Alexandre Toubiana

Introduction to model selection and hierarchical inference

PhD course, Milano-Bicocca, 15/05/2026

Plan of the lecture

- Bayesian model selection
- Hierarchical inference
- Examples of astrophysical inference
- Posterior predictive checking

Bayesian hypothesis testing

- The denominator in Bayes' Theorem

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}$$

is the **Bayesian evidence**

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta$$

Bayesian hypothesis testing

- The denominator in Bayes' Theorem

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}$$

is the **Bayesian evidence**

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta$$

dependence on the model

Bayesian hypothesis testing

- The denominator in Bayes' Theorem

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}$$

is the **Bayesian evidence**

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta$$

dependence on the model

- **Probability of observed data under given model**
  Can be used for model selection
  Posterior odds ratio:

$$\mathcal{O}_{12} = \frac{p(x|M_1)p(M_1)}{p(x|M_2)p(M_2)}$$

Bayesian hypothesis testing

- The denominator in Bayes' Theorem

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}$$

is the **Bayesian evidence**

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta$$

dependence on the model

- **Probability of observed data under given model**
  Can be used for model selection
  Posterior odds ratio:

$$\mathcal{O}_{12} = \frac{p(x|M_1)}{p(x|M_2)}\frac{p(M_1)}{p(M_2)}$$

Bayes Factor **Prior odds ratio.**

Gaussian example

- Gaussian likelihood:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

- We want to test the models

$$\mathrm{M}_1 : \theta = 0, \;\; \mathrm{M}_2 : \theta \in [-\Delta, \Delta]$$

Gaussian example

- Gaussian likelihood:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

- We want to test the models

$$\mathrm{M}_1 : \theta = 0, \;\; \mathrm{M}_2 : \theta \in [-\Delta, \Delta]$$

- Bayes' factor (assuming $\Delta \gg \sigma$)

$$B_{12} = \frac{2\Delta}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right].$$

Gaussian example

- Gaussian likelihood:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

- We want to test the models

$$\mathrm{M}_1 : \theta = 0,\ \ \mathrm{M}_2 : \theta \in [-\Delta, \Delta]$$

- Bayes' factor (assuming $\Delta \gg \sigma$)

Prior volume penalty

"Occam's razor"

$$B_{12} = \frac{2\Delta}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right].$$

data information

Gaussian example

- Gaussian likelihood:

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

- We want to test the models

$$\mathrm{M}_1 : \theta = 0, \;\; \mathrm{M}_2 : \theta \in [-\Delta, \Delta]$$

- Bayes' factor (assuming $\Delta \gg \sigma$)

Prior volume penalty

"Occam's razor"

$$B_{12} = \frac{2\Delta}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right].$$

data information

- After $N_{\rm obs}$ observations:

$$B_{12}^{(N_{\rm obs})} \approx \frac{2\Delta\sqrt{N_{\rm obs}}}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{N_{\rm obs}\bar{x}^2}{2\sigma^2}\right).$$

Interpreting Bayes' factors

- Interpretation is somewhat arbitrary
- Kass and Rafferty (1995) suggested the following scale:

Bayes Factor	Interpretation
< 3	No evidence of M_1 over M_2
> 3	Positive evidence for M_1
> 20	Strong evidence for M_1
> 150	Very strong evidence for M_1

Correspond to "p-values" of 0.25, 0.05, 0.007

- In practice, posterior odds ratios can also be used as a test statistic, with significance and power computed via simulation in the usual (frequentist) way.

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

Inefficient!

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N}\sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

Inefficient!

- **Harmonic mean estimator:**

$$\frac{1}{p(x|M)} = \int p(\theta|M)\frac{1}{p(x|M)}\;\mathrm{d}\theta$$

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

  Inefficient!

- **Harmonic mean estimator:**

$$\frac{1}{p(x|M)} = \int p(\theta|M)\frac{1}{p(x|M)}\ \mathrm{d}\theta = \int \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}\frac{1}{p(x|\theta, M)}\ \mathrm{d}\theta$$

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

Inefficient!

- **Harmonic mean estimator:**

$$\frac{1}{p(x|M)} = \int p(\theta|M)\frac{1}{p(x|M)}\,\mathrm{d}\theta = \int \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}\frac{1}{p(x|\theta, M)}\,\mathrm{d}\theta$$

$= p(\theta|x, M)$

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

Inefficient!

- **Harmonic mean estimator:**

$$\frac{1}{p(x|M)} = \int p(\theta|M)\frac{1}{p(x|M)}\,\mathrm{d}\theta = \int \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}\frac{1}{p(x|\theta, M)}\,\mathrm{d}\theta$$

$$\simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|x,M)} \frac{1}{p(x|\theta_i, M)}$$

$= p(\theta|x, M)$

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

Inefficient!

- **Harmonic mean estimator:**

$$\frac{1}{p(x|M)} = \int p(\theta|M)\frac{1}{p(x|M)}\,\mathrm{d}\theta = \int \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)} \frac{1}{p(x|\theta, M)}\,\mathrm{d}\theta$$

$$\simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|x, M)} \frac{1}{p(x|\theta_i, M)}$$

$$= p(\theta|x, M)$$

Few samples in low likelihood regions → unstable estimator

Computing Bayes' factors

- **Direct integration**:

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta \simeq \frac{1}{N}\sum_{\theta_i \sim p(\theta|M)} p(x|\theta_i, M)$$

  Inefficient!

- **Harmonic mean estimator:**

$$\frac{1}{p(x|M)} = \int p(\theta|M)\frac{1}{p(x|M)}\,\mathrm{d}\theta = \int \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}\frac{1}{p(x|\theta, M)}\,\mathrm{d}\theta$$

$$\simeq \frac{1}{N}\sum_{\theta_i \sim p(\theta|x,M)} \frac{1}{p(x|\theta_i, M)}$$

$$= p(\theta|x, M)$$

  Few samples in low likelihood regions → unstable estimator

- Other techniques:
  - **thermodynamic integration**
  - **nested sampling**
  - **reversible-jump MCMC**

Hierarchical Models

based on Mandel, Farr, Gair MNRAS 2018, Taylor, Gerosa, PRD 2018

Hierarchical Models

- Individual events are drawn from an underlying distribution, the population prior:

$$p(\theta|\Lambda)$$

Hierarchical Models

- Individual events are drawn from an underlying distribution, the population prior:

$$p(\theta|\Lambda)$$

- Hierarchical likelihood:

$$p(x|\Lambda) = \int p(x|\theta)p(\theta|\Lambda)\mathrm{d}\theta$$

Hierarchical Models

- Individual events are drawn from an underlying distribution, the population prior:

$$p(\theta|\Lambda)$$

- Hierarchical likelihood:

$$p(x|\Lambda) = \int p(x|\theta)p(\theta|\Lambda)\mathrm{d}\theta$$

- If independent events:

$$p(\{x\}|\Lambda) = \prod_i p(x_i|\Lambda)$$

Hierarchical Models

- Individual events are drawn from an underlying distribution, the population prior:

$$p(\theta|\Lambda)$$

- Hierarchical likelihood:

$$p(x|\Lambda) = \int p(x|\theta)p(\theta|\Lambda)\mathrm{d}\theta$$

- If independent events:

$$p(\{x\}|\Lambda) = \prod_i p(x_i|\Lambda)$$

- Bayes' theorem:

$$p(\Lambda|\{x\}) = \frac{p(\{x\}|\Lambda)\pi(\Lambda)}{p(\{x\}|\mathrm{pop\ model})}$$

Hierarchical Models

- Individual events are drawn from an underlying distribution, the population prior:

$$p(\theta|\Lambda)$$

- Hierarchical likelihood:

$$p(x|\Lambda) = \int p(x|\theta)p(\theta|\Lambda)\mathrm{d}\theta$$

- If independent events:

$$p(\{x\}|\Lambda) = \prod_i p(x_i|\Lambda)$$

- Bayes' theorem:

$$p(\Lambda|\{x\}) = \frac{p(\{x\}|\Lambda)\pi(\Lambda)}{p(\{x\}|\mathrm{pop\ model})}$$

- Likelihood ratio:

$$\frac{p(\{x\}|\Lambda_1)}{p(\{x\}|\Lambda_2)} = \frac{\int p(x|\theta)p(\theta|\Lambda_1)\mathrm{d}\theta}{\int p(x|\theta)p(\theta|\Lambda_2)\mathrm{d}\theta} = \mathcal{B}_{\Lambda_1\Lambda_2}$$

Hierarchical Models

- Individual events are drawn from an underlying distribution, the population prior:

$$p(\theta|\Lambda)$$

- Hierarchical likelihood:

$$p(x|\Lambda) = \int p(x|\theta)p(\theta|\Lambda)\mathrm{d}\theta$$

- If independent events:

$$p(\{x\}|\Lambda) = \prod_i p(x_i|\Lambda)$$

- Bayes' theorem:

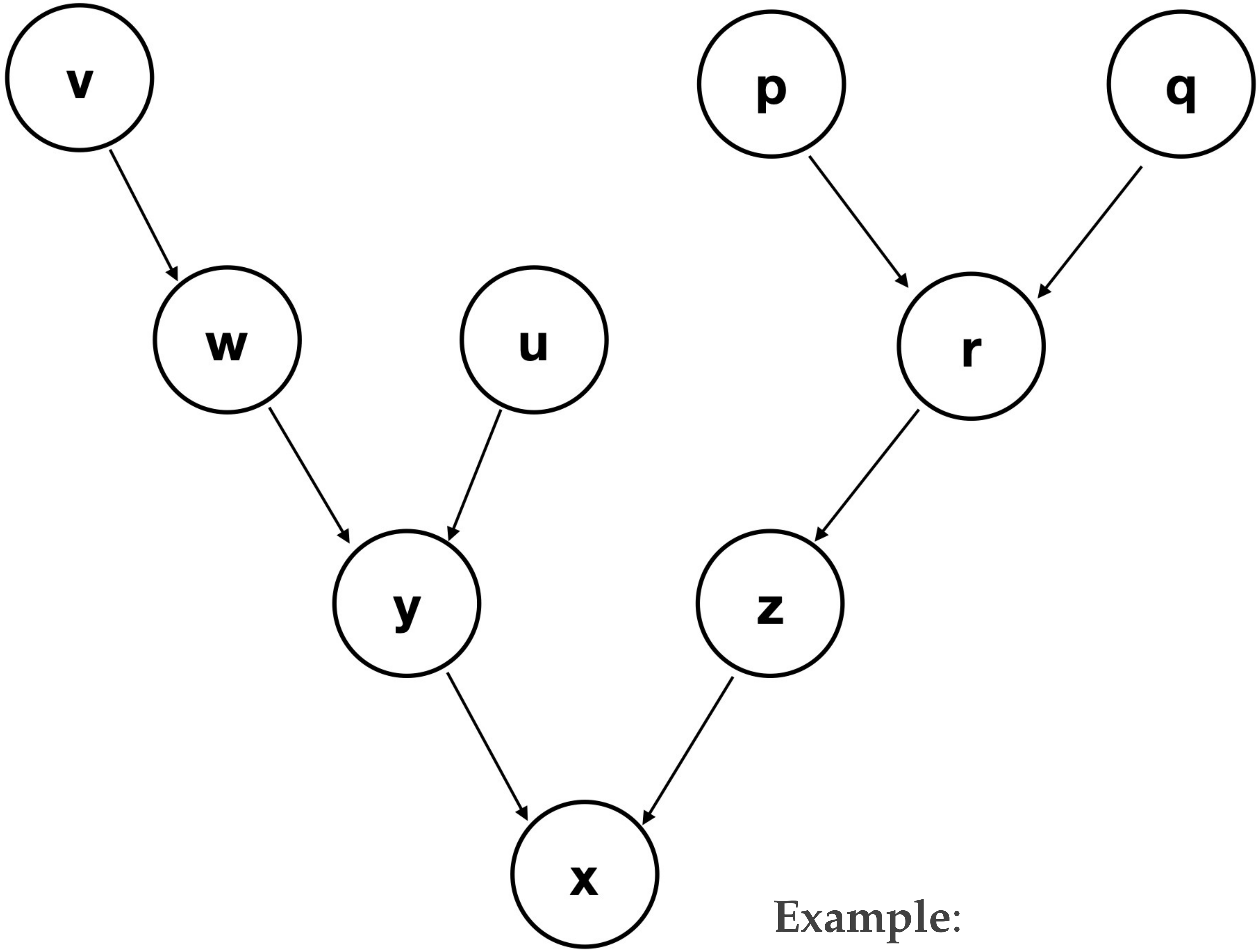
$$p(\Lambda|\{x\}) = \frac{p(\{x\}|\Lambda)\pi(\Lambda)}{p(\{x\}|\text{pop model})}$$

- Likelihood ratio:

$$\frac{p(\{x\}|\Lambda_1)}{p(\{x\}|\Lambda_2)} = \frac{\int p(x|\theta)p(\theta|\Lambda_1)\mathrm{d}\theta}{\int p(x|\theta)p(\theta|\Lambda_2)\mathrm{d}\theta} = \mathcal{B}_{\Lambda_1\Lambda_2}$$

"Continuous model selection"

Graphical Model



Example:

p(p, q, r, s, t, u, v, w, x, y, z) =

p(x | y, z) p(y | u, w) p(z | r) p(w | v) p(r | p, q) p(v) p(u) p(p) p(q)

Gaussian example

- Population prior:

$$p(\theta|\Lambda) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{pop}}}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2_{\mathrm{pop}}}\right] \qquad \Lambda = (\mu, \sigma^2_{\mathrm{pop}})$$

- Single-event likelihood:

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{PE}}}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2_{\mathrm{PE}}}\right]$$

Gaussian example

- Population prior:

$$p(\theta|\Lambda) = \frac{1}{\sqrt{2\pi\sigma_{\rm pop}^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma_{\rm pop}^2}\right] \qquad \Lambda = (\mu, \sigma_{\rm pop}^2)$$

- Single-event likelihood:

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma_{\rm PE}^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma_{\rm PE}^2}\right]$$

- Population likelihood:

$$p(x|\Lambda) = \frac{1}{\sqrt{2\pi(\sigma_{\rm PE}^2+\sigma_{\rm pop}^2)}} \exp\left[-\frac{(x-\mu)^2}{2(\sigma_{\rm PE}^2+\sigma_{\rm pop}^2)}\right]$$

Gaussian example

- Combining events:

$$p(\{x\}|\Lambda) = \frac{1}{(2\pi(\sigma^2_{\rm PE} + \sigma^2_{\rm pop}))^{N_{\rm obs}/2}} \exp\left[-\frac{N_{\rm obs}(\bar{x}-\mu)^2}{2(\sigma^2_{\rm PE} + \sigma^2_{\rm pop})}\right] \exp\left[-\frac{N_{\rm obs}\bar{s}^2}{2(\sigma^2_{\rm PE} + \sigma^2_{\rm pop})}\right]$$

$$\bar{x} = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} x_i \qquad \bar{s}^2 = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} (x_i - \bar{x})^2$$

Gaussian example

- Combining events:

$$p(\{x\}|\Lambda) = \frac{1}{(2\pi(\sigma_{\rm PE}^2 + \sigma_{\rm pop}^2))^{N_{\rm obs}/2}} \exp\left[-\frac{N_{\rm obs}(\bar{x}-\mu)^2}{2(\sigma_{\rm PE}^2 + \sigma_{\rm pop}^2)}\right] \exp\left[-\frac{N_{\rm obs}\bar{s}^2}{2(\sigma_{\rm PE}^2 + \sigma_{\rm pop}^2)}\right]$$

$$\bar{x} = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} x_i \qquad \bar{s}^2 = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} (x_i - \bar{x})^2$$

- Maximum likelihood: $\mu = \bar{x} \qquad \sigma_{\rm pop}^2 = \bar{s}^2 - \sigma_{\rm PE}^2$

Gaussian example

- Combining events:

$$p(\{x\}|\Lambda) = \frac{1}{(2\pi(\sigma_{\rm PE}^2 + \sigma_{\rm pop}^2))^{N_{\rm obs}/2}} \exp\left[-\frac{N_{\rm obs}(\bar{x}-\mu)^2}{2(\sigma_{\rm PE}^2+\sigma_{\rm pop}^2)}\right] \exp\left[-\frac{N_{\rm obs}\bar{s}^2}{2(\sigma_{\rm PE}^2+\sigma_{\rm pop}^2)}\right]$$

$$\bar{x} = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} x_i \qquad \bar{s}^2 = \frac{1}{N_{\rm obs}} \sum_{i=1}^{N_{\rm obs}} (x_i - \bar{x})^2$$

- Maximum likelihood: $\mu = \bar{x} \qquad \sigma_{\rm pop}^2 = \bar{s}^2 - \sigma_{\rm PE}^2$

- Expanding around maximum in the large $N_{\rm obs}$ limit ("Fisher approximation")
  Assuming broad priors:

$$p(\mu, \sigma_{\rm pop}^2|\{x\}) = \frac{N_{\rm obs}}{2\sqrt{2\pi}\,\bar{s}^3} \exp\left[-\frac{N_{\rm obs}}{2}\left(\frac{(\mu-\bar{x})^2}{\bar{s}^2} + \frac{(\sigma_{\rm pop}^2 - (\bar{s}^2 - \sigma_{\rm PE}^2))^2}{2\,\bar{s}^4}\right)\right]$$

Include selection effects “by hand”

- Analysis include only **detected** events, i.e such that $F(x) > \text{threshold}$

Include selection effects "by hand"

- Analysis include only **detected** events, i.e such that $F(x) > \text{threshold}$
- Renormalise likelihood to account for it

$$p(x|\Lambda, \text{det}) = \frac{p(x|\Lambda)}{\xi(\Lambda)}$$

$$\xi(\Lambda) = \int p(\text{det}|\theta)\, p(\theta|\Lambda)\, \text{d}\theta \qquad p(\text{det}|\theta) = \int_{F(x)>\text{threshold}} p(x|\theta)\text{d}x.$$

Include selection effects "by hand"

- Analysis include only **detected** events, i.e such that $F(x) > \text{threshold}$
- Renormalise likelihood to account for it

$$p(x|\Lambda, \text{det}) = \frac{p(x|\Lambda)}{\xi(\Lambda)}$$

$$\xi(\Lambda) = \int p(\text{det}|\theta)\, p(\theta|\Lambda)\, \text{d}\theta \qquad p(\text{det}|\theta) = \int_{F(x) > \text{threshold}} p(x|\theta) \text{d}x.$$

- Combining events

$$p(\{x\}|\Lambda) = \frac{1}{\xi(\Lambda)^{N_{\text{obs}}}} \prod_i \int p(x_i|\theta) p(\theta|\Lambda) \text{d}\theta$$

Include selection effects "by hand"

- Analysis include only **detected** events, i.e such that $F(x) > \text{threshold}$
- Renormalise likelihood to account for it

$$p(x|\Lambda, \text{det}) = \frac{p(x|\Lambda)}{\xi(\Lambda)}$$

$$\xi(\Lambda) = \int p(\text{det}|\theta)\, p(\theta|\Lambda)\, \text{d}\theta \qquad p(\text{det}|\theta) = \int_{F(x)>\text{threshold}} p(x|\theta)\text{d}x.$$

- Combining events

$$p(\{x\}|\Lambda) = \frac{1}{\xi(\Lambda)^{N_{\text{obs}}}} \prod_i \int p(x_i|\theta)p(\theta|\Lambda)\text{d}\theta$$

- Including rates

$$N_{\text{det}}(\Lambda) = \xi(\Lambda)N(\Lambda)$$

Include selection effects "by hand"

- Analysis include only **detected** events, i.e such that $F(x) > \text{threshold}$
- Renormalise likelihood to account for it

$$p(x|\Lambda, \text{det}) = \frac{p(x|\Lambda)}{\xi(\Lambda)}$$

$$\xi(\Lambda) = \int p(\text{det}|\theta)\, p(\theta|\Lambda)\, \text{d}\theta \qquad p(\text{det}|\theta) = \int_{F(x)>\text{threshold}} p(x|\theta)\text{d}x.$$

- Combining events

$$p(\{x\}|\Lambda) = \frac{1}{\xi(\Lambda)^{N_{\text{obs}}}} \prod_i \int p(x_i|\theta)p(\theta|\Lambda)\text{d}\theta$$

- Including rates

$$N_{\text{det}}(\Lambda) = \xi(\Lambda)N(\Lambda)$$

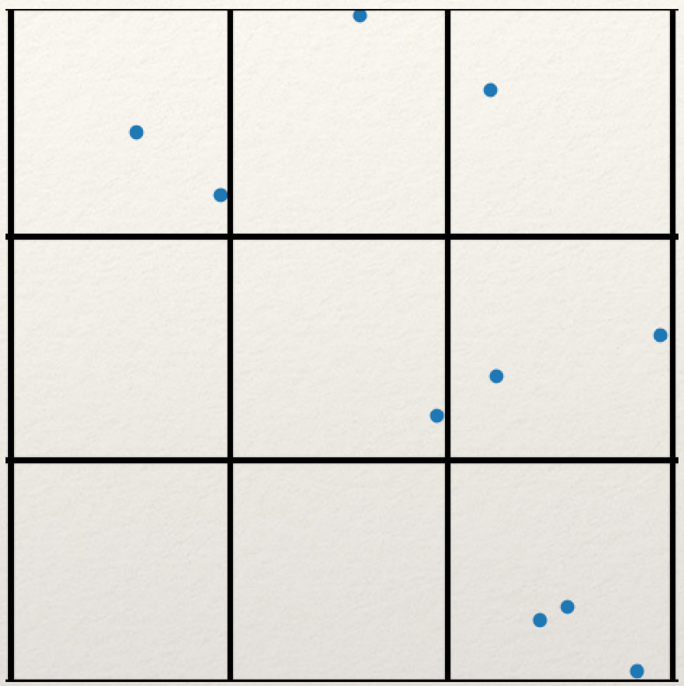
$$p(\{x\}|\Lambda) = \frac{N_{\text{det}}(\Lambda)^{N_{\text{obs}}}}{\xi(\Lambda)^{N_{\text{obs}}}} e^{-N_{\text{det}(\Lambda)}} \prod_i \int p(x_i|\theta)p(\theta|\Lambda)\text{d}\theta$$

$$= N(\Lambda)^{N_{\text{obs}}} e^{-N_{\text{det}(\Lambda)}} \prod_i \int p(x_i|\theta)p(\theta|\Lambda)\text{d}\theta$$

More general approach

Inhomogeneous Poisson process for a given binning scheme:

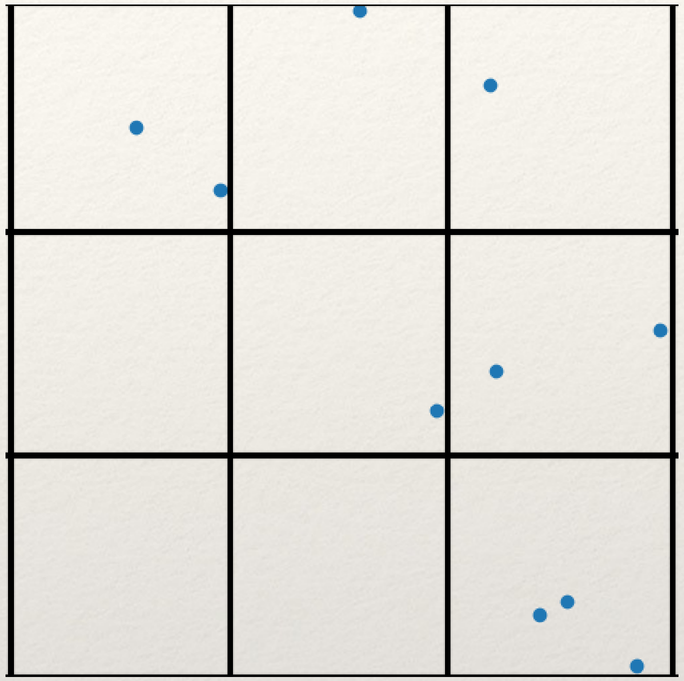
$$p(\{\theta\}|\Lambda) = \prod_{k=1}^{N_{\rm bins}} \frac{N_k(\Lambda)^{n_k} \exp[-N_k(\Lambda)]}{n_k!}$$



More general approach

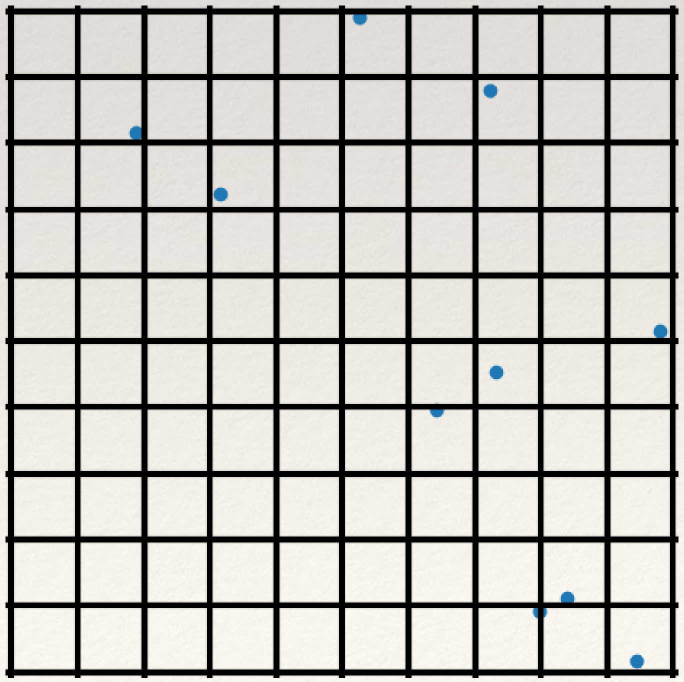
Inhomogeneous Poisson process for a given binning scheme:

$$p(\{\theta\}|\Lambda) = \prod_{k=1}^{N_{\rm bins}} \frac{N_k(\Lambda)^{n_k} \exp[-N_k(\Lambda)]}{n_k!}$$



Make bins infinitesimally small

$$n_k \to 0 \text{ or } 1$$



More general approach

Inhomogeneous Poisson process for a given binning scheme:

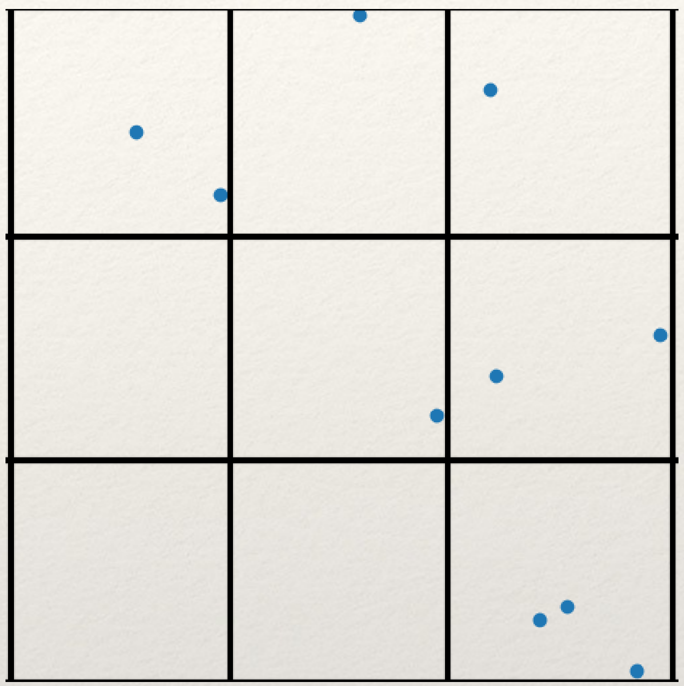
$$p(\{\theta\}|\Lambda) = \prod_{k=1}^{N_{\text{bins}}} \frac{N_k(\Lambda)^{n_k} \exp[-N_k(\Lambda)]}{n_k!}$$

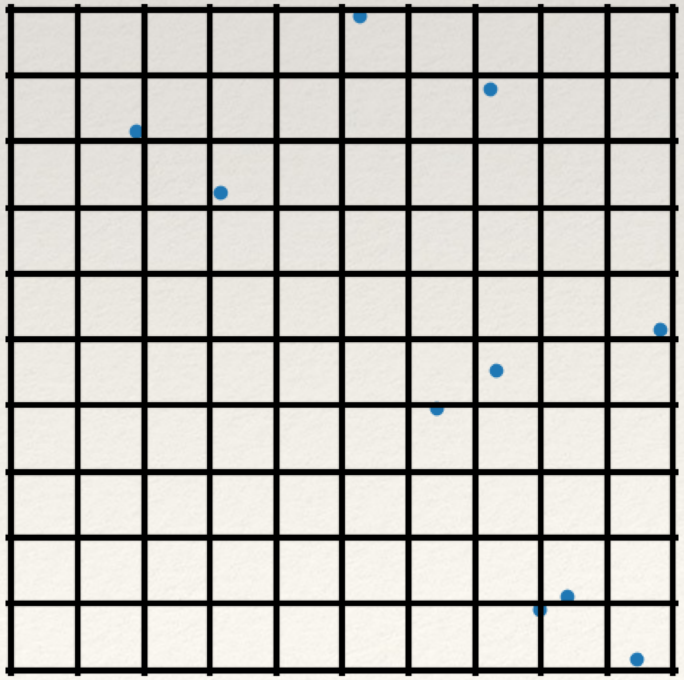
Make bins infinitesimally small

$$n_k \to 0 \text{ or } 1$$

$$p(\{\theta\} \mid \Lambda) = \exp[-N(\Lambda)] \prod_{k=1}^{N_{\text{obs}}} \frac{\mathrm{d}N}{\mathrm{d}\theta_k}(\Lambda)$$

$$N(\Lambda) = \int \frac{\mathrm{d}N}{\mathrm{d}\theta}(\Lambda)\mathrm{d}\theta \qquad \frac{\mathrm{d}N}{\mathrm{d}\theta} = N(\Lambda)p(\theta|\Lambda)$$





More general approach

- Joint likelihood for all data/parameter pairs, including both **detected** events (indexed by *i*) and **undetected** events (indexed by *j*)

$$p(\{x_i\},\{\theta_i\},\{x_j\},\{\theta_j\}|\Lambda) = \left[\prod_i^{N_{\rm obs}} p(x_i|\theta_i)\frac{\mathrm{d}N}{\mathrm{d}\theta_i}(\Lambda)\right]\left[\prod_j^{N_{\rm nobs}} p(x_j|\theta_j)\frac{\mathrm{d}N}{\mathrm{d}\theta_j}(\Lambda)\right]\exp[-N(\Lambda)]$$

More general approach

- Joint likelihood for all data/parameter pairs, including both **detected** events (indexed by *i*) and **undetected** events (indexed by *j*)

$$p(\{x_i\},\{\theta_i\},\{x_j\},\{\theta_j\}|\Lambda) = \left[\prod_i^{N_{\rm obs}} p(x_i|\theta_i)\frac{\mathrm{d}N}{\mathrm{d}\theta_i}(\Lambda)\right]\left[\prod_j^{N_{\rm nobs}} p(x_j|\theta_j)\frac{\mathrm{d}N}{\mathrm{d}\theta_j}(\Lambda)\right]\exp[-N(\Lambda)]$$

- Marginalising over the unobserved data:

$$p(\{x_i\},\{\theta_i\}|\Lambda) = \left[\prod_{i=1}^{N_{\rm obs}} p(x_i \mid \theta_i)\frac{\mathrm{d}N}{\mathrm{d}\theta_i}(\Lambda)\right] e^{-N(\Lambda)}\frac{N_{\rm ndet}(\Lambda)^{N_{\rm nobs}}}{N_{\rm nobs}!}.$$

$$N_{\rm ndet}(\Lambda) = \int_{F(x)<\text{threshold}}\int p(x|\theta)\,\frac{\mathrm{d}N}{\mathrm{d}\theta}(\Lambda)\,\mathrm{d}\theta\,\mathrm{d}d = N(\Lambda)(1-\xi(\Lambda))$$

More general approach

- Marginalising over the unknown number of unobserved events:

$$p(\{x_i\}, \{\theta_i\}|\Lambda) = e^{-N(\Lambda)} e^{N_{\rm ndet}(\Lambda)} \prod_{i=1}^{N_{\rm obs}} p(x_i|\theta_i) \frac{{\rm d}N}{{\rm d}\theta_i}(\Lambda)$$

More general approach

- Marginalising over the unknown number of unobserved events:

$$p(\{x_i\}, \{\theta_i\}|\Lambda) = e^{-N(\Lambda)} e^{N_{\rm ndet}(\Lambda)} \prod_{i=1}^{N_{\rm obs}} p(x_i|\theta_i) \frac{\mathrm{d}N}{\mathrm{d}\theta_i}(\Lambda)$$

- Marginalising over the parameters of observed events:

$$p(\{x\}|\Lambda) = e^{-N_{\rm det}(\Lambda)} \prod_{i=1}^{N_{\rm obs}} \int p(x_i|\theta) \frac{\mathrm{d}N}{\mathrm{d}\theta}(\Lambda)\, \mathrm{d}\theta$$

More general approach

- Marginalising over the unknown number of unobserved events:

$$p(\{x_i\},\{\theta_i\}|\Lambda) = e^{-N(\Lambda)} e^{N_{\mathrm{ndet}}(\Lambda)} \prod_{i=1}^{N_{\mathrm{obs}}} p(x_i|\theta_i) \frac{\mathrm{d}N}{\mathrm{d}\theta_i}(\Lambda)$$

- Marginalising over the parameters of observed events:

$$p(\{x\}|\Lambda) = e^{-N_{\mathrm{det}}(\Lambda)} \prod_{i=1}^{N_{\mathrm{obs}}} \int p(x_i|\theta) \frac{\mathrm{d}N}{\mathrm{d}\theta}(\Lambda)\,\mathrm{d}\theta$$

The selection function comes from marginalising over datasets with no detected events (and the parameters of undetected events)

Case of backgrounds

- Effectively, one dataset and no selection

$$p(x|\Lambda) = \int p(x|\Sigma)p(\Sigma|\Lambda)\mathrm{d}\theta$$

With:

$$p(\Sigma|\Lambda) = \sum_{n=0}^{+\infty} \int \mathrm{d}\{\theta\}_n \; p(\Sigma|\{\theta\}_n, n) \prod_{j=1}^{n} p(\theta_j|\Lambda) \frac{N(\Lambda)^n}{n!} e^{-n}$$

Ex: $p(\Sigma|\{\theta\}_n, n) = \delta(\Sigma^2 - \sum_{i=1}^{n} |h(\theta_i)|^2)$

Case of backgrounds

- Effectively, one dataset and no selection

$$p(x|\Lambda) = \int p(x|\Sigma)p(\Sigma|\Lambda)\mathrm{d}\theta$$

With:

$$p(\Sigma|\Lambda) = \sum_{n=0}^{+\infty} \int \mathrm{d}\{\theta\}_n \; p(\Sigma|\{\theta\}_n, n) \prod_{j=1}^{n} p(\theta_j|\Lambda) \frac{N(\Lambda)^n}{n!} e^{-n}$$

Ex: $p(\Sigma|\{\theta\}_n, n) = \delta(\Sigma^2 - \sum_{i=1}^{n} |h(\theta_i)|^2)$

- **Background + resolved**:
  - Callister et al. APJL 2020: multiply likelihoods, okay for current LVK
  - Toubiana, Gair 2601.04168: general case, including LISA Galactic binaries

Implementation

- Use Bayes theorem for each event:

$$p(\{x\}|\Lambda) = e^{-N_{\rm det}(\Lambda)} \left(\prod_{i=1}^{N_{\rm obs}} p(x_i)\right) \left(\prod_{i=1}^{N_{\rm obs}} \int \frac{p(\theta|x_i)}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta\right)$$

Implementation

- Use Bayes theorem for each event:

$$p(\{x\}|\Lambda) = e^{-N_{\rm det}(\Lambda)} \left(\prod_{i=1}^{N_{\rm obs}} p(x_i)\right) \left(\prod_{i=1}^{N_{\rm obs}} \int \frac{p(\theta|x_i)}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta\right)$$

- Monte-Carlo sampling

$$p(\{x\}|\Lambda) \propto e^{-N_{\rm det}(\Lambda)} \prod_{i=1}^{N_{\rm obs}} \frac{1}{N_s} \sum_{\theta\sim p(\theta|x_i)} \frac{1}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta$$

Implementation

- Use Bayes theorem for each event:

$$p(\{x\}|\Lambda) = e^{-N_{\mathrm{det}}(\Lambda)} \left(\prod_{i=1}^{N_{\mathrm{obs}}} p(x_i) \right) \left(\prod_{i=1}^{N_{\mathrm{obs}}} \int \frac{p(\theta|x_i)}{\pi_{\mathrm{PE}}(\theta)} \frac{\mathrm{d}N}{\mathrm{d}\theta}(\Lambda)\mathrm{d}\theta \right)$$

- Monte-Carlo sampling

$$p(\{x\}|\Lambda) \propto e^{-N_{\mathrm{det}}(\Lambda)} \prod_{i=1}^{N_{\mathrm{obs}}} \frac{1}{N_s} \sum_{\theta \sim p(\theta|x_i)} \frac{1}{\pi_{\mathrm{PE}}(\theta)} \frac{\mathrm{d}N}{\mathrm{d}\theta}(\Lambda)\mathrm{d}\theta$$

Prior reweighting

Implementation

- Use Bayes theorem for each event:

$$p(\{x\}|\Lambda) = e^{-N_{\rm det}(\Lambda)} \left(\prod_{i=1}^{N_{\rm obs}} p(x_i) \right) \left(\prod_{i=1}^{N_{\rm obs}} \int \frac{p(\theta|x_i)}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta \right)$$

- Monte-Carlo sampling

$$p(\{x\}|\Lambda) \propto e^{-N_{\rm det}(\Lambda)} \prod_{i=1}^{N_{\rm obs}} \frac{1}{N_s} \sum_{\theta \sim p(\theta|x_i)} \frac{1}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta$$

Prior reweighting

- Similarly for selection function:

$$N_{\rm det}(\Lambda) = \int_{F(x)>{\rm threshold}} \int p(x|\theta) \frac{p(\theta|\Lambda)}{\pi_{\rm inj}(\theta)} \pi_{\rm inj}(\theta)\ {\rm d}\theta {\rm d}x$$

$$N_{\rm det}(\Lambda) \simeq \frac{1}{N_{\rm inj}} \sum_{\theta \sim \pi_{\rm inj}(\theta),{\rm det}} \frac{1}{\pi_{\rm inj}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda)$$

Implementation

- Use Bayes theorem for each event:

$$p(\{x\}|\Lambda) = e^{-N_{\rm det}(\Lambda)} \left(\prod_{i=1}^{N_{\rm obs}} p(x_i)\right) \left(\prod_{i=1}^{N_{\rm obs}} \int \frac{p(\theta|x_i)}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta\right)$$

- Monte-Carlo sampling

$$p(\{x\}|\Lambda) \propto e^{-N_{\rm det}(\Lambda)} \prod_{i=1}^{N_{\rm obs}} \frac{1}{N_s} \sum_{\theta\sim p(\theta|x_i)} \frac{1}{\pi_{\rm PE}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda){\rm d}\theta$$

Prior reweighting

- Similarly for selection function:

$$N_{\rm det}(\Lambda) = \int_{F(x)>{\rm threshold}} \int p(x|\theta) \frac{p(\theta|\Lambda)}{\pi_{\rm inj}(\theta)} \pi_{\rm inj}(\theta)\ {\rm d}\theta {\rm d}x$$

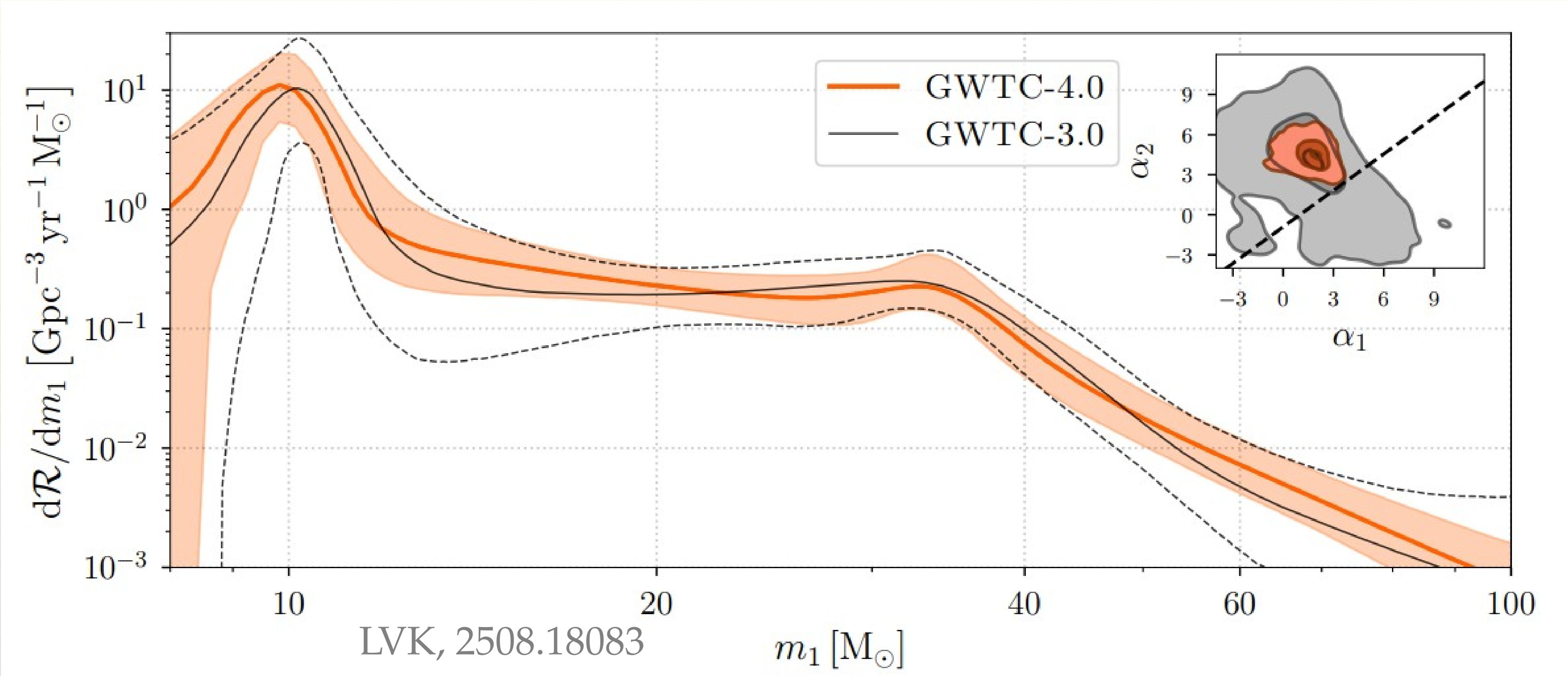
$$N_{\rm det}(\Lambda) \simeq \frac{1}{N_{\rm inj}} \sum_{\theta\sim\pi_{\rm inj}(\theta),{\rm det}} \frac{1}{\pi_{\rm inj}(\theta)} \frac{{\rm d}N}{{\rm d}\theta}(\Lambda)$$

Take care with efficiency of Monte-Carlo estimators!

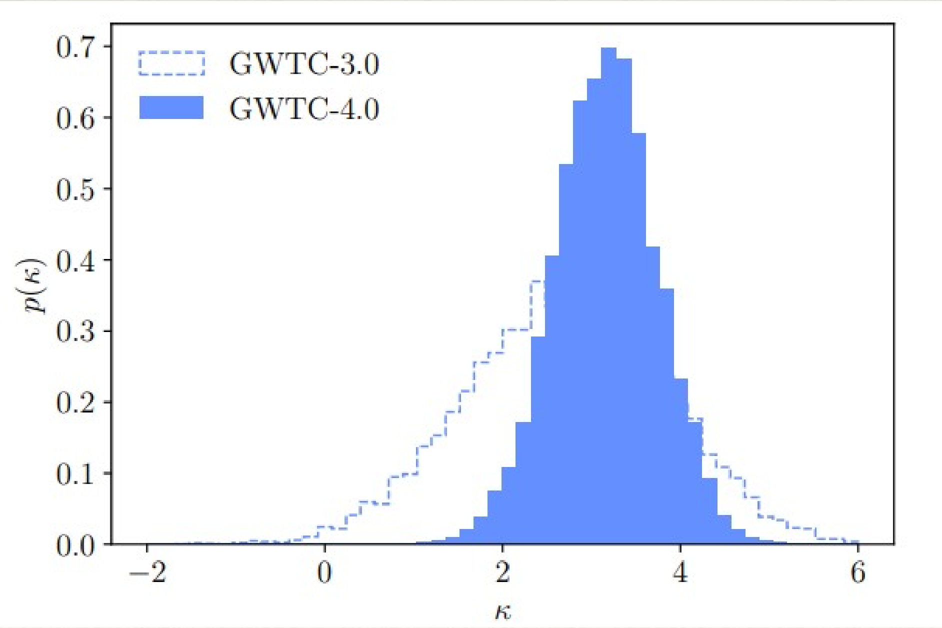
Astrophysical inference on GW data

- Assume a form for $p(\theta|\Lambda)$
  Can be:
  - Astrophysical
  - parametric
  - non-parametric

Parametric model (LVK)



Broken power-law + 2 peaks

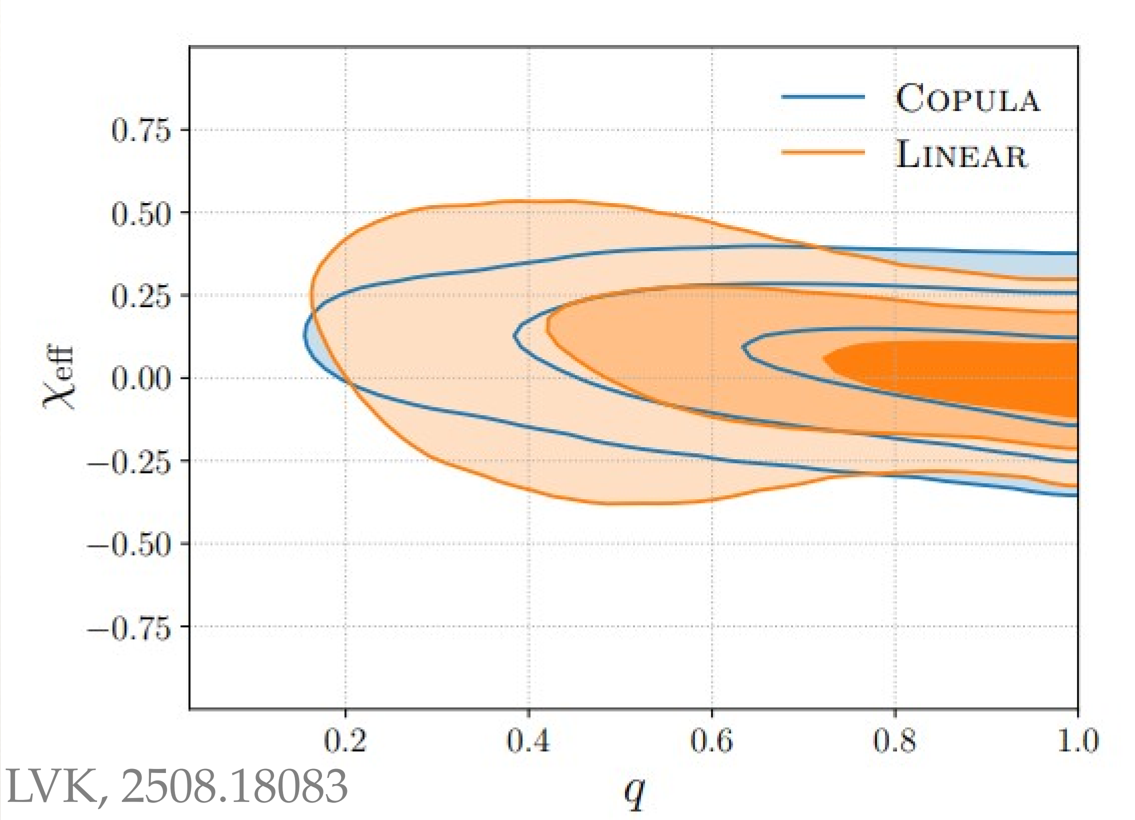


- Peaks at $\sim 10 M_\odot$ and $\sim 35 M_\odot$, steeper power-law at high mass
- Evidence for for rate evolution, with the rate higher in the past. $R(z) = R_0(1+z)^\kappa$

Parametric model 2D (LVK)

Linear:

$$p(\chi_{\rm eff}|q) \propto \exp\left[-\frac{(\chi_{\rm eff}-\mu(q))^2}{2\sigma(q)^2}\right]$$
$$\mu(q) = \mu_0 + \alpha(q-1)$$
$$\log_{10}\sigma(q) = \log_{10}(\sigma_0) + \beta(q-1)$$



LVK, 2508.18083

Copula:

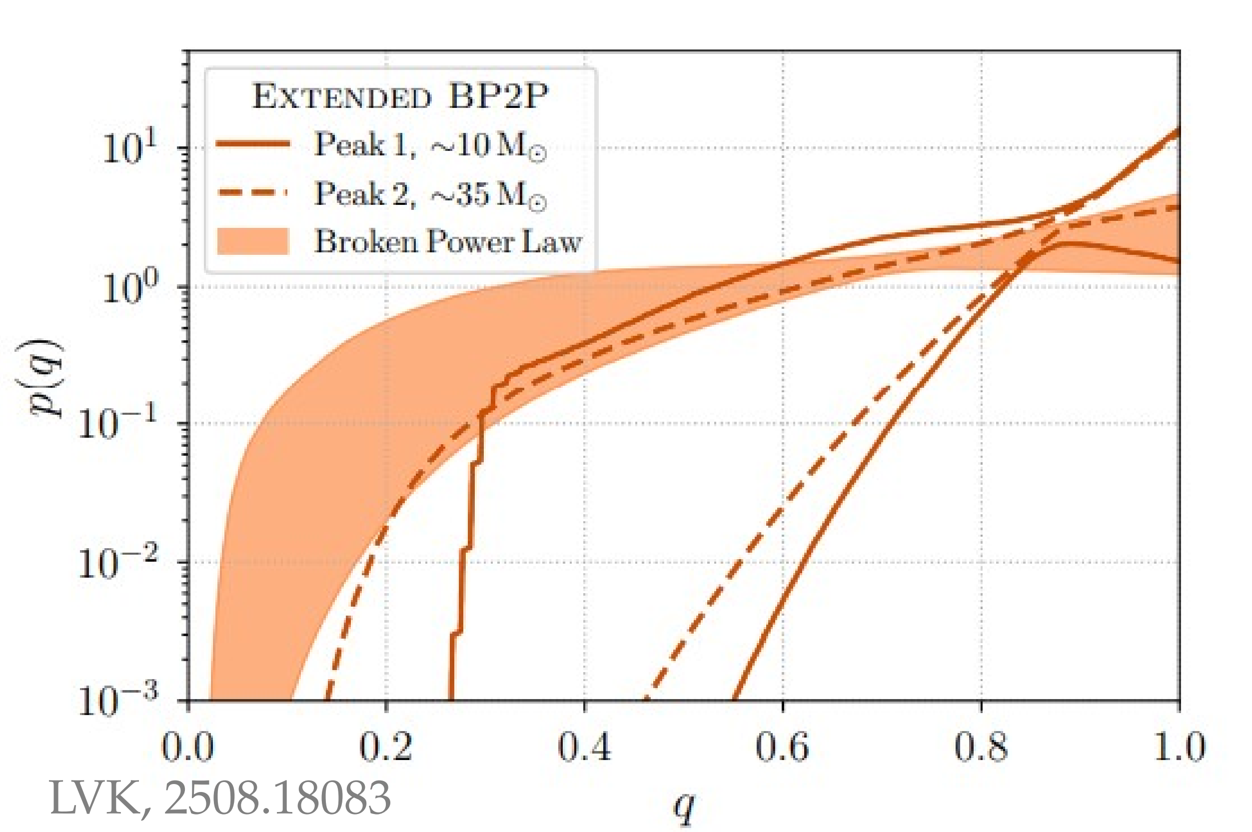
$$p(\chi_{\rm eff}, q) = p(\chi_{\rm eff})p(q)C(\chi_{\rm eff}, q)$$

$C(\chi_{\rm eff}, q)$ preserves the marginal

- Weaker evidence in GWTC-4 compared to GWTC-3

Parametric model nD (LVK)

$$p(m_1, q|\Lambda) = \sum_{i=1}^{n} \beta_i p_i(m_1|\Lambda) p_i(q|\Lambda)$$



LVK, 2508.18083

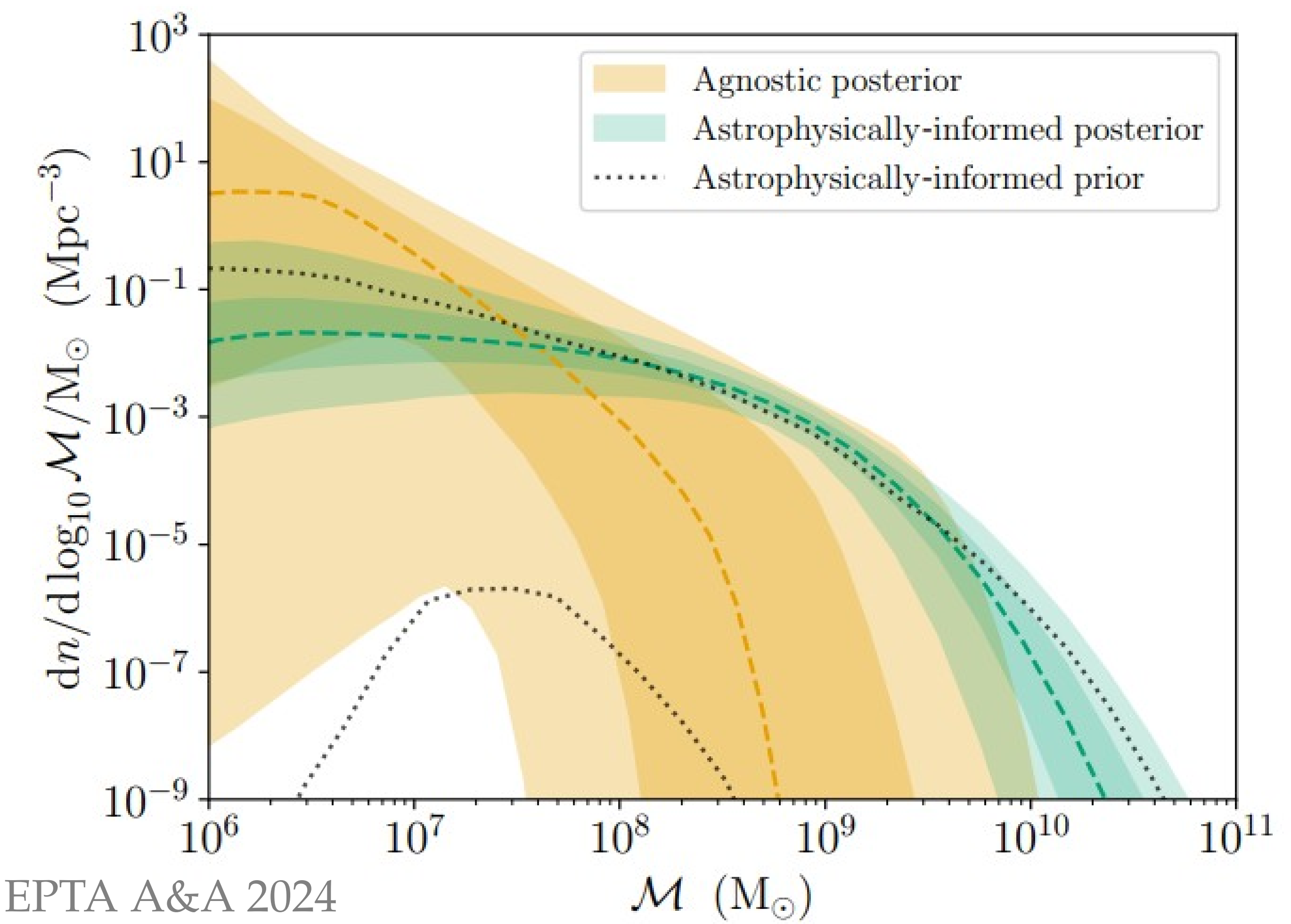
- Black holes in the $10 M_\odot$ peak may favour unequal mass ratios

- Number of components can be inferred with reversible-jump MCMC (Cheng et al. to appear)

Parametric model (PTA)

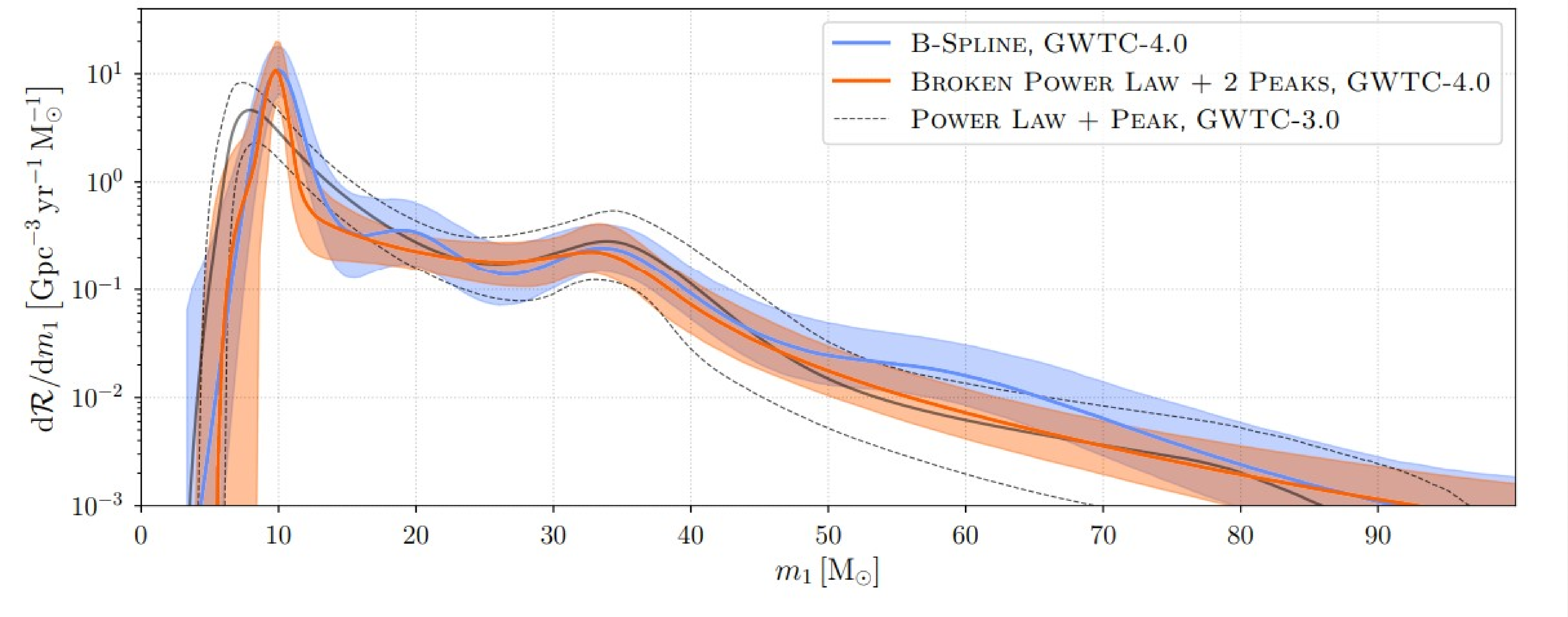
$$\frac{\mathrm{d}^2 n}{\mathrm{d}z\,\mathrm{d}\log_{10}\mathcal{M}} = \dot{n}_0 \left[\left(\frac{\mathcal{M}}{10^7 M_\odot} \right)^{-\alpha_\mathcal{M}} e^{-\mathcal{M}/\mathcal{M}_*} \right] \left[(1+z)^{\beta_z} e^{-z/z_0} \right] \frac{\mathrm{d}t_R}{\mathrm{d}z}$$

$$\mathcal{M} = \left(\frac{m_1^3 m_2^3}{m_1 + m_2} \right)^{1/5}$$



EPTA A&A 2024

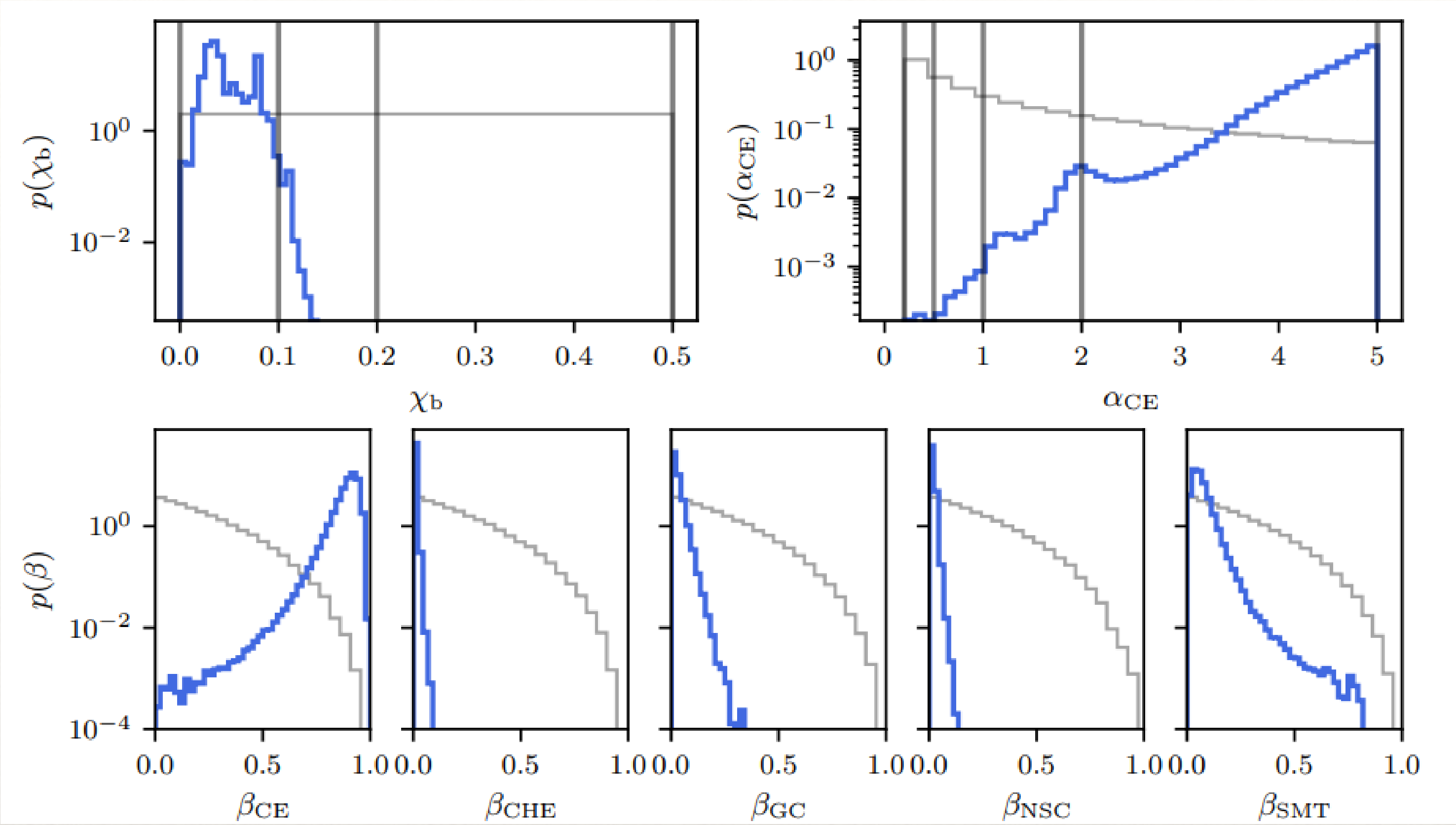
Non-parametric model (LVK)



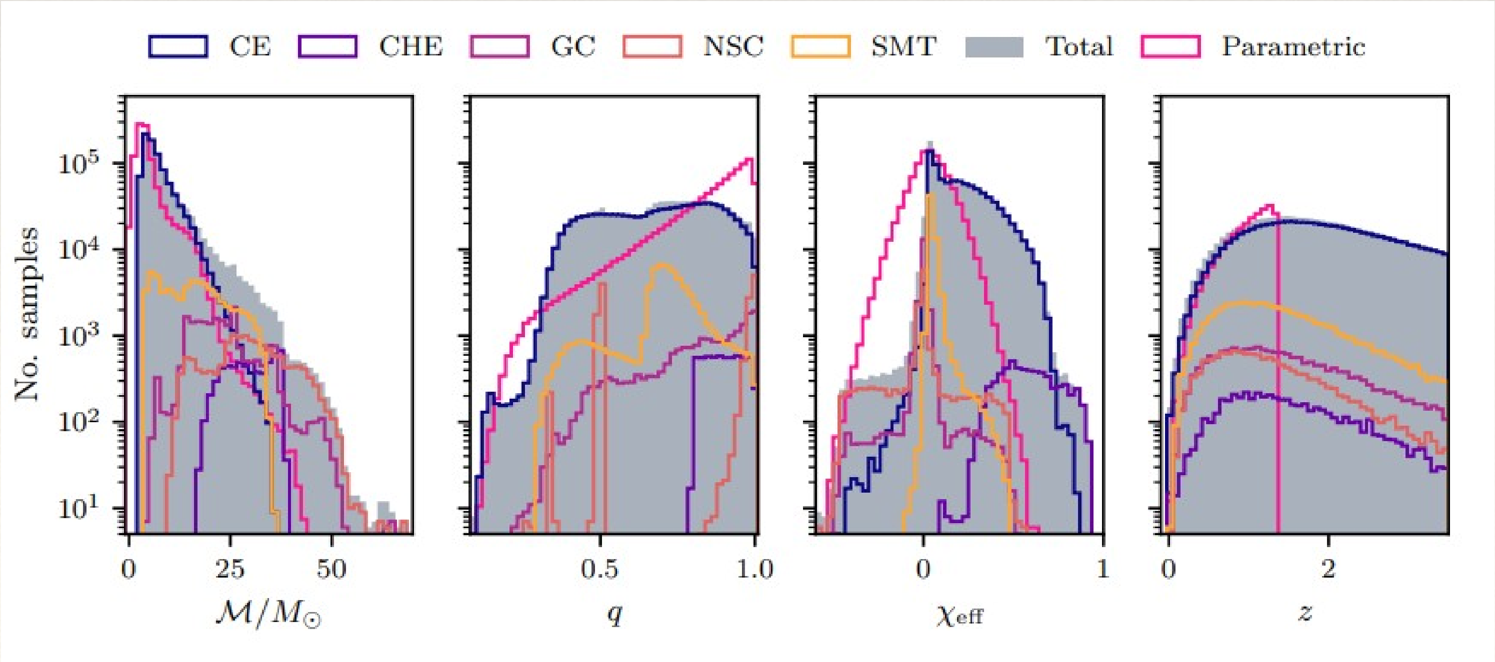
- Additional peak at $\sim 20 M_\odot$, excess at high mass?
- See also Toubiana et al. MNRAS 2023 for optimisation of number of knots with reversible-jump MCMC

Astrophysical (LVK)

$$p(\theta|\Lambda) = \sum_{i=1}^{5} \beta_i p_i(\theta|\Lambda)$$



- CE: common envelope
- CHE: chemically homogeneous evolution
- GC: globular cluster
- NSC: nuclear stellar cluster
- SMT: stable mass transfer



Colloms et al. A&A 2025

Favours formation through common envelope

Astrophysical inference on GW data

Astrophysical

- **Pros:**

  Gives direct information on astrophysical processes

- **Cons:**

  Huge uncertainty on astro models. We might not include all channels (see Cheng et al., 2307.03129, Raikman et al., 2310.10736)

  Requires some way to evaluate pdf from samples and to interpolate (see Toubiana et al., PRD 2021 for systematic errors)

Parametric

- **Pros:**

  Analytic pdfs, easy to evaluate

  Some astrophysical meaning

- **Cons:**

  Little flexibility

  Not so much astrophysical meaning

Non-parametric

- **Pros:**

  Very flexible

  Requires less a prior knowledge

- **Cons:**

  Parameters have no astrophysical meaning

  Complexity is a priori arbitrary (might be alleviated using reversible-jump MCMC)

"Mock LVK" toy model

- Mass distributed according to power-law+peak
- Known selection function
- Gaussian error on each measurement
- 100 detected events

Posterior predictive checks

- Test if the assumed model is a good fit to the data

Posterior predictive checks

- Test if the assumed model is a good fit to the data
- **Prior predictive distribution**:

$$p(x) = \int p(x \mid \Lambda)\, p(\Lambda)\, \mathrm{d}\Lambda$$

Distribution of data within the model assumed in the prior

Posterior predictive checks

- Test if the assumed model is a good fit to the data
- **Prior predictive distribution**:

$$p(x) = \int p(x \mid \Lambda)\, p(\Lambda)\, \mathrm{d}\Lambda$$

  Distribution of data within the model assumed in the prior

- **Posterior predictive distribution**:

$$p(x_{\mathrm{new}} \mid \{x\}_{\mathrm{old}}) = \int p(x_{\mathrm{new}} \mid \Lambda)\, p(\Lambda \mid \{x\}_{\mathrm{old}})\, \mathrm{d}\Lambda$$

Posterior predictive checks

- Test if the assumed model is a good fit to the data
- **Prior predictive distribution**:

$$p(x) = \int p(x \mid \Lambda)\, p(\Lambda)\, \mathrm{d}\Lambda$$

  Distribution of data within the model assumed in the prior

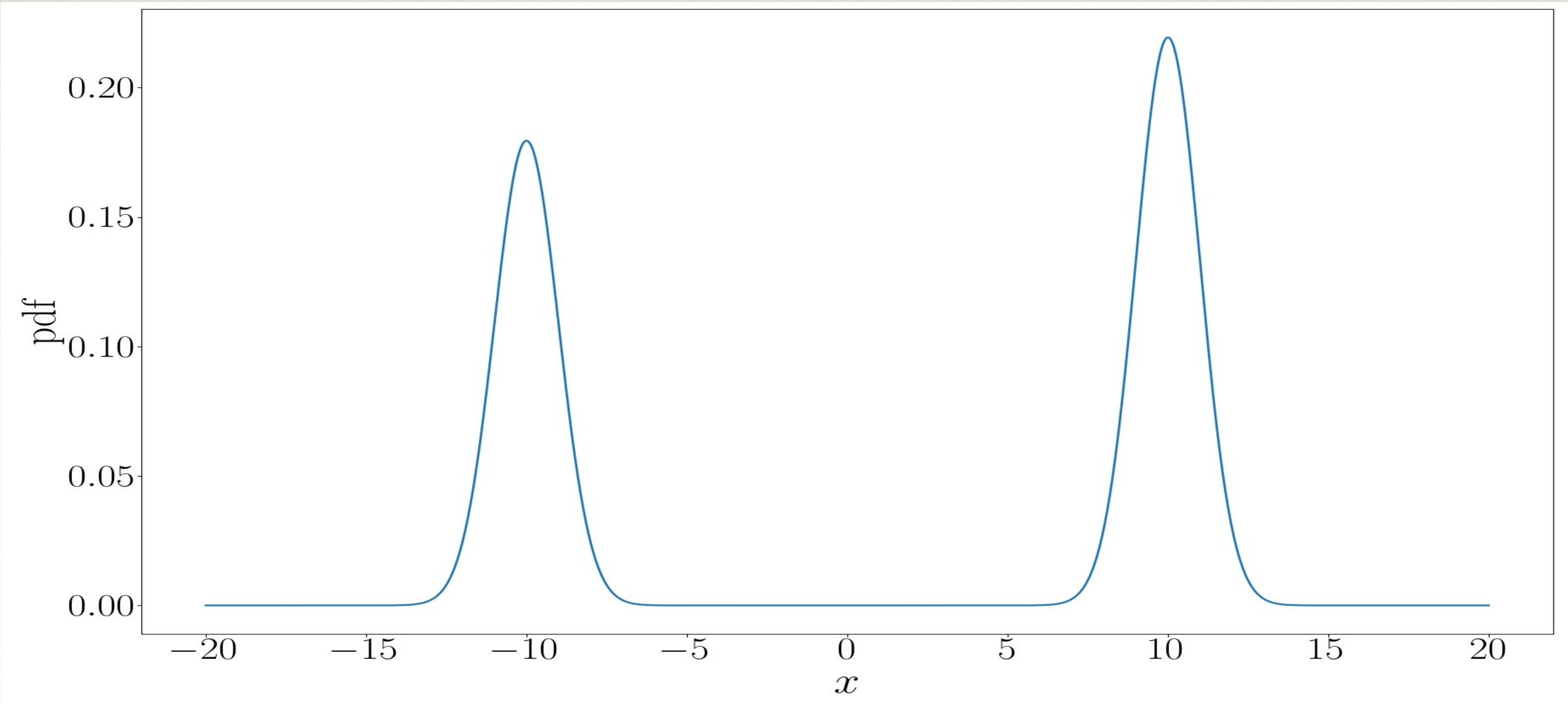
- **Posterior predictive distribution**:

$$p(x_{\mathrm{new}} \mid \{x\}_{\mathrm{old}}) = \int p(x_{\mathrm{new}} \mid \Lambda)\, p(\Lambda \mid \{x\}_{\mathrm{old}})\, \mathrm{d}\Lambda$$

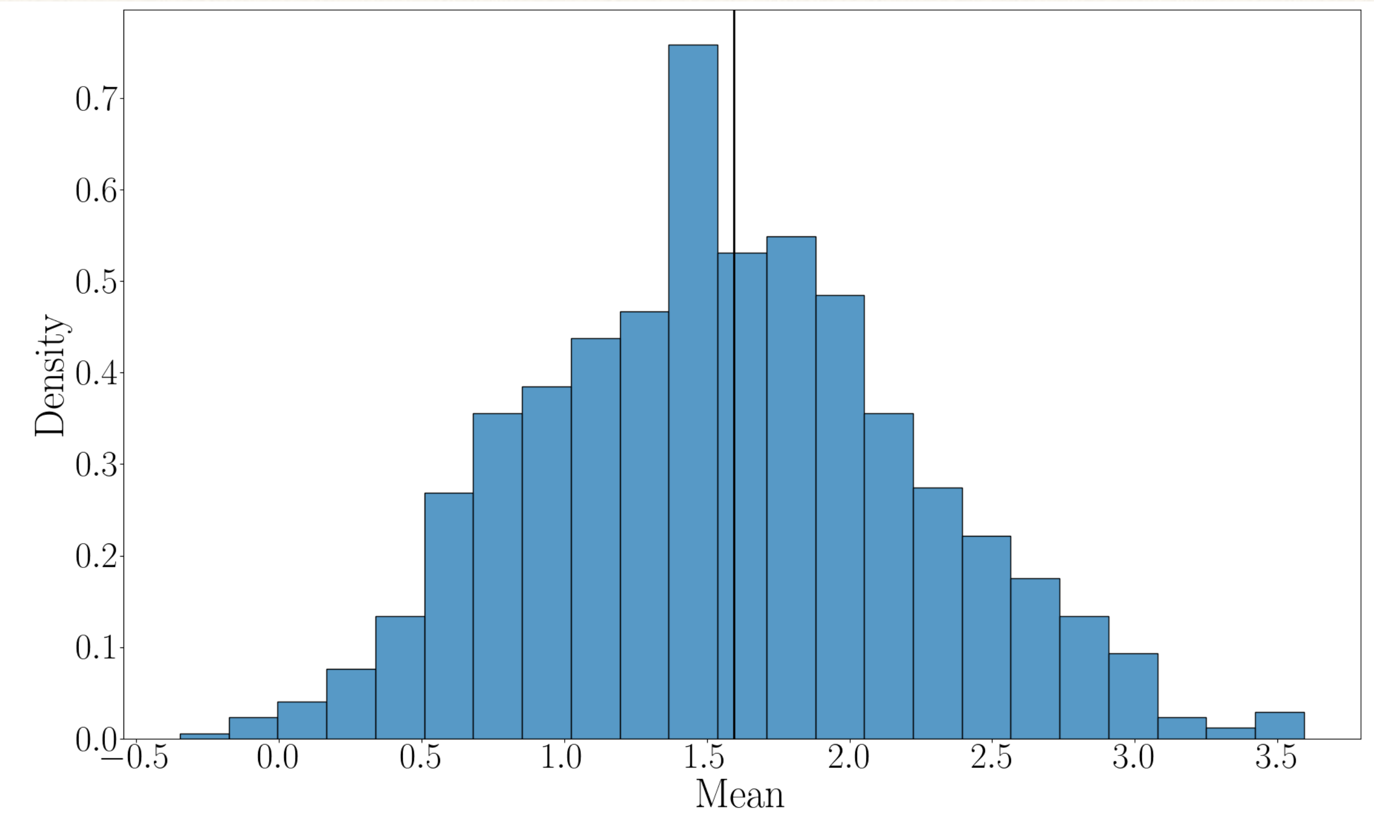
  Distribution of new datasets based on the model fitted to the data.
  Observed data should lie within this distribution if model is good.

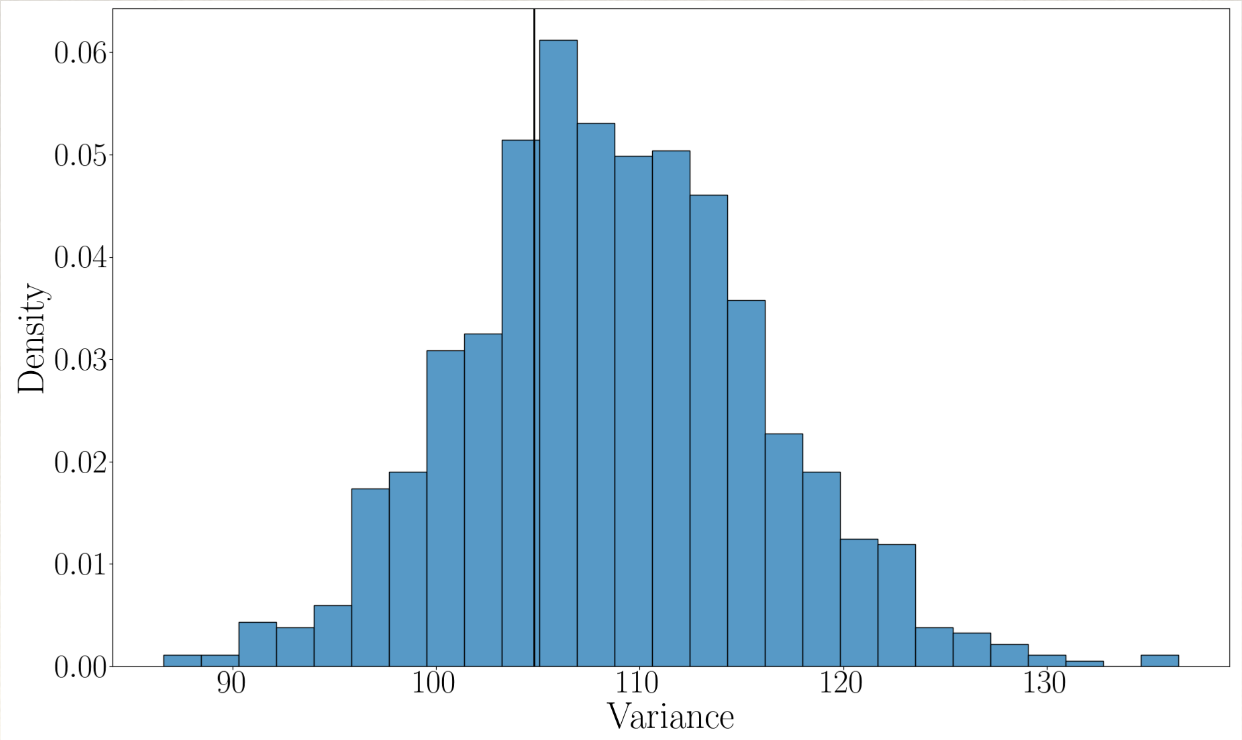
Example: gaussian fit

- The predictive distribution can be used to compute the distribution of summary quantities. The value of those summary quantities in the observed data can then be compared to these distributions.
- It is better to choose quantities that are somewhat "orthogonal" to what is adjusted to fit the data.
- Example: we try to fit a Gaussian to the following distribution:
- We assume a Gaussian measurement error of 2

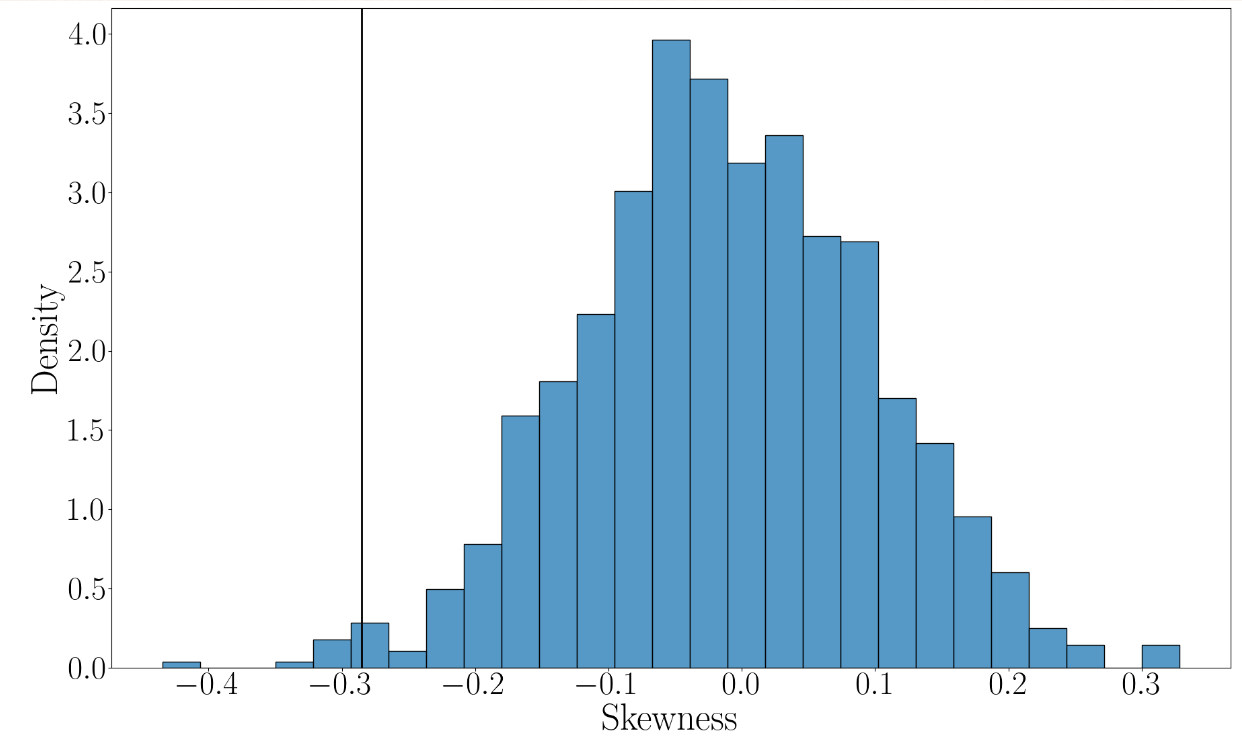


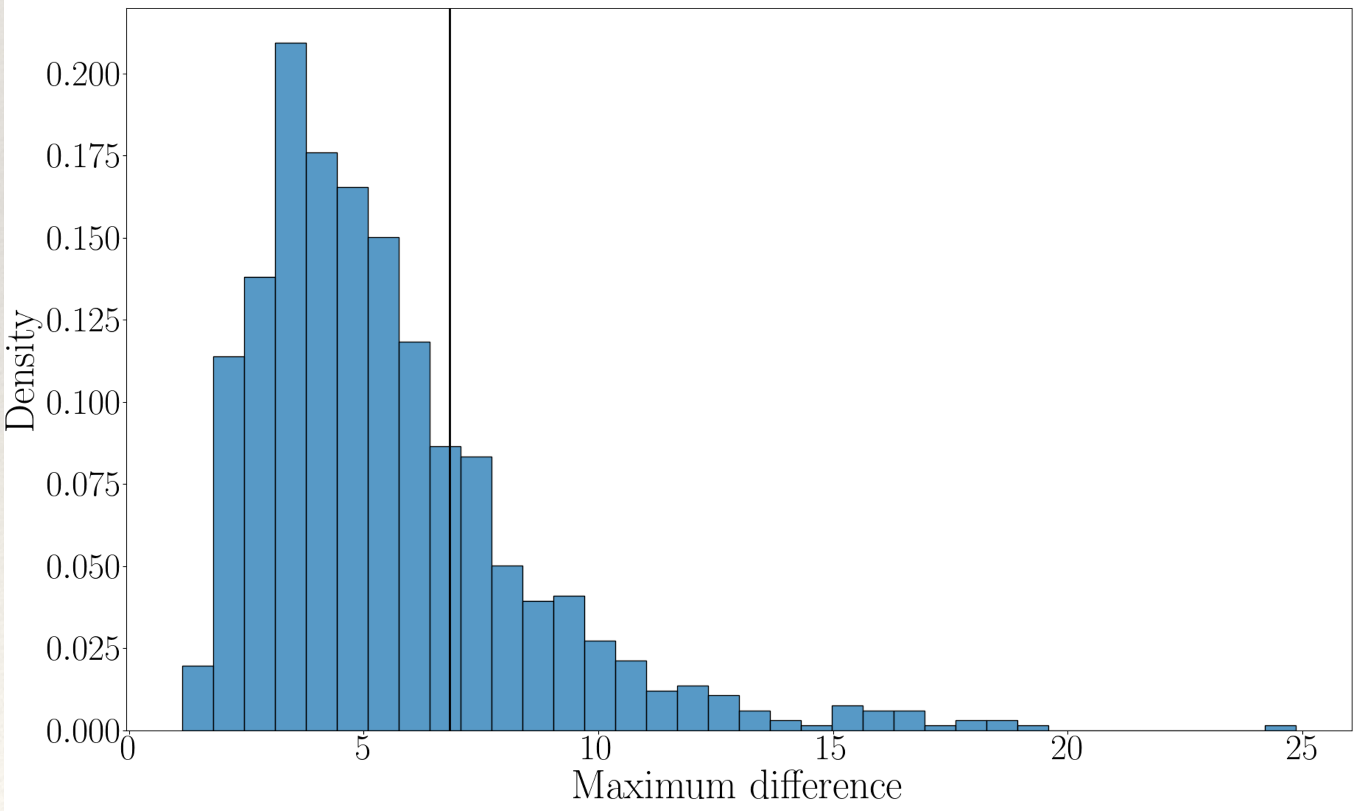
Example: gaussian fit





Example: gaussian fit





Posterior predictive checks

- Posterior predictive checks are “good practice”.
- Can help build intuition how to improve models.
- But are often computationally expensive...