \{0, 1\}$ and keep both (θ, ψ) in the state:

$$\pi(\theta, \psi \mid k) \;=\; \begin{cases} \pi_0(\theta)\, \tilde{\pi}(\psi), & k = 0, \\ \pi_1(\theta)\, \pi(\psi), & k = 1, \end{cases} \qquad \text{with model prior } \pi(k). \tag{31}$$

The joint target is

$$p(k, \theta, \psi \mid d) \;\propto\; \pi(k)\, \pi(\theta, \psi \mid k)\, \mathcal{L}(d \mid \theta, \psi, k),$$

with $\mathcal{L}(d \mid \theta, \psi, k{=}0) = \mathcal{L}(d \mid \theta, M_0)$ and $\mathcal{L}(d \mid \theta, \psi, k{=}1) = \mathcal{L}(d \mid \theta, \psi, M_1)$.

Product space: two nested models

- Jump between models (fixed parameters)

$$lr = \log \frac{\pi(k=1)}{\pi(k=0)} + \log \frac{\mathcal{L}(d \mid \theta, \psi,\ M_1)}{\mathcal{L}(d \mid \theta,\ M_0)} + \log \frac{\pi(\theta, \psi)}{\pi(\theta)\tilde{\pi}(\psi)} \tag{32}$$

- We accept $k = 1$ with probability $\propto$ Bernoulli$(\sigma(lr))$, where $\sigma(x) = (1 + e^{-x})^{-1}$ (sigmoid).
- in-model update. We use Metropolis-Hastings ratio. We update all parameters using a proposal $q(\theta_{t+1}, \psi_{t+1} | \theta_t, \psi_t, k_{t+1})$. Preferably using block-Gibbs update again:

$$\theta_t \rightarrow \theta_{t+1} | \theta_t, \psi_t, k_{t+1}, \qquad \psi_t \rightarrow \psi_{t+1} | \theta_{t+1}, \psi_t, k_{t+1} \tag{33}$$

  going to ψ_{t+1} if $k_{t+1} = 0$ using pseudo priors.
- Bayes factor is again:

$$\frac{\text{time in M}_1}{\text{time in M}_0} \approx \frac{p(M_1|d)}{p(M_0|d)} = \mathcal{B}_{1,0} \frac{p(M_1)}{p(M_0)}$$

- Choice of $\tilde{\pi}(\psi)$: close to the marginal posterior of ψ under M_1 (e.g., Gaussian or ...)

Global fit in LISA

- We use block-Gibbs update across population of components
- components: **noise**, **GB**, **MBHB**, **EMRI**, ...
- updates

$$\text{noise}_{t+1}|\text{GB}_t, \text{MBHB}_t, \text{EMRI}_t, ...$$

$$\text{GB}_{t+1}|\text{noise}_{t+1}, \text{MBHB}_t, \text{EMRI}_t, ...$$

$$\text{MBHB}_{t+1}|\text{noise}_{t+1}, \text{GB}_{t+1}, \text{EMRI}_t, ...$$

- Given a noise model, we compute likelihood based on the residuals:

$$d - \sum_i h_i^{\text{GB}} - \sum_j h_j^{\text{MBHB}} - \dots$$

- if sources are not correlated (narrow-band GW signals from GBs, short duration MBHB mergers) we can use "parallel Gibbs", update uncorrelated sources in parallel (not sequentially). Example odd and even frequency sub-bands for GBs
- Noise is correlated with everything (enters inner product): need to be updated sequentially.

Non-stationary noise

Consider two types of non-stationarity: (1) (slow) drift of the noise level in time (2) transient non-stationary features (gaps, glitches)

- Slow drift: can split data into segments where the noise is approximately stationary $\to S_n(\tau_k, f)$, where τ_k is a centre of a time segment.

$$\mathbb{E}\left[\tilde{n}_k(f)\, \tilde{n}^*_{k'}(f')\right] \;\simeq\; \tfrac{1}{2}\, S_n(\tau_k, f)\, \delta_{kk'}\, \delta(f - f').$$

- T-F Transform

$$\tilde{d}(\tau_k, f) = \int w_k(t - \tau_k)\, d(t)\, e^{-2\pi i f t}\, \mathrm{d}t, \quad \tilde{h}(\tau_k, f; \theta) = \int w_k(t - \tau_k)\, h(t; \theta)\, e^{-2\pi i f t}\, \mathrm{d}t.$$

  where $w(t, \tau_k)$ is a window centred at τ_k and we can use wavelets instead of short Fourier transforms.

- inner product:

$$(d \,|\, h)_{\mathrm{TF}} \;=\; \sum_{k=1}^{K} 4\, \mathrm{Re} \int_0^\infty \frac{\tilde{d}(\tau_k, f)\, \tilde{h}^*(\tau_k, f)}{S_n(\tau_k, f)\, \mathrm{d}f}.$$

- Discrete time & frequency likelihood:

$$-2 \log \mathcal{L}(d|\theta) \;\approx\; \sum_k 4 \sum_{m \geq 0} \frac{\left|\tilde{d}_{km} - \tilde{h}_{km}(\theta)\right|^2}{S_n(\tau_k, f_{km})}\, \Delta f_k \;+\; \text{const}.$$

Glitches, gaps

- Gaps: inpainting – gap filling, time-frequency likelihood
- Glitches
- Need to detect glitches (maximum likelihood?)
- if detected: (i) gating – remove the damaged data → gaps. (ii) model glitches (like GWs) using decomposition in some basis and infer *together* with GW signal (BayesWave)
- Is it GW of instrument/background