Blind Denoising with Random Greedy Pursuits

Manuel Moussallam, Alexandre Gramfort, Laurent Daudet, and Gaël Richard

***Abstract*—Denoising methods require some assumptions about the signal of interest and the noise. While most denoising procedures require some knowledge about the noise level, which may be unknown in practice, here we assume that the signal expansion in a given dictionary has a distribution that is more heavy-tailed than the noise. We show how this hypothesis leads to a stopping criterion for greedy pursuit algorithms which is independent from the noise level. Inspired by the success of ensemble methods in machine learning, we propose a strategy to reduce the variance of greedy estimates by averaging pursuits obtained from randomly subsampled dictionaries. We call this denoising procedure Blind Random Pursuit Denoising (BIRD). We offer a generalization to multidimensional signals, with a structured sparse model (S-BIRD). The relevance of this approach is demonstrated on synthetic and experimental MEG signals where, without any parameter tuning, BIRD outperforms state-of-the-art algorithms even when they are informed by the noise level. Code is available to reproduce all experiments.**

***Index Terms*— Please add index terms.**

I. Introduction

TIME series obtained from experimental measurements are always contaminated by noise. Separating the informative signal from the noise in such raw data is called *denoising* and requires some assumptions on the signals and/or noise, for instance imposing a sparse model on the discrete signal y, of finite size N:

$$y = \Phi\alpha + w,$$

where $\Phi \in \mathbb{R}^{N\times M}$ is a (usually overcomplete) dictionary of M elementary objects ϕ_m called atoms and assumed normalized (*i.e.* $\forall m, \|\phi_m\|_2 = 1$), α is a sparse vector (*i.e.* $\|\alpha\|_0 = k \ll M$), and w is the additive noise to be removed. This model thus expresses the informative part of the signal as a sparse expansion in Φ and implicitly states that the noise component has no such expansion. Under this assumption, a denoised estimate $\hat{y} = \Phi\hat{\alpha}$ of y can be obtained by solving:

$$\hat{\alpha} = \arg\min_{\alpha\in\mathbb{R}^M} \|\alpha\|_0 \text{ subject to } \|y - \Phi\alpha\|_2 \leq \epsilon. \tag{1}$$

The value ϵ must be chosen according to the noise level (*i.e.* the norm of w). As problem (1) is NP hard, it is approximately solved using greedy algorithms [1]–[5], or via convex relaxations (*e.g.* Basis Pursuit Denoising [6]). A greedy algorithm, such as Matching Pursuit (MP) [1] and variants, will iteratively build an estimate $\hat{y}$ by selecting atoms in Φ and updating a residual signal accordingly. This latter class of methods suffers from two main limitations: (i) choosing a good value for ϵ, *i.e.* in practice a stopping rule for the algorithm, requires some knowledge on the noise variance, and (ii) the obtained approximation strongly depends on the dictionary design.

The contributions of this paper are threefold. First, we derive a data-driven stopping criterion for greedy pursuits based on order statistics. This technique allows denoising without knowledge of the noise variance. Second, we show how randomized greedy pursuits can be combined to improve the performance and reduce the dependency on the dictionary choice. This new algorithm, called BIRD, can be generalized to the case of multidimensional signals (S-BIRD). Third, we use popular synthetic signals to compare the performance of the proposed method with state-of-the-art techniques (soft and hard thresholding using cycle spinning [7], stochastic MP [8] and randomized MP [9]). Results on experimental data obtained with magnetoencephalography (MEG) are presented.

II. Blind Denoising with Randomized Pursuit

A. Stopping Criterion for Greedy Denoising Methods

Greedy algorithms require a stopping criterion to control model complexity, and avoid under- or over-fitting. For denoising purposes, this stop ideally occurs when the residual equals the noise and all atoms selected so far only explain the signal. In practice, a clear distinction between signal atoms and noise atoms is not always available. In this context, an interesting measure with greedy approaches is the *normalized coherence* of a signal y in Φ as defined in [1]:

$$\lambda_\Phi(y) = \sup_{\phi\in\Phi} \frac{|\langle y, \phi\rangle|}{\|y\|_2}. \tag{2}$$

Let r^n be the residual signal at iteration n, it's energy decay can be expressed in terms of the normalized coherence by:

$$\frac{\|r^n\|_2^2}{\|r^{n-1}\|_2^2} = 1 - \lambda_\Phi^2(r^{n-1}). \tag{3}$$

This relation is essentially used to bound from above the convergence of the algorithm using the *coherence* of the dictionary $\Lambda(\Phi) = \inf_{x\in\mathbb{R}^N}(\lambda_\Phi(x))$. This value is useful to describe the worst case convergence scenario, *i.e.* the convergence rate for the signal that is least correlated with Φ. Considering the noise signal w as a realization of a stochastic process, one may be also interested in the value:

$$\Lambda_W(\Phi) = \mathbb{E}\left[\lambda_\Phi(w)\right]. \tag{4}$$

Denoising can then be achieved by selecting only atoms whose normalized coherence is significantly higher than this value [1], [3]. Estimating (4) is however uneasy, and is typically learned from a training set [3].

Manuscript received February 13, 2014; revised April 11, 2014; accepted June 19, 2014. Date of publication July 01, 2014; date of current version July 09, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chandra Sekhar Seelamantula.

M. Moussallam and L. Daudet are with the Institut Langevin - University Paris Diderot - ESPCI ParisTech, 75005 Paris, France (e-mail: manuel.moussallam@espci.fr; laurent.daudet@espci.fr).

A. Gramfort and G. Richard are with the Institut Mines-Telecom - Telecom ParisTech - CNRS/LTCI, 75014 Paris, France (e-mail: gael.richard@telecom-paristech.fr; alexandre.gramfort@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2334231

The novelty of our approach is to propose a closed form estimate of $\Lambda_W(\Phi)$ based on a stochastic argument and order statistics. Let us consider the projections of w over Φ as M realizations z_m^w of a random variable (RV) Z^w:

$$\forall m \in [1, M], z_m^w = \frac{|\langle w, \phi_m \rangle|}{\|w\|_2}. \tag{5}$$

In most cases, M is greater than N. It implies that the z_m are not independent: their joint distribution is intricate. However, for analytical simplifications, we will assume they are i.i.d. This simplification will allow us to derive a new bound that turns out to be near-optimal in the proposed experimental framework. When run on w (*i.e.* pure noise), a greedy algorithm such as MP, or Orthogonal MP (OMP [10]), will typically select the atom that maximizes (5). Let us denote by $Z_{(M)}^w$ the RV describing the maximum projection value among M samples of Z^w. It is also known as the last order statistic of Z^w, and its cumulative density function writes (see for instance [11]):

$$F_{(M)}^{Z^w}(z) = M \int_0^z \left(F_Z^w(z')^{M-1} f_Z^w(z') \right) dz', \tag{6}$$

where f_Z^w (respectively F_Z^w) is the probability (respectively cumulative) density function, PDF (respectively CDF) of Z^w. Given an assumed i.i.d. distribution of the noise projections in a dictionary of size M, (6) gives a closed form formula for the CDF of the maximum.

The intuition behind this work writes as follows: the value $p = 1 - F_{(M)}^{Z^w}(z)$ is the probability that the maximum correlation between a dictionary element and a pure noise is to be greater than z. Thus we need to design a dictionary such that unlikely observations indicate the presence of a signal.

Let us now make the assumption that the dictionary is designed such that: (i) the projections of the noise on its atoms are distributed according to a zero mean Gaussian distribution (GD) and (ii) the distribution of the projections of the informative part has a heavier tail than the GD. The GD model fits well a variety of practical situations (*e.g.* white noise in a windowed-Fourier dictionary) and is more general than the standard Gaussian noise hypothesis. A reasonable model for Z^w is thus a half-normal RV, for which (6) is easily computed. This allows us to replace the value in (4) by:

$$\Lambda_W(\Phi, p) = \frac{\sqrt{2}}{N} \sqrt{\left(1 - \frac{2}{\pi}\right)} \operatorname{erfinv} \left((1-p)^{\frac{1}{M}} \right), \tag{7}$$

where erfinv is the inverse error function.

The parameter p expresses the confidence in the model and thus controls how much an approximation shall fit the data. Large values of p can lead to overfitting while small values can be too conservative. In this sense, this parameter plays a similar role to the more classical approximation error in [1]. However, it is important to emphasize that p is set for a given dictionary independently of the noise level. Experimental results testing the sensibility of the method with respect to p are given in supplementary material.

B. Double Randomization

The underlying assumption of (1) is that the *sparsest* representation is the optimal choice. However, as shown by Elad *et al.* [9], a better strategy (in the sense of the mean squared error) is to sample a set of J random sparse approximations $\{\hat{y}^j\}_{j=1\cdots J}$ and average them. Such an approach would be named *ensemble method* in the statistical learning literature (see *e.g.* [12] chap. 16). The J randomized greedy decompositions are run in parallel with the following probabilistic selection procedure. Let r^{n-1} be the residual signal at iteration n. The n-th element index γ^n to be selected is chosen at random among the M columns ϕ_m of Φ with greater probabilities, *i.e.* with large inner products $|\langle \phi_{\gamma^n}, r^{n-1} \rangle|$.

Our strategy, based on the work in [13], extends this idea with a *Random Forest*-like approach [14]. At each iteration, the most correlated element from a random subset $\Phi_n \subset \Phi$ is selected. Here, Φ_n will be about 50 times smaller than Φ. This spares us the computation of the M inner products, while proper randomization scheme allows us to browse the whole dictionary across iterations, a strategy particularly interesting when dictionary elements are finely located in time and frequency (see [13]). Running J instances of this pursuit on a random sequence of subdictionaries (*i.e.* the equivalent of J random trees) yields a set $\{\hat{y}^j\}_{j=1\cdots J}$ of sparse approximations. They can then be averaged in order to obtain the denoised signal: $\tilde{y} = \frac{1}{J}\sum_j \hat{y}^j$. The complete algorithm is detailed in Algorithm 1.

Algorithm 1: Blind Random Pursuit Denoising (BIRD)

```
Input: $y$, $\Phi$, $J$, $\Lambda_W(\Phi, p)$
Output: $\tilde{y}$
for $j = 1 \cdots J$ do
    initialization: $n = 0, r^0 = y, \hat{y}^j = 0$;
    while condition do
        $n \leftarrow n + 1$;
        Draw at random $\Phi_n \subset \Phi$;
        Select $\phi_{\gamma^n} = \arg\max_{\phi \in \Phi_n} |\langle r^{n-1}, \phi \rangle|^2$;
        Update $\hat{y}^j \leftarrow \hat{y}^j + \langle r^{n-1}, \phi_{\gamma^n} \rangle \phi_{\gamma^n}$;
        and $r^n = y - \hat{y}^j$;
        condition $= \lambda_\Phi(r^{n-1}) > \Lambda_W(\Phi, p)$;
    end
end
$\tilde{y} = \frac{1}{J}\sum_j \hat{y}^j$;
```

III. Structured Sparse Model

In case of data acquired with multiple sensors, the sparse model can be extended to take the structure of the data into account. Let $Y \in \mathbb{R}^{N \times C}$ be the data matrix formed by stacking the signals recorded by C sensors. Given the same dictionary Φ, one seeks an approximate $\hat{Y} = \Phi \hat{A}$ of Y as a sparse expansion in Φ. The unstructured problem reads:

$$\hat{A} = \arg \min_{A \in \mathbb{R}^{M \times C}} \|A\|_0 \text{ subject to } \|Y - \Phi A\|_F \leq \epsilon, \tag{8}$$

where $\|\cdot\|_F$ stands for the Frobenius norm and $\|\cdot\|_0$ is the number of non-zero entries. The signal model for one sensor reads:

$$y_c = \Phi(\alpha_c \odot s) + w_c, \tag{9}$$

where y_c is the c-th column of Y, w_c is the noise recorded by sensor c, s is a binary sparse vector of zeroes or ones independent of the sensor, and α_c is a weight vector specific to the sensor that has the same support as s and whose values are all zeroes if c is not in the set Γ_C. The notation $\odot$ stands for the element-wise multiplication. In a matrix form, this writes $Y = \Phi A + W$ with A a matrix whose c-th column is full of zeroes if c is not in Γ_C and whose m-th row is full of zeroes if $s[m] = 0$. Such matrices typically arise when using mixed-norms for group-sparse approximation problems [15], such as:

$$\hat{A} = \arg \min_{A \in \mathbb{R}^{M \times C}} \|A\|_{2,1} \text{ subject to } \|Y - \Phi A\|_F^2 \leq \epsilon \quad (10)$$

where $\| \cdot \|_{2,1}$ is the $\ell_{2,1}$ norm as in [15]. Such problems are commonly addressed with greedy algorithms [16] or group soft-thresholding (Group-LASSO) [17]. It is known as a Multiple Measurement Vector (MMV) problem in the signal processing literature [16].

Adapting Algorithm 1 to the structured case requires a refined selection rule. Typically an atom is selected if it maximizes the sum of the projections over all sensors [16]. However, only a fraction of the sensors may simultaneously record the signal from a source. Let l, $0 < l \leq 1$, be this fraction. Let r_c^{n-1} be the residual signal of sensor c and let us write $p_{c,m}^{n-1} = |\langle \phi_m, r_c^{n-1} \rangle|^2$ and $p_{(c),m}^{n-1}$ the ordered projections of r_c^{n-1} on ϕ_m (*i.e.* $p_{(1),m}^{n-1} \leq \ldots \leq p_{(C),m}^{n-1}$). Let us denote:

$$\phi^n = \arg \max_{\phi \in \Phi} \frac{1}{\lfloor lC \rfloor} \sum_{c=\lfloor C(1-l) \rfloor}^{C} p_{(c),m}^{n-1}. \quad (11)$$

The selection procedure also yields a list Γ_c^n of $\lfloor lC \rfloor$ sensors containing the atom, and where an update is necessary.

In theory, it is again possible to use order statistics to model the sum in (11) and its maximum as RVs. In practice, a simple idea is to stop the decomposition once a given proportion of the most energetic signals have been denoised:

$$Cond\left[\lambda_\Phi(r_c^{n-1})\right] = \frac{1}{\lfloor lC \rfloor} \sum_{c=\lfloor C(1-l) \rfloor}^{C} \lambda_\Phi(r_{(c)}^{n-1}) > \Lambda_W(\Phi, p), \quad (12)$$

where the set $\{\lambda_\Phi(r_{(c)}^{n-1})\}$ corresponds to the $\lfloor lC \rfloor$ biggest values of $\lambda_\Phi(r_c^{n-1})$. This criterion is coherent with the selection rule (11). The complete procedure is summarized in Algorithm 2.

Algorithm 2 Structured Blind Random Pursuit Denoising (S-BIRD)

```
Input: $Y$, $\Phi$, $J$, $l$, $\Lambda_W(\Phi, p)$
Output: $\tilde{Y}$
for $j = 1 \cdots J$ do
    initialization: $n = 0, \forall c, r_c^0 = y_c, \hat{y}_c^j = 0$;
    while condition do
        $n \leftarrow n + 1$;
        Draw at random $\Phi_n \subset \Phi$;
        Select $(\phi_{\gamma^n}, \Gamma_c^n) =$
$\arg\max \frac{1}{\lfloor lC \rfloor} \sum_{c=\lfloor C(1-l) \rfloor}^{C} p_{(c),m}^{n-1}$;
        Update $\forall c \in \Gamma_c^n, \hat{y}_c^j \leftarrow \hat{y}_c^j + \langle r_c^{n-1}, \phi_{\gamma^n} \rangle \phi_{\gamma^n}$;
        and $r_c^n = y_c - \hat{y}_c^j$;
        condition $= Cond[\lambda_\Phi(r_c^{n-1})]$;
    end
end
$\tilde{Y} = \frac{1}{J} \sum_j \hat{Y}^j$;
```

IV. Experimental Validation

All the experiments and results shown in this section as well as those presented in the provided supplementary material can be reproduced using our Python code freely available online[1].

A. Synthetic Examples

A first set of experiments compares the proposed algorithm BIRD to existing mono-channel denoising techniques on synthetic signals and simulated MEG data. We present the performance of BIRD compared to various state-of-the art methods among which:

- Wavelet Shrinkage (WaveShrink) methods (with both soft and hard thresholding) using Daubechies wavelets and a Short-Time Fourier Transform (STFT) dictionaries. These dictionaries can be made shift-invariant using the Cycle-Spinning method [7], [18].
- Stochastic MP (SMP) as introduced in [8]. In this method, each of the J runs is performed on a subdictionary $\Phi_j \subset \Phi$, that is chosen at random once for each run, and kept unchanged in the whole decomposition.
- Randomized OMP (RandOMP) as introduced in [9]. In this method, atoms are selected at random in the complete dictionary Φ at every iteration of the J runs.

These methods require a stopping criterion, typically set by fixing the reconstruction error in accordance with the noise level. In contrast, our algorithms select an atom in a random subdictionary at each iteration of the J runs and derive their stopping criterion from the statistics of the projections as explained above. For comparison, we present the results obtained by WaveShrink (respectively SMP and RandOMP) methods in an *Oracle* case, that is when the true signal y is known and used to set the target reconstruction errors in order to minimize the errors.

For all greedy approaches, we use an overcomplete dictionary Φ built as a union of Modulated Discrete Cosine Transforms (MDCT) of 6 different scales. For each basis, atoms are further replicated and shifted so as to form a highly overcomplete, shift-invariant, dictionary Φ of size $M \gg N$. We set the overfitting probability to $p= 10^{-6} \sim 1/M$. This may seem a conservative value, but in practice the algorithm is not very sensitive to p (See Fig. 3 in supplementary materials).

One can verify in Fig. 1 that BIRD has some advantages over alternative methods. Two observations can be made. First, the double randomization scheme is valuable: with a limited number of runs J the resulting denoised signal with BIRD presents less disturbing artifacts than SMP or RandOMP methods. Second, the self-stopping criterion yields satisfying signal estimates in most cases. Note that no parameter is modified when varying the SNR. Given pure white noise as input, BIRD does not select any spurious atom. Given a pure sparse signal, BIRD will select atoms up to a very high reconstruction fidelity.

B. Simulated MEG Data

Publicly available software [?] has been used to simulate MEG signals. A controlled level of white or colored noise W (auto-regressive process fitted on real data) was added to a

[1] http://manuel.moussallam.net/birdcode

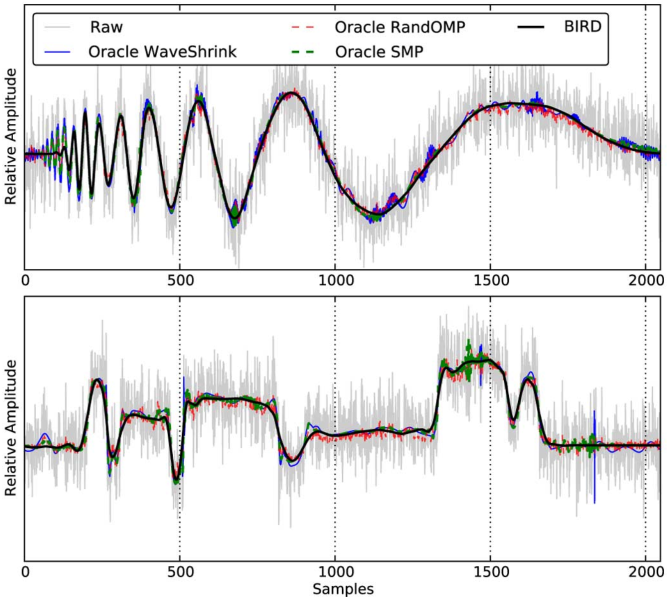

Fig. 1. Examples of denoising for synthetic signals *Doppler* and *Blocks*. SMP, RandOMP and BIRD are set with $J = 30$ runs and use a multiscale MDCT dictionary.

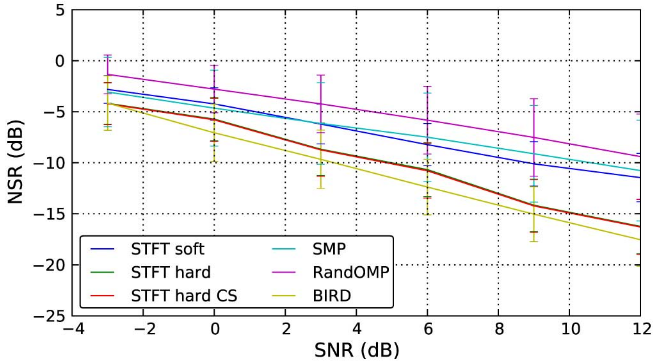

Fig. 2. NMSE for various denoising methods on simulated single-sensor MEG data (score averaged over 20 trials) as a function of the noise level.

collection of C smooth and oscillatory signals mimicking classical MEG evoked responses thanks to the use of a real forward solution. The recordings in the absence of noise are given in the form of a ground truth matrix X. Denoising methods can then be applied to $Y = X + W$ and compared using a Normalized Mean Squared Error (NMSE) ratio:

$$NMSE(\hat{Y}) = 10 \log_{10} \frac{\|X - \hat{Y}\|_F^2}{\|X\|_F^2}. \tag{13}$$

Fig. 2 illustrates the performance of the proposed approach in terms of reconstruction for a single-sensor signal. BIRD outperforms competitive methods, even informed by the oracle noise level.

Given the same test framework, we now evaluate the denoising capabilities on multichannel signals. We compare the performance of BIRD being applied independently to each sensor (*i.e.* not taking any structure into account) to the Structured version S-BIRD. In this simulation study, all channels contain information and we set $l = 1$. For comparison, we use group soft thresholding methods using wavelets (Daubechies wavelets with 3 vanishing moments) and STFT, and present the best results obtained while varying the threshold parameter (labelled Oracle Group WS). Finally, it is compared to SMP applied independently on each sensor. As shown on Fig. 3, by taking cross-sensor correlation into account at the tom selection level, S-BIRD improves over BIRD and outperforms other methods. This improvement is even more visible for colored noise and with real data (see supplementary material for additional figures).

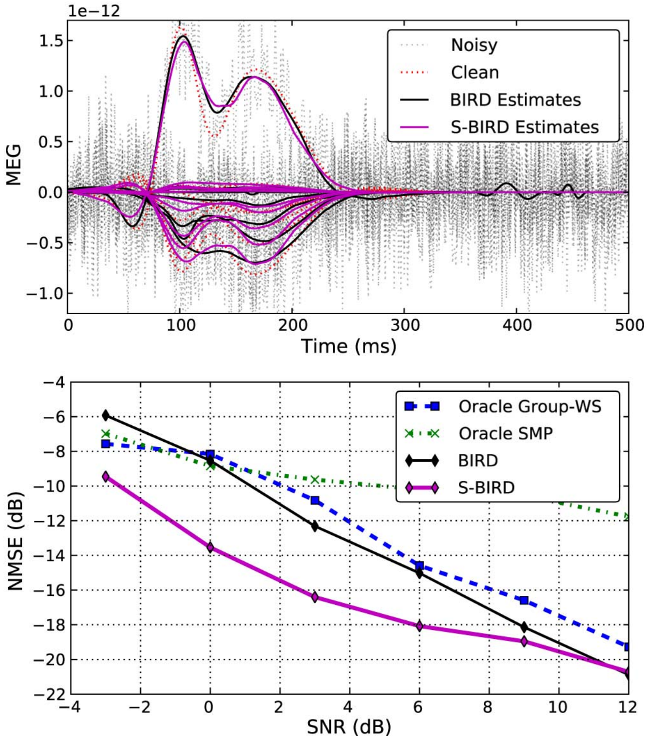

Fig. 3. Top: examples of BIRD and S-BIRD denoising for simulated multi-sensor MEG signals (evoked response corrupted by white noise). Bottom: NMSE (dB) for various methods.

V. Conclusion

We propose a greedy strategy that relies on averaging the results of multiple runs of random sequential pursuits, each of which can select a different number of atoms and reach a different approximation level that is determined by a signal-independent stopping criterion. The only parameter p depends solely on the dictionary.

The algorithm is fast as it avoids computing all projections while using FFT-based MDCT or wavelet dictionaries. An enhanced version S-BIRD, taking into account atom correlations between multiple sensors achieves even better results on simulated data with white or colored noise, as well as on MEG data. The multiple runs averaging strategy, also called *bagging* [19] in the machine learning literature, reduces the estimation variance of a single pursuit, and is a key ingredient of the BIRD algorithm.

References

[1] S. Mallat and Z. Zhang, “Matching pursuit with time-frequency dictionaries,” *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3414, 1993.

[2] O. Bertrand, J. Bohorquez, and J. Pernier, “Time-frequency digital filtering based on an invertible wavelet transform: An application to evoked potentials,” *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 77–88, Jan. 1994.

[3] P. J. Durka and K. J. Blinowska, “Analysis of EEG transients by means of matching pursuit,” *Ann. Biomed. Eng.*, vol. 23, no. 5, pp. 608–11, 1995.

[4] R. Quiroga and H. Garcia, “Single-trial event-related potentials with wavelet denoising,” *Clin. Neurophys.*, vol. 114, pp. 376–390, Feb. 2003.

[5] P. J. Durka, A. Matysiak, E. M. Montes, P. V. Sosa, and K. Blinowska, “Multichannel matching pursuit and EEG inverse solutions,” *J. Neurosci. Meth.*, vol. 148, pp. 49–59, Oct. 2005.

[6] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[7] R. Coifman and D. Donoho, Translation-invariant de-noising Dept. Statist., Stanford Univ., Stanford, CA, USA, Tech. Rep., 1995.

[8] P. Durka, D. Ircha, and K. Blinowska, “Stochastic time-frequency dictionaries for matching pursuit,” *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 507–510, Mar. 2001.

[9] M. Elad and I. Yavneh, “A plurality of sparse representations is better than the sparsest one alone,” *IEEE Trans. Inf. Theory*, vol. 55, pp. 4701–4714, Oct. 2009.

[10] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in *Proc. 27th Annu. Asilomar Conf. Signals, Systems, and Computers*, 1993, pp. 40–44.

[11] S. S. Wilks, “Order statistics,” *Bull. Amer. Math. Soc.*, vol. 54, pp. 6–50, 1948.

[12] T. Hastie, R. Tibshirani, and J. H. Friedman*, The Elements of Statistical Learning*. Berlin, Germany: Springer, Jul. 2003.

[13] M. Moussallam, L. Daudet, and G. Richard, “Matching pursuits with random sequential subdictionaries,” *Signal Process.*, vol. 92, pp. 2532–2544, 2012.

[14] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[15] M. Kowalski, “Sparse regression using mixed norms,” *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 303–324, 2009.

[16] J. Tropp and A. Gilbert, “Simultaneous sparse approximation via greedy pursuit,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2005, pp. 721–724.

[17] V. Roth and B. Fischer, “The group-lasso for generalized linear models: Uniqueness of solutions and efficient algorithms,” in *Int. Conf. Machine Learning*, 2008.

[18] U. Kamilov, E. Bostan, and M. Unser, “Wavelet shrinkage with consistent cycle spinning generalizes total variation denoising,” *IEEE Signal Process. Lett.*, vol. 19, pp. 187–190, Apr. 2012.

[19] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 140, pp. 123–140, 1996.