Variational Bayesian Approximation with scale mixture prior for inverse problems : a numerical comparison between three algorithms

Leila Gharsalli, Ali Mohammad-Djafari
Aurélia Fraysse, Thomas Rodet
Groupe Problèmes Inverses (GPI)
Laboratoire des signaux et systèmes (L2S)
CNRS-SUPELEC-PARIS SUD, 91192 Gif-sur-yvette, France
Email : Leila.GHARSALLI@lss.supelec.fr

22 novembre 2012

Summary

Introduction

Linear inverse problem (discrized) : $\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon}$.

- $\mathbf{f} = [f_1, f_2, ..., f_N]^t \in \mathcal{R}^N$: unknowns to be estimated.
- $\mathbf{g} = [g_1, g_2, ..., g_M]^t \in \mathcal{R}^M$: observed data.
- $\boldsymbol{\epsilon}$: errors of modelling and measurement.
- $\mathbf{H} \in \mathcal{M}_{M \times N}$: matrix of the system response with high dimensions $\Rightarrow$ ill-posed inverse problem.

Objective : Estimate $\mathbf{f} \rightarrow \hat{\mathbf{f}}$.

Tools :

1. Deterministic : Regularization (Tikhonov regularization [Tikhonov, 1963]).
2. Probabilistic : Bayesain Approach
   -MCMC [Robert and Casella, 1998] (computational cost).
   - Variational Bayesian Approach
   [Smídl and Quinn, 2005](faster approach).

General Bayesian inference with scale mixture prior

$\mathbf{g} = \mathbf{H}\mathbf{f} + \boldsymbol{\epsilon} \Rightarrow$ Bayes Rule :

$$p(\mathbf{f}|\mathbf{g};\mathcal{M}) = \frac{p(\mathbf{g}|\mathbf{f};\mathcal{M})\, p(\mathbf{f}|\mathcal{M})}{p(\mathbf{g}|\mathcal{M})}$$

Likelihood :

$$p(\mathbf{g}|\mathbf{f};\mathcal{M}) = (2\pi v_\epsilon)^{-M/2} \exp\left\{-\frac{\|\mathbf{g}-\mathbf{H}\mathbf{f}\|^2}{2v_\epsilon}\right\}$$

- Prior which can be used for sparsity enforcing [Mohammad-Djafari, 2012] :
  - Simple heavy tailed models.
  - Hierarchical mixture models.
- Scale Mixture model of Student-t :

$$\begin{cases} p(f_j|v_f,\alpha,\beta) = \int \mathcal{N}(f_j|0,\frac{v_f}{z_j})\mathcal{G}(z_j|\alpha,\beta)\, \mathrm{d}z_j \\ p(\mathbf{f}) = \prod_j p(f_j) \end{cases} \tag{1}$$

Bayesian framework

- Hierarchical represantation with hidden variables :

$$\begin{cases} p(f_j|z_j, v_f) & = \mathcal{N}(f_j|0, \frac{v_f}{z_j}) \\ p(z_j|\alpha_0, \beta_0) & = \mathcal{G}(z_j|\alpha_0, \beta_0) \end{cases}$$

$$\Rightarrow p(\mathbf{f}, \mathbf{z}|\mathbf{g}, v_\epsilon, v_f) \propto p(\mathbf{g}|\mathbf{f}, v_\epsilon) p(\mathbf{f}|\mathbf{z}/v_f)\, p(\mathbf{z}|\alpha_0, \beta_0)$$

Posterior distribution :

$$p(\mathbf{f}, \mathbf{z}|\mathbf{g}) \propto {v_\epsilon}^{-M/2} \exp\left\{\frac{-\|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2}{2v_\epsilon}\right\} \prod_{j=1}^{N} (z_j/v_f)^{1/2} \exp\left\{-\frac{z_j f_j^2}{2v_f}\right\}$$

$$\times \frac{\beta_j^{\alpha_j} {z_j}^{\alpha_j - 1} \exp\{-\beta_j z_j\}}{\Gamma(\alpha_j)}$$

- For a given model $\mathcal{M}$, the expression of $p(\mathbf{f}, \mathbf{z}|\mathbf{g}; \mathcal{M})$ is usually complex.

Variational Bayesian Approximation (VBA)

- Objective : Approximate $p(\mathbf{f}, \mathbf{z}|\mathbf{g})$ by a separable law $q(\mathbf{f}, \mathbf{z}) = q_1(\mathbf{f})\, q_2(\mathbf{z})$.
- Criterion :

$$\mathrm{KL}(q : p) = \int q \ln \frac{q}{p} = \left\langle \ln \frac{q}{p} \right\rangle_q$$

- Free energy : $\mathrm{KL}(q : p) = \ln p(\mathbf{g}|\mathcal{M}) - \mathcal{F}(q)$ where

$$p(\mathbf{g}|\mathcal{M}) = \int \int p(\mathbf{f}, \mathbf{z}, \mathbf{g}|\mathcal{M})\, \mathrm{d}\mathbf{f}\, \mathrm{d}\mathbf{z}$$

- $\mathcal{F}(q)$ is the free energy associated to q defined as :

$$\mathcal{F}(q) = \left\langle \ln \frac{p(\mathbf{f}, \mathbf{z}, \mathbf{g}|\mathcal{M})}{q(\mathbf{f}, \mathbf{z})} \right\rangle_q$$

- For a given model $\mathcal{M}$, minimizing $\mathrm{KL}(q : p)$ is equivalent to maximizing $\mathcal{F}(q)$ and when optimized, $\mathcal{F}(q^*)$ gives a lower bound for $\ln p(\mathbf{g}|\mathcal{M})$.

VBA : Alternate optimization

-Alternate optimization scheme

$$\left\{\begin{array}{ll} \hat{q_1} & = \arg\max_{q_1}\{\mathcal{F}(q_1\,\hat{q_2})\} \Rightarrow q_1(\mathbf{f}) = \frac{1}{K_1}\exp\left\{\langle \ln p(\mathbf{g},\mathbf{f},\mathbf{z})\rangle_{q_2}\right\} \\ \hat{q_2} & = \arg\max_{q_2}\{\mathcal{F}(\hat{q_1}\,q_2)\} \Rightarrow q_2(\mathbf{z}) = \frac{1}{K_2}\exp\left\{\langle \ln p(\mathbf{g},\mathbf{f},\mathbf{z})\rangle_{q_1}\right\} \end{array}\right.$$

-Conjugacy property

$$\left\{\begin{array}{lll} q_1^{(k)}(\mathbf{f}) & = \prod_j \mathcal{N}(f_j|\widetilde{m}_j^{(k)},\widetilde{v}_j^{(k)}) & = \mathcal{N}(\mathbf{f}|\widetilde{\mathbf{m}}^{(k)},\widetilde{\mathbf{v}}^{(k)}) \\ q_2^{(k)}(\mathbf{z}) & = \prod_j \mathcal{G}(z_j|\widetilde{\alpha}_j^{(k)},\widetilde{\beta}_j^{(k)}) & = \mathcal{G}(\mathbf{z}|\widetilde{\boldsymbol{\alpha}}^{(k)},\widetilde{\boldsymbol{\beta}}^{(k)}) \end{array}\right.$$

-Initialization

$$\left\{\begin{array}{lll} q_1^{(0)}(\mathbf{f}) & = \prod_j \mathcal{N}(f_j|m_j^{(0)},v_j^{(0)}) & = \mathcal{N}(\mathbf{f}|\mathbf{m}^{(0)},\mathrm{Diag}(\mathbf{v}^{(0)})) \\ q_2^{(0)}(\mathbf{z}) & = \prod_j \mathcal{G}(z_j|\alpha_j^{(0)},\beta_j^{(0)}) & = \mathcal{G}(\mathbf{z}|\boldsymbol{\alpha}^{(0)},\boldsymbol{\beta}^{(0)}) \end{array}\right.$$

VBA : Alternate optimization

Updating of Gamma distribution

$$\begin{cases} \widetilde{\alpha}_j^{(k+1)} & = \alpha_0 + 1/2 \\ \widetilde{\beta}_j^{(k+1)} & = \frac{{m_{jk}}^2 + v_{jk}}{2v_f} + \beta_j^k \end{cases}$$

Updating of Gaussian distribution

$$\begin{cases} \widetilde{v}_j^{(k+1)} & = \left(\frac{1}{v_f} \widetilde{\alpha}_j / \widetilde{\beta}_j^{(k+1)} + \frac{1}{v_\epsilon} \text{diag}\left(\mathbf{H}^t\mathbf{H}\right)_j \right)^{-1} \\ \widetilde{m}_j^{(k+1)} & = \frac{v_j^{(k+1)}}{v_\epsilon} \left(\left[\mathbf{H}^t(\mathbf{g} - \mathbf{H}\mathbf{m}^{(k)}) \right]_j - \text{diag}\left(\mathbf{H}^t\mathbf{H}\right)_j \mathbf{m}_j^{(k)} \right) \end{cases}$$

- $\widetilde{\alpha}_j^{(k+1)}$ does not depend on the iterations.
- $p(\mathbf{z}|\mathbf{f},\mathbf{g})$ is separable, thus all the z_j can be computed simultaneously when f_j is computed (Classical alternate algorithm).

New optimization algorithms

- Gradient based [Fraysse and Rodet, 2011] :
  -Use the structure of probability densities set.
  -Construct a new density using the previous one thanks to *Radon Nikodym* theorem [Rudin, 1987].

$$q^{(k+1)} = h\, q^{(k)} \text{ where } h \in L^1(q^{(k)}).$$

-Exponential Gradient [Kivinen, 1997] :

$$q^{(k+1)} = \exp\left\{\lambda_k \, \mathrm{d}\mathcal{F}(q^{(k)}, \mathbf{f})\right\} q^{(k)}$$

$\Rightarrow$

$$q^{(k+1)} = q^{(k)} \left(\frac{q^{(r)}}{q^{(k)}}\right)^{\lambda_{subopt}} .$$

- $q^{(r)}$ an intermediate mesure.
-λ_{subopt} a suboptimal step of descent.
-$\frac{q^{(r)}}{q^{(k)}}$ the differential of $\mathcal{F}$ with respect to q.

New optimization algorithms

- Conjugate gradient like

$$q^{(k+1)} = q^{(k)}\left(\frac{q^{(r)}}{q^{(k)}}\right)^{\lambda_{subopt}}\left(\frac{q^{(k)}}{q^{(k-1)}}\right)^{\lambda_{subopt}\beta}$$

-λ_{subopt} and β the parameters of the conjugate gradient (need to be determinated *via* a scalar product).
- $q^{(r)}$ an intermediate mesure, $\frac{q^{(k)}}{q^{(k-1)}}$ is the descent direction at the previous estimate.

  - Difficulties : Absence of scalar product in the fonctional space $\rightarrow$Impossible to obtain the parameter that allows to get the new descent direction.
  - $\beta = 0 \rightarrow$ previous case, $\beta = 1 \rightarrow$ corrections of Vignes / bisector.

Numerical comparison of algorithms

- Gradient based

  Intermediate measure $q^{(r)}$: Alternate optimization

$$\begin{cases} v_j^{(r)} & = \left(\frac{1}{v_f} \widetilde{\alpha}_j / \widetilde{\beta}_j^{(k+1)} + \frac{1}{v_\epsilon} \text{diag}\left(\mathbf{H}^t \mathbf{H}\right)_j \right)^{-1} \\ m_j^{(r)} & = \frac{v_j^{(r)}}{v_\epsilon} \left(\left[\mathbf{H}^t (\mathbf{g} - \mathbf{H}\mathbf{m}^{(k)}) \right]_j - \text{diag}\left(\mathbf{H}^t \mathbf{H}\right)_j \mathbf{m}_j^{(k)} \right) \end{cases}$$

  Finally

$$\begin{cases} \widetilde{v}_j^{(\lambda)} & = \dfrac{v_j^{(r)} v_j^{(k)}}{v_j^{(r)} + \lambda (v_j^{(k)} - v_j^{(r)})} \\ \widetilde{m}_j^{(\lambda)} & = \dfrac{m_j^{(k)} v_j^{(r)} + \lambda (m_j^{(r)} v_j^{(k)} - m_j^{(k)} v_j^{(r)})}{v_j^{(r)} + \lambda (v_j^{(k)} - v_j^{(r)})} \end{cases}$$

Numerical comparison of algorithms

- Approximate Conjugate Gradient

$$\begin{cases} \widetilde{v}_j^{(\lambda\beta)} & = \dfrac{\left(v^{(r)} v^{(k)} v^{(k-1)}\right)_j}{\left(v^{(r)} v^{(k-1)}\right)_j + \lambda \left(v^{(k-1)} \delta^{(k)} + \beta v^{(r)} \delta^{(k-1)}\right)_j} \\ \widetilde{m}_j^{(\lambda\beta)} & = \dfrac{\left(m^{(k)} v^{(k-1)} v^{(r)}\right)_j + \lambda \left(\beta v^{(r)} \Delta^{(k)} + v^{(k-1)} \Delta^{(k-1)}\right)_j}{\left(v^{(k-1)} v^{(r)}\right)_j + \lambda \left(v^{(k-1)} \delta^{(k)} + \beta v^{(r)} \delta^{(k-1)}\right)_j} \end{cases}$$

with
$\delta^{(k)} = v^{(k)} - v^{(r)}$, $\delta^{(k-1)} = v^{(k-1)} - v^{(k)}$,
$\Delta^{(k)} = m^{(k)} v^{(k-1)} - m^{(k-1)} v^{(k)}$, $\Delta^{(k-1)} = m^{(r)} v^{(k)} - m^{(k)} v^{(r)}$.

Implementation issues

$$m_j^{(r)} = \frac{v_j^{(r)}}{v_\epsilon} \left(\left[\mathbf{H}^t (\mathbf{g} - \mathbf{H}\mathbf{m}^{(k)}) \right]_j - \text{diag}\left(\mathbf{H}^t\mathbf{H}\right)_j \mathbf{m}_j^{(k)} \right)$$

- Operations :

$$\begin{cases} \hat{\mathbf{g}} & = \mathbf{H}\mathbf{m} \\ \delta\mathbf{g} & = \mathbf{g} - \hat{\mathbf{g}} \\ \delta\mathbf{f} & = \mathbf{H}^t \delta\mathbf{g} \end{cases}$$

- In inverse problems we do not have access to the matrix $\mathbf{H}$, but we can compute
  1. Forward operator : $\mathbf{H}\mathbf{m} \rightarrow \hat{\mathbf{g}}$, (CT : Radon).
  2. Adjoint operator : $\mathbf{H}^t \delta\mathbf{g} \rightarrow \delta\mathbf{f}$, (CT : Backprojection).
- We may also need to compute the diagonal elements of $\left[\mathbf{H}^t\mathbf{H}\right]$ by developing algorithms that provide this (extract the diagonal by doing some operations using the canonical basis).

Conclusion and perspectives

- Conclusions :
  - Application of Variational Bayesian Approximation with Student-t prior.
  - New optimization algorithm in the space of the probability density.
  - Lower computational cost than MCMC.
- Perspectives :
  - Development of these methods in non linear (bi-linear or multi-linear) cases : Diffraction wave tomography (Microwave, optical..)

Thank you for your attention.


Fraysse, A. and Rodet, T. (2011).
A gradient-like variational Bayesian algorithm.
In *SSP 2011*, number S17.5, pages 605 – 608, Nice, France.


Kivinen, J. (1997).
Exponentiated gradient versus gradient descent for linear predictors.
Information and Computation.


Mohammad-Djafari, A. (2012).
Bayesian approach with prior models which enforce sparsity in signal and image processing.
EURASIP Journal on Advances in Signal Processing, Special issue on Sparse Signal Processing.


Robert, C. P. and Casella, G. (1998).
Monte carlo statistical methods.


Rudin, W. (1987).
Real and complex analysis.
McGraw-Hill Book Co., New York.


Smídl, V. and Quinn, A. (2005).
The Variational Bayes Method in Signal Processing (Signals and Communication Technology).
Springer-Verlag New York, Inc., Secaucus, NJ, USA.


Tikhonov, A. (1963).
Regularization of incorrectly posed problems.
Soviet. Math. Dokl., 4 :1624–1627.