Bayesian uncertainty quantification

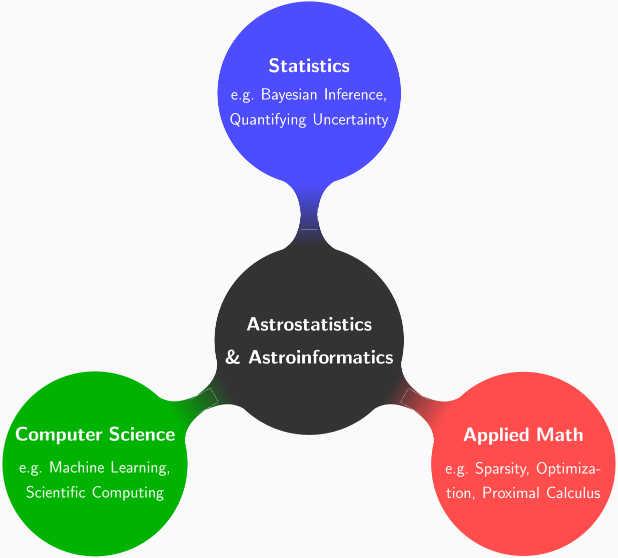
for radio interferometry and beyond

Jason D. McEwen
www.jasonmcewen.org

Mullard Space Science Laboratory (MSSL), UCL

April 2022

Merging paradigms



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

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

for parameters θ, model M and observed data y.

For **parameter estimation**, typically draw samples from the posterior by *Markov chain Monte Carlo (MCMC)* sampling.

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.

For **parameter estimation**, typically draw samples from the posterior by *Markov chain Monte Carlo (MCMC)* sampling.

$\rightarrow$ Challenging computational problem in high-dimensions.

Bayesian inference: model selection

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

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

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

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

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

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

Bayesian inference: model selection

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

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

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

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

→ Extremely challenging computational problem in high-dimensions.

Outline

1. Learnt harmonic mean estimator for Bayesian model comparison
2. Proximal nested sampling for high-dimensional Bayesian model comparison
3. High-dimensional Bayesian uncertainty quantification for extreme computation

Learnt harmonic mean estimator for Bayesian model comparison

Desirable properties for Bayesian evidence estimators

Seek estimator that is:

- **Agnostic to sampling method** and uses posterior samples.
- Potential to **scale to high-dimensions**.

Desirable properties for Bayesian evidence estimators

Seek estimator that is:

- **Agnostic to sampling method** and uses posterior samples.
- Potential to **scale to high-dimensions**.

Harmonic mean estimator has potential to meet these criteria but has serious shortcomings as originally posed.

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)

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

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)

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

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:

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 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** $\mathrm{P}(\theta \,|\, y)$.

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:

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

Importance sampling interpretation:

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

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 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)

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

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)

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

Re-targeted harmonic mean estimator

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).

Recall:

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

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).

Recall:

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

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:

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

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.

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:

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

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

Also develop strategy to estimate the variance of the estimator, its variance, and other sanity checks.

Learning the target distribution

Consider a **variety of machine learning approaches**:

- Uniform hyper-ellipsoid
- Kernel Density Estimation (KDE)
- Modified Gaussian mixture model (MGMM)

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

$$\min \ \hat{\sigma}^2 + \lambda R \quad \text{subject to} \quad \hat{\rho} = \hat{\mu}_1$$

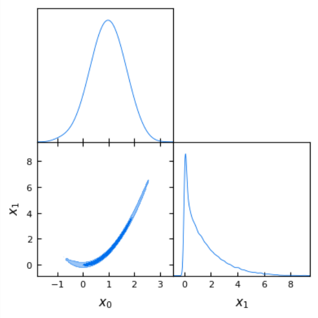
Solve by bespoke **mini-batch stochastic gradient descent**.

Cross-validation to select machine learning model and hyperparameters.

Rosenbrock example

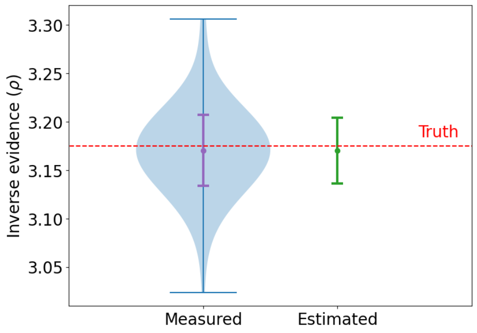
Rosenbrock function is the classical example of a **pronounced thin curving degeneracy**, with likelihood defined by

$$f(\theta) = \sum_{i=1}^{n-1} \left[(a - \theta_i)^2 + b(\theta_{i+1} - \theta_i^2)^2 \right] , \qquad \log(\mathcal{L}(\theta)) = -f(\theta) .$$



Posterior recovered by MCMC sampling.

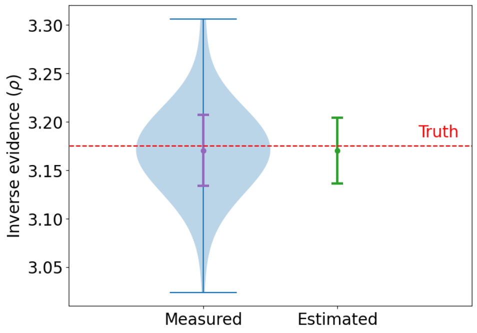
Rosenbrock example



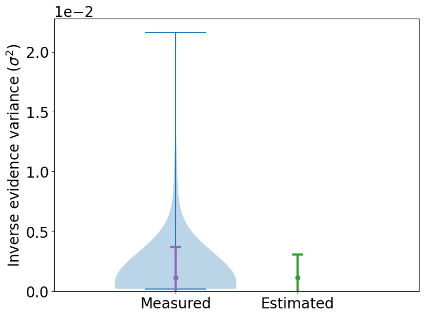
Reciprocal evidence

Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Rosenbrock example



Reciprocal evidence



Variance of reciprocal evidence

Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Normal-Gamma example

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

Normal-Gamma example

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

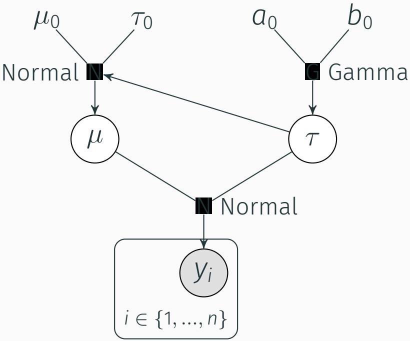
Data model:

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

Prior model:

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

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



Hierarchical Bayesian model of Normal-Gamma example.

Normal-Gamma example

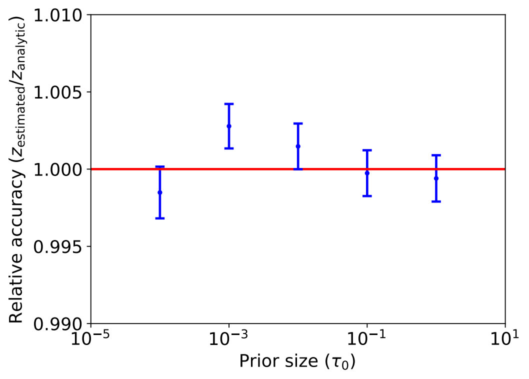
Analytic evidence:

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

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.

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

Radiata pine example

Radiata pine data-set has become **classical benchmark** for evaluating evidence estimators:

- maximum compression strength parallel to grain y_i,
- density x_i,
- density adjust for resin content z_i,

for $i \in \{1, \ldots, n\}$ where $n = 42$ specimens.

Is density or resin-adjusted density a better predictor of compression strength?

Radiata pine example

Gaussian linear models:

$$M_1: \qquad y_i = \alpha + \boxed{\beta(x_i - \bar{x})} + \epsilon_i , \qquad \epsilon_i \sim \mathrm{N}(0, \tau^{-1}) .$$

density

$$M_2: \qquad y_i = \gamma + \boxed{\delta(z_i - \bar{z})} + \eta_i , \qquad \eta_i \sim \mathrm{N}(0, \lambda^{-1}) .$$

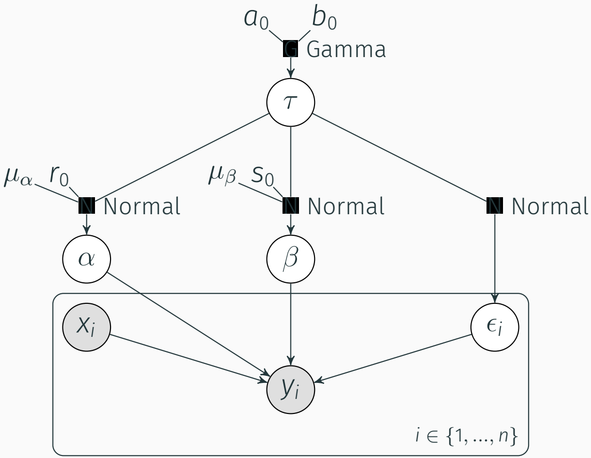
resin-adjusted density

Priors for model 1 (similar for model 2):

$$\alpha \sim \mathrm{N}\big(\mu_\alpha, (r_0\tau)^{-1}\big) ,$$
$$\beta \sim \mathrm{N}\big(\mu_\beta, (s_0\tau)^{-1}\big) ,$$
$$\tau \sim \mathrm{Ga}(a_0, b_0) ,$$

$(\mu_\alpha = 3000,\ \mu_\beta = 185,\ r_0 = 0.06,\ s_0 = 6,\ a_0 = 3,\ b_0 = 2 \times 300^2)$.

Radiata pine example



Hierarchical Bayesian model for Radiata pine example (for model 1; model 2 is similar).

Radiata pine example

Analytic evidence:

$$z = \pi^{-n/2} b_0^{a_0} \frac{\Gamma(a_0 + n/2)}{\Gamma(a_0)} \frac{|Q_0|^{1/2}}{|M|^{1/2}} \left(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} + \boldsymbol{\mu_0}^\mathsf{T} Q_0 \boldsymbol{\mu_0} - \boldsymbol{\nu_0}^\mathsf{T} M \boldsymbol{\nu_0} + 2b_0\right)^{-a_0 - n/2}$$

where $\boldsymbol{\mu_0} = (\mu_\alpha, \mu_\beta)^\mathsf{T}$, $Q_0 = \text{diag}(r_0, s_0)$, and $M = X^\mathsf{T} X + Q_0$.

Radiata pine example

Marginal likelihood values for Radiata Pine example.

	Model M_1 $\log(z_1)$	Model M_2 $\log(z_2)$	$\log \mathrm{BF}_{21}$ $= \log(z_2) - \log(z_1)$
Analytic	-310.12829	-301.70460	8.42368
Estimated	-310.12807	-301.70413	8.42394
	$\pm$ 0.00072	$\pm$ 0.00074	$\pm$ 0.00145
Error (learnt harmonic mean)	0.00022	0.00047	0.00026
Error (original harmonic mean)	–	–	-0.17372

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`

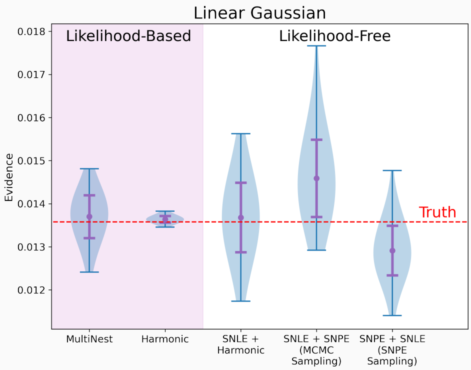
Docs: `https://astro-informatics.github.io/harmonic`

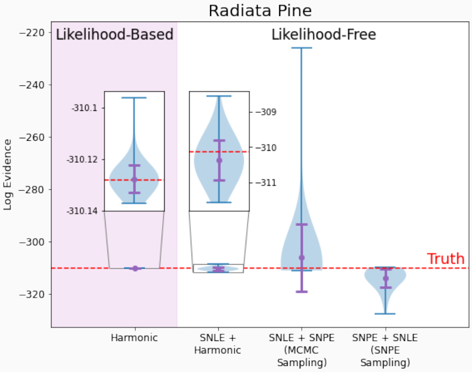
(Seamless integration with emcee.)

Code example

```
# Import packages
import numpy as np
import emcee
import harmonic

# Run MCMC sampler
sampler = emcee.EnsembleSampler(nchains, ndim, ln_posterior, args=[args])
sampler.run_mcmc(pos, samples_per_chain)
samples = np.ascontiguousarray(sampler.chain[:,nburn:,:])
lnprob = np.ascontiguousarray(sampler.lnprobability[:,nburn:])

# Set up chains
chains = harmonic.Chains(ndim)
chains.add_chains_3d(samples, lnprob)

# Fit model
chains_train, chains_test = harmonic.utils.split_data(chains, train_prop=0.05)
model = harmonic.model.KernelDensityEstimate(ndim, domain, hyper_parameters)
model.fit(chains_train.samples, chains_train.ln_posterior)

# Compute evidence
evidence = harmonic.Evidence(chains_test.nchains, model)
evidence.add_chains(chains_test)
ln_evidence, ln_evidence_std = evidence.compute_ln_evidence()
```

Model comparison for likelihood-free inference





Proximal nested sampling for high-dimensional Bayesian model comparison

Nested sampling

Nested sampling is a clever approach to efficiently evalute the evidence (Skilling 2006).

Consider $\Omega_{L*} = \{x|\mathcal{L}(x) \geq L^*\}$, which groups the parameter space Ω into a series of **nested subspaces**.

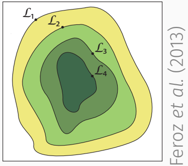
Define the prior volume ξ by $d\xi = \pi(x)dx$, where

$$\xi(L^*) = \int_{\Omega_{L*}} \pi(x)dx.$$

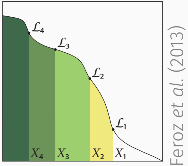
The marginal likelihood integral can then be rewritten as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(\xi)d\xi,$$

which is a **one-dimensional integral** over the prior volume ξ.



Nested subspaces



Reparameterised likelihood

Constrained sampling from the prior

To compute the marginal likelihood by nested sampling, thus require strategy to generate likelihoods L_i and associated prior volumes ξ_i.

Acheived by **sampling from the prior, subject the likelihood iso-contour constraint**, *i.e.* sampling from the prior $\pi(x)$, such that $\mathcal{L}(x) > L^*$

Constrained sampling from the prior

To compute the marginal likelihood by nested sampling, thus require strategy to generate likelihoods L_i and associated prior volumes ξ_i.

Acheived by **sampling from the prior, subject the likelihood iso-contour constraint**, *i.e.* sampling from the prior $\pi(x)$, such that $\mathcal{L}(x) > L^*$

This is the main difficulty in applying nested sampling to high-dimensional problems.

Exploit common structure

Many high-dimensional inverse problems are **log-convex**, *e.g.* inverse imaging problems with Gaussian data fidelity and sparsity-promoting prior.

$\rightarrow$ Exploit structure (log convexity) of the problem.

Constrained sampling formulation

Consider case where prior and likelihood of form

$$\underbrace{\boxed{\pi(x) = \exp(-f(x))}}_{\text{prior}}, \qquad \underbrace{\boxed{\mathcal{L}(x) = \exp(-g(x))}}_{\text{likelihood}},$$

where f and g are convex lower semicontinuous functions on Ω.

Let $\iota_{L^*}(x)$ and $\chi_{L^*}(x)$ be the indicator and characteristic functions:

$$\iota_{L^*}(x) = \begin{cases} 1, & \mathcal{L}(x) > L^*, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad \chi_{L^*}(x) = \begin{cases} 0, & \mathcal{L}(x) > L^*, \\ +\infty, & \text{otherwise}. \end{cases} \tag{1}$$

Then let $\boxed{\pi_{L^*}(x) = \pi(x)\iota_{L^*}(x)}$ represent the prior distribution with the hard likelihood constraint.

Constrained sampling formulation

Taking the logarithm, we can write

$$\boxed{-\log \pi_{L^*}(x) = f(x) + \chi_{\mathcal{B}_\tau}(x)},$$

where $\chi_{\mathcal{B}_\tau}(x)$ is the characteristic function associated with the convex set

$$\mathcal{B}_\tau := \{x | g(x) < \tau\},$$

for $\tau = -\log L^*$.

MCMC sampling with Langevin dynamics

Consider posteriors of the following form:

$$\mathrm{P}(x \mid y) = \pi(x) \propto \exp\bigl(-p(x)\bigr).$$

If $p(x)$ differentiable can adopt Langevin dynamics.

Based on **Langevin diffusion process** $\mathcal{L}(t)$, with π as stationary distribution:

$$\mathrm{d}\mathcal{L}(t) = \frac{1}{2}\nabla \log \pi\bigl(\mathcal{L}(t)\bigr)\mathrm{d}t + \mathrm{d}\mathcal{W}(t), \quad \mathcal{L}(0) = l_0$$

where $\mathcal{W}$ is Brownian motion.

MCMC sampling with Langevin dynamics

Consider posteriors of the following form:

$$\mathrm{P}(x \mid y) = \pi(x) \propto \exp\big(-p(x)\big).$$

If $p(x)$ differentiable can adopt Langevin dynamics.

Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

$$\mathrm{d}\mathcal{L}(t) = \frac{1}{2} \underbrace{\nabla \log \pi\big(\mathcal{L}(t)\big)}_{\text{gradient}} \mathrm{d}t + \mathrm{d}\mathcal{W}(t), \quad \mathcal{L}(0) = l_0$$

where $\mathcal{W}$ is Brownian motion.

Need gradients so **not directly applicable**.

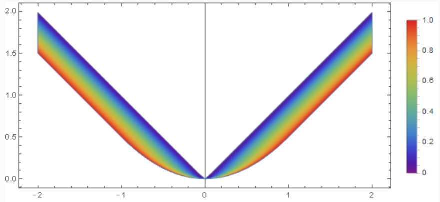
Moreau-Yosida approximation

Moreau-Yosida approximation (envelope) of f:

$$f^{\lambda}(x) = \inf_{u \in \mathbb{R}^N} f(u) + \frac{\|u - x\|^2}{2\lambda}$$

Important properties of $f^{\lambda}(x)$:

1. As $\lambda \rightarrow 0, f^{\lambda}(x) \rightarrow f(x)$
2. $\nabla f^{\lambda}(x) = (x - \text{prox}_f^{\lambda}(x))/\lambda$



Moreau-Yosida envelope of $|x|$ for varying λ [Credit: Stack exchange (ubpdqn)]

Proximal nested sampling

Proximal nested sampling (Cai, McEwen & Pereyra 2021; arXiv:2106.03646)

- Constrained sampling formulation
- Langevin MCMC sampling
- Moreau-Yosida approximation of constraint (and any non-differentiable prior)

Proximal nested sampling

Proximal nested sampling (Cai, McEwen & Pereyra 2021; arXiv:2106.03646)

- Constrained sampling formulation
- Langevin MCMC sampling
- Moreau-Yosida approximation of constraint (and any non-differentiable prior)

Proximal nested sampling Markov chain:

$$\boxed{x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\left[x^{(k)} - \mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right] + \sqrt{\delta}\, w^{(k+1)}}\,.$$

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta} w^{(k+1)}.$$

1. $x^{(k)}$ **is already in** $\mathcal{B}_\tau$: term $\left[x^{(k)} - \text{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.
2. $x^{(k)}$ **is not in** $\mathcal{B}_\tau$: a step is also taken in the direction $-\left[x^{(k)} - \text{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$, which moves the next iteration in the direction of the projection of $x^{(k)}$ onto the convex set $\mathcal{B}_\tau$. Acts to push the Markov chain back into the constraint set $\mathcal{B}_\tau$ if it wanders outside of it.

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta}w^{(k+1)}.$$

1. $x^{(k)}$ **is already in** $\mathcal{B}_\tau$: term $\left[x^{(k)} - \mathrm{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.
2. $x^{(k)}$ **is not in** $\mathcal{B}_\tau$: a step is also taken in the direction $-\left[x^{(k)} - \mathrm{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$, which moves the next iteration in the direction of the projection of $x^{(k)}$ onto the convex set $\mathcal{B}_\tau$. Acts to push the Markov chain back into the constraint set $\mathcal{B}_\tau$ if it wanders outside of it.

A subsequent Metropolis-Hastings step guarantees hard likelihood constraint is satisfied.

Proximal nested sampling intuition

Recall proximal nested sampling Markov chain:

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\boxed{\left[x^{(k)} - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]} + \sqrt{\delta}w^{(k+1)}.$$

1. $x^{(k)}$ **is already in** $\mathcal{B}_\tau$: term $\left[x^{(k)} - \text{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$ disappears and recover usual Langevin MCMC.
2. $x^{(k)}$ **is not in** $\mathcal{B}_\tau$: a step is also taken in the direction $-\left[x^{(k)} - \text{prox}^{\lambda}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\right]$, which moves the next iteration in the direction of the projection of $x^{(k)}$ onto the convex set $\mathcal{B}_\tau$. Acts to push the Markov chain back into the constraint set $\mathcal{B}_\tau$ if it wanders outside of it.

A subsequent Metropolis-Hastings step guarantees hard likelihood constraint is satisfied.

Many further details regarding explicit forms for common priors and likelihoods and how to compute proximity operators efficiently (Cai, McEwen & Pereyra 2021; arXiv:2106.03646).

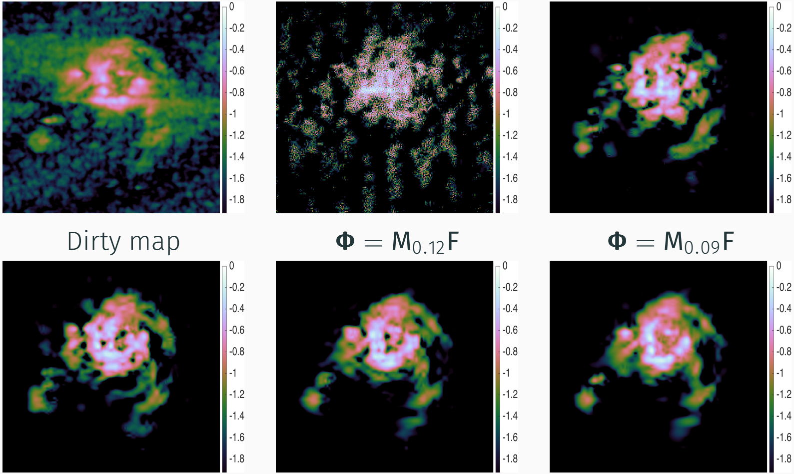
Measurement model misspecification experiment

Consider ground truth model $\mathbf{\Phi} = \mathsf{M}_{\text{truth}}\mathsf{F}$ to simulate observational data $\boldsymbol{y}$.

However, when solving the inverse problem consider misspecified models M_γ, where $\gamma > 0$ encodes the level of misspecification (mimics incorrectly specified wavelength).

Compute the model evidence using **proximal nested sampling**, using evidence to distinguish correct model.

Measurement model misspecification experiment



Dirty map

$\mathbf{\Phi} = \mathsf{M}_{0.12}\mathsf{F}$

$\mathbf{\Phi} = \mathsf{M}_{0.09}\mathsf{F}$

$\mathbf{\Phi} = \mathsf{M}_{0.06}\mathsf{F}$

$\mathbf{\Phi} = \mathsf{M}_{0.03}\mathsf{F}$

$\mathbf{\Phi} = \mathsf{M}_{\text{truth}}\mathsf{F}$

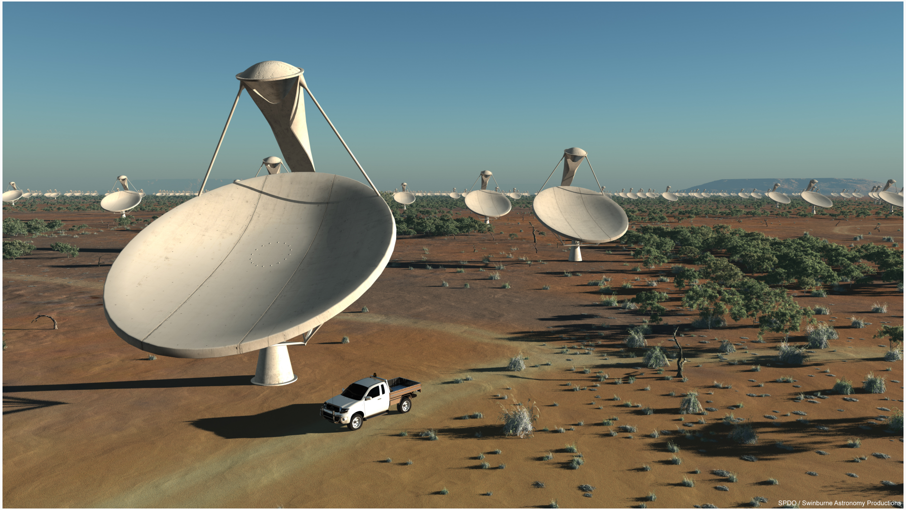
Measurement model misspecification experiment

Model	$\log \mathcal{Z}$	RMSE (Requires ground truth)
$\mathbf{\Phi} = \mathsf{M}_{\text{truth}}\mathsf{F}$	$-4.47 \times 10^3 \pm 0.08$	3.40
$\mathbf{\Phi} = \mathsf{M}_{0.03}\mathsf{F}$	$-4.88 \times 10^3 \pm 0.08$	7.85
$\mathbf{\Phi} = \mathsf{M}_{0.06}\mathsf{F}$	$-5.63 \times 10^3 \pm 0.08$	12.01
$\mathbf{\Phi} = \mathsf{M}_{0.09}\mathsf{F}$	$-9.21 \times 10^3 \pm 0.07$	15.71
$\mathbf{\Phi} = \mathsf{M}_{0.12}\mathsf{F}$	$-1.44 \times 10^4 \pm 0.08$	18.08

Evidence computed by proximal nested sampling correctly classifies models.

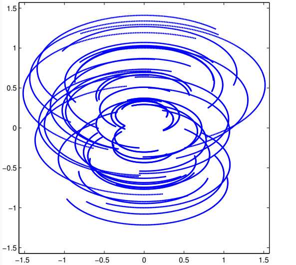
High-dimensional Bayesian uncertainty quantification for extreme computation

Square Kilometre Array (SKA)



Radio interferometric telescopes acquire “Fourier” measurements

“Fourier”
Measurements
$\Rightarrow$



Radio interferometric inverse problem

- Consider the ill-posed inverse problem of radio interferometric imaging:

$$\boxed{\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{n}},$$

where $\boldsymbol{y}$ are the measured visibilities, $\boldsymbol{\Phi}$ is the linear measurement operator, $\boldsymbol{x}$ is the underlying image and $\boldsymbol{n}$ is instrumental noise.

Radio interferometric inverse problem

- Consider the ill-posed inverse problem of radio interferometric imaging:

$$\boxed{y = \Phi x + n},$$

where y are the measured visibilities, Φ is the linear measurement operator, x is the underlying image and n is instrumental noise.

- Measurement operator, *e.g.* $\boxed{\Phi = \mathsf{GFA}}$, may incorporate:
  - primary beam A of the telescope;
  - Fourier transform F;
  - convolutional de-gridding G to interpolate to continuous *uv*-coordinates;
  - direction-dependent effects (DDEs)...

Radio interferometric inverse problem

- Consider the ill-posed inverse problem of radio interferometric imaging:

$$\boxed{\boldsymbol{y} = \boldsymbol{\Phi} \boldsymbol{x} + \boldsymbol{n}} \, ,$$

  where $\boldsymbol{y}$ are the measured visibilities, $\boldsymbol{\Phi}$ is the linear measurement operator, $\boldsymbol{x}$ is the underlying image and $\boldsymbol{n}$ is instrumental noise.

- Measurement operator, *e.g.* $\boxed{\boldsymbol{\Phi} = \mathbf{G}\mathbf{F}\mathbf{A}}$, may incorporate:
  - primary beam $\mathbf{A}$ of the telescope;
  - Fourier transform $\mathbf{F}$;
  - convolutional de-gridding $\mathbf{G}$ to interpolate to continuous *uv*-coordinates;
  - direction-dependent effects (DDEs)...

Interferometric imaging: recover an image from noisy and incomplete Fourier measurements.

Interferometric imaging and MAP estimation

Many interferometric imaging approaches are based on **regularisation**, *i.e.* minimising an objective function comprised of a data-fidelity penalty and a regularisation penalty.

From a Bayesian perspective this is **maximum a-posteriori (MAP) estimation**...

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$\mathrm{P}(x \,|\, y) \propto \mathrm{P}(y \,|\, x)\mathrm{P}(x) \,, \quad \textit{i.e.} \text{ posterior} \propto \text{likelihood} \times \text{prior}$$

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$\mathrm{P}(x \mid y) \propto \mathrm{P}(y \mid x)\mathrm{P}(x)\ , \quad \textit{i.e.}\ \text{posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$\mathrm{P}(\boldsymbol{x} \,|\, \boldsymbol{y}) \propto \mathrm{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) \mathrm{P}(\boldsymbol{x}) \;, \quad \textit{i.e.} \text{ posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$\boxed{\mathrm{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \exp\left(-\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / (2\sigma^2)\right)}$$

likelihood

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$\mathrm{P}(\boldsymbol{x} \,|\, \boldsymbol{y}) \propto \mathrm{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) \mathrm{P}(\boldsymbol{x}) \,, \quad \textit{i.e.} \text{ posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$\boxed{\mathrm{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \exp\Big(-\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / (2\sigma^2)\Big)} \qquad \boxed{\mathrm{P}(\boldsymbol{x}) \propto \exp\Big(-R(\boldsymbol{x})\Big)}$$

likelihood   prior

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$\mathrm{P}(\boldsymbol{x} \,|\, \boldsymbol{y}) \propto \mathrm{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) \mathrm{P}(\boldsymbol{x}) \ , \quad \textit{i.e.} \text{ posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$\mathrm{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \exp\Big(-\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / (2\sigma^2)\Big)$$

likelihood

$$\mathrm{P}(\boldsymbol{x}) \propto \exp\Big(-R(\boldsymbol{x})\Big)$$

prior

Consider log-posterior:

$$\log \mathrm{P}(\boldsymbol{x} \,|\, \boldsymbol{y}) = -\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / (2\sigma^2) - R(\boldsymbol{x}) + \text{const.}$$

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) \propto \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) \mathrm{P}(\boldsymbol{x}) , \quad \textit{i.e.}\ \text{posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$\mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) \propto \exp\left(-\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / (2\sigma^2)\right)$$

likelihood

$$\mathrm{P}(\boldsymbol{x}) \propto \exp\left(-R(\boldsymbol{x})\right)$$

prior

Consider log-posterior:

$$\log \mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = -\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / (2\sigma^2) - R(\boldsymbol{x}) + \text{const.}$$

MAP estimator:

$$\boldsymbol{x}_{\text{map}} = \arg\max_{\boldsymbol{x}} \left[\log \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x})\right]$$

MAP estimation and regularisation

Start with Bayes Theorem (ignore normalising evidence):

$$P(x \mid y) \propto P(y \mid x) P(x) \,, \quad \textit{i.e.}\ \text{posterior} \propto \text{likelihood} \times \text{prior}$$

Define likelihood (assuming Gaussian noise) and prior:

$$P(y \mid x) \propto \exp\left(-\|y - \mathbf{\Phi} x\|_2^2 / (2\sigma^2)\right)$$

likelihood

$$P(x) \propto \exp\left(-R(x)\right)$$

prior

Consider log-posterior:

$$\log P(x \mid y) = -\|y - \mathbf{\Phi} x\|_2^2 / (2\sigma^2) - R(x) + \text{const.}$$

MAP estimator:

$$x_{\text{map}} = \underset{x}{\arg\max} \left[\log P(y \mid x)\right] = \underset{x}{\arg\min} \left[\|y - \mathbf{\Phi} x\|_2^2 + \lambda R(x) \right]$$

data fidelity — regulariser

CLEAN and MEM as MAP estimators

- CLEAN

  Consider the sparse prior: $\mathrm{P}(\boldsymbol{x}) \propto \exp\left(-\beta \left\|\boldsymbol{x}\right\|_0\right)$.

CLEAN and MEM as MAP estimators

- CLEAN

  Consider the sparse prior: $\mathrm{P}(\boldsymbol{x}) \propto \exp\left(-\beta \left\|\boldsymbol{x}\right\|_0\right)$.

  Corresponding MAP estimator is:

$$\boldsymbol{x}_{\text{clean}} = \arg\min_{\boldsymbol{x}} \left[\left\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\right\|_2^2 + \lambda \left\|\boldsymbol{x}\right\|_0 \right]$$

  (Laplace prior $\mathrm{P}(\boldsymbol{x}) \propto \exp(-\beta \left\|\boldsymbol{x}\right\|_1)$ is good proxy)

CLEAN and MEM as MAP estimators

- CLEAN

  Consider the sparse prior: $P(x) \propto \exp\left(-\beta \left\|x\right\|_0\right)$.

  Corresponding MAP estimator is:

$$x_{\text{clean}} = \arg\min_x \left[\left\| y - \Phi x \right\|_2^2 + \lambda \left\| x \right\|_0 \right]$$

  (Laplace prior $P(x) \propto \exp(-\beta \left\|x\right\|_1)$ is good proxy)

- MEM

  Consider the entropic prior: $P(x) \propto \exp\left(-\beta \, x^\dagger \log x\right)$.

CLEAN and MEM as MAP estimators

- CLEAN

  Consider the sparse prior: $\text{P}(x) \propto \exp\left(-\beta \, \|x\|_0\right)$.

  Corresponding MAP estimator is:

$$x_{\text{clean}} = \arg\min_x \left[\|y - \Phi x\|_2^2 + \lambda \, \|x\|_0 \right]$$

  (Laplace prior $\text{P}(x) \propto \exp(-\beta \, \|x\|_1)$ is good proxy)

- MEM

  Consider the entropic prior: $\text{P}(x) \propto \exp\left(-\beta \, x^\dagger \log x\right)$.

  Corresponding MAP estimator is:

$$x_{\text{mem}} = \arg\min_x \left[\|y - \Phi x\|_2^2 + \lambda \, x^\dagger \log x \right]$$

CLEAN and MEM as MAP estimators

- CLEAN

  Consider the sparse prior: $\mathrm{P}(x) \propto \exp\left(-\beta \left\|x\right\|_0\right)$.

  Corresponding MAP estimator is:

$$x_{\mathrm{clean}} = \arg\min_x \left[\left\| y - \mathbf{\Phi} x \right\|_2^2 + \lambda \left\| x \right\|_0 \right]$$

  (Laplace prior $\mathrm{P}(x) \propto \exp\left(-\beta \left\|x\right\|_1\right)$ is good proxy)

- MEM

  Consider the entropic prior: $\mathrm{P}(x) \propto \exp\left(-\beta \, x^\dagger \log x\right)$.

  Corresponding MAP estimator is:

$$x_{\mathrm{mem}} = \arg\min_x \left[\left\| y - \mathbf{\Phi} x \right\|_2^2 + \lambda \, x^\dagger \log x \right]$$

(In practice some differences: CLEAN does not solve MAP problem exactly; MEM considered in RI imposes additional constraints.)

Sparse regularisation (*cf.* compressive sensing)

Sparse synthesis regularisation problem:

$$x_{\text{synthesis}} = \boldsymbol{\Psi} \times \arg\min_{\boldsymbol{\alpha}} \left[\left\| y - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 \right]$$

synthesis framework

where consider sparsifying (*e.g.* wavelet) representation of image: $x = \boldsymbol{\Psi}\boldsymbol{\alpha}$.

Sparse regularisation (*cf.* compressive sensing)

Sparse synthesis regularisation problem:

$$x_{\text{synthesis}} = \boldsymbol{\Psi} \times \underset{\boldsymbol{\alpha}}{\arg\min} \left[\left\| y - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 \right]$$

synthesis framework

where consider sparsifying (*e.g.* wavelet) representation of image: $x = \boldsymbol{\Psi}\boldsymbol{\alpha}$.

Sparse analysis regularisation problem (Elad *et al.* 2007, Nam *et al.* 2012):

$$x_{\text{analysis}} = \underset{x}{\arg\min} \left[\left\| y - \boldsymbol{\Phi} x \right\|_2^2 + \lambda \left\| \boldsymbol{\Psi}^\dagger x \right\|_1 \right]$$

analysis framework

Sparse regularisation (*cf.* compressive sensing)

Sparse **synthesis** regularisation problem:

$$x_{\text{synthesis}} = \mathbf{\Psi} \times \underset{\boldsymbol{\alpha}}{\arg\min} \left[\left\| y - \mathbf{\Phi}\mathbf{\Psi}\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 \right]$$

synthesis framework

where consider sparsifying (*e.g.* wavelet) representation of image: $x = \mathbf{\Psi}\boldsymbol{\alpha}$.

Sparse **analysis** regularisation problem (Elad *et al.* 2007, Nam *et al.* 2012):

$$x_{\text{analysis}} = \underset{x}{\arg\min} \left[\left\| y - \mathbf{\Phi} x \right\|_2^2 + \lambda \left\| \mathbf{\Psi}^\dagger x \right\|_1 \right]$$

analysis framework

More sophisticated extensions (*e.g.* overcomplete dictionaries, constrained vs unconstrained, re-weighting).

MAP estimation vs MCMC sampling

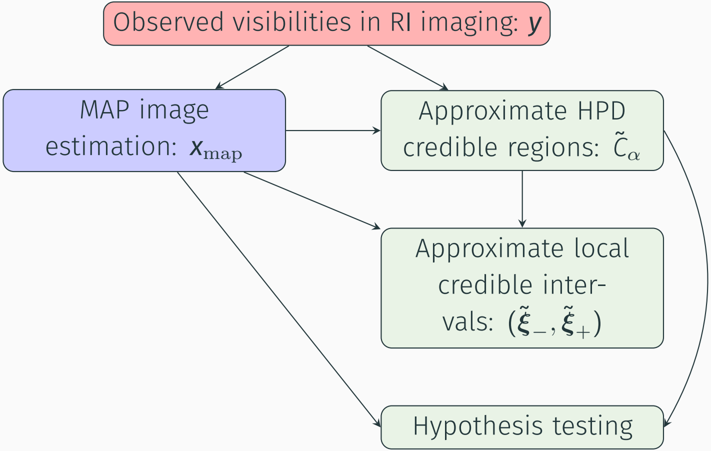
MAP estimation

- \+ Based on optimization so computationally efficient.
- – Does not traditionally provide uncertainties.

MCMC sampling

- – Based on sampling so computationally demanding.
- \+ Recover full posterior distribution.

MAP estimation and uncertainty quantification



Approximate Bayesian credible regions for MAP estimation

Combine **uncertainty quantification and scalable MAP estimation** (sparse regularisation) to scale to big-data (Cai, Pereyra & McEwen 2017, 2018; arXiv:1711.04819; arXiv:1811.02514).

Approximate Bayesian credible regions for MAP estimation

Combine uncertainty quantification and scalable MAP estimation (sparse regularisation) to scale to big-data (Cai, Pereyra & McEwen 2017, 2018; arXiv:1711.04819; arXiv:1811.02514).

Let C_α denote the **highest posterior density (HPD) Bayesian credible region** with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\mathbf{x} : g(\mathbf{x}) \leq \gamma_\alpha\}$.

Approximate Bayesian credible regions for MAP estimation

Combine **uncertainty quantification and scalable MAP estimation** (sparse regularisation) to scale to big-data (Cai, Pereyra & McEwen 2017, 2018; arXiv:1711.04819; arXiv:1811.02514).

Let C_α denote the **highest posterior density (HPD) Bayesian credible region** with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{x : g(x) \leq \gamma_\alpha\}$.

Analytic approximation of γ_α:

$$\tilde{\gamma}_\alpha = g(x^\star) + N(\tau_\alpha + 1)$$

where $\tau_\alpha = \sqrt{16 \log(3/\alpha)/N}$ and $\alpha \in (4\exp(-N/3), 1)$ (Pereyra 2016b). Define **approximate HPD regions** by $\tilde{C}_\alpha = \{x : g(x) \leq \tilde{\gamma}_\alpha\}$.

Compute $x^\star$ by MAP estimation (optimization), then **estimate local Bayesian credible intervals** and perform **uncertainty quantification** using approximate HPD regions.

Local Bayesian credible intervals for MAP estimation

(Cai, Pereyra & McEwen 2017, 2018; arXiv:1711.04819; arXiv:1811.02514)

Let Ω define the area (or pixel) over which to compute the credible interval $(\tilde{\xi}_-, \tilde{\xi}_+)$ and $\boldsymbol{\zeta}$ be an index vector describing Ω (*i.e.* $\zeta_i = 1$ if $i \in \Omega$ and 0 otherwise).

Consider the test image with the Ω region replaced by constant value ξ:

$$\boxed{\boldsymbol{x}' = \boldsymbol{x}^\star(\mathcal{I} - \boldsymbol{\zeta}) + \xi\boldsymbol{\zeta}} .$$

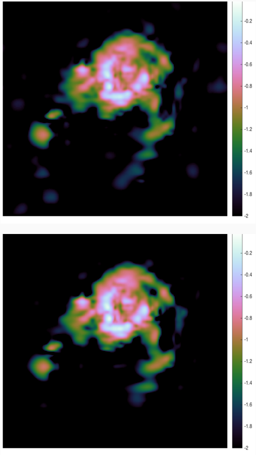
Given $\tilde{\gamma}_\alpha$ and $\boldsymbol{x}^\star$, compute the credible interval by

$$\boxed{\begin{aligned} \tilde{\xi}_- &= \min_{\xi} \left\{ \xi \mid g_{\boldsymbol{y}}(\boldsymbol{x}') \leq \tilde{\gamma}_\alpha, \ \forall \xi \in [-\infty, +\infty) \right\} , \\ \tilde{\xi}_+ &= \max_{\xi} \left\{ \xi \mid g_{\boldsymbol{y}}(\boldsymbol{x}') \leq \tilde{\gamma}_\alpha, \ \forall \xi \in [-\infty, +\infty) \right\} . \end{aligned}}$$

Local credible intervals for M31 experiment

Prox MCMC

MAP



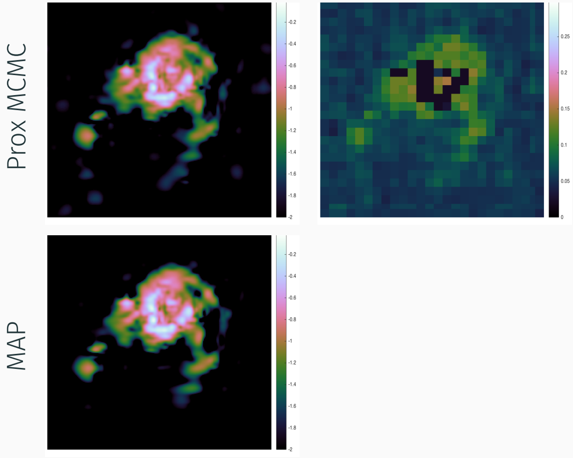
(a) point estimators

(b) local credible interval (grid size 10 × 10 pixels)

(c) local credible interval (grid size 20 × 20 pixels)

(d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for M31 experiment



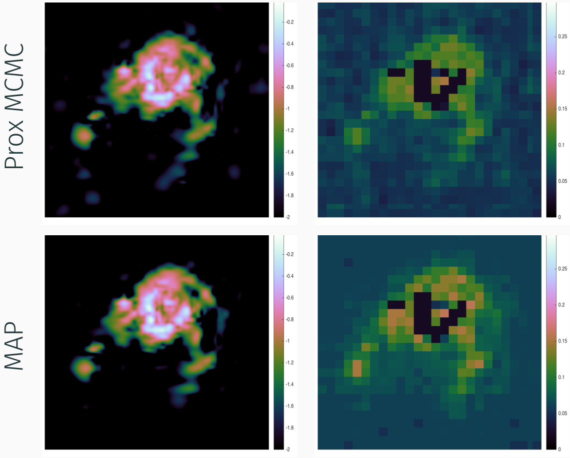
(a) point estimators

(b) local credible interval (grid size 10 × 10 pixels)

(c) local credible interval (grid size 20 × 20 pixels)

(d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for M31 experiment



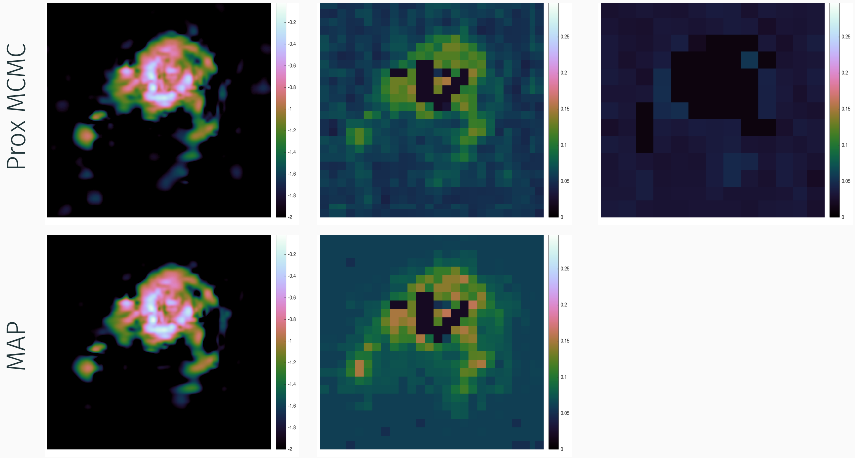
(a) point estimators

(b) local credible interval (grid size 10 × 10 pixels)

(c) local credible interval (grid size 20 × 20 pixels)

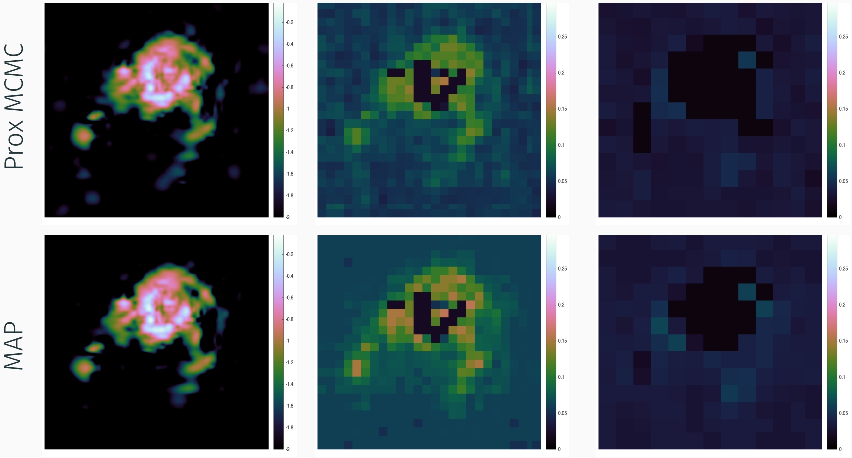
(d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for M31 experiment



(a) point estimators (b) local credible interval (grid size 10 × 10 pixels) (c) local credible interval (grid size 20 × 20 pixels) (d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for M31 experiment



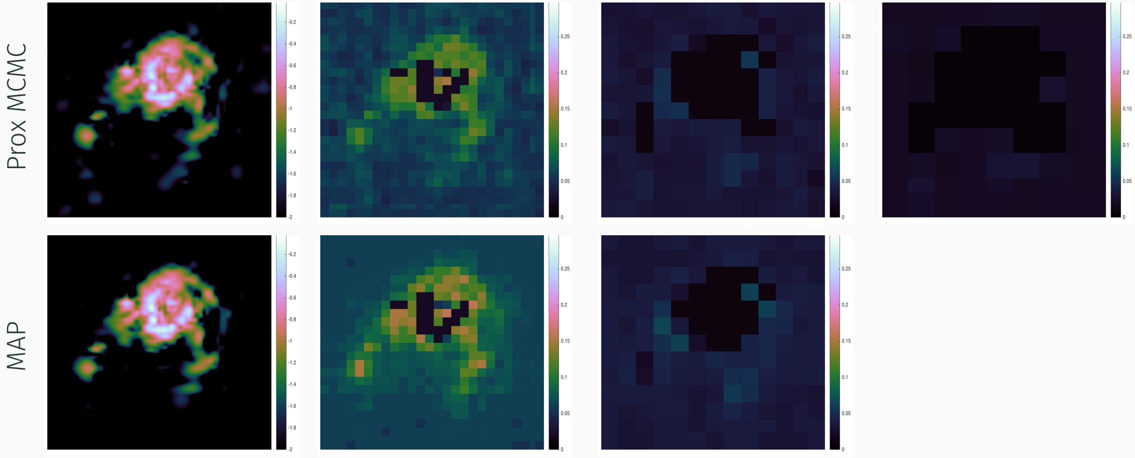
(a) point estimators

(b) local credible interval (grid size 10×10 pixels)

(c) local credible interval (grid size 20×20 pixels)

(d) local credible interval (grid size 30×30 pixels)

Local credible intervals for M31 experiment



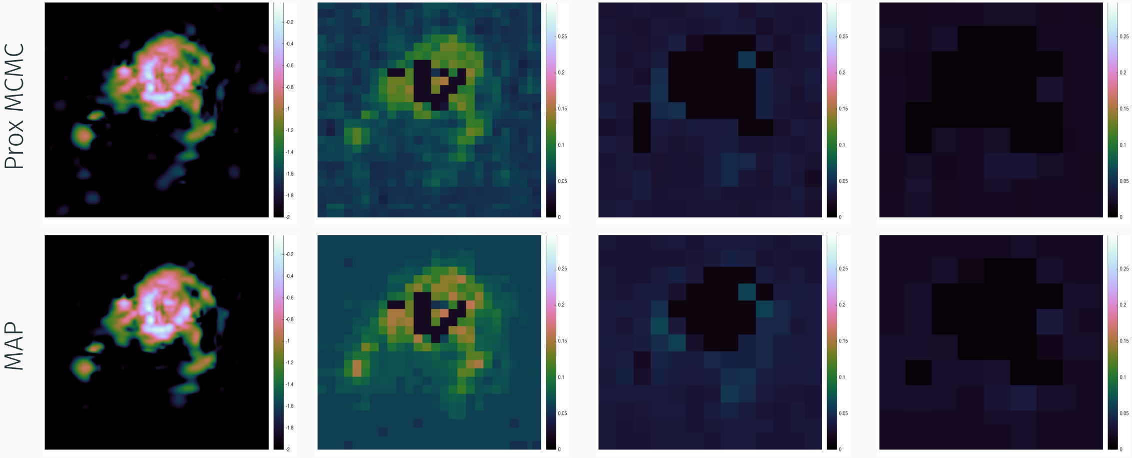
(a) point estimators

(b) local credible interval (grid size 10×10 pixels)

(c) local credible interval (grid size 20×20 pixels)

(d) local credible interval (grid size 30×30 pixels)

Local credible intervals for M31 experiment



(a) point estimators

(b) local credible interval (grid size 10 × 10 pixels)

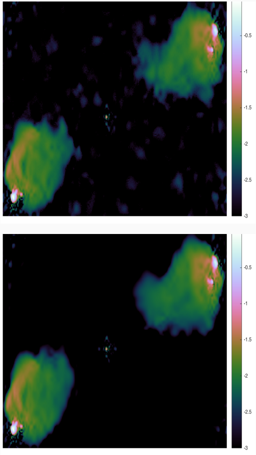
(c) local credible interval (grid size 20 × 20 pixels)

(d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for Cygnus A experiment

Prox MCMC

MAP

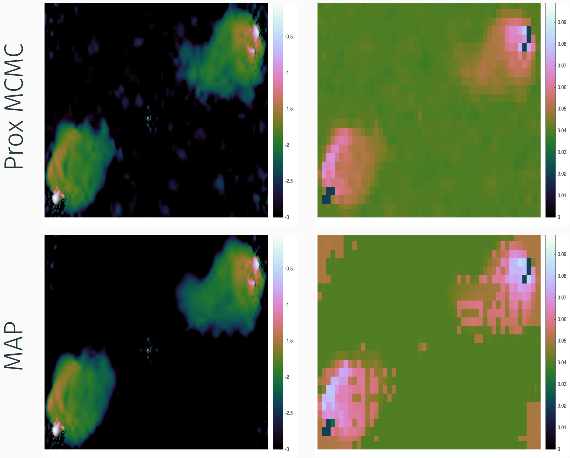


(a) point estimators

(b) local credible interval (grid size 10 × 10 pixels)

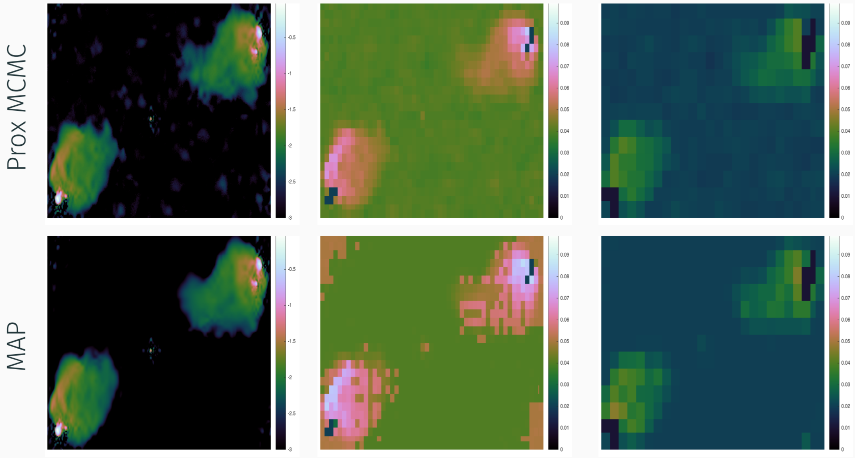
(c) local credible interval (grid size 20 × 20 pixels)

(d) local credible interval (grid size 30 × 30 pixels)



(a) point estimators (b) local credible interval (grid size 10 × 10 pixels) (c) local credible interval (grid size 20 × 20 pixels) (d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for Cygnus A experiment



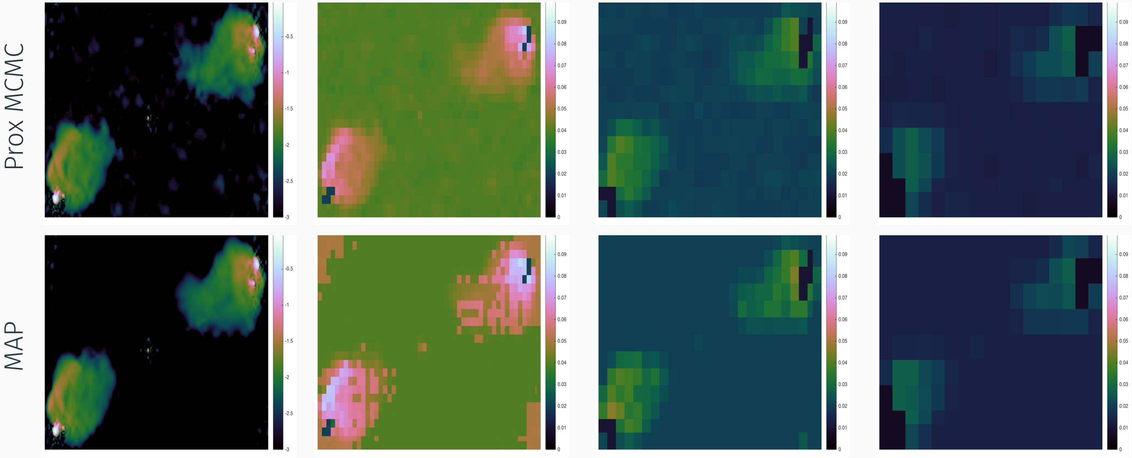
(a) point estimators

(b) local credible interval (grid size 10 × 10 pixels)

(c) local credible interval (grid size 20 × 20 pixels)

(d) local credible interval (grid size 30 × 30 pixels)

Local credible intervals for Cygnus A experiment



(a) point estimators (b) local credible interval (grid size 10 × 10 pixels) (c) local credible interval (grid size 20 × 20 pixels) (d) local credible interval (grid size 30 × 30 pixels)

Hypothesis testing

Is structure in an image physical or an artifact?

Hypothesis testing

Is structure in an image physical or an artifact?

Perform **hypothesis tests** of image structure using Bayesian credible regions (Pereyra 2016b).

Hypothesis testing

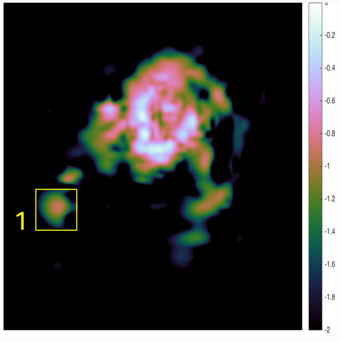
Is structure in an image physical or an artifact?

Perform **hypothesis tests** of image structure using Bayesian credible regions (Pereyra 2016b).

Hypothesis testing of physical structure

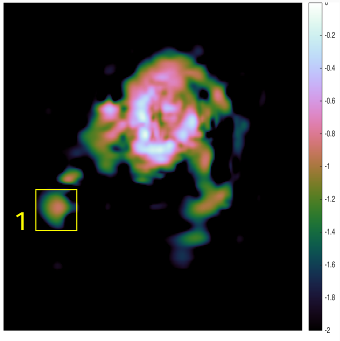
1. Remove structure of interest from recovered image $\boldsymbol{x}^\star$.
2. Inpaint background (noise) into region, yielding surrogate image $\boldsymbol{x}'$.
3. Test whether $\boldsymbol{x}' \in C_\alpha$:
    - If $\boldsymbol{x}' \notin C_\alpha$ then reject hypothesis that structure is an artifact with confidence $(1-\alpha)\%$, *i.e.* structure most likely physical.
    - If $\boldsymbol{x}' \in C_\alpha$ uncertainly too high to draw strong conclusions about the physical nature of the structure.

Hypothesis testing for M31 experiment

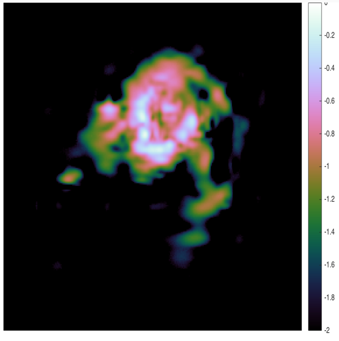


Recovered image

Hypothesis testing for M31 experiment

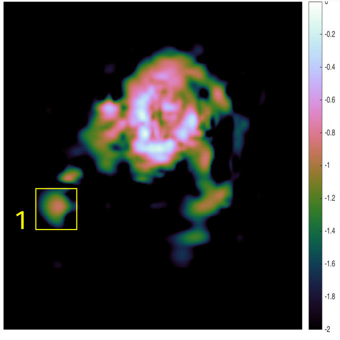


Recovered image

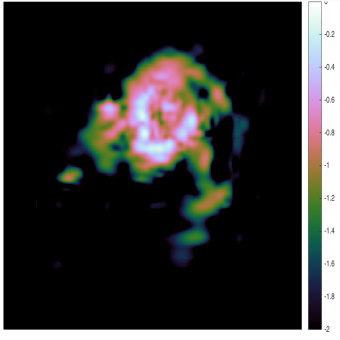


Surrogate with region removed

Hypothesis testing for M31 experiment



Recovered image

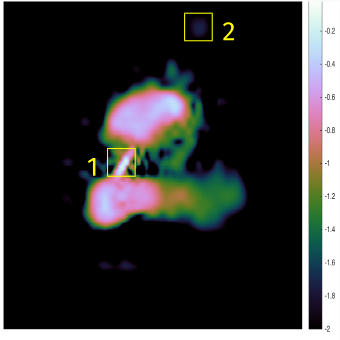


Surrogate with region removed

1. Reject null hypothesis

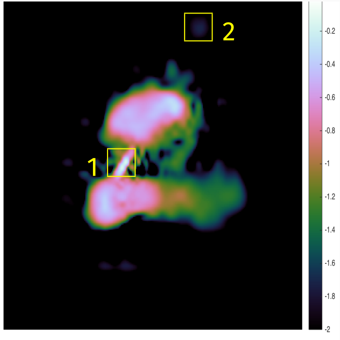
$\Rightarrow$ **structure physical**

Hypothesis testing for 3C288 experiment

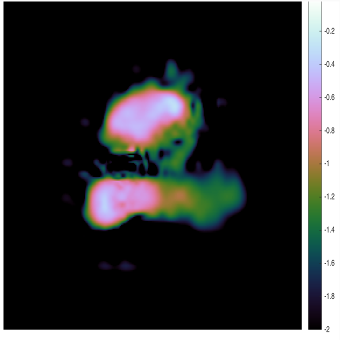


Recovered image

Hypothesis testing for 3C288 experiment

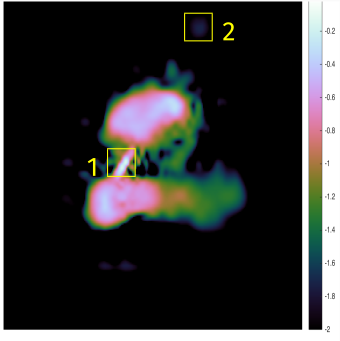


Recovered image

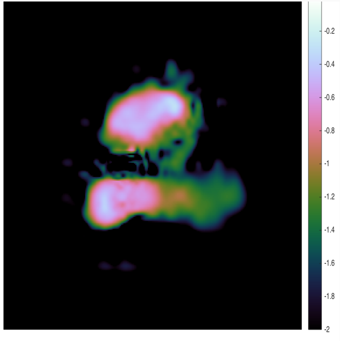


Surrogate with region removed

Hypothesis testing for 3C288 experiment



Recovered image



Surrogate with region removed

1. Reject null hypothesis

   $\Rightarrow$ **structure physical**

2. Cannot reject null hypothesis

   $\Rightarrow$ **cannot make strong statistical statement about origin of structure**

Computation time

CPU time in minutes for Proximal MCMC sampling and MAP estimation

Image	Method	CPU time	
		Analysis	Synthesis
Cygnus A	P-MALA	2274	1762
	MYULA	1056	942
	MAP	.07	.04
M31	P-MALA	1307	944
	MYULA	618	581
	MAP	.03	.02
3C288	P-MALA	1144	881
	MYULA	607	538
	MAP	.03	.02

Summary

1. Learnt harmonic mean estimator for Bayesian model comparison
   (McEwen *et al.* 2021; arXiv:2111.12720)

2. Proximal nested sampling for high-dimensional Bayesian model comparison
   (Cai, McEwen & Pereyra 2021; arXiv:2106.03646)

3. High-dimensional Bayesian uncertainty quantification for extreme computation
   (Cai, Pereyra & McEwen 2017, 2018; arXiv:1711.04819, arXiv:1811.02514)