Efficient Bayesian computation by Proximal Markov chain Monte Carlo: when Langevin meets Moreau.

Dr. Marcelo Pereyra
http://www.macs.hw.ac.uk/~mp71/

Maxwell Institute for Mathematical Sciences, Heriot-Watt University

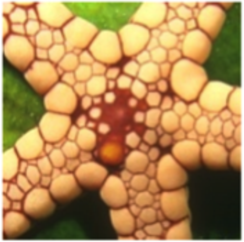
June 2020, Lisbon, Portugal (from Edinburgh).

HERIOT WATT UNIVERSITY

Outline

1. Bayesian inference in imaging inverse problems
2. Proximal Markov chain Monte Carlo
3. Uncertainty quantification in astronomical and medical imaging
4. Conclusion & Perspectives

Forward imaging problem

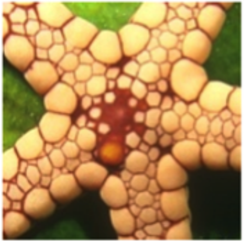


True scene

Imaging device

Observed image

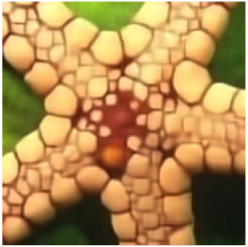
Inverse imaging problem



True scene

Imaging device

Observed image

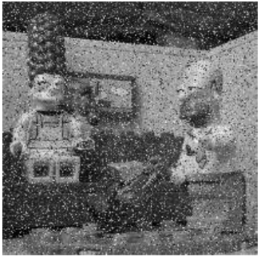


Restored image

Short-exposure imaging: forward problem

True scene

Imaging devise

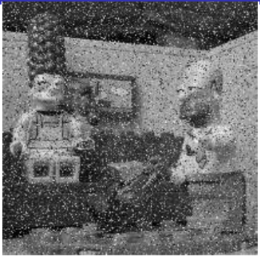


Observed image

Short-exposure imaging: inverse problem

True scene

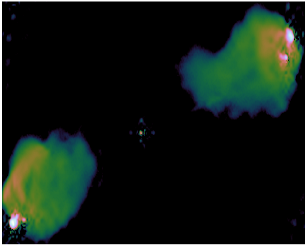
Imaging device



Observed image

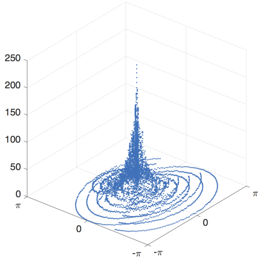
Restored image
(J. Delon and A. Desolneux (2013))

Radio-astronomy: forward problem



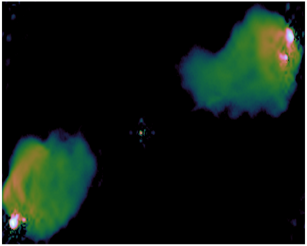
True scene

Imaging device



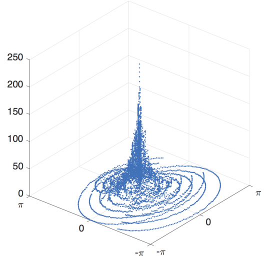
K-space data

Radio-astronomy: inverse problem

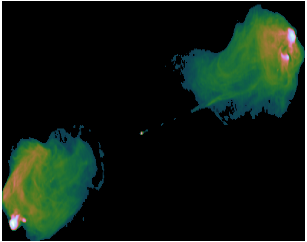


True scene

Imaging device



K-space data



Estimated image
(X. Cai et al. (2018))

Problem statement

- We are interested in an unknown image $x \in \mathbb{R}^d$.
- We measure y, related to x by some mathematical model.
- For example, in many imaging problems

$$y = Ax + w,$$

  for some operator A that is poorly conditioned or rank deficient, and an unknown perturbation or "noise" w.
- The recovery of x from y is often ill-posed or ill-conditioned, so we regularise the problem to make it well posed.

The Bayesian framework

Bayesian statistics is a mathematical framework for deriving inferences about x, from some observed data y <u>and</u> prior knowledge available.

Adopting a subjective probability approach, we represent x as a random quantity and use probability distributions to model expected properties.

To derive inferences about x from y we postulate a joint statistical model $p(x,y)$; typically specified via the decomposition $p(x,y) = p(y|x)p(x)$.

The Bayesian framework

The decomposition $p(x, y) = p(y|x)p(x)$ has two key ingredients:

The likelihood function: the conditional distribution $p(y|x)$ that models the data observation process (forward model).

The prior function: the marginal distribution $p(x) = \int p(x, y)\mathrm{d}y$ that models our knowledge about the solution x.

For example, for $y = Ax + w$, with $w \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$, we have

$$y \sim \mathcal{N}(Ax, \sigma^2 \mathbb{I}) \,,$$

or equivalently

$$p(y|x) \propto \exp\{-\|y - Ax\|_2^2 / 2\sigma^2\} \,.$$

Prior distribution

The prior distribution is usually of the form:

$$p(x) = \frac{1}{Z(\theta)} \mathrm{e}^{-\theta^\top \psi(x)} 1_\Omega(x) \,, \quad Z(\theta) = \int_\Omega \mathrm{e}^{-\theta^\top \psi(x)} \mathrm{d}x \,,$$

for some statistic $\psi : \mathbb{R}^d \to \mathbb{R}^m$, $\theta \in \mathbb{R}^m$, and constraint set $\Omega \subset \mathbb{R}^d$.

Often ψ and Ω are convex on $\mathbb{R}^d$ and $p(x)$ is log-concave.

Log-concave priors regularise the inverse problem by promoting solutions for which $\psi(x)$ is close to its expectation $\mathrm{E}(\psi|\theta)$, controlled by $\theta \in \mathbb{R}^p$.

Prior distribution

Formally, when ψ is convex we have concentration of probability mass on the typical set (see Bobkov and Madiman (2011))

$$\mathrm{P}\{\|\psi(x) - \mathrm{E}(\psi|\theta)\| > \eta|\theta\} < 3\exp\{-\eta^2 d/16\}, \quad \forall \eta \in (0,2) \tag{1}$$

Moreover, by differentiating $Z(\theta)$ and using Leibniz integral rule

$$\mathrm{E}(\psi|\theta) = \int_\Omega \psi(x)p(x)\mathrm{d}x = -\nabla_\theta \log Z(\theta), \tag{2}$$

hence $p(x|\theta)$ softly constrains $\psi(x) \approx -\nabla_\theta \log Z(\theta)$ when d is large.

$Z(\theta)$ is strongly log-concave, hence $\nabla_\theta \log Z(\theta)$ spans $\mathbb{R}^p$ (think duality).

Prior distribution

For example, priors of the form

$$p(x) \propto \mathrm{e}^{-\theta \|\Psi x\|_\dagger},$$

for some basis or dictionary $\Psi \in \mathbb{R}^{d \times p}$ and norm $\|\cdot\|_\dagger$, are encoding

$$\mathrm{E}(\|\Psi x\|_\dagger | \theta) = \frac{d}{\theta}.$$

See Pereyra et al. (2015); Fernandez-Vidal and Pereyra (2018) for more details and other examples.

Posterior distribution

We base our inferences on the posterior distribution $p(x|y)$.

We derive $p(x|y)$ from the likelihood $p(y|x)$ and the prior $p(x)$ by using

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

where $p(y) = \int p(y|x)p(x)\mathrm{d}x$ measures model-fit-to-data.

The conditional $p(x|y)$ models our knowledge about x after observing y.

In this talk we consider that $p(x|y)$ is log-concave; i.e.,

$$p(x|y) = \exp\{-\phi(x)\}/\int \exp\{-\phi(x)\}\mathrm{d}x\,,$$

where $\phi(x)$ is a convex function on $\mathbb{R}^d$.

Maximum-a-posteriori (MAP) estimation

The predominant Bayesian approach in imaging is MAP estimation

$$\begin{aligned} \hat{x}_{MAP} &= \underset{x \in \mathbb{R}^d}{\operatorname{argmax}}\, p(x|y), \\ &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, \phi(x). \end{aligned} \tag{3}$$

This Bayesian estimator is

1. efficiently computed by convex optimisation,
2. decision-theoretically optimal in the sense of the ϕ-Bregman error.

However, MAP estimation has some limitations, e.g.,

1. it provides little information about $p(x|y)$,
2. it struggles with unknown/partially unknown models.

See, e.g., Chambolle and Pock (2016); Pereyra (2016) for more details.

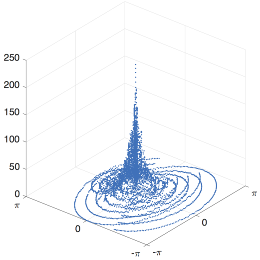
Illustrative example: astronomical image reconstruction

Recover $x \in \mathbb{R}^d$ from low-dimensional degraded observation

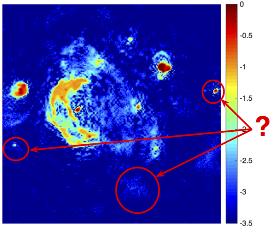
$$y = M\mathcal{F}x + w,$$

where $\mathcal{F}$ is the continuous Fourier transform, $M \in \mathbb{C}^{m\times d}$ is a measurement operator, Ψ is a wavelet basis, and $w \sim \mathcal{N}(0, \sigma^2\mathbb{I}_m)$. We use the model

$$p(x|y) \propto \exp\left(-\|y - M\mathcal{F}x\|^2/2\sigma^2 - \theta\|\Psi x\|_1\right)\mathbf{1}_{\mathbb{R}^n_+}(x). \tag{4}$$



y



$\hat{x}_{MAP}$

Figure: Radio-interferometric image reconstruction of the `W28 supernova`.

Outline

Inference by Markov chain Monte Carlo integration

Monte Carlo integration
Given a set of samples $X_1, \ldots, X_M$ distributed according to $p(x|y)$, we approximate posterior expectations and probabilities

$$\frac{1}{M}\sum_{m=1}^{M} h(X_m) \rightarrow \mathrm{E}\{h(x)|y\}, \quad \text{as } M \rightarrow \infty$$

Markov chain Monte Carlo:
Construct a Markov kernel $X_{m+1}|X_m \sim K(\cdot|X_m)$ such that the Markov chain $X_1, \ldots, X_M$ has $p(x|y)$ as stationary distribution.

MCMC simulation in high-dimensional spaces is very challenging.

Unadjusted Langevin algorithm

Suppose for now that $p(x|y) \in \mathcal{C}^1$. Then, we can generate samples by mimicking a Langevin diffusion process that converges to $p(x|y)$ as $t \to \infty$,

$$\boldsymbol{X}: \quad \mathrm{d}\boldsymbol{X}_t = \frac{1}{2}\nabla \log p\left(\boldsymbol{X}_t|y\right)\mathrm{d}t + \mathrm{d}W_t, \quad 0 \leq t \leq T, \quad \boldsymbol{X}(0) = x_0.$$

where W is the Brownian motion on $\mathbb{R}^d$.

Because solving $\boldsymbol{X}_t$ exactly is generally not possible, we use an Euler Maruyama approximation and obtain the "unadjusted Langevin algorithm"

$$\mathrm{ULA}: \; X_{m+1} = X_m + \delta\nabla \log p(X_m|y) + \sqrt{2\delta}Z_{m+1}, \;\; Z_{m+1} \sim \mathcal{N}(0, \mathbb{I}_n)$$

ULA is remarkably efficient when $p(x|y)$ is sufficiently regular.

Unfortunately, imaging models are often violate these regularity conditions.

Non-smooth models

Without loos of generality, suppose that

$$p(x|y) \propto \exp\{-f(x) - g(x)\} \tag{5}$$

where $f(x)$ and $g(x)$ are l.s.c. convex functions from $\mathbb{R}^d \to (-\infty, +\infty]$, f is L_f-Lipschitz differentiable, and $g \notin \mathcal{C}^1$.

For example,

$$f(x) = \frac{1}{2\sigma^2}\|y - Ax\|_2^2, \quad g(x) = \alpha\|Bx\|_\dagger + \mathbf{1}_{\mathcal{S}}(x)\,,$$

for some linear operators A, B, norm $\|\cdot\|_\dagger$, and convex set $\mathcal{S}$.

Unfortunately, such non-models are beyond the scope of ULA.

Idea: Regularise $p(x|y)$ to enable efficient Langevin sampling.

Approximation of $p(x|y)$

Moreau-Yoshida approximation of $p(x|y)$ (Pereyra, 2015):

Let $\lambda > 0$. We propose to approximate $p(x|y)$ with the density

$$p_\lambda(x|y) = \frac{\exp[-f(x) - g_\lambda(x)]}{\int_{\mathbb{R}^d} \exp[-f(x) - g_\lambda(x)]\mathrm{d}x},$$

where g_λ is the Moreau-Yoshida envelope of g given by

$$g_\lambda(x) = \inf_{u \in \mathbb{R}^d} \{g(u) + (2\lambda)^{-1} \|u - x\|_2^2\},$$

and where λ controls the approximation error involved.

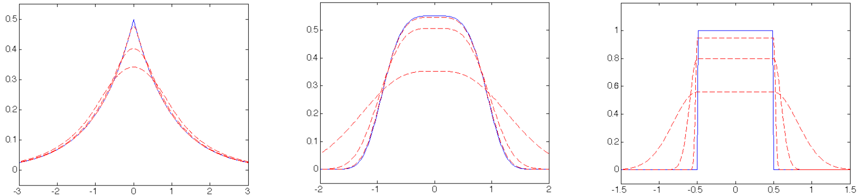
Moreau-Yoshida approximations

Key properties (Pereyra, 2015; Durmus et al., 2018):

1. $\forall \lambda > 0$, p_λ defines a proper density of a probability measure on $\mathbb{R}^d$.
2. *Convexity and differentiability*:
   - p_λ is log-concave on $\mathbb{R}^d$.
   - $p_\lambda \in \mathcal{C}^1$ even if p not differentiable, with
   $$\nabla \log p_\lambda(x|y) = -\nabla f(x) + \{\text{prox}_g^\lambda(x) - x\}/\lambda,$$
   and $\text{prox}_g^\lambda(x) = \text{argmin}_{u \in \mathbb{R}^\mathbb{N}}\; g(u) + \frac{1}{2\lambda}\|u - x\|^2$.
   - $\nabla \log p_\lambda$ is Lipchitz continuous with constant $L \leq L_f + \lambda^{-1}$.
3. *Approximation error between $p_\lambda(x|y)$ and $p(x|y)$*:
   - $\lim_{\lambda \to 0} \|p_\lambda - p\|_{TV} = 0$.
   - If g is L_g-Lipchitz, then $\|p_\lambda - p\|_{TV} \leq \lambda L_g^2$.

Illustration

Examples of Moreau-Yoshida approximations:



$p(x) \propto \exp\left(-|x|\right)$ $\quad$ $p(x) \propto \exp\left(-x^4\right)$ $\quad$ $p(x) \propto \mathbf{1}_{[-0.5,0.5]}(x)$

Figure: True densities (solid blue) and approximations (dashed red).

Proximal ULA

We approximate $\boldsymbol{X}$ with the "regularised" auxiliary Langevin diffusion

$$\boldsymbol{X}^{\lambda}: \quad \mathrm{d}\boldsymbol{X}_t^{\lambda} = \frac{1}{2}\nabla \log p_{\lambda}\left(\boldsymbol{X}_t^{\lambda}|y\right)\mathrm{d}t + \mathrm{d}W_t, \quad 0 \leq t \leq T, \quad \boldsymbol{X}^{\lambda}(0) = x_0,$$

which targets $p_{\lambda}(x|y)$. Remark: we can make $\boldsymbol{X}^{\lambda}$ arbitrarily close to $\boldsymbol{X}$.

Finally, an Euler Maruyama discretisation of $\boldsymbol{X}^{\lambda}$ leads to the (Moreau-Yoshida regularised) proximal ULA

$$\mathrm{MYULA}: \quad X_{m+1} = (1 - \tfrac{\delta}{\lambda})X_m - \delta\nabla f\{X_m\} + \tfrac{\delta}{\lambda}\,\mathrm{prox}_g^{\lambda}\{X_m\} + \sqrt{2\delta}Z_{m+1},$$

where we used that $\nabla g_{\lambda}(x) = \{x - \mathrm{prox}_g^{\lambda}(x)\}/\lambda$.

Convergence results

Non-asymptotic estimation error bound

Theorem 2.1 (Durmus et al. (2018))

Let $\delta_\lambda^{max} = (L_1 + 1/\lambda)^{-1}$. Assume that g is Lipchitz continuous. Then, there exist $\delta_\epsilon \in (0, \delta_\lambda^{max}]$ and $M_\epsilon \in \mathbb{N}$ such that $\forall \delta < \delta_\epsilon$ and $\forall M \geq M_\epsilon$

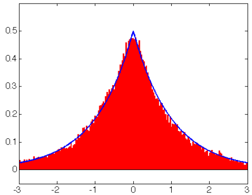
$$\|\delta_{x_0} Q_\delta^M - p\|_{TV} < \epsilon + \lambda L_g^2,$$

where Q_δ^M is the kernel assoc. with M iterations of MYULA with step δ.

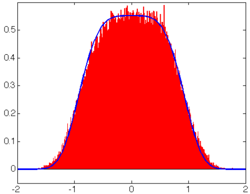
Note: δ_ϵ and M_ϵ are explicit and tractable. If $f + g$ is strongly convex outside some ball, then M_ϵ scales with order $\mathcal{O}(d \log(d))$. See Durmus et al. (2018) for other convergence results.

Illustration

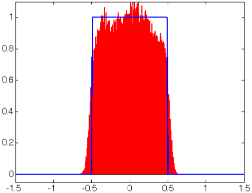
Illustrative examples:



$p(x) \propto \exp\left(-|x|\right)$



$p(x) \propto \exp\left(-x^4\right)$

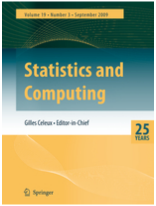


$p(x) \propto \mathbf{1}_{[-0.5,0.5]}(x)$

Figure: True densities (blue) and MC approximations (red histogram).

Modern Bayesian computation

Surveys on Bayesian computation...



25th anniversary special issue on Bayesian computation

P. Green, K. Latuszynski, M. Pereyra, C. P. Robert, "Bayesian computation: a perspective on the current state, and sampling backwards and forwards", Statistics and Computing, vol. 25, no. 4, pp 835-862, Jul. 2015.

Special issue on "Stochastic simulation and optimisation in signal processing"

M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin, "A Survey of Stochastic Simulation and Optimization Methods in Signal Processing" IEEE Sel. Topics in Signal Processing, vol. 10, no. 2, pp 224 - 241, Mar. 2016.

Outline

Uncertainty quantification in radio-interferometric imaging

Where does the posterior probability mass of x lie?

- A set C_α is a posterior credible region of confidence level $(1-\alpha)\%$ if

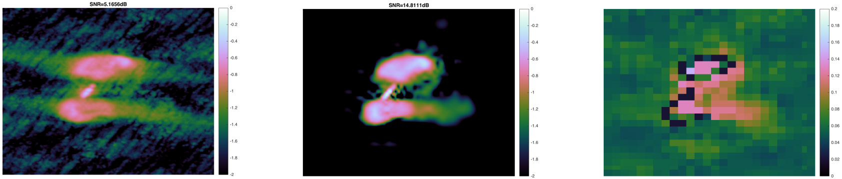
$$\mathrm{P}\left[x \in C_\alpha | y\right] = 1 - \alpha.$$

- The *highest posterior density* (HPD) region is decision-theoretically optimal (Robert, 2001)

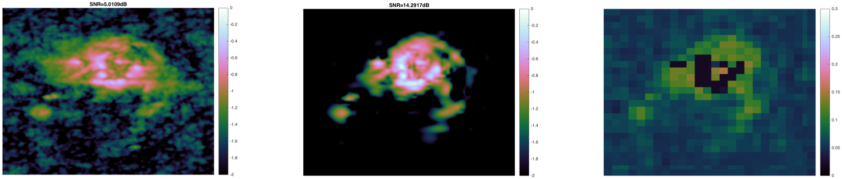
$$C^*_\alpha = \{x : \phi(x) \leq \gamma_\alpha\}$$

with $\gamma_\alpha \in \mathbb{R}$ chosen such that $\int_{C^*_\alpha} p(x|y)\mathrm{d}x = 1 - \alpha$ holds.

Astro-imaging experiment with redundant wavelet frame (Cai et al., 2017).



$\hat{x}_{penMLE}(y)$ $\quad$ $\hat{x}_{MAP}$ (by optimisation) $\quad$ credible intervals (scale 10×10)

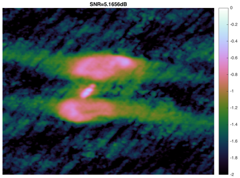


$\hat{x}_{penMLE}(y)$ $\quad$ $\hat{x}_{MAP}$ (by optimisation) $\quad$ credible intervals (scale 10×10)

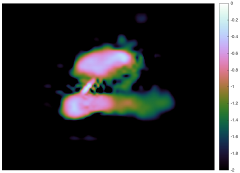
3C2888 and M31 radio galaxies (size 256×256 pixels).

Visualising uncertainty in radio-interferometric imaging

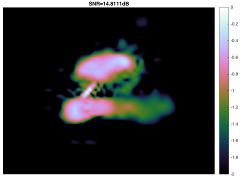
Astro-imaging experiment with redundant wavelet frame (Cai et al., 2017).



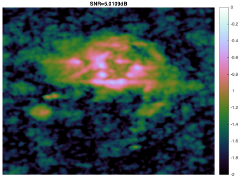
$\hat{x}_{penMLE}(y)$



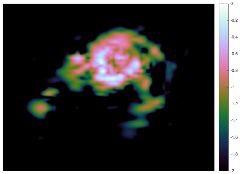
$\hat{x}_{MMSE} = \mathrm{E}(x|y)$



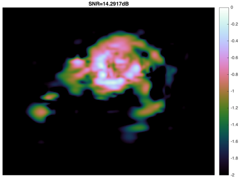
$\hat{x}_{MAP}$ (by optimisation)



$\hat{x}_{penMLE}(y)$



$\hat{x}_{MMSE} = \mathrm{E}(x|y)$



$\hat{x}_{MAP}$ (by optimisation)

3C2888 and M31 radio galaxies. Visual comparison MMSE and MAP estimation.

Hypothesis testing

Bayesian hypothesis test for specific image structures (e.g., lesions)

$$\mathrm{H}_0 : \text{The structure of interest is ABSENT in the true image}$$
$$\mathrm{H}_1 : \text{The structure of interest is PRESENT in the true image}$$

The null hypothesis H_0 is rejected with significance α if

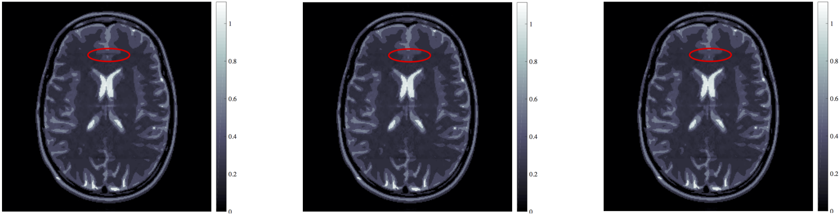
$$\mathsf{P}(\mathrm{H}_0|y) \leq \alpha.$$

Theorem (Repetti et al., 2018)
Let $\mathcal{S}$ denote the region of $\mathbb{R}^d$ associated with H_0, containing all images *without the structure* of interest. Then

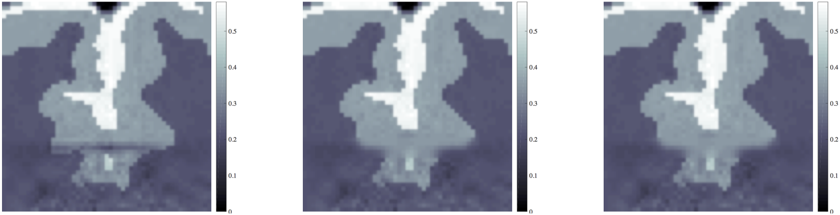
$$\mathcal{S} \cap \mathcal{C}_\alpha = \varnothing \implies \mathsf{P}(H_0|y) \leq \alpha\,.$$

If in addition $\mathcal{S}$ is convex, then checking $\mathcal{S} \cap \mathcal{C}_\alpha = \varnothing$ is a convex problem

$$\min_{\bar{x},\, \underline{x} \in \mathbb{R}^d} \|\bar{x} - \underline{x}\|_2^2 \quad \text{s.t.} \quad \bar{x} \in \mathcal{C}_\alpha\,, \quad \underline{x} \in \mathcal{S}\,.$$



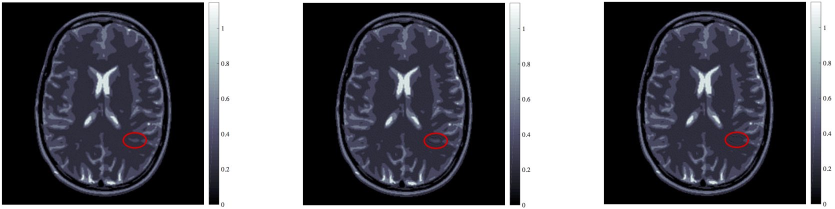
$\hat{x}_{MAP}$ $\quad$ $\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ $\quad$ $\underline{x} \in \mathcal{S}$



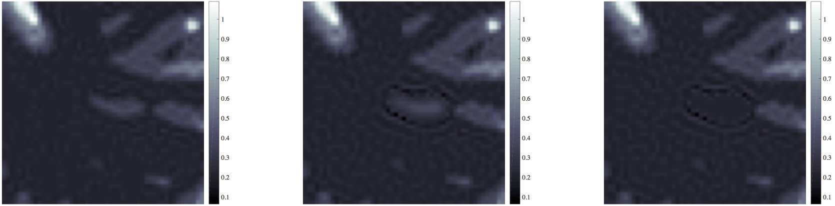
$\hat{x}_{MAP}$ (zoom) $\quad$ $\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ (zoom) $\quad$ $\underline{x} \in \mathcal{S}$ (zoom)

MRI experiment: test images $\bar{x} = \underline{x}$, hence we fail to reject H_0 and conclude that there is little evidence to support the observed structure.

Uncertainty quantification in MRI imaging



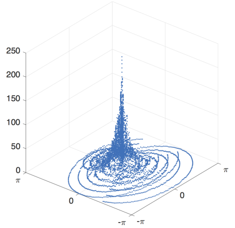
$\hat{x}_{MAP}$ $\quad$ $\bar{x} \in \mathcal{C}_{0.01}$ $\quad$ $\underline{x} \in \mathcal{S}_0$



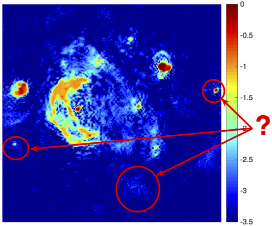
$\hat{x}_{MAP}$ (zoom) $\quad$ $\bar{x} \in \mathcal{C}_{0.01}$ (zoom) $\quad$ $\underline{x} \in \mathcal{S}_0$ (zoom)

MRI experiment: test images $\bar{x} \neq \underline{x}$, hence we reject H_0 and conclude that there is significant evidence in favour of the observed structure.

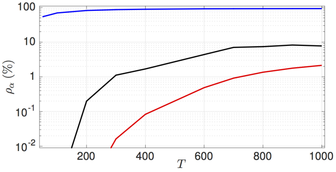
Quantification of minimum energy of different energy structures, at level $(1-\alpha) = 0.99$, as the number of measurements $T = \dim(y)/2$ increases.



$|y|$



$\hat{x}_{MAP}(T = 200)$



ρ_α, energy ratio preserved at $\alpha = 0.01$

Figure: Analysis of 3 structures in the `W28` supernova RI image.

Note: energy ratio calculated as

$$\rho_\alpha = \frac{\|\bar{x} - \underline{x}\|_2}{\|x_{MAP} - \tilde{x}_{MAP}\|_2}$$

where $\bar{x}, \underline{x}$ are computed with $\alpha = 0.01$, and $\tilde{x}_{MAP}$ is a modified version of x_{MAP} where the structure of interest has been carefully removed from the image.

Outline

Conclusion & Perspectives

- The challenges facing modern imaging sciences require a methodological paradigm shift to go beyond point estimation.
- The Bayesian framework can support this paradigm shift, but this requires significantly accelerating computation methods.
- We explored improving efficiency by integrating modern stochastic and variational approaches to construct proximal MCMC methods.
- MYULA has been superseded by more advanced proximal MCMC methods, e.g., the accelerated method of Pereyra et al. (2020).
- Future works should focus on improving frequentist coverage properties by using more accurate Bayesian priors; e.g., by integration with machine learning, plug-and-play, and scene-adapted approaches.

Bibliography:

Bobkov, S. and Madiman, M. (2011). Concentration of the information in data with log-concave distributions. *Ann. Probab.*, 39(4):1528–1543.

Cai, X., Pereyra, M., and McEwen, J. D. (2017). Uncertainty quantification for radio interferometric imaging II: MAP estimation. *ArXiv e-prints*.

Chambolle, A. and Pock, T. (2016). An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319.

Durmus, A., Moulines, E., and Pereyra, M. (2018). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci.*, 11(1):473–506.

Fernandez-Vidal, A. and Pereyra, M. (2018). Maximum likelihood estimation of regularisation parameters. In *Proc. IEEE ICIP 2018*.

Pereyra, M. (2015). Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*. open access paper, http://dx.doi.org/10.1007/s11222-015-9567-4.

Pereyra, M. (2016). Revisiting maximum-a-posteriori estimation in log-concave models: from differential geometry to decision theory. *ArXiv e-prints*.

Pereyra, M., Bioucas-Dias, J., and Figueiredo, M. (2015). Maximum-a-posteriori estimation with unknown regularisation parameters. In *Proc. Europ. Signal Process. Conf. (EUSIPCO) 2015*.

Pereyra, M., Mieles, L. V., and Zygalakis, K. C. (2020). Accelerating proximal markov chain monte carlo by using an explicit stabilized method. *SIAM J Imaging Sci*, 13(2):905–935.

Repetti, A., Pereyra, M., and Wiaux, Y. (2018). Scalable Bayesian uncertainty quantification in imaging inverse problems via convex optimisation. *ArXiv e-prints*.

Robert, C. P. (2001). *The Bayesian Choice (second edition)*. Springer Verlag, New-York.

Thank you!