Bayesian inference in imaging inverse problems - Part 1

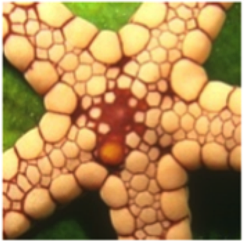
Marcelo Pereyra

Heriot-Watt University & Maxwell Institute for Mathematical Sciences

June 2025, Bologna, Italy.

Outline

Forward imaging problem

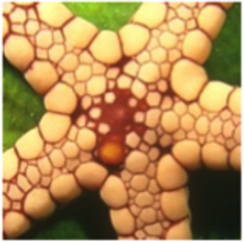


True scene

Imaging device

Observed image

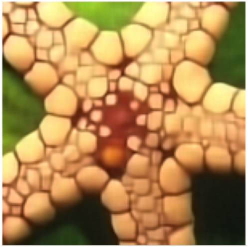
Inverse imaging problem



True scene

Imaging device

Observed image



Restored image

Imaging inverse problems

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

$$y = Ax^\star + w,$$

for some operator A that is poorly conditioned or rank deficient, and an unknown perturbation or "noise" w.

- The recovery of $x^\star$ from y is usually not well posed. Additional information is required in order to deliver meaningful solutions.

Mathematical imaging frameworks

- There are three main mathematical and computational frameworks for inference in imaging inverse problems:
  1. Mathematical analysis
  2. Bayesian statistics.
  3. Machine learning.
- These frameworks have complementary strengths and weaknesses.
- Our aim is to develop a unifying framework of theory, methods, and algorithms that inherits the benefits of each approach.

The Bayesian statistical approach

- Model $x^\star$ as a realisation of a r.v. $\mathbb{x}$ on $\mathbb{R}^d$. Use the distribution of $\mathbb{x}$ to regularise the problem and promote expected properties.
- The observation y is a realisation of a r.v. $(\mathbb{y}|\mathbb{x} = x^\star)$.
- Inferences about $x^\star$ from y are derived from the joint distribution of $(\mathbb{x}, \mathbb{y})$ - 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 $\mathbb{x}$.

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

For this tutorial, we assume a prior distribution of the form:

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

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.

The normalising constant $Z(\theta)$ is given by

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

so $\int_\Omega p(x)\mathrm{d}x = 1$. This will play a key role in model selection techniques.

The statistic ψ can be assumption-driven (e.g., a sparsity promoting norm), purely data-driven, or a combination of both.

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 $p(x|y)$ <u>models</u> our beliefs about $\mathbb{x}$ after observing $\mathbb{y} = y$.

In this first tutorial, 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 $\mathbb{R}^d$.

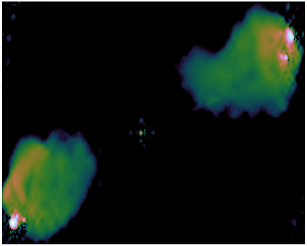
Illustrative example: astronomical image reconstruction

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

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

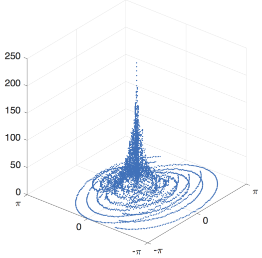
where $\mathcal{F}$ is the continuous Fourier transform, $M \in \mathbb{C}^{m \times d}$ is a measurement operator, Ψ is a wavelet basis, and $w \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_m)$. We use the model

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



$x^\star$

$y \sim \mathcal{N}(M\mathcal{F}x^\star, \sigma^2\mathbb{I})$



y

Figure: Radio-interferometric measurements of the `W28 supernova`.

Maximum-a-posteriori (MAP) estimation

The predominant Bayesian approach in imaging is MAP estimation

$$\begin{aligned}\hat{x}_{MAP} &= \underset{x \in \mathbb{R}^d}{\operatorname{argmax}}\, p(x|y), \\ &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, \phi(x).\end{aligned} \tag{2}$$

When $p(x|y)$ is log-concave, $\hat{x}_{MAP}$ is a convex optimisation problem. We are usually able to solve convex problems very efficiently (see Chambolle and Pock (2016))..

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

MAP estimation by proximal optimisation

To compute $\hat{x}_{MAP}$ we could use a proximal splitting algorithm. Let

$$f(x) = \|y - M\mathcal{F}x\|^2/2\sigma^2\,, \quad \text{and} \quad g(x) = \theta\|\Psi x\|_1 + -\log \mathbf{1}_{\mathbb{R}^n_+}(x)\,,$$

where f and g are l.s.c. convex on $\mathbb{R}^d$, and f is L_f-Lipschitz differentiable.

For example, we could use a **proximal gradient** iteration

$$x^{m+1} = \text{prox}_g^{L_f^{-1}}\{x^m + L_f^{-1}\nabla f(x^m)\},$$

converges to $\hat{x}_{MAP}$ at rate $O(1/m)$, with poss. acceleration to $O(1/m^2)$.

Definition For $\lambda > 0$, the λ-proximal operator of a convex l.s.c. function g is defined as (Moreau, 1962)

$$\text{prox}_g^{\lambda}(x) \triangleq \underset{u\in\mathbb{R}^{\mathbb{N}}}{\text{argmin}}\; g(u) + \frac{1}{2\lambda}\|u - x\|^2.$$

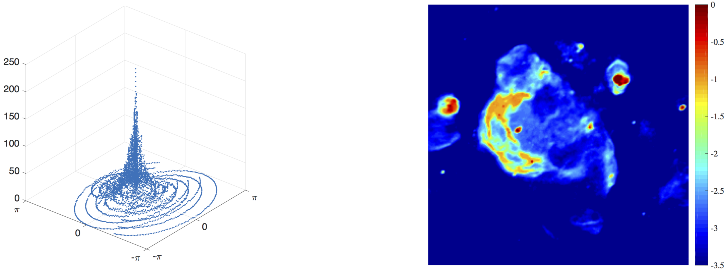
Illustrative example: astronomical image reconstruction

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

$$p(x|y) \propto \exp\left(-\|y - M\mathcal{F}x\|^2/2\sigma^2 - \theta\|\Psi x\|_1\right)\mathbf{1}_{\mathbb{R}^n_+}(x). \tag{3}$$



y $\hat{x}_{MAP}$

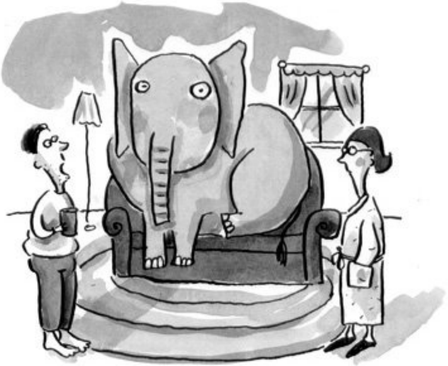
Figure: Radio-interferometric image reconstruction of the `W28 supernova`.

Summary

- Modern convex optimisation can compute $\hat{x}$ very efficiently...
- With parallelised and distributed algorithms...
- With theoretical convergence guarantees..
- And GPU implementations...
- With data-driven flavours based on input convex neural networks..
- Also non-convex extensions...

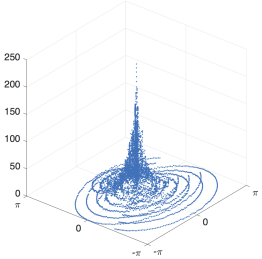
So the problem is quite solved, right?

Not really...

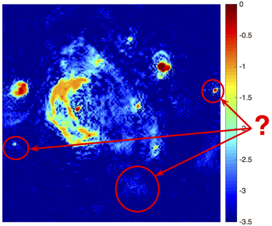


What elephant?

Elephant 1: what is the uncertainty about $\hat{x}$?



$|y|$



$\hat{x}_{MAP}$

How confident are we about all these structures in the image?

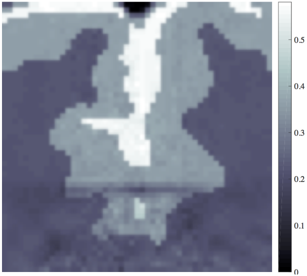
What is the error in their intensity, position, spectral properties?

Using $\hat{x}_{MAP}$ to derive physical quantities? what error bars should we put...

Illustrative example: magnetic resonance imaging

We use very similar techniques to produce magnetic resonance images...

$\hat{x}$

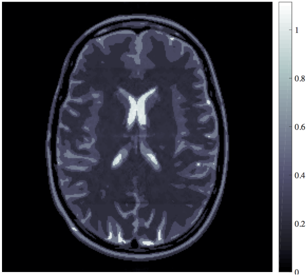


$\hat{x}$ (zoom)

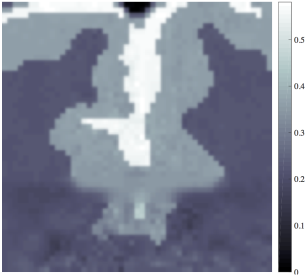
Figure: Magnetic resonance imaging of brain lesion.

How can we quantify our uncertainty about the brain lesion in the image?

What about this other solution to the problem, with no lesion?



$\hat{x}'$

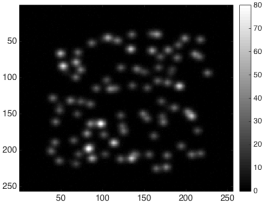


$\hat{x}'$ (zoom)

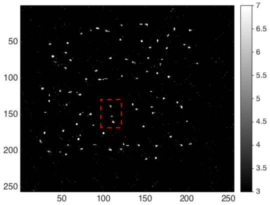
Figure: Magnetic resonance imaging of brain lesion.

Do we have any arguments to reject this solution?

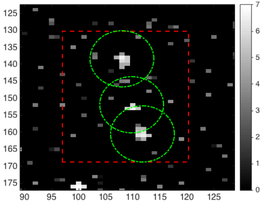
Another example related to sparse super-resolution in live-cell microscopy



y



$\hat{x}_{MAP}$



$\hat{x}_{MAP}$ (zoom)

Figure: Live-cell microscopy data (Zhu et al., 2012).

The image is sharpened to enhance molecule position measurements, but what is the precision of the procedure?

Elephant 2: multiple competing models

Two imaging scientists often formulate different models/cost functions to recover x

$$
\begin{aligned}
\hat{x}_1 &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \|y - A_1 x\|_2^2 + \theta_1 h_1(x), \\
\hat{x}_2 &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \|y - A_2 x\|_2^2 + \theta_2 h_2(x),
\end{aligned}
\tag{4}
$$

How can we compare them without ground truth available?

Can we use several models simultaneously?

Elephant 3: partially unknown models

Some of the model parameters might also be unknown; e.g., $\theta \in \mathbb{R}^+$ in

$$\hat{x}_1 = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \|y - A_1 x\|_2^2 + \theta h_1(x). \tag{5}$$

Then θ parametrises a class of models for $y \to x$.

How can we select θ without using ground truth?

Could we use all models simultaneously?

Outline

Bayesian decision theory

Given the following elements defining a decision problem:

1. Decision space Δ
2. Loss function $L(\delta, x) : \Delta \times \mathbb{R}^d \to \mathbb{R}$ quantifying the loss (or profit) related to taking action $\delta \in \Delta$ when the truth is $x \in \mathbb{R}^d$.
3. A model $p(x)$ representing probabilities for x.

What is the optimal decision $\delta^* \in \Delta$ when x is unknown?

Bayesian decision theory

Given the following elements defining a decision problem:

1. Decision space Δ
2. Loss function $L(\delta, x) : \Delta \times \mathbb{R}^d \to \mathbb{R}$ quantifying the loss (or profit) related to taking action $\delta \in \Delta$ when the truth is $x \in \mathbb{R}^d$.
3. A probability model $p(x)$ representing knowledge about x.

According to Bayesian decision theory (Robert, 2001), the optimal decision under uncertainty is

$$\delta^* = \operatorname*{argmin}_{\delta \in \Delta} \mathrm{E}\{L(\delta, \mathrm{x}) | y\} = \operatorname*{argmin}_{\delta \in \Delta} \int L(\delta, x) p(x) \mathrm{d}x.$$

Bayesian point estimators

Bayesian point estimators arise from the decision "what point $\hat{x} \in \mathbb{R}^d$ summarises $x|y$ best?". The optimal decision under uncertainty is

$$\hat{x}_L = \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathrm{E}\{L(u, \mathrm{x})|y\} = \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \int L(u, x) p(x|y) \mathrm{d}x$$

where the loss $L(u, x)$ measures the "dissimilarity" between u and x.

General desiderata:

1. $L(u, x) \geq 0, \; \forall \, u, x \in \mathbb{R}^d$,
2. $L(u, x) = 0 \iff u = x$,
3. L strictly convex w.r.t. its first argument (for estimator uniqueness).

Bayesian point estimators - MMSE estimation

Example: the squared Euclidean distance $L(u,x) = \|u - x\|^2$ defines the so-called minimum mean squared error estimator.

$$\hat{x}_{MMSE} = \operatorname*{argmin}_{u \in \mathbb{R}^d} \int \|u - x\|_2^2 \, p(x|y) \mathrm{d}x \,.$$

By differentiating w.r.t. to u and equating to zero we obtain that

$$\int (\hat{x}_{MMSE} - x) p(x|y) \mathrm{d}x = 0 \implies \hat{x}_{MMSE} \int p(x|y) \mathrm{d}x = \int x p(x|y) \mathrm{d}x \,,$$

hence $\hat{x}_{MMSE} = \mathrm{E}\{\mathbb{x}|y\}$ (recall that $\int p(x|y) \mathrm{d}x = 1$).

What about the MAP estimator?

Assume that $p(x|y) \propto \exp\{-\phi_y(x)\}$ is log-concave, i.e., ϕ_y convex on $\mathbb{R}^d$.

The Bayesian estimator that minimises the Bregman divergence loss $L(u,x) = D_\phi(u,x) \triangleq \phi(u) - \phi(x) - \nabla\phi(x)(u-x)$, is the MAP estimator

$$
\begin{aligned}
\hat{x}_L = \underset{u\in\mathbb{R}^d}{\operatorname{argmin}} \mathrm{E}\{L(u,\mathrm{x})|y\} &= \underset{u\in\mathbb{R}^d}{\operatorname{argmin}} \int D_\phi(u,x)p(x|y)\mathrm{d}x \\
&= \underset{x\in\mathbb{R}^d}{\operatorname{argmax}}\, p(x|y), \\
&= \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} -\phi_y(x), \\
&= \hat{x}_{MAP}
\end{aligned}
\tag{6}
$$

Interestingly, the "dual" estimator is $\hat{x}_{MMSE}$, i.e.,

$$
\begin{aligned}
\hat{x}_L = \underset{u\in\mathbb{R}^d}{\operatorname{argmin}} \mathrm{E}\{L(x,\mathrm{x})|y\} &= \underset{u\in\mathbb{R}^d}{\operatorname{argmin}} \int D_\phi(x,u))p(x|y)\mathrm{d}x \\
&= \hat{x}_{MMSE}
\end{aligned}
\tag{7}
$$

See Pereyra (2019) for proof and details.

Posterior credible regions

Where does the posterior probability mass of x lie?

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

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

For any $\alpha \in (0,1)$ there are infinitely many regions of the parameter space that verify this property.

The *highest posterior density* (HPD) region is decision-theoretically optimal in a compactness sense

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

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

Hypothesis testing

Hypothesis test split the solution space in two meaningful regions, e.g.,

$$\mathrm{H}_0 : x \in \mathcal{S}$$
$$\mathrm{H}_1 : x \notin \mathcal{S}$$

where $\mathcal{S} \subset \mathbb{R}^d$ contains all solutions with some characteristic of interest.

We can then assess the degree of support for H_0 vs. H_1 by computing

$$P(H_0|y) = \int_{\mathcal{S}} p(x|y)\mathrm{d}x, \quad P(H_1|y) = 1 - P(H_0|y)\,.$$

We can also reject H_0 in favour of H_1 with significance $\alpha \in [0, 1]$ if

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

Bayesian model selection

The Bayesian framework provides theory for comparing models objectively.

Given K alternative models $\{\mathcal{M}_j\}_{j=1}^K$ with posterior densities

$$\mathcal{M}_j : \quad p_j(x|y) = p_j(y|x)p_j(x))/p_j(y)\,,$$

we compute the (marginal) posterior probability of each model, i.e.,

$$p(\mathcal{M}_j|y) \propto p(y|\mathcal{M}_j)p(\mathcal{M}_j) \tag{8}$$

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

We then select for our inferences the "best" model, i.e.,

$$\mathcal{M}^* = \underset{j\in\{1,\ldots,K\}}{\operatorname{argmax}}\; p(\mathcal{M}_j|y).$$

Bayesian model calibration

Alternatively, given a continuous class of models $\{\mathcal{M}_\theta, \theta \in \theta\}$ with

$$\mathcal{M}_{\boldsymbol{\theta}} : \quad p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)} \,,$$

we compute the (marginal) posterior

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \tag{9}$$

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

We then calibrate our model with the "best" value of θ, i.e.,

$$\hat{\theta}_{MAP} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\theta|y).$$

Bayesian model averaging

We can also use all models simultaneously!

Given K alternative models $\{\mathcal{M}_j\}_{j=1}^K$ with posterior densities

$$\mathcal{M}_j : \quad p_j(x|y) = p_j(y|x)p_j(x))/p_j(y)\,,$$

we marginalise w.r.t. the model selector j, i.e.,

$$p(x|y) = \sum_{j=1}^{K} p(x, \mathcal{M}_j|y) = \sum_{j=1}^{K} p(x|y, \mathcal{M}_j)p(\mathcal{M}_j|y) \tag{10}$$

where the posterior probabilities $p(\mathcal{M}_j|y)$ control the relative importance of each model.

Bayesian model calibration

Similarly, given a continuous class of models $\{\mathcal{M}_\theta, \theta \in \Theta\}$ with

$$\mathcal{M}_{\boldsymbol{\theta}} : \quad p(x|y,\theta) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)}\,,$$

and a prior $p(\theta)$, we marginalise θ

$$\begin{aligned} p(x|y) &= \int_\Theta p(x,\theta|y)\mathrm{d}\theta, \\ &= \int_\Theta p(x|y,\theta)p(\theta|y)\mathrm{d}\theta \end{aligned}$$

where again $p(\theta|y)$ controls the relative contribution of each model.

My interpretation of log-concave prior distributions

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 $\theta \in \mathbb{R}^p$.

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

$$\mathrm{P}\{\|\psi(\mathbb{x}) - \mathrm{E}(\psi|\theta)\| > \eta|\theta\} < 3\exp\{-\eta^2 d/16\}, \quad \forall \eta \in (0,2) \tag{11}$$

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

$$\mathrm{E}(\psi(\mathbb{x})|\theta) = \int_\Omega \psi(x)p(x)\mathrm{d}x = -\nabla_\theta \log Z(\theta), \tag{12}$$

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).

My interpretation of log-concave prior distributions

For example, priors of the form

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

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

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

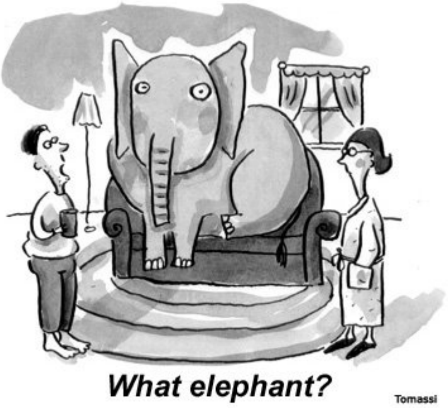
See Pereyra et al. (2015); Fernandez-Vidal and Pereyra (2018) for more details and other examples.

Summary

- The Bayesian statistical paradigm provides a power mathematical framework to solve imaging problems...
- It allows deriving optimal estimators for x..
- As well as quantifying the uncertainty in the solutions delivered...
- It supports hypothesis tests to inform decisions and conclusions...
- And allows operating with partially unknown models...
- And with several competing models...

So the problem is quite solved, right?

Not really...



How do we compute all these probabilities?

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

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

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 $t \to \infty$,

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

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 often violate these regularity conditions.

Non-smooth models

Without loss of generality, suppose that

$$p(x|y) \propto \exp\{-f(x) - g(x)\} \tag{13}$$

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

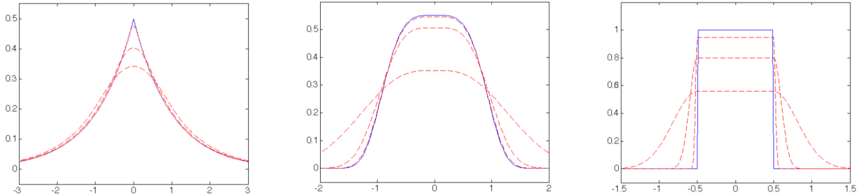
Key properties (Pereyra, 2015; Durmus et al., 2018):

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 $\text{prox}_g^\lambda(x) = \text{argmin}_{u \in \mathbb{R}^N} \, g(u) + \frac{1}{2\lambda}\|u - x\|^2$.
   - $\nabla \log p_\lambda$ 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 $\|p_\lambda - p\|_{TV} \leq \lambda L_g^2$.

Examples of Moreau-Yoshida approximations:



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

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

Proximal ULA

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

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

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

Non-asymptotic estimation error bound

Theorem 3.1 (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 $\forall M \geq M_\epsilon$

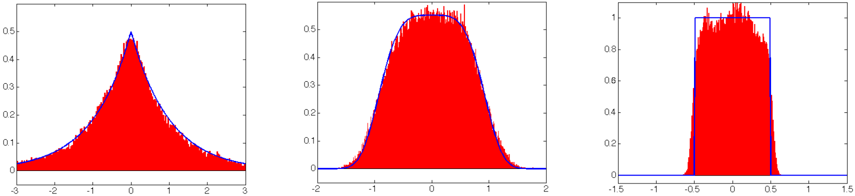
$$\|\delta_{x_0} Q_\delta^M - p\|_{TV} < \epsilon + \lambda L_g^2,$$

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

Note 1: 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.

Illustrative example 1

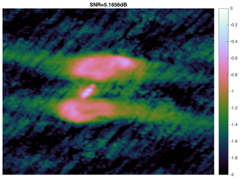
Three toy models:



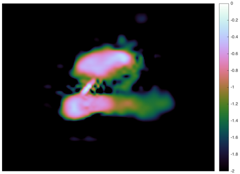
$p(x) \propto \exp\left(-|x|\right)$ $\quad$ $p(x) \propto \exp\left(-x^4\right)$ $\quad$ $p(x) \propto \mathbf{1}_{[-0.5,0.5]}(x)$

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

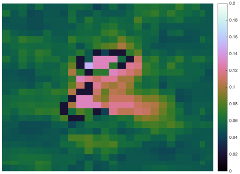
Astro-imaging experiment with redundant wavelet frame (Cai et al., 2017).



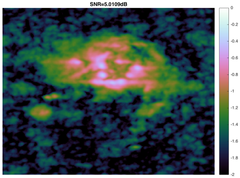
$\hat{x}_{penMLE}(y)$



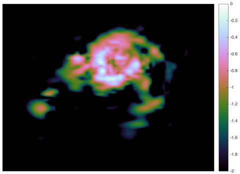
$\hat{x}_{MMSE} = \mathrm{E}(x|y)$



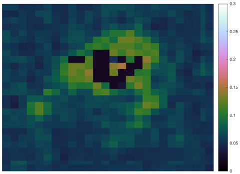
credible intervals (scale 10×10)



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



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



credible intervals (scale 10×10)

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

Bayesian Uncertainty quantification

Where does the posterior probability mass of x lie?

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

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

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

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

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

- Given a set of samples $X_1, \ldots, X_M$ distributed according to $p(x|y)$, we can estimate γ_α from (sample) quantiles of $\phi(X_1), \ldots, \phi(X_M)$.
- Alternatively, we can compute a bound $\gamma_\alpha \le \tilde{\gamma}_\alpha(\hat{x}_{MAP})$ analytically by using probability concentration results. See Pereyra (2016).

Bayesian Uncertainty quantification

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

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

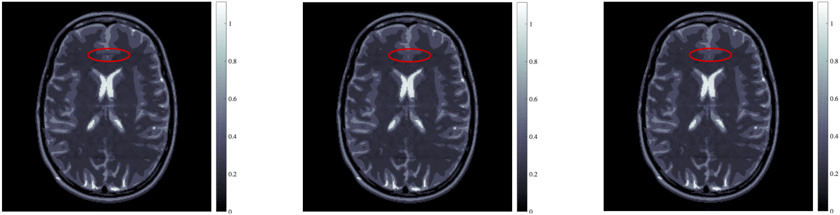
Theorem (Repetti et al., 2018)
Let $\mathcal{S}$ denote the region of $\mathbb{R}^d$ associated with H_0, containing all images *without the structure* of interest. Then

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

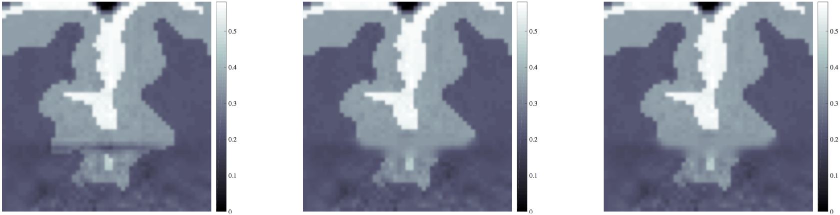
If in addition $\mathcal{S}$ is convex, then checking $\mathcal{S} \cap \mathcal{C}_\alpha = \varnothing$ is a convex problem

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

Uncertainty quantification in MRI imaging

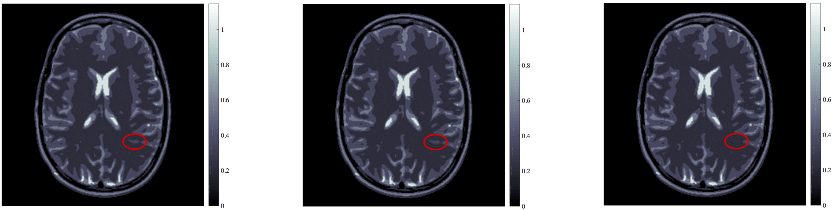


$\hat{x}_{MAP}$ $\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ $\underline{x} \in \mathcal{S}$

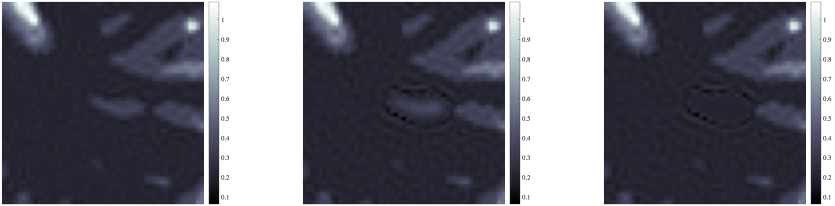


$\hat{x}_{MAP}$ (zoom) $\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ (zoom) $\underline{x} \in \mathcal{S}$ (zoom)

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.



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

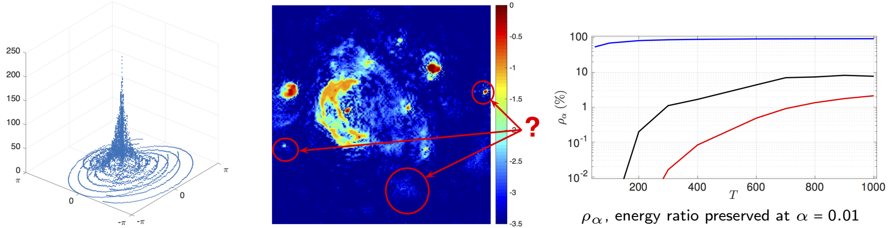


$\hat{x}_{MAP}$ (zoom) $\bar{x} \in \mathcal{C}_{0.01}$ (zoom) $\underline{x} \in \mathcal{S}_0$ (zoom)

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.

Uncertainty quantification in radio-interferometric imaging

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.

Bayesian Model Selection

The Bayesian framework provides theory for comparing models objectively.

Given K alternative models $\{\mathcal{M}_j\}_{j=1}^K$ with posterior densities

$$\mathcal{M}_j : \quad p_j(x|y) = p_j(y|x)p_j(x))/p_j(y)\,,$$

we compute the (marginal) posterior probability of each model, i.e.,

$$p(\mathcal{M}_j|y) \propto p(y|\mathcal{M}_j)p(\mathcal{M}_j) \tag{14}$$

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

We then select for our inferences the "best" model, i.e.,

$$\mathcal{M}^* = \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\, p(\mathcal{M}_j|y).$$

Experiment setup

We degrade the Boat image of size 256×256 pixels with a 5×5 uniform blur operator A^* and Gaussian noise $w \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_N)$ with $\sigma = 0.5$.

$$y = A^* x + w$$

We consider four alternative models to estimate x, given by

$$\mathcal{M}_j : \quad p_j(x|y) \propto \exp\left[-(\|y - A_j x\|^2 / 2\sigma^2) - \beta_j \phi_j(x)\right] \tag{15}$$

with fixed hyper-parameters σ and β, and where:

- $\mathcal{M}_1$: A_1 is the correct blur operator and $\phi_j(x) = TV(x)$.
- $\mathcal{M}_2$: A_2 is a mildly misspecified blur operator and $\phi_j(x) = TV(x)$.
- $\mathcal{M}_3$: A_3 is the correct blur operator and $\phi_j(x) = \|\Psi x\|_1$.
- $\mathcal{M}_4$: A_4 is a mildly misspecified blur operator and $\phi_j(x) = \|\Psi x\|_1$.

where Ψ is a wavelet frame and $TV(x) = \|\nabla_d x\|_{1-2}$ is the total-variation pseudo-norm. The β_j are adjusted automatically (see model calibration).

Monte Carlo strategy

To perform model selection we use MYULA to approximate the posterior probabilities $p(\mathcal{M}_j|y)$ for $j = 1, 2, 3, 4$ by Monte Carlo integration.

For each model we generate $n = 10^5$ samples $\{X_k^j\}_{k=1}^n \sim p(x|y, \mathcal{M}_j)$ and use the truncated harmonic mean estimator

$$p(y|\mathcal{M}_j) \approx \left(\sum_{k=1}^{n} \frac{\mathbf{1}_{\mathcal{S}^\star}(X_k^M)}{p(X_k^M, y|\mathcal{M}_j)} \right)^{-1} \mathrm{vol}(\mathcal{S}^\star) , \quad j = \{1, 2, 3, 4\} \tag{16}$$

where $\mathcal{S}^\star$ is a union of HPDs of $p(x|y, \mathcal{M}_j)$, also estimated from $\{X_k^j\}_{k=1}^n$. Computing time approx. 30 minutes per model.

MYULA can also be used within a nested sampling scheme that provides a faster and more reliable estimation of $p(y|\mathcal{M}_j)$. See Cai et al. (2022).

Numerical results

We obtain that $p(\mathcal{M}_1|y) \approx 0.68$ and $p(\mathcal{M}_3|y) \approx 0.27$ with the correct blur are the best models, $p(\mathcal{M}_2|y) < 0.05$ and $p(\mathcal{M}_4|y) < 0.01$ perform poorly.

$\mathcal{M}_1$ $\mathcal{M}_3$

y

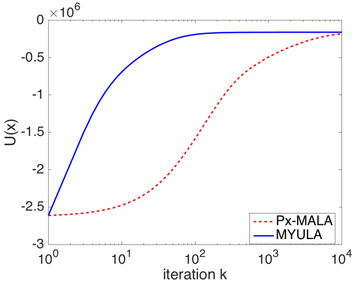
$\hat{x}_{MAP}$ (PSNR 34.1dB) $p(\mathcal{M}_1|y) \approx 0.68$

$\hat{x}_{MAP}$ (PSNR 32.9dB) $p(\mathcal{M}_3|y) \approx 0.27$

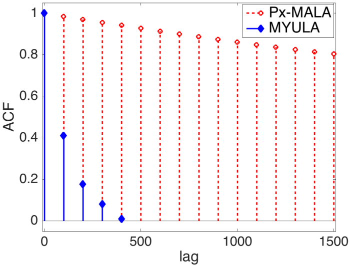
Figure: MAP estimation results for the `Boat` image deblurring experiment. (Note: error w.r.t. "exact" probabilities from Px-MALA approx. 0.5%.)

Numerical results

MYULA and Px-MALA efficiency comparison:



(a)



(b)

Figure: (a) Convergence of the chains to the typical set of $x|y$ under model $\mathcal{M}_1$, (b) chain autocorrelation function (ACF).)

Warning: I rarely use MYULA (or any other ULA)!

ULA samplers are based on a simple Euler-Maruyama approximation of the Langevin diffusion that behaves similarly to gradient descent optimisation.

These samplers are useful for introducing ideas, however, they have been superseded by accelerated samplers, see, e.g., Pereyra et al. (2020).

Outline

Problem statement

Consider the class of Bayesian models

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

parametrised by a regularisation parameter $\theta \in \Theta$. For example,

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\{-\theta\psi(x)\}, \quad p(y|x) \propto \exp\{-f_y(x)\},$$

with f_y and ψ convex l.s.c. functions, and f_y L-Lipschitz differentiable.

We assume that $p(x|\theta)$ is proper, i.e.,

$$Z(\theta) = \int_{\mathbb{R}^d} \exp\{-\theta\psi(x)\} \mathrm{d}x < \infty,$$

with $Z(\theta)$ unknown and generally intractable.

Maximum-a-posteriori estimation

Recall that when θ is fixed, the posterior $p(x|y,\theta)$ is log-concave and

$$\hat{x}_{MAP} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, f_y(x) + \theta\psi(x)$$

is a convex optimisation problem that can be often solved efficiently.

For example, one can consider the proximal gradient algorithm

$$x^{m+1} = \operatorname{prox}_{\psi}^{L^{-1}}\{x^m + L^{-1}\nabla f_y(x^m)\}$$

to iteratively compute $\hat{x}_{MAP}$.

However, θ is often unknown, significantly complicating the problem.

Regularisation parameter MLE

We have two option to tackle θ. We first consider an empirical Bayes approach based on the MLE

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmax}_{\theta \in \Theta} p(y|\theta)\,, \\ &= \operatorname*{argmax}_{\theta \in \Theta} \int_{\mathbb{R}^d} p(y, x|\theta)\mathrm{d}x\,,\end{aligned}$$

which we solve efficiently by using a stochastic gradient algorithm driven by two proximal MCMC kernels (see Fernandez-Vidal and Pereyra (2018)).

Given $\hat{\theta}$, we then straightforwardly compute

$$\hat{x}_{MAP} = \operatorname*{argmin}_{x \in \mathbb{R}^d} f_y(x) + \hat{\theta}\psi(x)\,. \tag{17}$$

Projected gradient algorithm

Assume that Θ is convex, and that $\hat{\theta}$ is the only root of $\nabla_\theta \log p(y|\theta)$ in Θ.

Then $\hat{\theta}$ is also the unique solution of the fixed-point equation

$$\theta = P_\Theta \left[\theta + \delta \nabla_\theta \log p(y|\theta)\right] .$$

where P_Θ is the projection operator on Θ and $\delta > 0$.

If $\nabla \log p(y|\theta)$ was tractable, we could compute $\hat{\theta}$ iteratively by using

$$\theta^{(t+1)} = P_\Theta \left[\theta^{(t)} + \delta_t \nabla_\theta \log p(y|\theta^{(t)})\right],$$

with sequence $\delta_t = \alpha t^{-\beta}$, $\alpha > 0$, $\beta \in [1/2, 1]$.

However, $\nabla \log p(y|\theta)$ is "doubly" intractable...

Stochastic projected gradient algorithm

To circumvent the intractability of $\nabla_\theta \log p(y|\theta)$ we use Fisher's identity

$$\begin{aligned}\nabla_\theta \log p(y|\theta) &= \mathrm{E}_{\mathbb{x}|y,\theta}\{\nabla_\theta \log p(\mathbb{x}, y|\theta)\}\,, \\ &= -\mathrm{E}_{\mathbb{x}|y,\theta}\{\psi(\mathbb{x}) + \nabla_\theta \log Z(\theta)\}\,,\end{aligned}$$

together with the identity

$$\nabla_\theta \log Z(\theta) = -\mathrm{E}_{\mathbb{x}|\theta}\{\psi(\mathbb{x})\}\,,$$

to obtain $\nabla_\theta \log p(y|\theta) = \mathrm{E}_{\mathbb{x}|\theta}\{\psi(\mathbb{x})\} - \mathrm{E}_{\mathbb{x}|y,\theta}\{\psi(\mathbb{x})\}$.

This leads to the equivalent fixed-point equation

$$\theta = P_\Theta\left(\theta + \delta \mathrm{E}_{\mathbb{x}|\theta}\{\psi(\mathbb{x})\} - \delta \mathrm{E}_{\mathbb{x}|y,\theta}\{\psi(\mathbb{x})\}\right), \tag{18}$$

which we solve by using a stochastic approximation algorithm.

SAPG algorithm driven by MCMC kernels

Initialisation $x^{(0)}, u^{(0)} \in \mathbb{R}^d$, $\theta^{(0)} \in \Theta$, $\delta_t = \delta_0 t^{-0.8}$.

for $t = 0$ to n

1. MCMC update $x^{(t+1)} \sim M_{x|y,\theta^{(t)}}(\cdot|x^{(t)})$ targeting $p(x|y,\theta^{(t)})$
2. MCMC update $u^{(t+1)} \sim K_{x|\theta^{(t)}}(\cdot|u^{(t)})$ targeting $p(x|\theta^{(t)})$
3. Stoch. grad. update

$$\theta^{(t+1)} = P_\Theta\left[\theta^{(t)} + \delta_t \psi(u^{(t+1)}) - \delta_t \psi(x^{(t+1)})\right].$$

end for

Output The iterates $\theta^{(t)} \to \hat{\theta}$ as $n \to \infty$.

SAPG algorithm driven MCMC kernels

Initialisation $x^{(0)}, u^{(0)} \in \mathbb{R}^d$, $\theta^{(0)} \in \Theta$, $\delta_t = \delta_0 t^{-0.8}$, $\lambda = 1/L$, $\gamma = 1/4L$.

for $t = 0$ to n

1. Coupled Proximal MCMC updates: generate $z^{(t+1)} \sim \mathcal{N}(0, \mathbb{I}_d)$

$$x^{(t+1)} = (1 - \frac{\gamma}{\lambda})x^{(t)} - \gamma \nabla f_y\left(x^{(t)}\right) + \frac{\gamma}{\lambda}\mathrm{prox}_{\psi}^{\theta\lambda}\left(x^{(t)}\right) + \sqrt{2\gamma}z^{(t+1)},$$
$$u^{(t+1)} = (1 - \frac{\gamma}{\lambda})u^{(t)} + \frac{\gamma}{\lambda}\mathrm{prox}_{\psi}^{\theta\lambda}\left(u^{(t)}\right) + \sqrt{2\gamma}z^{(t+1)},$$

2. Stochastic gradient update

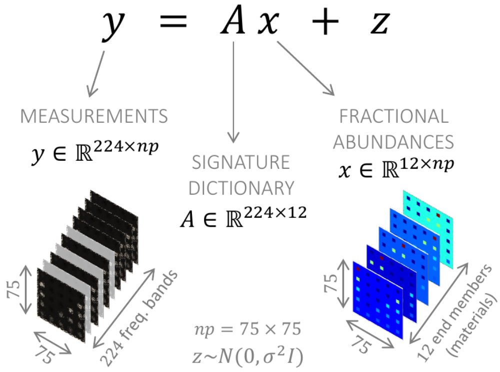
$$\theta^{(t+1)} = P_\Theta\left[\theta^{(t)} + \delta_t \psi(u^{(t+1)}) - \delta_t \psi(x^{(t+1)})\right].$$

end for

Output Averaged estimator $\bar{\theta} = n^{-1}\sum_{t=1}^{n} \theta^{(t+1)}$ converges approx. to $\hat{\theta}$.

Illustrative example: hyperspectral unmixing

- We seek to recover fractional abundances x from the mixed noisy spectral signatures y measured for every pixel.



Empirical Bayesian MAP estimation

We consider the prior of Iordache et al. (IEEE TGRS, 2012)

$$p(x|\theta) \propto \exp\{\theta_{TV}\, TV(x) + \theta_{L1}\, \|x\|_1\} \quad \text{s.t.}\, x \geq 0\,,$$

and use the SAPG scheme to estimate θ_{TV} and θ_{L1} by MMLE.

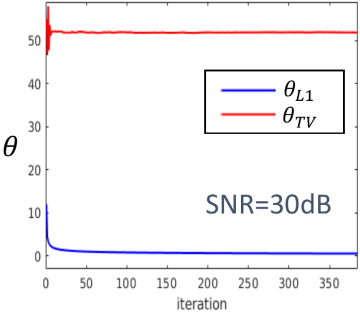
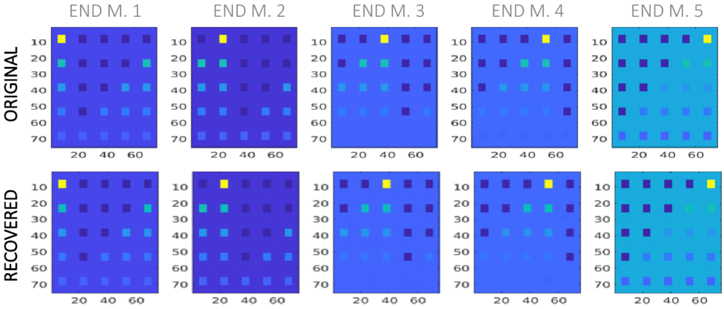


Figure: Evolution of the iterates associated to θ_{TV} and θ_{L1}.

Empirical Bayesian MAP estimation

Given $\hat{\theta}_{TV}$ and $\hat{\theta}_{L1}$, we compute the MAP solution by using the convex optimisation algorithm SUNSAL:

MAP estimates of fractional abundances for 5 end members



Hierarchical Bayesian treatment of unknown θ

Option 2: hierarchical Bayesian inference also allows estimating x without specifying the value of θ.

We incorporate θ to the model by assigning it an hyper-prior $p(\theta)$.

The extended model is

$$\begin{aligned} p(x, \theta | y) &= p(y|x)p(x|\theta)p(\theta)/p(y), \\ &\propto \frac{\exp\{-f_y(x) - \theta\psi(x) - \log p(\theta)\}}{Z(\theta)}, \end{aligned} \tag{19}$$

but $Z(\theta) = \int_{\mathbb{R}^d} \exp\{-\theta\psi(x)\}\mathrm{d}x$ is typically intractable!

If we had access to $Z(\theta)$ we could either estimate x and θ jointly, or alternatively marginalise θ followed by MAP inference on x.

Idea: Use MYULA to estimate $E[\psi(\mathbb{x})|\theta]$ over a θ-grid, and then approximate $\log Z(\theta)$ by using the identity $\frac{\mathrm{d}}{\mathrm{d}\theta}\log Z(\theta) = E[\psi(\mathbb{x})|\theta]$.

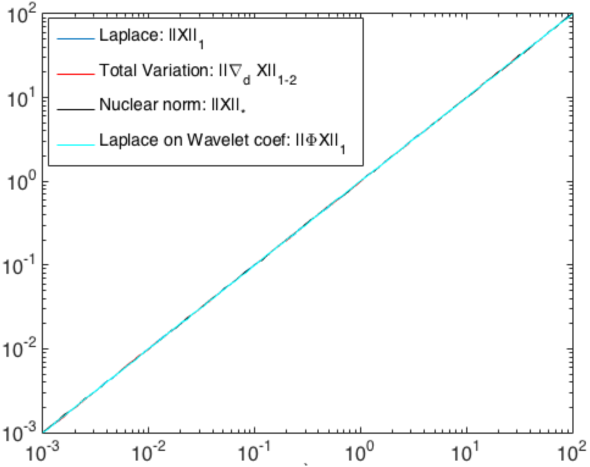


Figure: Monte Carlo approximations of $E[\psi(\mathbb{x})|\theta]/d$ Vs θ for 4 widely used prior distributions and for $\theta \in [10^{-3}, 10^{2}]$. **Surprise:** they all coincide!

Main theoretical result

Definition[***k***-homogeneity] The regulariser ψ is a k-homogeneous function if $\exists k \in \mathbb{R}^+$ such that

$$\psi(\eta x) = \eta^k \psi(x), \quad \forall x \in \mathbb{R}^d, \forall \eta > 0. \tag{20}$$

Theorem[Pereyra et al. (2015)] Suppose that ψ, the sufficient statistic of $p(x|\theta)$, is k-homogenous. Then the normalisation factor has the form

$$Z(\theta) = Z(1)\theta^{-d/k},$$

with (generally intractable) constant $Z(1)$ independent of θ.

Note: This result holds for all norms (e.g., ℓ_1, ℓ_2, total-variation, nuclear, etc.), composite norms (e.g., $\ell_1 - \ell_2$), and compositions of norms with linear operators (e.g., analysis terms of the form $\|\Psi x\|_1$)!

Marginal maximum-a-posteriori estimation of x

Knowledge of $Z(\theta)$ enables (for example) *marginal MAP estimation* of x

$$\begin{aligned} \hat{x}^{\dagger}_{MAP} &= \underset{x \in \mathbb{R}^d}{\operatorname{argmax}} \int_0^\infty p(x, \theta | y) \mathrm{d}\theta, \\ &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, f_y(x) + (d/k + \alpha) \log\{\psi(x) + \beta\}, \end{aligned} \tag{21}$$

where we have used the hyper-prior $\theta \sim \mathrm{Gamma}(\alpha, \beta)$.

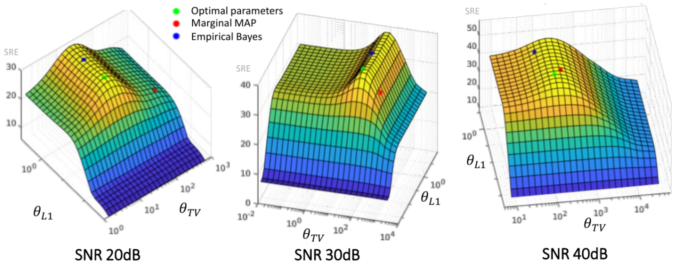
We can compute $\hat{x}^{\dagger}$ efficiently by *majorisation-minimisation* optimisation

$$\begin{aligned} x^{(t)} &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, f_y(x) + \theta^{(t-1)} \psi(x), \\ \theta^{(t)} &= \frac{d/k + \alpha}{\psi(x^{(t)}) + \beta}. \end{aligned} \tag{22}$$

which is also an *expectation-maximisation* algorithm.

Illustrative example: hyperspectral unmixing

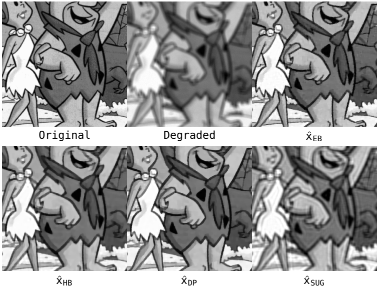
We use the hyperspectral unmixing example to develop an intuition for the strengths and drawbacks of the EB and HB approaches.

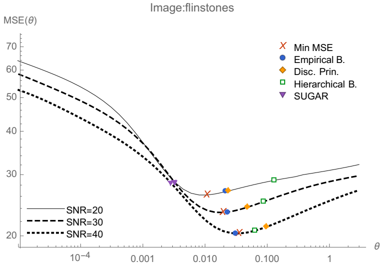


EB performs remarkably well in low SNR (high noise) regimes, close to the oracle performance. The benefits of HB kick-in in high SNR regimes.

Illustrative example: image deconvolution with a TV prior

	SNR=20dB		SNR=30		SNR=40	
	MSE	Time (min)	MSE	Time (min)	MSE	Time (min)
Best	23.29		21.39		19.06	
Emp. Bayes	23.50	0.86	21.46	0.85	19.24	0.85
Hier. Bayes	25.07	0.58	22.84	1.27	19.84	3.27
Disc. Princp.	23.73		21.87		19.78	
SUGAR	24.44	3.92	24.24	4.50	24.21	4.81



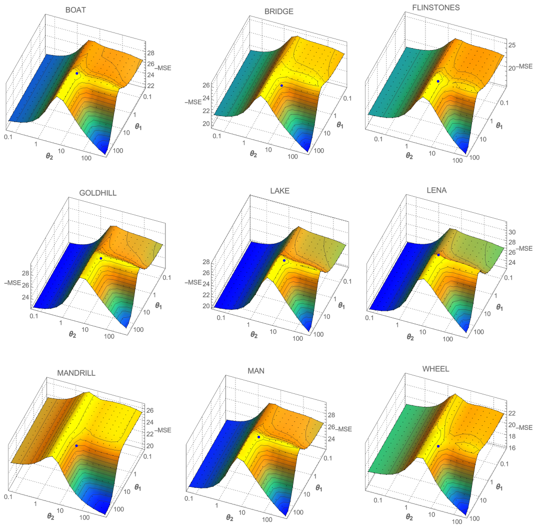


Denoising with Total Generalized Variation

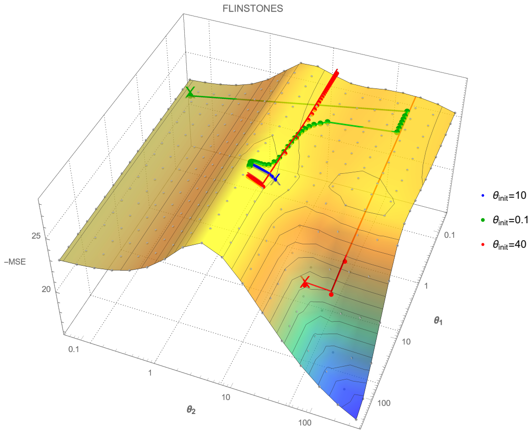
We consider $TGV_\theta^2(u) = \inf_{v \in BD(\Omega)} \theta_1 \int_\Omega |\nabla u - v| + \theta_2 \int_\Omega |\varepsilon(v)|$ with $k = 2$.

Figure: *Goldhill* image (Original-Degraded-Estimated MAP), SNR=12dB.

Denoising with Total Generalized Variation



Evolution of θ through iterations starting from different initial values:



Outline

Concluding remarks

- There are three main frameworks to solve imaging inverse problems - mathematical analysis, Bayesian statistics, and machine learning - with complementary strengths and drawbacks.
- This tutorial focused on Bayesian imaging methods that leverage a range of ideas and techniques from mathematical analysis to perform computations efficiently.
- We explored integrating modern stochastic and optimisation approaches to construct proximal MCMC methods and stochastic proximal gradient algorithms.
- The tutorials of tomorrow will focus on methods at that also incorporate a significant machine learning component.

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., J.D., M., and M., P. (2022). Proximal nested sampling for high-dimensional Bayesian model selection. *Statistics and Computing*, 37(87).

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*.

Moreau, J.-J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.*, 255:2897–2899.

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). Maximum-a-posteriori estimation with bayesian confidence regions. *SIAM J. Imaging Sci.*, 6(3):1665–1688.

Pereyra, M. (2019). Revisiting maximum-a-posteriori estimation in log-concave models. *SIAM J. Imaging Sci.*, 12(1):650–670.

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 Journal on Imaging Sciences*, 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.

Zhu, L., Zhang, W., Elnatan, D., and Huang, B. (2012). Faster STORM using compressed sensing. *Nat. Meth.*, 9(7):721–723.