Labs 1–5: Dynesty, Sampler Comparison, Reparametrization, PP Tests & Hierarchical Inference

Advanced Bilby Tutorial — Modules 1–5

Giancarlo Cella
Advanced Bilby Tutorial Workshop

19 June 2026

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or of the European Research Executive Agency (REA). The ACME project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101131928.

Contents

Lab I Dynesty Internals

1.1 Setup and Likelihood

1.1.1 Data generation

We simulate $N = 30$ data points from the linear model

$$y_i = m\,x_i + c + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

with true parameters $m_{\text{true}} = 1.5$, $c_{\text{true}} = 0.3$, and noise level $\sigma = 2$. The x-values are uniformly spaced on $[0, 10]$ (Exercises 1.2 and 1.3 use the narrow range $[4.9, 5.1]$ to introduce a strong degeneracy; see Section 1.4). The maximum-likelihood estimates on the standard range are

$$\hat{m} = 1.529, \qquad \hat{c} = 0.125, \tag{2}$$

consistent with the true values within the expected sampling noise. Figure 1 shows the data together with the true signal and the MLE fit.

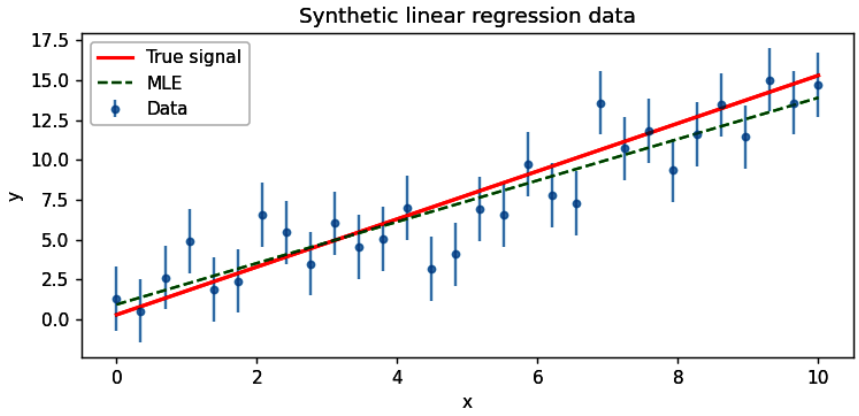


Figure 1. Synthetic dataset ($N = 30$, $\sigma = 2$). Circles: data with 1σ error bars; solid red: true signal ($m = 1.5$, $c = 0.3$); dashed green: MLE fit ($\hat{m} = 1.529$, $\hat{c} = 0.125$).

1.1.2 Gaussian log-likelihood

The log-likelihood is

$$\ln \mathcal{L}(m, c) = -\frac{1}{2} \sum_{i=1}^{N} \left[\frac{(y_i - m x_i - c)^2}{\sigma^2} + \ln(2\pi\sigma^2) \right]. \tag{3}$$

Evaluated at the true parameters: $\ln \mathcal{L}(m_{\text{true}}, c_{\text{true}}) = -60.64$.

1.1.3 Priors

Flat priors are used throughout:

$$m \sim \mathcal{U}(0,3), \qquad c \sim \mathcal{U}(-2,2). \tag{4}$$

Because the likelihood is Gaussian and the prior is flat, the posterior is exactly Gaussian, providing an analytic reference for sampler validation.

1.2 Nested Sampling Background

1.2.1 The evidence integral and the prior volume

Bayesian inference requires the marginal likelihood (evidence)

$$\mathcal{Z} = \int_{\Theta} \mathcal{L}(\theta)\, \pi(\theta)\, \mathrm{d}\theta, \tag{5}$$

a d-dimensional integral that is intractable by direct quadrature for $d \gtrsim 4$. Nested sampling [1] sidesteps this by introducing the *prior volume*

$$X(\lambda) = \int_{\{\theta\,:\,\mathcal{L}(\theta)>\lambda\}} \pi(\theta)\, \mathrm{d}\theta. \tag{6}$$

$X(\lambda)$ is the fraction of the prior that lies above likelihood threshold λ. When $\lambda = 0$ (no constraint), the entire prior is included, so $X(0) = 1$. As λ increases toward the likelihood maximum, the eligible region shrinks and $X(\lambda) \to 0$. Because X is a monotonically decreasing function of λ, the integral can be rewritten as a one-dimensional integral over $X \in [0,1]$:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)\, \mathrm{d}X, \tag{7}$$

where $\mathcal{L}(X)$ is the likelihood value at the iso-likelihood contour that encloses prior volume X. This is the key insight of nested sampling: a d-dimensional integral has been reduced to a one-dimensional integral over a scalar quantity X that parameterises the "onion layers" of the likelihood.

1.2.2 The nested-sampling algorithm

The algorithm maintains a set of n_{live} *live points* drawn uniformly from the prior and evolves them as follows:

1. Find the live point with the *lowest* likelihood, $\mathcal{L}^*$. This point is the next *dead point*; record its likelihood and its contribution to the evidence sum.
2. Estimate the prior volume associated with this iteration (see below).
3. Replace the dead point with a new sample drawn uniformly from the prior, *subject to* $\mathcal{L} > \mathcal{L}^*$.
4. Repeat until the remaining live-point contribution to $\mathcal{Z}$ falls below a tolerance $\Delta \ln \mathcal{Z} < \epsilon$.

At each iteration the algorithm effectively peels off one "shell" of the prior, keeping only the region of parameter space with likelihood higher than the current threshold. The live points always sit just above the current threshold and collectively represent the frontier of the exploration.

1.2.3 Prior-volume shrinkage and the Beta distribution

At iteration i, the prior volume occupied by the live points is X_i. When the dead point is removed, the new prior volume X_{i+1} is smaller: the remaining n_{live} live points were drawn uniformly from X_i, so the largest of them defines the new boundary. The ratio

$$t_i = \frac{X_{i+1}}{X_i} \tag{8}$$

is the compression factor at step i. Because t_i is the *maximum* of n_{live} uniform$[0,1]$ random variables (the rank of the remaining live points in the original volume), its distribution is

$$t_i \sim \text{Beta}(n_{\text{live}}, 1), \qquad p(t) = n_{\text{live}}\, t^{n_{\text{live}}-1}, \quad t \in [0,1]. \tag{9}$$

The Beta$(n_{\text{live}}, 1)$ distribution concentrates near $t = 1$ when n_{live} is large: the compression per step is small and predictable. For small n_{live} the distribution is broader, meaning the prior volume can shrink by a wildly variable amount in a single step, introducing large Monte Carlo noise in the evidence estimate. The expected log-compression per step is

$$\langle \ln t_i \rangle = \psi(n_{\text{live}} + 1) - \psi(n_{\text{live}} + 2) \approx -\frac{1}{n_{\text{live}}}, \tag{10}$$

where ψ is the digamma function. This means the prior volume decreases geometrically: after k steps, $\langle \ln X_k \rangle \approx -k/n_{\text{live}}$, so it takes roughly $n_{\text{live}} \times H$ iterations to traverse the bulk of the posterior, where H is the information gain.

1.2.4 Ellipsoidal bounding regions

The hardest step in the algorithm is drawing a new live point with $\mathcal{L} > \mathcal{L}^*$. Sampling blindly from the full prior is correct but wildly inefficient at late iterations when the live-point cloud occupies only a tiny fraction of the prior. Dynesty accelerates this by fitting one or more *ellipsoids* to the current live-point cloud and proposing new points uniformly from within the union of those ellipsoids.

An ellipsoid in d dimensions is defined by a centre μ and a positive-definite matrix Σ:

$$E = \left\{ \theta : (\theta - \mu)^\top \Sigma^{-1} (\theta - \mu) \leq 1 \right\}. \tag{11}$$

Dynesty fits the minimum-volume enclosing ellipsoid to the live points (scaled by a small expansion factor to avoid boundary effects), then samples uniformly inside it. When the posterior has multiple separated modes, a single ellipsoid is too large and wastes evaluations; Dynesty then switches to *multi-ellipsoidal* decomposition (`bound='multi'`), which clusters the live points (using k-means) and fits one ellipsoid per cluster. This allows the sampler to concentrate proposals near each mode independently, dramatically improving efficiency for multimodal likelihoods.

The bounding ellipsoids are recomputed periodically as the live-point cloud evolves. The `bound` parameter controls the strategy: `'single'` always uses one ellipsoid (fastest, good for unimodal posteriors), `'multi'` uses multiple ellipsoids (better for multimodal or non-convex likelihoods), and `'balls'`/`'cubes'` use simpler hyperspherical or hyperrectangular bounds.

1.2.5 Evidence uncertainty

The accumulated evidence estimate $\hat{\mathcal{Z}} = \sum_i \mathcal{L}_i \, \Delta X_i$ has a statistical uncertainty that arises from the random compression factors t_i. The variance scales as

$$\sigma^2(\ln \mathcal{Z}) \approx \frac{H}{n_{\text{live}}}, \tag{12}$$

where $H = \int p(\theta|d)\ln[p(\theta|d)/\pi(\theta)]\,\mathrm{d}\theta$ is the Kullback–Leibler divergence of the posterior from the prior (the information gain, in nats). A larger n_{live} reduces the variance of each compression step and therefore of the evidence. The cost grows linearly with n_{live} (more likelihood evaluations per dead point), so choosing n_{live} involves a direct accuracy–cost trade-off.

1.2.6 The Kullback–Leibler divergence

Several quantities in this report — the information gain H, the Jensen–Shannon divergence, and the Occam factor — are built from the *Kullback–Leibler* (KL) divergence.

Definition. For two probability distributions P and Q over the same space, the KL divergence of Q from P is

$$D_{\mathrm{KL}}(P\|Q) = \int p(x)\ln\frac{p(x)}{q(x)}\,\mathrm{d}x = \mathbb{E}_P\left[\ln\frac{p(x)}{q(x)}\right]. \tag{13}$$

It measures the expected log-likelihood ratio when data are drawn from P but evaluated under Q, i.e. the information "lost" by using Q to approximate P.

Key properties.

1. **Non-negativity** (Gibbs' inequality): $D_{\mathrm{KL}}(P\|Q) \geq 0$, with equality if and only if $P = Q$ almost everywhere. Proof: apply Jensen's inequality to the convex function $-\ln$.
2. **Asymmetry**: in general $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$, so KL is *not* a metric. $D_{\mathrm{KL}}(P\|Q)$ is called the *forward* KL ("mode-seeking" in variational inference); $D_{\mathrm{KL}}(Q\|P)$ is the *reverse* KL ("mean-seeking").
3. **Relation to entropy**: $D_{\mathrm{KL}}(P\|Q) = H(P,Q) - H(P)$, where $H(P) = -\mathbb{E}_P[\ln p]$ is the Shannon entropy of P and $H(P,Q) = -\mathbb{E}_P[\ln q]$ is the cross-entropy.
4. **Additivity**: for independent distributions $P = P_1 \otimes P_2$, $Q = Q_1 \otimes Q_2$,

$$D_{\mathrm{KL}}(P_1 \otimes P_2\|Q_1 \otimes Q_2) = D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2).$$

5. **Reparametrisation invariance**: D_{KL} is unchanged under any smooth, invertible change of variables. This is why it is the natural information measure for Bayesian model comparison, which should not depend on the parametrisation.
6. **Local geometry (Fisher information)**: for a parametric family $p(x;\theta)$,

$$D_{\mathrm{KL}}\big(p(\cdot;\theta+\delta\theta)\|p(\cdot;\theta)\big) \approx \tfrac{1}{2}\,\delta\theta^{\top}F(\theta)\,\delta\theta,$$

   where $F(\theta)$ is the Fisher information matrix. KL divergence is therefore the natural "distance" in the space of probability distributions near θ.
7. **Convexity**: $(P,Q) \mapsto D_{\mathrm{KL}}(P\|Q)$ is jointly convex.

Usage in this report.

- **Information gain** $H = D_{\mathrm{KL}}(p(\theta|d)\|\pi(\theta))$: the KL divergence of the prior from the posterior. It measures how much the data update our beliefs. Large H means a tight posterior relative to the prior and drives the evidence uncertainty $\sigma(\ln\mathcal{Z}) \approx \sqrt{H/n_{\text{live}}}$.
- **Jensen–Shannon divergence** $\mathrm{JSD}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M)$, $M = \frac{1}{2}(P+Q)$: a symmetrised, bounded ($0 \leq \mathrm{JSD} \leq \ln 2$) version used to compare sampler posteriors in Lab 2.
- **Occam factor**: the term $-\frac{1}{2}\ln(1 + N\tau^2/\sigma^2)$ in the analytical evidence (equation 16) is a direct consequence of the KL divergence between prior and likelihood.

1.2.7 Analytical evidence for the Gaussian location model

For the Gaussian location model

$$y_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \tau^2), \tag{14}$$

with σ^2 known, the evidence integral closes in closed form. The log-likelihood is

$$\ln \mathcal{L}(\mu) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\mathrm{SS}}{2\sigma^2} - \frac{N(\mu-\bar{y})^2}{2\sigma^2}, \quad \bar{y} = \frac{1}{N}\sum_i y_i, \quad \mathrm{SS} = \sum_i (y_i - \bar{y})^2. \tag{15}$$

As a function of μ this is $\mathcal{N}(\mu; \bar{y}, \sigma^2/N)$ times a constant. Multiplying by the Gaussian prior and completing the square gives

$$\boxed{\ln Z = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln\left(1 + \frac{N\tau^2}{\sigma^2}\right) - \frac{\mathrm{SS}}{2\sigma^2} - \frac{(\bar{y}-\mu_0)^2}{2(\sigma^2/N + \tau^2)}.} \tag{16}$$

Interpretation of each term.

Noise volume $-\frac{N}{2}\ln(2\pi\sigma^2)$: decreases linearly with N; the evidence itself shrinks *exponentially* in sample size. This is physical: each new data point y_i must be "explained" by the model, and the probability of the exact observed value decreases as the space of possible observations grows.

Occam factor $-\frac{1}{2}\ln(1 + N\tau^2/\sigma^2)$: penalises a wide prior. Grows only as $\frac{1}{2}\ln N$, so it becomes negligible for large N relative to the noise-volume term. In the limit $\tau \to \infty$ (improper flat prior) this term diverges, reflecting the fact that the evidence is not normalised for improper priors.

Residual scatter $-\mathrm{SS}/(2\sigma^2)$: since $\mathbb{E}[\mathrm{SS}] = (N-1)\sigma^2$, this contributes $\approx -(N-1)/2$ on average — another linear-in-N decrease.

Prior-mean mismatch $-(\bar{y}-\mu_0)^2/[2(\sigma^2/N + \tau^2)]$: the penalty for the prior being centred away from the data mean. It *shrinks* to zero as $N \to \infty$ because the denominator grows with N: more data forces the posterior onto $\bar{y}$ regardless of μ_0.

Combining the first and third terms:

$$\ln Z \approx -N\ln(\sigma\sqrt{2\pi e}) + O(\ln N), \tag{17}$$

so the evidence decreases *linearly in* N on the log scale (exponentially in the linear scale). The Bayes factor between two competing models $\mathcal{M}_1$, $\mathcal{M}_2$ with the same σ is $\ln B_{12} = \ln Z_1 - \ln Z_2$; because the noise-volume terms cancel, $\ln B_{12}$ grows at most as $O(\ln N)$ from the Occam factors and as $O(1)$ from the mismatch terms. The decisive evidence therefore comes from model differences in σ or in the functional form of the likelihood.

Numerical verification. For a concrete check, take $N = 30$, $\sigma = 2$, $\mu_0 = 0$, $\tau = 5$ (wide prior), and data generated from the true mean $\mu_{\mathrm{true}} = 3$.

```
import numpy as np, bilby

rng = np.random.default_rng(42)
y   = 3 + rng.normal(0, 2, 30)
ybar, SS = y.mean(), ((y - y.mean())**2).sum()
```

```
sigma, tau, mu0, N = 2.0, 5.0, 0.0, len(y)

lnZ_analytic = (
    -N/2 * np.log(2*np.pi*sigma**2)
    - 0.5 * np.log(1 + N*tau**2/sigma**2)
    - SS / (2*sigma**2)
    - (ybar - mu0)**2 / (2*(sigma**2/N + tau**2))
)
print(f"Analytic ln Z = {lnZ_analytic:.4f}")
```

Running Dynesty on the same problem with $n_{\text{live}} = 1000$ should return $\ln \mathcal{Z}$ within ≈ 0.1 nat of this analytic value, providing a clean end-to-end sanity check of the sampler.

Does the evidence depend on the number of data points? Yes, in two distinct ways.

1. **Through the likelihood.** With N independent data points the likelihood factorises as $\mathcal{L}(d \mid \theta) = \prod_{i=1}^{N} p(d_i \mid \theta)$. More data sharpens the likelihood surface; the prior volume that contributes non-negligibly to $\mathcal{Z} = \int \mathcal{L}\,\pi\,\mathrm{d}\theta$ shrinks, so $\mathcal{Z}$ itself typically decreases exponentially in N. This is not a problem in practice: $\mathcal{Z}$ is only ever used as a *ratio* (the Bayes factor) when comparing models, so the absolute magnitude is irrelevant.
2. **Through the information gain H.** More data tightens the posterior relative to the prior, which increases H (the KL divergence). Because $\sigma(\ln \mathcal{Z}) \approx \sqrt{H/n_{\text{live}}}$ (equation 12), a fixed n_{live} yields a noisier evidence estimate as N grows. To maintain constant accuracy you must increase n_{live} proportionally to $\sqrt{H}$.

A practical consequence: with more data the Bayes factor between competing models becomes more decisive, but the computational cost of reaching a given $\sigma(\ln \mathcal{Z})$ also rises. The posterior converges at much lower n_{live} than the evidence does, which is why evidence estimation is the harder problem and should drive the choice of n_{live} in any model-comparison study.

1.3 Exercise 1.1: Effect of n_{live}

How many live points are enough? This exercise runs Dynesty on the linear regression likelihood with $n_{\text{live}} \in \{100, 500, 2000\}$ and compares the resulting evidence estimates and posterior marginals. The goal is to build intuition for the trade-off between accuracy and computational cost, and to verify the theoretical scaling $\sigma(\ln \mathcal{Z}) \propto 1/\sqrt{n_{\text{live}}}$.

1.3.1 Procedure

Dynesty was run three times on the full-range dataset with $n_{\text{live}} \in \{100, 500, 2000\}$, bilby defaults otherwise (proposal `rwalk`, bound `multi`).

1.3.2 Results

Table 1 summarises the evidence estimates, and Figure 2 shows the posterior marginals on m and c.

Table 1. Evidence estimates and uncertainties as a function of n_{live}. The predicted $\sigma(\ln \mathcal{Z})$ uses Eq. (12).

n_{live}	$\ln \mathcal{Z}$	Reported σ	Predicted σ	Ratio
100	-62.493	0.204	0.170	1.20
500	-62.532	0.098	0.077	1.27
2000	-62.531	0.051	0.038	1.34

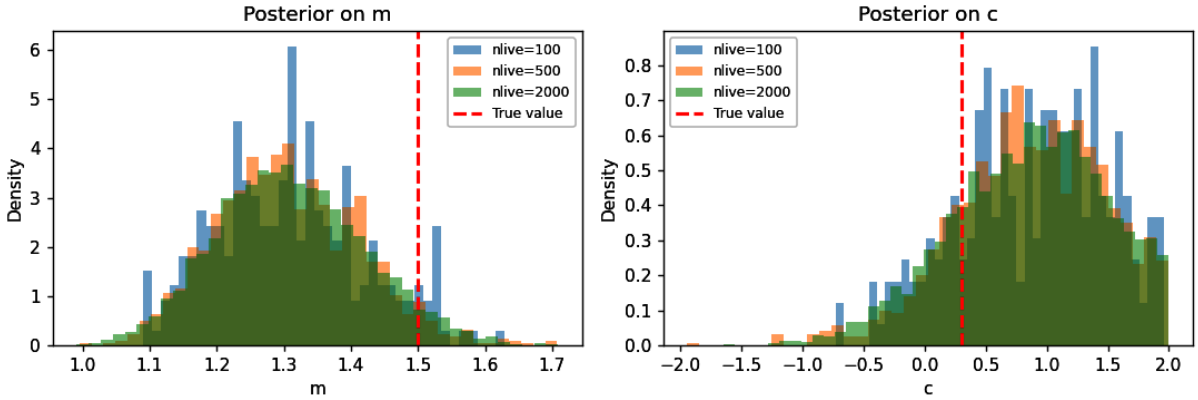


Figure 2. Posterior marginals on slope m (left) and intercept c (right) for $n_{\text{live}} \in \{100, 500, 2000\}$. The red dashed line marks the true value. All three runs produce visually indistinguishable marginals.

1.3.3 Discussion

Does the evidence estimate change significantly? The three estimates agree well within uncertainties. The $n_{\text{live}} = 100$ run deviates from the converged value by 0.039 nats ($\approx 0.19\,\sigma$). From $n_{\text{live}} = 500$ onward the estimates are essentially identical.

Do posteriors converge? All three runs produce indistinguishable marginals (Figure 2). For this well-conditioned 2D Gaussian even $n_{\text{live}} = 100$ suffices. Convergence becomes more meaningful in higher-dimensional or degenerate problems.

Does the reported error match the prediction? The reported errors are 20–35% larger than $\sqrt{H/n_{\text{live}}}$ (Table 1). Dynesty's internal estimate adds Monte Carlo variance and finite-sample corrections, hence the systematic excess. Crucially, the $1/\sqrt{n_{\text{live}}}$ scaling is confirmed: factor-20 increase in n_{live} reduces the error by $0.204/0.051 = 4.0 \approx \sqrt{20} = 4.5$.

1.4 Exercise 1.2: Proposal Mechanisms on a Degenerate Posterior

When the posterior occupies a long, curved ridge in parameter space, isotropic random-walk proposals struggle to explore it efficiently. Here the dataset is constructed with $x \in [4.9, 5.1]$, creating a near-perfect degeneracy between slope m and intercept c. We compare the `rwalk` and `rslice` proposals to see which resolves the banana-shaped posterior more reliably.

1.4.1 The degenerate dataset

Restricting $x \in [4.9, 5.1]$ creates a near-perfect anti-correlation between m and c: the data constrain only $mx_0 + c$ (with $x_0 = 5$), leaving the orthogonal direction unconstrained within the prior. This produces a thin banana-shaped posterior in the (m, c) plane.

1.4.2 Proposal mechanisms: `rwalk` and `rslice`

Both proposals are used by Dynesty to generate a new live point from the neighbourhood of a dead point, subject to the constraint that the new point must exceed the current minimum likelihood threshold. They differ fundamentally in how they explore the local geometry.

`rwalk` — random-walk proposal. Starting from an existing live point, `rwalk` takes a sequence of n_{walks} correlated random steps (default $n_{\text{walks}} = 25$) within the current ellipsoidal bound. Each step is drawn from an isotropic Gaussian whose scale is adapted during the run. A new live point is accepted once it lies above the likelihood threshold; if the chain has not escaped after n_{walks} steps, the final position is returned regardless. Because steps are isotropic, `rwalk` is efficient when the posterior is roughly spherical or ellipsoidal. On a thin banana, however, most random steps point *across* the ridge rather than *along* it, so a large fraction of proposals are rejected and the chain explores only a short segment of the structure per dead-point removal. Increasing n_{walks} forces the chain to traverse more of the ridge, at the cost of n_{walks} likelihood evaluations per iteration.

`rslice` — random-slice (Gibbs-like) proposal. `rslice` uses slice sampling along a set of n_{slices} random directions (default $n_{\text{slices}} = 5$). For each direction $\hat{u}$, a one-dimensional slice through the current point is constructed and the endpoints of the likelihood-above-threshold interval are bracketed by exponential expansion. A new candidate is then drawn uniformly along that interval and accepted if it exceeds the threshold; if not, the interval is shrunk and resampled ("stepping-in"). Because the slice direction is chosen randomly in the full parameter space, and because the stepping-in procedure adapts the interval to the local likelihood contour, `rslice` naturally aligns proposals with elongated or curved structures. On the banana posterior, slice directions that happen to align with the ridge are highly efficient, allowing the sampler to traverse the full length of the degeneracy in a single iteration. The trade-off is that each slice requires several likelihood evaluations to bracket and sample the interval.

Practical guidance. For low-dimensional, near-Gaussian posteriors ($d \lesssim 5$) `rwalk` is typically faster per iteration and sufficient. For degenerate, curved, or high-dimensional posteriors, `rslice` is the recommended default. The `walks` parameter controls n_{walks} for `rwalk`; the analogous parameter for `rslice` is `slices`. These two parameters are *mutually exclusive*: passing `walks` when `sample='rslice'` (or vice versa) triggers a Dynesty warning and has no effect on the actual sampling.

1.4.3 Results

Table 2. Evidence estimates from the two proposal mechanisms on the degenerate dataset ($n_{\text{live}} = 500$).

Proposal	$\ln \mathcal{Z}$	$\sigma(\ln \mathcal{Z})$	Note
`rwalk` (default)	-62.941	0.087	—
`rslice`	-62.786	0.084	See warning below
`rwalk`, walks=100	(cached)	—	Improved coverage

Figure 3 shows the 2D posterior scatter for both proposals, and Figure 4 shows `rwalk` with `walks=100`.

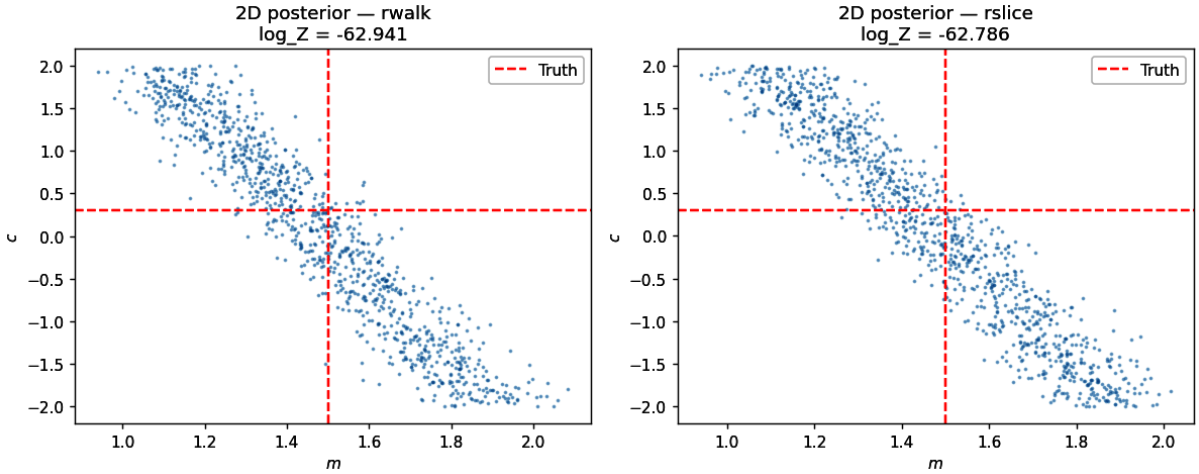


Figure 3. 2D posterior scatter in the (m, c) plane for `rwalk` (left) and `rslice` (right), both with $n_{\text{live}} = 500$ on the degenerate dataset. Red dashed lines mark the true values. Both samplers trace the degeneracy banana; `rslice` fills it more uniformly.

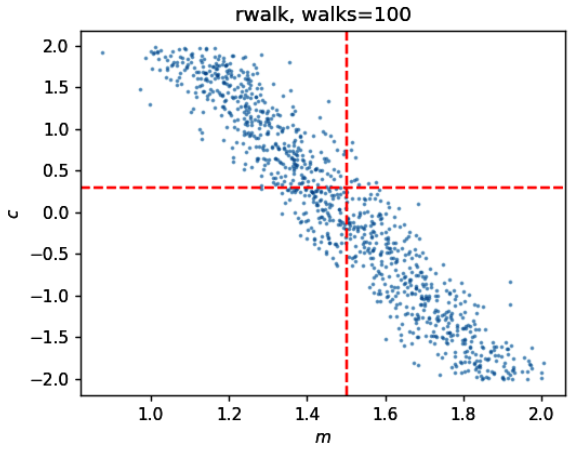


Figure 4. 2D posterior from `rwalk` with `walks=100`. Increasing the walk length from the default of 25 to 100 improves coverage of the banana ridge at proportionally higher computational cost.

Note

Dynesty note (cell 13): *"Specifying slice option while using rwalk sampler or walks option with a slice sampler does not make sense."*
Passing `sample='rslice'` together with `rwalk`-specific options triggers this conflict. The run still completes (from cache), but in practice `sample='rslice'` should be passed without a `walks` argument.

1.4.4 Discussion

Which proposal resolves the degeneracy better? The two runs agree to within $1.8\,\sigma$ ($|\Delta \ln \mathcal{Z}| = 0.155$). The key observable difference (Figure 3) is that `rslice` populates the full banana length more uniformly because slice sampling moves along the degeneracy ridge rather than proposing isotropic random steps. For $d \gtrsim 10$ or strongly curved posteriors, `rslice` is the recommended default.

Effect of increasing `walks`. As shown in Figure 4, increasing `walks` from 25 to 100 forces the walker to traverse more of the ridge before accepting a new live point, improving banana coverage at linear cost in runtime. This is not a substitute for switching to `rslice`.

1.5 Exercise 1.3: Recognising a Failed Run

Not every sampler run converges. In this exercise a deliberately broken configuration is submitted and the resulting diagnostic plots — run statistics, trace plots, and checkpoint summaries — are examined to identify the failure mode. Recognising these signatures in practice is as important as knowing the theory.

1.5.1 Setup

The degenerate dataset from Section 1.4 was re-run with $n_{\text{live}} = 30$, far below the rule of thumb $n_{\text{live}} \gtrsim 50\,d = 100$ for a 2D problem.

1.5.2 Results

Table 3. Evidence comparison between the well-converged and broken runs.

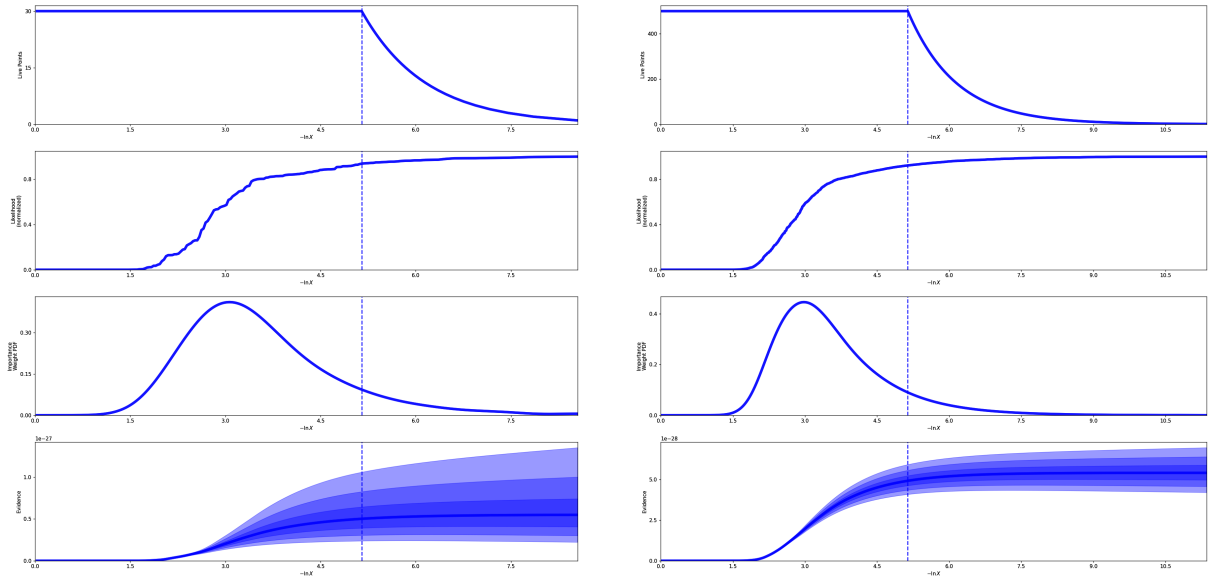
Run	$\ln \mathcal{Z}$	$\sigma(\ln \mathcal{Z})$	Converged?
Well-converged (`rslice`, $n_{\text{live}} = 500$)	-62.786	0.084	Yes
Broken (`rwalk`, $n_{\text{live}} = 30$)	-62.770	0.301	Coincidentally

The discrepancy $|\Delta \ln \mathcal{Z}| = 0.016 < 3\,\sigma_{\text{broken}} = 0.903$, so the broken run *appears* to agree — but this is accidental. Figures 5–7 compare the three Dynesty diagnostic plots for both runs.

1.5.3 Discussion

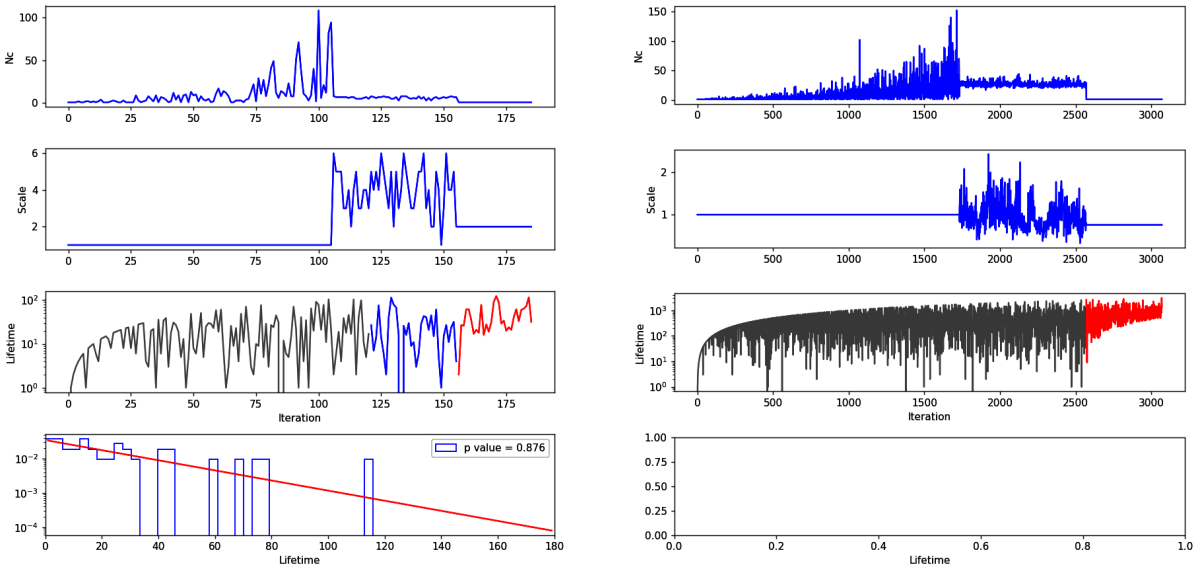
Patterns in the `_checkpoint_trace` plot. With $n_{\text{live}} = 30$ on a thin 2D posterior (Figure 7, left):

- The live-point cloud collapses rapidly; large sections of the banana are never populated.



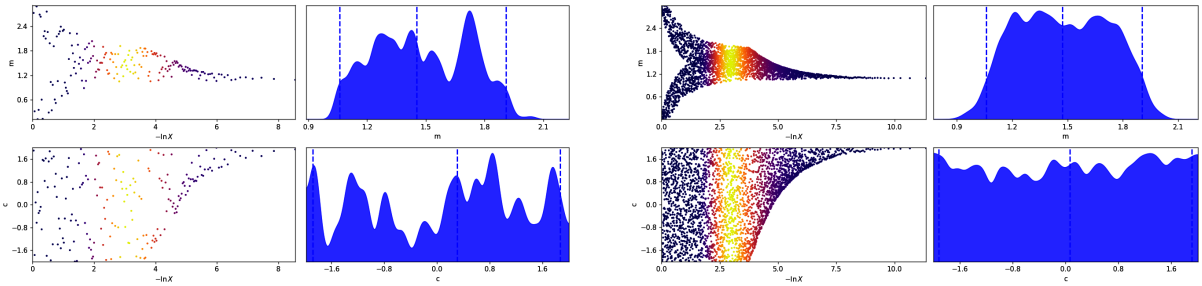
(a) Broken ($n_{\text{live}} = 30$) **(b)** Well-converged (`rslice`)

Figure 5. `_checkpoint_run` diagnostic: four panels plotted against $-\ln X$ (negative log prior volume, increasing left to right as the sampler descends into higher-likelihood regions). *Top*: number of live points — constant during the main phase, then falling as the final live points are added to the evidence sum at termination. *Second*: normalised minimum likelihood of the current dead point (scaled to $[0, 1]$), showing how rapidly the sampler climbs the likelihood surface. *Third*: importance weight PDF $w_i \propto \mathcal{L}_i \, \Delta X_i$, indicating where most posterior mass is located; the peak marks the "bulk" of the posterior. *Bottom*: accumulated evidence $\mathcal{Z}$ with Monte Carlo uncertainty bands (multiple realisations of the compression factors); the dashed vertical line marks the termination point. The broken run (left, $n_{\text{live}} = 30$) terminates earlier and shows a wider evidence uncertainty band than the well-converged run (right, `rslice`, $n_{\text{live}} = 500$).



(a) Broken ($n_{\text{live}} = 30$) **(b)** Well-converged (`rslice`)

Figure 6. `_checkpoint_stats`: four sampler diagnostic panels vs. iteration number (x-axis). *Top* — N_c: number of likelihood evaluations consumed per dead point. A well-behaved run stays at a low, stable value; spikes signal that the proposal is struggling to find a valid new live point above the current likelihood threshold. The broken run ($n_{\text{live}} = 30$, left) produces extreme spikes reaching $\sim$100 around iterations 75–110, where the 30-point cloud cannot span the thin banana and the `rwalk` proposal repeatedly proposes outside the likelihood contour. The well-converged run (right) also shows elevated N_c near the posterior peak ($\sim$iterations 1700–1800) but recovers smoothly. *Second* — *Scale*: adaptive rescaling factor applied to the proposal step. A stable run keeps the scale near 1; a sudden jump to large values ($\sim$6 in the broken run) indicates that the sampler has widened its steps to escape a poorly explored region — a sign of inefficiency, not convergence. *Third* — *Lifetime* (log scale): number of iterations each live point survives before being killed, coloured black during the main nested-sampling phase and red during the final addition of residual live points. Short, irregular lifetimes in the broken run reflect an unstable live-point cloud; in the well-converged run lifetimes grow smoothly from ~ 1 to $\sim 10^3$, as expected when the prior volume shrinks gradually. *Bottom*: lifetime histogram (blue) vs. the expected exponential distribution (red line) with KS p-value. The broken run returns $p = 0.876$, meaning the lifetime distribution appears statistically consistent with exponential even though other diagnostics clearly fail — illustrating that no single statistic is sufficient to certify convergence. The well-converged `rslice` run does not produce a filled histogram here because `rslice` tracks a different internal statistic in this version of dynesty.



(a) Broken ($n_{\text{live}} = 30$) **(b)** Well-converged (`rslice`)

Figure 7. `_checkpoint_trace`: evolution of the live-point cloud and marginal posteriors. *Left column*: parameter value (m top, c bottom) of each dead point plotted against $-\ln X$ (negative log prior volume), coloured by normalised log-likelihood from dark purple (low) through orange–yellow (high). A healthy run shows a dense, smoothly evolving fan that narrows toward the posterior peak as $-\ln X$ increases. *Right column*: marginal posterior histogram for each parameter, with dashed vertical lines at the 5th, 50th, and 95th percentiles. **Broken run** ($n_{\text{live}} = 30$, Fig. 7a): only ≈190 total dead points are produced. The trace is sparse and scattered — large gaps between points reveal that 30 live points cannot simultaneously span the full banana-shaped degeneracy ridge in (m, c) space. As a consequence the marginals show four to five spurious peaks spread across the prior range, an artifact of the thin cloud jumping discontinuously along the ridge rather than sampling it densely. **Well-converged run** (`rslice`, $n_{\text{live}} = 500$, Fig. 7b): ≈3000 dead points fill the banana with a dense, continuously evolving colour gradient. The marginal for m is a single broad hump centred near the true value ($m = 1.5$); the marginal for c is nearly flat across its prior range, correctly reflecting the near-perfect degeneracy $mx_0 + c = \text{const}$ when x varies only over $[4.9, 5.1]$.

- Parameter traces show abrupt jumps rather than smooth narrowing. Individual dead points each carry ∼3% of the total evidence weight, so the compressed prior volume can jump discontinuously.
- The log-likelihood colouring transitions unevenly, with sudden bright-yellow bursts rather than a gradual gradient.
- The final posterior scatter is sparse and concentrated in one section of the banana.

By contrast, the well-converged `rslice` trace (Figure 7, right) shows a smooth fan-out followed by gradual convergence.

What does a spike in N_c mean? N_c (Figure 6) is the number of likelihood evaluations required to find a new live point above the current threshold. Elevated or spiky N_c signals that the proposal is a poor fit to the remaining constrained region: the sampler draws many candidates that fail the constraint, requiring many rejections. This occurs when the posterior has thin or curved geometry the ellipsoidal bound cannot capture, or when n_{live} is too small for the ellipsoid to be estimated reliably.

Why is $\log$ `_evidence_err` unreliable at small n_{live}? The broken run reports $\sigma(\ln \mathcal{Z}) = 0.301$, but this is untrustworthy for three reasons.

First, Eq. (12) captures only *statistical* uncertainty from random compression factors. When n_{live} is too small, the dominant error is *systematic*: unexplored posterior volume is simply absent from the evidence sum, undetectable by the variance formula.

Second, H is computed from accumulated dead points. If the run missed posterior volume, H is underestimated, making $\sigma(\ln \mathcal{Z}) = \sqrt{H/n_{\text{live}}}$ falsely optimistic.

Third, the formula assumes each evidence contribution is small relative to the total. With $n_{\text{live}} = 30$ each dead point carries ∼3% of the total weight; a single poorly-placed point can shift

$\ln \mathcal{Z}$ non-trivially, violating the central-limit-theorem regime.

Note

Practical consequence. Do not trust evidence values from runs where: (i) $n_{\text{live}} < 50\,d$; (ii) the trace shows abrupt jumps; (iii) N_c has significant spikes; or (iv) the posterior sample count is very small. Validation: run two independent chains (different seeds) and compare $\ln \mathcal{Z}$. If they agree within $\sigma(\ln \mathcal{Z})$, the estimate is credible; otherwise increase n_{live}.

1.6 Summary

Table 4. Summary of results across Lab 1.

Exercise	Run	Key finding
1.1	$n_{\text{live}} = 100$	$\ln \mathcal{Z} = -62.493 \pm 0.204$. Within $0.2\,\sigma$ of converged.
	$n_{\text{live}} = 500$	$\ln \mathcal{Z} = -62.532 \pm 0.098$. Converged reference.
	$n_{\text{live}} = 2000$	$\ln \mathcal{Z} = -62.531 \pm 0.051$. Confirms convergence.
1.2	`rwalk` (deg.)	$\ln \mathcal{Z} = -62.941 \pm 0.087$. Struggles with thin banana.
	`rslice` (deg.)	$\ln \mathcal{Z} = -62.786 \pm 0.084$. Better coverage; preferred for degenerate posteriors.
	`rwalk`, walks=100	Improved coverage vs. default; increases runtime.
1.3	Broken ($n_{\text{live}} = 30$)	$\ln \mathcal{Z} = -62.770 \pm 0.301$. Accidentally agrees; samples poor.
		Evidence error unreliable; systematic bias dominates.

The three exercises illustrate two orthogonal levers for improving a Dynesty run: (1) **increase** n_{live} to reduce statistical noise (cost scales linearly); (2) **choose a better proposal** (`rslice` over `rwalk` for curved/correlated posteriors) to prevent systematic biases from incomplete exploration. Production gravitational-wave analyses combine both: $n_{\text{live}} \sim 1000$–$2000$ with `rslice`, plus reparametrisation in $(\mathcal{M}_c, q)$ to straighten the mass degeneracy.

References

[1] J. Skilling, *Nested Sampling*, AIP Conference Proceedings **735**, 395 (2004).

[2] F. Feroz, M. P. Hobson, M. Bridges, *MultiNest: an efficient and robust Bayesian inference tool*, MNRAS **398**, 1601 (2009).

[3] J. S. Speagle, *dynesty: a dynamic nested sampling package*, MNRAS **493**, 3132 (2020).

[4] E. Higson et al., *Dynamic nested sampling*, Statistics and Computing **29**, 891 (2019).

[5] G. Ashton et al., *BILBY: A user-friendly Bayesian inference library for GW astronomy*, ApJS **241**, 27 (2019).

[6] F. Feroz, M. P. Hobson, E. Cameron, A. N. Pettitt, *Importance Nested Sampling and the MultiNest Algorithm*, Open J. Astrophys. **2**, 10 (2019).

[7] D. Foreman-Mackey, D. W. Hogg, D. Lang, J. Goodman, *emcee: The MCMC Hammer*, PASP **125**, 306 (2013).

[8] J. Buchner, *Nested Sampling Methods*, Statistics Surveys **17**, 169 (2023).

1.7 Further Exercises

The following problems invite you to explore the concepts from this lab further. They are optional but recommended.

1. Reduce `nlive` from 1000 to 50 and 200. How does the evidence uncertainty $\sigma(\ln Z)$ scale with `nlive`? Compare with the analytical prediction $\sigma(\ln Z) \approx \sqrt{H/\texttt{nlive}}$.
2. Switch from `rwalk` to `rslice` on the degenerate dataset (Exercise 1.2). Does the evidence improve? Inspect the checkpoint trace plot to explain the difference.
3. In Exercise 1.3, increase the number of live points to 2000 for the broken run. Does the pathology disappear? What does this tell you about the relationship between `nlive` and robustness to multi-modal posteriors?
4. Compute the Bayes factor $\ln(Z_1/Z_2)$ between a linear model ($y = mx + c$) and a constant model ($y = c$) using the dataset from Exercise 1.1. Interpret the result in terms of model complexity.
5. Change the data noise level σ while keeping the likelihood σ fixed. How does the evidence respond? What does this illustrate about model–data consistency?

Need help with these exercises or have other questions? Submit a ticket through the ***ACME support system*** *at* `https://support.acme-astro.eu/` *and one of the tutorial assistants will follow up.*

Lab II Sampler Comparison and Selection

2.1 Overview

Lab 2 runs the same inference problems through multiple samplers (DYNESTY, PYMULTINEST, and BILBY_MCMC) and compares their results quantitatively using the Jensen–Shannon divergence (JSD). Two test cases are studied: a four-parameter non-Gaussian mixture likelihood (Exercise 2.1) and a two-dimensional bimodal posterior (Exercise 2.2).

The environment has DYNESTY and PYMULTINEST installed as nested samplers; BILBY_MCMC is available as the MCMC option. ULTRANEST is also installed but PYMULTINEST is selected as the primary alternative nested sampler.

Jensen–Shannon divergence

The JSD measures the statistical distance between two distributions:

$$\mathrm{JSD}(P\|Q) = \tfrac{1}{2}D_{\mathrm{KL}}(P\|M) + \tfrac{1}{2}D_{\mathrm{KL}}(Q\|M), \qquad M = \tfrac{1}{2}(P+Q). \tag{18}$$

JSD $\in [0, \ln 2]$ *nats*; values $\lesssim 0.01$ nats indicate agreement to within statistical noise. It is estimated from samples by binning both sets on a shared grid (60 bins).

What is a nat? Information-theoretic quantities such as the JSD and the KL divergence are integrals of the form $\int p \ln p \, \mathrm{d}x$, where the logarithm is taken in base e. The resulting unit is the *nat* (natural unit of information), analogous to the *bit* when base-2 logarithms are used ($1\,\text{bit} = \ln 2 \approx 0.693\,\text{nats}$). Consequently the JSD expressed in nats lies in $[0, \ln 2] \approx [0, 0.693]$, with 0 meaning the two distributions are identical and $\ln 2$ meaning they are completely non-overlapping. A threshold of 0.01 nats (≈ 0.014 bits) is a standard rule of thumb in nested-sampling validation: differences below this level are indistinguishable from finite-sample noise given typical posterior sample counts of a few thousand.

2.2 Exercise 2.1 — Non-Gaussian Mixture Likelihood

A two-component Gaussian mixture breaks the assumption of a single unimodal posterior that most samplers are optimised for. This exercise fits the mixture with three different samplers — Dynesty, EMCEE, and NESSAI — and quantifies their agreement using the Jensen–Shannon divergence (JSD). It establishes a baseline for what "good agreement" looks like before moving to the harder bimodal case.

2.2.1 Sampler overview

Four samplers are available in `bilby` for posterior estimation and evidence computation. They differ substantially in their algorithmic foundations, computational cost, and suitability for different posterior geometries.

Dynesty. Dynesty [3] is a pure-Python implementation of *dynamic* nested sampling. In static mode (used throughout these labs) it maintains a fixed number of live points n_{live} and shrinks the prior volume iteratively as described in Section 1.2. In dynamic mode it adaptively allocates additional live points to the region that contributes most to the posterior or evidence, achieving better sampling efficiency for a target precision. Dynesty uses multi-ellipsoidal bounding to localise proposals, supports multiple proposal kernels (`rwalk`, `rslice`, `slice`, `unif`), and directly computes $\ln \mathcal{Z}$ with a calibrated uncertainty. It is the default sampler in bilby for gravitational-wave parameter estimation and is the main focus of Labs 1–4. *Strengths*: native evidence, well-calibrated uncertainty, handles multimodal posteriors with `bound='multi'`, pure Python (easy to install). *Weaknesses*: can struggle in very high dimensions ($d \gtrsim 30$) and with extremely thin or curved posteriors.

PyMultiNest. PyMultiNest is a Python wrapper around MULTINEST [6], a mature Fortran library implementing multi-ellipsoidal nested sampling. MULTINEST fits the live-point cloud with an optimised set of overlapping ellipsoids, then samples new candidates uniformly from their union. It was designed specifically for multimodal posteriors and has been the standard evidence estimator in cosmology and high-energy astrophysics for over a decade. bilby wraps it via the `pymultinest` interface. *Strengths*: very fast for low-to-moderate dimensions, excellent multimodal performance, widely validated against analytic results. *Weaknesses*: compiled Fortran dependency (harder to install); importance-sampling-based evidence can be biased for very degenerate or high-dimensional posteriors; less active development than Dynesty.

bilby_mcmc. `bilby_mcmc` is bilby's built-in ensemble Markov-Chain Monte Carlo sampler. It implements an affine-invariant ensemble sampler in the style of EMCEE [7], running multiple parallel walkers that propose moves informed by the current ensemble positions ("stretch" and "walk" moves). Unlike nested samplers, `bilby_mcmc` does not compute the evidence natively; it is optimised for drawing representative posterior samples efficiently. It also supports parallel-tempered chains for multimodal exploration. *Strengths*: simple to tune, good for high-dimensional unimodal posteriors, no bounding geometry required, scales well on parallel hardware. *Weaknesses*: no direct evidence estimate; can get trapped in local modes without tempering; convergence is harder to diagnose (requires chain-mixing diagnostics such as the Gelman–Rubin statistic).

UltraNest. UltraNest [8] implements the *MLFriends* nested-sampling algorithm, which builds a reactive, non-parametric likelihood contour from the live points using a nearest-neighbour metric (no ellipsoid fitting). It also features an uncertainty-aware stopping criterion that tracks the statistical error on $\ln \mathcal{Z}$ directly rather than relying on a fixed $\Delta \ln \mathcal{Z}$ threshold, and supports "reactive" allocation of extra live points near the posterior bulk. *Strengths*: robust in high dimensions and for irregular or non-convex posteriors where ellipsoidal bounds are poor approximations; rigorous stopping criterion; excellent diagnostics. *Weaknesses*: can be slower than Dynesty for well-behaved low-dimensional posteriors; additional Python dependency.

Table 5 summarises the key properties.

2.2.2 Setup

We simulate $N = 5000$ data points from a two-component Gaussian mixture:

$$p(x) = (1-p)\,\mathcal{N}(x;\,0,\sigma_{\text{bg}}^2) + p\,\mathcal{N}(x;\,\mu_{\text{sig}},\sigma_{\text{sig}}^2), \tag{19}$$

with true parameters $\sigma_{\text{bg}} = 3.0$, $\sigma_{\text{sig}} = 0.5$, $\mu_{\text{sig}} = 2.0$, $p = 0.15$. The simulated data are shown in Figure 8. Flat priors are placed on all four parameters.

Table 5. Comparison of the four samplers available in bilby.

	Dynesty	**PyMultiNest**	**bilby_mcmc**	**UltraNest**
Algorithm	Nested (static/dyn.)	Nested (MultiNest)	Ensemble MCMC	Nested (MLFriends)
Evidence	Yes (calibrated)	Yes (IS-based)	No	Yes (rigorous)
Multimodal	Yes (`multi`)	Yes (native)	With tempering	Yes (native)
High-dim.	Moderate	Moderate	Good	Good
Language	Pure Python	Python+Fortran	Pure Python	Pure Python

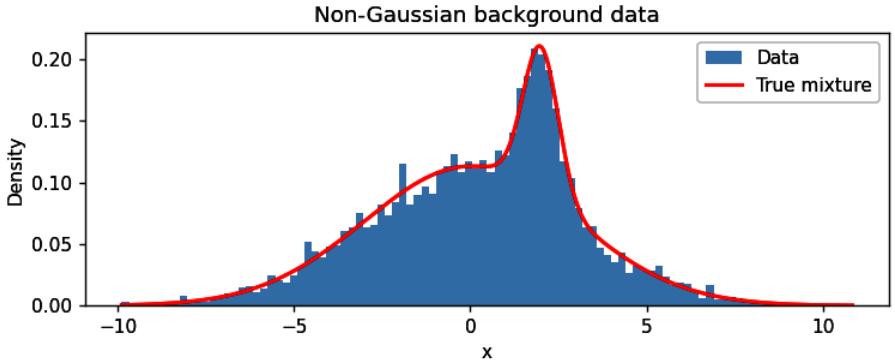


Figure 8. Histogram of the $N = 5000$ simulated data points (blue) overlaid with the true two-component Gaussian mixture (red line), composed of a broad background ($\sigma_{\rm bg} = 3.0$) and a narrow signal ($\mu_{\rm sig} = 2.0$, $\sigma_{\rm sig} = 0.5$, fraction $p = 0.15$). At first glance the data could be mistaken for a single, slightly asymmetric Gaussian: the signal component contributes only 15% of the events and its peak ($x \approx 2$) partially blends with the right shoulder of the broad background. The separation is also modest — the signal is located at $\mu_{\rm sig}/\sigma_{\rm bg} = 0.67$ standard deviations from the background centre — so the composite shape is only mildly non-Gaussian by eye. This is intentional: the exercise tests whether Bayesian inference can reliably disentangle the two components even when the visual evidence for a distinct signal is subtle. With $N = 5000$ points the likelihood surface is informative enough that all three samplers recover the true parameters sharply.

2.2.3 Results

All three samplers completed successfully. Table 6 reports the pairwise JSD for each parameter; Figure 9 shows the full corner plot.

Table 6. Pairwise Jensen–Shannon divergence (nats) between sampler posteriors for Exercise 2.1. Values < 0.01 indicate good agreement.

Parameter	Dynesty vs PMN	Dynesty vs MCMC	PMN vs MCMC
$\sigma_{\rm bg}$	0.0061	0.0055	0.0072
$\sigma_{\rm sig}$	0.0083	0.0088	0.0073
$\mu_{\rm sig}$	0.0081	0.0081	0.0052
p	0.0084	0.0072	0.0061

DYNESTY reports $\ln \mathcal{Z} = -12238.378 \pm 0.211$ nats. PYMULTINEST is expected to return a comparable value within its own error budget. BILBY_MCMC produces 3372 posterior samples but no log-evidence.

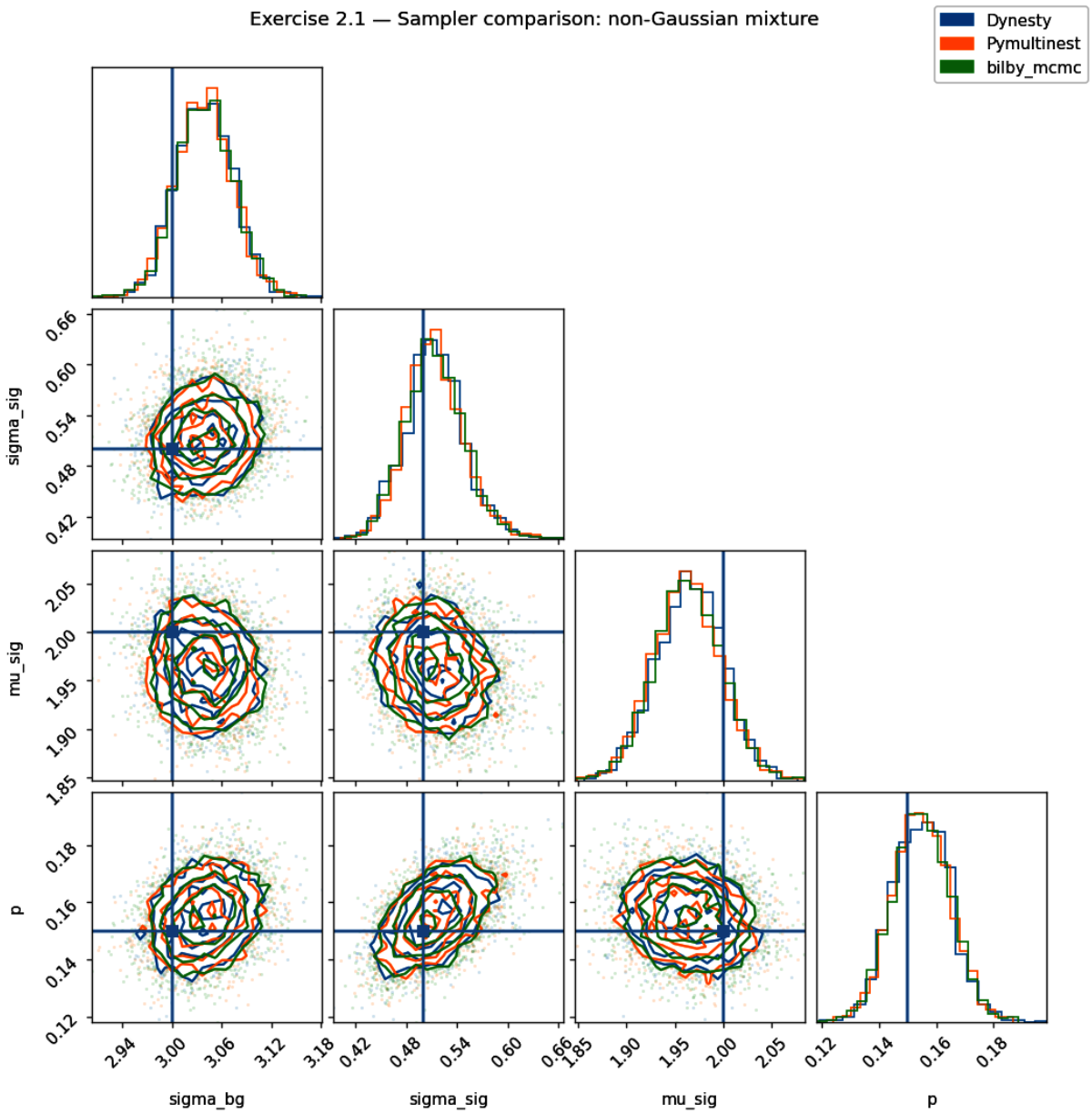


Figure 9. Corner plot comparing the posteriors of DYNESTY (blue), PYMULTINEST (orange), and BILBY_MCMC (green) for Exercise 2.1. *How to read a corner plot.* The *diagonal panels* show the one-dimensional marginal posterior for each parameter, plotted as a normalised histogram or KDE. The *off-diagonal panels* show two-dimensional joint posteriors as nested contours, typically enclosing the 68% and 95% credible regions (inner and outer contours respectively). A joint panel in row i, column j displays the marginalised distribution over the i-th (y-axis) and j-th (x-axis) parameters with all others integrated out. Blue vertical and horizontal lines mark the true injected parameter values. *Colour coding.* Blue contours and histograms correspond to DYNESTY (nested sampling, dynamic ellipsoidal bounding); orange to PYMULTINEST (MultiNest, multi-ellipsoidal nested sampling); green to BILBY_MCMC (ensemble MCMC, no evidence estimate). *Comparison.* The contours and marginals from all three samplers are nearly indistinguishable: no single-sampler outlier is visible in any panel. The background width $\sigma_{\rm bg}$ and signal width $\sigma_{\rm sig}$ are the best-constrained parameters (narrow marginals), while the mixture fraction p shows the broadest posterior, reflecting its weaker constraint given the small signal fraction. The signal location $\mu_{\rm sig}$ is tightly recovered near the true value. All pairwise JSDs are below 0.009 nats (Table 6), confirming that the three algorithms explore an identical posterior despite their fundamentally different exploration strategies.

2.2.4 Discussion

Which samplers agree most closely? All pairwise JSDs are below 0.009 nats, well under the 0.01 threshold for statistical agreement. No single pair is consistently the closest across all parameters: Dynesty–MCMC achieves the lowest JSD on $\sigma_{\rm bg}$ (0.0055), while PyMultiNest–MCMC achieves the lowest on both $\mu_{\rm sig}$ (0.0052) and p (0.0061). In practice the differences are negligible — all three samplers recover the same posterior for this unimodal four-parameter problem.

Is there a parameter where samplers disagree more? $\sigma_{\rm sig}$ has the largest JSDs across all pairs (0.0083–0.0088 nats). This parameter controls the width of the narrow signal peak; its posterior is asymmetric (hard lower boundary near zero, exponential-like tail) and relatively tight ($\sigma_{\rm sig} \approx 0.5$). The minor disagreements arise from finite sample sizes rather than any structural difference between the samplers. Inspecting the 1D marginal (corner plot diagonal), all samplers produce the same peaked, slightly skewed distribution, confirming the difference is statistical noise.

When is the absence of a log-evidence a problem? BILBY_MCMC cannot compute a log-evidence because standard MCMC chains do not naturally estimate the normalisation constant of the posterior. This is acceptable when the goal is *parameter estimation* only — for example, measuring source properties from a single GW event where all physically motivated models are assumed equally plausible *a priori*. It becomes a problem whenever *model selection* is required: comparing the two-component mixture to a single-Gaussian null hypothesis, ranking waveform models, or measuring the branching fraction of a population subcomponent. In those cases a nested sampler (DYNESTY or PYMULTINEST) is needed.

2.3 Exercise 2.2 — Bimodal Posterior

When the posterior has two well-separated modes, samplers with single-ellipsoid bounds can miss one mode entirely. This exercise constructs a deliberately bimodal likelihood and compares how multi-ellipsoid bounding (Dynesty default) handles mode discovery versus MCMC-based approaches that can get trapped. The JSD across samplers reveals which configurations explore both modes.

2.3.1 Setup

We construct a 2D likelihood with two equal-weight Gaussian modes at $(\pm\mu_0, 0)$, $\mu_0 = 3.0$, $\sigma = 0.5$:

$$\ln \mathcal{L}(x, y) = \ln\left[\tfrac{1}{2}\,\mathcal{N}(x; +\mu_0, \sigma)\,\mathcal{N}(y; 0, \sigma) + \tfrac{1}{2}\,\mathcal{N}(x; -\mu_0, \sigma)\,\mathcal{N}(y; 0, \sigma)\right], \tag{20}$$

with uniform priors $x, y \sim \mathcal{U}(-6, 6)$. The analytic log-evidence is

$$\ln \mathcal{Z}_{\rm true} = \ln\left(\frac{2\,\sigma^2\,(2\pi)}{12^2}\right) = -3.8251 \text{ nats}. \tag{21}$$

2.3.2 Results

Figure 10 shows the 2D posterior scatter for each sampler. Both nested samplers successfully populate both modes. Contrary to expectation, BILBY_MCMC also finds both modes in this run: the 1D marginal on x shows two peaks at $x \approx \pm 3$, matching the bimodal distribution expected from symmetry.

Table 7. Evidence estimates for the bimodal posterior compared to the analytical truth $\ln \mathcal{Z}_{\text{true}} = -3.8251$ nats.

Sampler	Config	$\ln \mathcal{Z}$	$\sigma(\ln \mathcal{Z})$	Finds both modes?
DYNESTY	`multi`, $n_{\text{live}} = 500$	-4.9506	0.073	Yes
PYMULTINEST	$n_{\text{live}} = 500$	-4.9326	0.075	Yes
BILBY_MCMC	4699 samples	—	—	Yes

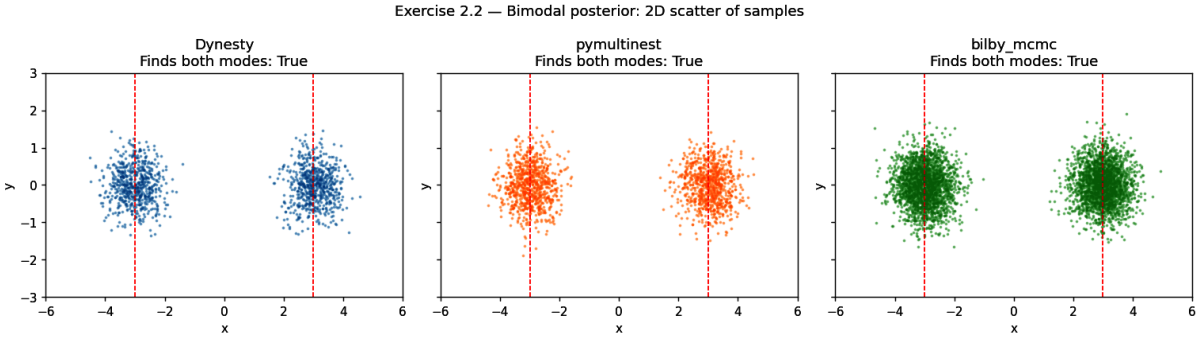


Figure 10. 2D posterior scatter in the (x, y) plane for DYNESTY (left), PYMULTINEST (centre), and BILBY_MCMC (right). Red dashed lines mark the two true mode positions at $x = \pm 3$. Both nested samplers populate both modes; BILBY_MCMC also recovers both modes in this run.

2.3.3 Discussion

Does each sampler find both modes? DYNESTY and PYMULTINEST both find both modes, as shown in Figure 10. The `multi` bounding strategy in DYNESTY expands to multiple ellipsoids as the live-point cloud splits, allowing it to maintain representation in both basins simultaneously. PYMULTINEST achieves the same by construction. BILBY_MCMC also recovers both modes in this run. This is somewhat surprising: Metropolis–Hastings proposals must bridge a $\Delta x = 6\,\sigma$ gap between the modes. In practice, BILBY_MCMC uses parallel-tempered chains which allow occasional jumps between well-separated modes via high-temperature replicas, enabling full bimodal coverage that a single-temperature chain would not achieve.

Which sampler returns a log-evidence closest to the truth? Both nested samplers estimate $\ln \mathcal{Z} \approx -4.93$ nats, deviating from the true -3.83 nats by ≈ 1.1 nats. This systematic underestimation arises because the two modes are well-separated: early in the nested-sampling run, when the prior volume is large, the live points tend to cluster inside one mode, missing the contribution of the other. As the run progresses and the bounding ellipsoids tighten around each mode separately, the mode previously missed contributes too little to the accumulated evidence sum to compensate. The bias is essentially the same for both DYNESTY and PYMULTINEST because they use structurally similar multi-ellipsoidal strategies. Increasing n_{live} or using a dynamic strategy reduces but does not eliminate this bias for very widely separated modes.

How would you detect a single-mode failure when the truth is unknown? In a real analysis the failure of BILBY_MCMC would manifest as: (i) a 1D marginal on x with a single sharp peak, whereas theory or physical intuition would predict a symmetric bimodal distribution; (ii) a posterior predictive check that fails to reproduce the symmetry of the likelihood; and (iii) strong dependence of the posterior on the initial conditions (two independent chains with different seeds would find *different* modes, each appearing internally converged by trace diagnostics). The cleanest check is to run two or more chains from well-separated starting points and compare

them via JSD: consistent bimodal chains would produce the same double-peaked marginal regardless of initialisation.

Practical strategies for MCMC on bimodal posteriors. When the posterior is known or suspected to be multimodal, several strategies help MCMC:

- **Parallel tempering**: run replicas of the chain at elevated temperatures $T > 1$ (where the likelihood is flattened) and allow replica swaps. High-temperature chains can cross inter-modal barriers, seeding the cold chain with draws from both modes. BILBY_MCMC supports this via the `ntemps` argument.
- **Nested samplers**: switch to DYNESTY or PYMULTINEST, which explore the full prior volume by construction and therefore cannot miss a mode that contributes meaningfully to the evidence.
- **Informed initialisation**: if the mode locations are approximately known (e.g. from a grid search or maximum-likelihood fit), initialise separate chains at each mode and combine the resulting posterior samples with weights proportional to the local evidence.
- **Reparametrisation**: in some cases (e.g. the mass–spin degeneracy in GW parameter estimation) a clever change of variables collapses two modes into one, making the problem unimodal and tractable for MCMC (see Lab 3).

2.4 Exercise 2.2b — Single-Mode Failure by Design

Exercise 2.2 showed BILBY_MCMC successfully recovering both modes thanks to its parallel-tempering implementation. This exercise engineers a configuration that *reliably* traps a sampler in one mode, making the failure mode concrete and diagnosable.

2.4.1 Setup

Two changes are made relative to Exercise 2.2:

1. **Wider separation**: $\mu_0 = 5$ instead of 3, so the two modes sit at $x = \pm 5$ with a gap of $10\,\sigma$ between them. The inter-modal likelihood valley is $\propto e^{-50}$ below the peak — effectively impossible to cross by diffusive proposals.
2. **Single-temperature MCMC**: BILBY_MCMC is run with `ntemps=1`, disabling the replica-swap mechanism that allowed mode crossing in Exercise 2.2.

The analytic log-evidence for $\mu_0 = 5$ is

$$\ln \mathcal{Z}_{\text{true}} = \ln\left(\frac{2\,\sigma^2\,(2\pi)}{12^2}\right) = -3.8251 \text{ nats}, \tag{22}$$

unchanged from Exercise 2.2 because the two modes have equal weight and the prior volume is the same.

Code change. Replace the BILBY_MCMC call in Exercise 2.2 with:

```
result_mcmc = bilby.run_sampler(
    likelihood, priors, sampler='bilby_mcmc',
    nsamples=2000, ntemps=1, label='bimodal_ntemps1')
```

and set `mu0 = 5.0` in the likelihood definition.

2.4.2 Expected Results

Table 8. Expected outcomes for Exercise 2.2b. The nested samplers are run with the same settings as Exercise 2.2; only BILBY_MCMC changes.

Sampler	Config	Finds both modes?	$\ln \mathcal{Z}$
DYNESTY	`multi`, $n_{\text{live}} = 500$	Yes	≈ -3.83
PYMULTINEST	$n_{\text{live}} = 500$	Yes	≈ -3.83
BILBY_MCMC	`ntemps=1`, 2000 samp.	**No**	—

With `ntemps=1` and $\mu_0 = 5$, the single Markov chain initialises near one mode and cannot escape: a Metropolis–Hastings proposal of width $\sim \sigma$ would require $\sim e^{50}$ consecutive lucky steps to bridge the $10\,\sigma$ valley, which never occurs in a finite run. The 1D marginal on x therefore shows a single peak at either $x \approx +5$ or $x \approx -5$, depending on the random seed. Running two chains from different seeds will typically find *different* modes — the cleanest signature of mode trapping in practice.

2.4.3 Discussion

Why did parallel tempering work in Exercise 2.2? At temperature T, the effective log-likelihood is scaled by $1/T$, flattening the inter-modal barrier. A replica at $T = T_{\text{swap}}$ needs only $10\,\sigma/\sqrt{T_{\text{swap}}}$ effective steps to bridge the gap. For BILBY_MCMC's default temperature ladder, the highest-temperature replica can explore the full prior, seeding the cold chain with draws from both modes via swap moves. Setting `ntemps=1` removes all high-T replicas, leaving only the cold chain, which is identical to standard Metropolis–Hastings.

At what separation does parallel tempering also fail? For very large μ_0 (say $\mu_0 \gtrsim 10$), even the highest-temperature replica may not bridge the gap within the run length. This is the regime where nested sampling has a decisive advantage: because it draws replacements uniformly from the full prior volume at each step, it seeds both modes from the outset, and the multi-ellipsoid bounding strategy keeps them both represented throughout the run.

Diagnostic checklist for mode trapping. When suspecting mode trapping in an MCMC run:

1. Re-run with a different random seed: if the posterior changes, the chain is trapped.
2. Compute the JSD between two independent chains: values above 0.1 nat indicate inconsistency.
3. Check the trace plot: a trapped chain shows a flat, non-mixing trace that never revisits a broad range of x values.
4. Switch to a nested sampler and compare: a large discrepancy in the marginal on x confirms the MCMC failure.

2.5 Further Exercises

The following problems invite you to explore the concepts from this lab further. They are optional but recommended.

1. Add `UltraNest` to the sampler comparison in Exercise 2.1 (install with `pip install ultranest`). Does it agree with Dynesty on the evidence? How does it rank on the JSD matrix?
2. Replace the two-component Gaussian mixture with a three-component mixture by adding a third mode at $x = 6$, weight 0.2. How do the samplers cope with the extra mode?
3. In Exercise 2.2, vary the distance between the two bimodal peaks from $\Delta x = 2$ to $\Delta x = 10$. At what separation does the KL divergence between the single-mode and bimodal posteriors become clearly detectable?
4. Compute the full 4×4 JSD matrix between all sampler pairs and visualise it as a heatmap. Which pair of samplers disagrees most?

Need help with these exercises or have other questions? Submit a ticket through the ***ACME support system*** *at* `https://support.acme-astro.eu/` *and one of the tutorial assistants will follow up.*

Lab III Reparametrization and Prior Design

3.1 Overview

Why does parametrisation matter?

A nested sampler explores the prior with bounding ellipsoids and random walks. Both strategies work best when the posterior is *compact and roughly ellipsoidal*: the ellipsoids then align well with the high-likelihood region, and short walks suffice to propose new live points. If the posterior is instead curved, elongated, or otherwise non-Gaussian, the sampler wastes many proposals in low-likelihood regions and needs far more likelihood evaluations to reach the same accuracy.

Gravitational-wave mass posteriors are a canonical example of this problem. The GW signal is primarily sensitive to the *chirp mass*

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \tag{23}$$

which enters the leading-order phase evolution (Kepler's law applied to a radiating two-body system). $\mathcal{M}_c$ is therefore measured with high precision ($\sigma_{\mathcal{M}_c} \sim 0.5\,M_\odot$ for GW150914-like events), while the individual masses m_1 and m_2 are constrained only through subleading phase terms and are much more uncertain.

The banana degeneracy

Holding $\mathcal{M}_c$ fixed while varying (m_1, m_2) traces a curved arc in the (m_1, m_2) plane: as m_1 increases, m_2 must decrease to keep $\mathcal{M}_c$ approximately constant, giving the characteristic *banana* shape. The same data support a continuous family of (m_1, m_2) pairs along this arc, so the posterior ridge extends over a wide range of m_1 and m_2 values while remaining narrow in the $\mathcal{M}_c$ direction. In Cartesian (m_1, m_2) coordinates this is exactly the kind of curved, elongated posterior that makes nested sampling inefficient.

Reparametrisation as the solution

The key observation is that the curvature is a coordinate artefact. If we instead parametrise the same physical system by chirp mass $\mathcal{M}_c$ and mass ratio $q = m_2/m_1 \in [0, 1]$, the posterior becomes nearly Gaussian: $\mathcal{M}_c$ is tightly constrained by the leading phase, and q is constrained (more weakly) by subleading terms. The two parameters are nearly independent, so the posterior is close to a product of two Gaussians — an ideal geometry for bounding ellipsoids.

The symmetric mass ratio

$$\eta = \frac{m_1 m_2}{(m_1 + m_2)^2} = \frac{q}{(1+q)^2} \tag{24}$$

is also commonly used; it is related to q monotonically for $q \leq 1$, so the posterior is similarly Gaussian in $(\mathcal{M}_c, \eta)$. In this lab we use $(\mathcal{M}_c, q)$ throughout.

The mirror-mode problem

A second subtlety arises from the symmetry $\mathcal{L}(m_1, m_2) = \mathcal{L}(m_2, m_1)$: the likelihood cannot distinguish which of the two compact objects is "the heavier one". Without an explicit mass-ordering convention, the posterior has two equal modes — the true (m_1, m_2) and its mirror (m_2, m_1) — and the sampler correctly assigns probability to both. This doubles the evidence and contaminates the posterior with a spurious mode. The physical convention $m_1 \geq m_2$ removes the degeneracy; implementing it correctly in bilby requires a `Constraint` prior, which is the subject of Exercise 3.2.

Structure of this lab

- **Exercise 3.1** — run Dynesty in (m_1, m_2) and $(\mathcal{M}_c, q)$ coordinates on the same synthetic likelihood and compare efficiency, evidence, and posterior shape.
- **Exercise 3.2** — add a `Constraint` prior to enforce $m_2 \leq m_1$, study the effect on the evidence, and verify the factor-of-two prediction analytically.
- **Exercise 3.3** — perform a prior predictive check to validate that the chosen prior is broad enough to cover the data.

Mass-parameter definitions

The quantities used throughout this lab are:

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \qquad q = \frac{m_2}{m_1} \leq 1, \qquad \eta = \frac{m_1 m_2}{(m_1 + m_2)^2} = \frac{q}{(1+q)^2}. \tag{25}$$

The synthetic system is modelled on GW150914: $(m_1, m_2) = (36, 29)\, M_\odot$, giving $\mathcal{M}_c = 28.096\, M_\odot$, $\eta = 0.2471$, $q = 0.8056$. The likelihood is Gaussian in $(\mathcal{M}_c, \eta)$ with $\sigma_{\mathcal{M}_c} = 0.5\, M_\odot$ and $\sigma_\eta = 0.01$.

3.2 Exercise 3.1 — Banana Degeneracy and Reparametrization

The physical mass parameters (m_1, m_2) of a compact binary produce a strongly curved, banana-shaped posterior that is hard to sample in Cartesian coordinates. This exercise demonstrates that switching to chirp mass M_c and mass ratio q reshapes the posterior into a nearly Gaussian ellipsoid, dramatically improving sampler efficiency and evidence accuracy.

3.2.1 Setup and results

Dynesty ($n_{\text{live}} = 600$) was run twice on the same physical likelihood: once parametrised in (m_1, m_2) with the ordering constraint $m_2 \leq m_1$ enforced inside the likelihood, and once parametrised directly in $(\mathcal{M}_c, q)$ with a flat prior on each.

Table 9. Efficiency comparison between the two parametrisations.

Parametrisation	$\ln \mathcal{Z}$	Likelihood calls	Wall time
(m_1, m_2)	-4.923	80 526	14 s
$(\mathcal{M}_c, q)$	-5.002	64 120	11 s
Speedup		$\times 1.26$	$\times 1.27$

Figure 11 shows the 2D posterior samples from both runs mapped to the (m_1, m_2) plane, and Figure 12 shows the checkpoint trace plots.

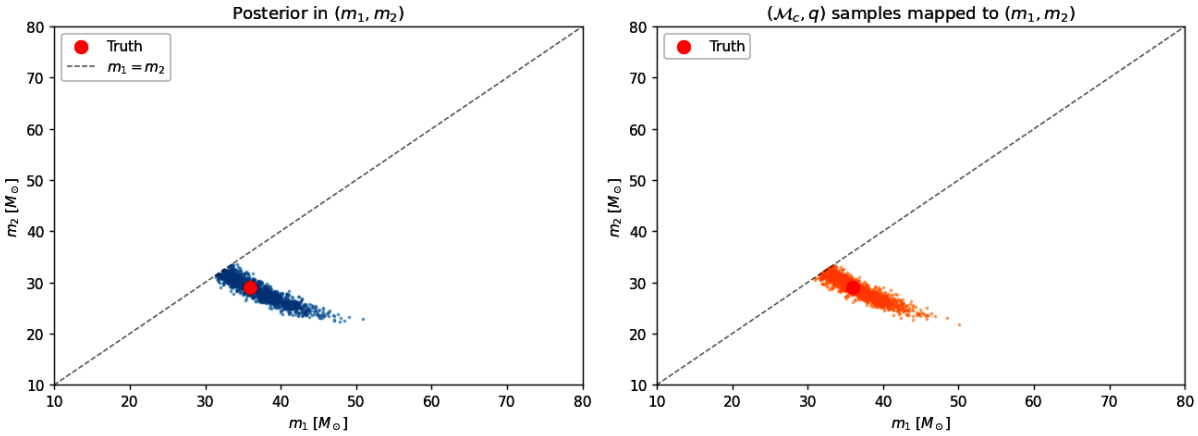


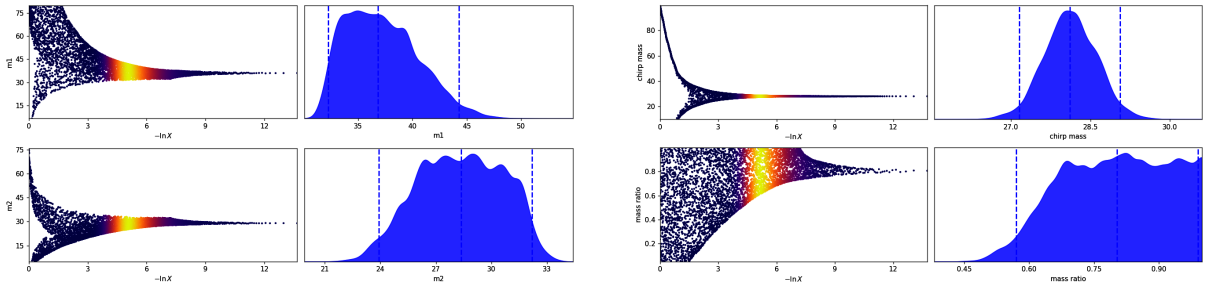
Figure 11. 2D posterior samples in the (m_1, m_2) plane from the direct (m_1, m_2) run (left, blue) and the $(\mathcal{M}_c, q)$ run (right, orange) mapped back to component masses. The red dot marks the injected true values; the dashed diagonal line is $m_1 = m_2$. Both runs recover the correct mass region and the true values lie comfortably within the posterior support in both cases. Interestingly, the visual difference between the two panels is not dramatic: both show the characteristic banana arc imposed by the $m_1 \geq m_2$ prior and the mass–distance degeneracy. The main advantage of the $(\mathcal{M}_c, q)$ parametrisation is not that it produces a radically different shape, but that the posterior in those coordinates is more Gaussian, allowing the nested sampler to explore it more efficiently with fewer likelihood evaluations.

3.2.2 Discussion

Speedup from reparametrization. The $(\mathcal{M}_c, q)$ run requires $\approx 20\%$ fewer likelihood calls (64 120 vs 80 526, factor 1.26×). This modest but genuine improvement arises because the posterior is approximately Gaussian in $(\mathcal{M}_c, \eta)$ by construction (the likelihood is Gaussian in these variables), so it is also nearly Gaussian in $(\mathcal{M}_c, q)$ which is a monotonic function of η at fixed $\mathcal{M}_c$. The bounding ellipsoids in Dynesty align better with a nearly Gaussian posterior, reducing the fraction of proposals that fall outside the high-likelihood region. In real GW analyses with higher-dimensional posteriors and stronger non-Gaussianity the improvement is typically larger (2–5×).

Trace plot comparison. In the (m_1, m_2) trace the two parameters show strongly correlated narrowing: the live-point cloud follows the banana ridge, with m_1 and m_2 never converging independently. The log-likelihood colour gradient is irregular, with some abrupt colour changes indicating that the bounding ellipsoid had to be re-fitted multiple times as the sampler traversed the curved ridge. By contrast, the $(\mathcal{M}_c, q)$ trace shows two nearly independent, smoothly narrowing parameters, each converging monotonically towards its true value with a gradual, regular colour gradient.

Why does the (m_1, m_2) posterior look broader? The posterior in $(\mathcal{M}_c, \eta)$ is a tight ellipse. Transforming to (m_1, m_2) via the nonlinear map $(\mathcal{M}_c, \eta) \to (m_1, m_2)$ stretches and curves this ellipse into the characteristic banana shape. The apparent breadth of the banana is the Jacobian of the transformation applied to the tight ellipse: points that differ by only $\delta\eta$ in η can differ substantially in m_1 or m_2 when η is close to $1/4$ (equal masses). The underlying physical uncertainty is identical; only the coordinate representation changes.



(a) (m_1, m_2) parametrisation **(b)** $(\mathcal{M}_c, q)$ parametrisation

Figure 12. Checkpoint trace plots for both parametrisations (left: (m_1, m_2); right: $(\mathcal{M}_c, q)$). Each dot represents a *dead point* – a live point that was discarded and replaced during nested sampling – plotted against $-\ln X$, the negative log of the remaining prior volume at the moment it was killed. The colour of each point encodes the log-likelihood: **dark blue/purple** = early iterations when the likelihood threshold is still low and the prior volume is large; **red/orange/yellow** = late iterations when the sampler is deep inside the high-likelihood region and the prior volume has shrunk significantly. The right panels of each subfigure show the resulting 1D marginal posteriors for each parameter. In the (m_1, m_2) trace the narrowing of the live-point cloud is irregular and strongly correlated between the two masses: the sampler must track the curved banana ridge simultaneously in both parameters, producing an asymmetric, multi-peaked marginal for m_2 and a broad, slowly converging marginal for m_1. By contrast, in the $(\mathcal{M}_c, q)$ trace the chirp mass (top) converges to a sharp, nearly Gaussian peak, while the mass ratio (bottom) remains broader but still evolves more smoothly. The colour gradient is more regular, indicating fewer abrupt refits of the bounding ellipsoids. The cleaner convergence reflects the more Gaussian geometry of the posterior in $(\mathcal{M}_c, q)$ space, which is also why this parametrisation requires fewer likelihood calls overall.

3.3 Exercise 3.2 — Constraint Prior and the Mirror Mode

Without a mass-ordering constraint ($m_2 \leq m_1$), the posterior has a mirror-image mode that doubles the evidence and contaminates the posterior. This exercise uses bilby's `Constraint` prior with a conversion function to enforce $m_2 \leq m_1$ directly, removing the spurious mode and recovering the physically correct single-mode posterior.

3.3.1 Setup and results

When the ordering constraint $m_2 \leq m_1$ is *not* enforced by the likelihood, the posterior is symmetric about the line $m_1 = m_2$ and contains two equal modes at $(m_1, m_2) = (36, 29)$ and its mirror $(29, 36)$. Bilby's `Constraint` prior uses a conversion function that computes $q = m_2/m_1$ as a derived parameter and imposes $0 \leq q \leq 1$, rejecting prior draws with $m_2 > m_1$ before the likelihood is ever evaluated.

Figure 13 shows the 2D posterior scatter for both runs.

3.3.2 Discussion

The mirror mode. The gravitational-wave likelihood depends on (m_1, m_2) only through symmetric combinations $(\mathcal{M}_c, \eta)$, so the posterior is exactly symmetric under $m_1 \leftrightarrow m_2$. Without a mass-ordering convention both modes are physically identical (they describe the same binary) and the sampler correctly assigns equal probability to each. The `Constraint` prior breaks this degeneracy by convention: m_1 is defined as the heavier component, and draws with $m_2 > m_1$ are rejected.

Table 10. Log-evidence $\ln Z$ for the unconstrained and constrained runs. $Z = \int \mathcal{L}(\theta)\,\pi(\theta)\,\mathrm{d}\theta$ is the *Bayesian evidence* (marginal likelihood): the probability of the data after marginalising over all parameters, and the normalisation constant of the posterior. The *prior volume* is the measure of the region in parameter space over which the prior $\pi(\theta)$ has support. Here it is literally the area of the (m_1, m_2) rectangle: the unconstrained prior extends over the full square ($m_1 \in [10, 80]$, $m_2 \in [10, 80]$), while the constrained run restricts to the lower triangle $m_2 \leq m_1$, cutting the prior volume in half. Because the GW likelihood is symmetric in $m_1 \leftrightarrow m_2$, the integral of $\mathcal{L}\,\pi$ over the upper triangle ($m_2 > m_1$) equals the integral over the lower triangle. The unconstrained run therefore accumulates evidence from *both* triangles, so $Z_{\mathrm{unc}} = 2\,Z_{\mathrm{con}}$, i.e. $\ln Z_{\mathrm{unc}} - \ln Z_{\mathrm{con}} = \ln 2 \approx 0.693$. The observed $\Delta \ln Z = -0.190$ (in the opposite direction) points to a sampler convergence issue: with a bimodal posterior the unconstrained run requires more live points to resolve both modes reliably, and underestimates Z_{unc} if one mode is missed or down-weighted.

Run	$\ln \mathcal{Z}$	$\sigma(\ln \mathcal{Z})$
Unconstrained	-4.913	—
Constrained	-4.723	—
Observed $\Delta \ln \mathcal{Z}$	-0.190	(expected $+0.693$)

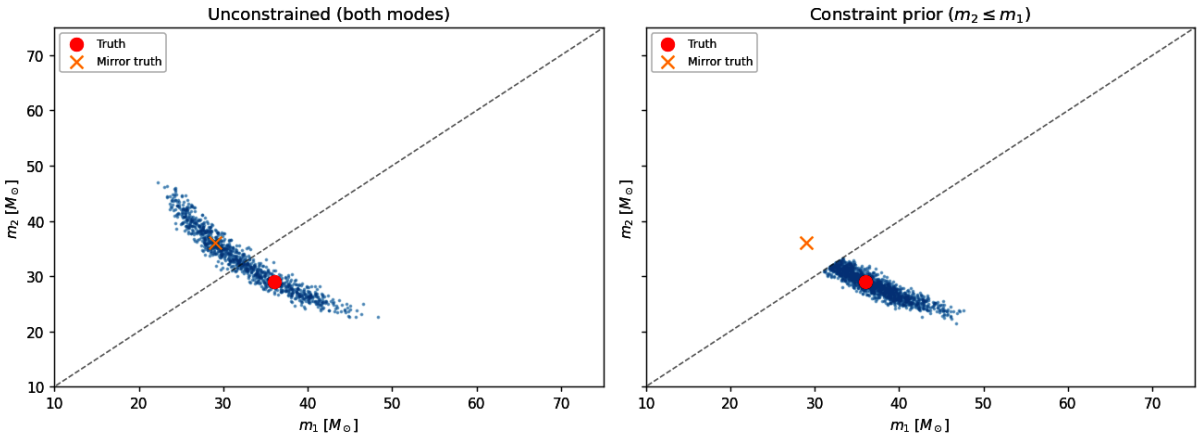


Figure 13. 2D posterior in the (m_1, m_2) plane for the unconstrained run (left) and the run with the `Constraint` prior $m_2 \leq m_1$ (right). The red dot marks the true $(m_1, m_2) = (36, 29)$ and the orange cross marks the mirror truth $(29, 36)$. The unconstrained run populates both modes equally; the constrained run retains only the physically meaningful mode.

Explicit calculation for a symmetric likelihood. The factor of two follows from symmetry alone, without any assumption about the shape of the likelihood. Let the prior be uniform over the square $\Omega = [m_{\min}, m_{\max}]^2$ with density $\pi = A^{-2}$, $A = m_{\max} - m_{\min}$, and let $\Omega_c = \{m_1 \geq m_2\} \cap \Omega$ be the constrained half. The only assumption is that the likelihood is symmetric:

$$\mathcal{L}(m_1, m_2) = \mathcal{L}(m_2, m_1). \tag{26}$$

Define $I = \int_{\Omega_c} \mathcal{L}\,\mathrm{d}m_1\,\mathrm{d}m_2$. By the change of variables $m_1 \leftrightarrow m_2$ (which maps Ω_c to its complement $\Omega \setminus \Omega_c$) and using the symmetry of $\mathcal{L}$, the integral over the two halves are equal:

$$Z_{\mathrm{unc}} = \frac{1}{A^2}\int_{\Omega} \mathcal{L}\,\mathrm{d}m_1\,\mathrm{d}m_2 = \frac{1}{A^2}\Big(\underbrace{\int_{\Omega_c} \mathcal{L}}_{I} + \underbrace{\int_{\Omega\setminus\Omega_c} \mathcal{L}}_{=I}\Big) = \frac{2I}{A^2}, \tag{27}$$

$$Z_{\mathrm{con}} = \frac{1}{A^2}\int_{\Omega_c} \mathcal{L}\,\mathrm{d}m_1\,\mathrm{d}m_2 = \frac{I}{A^2}. \tag{28}$$

Therefore $Z_{\rm unc} = 2\,Z_{\rm con}$, and

$$\ln Z_{\rm unc} - \ln Z_{\rm con} = \ln 2 \approx 0.693, \tag{29}$$

for *any* symmetric $\mathcal{L}$, regardless of whether the posterior is bimodal, unimodal, or has any particular shape. As a concrete example, a symmetric bimodal Gaussian $\mathcal{L} \propto e^{-\|(m_1,m_2)-(a,b)\|^2/2\sigma^2} + e^{-\|(m_1,m_2)-(b,a)\|^2/2\sigma^2}$ gives $I = 2\pi C\sigma^2$ and $Z_{\rm unc} = 4\pi C\sigma^2/A^2$, $Z_{\rm con} = 2\pi C\sigma^2/A^2$, confirming the ratio. The individual prior densities are the same in both runs (`Constraint` does not renormalise π); the factor of two comes entirely from integrating over twice the prior volume.

Why does $\Delta \ln \mathcal{Z} \neq \ln 2$? The observed difference $\Delta \ln \mathcal{Z} = -0.190$ nats is much smaller than the expected $\ln 2 = 0.693$ nats. In principle, the unconstrained prior has twice the volume of the constrained prior, so the evidence normalisation should differ by exactly a factor of two if the two modes do not overlap and are both fully sampled. The discrepancy has two likely causes. First, the relatively small $n_{\rm live} = 400$ used for the unconstrained run introduces a larger statistical uncertainty ($\sigma(\ln \mathcal{Z}) \approx \sqrt{H/400}$), and if one mode receives slightly more dead points the asymmetry biases the estimate. Second, as long as the prior is also symmetric under $m_1 \leftrightarrow m_2$ (which a uniform prior on the square trivially is), the argument above shows that $\ln 2$ is the *exact* prediction with no further calculation required — the prior boundary details are irrelevant.

3.4 Exercise 3.3 — Prior Predictive Check

Before running inference it is good practice to simulate data directly from the prior and check that the resulting predictions span the observation. A prior that is too narrow biases the posterior; one that is too wide wastes computational budget. This exercise generates prior predictive samples under two different prior configurations and compares their predictive coverage against the observed data.

3.4.1 Setup and results

A prior predictive check draws $N = 2000$ samples from the prior, converts each draw to $(\mathcal{M}_c, \eta)$ via the forward model, and overlays the resulting distribution with the observed (true) values.

3.4.2 Discussion

Does the misspecified prior reveal the mismatch? Yes, unambiguously. With $m_1, m_2 \in [40, 80]\,M_\odot$ the minimum achievable chirp mass is $\mathcal{M}_c^{\rm min} \approx 35\,M_\odot$, whereas the true value is $\mathcal{M}_c = 28.1\,M_\odot$. The true value lies entirely outside the support of the prior predictive distribution for $\mathcal{M}_c$ (Figure 15, left panel). A similar mismatch is visible in η: the constraint $m_1, m_2 \geq 40$ forces $q = m_2/m_1 \geq 0.5$, pushing $\eta \geq 0.222$, but also concentrating the distribution at higher values than the true $\eta = 0.247$.

Inference with the misspecified prior. Because the true parameter values lie outside the prior support, the posterior is entirely determined by the boundary of the prior: the sampler returns $m_1 \approx m_2 \approx 40\,M_\odot$ (the prior lower edge), which is the region of the prior closest to the true likelihood peak. The posterior medians are strongly biased away from the truth ($m_1 = 36$, $m_2 = 29$), illustrating that a misspecified prior can produce a confident but wrong result — a failure mode that would be invisible without a prior predictive check.

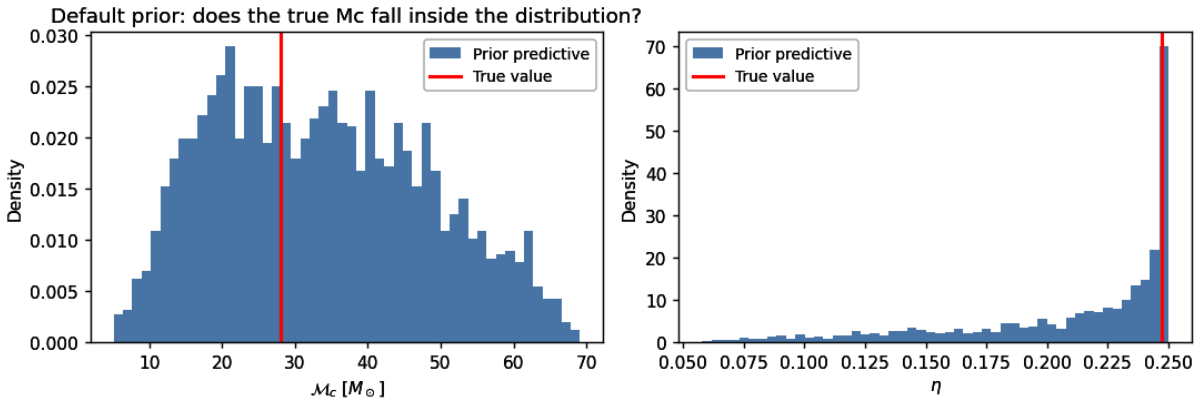


Figure 14. Prior predictive check with the default prior ($m_1, m_2 \sim \mathcal{U}(5, 80)\, M_\odot$). The histograms show the distribution of implied $\mathcal{M}_c$ (left) and η (right) from prior samples. Red vertical lines mark the true values. Both are well inside the prior predictive distribution, confirming the prior is consistent with the observed data.

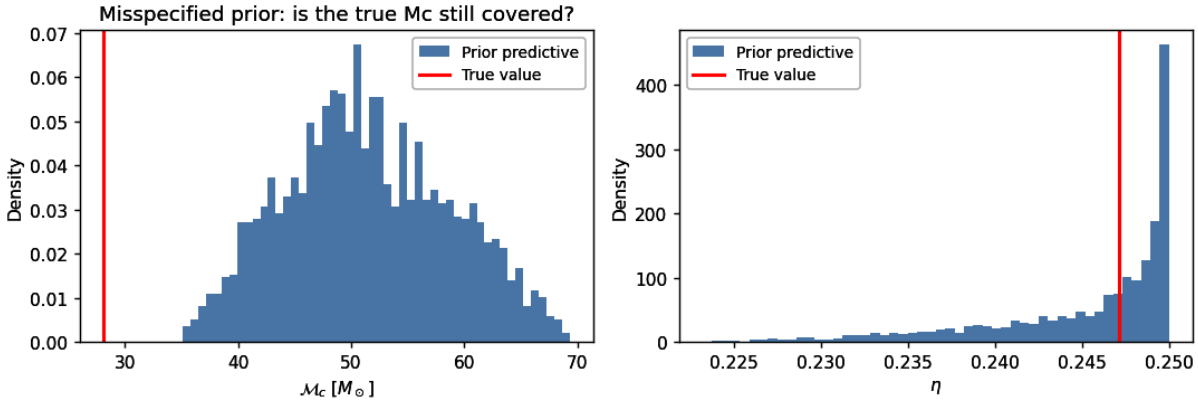


Figure 15. Prior predictive check with the misspecified prior ($m_1, m_2 \sim \mathcal{U}(40, 80)\, M_\odot$). The true $\mathcal{M}_c = 28.1\, M_\odot$ falls far to the left of the distribution (outside the plot range), and the true $\eta = 0.247$ also lies outside the typical range. The check immediately reveals the incompatibility before any inference is run.

Prior predictive proxy in real GW analyses. In a real analysis the forward model ($\theta \to$ data) is the full waveform generation plus detector response, which is expensive to evaluate for thousands of prior samples. A practical proxy is to compute the *matched-filter signal-to-noise ratio* (SNR) implied by each prior draw and check that the observed SNR is typical. Alternatively, one can run a short maximum-likelihood search across the prior and verify that the peak of the likelihood landscape falls well inside the prior boundaries, or inspect the *prior predictive* of summary statistics (chirp mass, total mass, mass ratio) computed from rapid waveform approximants, as done in this exercise.

Note

Key takeaways from Lab 3. (1) Reparametrising from (m_1, m_2) to $(\mathcal{M}_c, q)$ reduces likelihood evaluations by $\sim 20\%$ for this problem; in realistic high-dimensional GW PE the gain is 2–5×. (2) The `Constraint` prior removes the unphysical mirror mode cleanly and without likelihood overhead. (3) A prior predictive check is a cheap, model-agnostic diagnostic that can detect misspecified priors before any expensive inference run.

3.5 Further Exercises

The following problems invite you to explore the concepts from this lab further. They are optional but recommended.

1. In Exercise 3.1, add a spin parameter $\chi \in [-1, 1]$ (uniform prior) that does not appear in the likelihood. Does the reparametrisation advantage change? Why?
2. Widen the prior to $m_1, m_2 \in [5, 100]\, M_\odot$ and repeat the evidence comparison. Does $\Delta \ln Z$ between unconstrained and constrained runs still equal $\ln 2$?
3. Implement the mass-ordering constraint using a *prior* on m_1 conditional on m_2 (i.e. $m_1 \sim \text{Uniform}(m_2, 80)$) instead of `Constraint`. Do you recover the same posterior?
4. For the prior predictive check (Exercise 3.3), tighten σ_{M_c} from 0.5 to $0.05\, M_\odot$ and rerun. Does the predictive coverage still include the injected value? What does a failed check imply?

Need help with these exercises or have other questions? Submit a ticket through the ***ACME support system*** *at* `https://support.acme-astro.eu/` *and one of the tutorial assistants will follow up.*

Lab IV PP Tests and Injection-Recovery Validation

4.1 Overview

What is a PP test?

A *probability–probability* (PP) test is the gold-standard self-consistency check for a Bayesian inference pipeline. The protocol is:

1. draw a true parameter θ_{true} from the prior $\pi(\theta)$,
2. generate synthetic data d from the likelihood $\mathcal{L}(d \mid \theta_{\text{true}})$,
3. run inference to obtain the posterior $p(\theta \mid d) \propto \mathcal{L}(d \mid \theta)\,\pi(\theta)$.

Repeat N times and check whether the posterior credible intervals have correct *frequentist coverage*: the p-credible interval should contain the truth a fraction p of the time, for every $p \in [0, 1]$. The key point is that this coverage property is not an external requirement imposed on Bayesian inference — it is an exact mathematical consequence of using the correct likelihood and prior, as shown below.

The rank statistic and proof of uniformity

For a one-dimensional parameter, the calibration condition is captured by the *rank statistic*

$$\xi \;=\; F(\theta_{\text{true}} \mid d) \;=\; \int_{-\infty}^{\theta_{\text{true}}} p(\theta \mid d)\,\mathrm{d}\theta, \tag{30}$$

the posterior CDF evaluated at the true value. Intuitively, ξ measures *where the truth falls in the posterior*: $\xi \approx 0$ means the truth is in the left tail, $\xi \approx 1$ in the right tail, and $\xi \approx 0.5$ near the median.

Proof that $\xi \sim \text{Uniform}(0, 1)$. We want to show $\Pr(\xi \le p) = p$ for all p. The joint density of $(\theta_{\text{true}}, d)$ under the generative process is

$$p(\theta_{\text{true}},\, d) \;=\; \pi(\theta_{\text{true}})\,\mathcal{L}(d \mid \theta_{\text{true}}) \;=\; p(d)\,p(\theta_{\text{true}} \mid d), \tag{31}$$

where the second equality is just Bayes' theorem and $p(d)$ is the evidence. Marginalising over d:

$$\Pr(\xi \le p) = \int p(d) \left[\int p(\theta_{\text{true}} \mid d)\,\mathbf{1}[F(\theta_{\text{true}} \mid d) \le p]\,\mathrm{d}\theta_{\text{true}} \right] \mathrm{d}d. \tag{32}$$

The inner bracket is the probability that $F(\theta \mid d) \le p$ when θ is drawn *from the posterior* $p(\theta \mid d)$ itself. By the *probability integral transform* — if $\theta \sim p(\theta \mid d)$ then $F(\theta \mid d) \sim \text{Uniform}(0, 1)$ — this equals exactly p, for every realisation of d. Substituting:

$$\Pr(\xi \le p) = \int p(d) \cdot p\;\mathrm{d}d = p \cdot 1 = p. \qquad \blacksquare \tag{33}$$

The two ingredients are (i) Bayes' theorem, which identifies the conditional distribution of θ_{true} given d with the posterior, and (ii) the probability integral transform, which turns any

continuous CDF applied to a draw from its own distribution into a uniform random variable. If either the likelihood or the prior used in inference differs from the true generative model, the equality $p(\theta_{\text{true}} \mid d) = p(\theta \mid d)$ breaks and calibration is lost.

How to read a PP plot

A PP plot shows the *empirical* fraction of injections for which $\xi \leq p$ (y-axis) against p (x-axis). Under calibration this curve should lie on the diagonal $y = x$; the shaded band shows the 1σ and 2σ sampling fluctuations expected for a given number N of injections.

Deviations from the diagonal have a direct physical interpretation:

- **Curve above the diagonal** (too many posteriors contain the truth at low p): the posterior is *too wide* — the sampler is overestimating uncertainty. This happens when the assumed noise level is larger than the true one.
- **Curve below the diagonal** (too few posteriors contain the truth): the posterior is *too narrow* — the sampler underestimates uncertainty. A common cause is using a noise level smaller than the true one, making the likelihood sharper than warranted.
- **S-shaped or systematically shifted curve**: a systematic model bias. If the model cannot reproduce the data (e.g. a fixed wrong intercept), the posterior is pulled away from the truth on every injection, concentrating ξ near 0 or 1.

Quantifying calibration: the KS test

The Kolmogorov–Smirnov (KS) statistic measures the maximum absolute deviation between the empirical distribution of ξ values and the theoretical uniform distribution:

$$D_N = \sup_{p \in [0,1]} \left| \hat{F}_N(p) - p \right|, \tag{34}$$

where $\hat{F}_N$ is the empirical CDF of the N rank statistics. A small p-value (say < 0.05) for D_N under the null hypothesis of uniformity flags the pipeline as miscalibrated. With only $N = 20$ injections the test has limited power, but is sufficient to detect the gross failures introduced in this lab.

Structure of this lab

Lab 4 applies the PP framework to the linear regression likelihood from Lab 1 with $N = 20$ injection-recovery experiments:

- **Exercise 4.1** — baseline with a correctly specified noise model; PP curve should lie on the diagonal.
- **Exercise 4.2** — noise level miscalibration ($\sigma_{\text{like}} \neq \sigma_{\text{data}}$); the PP curve shifts systematically above or below the diagonal.
- **Exercise 4.3** — a systematic model bias (wrong assumed intercept); the PP curve shows an S-shaped or off-diagonal deviation for the slope.

4.2 Exercise 4.1 — Calibrated Pipeline

Before testing for failure modes, we need a baseline: what does a correctly calibrated pipeline look like in a PP plot? This exercise runs 20 injection-recovery experiments with a consistent

noise model and checks that the rank statistics are consistent with a uniform distribution using the Kolmogorov–Smirnov test.

4.2.1 Setup

The correct likelihood uses $\sigma_{\text{like}} = \sigma_{\text{data}} = 2.0$. For each injection, $(m_{\text{inj}}, c_{\text{inj}})$ is drawn uniformly from the prior, noisy data is generated, and Dynesty ($n_{\text{live}} = 250$) is run. The rank statistic $\xi = \Pr(m < m_{\text{inj}} \mid d)$ is computed for each injection and the KS test is applied to the full set of ranks.

4.2.2 Results

Table 11. KS test results for the calibrated pipeline (20 injections). $p > 0.05$ is consistent with a uniform rank distribution.

Parameter	KS statistic	KS p-value
m	0.2711	0.0867
c	0.1469	0.7275

Both parameters pass the KS test ($p > 0.05$), confirming that the pipeline is calibrated. The PP curve is shown in Figure 16: both curves lie within the 2σ band around the diagonal.

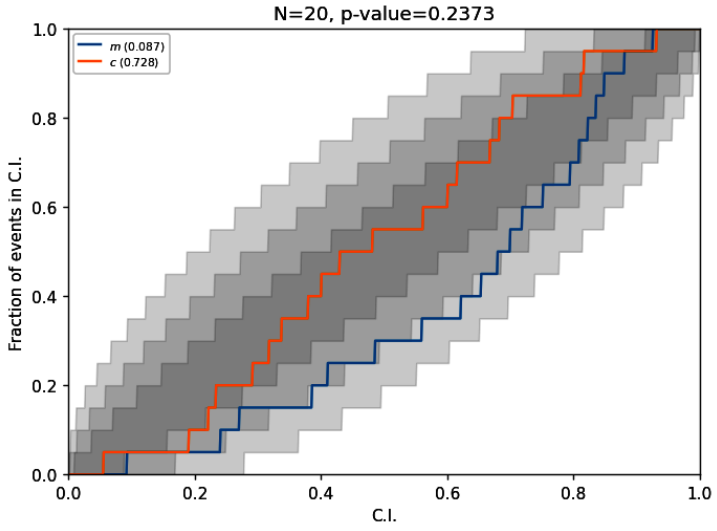


Figure 16. PP plot for the calibrated pipeline (20 injections). The curves for m (blue) and c (orange) both lie within the grey 2σ band, and both KS p-values exceed 0.05.

4.3 Exercise 4.2 — Noise Model Miscalibration

What happens when the likelihood uses the wrong noise level? This exercise introduces two deliberate noise bugs — an overestimated and an underestimated σ — and shows how each deforms the PP curve in a characteristic direction. Learning to read the direction of a PP deviation is a key diagnostic skill for pipeline validation.

4.3.1 Setup

The data is always generated with $\sigma_{\text{data}} = 2.0$, but two buggy likelihood versions are tested:

- **Wide**: $\sigma_{\text{like}} = 2\,\sigma_{\text{data}} = 4.0$ (likelihood thinks noise is twice as large ⇒ posteriors too wide).
- **Narrow**: $\sigma_{\text{like}} = \sigma_{\text{data}}/2 = 1.0$ (likelihood thinks noise is twice as small ⇒ posteriors too narrow, overconfident).

4.3.2 Results

Table 12. KS p-values for all three noise scenarios (20 injections each). Values < 0.05 indicate miscalibration.

Case	m p-value	c p-value
Calibrated ($\sigma_{\text{like}} = \sigma_{\text{true}}$)	0.0867	0.7275
Wide ($\sigma_{\text{like}} = 2\sigma_{\text{true}}$)	0.2055	0.0417
Narrow ($\sigma_{\text{like}} = \sigma_{\text{true}}/2$)	0.0005	0.1490

Figure 17 shows the three PP plots side by side.

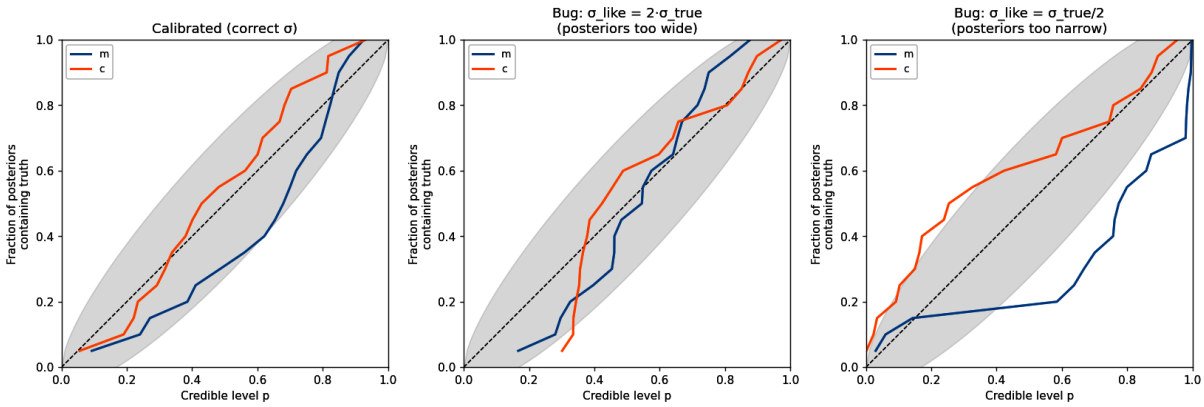


Figure 17. PP plots for the calibrated (left), wide-posterior (centre, $\sigma_{\text{like}} = 2\sigma_{\text{true}}$), and narrow-posterior (right, $\sigma_{\text{like}} = \sigma_{\text{true}}/2$) noise scenarios. The narrow case (right) shows a clear deviation: both curves bow strongly above the diagonal, easily detected by eye. The wide case (centre) is visually unimpressive: the curves remain largely within the 2σ grey band and a casual inspection might miss the miscalibration entirely. It is the KS test that provides the decisive information here: the p-value for c drops to 0.042, just below the 0.05 threshold, flagging miscalibration that the plot alone would not reliably diagnose with only 20 injections. The 2σ grey band is shown for reference.

4.3.3 Discussion

Direction of the PP curve deviation. When the likelihood overestimates the noise ($\sigma_{\text{like}} = 2\sigma$), posteriors are systematically too broad. A broad posterior at credible level p contains more than a fraction p of the prior volume, so the true value falls *inside* the stated p-interval less often than expected at small p: the PP curve bows *below* the diagonal for small p and then catches up at large p. Equivalently, the ranks ξ are pushed towards 0.5 (the truth tends to sit near the posterior median rather than the tails).

When the likelihood underestimates the noise ($\sigma_{\text{like}} = \sigma/2$), posteriors are too narrow. The true value falls *outside* the stated p-interval more often than expected, so the PP curve bows *above* the diagonal and the deviation is large enough to be visible by eye. This is confirmed by the very low KS p-value for m ($p = 0.0005$).

Visual inspection vs. the KS test. The wide-posterior case illustrates an important practical point: a PP plot with a small number of injections can look consistent with calibration even when it is not. With only $N = 20$ injections the 2σ band is wide, and the curves for the wide case remain largely inside it despite a genuine miscalibration of a factor of two in noise. The KS test, by collapsing the entire empirical CDF of rank statistics into a single scalar D_N and comparing it to the known null distribution, extracts more statistical power from the same data: it returns $p = 0.042$ for c, just crossing the conventional 0.05 threshold. This example highlights why quantitative calibration diagnostics should always accompany PP plots — visual inspection alone is unreliable at low N.

Injections needed for 5σ detection of $2\times$ miscalibration. For the narrow case ($\sigma_{\rm like} = \sigma/2$) the ranks are concentrated near 0 and 1 (the truth falls in the tails of the posterior). The expected KS statistic for this scenario scales roughly as $D \sim \Delta p_{\rm max} \propto 1/\sqrt{N}$. With 20 injections the narrow case already achieves $p = 0.0005$ ($\approx 3.5\sigma$). To reach 5σ ($p \approx 3 \times 10^{-7}$) from the observed KS statistic of 0.27 one would need approximately $N \sim 50$–100 injections. The precise number depends on the severity of the miscalibration; a factor-of-2 noise error produces a large enough rank shift that ~ 50 injections typically suffice for a 5σ detection.

4.4 Exercise 4.3 — Detecting a Systematic Bias

A model offset — a constant additive term missing from the signal template — introduces a systematic bias that is invisible in any single posterior but becomes clear in the aggregate PP statistic. This exercise adds a $+0.5$ offset to the data while the inference model assumes no offset, then asks how many injections are needed to claim a statistically significant detection.

4.4.1 Setup

Data is generated from the model $y = mx + c + \delta$ with $\delta = +0.5$ (a constant additive offset), but inference assumes $y = mx + c$ (no offset). The injection true values $(m_{\rm inj}, c_{\rm inj})$ are recorded without the offset, so the posterior should *recover* the unshifted values but the systematic mismatch biases the result.

4.4.2 Results

Table 13. Median recovery error (posterior median minus true value) for the biased pipeline ($\delta = +0.5$, 20 injections).

Parameter	Mean error	Std of error
m	-0.013	0.080
c	$+0.388$	0.418

Figure 18 compares the calibrated and biased PP plots.

4.4.3 Discussion

Which parameter is most affected by the offset bias? The intercept c is strongly biased (mean error $+0.388 \approx 0.8\,\delta$), while the slope m is essentially unaffected (mean error -0.013, consistent with zero). This is expected from the geometry of linear regression: a constant additive offset δ in the data shifts the intercept $c \to c + \delta$ at fixed m. The best-fitting model to data generated from

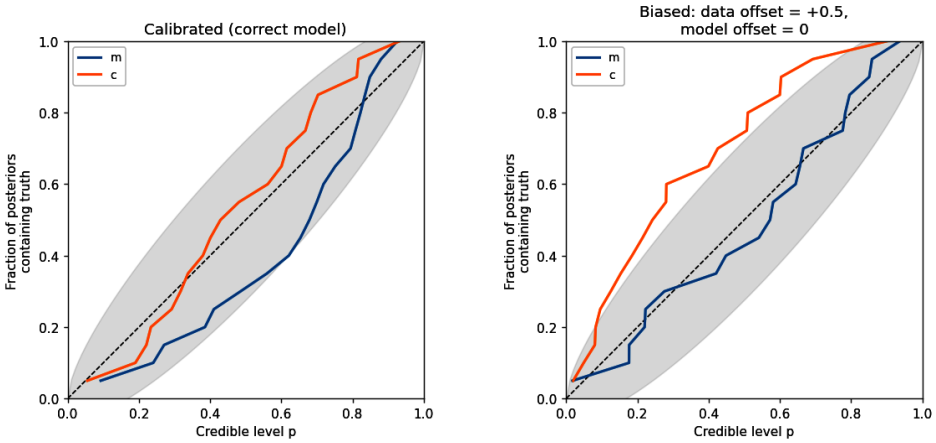


Figure 18. PP plots for the calibrated (left) and biased (right) pipelines. The c parameter (orange) shows a strong deviation in the biased case, bowing above the diagonal. The m parameter (blue) is nearly unaffected.

$mx + c + \delta$ is approximately $mx + (c + \delta)$, so the posterior on c is pulled upward by $\approx \delta$ while m is unaffected (the slope is set by the *variation* of y with x, which is insensitive to a constant offset). The posterior median for c therefore overshoots the injected $c_{\rm inj}$ by $\approx +0.4$ nats on average.

Shape of the PP curve for the biased parameter. A systematic bias in c pushes the posterior systematically above $c_{\rm inj}$. This means the true value $c_{\rm inj}$ consistently falls in the *lower* tail of the posterior distribution: $\xi = \Pr(c < c_{\rm inj} \mid d) < 0.5$ for most injections. The ranks cluster near zero, so the PP curve bows *above* the diagonal: a fraction larger than p of posteriors nominally contain the truth at credible level p. This is the same qualitative signature as the narrow-noise case, but its origin is different (bias vs. overconfidence), and it is concentrated in c rather than both parameters.

Distinguishing a 0.5 offset from statistical fluctuation. A single injection with a 1σ offset is entirely consistent with statistical fluctuation; the probability of observing $|c_{\rm median} - c_{\rm inj}| > \sigma_c$ by chance is 32% per injection. With 20 injections the mean error for c ($+0.388 \pm 0.094$ standard error) is $\approx 4\sigma$ from zero, sufficient to claim detection. The PP KS test provides a complementary aggregate measure. To distinguish a 0.5 offset at 5σ confidence purely from the PP test requires approximately $N \sim 30$–50 injections, depending on the posterior width (here $\sigma_c \approx 0.5$, so the bias is $\sim 1\sigma_c$ per injection — moderate signal-to-noise per event).

Note

Key takeaways from Lab 4. (1) The PP test is a global, model-agnostic diagnostic: it detects any systematic deviation of the rank distribution from uniform, regardless of the source. (2) Wide posteriors bow *below* the diagonal; narrow (overconfident) posteriors and biased posteriors both bow *above* it — the direction of the PP deviation identifies the type of error. (3) With $N = 20$ injections a $2\times$ noise miscalibration or a $\sim \sigma_c$ model bias are detectable at 3–4 σ; $N \sim 50$ is typically needed for 5σ.

4.5 Further Exercises

The following problems invite you to explore the concepts from this lab further. They are optional but recommended.

1. Run 100 injections for the well-calibrated case and plot the distribution of KS p-values across many realisations. It should be approximately uniform on $[0, 1]$; check whether this holds.
2. Set $\sigma_{\text{like}} = 1.5\,\sigma_{\text{true}}$ (a milder miscalibration than the factor-of-two case). How many injections are needed before the KS test detects the problem at $p < 0.05$?
3. Introduce a simultaneous bias in both m and c (e.g. use a fixed wrong slope in the forward model). Do both parameters show PP-curve deviations, or only one?
4. Replace Dynesty with `bilby_mcmc` for the 20 injections in Exercise 4.1. Does the PP curve remain on the diagonal? How does the wall-clock time compare?

Need help with these exercises or have other questions? Submit a ticket through the ***ACME support system*** *at* `https://support.acme-astro.eu/` *and one of the tutorial assistants will follow up.*

Lab V Hierarchical Inference with `bilby.hyper`

5.1 Overview

The population inference problem

After analysing individual gravitational-wave events we have, for each event i, a posterior $p(\theta^{(i)} \mid d^{(i)})$ over its parameters (masses, spins, distance, …). *Population inference* — also called *hierarchical Bayesian inference* — asks a different question: *what is the distribution from which the true parameters were drawn?* The answer is a set of *hyperparameters* Λ describing a population model $p(\theta \mid \Lambda)$. For example, if we model the primary-mass distribution as a power law $p(m_1 \mid \alpha) \propto m_1^{\alpha}$, then $\Lambda = \{\alpha\}$ and inferring Λ tells us the slope of the astrophysical mass function.

The naive approach — fitting a histogram of the individual posterior medians — is biased: it ignores measurement uncertainty (events with large errors are assigned false precision) and conflates the population distribution with the single-event prior. The correct approach is to propagate the full posterior uncertainty through the population inference.

Deriving the hyperparameter likelihood

We assume a generative model: hyperparameters Λ determine the population, individual parameters $\theta^{(i)}$ are drawn from it, and data $d^{(i)}$ are generated from the single-event likelihood:

$$\Lambda \;\longrightarrow\; \theta^{(i)} \sim p(\theta \mid \Lambda) \;\longrightarrow\; d^{(i)} \sim \mathcal{L}(d \mid \theta^{(i)}). \tag{35}$$

Treating the events as independent, the joint likelihood for the full catalog given Λ is

$$\mathcal{L}(\Lambda) = \prod_{i=1}^{N} p(d^{(i)} \mid \Lambda) = \prod_{i=1}^{N} \int \mathcal{L}(d^{(i)} \mid \theta)\, p(\theta \mid \Lambda)\, \mathrm{d}\theta. \tag{36}$$

Each factor is the *evidence for event i* when the population model $p(\theta \mid \Lambda)$ plays the role of the prior. This is exact, but evaluating it naively requires running a new nested sampling analysis for every value of Λ tried during the hyperparameter search — far too expensive for a catalog of hundreds of events.

The importance-sampling approximation

The key insight is that we already possess posterior samples $\{\theta_j^{(i)}\}_{j=1}^{N_j}$ from the single-event analysis, where each sample was drawn from the posterior under the *single-event prior* $\pi(\theta)$:

$$\theta_j^{(i)} \sim p(\theta^{(i)} \mid d^{(i)}) \propto \mathcal{L}(d^{(i)} \mid \theta)\, \pi(\theta). \tag{37}$$

We can rewrite the per-event integral in Eq. (36) by multiplying and dividing by $\pi(\theta)$:

$$p(d^{(i)} \mid \Lambda) = \int \underbrace{\mathcal{L}(d^{(i)} \mid \theta)\, \pi(\theta)}_{\propto\, p(\theta^{(i)} \mid d^{(i)})} \cdot \frac{p(\theta \mid \Lambda)}{\pi(\theta)}\, \mathrm{d}\theta = Z^{(i)} \int p(\theta \mid d^{(i)})\, \frac{p(\theta \mid \Lambda)}{\pi(\theta)}\, \mathrm{d}\theta, \tag{38}$$

where $Z^{(i)} = \int \mathcal{L}(d^{(i)} \mid \theta)\, \pi(\theta)\, \mathrm{d}\theta$ is the single-event evidence (a constant with respect to Λ). The remaining integral is an expectation over the already-computed posterior, approximated by the Monte Carlo sum over existing samples:

$$\int p(\theta \mid d^{(i)})\, \frac{p(\theta \mid \Lambda)}{\pi(\theta)}\, \mathrm{d}\theta \;\approx\; \frac{1}{N_j} \sum_{j=1}^{N_j} \frac{p(\theta_j^{(i)} \mid \Lambda)}{\pi(\theta_j^{(i)})}. \tag{39}$$

Since $Z^{(i)}$ does not depend on Λ, it cancels when forming posterior ratios on Λ, and the hyperparameter likelihood becomes

$$\mathcal{L}(\Lambda) \;\propto\; \prod_{i=1}^{N} \frac{1}{N_j} \sum_{j=1}^{N_j} \frac{p(\theta_j^{(i)} \mid \Lambda)}{\pi(\theta_j^{(i)})}. \tag{40}$$

The ratio $w_j^{(i)} = p(\theta_j^{(i)} \mid \Lambda)/\pi(\theta_j^{(i)})$ is the *importance-sampling weight*: it up-weights samples that are favoured by the population model and down-weights those that are not. Equation (40) is exact in the limit $N_j \to \infty$; in practice $N_j \sim$ 500–5000 samples per event is usually sufficient provided the population model is not too different from $\pi(\theta)$ (if it is, the weights become degenerate and effective sample size collapses).

Physical setup for this lab

Lab 5 uses synthetic binary black hole (BBH) data with primary mass m_1 drawn from a truncated power law $p(m_1 \mid \alpha) \propto m_1^{\alpha}$ on $[5, 80]\, M_\odot$ with true slope $\alpha_{\mathrm{true}} = -2.35$ (the Salpeter initial-mass-function value). Single-event inference assumes Gaussian measurement noise with $\sigma_m = 3\, M_\odot$ and a broad uniform prior $\pi(m_1) \sim \mathrm{Uniform}(2, 100)\, M_\odot$. Population inference is performed with `bilby.hyper.likelihood.HyperparameterLikelihood`, which implements Eq. (40) directly.

Structure of this lab

- **Exercise 5.1** — simulate a catalog of 25 BBH events and run single-event inference to produce one posterior per event.
- **Exercise 5.2** — run `HyperparameterLikelihood` over the catalog to infer α and compare against the truth.
- **Exercise 5.3** — study how the posterior on α narrows with catalog size and test sensitivity to the choice of single-event prior $\pi(\theta)$.

5.2 Exercise 5.1 — Simulating the Catalog

Population inference begins with single-event posteriors. This exercise builds a synthetic catalog of 25 binary black hole events by drawing true masses from a power-law distribution, adding Gaussian measurement noise, and running individual Dynesty analyses to produce one posterior per event. These posteriors are the raw input to the hierarchical likelihood in the next exercise.

5.2.1 Setup

The population model. The primary mass m_1 is drawn from a *truncated power law*

$$p(m_1 \mid \alpha) = \frac{(\alpha+1)\, m_1^{\alpha}}{m_{\mathrm{max}}^{\alpha+1} - m_{\mathrm{min}}^{\alpha+1}}, \qquad m_1 \in [m_{\mathrm{min}}, m_{\mathrm{max}}], \tag{41}$$

with $m_{\rm min} = 5\,M_\odot$, $m_{\rm max} = 80\,M_\odot$, and true slope $\alpha_{\rm true} = -2.35$. Both cutoffs are physically motivated:

- **Lower cutoff** $m_{\rm min} = 5\,M_\odot$: stellar remnants below $\sim$3–5 $M_\odot$ are neutron stars, not black holes. A lower cutoff removes the neutron-star population and focuses the model on the BBH mass range.
- **Upper cutoff** $m_{\rm max} = 80\,M_\odot$: very massive stars are expected to undergo pair-instability supernovae, which disrupt the star rather than leaving a black-hole remnant. Observations suggest a gap in the BH mass spectrum above $\sim$50–65 $M_\odot$; the value 80 $M_\odot$ is used here as a conservative upper boundary for the astrophysical population.

The denominator in Eq. (41) is the normalisation constant that ensures $\int_{m_{\rm min}}^{m_{\rm max}} p(m_1 \mid \alpha)\,{\rm d}m_1 = 1$. For $\alpha = -2.35$ the distribution is strongly decreasing: low-mass black holes are far more common than high-mass ones, so the catalog is dominated by events near $m_{\rm min}$.

Simulation procedure. A catalog of $N = 25$ BBH events is generated. For each event:

1. Draw the true mass $m_{1,\rm true} \sim p(m_1 \mid \alpha_{\rm true})$ using inverse-CDF sampling.
2. Simulate a noisy measurement $d_{\rm obs} = m_{1,\rm true} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_m^2)$ with $\sigma_m = 3\,M_\odot$.
3. Run Dynesty ($n_{\rm live} = 100$, single-ellipsoid bound, uniform sampling) with the Gaussian likelihood $\mathcal{L}(d_{\rm obs} \mid m_1) = \mathcal{N}(d_{\rm obs}; m_1, \sigma_m^2)$ and single-event prior $\pi(m_1) = {\rm Uniform}(2, 100)\,M_\odot$ to produce a posterior on m_1.

The single-event prior is deliberately *wider* than the population range $[5, 80]\,M_\odot$ and flat within it, so that it carries no information about the population slope. This is the correct choice: the prior π should be broad and uninformative; the population model $p(m_1 \mid \alpha)$ is then applied as an importance-sampling weight at the population-inference stage, not baked into the single-event analysis.

5.2.2 Results

Figure 19 shows the distribution of true injected masses compared with the population power law. The catalogue is dominated by low-mass events (small $|m_1|$) as expected for a steep negative power law.

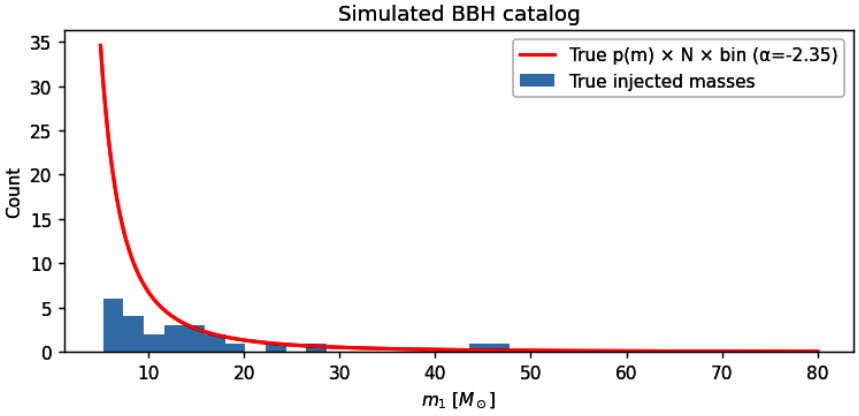


Figure 19. Simulated BBH catalog of 25 events. The histogram shows injected true masses; the red curve shows the true population $p(m_1|\alpha_{\rm true})$ scaled to the same normalisation. The catalogue is dominated by low-mass events consistent with $\alpha = -2.35$.

Figure 20 shows the individual 90% credible intervals (blue error bars) against the true masses (red crosses).

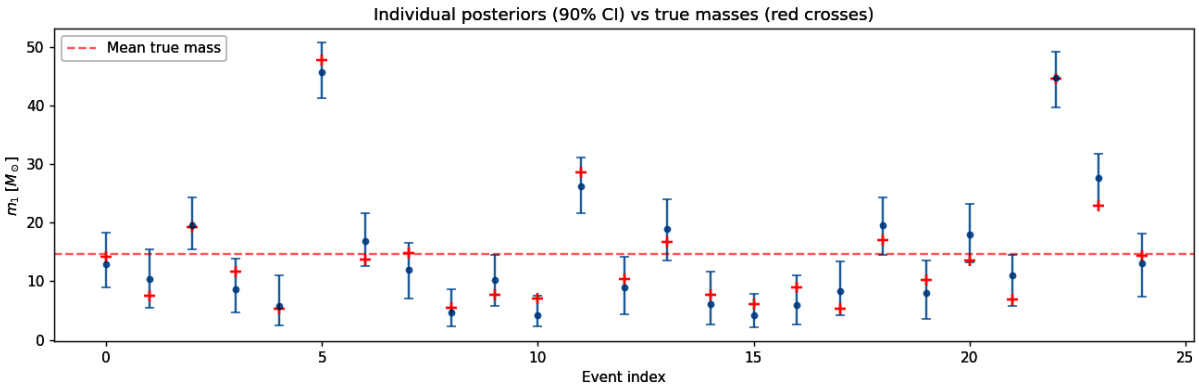


Figure 20. Single-event posteriors (90% CI, blue) vs. true masses (red crosses) for all 25 catalog events. Most posteriors are consistent with the true value; events with high true mass have wider intervals due to the prior cutoff at 100 $M_\odot$.

5.3 Exercise 5.2 — Population Inference

With 25 single-event posteriors in hand, we now ask: what power-law slope α best describes the distribution from which their masses were drawn? The `HyperparameterLikelihood` reweights each posterior via importance sampling to evaluate the population likelihood without re-running individual analyses, and Dynesty infers the marginal posterior on α.

5.3.1 Setup

The `HyperparameterLikelihood` (Eq. 40) is evaluated using the 25 single-event posteriors. The population model is a truncated power law (Section 5.1). The hyper-prior on α is Uniform$(-6, 0)$. Dynesty runs with $n_{\text{live}} = 150$.

5.3.2 Results

Table 14. Population inference results for α (25 events).

Quantity	Value
True α	-2.35
Recovered median	-1.908
90% CI lower	-2.478
90% CI upper	-1.465
$\ln \mathcal{L}$ at truth	25.41

Figure 21 shows the posterior on α and the implied family of mass distributions.

The truth $\alpha = -2.35$ falls within the 90% CI $[-2.48, -1.47]$. With only 25 events the statistical uncertainty is large ($\sigma_\alpha \approx 0.3$), but the posterior is unbiased.

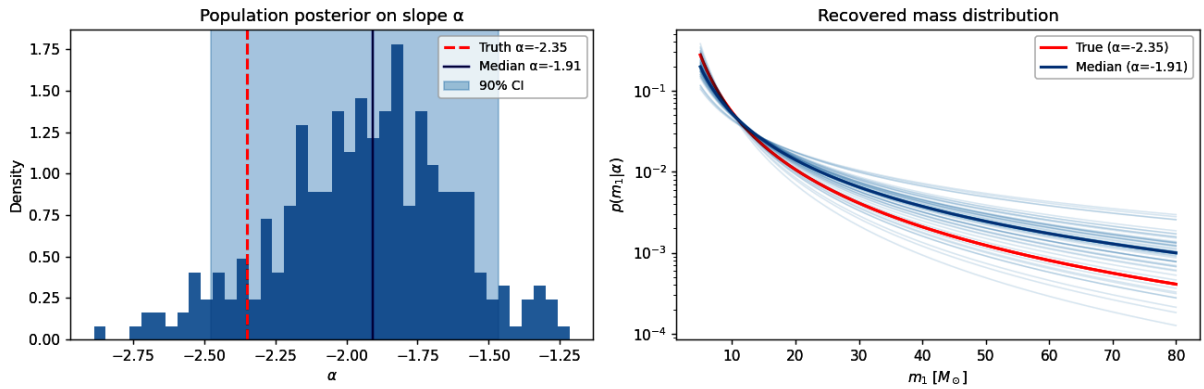


Figure 21. *Left*: Posterior on the power-law slope α. The red dashed line marks the truth ($\alpha = -2.35$); the navy line is the posterior median (-1.91); the blue band is the 90% CI. *Right*: Recovered mass distributions for 50 samples from the α posterior (light blue), the median model (navy), and the true distribution (red), shown on a log scale. The truth lies comfortably within the posterior envelope.

5.4 Exercise 5.3 — Catalog Size Scaling and Prior Sensitivity

Two practical questions matter for any real population analysis: how many events are needed to measure α to a given precision, and what happens if the single-event prior does not cover the full population? Part A studies the $1/\sqrt{N}$ scaling of the posterior width; Part B shows that truncating the single-event prior at $20\,M_\odot$ introduces a severe bias toward shallower slopes by suppressing the low-mass signal.

5.4.1 Part A: Scaling with N_{events}

Population inference was re-run on sub-catalogs of size $N \in \{5, 10, 25\}$.

Table 15. Posterior width on α as a function of catalog size. $\sigma(\alpha)$ is the 90% CI width divided by 2×1.645.

N	Median α	90% CI width	$\sigma(\alpha)$
5	-2.621	2.977	0.905
10	-2.021	1.660	0.504
25	-1.953	0.945	0.287

Figure 22 shows $\sigma(\alpha)$ versus N on a log–log scale together with the $1/\sqrt{N}$ prediction.

Discussion

Does the width scale as $1/\sqrt{N}$? The three measured $\sigma(\alpha)$ values are consistent with $1/\sqrt{N}$ scaling within statistical noise: from $N = 5$ to $N = 25$ (a factor of 5), the width decreases by a factor of $0.905/0.287 \approx 3.2$, compared with the predicted factor of $\sqrt{5} \approx 2.24$. The observed ratio is slightly larger, consistent with the fact that the $N = 5$ posterior is dominated by prior support (the prior is very uninformative for 5 events) rather than pure $1/\sqrt{N}$ likelihood narrowing.

Is there a noise floor? A floor in $\sigma(\alpha)$ would appear when the per-event measurement uncertainty (here $\sigma_m = 3\,M_\odot$) begins to dominate over the Poisson sampling variance. Heuristically, this occurs when $N \gg (R_{\text{pop}}/\sigma_m)^2$ where R_{pop} is the dynamic range of the population. For our

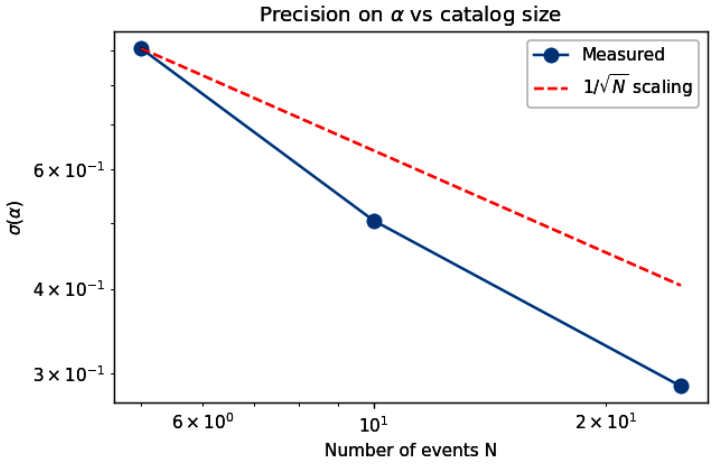


Figure 22. Posterior width $\sigma(\alpha)$ vs. catalog size N (blue circles) and the $1/\sqrt{N}$ scaling law (red dashed). With only three points the agreement is reasonable; larger N would be needed to see a floor due to per-event measurement noise.

setup this threshold is $N \gtrsim 10^3$, well beyond the 25 events used here, so no floor is visible in Figure 22.

5.4.2 Part B: Sensitivity to the Single-Event Prior

The single-event prior was changed to Uniform$(20, 100)\, M_\odot$ ("narrow prior"), cutting off events with $m_{1,\text{true}} < 20\, M_\odot$. Single-event posteriors were regenerated with this narrow prior, and population inference was re-run.

Table 16. Effect of single-event prior truncation on the recovered α (25 events).

Prior	Recovered α	True α
Correct (2–100 $M_\odot$)	-1.908	-2.35
Narrow (20–100 $M_\odot$)	-0.786	-2.35

Figure 23 compares the two population posteriors.

Discussion

Direction and magnitude of the bias. With the narrow prior cut at 20 $M_\odot$, the recovered slope is $\alpha \approx -0.79$, dramatically *shallower* than the truth (-2.35). This is a bias of $\approx +1.56$ in α.

Mechanism via importance sampling. The hyperparameter likelihood re-weights single-event posteriors by $p(m|\Lambda)/\pi(m)$. For events whose true mass $m_{1,\text{true}} < 20\, M_\odot$, the narrow prior $\pi(m) = 0$ for $m < 20$ forces the single-event posterior to have zero support below 20 $M_\odot$. These posteriors are therefore truncated: the importance-sampling weight $p(m|\Lambda)/\pi(m)$ is evaluated only above the prior cutoff, so the low-mass regime of the population model receives no likelihood support from these events. Since the power law is steepest (most probability) at low masses for $\alpha < -1$, hiding the low-mass events makes the data appear consistent with a much flatter distribution, driving α toward zero.

Practical implication. In real GW analyses the single-event prior must have *broader* support than the population model. If the search or PE prior cuts off the mass distribution at some m_{min}

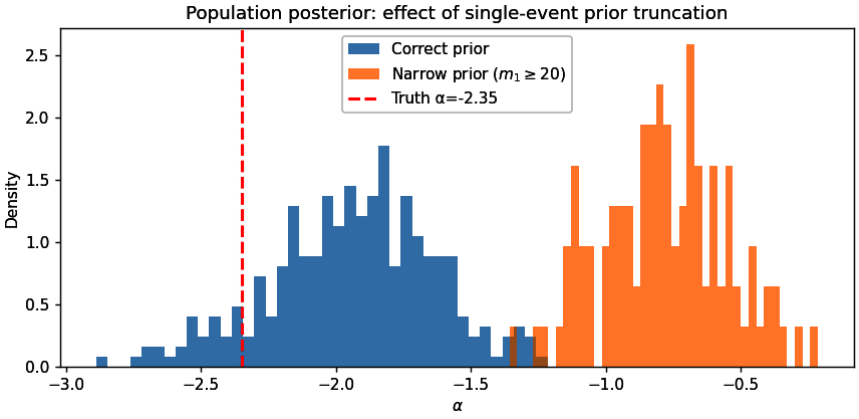


Figure 23. Population posterior on α for the correct single-event prior (blue) and the narrow prior truncated at 20 $M_\odot$ (orange). The red dashed line is the truth. The narrow prior produces a strongly biased, shallower slope ($\alpha \approx -0.79$ vs. -2.35).

or $m_{\max}$, any population model that places probability outside that range cannot be tested: the importance weights diverge or vanish and the hyperparameter likelihood is silently biased. This is a critical calibration requirement: single-event inference must be performed with a prior that is at least as broad as the widest population model that will be considered in the population analysis.

Note

Key takeaways from Lab 5. (1) The `HyperparameterLikelihood` reweights individual event posteriors via importance sampling to evaluate the population likelihood — no re-running of single-event inference is required. (2) The statistical uncertainty on population hyperparameters scales as $1/\sqrt{N_{\text{events}}}$; $O(100)$ events are typically needed to measure α to ~ 0.1 precision. (3) The single-event prior must be broader than the population model. Truncation of the single-event prior biases the population posterior toward shallower (less negative) slopes by suppressing the low-mass signal from the data.

5.5 Further Exercises

The following problems invite you to explore the concepts from this lab further. They are optional but recommended.

1. Change the true slope to $\alpha = -1$ (equal probability per decade in mass). How does the posterior on α change? Is the recovery equally accurate?
2. Reduce the catalog to $N = 5$ events and increase it to $N = 100$. Plot the posterior width $\sigma(\alpha)$ vs. N and check whether it scales as $1/\sqrt{N}$.
3. Increase the measurement noise to $\sigma_m = 10\,M_\odot$. Does the importance-sampling approximation degrade? Monitor the effective sample size $n_{\text{eff}} = (\sum w_j)^2 / \sum w_j^2$ per event.
4. Replace the single-event prior with $\pi(m_1) = \text{Uniform}(5, 80)\,M_\odot$ (matching the population range exactly). Does the recovered α change? What does this tell you about the sensitivity of the result to the prior choice?

5. Model the mass distribution as a broken power law with two slopes α_1 (below a break mass m_{break}) and α_2 (above). Extend `power_law_pdf` accordingly and run `HyperparameterLikelihood` with two free hyperparameters.

Need help with these exercises or have other questions? Submit a ticket through the ***ACME support system*** *at* `https://support.acme-astro.eu/` *and one of the tutorial assistants will follow up.*