Maximum a posteriori estimates in Bayesian inversion

Martin Burger and Tapio Helin

Department of Mathematics and Statistics

University of Helsinki

Tsukuba, January 19, 2016

Bayesian inversion

Logic in Bayesian inference:

$$\text{Prior modeling} \quad \underset{Measurement}{\Longrightarrow} \quad \text{Posterior modeling}$$

Fundamental implications to inverse problems:

- All variables included in the model are represented by random variables.
- The degree of information concerning these values is coded into their distributions.
- The solution of the problem is the posterior probability distribution.

Hence the Bayesian paradigm asks what is our information about the unknown?

Bayesian solution to an inverse problem

Problem setting changes

$$m = Au + e \quad \Rightarrow \quad M = AU + E$$

where the capital letters M, U and E stand for random variables.

Bayesian solution to an inverse problem is then the probability distribution of U conditioned on a sample of M, i.e., the measurement. The probability measure

$$\mathbb{P}(U \in \mathcal{U} \mid M = m)$$

is called the posterior probability. Here $\mathcal{U}$ denotes some set of possible values of the unknown U.

The Bayes formula in finite dimensions

Suppose all related random variables are $\mathbb{R}^n$-valued and their distributions are absolutely continuous with respect to the Lebesgue measure.

Prior density $\pi_{pr}(u)$ expresses all prior information independent of the measurement.

Likehood density $\pi(m \mid u)$ is the likelihood of a measurement outcome m given $U = u$.

Bayes formula:

$$\pi_{post}(u) = \pi(u \mid m) = \frac{\pi_{pr}(u)\pi(m \mid u)}{\pi(m)}$$

Typical point estimators

Classical inversion methods produce single estimates of the unknown. In statistical approach one can calculate point estimates and confidence or interval estimates.

Maximum a posteriori estimate (MAP):

$$u_{MAP} = \arg\max_{u \in \mathbb{R}^n} \pi(u \mid m)$$

Conditional mean estimate (CM):

$$u_{CM} = \mathbb{E}(u \mid M = m) = \int_{\mathbb{R}^n} u' \pi(u' \mid m) \mathrm{d}u'$$

The Gaussian case

Example. Let $M = AU + E$ with $E \sim \mathcal{N}(0, C_e)$ and $U \sim \mathcal{N}(0, C_u)$. In this case, the posteriori density function is

$$\begin{aligned} \pi(u \mid m) &\propto \pi_{pr}(u)\pi(m \mid u) \\ &\propto \exp\left(-\frac{1}{2}\left(\left\|C_u^{-1/2}u\right\|_2^2 + \left\|C_e^{-1/2}(m - Au)\right\|_2^2\right)\right). \end{aligned}$$

The MAP estimate for this posteriori distribution

$$\arg\max_{u \in \mathbb{R}^n} \pi(u \mid m) \quad \Leftrightarrow \quad \arg\min_{u \in \mathbb{R}^n} \left(\|u\|_{C_u^{-1}}^2 + \|m - Au\|_{C_e^{-1}}^2\right).$$

In fact, for this example the MAP and the CM estimates coincide. In general, they can be worlds apart.

Consider a toy problem:

$$\hat{m}(t) = Au(t) + e(t),$$

where A is a convolution operator as follows:

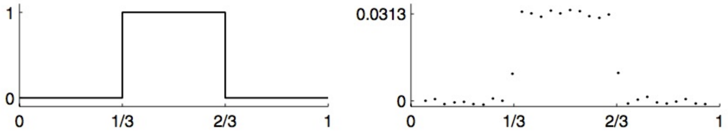

Figure 2. Left: simulated intensity distribution $u(t)$. Right: simulated noisy measurement $\hat{m}$. The dots are plotted at centre points of pixels.

Total Variation prior (Lassas–Siltanen 2004)

Total Variation prior in Bayesian inversion is formally defined as

$$\pi_{prior}(u) \propto \exp\left(-\alpha_n \int |\nabla u| dt\right)$$

to emulate the effect of regularization by BV-norm. Therefore, the posterior density is

$$\pi_{post}(u) \propto \exp\left(-\frac{1}{2}|Au - m|^2 - \alpha_n \int |\nabla u| dt\right)$$

Recall that

$$\pi_{prior}(u) \propto \exp\left(-\alpha_n \int |\nabla u| dt\right)$$

It turns out that TV prior is asymptotically unstable. The picture on the right is taken from

M. Lassas and S. Siltanen, Inverse problems 20(5), 2004.

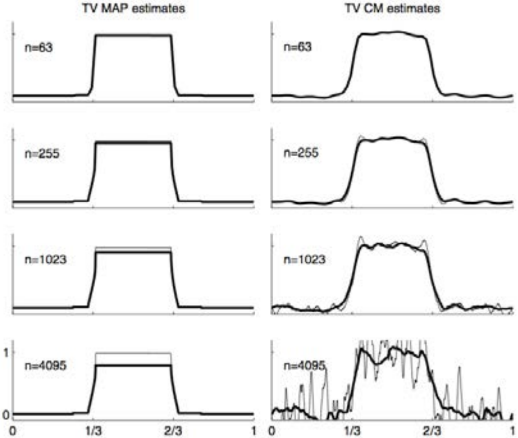

Figure 4. In all the plots in this figure, the coordinate axis limits are the same to allow easy comparison. Left column: MAP estimates for the TV prior with parameter $\alpha_n = 135$ (thin line) and $\alpha_n = 16.875\sqrt{n+1}$ (thick line). Right column: CM estimates for the TV prior with parameter $\alpha_n = 135$ (thin line) and $\alpha_n = 16.875\sqrt{n+1}$ (thick line).

Discretization invariance

$$\begin{array}{ccccccl} M & = & AU & + & E & & \text{Theoretical model} \\ \downarrow & & & & & & \\ M_k & = & A_k U & + & E_k & & \text{Measurement model} \\ & & \downarrow & & & & \\ M_{kn} & = & A_k U_n & + & E_k & & \text{Computational model} \end{array}$$

Challenges with infinite dimensions

(1) No uniform translation-invariant measure available (Lebesgue measure) $\Rightarrow$ working with the Bayes formula is more cumbersome

(2) Point estimators are problematic (CM is well-defined but difficult analyse, what is MAP?)

(3) Very few results on non-Gaussian models (Besov or hierarchical priors)

The Bayes formula in infinite-dimensional space

We consider the following measurement setting:

(1) a linear inverse problem $M = AU + E$, where $A : X \to \mathbb{R}^d$ is bounded,

(2) the prior distribution λ is a probability distribution on $(X, \mathcal{B}(X))$ and the noise satisfies $E \sim \mathcal{N}(0, I)$

Then a conditional distribution of U given M exists and

$$\mu_{post}(\mathcal{U} \mid m) = \frac{1}{Z} \int_{\mathcal{U}} \exp\left(-\frac{1}{2} \|Au - m\|_{\mathbb{R}^d}^2\right) \lambda(du) \quad \mathcal{U} \in \mathcal{B}(X),$$

for almost every $m \in \mathbb{R}^d$.

Some of the existing infinite-dimensional literature

- Behavior of Gaussian distributions is well-known (Mandelbaum (1984), Luschgy (1995), Lasanen (2002), Stuart)
- Posterior consistency i.e. noise converges to delta distribution (Pikkarainen-Neubauer (2008), Stuart, Agapiou, Kekkonen and many others)
- Non-Gaussian phenomena (Siltanen et al. (2004, 2009, 2011), Burger-Lucka (2014))
- Discretization invariance (Siltanen et al. (2004, 2009), Lasanen 2012)
- How to define a MAP estimate (Hegland (2007), Dashti et al. (2013), H-Burger (2015))

Differentiability of measures

The following concept originating to papers by Sergei Fomin in the 1960s.

Definition

A measure μ on X is called Fomin differentiable along the vector h if, for every set $\mathcal{A} \in \mathcal{B}(X)$, there exists a finite limit

$$d_h\mu(\mathcal{A}) = \lim_{t \to 0} \frac{\mu(\mathcal{A} + th) - \mu(\mathcal{A})}{t}$$

The set function $d_h\mu$ is a countably additive signed measure on $\mathcal{B}(X)$ and has bounded variation due to the Nikodym theorem.

We denote the domain of differentiability by

$$D(\mu) = \{h \in X \mid \mu \text{ is Fomin differentiable along } h\}$$

Differentiability of measures

By considering function $f(t) = \mu(\mathcal{A} + th)$ and its derivative at zero, we see that $d_h\mu$ is absolutely continuous with respect to μ.

Definition

The Radon–Nikodym density of the measure $d_h\mu$ with respect to μ is denoted by β_h^μ and is called the logarithmic derivative of μ along h.

Consequently, for all $\mathcal{A} \in \mathcal{B}(X)$ the logarithmic gradient β_h^μ satisfies

$$d_h\mu(\mathcal{A}) = \int_{\mathcal{A}} \beta_h^\mu(u)\mu(du)$$

and, in particular, we have $d_h\mu(X) = 0$ for any $h \in D(\mu)$ by definition. Moreover, $\beta_{sh}^\mu = s \cdot \beta_h^\mu$ for any $s \in \mathbb{R}$.

Finite dimensional example

Example. Suppose the posterior is of the form

$$\pi_{post}(u \mid m) \propto \exp\left(-\frac{1}{2}|Au - m|_2^2 - J(u)\right)$$

with differentiable J (e.g. $J(u) = |C_u^{-1/2} u|_2^2$ for a Gaussian prior). Then the logarithmic derivate satisfies

$$\beta_h^\mu(u) = -\langle A^*(Au - m) + J'(u), h\rangle.$$

Formally, we want study the zero points of β_h^μ in the infinite-dimensional case (see Hegland 2007).

Gaussian example

Example. Suppose

- X is a separable Hilbert space
- T is a non-negative self-adjoint Hilbert–Schmidt operator on X and
- γ is a zero-mean Gaussian measure on $(X, \mathcal{B}(X))$ with mean u_0 and covariance T^2,

then the Cameron–Martin space of γ is defined by

$$H(\gamma) := T(X), \qquad \langle h_1, h_2 \rangle_{H(\gamma)} = (T^{-1}h_1, T^{-1}h_2)_X.$$

and the logarithmic derivative of γ satisfies

$$\beta_h^\gamma(u) = -\langle h, u - u_0 \rangle_{H(\gamma)} \quad \text{for any} \quad h \in D(\gamma) = H(\gamma).$$

Pitfall: The expression for β_h^γ should be understood as a measurable extension.

MAP estimate by Dashti-Law-Stuart-Voss (2013)

Definition

Let $M^\epsilon = \sup_{u \in X} \mu(B_\epsilon(u))$. Any point $\hat{u} \in X$ satisfying

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

is a MAP estimate for the measure μ.

We remark that $\lim_\epsilon \left(\mu(B_\epsilon(u))/M^\epsilon\right) \leq 1$ holds for any $u \in X$. Dashti and others showed that for certain non-linear F, the MAP estimate for Gaussian noise ρ and prior λ satisfies

$$\hat{u} = \mathrm{argmin}_{u \in X} \left(\|F(u) - m\|^2_{CM(\rho)} + \|u\|^2_{CM(\lambda)} \right).$$

How to generalize for non-Gaussian priors?

Generalized Onsager–Machlup functional

Theorem (Bogachev)

Suppose μ is a Radon measure on a locally convex space X and is Fomin differentiable along a vector $h \in X$. Moreover, if, $\exp(\epsilon|\beta_h^\mu(\cdot)|) \in L^1(\mu)$ for some $\epsilon > 0$, then

$$\frac{d\mu_h}{d\mu}(u) = \exp\left(\int_0^1 \beta_h^\mu(u - sh)ds\right) \quad \text{in } L^1(\mu).$$

We also need to require that

(A1) for any $h \in E$ there exists $\epsilon > 0$ such that the prior probability measure λ satisfies $\exp(\epsilon|\beta_h^\lambda(\cdot)|) \in L^1(\lambda)$.

Main tools

We need to assume that

(A2) there exists a separable Banach space $E \subset D(\mu)$ such that E is topologically dense in X and $\beta_h^\mu \in C(X)$ for any $h \in E$ that is β_h^μ has a continuous representative.

Lemma

Assume that $\mu_h \ll \mu$ and denote $r_h = \frac{d\mu_h}{d\mu} \in L^1(\mu)$. Suppose r_h has a continuous representative $\tilde{r}_h \in C(X)$, i.e., $r_h - \tilde{r}_h = 0$ in $L^1(\mu)$. Then it holds that

$$\lim_{\epsilon \to 0} \frac{\mu_h(B_\epsilon(u))}{\mu(B_\epsilon(u))} = \tilde{r}_h(u)$$

for any $u \in X$.

Weak MAP estimate

Definition (H–Burger)

We call a point $\hat{u} \in X$, $\hat{u} \in \mathrm{supp}(\mu)$, a weak MAP (wMAP) estimate if

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

for all $h \in E$.

Every MAP is a wMAP

Lemma

Every MAP estimate $\hat{u}$ is a weak MAP estimate.

Proof.

The claim is trivial since

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

for any $h \in E$. □

Connection to a variational problem

A probability measure λ on $\mathcal{B}(X)$ is called convex if, for all sets $\mathcal{A}, \mathcal{B} \subset \mathcal{B}(X)$ and all $t \in [0,1]$, one has

$$\lambda(t\mathcal{A} + (1-t)\mathcal{B}) \geq \lambda(\mathcal{A})^t \lambda(\mathcal{B})^{1-t}.$$

Theorem

(1) *If $\hat{u} \in X$ is a weak MAP estimate of μ, then $\beta_h^\mu(\hat{u}) = 0$ for all $h \in E$.*

(2) *Suppose that μ is convex and there exists $\tilde{u} \in X$ such that $\beta_h^\mu(\tilde{u}) = 0$ for all $h \in E$. Then $\tilde{u}$ is a weak MAP estimate.*

Theorem

If $\hat{u} \in X$ is a weak MAP estimate of μ, then $\beta_h^\mu(\hat{u}) = 0$ for all $h \in E$.

Proof.

It follows from $\frac{d\mu_h}{d\mu}(\hat{u}) \leq 1$ and identity generalized Onsager–Machlup formula that

$$\int_0^t \beta_h^\mu(\hat{u} - sh)ds = \int_0^1 \beta_{th}^\mu(\hat{u} - s' \cdot th)ds' \leq 0$$

for all $h \in E$ and $t \in \mathbb{R}$. By continuity we then have $\beta_h^\mu(\hat{u}) \leq 0$. Now since $h, -h \in E \subset D(\mu)$ and by similar reasoning $\beta_{-h}^\mu(\hat{u}) \leq 0$, we must have

$$0 \leq -\beta_{-h}^\mu(\hat{u}) = \beta_h^\mu(\hat{u}) \leq 0$$

and the claim follows. □

So what about our posterior measure?

Recall that

$$\mu_{post}(\mathcal{U} \mid m) = \frac{1}{Z} \int_{\mathcal{U}} \exp\left(-\frac{1}{2}|Au - m|^2\right) \lambda(du)$$

- λ is convex $\Rightarrow \mu_{post}$ is convex
- Also, $D(\lambda) \subset D(\mu_{post})$ and

$$\begin{aligned} d_h \mu_{post} &= f \cdot d_h \lambda + \partial_h f \cdot \lambda \\ &= \left(\beta_h^\lambda(\cdot) - \langle A \cdot - m, Ah \rangle_{\mathbb{R}^d}\right) f \lambda \\ &= \beta_h^{\mu_{post}} \mu_{post} \end{aligned}$$

Theorem

If λ satisfies (A1) and (A2), then so does μ_{post}.

Proof.

(A1) is clear. For (A2) we have

$$\begin{aligned}
&\left\|\exp(\epsilon|\beta_h^{\mu_{post}}(\cdot)|)\right\|_{L^1(\mu)} \\
&\leq C\int_X \exp(\epsilon(C_1|Au-m|_{\mathbb{R}^d}+|\beta_h^\lambda(u)|))\exp\left(-\frac{1}{2}|Au-m|^2\right)\lambda(du) \\
&\qquad \leq \widetilde{C}\int_X \exp\left(-(|Au-m|_{\mathbb{R}^d}-C_2)^2\right)\exp(\epsilon|\beta_h^\lambda(u)|)\lambda(du) \\
&\qquad\qquad \leq \widetilde{C}\left\|\exp(\epsilon|\beta_h^\lambda(\cdot)|)\right\|_{L^1(\lambda)},
\end{aligned}$$

for suitable $\epsilon > 0$ and constants $C, \widetilde{C}, C_1, C_2 > 0$. □

Weak MAP in Bayesian inversion

Corollary

Let us assume that μ_{post} *and* λ *are as earlier. Moreover, we assume that the prior distribution* λ *is a convex measure and there is an (unbounded) convex functional* $J : X \to [0, \infty]$, *which is Frechet differentiable everywhere in its domain* $D(J)$ *and* $J'(u)$ *has a bounded extension* $J'(u) : E \to \mathbb{R}$ *such that*

$$\beta_h^\lambda(u) = J'(u)h$$

for any $h \in E$ *and any* $u \in X$. *Then a point* $\hat{u}$ *is a weak MAP estimate if and only if* $\hat{u} \in \arg\min_{u \in X} F(u)$ *where*

$$F(u) = \frac{1}{2}|Au - m|^2 + J(u). \tag{1}$$

Shortly about Besov spaces

Suppose $\{\psi_\ell\}_{\ell=1}^\infty$ form an orthonormal wavelet basis for $L^2(\mathbb{T}^d)$. We define $B^s_{pq}(\mathbb{T}^d)$ as follows: the series

$$f(x) = \sum_{\ell=1}^{\infty} c_\ell \psi_\ell(x) \tag{2}$$

belongs to $B^s_{pq}(\mathbb{T}^d)$ if and only if

$$2^{js} 2^{j(\frac{1}{2}-\frac{1}{p})} \left(\sum_{\ell=2^j}^{2^{j+1}-1} |c_\ell|^p \right)^{1/p} \in \ell^q(\mathbb{N}). \tag{3}$$

We write $B^s_p = B^s_{pp}$.

Definition

Let $1 \leq p < \infty$ and let $(X_\ell)_{\ell=1}^{\infty}$ be independent identically distributed real-valued random variables with the probability density function

$$\pi_X(x) = \sigma_p \exp(-|x|^p) \quad \text{with} \quad \sigma_p = \left(\int_{\mathbb{R}} \exp(-|x|^p) dx \right)^{-1}. \tag{4}$$

Let U be the random function

$$U(x) = \sum_{\ell=1}^{\infty} \ell^{-\frac{s}{d} - \frac{1}{2} + \frac{1}{p}} X_\ell \psi_\ell(x), \quad x \in \mathbb{T}^d.$$

Then we say that U is distributed according to a B_p^s prior.

Besov priors

Theorem

It holds that

(1) $D(\lambda) = B_2^{s+(\frac{1}{2}-\frac{1}{p})d}(\mathbb{T}^d)$ *for* $p > 1$,

(2) $\exp(|\beta_h^\lambda|) \in L^1(\lambda)$ *for any* $h \in E = B_p^{ps-(p-1)t}(\mathbb{T}^d)$ *and*

(3) $\tilde{\beta}_h^\lambda \in C(B_p^t(\mathbb{T}^d))$ *for any* $h \in B_p^{ps-(p-1)t}(\mathbb{T}^d)$ *and* $1 < p \leq 2$.

Moreover, the weak MAP estimate of the inverse problem is obtained by minimizing functional

$$F_{Besov}(u) = \frac{1}{2}|Au - m|^2 + \|u\|_{B_p^s}^p .$$

Conclusions

- Infinite-dimensional Bayesian inverse problems contain many big open questions
- Studying differentiability of the posterior opens new avenues of research
- MAP estimates can be solved for non-Gaussian priors with certain differentiability

For more details:

Helin, T. and Burger, M.: *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems 31(8) (2015).