Sparse Bayesian Inference & Uncertainty Quantification for Inverse Imaging Problems

Felix Lucka
Centrum Wiskunde & Informatica
University College London
Felix.Lucka@cwi.nl

Statistics for Structures Seminar
Leiden

October 20, 2017
Big Picture: From Qualitative to Quantitative Imaging

Traditional task: Produce results to be interpreted by trained experts
⇒ Qualitative usage of the reconstructed information.

Example: Conventional computer tomography (CT).

Source: Wikimedia Commons
Felix.Lucka@cwi.nl - Sparse Bayesian Inference & UQ for Inverse Imaging Problems
Big Picture: From Qualitative to Quantitative Imaging

**Traditional task:** Produce results to be interpreted by trained experts
- $\Rightarrow$ *Qualitative* usage of the reconstructed information.

**New demand:** Produce results for automatized analysis procedures / hypothesis testing; multimodal imaging.
- $\Rightarrow$ *Quantitative* usage of the reconstructed information.

**Example:** *Dynamical causal modeling (DCM).*

Source: Andre C. Marreiros et al. (2010), Scholarpedia, 5(7):9568.
Bayesian Inversion and Uncertainty Quantification

Noisy, ill-posed inverse problems:

\[ f = N(A(u), \varepsilon) \]

Example: \( f = Au + \varepsilon \)

\[ \text{p}_{\text{like}}(f|u) \propto \exp \left( -\frac{1}{2} \| f - Au \|^2 \right) \]

\[ \text{p}_{\text{prior}}(u) \propto \exp \left( -\lambda \| D^T u \|^2 \right) \]

\[ \text{p}_{\text{post}}(u|f) \propto \exp \left( -\frac{1}{2} \| f - Au \|^2 - \lambda \| D^T u \|^2 \right) \]

Probabilistic representation allows for rigorous quantification of solution's uncertainties.
Bayesian Inversion and Uncertainty Quantification

Noisy, ill-posed inverse problems:

\[ f = N (A(u), \varepsilon) \]

Example: \( f = Au + \varepsilon \)

\[ p_{\text{like}}(f|u) \propto \exp \left( -\frac{1}{2} \| f - Au \|_2^2 \right) \]

\[ p_{\text{prior}}(u) \propto \exp \left( -\lambda \| D^T u \|_1 \right) \]

\[ p_{\text{post}}(u|f) \propto \exp \left( -\frac{1}{2} \| f - Au \|_2^2 - \lambda \| D^T u \|_1 \right) \]

Probabilistic representation allows for rigorous quantification of solution’s uncertainties.
Sparsity / Compressible Representation

(a) 100%
(b) 10%
(c) 1%

**Sparsity** as a-priori constraints are used in variational regularization, compressed sensing and variable selection:

\[
\hat{u}_\lambda = \arg\min_u \left\{ \frac{1}{2} \| f - Au \|_2^2 + \lambda \| D^T u \|_1 \right\}
\]

(e.g. total variation, wavelet shrinkage, LASSO,...)
Sparsity as a-priori constraints are used in variational regularization, compressed sensing and variable selection:

\[
\hat{u}_\lambda = \arg\min_u \left\{ \frac{1}{2} \| f - Au \|_2^2 + \lambda \| D^T u \|_1 \right\}
\]

(e.g. total variation, wavelet shrinkage, LASSO, ...)

Sparse Bayesian inversion?
Uncertainty Quantification for Sparse Bayesian Inversion

- How to model sparsity?
  - $\ell_1$-norm priors.
  - Gaussian scale mixture (hierarchical Bayesian)
  - $\ell_p$-norm scale mixture (hierarchical Bayesian)

- How to compute estimators / UQ measures?

- What can we say about estimators?

- Meaningful UQ measures for sparse inversion/imaging?
Efficient MCMC for Sparse Image Reconstruction

Task: Monte Carlo integration by samples from

\[ p_{post}(u|f) \propto \exp \left( -\frac{1}{2} \| f - Au \|_{\Sigma^{-1}}^2 - \lambda \| D(u) \|_1 \right) \]

Problem: Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large \( n \) or \( \lambda \).
Efficient MCMC for Sparse Image Reconstruction

**Task:** Monte Carlo integration by samples from

\[ p_{\text{post}}(u|f) \propto \exp \left( -\frac{1}{2} \| f - A u \|_2^2 \Sigma^{-1} - \lambda \| D(u) \|_1 \right) \]

**Problem:** Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large \( n \) or \( \lambda \).

**Contributions:**

- Development of different Gibbs samplers.
- Efficient for high-dim. imaging \((n > 10^6)\).


Efficient MCMC for Sparse Image Reconstruction

**Task:** Monte Carlo integration by samples from

\[ p_{post}(u|f) \propto \exp \left( -\frac{1}{2} \| f - A u \|_{\Sigma^{-1}}^2 - \lambda \| D(u) \|_1 \right) \]

**Problem:** Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large \( n \) or \( \lambda \).

**Work by Marcelo Pereyra et al.:**

- Unadjusted Langevin algorithm applied to Moreau-Yoshida envelopes of posterior energy.
- As easy to implement as proximal gradient descent.

Point Estimators in Bayesian Inference for Imaging

\[ \hat{u}_{\text{MAP}} := \arg \max_{u \in \mathbb{R}^n} \{ p_{\text{post}}(u|f) \} \quad \text{vs.} \quad \hat{u}_{\text{CM}} := \int u \, p_{\text{post}}(u|f) \, du \]

State in imaging \( \sim 5 \) years ago:

- CM preferred in theory, inaccessible in practice.
- MAP discredited by theory, accessible in practice.
Point Estimators in Bayesian Inference for Imaging

\[ \hat{u}_{MAP} := \arg \max_{u \in \mathbb{R}^n} \{ p_{post}(u | f) \} \quad \text{vs.} \quad \hat{u}_{CM} := \int u \ p_{post}(u | f) \, du \]

State in imaging \( \sim \) 5 years ago:

- CM preferred in theory, inaccessible in practice.
- MAP discredited by theory, accessible in practice.

However:

- MAP results looks/perform better or similar to CM.
- Gaussian priors: MAP = CM. Funny coincidence?
- Theoretical argument has a logical flaw.
Point Estimators in Bayesian Inference for Imaging

\[ \hat{u}_{\text{MAP}} := \arg\max_{u \in \mathbb{R}^n} \{ p_{\text{post}}(u|f) \} \quad \text{vs.} \quad \hat{u}_{\text{CM}} := \int u \ p_{\text{post}}(u|f) \ du \]

State in imaging \(\sim\) 5 years ago:
- CM preferred in theory, inaccessible in practice.
- MAP discredited by theory, accessible in practice.

Contributions:
- Theoretical rehabilitation of MAP.
- Key: Bayes cost based on Bregman distances.
- Gaussian case consistent in this framework.


Experimental Data: Limited-Angle CT

- Cooperation with Samuli Siltanen, Esa Niemi et al.
- Besov and TV prior; non-negativity constraints.
- Stochastic noise modeling.
- Uncertainty quantification for limited angle CT.

Use the data set for your own work: arXiv:1502.04064)
Walnut-CT with TV Prior: Full vs. Limited Angle

(a) MAP, full
(b) CM, full
(c) CStd, full
(d) MAP, limited
(e) CM, limited
(f) CStd, limited
TV Prior, Non-Negativity Constraints, Limited Angle

(a) CM, uncon  
(b) CM, non-neg  
(c) CStd, uncon  
(d) CStd, non-neg
However...

(a) CStd, full
(b) CStd, limited

- What does it really tell me?
- Does the uncertainty decrease?!
Hierarchical Bayesian Modeling (HBM) of Sparsity

Gaussian increment prior:

\[ p_{\text{prior}}(u) \propto \prod_i \exp \left( -\frac{(u_{i+1} - u_i)^2}{\gamma} \right) \]

- Gaussian variables live on characteristic scale, determined by \( \gamma \).
- Similar amplitudes are likely, sparsity (= outliers) is unlikely.
Hierarchical Bayesian Modeling (HBM) of Sparsity

Conditionally Gaussian increment prior:

\[ p_{\text{prior}}(u|\gamma) \propto \prod_i \exp \left( -\frac{(u_{i+1} - u_i)^2}{\gamma_i} \right) \]

Scale-invariant hyperprior to approximate un-informative \( \gamma_i^{-1} \) prior:

\[ p_{\text{hyper}}(\gamma_i) \propto \gamma_i^{-(\alpha+1)} \exp \left( -\frac{\beta}{\gamma_i} \right), \quad \text{inverse gamma distribution} \]
The Implicit Energy Functional behind HBM

Implicit prior is a Student’s $t$-prior with $\nu = 2\alpha$, $\theta = \beta/(2\alpha)$:

$$p_{\text{prior}}(u) \propto \prod_i \left(1 + \frac{u_i^2}{\nu \theta}\right)^{-\frac{\nu-1}{2}}$$

$$p_{\text{post}}(u|f) \propto \exp \left(-\frac{1}{2} \|f - A u\|^2_{\Sigma^{-1}} - \frac{\nu-1}{2} \sum_i \log \left(1 + \frac{u_i^2}{\nu \theta}\right)\right)$$
Prior Samples

(a) $\ell_2$

(b) $\ell_1$

(c) $\ell_{1/2}$

(d) Cauchy

$p_{\text{prior}}(u_i) \propto \exp(-|u_i|^p)$ \hspace{1cm} vs. \hspace{1cm} $p_{\text{prior}}(u_i) \propto \frac{1}{1 + u_i^2}$
Why HBM? EEG/MEG Source Reconstruction

**Aim:** Reconstruction of brain activity by **non-invasive** measurement of induced electromagnetic fields outside of skull.

Source: Wikimedia Commons

source: Wikimedia Commons
Why HBM? EEG/MEG Source Reconstruction

**Aim:** Reconstruction of brain activity by non-invasive measurement of induced electromagnetic fields outside of skull.

Notoriously ill-posed problem!
HBM for EEG/MEG Source Reconstruction

- Inversion with log-concave priors (e.g., $\ell_1$-type) suffers from systematic depth miss-localization, HBM does not.

- HBM shows promising results for focal brain networks with simulated and real data and EEG-MEG combination.

### Comparison: Two Approaches to Sparsity

<table>
<thead>
<tr>
<th>feature</th>
<th>$\ell_p$ prior</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J(u)$</td>
<td>$|u|_p^p$</td>
<td>$\frac{\nu+1}{2} \sum \log \left(1 + \frac{u^2}{\nu \theta}\right)$</td>
</tr>
<tr>
<td>sparsifying parameter</td>
<td>$p &gt; 0$</td>
<td>$\nu &gt; 0$</td>
</tr>
<tr>
<td>quadratic limit</td>
<td>$p = 2$</td>
<td>$\nu \to \infty$</td>
</tr>
<tr>
<td>sparse limit</td>
<td>$p \to 0$</td>
<td>$\nu \to 0$</td>
</tr>
<tr>
<td>limit functional</td>
<td>$</td>
<td>u</td>
</tr>
<tr>
<td>solutions</td>
<td>sparse</td>
<td>compressible</td>
</tr>
<tr>
<td>differentiable</td>
<td>$p &gt; 1$</td>
<td>always</td>
</tr>
<tr>
<td>convex</td>
<td>everywhere for $p \geq 1$</td>
<td>$|u|_\infty &lt; \sqrt{\nu \theta}$</td>
</tr>
<tr>
<td>homogeneous</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
Comparison: Two Approaches to Sparsity

<table>
<thead>
<tr>
<th>feature</th>
<th>$\ell_p$ prior</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J(u)$</td>
<td>$|u|_p^p$</td>
<td>$\frac{\nu+1}{2} \sum \log \left(1 + \frac{u^2}{\nu \theta}\right)$</td>
</tr>
<tr>
<td>sparsifying parameter</td>
<td>$p &gt; 0$</td>
<td>$\nu &gt; 0$</td>
</tr>
<tr>
<td>quadratic limit</td>
<td>$p = 2$</td>
<td>$\nu \to \infty$</td>
</tr>
<tr>
<td>sparse limit</td>
<td>$p \to 0$</td>
<td>$\nu \to 0$</td>
</tr>
<tr>
<td>limit functional</td>
<td>$</td>
<td>u</td>
</tr>
<tr>
<td>solutions</td>
<td>sparse</td>
<td>compressible</td>
</tr>
<tr>
<td>differentiable</td>
<td>$p &gt; 1$</td>
<td>always</td>
</tr>
<tr>
<td>convex</td>
<td>everywhere for $p \geq 1$</td>
<td>$|u|_\infty &lt; \sqrt{\nu \theta}$</td>
</tr>
<tr>
<td>homogeneous</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Combine them to get best (worst?!?) of both worlds?
$\ell_p$-hypermodels with generalized Gamma hyperpriors

$$p_{\text{prior}}(u, \gamma) \propto \exp \left( - \sum_i \left( \frac{|D_i^T u|^p}{\gamma_i} + \frac{\gamma_i^r}{\beta} - (r\alpha - 1 - 1/p) \log(\gamma_i) \right) \right)$$

Implicit prior with inverse gamma hyperprior:

$$\prod_i \left( 1 + \frac{|D_i^T u|^p}{\beta} \right)^{-\alpha - 1/p}$$

(a) $p = 2$  
(b) $p = 1$
$\ell_p$-Hypermodels & Majorization-Minimization

Posterior with gamma hyperprior ($r = 1$), $p = 1$, and $\alpha = 2$:

$$p_{post}(u|f) \propto \exp \left( -\frac{1}{2} \| f - A u \|_2^2 - \sum_i \left( \frac{|D_i^T u|}{\gamma_i} + \frac{\gamma_i}{\beta} \right) \right)$$

Computational scheme for full-MAP estimation equivalent to majorization-minimization scheme for $\ell_{1/2}$ regularization (Adaptive Lasso):

$$u^{(k)} = \arg\min_u \left\{ \frac{1}{2} \| f - A u \|_{\Sigma^{-1}}^2 + \frac{1}{\sqrt{\beta}} \sum_i \frac{|D_i^T u|}{\sqrt{|D_i^T u|^{(k-1)}}} \right\}$$

Uncertainty Quantification for Non-Convex Sparse Recovery

Severely under-determined problems $f = Au$:

- Many sparse solutions consistent with data!

- Log-concave priors erase this ambiguity and yield single result.
- HBM posteriors get multi-modal.
- Traditional UQ measure do not capture these aspects.
- Can we preserve but quantify, structure and visualize ambiguity?
Mode Analysis with MCMC & Optimization

- Generate MCMC chain of posterior samples.
- Use every sample as initialization of gradient-based optimization.
- Analyse resulting chain of modes.
Sparse Source Network Analysis for EMEG Auditory Data

all 364 EEG+MEG

all 306 MEG

182 MEG+EEG
Summary, Outlook & Open Questions

- $\ell_p$-norm and HBM road to sparsity: Neither perfect but (somewhat) computationally tractable. $\rightsquigarrow$ spike-and-slab priors?

- MAP estimates are proper Bayes estimators, modes are meaningful.

- However: Everything beyond point estimation is what’s really interesting.

- Meaningful and interpretable UQ measures for sparse inversion / imaging that can complement variational approaches?

- Does it really make sense?
  (over confidence in ill-posed problems, prior domination)


MAP vs. CM Estimates: The Classical View

A theoretical argument "decides" the conflict: The Bayes cost formalism.

- An estimator is a random variable, as it relies on $f$ and $u$.
- How does it perform on average? Which estimator is "best"?
- $\Rightarrow$ Define a cost function $\Psi(u, v)$.
- Bayes cost is the expected cost:

$$BC(\hat{u}) = \int \int \Psi(u, \hat{u}(f)) p_{\text{like}}(f|u) \, df \, p_{\text{prior}}(u) \, du$$

- Bayes estimator $\hat{u}_{BC}$ for given $\Psi$ minimizes Bayes cost. Turns out:

$$\hat{u}_{BC}(f) = \arg \min_{\hat{u}} \left\{ \int \Psi(u, \hat{u}(f)) p_{\text{post}}(u|f) \, du \right\}$$
MAP vs. CM Estimates: The Classical View

Main classical arguments pro CM and contra MAP estimates:

- CM is Bayes estimator for \( \Psi(u, \hat{u}) = \|u - \hat{u}\|_2^2 \) (MSE).
- Also the minimum variance estimator.
- The mean value is intuitive, it is the "center of mass", the known "average".
- MAP estimate can be seen as an asymptotic Bayes estimator of

\[
\Psi_\epsilon(u, \hat{u}) = \begin{cases} 
0, & \text{if } \|u - \hat{u}\|_\infty \leq \epsilon \\
1, & \text{otherwise},
\end{cases}
\]

for \( \epsilon \to 0 \) (uniform cost). \( \implies \) It is not a proper Bayes estimator.

- MAP and CM seem theoretically and computationally fundamentally different \( \implies \) one should decide.
- "A real Bayesian would not use the MAP estimate"
- People feel "ashamed" when they have to compute MAP estimates (even when their results are good).
A False Conclusion

“A real Bayesian would not use the MAP estimate as it is not a proper Bayes estimator”.

"MAP estimate can be seen as an asymptotic Bayes estimator of

\[
\Psi_\epsilon(u, \hat{u}) = \begin{cases} 
0, & \text{if} \quad \|u - \hat{u}\|_\infty < \epsilon \\
1 & \text{otherwise},
\end{cases}
\]

for \( \epsilon \to 0 \).

It is not a proper Bayes estimator."

"MAP estimator is asymptotic Bayes estimator for some degenerate \( \Psi \)”

⇒ “MAP can’t be Bayes estimator for some proper \( \Psi \)” !!!!
Two New Bayes Cost Functions

Define

(a) \( \Psi_{LS}(u, \hat{u}) := \|A(\hat{u} - u)\|_2^2 \Sigma_{\epsilon}^{-1} + \beta \|L(\hat{u} - u)\|_2^2 \)

(b) \( \Psi_{Brg}(u, \hat{u}) := \|A(\hat{u} - u)\|_2^2 \Sigma_{\epsilon}^{-1} + \lambda D_J(\hat{u}, u) \)

for a regular \( L \) and \( \beta > 0 \).

Properties:

- Proper, convex cost functions
- For \( J(u) = \beta/\lambda \|Lu\|_2^2 \) (Gaussian case!) we have \( \lambda D_J(\hat{u}, u) = \beta \|L(\hat{u} - u)\|_2^2 \), and \( \Psi_{LS}(u, \hat{u}) = \Psi_{Brg}(u, \hat{u}) \)

Theorems:

(I) The CM estimate is the Bayes estimator for \( \Psi_{LS}(u, \hat{u}) \)

(II) The MAP estimate is the Bayes estimator for \( \Psi_{Brg}(u, \hat{u}) \)
Bregman distances

For a proper, convex functional $\Psi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$, the Bregman distance $D^p_\Psi(f, g)$ between $f, g \in \mathbb{R}^n$ for a subgradient $p \in \partial \Psi(g)$ is defined as

$$D^p_\Psi(f, g) = \Psi(f) - \Psi(g) - \langle p, f - g \rangle, \quad p \in \partial \Psi(g)$$

 Basically, $D_\Psi(f, g)$ measures the difference between $\Psi$ and its linearization in $f$ at another point $g$.