Approches rapides du bayésien variationnel pour des problèmes de grande dimension

Yuling Zheng, Thomas Rodet, Aurélia Fraysse,

Laboratoire des Signaux et Systèmes
CNRS, Université Paris Sud, Supélec, 3 rue Joliot Curie, 91190 Gif Sur Yvette, France
{zheng,rodet,fraysse}@lss.supelec.fr

Résumé – Dans cet article nous proposons deux algorithmes d'estimation non supervisés basés sur la méthodologie du bayésien variationnel. Nous montrons aussi l'application pratique de ces algorithmes à un problème de super résolution d'images via un *a priori* TV. Dans ce cadre, nos algorithmes sont comparés à une méthode récente de reconstruction. Nous montrons ainsi que grâce aux méthodes développées nous pouvons obtenir une qualité de reconstruction semblable à l'état de l'art avec un temps de calcul nettement amélioré.

Abstract – In this paper, we propose two unsupervised algorithms based on variational Bayesian methodology. We demonstrate the pratical application of these algorithms through a super-resolution problem using a TV *a priori*. These algorithms are also compared with a recently proposed reconstruction method. The comparison shows that thanks to the proposed methods, we could obtain a reconstruction quality similar to the state of art while significantly reducing the computation time.

1 Introduction

L'objectif de ce travail est de résoudre efficacement des problèmes inverses de grande taille de manière non-supervisée. Pour cela, nous nous plaçons dans une approche entièrement bayésienne où les paramètres d'intérêts ainsi que les hyperparamètres, c'est à dire les paramètres de réglages de la méthode, sont estimés conjointement. Dans ce cas, la loi *a posteriori* nécessaire pour construire des estimateurs efficaces est en général complexe et ne peut être utilisée directement. Une manière de contourner ce problème est d'utiliser les méthodes stochastiques de type Monte Carlo par Chaîne de Markov (MCMC) [7]. Ces approches sont cependant mal adaptées aux problèmes de grandes dimensions car trop coûteuses en temps de calcul.

Nous nous sommes donc intéressés aux approches bayésiennes variationnelles, [9], dont le principe est de faire une approximation analytique de la loi *a posteriori* par des densités approchantes plus simples, séparables par exemple. Ces lois approchantes sont déterminées en minimisant la divergence de Kullback-Leibler avec l'*a posteriori*, donc en résolvant un problème d'optimisation dans l'espace fonctionnel des densités de probabilités (d.d.p.). Le problème d'inférence statistique initial est donc résolu grâce à ce problème d'optimisation fonctionnelle. Celui-ci admet une solution analytique donnée par une équation implicite. Cette solution est alors approchée par des algorithmes de minimisation alternée suivant les composantes séparables de la loi approchante, algorithmes qui peuvent s'avérer peu pratiques en très grande dimension. Dans un travail précédent, [4], nous avions proposé une méthode itérative permettant de résoudre le problème du bayésien variationnel en un temps de calcul raisonnable, même en grande dimension. Pour cela nous avions transposé un algorithme de descente de gradient à l'optimisation des densités de probabilité.

L'article que nous présentons aujourd'hui vise à améliorer cette approche en considérant des directions de descente plus efficaces que celle donnée par le gradient. Nous nous sommes ainsi intéressés aux méthodes de sous-espaces décrits dans [3]. Dans un second temps nous nous intéressons aussi à la mise en œuvre pratique de ces méthodes à un problème de super-résolution, dont l'objectif est de construire une image haute-résolution à partir de plusieurs images basse-résolution représentant la même scène. Pour cela nous considérons un *a priori* donné par la norme de la variation totale, qui favorise les solutions régulières par morceaux.

2 Méthodes proposées

Les méthodes que nous développons dans cette partie s'appuient sur les méthodes classiques d'optimisation convexe par directions de descente dans un espace de Hilbert. Nous considérons ici que la direction de descente vit dans un sous-espace de dimension deux. Ce type d'approches, [3], permet plus de liberté et donc de meilleurs résultats que la descente suivant une direction fixée tout

en gardant un temps de calcul raisonnable. Dans ce cas chaque itération s'écrit :

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \delta^k = \mathbf{x}^k + s_1\mathbf{d}_1^k + s_2\mathbf{d}_2^k, \quad (1)$$

où δ^k est la direction de descente appartenant à l'espace engendré par les vecteurs $\mathbf{d}_1^k$ et $\mathbf{d}_2^k$. Ici s_1 et s_2 représentent le pas dans chaque direction.

Dans la suite, nous définissons nos approches. On définit tout d'abord les notations utilisées : $\mathbf{x}$ et $\mathbf{y}$ représentent respectivement le vecteur de paramètres à estimer et le vecteur regoupant les données tandis que $p(\mathbf{y}, \mathbf{x})$, $p(\mathbf{x}|\mathbf{y})$ et $q(\mathbf{x})$ sont la distribution jointe, la distribution *a posteriori* et son approximation.

Dans le cadre bayésien variationnel, nous supposons que q est séparable. Plusieurs choix de séparabilité peuvent être utilisés. Il peut s'agir d'une séparabilité partielle, par exemple la séparabilité entre les variables inconnues et les variables cachées, qui induit alors des opérations sur des matrices de grande dimension ou d'une séparabilité totale entre l'ensemble des éléments de $\mathbf{x}$, plus facile à manipuler mais moins précise.

L'approximation optimale est déterminée en minimisant la divergence de Kullback-Leibler de q par rapport à l'*a posteriori* vraie $p(\mathbf{x}|\mathbf{y})$. Néanmoins, en pratique, cette divergence n'est pas calculable car elle dépend de l'*a posteriori* dont la fonction de partition n'est pas connue. Heureusement, ce problème d'optimisation est équivalent à la maximisation de l'énergie libre négative [2] qui est définie par

$$\mathcal{F}(q) = \int_{\mathbb{R}^N} q(\mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{x})}{q(\mathbf{x})} d\mathbf{x}. \quad (2)$$

Nous voyons que cette énergie dépend de la loi jointe qui peut être facilement calculée. Le problème bayésien variationnel peut être formulé comme suit :

$$q^{opt} = \arg\max_q \mathcal{F}(q), \quad t.q. \ \ q \text{ d.d.p. séparable} \quad (3)$$

Nos approches s'appuient sur les méthodes de sous-espaces explicitées au dessus ainsi que sur la structure du gradient exponentialisé utilisée dans le cadre du bayésien variationnel dans [4], où, à l'itération $k+1$, la mise à jour de la loi approchante q^{k+1} est obtenue en prenant

$$q^{k+1}(\mathbf{x}) = q^k(\mathbf{x}) h^k(\mathbf{x}), \quad (4)$$

où $h^k \in L^1(q^k)$ est une fonction positive donnée dans nos approches par

$$h^k(\mathbf{x}) = K_k \exp(\delta^k(\mathbf{x})) = K_k \exp(s_1 d_1^k(\mathbf{x}) + s_2 d_2^k(\mathbf{x})), \quad (5)$$

et $(d_1^k(\mathbf{x}), d_2^k(\mathbf{x}))$ sont les fonctions permettant de définir notre espace tandis que K_k est une constante de normalisation.

Dans cet article nous considérons deux types de sous-espaces, [3] :

$$SG: \ d_1^k(\mathbf{x}) = \mathrm{d}f(q^k, \mathbf{x}), d_2^k(\mathbf{x}) = \mathrm{d}f(q^{k-1}, \mathbf{x}), \quad (6)$$

$$GM: \ d_1^k(\mathbf{x}) = \mathrm{d}f(q^k, \mathbf{x}), d_2^k(\mathbf{x}) = \delta^{k-1}(\mathbf{x}), \quad (7)$$

où $\mathrm{d}f(q^k, \mathbf{x})$ et $\mathrm{d}f(q^{k-1}, \mathbf{x})$ sont obtenus en considérant la différentielle au sens de Gateaux de $\mathcal{F}$ en q^k et en q^{k-1} tandis que $\delta^{k-1}(\mathbf{x})$ est la direction à l'itération précédente. Le sous-espace (6), appelé Sous-espace Gradient (SG), a été défini dans [8] tandis que le sous-espace (7), le Gradient à Mémoire (GM), a été introduit dans [6] comme une généralisation de la méthode du gradient conjugué.

En réinjectant la définition des sous-espaces (6) et (7) dans la structure (5), nous déterminons une densité de probabilité estimée à chaque itération qui ne dépend que des pas s_1 et s_2. Nous choisissons alors des pas proches des pas optimaux afin d'accélérer encore la méthode.

En définissant un pas multi-dimensionnel $\mathbf{s} = (s_1, s_2)$ et $f^k(\mathbf{s}) = \mathcal{F}(K_k q^k \exp(s_1 d_1^k(\mathbf{x}) + s_2 d_2^k(\mathbf{x})))$, on peut définir le pas optimal par :

$$(\mathbf{s}^{opt})^k = \arg\max_{\mathbf{s}\in\mathbb{R}^2} f^k(\mathbf{s}). \quad (8)$$

La détermination de ce pas optimal étant généralement coûteuse, nous proposons ici un pas sous-optimal obtenu en utilisant le développement de Taylor à l'ordre deux de $f^k(\mathbf{s})$ en zéro. Ce pas est ainsi donné par :

$$(\mathbf{s}^{subopt})^k = -\left(\mathbf{H}^k|_{\mathbf{s}=\mathbf{0}}\right)^{-1} \left.\frac{\partial f^k}{\partial \mathbf{s}}\right|_{\mathbf{s}=\mathbf{0}}, \quad (9)$$

où $\frac{\partial f^k}{\partial \mathbf{s}}$ et $\mathbf{H}^k$ représentent respectivement la dérivée du premier ordre et la matrice hessienne de $f^k(\mathbf{s})$.

3 Application aux problèmes inverses

3.1 Modèle utilisé

Dans la suite nous nous intéresserons au modèle linéaire :

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}, \quad (10)$$

où $\mathbf{y} \in \mathbb{R}^M$ et $\mathbf{x} \in \mathbb{R}^N$ représentent respectivement les données acquises et l'objet à estimer. L'opérateur $\mathbf{A}$ est une matrice connue de taille $M \times N$ et $\boldsymbol{\epsilon}$ est un bruit blanc gaussien $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \gamma_{\boldsymbol{\epsilon}}^{-1}\mathbf{I})$, où $\gamma_{\boldsymbol{\epsilon}}$ est l'inverse de la variance du bruit.

Concernant la distribution *a priori* de $\mathbf{x}$, nous considérons une densité basée sur la norme de la variation totale. Comme la constante de normalisation correspondante n'est pas calculable, nous considérons plutôt son approximation analytique donnée dans [1] :

$$p(\mathbf{x}|\gamma_p) \approx \tilde{p}(\mathbf{x}|\gamma_p) = c\gamma_p^{N/2} \exp\left[-\gamma_p TV(\mathbf{x})\right], \quad (11)$$

où c est une constante et

$$TV(\mathbf{x}) = \sum_{i=1}^{N} \sqrt{(\triangle_i^h(\mathbf{x}))^2 + (\triangle_i^v(\mathbf{x}))^2}. \quad (12)$$

Ici $\triangle_i^h$ et $\triangle_i^v$ représentent les différences d'ordre un, respectivement horizontalement et verticalement, pour le pixel i.

Afin d'obtenir des méthodes non supervisées, nous estimons aussi les hyperparamètres $\gamma_{\boldsymbol{\epsilon}}$ et γ_p grâce à un *a priori* non informatif de Jeffreys.

3.2 Difficulté liée à l'*a priori* TV

La présence de la norme TV dans l'*a priori* ne permet malheureusement pas de calculer l'énergie libre de nos lois approchantes directement. C'est pourquoi, comme dans [1] nous considérons des méthodes de type Minoration-Maximisation (MM) [5] dans lesquelles on maximise une loi minorante plus facile à manipuler :

$$\tilde{p}(\mathbf{x}|\gamma_p) \geq M(\mathbf{x}, \gamma_p; \boldsymbol{\lambda}) = c\gamma_p^{N/2} \times \exp\left[-\gamma_p \sum_{i=1}^{N} \frac{(\triangle_i^h(\mathbf{x}))^2 + (\triangle_i^v(\mathbf{x}))^2 + \lambda_i}{2\sqrt{\lambda_i}}\right], \quad (13)$$

où $(\lambda_i)_{i=1,\ldots,N}$ sont des variables auxiliaires positives.

Grâce à (13), nous obtenons une borne inférieure de $\mathcal{F}$:

$$\mathcal{F}(q(\mathbf{x}, \gamma_{\boldsymbol{\epsilon}}, \gamma_p)) \geq \mathcal{F}^L(q(\mathbf{x}, \gamma_{\boldsymbol{\epsilon}}, \gamma_p)) = \int q(\mathbf{x}, \gamma_{\boldsymbol{\epsilon}}, \gamma_p) \log\left(\frac{L(\mathbf{x}, \gamma_p, \gamma_{\boldsymbol{\epsilon}}, \mathbf{y}; \boldsymbol{\lambda})}{q(\mathbf{x}, \gamma_{\boldsymbol{\epsilon}}, \gamma_p)}\right) d\mathbf{x} d\gamma_{\boldsymbol{\epsilon}} d\gamma_p. \quad (14)$$

Ici, $L(\mathbf{x}, \gamma_p, \gamma_{\boldsymbol{\epsilon}}, \mathbf{y}; \boldsymbol{\lambda}) = p(\mathbf{y}|\mathbf{x}, \gamma_{\boldsymbol{\epsilon}}) M(\mathbf{x}, \gamma_p; \boldsymbol{\lambda}) p(\gamma_{\boldsymbol{\epsilon}}) p(\gamma_p)$ est une borne inférieure de la distribution jointe. Une maximisation alternée permet alors de résoudre le problème d'optimisation initial.

3.3 L'algorithme

Dans la mise en œuvre de nos méthodes, nous avons utilisé l'hypothèse de séparation suivant : $q(\mathbf{x}, \gamma_{\boldsymbol{\epsilon}}, \gamma_p) = \prod_i q_i(x_i) q_{\boldsymbol{\epsilon}}(\gamma_{\boldsymbol{\epsilon}}) q_p(\gamma_p)$ tandis que l'approche dans [1] ne suppose que la séparabilité entre $\mathbf{x}$, $\gamma_{\boldsymbol{\epsilon}}$ et γ_p.

Le fait d'avoir utilisé des lois conjuguées pour la distribution *a priori* permet d'obtenir des lois approchantes gaussiennes pour chaque x_i et des lois approchantes Gamma pour γ_p et $\gamma_{\boldsymbol{\epsilon}}$. La mise à jour de ces lois approchantes est ramenée à la réactualisation de leurs paramètres. La loi approchante de $\boldsymbol{x}$ est déterminée par notre approche. Pour ce qui est des variables $q_{\boldsymbol{\epsilon}}$ et q_p, l'*a posteriori* est séparable par rapport à ces variables. On peut donc utiliser directement l'algorithme du bayésien variationnel (BV) classique dans ce cas. Enfin les variables auxiliaires $\boldsymbol{\lambda}$ sont mises à jour à chaque itération en différentiant la fonction $\mathcal{F}^L$. Les étapes précédentes sont résumées dans l'Algorithme 1.

Algorithm 1 Algorithme proposé

1. Initialiser $(q_i^0)_{i=1,\ldots,N}$, $q_{\boldsymbol{\epsilon}}^0$, q_p^0 et $(\lambda_i^0)_{i=1,\ldots,N}$
2. Mettre à jour les paramètres de q_i^{k+1} pour $i = 1, \ldots, N$
   a. Déterminer les sous-espaces selon (6) ou (7)
   b. Calculer les pas sous-optimaux en utilisant (9)
   c. Mettre à jour la moyenne et la variance de q_i^{k+1} en utilisant (4)
3. Mettre à jour les paramètres de $q_{\boldsymbol{\epsilon}}^{k+1}$ en utilisant l'algorithme BV classique
4. Mettre à jour les paramètres de q_p^{k+1} en utilisant l'algorithme BV classique
5. Mettre à jour les variables auxiliaires $(\lambda_i^{k+1})_{i=1,\ldots,N}$
6. Retourner à 2 jusqu'à convergence

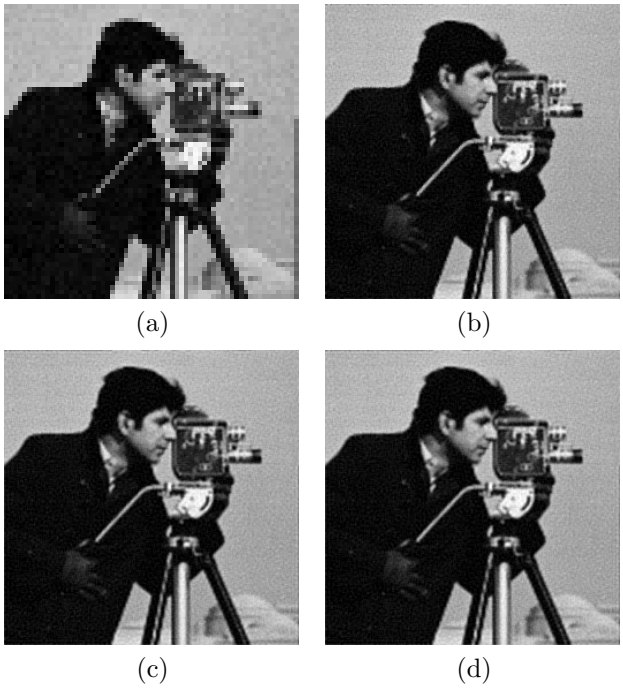

FIGURE 1 – Les images sont présentées dans un même niveau de gris. (a) Une des images basse-résolution (64×64), et les images haute-résolution (256×256) reconstruites par (b) Babacan [1], (c) SG, (d) GM.

4 Résultats

Afin de prouver l'efficacité de nos méthodes nous les avons implémentées sur un problème de super résolution d'images et comparées avec la méthode récente introduite par Babacan *et al.* dans [1]. Nous avons pris différentes images test, *Cameraman* de dimension 256×256 et *Lena* en 512×512. Nous avons construit des images basse-résolution grâce à un noyau de convolution de taille 3×3, décimées d'un facteur 4 à la fois horizontalement et verticalement, auxquelles nous avons rajouté du bruit à 5 dB, 25 dB, 45 dB.

Nous avons alors reconstruit l'image originale avec les différentes méthodes, initialisées de la même façon : $\mathbf{x}_0 = \mathbf{A}^T\mathbf{y}$ pour la moyenne et 100 pour la variance de l'image haute-résolution, les initialisations des hyperparamètres et des variables auxiliaires sont calculées à partir de $\mathbf{x}_0$. Nous montrons dans la FIGURE 1 une des images basse-résolution (FIGURE 1(a)) et les reconstructions (FIGURE 1(b-d)) obtenues par les trois méthodes implémentées pour le *Cameraman* dans le cas $SNR = 25$dB. Ces images nous permettent de conclure que les trois méthodes utilisées

améliorent les qualités de l'image et permettent d'obtenir des images de qualité similaire.

TABLE 1 – PERFORMANCES DE [1] ET DE NOS APPROCHES EN TERMES DE NOMBRE D'ITÉRATIONS/TEMPS CPU(S).

Données		PSNR	Babacan	SG	GM
Camera -man	5dB	11.64	30/15.7	111/9.3	**64/5.5**
	25dB	30.59	15/14.9	71/5.9	**49/4.0**
	45dB	40.62	31/86.8	139/11.3	**91/7.3**
Lena	5dB	14.81	26/56.6	134/43.1	**68/21.7**
	25dB	33.42	12/46.8	60/20.0	**60/19.4**
	45dB	38.72	29/296.4	201/67.8	**104/33.7**

Nous donnons aussi un tableau (TABLE 1) qui résume le temps de calcul obtenu avec chaque méthode pour arriver au même PSNR. Toutes les expériences ont été effectuées sur Intel(R) Core(TM) i5 CPU (3.33GHz) avec 8.0 GB RAM. En comparant les temps de calcul, nous pouvons voir que l'algorithme basé sur le sous-espace GM est, en moyenne, 4 fois plus rapide que celui utilisant l'algorithme bayésien variationnel classique [1] (nommé Babacan dans TABLE 1) tandis que l'algorithme basé sur le sous-espace SG est 2.7 fois plus rapide que [1].

Afin de montrer la convergence des algorithmes utilisés, nous donnons dans la FIGURE 2 les courbes de PSNR en fontion du temps de calcul pour *Cameraman* au cas $SNR = 45$dB. Ces courbes nous montrent que les trois méthodes implémentées convergent aux résultats de PSNR proches. On voit aussi clairement que nos deux approches convergent plus vite que [1]. Bien que la méthode dans [1] met moins d'itérations à converger, elle est moins efficace car chaque itération prend beaucoup plus de temps que nos deux approches.

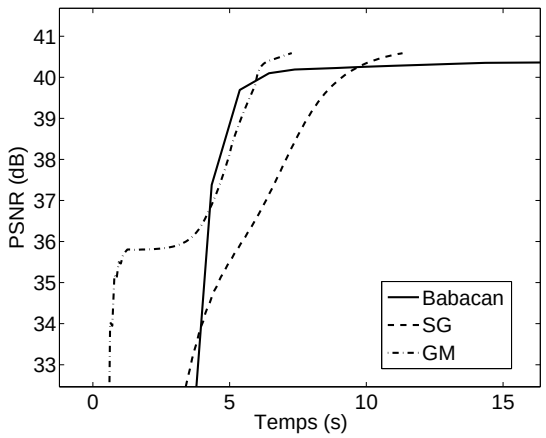

FIGURE 2 – Courbes de PSNR en fonction du temps de calcul (seconde) pour les trois méthodes implémentées.

5 Conclusion

Les approches efficaces applicables aux problèmes de grande taille ont été l'objectif de cet article. En transposant les méthodes de sous-espace dans l'espace fonctionnel, nous avons proposé deux approches itératives rapides pour la méthodologie du Bayésien variationnel. Ces approches sont appliquées à un problème de super-résolution en utilisant un a priori TV. Les comparaisons avec les méthode de l'état de l'art ont montré que nos approches, et plus particulièrement la méthode utilisant le sous-espace Gradient à Mémoire, sont plus efficaces que les approches classiques.

Références

[1] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Variational Bayesian super resolution. *IEEE Trans. Image Process.*, 20(4) :984–999, 2011.

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

[3] E. Chouzenoux, J. Idier, and S. Moussaoui. A Majorize-Minimize strategy for subspace optimization applied to image restoration. *IEEE Trans. Image Process.*, 20(18) :1517–1528, 2011.

[4] A. Fraysse and T. Rodet. A measure-theoretic variational Bayesian algorithm for large dimensional problems. Technical report, 2012. http ://hal.archives-ouvertes.fr/docs/00/70/22/59/PDF/var_bayV8.pdf.

[5] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *Am. Stat.*, 58(1) :30–37, 2004.

[6] A. Miele and J. W. Cantrell. Study on a memory gradient method for the minimization of functions. *J. Optimiz. Theory App.*, 3(6) :459–470, 1969.

[7] C. P. Robert and G. Casella. *Monte-Carlo Statistical Methods.* Springer Texts in Statistics. Springer, New York, 2000.

[8] Z. J. Shi and J. Shen. A new super-memory gradient method with curve search rule. *Appl. Math. Comput.*, 170(1) :1–16, 2005.

[9] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing.* Springer, 2006.