Modes of posterior measure for Bayesian inverse problems with a class of non-Gaussian priors

Masoumeh Dashti

University of Sussex

UQ for inverse problems in complex systems

Isaac Newton Institute, April 12th 2018

Based on joint works with

S Agapiou (Cyprus), M Burger (Münster), T Helin (Helsinki)

Introduction

Let μ be a probability measure on Banach space X with

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = f(u), \quad (\mu_0 \text{ and } f \text{ are given})$$

Introduction

Let μ be a probability measure on Banach space X with

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = f(u), \quad (\mu_0 \text{ and } f \text{ are given})$$

- If $X = \mathbb{R}^n$, μ has a Lebesgue density

  $$\rho_\mu(u) = f(u)\rho_{\mu_0}(u)$$

  and modes of μ are maximisers of $\rho_\mu : \mathbb{R}^n \to \mathbb{R}$
- When X is a function space (μ does not have a Lebesgue density)

  how do we define and characterise the most likely functions?

Introduction

Let μ be a probability measure on Banach space X with

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = f(u), \quad (\mu_0 \text{ and } f \text{ are given})$$

- If $X = \mathbb{R}^n$, μ has a Lebesgue density

$$\rho_\mu(u) = f(u)\rho_{\mu_0}(u)$$

  and modes of μ are maximisers of $\rho_\mu : \mathbb{R}^n \to \mathbb{R}$

- When X is a function space (μ does not have a Lebesgue density)

  how do we define and characterise the most likely functions?

In the context of Bayesian approach to inverse problems

μ_0: prior measure $\quad$ μ: posterior measure and

modes of μ are called maximum a posterior (MAP) estimators

Outline

1. Weak and strong MAP estimators
2. Besov priors
3. Posterior consistency

Outline

1. Weak and strong MAP estimators
2. Besov priors
3. Posterior consistency

MAP estimates

$(X, \|\cdot\|)$ *a separable Banach space;* $\mu(X) = 1$

Let $B^\delta(z)$ be a ball of radius δ and centre z in X.

- Fix δ and find z^δ such that $B^\delta(z^\delta)$ has maximal probability
- Look at the 'limit' of $\{z^\delta\}_\delta$ as δ shrinks to zero

MAP estimates

$(X, \|\cdot\|)$ *a separable Banach space;* $\mu(X) = 1$

Let $B^\delta(z)$ be a ball of radius δ and centre z in X.

- Fix δ and find z^δ such that $B^\delta(z^\delta)$ has maximal probability
- Look at the 'limit' of $\{z^\delta\}_\delta$ as δ shrinks to zero

Definition (D., Law, Stuart, Voss 2013)

Let

$$M^\delta = \sup_{z \in X} \mu(B^\delta(z)).$$

Any point $\tilde{z} \in X$ *satisfying*

$$\lim_{\delta \to 0} \frac{M^\delta}{\mu(B^\delta(\tilde{z}))} = 1,$$

is a mode (MAP estimator) of μ.

Onsager-Machlup functional

Suppose $\exists$ dense subspace $F \subset X$ and $I : F \to \mathbb{R}$ such that for any $z_1, z_2 \in I$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(z_2))}{\mu(B_\epsilon(z_1))} = \exp(I(z_1) - I(z_2)),$$

Onsager-Machlup functional

Suppose $\exists$ dense subspace $F \subset X$ and $I : F \to \mathbb{R}$ such that for any $z_1, z_2 \in I$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(z_2))}{\mu(B_\epsilon(z_1))} = \exp(I(z_1) - I(z_2)),$$

- Example: $\mu \sim \mathcal{N}(0, \mathcal{C}_0)$ on a Hilbert space X

$$I(u) = \tfrac{1}{2}\|u\|_F^2 \text{ with } F = \mathcal{C}_0^{\frac{1}{2}} X \text{ and } \mu(F) = 0$$

Onsager-Machlup functional

Suppose $\exists$ dense subspace $F \subset X$ and $I : F \to \mathbb{R}$ such that for any $z_1, z_2 \in I$

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(z_2))}{\mu(B_\epsilon(z_1))} = \exp(I(z_1) - I(z_2)),$$

- Example: $\mu \sim \mathcal{N}(0, \mathcal{C}_0)$ on a Hilbert space X

$$I(u) = \frac{1}{2}\|u\|_F^2 \text{ with } F = \mathcal{C}_0^{\frac{1}{2}} X \text{ and } \mu(F) = 0$$

Do minimisers of I characterise MAP estimators?

- *MAP estimators are 'limits' of sequences in X, while Onsager-Machlup functional is finite only on $F \subset X$.*

Gaussian prior

Recall: μ satisfies $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \propto \exp(-\Phi(u))$,

Gaussian prior

Recall: μ satisfies $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \propto \exp(-\Phi(u))$,
Let $\mu_0 \sim \mathcal{N}(0, \mathcal{C}_0)$

For locally Lipschitz $\Phi : X \to \mathbb{R}_+$ (D., Law, Stuart, Voss 2013)

MAP estimators are characterised by the minimisers of I

$$I(u) = \Phi(u) + \frac{1}{2}\|u\|_F^2$$

(with F: Cameron-Martin space of μ_0)

Weak MAP estimates

Definition (T. Helin & M. Burger 2015)

Let E be a dense subspace of X. We call a point $\hat{u} \in X$, a *E-weak mode* of μ if

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(\hat{u} - h))}{\mu(B_\epsilon(\hat{u}))} \leq 1,$$

for all $h \in E$.

Weak MAP estimates

Definition (T. Helin & M. Burger 2015)

Let E be a dense subspace of X. We call a point $\hat{u} \in X$, a *E-weak mode* of μ if

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(\hat{u} - h))}{\mu(B_\epsilon(\hat{u}))} \leq 1,$$

for all $h \in E$.

T. Helin & M. Burger 2015: *weak MAPs coincide with minimisers of I for smooth enough convex measures*

H.C. Lie & T.J. Sullivan 2018: *for μ nonatomic Borel probability measure on X, weak and strong modes coincide.*

Differentiable measures

- μ is differentiable along *h* if

$$\nu_h(A) := d_h\mu(A) = \lim_{t\to 0} \frac{\mu(A+th)-\mu(A)}{t} < \infty \quad \text{for all } A \in \mathcal{B}(X)$$

- $D(\mu) = \{h \in X : \mu \text{ is Fomin diff'ble along } h\}$.

Example:

– Standard Gaussian on $\mathbb{R}$, $\mu(A+th) = \frac{1}{\sqrt{\pi}} \int_A \mathrm{e}^{-\frac{1}{2}|u+th|^2}\,\mathrm{d}u$

$$\nu_h(A) = \frac{1}{\sqrt{\pi}} \int_A (-hu)\,\mathrm{e}^{-\frac{1}{2}|u|^2}\,\mathrm{d}u, \quad D(\mu) = \mathbb{R}$$

– Gaussian measure on Banach space *X*,

$$\nu_h(A) = \int_A \langle -h, u\rangle_{D(\mu)}\,\mathrm{d}\mu(u), \quad D(\mu) = F \subset X$$

Differentiable measures

- μ is differentiable along h if

$$\nu_h(A) := d_h\mu(A) = \lim_{t\to 0} \frac{\mu(A+th) - \mu(A)}{t} < \infty \quad \text{for all } A \in \mathcal{B}(X)$$

- $D(\mu) = \{h \in X : \mu \text{ is Fomin diff'ble along } h\}$.

Example:

– Standard Gaussian on $\mathbb{R}$, $\mu(A+th) = \frac{1}{\sqrt{\pi}} \int_A \mathrm{e}^{-\frac{1}{2}|u+th|^2}\,\mathrm{d}u$

$$\nu_h(A) = \frac{1}{\sqrt{\pi}} \int_A (-hu)\,\mathrm{e}^{-\frac{1}{2}|u|^2}\,\mathrm{d}u, \quad D(\mu) = \mathbb{R}$$

– Gaussian measure on Banach space X,

$$\nu_h(A) = \int_A \langle -h, u\rangle_{D(\mu)}\,\mathrm{d}\mu(u), \quad D(\mu) = F \subset X$$

- Let

$$\beta_h(u) = \frac{\mathrm{d}\nu_h}{\mathrm{d}\mu}(u), \quad \textit{(logarithmic derivative of } \mu\textit{)}$$

Modes of convex differentiable measures

(*Helin* and *Burger* 2015):

- zero points of β_h coincide with weak modes of μ (*under certain conditions*)*
- For a linear forward operator, zero points of β_h are minimisers of Onsager-Machlup functional as well.

Modes of convex differentiable measures

(*Helin* and *Burger* 2015):

- zero points of β_h coincide with weak modes of μ (*under certain conditions*)*
- For a linear forward operator, zero points of β_h are minimisers of Onsager-Machlup functional as well.

$*$ conditions on μ:

μ is a convex measure

with β_h that has a continuous representative:

$$\exists E \subset D(\mu) : \beta_h|_X \text{ is continuous for any } h \in E$$

Outline

1. Weak and strong MAP estimators
2. Besov priors
3. Posterior consistency

Besov priors

Define μ_0 through Karhunen-Loève expansion of its draws:

$$u(x) = \sum_{j \in \mathbb{N}} \alpha_j^{-1} \, \xi_j \, \psi_j(x)$$

$\{\psi_j\}$ orthonormal basis in $L^2(\mathbb{T}^d)$,

ξ_j i.i.d random variables,

$\{1/\alpha_j\}$ decreasing sequence determining smoothness of u

Besov priors

Define μ_0 through Karhunen-Loève expansion of its draws:

$$u(x) = \sum_{j \in \mathbb{N}} \alpha_j^{-1} \, \xi_j \, \psi_j(x)$$

$\{\psi_j\}$ orthonormal basis in $L^2(\mathbb{T}^d)$,

ξ_j i.i.d random variables,

$\{1/\alpha_j\}$ decreasing sequence determining smoothness of u

Gaussian	Besov (Lassas, Saksman, Siltanen 09)
$\xi_j \sim c \, \exp(-\frac{1}{2}\|x\|^2)$	$\xi_j \sim c_q \, \exp(-\|x\|^q), \quad q \geq 1$
$\{\psi_j\}$ an orthonormal basis	$\{\psi_j\}$ orthonormal wavelet basis

$q = 1$ (Besov-1 prior) especially interesting

Besov priors

For B_1^s-Besov prior: draws of μ_0 can be expressed as

$$u(x) = \sum_{j \in \mathbb{N}} \alpha_j^{-1} \, \xi_j \, \psi_j(x), \qquad \xi_j \sim \frac{1}{2} \mathrm{e}^{-|x|}, \quad \alpha_j = j^{\frac{s}{d} - \frac{1}{2}}$$

$\mu_0(B_1^t(\mathbb{T}^d)) = 1$ for any $t < s - d$.

Besov priors

For B^s_1-Besov prior: draws of μ_0 can be expressed as

$$u(x) = \sum_{j\in\mathbb{N}} \alpha_j^{-1}\,\xi_j\,\psi_j(x), \qquad \xi_j \sim \frac{1}{2}\mathrm{e}^{-|x|}, \quad \alpha_j = j^{\frac{s}{d}-\frac{1}{2}}$$

$\mu_0(B^t_1(\mathbb{T}^d)) = 1$ for any $t < s - d$.

- useful when we expect u to be smooth with a few local irregularities – promotes sparsity
- We have

$$\beta_h(u) = -\sum_{j=1}^{\infty} \alpha_j\,\mathrm{sign}(u_j)\,h_j,$$

  which does not have a continuous representative.

Recall: $\hat{u} \in X$ is a *E-weak mode* of μ if

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(\hat{u} - h))}{\mu(B_\epsilon(\hat{u}))} = \lim_{\epsilon \to 0} \frac{\mu^h(B_\epsilon(\hat{u}))}{\mu(B_\epsilon(\hat{u}))} \leq 1 \quad \forall h \in E$$

with $\mu^h(\cdot) = \mu(\cdot - h)$.

Recall: $\hat{u} \in X$ is a *E-weak mode* of μ if

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(\hat{u} - h))}{\mu(B_\epsilon(\hat{u}))} = \lim_{\epsilon \to 0} \frac{\mu^h(B_\epsilon(\hat{u}))}{\mu(B_\epsilon(\hat{u}))} \leq 1 \quad \forall h \in E$$

with $\mu^h(\cdot) = \mu(\cdot - h)$.

If Radon-Nikodym derivative $R_h = \mathrm{d}\mu^h / \mathrm{d}\mu$ is continuous over X for $h \in E$ then

$$\lim_{\epsilon \to 0} \frac{\mu^h(B_\epsilon(u))}{\mu(B_\epsilon(u))} = R_h(u) \quad \text{in } X$$

(Note that $E \subset F$)

- Continuity of R_h is enough to show equivalence of weak MAPs with minimisers of Onsager-Machlup functional.

(Agapiou, Burger, D, Helin 2017)

For μ_0 B^s_1-Besov prior

- For any $h \in B_2^{s-\frac{d}{2}}(\mathbb{T}^d)$

$$R^0_h(u) := \frac{d\mu_0^h}{d\mu_0}(u) = \lim_{N\to\infty} \exp \sum_{j=1}^{N} \left(-\alpha_j |h_j - u_j| + \alpha_j |u_j| \right)$$

- If $h \in E = B^r_1$ for $r > s$,

$R^0_h : B^t_1 \to \mathbb{R}$ is continuous for any $t < s - d$.

For μ_0 B^s_1-Besov prior

- For any $h \in B_2^{s-\frac{d}{2}}(\mathbb{T}^d)$

$$R^0_h(u) := \frac{d\mu_0^h}{d\mu_0}(u) = \lim_{N\to\infty} \exp \sum_{j=1}^{N} \big(-\alpha_j |h_j - u_j| + \alpha_j |u_j| \big)$$

- If $h \in E = B^r_1$ for $r > s$,

$R^0_h : B^t_1 \to \mathbb{R}$ is continuous for any $t < s - d$.

- Onsager-Machlup functional: For any $z_1, z_2 \in B^s_1$:

$$\lim_{\epsilon\to 0} \frac{\mu_0(B_\epsilon(z_1))}{\mu_0(B_\epsilon(z_2))} = \exp(-\|z_1\|_{B^s_1} + \|z_2\|_{B^s_1})$$

And hence for μ, for any $u \in B^s_1$,

$$I(u) = \Phi(u) + \|u\|_{B^s_1}$$

MAP estimators for Besov-1 prior

Theorem

Let μ satisfy $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \propto \exp(-\Phi(u))$ with μ_0 a B_1^s-Besov measure and $\Phi : B_1^t \to \mathbb{R}_+$ locally Lipschitz

- *$u \in B_1^s$ minimises I if and only if it is a MAP estimator of μ.*
- *For all δ $\exists z^\delta \in X$ s.t. $z^\delta = \arg\max_{z \in X} \mu(B^\delta(z))$*

  $\exists \bar{z} \in F$ and a subseq of $\{z^\delta\}$ which converges to $\bar{z}$ strongly in X

  $\bar{z}$ is a MAP estimate.

Outline

Weak posterior consistency

As data we have the following n random vectors

$$y_j = \mathcal{G}(u^\dagger) + \eta_j, \quad j = 1, \ldots, n$$

with $y_j \in \mathbb{R}^K$,

$$\mathcal{G} : X \to \mathbb{R}^K, \quad \text{and} \quad \eta_j \sim \mathcal{N}(0, \mathcal{C}_1), \quad \text{i.i.d.}$$

Weak posterior consistency

As data we have the following n random vectors

$$y_j = \mathcal{G}(u^\dagger) + \eta_j, \quad j = 1, \ldots, n$$

with $y_j \in \mathbb{R}^K$,

$$\mathcal{G} : X \to \mathbb{R}^K, \quad \text{and} \quad \eta_j \sim \mathcal{N}(0, \mathcal{C}_1), \quad \text{i.i.d.}$$

Given B_1^s-Besov prior μ_0, we have

$$\frac{\mathrm{d}\mu^{y_1,\ldots,y_n}}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}\sum_{j=1}^{n} \left|y_j - \mathcal{G}(u)\right|_{\mathcal{C}_1}^2\right).$$

with MAP estimates

$$u_n := \underset{B_1^s}{\operatorname{argmin}} \|u\|_{B_1^s} + \frac{1}{2}\sum_{j=1}^{n} |y_j - \mathcal{G}(u)|_{\mathcal{C}_1}^2.$$

Theorem. *Assume that*

$\mathcal{G}\colon X \to \mathbb{R}^K$ *is locally Lipschitz and*

$u^\dagger \in B_1^s$.

Then

- $\mathcal{G}(u_n) \to \mathcal{G}(u^\dagger)$ *in probability.*
- *If* $\mathcal{G}$ *is injective* $\|u_n - u^\dagger\|_X \to 0$ *in probability.*

  Otherwise, $\exists u^* \in B_1^s$ *and a subseq of* $\{u_n\}_{n\in\mathbb{N}}$ *such that* $\|u_n - u^\dagger\|_X \to 0$ *in probability. For any such* u^*, $\mathcal{G}(u^*) = \mathcal{G}(u^\dagger)$.

Final remarks

- *Posterior consistency* with contraction rates

  for Bayesian direct problem with Besov prior

  forthcoming work Agapiou, D, Helin

  using Ghosal & van der Vaart 2007

  van der Vaart & van Zanten 2008

  $\to$ linear inverse problem Knapik & Salomond 2014

- weak MAPs for *non-convex prior*

  local MAPs for locally convex priors (Cauchy)