Bayesian inference and mathematical imaging. Part III: probability & convex optimisation.

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

Maxwell Institute for Mathematical Sciences, Heriot-Watt University

January 2019, CIRM, Marseille.

Outline

Imaging inverse problems

- We are interested in an unknown image $x \in \mathbb{R}^d$.
- We measure y, related to x by a statistical model $p(y|x)$.
- The recovery of x from y is ill-posed or ill-conditioned, resulting in significant uncertainty about x.
- For example, in many imaging problems

$$y = Ax + w,$$

for some operator A that is rank-deficient, and additive noise w.

The Bayesian framework

- We use priors to reduce uncertainty and deliver accurate results.
- Given the prior $p(x)$, the posterior distribution of x given y

$$p(x|y) = p(y|x)p(x)/p(y)$$

models our knowledge about x after observing y.

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

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

where $\phi(x)$ is a convex function and $Z = \int \exp\{-\phi(x)\}\mathrm{d}x$.

Maximum-a-posteriori (MAP) estimation

The predominant Bayesian approach in imaging is MAP estimation

$$\begin{aligned} \hat{x}_{MAP} &= \operatorname*{argmax}_{x \in \mathbb{R}^d} p(x|y), \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^d} \phi(x), \end{aligned} \tag{1}$$

computed efficiently, even in very high dimensions, by (proximal) convex optimisation (Green et al., 2015; Chambolle and Pock, 2016).

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

1. it provides little information about $p(x|y)$,
2. it is not theoretically well understood (yet),
3. it struggles with unknown/partially unknown models.

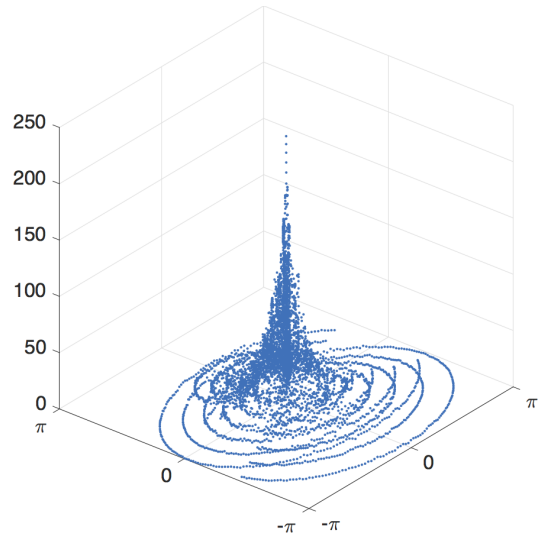
Illustrative example: astronomical image reconstruction

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

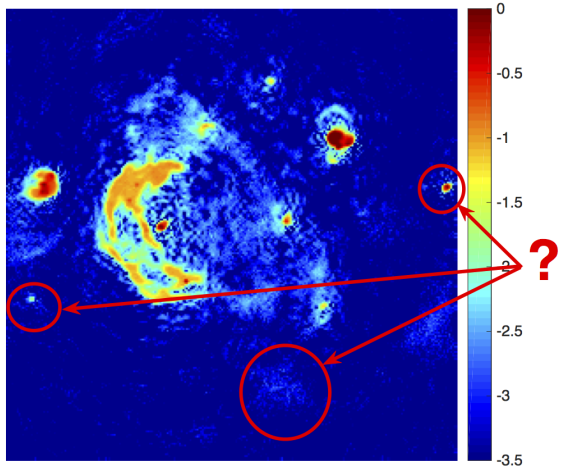
$$y = M\mathcal{F}x + w,$$

where $\mathcal{F}$ is the continuous Fourier transform, $M \in \mathbb{C}^{m \times d}$ is a measurement mask operator, and w is Gaussian noise. We use the model

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



$|y|$



$\hat{x}_{MAP}$

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

Outline

Outline

Posterior credible regions

Where does the posterior probability mass of x lie?

Recall that C_α is a $(1-\alpha)\%$ posterior credible region if

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

and the decision-theoretically optimum is the HPD region (Robert, 2001)

$$C_\alpha^* = \left\{x : \phi(x) \leq \gamma_\alpha\right\},$$

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

We could estimate C_α^* by MCMC sampling, but in high-dimensional log-concave models this is not necessary because something beautiful happens...

A concentration phenomenon!

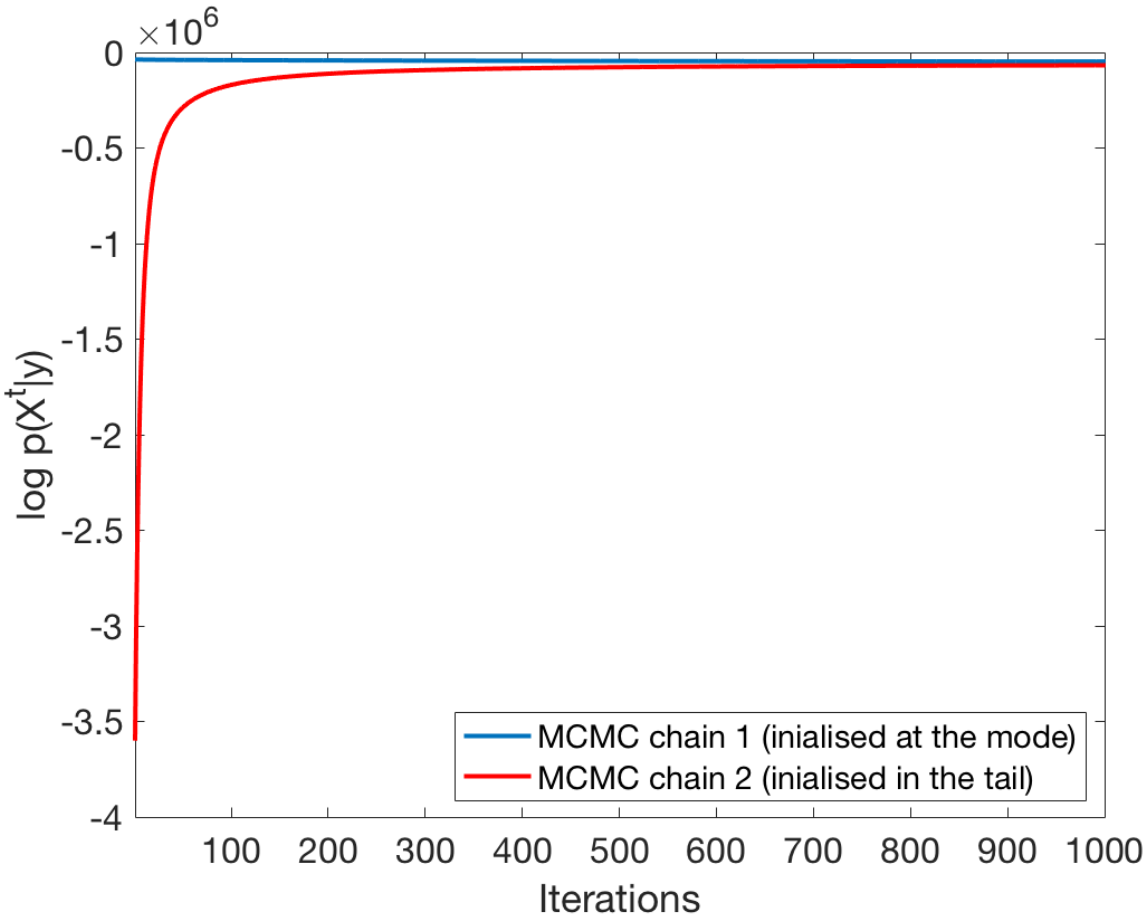


Figure : Convergence to "typical" set $\{x : \log p(x|y) \approx \mathrm{E}[\log p(x|y)]\}$.

Proposed approximation of C_α^*

Theorem 2.1 (Pereyra (2016))

Suppose that the posterior $p(x|y) = \exp\{-\phi(x)\}/Z$ is log-concave on $\mathbb{R}^d$. Then, for any $\alpha \in (4\exp(-n/3), 1)$, the HPD region C_α^* is contained by

$$\tilde{C}_\alpha = \{x : \phi(x) \leq \phi(\hat{x}_{MAP}) + \sqrt{d}\tau_\alpha + d)\},$$

with universal positive constant $\tau_\alpha = \sqrt{16\log(3/\alpha)}$ independent of $p(x|y)$.

Remark 1: $\tilde{C}_\alpha$ is a conservative approximation of C_α^*, i.e.,

$$x \notin \tilde{C}_\alpha \implies x \notin C_\alpha^*.$$

Remark 2: $\tilde{C}_\alpha$ is available as a by-product in any convex inference problem that is solved by MAP estimation!

Approximation error bounds

Is $\tilde{C}_\alpha$ a reliable approximation of C^*_α?

Theorem 2.2 (Finite-dimensional error bound (Pereyra, 2016))

Let $\tilde{\gamma}_\alpha = \phi(\hat{x}_{MAP}) + \sqrt{d}\tau_\alpha + d$. If $p(x|y)$ is log-concave on $\mathbb{R}^d$, then

$$0 \leq \frac{\tilde{\gamma}_\alpha - \gamma_\alpha}{d} \leq 1 + \eta_\alpha d^{-1/2},$$

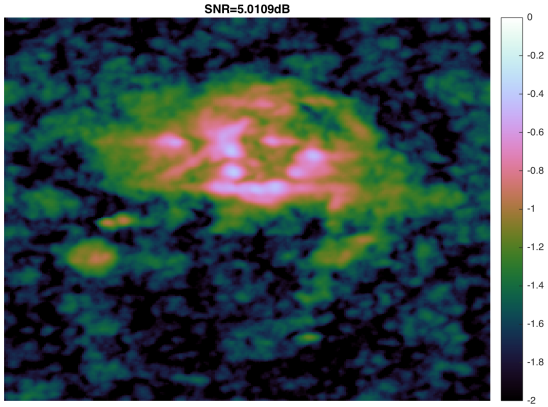
with universal positive constant $\eta_\alpha = \sqrt{16\log(3/\alpha)} + \sqrt{1/\alpha}$.

Remark 3: $\tilde{C}_\alpha$ is stable (as d becomes large, the error $(\tilde{\gamma}_\alpha - \gamma_\alpha)/d \lessapprox 1$).
Remark 4: The lower and upper bounds are asymptotically tight w.r.t. d.

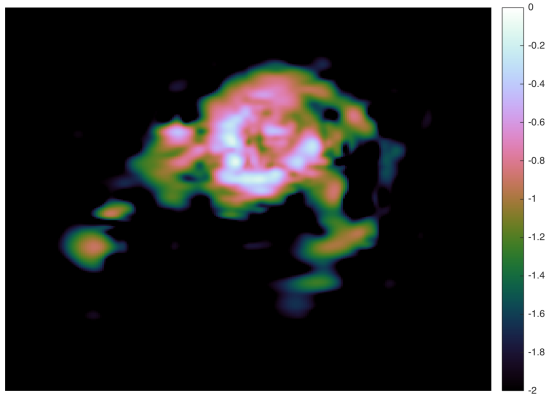
Outline

Example: Uncertainty visualisation in astro-imaging

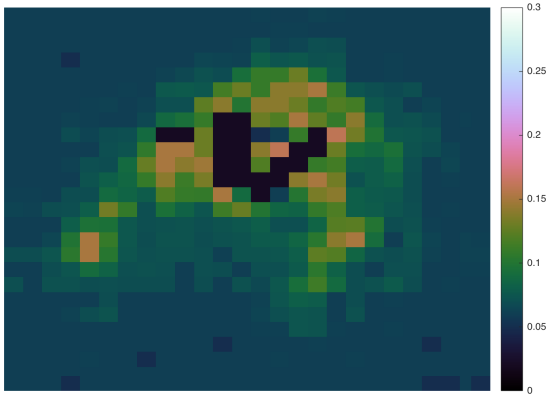
Radio-interferometry with redundant wavelet frame (Cai et al., 2017).



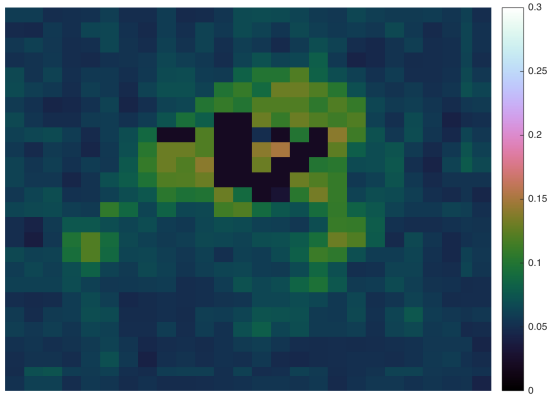
dirty image $\hat{x}_{penML}(y)$



$\hat{x}_{MAP}$



approx. credible intervals (scale 10×10 pixels)

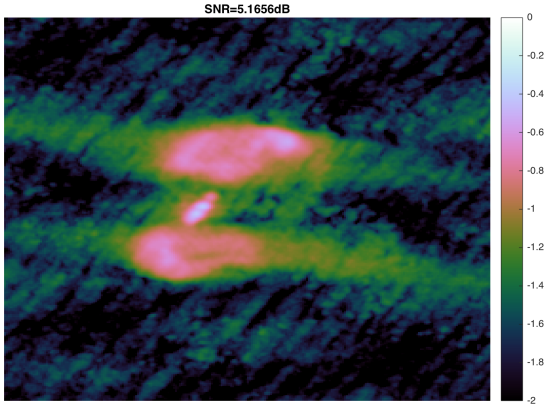


"exact" intervals (MCMC, minutes)

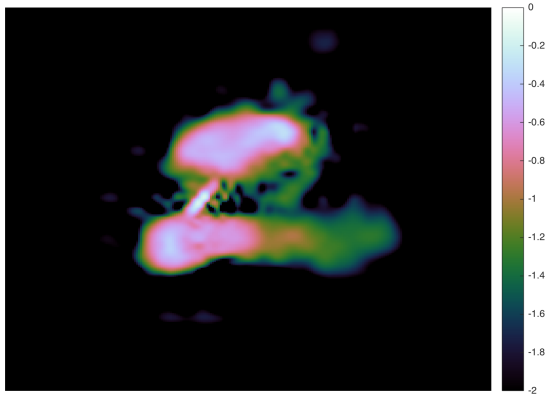
M31 radio galaxy (size 256×256 pixels, comp. time 1.8 secs.)

Example: Uncertainty visualisation in astro-imaging

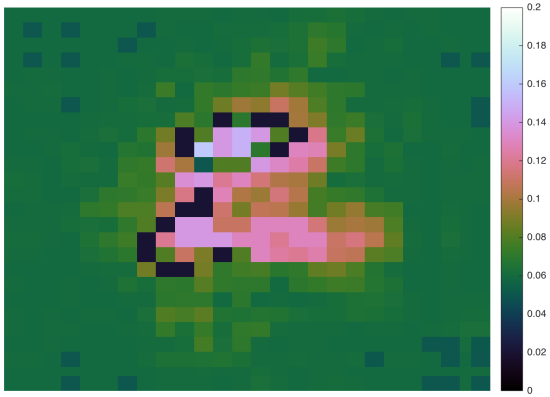
Radio-interferometry with redundant wavelet frame (Cai et al., 2017).



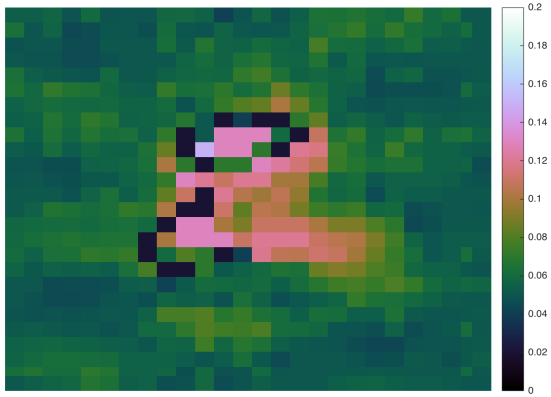
dirty image $\hat{x}_{penML}(y)$



$\hat{x}_{MAP}$



approx. credible intervals (scale 10×10 pixels)

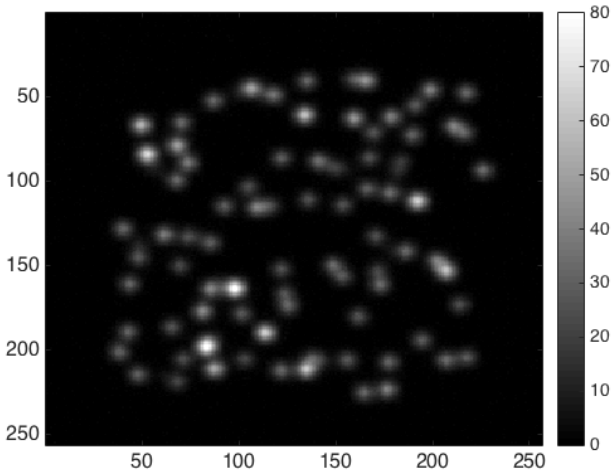


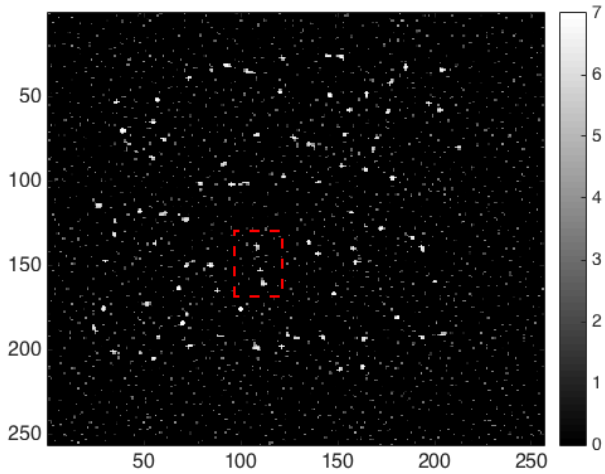
"exact" intervals (MCMC, minutes)

3C2888 radio galaxy (size 256×256 pixels, comp. time 1.8 secs.)

Example: uncertainty visualisation in microscopy

Live cell microscopy sparse super-resolution (Zhu et al., 2012):





$$y = Ax + w$$

(A is a blur operator)

$$\hat{x}_{MAP} = \operatorname{argmin}_{x \in \mathbb{R}^d} \|y - Hx\|^2 / 2\sigma^2 + \lambda \|x\|_1$$

(log-scale)

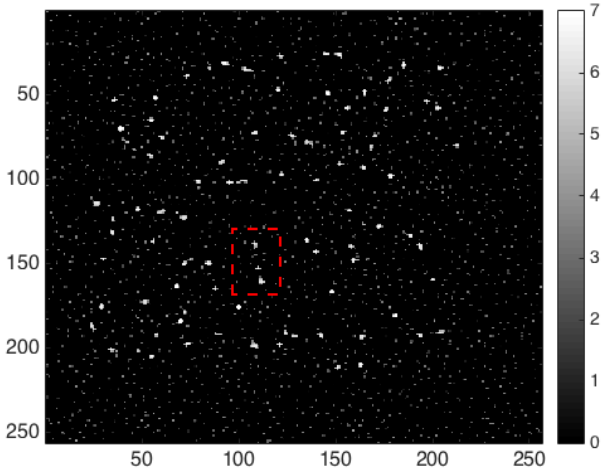
Consider the molecular structure in the highlighted region:

- Are we confident about this structure (its presence, position, etc.)?
- **Idea**: use $\tilde{C}_\alpha$ to explore/quantify the uncertainty about this structure.

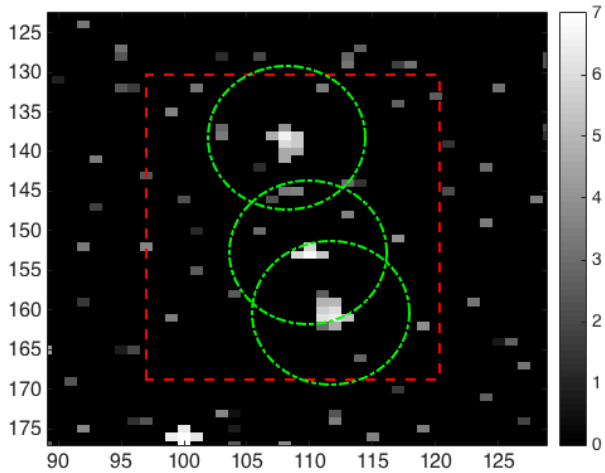
Example: uncertainty visualisation in microscopy

Position uncertainty quantification

Find maximum molecule displacement within $\tilde{C}_{0.05}$:



$$\hat{x}_{MAP} = \text{argmin}_{x\in\mathbb{R}^d} \|y - Ax\|^2/2\sigma^2 + \lambda\|x\|_1$$



Molecule position uncertainty
$(\pm 93nm \times \pm 140nm)$

Note: Uncertainty analysis $(\pm 93nm \times \pm 140nm)$ in agreement with MCMC estimates $(\pm 78nm \times \pm 125nm$ - approx. error of order of 1 pixel), and with experimental results (average precision $\pm 80nm$) of Zhu et al. (2012).

Outline

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

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

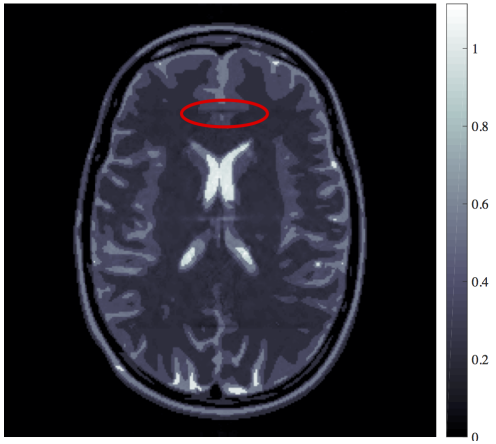
Theorem (Repetti et al., 2018)
Let $\mathcal{S}$ denote the region of $\mathbb{R}^d$ associated with H_0, containing all images *without the structure* of interest. Then

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

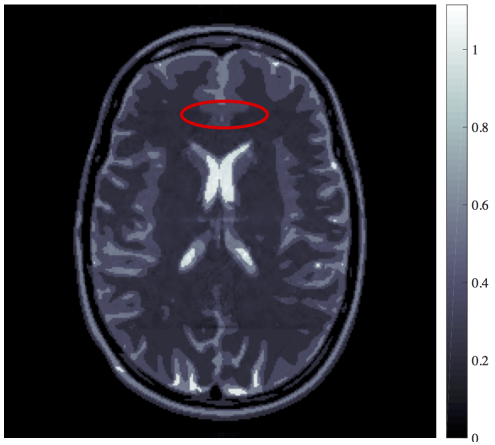
If in addition $\mathcal{S}$ is convex, then checking $\mathcal{S} \cap \widetilde{\mathcal{C}}_\alpha = \varnothing$ is a convex problem

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

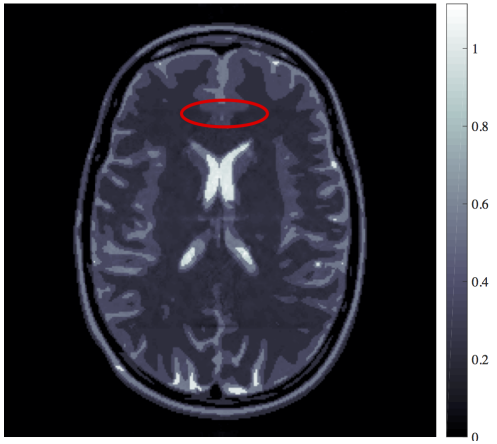
Uncertainty quantification in MRI imaging



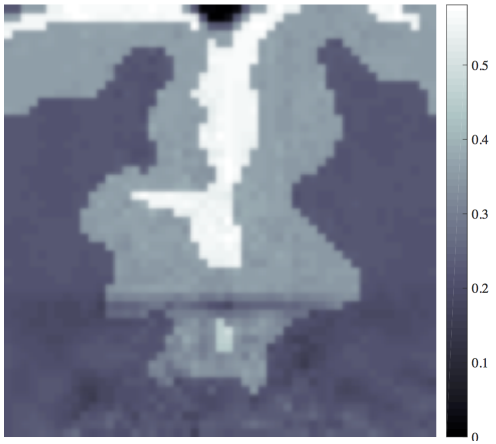
$\hat{x}_{MAP}$



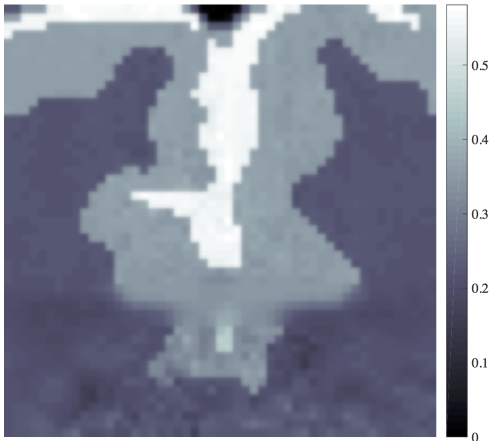
$\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$



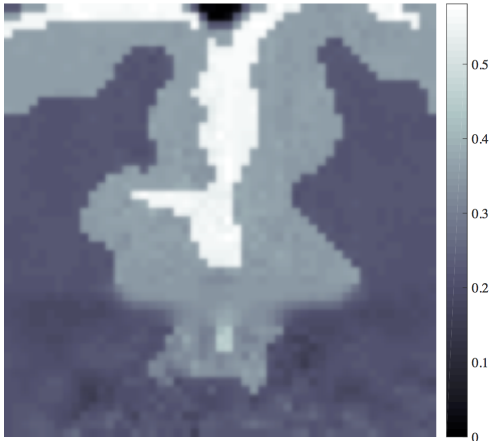
$\underline{x} \in \mathcal{S}$



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



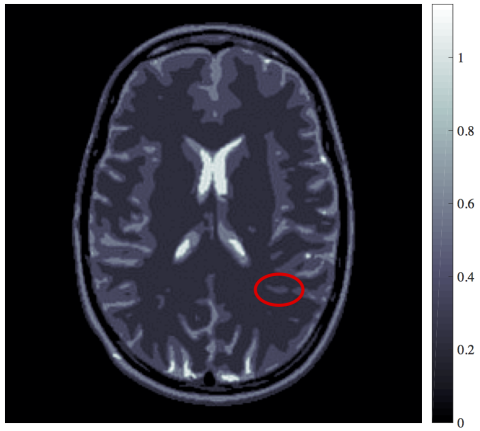
$\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ (zoom)



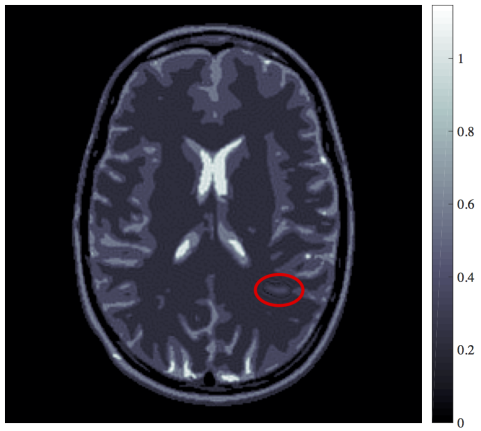
$\underline{x} \in \mathcal{S}$ (zoom)

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.

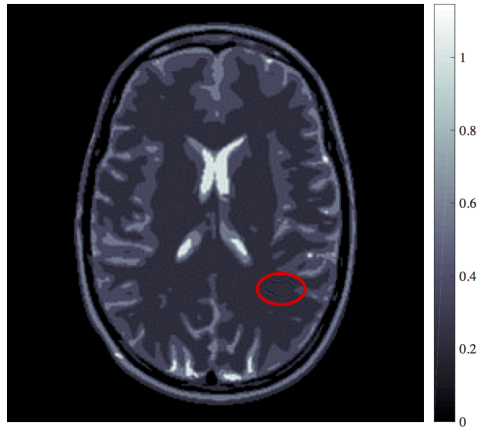
Uncertainty quantification in MRI imaging



$\hat{x}_{MAP}$

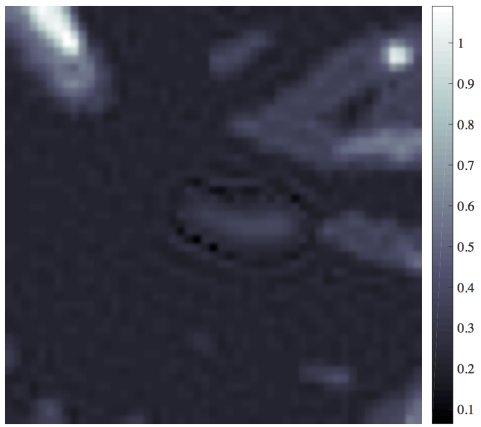


$\bar{x} \in \mathcal{C}_{0.01}$

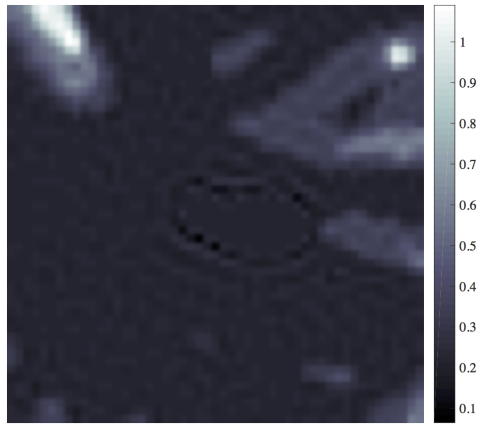


$\underline{x} \in \mathcal{S}_0$

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



$\bar{x} \in \mathcal{C}_{0.01}$ (zoom)

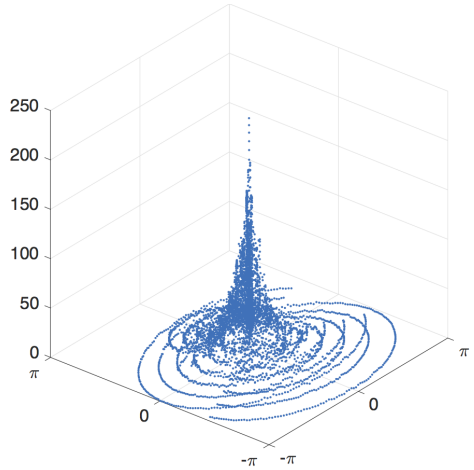


$\underline{x} \in \mathcal{S}_0$ (zoom)

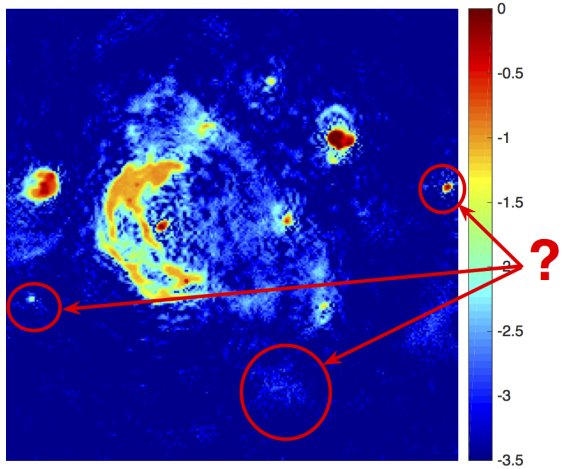
MRI experiment: test images $\bar{x} \neq \underline{x}$, hence we reject H_0 and conclude that there is significant evidence in favour of the observed structure.

Uncertainty quantification in radio-interferometric imaging

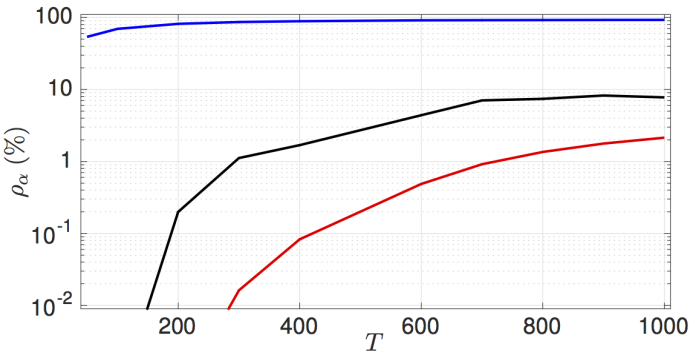
Quantification of minimum energy of different energy structures, at level $(1-\alpha) = 0.99$, as the number of measurements $T = \dim(y)/2$ increases.



$|y|$



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



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.

Outline

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

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

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

Example: Euclidean setting $L(u,x) = \|u - x\|^2$ and $\hat{x}_L = \hat{x}_{MMSE} = \mathrm{E}\{x|y\}$.

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

Does the convex geometry of $p(x|y)$ define an interesting loss $L(u, x)$?

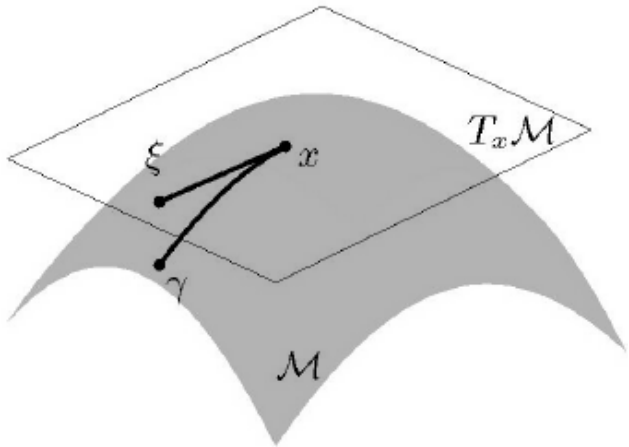
We use differential geometry to relate the convexity of $p(x|y)$, the geometry of the parameter space, and the loss L to perform estimation.

Differential geometry

A Riemannian manifold $\mathcal{M} = (\mathbb{R}^d, g)$, with metric $g : \mathbb{R}^d \to \mathcal{S}^d_{++}$ and global coordinate system x, is a vector space that is locally Euclidean.

For any point $x \in \mathbb{R}^d$ we have an Euclidean tangent space $\mathcal{T}_x\mathbb{R}^d$ with inner product $\langle u, x \rangle = u^\top g(x) x$ and norm $\|x\| = \sqrt{x^\top g(x) x}$.

This geometry is local and may vary smoothly from $\mathcal{T}_x\mathbb{R}^d$ to $\mathcal{T}_{x'}\mathbb{R}^d$ following the affine connection $\Gamma \in \mathbb{R}^{d \times d \times d}$, given by $\Gamma_{ij,k}(x) = \partial_k g_{i,j}(x)$.



Divergence functions

Similarly to Euclidean spaces, the manifold $(\mathbb{R}^d, g)$ supports divergences:

Definition 1 (Divergence functions)

A function $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a divergence function on $\mathbb{R}^d$ if the following conditions hold for any $u, x \in \mathbb{R}^d$:

- $D(u, x) \geq 0, \ \forall \ u, x \in \mathbb{R}^d$,
- $D(u, x) = 0 \iff x = u$,
- $D(u, x)$ is strongly convex w.r.t. u, and $\mathcal{C}^2$ w.r.t u and x.

Canonical divergence

We focus on the *canonical divergence* on $(\mathbb{R}^d, g)$, a generalisation of the Euclidean squared distance to this kind of manifold:

Definition 2 (Canonical divergence (Ay and Amari, 2015))

For any $(u, x) \in \mathbb{R}^d \times \mathbb{R}^d$, the canonical divergence on $(\mathbb{R}^d, g)$ is given by

$$D(u, x) = \int_0^1 t\dot{\gamma}_t^\top g(\gamma_t)\dot{\gamma}_t \mathrm{d}t \tag{3}$$

where γ_t is the Γ-geodesic from u to x and $\dot{\gamma}_t = \mathrm{d}/\mathrm{d}t\, \gamma_t$.

1. D fully species $(\mathbb{R}^d, g)$ and vice-versa.
2. $D(x + dx, x) = \|dx\|^2/2 + o(\|dx\|^2)$ where $\|\cdot\|$ is the norm on $\mathcal{T}_x\mathbb{R}^d$.
3. For Euclidean space with $\langle u, x\rangle = u^\top g x$, $D(u, x) = \frac{1}{2}(u - x)^\top g(u - x)$.

Differential-geometric derivation of $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$

Theorem 3 (Canonical Bayesian estimators - Part 1 (Pereyra, 2016))

Suppose that $\phi(x) = -\log p(x|y)$ *is strongly convex, continuous, and* $\mathcal{C}^3$ *on* $\mathbb{R}^d$. *Let* $(\mathbb{R}^d, g)$ *be the manifold induced by* ϕ, *i.e.,* $g_{i,j}(x) = \partial_i \partial_j \phi(x)$. *Then, the canonical divergence on* $(\mathbb{R}^d, g)$ *is the* ϕ*-Bregman divergence*

$$D_\phi(u, x) = \phi(u) - \phi(x) - \nabla\phi(x)(u - x).$$

Remark: Because ϕ is strongly convex, then $\phi(u) > \phi(x) - \nabla\phi(x)(u - x)$ for any $u \neq x$. The divergence $D_\phi(u, x)$ quantifies this gap, related to the length of the affine geodesic from u to x on the space induced by $p(x|y)$.

Differential-geometric derivation of $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$

Theorem 4 (Canonical Bayesian estimators - Part 2 (Pereyra, 2016))

The Bayesian estimator associated with $D_\phi(u,x)$ is unique and is given by

$$\begin{aligned}\hat{x}_{D_\phi} &\triangleq \operatorname*{argmin}_{u\in\mathbb{R}^d} \mathrm{E}_{x|y}[D_\phi(u,x)]\,,\\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d} \phi(x)\,,\\ &= \hat{x}_{MAP}\,.\end{aligned}$$

Remark2: $\hat{x}_{MAP}$ stems from Bayesian decision theory, and hence it stands on the same theoretical footing as the core Bayesian methodologies.

Remark3: The definition of the MAP estimator as the maximiser $\hat{x}_{MAP} = \operatorname{argmax}_{x\in\mathbb{R}^d} p(x|y)$ is mainly algorithmic for these models.

Differential-geometric derivation of $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$

Theorem 5 (Canonical Bayesian estimators - Part 3 (Pereyra, 2016))

Moreover, the Bayesian estimator associated with the dual canonical divergence $D^*_\phi(u,x) = D_\phi(x,u)$ is also unique and is given by

$$\begin{aligned}\hat{x}_{D^*_\phi} &\triangleq \operatorname*{argmin}_{u\in\mathbb{R}^d} \mathrm{E}_{x|y}[D^*_\phi(u,x)]\,,\\ &= \int_{\mathbb{R}^d} x p(x|y)\mathrm{d}x\,,\\ &= \hat{x}_{MMSE}\,.\end{aligned}$$

Remark 4: $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$ exhibit a surprising duality, arising from the asymmetry of the canonical divergence that $p(x|y)$ induces on $\mathbb{R}^d$.

Remark 5: These results carry partially to models that are not strongly convex, not smooth, or that involve constraints on the parameter space.

Expected estimation error bound

Are $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$ "good" estimators of $x|y$?

Proposition 3.1 (Expected canonical error bound)

Suppose that $\phi(x) = -\log \pi(x|y)$ is convex on $\mathbb{R}^d$ and $\mathcal{C}^1$. Then,

$$\mathrm{E}_{x|y}\left[D^*_\phi(\hat{x}_{MMSE}, x)/d\right] \leq \mathrm{E}_{x|y}\left[D^*_\phi(\hat{x}_{MAP}, x)/d\right] \leq 1.$$

Proposition 3.2 (Expected error w.r.t. regularisation function)

Also assume that the regularisation $h(x) = -\log p(x)$ is convex. Then,

$$\mathrm{E}_{x|y}\left[D^*_h(\hat{x}_{MMSE}, x)/d\right] \leq \mathrm{E}_{x|y}\left[D^*_h(\hat{x}_{MAP}, x)/d\right] \leq 1.$$

Remark 6: These are high-dimensional stability results for $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$; the estimation error cannot grow faster than the number of pixels.

Example 1: denoising with wavelet shrinkage prior

Consider a linear problem of the form $y = Ax + w$ and a shrinkage prior on the wavelet coefficients $z = Wx$. We consider the smoothed Laplace prior

$$p(z) \propto \exp\{-\sum_{i=1}^{d} \lambda \sqrt{z_i^2 + \gamma^2}\}$$

where $\lambda \in \mathbb{R}^+$ and $\gamma \in \mathbb{R}^+$ are scale and shape parameters.

The likelihood is $p(y|z) \propto \exp\{-\frac{1}{2\sigma^2}\|y - AW^\top z\|_2^2$ and hence

$$p(z|y) \propto \exp\{-\frac{1}{2\sigma^2}\|y - AW^\top z\|_2^2 - \sum_{i=1}^{d} \lambda \sqrt{z_i^2 + \gamma^2}\}$$

This model is $\mathcal{C}^\infty$ and strongly log-concave, and hence the theory applies.

Example 1: denoising with wavelet shrinkage prior

To analyse the geometry induced by $\phi(z) = -\log p(z|y)$ we suppose that $A = \mathbb{I}$ and $W^\top W = \mathbb{I}$, and obtain $D_\phi(u, z) = \sum_{i=1}^d D_\psi(u_i, z_i)$ with

$$D_\psi(u_i, z_i) = \frac{1}{2\sigma^2}(u_i - z_i)^2 + \lambda \frac{\sqrt{z_i^2 + \gamma^2}\sqrt{u_i^2 + \gamma^2} - z_i u_i - \gamma^2}{\sqrt{z_i^2 + \gamma^2}}.$$

The non-quadratic term introduces additional shrinkage and leads to the differences between x_{MMSE} and x_{MAP}.

To develop an intuition for this behaviour we analyse $z_i \ll \gamma$ and $z_i \gg \gamma$.

Example 1: denoising with wavelet shrinkage prior

When $z_i \gg \gamma$ the non-quadratic term vanishes, hence

$$D_\psi(u_i, z_i) \approx \frac{1}{2\sigma^2}(u_i - z_i)^2 .$$

Hence, when $p(z_i|y)$ has most of its mass in large values of z_i, the MAP estimate for z_i will agree with the MMSE estimate $\mathrm{E}(z_i|y)$.

In this case there is no additional shrinkage from the estimator.

Example 1: denoising with wavelet shrinkage prior

Conversely, when $z_i \ll \gamma$ we have

$$\begin{aligned} D_\psi(u_i, z_i) &\approx \tfrac{1}{2\sigma^2}(u_i - z_i)^2 + \lambda|u_i|\,, \\ &\approx \tfrac{1}{2\sigma^2} u_i^2 + \lambda|u_i|\,, \end{aligned}$$

for $u_i \gg \gamma$, and for $u_i \ll \gamma$ we obtain

$$D_\psi(u_i, z_i) \approx \tfrac{1}{2\sigma^2}(u_i - z_i)^2 + \lambda\left[\frac{u_i^2}{2\gamma} + \frac{z_i^2}{2\alpha} - \frac{-z_i u_i}{\gamma}\right] = \left(\frac{1}{2\sigma^2} + \frac{\lambda}{2\gamma}\right)(u_i - z_i)_2^2$$

In these two cases D_ψ boosts the effect of the shrinkage prior by promoting u_i values that are close to zero (explicitly via the penalty $\lambda|u_i|$, or by amplifying the constant of the quadratic loss from $1/2\sigma^2$ to $1/2\sigma^2 + \lambda/2\gamma$).

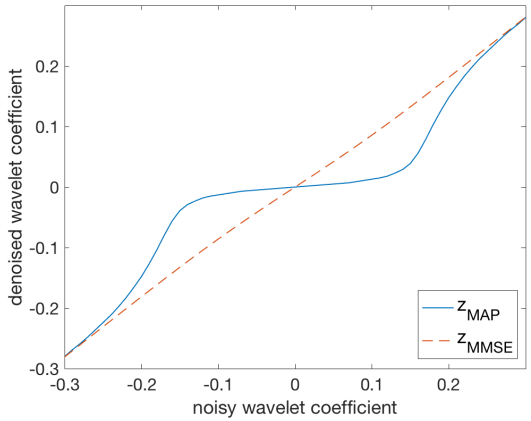
Example 1: denoising with wavelet shrinkage prior

Illustration with the `Flinstones` image ($\sigma = 0.08$, $\lambda = 12$ and $\gamma = 0.01$).

noisy image y (SNR 17.6dB)

$\hat{x}_{MAP}$ (SNR 19.8dB)

$\hat{x}_{MMSE}$ (SNR 17.7dB)

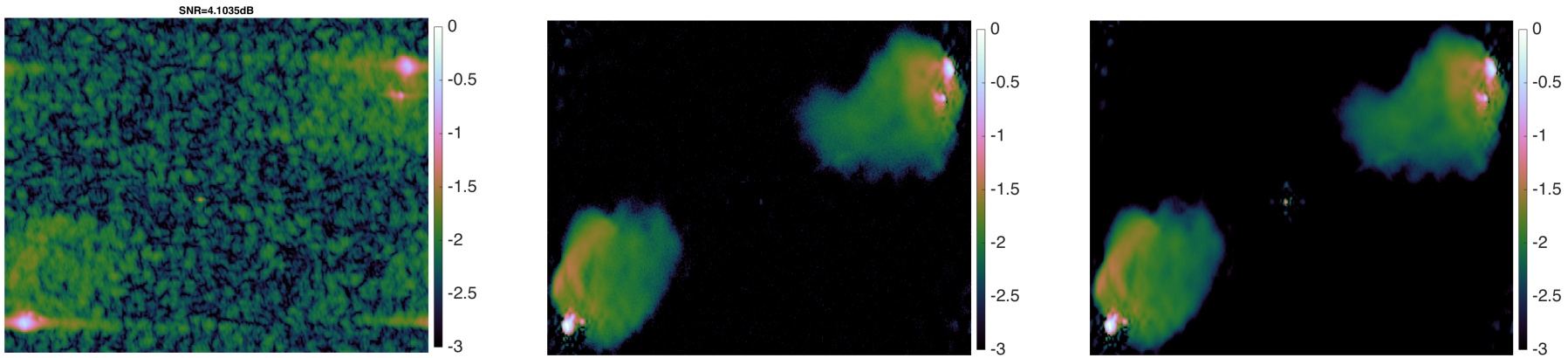


denoising functions for $\hat{z}_{MAP}$ and $\hat{z}_{MMSE}$

Illustrative example of a model where the action of the shrinkage prior acts predominantly via D_ψ (Note: setting $\gamma = 0$ leads to $\hat{x}_{MAP}$ with SNR 18.8dB).

Illustrative example: astronomical image reconstruction

Generalisation warning: shrinkage priors can also act predominantly via the model (not D_ψ), producing similar $\hat{x}_{MAP}$ and $\hat{x}_{MMSE}$ results; e.g.,



dirty image $\hat{x}_{penML}(y)$ $\hat{x}_{MAP}$ $\hat{x}_{MMSE}$

Radio-interferometric imaging of the `Cygnus A` galaxy **?**.

Outline

Problem statement

We consider the class of priors of the form

$$p(x|\theta) = \exp\{-\theta h(x)\}/C(\lambda)$$

where $h : \mathbb{R}^n \to [0, \infty]$ promotes expected properties of x, and $\theta \in \mathbb{R}^+$ is a "regularisation" parameter controlling the strength of the prior.

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

$$\hat{x}^{(\theta)}_{MAP} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, g_y(x) + \theta h(x) - \log C(\theta) - \log p(y),$$

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

Here we consider the infamous problem of not specifying the value of θ.

Hierarchical Bayesian treatment of unknown θ

Hierarchical Bayesian inference allows estimating x without specifying θ.

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

The extended model is

$$\begin{aligned} p(x, \theta|y) &= p(y|x)p(x|\theta)p(\theta)/p(y), \\ &\propto \frac{\exp\{-g_y(x) - \theta h(x) - \log p(\theta)\}}{C(\theta)}, \end{aligned} \tag{4}$$

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

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

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

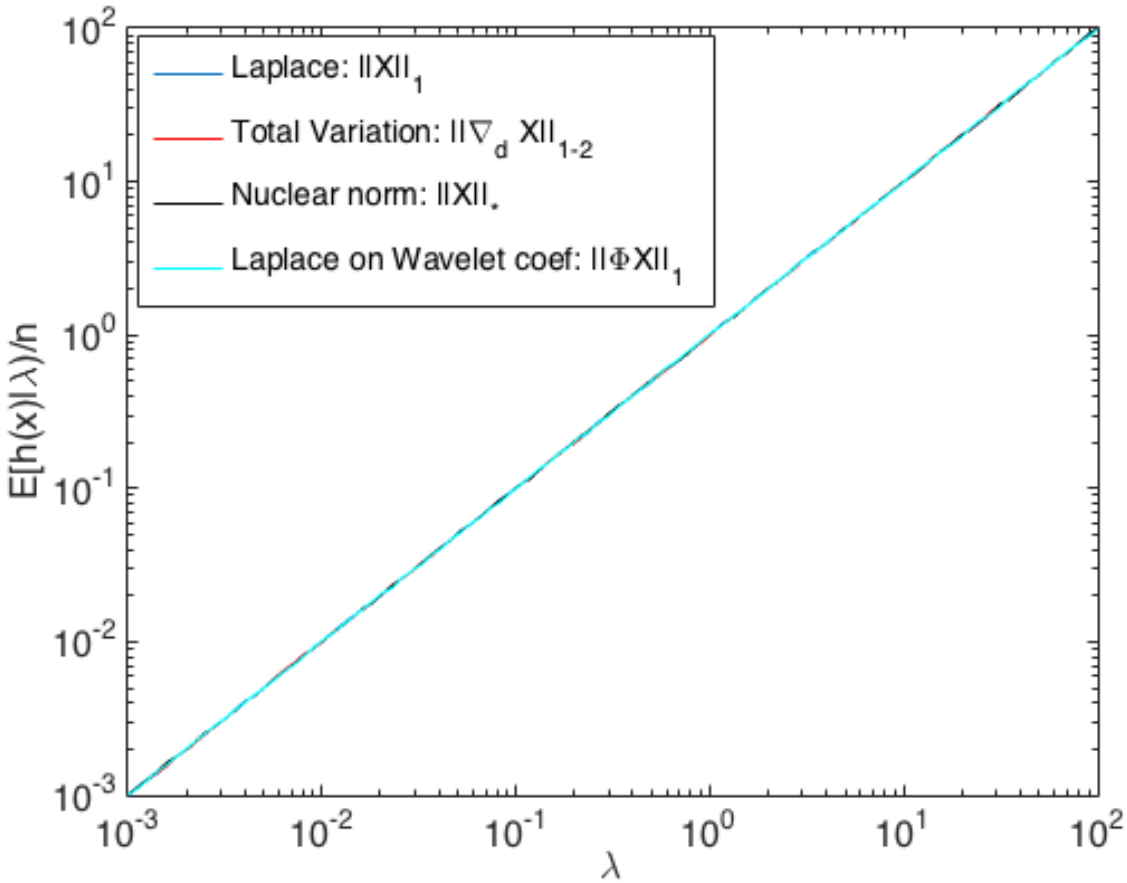


Figure : Monte Carlo approximations of $E[h(x)|\theta]$ for 4 widely used prior distributions and for $\theta \in [10^{-3}, 10^2]$. **Surprise:** they all coincide!

Main theoretical result

Definition 4.1 (**k**-homogeneity)

The regulariser h is a k-homogeneous function if $\exists k \in \mathbb{R}^+$ such that

$$h(\eta x) = \eta^k h(x), \quad \forall x \in \mathbb{R}^d, \forall \eta > 0. \tag{5}$$

Theorem 4.1 (Pereyra et al. (2015))

Suppose that h, the sufficient statistic of $p(x|\theta)$, is k-homogenous. Then the normalisation factor has the form

$$C(\theta) = D\theta^{-d/k},$$

with (generally intractable) constant $D = C(1)$ 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 $C(\lambda)$ enables (for example) *marginal MAP estimation* of x

$$\begin{aligned}\hat{x}^{\dagger}_{MAP} &= \operatorname*{argmax}_{x\in\mathbb{R}^d} \int_0^{\infty} p(x,\lambda|y)\mathrm{d}\lambda, \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d} g_y(x) + (d/k+\alpha)\log\{h(x)+\beta\},\end{aligned} \tag{6}$$

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

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

$$\begin{aligned}x^{(t)} &= \operatorname*{argmin}_{x\in\mathbb{R}^d} g_y(x) + \theta^{(t-1)}h(x), \\ \theta^{(t)} &= \frac{d/k+\alpha}{h(x^{(t)})+\beta}.\end{aligned} \tag{7}$$

which is also an *expectation-maximisation* algorithm.

Compressive sensing with ℓ_1-wavelet analysis prior

Recover an original image $x \in \mathbb{R}^d$ of size $n = 512 \times 512$ from a compressed and noisy measurement

$$y = \Phi x + w,$$

of size $p = d/2$, where $\Phi \in \mathbb{R}^{p \times d}$ is a compressive sensing random matrix and $w \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_p)$ is Gaussian noise with $\sigma^2 = 10$.

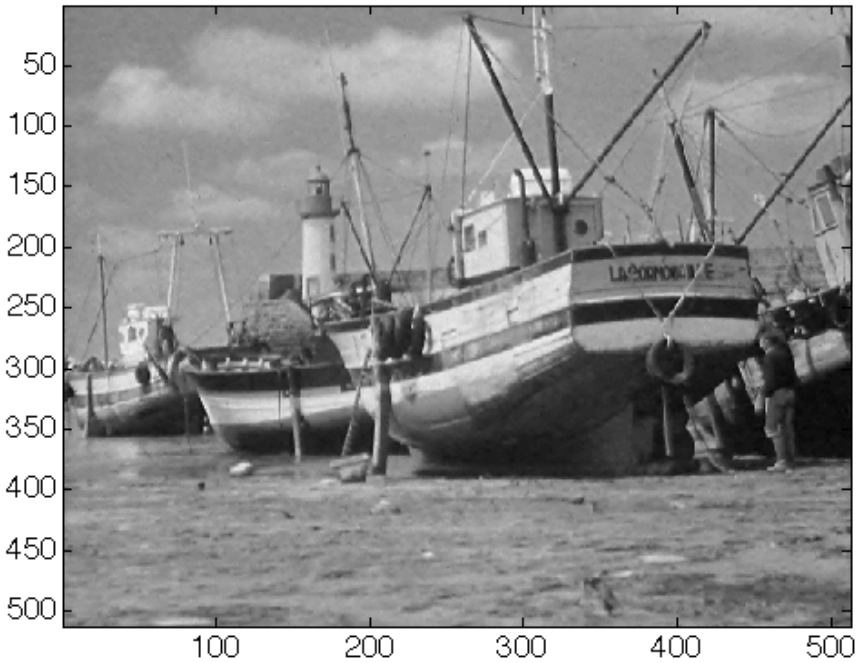
We use the *analysis* prior

$$p(x|\theta) = \exp\{-\theta \|\Psi x\|_1\}/C(\theta)$$

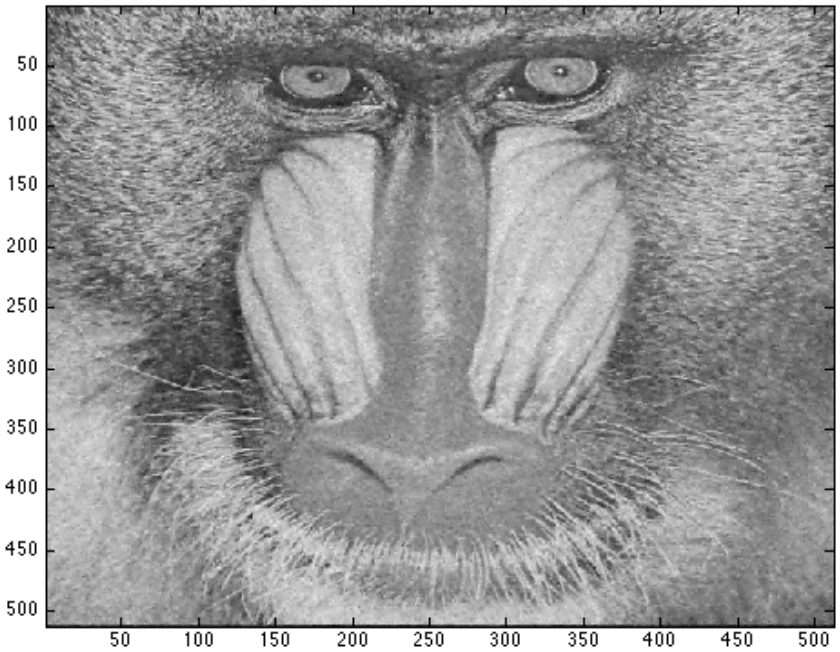
where Ψ is a *Daubechies 4* wavelet frame.

Note: $\|\Psi(x)\|_1$ is k-homogenous with $k = 1$.

Experiment with the Boat and Mandrill test images



(θ^* = 56.4, PSNR=33.4)



(θ^* = 2.04, PSNR=25.3)

Figure : Compressive sensing experiment with the Boat and Mandrill test images, with automatic selection of θ by marginalisation - see (7).

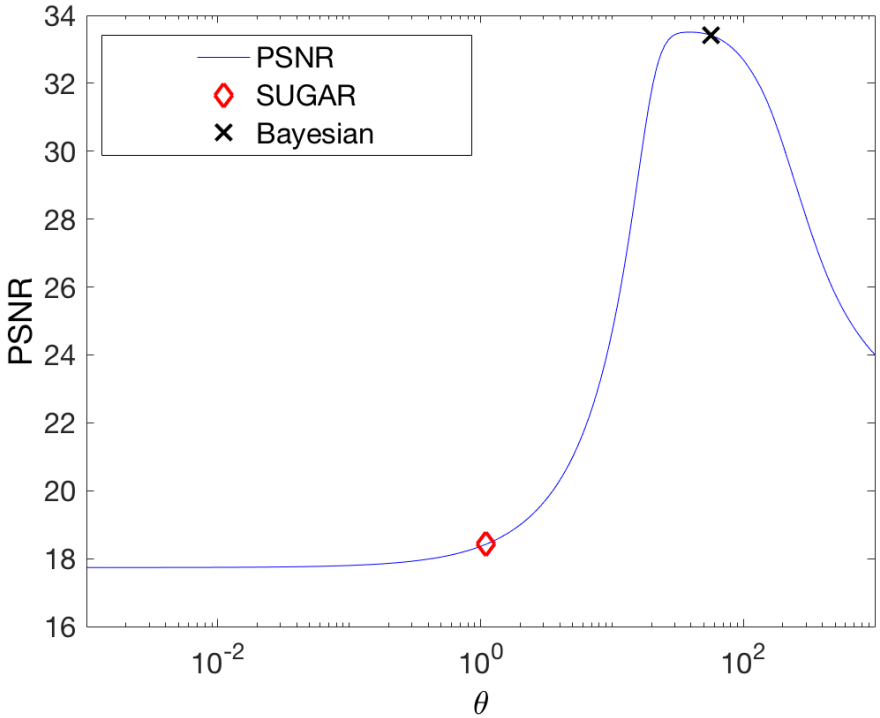
Comparison with SUGAR (Deledalle et al., 2014a) and with the MSE oracle.

	θ	PSNR	SSIM	time [sec]
Marginal MAP	56.4	33.4	0.96	299
SUGAR	1.10	18.4	0.55	1137
MSE Oracle	38.2	33.5	0.96	n/a
Least-squares	n/a	17.7	0.52	0.04

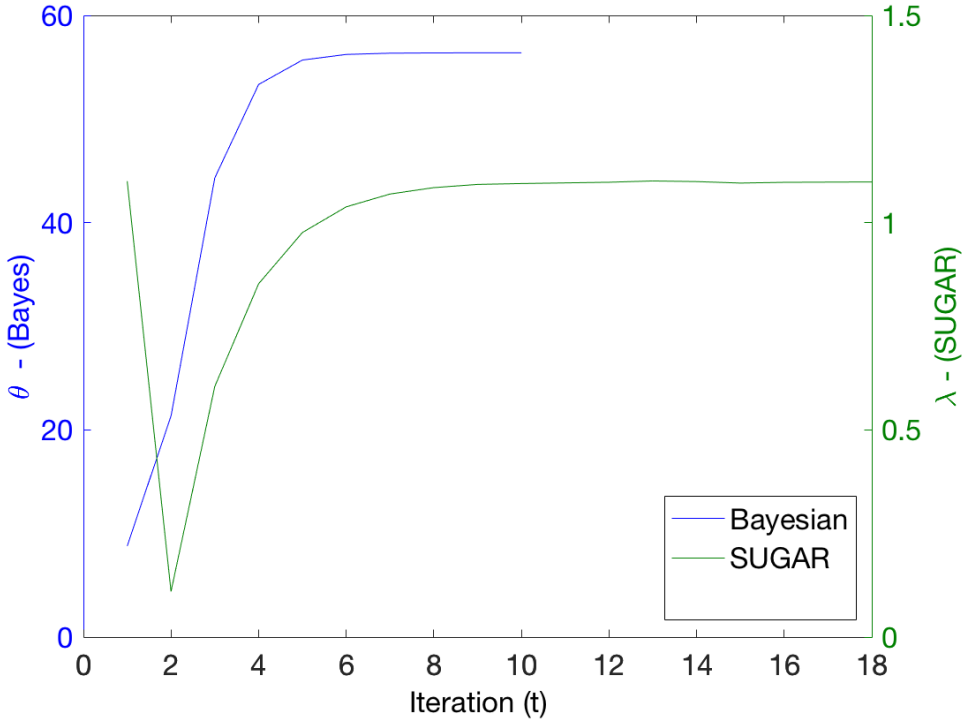
Table : `Boat`.

	θ	PSNR	SSIM	time [sec]
Marginal MAP	2.04	25.3	0.87	229
SUGAR	0.95	22.9	0.80	984
MSE Oracle	4.65	26.0	0.90	n/a
Least-squares	n/a	18.6	0.22	0.04

Table : `Mandrill`.



PSNR vs θ



Iterates $\theta^{(t)}$

Figure : [Left] Estimation PSNR as a function of θ. [Right] Evolution of the iterates $\theta^{(t)}$ for the Bayesian method (left axis) and for SUGAR (right axis).

Illustrative example - Image deblurring with TV prior

In a manner akin to Fernandez-Vidal and Pereyra (2018), we also apply the method to the Bayesian image deblurring model

$$p(x|y,\theta) \propto \exp\left(-\|y - Ax\|^2/2\sigma^2 - \theta\|\nabla_d x\|_{1-2}\right),$$

and compute $\hat{\theta} = \text{argmax}_{\theta\in\mathbb{R}^+}\, p(y|\theta)$.

We obtain the following results:

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	Avg. MSE	Avg. Time	Avg. MSE	Avg. Time	Avg. MSE	Avg. Time
θ^* (*Oracle*)	22.95 ± 3.10	–	21.05 ± 3.19	–	18.76 ± 3.19	–
Marginalization	24.67 ± 3.08	17.27	22.39 ± 3.07	6.31	19.44 ± 3.26	6.77
Empirical Bayes	23.24 ± 3.23	43.01	21.16 ± 3.24	41.50	18.90 ± 3.39	42.85
SUGAR	24.14 ±3.19	15.74	23.96 ± 3.26	20.87	23.94± 3.27	20.59

Comparison with the empirical Bayesian method (Fernandez-Vidal and Pereyra, 2018), the SUGAR method (Deledalle et al., 2014b), and an oracle that knows the optimal value of θ. Average values over 10 test images of size 512×512 pixels.

An exhaustive evaluation comparing different methods on a range of imaging problems will be reported soon.

Outline

Conclusion

- The challenges facing modern imaging sciences require a methodological paradigm shift to go beyond point estimation.
- In Part I we discussed how the Bayesian framework can support this paradigm shift, provided we significantly accelerate computations.
- In Part II we considered efficiency improvements by integrating modern stochastic and variational computation approaches.
- In Part III we explored methods based on convex optimisation and probability, and developed theory for MAP estimation.

Conclusion

In the next lecture...

We will explore Bayesian models and stochastic computation algorithms for problems that are significantly more difficult, and where deterministic approaches fail.

Thank you!

Bibliography:

Ay, N. and Amari, S.-I. (2015). A novel approach to canonical divergences within information geometry. *Entropy*, 17(12):7866.

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.

Deledalle, C., Vaiter, S., Peyré, G., and Fadili, J. (2014a). Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM J. Imaging Sci.*, 7(4):2448–2487.

Deledalle, C.-A., Vaiter, S., Fadili, J., and Peyré, G. (2014b). Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487.

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

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Pereyra, M. (2016). Maximum-a-posteriori estimation with bayesian confidence regions. *SIAM J. Imaging Sci.*, 6(3):1665–1688.

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*.

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.