High-dimensional uncertainty quantification in astrophysics

(with application to radio interferometric imaging, weak lensing, and beyond)

Jason McEwen
www.jasonmcewen.org
@jasonmcewen

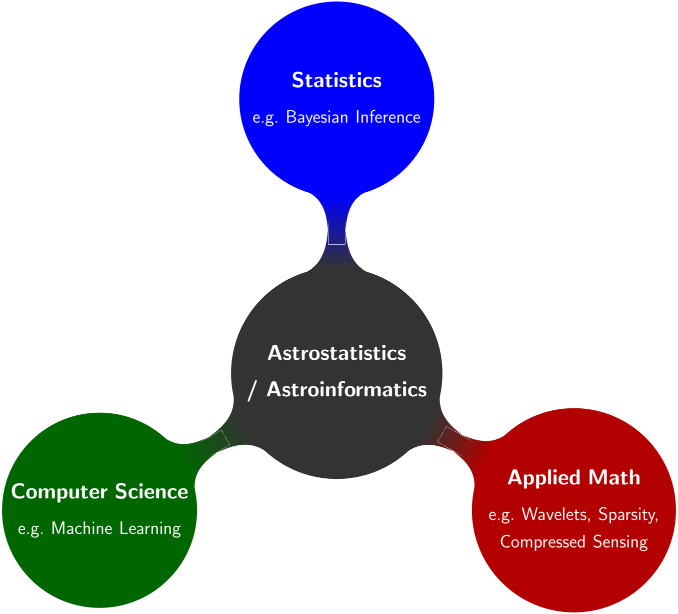
Mullard Space Science Laboratory (MSSL)
University College London (UCL)

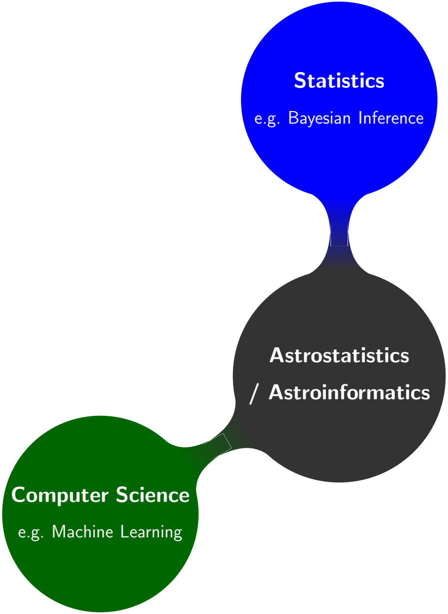
Cai, Pereyra & McEwen (2017a): arXiv:1711.04818
Cai, Pereyra & McEwen (2017b): arXiv:1711.04819
Cai, Pereyra & McEwen (2018): arXiv:1811.02514

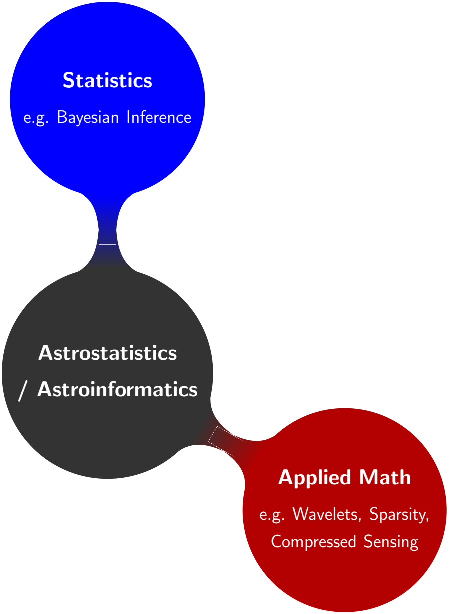
Price, McEwen, Cai, Kitching, Wallis (2018a): arXiv:1812.04014
Price, Cai, McEwen, Pereyra, Kitching (2018b): arXiv:1812.04017
Price, McEwen, Cai, Kitching (2018c): arXiv:1812.04018

Astrophysics Seminar, Imperial College

December 2018


Statistics
e.g. Bayesian Inference
Astrostatistics / Astroinformatics
Computer Science
e.g. Machine Learning
Applied Math
e.g. Wavelets, Sparsity, Compressed Sensing


Statistics
e.g. Bayesian Inference
Astrostatistics
/ Astroinformatics
Computer Science
e.g. Machine Learning

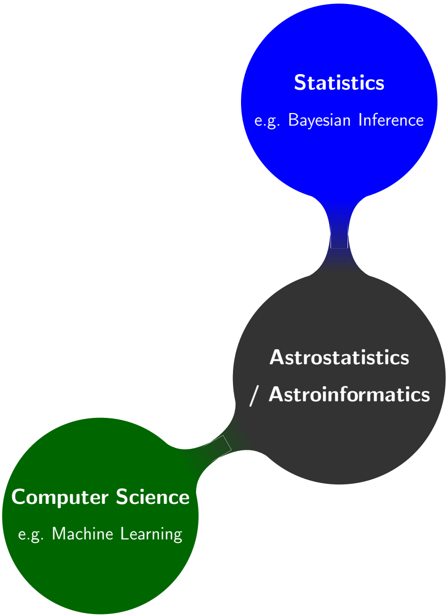

Statistics
e.g. Bayesian Inference
Astrostatistics
/ Astroinformatics
Applied Math
e.g. Wavelets, Sparsity,
Compressed Sensing

Outline

1. Learnt harmonic mean estimator
2. Radio interferometric imaging
3. Proximal MCMC sampling and uncertainty quantification
4. MAP estimation and uncertainty quantification
5. Mass-mapping via weak gravitational lensing

Outline

1. Learnt harmonic mean estimator
2. Radio interferometric imaging
3. Proximal MCMC sampling and uncertainty quantification
4. MAP estimation and uncertainty quantification
5. Mass-mapping via weak gravitational lensing


Statistics
e.g. Bayesian Inference
Astrostatistics / Astroinformatics
Computer Science
e.g. Machine Learning

Bayesian inference

Parameter estimation

Bayes' theorem

$$\mathsf{P}(\theta \,|\, \boldsymbol{y}, M) = \frac{\mathsf{P}(\boldsymbol{y} \,|\, \theta, M)\, \mathsf{P}(\theta \,|\, M)}{\mathsf{P}(\boldsymbol{y} \,|\, M)} \; ,$$

for parameters θ, model M and observed data $\boldsymbol{y}$.

Bayesian inference

Parameter estimation

Bayes' theorem

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

for parameters θ, model M and observed data $\boldsymbol{y}$.

Bayesian inference
Parameter estimation

Bayes' theorem

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

for parameters θ, model M and observed data $\boldsymbol{y}$.

Shorthand notation:

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

Bayesian inference

Parameter estimation

Bayes' theorem

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

for parameters θ, model M and observed data $\boldsymbol{y}$.

Shorthand notation:

$$\underbrace{\boxed{\mathrm{P}(\theta \mid \boldsymbol{y})}}_{\text{posterior}} = \frac{\overbrace{\boxed{\mathcal{L}(\theta)}}^{\text{likelihood}} \overbrace{\boxed{\pi(\theta)}}^{\text{prior}}}{\underbrace{\boxed{z}}_{\text{constant}}},$$

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

Bayesian inference
Model selection

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

$$\frac{\mathrm{P}(M_1 \mid \boldsymbol{y})}{\mathrm{P}(M_2 \mid \boldsymbol{y})} = \frac{\mathrm{P}(M_1)}{\mathrm{P}(M_2)} \times \frac{\mathrm{P}(\boldsymbol{y} \mid M_1)}{\mathrm{P}(\boldsymbol{y} \mid M_2)} .$$

Bayesian inference

Model selection

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

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

Bayesian inference

Model selection

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

$$\boxed{\frac{\mathrm{P}(M_1 \mid \boldsymbol{y})}{\mathrm{P}(M_2 \mid \boldsymbol{y})}} = \boxed{\frac{\mathrm{P}(M_1)}{\mathrm{P}(M_2)}} \times \boxed{\frac{\mathrm{P}(\boldsymbol{y} \mid M_1)}{\mathrm{P}(\boldsymbol{y} \mid M_2)}} .$$

posterior odds prior odds Bayes factor

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

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

Bayesian inference

Model selection

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

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

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

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

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

Bayesian inference

Model selection

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

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

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

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

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

Variety of powerful methods exist:

- Nested sampling (Skilling 2004), *e.g.* MultiNest (Feroz, Hobson, Bridges 2008), PolyCord (Handley, Hobson, Lasenby 2015)
- Heavens *et al.* (2017)

Desirable properties for Bayesian evidence estimators

Seek estimator that is:

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

Desirable properties for Bayesian evidence estimators

Seek estimator that is:

- Agnostic to sampling method and **uses posterior samples**.
- Scales 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)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

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

$$= \frac{1}{z}$$

Original harmonic mean estimator (Newton & Raftery 1994)

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

Original harmonic mean estimator

Harmonic mean relationship (Newton & Raftery 1994)

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

Original harmonic mean estimator (Newton & Raftery 1994)

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

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

Original harmonic mean estimator

Importance sampling interpretation

Alternative derivation of harmonic mean relationship:

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

Original harmonic mean estimator

Importance sampling interpretation

Alternative derivation of harmonic mean relationship:

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

Importance sampling interpretation:

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

Original harmonic mean estimator

Importance sampling interpretation

Alternative derivation of harmonic mean relationship:

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

Importance sampling interpretation:

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

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

Original harmonic mean estimator

Importance sampling interpretation

Alternative derivation of harmonic mean relationship:

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

Importance sampling interpretation:

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

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

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

Original harmonic mean estimator

Simulation pseudo bias

Simulation pseudo bias (Lenk 2009)

In practice posterior simulation support Ω is a subset of the prior support Θ, hence do not fully capture prior (target distribution).

Original harmonic mean estimator

Simulation pseudo bias

Simulation pseudo bias (Lenk 2009)

In practice posterior simulation support Ω is a subset of the prior support Θ, hence do not fully capture prior (target distribution).

Corrected harmonic mean estimator (Lenk 2009)

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

where $\mathrm{P}(\Omega)$ is the prior probability of the posterior simulation support $\Omega \subset \Theta$.

Original harmonic mean estimator

Simulation pseudo bias

Simulation pseudo bias (Lenk 2009)

In practice posterior simulation support Ω is a subset of the prior support Θ, hence do not fully capture prior (target distribution).

Corrected harmonic mean estimator (Lenk 2009)

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

where $\mathrm{P}(\Omega)$ is the prior probability of the posterior simulation support $\Omega \subset \Theta$.

Mitigates simulation pseudo bias but does not eliminate.

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)

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

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)

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

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

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)

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

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)

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

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

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

Re-targeted harmonic mean estimator

Importance sampling interpretation

Importance sampling derivation:

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

Re-targeted harmonic mean estimator

Importance sampling interpretation

Importance sampling derivation:

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

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

Re-targeted harmonic mean estimator

Importance sampling interpretation

Importance sampling derivation:

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

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

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

Re-targeted harmonic mean estimator

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)

Re-targeted harmonic mean estimator

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:

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

(resulting estimator has zero variance).

Re-targeted harmonic mean estimator

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:

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

(resulting estimator has zero variance).

Recall:

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

Re-targeted harmonic mean estimator

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:

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

(resulting estimator has zero variance).

Recall:

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

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

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

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

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.

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

Learnt harmonic mean estimator

Learning the target distribution

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

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

Modify learning objective function to include **variance penalty and regularisation**.

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

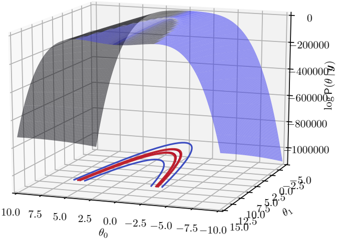
Cross-validation to select machine learning approach and hyperparameters.

Rosenbrock example

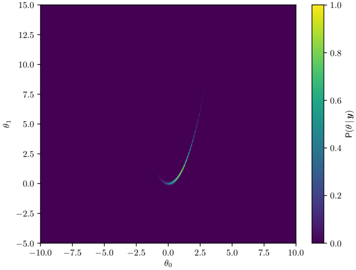
Posterior

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

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



(a) Log-Posterior



(b) Posterior

Figure: Rosenbrock posterior evaluated on grid.

Rosenbrock example

MCMC sampling and learning the target distribution φ

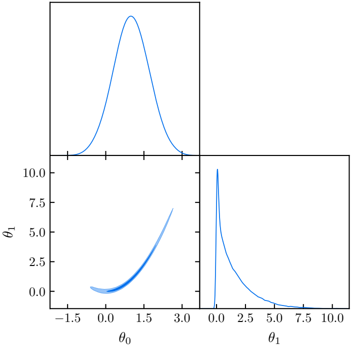


Figure: Posterior recovered by MCMC sampling.

Rosenbrock example

MCMC sampling and learning the target distribution φ

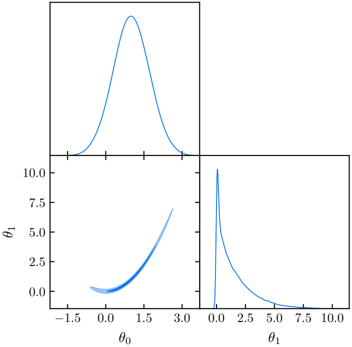


Figure: Posterior recovered by MCMC sampling.

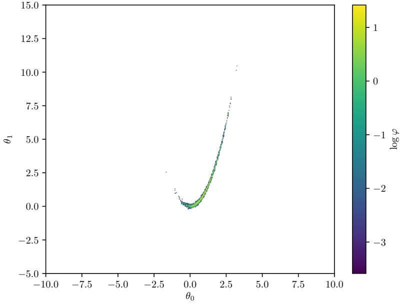
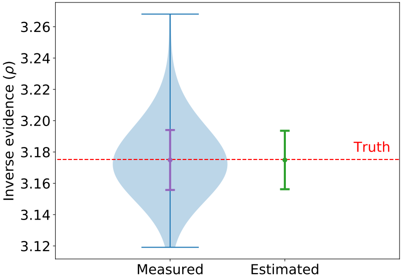


Figure: Learnt target distribution φ (by KDE).

Rosenbrock example

Accuracy of learnt harmonic mean estimator

- Compare to Monte Carlo simulations, repeating entire analysis.
- Also estimate the variance of the estimator and its variance.



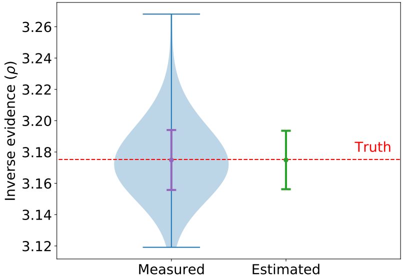
(a) **Inverse evidence**

Figure: Accuracy of learnt harmonic mean estimator for Rosenbrock example.

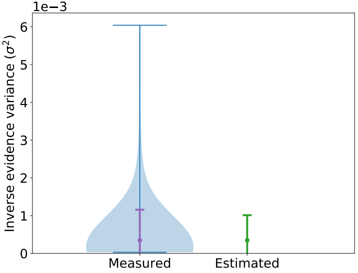
Rosenbrock example

Accuracy of learnt harmonic mean estimator

- Compare to Monte Carlo simulations, repeating entire analysis.
- Also estimate the variance of the estimator and its variance.



(a) Inverse evidence



(b) Variance of inverse evidence

Figure: Accuracy of learnt harmonic mean estimator for Rosenbrock example.

Normal-Gamma example
Model

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

Normal-Gamma example
Model

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

Data model:

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

Prior model:

$$\text{Mean: } \mu \sim \mathrm{N}\big(\mu_0, (\tau_0\tau)^{-1}\big)$$
$$\text{Precision: } \tau \sim \mathrm{Ga}(a_0, b_0)$$

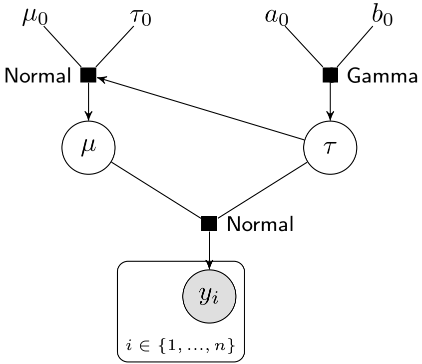


Figure: Graph of hierarchical Bayesian model of Normal-Gamma example.

Normal-Gamma example

Analytic evidence

Analytic evidence:

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

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

Accuracy of learnt harmonic mean estimator and sensitivity to prior

Table: Analytic and estimated evidence for various prior sizes τ_0.

Prior size τ_0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Analytic $\log(z)$	-160.3888	-159.2375	-158.0863	-156.9359	-155.7935
Estimated $\log(\hat{z})$	-160.3883	-159.2370	-158.0851	-156.9359	-155.7921
Error (learnt harmonic mean)	-0.0005	-0.0005	-0.0012	0.0000	-0.0014
Error (original harmonic mean)*	-12.2100	–	-9.7900	-8.5000	-7.1000

*Friel & Wyse (2012)

Normal-Gamma example

Accuracy of learnt harmonic mean estimator and sensitivity to prior

Table: Analytic and estimated evidence for various prior sizes τ_0.

Prior size τ_0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Analytic $\log(z)$	-160.3888	-159.2375	-158.0863	-156.9359	-155.7935
Estimated $\log(\hat{z})$	-160.3883	-159.2370	-158.0851	-156.9359	-155.7921
Error (learnt harmonic mean)	-0.0005	-0.0005	-0.0012	0.0000	-0.0014
Error (original harmonic mean)*	-12.2100	–	-9.7900	-8.5000	-7.1000

*Friel & Wyse (2012)

Normal-Gamma example

Accuracy of learnt harmonic mean estimator and sensitivity to prior

Table: Analytic and estimated evidence for various prior sizes τ_0.

Prior size τ_0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Analytic $\log(z)$	-160.3888	-159.2375	-158.0863	-156.9359	-155.7935
Estimated $\log(\hat{z})$	-160.3883	-159.2370	-158.0851	-156.9359	-155.7921
Error (learnt harmonic mean)	**-0.0005**	**-0.0005**	**-0.0012**	**0.0000**	**-0.0014**
Error (original harmonic mean)*	-12.2100	–	-9.7900	-8.5000	-7.1000

*Friel & Wyse (2012)

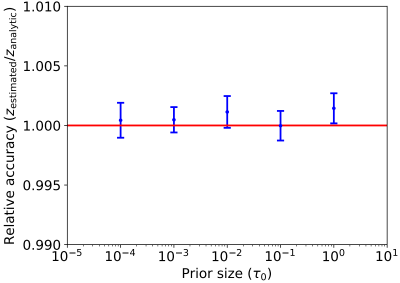


Figure: Accuracy for various prior sizes τ_0.

Normal-Gamma example

Accuracy of learnt harmonic mean estimator and sensitivity to prior

Table: Analytic and estimated evidence for various prior sizes τ_0.

Prior size τ_0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Analytic $\log(z)$	-160.3888	-159.2375	-158.0863	-156.9359	-155.7935
Estimated $\log(\hat{z})$	-160.3883	-159.2370	-158.0851	-156.9359	-155.7921
Error (learnt harmonic mean)	-0.0005	-0.0005	-0.0012	0.0000	-0.0014
Error (original harmonic mean)*	-12.2100	–	-9.7900	-8.5000	-7.1000

*Friel & Wyse (2012)

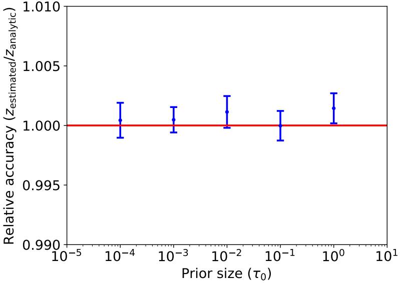


Figure: Accuracy for various prior sizes τ_0.

Non-nested linear regression: Radiata pine example

Data

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

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

Non-nested linear regression: Radiata pine example

Data

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?

Non-nested linear regression: Radiata pine example

Models

Gaussian linear models:

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

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

Priors for model 1 (similar for model 2):

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

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

Non-nested linear regression: Radiata pine example

Models

Gaussian linear models:

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

$$M_2 : \quad y_i = \gamma + \underbrace{\boxed{\delta(z_i - \bar{z})}}_{\text{Resin-adjusted density}} + \eta_i \, , \qquad \eta_i \sim \mathsf{N}(0, \lambda^{-1}) \, .$$

Priors for model 1 (similar for model 2):

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

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

Non-nested linear regression: Radiata pine example

Models

Gaussian linear models:

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

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

Priors for model 1 (similar for model 2):

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

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

Non-nested linear regression: Radiata pine example

Models

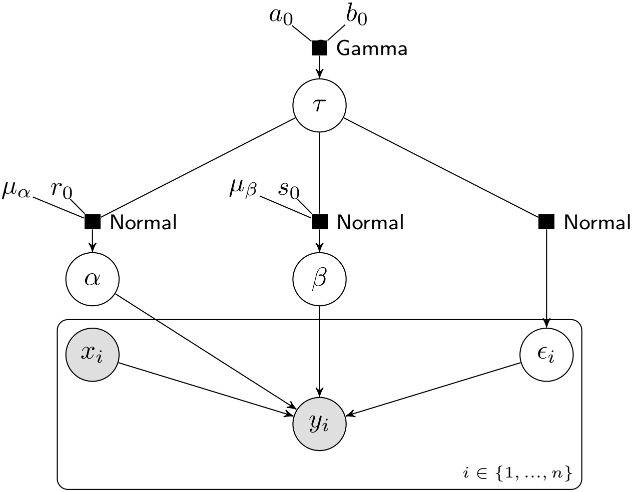


Figure: Graph of hierarchical Bayesian model for Radiata pine example (for model 1; model 2 is similar).

Non-nested linear regression: Radiata pine example

Analytic evidence

Analytic evidence:

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

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

Non-nested linear regression: Radiata pine example

Accuracy of learnt harmonic mean estimator

Table: Analytic and estimated evidence.

	Model M_1 $\log(z_1)$	Model M_2 $\log(z_2)$	$\log$ **BF**$_{21}$ $= \log(z_2) - \log(z_1)$
Analytic	-310.12833	-301.70460	8.42368
Estimated	-310.12839	-301.70489	8.42350
Error (learnt harmonic mean)	0.00006	0.00029	0.00018
Error (original harmonic mean)*	–	–	0.17372

*Friel & Wyse (2012)

Non-nested linear regression: Radiata pine example

Accuracy of learnt harmonic mean estimator

Table: Analytic and estimated evidence.

	Model M_1 $\log(z_1)$	Model M_2 $\log(z_2)$	$\log \mathsf{BF}_{21}$ $= \log(z_2) - \log(z_1)$
Analytic	-310.12833	-301.70460	8.42368
Estimated	-310.12839	-301.70489	8.42350
Error (learnt harmonic mean)	0.00006	0.00029	0.00018
Error (original harmonic mean)*	–	–	0.17372

*Friel & Wyse (2012)

Non-nested linear regression: Radiata pine example

Accuracy of learnt harmonic mean estimator

Table: Analytic and estimated evidence.

	Model M_1 $\log(z_1)$	Model M_2 $\log(z_2)$	$\log \textbf{BF}_{21} = \log(z_2) - \log(z_1)$
Analytic	-310.12833	-301.70460	8.42368
Estimated	-310.12839	-301.70489	8.42350
Error (learnt harmonic mean)	**0.00006**	**0.00029**	**0.00018**
Error (original harmonic mean)*	–	–	0.17372

*Friel & Wyse (2012)

Non-nested linear regression: Radiata pine example

Accuracy of learnt harmonic mean estimator

Table: Analytic and estimated evidence.

	Model M_1 $\log(z_1)$	Model M_2 $\log(z_2)$	$\log \mathbf{BF}_{21} = \log(z_2) - \log(z_1)$
Analytic	-310.12833	-301.70460	8.42368
Estimated	-310.12839	-301.70489	8.42350
Error (learnt harmonic mean)	**0.00006**	**0.00029**	**0.00018**
Error (original harmonic mean)*	–	–	**0.17372**

*Friel & Wyse (2012)

Code

Python package: **harmonic**

Harmonic python package implementing *learnt* harmonic mean estimator.

User-facing features:

- **Ease of use** (modular python package).
- Follow **software engineering best-practice** (*e.g.* well documented, extensive test suite, CI).
- Cython for **speed**.
- **Flexible** choice of sampler (we use **emcee**).
- Bespoke integrated **cross-validation** to select machine learning algorithm and hyperparameters.

Under the hood:

- Bespoke objective functions with **variance penalty** and **regularisation**.
- Solve by bespoke **mini-batch stochastic gradient descent**.

Code

Python package: **harmonic**

Harmonic python package implementing *learnt* harmonic mean estimator.

User-facing features:

- **Ease of use** (modular python package).
- Follow **software engineering best-practice** (*e.g.* well documented, extensive test suite, CI).
- Cython for **speed**.
- **Flexible** choice of sampler (we use **emcee**).
- Bespoke integrated **cross-validation** to select machine learning algorithm and hyperparameters.

Under the hood:

- Bespoke objective functions with **variance penalty** and **regularisation**.
- Solve by bespoke **mini-batch stochastic gradient descent**.

Code

Python package: **harmonic**

Harmonic python package implementing *learnt* harmonic mean estimator.

User-facing features:

- **Ease of use** (modular python package).
- Follow **software engineering best-practice** (*e.g.* well documented, extensive test suite, CI).
- Cython for **speed**.
- **Flexible** choice of sampler (we use **emcee**).
- Bespoke integrated **cross-validation** to select machine learning algorithm and hyperparameters.

Under the hood:

- Bespoke objective functions with **variance penalty** and **regularisation**.
- Solve by bespoke **mini-batch stochastic gradient descent**.

Code

Pseudo code example

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

Code

Pseudo 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:])
```

Code

Pseudo 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)
```

Code

Pseudo 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)
```

Code

Pseudo 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()
```

Summary and future work

Problems of harmonic mean estimator can be fixed by re-targeting.

Apply machine learning to approximate optimal importance sampling target.

$\Rightarrow$ ***Learnt* harmonic mean estimator**

Future work:

- Finalising paper.
- Numerical optimisations.
- Apply to more examples and push to higher dimensions.
- Make code public.
- Extend general approach to other statistical problems (*e.g.* learnt importance sampling distributions, learnt proposal distributions).

Summary and future work

Problems of harmonic mean estimator can be fixed by re-targeting.

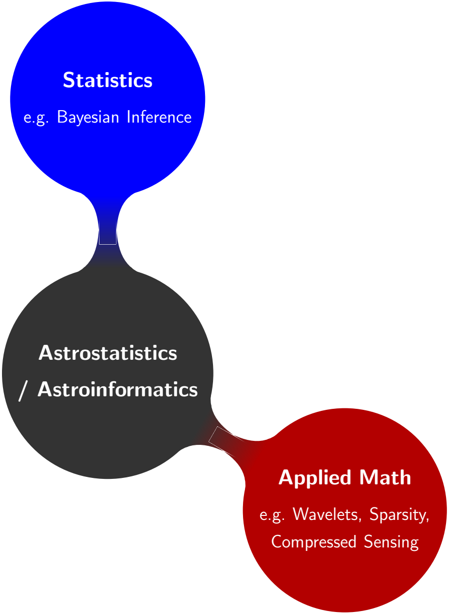
Apply machine learning to approximate optimal importance sampling target.

$\Rightarrow$ ***Learnt* harmonic mean estimator**

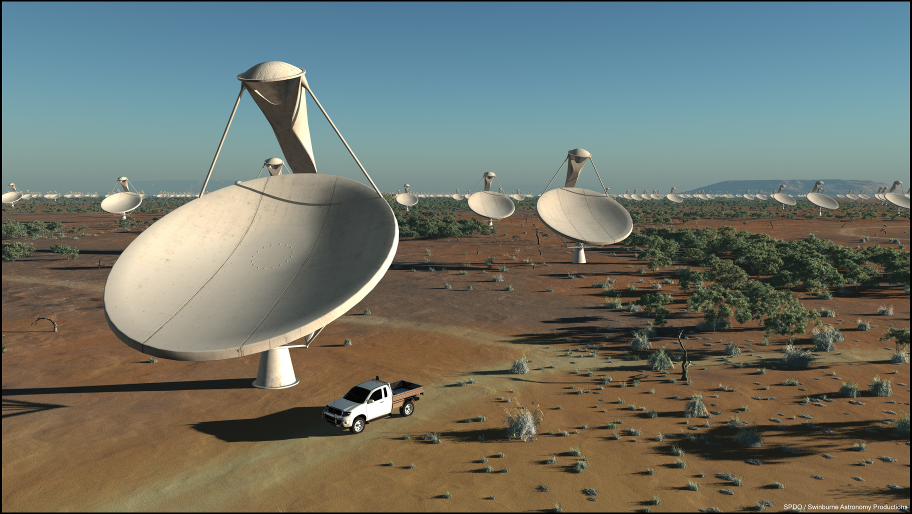
Future work:

- Finalising paper.
- Numerical optimisations.
- Apply to more examples and push to higher dimensions.
- Make code public.
- Extend general approach to other statistical problems (*e.g.* learnt importance sampling distributions, learnt proposal distributions).

Outline


Statistics
e.g. Bayesian Inference
Astrostatistics
/ Astroinformatics
Applied Math
e.g. Wavelets, Sparsity,
Compressed Sensing

Square Kilometre Array (SKA)

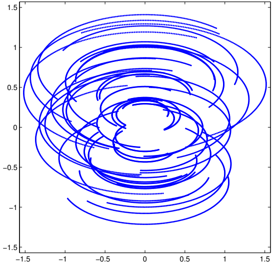


The SKA poses a considerable big-data challenge

The SKA poses a considerable big-data challenge

Radio interferometric telescopes acquire "Fourier" measurements

"Fourier"
Measurements



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.* $\boxed{\boldsymbol{\Phi} = \mathbf{GFA}}$, 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.

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.* $\boxed{\boldsymbol{\Phi} = \mathbf{GFA}}$, 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.

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.* $\boxed{\boldsymbol{\Phi} = \mathbf{GFA}}$, 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.

Sparse regularisation

Synthesis and analysis frameworks

- Sparse synthesis regularisation problem:

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

Synthesis framework

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

- Typically sparsity assumption justified by analysing example signals in transformed domain.
- Different to synthesising signals.
- Suggests sparse analysis regularisation problem (Elad *et al.* 2007, Nam *et al.* 2012):

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

Analysis framework

(For orthogonal bases the two approaches are identical but otherwise very different.)

Sparse regularisation

Synthesis and analysis frameworks

- Sparse synthesis regularisation problem:

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

Synthesis framework

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

- Typically sparsity assumption justified by analysing example signals in transformed domain.
- Different to synthesising signals.
- Suggests sparse analysis regularisation problem (Elad *et al.* 2007, Nam *et al.* 2012):

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

Analysis framework

(For orthogonal bases the two approaches are identical but otherwise very different.)

Sparse regularisation

Synthesis and analysis frameworks

- Sparse synthesis regularisation problem:

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

Synthesis framework

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

- Typically sparsity assumption justified by analysing example signals in transformed domain.
- Different to synthesising signals.
- Suggests sparse analysis regularisation problem (Elad *et al.* 2007, Nam *et al.* 2012):

$$\boxed{\boldsymbol{x}_{\mathsf{analysis}} = \arg\min_{\boldsymbol{x}} \left[\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1 \right]}$$

Analysis framework

(For orthogonal bases the two approaches are identical but otherwise very different.)

Sparse regularisation
SARA algorithm

- Sparsity averaging reweighted analysis (SARA)
  (Carrillo, McEwen & Wiaux 2012; Carrillo, McEwen, Van De Ville, Thiran & Wiaux 2013).

- Overcomplete dictionary composed of a concatenation of orthonormal bases:

$$\boxed{\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \ldots, \boldsymbol{\Psi}_q]}$$

  with following bases: Dirac (*i.e.* pixel basis); Haar wavelets (promotes gradient sparsity); Daubechies wavelets two to eight $\Rightarrow$ concatenation of 9 bases.

- Promote average sparsity by solving the constrained reweighted ℓ_1 analysis problem:

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{R}^N} \|\mathsf{W}\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \epsilon \quad \text{and} \quad \boldsymbol{x} \geq 0} \quad \text{SARA}$$

Sparse regularisation

SARA algorithm

- Sparsity averaging reweighted analysis (SARA)
  (Carrillo, McEwen & Wiaux 2012; Carrillo, McEwen, Van De Ville, Thiran & Wiaux 2013).

- Overcomplete dictionary composed of a concatenation of orthonormal bases:

$$\boxed{\boldsymbol{\Psi} = \left[\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \ldots, \boldsymbol{\Psi}_q\right]}$$

  with following bases: Dirac (*i.e.* pixel basis); Haar wavelets (promotes gradient sparsity); Daubechies wavelets two to eight $\Rightarrow$ concatenation of 9 bases.

- Promote average sparsity by solving the constrained reweighted ℓ_1 analysis problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^N} \|\mathbf{W}\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \epsilon \quad \text{and} \quad \boldsymbol{x} \geq 0$$

SARA

Sparse regularisation
SARA algorithm

- Sparsity averaging reweighted analysis (SARA)
  (Carrillo, McEwen & Wiaux 2012; Carrillo, McEwen, Van De Ville, Thiran & Wiaux 2013).

- Overcomplete dictionary composed of a concatenation of orthonormal bases:

$$\boxed{\boldsymbol{\Psi} = \left[\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \dots, \boldsymbol{\Psi}_q\right]}$$

  with following bases: Dirac (*i.e.* pixel basis); Haar wavelets (promotes gradient sparsity); Daubechies wavelets two to eight $\Rightarrow$ concatenation of 9 bases.

- Promote average sparsity by solving the constrained reweighted ℓ_1 analysis problem:

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^N} \|\mathbf{W}\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \epsilon \quad \text{and} \quad \boldsymbol{x} \geq 0} \quad \text{SARA}$$

Distributed and parallelised convex optimisation

- Solve resulting convex optimisation problems by proximal splitting.
- Block inexact ADMM algorithm to split data and measurement operator:
  (Carrillo, McEwen & Wiaux 2014; Onose, Carrillo, Repetti, McEwen, Thiran, Pesquet, & Wiaux 2016)

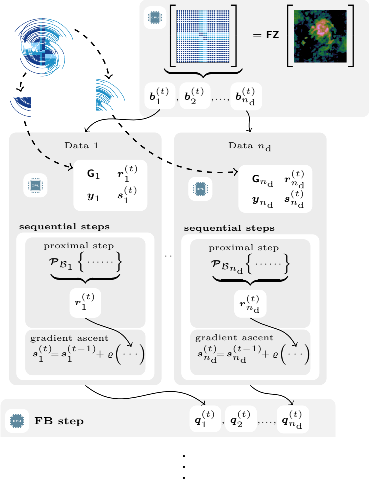
$$\boxed{\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_{n_\mathrm{d}} \end{bmatrix}}, \quad \boxed{\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 \\ \vdots \\ \boldsymbol{\Phi}_{n_\mathrm{d}} \end{bmatrix} = \begin{bmatrix} \mathsf{G}_1 \mathsf{M}_1 \\ \vdots \\ \mathsf{G}_{n_\mathrm{d}} \mathsf{M}_{n_\mathrm{d}} \end{bmatrix} \mathsf{FZ}} .$$

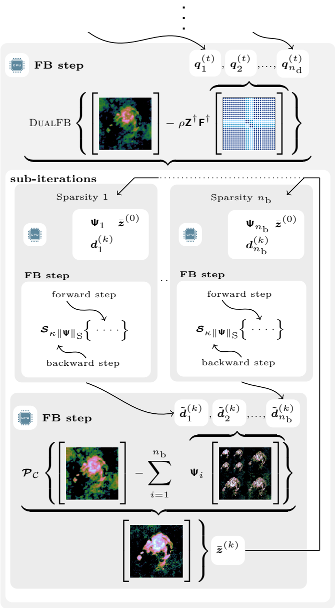
Distributed and parallelised convex optimisation

- Solve resulting convex optimisation problems by proximal splitting.
- Block inexact ADMM algorithm to split data and measurement operator:
  (Carrillo, McEwen & Wiaux 2014; Onose, Carrillo, Repetti, McEwen, Thiran, Pesquet, & Wiaux 2016)

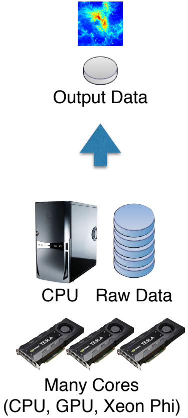
$$\boxed{\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_{n_\mathrm{d}} \end{bmatrix}}, \quad \boxed{\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 \\ \vdots \\ \boldsymbol{\Phi}_{n_\mathrm{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 \mathbf{M}_1 \\ \vdots \\ \mathbf{G}_{n_\mathrm{d}} \mathbf{M}_{n_\mathrm{d}} \end{bmatrix} \mathbf{F}\mathbf{Z}}\,.$$

Distributed and parallelised convex optimisation

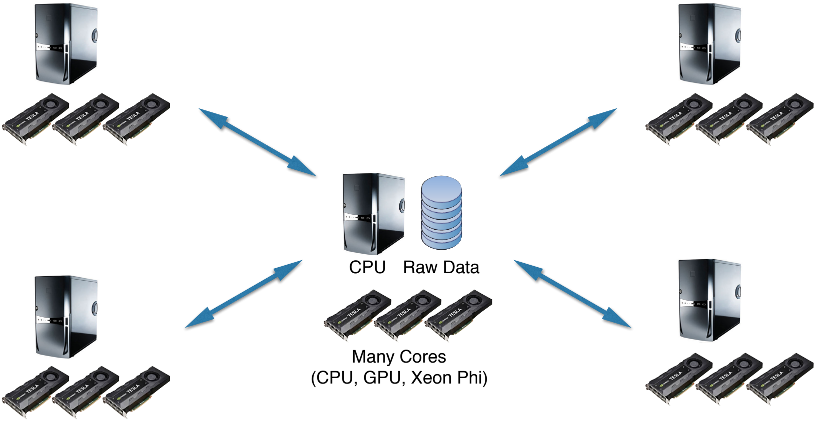




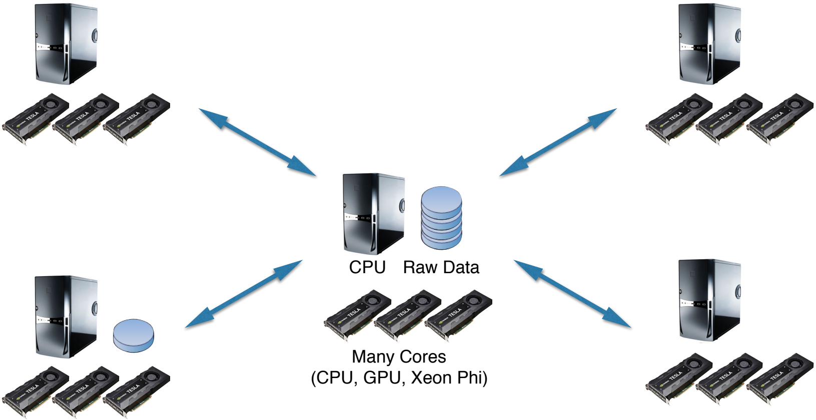
Standard algorithms



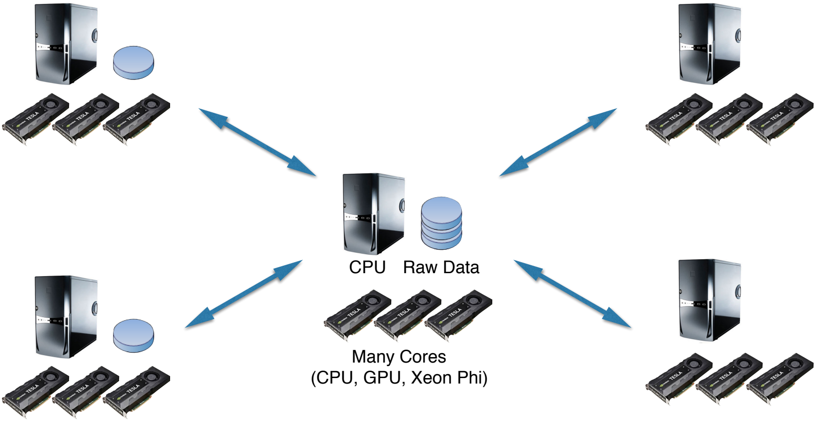
Highly distributed and parallelised algorithms



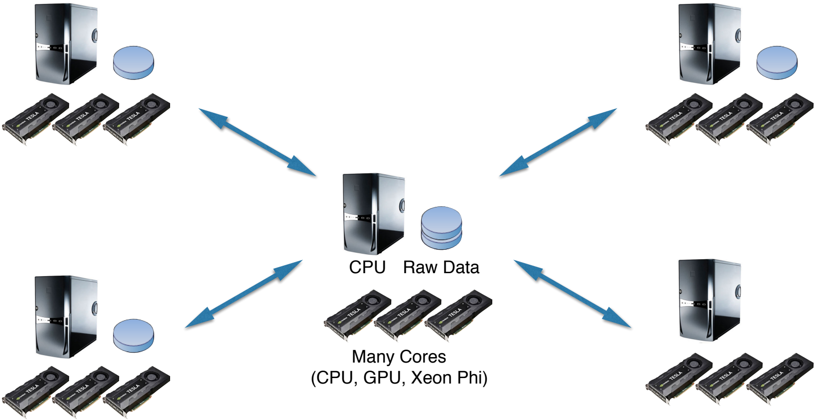
Highly distributed and parallelised algorithms



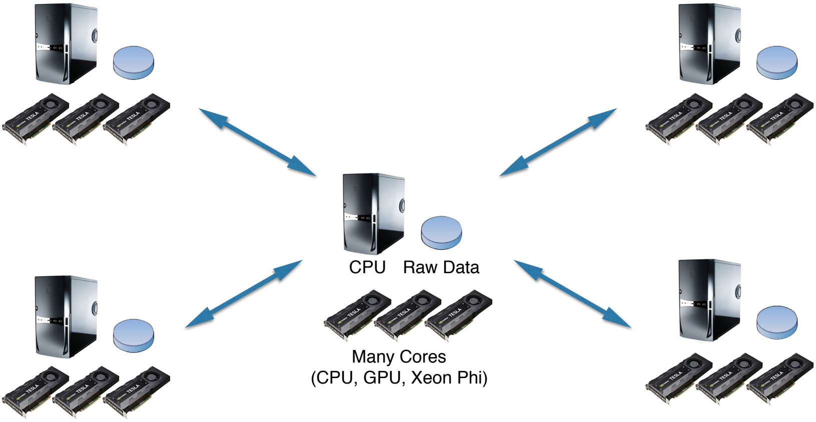
Highly distributed and parallelised algorithms



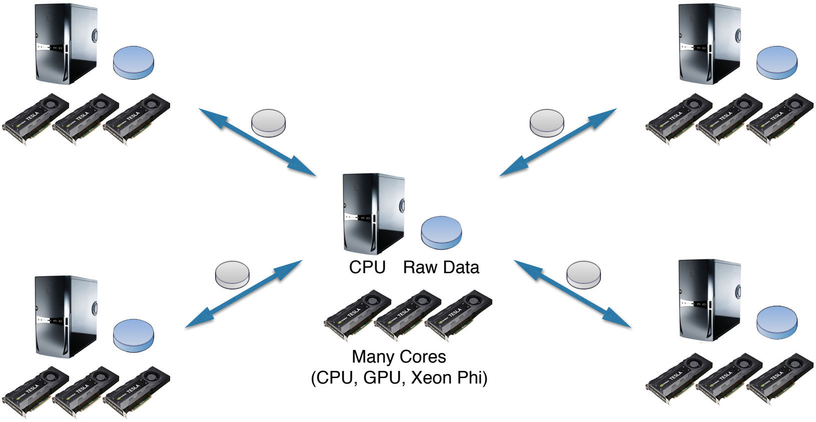
Highly distributed and parallelised algorithms



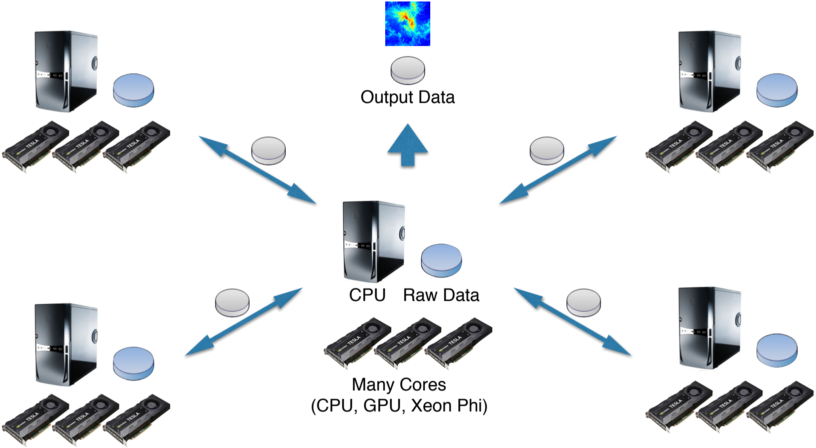
Highly distributed and parallelised algorithms



Highly distributed and parallelised algorithms



Highly distributed and parallelised algorithms

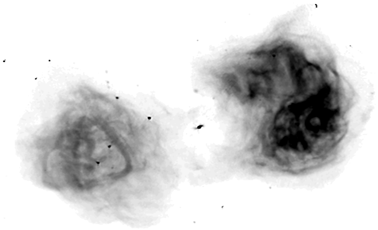


Highly distributed and parallelised algorithms

Reconstruction

- Hybrid w-stacking and w-projection distributed and parallelised reconstruction (Pratley, Johnston-Hollitt & McEwen 2018)
  - 100 millions visibilities (measurements)
  - 4096$\times$4096 pixel image ($\sim$17 million pixels)
  - 17° field of view
  - w-terms of $\pm$300 wavelengths (to account for wide fields)

Imaging with exact wide-field corrections for 100 million visibilities in 30 minutes.



Public open-source codes

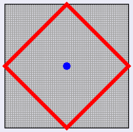
PURIFY code http://basp-group.github.io/purify/

Next-generation radio interferometric imaging

Carrillo, McEwen, Wiaux, Pratley, d'Avezac

PURIFY is an open-source code that provides functionality to perform radio interferometric imaging, leveraging recent developments in the field of compressive sensing and convex optimisation.

SOPT code http://basp-group.github.io/sopt/



Sparse OPTimisation

Carrillo, McEwen, Wiaux, Kartik, d'Avezac, Pratley, Perez-Suarez

SOPT is an open-source code that provides functionality to perform sparse optimisation using state-of-the-art convex optimisation algorithms.

Imaging observations from the VLA and ATCA with PURIFY

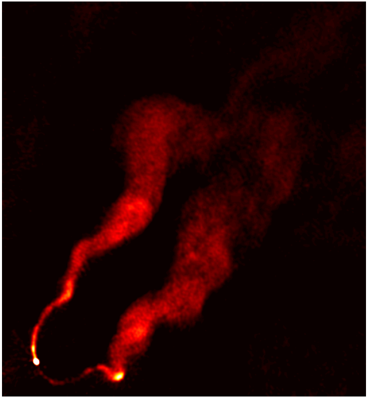
(a) NRAO Very Large Array (VLA)

(b) Australia Telescope Compact Array (ATCA)

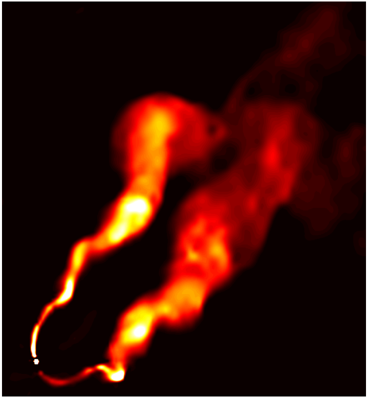
Figure: Radio interferometric telescopes considered

PURIFY reconstruction

VLA observation of 3C129



(a) CLEAN (uniform)

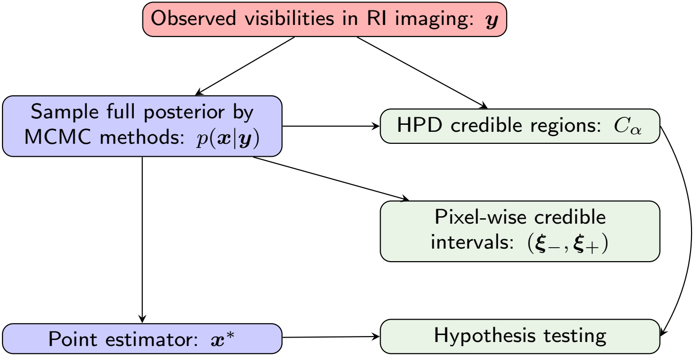


(b) PURIFY

Figure: 3C129 recovered images (Pratley, McEwen, *et al.* 2016)

Outline

MCMC sampling and uncertainty quantification



MCMC sampling the full posterior distribution

- Sample full posterior distribution $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y})$.
- MCMC methods for high-dimensional problems (like interferometric imaging):
  - Gibbs sampling (sample from conditional distributions)
  - Hamiltonian MC (HMC) sampling (exploit gradients)
  - Metropolis adjusted Langevin algorithm (MALA) sampling (exploit gradients)

Require MCMC approach to support sparsity priors, which shown to be highly effective.

MCMC sampling the full posterior distribution

- Sample full posterior distribution $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y})$.
- MCMC methods for high-dimensional problems (like interferometric imaging):
  - Gibbs sampling (sample from conditional distributions)
  - Hamiltonian MC (HMC) sampling (exploit gradients)
  - Metropolis adjusted Langevin algorithm (MALA) sampling (exploit gradients)

Require MCMC approach to support sparsity priors, which shown to be highly effective.

MCMC sampling the full posterior distribution

- Sample full posterior distribution $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y})$.
- MCMC methods for high-dimensional problems (like interferometric imaging):
  - Gibbs sampling (sample from conditional distributions)
  - Hamiltonian MC (HMC) sampling (exploit gradients)
  - Metropolis adjusted Langevin algorithm (MALA) sampling (exploit gradients)

Require MCMC approach to support sparsity priors, which shown to be highly effective.

MCMC sampling with gradients

Langevin dynamics

- Consider posteriors of the following form:

$$\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \underbrace{\boxed{\pi(\boldsymbol{x})}}_{\text{Posterior}} \propto \exp(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Smooth}})$$

- If $g(x)$ differentiable can adopt MALA (Langevin dynamics).
- Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

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

  where $\mathcal{W}$ is Brownian motion.
- Need gradients so cannot support sparse priors.

MCMC sampling with gradients

Langevin dynamics

- Consider posteriors of the following form:

$$\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \underbrace{\boxed{\pi(\boldsymbol{x})}}_{\text{Posterior}} \propto \exp\big(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Smooth}}\big)$$

- If $g(\boldsymbol{x})$ differentiable can adopt MALA (Langevin dynamics).
- Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

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

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

- Need gradients so cannot support sparse priors.

MCMC sampling with gradients

Langevin dynamics

- Consider posteriors of the following form:

$$\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \underbrace{\boxed{\pi(\boldsymbol{x})}}_{\text{Posterior}} \propto \exp\big(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Smooth}}\big)$$

- If $g(\boldsymbol{x})$ differentiable can adopt MALA (Langevin dynamics).
- Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

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

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

- Need gradients so cannot support sparse priors.

MCMC sampling with gradients
Langevin dynamics

- Consider posteriors of the following form:

$$\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \underbrace{\boxed{\pi(\boldsymbol{x})}}_{\text{Posterior}} \propto \exp\big(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Smooth}}\big)$$

- If $g(\boldsymbol{x})$ differentiable can adopt MALA (Langevin dynamics).
- Based on Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution:

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

  where $\mathcal{W}$ is Brownian motion.
- Need gradients so cannot support sparse priors.

Proximity operators
A brief aside

- Define proximity operator:

$$\boxed{\mathrm{prox}_g^\lambda(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left[g(\boldsymbol{u}) + \|\boldsymbol{u} - \boldsymbol{x}\|^2 / 2\lambda \right]}$$

- Generalisation of projection operator:

$$\mathcal{P}_\mathcal{C}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left[\imath_\mathcal{C}(\boldsymbol{u}) + \|\boldsymbol{u} - \boldsymbol{x}\|^2 / 2 \right],$$

where $\imath_\mathcal{C}(\boldsymbol{u}) = \infty$ if $\boldsymbol{u} \notin \mathcal{C}$ and zero otherwise.

Proximity operators

A brief aside

- Define proximity operator:

$$\boxed{\operatorname{prox}_g^\lambda(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left[g(\boldsymbol{u}) + \|\boldsymbol{u} - \boldsymbol{x}\|^2 / 2\lambda \right]}$$

- Generalisation of projection operator:

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left[\imath_{\mathcal{C}}(\boldsymbol{u}) + \|\boldsymbol{u} - \boldsymbol{x}\|^2 / 2 \right],$$

where $\imath_{\mathcal{C}}(\boldsymbol{u}) = \infty$ if $\boldsymbol{u} \notin \mathcal{C}$ and zero otherwise.

Proximity operators

A brief aside

- Define proximity operator:

$$\boxed{\operatorname{prox}_g^{\lambda}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left[g(\boldsymbol{u}) + \|\boldsymbol{u} - \boldsymbol{x}\|^2 / 2\lambda \right]}$$

- Generalisation of projection operator:

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left[\imath_{\mathcal{C}}(\boldsymbol{u}) + \|\boldsymbol{u} - \boldsymbol{x}\|^2 / 2 \right],$$

where $\imath_{\mathcal{C}}(\boldsymbol{u}) = \infty$ if $\boldsymbol{u} \notin \mathcal{C}$ and zero otherwise.

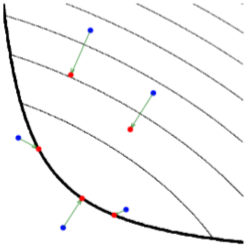


Figure: Illustration of proximity operator [Credit: Parikh & Boyd (2013)]

Proximal MCMC methods

- Exploit proximal calculus.
- "Replace gradients with sub-gradients".

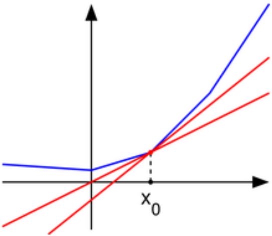


Figure: Illustration of sub-gradients [Credit: Wikipedia (Maksim)]

Proximal MALA

Moreau approximation

- Moreau approximation of $f(\boldsymbol{x}) \propto \exp(-g(\boldsymbol{x}))$:

$$\boxed{f_\lambda^{\mathsf{MA}}(\boldsymbol{x}) = \sup_{\boldsymbol{u} \in \mathbb{R}^N} f(\boldsymbol{u}) \exp\left(-\frac{\|\boldsymbol{u}-\boldsymbol{x}\|^2}{2\lambda}\right)}$$

- Important properties of $f_\lambda^{\mathsf{MA}}(\boldsymbol{x})$:

1. As $\lambda \rightarrow 0$, $f_\lambda^{\mathsf{MA}}(\boldsymbol{x}) \rightarrow f(\boldsymbol{x})$
2. $\nabla \log f_\lambda^{\mathsf{MA}}(\boldsymbol{x}) = (\mathrm{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x})/\lambda$

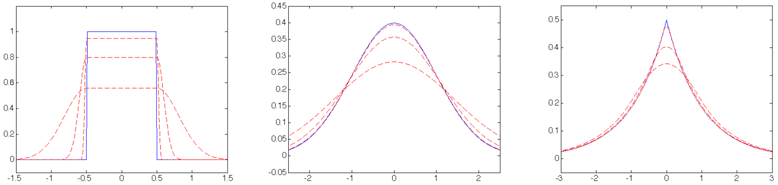


Figure: Illustration of Moreau approximations [Credit: Pereyra 2016a]

Proximal MALA

Moreau approximation

- Moreau approximation of $f(\boldsymbol{x}) \propto \exp(-g(\boldsymbol{x}))$:

$$\boxed{f_\lambda^{\mathsf{MA}}(\boldsymbol{x}) = \sup_{\boldsymbol{u}\in\mathbb{R}^N} f(\boldsymbol{u}) \exp\left(-\frac{\|\boldsymbol{u}-\boldsymbol{x}\|^2}{2\lambda}\right)}$$

- Important properties of $f_\lambda^{\mathsf{MA}}(\boldsymbol{x})$:

1. As $\lambda \to 0$, $f_\lambda^{\mathsf{MA}}(\boldsymbol{x}) \to f(\boldsymbol{x})$
2. $\nabla \log f_\lambda^{\mathsf{MA}}(\boldsymbol{x}) = (\mathrm{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x})/\lambda$

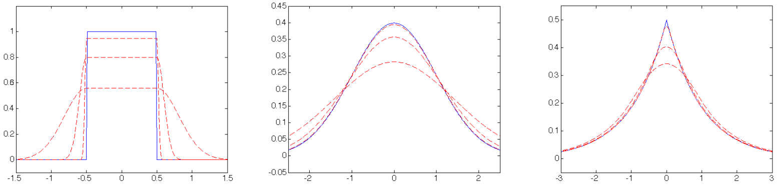


Figure: Illustration of Moreau approximations [Credit: Pereyra 2016a]

Proximal MALA

MCMC sampling

Proximal Metropolis adjusted Langevin algorithm (P-MALA)

Pereyra (2016a)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Convex}})$.
- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau approximation to π:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2}\boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta}\boldsymbol{w}^{(m)} \; .$$

$$\nabla \log \pi_\lambda(\boldsymbol{x}) = (\mathrm{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x})/\lambda$$

- Metropolis-Hastings accept-reject step.

Proximal MALA

MCMC sampling

Proximal Metropolis adjusted Langevin algorithm (P-MALA)

Pereyra (2016a)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Convex}})$.
- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau approximation to π:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2}\boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta}\boldsymbol{w}^{(m)} \ .$$

$$\nabla \log \pi_\lambda(\boldsymbol{x}) = (\mathrm{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x})/\lambda$$

- Metropolis-Hastings accept-reject step.

Proximal MALA

MCMC sampling

Proximal Metropolis adjusted Langevin algorithm (P-MALA)

Pereyra (2016a)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp\big(-\boxed{g(\boldsymbol{x})}\big)$. Convex
- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau approximation to π:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2}\boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta}\boldsymbol{w}^{(m)} \ .$$

$$\nabla \log \pi_\lambda(\boldsymbol{x}) = (\mathrm{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x})/\lambda$$

- Metropolis-Hastings accept-reject step.

Proximal MALA

MCMC sampling

Proximal Metropolis adjusted Langevin algorithm (P-MALA)

Pereyra (2016a)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp\big(-\underbrace{\boxed{g(\boldsymbol{x})}}_{\text{Convex}}\big)$.

- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau approximation to π:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2} \boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta} \boldsymbol{w}^{(m)} \, .$$

$$\nabla \log \pi_\lambda(\boldsymbol{x}) = (\mathrm{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x})/\lambda$$

- Metropolis-Hastings accept-reject step.

Proximal MALA

Computing proximity operators for the analysis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\bar{g}(\boldsymbol{x}) = \bar{f}_1(\boldsymbol{x}) + \bar{f}_2(\boldsymbol{x})$, where $\boxed{\bar{f}_1(\boldsymbol{x}) = \mu\|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1}$ and $\boxed{\bar{f}_2(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2/2\sigma^2}$.

  Prior Likelihood

- Must solve an optimisation problem for each iteration!

$$\boxed{\operatorname{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) = \underset{\boldsymbol{u}\in\mathbb{R}^N}{\operatorname{argmin}} \left\{ \mu\|\boldsymbol{\Psi}^\dagger \boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{x}\|_2^2}{\delta} \right\}}.$$

- Taylor expansion at point $\boldsymbol{x}$: $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2(\boldsymbol{u}-\boldsymbol{x})^\top \boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\boxed{\operatorname{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \operatorname{prox}_{\bar{f}_1}^{\delta/2}\left(\boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2\right)}.$$

Single forward-backward iteration

- Analytic approximation:

$$\boxed{\operatorname{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \bar{\boldsymbol{v}} + \boldsymbol{\Psi}\left(\operatorname{soft}_{\mu\delta/2}(\boldsymbol{\Psi}^\dagger\bar{\boldsymbol{v}}) - \boldsymbol{\Psi}^\dagger\bar{\boldsymbol{v}}\right)}, \quad \text{where } \bar{\boldsymbol{v}} = \boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2.$$

Proximal MALA

Computing proximity operators for the analysis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.

- Let $\bar{g}(\boldsymbol{x}) = \bar{f}_1(\boldsymbol{x}) + \bar{f}_2(\boldsymbol{x})$, where $\boxed{\bar{f}_1(\boldsymbol{x}) = \mu \|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1}$ and $\boxed{\bar{f}_2(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / 2\sigma^2}$.

  Prior Likelihood

- Must solve an optimisation problem for each iteration!

$$\boxed{\mathrm{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathbb{R}^N}{\mathrm{argmin}} \left\{ \mu \|\boldsymbol{\Psi}^\dagger \boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{x}\|_2^2}{\delta} \right\}}.$$

- Taylor expansion at point $\boldsymbol{x}$: $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{x})^\top \boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})$.

- Then proximity operator approximated by

$$\boxed{\mathrm{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \mathrm{prox}_{\bar{f}_1}^{\delta/2}\left(\boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2\right)}.$$

  Single forward-backward iteration

- Analytic approximation:

$$\boxed{\mathrm{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \bar{\boldsymbol{v}} + \boldsymbol{\Psi}\left(\mathrm{soft}_{\mu\delta/2}(\boldsymbol{\Psi}^\dagger \bar{\boldsymbol{v}}) - \boldsymbol{\Psi}^\dagger \bar{\boldsymbol{v}}\right)}, \quad \text{where } \bar{\boldsymbol{v}} = \boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2.$$

Proximal MALA

Computing proximity operators for the analysis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\bar{g}(\boldsymbol{x}) = \bar{f}_1(\boldsymbol{x}) + \bar{f}_2(\boldsymbol{x})$, where $\underbrace{\boxed{\bar{f}_1(\boldsymbol{x}) = \mu \|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1}}_{\text{Prior}}$ and $\underbrace{\boxed{\bar{f}_2(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / 2\sigma^2}}_{\text{Likelihood}}$.
- Must solve an optimisation problem for each iteration!

$$\boxed{\operatorname{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ \mu \|\boldsymbol{\Psi}^\dagger \boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{x}\|_2^2}{\delta} \right\}}\,.$$

- Taylor expansion at point $\boldsymbol{x}$: $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{x})^\top \boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\underbrace{\boxed{\operatorname{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \operatorname{prox}_{\bar{f}_1}^{\delta/2}\left(\boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2\right)}}_{\text{Single forward-backward iteration}}\,.$$

- Analytic approximation:

$$\boxed{\operatorname{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \bar{\boldsymbol{v}} + \boldsymbol{\Psi}\left(\operatorname{soft}_{\mu\delta/2}(\boldsymbol{\Psi}^\dagger \bar{\boldsymbol{v}}) - \boldsymbol{\Psi}^\dagger \bar{\boldsymbol{v}}\right)}\,, \quad \text{where } \bar{\boldsymbol{v}} = \boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2.$$

Proximal MALA

Computing proximity operators for the analysis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\bar{g}(\boldsymbol{x}) = \bar{f}_1(\boldsymbol{x}) + \bar{f}_2(\boldsymbol{x})$, where $\boxed{\bar{f}_1(\boldsymbol{x}) = \mu\|\boldsymbol{\Psi}^\dagger\boldsymbol{x}\|_1}$ and $\boxed{\bar{f}_2(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2/2\sigma^2}$.

  Prior Likelihood

- Must solve an optimisation problem for each iteration!

$$\boxed{\mathrm{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) = \underset{\boldsymbol{u}\in\mathbb{R}^N}{\operatorname{argmin}}\left\{\mu\|\boldsymbol{\Psi}^\dagger\boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{x}\|_2^2}{\delta}\right\}}\,.$$

- Taylor expansion at point $\boldsymbol{x}$: $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{x})^\top\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\boxed{\mathrm{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \mathrm{prox}_{\bar{f}_1}^{\delta/2}\left(\boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2\right)}\,.$$

Single forward-backward iteration

- Analytic approximation:

$$\boxed{\mathrm{prox}_{\bar{g}}^{\delta/2}(\boldsymbol{x}) \approx \bar{\boldsymbol{v}} + \boldsymbol{\Psi}\left(\mathrm{soft}_{\mu\delta/2}(\boldsymbol{\Psi}^\dagger\bar{\boldsymbol{v}}) - \boldsymbol{\Psi}^\dagger\bar{\boldsymbol{v}}\right)}\,, \quad \text{where } \bar{\boldsymbol{v}} = \boldsymbol{x} - \delta\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})/2\sigma^2.$$

Proximal MALA

Computing proximity operators for the synthesis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\hat{g}(\boldsymbol{x}(\boldsymbol{a})) = \hat{f}_1(\boldsymbol{a}) + \hat{f}_2(\boldsymbol{a})$, where $\boxed{\hat{f}_1(\boldsymbol{a}) = \mu\|\boldsymbol{a}\|_1}$ and $\boxed{\hat{f}_2(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\Phi\Psi a}\|_2^2/2\sigma^2}$.

  Prior Likelihood

- Must solve an optimisation problem for each iteration!

$$\text{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) = \underset{\boldsymbol{u}\in\mathbb{R}^L}{\text{argmin}}\left\{\mu\|\boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi\Psi u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{a}\|_2^2}{\delta}\right\}.$$

- Taylor expansion at point $\boldsymbol{a}$: $\|\boldsymbol{y} - \boldsymbol{\Phi\Psi u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi\Psi a}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{a})^\top \boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi\Psi a} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\text{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \text{prox}_{\hat{f}_1}^{\delta/2}\left(\boldsymbol{a} - \delta\boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi\Psi a} - \boldsymbol{y})/2\sigma^2\right).$$

Single forward-backward iteration

- Analytic approximation:

$$\text{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \text{soft}_{\mu\delta/2}\left(\boldsymbol{a} - \delta\boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi\Psi a} - \boldsymbol{y})/2\sigma^2\right).$$

Proximal MALA

Computing proximity operators for the synthesis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\hat{g}(\boldsymbol{x}(\boldsymbol{a})) = \hat{f}_1(\boldsymbol{a}) + \hat{f}_2(\boldsymbol{a})$, where $\boxed{\hat{f}_1(\boldsymbol{a}) = \mu \|\boldsymbol{a}\|_1}$ and $\boxed{\hat{f}_2(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a}\|_2^2 / 2\sigma^2}$.

  Prior Likelihood
- Must solve an optimisation problem for each iteration!

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) = \underset{\boldsymbol{u} \in \mathbb{R}^L}{\operatorname{argmin}} \left\{ \mu \|\boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{a}\|_2^2}{\delta} \right\}} .$$

- Taylor expansion at point $\boldsymbol{a}$: $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{a})^\top \boldsymbol{\Psi}^\dagger \boldsymbol{\Phi}^\dagger (\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \operatorname{prox}_{\hat{f}_1}^{\delta/2} \left(\boldsymbol{a} - \delta \boldsymbol{\Psi}^\dagger \boldsymbol{\Phi}^\dagger (\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{y}) / 2\sigma^2 \right)} .$$

Single forward-backward iteration

- Analytic approximation:

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \operatorname{soft}_{\mu\delta/2} \left(\boldsymbol{a} - \delta \boldsymbol{\Psi}^\dagger \boldsymbol{\Phi}^\dagger (\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{y}) / 2\sigma^2 \right)} .$$

Proximal MALA

Computing proximity operators for the synthesis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\hat{g}(\boldsymbol{x}(\boldsymbol{a})) = \hat{f}_1(\boldsymbol{a}) + \hat{f}_2(\boldsymbol{a})$, where $\underbrace{\boxed{\hat{f}_1(\boldsymbol{a}) = \mu\|\boldsymbol{a}\|_1}}_{\text{Prior}}$ and $\underbrace{\boxed{\hat{f}_2(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a}\|_2^2/2\sigma^2}}_{\text{Likelihood}}$.
- Must solve an optimisation problem for each iteration!

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) = \underset{\boldsymbol{u}\in\mathbb{R}^L}{\operatorname{argmin}} \left\{ \mu\|\boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{a}\|_2^2}{\delta} \right\}} \, .$$

- Taylor expansion at point $\boldsymbol{a}$: $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{a})^\top \boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\underbrace{\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \operatorname{prox}_{\hat{f}_1}^{\delta/2}\left(\boldsymbol{a} - \delta\boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{y})/2\sigma^2\right)}}_{\text{Single forward-backward iteration}} \, .$$

- Analytic approximation:

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \operatorname{soft}_{\mu\delta/2}\left(\boldsymbol{a} - \delta\boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{y})/2\sigma^2\right)} \, .$$

Proximal MALA

Computing proximity operators for the synthesis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\hat{g}(\boldsymbol{x}(\boldsymbol{a})) = \hat{f}_1(\boldsymbol{a}) + \hat{f}_2(\boldsymbol{a})$, where $\boxed{\hat{f}_1(\boldsymbol{a}) = \mu\|\boldsymbol{a}\|_1}$ and $\boxed{\hat{f}_2(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\Phi\Psi a}\|_2^2/2\sigma^2}$.

  Prior Likelihood
- Must solve an optimisation problem for each iteration!

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) = \underset{\boldsymbol{u}\in\mathbb{R}^L}{\operatorname{argmin}} \left\{ \mu\|\boldsymbol{u}\|_1 + \frac{\|\boldsymbol{y} - \boldsymbol{\Phi\Psi u}\|_2^2}{2\sigma^2} + \frac{\|\boldsymbol{u} - \boldsymbol{a}\|_2^2}{\delta} \right\}}\,.$$

- Taylor expansion at point $\boldsymbol{a}$: $\|\boldsymbol{y} - \boldsymbol{\Phi\Psi u}\|_2^2 \approx \|\boldsymbol{y} - \boldsymbol{\Phi\Psi a}\|_2^2 + 2(\boldsymbol{u} - \boldsymbol{a})^\top \boldsymbol{\Psi}^\dagger \boldsymbol{\Phi}^\dagger (\boldsymbol{\Phi\Psi a} - \boldsymbol{y})$.
- Then proximity operator approximated by

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \operatorname{prox}_{\hat{f}_1}^{\delta/2}\left(\boldsymbol{a} - \delta\boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi\Psi a} - \boldsymbol{y})/2\sigma^2\right)}\,.$$

Single forward-backward iteration

- Analytic approximation:

$$\boxed{\operatorname{prox}_{\hat{g}}^{\delta/2}(\boldsymbol{a}) \approx \operatorname{soft}_{\mu\delta/2}\left(\boldsymbol{a} - \delta\boldsymbol{\Psi}^\dagger\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi\Psi a} - \boldsymbol{y})/2\sigma^2\right)}\,.$$

MYULA

Moreau-Yosida approximation

- Moreau-Yosida approximation (Moreau envelope) of f:

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

- Important properties of $f_\lambda^{\mathsf{MY}}(\boldsymbol{x})$:
  1. As $\lambda \rightarrow 0$, $f_\lambda^{\mathsf{MY}}(\boldsymbol{x}) \rightarrow f(\boldsymbol{x})$
  2. $\nabla f_\lambda^{\mathsf{MY}}(\boldsymbol{x}) = (\boldsymbol{x} - \mathrm{prox}_f^\lambda(\boldsymbol{x}))/\lambda$

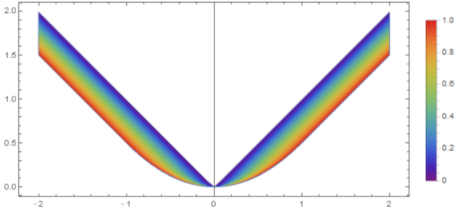


Figure: Illustration of Moreau-Yosida envelope of $|x|$ for varying λ [Credit: Stack exchange (ubpdqn)]

MYULA
Moreau-Yosida approximation

- Moreau-Yosida approximation (Moreau envelope) of f:

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

- Important properties of $f_\lambda^{\mathsf{MY}}(\boldsymbol{x})$:

1. As $\lambda \to 0$, $f_\lambda^{\mathsf{MY}}(\boldsymbol{x}) \to f(\boldsymbol{x})$
2. $\nabla f_\lambda^{\mathsf{MY}}(\boldsymbol{x}) = (\boldsymbol{x} - \mathrm{prox}_f^\lambda(\boldsymbol{x}))/\lambda$

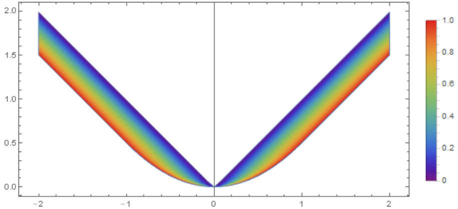


Figure: Illustration of Moreau-Yosida envelope of $|x|$ for varying λ [Credit: Stack exchange (ubpdqn)]

MYULA
MCMC sampling

Moreau-Yosida unadjusted Langevin algorithm (MYULA)
Durmus, Moulines & Pereyra (2016)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$, where

$$g(\boldsymbol{x}) = \underbrace{\boxed{f_1(\boldsymbol{x})}}_{\text{Convex}} + \underbrace{\boxed{f_2(\boldsymbol{x})}}_{\text{Smooth}} .$$

- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau-Yosida approximation to f_1:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2}\boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta}\boldsymbol{w}^{(m)} .$$

$$\nabla \log \pi(\boldsymbol{x}) \approx \big(\mathrm{prox}_{f_1}^{\lambda}(\boldsymbol{x}) - \boldsymbol{x}\big)/\lambda - \nabla f_2(\boldsymbol{x})$$

- No Metropolis-Hastings accept-reject step. Converges geometrically fast, where bias can be made arbitrarily small. To achieve precision target ϵ requires:
  - Worst case: order $N^5 \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.
  - Strong convexity worst case: order $N \log(N) \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.

MYULA

MCMC sampling

Moreau-Yosida unadjusted Langevin algorithm (MYULA)

Durmus, Moulines & Pereyra (2016)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$, where

$$g(\boldsymbol{x}) = \underbrace{\boxed{f_1(\boldsymbol{x})}}_{\text{Convex}} + \underbrace{\boxed{f_2(\boldsymbol{x})}}_{\text{Smooth}} .$$

- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau-Yosida approximation to f_1:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2}\boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta}\boldsymbol{w}^{(m)} .$$

$$\nabla \log \pi(\boldsymbol{x}) \approx \big(\mathrm{prox}_{f_1}^{\lambda}(\boldsymbol{x}) - \boldsymbol{x}\big)/\lambda - \nabla f_2(\boldsymbol{x})$$

- No Metropolis-Hastings accept-reject step. Converges geometrically fast, where bias can be made arbitrarily small. To achieve precision target ϵ requires:
  - Worst case: order $N^5 \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.
  - Strong convexity worst case: order $N \log(N) \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.

MYULA

MCMC sampling

Moreau-Yosida unadjusted Langevin algorithm (MYULA)

Durmus, Moulines & Pereyra (2016)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$, where

$$g(\boldsymbol{x}) = \underbrace{\boxed{f_1(\boldsymbol{x})}}_{\text{Convex}} + \underbrace{\boxed{f_2(\boldsymbol{x})}}_{\text{Smooth}} .$$

- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau-Yosida approximation to f_1:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2} \boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta}\boldsymbol{w}^{(m)} .$$

$$\nabla \log \pi(\boldsymbol{x}) \approx \big(\mathrm{prox}^{\lambda}_{f_1}(\boldsymbol{x}) - \boldsymbol{x}\big)/\lambda - \nabla f_2(\boldsymbol{x})$$

- No Metropolis-Hastings accept-reject step. Converges geometrically fast, where bias can be made arbitrarily small. To achieve precision target ϵ requires:
  - Worst case: order $N^5 \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.
  - Strong convexity worst case: order $N \log(N) \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.

MYULA

MCMC sampling

Moreau-Yosida unadjusted Langevin algorithm (MYULA)

Durmus, Moulines & Pereyra (2016)

- Consider log-convex posteriors: $\mathrm{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \pi(\boldsymbol{x}) \propto \exp\bigl(-g(\boldsymbol{x})\bigr)$, where

$$g(\boldsymbol{x}) = \underbrace{\boxed{f_1(\boldsymbol{x})}}_{\text{Convex}} + \underbrace{\boxed{f_2(\boldsymbol{x})}}_{\text{Smooth}} .$$

- Langevin diffusion process $\mathcal{L}(t)$, with π as stationary distribution ($\mathcal{W}$ Brownian motion):

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

- Euler discretisation and apply Moreau-Yosida approximation to f_1:

$$\boldsymbol{l}^{(m+1)} = \boldsymbol{l}^{(m)} + \frac{\delta}{2} \boxed{\nabla \log \pi(\boldsymbol{l}^{(m)})} + \sqrt{\delta} \boldsymbol{w}^{(m)} .$$

$$\nabla \log \pi(\boldsymbol{x}) \approx \bigl(\mathrm{prox}_{f_1}^{\lambda}(\boldsymbol{x}) - \boldsymbol{x}\bigr)/\lambda - \nabla f_2(\boldsymbol{x})$$

- No Metropolis-Hastings accept-reject step. Converges geometrically fast, where bias can be made arbitrarily small. To achieve precision target ϵ requires:
  - Worst case: order $N^5 \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.
  - Strong convexity worst case: order $N \log(N) \log^2(\epsilon^{-1})\epsilon^{-2}$ iterations.

MYULA

Computing proximity operators for the analysis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\bar{g}(\boldsymbol{x}) = \bar{f}_1(\boldsymbol{x}) + \bar{f}_2(\boldsymbol{x})$, where $\boxed{\bar{f}_1(\boldsymbol{x}) = \mu \|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1}$ and $\boxed{\bar{f}_2(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / 2\sigma^2}$.

  Prior — Likelihood
- Only need to compute proximity operator of $\bar{f}_1$, which can be computed analytically without any approximation:

$$\boxed{\mathrm{prox}_{\bar{f}_1}^{\delta/2}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\Psi}\left(\mathrm{soft}_{\mu\delta/2}(\boldsymbol{\Psi}^\dagger \boldsymbol{x}) - \boldsymbol{\Psi}^\dagger \boldsymbol{x}\right)}.$$

MYULA

Computing proximity operators for the analysis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\bigl(-g(\boldsymbol{x})\bigr)$.
- Let $\bar{g}(\boldsymbol{x}) = \bar{f}_1(\boldsymbol{x}) + \bar{f}_2(\boldsymbol{x})$, where $\underbrace{\boxed{\bar{f}_1(\boldsymbol{x}) = \mu \|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\|_1}}_{\text{Prior}}$ and $\underbrace{\boxed{\bar{f}_2(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 / 2\sigma^2}}_{\text{Likelihood}}$.
- Only need to compute proximity operator of f_1, which can be computed analytically without any approximation:

$$\boxed{\mathrm{prox}_{\bar{f}_1}^{\delta/2}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\Psi}\left(\mathrm{soft}_{\mu\delta/2}(\boldsymbol{\Psi}^\dagger \boldsymbol{x}) - \boldsymbol{\Psi}^\dagger \boldsymbol{x}\right)}\,.$$

MYULA

Computing proximity operators for the synthesis case

- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\hat{g}(\boldsymbol{x}(\boldsymbol{a})) = \hat{f}_1(\boldsymbol{a}) + \hat{f}_2(\boldsymbol{a})$, where $\underbrace{\boxed{\hat{f}_1(\boldsymbol{a}) = \mu \|\boldsymbol{a}\|_1}}_{\text{Prior}}$ and $\underbrace{\boxed{\hat{f}_2(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\Phi\Psi a}\|_2^2 / 2\sigma^2}}_{\text{Likelihood}}$.
- Only need to compute proximity operator of f_1, which can be computed analytically without any approximation:

$$\boxed{\operatorname{prox}_{\hat{f}_1}^{\delta/2}(\boldsymbol{a}) = \operatorname{soft}_{\mu\delta/2}(\boldsymbol{a})}.$$

MYULA

Computing proximity operators for the synthesis case

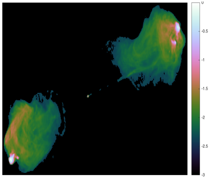
- Recall posterior: $\pi(\boldsymbol{x}) \propto \exp\big(-g(\boldsymbol{x})\big)$.
- Let $\hat{g}(\boldsymbol{x}(\boldsymbol{a})) = \hat{f}_1(\boldsymbol{a}) + \hat{f}_2(\boldsymbol{a})$, where $\boxed{\hat{f}_1(\boldsymbol{a}) = \mu \|\boldsymbol{a}\|_1}$ and $\boxed{\hat{f}_2(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{a}\|_2^2 / 2\sigma^2}$.

  Prior Likelihood
- Only need to compute proximity operator of f_1, which can be computed analytically without any approximation:

$$\boxed{\operatorname{prox}_{\hat{f}_1}^{\delta/2}(\boldsymbol{a}) = \operatorname{soft}_{\mu\delta/2}(\boldsymbol{a})}\,.$$

Numerical experiments

MYULA with analysis model

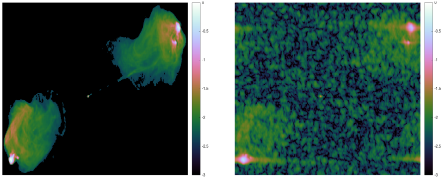


(a) Ground truth

Figure: Cygnus A

Numerical experiments

MYULA with analysis model

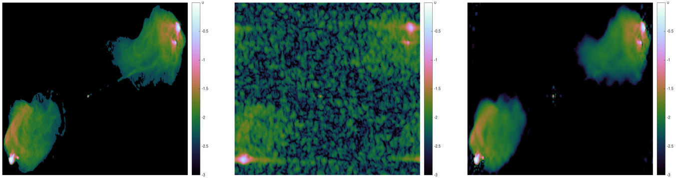


(a) Ground truth (b) Dirty image

Figure: Cygnus A

Numerical experiments

MYULA with analysis model

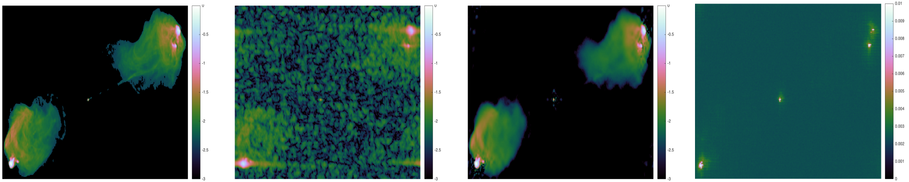


(a) Ground truth (b) Dirty image (c) Mean recovered image

Figure: Cygnus A

Numerical experiments

MYULA with analysis model

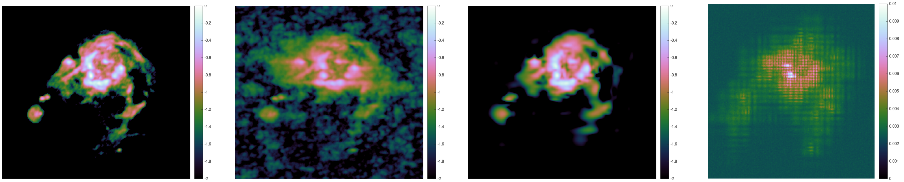


(a) Ground truth (b) Dirty image (c) Mean recovered image (d) Credible interval length

Figure: Cygnus A

Numerical experiments

MYULA with analysis model

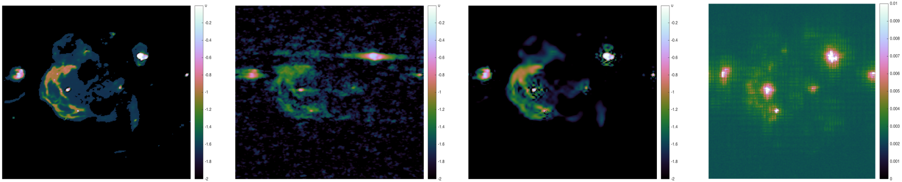


(a) Ground truth (b) Dirty image (c) Mean recovered image (d) Credible interval length

Figure: HII region of M31

Numerical experiments

MYULA with analysis model

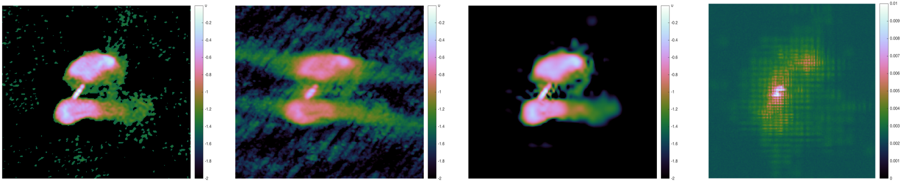


(a) Ground truth (b) Dirty image (c) Mean recovered image (d) Credible interval length

Figure: W28 Supernova remnant

Numerical experiments

MYULA with analysis model



(a) Ground truth (b) Dirty image (c) Mean recovered image (d) Credible interval length

Figure: 3C288

Numerical experiments
Computation time

Table: CPU time in minutes for Proximal MCMC sampling

Image	Method	CPU time (min) Analysis	CPU time (min) Synthesis
Cygnus A	P-MALA	2274	1762
	MYULA	1056	942
M31	P-MALA	1307	944
	MYULA	618	581
W28	P-MALA	1122	879
	MYULA	646	598
3C288	P-MALA	1144	881
	MYULA	607	538

Hypothesis testing
Method

- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

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.
   - If $\boldsymbol{x}' \in C_\alpha$ uncertainty too high to draw strong conclusions about the physical nature of the structure.

Hypothesis testing
Method

- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

Hypothesis testing of physical structure

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

Hypothesis testing
Method

- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

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
Method

- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

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

Method

- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

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

Method

- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

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
Method

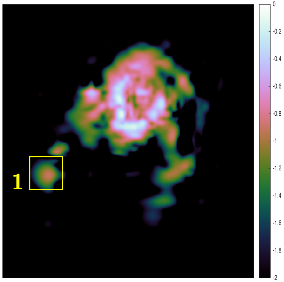
- Perform hypothesis tests of image structure using Bayesian credible regions (Pereyra 2016b).
- Let C_α denote the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.

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

Numerical experiments

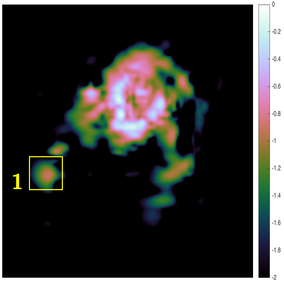


(a) Recovered image

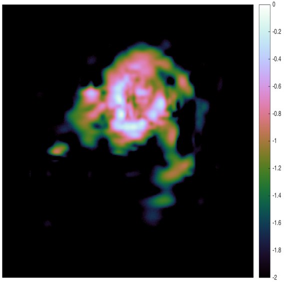
Figure: HII region of M31

Hypothesis testing

Numerical experiments



(a) Recovered image

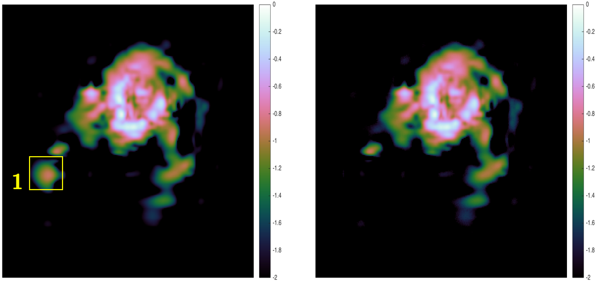


(b) Surrogate with region removed

Figure: HII region of M31

Hypothesis testing

Numerical experiments



(a) Recovered image

(b) Surrogate with region removed

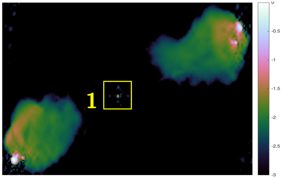
Figure: HII region of M31

1. Reject null hypothesis

   ⇒ structure physical

Hypothesis testing

Numerical experiments

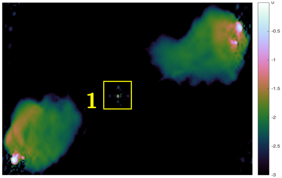


(a) Recovered image

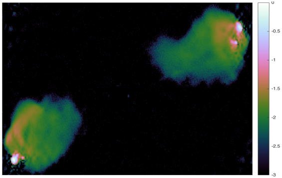
Figure: Cygnus A

Hypothesis testing

Numerical experiments



(a) Recovered image

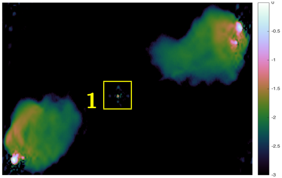


(b) Surrogate with region removed

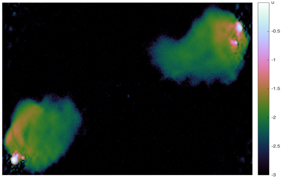
Figure: Cygnus A

Hypothesis testing

Numerical experiments



(a) Recovered image



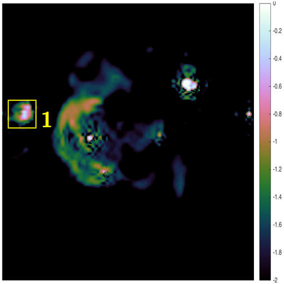
(b) Surrogate with region removed

Figure: Cygnus A

1. Cannot reject null hypothesis

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

Hypothesis testing
Numerical experiments

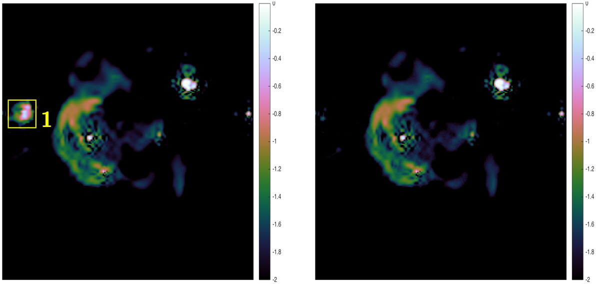


(a) Recovered image

Figure: Supernova remnant W28

Hypothesis testing

Numerical experiments

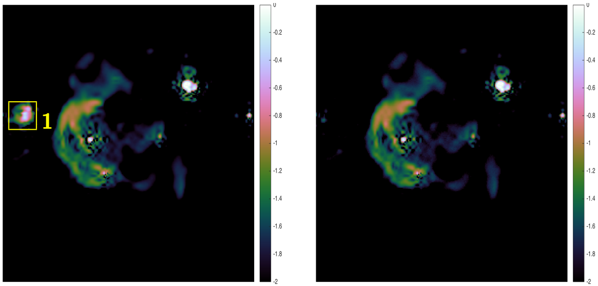


(a) Recovered image (b) Surrogate with region removed

Figure: Supernova remnant W28

Hypothesis testing

Numerical experiments



(a) Recovered image

(b) Surrogate with region removed

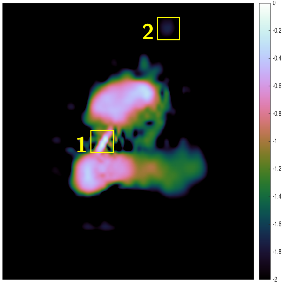
Figure: Supernova remnant W28

1. Reject null hypothesis

$\Rightarrow$ structure physical

Hypothesis testing

Numerical experiments

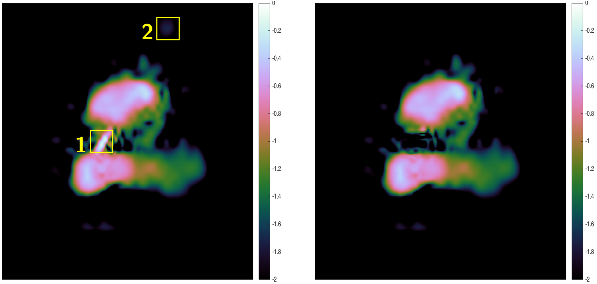


(a) Recovered image

Figure: 3C288

Hypothesis testing

Numerical experiments

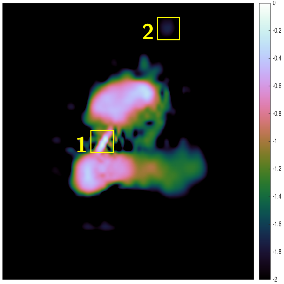


(a) Recovered image (b) Surrogate with region removed

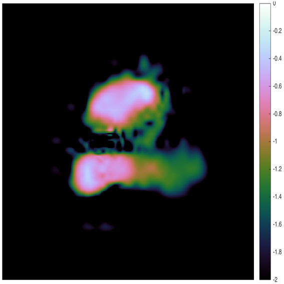
Figure: 3C288

Hypothesis testing

Numerical experiments



(a) Recovered image



(b) Surrogate with region removed

Figure: 3C288

1. Reject null hypothesis

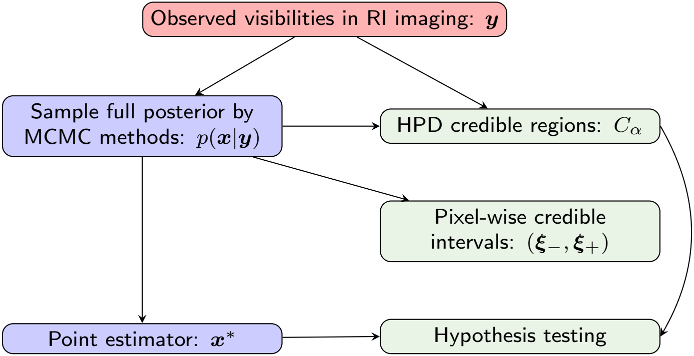
   $\Rightarrow$ structure physical

2. Cannot reject null hypothesis

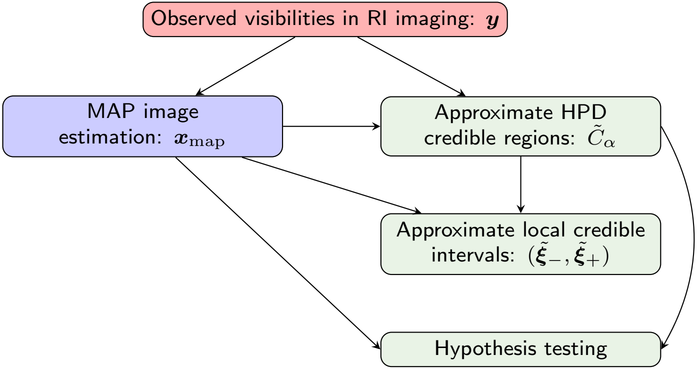
   $\Rightarrow$ cannot make strong statistical statement about origin of structure

Outline

Proximal MCMC sampling and uncertainty quantification



MAP estimation and uncertainty quantification



Approximate Bayesian credible regions for MAP estimation

- Combine uncertainty quantification with fast sparse regularisation to scale to big-data.
- Recall C_α denotes the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.
- Analytic approximation of γ_α:

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

where $\tau_\alpha = \sqrt{16 \log(3/\alpha)/N}$ and $\alpha \in (4\exp(-N/3), 1)$ (Pereyra 2016b).

- Define approximate HPD regions by $\tilde{C}_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \tilde{\gamma}_\alpha\}$.
- Compute $\boldsymbol{x}^\star$ by sparse regularisation, then estimate local Bayesian credible intervals and perform hypothesis testing using approximate HPD regions.

Approximate Bayesian credible regions for MAP estimation

- Combine uncertainty quantification with fast sparse regularisation to scale to big-data.
- Recall C_α denotes the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.
- Analytic approximation of γ_α:

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

where $\tau_\alpha = \sqrt{16 \log(3/\alpha)/N}$ and $\alpha \in (4\exp(-N/3), 1)$ (Pereyra 2016b).

- Define approximate HPD regions by $\tilde{C}_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \tilde{\gamma}_\alpha\}$.
- Compute $\boldsymbol{x}^\star$ by sparse regularisation, then estimate local Bayesian credible intervals and perform hypothesis testing using approximate HPD regions.

Approximate Bayesian credible regions for MAP estimation

- Combine uncertainty quantification with fast sparse regularisation to scale to big-data.
- Recall C_α denotes the highest posterior density (HPD) Bayesian credible region with confidence level $(1-\alpha)\%$ defined by posterior iso-contour: $C_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \gamma_\alpha\}$.
- Analytic approximation of γ_α:

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

where $\tau_\alpha = \sqrt{16\log(3/\alpha)/N}$ and $\alpha \in (4\mathrm{exp}(-N/3), 1)$ (Pereyra 2016b).

- Define approximate HPD regions by $\tilde{C}_\alpha = \{\boldsymbol{x} : g(\boldsymbol{x}) \leq \tilde{\gamma}_\alpha\}$.
- Compute $\boldsymbol{x}^\star$ by sparse regularisation, then estimate local Bayesian credible intervals and perform hypothesis testing using approximate HPD regions.

Local Bayesian credible intervals for MAP estimation

Local Bayesian credible intervals for sparse reconstruction
(Cai, Pereyra & McEwen 2017b, 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 Bayesian credible intervals for MAP estimation

Local Bayesian credible intervals for sparse reconstruction

(Cai, Pereyra & McEwen 2017b, 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

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

Local Bayesian credible intervals for MAP estimation

Local Bayesian credible intervals for sparse reconstruction

(Cai, Pereyra & McEwen 2017b, 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).

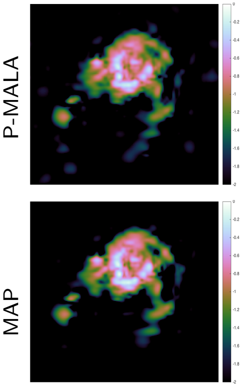
Consider the test image with the Ω region replaced by constant value ξ:

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

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

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

Numerical experiments



(a) point estimators

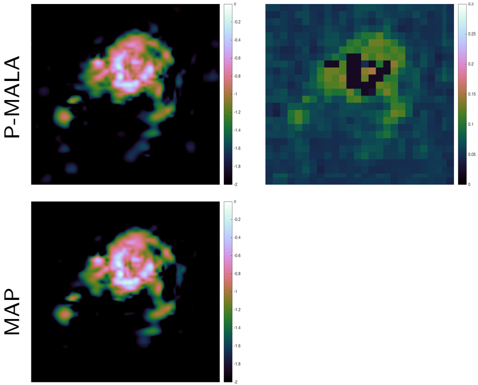
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

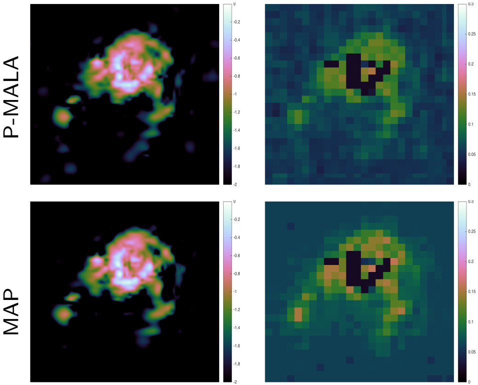
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

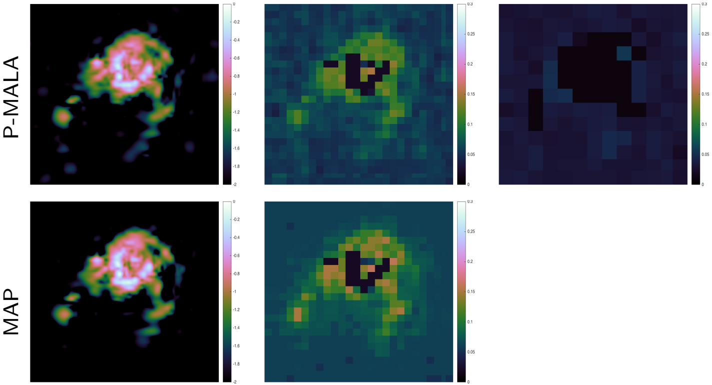
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

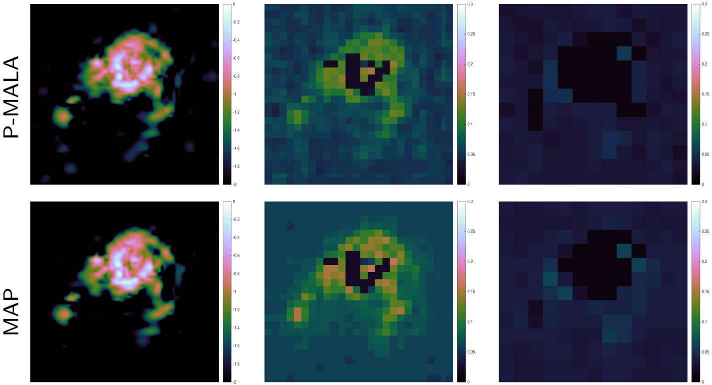
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

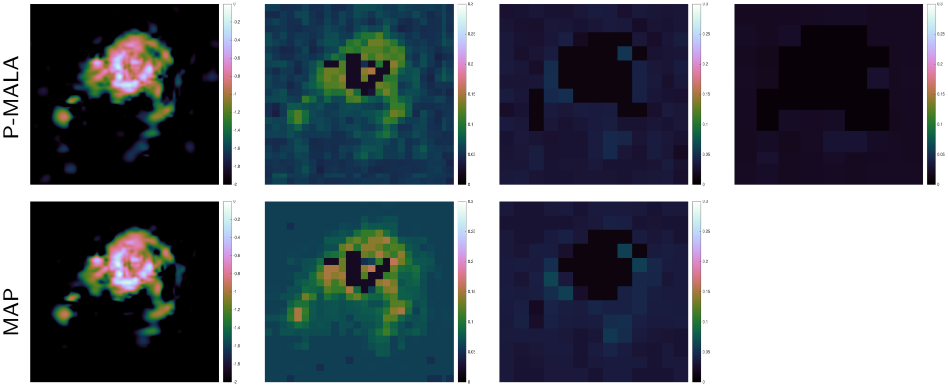
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

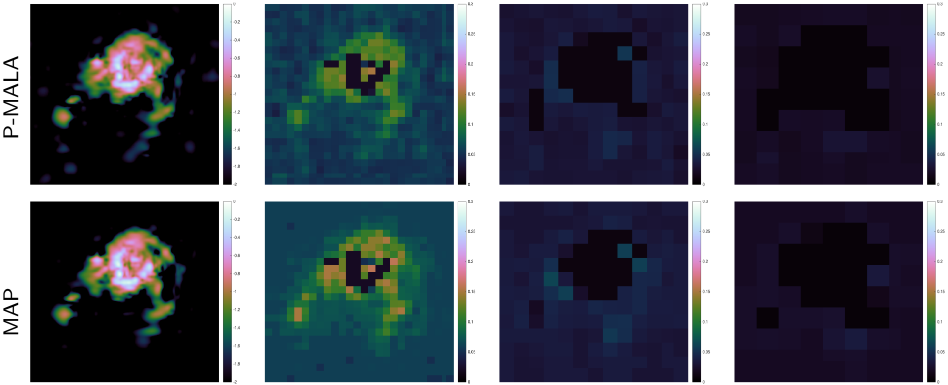
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

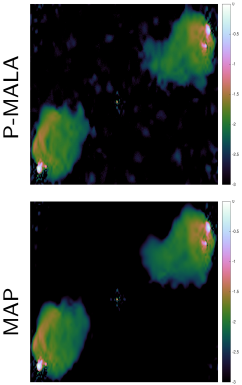
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for M31 for the analysis model.

Numerical experiments



(a) point estimators

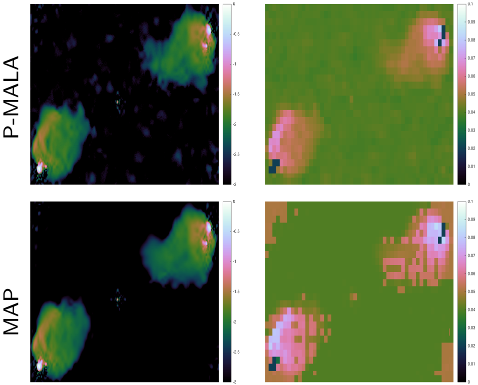
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for Cygnus A for the analysis model.

Numerical experiments



(a) point estimators

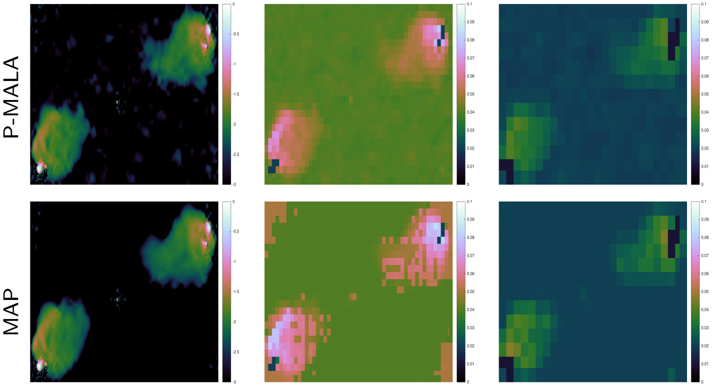
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for Cygnus A for the analysis model.

Numerical experiments



(a) point estimators

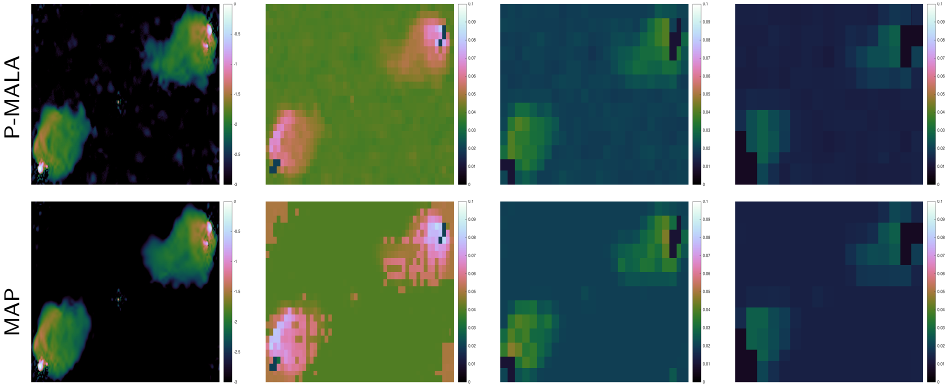
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for Cygnus A for the analysis model.

Numerical experiments



(a) point estimators

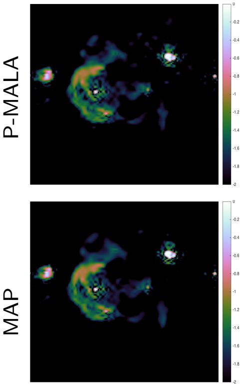
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for Cygnus A for the analysis model.

Numerical experiments



(a) point estimators

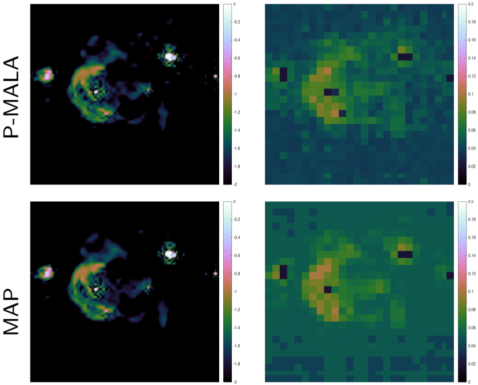
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for W28 for the analysis model.

Numerical experiments



(a) point estimators

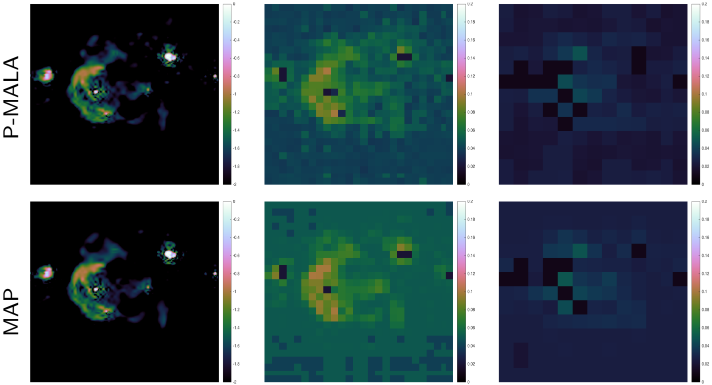
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for W28 for the analysis model.

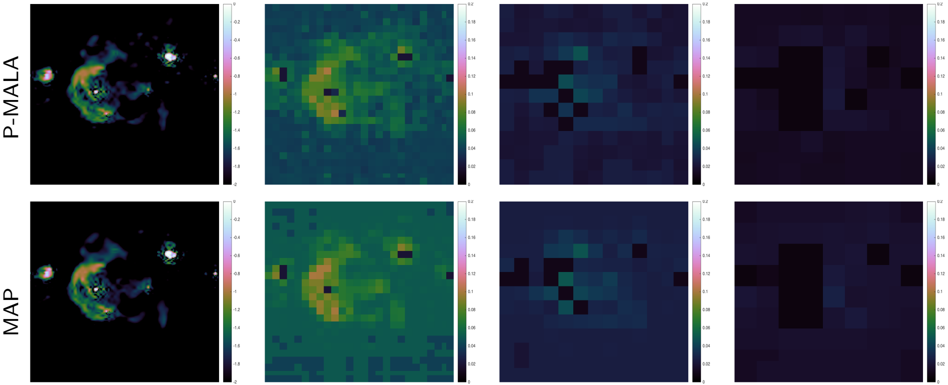
Numerical experiments



(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)

Figure: Length of local credible intervals for W28 for the analysis model.

Numerical experiments



(a) point estimators

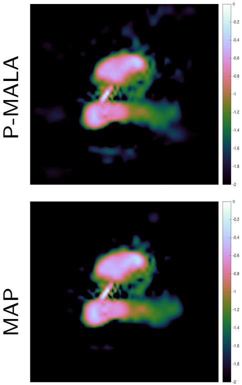
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for W28 for the analysis model.

Numerical experiments



(a) point estimators

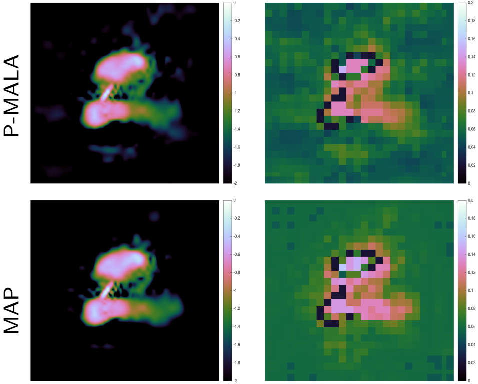
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for 3C288 for the analysis model.

Numerical experiments



(a) point estimators

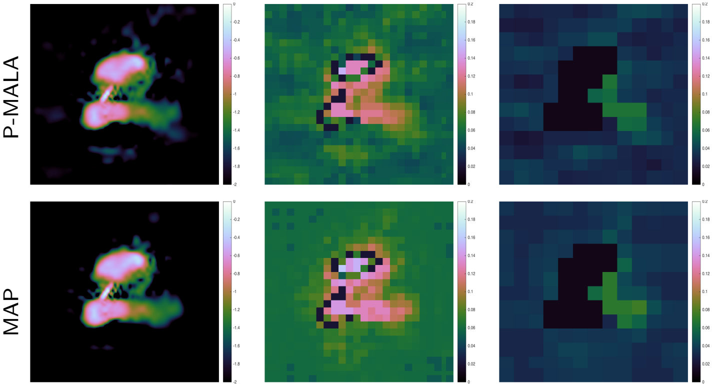
(b) local credible interval (grid size 10×10 pixels)

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

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

Figure: Length of local credible intervals for 3C288 for the analysis model.

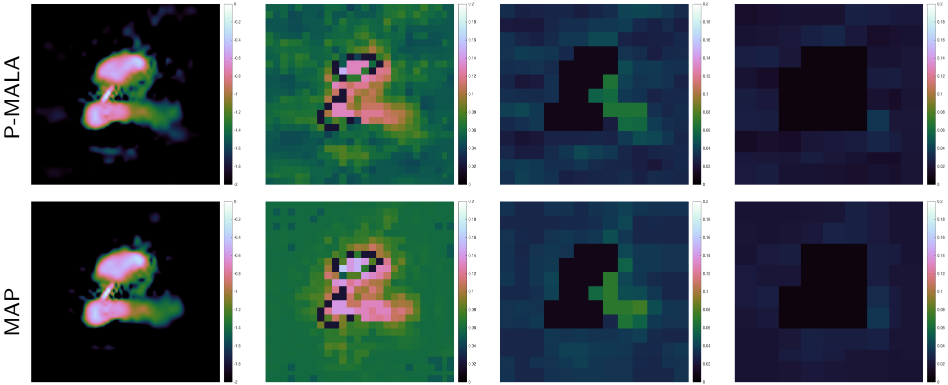
Numerical experiments



(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)

Figure: Length of local credible intervals for 3C288 for the analysis model.

Numerical experiments



(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)

Figure: Length of local credible intervals for 3C288 for the analysis model.

Computation time

Table: CPU time in minutes for Proximal MCMC sampling and MAP estimation

Image	Method	CPU time Analysis	CPU time Synthesis
Cygnus A	P-MALA	2274	1762
	MYULA	1056	942
	MAP	.07	.04
M31	P-MALA	1307	944
	MYULA	618	581
	MAP	.03	.02
W28	P-MALA	1122	879
	MYULA	646	598
	MAP	.06	.04
3C288	P-MALA	1144	881
	MYULA	607	538
	MAP	.03	.02

Hypothesis testing

Comparison of numerical experiments

Table: Comparison of hypothesis tests for different methods for the analysis model.

Image	Test area	Ground truth	Method	Hypothesis test
M31	1	✓	P-MALA	✓
			MYULA	✓
			MAP	✓
Cygnus A	1	✓	P-MALA	✗
			MYULA*	✗
			MAP	✗
W28	1	✓	P-MALA	✓
			MYULA	✓
			MAP	✓
3C288	1	✓	P-MALA	✓
			MYULA	✓
			MAP	✓
	2	✗	P-MALA	✗
			MYULA	✗
			MAP	✗

(* Can correctly detect physical structure if use median point estimator.)

Outline

Mass-mapping via weak gravitational lensing

Model

- Let $\gamma \in \mathbb{C}^M$ be the discretized complex shear field extracted from an underlying discretized convergence field $\kappa \in \mathbb{C}^N$ by a **measurement operator**

$$\boxed{\boldsymbol{\Phi} \in \mathbb{C}^{M \times N} : \kappa \mapsto \gamma} .$$

- In the planar setting $\boldsymbol{\Phi}$ can be modelled by

$$\boxed{\boldsymbol{\Phi} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}} .$$

- The planar forward model in Fourier space:

$$\hat{\gamma}(k_x, k_y) = \mathbf{D}_{k_x,k_y} \hat{\kappa}(k_x, k_y) ,$$

with the mapping operator

$$\mathbf{D}_{k_x,k_y} = \frac{k_x^2 - k_y^2 + 2ik_xk_y}{k_x^2 + k_y^2} .$$

Mass-mapping via weak gravitational lensing

Model

- Let $\gamma \in \mathbb{C}^M$ be the discretized complex shear field extracted from an underlying discretized convergence field $\kappa \in \mathbb{C}^N$ by a **measurement operator**

$$\boxed{\mathbf{\Phi} \in \mathbb{C}^{M \times N} : \kappa \mapsto \gamma} \, .$$

- In the planar setting $\mathbf{\Phi}$ can be modelled by

$$\boxed{\mathbf{\Phi} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}} \, .$$

- The planar forward model in Fourier space:

$$\hat{\gamma}(k_x, k_y) = \mathbf{D}_{k_x,k_y} \hat{\kappa}(k_x, k_y) \, ,$$

with the mapping operator

$$\mathbf{D}_{k_x,k_y} = \frac{k_x^2 - k_y^2 + 2ik_xk_y}{k_x^2 + k_y^2} \, .$$

Mass-mapping via weak gravitational lensing

Model

- Let $\gamma \in \mathbb{C}^M$ be the discretized complex shear field extracted from an underlying discretized convergence field $\kappa \in \mathbb{C}^N$ by a **measurement operator**

$$\boxed{\mathbf{\Phi} \in \mathbb{C}^{M\times N} : \kappa \mapsto \gamma} \, .$$

- In the planar setting $\mathbf{\Phi}$ can be modelled by

$$\boxed{\mathbf{\Phi} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}} \, .$$

- The planar forward model in Fourier space:

$$\hat{\gamma}(k_x, k_y) = \mathbf{D}_{k_x,k_y} \hat{\kappa}(k_x, k_y) \, ,$$

with the mapping operator

$$\mathbf{D}_{k_x,k_y} = \frac{k_x^2 - k_y^2 + 2ik_xk_y}{k_x^2 + k_y^2} \, .$$

Mass-mapping via weak gravitational lensing

Bayesian MAP estimation by sparse regularisation

- Spare Bayesian mass-mapping framework (Price, McEwen, Cai, Kitching, Wallis 2018a: arXiv:1812.04014).
- Consider **posterior** $\boxed{\mathrm{P}(\kappa \,|\, \gamma) \propto \mathrm{P}(\gamma \,|\, \kappa)\, \mathrm{P}(\kappa)}$.
- Likelihood:

$$\boxed{\mathrm{P}(\gamma \,|\, \kappa) \propto \exp\left[- \frac{(\mathbf{\Phi}\kappa - \gamma)^\dagger \Sigma^{-1} (\mathbf{\Phi}\kappa - \gamma)}{2} \right]}.$$

- General (non-Gaussian) wavelet Laplacian prior:

$$\boxed{\mathrm{P}(\kappa) \propto \exp\left(- \mu \| \mathbf{\Psi}^\dagger \kappa \|_1 \right)}.$$

- Maximum a posterior (**MAP**) solution given by solving (convex) optimisation problem (*cf.* GLIMPSE of Lanusse *et al.* 2016):

$$\boxed{\kappa^{\mathrm{map}} = \underset{\kappa}{\operatorname{argmin}} \left[\mu \| \mathbf{\Psi}^\dagger \kappa \|_1 + \frac{\| \mathbf{\Phi}\kappa - \gamma \|_2^2}{2\sigma_n^2} \right]},$$

Mass-mapping via weak gravitational lensing

Bayesian MAP estimation by sparse regularisation

- Spare Bayesian mass-mapping framework (Price, McEwen, Cai, Kitching, Wallis 2018a: arXiv:1812.04014).
- Consider **posterior** $\boxed{\mathrm{P}(\kappa \,|\, \gamma) \propto \mathrm{P}(\gamma \,|\, \kappa)\, \mathrm{P}(\kappa)}$.
- Likelihood:

$$\boxed{\mathrm{P}(\gamma \,|\, \kappa) \propto \exp\left[- \frac{(\boldsymbol{\Phi}\kappa - \gamma)^\dagger \Sigma^{-1} (\boldsymbol{\Phi}\kappa - \gamma)}{2} \right]}.$$

- General (non-Gaussian) wavelet Laplacian prior:

$$\boxed{\mathrm{P}(\kappa) \propto \exp\left(- \mu \|\boldsymbol{\Psi}^\dagger \kappa\|_1 \right)}.$$

- Maximum a posterior (**MAP**) solution given by solving (convex) optimisation problem (*cf.* GLIMPSE of Lanusse *et al.* 2016):

$$\boxed{\kappa^{\mathsf{map}} = \underset{\kappa}{\operatorname{argmin}} \left[\mu \|\boldsymbol{\Psi}^\dagger \kappa\|_1 + \frac{\|\boldsymbol{\Phi}\kappa - \gamma\|_2^2}{2\sigma_n^2} \right]},$$

Mass-mapping via weak gravitational lensing

Bayesian MAP estimation by sparse regularisation

- Spare Bayesian mass-mapping framework
  (Price, McEwen, Cai, Kitching, Wallis 2018a: arXiv:1812.04014).

- Consider **posterior** $\boxed{\mathrm{P}(\kappa \,|\, \gamma) \propto \mathrm{P}(\gamma \,|\, \kappa)\, \mathrm{P}(\kappa)}$.

- Likelihood:

$$\boxed{\mathrm{P}(\gamma \,|\, \kappa) \propto \exp\left[- \frac{(\boldsymbol{\Phi}\kappa - \gamma)^\dagger \Sigma^{-1} (\boldsymbol{\Phi}\kappa - \gamma)}{2} \right]} .$$

- General (non-Gaussian) wavelet Laplacian prior:

$$\boxed{\mathrm{P}(\kappa) \propto \exp\left(- \mu \|\boldsymbol{\Psi}^\dagger \kappa\|_1 \right)} .$$

- Maximum a posterior (**MAP**) solution given by solving (convex) optimisation problem (*cf.* GLIMPSE of Lanusse *et al.* 2016):

$$\boxed{\kappa^{\text{map}} = \underset{\kappa}{\operatorname{argmin}} \left[\mu \|\boldsymbol{\Psi}^\dagger \kappa\|_1 + \frac{\|\boldsymbol{\Phi}\kappa - \gamma\|_2^2}{2\sigma_n^2} \right]},$$

Mass-mapping via weak gravitational lensing

Bayesian MAP estimation by sparse regularisation

- Spare Bayesian mass-mapping framework
  (Price, McEwen, Cai, Kitching, Wallis 2018a: arXiv:1812.04014).

- Consider **posterior** $\boxed{\mathrm{P}(\kappa \,|\, \gamma) \propto \mathrm{P}(\gamma \,|\, \kappa)\, \mathrm{P}(\kappa)}$.

- Likelihood:

$$\boxed{\mathrm{P}(\gamma \,|\, \kappa) \propto \exp\left[- \frac{(\boldsymbol{\Phi}\kappa - \gamma)^\dagger \Sigma^{-1} (\boldsymbol{\Phi}\kappa - \gamma)}{2} \right]} .$$

- General (non-Gaussian) wavelet Laplacian prior:

$$\boxed{\mathrm{P}(\kappa) \propto \exp\big(- \mu \| \boldsymbol{\Psi}^\dagger \kappa \|_1 \big)} .$$

- Maximum a posterior (**MAP**) solution given by solving (convex) optimisation problem
  (*cf.* GLIMPSE of Lanusse *et al.* 2016):

$$\boxed{\kappa^{\textsf{map}} = \underset{\kappa}{\operatorname{argmin}} \left[\mu \| \boldsymbol{\Psi}^\dagger \kappa \|_1 + \frac{\| \boldsymbol{\Phi}\kappa - \gamma \|_2^2}{2\sigma_n^2} \right]} ,$$

Mass-mapping via weak gravitational lensing

Selection of the regularisation parameter

- How set regularisation parameter μ?
- Set up **gamma-type hyper-prior** (typical hyper-prior for scale-parameters) following Pereyra *et al.* (2015):

$$\boxed{\mathrm{P}(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu} \mathbb{I}_{\mathbb{R}_+}(\mu)} ,$$

  where without loss of generality $\alpha = \beta = 1$ (results highly insensitive to choice of α and β).
- Compute the **joint MAP estimator** $(\kappa^{\text{map}}, \mu^{\text{map}})$, which maximizes $\mathrm{P}(\kappa, \mu \,|\, \gamma)$ such that

$$\mathbf{0}_{N+1} \in \partial_{\kappa,\mu} \log p(\kappa^{\text{map}}, \mu^{\text{map}} \,|\, \gamma) .$$

- Yields textbfanalytic update for μ estimator (Pereyra *et al.* 2015):

$$\boxed{\mu^{t+1} = \frac{Nk^{-1} + \alpha - 1}{f(\kappa^t) + \beta}} ,$$

  where $f(\cdot)$ is the log-prior.

Mass-mapping via weak gravitational lensing

Selection of the regularisation parameter

- How set regularisation parameter μ?
- Set up **gamma-type hyper-prior** (typical hyper-prior for scale-parameters) following Pereyra *et al.* (2015):

$$\boxed{\mathsf{P}(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu} \mathbb{I}_{\mathbb{R}_+}(\mu)} ,$$

  where without loss of generality $\alpha = \beta = 1$ (results highly insensitive to choice of α and β).

- Compute the **joint MAP estimator** $(\kappa^{\text{map}}, \mu^{\text{map}})$, which maximizes $\mathsf{P}(\kappa, \mu \,|\, \gamma)$ such that

$$\mathbf{0}_{N+1} \in \partial_{\kappa,\mu} \log p(\kappa^{\text{map}}, \mu^{\text{map}} \,|\, \gamma) .$$

- Yields textbfanalytic update for μ estimator (Pereyra *et al.* 2015):

$$\boxed{\mu^{t+1} = \frac{Nk^{-1} + \alpha - 1}{f(\kappa^t) + \beta}} ,$$

  where $f(\cdot)$ is the log-prior.

Mass-mapping via weak gravitational lensing

Selection of the regularisation parameter

- How set regularisation parameter μ?
- Set up **gamma-type hyper-prior** (typical hyper-prior for scale-parameters) following Pereyra *et al.* (2015):

$$\boxed{\mathrm{P}(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu} \mathbb{I}_{\mathbb{R}_+}(\mu)} ,$$

where without loss of generality $\alpha = \beta = 1$ (results highly insensitive to choice of α and β).

- Compute the **joint MAP estimator** $(\kappa^{\textsf{map}}, \mu^{\textsf{map}})$, which maximizes $\mathrm{P}(\kappa, \mu \,|\, \gamma)$ such that

$$\mathbf{0}_{N+1} \in \partial_{\kappa,\mu} \log p(\kappa^{\textsf{map}}, \mu^{\textsf{map}} \,|\, \gamma) .$$

- Yields textbfanalytic update for μ estimator (Pereyra *et al.* 2015):

$$\boxed{\mu^{t+1} = \frac{Nk^{-1} + \alpha - 1}{f(\kappa^t) + \beta}} ,$$

where $f(\cdot)$ is the log-prior.

Mass-mapping via weak gravitational lensing

Selection of the regularisation parameter

- How set regularisation parameter μ?
- Set up **gamma-type hyper-prior** (typical hyper-prior for scale-parameters) following Pereyra *et al.* (2015):

$$\boxed{\mathrm{P}(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu} \mathbb{I}_{\mathbb{R}_+}(\mu)} \, ,$$

  where without loss of generality $\alpha = \beta = 1$ (results highly insensitive to choice of α and β).
- Compute the **joint MAP estimator** $(\kappa^{\textsf{map}}, \mu^{\textsf{map}})$, which maximizes $\mathrm{P}(\kappa, \mu \,|\, \gamma)$ such that

$$\mathbf{0}_{N+1} \in \partial_{\kappa,\mu} \log p(\kappa^{\textsf{map}}, \mu^{\textsf{map}} \,|\, \gamma) \, .$$

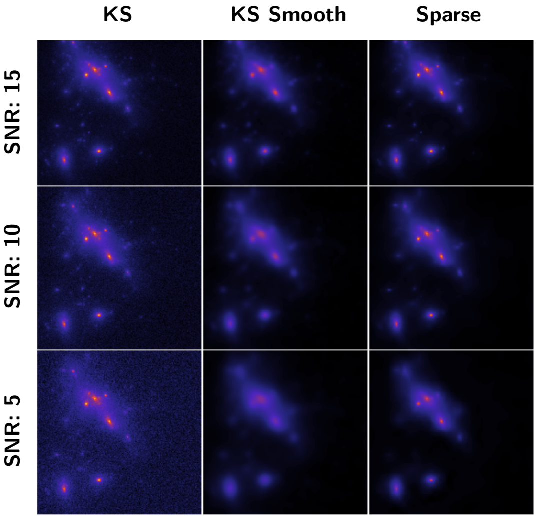
- Yields textbfanalytic update for μ estimator (Pereyra *et al.* 2015):

$$\boxed{\mu^{t+1} = \frac{Nk^{-1} + \alpha - 1}{f(\kappa^t) + \beta}} \, ,$$

  where $f(\cdot)$ is the log-prior.

Bayesian sparse mass-mapping

Recovering mass-maps from simulations



Bayesian sparse mass-mapping

Recovering mass-maps from simulations

SNR (dB)				
Input SNR	**KS**	**KS Smooth**	**Sparse**	**Difference**
20.0	3.986	3.988	9.298	+ 5.310
15.0	3.844	3.912	9.906	+ 5.993
10.0	3.480	3.831	9.230	+ 5.399
5.0	2.670	3.0305	7.296	+ 4.265

Hypothesis testing of structure

Single object structure

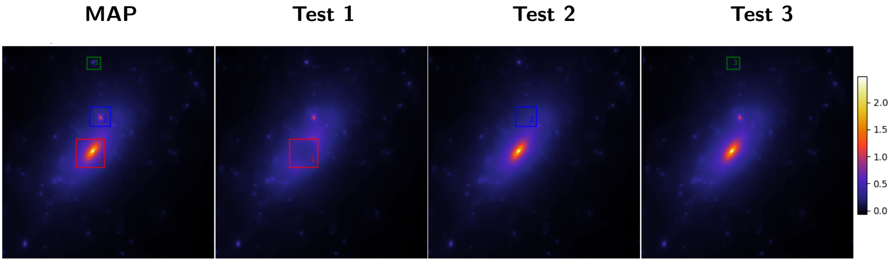


Figure: Hypothesis testing of three selected structures in the Bolshoi-1 cluster convergence field. The SNR of added Gaussian noise was 20 dB. The SNR of the sparse recovery was $\sim$ 6 dB (an increase in SNR of ~ 3.5 dB over the KS reconstruction). We correctly determine that region 1 (*red*) is physical with 99% confidence. Regions 2 (*blue*) and 3 (*green*) remain within the HPD region and are therefore inconclusive, given the data and noise level.

Hypothesis testing of structure

Multiple object structure

MAP **Test 1** **Test 2** **Test 3**

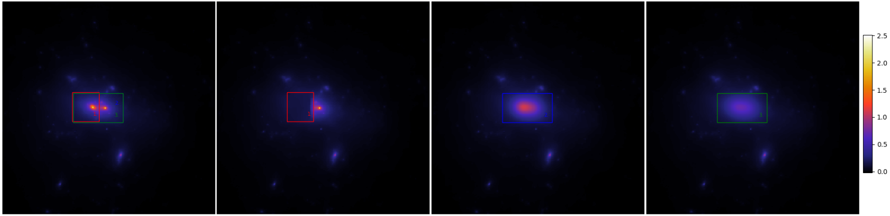


Figure: Hypothesis testing of three selected structures in the Bolshoi-2 cluster convergence field. The SNR of added Gaussian noise was 20 dB. The SNR of the sparse recovery was $\sim$ 12 dB (an increase in SNR of $\sim$ 7 dB over the KS reconstruction). We correctly determine that all three null hypothesis' (*red*, *blue* and *green*) are rejected at 99% confidence. In test 1 the conclusion is that the left hand peak was statistically significant. In tests 2 and 3 the conclusions is that an image with the two peaks merged it unacceptable, and therefore the peaks are distinct at 99% confidence.

Hypothesis testing of structure

Complex structure

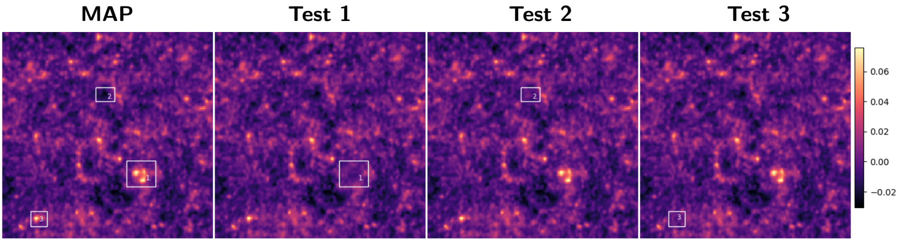


Figure: Hypothesis testing of structure in an $\sim 1.2\,\mathrm{deg}^2$ planar Buzzard extract. Both over-densities 1 and 3 are deemed to be physical, whereas the void structure 2 is inconclusive.

Analysis of A520 cluster

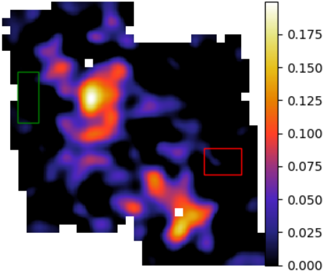
Evidence for self-interacting dark matter?

- Some controversy over peaks recovered from observations of A520 cluster (Jee *et al.* 2012, 2014, Clowe *et al.* 2012).
- A small, central convergence peak detected (J12, J14), with a notably large mass-to-light ratio, which **could indicate the possibility of self-interacting dark matter**.
- Peel *et al.* (2017) concluded that peak existed in the J14 dataset but not in the C12 dataset (using GLIMPSE; Lanusse *et al.* 2016) but cannot confirm its existence or otherwise.

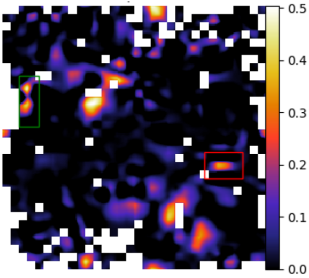
Analysis of A520 cluster

Evidence for self-interacting dark matter?

- Recovered mass-maps and perform local and global hypothesis tests.
- Data-sets are **globally consistent** at 99% credible level.
- Peak in question is detected in J14 but determined **not statistically significant**.
- Also discover some **new peaks** but they are also **not statistically significant**.



(a) J14



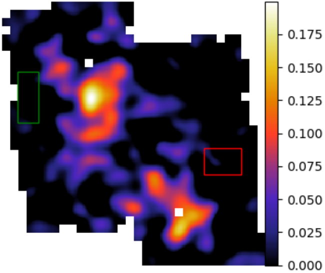
(b) C12

Figure: Recovered mass maps

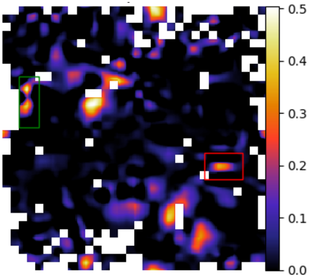
Analysis of A520 cluster

Evidence for self-interacting dark matter?

- Recovered mass-maps and perform local and global hypothesis tests.
- Data-sets are **globally consistent** at 99% credible level.
- Peak in question is detected in J14 but determined **not statistically significant**.
- Also discover some **new peaks** but they are also **not statistically significant**.



(a) J14



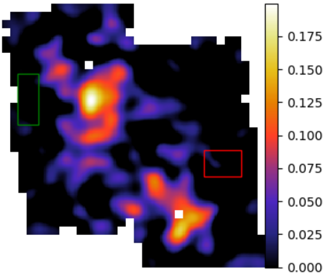
(b) C12

Figure: Recovered mass maps

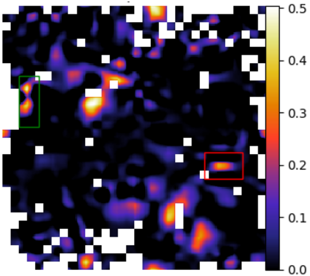
Analysis of A520 cluster

Evidence for self-interacting dark matter?

- Recovered mass-maps and perform local and global hypothesis tests.
- Data-sets are **globally consistent** at 99% credible level.
- Peak in question is detected in J14 but determined **not statistically significant**.
- Also discover some **new peaks** but they are also **not statistically significant**.



(a) J14



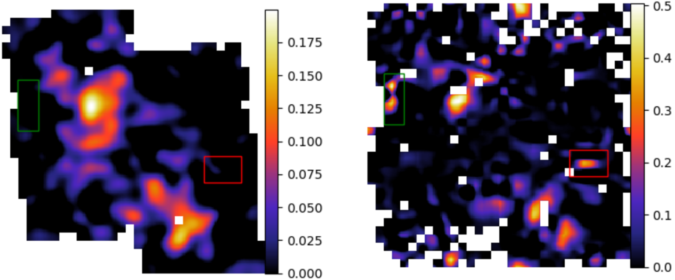
(b) C12

Figure: Recovered mass maps

Analysis of A520 cluster

Evidence for self-interacting dark matter?

- Recovered mass-maps and perform local and global hypothesis tests.
- Data-sets are **globally consistent** at 99% credible level.
- Peak in question is detected in J14 but determined **not statistically significant**.
- Also discover some **new peaks** but they are also **not statistically significant**.



(a) J14 (b) C12

Figure: Recovered mass maps

Local Bayesian credible intervals

Bolshoi simulation

- Recover local credible intervals from MAP solution and compare to MCMC reconstructions (Price, Cai, McEwen, Pereyra, Kitching 2018b: arXiv:1812.04017).

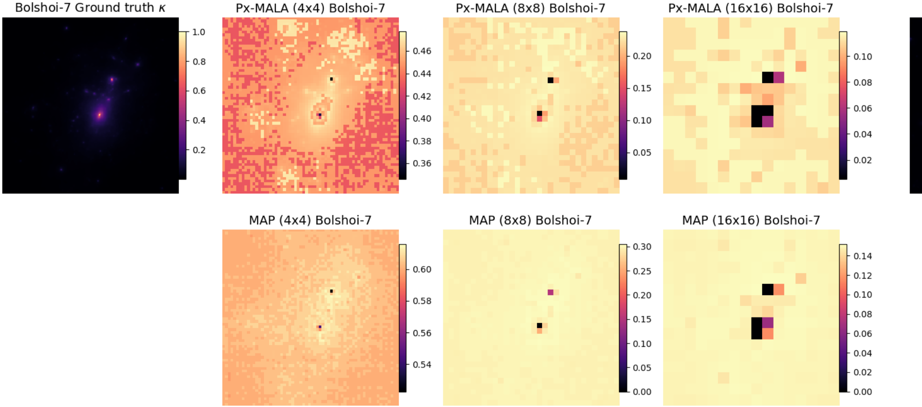


Figure: Length of local credible intervals at 99% credible level for Bolshoi simulation.

Local Bayesian credible intervals
Buzzard simulation

- Recover local credible intervals from MAP solution and compare to MCMC reconstructions (Price, Cai, McEwen, Pereyra, Kitching 2018b: arXiv:1812.04017).

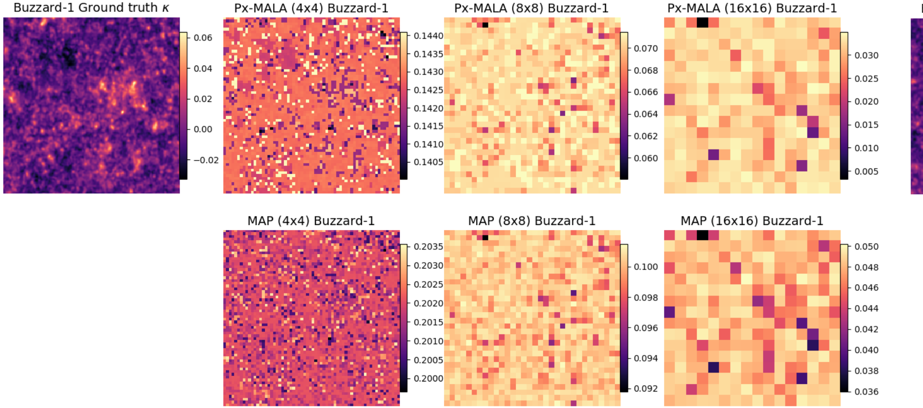


Figure: Length of local credible intervals at 99% credible level for Buzzard simulation.

Feature locations and peak statistics

- Quantify uncertainties associated with peak locations and counts (Price, McEwen, Cai, Kitching 2018c: arXiv:1812.04018).

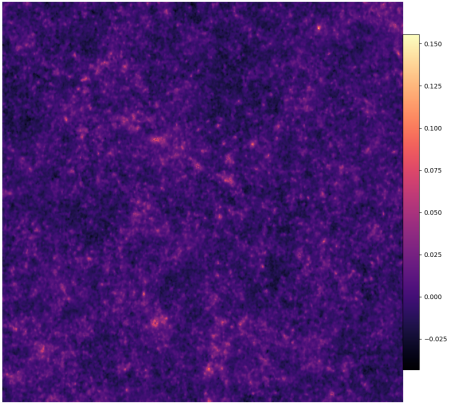
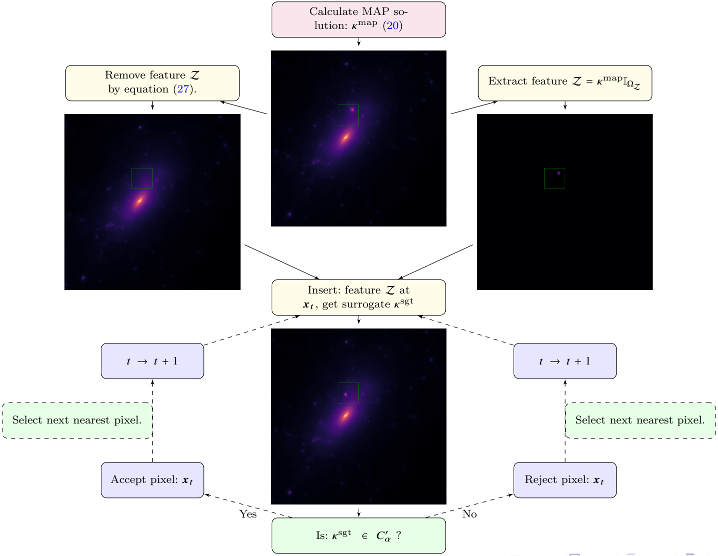


Figure: Input 2048×2048 convergence map extracted from the Buzzard N-body simulation.

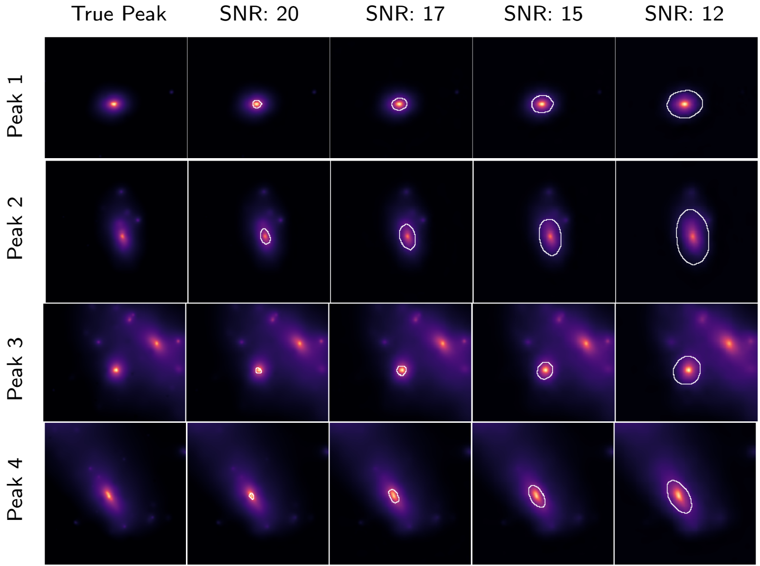
Feature locations

Procedure



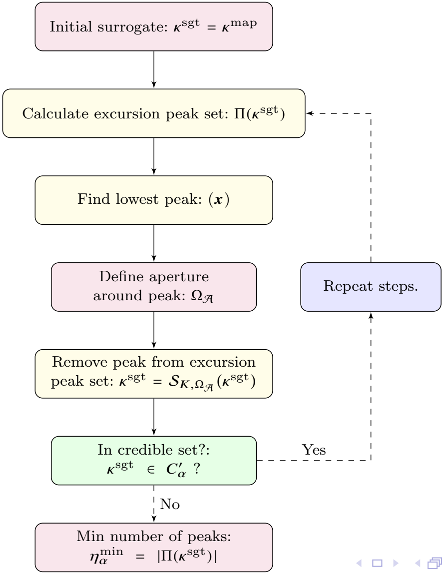
Feature locations

Results



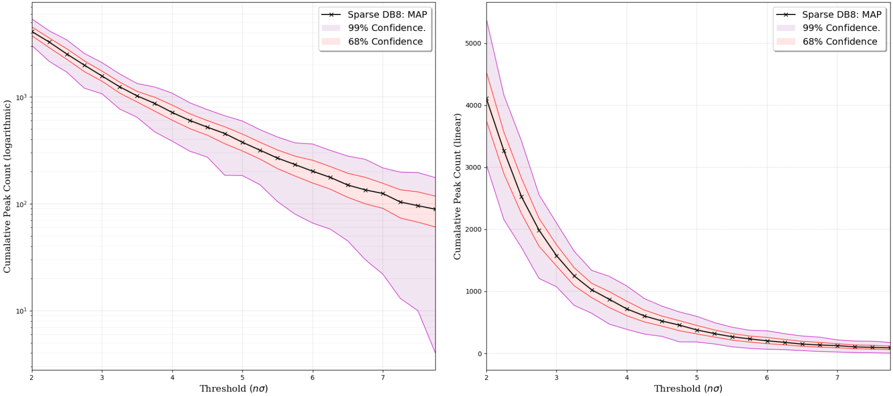
Peak statistics

Procedure



Peak statistics

Results



Conclusions

- **Uncertainty quantification** of increasing importance for principled, robust scientific inference of large, complex data-sets.
- **Multidisciplinary techniques** to quantify uncertainties in high-dimensional settings.
  - Machine learning assisted **Bayesian evidence** computation
  - **Proximal MCMC** sampling can support sparse priors in full Bayesian framework.
  - **Sparse regularisation by MAP estimation** with **approximate uncertainty quantification**.
- Numerous uses in astronomy and beyond.
  - **Radio interferometric imaging**.
  - **Mass-mapping** via weak gravitational lensing.

Supported by:

Conclusions

- **Uncertainty quantification** of increasing importance for principled, robust scientific inference of large, complex data-sets.
- **Multidisciplinary techniques** to quantify uncertainties in high-dimensional settings.
  - Machine learning assisted **Bayesian evidence** computation
  - **Proximal MCMC** sampling can support sparse priors in full Bayesian framework.
  - **Sparse regularisation by MAP estimation** with **approximate uncertainty quantification**.
- Numerous uses in astronomy and beyond.
  - **Radio interferometric imaging**.
  - **Mass-mapping** via weak gravitational lensing.

Supported by:

Conclusions

- **Uncertainty quantification** of increasing importance for principled, robust scientific inference of large, complex data-sets.
- **Multidisciplinary techniques** to quantify uncertainties in high-dimensional settings.
  - Machine learning assisted **Bayesian evidence** computation
  - **Proximal MCMC** sampling can support sparse priors in full Bayesian framework.
  - **Sparse regularisation by MAP estimation** with **approximate uncertainty quantification**.
- Numerous uses in astronomy and beyond.
  - **Radio interferometric imaging**.
  - **Mass-mapping** via weak gravitational lensing.

Supported by:

Extra Slides

Analysis vs synthesis

Bayesian interpretations

Distribution and parallelisation

PURIFY reconstructions

Extra Slides

Analysis vs synthesis

Analysis vs synthesis

- Typically sparsity assumption is justified by analysing example signals in terms of atoms of the dictionary.
- Different to synthesising signals from atoms.
- Suggests an analysis-based framework (Elad *et al.* 2007, Nam *et al.* 2012):

$$\boxed{\boldsymbol{x}^\star = \underset{\boldsymbol{x}}{\arg\min} \, \|\boldsymbol{\Omega}\boldsymbol{x}\|_1 \ \text{ subject to } \ \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2 \leq \epsilon \, .}$$

analysis

- Contrast with synthesis-based approach:

$$\boxed{\boldsymbol{x}^\star = \Psi \cdot \underset{\boldsymbol{\alpha}}{\arg\min} \, \|\boldsymbol{\alpha}\|_1 \ \text{ subject to } \ \|\boldsymbol{y} - \Phi\Psi\boldsymbol{\alpha}\|_2 \leq \epsilon \, .}$$

synthesis

- For orthogonal bases $\boldsymbol{\Omega} = \Psi^\dagger$ and the two approaches are identical.

Analysis vs synthesis

Comparison

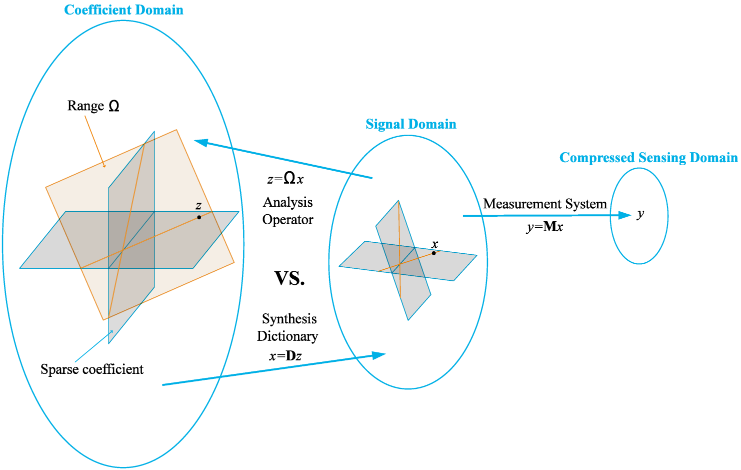


Figure: Analysis- and synthesis-based approaches **[Credit: Nam *et al.* (2012)]**.

Analysis vs synthesis
Comparison

- Synthesis-based approach is more general, while analysis-based approach more restrictive.
- More restrictive analysis-based approach may make it more robust to noise.
- The greater descriptive power of the synthesis-based approach may provide better signal representations (too descriptive?).

Extra Slides

Bayesian interpretations

Bayesian interpretations

One Bayesian interpretation of the synthesis-based approach

- Consider the inverse problem:
$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha} + \boldsymbol{n}\,.$$

- Assume Gaussian noise, yielding the likelihood:
$$\mathsf{P}(\boldsymbol{y}\,|\,\boldsymbol{\alpha}) \propto \exp\Big(\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha}\|_2^2/(2\sigma^2)\Big)\,.$$

- Consider the Laplacian prior:
$$\mathsf{P}(\boldsymbol{\alpha}) \propto \exp\Big(-\beta\|\boldsymbol{\alpha}\|_1\Big)\,.$$

- The maximum *a-posteriori* (MAP) estimate (with $\lambda = 2\beta\sigma^2$) is

$$\boxed{\boldsymbol{x}^\star_{\text{MAP-synthesis}} = \Psi \cdot \underset{\boldsymbol{\alpha}}{\text{arg max}}\, \mathsf{P}(\boldsymbol{\alpha}\,|\,\boldsymbol{y}) = \Psi \cdot \underset{\boldsymbol{\alpha}}{\text{arg min}}\, \|\boldsymbol{y} - \Phi\Psi\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\,.}$$

synthesis

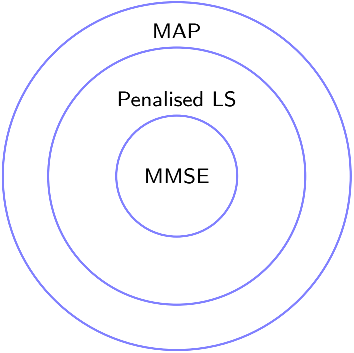
- One possible Bayesian interpretation!

- Signal may be ℓ_0-sparse, then solving ℓ_1 problem finds the correct ℓ_0-sparse solution!

Bayesian interpretations

Other Bayesian interpretations of the synthesis-based approach

- Other Bayesian interpretations are also possible (Gribonval 2011).
- Minimum mean square error (MMSE) estimators
  - $\subset$ synthesis-based estimators with appropriate penalty function, *i.e.* penalised least-squares (LS)
  - $\subset$ MAP estimators



Bayesian interpretations

One Bayesian interpretation of the analysis-based approach

- Analysis-based MAP estimate is

$$\boxed{\boldsymbol{x}^{\star}_{\text{MAP-analysis}} = \boldsymbol{\Omega}^{\dagger} \cdot \underset{\boldsymbol{\gamma} \in \textbf{column space } \boldsymbol{\Omega}}{\arg\min} \|\boldsymbol{y} - \Phi \boldsymbol{\Omega}^{\dagger} \boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \, .}$$

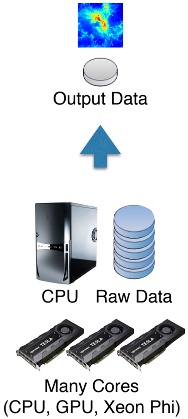
analysis

- Different to synthesis-based approach if analysis operator $\boldsymbol{\Omega}$ is not an orthogonal basis.
- Analysis-based approach more restrictive than synthesis-based.
- Similar ideas promoted by Maisinger, Hobson & Lasenby (2004) in a Bayesian framework for wavelet MEM (maximum entropy method).

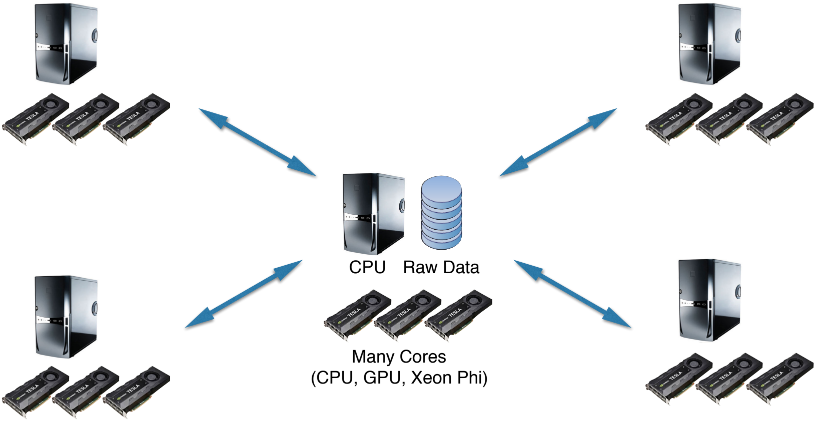
Extra Slides

Distribution and parallelisation

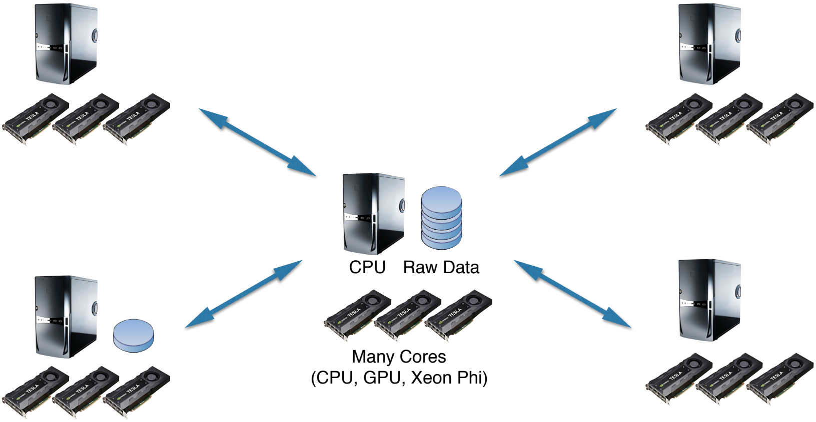
Standard algorithms



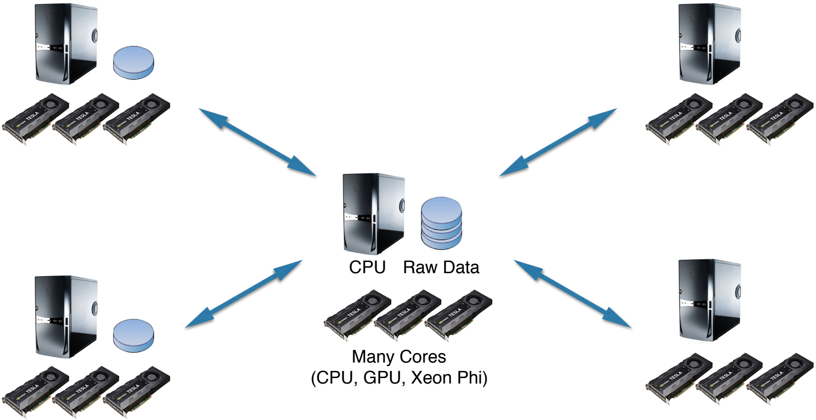
Highly distributed and parallelised algorithms



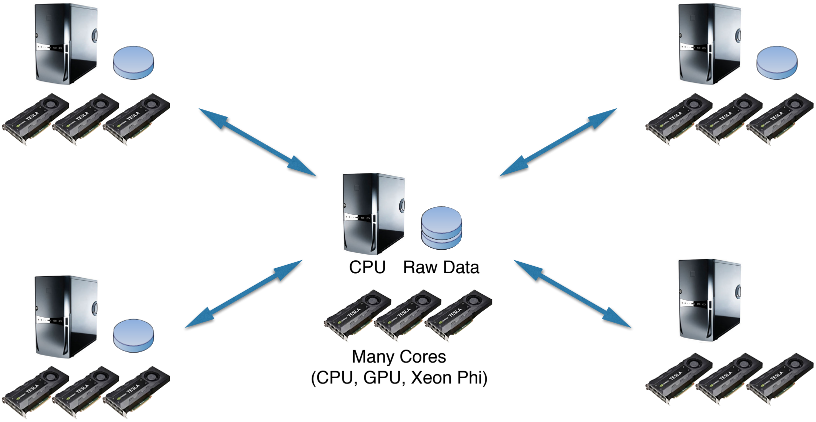
Highly distributed and parallelised algorithms



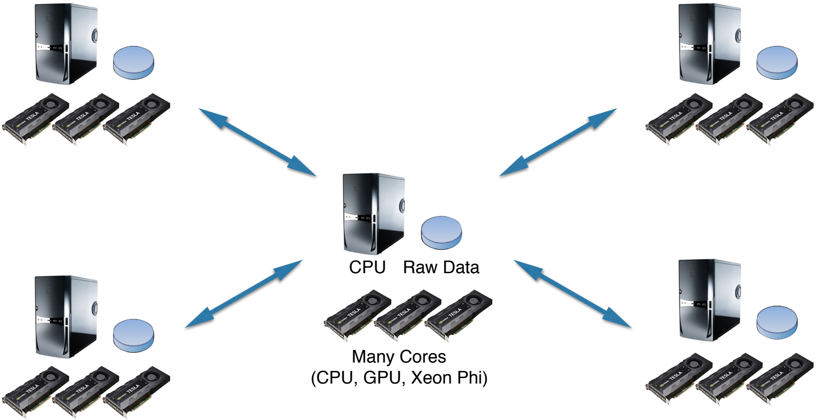
Highly distributed and parallelised algorithms



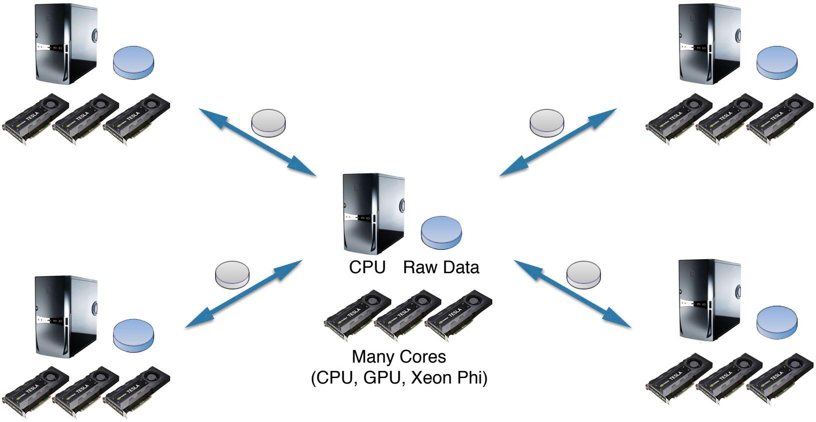
Highly distributed and parallelised algorithms



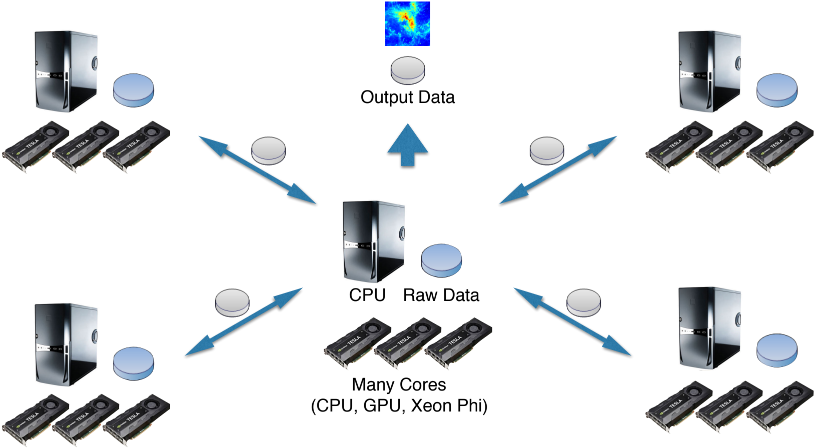
Highly distributed and parallelised algorithms



Highly distributed and parallelised algorithms



Highly distributed and parallelised algorithms



Extra Slides

PURIFY reconstructions

PURIFY reconstruction

VLA observation of 3C129

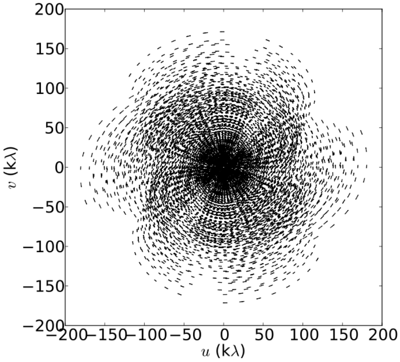
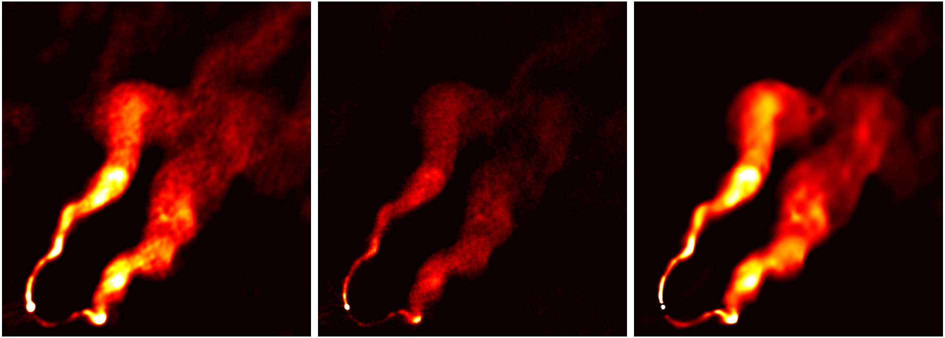


Figure: VLA visibility coverage for 3C129

PURIFY reconstruction
VLA observation of 3C129

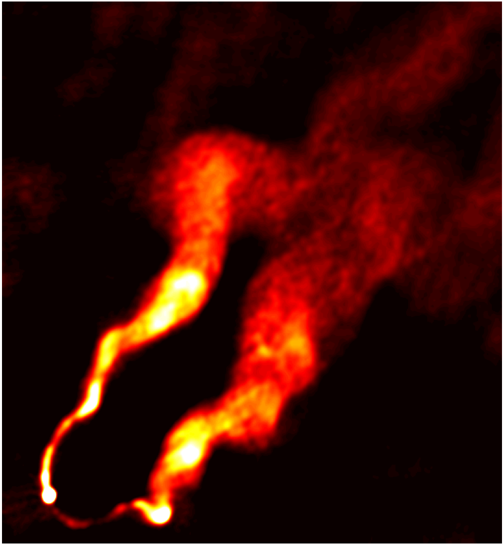


(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: 3C129 recovered images (Pratley, McEwen, *et al.* 2016)

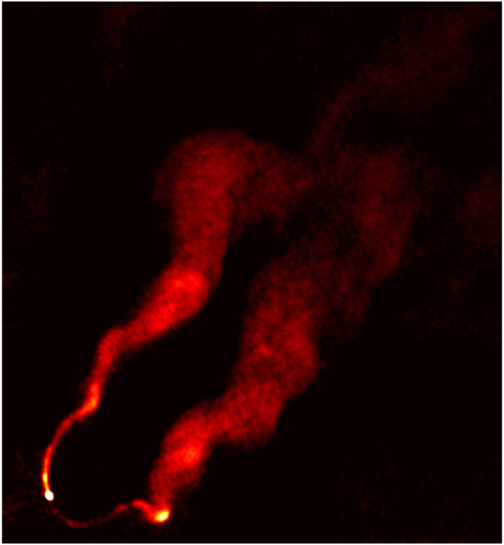
PURIFY reconstruction

VLA observation of 3C129 imaged by CLEAN (natural)



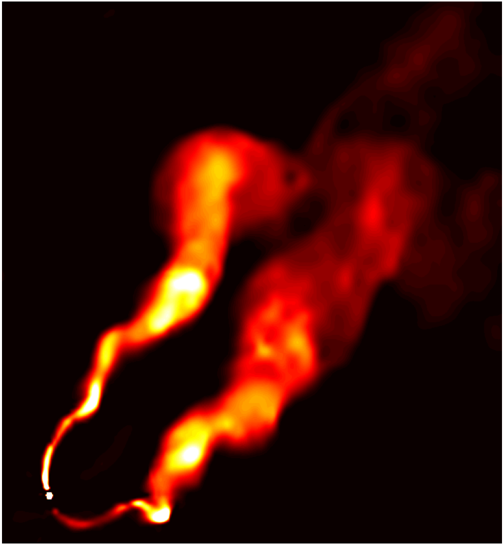
PURIFY reconstruction

VLA observation of 3C129 images by CLEAN (uniform)



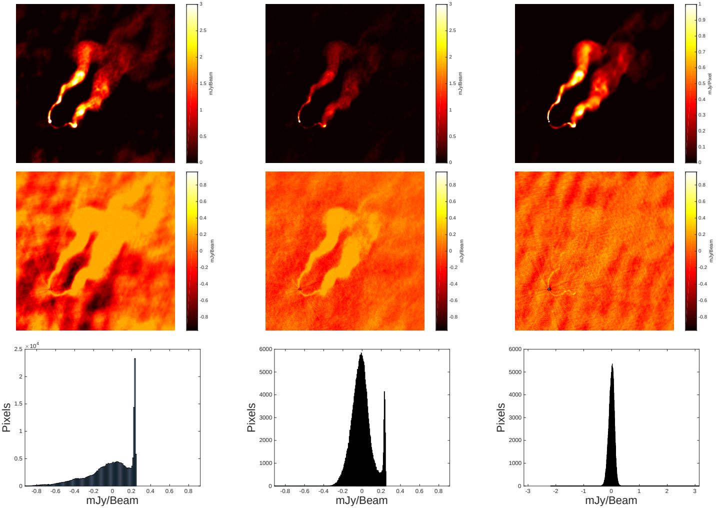
PURIFY reconstruction

VLA observation of 3C129 images by PURIFY



PURIFY reconstruction

VLA observation of 3C129



(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: 3C129 recovered images and residuals (Pratley, McEwen, *et al.* 2016)

PURIFY reconstruction
VLA observation of Cygnus A

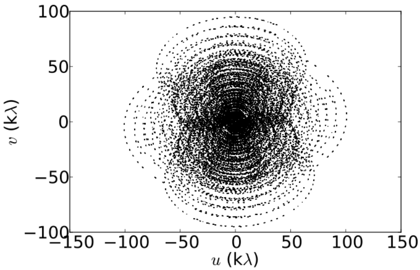
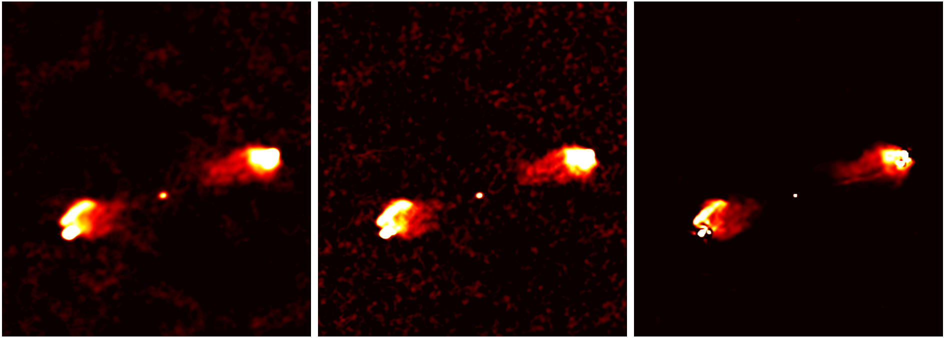


Figure: VLA visibility coverage for Cygnus A

PURIFY reconstruction

VLA observation of Cygnus A

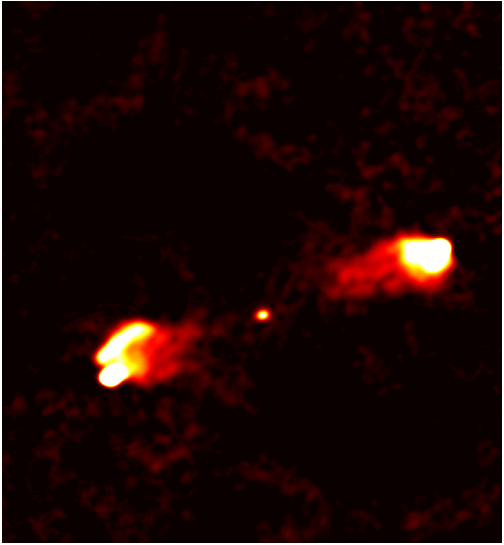


(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: Cygnus A recovered images (Pratley, McEwen, *et al.* 2016)

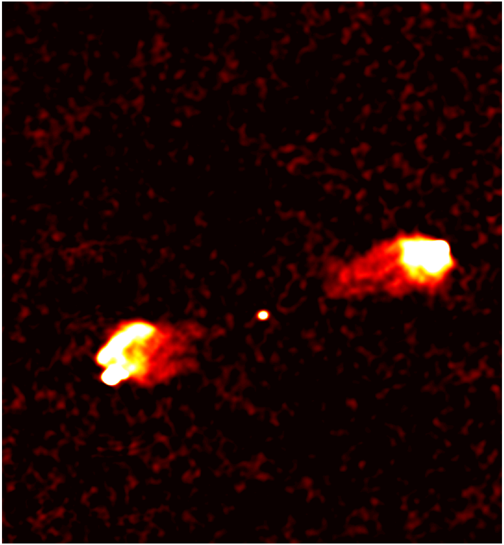
PURIFY reconstruction

VLA observation of Cygnus A imaged by CLEAN (natural)



PURIFY reconstruction

VLA observation of Cygnus A images by CLEAN (uniform)

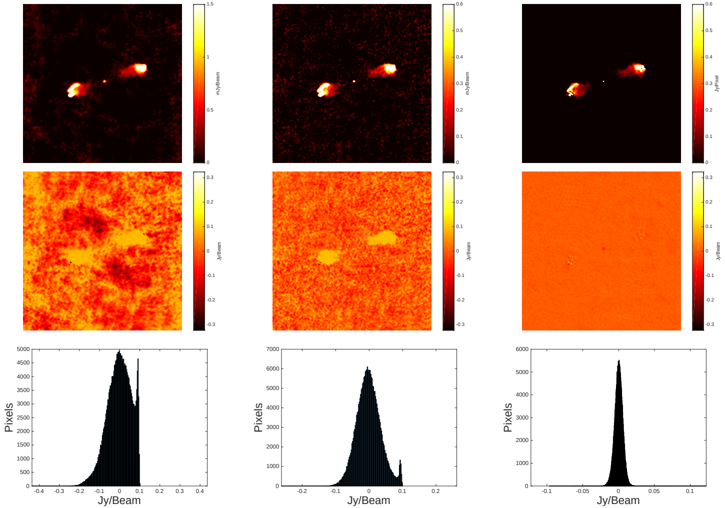


PURIFY reconstruction

VLA observation of Cygnus A images by PURIFY

PURIFY reconstruction

VLA observation of Cygnus A



(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: Cygnus A recovered images and residuals (Pratley, McEwen, *et al.* 2016)

PURIFY reconstruction

ATCA observation of PKS J0334-39

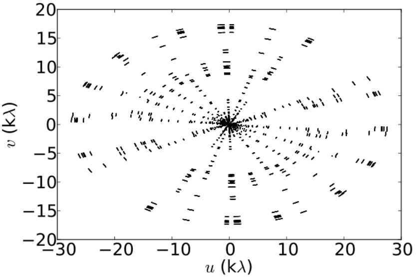
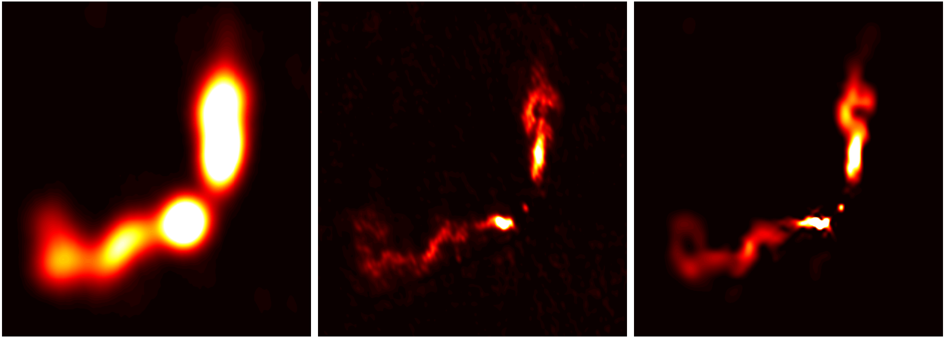


Figure: VLA visibility coverage for PKS J0334-39

PURIFY reconstruction

ATCA observation of PKS J0334-39

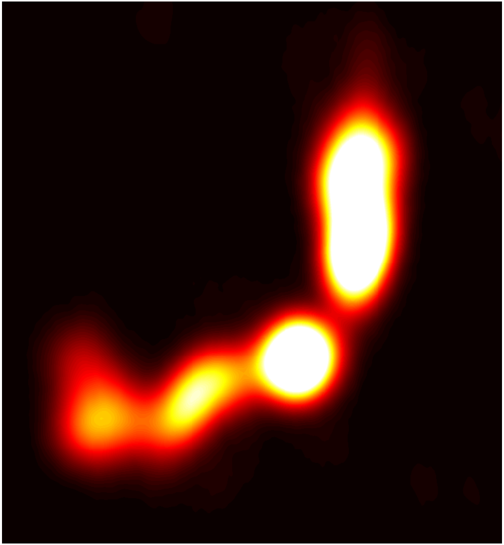


(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: PKS J0334-39 recovered images (Pratley, McEwen, *et al.* 2016)

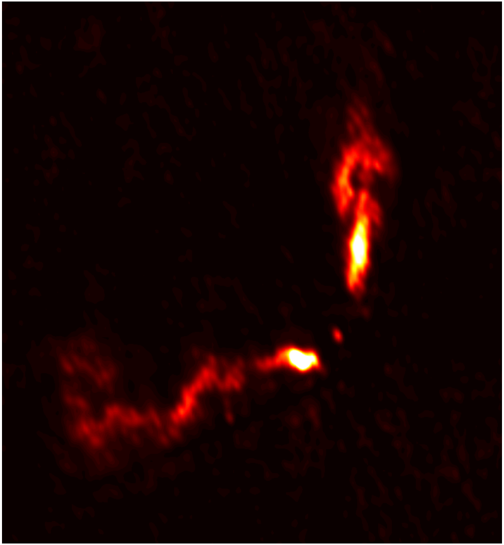
PURIFY reconstruction

VLA observation of PKS J0334-39 imaged by CLEAN (natural)



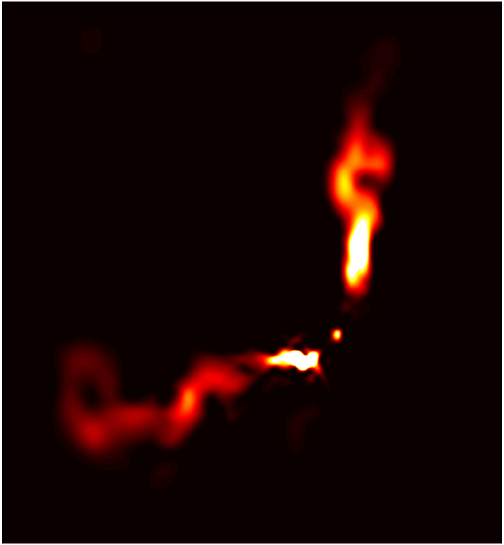
PURIFY reconstruction

VLA observation of PKS J0334-39 images by CLEAN (uniform)



PURIFY reconstruction

VLA observation of PKS J0334-39 images by PURIFY



PURIFY reconstruction

ATCA observation of PKS J0334-39

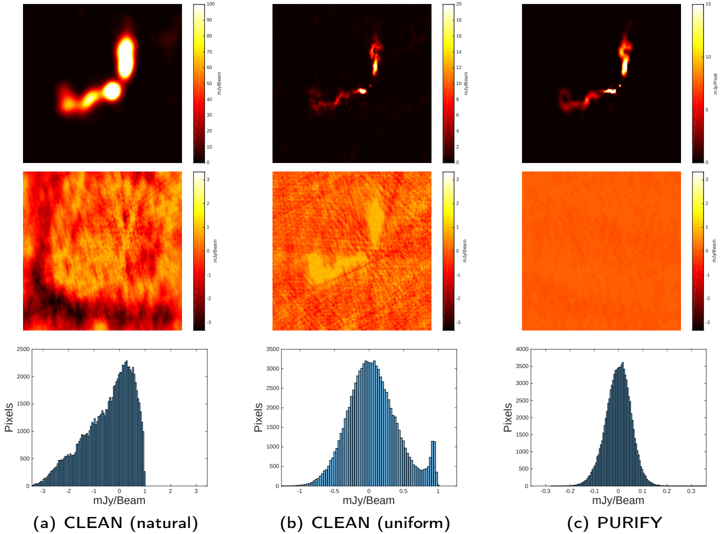


Figure: PKS J0334-39 recovered images and residuals (Pratley, McEwen, *et al.* 2016)

PURIFY reconstruction

ATCA observation of PKS J0116-473

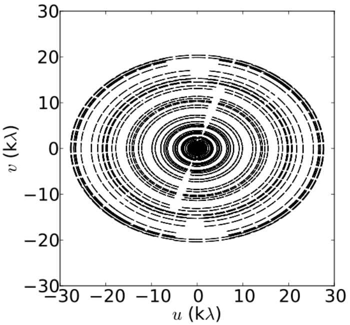
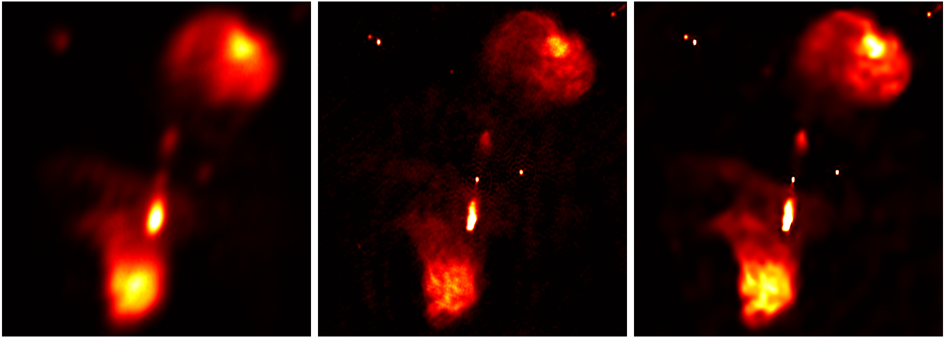


Figure: ATCA visibility coverage for Cygnus A

PURIFY reconstruction

ATCA observation of PKS J0116-473

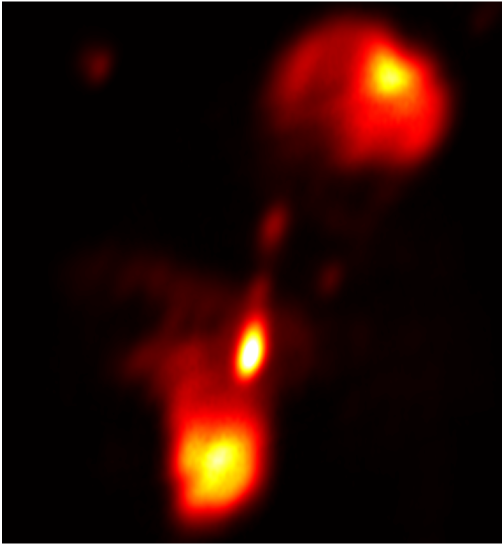


(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: PKS J0116-473 recovered images (Pratley, McEwen, *et al.* 2016)

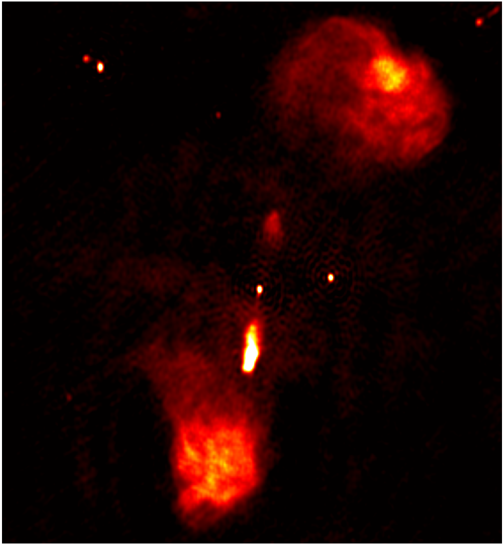
PURIFY reconstruction

VLA observation of PKS J0116-473 imaged by CLEAN (natural)



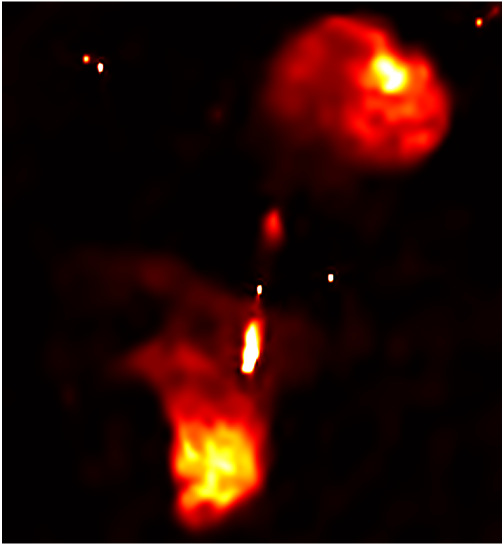
PURIFY reconstruction

VLA observation of PKS J0116-473 images by CLEAN (uniform)



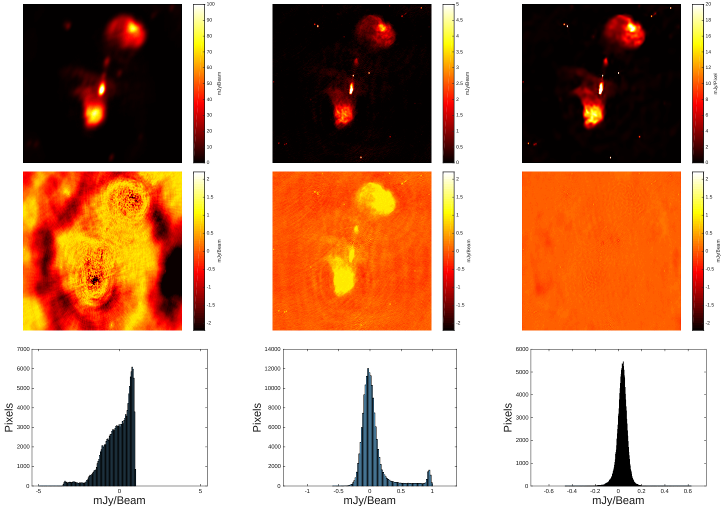
PURIFY reconstruction

VLA observation of PKS J0116-473 images by PURIFY



PURIFY reconstruction

ATCA observation of PKS J0116-473



(a) CLEAN (natural) (b) CLEAN (uniform) (c) PURIFY

Figure: PKS J0116-473 recovered images and residuals (Pratley, McEwen, *et al.* 2016)

PURIFY reconstructions

Table: Root-mean-square of residuals of each reconstruction (units in mJy/Beam)

Observation	PURIFY	CLEAN (natural)	CLEAN (uniform)
3C129	0.10	0.23	0.11
Cygnus A	6.1	59	36
PKS J0334-39	0.052	1.00	0.37
PKS J0116-473	0.054	0.88	0.24