MAP Estimators for Laplace-type priors in Bayesian Inverse Problems

Sergios Agapiou

Department of Mathematics and Statistics, University of Cyprus

Algorithms and Computationally Intensive Inference seminar
23 June 2017, Warwick

University of Cyprus

Outline

Outline

Inverse Problem

$$y = \mathcal{G}(u) + \xi$$

- $u \in X$ unknown function, X separable Banach space
- $y \in \mathbb{R}^J$ finite-dim observation
- $\xi \sim N(0, \Sigma)$, $\Sigma \in \mathbb{R}^{J \times J}$ positive definite, observational noise
- $\mathcal{G} : X \to \mathbb{R}^J$ possibly nonlinear forward operator, locally Lipschitz

Bayesian Formulation

- Prior: $u \sim \mu_0$
- Likelihood: $y|u \sim N(\mathcal{G}(u), \Sigma)$
- Posterior: $u|y \sim \mu^y$

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\big(-\Phi(u;y)\big),$$

$$\Phi(u;y) = \frac{1}{2}\Big|\Sigma^{-\frac{1}{2}}(y - \mathcal{G}(u))\Big|^2.$$

M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, Handbook of Uncertainty Quantification, 2015.

Edge-preserving and Sparsity-promoting Priors

Blocky structure and **sparsity** in an appropriate expansion

- 1-Besov priors
  - M. Lassas, E. Saksman and S. Siltanen, *Discretization-invariant Bayesian inversion and Besov space priors*, 2009
  - M. Dashti, S. Harris and A. Stuart, *Besov priors for Bayesian inverse problems*, 2013
- Infinitely divisible and heavy tailed priors, e.g. Cauchy priors
  - T. Sullivan, *Well-posed Bayesian inverse problems and heavy-tailed stable Banach space priors*, 2016
  - B. Hosseini, *Well-posed Bayesian inverse problems with infinitely-divisible and heavy-tailed prior measures*, 2017

Need for MAP

- Probing the posterior can be difficult for such priors
- Sampling via MCMC can get prohibitively expensive if $\mathcal{G}$ difficult to evaluate
- Function space MCMC algorithms not yet fully developed for such priors
- Maximum a posteriori (MAP) estimators
  - understood as modes of posterior μ^y
  - Attractive because they require the solution of a single optimization problem
  - **We study MAP estimates in this non-parametric setting with 1-Besov priors**

http://www.sergiosagapiou.com/

S. Agapiou, M. Burger, M. Dashti and T. Helin, *Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems*, arXiv:1705.03286

Build on

M. Dashti, K. Law, A. Stuart and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 2013

MAP for Gaussian priors

T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems, 2015

wMAP theory using differential calculus of measures, does not cover 1-Besov priors, basis for tackling Cauchy

Outline

Finite-dimensional Intuition

- Assume $X = \mathbb{R}^N$ and prior has Lebesgue density

$$\pi(u) \propto \exp(-W(u))$$

- Posterior Lebesgue density

$$\pi^y(u) \propto \exp(-I(u; y)),$$

where

$$I(u; y) = \Phi(u; y) + W(u).$$

- MAP estimators maximize posterior density, i.e. minimize I
- MAP estimators correspond to Tikhonov regularized estimate
  - fidelity term $\longleftrightarrow$ likelihood
  - penalty term $\longleftrightarrow$ prior

Modes in Infinite-dimensions

- In infinite dimensions no uniform measure. Modes of measure μ on function space X?

> - consider $\mu(B_\epsilon(u))$ for $u \in X$
> - send $\epsilon \to 0$
> - $\hat{u}$ mode of μ if it maximizes the limiting small ball probabilities in a specific sense

- strong mode: max probability among all centres in X, *Dashti et al '13*
- weak mode: max probability among all shifts of the ball by elements in a dense subspace $E \subset X$, *Helin and Burger '15*

MAP and wMAP

Definition (Dashti et al '13)

Let $M^\epsilon = \sup_{u\in X} \mu(B_\epsilon(u))$. $\hat{u} \in X$ is a mode of μ, if

$$\lim_{\epsilon\to 0} \frac{\mu(B_\epsilon(\hat{u}))}{M^\epsilon} = 1.$$

Definition (Helin and Burger '15)

Let E dense subspace of X. $\hat{u} \in X$ weak mode of μ if

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

A MAP (resp. wMAP) estimate is a mode (resp. weak mode) of μ^y.

Remarks

- Weak mode allows flexibility of choosing E.
- Any strong mode is a weak mode: e.g. choose $E = X$ and bound $\mu(B_\epsilon(\hat{u} - h)) \leq M^\epsilon$.
- Weak modes interesting when small ball probabilities available only in some subspace of translations h, E. Typically E has measure zero.
- For convex μ set of modes is convex for both notions (typically fails for posterior).
- Can define local strong and weak modes (hence MAP estimates), which maximize small ball probabilities in the respective senses only locally.

AIM: establish variational characterization of wMAP and MAP in ∞-dim BIP.

Onsager-Machlup Functional

- Suppose can find $I : F \to \mathbb{R}$ s.t.

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

- F dense subspace of X.
- Fix $z_1 \in F$. A $z_2 \in F$ minimizing I is a potential mode of μ.
- Such I is called the (generalized) Onsager-Machlup functional of μ.

Onsager-Machlup Functional

- For $X = \mathbb{R}^N$, posterior density $\pi^y(u) \propto \exp(-I)$, by Lebesgue differentiation theorem

$$\lim_{\epsilon \to 0} \frac{\mu^y(B_\epsilon(z_2))}{\mu^y(B_\epsilon(z_1))} = \lim_{\epsilon \to 0} \frac{\int_{B_\epsilon(z_2)} \pi^y(du)}{\int_{B_\epsilon(z_1)} \pi^y(du)} = \frac{\pi^y(z_2)}{\pi^y(z_1)} = \exp(I(z_1) - I(z_2))$$

- $I = \Phi + W$ is the O-M functional of μ^y.

Strategy: crucial first step

- For μ measure, define $\mu_h(\cdot) = \mu(\cdot - h)$.
- For h such that $\mu_h \ll \mu$, denote

$$R_h^\mu(u) = \frac{d\mu_h}{d\mu}(u).$$

Lemma (Helin and Burger '15)

$$\lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(u-h))}{\mu(B_\epsilon(u))} = R_h^\mu(u),$$

for all h such that R_h^μ is continuous for $u \in X$.

Proof.

$$\inf_{v \in B_\epsilon(u)} R_h^\mu(v) \leq \frac{\mu_h(B_\epsilon(u))}{\mu(B_\epsilon(u))} = \frac{\int_{B_\epsilon(u)} R_h^\mu(z)\mu(dz)}{\mu(B_\epsilon(u))} \leq \sup_{v \in B_\epsilon(u)} R_h^\mu(v),$$

for all $\epsilon > 0$ and $u \in X$. Take $\epsilon \to 0$ and use ctty. □

Strategy

- Choose E sufficiently regular s.t. $R_h^{\mu_0}$ continuous in $u \in X$, for all $h \in E$.
- Straightforward to determine O-M functional I_0 of μ_0, defined on F containing E.
- Straightforward to determine O-M functional I of $\mu^y \propto e^{-\Phi}\mu_0$, defined on F.
- Straightforward to establish equivalence of (E-)wMAP and minimizers of I.
- Considerably harder to establish equivalence of MAP and minimizers of I, due to smallness of $F \subset X$ wrt prior.

Strategy: establishing continuity of R_h^μ

- In HB15 ctty of R_h^μ established via differential calculus of measures assuming ctty of log-derivative β_h^μ for h in a sufficiently smooth space E.
  - For ν probability measure on $\mathbb{R}$ with differentiable pdf π, log-derivative is $\beta^\nu = \frac{\pi'}{\pi}$.
  - For product measures $\mu = \otimes_{\ell=1}^\infty \mu_\ell$, log-derivative in direction $h = \{h_\ell\}$ is

$$\beta_h^\mu = \sum_{\ell=1}^\infty \beta_{h_\ell}^{\mu_\ell} = \sum_{\ell=1}^\infty h_\ell \frac{\pi'_\ell}{\pi_\ell}$$

  - Works for eg p-Besov priors with $1 < p \leq 2$ and Cauchy priors but not for 1-Besov prior whose log-derivative is inherently discontinuous.

- For product measures may be possible to get an explicit expression for R_h^μ via Kakutani-Hellinger theory and study its ctty analytically (works for 1-Besov priors)

Outline

Periodic Besov Spaces

- Let $\{\psi_\ell\}_{\ell=1}^{\infty}$ orthonormal wavelet basis for $L^2(\mathbb{T})$. Define $f : \mathbb{T} \to \mathbb{R}$ by

$$f(x) = \sum_{\ell=1}^{\infty} c_\ell \psi_\ell(x).$$

- $f \in B_p^s(\mathbb{T})$ iff

$$\|f\|_{B_p^s(\mathbb{T})} = \left(\sum_{\ell=1}^{\infty} \ell^{p(s+\frac{1}{2})-1} |c_\ell|^p \right)^{\frac{1}{p}} < \infty.$$

- p integrability, s smoothness parameter.
- For $p = 2$, Sobolev spaces of functions with s square integrable derivatives

$$\|f\|_{B_2^s(\mathbb{T})} = \left(\sum_{\ell=1}^{\infty} \ell^{2s} |c_\ell|^2 \right)^{\frac{1}{2}}.$$

- For $p = 1$

$$\|f\|_{B_1^s(\mathbb{T})} = \sum_{\ell=1}^{\infty} \ell^{s-\frac{1}{2}} |c_\ell|.$$

1-Besov Priors

Definition (Lassas et al '09)

Let $X_\ell \overset{iid}{\sim} \frac{1}{2}\exp(-|x|)$ and $\alpha_\ell = \ell^{s-\frac{1}{2}}$. The random function

$$U(x) = \sum_{\ell=1}^{\infty} \alpha_\ell^{-1} X_\ell \psi_\ell(x), \quad x \in \mathbb{T},$$

is said to be distributed according to a B_1^s-Besov prior, λ.

- Formally, U has density $\pi(u) \propto \exp(-\|u\|_{B_1^s})$ since $\alpha_\ell^{-1} X_\ell \sim \frac{\alpha_\ell}{2}\exp(-\alpha_\ell|x|)$.
- $\|U\|_{B_1^t} < \infty$, almost surely for all $t < s - 1$ (LSS09).
- $X = B_1^t$, $t < s - 1$.
- β_h^λ discontinuous, since derivative of $\frac{1}{2}\exp(-|x|)$ has a jump at the origin.

1-Besov Priors, MAP and wMAP

- Consider BIP with $\mu_0 = \lambda$
$$\mu^y(du) \propto \exp(-\Phi(u;y))\lambda(du).$$
- Let

$$I(u;y) = \Phi(u;y) + \|u\|_{B_1^s}, \quad I : B_1^s \to [0,\infty).$$

- Finite-dim intuition suggests that minimizers of I are maximizers of posterior since formally
$$"\mu^y(du) \propto \exp(-\Phi(u;y) - \|u\|_{B_1^s})du".$$

We make this rigorous.

Kakutani-Hellinger Theory

- For μ, ν measures both absolutely continuous wrt ζ, define Hellinger integral

$$H(\mu, \nu) = \int \sqrt{\frac{d\mu}{d\zeta}\frac{d\nu}{d\zeta}} d\zeta, \qquad H(\mu, \nu) \in [0, 1].$$

- If $H(\mu, \nu) = 0$ then μ, ν singular.
- In general if $H(\mu, \nu) > 0$ then μ, ν not necessarily equivalent, but this happens for products of equivalent measures.

Kakutani-Hellinger Theory

If $\mu = \otimes_{\ell=1}^{\infty} \mu_\ell$, $\nu = \otimes_{\ell=1}^{\infty} \nu_\ell$, then

$$H(\mu, \nu) = \prod_{\ell=1}^{\infty} H(\mu_\ell, \nu_\ell).$$

Theorem (Kakutani)

Let μ, ν product measures, where μ_ℓ, ν_ℓ equivalent for all $\ell \in \mathbb{N}$. Then μ and ν equivalent iff $H(\mu, \nu) > 0$, and if equivalent

$$\frac{d\mu}{d\nu}(u) = \lim_{N\to\infty} \prod_{\ell=1}^{N} \frac{d\mu_\ell}{d\nu_\ell}(u_\ell), \quad \text{in } L^1(\mathbb{R}^\infty, \mu).$$

G. Da Prato, *An Introduction to Infinite-Dimensional Analysis*, Springer, 2006.

1-Besov Priors, R_h^λ

Lemma (A., Burger, Dashti and Helin '17)

- We have $\lambda_h \sim \lambda$ if and only if $h \in B_2^{s-\frac{1}{2}}$.
- For $h \in B_2^{s-\frac{1}{2}}$

$$\frac{d\lambda_h}{d\lambda}(u) = \lim_{N\to\infty} \exp \sum_{\ell=1}^{N} (-\alpha_\ell |h_\ell - u_\ell| + \alpha_\ell |u_\ell|).$$

- For $h \in B_1^r$, $r > s$, the limit on rhs is continuous in $u \in X = B_1^t$, $t < s - 1$.

1-Besov Priors, R_h^λ

Proof.

- By Kakutani theorem suffices to compute Hellinger integral and check its positivity.
- Kakutani theorem also gives that

$$\frac{d\lambda_h}{d\lambda}(u) = \lim_{N\to\infty} \prod_{\ell=1}^{N} \frac{d\lambda_{h,\ell}}{d\lambda_\ell}(u) = \lim_{N\to\infty} \prod_{\ell=1}^{N} \frac{e^{-\alpha_\ell|h_\ell - u_\ell|}}{e^{-\alpha_\ell|u_\ell|}}.$$

- For ctty, technical explicit proof showing that $|R_h^\mu(u) - R_h^\mu(v)| \to 0$ as $\|u-v\|_{B_1^t} \to 0$ by examining all combinations of signs.

□

1-Besov priors, Small Ball Probabilities

Corollary (A., Burger, Dashti and Helin '17)

$$\lim_{\epsilon \to 0} \frac{\lambda(B_\epsilon(u-h))}{\lambda(B_\epsilon(u))} = \exp \sum_{\ell=1}^{\infty} (-\alpha_\ell |h_\ell - u_\ell| + \alpha_\ell |u_\ell|),$$

for $h \in B_1^r$, $r > s$.

- Choose $E = B_1^r$, $r > s$ in wMAP definition (E dense in X).
- To get Onsager-Machlup functional of λ need a bit more work.

1-Besov priors, Small Ball Probabilities

Corollary (A., Burger, Dashti and Helin '17)

$$\lim_{\epsilon \to 0} \frac{\lambda(B_\epsilon(u-h))}{\lambda(B_\epsilon(u))} = \exp \sum_{\ell=1}^{\infty} (-\alpha_\ell |h_\ell - u_\ell| + \alpha_\ell |u_\ell|),$$

for $h \in B_1^r, r > s$.

- Choose $E = B_1^r$, $r > s$ in wMAP definition (E dense in X).
- To get Onsager-Machlup functional of λ need a bit more work.
- For $u = 0$

$$\lim_{\epsilon \to 0} \frac{\lambda_h(B_\epsilon(0))}{\lambda(B_\epsilon(0))} = \exp \sum_{\ell=1}^{\infty} (-\alpha_\ell |h_\ell|) = e^{-\|h\|_{B_1^s}},$$

for $h \in B_1^r, r > s$.

1-Besov priors, Onsager-Machlup Functional of λ

Theorem (A., Burger, Dashti and Helin '17)

For $h \in B_1^s$,

$$\lim_{\epsilon \to 0} \frac{\lambda_h(B_\epsilon(0))}{\lambda(B_\epsilon(0))} = e^{-\|h\|_{B_1^s}}.$$

Proof.

Take $h^j \subset B_1^{s+1}$ s.t. $h^j \to h$ in B_1^s. □

1-Besov priors, Onsager-Machlup Functional of λ

Theorem (A., Burger, Dashti and Helin '17)

For $h \in B_1^s$,

$$\lim_{\epsilon \to 0} \frac{\lambda_h(B_\epsilon(0))}{\lambda(B_\epsilon(0))} = e^{-\|h\|_{B_1^s}}.$$

Proof.

Take $h^j \subset B_1^{s+1}$ s.t. $h^j \to h$ in B_1^s. □

- For $z_1, z_2 \in B_1^s$

$$\lim_{\epsilon \to 0} \frac{\lambda(B_\epsilon(z_2))}{\lambda(B_\epsilon(z_1))} = \lim_{\epsilon \to 0} \frac{\lambda(B_\epsilon(z_2))}{\lambda(B_\epsilon(0))} \frac{\lambda(B_\epsilon(0))}{\lambda(B_\epsilon(z_1))} = \exp(\|z_1\|_{B_1^s} - \|z_2\|_{B_1^s}).$$

- The Onsager-Machlup functional of λ is $I_0 : B_1^s \to [0, \infty)$, $I_0(z) = \|z\|_{B_1^s}$.

1-Besov priors, Onsager-Machlup Functional of μ^y

- Recall

$$I(u;y) = \Phi(u;y) + \|u\|_{B_1^s}, \quad I : B_1^s \to [0,\infty).$$

- Using local Lipschitz ctty of Φ, can show:

Proposition (A., Burger, Dashti and Helin '17)

I is the Onsager-Machlup functional for μ^y:

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

for all $z_1, z_2 \in B_1^s$.

- Calculus of variations shows that I has a minimizer $\hat{u} \in B_1^s$ (ABDH17)

1-Besov Priors, MAP and wMAP

Theorem (A., Burger, Dashti and Helin '17)

Both wMAP and MAP estimates of the posterior μ^y are identified with the minimizers of I.

Proof

- wMAP:
  - Assume $u_{min} \in B_1^s$ minimizer of I.

    For any $h \in B_1^s$, $r > s$ let $z_1 = u_{min} \in B_1^s$, $z_2 = u_{min} - h \in B_1^s$

$$\lim_{\epsilon \to 0} \frac{\mu^y(B_\epsilon(u_{min} - h))}{\mu^y(B_\epsilon(u_{min}))} \stackrel{prop}{=} \exp(I(u_{min}) - I(u_{min} - h)) \leq 1.$$

    Hence u_{min} is wMAP for $E = B_1^r$, $r > s$.

1-Besov Priors, MAP and wMAP

Proof.

- Assume $\hat{u}$ wMAP for $E = B_1^r$.

  Exclude $\hat{u} \in B_1^t \setminus B_1^s$ by showing contradiction to ctty of Φ.

  Thus $\hat{u} \in B_1^s$ and by proposition

$$1 \geq \frac{\mu^y(B_\epsilon(\hat{u}-h))}{\mu^y(B_\epsilon(\hat{u}))} \overset{prop}{=} \exp(I(\hat{u}) - I(\hat{u}-h)), \ \forall h \in B_1^r, r > s,$$

hence

$$I(\hat{u}) \geq I(\hat{u}-h), \ \forall h \in B_1^r, r > s.$$

By ctty of I in B_1^s and density of B_1^r in B_1^s, we get that $\hat{u}$ minimizer of I.

□

1-Besov Priors, MAP and wMAP

Proof.

- Assume $\hat{u}$ wMAP for $E = B_1^r$.

  Exclude $\hat{u} \in B_1^t \setminus B_1^s$ by showing contradiction to ctty of Φ.

  Thus $\hat{u} \in B_1^s$ and by proposition

$$1 \geq \frac{\mu^y(B_\epsilon(\hat{u} - h))}{\mu^y(B_\epsilon(\hat{u}))} \overset{prop}{=} \exp(I(\hat{u}) - I(\hat{u} - h)), \ \forall h \in B_1^r, r > s,$$

  hence

$$I(\hat{u}) \geq I(\hat{u} - h), \ \forall h \in B_1^r, r > s.$$

  By ctty of I in B_1^s and density of B_1^r in B_1^s, we get that $\hat{u}$ minimizer of I.

- MAP: Very technical, requires many new small ball probability estimates for centres in B_1^t.

□

1-Besov Priors, Consistency of MAP estimates

- Consider frequentist setup

$$y_j = \mathcal{G}(u^\dagger) + \xi_j,$$

  for fixed underlying $u^\dagger \in X$ and $\xi_j \overset{i.i.d.}{\sim} N(0, \Sigma)$.

- Sequence of posteriors

$$\frac{d\mu^{y_1,\dots,y_n}}{d\lambda}(u) \propto \exp\Big(-\frac{1}{2}\sum_{j=1}^{n} |\Sigma^{-\frac{1}{2}}(y_j - \mathcal{G}(u))|^2\Big)$$

- Previous result shows that MAP (and wMAP) estimates coincide with minimizers of

$$I_n(u) = \|u\|_{B_1^s} + \frac{1}{2}\sum_{j=1}^{n} |\Sigma^{-\frac{1}{2}}(y_j - \mathcal{G}(u))|^2.$$

1-Besov Priors, Consistency of MAP estimates

- Let $\{u_n\}$ be a sequence of MAP estimates corresponding to $\mu^{y_1,\dots,y_n}$.
- We investigate whether as $n \to \infty$, $\{u_n\}$ recovers $u^\dagger$ in some sense.
- Cannot expect to recover $u^\dagger$ fully, unless e.g. $\mathcal{G}$ is injective.

Theorem

Assume $u^\dagger \in B_1^s$. Then there exists $u^\star \in B_1^s$ and a subsequence of $\{u_n\}$ such that $u_n \to u^\star$ in $B_1^{\tilde{s}}$ a.s., for any $\tilde{s} < s$. For any such $u^\star$ we have $\mathcal{G}(u^\star) = \mathcal{G}(u^\dagger)$.

- Under the milder assumption $u^\dagger \in B_1^t$, only get a.s. convergence of $\mathcal{G}(u_n) \to \mathcal{G}(u^\dagger)$.

Outline

Conclusion and Future Work

- Wealth of new function-space priors, giving rise to many interesting questions.
- Discussed MAP estimates in this context, have a complete picture for Gaussian and 1-Besov priors, developing the picture for other priors.
- Other interesting questions:
  - When do MAP and wMAP coincide?
  - Local MAP and their theory
  - Posterior contraction rates for the new priors (work in progress with M. Dashti and T. Helin for Besov priors)

http://www.sergiosagapiou.com/

S. Agapiou, M. Burger, M. Dashti and T. Helin, *Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems*, arXiv:1705.03286

M. Dashti, K. Law, A. Stuart and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 2013

T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems, 2015

M. Dunlop and A. M. Stuart, *MAP estimators for piecewise continuous inversion*, Inverse Problems, 2016.

M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, Handbook of Uncertainty Quantification, 2015.