

Une véritable approche ℓ_0 pour l'apprentissage de dictionnaire

Yuan LIU, Stéphane CANU, Paul HONEINE, Su RUAN

Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS;
Avenue de l'Université, 76801 Saint-Étienne-du-Rouvray Cedex, France
yuan.liu@insa-rouen.fr, stephane.canu@insa-rouen.fr
paul.honeine@univ-rouen.fr, su.ruan@univ-rouen.fr

Résumé – Ces derniers temps, les modèles parcimonieux ont suscité un vif intérêt de part leur capacité à sélectionner automatiquement un modèle simple parmi une grande collection. Ils se sont notamment révélés être utiles pour l'apprentissage de dictionnaire. La manière classique d'exprimer la parcimonie dans ce cadre, consiste à utiliser la norme ℓ_0 qui permet de compter, et donc de contrôler, le nombre de composantes d'un modèle. Malheureusement, les problèmes d'optimisation associés s'avèrent être non convexes et NP-difficiles, ce qui a justifié la recherche de relaxations pour obtenir une bonne approximation de la solution globale du problème. A l'inverse, nous montrons dans cet article que, en dépit de sa complexité, ce problème l'apprentissage de dictionnaire avec la norme ℓ_0 peut aujourd'hui être traité efficacement. L'idée est de reformuler le problème comme un programme quadratique mixte en nombres entiers (MIQP) et d'utiliser un logiciel d'optimisation pour obtenir l'optimum global du problème. La principale difficulté de cette approche étant le temps de calcul, nous proposons deux méthodes pour le réduire. L'application de notre méthode d'apprentissage de dictionnaire MIQP à un problème de débruitage d'images démontre sa faisabilité.

Abstract – Sparse representation learning has recently gained a great success in signal and image processing, thanks to recent advances in dictionary learning. To this end, the ℓ_0 -norm is often used to control the sparsity level. Nevertheless, optimization problems based on the ℓ_0 -norm are non-convex and NP-hard. For these reasons, relaxation techniques have been attracting much attention of researchers, by priorly targeting approximation solutions (e.g. ℓ_1 -norm, pursuit strategies). On the contrary, this paper considers the exact ℓ_0 -norm optimization problem and proves that it can be solved effectively, despite of its complexity. The proposed method reformulates the problem as a Mixed-Integer Quadratic Program (MIQP) and gets the global optimal solution by applying existing optimization software. Because the main difficulty of this approach is its computational time, two techniques are introduced that improve the computational speed. Finally, our method is applied to image denoising which shows its feasibility and relevance compared to the state-of-the-art.

1 Introduction

L'apprentissage de représentations parcimonieuses a été largement considéré avec succès dans les domaines du traitement du signal, des images et en vision, notamment pour des applications de débruitage d'image [6, 3], d'inpainting [11] et de classification [15], pour n'en citer que quelques-unes. Il consiste à modéliser les données par une combinaison linéaire de quelques éléments d'un dictionnaire. Au delà d'un dictionnaire prédéfini, nous considérons ici l'apprentissage du dictionnaire pour une représentation adaptée aux données disponibles.

L'apprentissage nécessite alors l'estimation jointe des éléments du dictionnaire et de leur coefficient de pondération. En opérant une procédure d'optimisation alternée, les éléments du dictionnaire peuvent être estimés facilement à chaque itération par moindres carrés ou descente de gradient stochastique [11]. L'estimation des coefficient de pondération, dite codage parcimonieux (*sparse coding* en anglais), est un problème non convexe et NP-difficile à cause de la contrainte de type ℓ_0 pour imposer la parcimonie. Afin de surmonter cette difficulté, deux principales approches ont été mises en œuvre pour relaxer la contrainte ℓ_0 . La première considère une solution approchée par poursuite séquentielle [6, 9]. La seconde approche opère en

remplaçant la norme ℓ_0 par sa relaxation convexe : la norme ℓ_1 . Les méthodes les plus connues sont la méthode bayésienne [13], la méthode *K-SVD* [1] et la méthode proximale [3].

Le présent article vise à résoudre d'une manière exacte le problème de l'apprentissage de dictionnaire, c'est à dire en considérant la norme ℓ_0 sans aucune relaxation. Pour ce faire, nous reformulons le problème d'optimisation pour le résoudre par programme quadratique mixte en nombres entiers (MIQP). Ainsi, les récentes avancées théoriques en optimisation sont-elles exploitées avec les améliorations d'implémentation qui les accompagnent. Nous proposons aussi deux techniques pour réduire le coût calculatoire, d'une part avec la mise en place de nouvelles contraintes pour renforcer la formulation et d'autre part en initialisant par une méthode proximale. Nous montrons que ces différentes contributions permettent la résolution exacte de l'apprentissage parcimonieux pour le débruitage d'images. Les résultats obtenus corroborent une récente étude sur la tolérance élevée de MIQP à la présence du bruit [4].

L'article est organisé comme suit. Le problème de représentation parcimonieuse est présenté et la méthode d'optimisation exacte est décrite dans la section 2. La section 3 montre la pertinence de la méthode proposée en débruitage d'image, et la dernière section conclut l'article.

2 Représentation parcimonieuse

2.1 Énoncé du problème

Soit $Y = [\mathbf{y}_1, \dots, \mathbf{y}_\ell] \in \mathbb{R}^{n \times \ell}$, une matrice contenant ℓ signaux \mathbf{y}_i , $i = 1, \dots, \ell$, de dimension n . On suppose que $Y = M + B$ où M et B sont deux matrices modélisant respectivement les parts d'information et de bruit inconnu, contenues des signaux. La représentation parcimonieuse de Y , consiste à trouver une matrice $X = [\mathbf{x}_1, \dots, \mathbf{x}_\ell] \in \mathbb{R}^{p \times \ell}$ parcimonieuse (i.e., avec seulement quelques termes non nuls) et un dictionnaire $D = [\mathbf{d}_1, \dots, \mathbf{d}_p] \in \mathbb{R}^{n \times p}$ tels que $M = DX$. Les éléments \mathbf{d}_i , $i = 1, \dots, p$, sont appelés atomes et D appartient à l'espace $\mathcal{D} = \{D \in \mathbb{R}^{n \times p}, \mathbf{d}_j^T \mathbf{d}_j \leq 1, j = 1, \dots, p\}$. L'estimation jointe de X et de D peut s'écrire comme un problème de minimisation du risque empirique régularisé :

$$\min_{\substack{D \in \mathcal{D} \\ X \in \mathbb{R}^{p \times \ell}}} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{1}{2} \|\mathbf{y}_i - D\mathbf{x}_i\|_2^2 + \lambda \Omega(\mathbf{x}_i) \right). \quad (1)$$

Le premier terme représente l'erreur de reconstruction et le second la régularisation, qui comprend un opérateur de régularisation $\Omega(\mathbf{x}_i)$. Le paramètre $\lambda > 0$ contrôle le compromis entre la fidélité aux données et la parcimonie. Une manière classique d'aborder ce problème (1) d'estimation jointe de D et X consiste à utiliser une procédure de relaxation alternée en deux phases [1]. La première phase, dite de codage parcimonieux (*sparse coding*), consiste à estimer X en supposant D connu. Les méthodes les plus utilisées sont celle de Gauss Seidel [14] et de descente de gradient [10]. La seconde phase, dite d'apprentissage de dictionnaire, consiste à estimer D en supposant X connu. Les algorithmes les plus utilisés sont ceux des moindres carrés et de descente de gradient stochastique [11].

Dans ce travail nous nous intéressons au cas où la régularisation est de type ℓ_0 , ($\Omega(\mathbf{x}) = \|\mathbf{x}\|_0$), ce qui permet de contrôler explicitement le nombre de termes non nuls du vecteur \mathbf{x} . En pratique, deux formulations de minimisation sous contraintes analogues à (1) peuvent être utilisées. La première s'écrit, pour un $T > 0$ majorant le nombre de composantes non nulles :

$$\min_{\substack{D \in \mathcal{D} \\ \mathbf{x}_i \in \mathbb{R}^p}} \frac{1}{2} \|\mathbf{y}_i - D\mathbf{x}_i\|_2^2 \text{ avec } \|\mathbf{x}_i\|_0 \leq T, \quad i = 1, \dots, n. \quad (2)$$

La seconde s'écrit, pour $\varepsilon > 0$ représentant le niveau de bruit :

$$\min_{\substack{D \in \mathcal{D} \\ \mathbf{x}_i \in \mathbb{R}^p}} \|\mathbf{x}_i\|_0 \text{ avec } \frac{1}{2} \|\mathbf{y}_i - D\mathbf{x}_i\|_2^2 \leq \varepsilon, \quad i = 1, \dots, n. \quad (3)$$

Quelle que soit la formulation choisie, (1), (2) ou (3), dans le cadre de l'approche de minimisation alternée que nous nous proposons d'utiliser, à cause de la la norme ℓ_0 , le problème d'optimisation lié à la phase de codage parcimonieux est non-convexe et NP-difficile. Constatant cette difficulté, la plupart des travaux dans le domaine proposent de relaxer le problème en remplaçant la norme ℓ_0 par un terme convexe comme la norme ℓ_1 [11] ou d'utiliser un algorithme approché comme celui de poursuite [6]. Dans cet article, nous proposons un nouvel algorithme basé sur la programmation quadratique mixte binaire, permettant de résoudre de manière exacte le problème de codage parcimonieux avec la norme ℓ_0 associé à l'équation (2).

2.2 Optimisation globale du problème

2.2.1 Programmation quadratique mixte (MIQP)

Une manière de traiter la norme ℓ_0 qui apparaît dans la phase de *sparse coding* associée au problème d'optimisation (2), consiste à réécrire le problème sous une forme standard que l'on sait résoudre avec les logiciels d'optimisation d'aujourd'hui.

Cette réécriture peut s'effectuer grâce à l'introduction de variables binaires. L'idée est d'associer à tous les éléments x_i du vecteur \mathbf{x} , une variable binaire z_i égale à 0 si $x_i = 0$ et égale à un sinon. Cela revient à imposer la relation logique suivante, composante par composante

$$z_i = 0 \iff x_i = 0, \quad i = 1, \dots, p. \quad (4)$$

A l'aide de cette nouvelle variable binaire, la contrainte de parcimonie $\|\mathbf{x}\|_0 \leq T$ peut s'exprimer sous la forme

$$\sum_{i=1}^p z_i \leq T, \quad (5)$$

de sorte que le problème de *sparse coding* associé à (2) se réécrit :

$$\begin{aligned} \min_{\substack{\mathbf{x} \in \mathbb{R}^p \\ \mathbf{z} \in \{0,1\}^p}} \quad & \frac{1}{2} \|\mathbf{y} - D\mathbf{x}\|_2^2 \\ \text{avec} \quad & z_i = 0 \iff x_i = 0, \quad i = 1, \dots, p \\ & \mathbf{1}_p^T \mathbf{z} \leq T, \end{aligned} \quad (6)$$

où $\mathbf{1}_p$ est un vecteur de 1 de dimension p .

Si la relation logique (4) est traitée par certains logiciels, il est parfois préférable de l'éliminer explicitement. Ce peut être réalisé par l'intermédiaire de l'introduction d'une contrainte de type *big M*. Supposons que l'on connaisse un réel $M > 0$ suffisamment grand, de sorte que, si \mathbf{x}^* est solution du problème (2), alors $\|\mathbf{x}^*\|_\infty < M$. Dans ce cas, imposer les contraintes (4) revient à poser

$$-z_i M \leq x_i \leq z_i M, \quad i = 1, \dots, p. \quad (7)$$

La formulation *big M* du problème (2) est donc, pour $M > 0$ et $T > 0$ donnés :

$$\begin{aligned} \min_{\substack{\mathbf{x} \in \mathbb{R}^p \\ \mathbf{z} \in \{0,1\}^p}} \quad & \frac{1}{2} \|\mathbf{y} - D\mathbf{x}\|_2^2 \\ \text{avec} \quad & -\mathbf{z}M \leq \mathbf{x} \leq \mathbf{z}M \\ & \mathbf{1}_p^T \mathbf{z} \leq T. \end{aligned} \quad (8)$$

Selon [4], la contrainte dans (8) est équivalent à celle de (2). De plus, D doit satisfaire la propriété d'*Unique Representation Property* [7] qui assure l'unicité de la solution. Analysons maintenant ce problème d'optimisation. D'abord, sa fonction objectif est quadratique. Ensuite, il comporte deux types des variables \mathbf{x} et \mathbf{z} respectivement continues et entières : c'est ce qu'on appelle un problème d'optimisation mixte en nombre binaires (ou plus généralement en nombre entiers). Enfin, les contraintes, quand à elles, sont linéaires. Ce type de problème est connu sous le nom de programme quadratique mixte en nombre binaires ou en anglais *mixed binary (integer) quadratic programming* (MIQP). Ce problème MIQP (8) peut être résolu exactement sur les images qui nous intéressent en utilisant un logiciel d'optimisation comme CPLEX ou GUROBI.

2.2.2 L'introduction de contraintes complémentaires

En programmation mixte, il est bien connu qu'une « bonne » formulation des contraintes peut grandement accélérer les performances d'un solveur [12]. Notamment, si dans le meilleur des cas on arrive à exprimer des contraintes définissant l'enveloppe convexe du domaine admissible, alors la solution optimale d'un programme mixte est la même que celle de sa relaxation continue [8]. Malheureusement, l'obtention de cette enveloppe convexe est un problème NP-difficile. En revanche, l'enveloppe convexe des contraintes sur les variables continues

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^p \mid \mathbf{z} \in \{0, 1\}^p, \sum_{j=1}^p z_j \leq T, |\mathbf{x}_j| \leq z_j T, \right\},$$

peut s'exprimer à l'aide des normes un et infinies de \mathbf{x} comme :

$$\left\{ \mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_1 \leq TM, \|\mathbf{x}\|_\infty \leq M \right\}.$$

Nous proposons d'ajouter ces contraintes au problème (8) pour obtenir le problème suivant équivalent mais mieux structuré :

$$\begin{aligned} \min_{\substack{\mathbf{x} \in \mathbb{R}^p \\ \mathbf{z} \in \{0, 1\}^p}} \quad & \frac{1}{2} \|\mathbf{y} - D\mathbf{x}\|_2^2 \\ \text{avec} \quad & -\mathbf{z}M \leq \mathbf{x} \leq \mathbf{z}M \\ & \mathbf{1}_p^T \mathbf{z} \leq T \\ & \|\mathbf{x}\|_1 \leq TM \\ & \|\mathbf{x}\|_\infty \leq M. \end{aligned} \quad (9)$$

Cette formulation permet au solveur d'obtenir la même solution que (8), mais plus rapidement.

Il reste que les performances des solveurs sur cette formulation sont très sensibles au choix de la constante M . Nous allons maintenant voir comment régler ce paramètre en utilisant une procédure proximale du premier ordre permettant en plus, d'obtenir une bonne initialisation des variables à optimiser.

2.2.3 Initialisation par la méthode du gradient proximal

L'algorithme du gradient proximal est une méthode du premier ordre permettant d'obtenir rapidement une solution locale du problème (2). Cette solution peut être utilisée comme une bonne initialisation des variables à optimiser et du paramètre M , permettant aux solveurs d'accéder plus rapidement au minimum globale du problème [2]. L'approche proximale consiste à minimiser itérativement une succession de majorations de la fonction objectif. Elle est construite à partir de l'opérateur proximal associé à la contrainte $\|\mathbf{x}\|_0 \leq T$:

$$\begin{aligned} \text{prox}_T : \mathbb{R}^p &\longrightarrow \mathbb{R}^p \\ \mathbf{x} &\longmapsto \text{prox}_T(\mathbf{x}) = \arg \min_{\|\mathbf{u}\|_0 \leq T} \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2. \end{aligned}$$

Il est facile de voir que la solution de ce problème est donnée par les T plus grandes valeurs absolues des composantes du vecteur \mathbf{x} , soit

$$\text{prox}_T(\mathbf{x}) = \begin{cases} x_j & \text{si } j \in \{(1), \dots, (T)\} \\ 0 & \text{sinon,} \end{cases}$$

où (j) est la suite d'indices tels que $|x_{(1)}| \geq \dots \geq |x_{(p)}|$. L'algorithme de descente de gradient proximal consiste alors à mettre en œuvre, pour un pas ρ , les itérations suivantes [3] :

$$\mathbf{x}^{k+1} \in \text{prox}_T(\mathbf{x}^k - \rho D^T(D\mathbf{x}^k - \mathbf{y})).$$

Lorsque le pas ρ est bien choisi, il est possible de démontrer que l'algorithme proximal converge vers un minimum local. Soulignons que, comme pour la plupart des méthodes itératives, il existe de nombreuses variantes de l'algorithme permettant d'accélérer la convergence.

Connaissant \mathbf{x}^* le point de convergence de l'algorithme proximal, il est possible d'en déduire une initialisation pour le vecteur \mathbf{z} et le paramètre M . Par exemple, pour un $\varepsilon > 0$ donné, on peut initialiser \mathbf{z} avec $z_j = 0$ si $|x_j^*| \leq \varepsilon$ et $z_j = 1$ sinon. La constante M peut être choisie telle que $M = (1 + \alpha)\|\mathbf{x}^*\|_\infty$ avec $\alpha > 0$ choisi le plus petit possible.

Par la résolution exacte de (2) pour déterminer le codage parcimonieux, la convergence de notre algorithme peut être assuré selon l'analyse conduite dans [1].

3 Résultats expérimentaux

Le but de nos expériences est de comparer les performances de notre méthode MIQP avec celles des algorithmes de référence sur une tâche de débruitage. Nous avons travaillé sur cinq images naturelles de bonne qualité fréquemment utilisées et extraites de *Miscellaneous volumes of the USC-SIPI Image Database*¹ (Barbara, Cameraman, Elaine, Lena et Men).

Nous avons construit la matrice Y des données d'apprentissage à l'aide des cinq images simultanément en utilisant, comme dans la littérature [6], des imagerie de taille 8×8 se chevauchant. Notre matrice Y a donc pour dimension $n = 64$ et $\ell > 3,5 \times 10^4$. Nous avons fixé expérimentalement le nombre d'atomes du dictionnaire à $p = 100$ et le coefficient de parcimonie à $T = 20$. Nous avons testé différents niveaux de bruit additif gaussien sur les images en utilisant trois différents niveaux de bruit avec des valeurs d'écart type $\sigma = 10, 20$ et 50 . Pour chaque expérience, nous avons utilisé la procédure de relaxation alternée et nous avons itéré 30 fois les deux phases successives de codage parcimonieux et d'apprentissage de dictionnaire.

Pour la phase de codage parcimonieux, nous avons comparé notre approche MIQP avec les deux méthodes références de la littérature, K-SVD [6] et la méthode proximale [3], toutes choses égales par ailleurs. Nous avons aussi comparé deux méthodes de reconstruction : la méthode directe où l'image est reconstruite par $D\mathbf{x}$ et l'approche proposée par Elad *et al.* [6] où elle est estimée par une combinaison linéaire entre Y et $D\mathbf{x}$.

Nous avons réalisé nos expériences en Matlab sur un PC Dell T5500 à 8 cœurs. Le nombre d'itérations maximal de la méthode proximale a été fixé à 200. Nous avons utilisé GUROBI 7.0 pour résoudre les MIQP avec un temps d'exécution maximal de 50 secondes, un nombre d'itération maximal de 200. Nous avons aussi fixé $\alpha = 1,5$ pour des raisons de stabilité.

1. <http://sipi.usc.edu/database/database.php?volume=misc>

La TABLE 1 résume nos résultats. La première remarque est que, pour un fort niveau de bruit ($\sigma = 50$) et sur toutes les images testées, notre méthode MIQP donne de meilleurs résultats (le rapport signal sur bruit est plus grand) que les autres approches de codage parcimonieux (K-SVD et l'algorithme proximal seul). L'amélioration peut être quantifiée en moyenne par une augmentation de 1,79 par rapport à la méthode proximale et de 3,73 par rapport à K-SVD, soit un gain de près de 20 %. Cela reste vrai quelle que soit la méthode de reconstruction utilisée. Nous avons aussi constaté que la méthode de reconstruction proposée par [6] donne systématiquement de meilleurs résultats. Cependant, pour un faible niveau de bruit ($\sigma = 10$), notre méthode ne réussit pas à améliorer les résultats de K-SVD alors que, pour un niveau intermédiaire ($\sigma = 20$) les résultats sont plus contrastés et dépendent de l'image considérée. Nous constatons que MIQP a tendance à moins bien se comporter sur des images de bonne qualité. D'une certaine manière, MIQP montre une meilleure capacité à éliminer le bruit via la parcimonie que les autres approches.

Quand nous comparons les résultats obtenus par chaque méthode pour différents niveaux de bruit, nous constatons aussi que ce sont ceux de MIQP qui diminuent le plus lentement que les autres méthodes lorsque σ augmente. Soulignons enfin que, contrairement aux algorithmes de référence de la littérature [6, 5, 3] pour lesquels le niveau de bruit doit être connu, la méthode proposée débruite l'image sans des connaissances a priori autre que le niveau de parcimonie souhaité.

4 Conclusion

Dans cet article nous avons proposé, pour résoudre un problème de débruitage, une véritable modélisation ℓ_0 de la parcimonie et la reformulation du problème associé sous la forme d'un programme quadratique mixte en nombre entiers (MIQP). Nous avons montré que les logiciels d'optimisation disponibles aujourd'hui pouvaient donner la solution globale du problème en un temps raisonnable, ce qui nous a permis de traiter des problèmes de débruitage sur de vraies images par une méthode itérative d'apprentissage de dictionnaire, exigeant à chaque itération la résolution de plus de 35 000 MIQP. Pour arriver à ce résultat, nous avons proposé deux techniques d'accélération du traitement des MIQP, la reformulation des contraintes pour mieux structurer le problème, et l'initialisation efficace de la procédure grâce à un algorithme proximal.

Nos résultats démontent d'abord la faisabilité de notre approche. Les progrès conjugués des logiciels, du matériel et de la modélisation (notre compréhension de la nature du problème), permettent aujourd'hui d'utiliser la programmation mixte en nombre entiers pour résoudre des problèmes de traitement d'image. Cela ouvre la porte à une nouvelle approche des problèmes de modélisation de la parcimonie, puisqu'il est maintenant possible de la gérer explicitement grâce aux programmes mixtes dont on est en mesure de calculer la solution globale, en dépit de leur caractère non convexe et NP difficile.

TABLE 1: Résultats en terme de rapport signal sur bruit (PSNR) avec reconstruction standard de l'image (à gauche) et avec la reconstruction proposée dans Elad *et al.* [6] (à droite). Les meilleurs résultats de chaque expérience sont en rouge.

σ image	method	PSNR			PSNR [6]		
		10	20	50	10	20	50
Barbara	K-SVD	32,15	27,48	19,71	33,6	28,25	20,03
	proximal	31,49	27,36	20,71	32,98	28,13	21,03
	MIQP	26,42	25,72	22,73	27,91	26,50	23,05
Cameraman	K-SVD	29,53	26,45	19,46	31,02	27,23	19,78
	proximal	28,80	26,75	21,11	30,30	27,53	21,43
	MIQP	25,90	25,25	22,30	27,39	26,03	22,62
Elaine	K-SVD	32,93	27,45	19,73	34,42	28,52	20,05
	proximal	33,14	28,99	22,87	34,63	29,77	23,19
	MIQP	30,88	29,09	24,20	32,38	29,87	24,52
Lena	K-SVD	33,61	27,91	19,79	35,10	28,69	20,11
	proximal	34,08	29,52	22,12	35,57	30,30	22,44
	MIQP	30,82	29,07	24,20	32,31	29,85	24,52
Men	K-SVD	31,95	27,36	19,68	33,45	28,14	20,00
	proximal	31,62	28,20	21,26	33,11	28,98	21,58
	MIQP	28,47	27,38	23,59	29,97	28,16	23,91

Références

- [1] M. Aharon, M. Elad, and A. Bruckstein. *k*-svd : An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11) :4311–4322, 2006.
- [2] A. Atamtürk and M. W. Savelsbergh. Integer-programming software systems. *Annals of Operations Research*, 140(1) :67–124, 2005.
- [3] C. Bao, H. Ji, Y. Quan, and Z. Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3858–3865, 2014.
- [4] S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau. Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance. *IEEE Transactions on Signal Processing*, 64(6) :1405–1419, 2016.
- [5] H. P. Dang and P. Chainais. Towards dictionaries of optimal size : A bayesian non parametric approach. *Journal of Signal Processing Systems*, pages 1–12, 2016.
- [6] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12) :3736–3745, 2006.
- [7] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using focuss : A re-weighted minimum norm algorithm. *IEEE Transactions on signal processing*, 45(3) :600–616, 1997.
- [8] K. L. Hoffman and T. K. Ralphs. Integer and combinatorial optimization. In *Encyclopedia of Operations Research and Management Science*, pages 771–783. Springer, 2013.
- [9] P. Honeine. Analyzing sparse dictionaries for online learning with kernels. *IEEE Transactions on Signal Processing*, 63(23) :6343–6353, December 2015.
- [10] J. Mairal, F. Bach, J. Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3) :85–283, 2014.
- [11] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM, 2009.
- [12] A. Neumaier and O. Shcherbina. Safe bounds in linear and mixed-integer linear programming. *Mathematical Programming*, 99(2) :283–296, 2004.
- [13] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set : A strategy employed by v1? *Vision research*, 37(23) :3311–3325, 1997.
- [14] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12) :4655–4666, 2007.
- [15] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.