Une brève histoire de la parcimonie : du traitement de signal à l'apprentissage profond

Rémi GRIBONVAL[1] Elisa RICCIETTI[2]

[1]Inria, CNRS, ENS de Lyon, Université Claude Bernard Lyon 1, LIP, UMR 5668, 69342, Lyon cedex 07, France

[2]ENS de Lyon, CNRS, Inria, Université Claude Bernard Lyon 1, LIP, UMR 5668, 69342, Lyon cedex 07, France

Résumé – La notion de parcimonie joue un rôle transverse en traitement du signal et de l'image: elle permet d'aborder des tâches en apparence aussi diverses que la compression, le débruitage, la séparation de sources, l'acquisition compressée, et plus généralement les problèmes inverses. Son histoire illustre en quoi cette notion, avant tout naturelle quand on cherche à compresser des signaux, s'est aussi avérée une propriété précieuse quand on cherche à les reconstruire à partir d'observations incomplètes. Le savoir-faire associé a donné lieu à des algorithmes combinant garanties de performance et complexité bornée pour les problèmes inverses, mais comme nous allons le voir leur extension dans un cadre de réseaux profonds pour l'apprentissage frugal réserve aussi quelques surprises et soulève de nouveaux défis scientifiques.

Abstract – The notion of sparsity plays a transveral role in signal and image processing: it allows to tackle tasks as seemingly diverse as compression, denoising, source separation, compressed sening and, more generally, inverse problems. Its history illustrates how sparsity, which appears as a natural objective for data compression, is also a valuable *prior* for data reconstruction from incomplete observations. The associated know-how has given rise to algorithms combining guaranteed performance and bounded complexity for inverse problems, but as we shall see, their extension into a deep learning context for frugal learning also brings a few surprises and raises challenging new research questions.

1 Parcimonie: les origines

Un vecteur est dit parcimonieux si la plupart de ses coordonnées sont nulles. Un vecteur avec une telle propriété est naturellement compressible, puisqu'il peut être décrit en indiquant simplement les quelques indices associés à des coordonnées non nulles, ainsi que les valeurs de ces coordonnées. La parcimonie de représentations de signaux et d'images dans des domaines adaptés s'avère donc une sorte de don de la nature pour faciliter leur compression. Ainsi, les techniques de compression en ondelettes développées dans les années 90 s'appuient sur le fait qu'on peut obtenir de très bonnes approximations des signaux réguliers par morceaux avec des combinaisons linéaires d'un faible nombre d'ondelettes, exploitant la parcimonie approximative du vecteur de représentation en ondelettes. On retrouve le même phénomène pour la compression de signaux audio via des représentations temps-fréquence (Figure 1.1).

De la compression au débruitage. Lorsqu'un signal ou une image avec une telle représentation parcimonieuse est contaminé par du bruit additif Gaussien, on peut aussi exploiter cette propriété de parcimonie pour le débruiter, en seuillant les coefficients d'ondelettes du signal bruité pour ne garder que les plus grands avant de procéder à la transformée en ondelettes inverse. Comme établi par D. Donoho et I. Johnstone [12], un tel procédé de débruitage est statistiquement quasi-optimal.

Parcimonie dans un dictionnaire. Une avancée majeure, qui ouvre la voie à tout un pan de travaux exploitant la parcimonie pour les problèmes inverses et l'échantillonnage compressé (nous y reviendrons), est le passage de représentations de signaux dans des bases (souvent orthogonales) à des représentations dans des familles redondantes, appelées dictionnaires [27], et dont les éléments sont appelés des atomes. Alors que la représentation d'un vecteur comme combinaison linéaire d'éléments d'une base est toujours unique, il y a une infinité de représentations dans un dictionnaire : cela offre en effet une flexibilité accrue pour choisir une représentation aussi propice que possible à la tâche considérée (compression, débruitage ...).

Algorithmes de décomposition parcimonieuse. L'idée d'exploiter de tels dictionnaires redondants pour calculer des approximations non-linéaires, ou des décompositions atomiques, apparaît dans les années 90 en traitement du signal et de l'image sous deux formes complémentaires : dans le travail de S. Mallat et Z. Zhang sur Matching Pursuit [27], et dans celui de S. Chen et D. Donoho sur Basis Pursuit [5].

Matching Pursuit –de même que sa variante Orthonormal Matching Pursuit (OMP), présente en filigrane dès le travail de S. Mallat et Z. Zhang– est un algorithme qui sélectionne itérativement des atomes pour minimiser l'erreur résiduelle. Cette approche gloutonne est inspirée de Projection Pursuit [16], une technique de régression introduite en statistiques par J. Friedmann et J. Tukey dans les années 70, également très liée à l'algorithme CLEAN de J. Högborn pour la déconvolution en radio-astronomie [23]. Laurent Jacques mentionne aussi dans une note de blog [1] des travaux des années 30 où des idées similaires apparaissent pour résoudre « à la main » de façon approchée des systèmes linéaires, avant l'apparition des premiers ordinateurs programmables.

1. Matching Pursuit Before Computer Science, June 2008, https://laurentjacques.gitlab.io/post/matching-pursuit-before-computer-science/

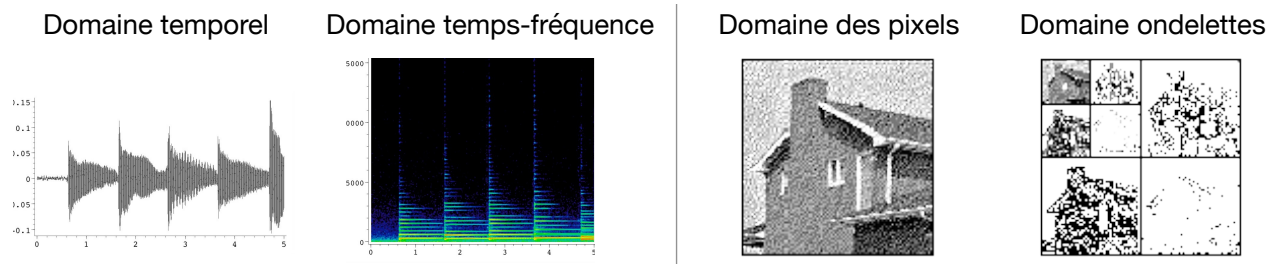

Figure 1.1 – Exemples de signaux et images naturels illustrant la parcimonie de leur représentation dans un domaine adapté.

Plus qu'un algorithme, *Basis Pursuit* (ainsi que son jumeau LASSO, introduit de façon quasi-concomitante en statistiques par R. Tibshirani [34]) formule la sélection d'atomes de façon indirecte comme un problème convexe de minimisation de norme L1 —une idée alternative aux moindres carrés qu'on trouve déjà dans les travaux de J. Claerbout et F. Muir [6] dans les années 70— et nécessite donc de faire appel à des algorithmes d'optimisation sur lesquels nous reviendrons. En présence de bruit, Basis Pursuit denoising formule un problème de minimisation comportant deux termes : un terme d'attache aux données quadratique, auquel s'ajoute un terme de pénalité L1 dont les propriétés mathématiques favorisent des solutions parcimonieuses.

2 Des heuristiques aux garanties

Étant donné un vecteur z et une matrice $\mathbf{A}$, le problème idéalisé de représentation parcimonieuse consiste à chercher le vecteur x le plus parcimonieux tel que $z = \mathbf{A}x$. Des variantes existent dans le cas de représentations approchées, mais bien que l'on réalise vite que tous ces problèmes sont essentiellement combinatoires et NP-difficiles [28], le comportement empirique de Basis Pursuit et de Matching Pursuit ne manque pas d'intriguer [11] : dans un cadre synthétique où un signal est généré à partir d'une représentation parcimonieuse « de référence » x_0 sous la forme $z = \mathbf{A}x_0$, ces approches sont souvent capables de retrouver *exactement* ladite représentation. Ces phénomènes empiriques sont rapidement exploités en séparation de sources [2, 20], avec notamment la notion d'analyse en composantes parcimonieuses ou morphologiques [33, 21]. L'idée générale est que si on observe la superposition $z = z_1 + z_2$ de deux signaux admettant chacun une représentation parcimonieuse, $z_i = \mathbf{A}_i x_i$, alors sous des conditions empiriquement favorables on peut identifier x_1, x_2 (et donc reconstruire z_1, z_2) en cherchant la représentation *la plus parcimonieuse* de $z = [\mathbf{A}_1, \mathbf{A}_2]\binom{x_1}{x_2}$.

Garanties pour les problèmes inverses parcimonieux. Le deuxième jalon clé du domaine est, au début des années 2000, la caractérisation des premières conditions formelles garantissant l'identification de représentations parcimonieuses via le principe de minimisation L1 [11] ou via l'algorithme OMP [36]. Il semble que ce soit aussi le moment où émerge la notion de « pseudo-norme » L0 d'un vecteur z : cette « norme » correspond simplement au nombre de composantes non nulles de z, et la notation est analogue aux classiques normes Lp, $\|x\|_p^p := \sum_i |x_i|^p$, puisque $\|x\|_0 = \sum_i |x_i|^0$ avec la convention que si $x_i = 0$ on a $|x_i|^0 = 0$, tandis que $|x_i|^0 = 1$ pour $x_i \neq 0$. Le problème idéalisé de représentation parcimonieuse, dont la minimisation L1 est la *relaxation convexe*, s'écrit ainsi

$$\min_x \|x\|_0 \text{ s.t. } z = \mathbf{A}x$$

Rôle de l'optimisation proximale. On peut raisonnablement supposer que la mise en lumière des garanties de reconstruction parcimonieuse par régularisation L1 –avec des techniques d'analyse déjà présentes dans les travaux précurseurs de Jean-Jacques Fuchs [17]– joue un rôle d'aiguillon pour motiver le développement d'algorithmes proximaux [8]. Ces algorithmes rendent facilement accessible la minimisation L1 pénalisée de la formulation *Basis Pursuit*/LASSO,

$$\min_x \tfrac{1}{2}\|z - \mathbf{A}x\|_2^2 + \lambda\|x\|_1,$$

jusqu'alors réservée aux experts car nécessitant de faire appel à de coûteux algorithmes génériques de programmation linéaire ou quadratique, avec des paramètres complexes à régler. Les premières garanties en reconstruction parcimonieuse sont vite améliorées et étendues, elles donnent par ailleurs un cadre théorique et algorithmique flexible et solide [14, 3] dont le déploiement couvre aujourd'hui une très large gamme de problèmes inverses sous-déterminés [32].

Naissance de l'échantillonnage compressif. Une conséquence majeure des garanties de reconstruction sous hypothèse de parcimonie est l'apparition du concept d'échantillonnage compressif (*compressed sensing*, terminologie due à Donoho [10] pour des idées —notamment l'échantillonnage aléatoire dans le domaine de Fourier-– attribuées au travaux pionniers de Candès-Romberg-Tao [4]). Il s'agit de s'affranchir des contraintes inhérentes aux problèmes inverses linéaires classiques, dans lesquels la matrice $\mathbf{A}$ est en quelque sorte *subie*, et peut donc avoir des caractéristiques défavorables limitant de facto les niveaux de parcimonie pour lesquels des garanties de reconstruction sont possibles.

L'idée centrale de l'échantillonnage compressif, est la suivante : quand on conçoit un dispositif physique pour effectuer des mesures linéaires de signaux analogiques (par exemple, un radiotélescope, ou un scanner IRM ...), pourquoi ne pas profiter des degrés de liberté dont on dispose pour s'assurer que la matrice correspondante ait de bonnes propriétés vis-à-vis de la régularisation parcimonieuse ? L'échantillonnage compressif consiste ainsi à guider la conception du dispositif d'acquisition des données, modélisé par $\mathbf{A}$, pour en améliorer les performances de reconstruction sous hypothèse de parcimonie et réduire significativement les temps ou les coûts d'acquisition.

Matrices aléatoires et échantillonnage compressif. De façon a priori surprenante, l'optimisation des propriétés de $\mathbf{A}$ passe par l'exploitation volontaire d'aléa dans sa conception, un principe qui a une longue histoire en informatique théorique – depuis les travaux pionniers de Ph. Flajolet jusqu'à ceux de A. Gilbert, Y. Avridis, S. Muthukrishnan et M. Strauss [18], qui ont de fait directement inspiré le développement de l'échantillonnage compressif – mais aussi en imagerie par rayons X avec les techniques de masques à ouverture codée [1].

La parcimonie dans tous ses états. De ses origines philosophiques – le rasoir d'Occam – la notion de parcimonie n'a cessé d'être un concept évolutif, souvent flou et qualitatif [13], et toujours en perpétuel renouvellement. De fait, la parcimonie prend aujourd'hui de nombreux visages nouveaux, où la représentation explicite via des vecteurs parcimonieux dans un domaine connu (par exemple en ondelettes) ou inconnu (via la notion d'*apprentissage de dictionnaire* [35] – un précurseur des *auto-encodeurs* et de l'*apprentissage de représentation*) a donné naissance à de nombreuses variantes. Il s'agit notamment d'approches rendant compte d'une diversité de connaissances a priori sur les données, qui vont notamment de la parcimonie structurée [24] ou sociale [25] à la parcimonie continue [9], et des modèles de rang faible à la *sparse PCA* [29] ou au *Graphical LASSO* [15]. Mais au-delà, les avatars modernes de la parcimonie capturent en essence la faible dimension intrinsèque d'un modèle structuré de données : même si celles-ci sont au premier abord dans des espaces de très grande dimension, il existe souvent des indices comme quoi elles peuvent être exprimées avec un nombre réduit de paramètres ou degrés de liberté, une idée que l'on retrouve notamment dans les modèles génératifs de type GAN (*generalized adversarial networks*).

3 Parcimonie et apprentissage frugal?

En complément de la capacité à *compresser* des données via la parcimonie *de leurs représentations*, le concept de parcimonie apparaît aussi en filigrane en traitement du signal et de l'image pour *traiter efficacement* ces donnés via des transformées rapides (Fourier, ondelettes): l'efficacité de ces dernières est en effet intrinsèquement liée à l'existence d'une *factorisation creuse* des matrices correspondantes (voir Figure 3.2).

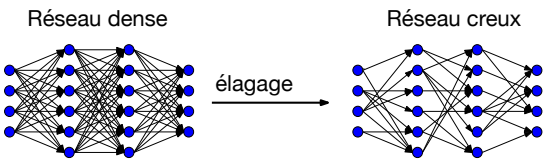

Figure 3.1 – L'élagage des couches denses d'un réseau de neurones est un objectif naturel pour en réduire l'empreinte mémoire mais aussi le coût de calcul, de façon analogue aux transformées rapides de la Figure 3.2.

Il est bien sûr tentant d'exploiter ces idées dans le contexte de l'apprentissage profond (voir Figure 3.1), où le besoin n'a jamais été aussi grand de mieux contrôler les compromis entre complexité et performance des modèles neuronaux devenus de taille astronomique. Peut-on s'appuyer pour cela sur le savoir-faire théorique et algorithmique en régularisation parcimonieuse pour les problèmes linéaires inverses ?

Cette question a été abordée dans un recent tutoriel sur la parcimonie profonde [22], qui en étudie les malédictions et les bénédictions. Les auteurs tirent en particulier trois leçons des problèmes inverses, et montrent que celles-ci ne sont en fait plus valides en contexte profond:

1. *la minimisation $L1$ (resp. $L2$) induit des solutions creuses (resp. "lisses"):* en apprentissage profond, la minimisation L2 correspond au *weight decay*, et celui-ci favorise des solutions *creuses ou de rang faible* [31, 30] plutôt que lisses.
2. *les approches gloutonnes ou par seuillage permettent d'identifier les coefficients significatifs:* à cause des symétries de remise à l'échelle dans les réseaux ReLU, la notion de "grand" coefficient est mal définie, rendant inopérantes les approches du type glouton. Un exemple qui explicite bien ce concept est celui de l'élagage (ou pruning): si l'on met à zéro des poids dans un réseau via la technique de *iterative magnitude pruning* (IMP), une simple remise à l'échelle adversaire des poids peut complètement changer le résultat.
3. *retrouver un vecteur creux est un simple problème de moindres carrés si son support (les indices des coordonnées non nulles) est connu, la difficulté essentielle est d'identifier ce support:* en contexte profond le probleme reste difficile même à support fixé, et ceci déjà dans le cas le plus simple de la factorisation de matrice à deux facteurs. La difficulté n'est pas seulement dûe à la présence possible de minima locaux qui compliquent l'optimisation, celle-ci peut même être intrinsèquement mal posée et forcer la divergence à l'infini des coefficients au cours des itérations de descente de gradient.

Pour surmonter ces difficultés inattendues, une piste féconde consiste à chercher des formes *structurées* de parcimonie profonde susceptibles d'en garder les bénéfices en termes d'économie de ressources tout en garantissant des propriétés favorables des problèmes d'optimisation associés. Si de premières pistes [7, 26, 19] ont mis en lumière les promesses de la structure *papillon* sous-jacente à la transformée de Fourier rapide, il reste encore beaucoup à comprendre dans ce nouveau champ de recherche.

Remerciements Ce travail a été soutenu par les projets ANR AllegroAssai ANR-19-CHIA-0009 et SHARP ANR-23-PEIA-0008 (financés par France 2030), et MEPHISTO ANR-24-CE23-7039, ainsi que le projet MOMIGS du GdR ISIS. Les auteurs remercient Léon Zheng et Quoc-Tung Le pour les échanges sur les sujets couverts ici.

References

[1] J. G. Ables. Fourier Transform Photography: A New Method for X-Ray Astronomy. *Pub. Astron. Soc. Australia*, 1(4):172–173, December 1968.

[2] L. Benaroya, R. Gribonval, and F. Bimbot. Représentations parcimonieuses pour la séparation de sources avec un seul capteur. In *Colloque GRETSI 2001*, Toulouse, France, 2001.

[3] Q. Bertrand, Q. Klopfenstein, P.-A. Bannier, G. Gidel, and M. Massias. Beyond L1: Faster and Better Sparse Models with skglm. In *NeurIPS 2022*, New Orleans, United States, November 2022.

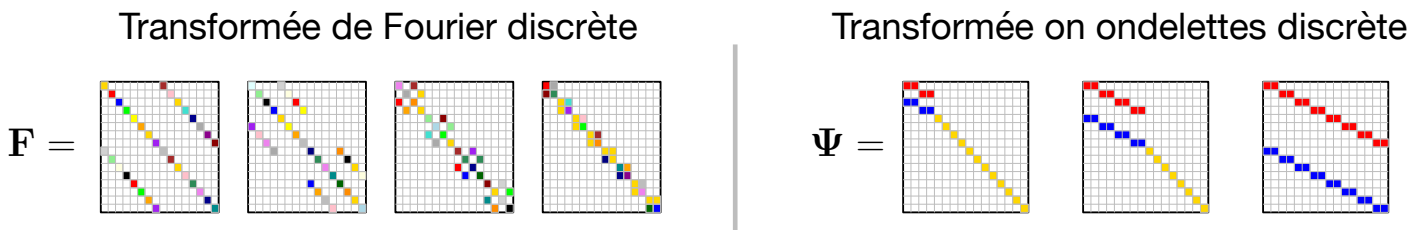

Figure 3.2 – Parcimonie cachée des transformées rapides usuelles. La matrice $n \times n$ de la transformée de Fourier discrète est un produit de $\log_2 n$ facteurs à structure "papillon", chacun avec deux coefficients non nuls par ligne et par colonne, d'où la complexité $O(n \log_2 n)$ de la FFT. La matrice de la transformée en ondelettes discrète est un produit de facteurs creux associés à des filtres décimés, d'où une complexité $O(n)$. Le nombre de multiplications est borné par le nombre total de coefficients non nuls.

[4] E. J. Candès, J. K. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Th.*, 52(2):480–509, 2006.

[5] S. Chen and D. Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44 vol.1, October 1994. ISSN: 1058-6393.

[6] J. F. Claerbout and F. Muir. Robust modeling with erratic data. *GEOPHYSICS*, 38(5):826–844, October 1973.

[7] T. Dao, B. Chen, N. Sohoni, A. Desai, M. Poli, J. Grogan, A. Liu, A. Rao, A. Rudra, and C. Ré. Monarch: Expressive structured matrices for efficient and accurate training. In *ICML*, 2022.

[8] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm Pure & Appl. Math.*, 57(11):1413–1457, 2004.

[9] Y. de Castro and F. Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and Applications*, 395(1):336–354, November 2012.

[10] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Th.*, 52(4):1289–1306, 2006.

[11] D. L. Donoho and X. Huo. Uncertainty Principles and Ideal Atomic Decompositions. *IEEE Trans. Inf. Th.*, 47(7):2845–2862, 2001.

[12] D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes-rendus Acad. Sc. Série I, Mathématique*, 319:1317–1322, 1994.

[13] G. A. Ferguson. The concept of parsimony in factor analysis. *Psychometrika*, 19(4):281–290, 1954.

[14] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, New York, NY, 2013.

[15] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[16] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Computers*, C-23:881–889, 1974.

[17] J. J. Fuchs. Une approche à l'estimation et l'identification simultanées. In *Colloque GRETSI 1997*, Grenoble.

[18] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *VLDB*, pages 79–88, 2001.

[19] Antoine Gonon, Léon Zheng, Pascal Carrivain, and Quoc-Tung Le. Fast Inference with Kronecker-Sparse Matrices. In *ICML 2025*, Vancouver (BC), Canada, July 2025.

[20] R. Gribonval. Sparse decomposition of stereo signals with Matching Pursuit and application to blind separation of more than two sources from a stereo mixture. In *ICASSP 2002.*, pages III/3057–III/3060, Orlando, Florida, 2002.

[21] R. Gribonval and S. Lesage. A survey of Sparse Component Analysis for blind source separation: principles, perspectives, and new challenges. In *ESANN'06*, Bruges (Belgium), 2006.

[22] R. Gribonval, E. Riccietti, Q.-T. Le, and L. Zheng. Rapture of the deep: highs and lows of sparsity in a world of depths, February 2025. preprint hal-04954574.

[23] J. A. Högbom. Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *Astronomy and Astrophysics Supplement Series*, 15:417, June 1974.

[24] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured Variable Selection with Sparsity-Inducing Norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[25] M. Kowalski, K. Siedenburg, and M. Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *IEEE Trans. Sig. Proc.*, 2013.

[26] Q.-T. Le, L. Zheng, E. Riccietti, and R. Gribonval. Butterfly factorization with error guarantees, November 2024. arXiv:2411.04506.

[27] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Sig. Proc.*, 41(12):3397–3415, 1993.

[28] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 25(2):227–234, 1995.

[29] D. Papailiopoulos, A. Dimakis, and S. Korokythakis. Sparse PCA through Low-rank Approximations. In *ICML 2013*, May 2013.

[30] R. Parhi and R. D. Nowak. Deep Learning Meets Sparse Regularization: A signal processing perspective. *IEEE Signal Processing Magazine*, 40(6):63–74, September 2023.

[31] M. Pilanci and T. Ergen. Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks. In *ICML 2020*.

[32] N. Pustelnik and C. Chaux. Evolution de la résolution de problèmes inverses en imagerie. In *Colloque GRETSI 2023*, Grenoble.

[33] J.-L. Starck, Y. Moudden, J. Bobin, M. Elad, and D. L. Donoho. Morphological Component Analysis. In *SPIE Conf. Wavelets*, 2005.

[34] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *J. Royal Stat. Soc. Series B (Method.)*, 58(1):267–288, 1996.

[35] I. Tosic and P. Frossard. Dictionary Learning. *IEEE Signal Processing Magazine*, 28(2):27–38, March 2011.

[36] J. A. Tropp. Greed is Good: Algorithmic Results for Sparse Approximation. *IEEE Trans. Inf. Th.*, 50(10):2231–2242, 2004.