

Estimation d'hyperparamètres pour la résolution de problèmes inverses à l'aide d'ondelettes *

Caroline CHAUX¹, Laure BLANC-FÉRAUD²

¹Université Paris-Est,
Laboratoire d'Informatique Gaspard Monge, UMR CNRS 8049
5 Bd Descartes, 77454 Marne-la-Vallée Cedex 2, France

²ARIANA I3S-INRIA
2004 Route des lucioles, 06902 Sophia Antipolis Cedex, France
caroline.chaux@univ-paris-est.fr, Laure.Blanc-Feraud@sophia.inria.fr

Résumé – Nous nous intéressons à l'estimation des paramètres de régularisation pour la restauration d'image floue et bruitée. Dans l'approche variationnelle, la restauration consiste à minimiser un critère convexe composé d'un terme de rappel aux données (quadratique) et d'un terme de régularisation (norme ℓ^1) opérant dans le domaine ondelettes. Nous proposons une méthode d'estimation des paramètres de régularisation (hyperparamètres, un par sous-bande) par maximum de vraisemblance, à partir de la seule image observée. La difficulté de l'estimation en données incomplètes est de pouvoir échantillonner des lois sur des champs de variables aléatoires dont les interactions entre voisins sont étendues, du fait de l'opérateur linéaire de flou. Nous proposons une méthode qui permet de calculer ces échantillons par MCMC (échantillonnage de Gibbs et Metropolis-Hastings). Pour l'estimation, nous utilisons une méthode de gradient. Les résultats de simulation obtenus montrent la faisabilité de la méthode et ses bonnes performances en terme d'estimation.

Abstract – When image restoration (blur + noise) purpose is considered, a variational approach can be adopted thus consisting in minimizing a convex criterion often splitted in two parts: a data fidelity term (quadratic) and a regularization term (ℓ^1 norm) expressed in the wavelet domain. We propose a method allowing to estimate the regularization hyperparameters (one per subband) based on a maximum likelihood estimator (only knowing the observed (degraded) image). The main difficulty is to be able to sample according to a priori and a posteriori distributions because of pixel interactions introduced by the blur operator. The proposed method allows to compute sampling with MCMC (Gibbs sampling and Metropolis-Hastings) and then a gradient method is used to estimate the hyperparameters. Simulation results demonstrate the good performance and behavior of the proposed approach.

1 Introduction

Dans de nombreux domaines (spatial, médical, ...) les images acquises sont sujettes à des dégradations dues aux imperfections de l'instrument optique et au mode d'acquisition en général. Le modèle d'observation considéré dans ce travail et largement utilisé en restauration d'image est le suivant :

$$g = Au + n.$$

où g (vecteur de dimension k) est l'image observée, dégradée par un opérateur de flou A connu et un bruit additif supposé blanc et gaussien $n \sim \mathcal{N}(0, \sigma^2 I)$ (σ connu). Afin d'estimer l'image u à partir d'une observation dégradée g , une approche variationnelle peut être adoptée avec une régularisation définie par une norme ℓ^1 sur des coefficients en ondelettes [1, 2, 3, 4, 5, 6]. Le critère $J(u)$ qui est mi-

nimisé est :

$$J(u) = \frac{\|g - Au\|^2}{2\sigma^2} + \sum_{\mathbf{m}} \lambda_{\mathbf{m}} \|F_{\mathbf{m}}u\|_1 \quad (1)$$

où $F_{\mathbf{m}}u$ désigne une sous-bande et F représente l'opérateur de transformée en ondelettes. Nous considérons dans ce travail une transformée en ondelettes sur une base orthogonale [7].

Cependant, ces approches variationnelles requièrent la connaissance de $\lambda = (\lambda_{\mathbf{m}})_{\mathbf{m}}$ et la question est de savoir comment fixer ces hyperparamètres. Dans certaines applications et lorsqu'ils ne sont pas trop nombreux, les hyperparamètres sont réglés empiriquement par l'utilisateur. Néanmoins, lorsque la convergence de l'algorithme est lente ou que le nombre de paramètres à déterminer est important, cette solution semble difficilement envisageable. Notre objectif est donc l'estimation automatique des hyperparamètres. Estimer λ ne connaissant que l'observation dégradée g est le problème difficile de l'estima-

*Ce travail a été partiellement financé par l'ANR DETECFINE (ANR-06-BLAN-0388)

tion de paramètres en données incomplètes [8]. Parmi les méthodes les plus utilisées on peut citer l'estimation au sens du maximum de vraisemblance [8, 9], les algorithmes du type EM (*Expectation Maximization*) [10, 3] ou les méthodes de validation croisée [11]. Plus récemment dans [5], les auteurs proposent d'utiliser le principe de Stein afin de déterminer adaptivement le paramètre de régularisation. Une autre approche consiste à considérer les paramètres non pas comme des variables déterministes mais comme des variables aléatoires, de leur associer une loi de probabilité, et de faire leur estimation, en utilisant des méthodes d'échantillonnage par Chaînes de Markov à dynamique de Monte Carlo (MCMC) [12, 13].

Dans ce travail, nous proposons de considérer les $(\lambda_{\mathbf{m}})_{\mathbf{m}}$ comme des paramètres déterministes à estimer, car nous n'avons d'*a priori* sur ces paramètres, et nous les estimons au sens du maximum de vraisemblance. Ceci permet d'une part, d'exploiter les bonnes propriétés de cet estimateur et d'autre part, λ étant déterministe, d'éviter d'introduire de nouveaux hyper hyperparamètres qui auraient été nécessaires si des lois avaient été fixées pour $(\lambda_{\mathbf{m}})_{\mathbf{m}}$.

2 Estimation par maximum de vraisemblance

2.1 Difficultés

Dans le but d'estimer les paramètres, le critère (1) est interprété au sens stochastique. Il correspond à l'énergie de la probabilité *a posteriori* dans un cadre bayésien où le premier terme est l'énergie de la probabilité $p(g|u)$ donnée par la probabilité du bruit n supposé de loi gaussienne

$$p(g|u) = \frac{1}{K_{\sigma}} \exp\left(-\frac{\|g - Au\|^2}{2\sigma^2}\right) \quad (2)$$

où la constante vaut $K_{\sigma} = (2\pi)^{k/2}\sigma^k$ et le deuxième terme correspond à l'énergie du modèle *a priori* donnée par

$$p_{\lambda}(u) = \frac{1}{Z_{\lambda}} \prod_{\mathbf{m}} \exp(-\lambda_{\mathbf{m}}\|F_{\mathbf{m}}u\|_1) \quad (3)$$

où la constante Z_{λ} est la constante de normalisation définie par

$$Z_{\lambda} = \int_u \exp\left(-\sum_{\mathbf{m}} \lambda_{\mathbf{m}}\|F_{\mathbf{m}}u\|_1\right) du \quad (4)$$

La variable u est un champ discret de k variables aléatoires réelles. Ainsi la variable u prend ses valeurs dans \mathbb{R}^k , qui est le domaine d'intégration de (4). F étant un changement de base, $\text{Ker}(F) = \{0\}$ et l'intégrale (4) converge en supposant que $\exists \mathbf{m}, \lambda_{\mathbf{m}} \geq \epsilon > 0$.

L'estimation de λ au sens du maximum de vraisemblance consiste à maximiser $p_{\lambda}(g)$ en λ . Par inférence puis dérivation, l'estimateur peut être calculé par une méthode

de gradient [8]. On a :

$$p_{\lambda}(g) = \int_u p_{\lambda}(g, u) du = \int_u p(g|u) p_{\lambda}(u) du = \frac{Z_{\sigma, \lambda}(g)}{K_{\sigma} Z_{\lambda}} \quad (5)$$

où $Z_{\sigma, \lambda}(g) = \int_u \exp\left(-\frac{\|g - Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}} \lambda_{\mathbf{m}}\|F_{\mathbf{m}}u\|_1\right) du$.

Pour maximiser $p_{\lambda}(g)$ en λ on cherche la dérivée de (5) par rapport à chaque $\lambda_{\mathbf{m}}$. On note que, étant donné qu'il n'est pas possible de montrer la convexité de $p_{\lambda}(g)$ en λ , un algorithme de gradient donnera un minimum local, dépendant des valeurs initiales de λ .

La dérivée de (5) par rapport à chaque $\lambda_{\mathbf{m}}$ est (en appliquant le théorème de convergence dominée de Lebesgue afin d'inverser dérivée et signe intégral) :

$$\frac{\partial \log p_{\lambda}(g)}{\partial \lambda_{\mathbf{m}}} = \frac{\partial \log Z_{\sigma, \lambda}(g)}{\partial \lambda_{\mathbf{m}}} - \frac{\partial \log(Z_{\lambda})}{\partial \lambda_{\mathbf{m}}}$$

et

$$\begin{aligned} \frac{\partial \log Z_{\sigma, \lambda}(g)}{\partial \lambda_{\mathbf{m}}} &= \frac{-1}{Z_{\sigma, \lambda}(g)} \int_u \|F_{\mathbf{m}}u\|_1 \exp\left(-\frac{\|g - Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}} \lambda_{\mathbf{m}}\|F_{\mathbf{m}}u\|_1\right) du \\ &= -\mathbb{E}_{\sigma, \lambda}[\|F_{\mathbf{m}}u\|_1] \end{aligned} \quad (6)$$

$$\begin{aligned} \frac{\partial \log(Z_{\lambda})}{\partial \lambda_{\mathbf{m}}} &= \frac{-1}{Z_{\lambda}} \int_u \|F_{\mathbf{m}}u\|_1 \exp\left(-\sum_{\mathbf{m}} \lambda_{\mathbf{m}}\|F_{\mathbf{m}}u\|_1\right) du \\ &= -\mathbb{E}_{\lambda}[\|F_{\mathbf{m}}u\|_1] \end{aligned} \quad (7)$$

La première espérance, en (6), dépend de la loi *a posteriori* $p_{\lambda}(u|g)$ et la seconde espérance, en (7) dépend de la loi *a priori* $p_{\lambda}(u)$.

Pour mettre en œuvre une méthode de gradient, il faut savoir approcher numériquement les espérances (6) et (7). La difficulté vient du terme d'espérance selon la loi de probabilité *a posteriori* $p_{\lambda}(u|g)$ dont l'énergie est (1). Cette espérance n'est pas calculable analytiquement et il faut l'approcher par une moyenne empirique, calculée sur des échantillons tirés selon cette loi. Mais l'échantillonnage par un algorithme de Gibbs ou de Metropolis-Hastings est difficile d'un point de vue pratique (coût de calcul élevé) car l'opérateur de flou A induit une interaction des pixels entre eux sur un voisinage très large [9]. Une manière d'éliminer cette difficulté est de passer dans le domaine de Fourier (ou transformée en Cosinus) afin de diagonaliser l'opérateur A . Le problème est que F ne se diagonalise pas par transformée de Fourier, mais devient au contraire un opérateur agissant sur tous les pixels. C'est alors le terme de régularisation (norme ℓ^1) qui pose problème pour l'échantillonnage. La difficulté vient de la présence simultanée des opérateurs A et F qui ne sont pas diagonalisables dans le même espace. Il faut donc dissocier les deux opérateurs dans l'énergie de la probabilité.

2.2 Introduction d'une variable intermédiaire

La dissociation des deux opérateurs peut être obtenue en introduisant une variable intermédiaire w . On considère alors un nouveau critère $J(u, w)$ donné par [4]

$$J(u, w) = \frac{1}{2\sigma^2\mu} \left(\|u - w\|^2 + \langle Cw, w \rangle \right) + \frac{1}{2\sigma^2} \left(\|g\|^2 - 2\langle Au, g \rangle \right) + \sum_{\mathbf{m}} \lambda_{\mathbf{m}} \|F_{\mathbf{m}}u\|_1$$

où $C = B(I - B)^{-1}$ et $B = \mu A^*A$ (μ tel que $\mu \|A^*A\| < 1$). Il est équivalent de minimiser $J(u)$ en u et $J(u, w)$ en (u, w) [4]. L'opérateur linéaire peut ainsi se trouver appliqué à g et non à u puisque $\langle Au, g \rangle = \langle u, A^*g \rangle$.

Afin d'interpréter ce critère en terme de probabilités, nous le réécrivons de manière équivalente :

$$J(u, w) = \frac{1}{2\sigma^2\mu} \left((w - (I + C)^{-1}u)^T (I + C) (w - (I + C)^{-1}u) \right) + \frac{1}{2\sigma^2} \left(\|g - Au\|^2 \right) + \sum_{\mathbf{m}} \lambda_{\mathbf{m}} \|F_{\mathbf{m}}u\|_1$$

puisque $(I + C)^{-1} = I - \mu A^*A$.

On remarque que les variables g et w sont indépendantes conditionnellement à u et

$$p(w|u) = \frac{1}{K_{\mu}} \times \exp \left(- \frac{(w - (I + C)^{-1}u)^T (I + C) (w - (I + C)^{-1}u)}{2\sigma^2\mu} \right) \quad (8)$$

avec $K_{\mu} = (2\pi\sigma^2\mu)^{k/2} (\det(I + C))^{-1/2}$, est une loi gaussienne $\mathcal{N}((I + C)^{-1}u, \sigma^2\mu(I + C)^{-1})$.

A présent, nous allons procéder de manière identique à la section précédente. Afin d'estimer les paramètres au sens du maximum de vraisemblance, il faut maximiser en λ la loi de probabilité

$$\begin{aligned} p_{\lambda}(g) &= \int_{u,w} p_{\lambda}(g, u, w) du dw \\ &= \int_{u,w} p(g|u) p(w|u) p_{\lambda}(u) du dw \\ &= \frac{1}{K_{\sigma} K_{\mu} Z_{\lambda}} \int_{u,w} \exp \left(- \frac{1}{2\sigma^2\mu} \left(\|u - w\|^2 + \langle Cw, w \rangle \right) - \frac{1}{2\sigma^2} \left(\|g\|^2 - 2\langle u, A^*g \rangle \right) - \sum_{\mathbf{m}} \lambda_{\mathbf{m}} \|F_{\mathbf{m}}u\|_1 \right) du dw \end{aligned}$$

Calculons les dérivées partielles de la log-vraisemblance par rapport à chaque paramètre $\lambda_{\mathbf{m}}$ en invoquant de nouveau de théorème de convergence dominée de Lebesgue :

$$\frac{\partial \log p_{\lambda}(g)}{\partial \lambda_{\mathbf{m}}} = \mathbb{E}_{\lambda} [\|F_{\mathbf{m}}u\|_1] - \mathbb{E}_{\sigma, \lambda, \mu} [\|F_{\mathbf{m}}u\|_1].$$

La première espérance dépend de la loi *a priori* $p_{\lambda}(u)$ qui est une loi de Laplace (3) dans les sous-bandes et donc

l'espérance se calcule facilement et vaut $1/\lambda_{\mathbf{m}}$. La seconde espérance dépend de la loi *a posteriori* $p_{\lambda}(u, w|g)$. Contrairement au cas précédant où l'opérateur A était appliqué à u , on se rend compte qu'il est cette fois-ci appliqué à g ce qui va nous permettre de pouvoir échantillonner plus facilement selon cette loi et comme décrit dans le paragraphe suivant.

2.3 Algorithme de gradient

Afin de calculer les paramètres $\lambda_{\mathbf{m}}$ au sens du maximum de vraisemblance, on utilise une méthode de gradient [8] :

$$\lambda_{\mathbf{m}}^{(n+1)} = \lambda_{\mathbf{m}}^{(n)} + \alpha_n \left[1/\lambda_{\mathbf{m}}^{(n)} - \frac{2}{I} \sum_{i=I/2+1}^I \|F_{\mathbf{m}}(u^{(n)})_{\sigma, \lambda^{(n)}, \mu}_i\|_1 \right]$$

où $(u^{(n)})_{\sigma, \lambda^{(n)}, \mu}_i$ est le i ème échantillon tiré selon la loi de probabilité *a posteriori* $p_{\lambda^{(n)}}(u, w|g)$. α_n est ici le pas de l'algorithme et il peut varier au cours des itérations.

L'échantillonnage devra donc se faire suivant la loi $p_{\lambda}(u, w|g)$ pour $\lambda = \lambda^{(n)}$. Pour cela, on tirera (échantillonneur de Gibbs) successivement $p(w|u)$ donnée par (8) (loi gaussienne) et $p_{\lambda}(u|w, g)$ où :

$$p_{\lambda}(u|w, g) \propto \exp \left(- \frac{1}{2\sigma^2\mu} \|u - w\|^2 + \frac{1}{\sigma^2} \langle u, A^*g \rangle - \sum_{\mathbf{m}} \lambda_{\mathbf{m}} \|F_{\mathbf{m}}u\|_1 \right) \quad (9)$$

Concernant le tirage selon $p(w|u)$, on effectue le tirage dans le domaine de Fourier car la matrice de covariance de la loi se diagonalise facilement. Pour le tirage selon $p_{\lambda}(u|w, g)$, on utilise un algorithme de type Metropolis-Hastings directement dans le domaine ondelettes ce qui revient exprimer (9) comme

$$p_{\lambda}(u|w, g) \propto \exp \left(- \frac{1}{2\sigma^2\mu} \|Fu - Fw\|^2 + \frac{1}{\sigma^2} \langle Fu, FA^*g \rangle - \sum_{\mathbf{m}} \lambda_{\mathbf{m}} \|F_{\mathbf{m}}u\|_1 \right)$$

Cette opération est effectuée I fois et une moyenne sur les $I/2$ derniers échantillons générés est prise en compte.

3 Simulations numériques

Afin de tester les performances de notre algorithme, nous avons généré aléatoirement des coefficients d'ondelettes suivant la loi *a priori* (3) avec des paramètres $\lambda_{\mathbf{m}}$ fixés par sous-bande, réalistes pour des images naturelles. Nous avons utilisé des Symlets [14] de longueur 8 sur $J = 2$ niveaux de résolution et chaque sous-bande \mathbf{m} obtenue est représentée par le triplet (j, l, c) où j est l'indice du niveau de résolution et $(l, c)_{l \in \{0,1\}, c \in \{0,1\}}$ représente les

sous-bandes filtrées passe-bas/passe-haut (le couple (1, 1) représente donc les détails diagonaux). La taille de l'image générée correspondante est 128×128 . Nous appliquons notre méthode sur l'image dégradée g où A est un flou gaussien d'écart-type 0.5 et le bruit est de variance $\sigma^2 = 25$ (le SNR initial résultant est de 10.68 dB). L'algorithme de gradient est lancé sur 2000 itérations, il est initialisé en appliquant un filtre de Wiener sur g et en estimant les paramètres par maximum de vraisemblance sur l'image résultante. Les valeurs données dans le Tab. 1 sont : la valeur théorique, la valeur estimée par maximum de vraisemblance sur l'image originale non dégradée (données complètes) et pour la méthode proposée, la valeur moyenne sur les 500 dernières itérations du gradient.

Sous-bande \mathbf{m}	(1, 0, 1)	(1, 1, 0)	(1, 1, 1)
valeur théo.	0.0611	0.0547	0.0551
maxi vrais.	0.0609	0.0548	0.0555
meth. prop.	0.0606	0.0544	0.0555

Sous-bande \mathbf{m}	(2, 0, 0)	(2, 0, 1)	(2, 1, 0)	(2, 1, 1)
valeur théo.	0.0124	0.0256	0.0247	0.0255
maxi vrais.	0.0123	0.0255	0.0251	0.0266
meth. prop.	0.0123	0.0256	0.0250	0.0265

TABLE 1 – Résultats numériques.

Les résultats obtenus dans ce tableau sont encourageants puisque les valeurs estimées avec l'algorithme proposé sont proches de celles estimées par maximum de vraisemblance sur l'image originale. Par ailleurs, si l'on regarde la figure 1, on constate que l'algorithme se stabilise assez vite proche de la valeur théorique. Néanmoins, il reste légèrement oscillant et c'est pour cette raison que le paramètre estimé est obtenu en moyennant sur les 500 dernières itérations.

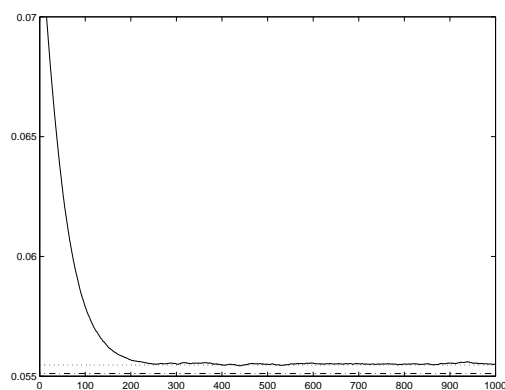


FIGURE 1 – Évolution du paramètre $\lambda_{\mathbf{m}}$ (sur les 1000 premières itérations) associé à la sous-bande diagonale au niveau de résolution 1 ($\mathbf{m} = (1, 1, 1)$). En point-pointillé la valeur théorique, en pointillé la valeur obtenue par maximum de vraisemblance en données complètes et en continu, le paramètre estimé au cours des itérations.

4 Conclusion

Une méthode d'estimation des paramètres de régularisation est proposée dans un cadre variationnel. Les paramètres à estimer sont considérés déterministes et un estimateur au sens du maximum de vraisemblance est utilisé. Néanmoins, il serait possible *a priori* d'inclure l'échantillonnage de λ et σ en leur attribuant des lois et hyper hyperparamètres adaptés. Par ailleurs, sous certaines conditions, cette approche peut-être étendue à d'autres fonctions de régularisation et l'application de la méthode proposée à des images réelles est en cours.

Remerciements : Nous adressons nos remerciements au Prof. J.-C. Pesquet pour ses conseils avisés.

Références

- [1] I. Daubechies, M. DeFrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Applied Math.*, 57 :1413–1457, 2004.
- [2] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4) :564–574, December 2007.
- [3] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. on Image Proc.*, 12(8) :906–916, Aug. 2003.
- [4] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle. A l^1 -unified variational framework for image restoration. In T. Pajdla and J. Matas, editors, *Proc. European Conference on Computer Vision (ECCV)*, volume LNCS 3024, pages 1–13, Prague, Czech Republic, May 2004. Springer.
- [5] C. Vonesch, S. Ramani, and M. Unser. Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constrain. In *Proc. of the 2008 IEEE International Conf. on Image Proc. (ICIP'08)*, pages 665–668, San Diego CA, USA, October 12-15 2008.
- [6] C. Chau, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs. A variational formulation for frame based inverse problems. *Inverse Problems*, 23 :1495–1518, June 2007.
- [7] S. Mallat. *A wavelet tour of signal processing*. Academic Press, San Diego, USA, 1998.
- [8] L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82 :625–645, 1989. Springer-Verlag.
- [9] A. Jalobeanu, L. Blanc-Féraud, and J. Zerubia. Hyperparameter estimation for satellite image restoration using a MCMC maximum likelihood method. *Pattern Recognition*, 35(2) :341–352, 2002.
- [10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 (series B) :1–38, 1977.
- [11] M. Jansen and A. Bultheel. Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise. *IEEE Trans. on Image Proc.*, 8(7) :947–953, July 1999.
- [12] M. M. Ichir and A. Mohammad-Djafari. Hidden markov models for wavelet-based blind source separation. *IEEE Trans. on Image Proc.*, 15(7) :1887–1899, July 2006.
- [13] D. Leporini and J.-C. Pesquet. Bayesian wavelet denoising : Besov priors and non-Gaussian noises. *Signal processing*, 81(1) :55–67, 2001.
- [14] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF, SIAM Lecture Series, Philadelphia, PA, USA, 1992.