Maximum a posteriori estimates in Bayesian inversion

Tapio Helin

Department of Mathematics and Statistics
University of Helsinki

Warwick, June 2, 2015
1. Shortly on Bayesian inversion

2. Towards a general MAP estimate

3. Example: Besov priors

4. Bayes cost of MAP
1. Shortly on Bayesian inversion

2. Towards a general MAP estimate

3. Example: Besov priors

4. Bayes cost of MAP
The Bayes formula in finite dimensions

Consider a linear statistical inverse problem

\[ M = AU + E \]

where \( M, U \) and \( E \) stand for the random variables describing measurement, unknown and noise, respectively.

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

- **Likelihood density** \( \pi(m \mid u) \) is the likelihood of a measurement outcome \( m \) given \( U = u \).

- **Bayes formula:**

\[
\pi_{post}(u) = \pi(u \mid m) = \frac{\pi_{like}(m \mid u)\pi_{pr}(u)}{\pi(m)}
\]
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): 

\[ u_{MAP} = \arg \max_{u \in \mathbb{R}^d} \pi(u | m) \]

Conditional mean estimate (CM): 

\[ u_{CM} = \mathbb{E}(u | M = m) = \int_{\mathbb{R}^d} u' \pi(u' | m) du' \]
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

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

The MAP estimate for this posteriori distribution

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

In fact, for this example the MAP and the CM estimates coincide.
Total Variation prior in Bayesian inversion is formally defined as

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

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

Discretization invariance (Siltanen et al. 2004, 2009)

\[
M = AU + E \quad \text{Theoretical model}
\]

\[
M_k = A_k U + E_k \quad \text{Measurement model}
\]

\[
M_{kn} = A_k U_n + E_k \quad \text{Computational model}
\]
Challenges with infinite dimensions

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

(2) Point estimators are problematic (CM is well-defined but difficult to 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 \( \lambda \) 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

\[
\mu_{\text{post}}(U \mid m) = \frac{1}{Z} \int_U \exp \left( -\frac{1}{2} |Au - m|^2 \right) \lambda(du) \quad U \in \mathcal{B}(X),
\]

for almost every \( m \in \mathbb{R}^d \).
Some of the existing infinite-dimensional literature

- Behavior of Gaussian distributions is well-understood (Mandelbaum (1984), Somersalo et al. (1989), Stuart and others)
- Posterior consistency i.e. noise converges to delta distribution (Pikkarainen et al. (2008), van der Vaart, Stuart, Agapiou, Kekkonen and others)
- Non-Gaussian phenomena (Siltanen et al. (2004, 2009), Dashti et al. (2011), Burger-Lucka (2014))
- Discretization invariance (Siltanen et al. (2004, 2009), Cotter et al. (2010), Lasanen 2012)
- How to define a MAP estimate (Hegland (2007), Dashti et al. (2013), H-Burger (2015))
1. Shortly on Bayesian inversion
2. Towards a general MAP estimate
3. Example: Besov priors
4. Bayes cost of MAP
Differentiability of measures

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

**Definition**

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

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

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

**Definition**

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

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

$$d_h\mu(A) = \int_A \beta^\mu_h(u)\mu(du)$$

and, in particular, we have $d_h\mu(X) = 0$ for any $h \in D(\mu)$ by definition. Moreover, $\beta^\mu_{sh} = s \cdot \beta^\mu_h$ for any $s \in \mathbb{R}$. 
Finite dimensional example

Suppose the posterior is of the form

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

with differentiable \( J \). Then the logarithmic derivative satisfies

\[ \beta^\mu_h(u) = -\langle A^*(Au - m) + J'(u), h \rangle. \]

Formally, we want to study the zero points of \( \beta^\mu_h \) in the infinite-dimensional case (see Hegland 2007).
Suppose

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

then the Cameron–Martin space of $\gamma$ is defined by

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

and the logarithmic derivative of $\gamma$ satisfies

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

**Pitfall:** The formula for $\beta^\gamma_h$ should be understood as a measurable extension.
**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 \( \mu \).

Dashti and others showed that for certain non-linear \( F \), the MAP estimate for Gaussian noise \( \rho \) and prior \( \lambda \) satisfies

\[
\hat{u} = \arg\min_{u \in X} \left( \|F(u) - m\|_{CM(\rho)}^2 + \|u\|_{CM(\lambda)}^2 \right).
\]

How to generalize for non-Gaussian priors?
Theorem (Bogachev 2010)

Suppose $\mu$ 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

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

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 $u \in X$. 
Assumptions:

(A1) there exists a separable Banach space $E \subset D(\mu)$ such that $E$ is dense in $X$ and $\beta_h^\mu \in C(X)$ for any $h \in E$ and

(A2) for any $h \in E$ there exists $\epsilon > 0$ such that $\exp(\epsilon |\beta_h^\mu(\cdot)|) \in L^1(\mu)$.

Definition (H–Burger)

We call a point $\hat{u} \in X$, $\hat{u} \in \text{supp}(\mu)$, 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 \( h \in E \).
A probability measure $\lambda$ on $\mathcal{B}(X)$ is called convex if, for all sets $A, B \subset \mathcal{B}(X)$ and all $t \in [0, 1]$, one has

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

If \( \hat{u} \in X \) is a weak MAP estimate of \( \mu \), then \( \beta^\mu_h(\hat{u}) = 0 \) 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

\[
\int_0^t \beta^\mu_h(\hat{u} - sh)ds = \int_0^1 \beta^\mu_{th}(\hat{u} - s' \cdot th)ds' \leq 0
\]

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

\[
0 \leq -\beta^\mu_{-h}(\hat{u}) = \beta^\mu_h(\hat{u}) \leq 0
\]

and the claim follows.
So what about our posterior measure?

Recall that

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

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

$$
d_h \mu_{\text{post}} = f \cdot d_h \lambda + \partial_h f \cdot \lambda
= \left( \beta_h^\lambda(\cdot) - \langle A \cdot -m, Ah \rangle_{\mathbb{R}^d} \right) f \lambda
= \beta_{\mu_{\text{post}}}^h \mu_{\text{post}}
$$
Theorem

If \( \lambda \) satisfies (A1) and (A2), then so does \( \mu_{\text{post}} \).

Proof.

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

\[
\left\| \exp(\epsilon |\beta_h^\mu_{\text{post}}(\cdot)|) \right\|_{L^1(\mu)} 
\leq C \int_X \exp(\epsilon (C_1 |Au - m|_{\mathbb{R}^d} + |\beta_h^\lambda(u)|)) \exp \left( -\frac{1}{2} |Au - m|^2 \right) \lambda(du) 
\leq \tilde{C} \int_X \exp \left( -(|Au - m|_{\mathbb{R}^d} - C_2)^2 \right) \exp(\epsilon |\beta_h^\lambda(u)|) \lambda(du) 
\leq \tilde{C} \left\| \exp(\epsilon |\beta_h^\lambda(\cdot)|) \right\|_{L^1(\lambda)},
\]

for suitable \( \epsilon > 0 \) and constants \( C, \tilde{C}, C_1, C_2 > 0 \).
Theorem (H–Burger)

Suppose \( \lambda \) is convex and satisfies (A1) and (A2). Moreover, assume that there is a convex functional \( J : X \to \mathbb{R} \cup \{\infty\} \), which is Fréchet differentiable 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 \( \hat{u} \) is wMAP if and only if \( \hat{u} \in \arg \min_{u \in X} F(u) \) where

\[
F(u) = \frac{1}{2} |Au - m|^2 + J(u). \tag{1}
\]
1. Shortly on Bayesian inversion
2. Towards a general MAP estimate
3. Example: Besov priors
4. Bayes cost of MAP
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

\[
f(x) = \sum_{\ell=1}^{\infty} c_\ell \psi_\ell(x)
\]

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

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

We write \( B^s_p = B^s_{pp} \).
Besov priors

**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

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

(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 Besov prior in $B^s_p$. 
Theorem

Assume that $\lambda$ is a Besov prior in $B_p^s$. It holds that

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

2. $\exp(|\beta^\lambda_h|) \in L^1(\lambda)$ for any $h \in B_p^{ps-(p-1)t}(\mathbb{T}^d)$ and

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

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

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

2. Towards a general MAP estimate

3. Example: Besov priors

4. Bayes cost of MAP
Bayes Cost Formalism

Consider a general estimator \( \hat{u} : m \mapsto \hat{u}(m) \) for the inverse problem. The Bayes cost is defined by the expected cost, i.e., the average performance

\[
BC_\Psi(\hat{u}) := \mathbb{E}[\Psi(u, \hat{u}(m))] = \int \int \Psi(u, \hat{u}(m)) \pi_{post}(u \mid m) du \pi(m) dm.
\]

The Bayes estimator \( \hat{u}_\Psi \) is the estimator which minimizes \( BC_\Psi(\hat{u}) \).

- The CM estimate minimizes mean squared error
  \[
  \Psi_{MSE}(u, \hat{u}) = |u - \hat{u}|_2^2.
  \]

- The MAP is asymptotic Bayes estimator of
  \[
  \Psi_\delta(u, \hat{u}) = \begin{cases} 
  0, & \text{if } |u - \hat{u}|_\infty \leq \delta, \\
  1, & \text{otherwise.}
  \end{cases}
  \]
MAP as a proper Bayes estimator

**Definition**

For a convex functional $J : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$, the Bregman distance $D^q_J(u, v)$ between $u, v \in \mathbb{R}^n$ for a subgradient $q \in \partial J(v)$ is defined as

$$D^q_J(u, v) = J(u) - J(v) - \langle q, u - v \rangle, \quad q \in \partial J(v).$$

Suppose $\pi_{pr}(u) \propto \exp(-J(u))$, where $J : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is a Lipschitz-continuous convex functional and $u \mapsto |Au|^2_2 + J(u)$ has at least linear growth at infinity. Define

$$\Psi_{Brg}(u, \hat{u}) = \frac{1}{2} |A(\hat{u} - u)|^2_2 + D_J(\hat{u}, u).$$

**Theorem (Burger–Lucka 2014)**

The MAP estimate is a Bayes estimator for $\Psi_{Brg}$. 
Define homogeneous Bregman distance

\[ \tilde{D}_J(u, \cdot) = J(u) + \beta_u^\lambda(\cdot) \quad \text{in} \ L^1(\lambda) \]

for any \( u \in D(\lambda) \cap D(J) \).

**Theorem**

Assume that \( \mu_{\text{post}} \) and \( \lambda \) are as in previous theorem. Then \( \hat{u} \) is a weak MAP estimate if and only if it minimizes the functional

\[
G(u) = \int_X \left( \frac{1}{2} |Au - Av|^2_2 + \tilde{D}_J(u, v) \right) \mu(dv).
\]

Moreover, we have

\[
\int_X \tilde{D}_J(u_{\text{MAP}}, v) \mu(dv) \leq \int_X \tilde{D}_J(u_{\text{CM}}, v) \mu(dv).
\]
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 rigorously defined for certain class of non-Gaussian priors

For more details: