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

Leila Gharsalli, Ali Mohammad-Djafari,
Aurélia Fraysse and Thomas Rodet

Laboratoire des Signaux et Systèmes,
Unité mixte de recherche 8506 (CNRS-SUPELEC-UNIV PARIS SUD)
Supélec, Plateau de Moulon, 3 rue Juliot-Curie, 91192 Gif-sur-Yvette, France

Abstract. Our aim is to solve a linear inverse problem using various methods based on the Variational Bayesian Approximation (VBA). We choose to take sparsity into account *via* a scale mixture prior, more precisely a student-t model. The joint posterior of the unknown and hidden variable of the mixtures is approximated via the VBA. To do this approximation, classically the alternate algorithm is used. But this method is not the most efficient. Recently other optimization algorithms have been proposed; indeed classical iterative algorithms of optimization such as the steepest descent method and the conjugate gradient have been studied in the space of the probability densities involved in the Bayesian methodology to treat this problem. The main object of this work is to present these three algorithms and a numerical comparison of their performances.

INTRODUCTION

In many inverse problems in signal processing we want to infer on an unknown signal through an observed one related between them through any linear or non linear transformation. In the case of linear relation and when discretized, we arrive to the following model of observation:

$$g = Hf + \varepsilon \tag{1}$$

where $f = [f_1, f_2, ..., f_N] \in \mathcal{R}^N$ represents the unknowns to be estimated, $g = [g_1, g_2, ..., g_M] \in \mathcal{R}^M$ the observed data, ε the errors of modelling and measurement and $H \in \mathcal{M}_{N\times M}$ the matrix of the system response. In practice, the dimension of this matrix is huge thus the computation of the direct solution $\widehat{f} = H^{-1}g$ even if it exists, is too high. Hence the character of ill-posed problem. A classical approach when dealing with these ill-posed problems is to introduce additional information. This is the shape of the Bayesian framework [9] which relates the probability of the hypothesis from both the currently achieved information (data) and previous knowldege (prior distribution) in order to compute the posterior distribution of parameters.

To obtain the posterior distribution, we need to relate the likelihood and the prior distribution according to the Bayes formula

$$p(f|g) = \frac{p(f,g)}{p(g)} = \frac{p(g|f)\,p(f)}{p(g)} \tag{2}$$

where $p(f|g)$ is the posterior distribution and $p(g|f)$ is the likelihood.

In practice, most of the Bayesian approaches use Markov Chain Monte Carlo (MCMC) [15] algorithms to estimate the posterior mean. However, for large dimensional problems involving a complicated covariance matrix, MCMC methods does not give a good results, in addition to their high computational cost.

Therefore, an alternative methodology called Variational Bayesian (VB) was proposed in 1995 by D. Mackay [12], having applications in computer sciences [10], signal processing for different applications such as recursives methods [3], deconvolution [17] and source separation [8] etc. The main idea is to approximate the joint posterior of the unknown by a separable density. This method could be more effecient than MCMC in large dimensional cases especially when the covariance matrix is no longer convertible. Besides, as the computation are analytical, the rate of convergence is better than MCMC methods. This VB methodology is the main point of this paper.

In the following, based on our model (1) and assuming that the noise is an *iid* Gaussian vector $\mathcal{N}(0,\sigma_\varepsilon^2 I)$, then thanks to (2) we obtain

$$p(g|f) = (2\pi v_\varepsilon)^{-N/2} \exp\left\{-\frac{\|g-Hf\|^2}{2v_\varepsilon}\right\} \tag{3}$$

We introduce prior information in the data and we take sparsity into account. In fact, different prior model have been used to enforce sparsity [13], such as the student model where the most general case is given by Gaussian Vector Scale Mixture (GVSM) [1]. We consider then $\forall j \in \{1,...,N\}$, f_j destributed following a separable heavy tailed distribution. We suppose that $f_j \sim u_j/\sqrt{z_j}$ where $u_j \sim \mathcal{N}(u_j|0,\sigma_f^2 I)$ and z_j a precision parameter chosen as a Gamma random variable i.e $z_j \sim \mathcal{G}(z_j|\alpha,\beta)$. Then $\forall j \in \{1,...,N\}$ we have $p(f_j,z_j) \sim \mathcal{N}(f_j|0,\sigma_f/z_j)\,\mathcal{G}(z_j|\alpha,\beta)$. And $\forall j \in \{1,...,N\}$ the density of f_j is given by

$$p(f_j) = \int_{\mathcal{R}} \frac{\sqrt{z_j}}{(2\pi v_f)^{1/2}} \exp\left\{-\frac{z_j f_j^2}{2v_f}\right\} z_j^{\alpha_j-1} \frac{\beta_j^{\alpha_j}\exp\{-\beta_j z_j\}}{\Gamma(\alpha_j)}\,\mathrm{d}z_j \tag{4}$$

Thereby we take the hidden variable z into account and thanks to (3) and (4) the student-t distribution is conjugate with the Gaussian likelihood which leads to the following posterior distribution

$$\begin{aligned} p(f,z|g) \propto & v_\varepsilon^{M/2} \exp\left\{\frac{-\|g-Hf\|^2}{2v_\varepsilon}\right\} \prod_{j=1}^{N} (z_j/v_f)^{1/2} \exp\left\{-\frac{z_j f_j^2}{2v_f}\right\} \\ & \times \frac{\beta_j^{\alpha_j} z_j^{\alpha_j-1}\exp\{-\beta_j z_j\}}{\Gamma(\alpha_j)} \end{aligned} \tag{5}$$

However this posterior distribution is not tractable analytically due to two main drawbacks: the first one is the link between f and z which can be resolved using the VB approach presented later in the paper. The second drawback is related to the dimension

of the vector f. When the dimension increases the correlation matrix becomes too large to be inverted effenciently. This problem can be solved by the two new algorithms presented in next sections.

The outline of the rest of this paper is as follows: in section 2 we briefly review the VBA and in section 3 and 4 we present the new algorithms. Finally we conclude by a numerical comparison between the three methods and present some perspectives of this work.

VARIATIONNAL BAYESIAN METHOD

The purpose of the classical VBA [6, 12] is to find the best approximating law within a class of approximating laws to the exact posterior distribution by minimizing the Kullback Leibler divergence. However, minimizing the Kullback divergence is equivalent to maximizing the free negative energy $\mathcal{F}(q)$ (a term derived from statistical physics) which is given by

$$\mathcal{F}(q) = \int_{\mathcal{R}^N} q(f,z) \ln \frac{p(g,f,z)}{q(f,z)} \, \mathrm{d}f \, \mathrm{d}z \tag{6}$$

where q is an approximate posterior probability density (pdf) taken as a separable law $q(f,z) = \prod q_1(f)\, q_2(z) = \prod_j q_{1j}(f_j) \prod_j q_{2j}(z_j)$.

We can thus resume the objective of the VBA method by finding

$$q^{opt} = \arg\max_q \mathcal{F}(q) \tag{7}$$

And assuming the separability we can obtain an analytic form for q (see [6] for variational calculus):

$$q_j{}^{k+1}(f_j) = \frac{1}{K_j} \exp\left\{ \langle \ln p(g,f,z) \rangle_{\prod_{j\neq i} q_i^k(f_i) q^{(k+1)}(z)} \right\} \tag{8}$$

where K_j is the normalizing factor.

Although the solution is obtained analytically, (8) does not have an explicit form. This solution is hardly tractable in practice, and it is approximated using iterative methods that impose the use of conjugate prior to obtain the posterior law belonging to a known family. Thus optimizing the posterior turns out to be an optimization of its distribution parameters.

Considering the model (1), we choose a conjugate prior for f and z and the optimal approximating distribution of f is known to belong to a Gaussian family, and a Gamma one for the prior probability density function of z. So we initialize the two densities as

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

where $m^{(0)} = [m_1^{(0)}, m_2^{(0)}, ..., m_N^{(0)}]' \in \mathcal{R}^N$ is the vector of means, $v^{(0)} \in \mathcal{R}^N$ is the vector of initial variances and $\mathrm{Diag}(v^{(0)})$ is a diagonal matrix with $(v^{(0)})$ on its diagonal.

Thanks to the conjugacy property, if we denote by k the number of iterations then during the iterations, $q_1^{(k)}(f)$ and $q_2^{(k)}(z)$ will stay in the same families and we obtain

$$\left\{\begin{array}{ll} q_1^{(k)}(f) & = \prod_j \mathcal{N}(f_j|\widetilde{m}_j^{(k)},\widetilde{v}_j^{(k)}) \quad = \mathcal{N}(f|\widetilde{m}^{(k)},\widetilde{v}^{(k)}) \\ q_2^{(k)}(z) & = \prod_j \mathcal{G}(z_j|\widetilde{\alpha}_j^{(k)},\widetilde{\beta}_j^{(k)}) \quad = \mathcal{G}(z|\widetilde{\alpha}^{(k)},\widetilde{\beta}^{(k)}) \end{array}\right. \tag{9}$$

Based on the equation (7) we deduce that the optimization problem can be tackled following the alternating optimization scheme

$$\left\{\begin{array}{ll} \hat{q_1} & = \arg\max_{q_1}\{\mathcal{F}(q_1\,\hat{q_2})\} \\ \hat{q_2} & = \arg\max_{q_2}\{\mathcal{F}(\hat{q_1}\,q_2)\} \end{array}\right. \tag{10}$$

For the approximation of $q_2(z)$ it is easy to see from (5) that $p(z|f,g)$ is separable, thus all the z_i can be computed simultaneously knowing only $q_1(f)$. And we obtain:

$$\left\{\begin{array}{ll} \widetilde{\alpha}_j^{(k+1)} & = \alpha_j + 1/2 \\ \widetilde{\beta}_j^{(k+1)} & = \frac{{m_j^{(k)}}^2 + v_j^{(k)}}{2v_f} + \beta_j \end{array}\right. \tag{11}$$

However, for the approximation of $q_1(f)$ we use the updating equations (8) and (5), that eventually result in the following updating equations (see [7] for details):

$$\left\{\begin{array}{ll} \widetilde{v}_j^{(k+1)} & = \left(\frac{1}{v_f}\widetilde{\alpha}_j/\widetilde{\beta}_j^{(k+1)} + \frac{1}{v_\varepsilon}\mathrm{diag}\,(H^tH)_j\right)^{-1} \\ \widetilde{m}_j^{(k+1)} & = \frac{v_j^{(k+1)}}{v_\varepsilon}\left(\left[H^t(g-Hm^{(k)})\right]_j - \mathrm{diag}\,(H^tH)_j\,m_j^{(k)}\right). \end{array}\right. \tag{12}$$

In the following we present methods meant to solve the functional optimization problem given by the VBA more efficiently than the approaches suggested by (8).

VARIATIONAL BAYESIAN DESCENT GRADIENT LIKE ALGORITHM

The maximization problem (7) is considered as an infinite dimensional concave problem subject to the positivity assumption and the normalization constraint; $\forall i,\ \int q_i(f_i)\,\mathrm{d}f_i = 1$. It would be convenient to determine the approximating density using the gradient descent method. Therefore, choosing a workspace adapted for Variational Bayesian methodolgy where classical optimization algorithms still hold is an important point. In [14], such a problem was treated, then inspired by that and by the *exponentiated gradient* descent in [11] a new iterative algorithm was proposed in [7]. It is essentially based on a measure theory result which is the Radon-Nykodym theorem, see for instance [16]. The main idea is that a measurable function $h^{(k)}$ exists, precisely $h^{(k)} \in L^1(q^{(k)})$ such that

$$q^{(k+1)}(\,\mathrm{d}f) = h^{(k)}(f)q^{(k)}(\,\mathrm{d}f) \tag{13}$$

where $f \in \mathcal{R}^N$. Likewise the classical scheme given by the gradient descent method [11] leads to the construction $h^{(k)}$ as the Frechet derivate of $\mathcal{F}$ at $q^{(k)}$ [4]. So we obtain

$$q^{(k+1)} = K \exp\left\{\lambda_k \, \mathrm{d}J(q^{(k)}, f)\right\} q^{(k)} \tag{14}$$

where λ_k is the algorithm step size at the iteration k which corresponds to the step of descent whereas $\mathrm{d}J(q^{(k)}, f)$ is given by

$$\forall j \in \{1, ..., N\}, \, \forall f \in \mathcal{R}^N, \, d_j J(q_j, f) = \langle \ln p(g, f) \rangle_{\prod_{j \neq i} q_i(f_i)} - \ln q_j - 1 \tag{15}$$

Then, let $k \geq 0$ be given and $q^{(k)}$ be constructed, we consider $q^{(\lambda)}$ given for $\lambda \geq 0$ by

$$\begin{aligned}
\forall f \in \mathcal{R}^n, \, q_1^{(\lambda)}(f) &\propto q_1^{(k)}(f) \left(\prod_j \exp\left\{ \mathrm{d}_j J(q_j^{(k)}, f_j) \right\} \right)^{\lambda_k} \\
&\propto q_1^{(k)}(f) \left(\prod_j \frac{\exp\left\{ \langle \ln p(g, f) \rangle_{\prod_{j \neq i} q_i^{(k)}(f_i)} \right\}}{q_{1j}^{(k)}(f_j)} \right)^{\lambda_k} \\
&\propto q_1^{(k)}(f) \left(\prod_j \frac{q_j^{(r)}(f_j)}{q_{1j}^{(k)}(f_j)} \right)^{\lambda_k}
\end{aligned} \tag{16}$$

Finally we obtain

$$\left\{ \begin{array}{ll}
\widetilde{v}_j^{(\lambda)} & = \dfrac{v_j^{(r)} v_j^{(k)}}{v_j^{(r)} + \lambda (v_j^{(k)} - v_j^{(r)})} \\
\widetilde{m}_j^{(\lambda)} & = \dfrac{m_j^{(k)} v_j^{(r)} + \lambda (m_j^{(r)} v_j^{(k)} - m_j^{(k)} v_j^{(r)})}{v_j^{(r)} + \lambda (v_j^{(k)} - v_j^{(r)})}
\end{array} \right. \tag{17}$$

where

$$\left\{ \begin{array}{ll}
v_j^{(r)} & = \left(\frac{1}{v_f} \widetilde{\alpha}_j / \widetilde{\beta}_j^{(k+1)} + \frac{1}{v_\varepsilon} \mathrm{diag}\left(H^t H\right)_j \right)^{-1} \\
m_j^{(r)} & = \frac{v_j^{(r)}}{v_\varepsilon} \left(\left[H^t (g - H m^{(k)}) \right]_j - \mathrm{diag}\left(H^t H\right)_j m_j^{(k)} \right)
\end{array} \right. \tag{18}$$

This new algorithm allows to minimize jointly all (q_j) unlike the classical Variational Bayesian algorithm. Besides, a suboptimal stepsize λ_{subopt} can be chosen in order to optimize the convergence rate. See [7] for more details on this method.

VARIATIONAL BAYESIAN CONJUGATE GRADIENT LIKE ALGORITHM

As the gradient descent algorithm is known for its slow rate of convergence, a natural idea to extend the previous method was to consider the case of the conjugate gradient algorithm [2]. However, its use in the metric space of the probability density encountered problems, because of the absence of the notion of inner product in that space which allows in a normal situation to have the conjugate directions.

But we can still note, if we follow the previous construction steps and thanks to the Radon-Nykodym theorem, that we can arrive to the following update equation

$$q_1^{(\lambda\beta)}(f) = q_1^{(k)}(f) \left(\frac{q_1^{(r)}(f)}{q_1^{(k)}(f)} \right)^{(\lambda)} \left(\frac{q_1^{(k)}(f)}{q_1^{(k-1)}(f)} \right)^{(\lambda\beta)} \tag{19}$$

where $\frac{q_1^{(k)}}{q_1^{(k-1)}}$ is the descent direction at the previous estimate, $q_1(r)$ is determined by (18) and β is the conjugate parameter which is difficult to determinate. That's why in our application later, we are going to consider only the case where $\beta = 1$ which is a particular case known for the correction of Vignes [5]. We notice also that for $\beta = 0$ we are simply in the domain of the algorithm presented in [7].

Applying this algorithm to estimate $q_2(f)$ we obtain

$$\begin{cases} \widetilde{v}_j^{(\lambda\beta)} & = \dfrac{\left(v^{(r)}v^{(k)}v^{(k-1)}\right)_j}{\left(v^{(r)}v^{(k-1)}\right)_j + \lambda\left(v^{(k-1)}\delta^{(k)} + \beta v^{(r)}\delta^{(k-1)}\right)_j} \\ \widetilde{m}_j^{(\lambda\beta)} & = \dfrac{\left(m^{(k)}v^{(k-1)}v^{(r)}\right)_j + \lambda\left(\beta v^{(r)}\Delta^{(k)} + v^{(k-1)}\Delta^{(k-1)}\right)_j}{\left(v^{(k-1)}v^{(r)}\right)_j + \lambda\left(v^{(k-1)}\delta^{(k)} + \beta v^{(r)}\delta^{(k-1)}\right)_j} \end{cases} \tag{20}$$

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

Analyzing (12), (18) and (20) we notice that the common part of those three algorithms is the alternating optimzation scheme and

$$\begin{cases} \hat{g} & = Hm \\ \delta g & = g - \hat{g} \\ \delta f & = H^t \delta g \end{cases} \tag{21}$$

are the most time consuming operators that can cause some implementation issues. But as in many inverse problems where we do not have access to the matrix H, we can still compute forward operator: $Hm \to \hat{g}$, which corresponds to the Radon operator in Computed Tomography (CT) for example , and the adjoint operator: $H^t\delta g \to \delta f$, like Backprojection in CT. We may also need to compute the diagonal elements of $[H^tH]$ by developing algorithms that provide this.

RESULTS

To illustrate our results, we choose to treat the linear inverse problem given by (1) with a non invertible matrix H coming from a small tomographic problem. The test image (64×64) and the collected data are displayed in Fig. 1. The simulation parameters are presented in Tab. 1. There are 32 various angular positions uniformly spaced over $[0, 180[$. For each projection, 95 detector cells are performed. And we add to data a white Gaussian noise (iid) with standard deviation equal to 0.3. The programming language used to implement algorithms is MATLAB.

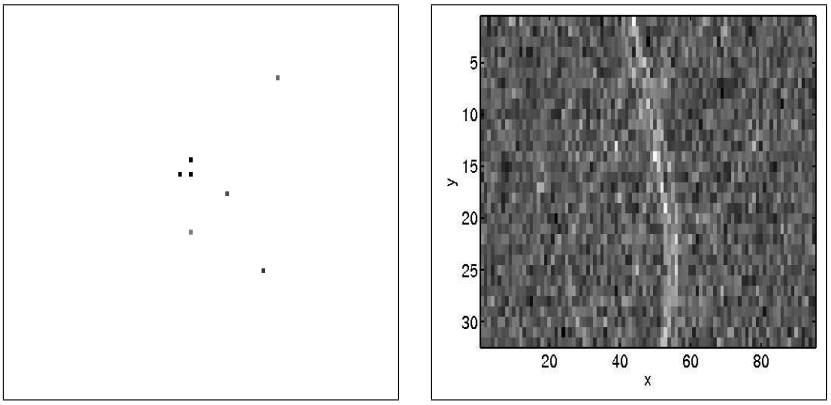
TABLE 1. Simulation parameters

Image test	Peaks on the image	Angles of projection	Detector cells	Noise
Sparse phantom	7	32	95	0.3

The three approaches are initialized with a zero mean and a variance equal to one, the hyperparameters v_e, v_f and α are respectively fixed to 1, 0.05, and 0.1. The reconstruction results collected in Fig. 3. Furthermore, Tab. 2 show us that the two new algorithms have almost the same reconstruction quality than the classical VB. Besides, by comparing the exucution time we see that the two new algorithms are faster than the classical VB. Fig. 2 shows also the performance of the free negative energy criterion for the new methods and clearly demonstrates their interesting rate of convergence for 150 iterations of the algorithms. We can also check other results in the referenced paper [7] of these methods and their reconstruction properties on a large dimensional dictionary application problem.

TABLE 2. Execution Time

Methods	FBP	Classical VB	VB Grad	VB Conjugate Grad ($\beta = 1$)
Iterations	1	15	150	150
SNR (db)	-2.04	5.69	2.19	7.11
Time cpu (s)	0.036	579.6	8.77	8.93



(a) (b)

FIGURE 1. (a) True image of 7 peaks (b) Sinogram

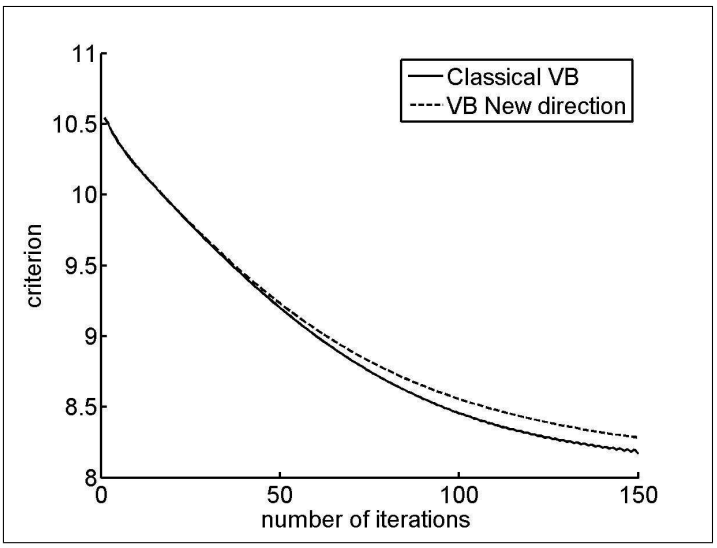


FIGURE 2. Performance of the free negative energy criterion for the two new algorithms

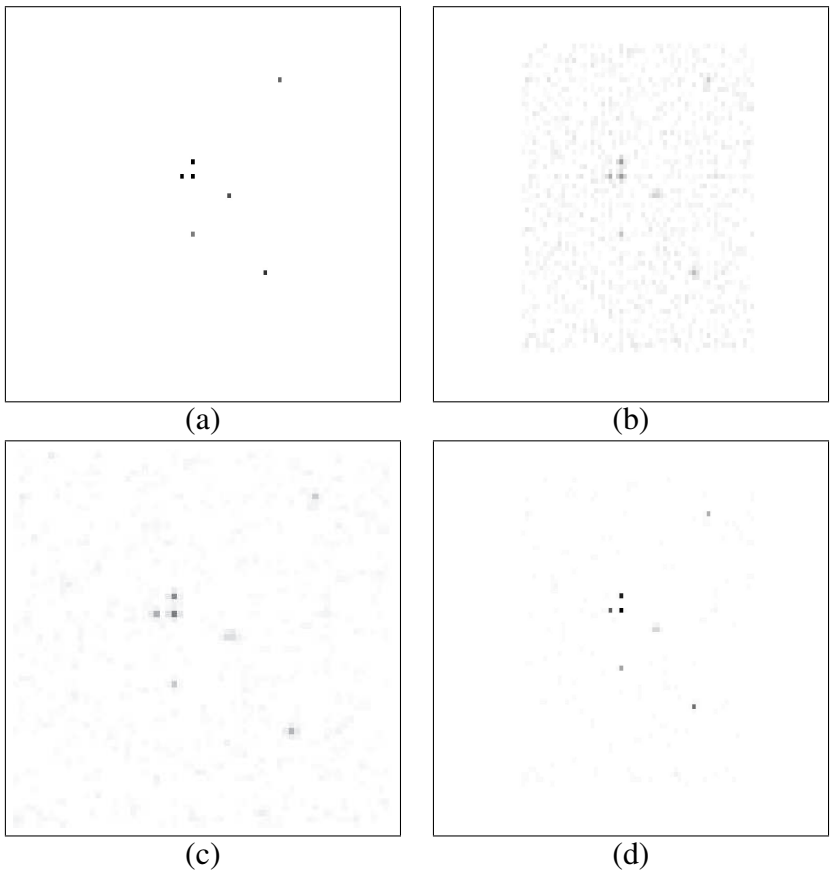


FIGURE 3. (a) Original image with 7 peaks (b) VB Classique (c) VB Grad (d) VB Conjugate Grad ($\beta = 1$)

CONCLUSIONS AND PERSPECTIVES

The sparsity is requiered property in many signal and image processing applications. In this paper first we reviewed the main idea of the Variational Bayesian Approximation for ill-posed inverse problems. The student-t model was used to enforce sparsity. Then we presented two new algorithms involved in the Bayesian methodology and able to provide better results than the classical methods. In a next step we want to test the performences of these methods in non linear (bi-linear or multi-linear) cases like diffraction wave tomography (Microwave, optical..).

REFERENCES

1. Andrews, D. F. and Mallows, C. L. (1974). Mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102.
2. Aude, T. and Sébastien, T. (2010). *Introduction à l'optimisation numérique*. Cambridge.
3. Babacan, D., Molina, R., and Katsaggelos, A. (2009). Variational bayesian blind deconvolution using a total variation prior. *IEEE Transactions on Image Processing*, 18(1):12–26.
4. Behmardi, D. and Nayeri, E. (2008). Introduction of fréchet and gâteaux derivative. *Applied Mathematical Sciences*, 2(20):975–980.
5. Brette, S., Carfantan, H., Giovannelli, J.-F., Martin, T., Bercher, J.-F., Heinrich, C., Idier, J., and Soussen, C. (2006). Gradient à pas adaptatif avec corrections, une mise en oeuvre matlab : GPAC.m. Technical report, GPI – L2S, France.
6. Choudrey, R. A. (2002). *Variational Methods for Bayesian Independent Component Analysis*. PhD thesis, University of Oxford.
7. Fraysse, A. and Rodet, T. (2012). A measure-theoretic variational Bayesian algorithm for large dimensional problems. Technical report. http://hal.archives-ouvertes.fr/docs/00/70/22/59/PDF/var_bayV8.pdf.
8. Ichir, M. M. and Mohammad-djafari, A. (2005). A Mean field approximation approach to blind source separation with Lp priors. *In Eusipco, September, Antalya, Turkey*.
9. Idier, J., editor (2001). *Approche bayésienne pour les problèmes inverses*. Traité IC2, Série traitement du signal et de l'image, Hermès, Paris.
10. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.
11. Kivinen, J. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63.
12. Mackay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.
13. 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.
14. Molchanov, I. (2000). Tangent sets in the space of measures: with applications to variational analysis. *J.Math.Anal Appl*, 249:539–552.
15. Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
16. Rudin, W. (1987). *Real and complex analysis*. McGraw-Hill Book Co., New York.
17. 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.