Sur un nouvel algorithme bayésien variationnel

Aurélia Fraysse, Thomas Rodet,

Laboratoire des Signaux et Systèmes
CNRS, Université Paris Sud, Supélec, 3 rue Joliot Curie, 91192 Gif Sur Yvette, France
Aurelia.Fraysse@lss.supelec.fr,Thomas.Rodet@lss.supelec.fr

Résumé – Nous nous intéressons ici aux problèmes inverses mal posés en très grande dimension. En pratique, les approches non supervisées classiques basées sur la méthodologie bayésienne nécessitent un temps de calcul trés important. C'est pourquoi nous proposons ici un algorithme non supervisé de reconstruction, plus rapide que les méthodes existantes, basé sur la méthodologie du bayésien variationnel. Les performances de cet algorithme sont évaluées sur un problème inverse où l'information *a priori* met en valeur la parcimonie de l'image. Une comparaison avec les méthodes bayésiennes standard montre que notre approche permet d'avoir une très bonne qualité de reconstruction en un temps relativement faible.

Abstract – This paper is devoted to the definiton and application of an unsupervised reconstruction algorithm for ill-posed problems. The main goal is to determine an efficient method for large dimensional datasets. The main tool involved is given by the Variational Bayesian methodology. In a first part of this paper we define our method whereas in a second part we enhance the performances of our algorithm by comparating it with classical methods on a sparse image. The simulation results show that our approach gives a good reconstruction in a few time.

1 Introduction

Les dernières avancées technologiques ont fait apparaître des masses de données de tailles de plus en plus importantes. Une approche classique pour étudier de tels signaux est alors de prendre en compte l'information inhérente au problème, via l'estimation bayésienne, et de modéliser la source par une variable aléatoire suivant une loi donnée, [3]. L'avantage de telles méthodes est qu'elles sont non supervisées, au sens où elles permettent l'ajustement automatique des paramètres liés au modèle, au travers du compromis entre l'information *a priori*, imposée à la source, et l'*a posteriori* venant des données. La principale difficulté dans ce cas est alors de déterminer exactement, à partir d'un *a priori* donné, la distribution *a posteriori* correspondante. En pratique, l'*a posteriori* a une structure telle qu'elle ne peut être déterminée directement. Une étape d'approximation est donc nécéssaire. Cette approximation peut être stochastique par le biais des algorithmes MCMC (Monte Carlo Markov Chain), voir [4] par exemple, ou analytiques dans les méthodes du Bayésien variationnel, [6, 2].

Dans cet article, nous nous proposons de définir un nouvel algorithme bayésien basé sur le principe du bayésien variationnel. L'intérêt de cet algorithme est qu'il permet, en un nombre limité d'itérations, de trouver une approximation de la loi *a posteriori* pour des problèmes en grande dimension. Le principe de la méthode proposée est d'adapter une méthode classique d'optimisation convexe à l'espace des densités de probabilités. A titre d'exemple, nous montrons la mise en œuvre de cette méthode sur un problème de tomographie où l'information *a priori* utilisée favorise les solutions parcimonieuses (images composées de pics).

2 Présentation de l'algorithme

2.1 Le Bayésien variationnel

Pour une meilleure compréhension de la méthode proposée dans cet article, nous commençons par rappeler les concepts de base du bayésien variationnel introduit dans [6]. Dans la suite, nous noterons $\boldsymbol{Y} \in \mathbb{R}^M$ le vecteur des données et $\boldsymbol{W} \in \mathbb{R}^N$ le vecteur des sources que l'on cherche à estimer. Le principe de la méthode est d'approcher la distribution *a posteriori* $p(\mathbf{W}|\mathbf{Y})$ par une densité de probabilité séparable en minimisant la divergence de Kullback-Leibler entre les deux. Cependant, trouver un minimum dans ce cas peut s'avérer relativement compliqué. C'est pourquoi on préfère utiliser la relation suivante, [2] :

$$\log p(\mathbf{Y}) = F(q(\mathbf{W})) + \mathcal{KL}[q(\mathbf{W})||p(\mathbf{W}|\mathbf{Y}))], \quad (1)$$

où F est l'energie libre donnée par

$$F[\mathbf{W}] = < \log p(\mathbf{Y}, \mathbf{W}) >_{q(\mathbf{W})} + \mathcal{H}(\mathbf{W}) \quad (2)$$

Dans l'équation précédente, $\mathcal{H}(\mathbf{W})$ est l'entropie de $\boldsymbol{W}$ sous la loi de q tandis que

$$< \log p(\mathbf{Y}, \mathbf{W}) >_{q(\mathbf{W})} = \int \log(p(\mathbf{Y}, \mathbf{W}))q(\mathbf{W}).$$

Par conséquent, minimiser la divergence revient à maximiser l'énergie libre.

Dans la suite nous cherchons la densité de probabilité séparable q qui maximise F. Comme F est une fonctionnelle concave, cette minimisation peut se faire par calcul des variations, [2], et nous donne, pour tout $i = 1, \ldots, N$

$$q_i(w_i) = \frac{1}{Z_i} \exp\left(< \log p(\mathbf{Y}, \mathbf{W}) >_{\prod_{j \neq i} q_j(w_j)} \right). \quad (3)$$

Cette solution analytique est cependant inutilisable en pratique puisqu'elle n'a pas de forme explicite. C'est pourquoi des algorithmes itératifs sont mis en œuvre pour calculer l'optimum à partir de (3). Le plus classique est d'utiliser les algorithmes de minimisation alternée. Malheureusement l'utilisation de ces algorithmes augmente considérablement le temps de calcul. Elle ne permet donc pas de traiter des données en très grande dimension. C'est pourquoi nous proposons une méthode différente plus rapide que les méthodes existantes.

2.2 Méthode proposée

Le principe de notre algorithme est d'utiliser, dans la définition du problème d'optimisation précédent, la structure de l'espace de dimension infinie sous-jacent, à savoir l'ensemble des densités de probabilités. Pour cela, plusieurs approches peuvent être considérées. La première consisterait à définir cet ensemble comme un sous-espace de l'ensemble des fonctions intégrables. Dans ce cas, la contrainte de masse totale doit être prise en compte et l'algorithme le plus adapté serait une méthode de type gradient projeté. L'autre point de vue, qui est celui adopté ici, est de considérer cet ensemble comme un sous ensemble de l'ensemble des mesures de probabilités. C'est pourquoi nous proposons d'utiliser une version de la descente de gradient, adapté à la structure de cet ensemble. Cette méthode d'optimisation se rapproche de la méthode du "gradient exponentialisé", voir [5], introduit dans la communauté du « Machine Learning ».

Pour cela, supposons qu'à l'itération k, où $k \geq 0$, on ait construit $\{q_1^k, \ldots, q_N^k\}$ et que $q^k(\mathbf{W}) = \prod q_i^k(w_i)$. On veut alors que q^{k+1} soit une densité de probabilités telle que $q^{k+1}d\lambda$, où λ est la mesure de Lebesgue sur $\mathbb{R}^N$, soit absolument continue par rapport à $q^k d\lambda$. Dans ce cas, le théorème de Radon-Nikodym, voir [7], nous offre un cadre naturel pour mettre à jour notre loi en considérant :

$$q^{k+1} = hq^k \quad (4)$$

où $h \in L^1(q^k)$ est une fonction à valeurs dans $\mathbb{R}^+$. Il nous reste alors à déterminer la fonction h. Pour cela, on s'inspire de la méthode de descente de gradient. Autrement dit on choisit

$$h(W) = \exp(\nabla F(q(W)))^\alpha. \quad (5)$$

où ∇F représente la différentielle de F au sens de Fréchet et où $\alpha > 0$ est le pas de l'algorithme. Cette expression de h est la plus à même de nous assurer que $F(q^k)$ est une fonctionnelle croissante.

Or

$$\forall i = 1, \ldots, N; \frac{\partial F}{\partial q_i} = < \log p(\mathbf{Y}, \mathbf{W}) >_{\prod_{j \neq i} q_j^k(w_j)} - \log q_i - 1.$$

Ce qui nous donne finalement pour tout $i = 1, \ldots, N$,

$$q_i^k(w_i) = q_i^k(w_i) \left(\frac{1}{K_i} \frac{\exp\left(< \log p(\mathbf{Y}, \mathbf{W}) >_{\prod_{j \neq i} q_j^k(w_j)} \right)}{q_i^k} \right)^\alpha. \quad (6)$$

On peut constater que dans ce cas, la suite $(F(q^k))_{k \in \mathbb{N}}$ est bien une suite croissante.

3 Application à la tomographie impulsionnelle

3.1 Modélisation du problème

Nous mettons ici en œuvre notre méthode sur un problème classique de tomographie en utilisant un *a priori* favorisant les images parcimonieuses. Pour cela, nous écrivons le modèle direct linéaire :

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b}, \quad (7)$$

où $\mathbf{H}$ est une matrice de projection associée à la transformée de Radon, $\boldsymbol{y}$ est le vecteur des données, $\boldsymbol{x}$ est le vecteur des inconnues et $\mathbf{b} \in \mathbb{R}^M$ est une réalisation d'un bruit Gaussien iid de variance σ_b^2. De plus, on introduit un *a priori* sur $\boldsymbol{x}$ suivant une loi de Student-t qui peut favoriser naturellement les solutions parcimonieuses. La mise en oeuvre d'un *a priori* de Student-t dans des méthodes bayésiennes variationnelles a été notamment décrite dans [1].

La distribution de Student-t peut être représentée sous forme d'un mélange scalaire de gaussiennes (GSM), voir [8], où la loi de chaque coordonnée X_i sera Gaussienne sachant une variable cachée Z_i, ce qui s'écrit :

$$\forall i \in \{1, \ldots, N\},$$
$$p(x_i) = \frac{\tilde{b}_i^{\tilde{a}_i}}{\Gamma(\tilde{a}_i)} \int_{\mathbb{R}} \frac{\sqrt{z_i}}{(2\pi)^{N/2} |\sigma_1^2|^{1/2}} e^{-\frac{z_i x_i^2}{2\sigma_1^2}} z_i^{\tilde{a}_i - 1} e^{-z_i \tilde{b}_i} dz_i. \quad (8)$$

Où, $Z_i \sim \mathcal{G}(\tilde{a}_i, \tilde{b}_i)$ suit une loi Gamma, et $X_i \sim \mathcal{N}(0, \sigma_1^2/z_i)$ suit une loi Gaussienne.

A partir de ce modèle *a priori*, nous voulons mettre en oeuvre une approche non-supervisée nous allons donc aussi estimer la variance du bruit σ_b^2 et la variance de l'*a priori* sur $\boldsymbol{X}$, à savoir σ_1^2. Ainsi nous réglerons le compromis entre l'information parcellaire contenue dans les

données et l'information *a priori* de parcimonie. Comme nous suivons une démarche bayésienne variationnelle nous introduisons des *a priori* conjugués avec la vraisemblance, suivant donc des lois inverse Gamma sur ces deux variances. En rassemblant toutes les lois on détermine la loi *a posteriori* jointe à une constante près,

$$p(\boldsymbol{x},\boldsymbol{z},\sigma_b^2,\sigma_1^2|\boldsymbol{y}) \propto \sigma_b^{-M}\exp\left[-\frac{\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2}{2\sigma_b^2}\right]\sigma_1^{-N} \prod_i \sqrt{z_i}\exp\left[-\frac{z_i x_i^2}{2\sigma_1^2}\right]\frac{\tilde{b}_i^{\tilde{a}_i} z_i^{\tilde{a}_i-1}e^{-z_i\tilde{b}_i}}{\Gamma(\tilde{a}_i)} \times \frac{\tilde{b}_b^{\tilde{a}_b}\sigma_b^{-2(\tilde{a}_b-1)}e^{-\frac{\tilde{b}_b}{\sigma_b^2}}}{\Gamma(\tilde{a}_b)}\frac{\tilde{b}_1^{\tilde{a}_1}\sigma_1^{2(\tilde{a}_1-1)}e^{-\frac{\tilde{b}_1}{\sigma_1^2}}}{\Gamma(\tilde{a}_1)}. \quad (9)$$

3.2 Approche bayésienne variationnelle

Appliquons la méthodologie bayésienne variationnelle décrite par [6] à notre densité *a posteriori* eq. (9). Choisissons tout d'abord le degré de séparabilité de la loi approchante q. Comme l'on veut un algorithme rapide permettant de résoudre des problèmes de grande dimension, nous choisissons une loi séparable selon toutes les dimensions $q(\boldsymbol{x},\boldsymbol{z},\sigma_b^2,\sigma_1^2) = \prod_i q_i(x_i)\prod_j \tilde{q}_j(z_j)q(\sigma_b^2)q(\sigma_1^2)$. L'hypothèse faite sur les lois *a priori*, en l'occurence que ce soit des lois conjuguées, implique que les lois approchantes minimisant la divergence de Kullback-Leiber appartiennent à des familles connues :

$$\begin{aligned} q^0(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{m}_0, \mathrm{Diag}(\boldsymbol{\sigma}_0^2)) \\ \tilde{q}^0(\boldsymbol{z}) &= \prod_j \mathcal{G}(a_j^0, b_j^0) \\ q_b(\sigma_b^2) &= \mathcal{IG}(a_b^0, b_b^0) \\ q_1(\sigma_1^2) &= \mathcal{IG}(a_1^0, b_1^0) \end{aligned}$$

où $\boldsymbol{\sigma}_0^2$ est le vecteur des N variances initiales et l'opérateur Diag transforme un vecteur en une matrice diagonale.

L'optimum est déterminé par le schéma de minimisation suivant :

$$\begin{aligned} \tilde{q}^{k+1}(\boldsymbol{z}) &= \arg\max_{\tilde{q}(\boldsymbol{z})} F\big(q^k(\boldsymbol{x})\tilde{q}(\boldsymbol{z})q_b^k(\sigma_b^2)q_1^k(\sigma_1^2)\big) \\ q^{k+1}(\boldsymbol{x}) &= \arg\max_{q(\boldsymbol{x})} F\big(q(\boldsymbol{x})\tilde{q}^{k+1}(\boldsymbol{z})q_b^k(\sigma_b^2)q_1^k(\sigma_1^2)\big) \\ q_b^{k+1}(\sigma_b^2) &= \arg\max_{q(\sigma_b^2)} F\big(q^{k+1}(\boldsymbol{x})\tilde{q}^{k+1}(\boldsymbol{z})q(\sigma_b^2)q_1^k(\sigma_1^2)\big) \\ q_1^{k+1}(\sigma_1^2) &= \arg\max_{q(\sigma_1^2)} F\big(q^{k+1}(\boldsymbol{x})\tilde{q}^{k+1}(\boldsymbol{z})q_b^{k+1}(\sigma_b^2)q(\sigma_1^2)\big) \end{aligned}$$

Notre algorithme est mis en œuvre uniquement pour la deuxième maximisation, la loi conditionnelle *a posteriori* sur le vecteur $\boldsymbol{Z}$ étant déjà séparable et permettant donc une utilisation rapide de l'approche bayésienne variationnel standard.

4 Expérimentation

Nous exposons ici des premiers résultats sur données simulées. Nous avons considéré un problème mal posé de petite taille, 64× 64, afin de comparer notre méthode avec des méthodes existantes (la rétroprojection filtrée FBP, l'algorithme classique Bayésien Variationnel BV, et l'approche MCMC avec un échantillonneur de Gibbs). L'image simulée est composée de 7 pics d'amplitudes différentes, comprises entre 0.5 et 1 (voir fig. 1(a)).

Les résultats sont résumés dans le tableau 1 et sur la figure 1. Nous avons appelé BVGPI notre approche avec les hyperparamètres, à savoir variance du bruit et paramètres de la loi *a priori*, fixés et BVGPI non supervisé (BVGPI-NS) la même méthode avec conjointement, l'estimation de tous les paramètres. Les approches concurrentes ont été initialisées avec les même images et avec les mêmes hyperparamètres que notre approche BVGPI. On remarque clairement que fixer les hyperparamètres à des valeurs favorisant la parcimonie permet d'améliorer nettement la qualité des reconstructions (voir fig. 1 (d), (e) et (f)). On observe aussi sur la fig. 1 (c) que la méthode MCMC est la méthode bayésienne qui donne les moins bon résultats, ce qui est en contradiction avec les résultats théoriques. Cette qualité moindre vient essentiellement de la faible vitesse de convergence de la méthode qui induit un temps de calcul trop important. Autrement, on voit sur la fig. 1(e) que notre méthode (BVGPI) permet d'obtenir des résultats sensiblement de même qualité que l'approche classique (voir fig. 1(d)) pour un temps de calculs 13 fois plus faible. Elle offre aussi une réelle amélioration en terme de qualité d'image par rapport aux MCMC, tout en gagnant un facteur pratiquement égal à 100 en temps de calculs.

Enfin, nous avons développé une méthode non supervisée où le compromis entre l'information *a priori* et la vraisemblance est réglé automatiquement par la méthode. Les résultats sont de meilleurs qualités que dans l'approche supervisé (voir fig. 1 (f)). On peut en conclure que nous avons pas fait suffisamment d'essais avec l'approche non-supervisé pour avoir un très bon réglage des hyperparamètres.

TABLE 1 – Performances des différentes approches.

Méthodes	FBP	BV	BVGPI	MCMC	BVGPI-NS
CPU (s)	0,05	586,2	44,77	37079	87,34
nb d'iter.	1	15	500	1000	1000
err rel L_2	1,6%	0,26%	0,25%	1,15%	0,67%
err rel L_1	22%	7,95%	7,76%	19,1%	14,8%

5 Conclusions

Nous avons exposé une nouvelle approche pour résoudre des problèmes inverses basée sur le Bayésien variationnel

dont on a montré en pratique qu'elle est nettement plus efficace que les principales approches entièrement bayésiennes concurrentes. La méthode développée ici ne fait pas intervenir d'inversion de matrices, contrairement aux méthodes classiques de MCMC et de bayésien variationnel. Même sur un exemple de petite dimension, nous pouvons voir qu'elle améliore déjà le temps de calcul. De plus, cette approche est suffisamment générique pour développer des approches non-supervisées avec un coût calculatoire relativement faible par rapport aux méthodes classiques.

Références

[1] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders. Variational Bayesian image restoration based on a product of t-distributions image prior. *IEEE Trans. Image Processing*, 17(10) :1795–1805, October 2008.

[2] R. A. Choudrey. *Variational Methods for Bayesian Independent Component Analysis*. PhD thesis, University of Oxford, 2002.

[3] G. Demoment. Image reconstruction and restoration : Overview of common estimation structure and problems. *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-37(12) :2024–2036, December 1989.

[4] J.-F. Giovannelli. Unsupervised bayesian convex deconvolution based on a field with an explicit partition function. *IEEE Trans. Image Processing*, 17(1) :16–26, January 2008.

[5] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1) :1–63, 1997.

[6] D. J. C. MacKay. Ensemble learning and evidence maximization. http ://citeseerx.ist.psu.edu/viewdoc/summary ?doi=10.1.1.54.4083, 1995.

[7] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, 1987.

[8] M. Wainwright and E. Simoncelli. Scale Mixtures of Gaussians and the statistics of natural images. *NIPS*, 12, 2000.

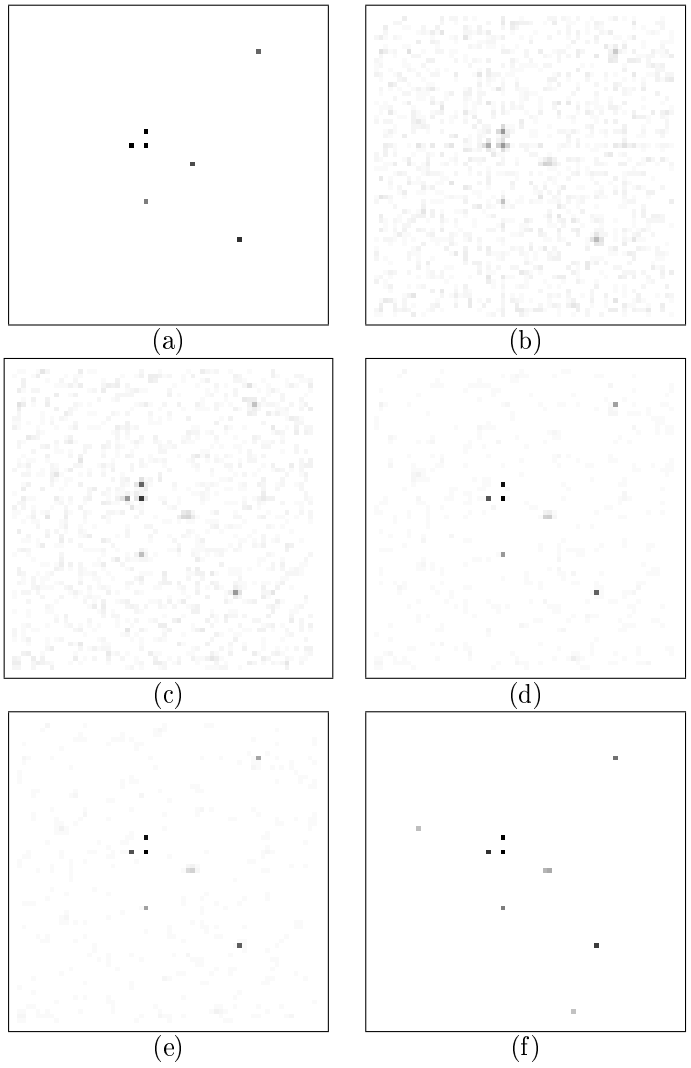

FIGURE 1 – Les images sont représentées avec la même échelle de gris : (a) l'image vrai composée de 7 pics, (b) La méthode de rétroprojection filtrée (FBP), (c) approche MCMC avec un échantillonneur de Gibbs, (d) Approche bayésienne variationnelle classique, (e) notre approche dans le cas où les hyper paramètres sont fixés arbitrairement, (f) notre approche non-supervisée.