Learnt harmonic mean estimator for Bayesian model comparison

arXiv:2111.12720; github.com/astro-informatics/harmonic

Jason D. McEwen*, Christopher G. R. Wallis, Matthew A. Price, Matthew M. Docherty

* www.jasonmcewen.org

Mullard Space Science Laboratory (MSSL), University College London (UCL)

April 2022

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{\mathrm{P}(\theta \mid y, M)}_{\text{posterior}} = \frac{\overbrace{\mathrm{P}(y \mid \theta, M)}^{\text{likelihood}} \; \overbrace{\mathrm{P}(\theta \mid M)}^{\text{prior}}}{\underbrace{\mathrm{P}(y \mid M)}_{\text{evidence}}}$$

for parameters θ, model M and observed data y.

Bayesian inference: parameter estimation

Bayes' theorem

$$\underbrace{\mathrm{P}(\theta \mid y, M)}_{\text{posterior}} = \frac{\overbrace{\mathrm{P}(y \mid \theta, M)}^{\text{likelihood}} \; \overbrace{\mathrm{P}(\theta \mid M)}^{\text{prior}}}{\underbrace{\mathrm{P}(y \mid M)}_{\text{evidence}}} = \frac{\overbrace{\mathcal{L}(\theta)}^{\text{likelihood}} \; \overbrace{\pi(\theta)}^{\text{prior}}}{\underbrace{Z}_{\text{evidence}}},$$

for parameters θ, model M and observed data y.

Bayesian inference: model selection

For **model selection**, consider the posterior model probabilities:

$$\underbrace{\frac{\mathrm{P}(M_1 \mid y)}{\mathrm{P}(M_2 \mid y)}}_{\text{posterior odds}} = \underbrace{\frac{\mathrm{P}(M_1)}{\mathrm{P}(M_2)}}_{\text{prior odds}} \times \underbrace{\frac{\mathrm{P}(y \mid M_1)}{\mathrm{P}(y \mid M_2)}}_{\text{Bayes factor}} .$$

Bayesian inference: model selection

For **model selection**, consider the posterior model probabilities:

$$\underbrace{\frac{P(M_1 \mid y)}{P(M_2 \mid y)}}_{\text{posterior odds}} = \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}} \times \underbrace{\frac{P(y \mid M_1)}{P(y \mid M_2)}}_{\text{Bayes factor}} .$$

Must compute the **Bayesian evidence** or **marginal likelihood** given by the normalising constant

$$z = P(y \mid M) = \int \mathrm{d}\theta \, \mathcal{L}(\theta) \, \pi(\theta) .$$

Bayesian inference: model selection

For **model selection**, consider the posterior model probabilities:

$$\underbrace{\frac{P(M_1 \mid y)}{P(M_2 \mid y)}}_{\text{posterior odds}} = \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}} \times \underbrace{\frac{P(y \mid M_1)}{P(y \mid M_2)}}_{\text{Bayes factor}} .$$

Must compute the **Bayesian evidence** or **marginal likelihood** given by the normalising constant

$$z = P(y \mid M) = \int \mathrm{d}\theta\, \mathcal{L}(\theta)\, \pi(\theta) .$$

→ Extremely challenging computational problem in high-dimensions.

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \mid y)}\left[\frac{1}{\mathcal{L}(\theta)}\right]$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)} \left[\frac{1}{\mathcal{L}(\theta)} \right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y)$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y)$$

$$= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\begin{aligned}\rho = \mathbb{E}_{\mathrm{P}(\theta\,|\,y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] &= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta\,|\,y) \\ &= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} \\ &= \frac{1}{z}\end{aligned}$$

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)} \left[\frac{1}{\mathcal{L}(\theta)} \right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y)$$
$$= \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$
$$= \frac{1}{z}$$

Original harmonic mean estimator (Newton & Raftery 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)} , \quad \theta_i \sim \mathrm{P}(\theta \,|\, y)$$

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta\,|\,y)}\left[\frac{1}{\mathcal{L}(\theta)}\right] = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta\,|\,y) = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} = \frac{1}{z}$$

Original harmonic mean estimator (Newton & Raftery 1994)

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)}, \quad \theta_i \sim \mathrm{P}(\theta\,|\,y)$$

Very simple approach but can fail catastrophically (Neal 1994).

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\pi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y) \, .$$

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \underbrace{\boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y)}}_{\text{importance sampling}} .$$

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

importance sampling

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y)} \, .$$

Importance sampling interpretation:

- Importance **sampling target distribution is prior** $\pi(\theta)$.
- Importance **sampling density is posterior** $\mathrm{P}(\theta \,|\, y)$.

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

importance sampling

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y)} .$$

Importance sampling interpretation:

- Importance **sampling target distribution is prior** $\pi(\theta)$.
- Importance **sampling density is posterior** $\mathrm{P}(\theta \,|\, y)$.

For importance sampling, want sampling density to have fatter tails than target.

Importance sampling interpretation of harmonic mean estimator

Alternative interpretation of harmonic mean relationship:

$$\rho = \int \mathrm{d}\theta \frac{1}{\mathcal{L}(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \underbrace{\boxed{\int \mathrm{d}\theta \frac{\pi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y)}}_{\text{importance sampling}} .$$

Importance sampling interpretation:

- Importance **sampling target distribution is prior** $\pi(\theta)$.
- Importance **sampling density is posterior** $\mathrm{P}(\theta \,|\, y)$.

For importance sampling, want sampling density to have fatter tails than target.

Not the case when importance sampling density is posterior and target is the prior.

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \mid y)} \left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right]$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y)$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y)$$

$$= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$
\begin{aligned}
\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] &= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y) \\
&= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} \\
&= \frac{1}{z}
\end{aligned}
$$

Re-targeted harmonic mean estimator

Introduce an arbitrary importance sampling target $\varphi(\theta)$ (which must be normalised).

Re-targeted **harmonic mean relationship** (Gelfand & Dey 1994)

$$\rho = \mathbb{E}_{\mathrm{P}(\theta \,|\, y)}\left[\frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)}\right] = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y)$$
$$= \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$
$$= \frac{1}{z}$$

Re-targeted **harmonic mean estimator** (Gelfand & Dey 1994)

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)}, \quad \theta_i \sim \mathrm{P}(\theta \,|\, y)$$

Re-targeted harmonic mean estimator

Importance sampling interpretation:

$$\rho = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y) \, .$$

Re-targeted harmonic mean estimator

Importance sampling interpretation:

$$\rho = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y) \,.$$

Ensure importance sampling target $\varphi(\theta)$ does **not** have fatter tails than posterior $\mathrm{P}(\theta \,|\, y)$ (importance sampling density).

Re-targeted harmonic mean estimator

Importance sampling interpretation:

$$\rho = \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathrm{P}(\theta \,|\, y) = \frac{1}{z} \int \mathrm{d}\theta \frac{\varphi(\theta)}{\mathrm{P}(\theta \,|\, y)} \mathrm{P}(\theta \,|\, y) \,.$$

Ensure importance sampling target $\varphi(\theta)$ does **not** have fatter tails than posterior $\mathrm{P}(\theta \,|\, y)$ (importance sampling density).

$\rightarrow$ How set importance sampling target distribution $\varphi(\theta)$?

How set importance sampling target distribution $\varphi(\theta)$?

Variety of cases been considered:

- Multi-variate Gaussian (*e.g.* Chib 1995)
- Indicator functions (*e.g.* Robert & Wraith 2009, van Haasteren 2009)

How set importance sampling target distribution $\varphi(\theta)$?

Variety of cases been considered:

- Multi-variate Gaussian (*e.g.* Chib 1995)
- Indicator functions (*e.g.* Robert & Wraith 2009, van Haasteren 2009)

Optimal target:

$$\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

(resulting estimator has zero variance).

How set importance sampling target distribution $\varphi(\theta)$?

Variety of cases been considered:

- Multi-variate Gaussian (*e.g.* Chib 1995)
- Indicator functions (*e.g.* Robert & Wraith 2009, van Haasteren 2009)

Optimal target:

$$\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}$$

(resulting estimator has zero variance).

But clearly **not feasible** since requires knowledge of the evidence z (recall the target must be normalised) $\rightarrow$ requires problem to have been solved already!

Learnt harmonic mean estimator

Propose the ***learnt* harmonic mean estimator** (McEwen *et al.* 2021; arXiv:2111.12720).

Learn an approximation of the optimal target distribution:

$$\boxed{\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}} .$$

Learnt harmonic mean estimator

Propose the ***learnt* harmonic mean estimator** (McEwen *et al.* 2021; arXiv:2111.12720).

Learn an approximation of the optimal target distribution:

$$\boxed{\varphi(\theta) \overset{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}}.$$

- Approximation not required to be highly accurate.
- Must not have fatter tails than posterior.

Learning the target distribution

Fit model by **minimising variance of resulting estimator**, while ensuring unbiased, with possible regularisation.

Solve by bespoke **mini-batch stochastic gradient descent**.

Develop strategy to estimate the **variance of the estimator**, its variance, and other sanity checks.

Cross-validation to select machine learning model and hyperparameters.

Normal-Gamma example

Pathological example (Friel & Wyse 2012) where original harmonic mean estimator fails.

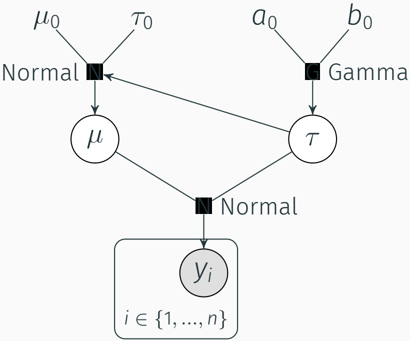
Data model:

$$y_i \sim \mathrm{N}(\mu, \tau^{-1})$$

Prior model:

Mean: $\mu \sim \mathrm{N}\big(\mu_0, (\tau_0\tau)^{-1}\big)$

Precision: $\tau \sim \mathrm{Ga}(a_0, b_0)$



Hierarchical Bayesian model of Normal-Gamma example.

Normal-Gamma example

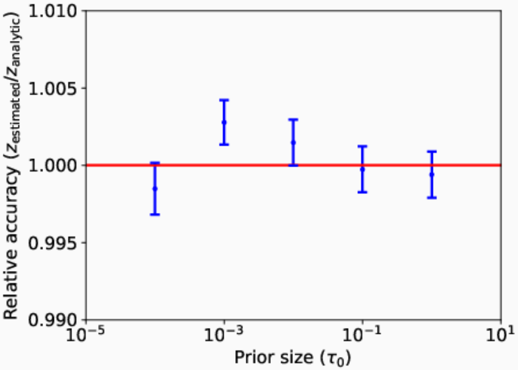
Analytic evidence:

$$z = (2\pi)^{-n/2} \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_n^{a_n}} \left(\frac{\tau_0}{\tau_n} \right)^{1/2}$$

where

$$\tau_n = \tau_0 + n \, , \quad a_n = a_0 + n/2 \, , \quad b_n = b_0 + \frac{1}{2} \sum_{i=1}^{n} (y_i - \bar{y})^2 + \frac{\tau_0 n (\bar{y} - \mu_0)^2}{2(\tau_0 + n)} \, .$$

Normal-Gamma example



Comparison of marginal likelihood values computed to truth for varying prior.

Normal-Gamma example

Marginal likelihood values for Normal-Gamma example with varying prior.

Prior size (τ_0)	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Analytic $\log(z)$	-144.5530	-143.4017	-142.2505	-141.0999	-139.9552
Estimated $\log(\hat{z})$	-144.5545	-143.3990	-142.2490	-141.1001	-139.9558
Error (learnt harmonic mean)	-0.0015	0.0027	0.0015	-0.0011	-0.0006
Error (original harmonic mean)	12.2100	—	9.7900	8.5000	7.1000

Harmonic code

GitHub harmonic | Tests passing | docs passing | codecov 96% | pypi package 1.0.3 | License GPL | arXiv 2111.12720

Harmonic

`harmonic` is an open source, well tested and documented Python implementation of the *learnt harmonic mean estimator* (McEwen et al. 2021) to compute the marginal likelihood (Bayesian evidence), required for Bayesian model selection.

Github: `https://github.com/astro-informatics/harmonic`

Docs: `https://astro-informatics.github.io/harmonic`

(Seamless integration with emcee.)

Summary

1. Learnt harmonic mean estimator is agnostic to sampling strategy.
2. Get model evidence (marginal likelihood) almost for free.
3. Professional code that's easy to use!

arXiv:2111.12720; github.com/astro-informatics/harmonic