Computational Statistics and Data Analysis 134 (2019) 36–57

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simultaneous nonparametric regression in RADWT dictionaries

Daniela De Canditiis [a,*], Italia De Feis [b]

[a] Istituto per le Applicazioni del Calcolo "M. Picone", Rome, Italy
[b] Istituto per le Applicazioni del Calcolo "M. Picone", Naples, Italy

ARTICLE INFO

Article history:
Received 9 April 2018
Received in revised form 23 November 2018
Accepted 29 November 2018
Available online 10 December 2018

Keywords:
RADWT
Grouped LASSO
Multichannel

ABSTRACT

A new technique for nonparametric regression of multichannel signals is presented. The technique is based on the use of the Rational-Dilation Wavelet Transform (RADWT), equipped with a tunable Q-factor able to provide sparse representations of functions with different oscillations persistence. In particular, two different frames are obtained by two RADWT with different Q-factors that give sparse representations of functions with low and high resonance. It is assumed that the signals are measured simultaneously on several independent channels and that they share the low resonance component and the spectral characteristics of the high resonance component. Then, a regression analysis is performed by means of the grouped lasso penalty. Furthermore, a result of asymptotic optimality of the estimator is presented using reasonable assumptions and exploiting recent results on group-lasso like procedures. Numerical experiments show the performance of the proposed method in different synthetic scenarios as well as in a real case example for the analysis and joint detection of sleep spindles and K-complex events for multiple electroencephalogram (EEG) signals.

1. Introduction

This paper deals with the problem of simultaneously recovering K different signals independently or simultaneously recorded under the hypothesis that these signals share common characteristics. Indeed, when drawing K independent or simultaneous experiments over the same (unknown) causal relation among variables, we expect that changing the experiment should not affect the causal relation but only some experiment specific characteristics. This situation is typical in the biological field, where scientists make experiments with more replicas because they assume a causal relationship between genes and response common to all replicas, while retaining a replicate-specific variability, see He et al. (2016), Ruffalo et al. (2017), Yuan et al. (2016) and Deun et al. (2011); but common characteristics are also expected in the medical field to model special EEG data, where one waits the simultaneous signals derived from the electrodes located in the subject's scalp at specific areas, see Selesnick (2011), Barros et al. (2000) and Parekh et al. (2017). See Bobin et al. (2009) for many other examples applied to different signal and image processing problems.

Such kind of problem is addressed in many different research areas: in the machine learning community it is well known as the multi-task learning problem (Liu et al., 2008; Lozano and Swirszcz, 2012; Argyriou et al., 2008), in the signal and image processing community as the multi-channel recovering problem (Rakotomamonjy, 2011), in econometrics as the panel-data

* Correspondence to: Via dei Taurini 19, 00185 Roma, Italy.
E-mail address: d.decanditiis@iac.cnr.it (D. De Canditiis).

https://doi.org/10.1016/j.csda.2018.11.003
0167-9473/

problem, in the approximation theory as the conjoint analysis as well as in the mathematical statistics community it is a special case of the multivariate regression problem. The enormous interest which is growing around this problem is due to its flexibility in modeling different situations and in the possibility of using fast algorithm to solve it.

In this paper we propose to treat the problem of simultaneous nonparametric regression from a new perspective by combining results from signal processing and statistical high-dimensional data analysis. In signal processing it is now well understood that orthogonal basis decompositions are not appropriate for signal recovery, since they can often fail to represent a particular function of interest efficiently, (Donoho and Elad, 2003). As a result, overcomplete representations such as wavelets and windowed Fourier expansions became mainstays of modern statistics and signal processing. Such representations are formalized through the theory of frames. Frames can be generated by the action of operators on a template function (mother wavelet or Gabor atom), or be unstructured and random (as in compressive sensing). Here we use results about RADWT (Selesnick, 2011), which is a modern and fast computational tool for analyzing a very general class of signals. In statistical high-dimensional data analysis it is established that the grouped-Lasso technique (Yuan and Lin, 2006) for the selection and estimation of grouped variables is very effective to identify the dictionary elements that guarantee efficient estimation of the unknown regression function. The advantage of this approach is twofold. First, from a theoretical point of view, it is possible to control the estimation error by the so called *oracle inequalities*, and the error rate becomes nearly parametric providing the function of interest can be represented via a linear combination of just few dictionary elements satisfying certain assumptions. Second, from a computational point of view, the group gradient descendent method permits a very fast implementation of the optimization algorithm to find the optimal path.

The remainder of the paper is organized as follows. Section 2 describes the data model we are considering with the working hypothesis. Section 3 presents and discusses the inference procedure within the paradigm of group-lasso procedures, enlightening the connections with other existing procedures. Section 4 provides convergence results, while Section 5 shows numerical experiments.

2. The data model

Consider the problem of recovering $K+1$ deterministic vectors $\mathbf{c}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)} \in \mathbb{R}^{n\times 1}$ from the following data

$$\mathbf{y}^{(k)} = \mathbf{c} + \mathbf{u}^{(k)} + \boldsymbol{\varepsilon}^{(k)} \qquad k = 1, \dots, K \quad \text{and} \quad \boldsymbol{\varepsilon}^{(k)} \sim N(\mathbf{0}, \sigma^2 I), \tag{2.1}$$

where vector $\mathbf{y}^{(k)}$ represents n-equispaced observations of function $c(t) + u^{(k)}(t)$ over the equispaced grid design $t_1 < t_2 < \cdots < t_n$ for each channel $k = 1, \dots, K$, i.e. $\mathbf{y}^{(k)} \in \mathbb{R}^{n\times 1}$. The grid can be thought to be sampled in time, in space, in radiation, in genome locations or in any other unit of measure according to the physical phenomena. The data model (2.1) represents the situation where the samples share a common effect, here represented by function $\mathbf{c}(t)$ which eventually can be zero, plus a functional component $u^{(k)}(t)$ which can be different across samples while sharing some common characteristics to be specified later. We do not hypothesize functions $c(t)$ and $u^{(k)}(t)$ belong to some functional Sobolev space $H^s_{p,q}[a, b]$ as it is usually done in functional nonparametric regression setting, instead we let these functions to be much more general and we restrict our attention to their finite-dimensional representation. Since many physiological and physical signals are not only non-stationary but also exhibit a mixture of oscillatory and non-oscillatory transient behaviors (for example, speech, stock-market, biomedical EEG, etc.) we suppose that each signal in each channel is the sum of a 'high-resonance' and a 'low-resonance' component. By a high-resonance component, we mean a signal consisting of multiple simultaneous sustained oscillations, in contrast, by a low-resonance component, we mean a signal consisting of non-oscillatory transients of unspecified shape and duration. We stress that the high and low resonance components of a signal cannot be extracted from its high and low frequencies components in a time-scale decomposition, but they can be well represented by a high-Q factor RADWT and a low-Q factor RADWT respectively as very well explained in Selesnick (2011). The RADWT is a normalized tight frame of $L_2(R)$ defined as $\left\{(\frac{q}{p})^{k/2}\psi\left((\frac{q}{p})^k t + \frac{sp}{q}l\right)\right\}_{k,l\in Z}$ where ψ is a wavelet function and (p, q, s) is a triplet of parameters which gives the time-scale characteristic of the frame. In particular the ratio $q/p > 1$ is closely related to the scale (or frequency) dilatation factor, the parameter s is closely related to the time dilatation factor and $\frac{p}{s(q-p)}$ is the redundant factor. The Q-factor depends on these parameters although there is not an explicit formula, in particular setting the dilatation factor q/p between 1 and 2 and $s > 1$ gives a RADWT with high Q-factor, while setting $s = 1$ we obtain a low Q-factor RADWT with time-scale characteristic similar to the dyadic wavelet transform. In particular, when $q = 2$, $p = 1$ and $s = 1$ the frame reduces to the classical wavelet basis. Given a finite energy signal $\mathbf{x}$ of length n and $J \in \mathbb{N}$ levels of decomposition, the RADWT transform is obtained by a sequence of proper down-sampling operations and fast Fourier transforms; it ends up with $\lceil \frac{np^J}{q^J} \rceil$ scaling coefficients (low-pass filtering) and $\lceil \frac{np^j}{q^j s} \rceil$ wavelet coefficients (high-pass filtering) at each level $j = 1, .. J$. See Bayram and Selesnick (2009) for details on fast analysis and synthesis schemes. In this paper we use these results of signal processing in order to formulate our working hypothesis. Let $\boldsymbol{\Psi} \in \mathbb{R}^{n\times d1}$ be the finite matrix representation of the low Q-factor analysis filter and let $\boldsymbol{\Phi} \in \mathbb{R}^{n\times d2}$ be the finite matrix representation of the high Q-factor analysis filter (the synthesis operators being just the transpose matrices), then our working hypothesis is the following:

(H1) signal $\mathbf{c}$ is sparse in $\boldsymbol{\Psi}$, i.e. setting $\boldsymbol{\alpha}_0 = \Psi^t \mathbf{c}$ we have that $\left|S_0^{\boldsymbol{\alpha}}\right| = \left|\{j : \alpha_{0_j} \neq 0\}\right| \ll d1$;

(H2) signals $\mathbf{u}^{(k)}$ have a **jointly** sparse representation in $\boldsymbol{\Phi}$, i.e. setting $\boldsymbol{\beta}_0^{(k)} = \boldsymbol{\Phi}^t \mathbf{u}^{(k)}$ and $S_0^{(k),\boldsymbol{\beta}} = \{j : \beta_{0_j}^{(k)} \neq 0\}$ we have that $S_0^{(1),\boldsymbol{\beta}} = \cdots = S_0^{(K),\boldsymbol{\beta}}$, with the common cardinality denoted by $\left|S_0^{\boldsymbol{\beta}}\right| \ll d2$.

(H3) the columns of matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ are normalized to have norm 1.

Finally it is worth to observe that the role of $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ in this model can be interchanged to accomplish cases where the common effect $\mathbf{c}$ has a high Q-factor behavior as opposed to the sample specific effect which has a low Q-factor behavior.

3. Inference

The linear model in (2.1) can be rewritten in terms of RADWT coefficients as follows

$$\begin{cases} \mathbf{y}^{(1)} &= \boldsymbol{\Psi}\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon}^{(1)} \\ \mathbf{y}^{(2)} &= \boldsymbol{\Psi}\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{\beta}^{(2)} + \boldsymbol{\varepsilon}^{(2)} \\ &\vdots \\ \mathbf{y}^{(K)} &= \boldsymbol{\Psi}\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{\beta}^{(K)} + \boldsymbol{\varepsilon}^{(K)}, \end{cases} \tag{3.2}$$

which turns out to be a classical multiple regression model with a special common design matrix. A first and somewhat naive approach would consist in treating separately each channel ignoring the underlying common structure; however this is obviously suboptimal. This is the reason why such kind of problem is reformulated in terms of a unique regression problem in the following form:

$$\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(K)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{\Phi} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{\Psi} & \mathbf{0} & \boldsymbol{\Phi} & \cdots & \mathbf{0} \\ \cdots & & & \cdots & \\ \boldsymbol{\Psi} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Phi} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \\ \vdots \\ \boldsymbol{\beta}^{(K)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}^{(1)} \\ \boldsymbol{\varepsilon}^{(2)} \\ \vdots \\ \boldsymbol{\varepsilon}^{(K)} \end{bmatrix} = X\,\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{3.3}$$

with obvious correspondence between elements of the two expression. So, $\mathbf{y}$ is a column vector of nK response variables, $\boldsymbol{X}$ a design matrix of dimension $nK \times d_1 + Kd_2$, $\boldsymbol{\theta}$ an unknown regression coefficients column vector of length $d_1 + Kd_2$ consisting of a first sub vector $\boldsymbol{\alpha} \in \mathbb{R}^{d_1 \times 1}$ and a second sub vector $\boldsymbol{\beta} = \left[\left(\boldsymbol{\beta}^{(1)}\right)^t, \ldots, \left(\boldsymbol{\beta}^{(K)}\right)^t\right]^t \in \mathbb{R}^{Kd_2 \times 1}$ and, finally, we let $\boldsymbol{\varepsilon}$ be a nK-variate Gaussian random column vector with zero mean and covariance matrix $\sigma^2 \boldsymbol{I}_{nK}$. Under the working hypothesis (H1) and (H2), we expect the coefficients of the common part $\boldsymbol{\alpha}$ to be sparse into the dictionary $\boldsymbol{\Psi}$, while on the remaining part of coefficient vector $\boldsymbol{\beta}$ we exploit the joint sparsity assumption, i.e. for all $j = 1, \ldots, d_2$ we know that $\beta_j^{(k)} = 0$, for all $k = 1, \ldots, K$ or $\beta_j^{(k)} \neq 0$ for all $k = 1, \ldots, K$. This provides the following non-overlapping group structure for the whole vector $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}$:

$$\{1, 2..., d_1 + Kd_2\} = \{1\} \cup \cdots \cup \{d_1\} \cup G_1 \cup \cdots \cup G_{d_2}, \tag{3.4}$$

with

$$G_j = \{d_1 + j, d_1 + j + d_2, d_1 + j + 2d_2 \ldots, d_1 + j + (K-1)d_2\}, j = 1, \ldots, d_2,$$

group of size K. Let $G^\star = \frac{d_1 + Kd_2}{d1 + d2}$ denote the average group size and let us denote

$$\|\boldsymbol{\theta}\|_{2,1} = \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_{2,1} = \sqrt{\frac{1}{G^\star}} \sum_{j=1}^{d_1} |\alpha_j| + \sqrt{\frac{K}{G^\star}} \sum_{j=1}^{d_2} \|\boldsymbol{\beta}(G_j)\|_2,$$

the l_1/l_2-norm, with $\boldsymbol{\beta}(G_j)$ denoting the reduction of vector $\boldsymbol{\beta}$ to the subset of index G_j, then we can consider the following group lasso problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{(d1 + Kd_2) \times 1}}{argmin} \left\{ \frac{1}{nK} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \sqrt{G^\star} \|\boldsymbol{\theta}\|_{2,1}. \right\} \tag{3.5}$$

Finally, we consider as our estimator the following reconstructions:

$$\hat{\boldsymbol{c}} = \boldsymbol{\Psi}\hat{\boldsymbol{\alpha}}; \quad \hat{\boldsymbol{u}}^{(k)} = \boldsymbol{\Phi}\hat{\boldsymbol{\beta}}^{(k)}, \quad k = 1, \ldots, K, \tag{3.6}$$

where $\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \left[\hat{\boldsymbol{\alpha}}^t, \left(\hat{\boldsymbol{\beta}}^{(1)}\right)^t, \ldots, \left(\hat{\boldsymbol{\beta}}^{(K)}\right)^t\right]^t$ is the solution of the optimization problem (3.5).

3.1. Algorithm

As already mentioned in the introduction one of the great advantages of the grouped Lasso penalization consists in the availability of efficient algorithms for its solution.

In particular, the most efficient algorithms in the modern statistics literature are the Group Descendent Algorithm, presented in Breheny and Huang (2015, 2009) and implemented in the R package `grpreg` available at https://cran.r-project.org/web/packages/grpreg/, and the Groupwise Majorization Descendent Algorithm presented in Yang and Zou (2015) and implemented in the R package `gglasso` available at https://cran.r-project.org/web/packages/gglasso/.

Both algorithms work groupwise by using the separability of model (3.5), i.e. update each group of variables iteratively until convergence. The main difference between the two algorithms is the updating of each group of variables: in `grpreg` it occurs through the solution of a single-group lasso, i.e. with a multivariate soft-thresholding operator, under the assumption of "orthonormal group", while in `gglasso` each group of variable is updated as the solution of a quadratic majorization problem. We stress that the "orthonormal group" property refers to the condition $\boldsymbol{X}(G_j)^t\boldsymbol{X}(G_j) = I$, not that groups $\boldsymbol{X}(G_j)$ and $\boldsymbol{X}(G_k)$ are orthogonal each other. When this condition is not satisfied the `grpreg` automatically orthonormalizes the design matrix, but this practice leads to a slight modification of the l_1/l_2-norm contained in the penalty, as pointed out in Huang et al. (2012) and Simon and Tibshirani (2012). This is not our case, because the design matrix defined in Eq. (3.3) satisfies the "orthonormal group" property and we can take complete advantage of the Group Descendent Algorithm in the `grpreg` package to solve problem (3.5) exactly.

Let us reorganize the design matrix $\boldsymbol{X}$ defined in Eq. (3.3) so that the group memberships are consecutive. From the group structure defined in Eq. (3.4) we have that the group membership vector I_g contains only one element for $g = 1, 2, \dots, d_1$, and K elements for $g = d_1 + j$ with $j = 1, \dots, d_2$. Hence, in the latter case the sub matrix $\boldsymbol{X}_{I_g}$, for $g = 1, \dots, d_1$, is a one-column matrix defined as

$$\boldsymbol{X}_{I_g} = \begin{bmatrix} \boldsymbol{\Psi}^{(g)} \\ \vdots \\ \boldsymbol{\Psi}^{(g)} \end{bmatrix} \in R^{nK\times 1},$$

where $\boldsymbol{\Psi}^{(g)}$ is the gth column of matrix $\boldsymbol{\Psi}$; while in the last case, for $g = d_1 + j$ with $j = 1, \dots, d_2$, the sub matrix $\boldsymbol{X}_{I_g}$ is a K-column matrix where each column is a shifted version of the jth column of matrix $\boldsymbol{\Phi}$ as in the following scheme

$$\boldsymbol{X}_{I_g} = \begin{bmatrix} \boldsymbol{\Phi}^{(j)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}^{(j)} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Phi}^{(j)} \end{bmatrix} \in R^{nK\times K}.$$

Finally, it is easy to verify the "orthonormal group" property, i.e. $\boldsymbol{X}_{I_g}^t\boldsymbol{X}_{I_g} = I$ for all $g = 1, \dots, d_1 + d_2$.

3.2. Connections with literature

As already stated in the introduction, multi-channel regression and equivalent problems have been investigated by diverse communities and a lot of literature is available on that.

Problem (3.2) is a particular case of the so-called Simultaneous Sparse Approximation (*SSA*) (Rakotomamonjy, 2011; Jenatton et al., 2011; Tropp et al., 2006; Tropp, 2006), defined as follows. Suppose that we have measured K signals $\{\mathbf{s_i}\}_{i=1}^K$, where each signal is of the form $\mathbf{s}_i = \boldsymbol{\Omega}\boldsymbol{c}_i + \boldsymbol{\epsilon}^{(i)}$, where $\{\mathbf{s}_i\} \in \mathbb{R}^{n\times 1}$, $\boldsymbol{\Omega} \in \mathbb{R}^{n\times m}$ is a matrix of unit-norm elementary functions, $\boldsymbol{c}_i \in \mathbb{R}^{m\times 1}$ is a weighting vector and $\boldsymbol{\epsilon}^{(i)}$ is a noise vector for each $i = 1, \dots, K$. The overall measurements can be written as

$$\mathbf{S} = \boldsymbol{\Omega}\boldsymbol{C} + \boldsymbol{\epsilon}, \tag{3.7}$$

where $S = [\mathbf{s}_1, \dots, \mathbf{s}_K]$ is a signal matrix, $C = [\mathbf{c}_1, \dots, \mathbf{c}_K]$ a coefficient matrix and $\boldsymbol{\epsilon}$ a noise matrix. For the SSA problem, the goal is then to recover the matrix $\mathbf{C}$ given the signal matrix $\mathbf{S}$ and the dictionary $\boldsymbol{\Omega}$ under the hypothesis that all signals $\mathbf{s}_i$ share the same sparsity profile. This latter hypothesis can be translated into the request that the coefficient matrix $\mathbf{C}$ has a minimal number of non-zero rows, i.e. solving the following problem

$$\min_{\mathbf{c}} \frac{1}{2}\|\mathbf{S} - \boldsymbol{\Omega}\boldsymbol{C}\|_F^2 \quad s.t. \quad \|\mathbf{C}\|_{row-0} \le T,$$

where

$$\|\mathbf{C}\|_{row-0} = \left|\{i \in [1, \dots, m] : c_{ij} \neq 0 \text{ for some j}\}\right|,$$

T is some parameter defined by the user to control the sparsity and $\|\cdot\|_F$ indicates the Frobenius norm.

This problem is not convex, but efficient greedy algorithms have been proposed to get an approximate solution. In particular, in Tropp et al. (2006), the author proposes the Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm, which selects, at each iteration, an element from the dictionary maximizing the sum of the absolute correlation between

the dictionary elements and the signal residual. As shown in Rakotomamonjy (2011), this greedy algorithm is actually one of the most efficient to solve the problem.

Another possibility to solve the minimization problem is to relax the constraint by replacing $\|\cdot\|_{row-0}$ with a more tractable row-sparsity measure. A large class of relaxed version of $\|\cdot\|_{row-0}$ consider the following constraint

$$J_{p,q}(\mathbf{C}) = \sum_i \left\|\mathbf{c}_{i,\cdot}\right\|_q^p \quad \text{with} \quad \left\|\mathbf{c}_{i,\cdot}\right\|_q = \left(\sum_j |c_{i,j}|^q\right)^{1/q},$$

where typically $p \le 1$ and $q \ge 1$.

Such kind of relaxed problems can be solved in different ways and a deep survey and comparison analysis can be found in Tropp (2006) and Rakotomamonjy (2011).

In particular, the case $p = 1$ and $q = 2$ can be efficiently solved by the Block Coordinate Descent (BCD) algorithm and has a strong connection with the group-lasso regression. Indeed, our problem (2.1) falls in this relaxed version, considering

$$\mathbf{S} = \boldsymbol{Y} = \left[\mathbf{y}^{(1)} \dots \mathbf{y}^{(K)}\right], \quad \boldsymbol{\Omega} = [\Psi, \Phi] \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} \boldsymbol{\alpha}^{(1)} & & \boldsymbol{\alpha}^{(K)} \\ \boldsymbol{\beta}^{(1)} & \cdots & \boldsymbol{\beta}^{(K)} \end{bmatrix}.$$

Moreover, there is also a connection with structured variable selection and structural penalties in the vector formulation of Eq. (3.3). In fact, the penalty we used in Eq. (3.5) is a particular case of Eq. (1), Section 2, described in Jenatton et al. (2011), and this permits to use all the optimization algorithms based on the proximal methods.

Finally, it is important to stress a fundamental difference with the proposed methodology, i.e. all reviewed methods do not take properly into account the constraint of a common low-component $\left(\boldsymbol{\alpha}^{(1)} = \dots = \boldsymbol{\alpha}^{(K)}\right)$, hence any multichannel reconstruction returns different low-resonance components for different channels, loosing in terms of estimation error as it will be shown in the numerical section.

4. Theoretical properties

The following results are obtained adapting results of Chapter 8 in Bühlmann and van de Geer (2011).

Let estimator $\left[\hat{\boldsymbol{c}}^t, \left(\hat{\boldsymbol{u}}^{(1)}\right)^t, \dots, \left(\hat{\boldsymbol{u}}^{(K)}\right)^t\right]^t$ be given by Eq. (3.6); in order to derive an oracle inequality for its error, we introduce the following notations and assumptions.

Notations. for any subset of indices $S \subseteq \mathcal{P} = \{1, \dots, d_1\} \cup \{d_1 + 1, \dots, d_1 + d_2\}$, we denote $S^{\boldsymbol{\alpha}} = S \cap \{1, \dots, d_1\}$ and $S^{\boldsymbol{\beta}} = \{j : 1 \le j \le d_2 \text{ and } d_1 + j \in S\}$, moreover subset S^c is its complement in $\mathcal{P}$ and $|S|$ is its cardinality, so that $|\mathcal{P}| = d_1 + d_2$. Let us abuse of notations writing $d_1 + S^{\boldsymbol{\beta}} = \{d_1 + j : j \in S^{\boldsymbol{\beta}}\}$. If $S = S^{\boldsymbol{\alpha}} \cup \{d_1 + S^{\boldsymbol{\beta}}\} \subseteq \mathcal{P}$ and $\theta \in \mathbb{R}^{d_1 + Kd_2 \times 1}$, then $\boldsymbol{\theta}(S) = \left[\boldsymbol{\alpha}\left(S^{\boldsymbol{\alpha}}\right) \ \boldsymbol{\beta}\left(S^{\boldsymbol{\beta}}\right)\right]$ denotes reduction of vector $\boldsymbol{\theta}$ to the subset of group index S, as $\boldsymbol{\alpha}\left(S^{\boldsymbol{\alpha}}\right) \in \mathbb{R}^{|S^{\boldsymbol{\alpha}}| \times 1}$ denotes reduction of vector $\boldsymbol{\alpha}$ to the subset of variable index $S^{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}\left(S^{\boldsymbol{\beta}}\right) = \left[\left(\boldsymbol{\beta}^{(1)}\left(S^{\boldsymbol{\beta}}\right)\right)^t, \ \dots, \ \left(\boldsymbol{\beta}^{(K)}\left(S^{\boldsymbol{\beta}}\right)\right)^t\right]^t$ is such that $\boldsymbol{\beta}^{(k)}\left(S^{\boldsymbol{\beta}}\right) \in \mathbb{R}^{\left|S^{\boldsymbol{\beta}}\right| \times 1}$ denotes reduction of vector $\boldsymbol{\beta}^{(k)}$ to the subset of variables index $S^{\boldsymbol{\beta}}$ for all $k = 1, \dots, K$.

Assumptions.

(A1) The linear model in Eq. (3.3) holds exactly with some *true parameter value* $\boldsymbol{\theta}_0 = \left[\boldsymbol{\alpha}_0^t, \left(\boldsymbol{\beta}_0^{(1)}\right)^t, \dots, \left(\boldsymbol{\beta}_0^{(K)}\right)^t\right]^t$, $S_0 = S_0^{\boldsymbol{\alpha}} \cup \{d_1 + S_0^{\boldsymbol{\beta}}\}$ being the true active set of groups.

(A2) The **compatibility condition** holds for the group index set $S_0 = S_0^{\boldsymbol{\alpha}} \cup \{d_1 + S_0^{\boldsymbol{\beta}}\}$ with constant $\phi(S_0) > 0$, if for all $\boldsymbol{\theta} \in \mathbb{R}^{d_1 + Kd_2 \times 1}$ such that $\|\boldsymbol{\theta}(S_0^c)\|_{2,1} \le 3\|\boldsymbol{\theta}(S_0)\|_{2,1}$, it holds that

$$G^\star \ \|\boldsymbol{\theta}(S_0)\|_{2,1}^2 \ \le \ \|\boldsymbol{X}\boldsymbol{\theta}\|_2^2 \ G^\star \ |S_0| \ / \ nK \ \phi(S_0)^2. \tag{4.8}$$

Note that Assumption **(A1)** means that the true signals $\boldsymbol{c} + \boldsymbol{u}^{(k)}$, for $k = 1, \dots, K$ are exact linear combination of the columns of matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ which simplifies the proof, however this assumption can be relaxed and the following theorem is stated for the best linear approximation of the unknown signals into the span of columns of matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$. Moreover, note that in Assumption **(A2)** $G^\star \ |S_0|$ is the average group size times the active number of groups and plays the role of the number of active variables into the compatibility condition. As often observed the compatibility constant $\phi(S_0)$ is linked to a condition on the smallest eigenvalue of the matrix $\boldsymbol{X}^t\boldsymbol{X}/n$ which turns out to be linked to the product $\boldsymbol{\Phi}^t\boldsymbol{\Psi}$ which in signal processing is the coherence between the two filters.

We can now prove the following main result:

Theorem 1. *Let $\hat{\boldsymbol{\theta}}$ be one solution of Eq. (3.5) and let assumptions* **(A1)**–**(A2)** *hold; then, for any $x > 0$ and any $\lambda \ge 2\lambda_0$, with probability at least $1 - 2e^{-x^2/2} - e^{-x}$, it holds that*

$$\frac{1}{nK}\left\|X(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\|_2^2+\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right\|_{2,1}\leq 4\,\lambda^2\,G^\star\,|S_0|/\phi(S_0)^2, \tag{4.9}$$

where $\lambda_0=max\left\{\lambda_0^{\alpha},\lambda_0^{\beta}/\sqrt{K}\right\}$, *with*

$$\lambda_0^{\alpha}=\frac{2\,\sigma}{\sqrt{nK}}\sqrt{x^2+2\,\log(d_1)},$$

and

$$\lambda_0^{\beta}=\frac{2\,\sigma}{\sqrt{nK}}\left(1+\sqrt{(4x+4\,\log(d_2))/K}+(4x+4\,\log(d_2))/K\right).$$

Proof is given in the Appendix.

The theorem proves the so called *oracle inequality* for the group lasso estimator and it directly gives a bound on the prediction error, indeed if λ is chosen as claimed in the theorem, it follows with high probability

$$\frac{1}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\|_2^2\sim\frac{\log(d)\,\sigma^2\,G^\star}{nK}\,|S_0|,$$

with $\log(d)=\max\left\{\log(d_1),\log^2(d_2)/K^2\right\}$ so that the price for not knowing the true active index groups S_0 is of the order $\log(d)$.

5. Simulations and real examples

In order to show the performance of the proposed methodology, a number of experiments were run on synthetic datasets and on a real EEG dataset, the first being an ideal modelization of the second.

For all results reported in this section, we used the `grpreg` package, that implements efficient algorithms for fitting the regularization path of linear or logistic regression models with different grouped penalties. It includes group selection methods such as group LASSO (referred to as `grlasso` in the following), group MCP, and group SCAD as well as bi-level selection methods such as the group exponential LASSO, the composite MCP, and the group bridge. The smoothing parameter λ can be estimated by BIC, AIC, GCV and CV.

We used the group LASSO to solve the penalized regression and the V-fold CV criterion to choose the smoothing parameter λ.

All the codes that have been used to produce the following results are freely available at http://www.iac.cnr.it/~danielad/software.html.

5.1. Synthetic data

In this section we present results obtained using synthetic data representing different sparse scenarios and different noise levels. We generated data according to model (3.2)

$$\mathbf{y}^{(k)}=\mathbf{c}+\mathbf{u}^{(k)}+\boldsymbol{\varepsilon}^{(k)}=\boldsymbol{\Psi}\boldsymbol{\alpha}+\boldsymbol{\Phi}\boldsymbol{\beta}^{(k)}+\boldsymbol{\varepsilon}^{(k)}\quad k=1,\dots,K,$$

using three channels ($K=3$) and $n=256$ observations in each channel. Matrix Ψ was generated using the following choice $p_{low}=1$, $q_{low}=2$, $s_{low}=1$, $J_{low}=4$ and matrix Φ was generated using $p_{high}=8$, $q_{high}=9$, $s_{high}=3$, $J_{high}=10$. These matrices represent RADWT with Q-factor almost 1 and 5 respectively, the first frame resembles the dyadic wavelet transform and its mother wavelet has almost one pulse, while the second frame has a mother wavelet with almost 5 pulses, as very well explained in Fig. 1 of Selesnick (2011). We considered three scenarios with different sparsity level:

Scenario 1: low sparsity, corresponding to $|S_{\boldsymbol{\alpha}}|=24$ and $\left|S_{\beta}\right|=24$;

Scenario 2: medium sparsity, corresponding to $|S_{\boldsymbol{\alpha}}|=12$ and $\left|S_{\beta}\right|=12$;

Scenario 3: high sparsity, corresponding to $|S_{\boldsymbol{\alpha}}|=6$ and $\left|S_{\beta}\right|=6$;

and for each scenario we used three signal to noise ratios (SNR): 1.5, 3, 6, defined as

$$\mathrm{SNR}=\frac{\frac{1}{K}\sum_{i=1}^{K}\mathrm{Var}(\boldsymbol{\Psi}\boldsymbol{\alpha}+\boldsymbol{\Phi}\boldsymbol{\beta}^{(k)})}{\sigma^2_{SNR}}.$$

Data were generated in each channel, using $\alpha_{0_j}=1, j\in S_0^{\alpha}$, and $\beta_j^{(k)}\sim Uniform(0,M)$, with $M=\|\mathbf{c}\|_{\infty}/\left\|\boldsymbol{\Phi}(S_0^{\beta})\right\|_{\infty}$, and $\boldsymbol{\varepsilon}^{(k)}\sim N(0,\sigma^2_{\mathrm{SNR}}\mathbf{I})$.

In all test cases the proposed procedure, indicated hereafter as `multi-c`, has been compared with the `single-c` procedure, i.e. the procedure where in each channel, the estimator $\hat{\mathbf{f}}^{(k)}=\boldsymbol{\Psi}\hat{\boldsymbol{\alpha}}+\boldsymbol{\Phi}\hat{\boldsymbol{\beta}}^{(k)}$ is obtained independently from

the other channels by the following minimization:

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}}^{(k)} \end{pmatrix} = \underset{\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \in \mathbb{R}^{d_1+d_2\times 1}}{argmin} \left\{ \frac{1}{n} \left\| \mathbf{y}^{(k)} - [\,\boldsymbol{\Psi}\ \boldsymbol{\Phi}] \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \right\|_2^2 + \lambda \left(\sum_{i=1}^{d1} |\alpha_i| + \sum_{i=1}^{d2} |\beta_i| \right) \right\},$$

$k = 1, \ldots, K$.

Performance was evaluated by computing the following indicators:

- Root Mean Square Error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}^{(k)}(t_i) - f^{(k)}(t_i) \right)^2},\ k = 1, \ldots, K;$$

  with $\mathbf{f}^{(k)} = \mathbf{c} + \mathbf{u}^{(k)}$ and $\hat{\mathbf{f}}^{(k)} = \boldsymbol{\Psi}\hat{\boldsymbol{\alpha}} + \boldsymbol{\Phi}\hat{\boldsymbol{\beta}}^{(k)}$ its estimate;
- Root Mean Square Error for the low resonance component (RMSE_{low}) defined as

$$\text{RMSE}_{low} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{c}(t_i) - c(t_i) \right)^2};$$

- Root Mean Square Error for the high resonance component (RMSE_{high}) defined as

$$\text{RMSE}_{high} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{u}^{(k)}(t_i) - u^{(k)}(t_i) \right)^2},\ k = 1, \ldots, K;$$

  RMSE_{low} and RMSE_{high} aim at evaluating a component wise accuracy.

With the aim of exploring the variable selection properties of the considered procedures, we also computed the following indicators:

- True positives for the low resonance component (TP_{low}) defined as

$$\text{TP}_{low} := \left| \hat{S}_0^{\boldsymbol{\alpha}} \right|,\ \hat{S}_0^{\boldsymbol{\alpha}} = \{ j : \hat{\alpha}_j \neq 0 \text{ and } \alpha_{0_j} \neq 0 \}.$$

- False negatives for the low resonance component (FN_{low}) defined as

$$\text{FN}_{low} := \left| \hat{S}_0^{\boldsymbol{\alpha},n} \right|,\ \hat{S}_0^{\boldsymbol{\alpha},n} := \left\{ j : \hat{\alpha}_j = 0 \text{ and } \alpha_{0_j} \neq 0 \right\}.$$

  For the `single-c` procedure TP_{low} and FN_{low} will be dependent on the channels, while for the `multi-c` procedure they will not.
- True positives for the high resonance component (TP_{high}) defined as

$$\text{TP}_{high} := \left| \hat{S}_0^{\boldsymbol{\beta}} \right| = \left| \hat{S}_0^{(k),\boldsymbol{\beta}} \right|,\ \hat{S}_0^{(k),\boldsymbol{\beta}} = \left\{ j : \hat{\beta}_j^{(k)} \neq 0 \text{ and } \beta_{0_j}^{(k)} \neq 0 \right\},$$

  $\forall k = 1, \ldots, K$.
- False negatives for the high resonance component (FN_{high}) defined as

$$\text{FN}_{high} := \left| \hat{S}_0^{\boldsymbol{\beta},n} \right| = \left| \hat{S}_0^{(k),\boldsymbol{\beta},n} \right|,\ \hat{S}_0^{(k),\boldsymbol{\beta},n} := \left\{ j : \hat{\beta}_j^{(k)} = 0 \text{ and } \beta_{0_j}^{(k)} \neq 0 \right\},$$

  $\forall k = 1, \ldots, K$.
  For the `multi-c` procedure the sets $\hat{S}_0^{(k),\boldsymbol{\beta}}$ and $\hat{S}_0^{(k),\boldsymbol{\beta},n}$ are all equal, while for the `single-c` procedure the sets depend on the channels.

Note that in general the following relationships hold: $\text{TP} = \text{NS} - \text{FP}$ and $\text{TP} + \text{FN} = \text{NS} - \text{FP} + \text{FN} = p_{\text{active}}$, where NS indicates the number of selected variables, FP indicates the number of false positives and p_{active} is the true number of active variables.

To be robust with respect to the particular realization in generating synthetic data (and corresponding noise), each experiment was run several times, in particular we set $Nrun = 100$ and we evaluated the averaged indicators.

Table 1 shows the results for RMSE, RMSE_{low} and RMSE_{high} for *Scenario 1* and SNR = 1.5, 3 and 6 respectively, for all the 3 channels indicated as ch1, ch2, ch3; standard deviation is displayed in parentheses. Table 2 shows the performance indicators TP and FN for the low resonance component and high resonance component.

Tables 3–5 contain the results for RMSE, RMSE_{low} and RMSE_{high} for *Scenario 2* and *Scenario 3*, respectively; analogously Tables 4–6 illustrate the performance indicators TP and FN for the same scenarios.

Table 1
Average values (standard deviation between parentheses) of RMSE, RMSE_{low} and RMSE_{high} based on 100 simulations with different noise realizations. Experiment carried out on Scenario 1 with SNR = 1.5, 3 and 6.

	RMSE		RMSE_{low}		RMSE_{high}	
	single-c	multi-c	single-c	multi-c	single-c	multi-c
SNR = 1.5						
ch1	0.2897 (0.0348)	0.2216 (0.0191)	0.2206 (0.0129)	0.1728 (0.0127)	0.2178 (0.0232)	0.2284 (0.0154)
ch2	0.3004 (0.0355)	0.2226 (0.0194)	0.2249 (0.0123)	0.1728 (0.0127)	0.2314 (0.0251)	0.2457 (0.0197)
ch3	0.2968 (0.0379)	0.2130 (0.0187)	0.2244 (0.0145)	0.1728 (0.0127)	0.2236 (0.0227)	0.2337 (0.0194)
SNR = 3						
ch1	0.2242 (0.0290)	0.1608 (0.0118)	0.1882 (0.0170)	0.1446 (0.0106)	0.1715 (0.0156)	0.1852 (0.0129)
ch2	0.2277 (0.0297)	0.1628 (0.0113)	0.1926 (0.0161)	0.1446 (0.0106)	0.1842 (0.0175)	0.2024 (0.0144)
ch3	0.2322 (0.0309)	0.1560 (0.0111)	0.1924 (0.0165)	0.1446 (0.0106)	0.1815 (0.0175)	0.1913 (0.0148)
SNR = 6						
ch1	0.1611 (0.0234)	0.1153 (0.0092)	0.1457 (0.0149)	0.1199 (0.0096)	0.1329 (0.0140)	0.1501 (0.0120)
ch2	0.1673 (0.0215)	0.1169 (0.0101)	0.1554 (0.0129)	0.1199 (0.0096)	0.1479 (0.0114)	0.1615 (0.0138)
ch3	0.1613 (0.0233)	0.1117 (0.0072)	0.1468 (0.0156)	0.1199 (0.0096)	0.1357 (0.0125)	0.1514 (0.0121)

Table 2
Fraction of correctly retrieved variables $\left(\text{TP}_{low}/\left|S_0^{\alpha}\right|\right)$ and incorrectly retrieved variables $\left(\text{FN}_{low}/\left|S_0^{\alpha}\right|\right)$ for the estimated low resonance signal component. Fraction of correctly retrieved variables $\left(\text{TP}_{high}/\left|S_0^{\beta}\right|\right)$ and incorrectly retrieved variables $\left(\text{FN}_{high}/\left|S_0^{\beta}\right|\right)$ for the estimated high resonance signal component. Values are based on 100 simulations with different noise realizations for Scenario 1 and SNR = 1.5, 3 and 6.

	$(\text{TP}_{low})/\text{p}_{low}$		$\text{FN}_{low}/\text{p}_{low}$		$(\text{TP}_{high})/\text{p}_{high}$		$\text{FN}_{high}/\text{p}_{high}$	
	single-c	multi-c	single-c	multi-c	single-c	multi-c	single-c	multi-c
SNR = 1.5								
ch1	0.4029	0.8696	0.5971	0.1304	0.4500	0.6508	0.5500	0.3492
ch2	0.3450	0.8696	0.6550	0.1304	0.4129	0.6508	0.5871	0.3492
ch3	0.3629	0.8696	0.6371	0.1304	0.4154	0.6508	0.5846	0.3492
SNR = 3								
ch1	0.6800	0.9546	0.3200	0.0454	0.5937	0.8613	0.4063	0.1387
ch2	0.6421	0.9546	0.3579	0.0454	0.5767	0.8613	0.4233	0.1387
ch3	0.6662	0.9546	0.3338	0.0454	0.5742	0.8613	0.4258	0.1387
SNR = 6								
ch1	0.8808	0.9912	0.1193	0.0088	0.7137	0.9450	0.2863	0.0550
ch2	0.8487	0.9912	0.1513	0.0088	0.6833	0.9450	0.3167	0.0550
ch3	0.8775	0.9912	0.1225	0.0088	0.7333	0.9450	0.2667	0.0550

Table 3
Average values (standard deviations between parentheses) of RMSE, RMSE_{low} and RMSE_{high} based on 100 simulations with different noise realizations. Experiment carried out on Scenario 2 with SNR = 1.5, 3 and 6.

	RMSE		RMSE_{low}		RMSE_{high}	
	single-c	multi-c	single-c	multi-c	single-c	multi-c
SNR = 1.5						
ch1	0.2151 (0.0187)	0.1662 (0.0143)	0.1664 (0.0105)	0.1122 (0.0105)	0.1619 (0.0187)	0.1486 (0.0163)
ch2	0.2249 (0.0225)	0.1783 (0.0180)	0.1646 (0.0116)	0.1122 (0.0105)	0.1786 (0.0206)	0.1660 (0.0188)
ch3	0.2175 (0.0197)	0.1627 (0.0152)	0.1644 (0.0108)	0.1122 (0.0105)	0.1598 (0.0190)	0.1447 (0.0169)
SNR = 3						
ch1	0.1692 (0.0192)	0.1209 (0.0114)	0.1396 (0.0142)	0.0826 (0.0078)	0.1239 (0.0154)	0.1099 (0.0130)
ch2	0.1748 (0.0184)	0.1302 (0.0130)	0.1421 (0.0125)	0.0826 (0.0078)	0.1370 (0.0149)	0.1239 (0.0150)
ch3	0.1679 (0.0163)	0.1154 (0.0099)	0.1378 (0.0013)	0.0826 (0.0078)	0.1202 (0.0128)	0.1038 (0.0120)
SNR = 6						
ch1	0.1237 (0.0143)	0.0881 (0.0082)	0.1059 (0.0126)	0.0606 (0.0052)	0.0944 (0.0098)	0.0833 (0.0095)
ch2	0.1254 (0.0151)	0.0941 (0.0089)	0.1078 (0.0122)	0.0606 (0.0052)	0.1047 (0.0102)	0.0924 (0.0090)
ch3	0.1182 (0.0142)	0.0825 (0.0074)	0.1004 (0.0138)	0.0606 (0.0052)	0.0891 (0.0093)	0.0761 (0.0087)

`Multi-c` procedure always outperforms `single-c` procedure in terms of RMSE with a consistently lower standard deviation. This is not surprising because `multi-c` procedure exploits the joint information among the channels leading to a more precise (mean) and robust (std) estimation error. We also note that, in almost all scenarios and SNRs, `multi-c` outperforms `single-c` reconstructing the two components except for *Scenario 1* where the low and high resonance components share pieces of signals (see Fig. 1). This is again not surprising, since the two procedures aim to regress $\mathbf{f} = \mathbf{c} + \mathbf{u}$

Table 4
Fraction of correctly retrieved variables $\left(\mathrm{TP}_{low}/\left|S_0^{\alpha}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{low}/\left|S_0^{\alpha}\right|\right)$ for the estimated low resonance signal component. Fraction of correctly retrieved variables $\left(\mathrm{TP}_{high}/\left|S_0^{\beta}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{high}/\left|S_0^{\beta}\right|\right)$ for the estimated high resonance signal component. Values are based on 100 simulations with different noise realizations for Scenario 2 and SNR = 1.5, 3 and 6.

	$(\mathrm{TP}_{low})/\mathrm{p}_{low}$		$\mathrm{FN}_{low}/\mathrm{p}_{low}$		$(\mathrm{TP}_{high})/\mathrm{p}_{high}$		$\mathrm{FN}_{high}/\mathrm{p}_{high}$	
	single-c	multi-c	single-c	multi-c	single-c	multi-c	single-c	multi-c
SNR = 1.5								
ch1	0.4708	0.9508	0.5292	0.0492	0.5708	0.7675	0.4292	0.2325
ch2	0.5017	0.9508	0.4983	0.0492	0.6142	0.7675	0.3858	0.2325
ch3	0.4908	0.9508	0.5092	0.0492	0.5758	0.7675	0.4242	0.2325
SNR = 3								
ch1	0.7867	0.9983	0.2133	0.0017	0.6208	0.8150	0.3792	0.1850
ch2	0.7650	0.9983	0.2350	0.0017	0.6675	0.8150	0.3325	0.1850
ch3	0.8242	0.9983	0.1758	0.0017	0.6608	0.8150	0.3392	0.1850
SNR = 6								
ch1	0.9800	1	0.0200	0	0.6683	0.8508	0.3317	0.1492
ch2	0.9500	1	0.0500	0	0.6808	0.8508	0.3192	0.1492
ch3	0.9725	1	0.0275	0	0.7017	0.8508	0.2983	0.1492

Table 5
Average values (standard deviations between parentheses) of RMSE, RMSE_{low} and RMSE_{high} based on 100 simulations with different noise realizations. Experiment carried out on Scenario 3 with SNR = 1.5, 3 and 6.

	RMSE		RMSE_{low}		RMSE_{high}	
	single-c	multi-c	single-c	multi-c	single-c	multi-c
SNR = 1.5						
ch1	0.0437 (0.0104)	0.0294 (0.0036)	0.0337 (0.0097)	0.0172 (0.0023)	0.0285 (0.0060)	0.0258 (0.0039)
ch2	0.0426 (0.0977)	0.0260 (0.0030)	0.0347 (0.0093)	0.0172 (0.0023)	0.0259 (0.0044)	0.0218 (0.0034)
ch3	0.0459 (0.0103)	0.0329 (0.0045)	0.0341 (0.0088)	0.0172 (0.0023)	0.0322 (0.0066)	0.0302 (0.0047)
SNR = 3						
ch1	0.0298 (0.0053)	0.0202 (0.0024)	0.0228 (0.0045)	0.0121 (0.0017)	0.0205 (0.0039)	0.0180 (0.0029)
ch2	0.0283 (0.0048)	0.0180 (0.0022)	0.0225 (0.0047)	0.0121 (0.0017)	0.0185 (0.0031)	0.0151 (0.0028)
ch3	0.0307 (0.0052)	0.0223 (0.0027)	0.0226 (0.0043)	0.0121 (0.0017)	0.0223 (0.0041)	0.0207 (0.0031)
SNR = 6						
ch1	0.0201 (0.0033)	0.0141 (0.0020)	0.0151 (0.0029)	0.0084 (0.0013)	0.0140 (0.0027)	0.0126 (0.0023)
ch2	0.0204 (0.0029)	0.0130 (0.0015)	0.0156 (0.0026)	0.0084 (0.0013)	0.0138 (0.0022)	0.0113 (0.0017)
ch3	0.0217 (0.0032)	0.0167 (0.0020)	0.0155 (0.0028)	0.0084 (0.0013)	0.0160 (0.0027)	0.0151 (0.0022)

Table 6
Fraction of correctly retrieved variables $\left(\mathrm{TP}_{low}/\left|S_0^{\alpha}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{low}/\left|S_0^{\alpha}\right|\right)$ for the estimated low resonance signal component. Fraction of correctly retrieved variables $\left(\mathrm{TP}_{high}/\left|S_0^{\beta}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{high}/\left|S_0^{\beta}\right|\right)$ for the estimated high resonance signal component. Values are based on 100 simulations with different noise realizations for Scenario 3 and SNR = 1.5, 3 and 6.

	$(\mathrm{TP}_{low})/\mathrm{p}_{low}$		$\mathrm{FN}_{low}/\mathrm{p}_{low}$		$(\mathrm{TP}_{high})/\mathrm{p}_{high}$		$\mathrm{FN}_{high}/\mathrm{p}_{high}$	
	single-c	multi-c	single-c	multi-c	single-c	multi-c	single-c	multi-c
SNR = 1.5								
ch1	0.9767	1	0.0233	0	0.6500	0.9833	0.3500	0.0167
ch2	0.9850	1	0.0150	0	0.4300	0.9833	0.5700	0.0167
ch3	0.9800	1	0.0200	0	0.8400	0.9833	0.1600	0.0167
SNR = 3								
ch1	1	1	0	0	0.7633	1	0.2367	0
ch2	1	1	0	0	0.5867	1	0.4133	0
ch3	1	1	0	0	0.9933	1	0.0067	0
SNR = 6								
ch1	1	1	0	0	0.8333	1	0.1667	0
ch2	1	1	0	0	0.6633	1	0.3367	0
ch3	1	1	0	0	1	1	0	0

and not the single components (as in Morphological Component Analysis). Hence, when the two components low resonance ($\mathbf{c}$) and high resonance ($\mathbf{u}$) are confounding `single-c` can have some advantage with respect to `multi-c`, remaining the latter more effective in reconstructing the whole signal f. The advantage of `multi-c` with respect to `single-c` is more

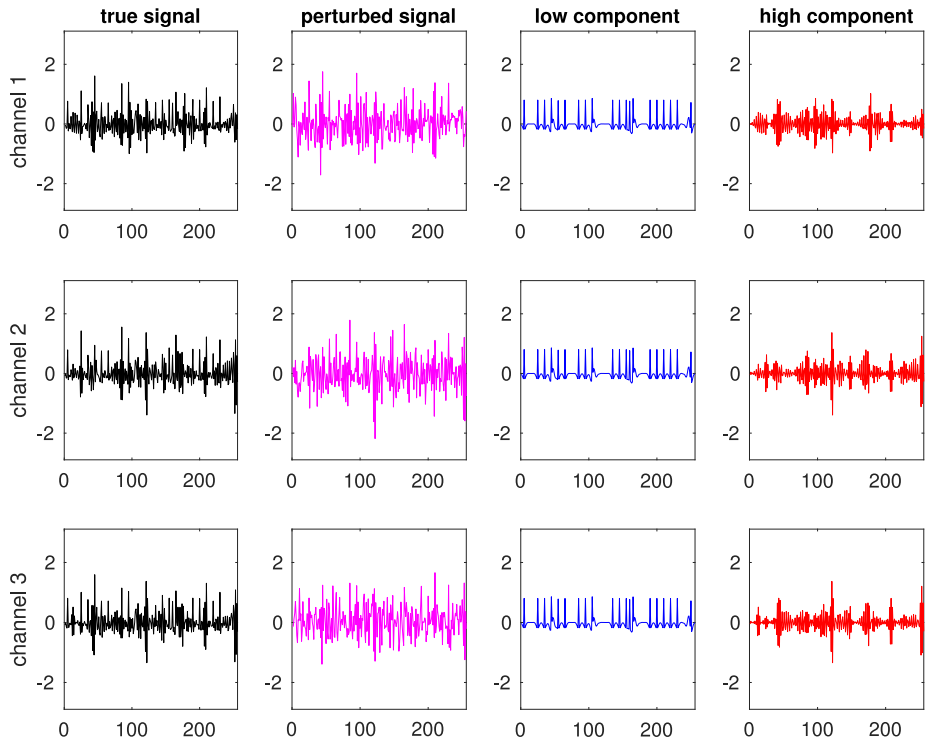

Fig. 1. True signal (first column), perturbed signal for a particular noise realization (second column), true low component (third column), true high component (fourth column) for each channel for *Scenario 1* and SNR = 1.5.

evident looking at the selecting capabilities of the procedure, with a good control of both false positives and false negatives. Of course performance improves when both SNR and sparsity increase.

For the sake of brevity, we only show the plots of the shape of the unknown signals and the goodness of reconstructions for the two extreme cases, i.e. *Scenario 1* with SNR = 1.5 and *Scenario 3* with SNR = 6, see Figs. 1–4.

5.2. Comparisons and further studies

For completeness in this section we compare our method with two competitors, namely BCD and SOMP. These techniques handle multi-task learning problems and their effectiveness has been shown in diverse survey papers, see Rakotomamonjy (2011) and Tropp (2006).

The routines mexSOMP and mexL1L2BCD contained in the Matlab SPAMS package (http://spams-devel.gforge.inria.fr/) were used to produce the presented results. The synthetic data were generated using the same numerical setting of the previous experiment, but we relaxed Hypothesis (H2), setting $\boldsymbol{\beta}^{(3)} = 0$. This allowed the data to be different from the correct RADWT model to test the robustness of the method.

Tables 7, 9 and 11 show the results of RMSE, RMSE_{low} and RMSE_{high} considering SNR = 1.5, 3 and 6, for *Scenario 1*, *Scenario 2* and *Scenario 3* respectively. The `multi-c` procedure gets a quite significant improvement in terms of RMSE, especially for severe noise condition, mostly due to the good estimation of the low-resonance component. This is not surprising since `multi-c` takes into proper account the equality constraint on the low-component ($\boldsymbol{\alpha}^{(1)} = \cdots = \boldsymbol{\alpha}^{(K)}$). It is also very interesting to note that the `multi-c` procedure outperforms BCD and SOMP in the retrieval of the high-component of the third channel (which is zero by construction), in fact it gives very low coefficients $\hat{\beta}^{(3)}$ as properly expected.

Finally, consistently with the previous analyses, Tables 8, 10 and 12 show the performance indicators TP and FN for the low resonance component and high resonance components for *Scenario 1*, *Scenario 2* and *Scenario 3* respectively. Note that indicators TP and FN are reported only for the first two channels, while for the third channel (which is zero) only the number of falsely non zero retrieved coefficients is reported. It is obvious that, this last index is minimum for the `single-c` procedure which works on the third channel independently from the other two, however `multi-c` is comparable with SOMP and does a good job with respect to BCD, especially for more severe level of noise.

5.3. Real data

To illustrate our procedure in a real case, we considered the problem of separating the transient and the oscillatory component in human sleep EEG data. This problem is actually a very hot topic in neuroscience, because several studies have

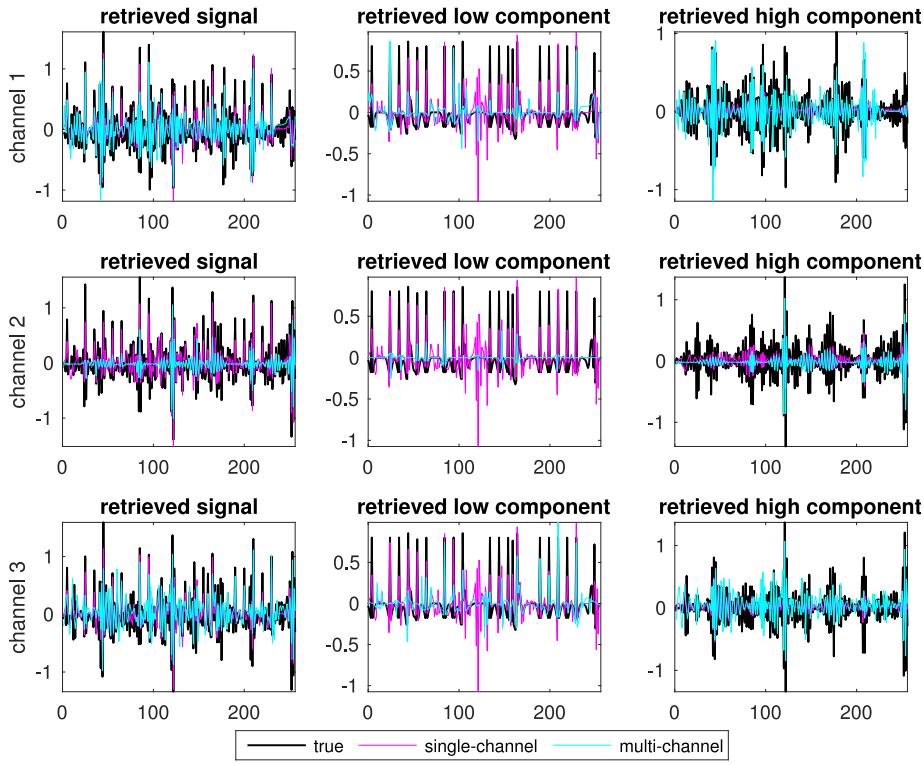

Fig. 2. Retrieved signal (first column), low component of the retrieved signal (second column), high component of the retrieved signal (third column) for each channel for a particular noise realization, for *Scenario 1* and SNR = 1.5. Black line refers to the true signal, cyan line refers to single channel retrieval, magenta line refers to multi channel retrieval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

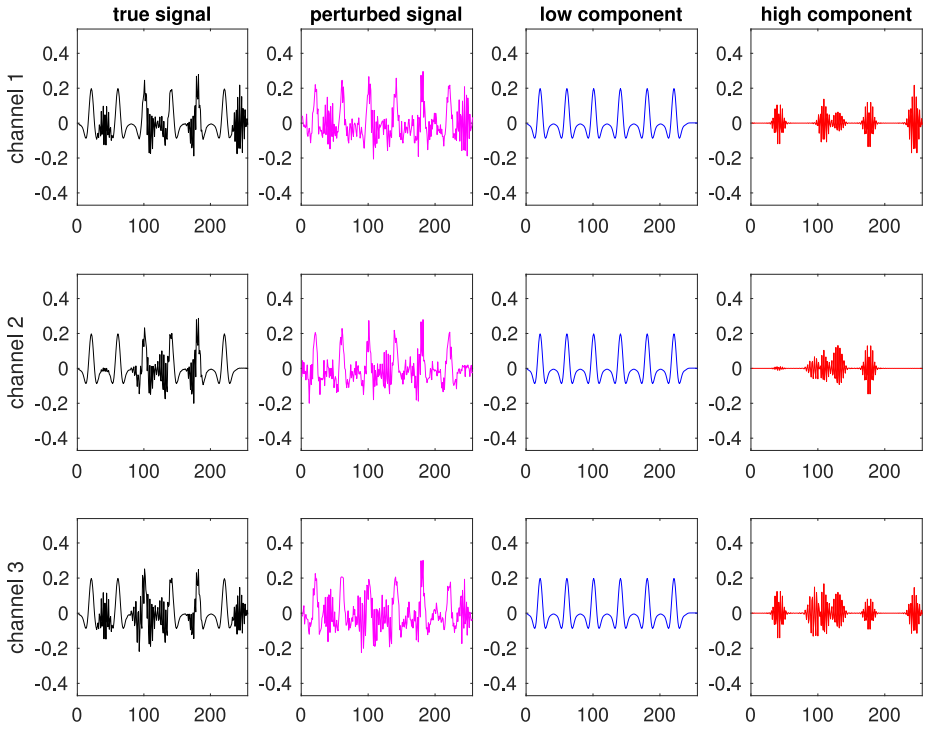

Fig. 3. True signal (first column), perturbed signal for a particular noise realization (second column), true low component (third column), true high component (fourth column) for each channel for *Scenario 3* and SNR = 6.

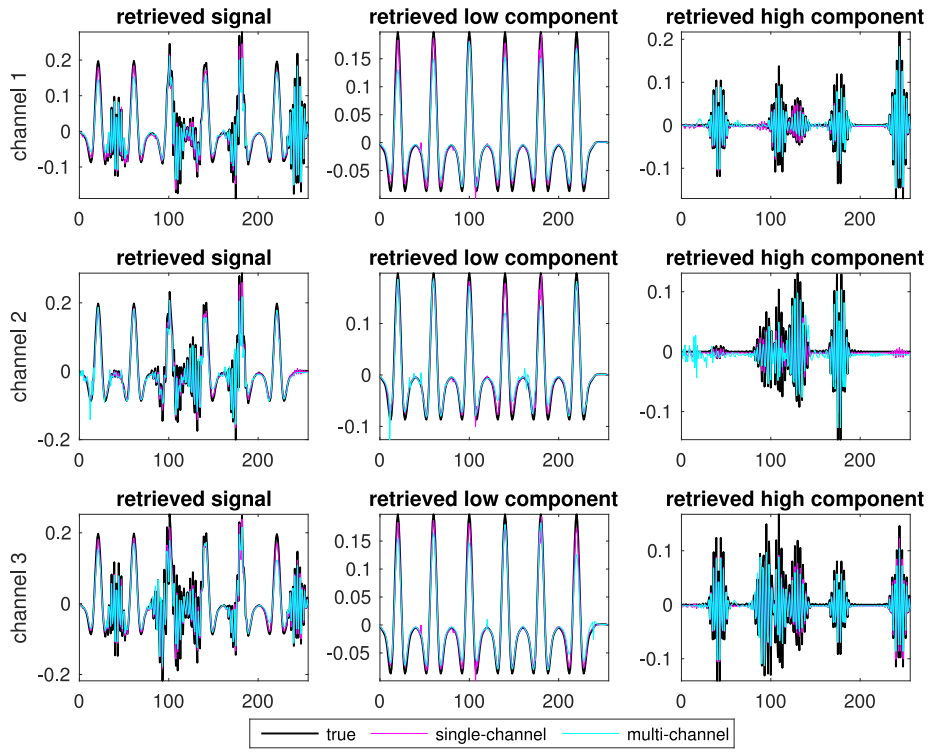

Fig. 4. Retrieved signal (first column), low component of the retrieved signal (second column), high component of the retrieved signal (third column) for each channel for a particular noise realization, for *Scenario 3* and SNR = 6. Black line refers to the true signal, cyan line refers to single channel retrieval, magenta line refers to multi channel retrieval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Average values (standard deviations between parentheses) of RMSE, RMSE_{low} and RMSE_{high} based on 100 simulations with different noise realizations. Experiment carried out on Scenario 1 with SNR = 1.5, 3 and 6.

	RMSE				RMSE_{low}				RMSE_{high}			
	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP
SNR = 1.5												
ch1	0.2468 (0.0296)	0.1961 (0.0200)	0.2035 (0.0097)	0.2334 (0.0175)	0.2003 (0.0159)	0.1172 (0.0092)	0.1640 (0.0109)	0.2076 (0.0298)	0.1790 (0.0185)	0.1838 (0.0204)	0.1782 (0.0115)	0.2147 (0.0281)
ch2	0.2624 (0.0348)	0.1953 (0.0169)	0.1972 (0.0101)	0.2430 (0.0157)	0.2099 (0.0172)	0.1172 (0.0092)	0.1627 (0.0104)	0.2190 (0.0283)	0.1940 (0.0202)	0.1864 (0.0167)	0.1634 (0.0131)	0.2306 (0.0298)
ch3	0.2280 (0.0258)	0.1218 (0.0094)	0.1763 (0.0102)	0.2268 (0.0171)	0.2309 (0.0220)	0.1172 (0.0092)	0.1574 (0.0112)	0.1959 (0.0223)	0.0237 (0.0283)	0.0455 (0.0135)	0.1197 (0.0109)	0.1941 (0.0226)
SNR = 3												
ch1	0.1893 (0.0239)	0.1374 (0.0129)	0.1575 (0.0089)	0.1577 (0.0109)	0.1654 (0.0172)	0.0870 (0.0062)	0.1258 (0.0085)	0.1330 (0.0179)	0.1430 (0.0125)	0.1330 (0.0133)	0.1443 (0.0108)	0.1402 (0.0181)
ch2	0.1995 (0.0282)	0.1420 (0.0129)	0.1435 (0.0088)	0.1598 (0.0123)	0.1727 (0.0163)	0.0870 (0.0062)	0.1230 (0.0099)	0.1388 (0.0206)	0.1561 (0.0152)	0.1419 (0.0138)	0.1230 (0.0100)	0.1474 (0.0198)
ch3	0.1874 (0.0381)	0.0925 (0.0065)	0.1188 (0.0071)	0.1519 (0.0094)	0.1926 (0.0360)	0.0870 (0.0062)	0.1203 (0.0085)	0.1260 (0.0170)	0.0319 (0.0237)	0.0462 (0.0125)	0.0709 (0.0080)	0.1285 (0.0172)
SNR = 6												
ch1	0.1317 (0.0181)	0.0968 (0.0072)	0.1396 (0.0077)	0.1061 (0.0081)	0.1213 (0.0141)	0.0626 (0.0049)	0.1048 (0.0072)	0.0874 (0.0132)	0.1099 (0.0104)	0.0968 (0.0083)	0.1255 (0.0086)	0.0906 (0.0131)
ch2	0.1361 (0.0211)	0.1024 (0.0086)	0.1247 (0.0077)	0.1088 (0.0083)	0.1242 (0.0151)	0.0626 (0.0049)	0.1020 (0.0067)	0.0890 (0.0146)	0.1176 (0.0116)	0.1059 (0.0096)	0.1029 (0.0087)	0.0963 (0.0156)
ch3	0.1139 (0.0259)	0.0676 (0.0061)	0.0943 (0.0062)	0.1039 (0.0074)	0.1189 (0.0254)	0.0626 (0.0049)	0.1010 (0.0072)	0.0850 (0.0121)	0.0401 (0.0198)	0.0409 (0.0089)	0.0442 (0.0058)	0.0860 (0.0107)

pointed out the benefit of separating the transients and oscillations before spindle detection, see Coppieters et al. (2016) and Parekh et al. (2015). There exist already several methods for separating transients and oscillations in EEG data, but

Table 8
Fraction of correctly retrieved variables $\left(\mathrm{TP}_{low}/\left|S_0^{\alpha}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{low}/\left|S_0^{\alpha}\right|\right)$ for the estimated low resonance signal component. Fraction of correctly retrieved variables $\left(\mathrm{TP}_{high}/\left|S_0^{\beta}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{high}/\left|S_0^{\beta}\right|\right)$ for channel 1 and 2 and false positives FP_{high} for channel 3 for the estimated high resonance signal component. Values are based on 100 simulations with different noise realizations for Scenario 1 and SNR = 1.5, 3 and 6.

	$(\mathrm{TP}_{low})/\mathrm{p}_{low}$				$\mathrm{FN}_{low}/\mathrm{p}_{low}$				$(\mathrm{TP}_{high})/\mathrm{p}_{high}$				$\mathrm{FN}_{high}/\mathrm{p}_{high}$				FP_{high}			
	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP
SNR = 1.5																				
ch1	0.5713	0.9517	0.9342	0.7792	0.4287	0.0483	0.0658	0.2208	0.4796	0.6667	0.7925	0.5763	0.5204	0.3333	0.2075	0.4238	–	–	–	–
ch2	0.5075	0.9517	0.9342	0.7792	0.4925	0.0483	0.0658	0.2208	0.4638	0.6667	0.7925	0.5763	0.5363	0.3333	0.2075	0.4238	–	–	–	–
ch3	0.2592	0.9517	0.9342	0.7792	0.7408	0.0483	0.0658	0.2208	–	–	–	–	–	–	–	–	7.2400	42.8300	141.4500	35.5200
SNR = 3																				
ch1	0.8000	0.9938	0.9650	0.9383	0.2000	0.0062	0.0350	0.0617	0.5946	0.8075	0.8492	0.7433	0.4054	0.1925	0.1508	0.2567	–	–	–	–
ch2	0.7896	0.9938	0.9650	0.9383	0.2104	0.0062	0.0350	0.0617	0.5929	0.8075	0.8492	0.7433	0.4071	0.1925	0.1508	0.2567	–	–	–	–
ch3	0.6358	0.9938	0.9650	0.9383	0.3642	0.0062	0.0350	0.0617	–	–	–	–	–	–	–	–	14.6500	64.7200	101.4200	35.7100
SNR = 6																				
ch1	0.9238	0.9996	0.9888	0.9888	0.0762	0.0004	0.0113	0.0113	0.7438	0.9062	0.8896	0.8554	0.2563	0.0938	0.1104	0.1446	–	–	–	–
ch2	0.9450	0.9996	0.9888	0.9888	0.0550	0.0004	0.0113	0.0113	0.7200	0.9062	0.8896	0.8554	0.2800	0.0938	0.1104	0.1446	–	–	–	–
ch3	0.9446	0.9996	0.9888	0.9888	0.0554	0.0004	0.0113	0.0113	–	–	–	–	–	–	–	–	34.0200	80.3600	62.5500	35.7900

Table 9
Average values (standard deviations between parentheses) of RMSE, $RMSE_{low}$ and $RMSE_{high}$ based on 100 simulations with different noise realizations. Experiment carried out on Scenario 2 with SNR = 1.5, 3 and 6.

	RMSE				$RMSE_{low}$				$RMSE_{high}$			
	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP
SNR = 1.5												
ch1	0.1848	0.1560	0.1555	0.1520	0.1478	0.0834	0.1180	0.1093	0.1497	0.1521	0.1301	0.1275
	(0.0183)	(0.0167)	(0.0108)	(0.0159)	(0.0127)	(0.0082)	(0.0110)	(0.0210)	(0.0156)	(0.0179)	(0.0137)	(0.0202)
ch2	0.1795	0.1360	0.1494	0.1486	0.1455	0.0834	0.1122	0.1102	0.1318	0.1237	0.1204	0.1212
	(0.0211)	(0.0126)	(0.0097)	(0.0141)	(0.0129)	(0.0082)	(0.0109)	(0.0201)	(0.0160)	(0.0135)	(0.0105)	(0.0176)
ch3	0.1673	0.0898	0.1353	0.1439	0.1674	0.0834	0.1111	0.1077	0.0158	0.0414	0.0878	0.1114
	(0.0188)	(0.0087)	(0.0106)	(0.0165)	(0.0187)	(0.0082)	(0.0111)	(0.0215)	(0.0194)	(0.0115)	(0.0096)	(0.0184)
SNR = 3												
ch1	0.1374	0.1094	0.1113	0.1027	0.1165	0.0603	0.0851	0.0723	0.1103	0.1069	0.0964	0.0808
	(0.0153)	(0.0105)	(0.0084)	(0.0117)	(0.0122)	(0.0061)	(0.0077)	(0.0151)	(0.0111)	(0.0112)	(0.0108)	(0.0149)
ch2	0.1307	0.1001	0.1079	0.1020	0.1106	0.0603	0.0819	0.0720	0.0995	0.0938	0.0905	0.0825
	(0.0155)	(0.0101)	(0.0091)	(0.0102)	(0.0131)	(0.0061)	(0.0079)	(0.0130)	(0.0106)	(0.0105)	(0.0091)	(0.0127)
ch3	0.1214	0.0659	0.0844	0.0977	0.1222	0.0603	0.0792	0.0710	0.0204	0.0347	0.0483	0.0745
	(0.0227)	(0.0060)	(0.0087)	(0.0091)	(0.0231)	(0.0061)	(0.0099)	(0.0120)	(0.0165)	(0.0070)	(0.0064)	(0.0117)
SNR = 6												
ch1	0.0945	0.0790	0.0951	0.0692	0.0822	0.0442	0.0703	0.0502	0.0794	0.0793	0.0802	0.0536
	(0.0103)	(0.0077)	(0.0077)	(0.0068)	(0.0094)	(0.0046)	(0.0067)	(0.0083)	(0.0086)	(0.0081)	(0.0083)	(0.0088)
ch2	0.0912	0.0731	0.0934	0.0702	0.0791	0.0442	0.0670	0.0500	0.0755	0.0713	0.0775	0.0555
	(0.0090)	(0.0075)	(0.0073)	(0.0076)	(0.0097)	(0.0046)	(0.0071)	(0.0090)	(0.0073)	(0.0083)	(0.0073)	(0.0081)
ch3	0.0804	0.0483	0.0645	0.0688	0.0780	0.0442	0.0637	0.0493	0.0261	0.0276	0.0278	0.0532
	(0.0152)	(0.0047)	(0.0069)	(0.0077)	(0.0159)	(0.0046)	(0.0076)	(0.0096)	(0.0183)	(0.0061)	(0.0053)	(0.0092)

here we refer to Lajnef et al. (2015) where the joint detection of sleep spindles and K-complex events are obtained using a Morphological Component Analysis (MCA) and two different RADWT with respectively high and low Q-factors, as supposed in this paper. On the other hand, although the American Academy of Sleep Medicine (AASM) manual recommends using more the one channel for scoring sleep and associated events, actually only few available methods advocate the use of multichannel EEG (Barros et al., 2000; Parekh et al., 2017), then our procedure can be considered a possible alternative in this respect.

In particular in this section we show results obtained by applying our proposed multichannel procedure to one publicly sleep EEG database, the DREAMS Sleep Spindles Database available at www.tcts.fpms.ac.be/~devuyst/Databases/DatabaseSpindles/. This database has been produced by the University of MONS — TCTS Laboratory (Stéphanie Devuyst, Thierry Dutoit) and the Université Libre de Bruxelles — CHU de Charleroi Sleep Laboratory (Myriam Kerkhofs).

These data were acquired in a sleep laboratory of a Belgium hospital using a digital 32-channel polygraph (BrainnetTM System of MEDATEC, Brussels, Belgium). They consist of height polysomnographic recordings coming from patients with different pathologies (dysomnia, restless legs syndrome, insomnia, apnoea/hypopnoea syndrome). Two EOG channels (P8-A1, P18-A1), three EEG channels (CZ-A1 or C3-A1, FP1-A1 and O1-A1) and one submental EMG channel were recorded. The standard European Data Format (EDF) was used for storing. The sampling frequency was 200 Hz, 100 Hz or 50 Hz. A segment of 30 min of the central EEG channel was extracted from each whole-night recording for spindles scoring, giving origin to 8 excerpts of 30 min. No effort was made to select good spindle epochs or noise free epochs, in order to reflect reality as much as possible. These excerpts were given independently to two experts for sleep spindles scoring.

In particular we focus on excerpt2 sampled at 200 Hz extracted from 00:00:00 to 00:30:00 with annotated EEG channels CZ-A1, FP1-A1 and O1-A1, belonging to a 40-years man, i.e. 3 signals, one for channel, formed by 360000 time points.

We segmented each signal in 360 segments of length 1000 time points, corresponding to 5 s, and we concentrate only on the 200 segments corresponding to sleep phase 2. In particular we focused on two consecutive segments: 25–30 s and 30–35 s, see Figs. 5–7 respectively. In both the segments the two experts annotated visually spindles events at same times. Indeed, in the first segment, the first expert annotated a spindle event at 26.09 s of length 1.28 s and the second expert annotated the event at 26.12 s with length 1 s; in the second segment, the first expert annotated a spindle event at 31.5 s of length 0.74 s and the second expert annotated the event at 31.515 s with length 1 s.

Following Lajnef et al. (2015), we suppose the oscillatory part to be well described by an RADWT with Q-factor=5 (which roughly corresponds to the choice $p = 8, q = 9, s = 3, J = 10$) and the transient part to be well represented by an RADWT with Q-factor=1 (which roughly corresponds to the choice $p = 1, q = 2, s = 1, J = 4$). Moreover we suppose that hypothesis (H1) is true, since we are considering sleep data where the epochs containing electrode artifacts due to lead and other body movements are not analyzed, hence we expect the 3 channels share the same underground/transient activity; we also suppose that hypothesis (H2) is true, since the spindles events, which represent the major and also the most interesting contribution to the oscillating part, simultaneously activate in the 3 channels, as widely discussed in Parekh et al. (2017).

Figs. 6–8 show the retrieval of the transient and oscillatory components for the two considered segments, 25–30 s and 30–35 s respectively. From the figures we can see how the transient part is really faithful to the underlying trend of the three

Table 10
Fraction of correctly retrieved variables $\left(\mathrm{TP}_{low}/\left|S_0^{\alpha}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{low}/\left|S_0^{\alpha}\right|\right)$ for the estimated low resonance signal component. Fraction of correctly retrieved variables $\left(\mathrm{TP}_{high}/\left|S_0^{\beta}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{high}/\left|S_0^{\beta}\right|\right)$ for channel 1 and 2 and false positives FP_{high} for channel 3 for the estimated high resonance signal component. Values are based on 100 simulations with different noise realizations for Scenario 2 and SNR = 1.5, 3 and 6.

	$(\mathrm{TP}_{low})/\mathrm{p}_{low}$				$\mathrm{FN}_{low}/\mathrm{p}_{low}$				$(\mathrm{TP}_{high})/\mathrm{p}_{high}$				$\mathrm{FN}_{high}/\mathrm{p}_{high}$				FP_{high}			
	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP
SNR = 1.5																				
ch1	0.6983	0.9992	0.9925	0.9550	0.3017	0.0008	0.0075	0.0450	0.7825	0.8575	0.8308	0.7692	0.2175	0.1425	0.1692	0.2308	–	–	–	–
ch2	0.7242	0.9992	0.9925	0.9550	0.2758	0.0008	0.0075	0.0450	0.5642	0.8575	0.8308	0.7692	0.4358	0.1425	0.1692	0.2308	–	–	–	–
ch3	0.4233	0.9992	0.9925	0.9550	0.5767	0.0008	0.0075	0.0450	–	–	–	–	–	–	–	–	5.0300	35.9400	120.7500	18.8000
SNR = 3																				
ch1	0.9333	1.0000	1.0000	0.9983	0.0667	0.0000	0.0000	0.0017	0.8158	0.9050	0.8350	0.8758	0.1842	0.0950	0.1650	0.1242	–	–	–	–
ch2	0.9608	1.0000	1.0000	0.9983	0.0392	0.0000	0.0000	0.0017	0.6567	0.9050	0.8350	0.8758	0.3433	0.0950	0.1650	0.1242	–	–	–	–
ch3	0.9133	1.0000	1.0000	0.9983	0.0867	0.0000	0.0000	0.0017	–	–	–	–	–	–	–	–	11.6100	38.2500	73.4700	18.8700
SNR = 6																				
ch1	0.9983	1.0000	1.0000	1.0000	0.0017	0.0000	0.0000	0.0000	0.8292	0.9267	0.8350	0.9325	0.1708	0.0733	0.1650	0.0675	–	–	–	–
ch2	0.9983	1.0000	1.0000	1.0000	0.0017	0.0000	0.0000	0.0000	0.6908	0.9267	0.8350	0.9325	0.3092	0.0733	0.1650	0.0675	–	–	–	–
ch3	0.9983	1.0000	1.0000	1.0000	0.0017	0.0000	0.0000	0.0000	–	–	–	–	–	–	–	–	27.5800	42.7800	35.3700	18.9900

Table 11
Average values (standard deviations between parentheses) of RMSE, $RMSE_{low}$ and $RMSE_{high}$ based on 100 simulations with different noise realizations. Experiment carried out on Scenario 3 with SNR = 1.5, 3 and 6.

	RMSE				$RMSE_{low}$				$RMSE_{high}$			
	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP
SNR = 1.5												
ch1	0.0382	0.0296	0.0291	0.0294	0.0299	0.0154	0.0195	0.0197	0.0248	0.0262	0.0225	0.0225
	(0.0077)	(0.0038)	(0.0029)	(0.0036)	(0.0071)	(0.0023)	(0.0032)	(0.0047)	(0.0049)	(0.0040)	(0.0032)	(0.0045)
ch2	0.0395	0.0259	0.0259	0.0305	0.0319	0.0154	0.0194	0.0195	0.0244	0.0217	0.0180	0.0238
	(0.0077)	(0.0029)	(0.0028)	(0.0039)	(0.0071)	(0.0023)	(0.0031)	(0.0052)	(0.0042)	(0.0029)	(0.0028)	(0.0050)
ch3	0.0324	0.0162	0.0204	0.0281	0.0321	0.0154	0.0192	0.0195	0.0050	0.0038	0.0081	0.0198
	(0.0072)	(0.0023)	(0.0028)	(0.0035)	(0.0076)	(0.0023)	(0.0031)	(0.0055)	(0.0048)	(0.0021)	(0.0019)	(0.0050)
SNR = 3												
ch1	0.0276	0.0206	0.0260	0.0195	0.0213	0.0111	0.0164	0.0129	0.0183	0.0183	0.0203	0.0147
	(0.0051)	(0.0029)	(0.0023)	(0.0031)	(0.0045)	(0.0013)	(0.0026)	(0.0039)	(0.0033)	(0.0031)	(0.0024)	(0.0035)
ch2	0.0283	0.0193	0.0235	0.0207	0.0222	0.0111	0.0167	0.0132	0.0183	0.0167	0.0166	0.0163
	(0.0046)	(0.0020)	(0.0019)	(0.0026)	(0.0042)	(0.0013)	(0.0023)	(0.0038)	(0.0030)	(0.0021)	(0.0019)	(0.0028)
ch3	0.0236	0.0120	0.0166	0.0195	0.0235	0.0111	0.0164	0.0126	0.0030	0.0041	0.0036	0.0147
	(0.0045)	(0.0014)	(0.0026)	(0.0027)	(0.0047)	(0.0013)	(0.0027)	(0.0038)	(0.0028)	(0.0015)	(0.0013)	(0.0037)
SNR = 6												
ch1	0.0194	0.0150	0.0256	0.0136	0.0145	0.0076	0.0160	0.0090	0.0135	0.0136	0.0200	0.0103
	(0.0031)	(0.0019)	(0.0018)	(0.0017)	(0.0029)	(0.0011)	(0.0016)	(0.0023)	(0.0024)	(0.0022)	(0.0019)	(0.0023)
ch2	0.0200	0.0146	0.0228	0.0140	0.0147	0.0076	0.0159	0.0090	0.0139	0.0131	0.0163	0.0108
	(0.0032)	(0.0014)	(0.0017)	(0.0021)	(0.0030)	(0.0011)	(0.0018)	(0.0026)	(0.0022)	(0.0017)	(0.0015)	(0.0025)
ch3	0.0170	0.0084	0.0160	0.0136	0.0170	0.0076	0.0159	0.0092	0.0023	0.0033	0.0021	0.0101
	(0.0035)	(0.0011)	(0.0019)	(0.0018)	(0.0036)	(0.0011)	(0.0019)	(0.0025)	(0.0023)	(0.0013)	(0.0008)	(0.0025)

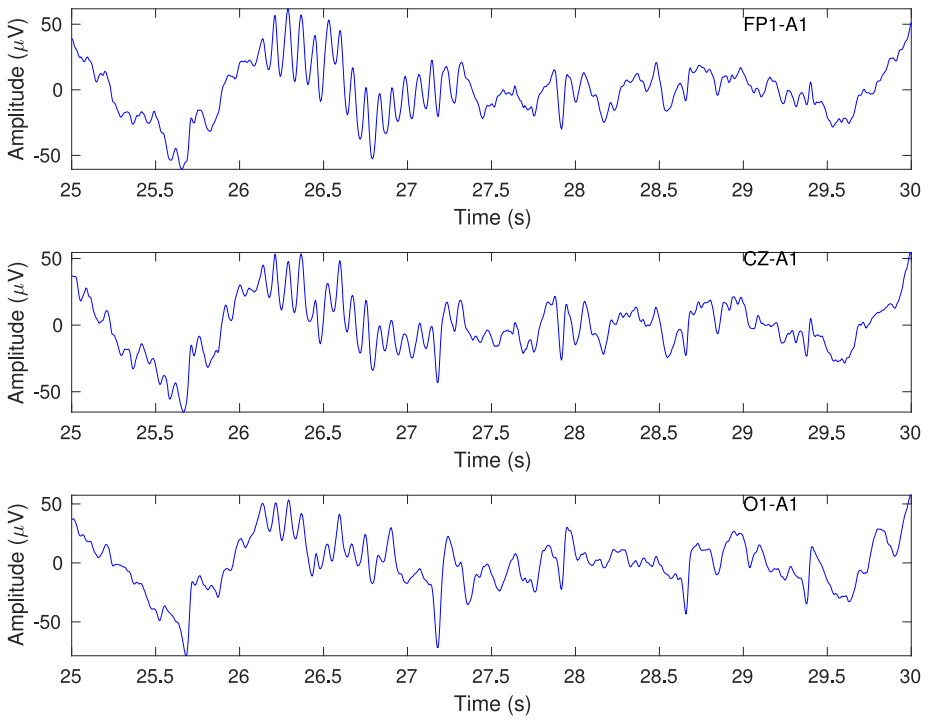

Fig. 5. EEG segment 6 corresponding to time interval −30 s for the 3 EEG channels, FP1-A1 , CZ-A1 and O1-A1.

channels, it keeps some oscillations that do not persist in time; moreover we can appreciate the 3 oscillatory components, in which similar but not equal oscillations resonate in the same time intervals. These phenomena correspond to the spindle events that most likely occur contemporaneously on the three EEG channels with similar characteristics being not exactly the same. Of course this procedure must be considered as a preprocessing step for an automatic spindles detection, which in this case appears very clear around sec. 26 in the first excerpt (visually inspecting Fig. 6) and around sec. 32 in the second

Table 12
Fraction of correctly retrieved variables $\left(\mathrm{TP}_{low}/\left|S_0^{\alpha}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{low}/\left|S_0^{\alpha}\right|\right)$ for the estimated low resonance signal component. Fraction of correctly retrieved variables $\left(\mathrm{TP}_{high}/\left|S_0^{\beta}\right|\right)$ and incorrectly retrieved variables $\left(\mathrm{FN}_{high}/\left|S_0^{\beta}\right|\right)$ for channel 1 and 2 and false positives FP_{high} for channel 3 for the estimated high resonance signal component. Values are based on 100 simulations with different noise realizations for Scenario 3 and SNR = 1.5, 3 and 6.

	$(\mathrm{TP}_{low})/\mathrm{p}_{low}$				$\mathrm{FN}_{low}/\mathrm{p}_{low}$				$(\mathrm{TP}_{high})/\mathrm{p}_{high}$				$\mathrm{FN}_{high}/\mathrm{p}_{high}$				FP_{high}			
	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP	single-c	multi-c	BCD	SOMP
SNR = 1.5																				
ch1	0.9900	1.0000	1.0000	1.0000	0.0100	0.0000	0.0000	0.0000	0.5467	0.6033	0.8483	0.6467	0.4533	0.3967	0.1517	0.3533	–	–	–	–
ch2	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.4983	0.6033	0.8483	0.6467	0.5017	0.3967	0.1517	0.3533	–	–	–	–
ch3	0.9967	1.0000	1.0000	1.0000	0.0033	0.0000	0.0000	0.0000	–	–	–	–	–	–	–	–	7.3300	8.3600	35.7400	8.4200
SNR = 3																				
ch1	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.6517	0.7950	0.8717	0.7933	0.3483	0.2050	0.1283	0.2067	–	–	–	–
ch2	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.6633	0.7950	0.8717	0.7933	0.3367	0.2050	0.1283	0.2067	–	–	–	–
ch3	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	–	–	–	–	–	–	–	–	4.7100	13.9000	9.2600	9.0000
SNR = 6																				
ch1	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.7433	0.9200	0.8500	0.9133	0.2567	0.0800	0.1500	0.0867	–	–	–	–
ch2	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.7833	0.9200	0.8500	0.9133	0.2167	0.0800	0.1500	0.0867	–	–	–	–
ch3	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	–	–	–	–	–	–	–	–	5.9600	13.6600	5.6400	9.2700

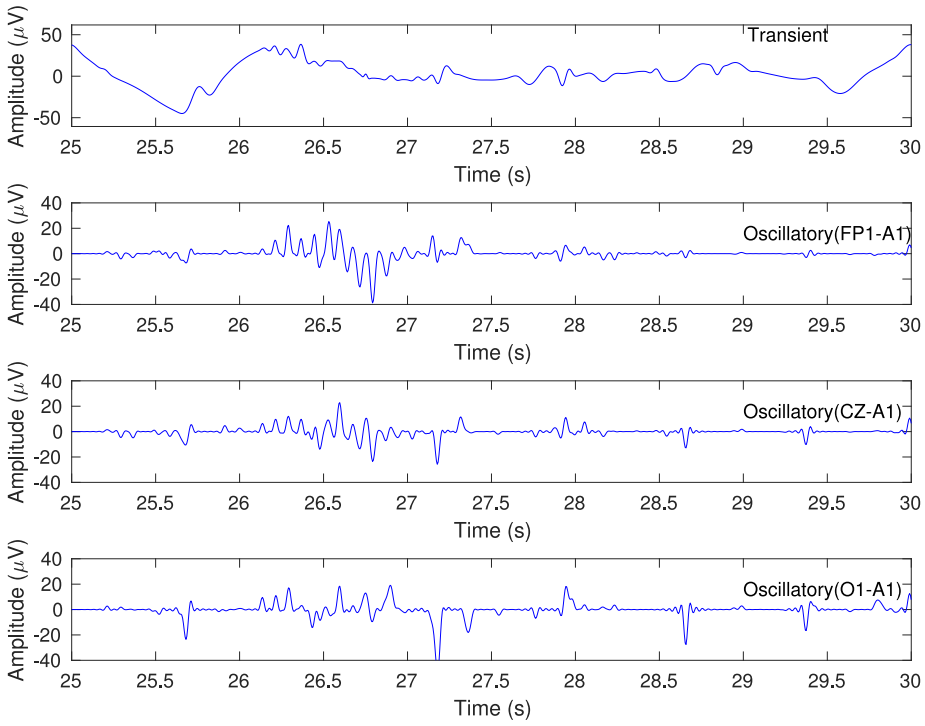

Fig. 6. Retrieved low-transient and high-oscillatory components of segment 6 corresponding to time interval 25–30 s for the 3 EEG channels, FP1-A1 , CZ-A1 and O1-A1.

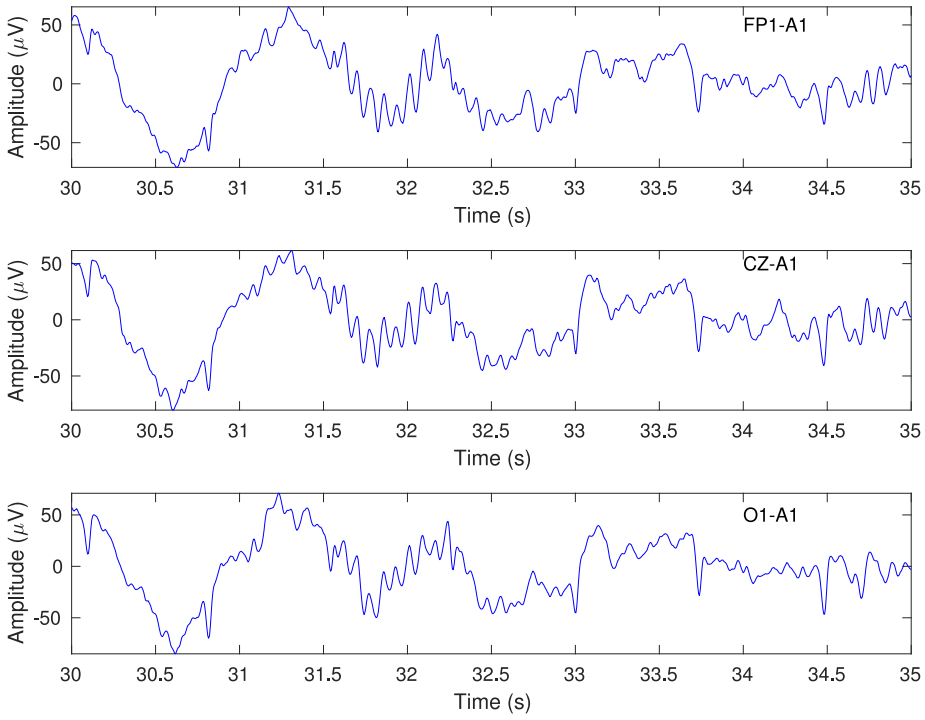

Fig. 7. EEG segment 7 corresponding to time interval 30–35 s for the 3 EEG channels, FP1-A1 , CZ-A1 and O1-A1.

excerpt (visually inspecting Fig. 8). The analyzed segment 25–30 s corresponds to the segment analyzed in paper Parekh et al. (2017), see Fig. 5, and it can be seen that the position of spindles coincides.

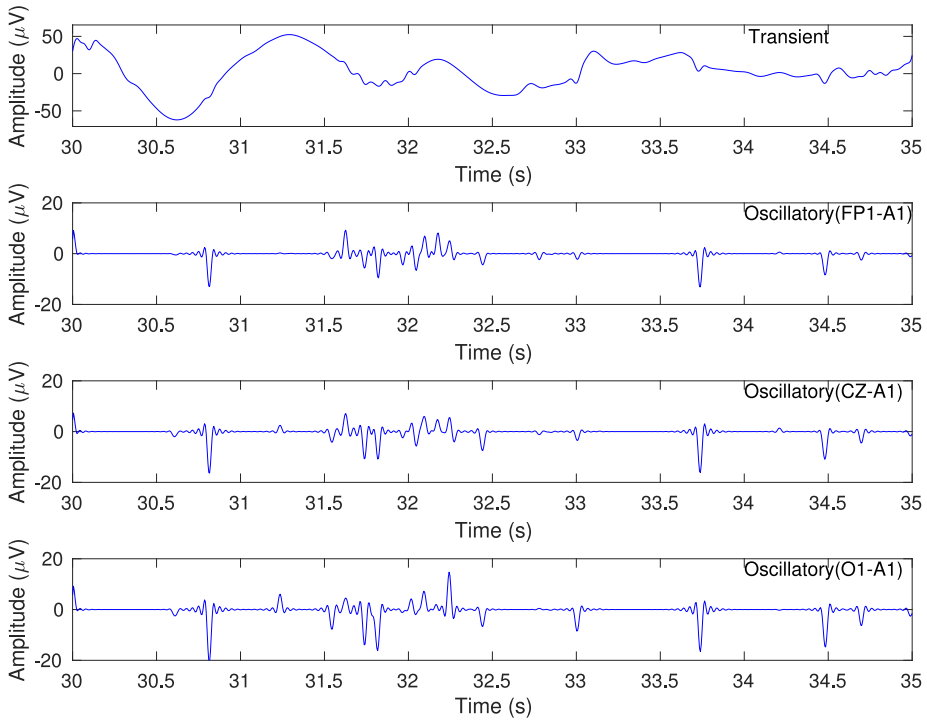

Fig. 8. Retrieved low-transient and high-oscillatory components of segment 7 corresponding to time interval 25–30 s for the 3 EEG channels, FP1-A1 , CZ-A1 and O1-A1.

6. Conclusions

In this paper we presented a method for nonparametric regression analysis of multichannel signals under a structural hypothesis on the underlying signals covering some specific real life situations. The method leverages on a complete filter bank (RADWT) that defines a frame in $L_2(R)$ which guarantees a perfect reconstruction property and a tunable Q-factor. In our work we used two frames, one with low Q-factor and one with high Q-factor, able to represent sparsity of signals with low and high resonance respectively. The structural hypothesis on the underlying signals explicitly states that in each channel the signal is a sum of two contributions, one (the low resonance signal) is common to all channels, while the other (the high resonance signal) is channel-specific but retains the same spectral properties in each channel, i.e. the positions of non-zero RADWT coefficients. We showed the connections with the SSA problem, stressing the difference between our proposal and the existing literature.

Firstly, we applied the method on a set of synthetic data satisfying the mathematical hypotheses, showing its ability in retrieving the signal in each channel, as expected from its asymptotic properties. We also compared its performance with other two techniques proposed in the literature, namely SOMP and BCD, considering a second synthetic dataset from a non correct RADWT generative model to test the robustness. Moreover, we displayed its skill in reconstructing the individual components and in controlling the sparsity of the model too. Finally, the proposed technique was tested on human sleep EEG data, confirming some results already studied in the literature.

Future research is devoted to the improvement of the algorithm in pursuing component specific results.

Acknowledgments

Daniela De Canditiis was partially supported by grant INdAM-GNCS Project 2018. We wish to thank the anonymous reviewers, the Associate Editor and the Editor for their helpful comments which have led to substantial improvements in the paper. In particular, we are extremely grateful to professor Anestis Antoniadis for providing valuable and detailed suggestions.

Appendix

Before proving Theorem 1, let us present some preliminary results.

For each $j = 1, \dots, d_1$, define the random variables

$$u_j = \frac{\boldsymbol{\varepsilon}^t \boldsymbol{X}^{(j)}}{\sqrt{nK}} \qquad \text{with} \quad \boldsymbol{X}^{(j)} = \left[\underbrace{\left(\boldsymbol{\Psi}^{(j)}\right)^t, \dots, \left(\boldsymbol{\Psi}^{(j)}\right)^t}_{K \text{ times}} \right]^t, \tag{A.10}$$

where $\boldsymbol{X}^{(j)}$ is the jth column of matrix $\boldsymbol{X}$ and $\boldsymbol{\Psi}^{(j)}$ the jth column of matrix $\boldsymbol{\Psi}$.

Proposition 1. *For the random variables u_j it holds for any $x > 0$*

$$P\left(\max_{1 \le j \le d_1} 2|u_j| < \sqrt{nK}\lambda_0^{\alpha} \right) \ge 1 - 2e^{-x^2/2}, \tag{A.11}$$

where

$$\lambda_0^{\alpha} = \frac{2\,\sigma}{\sqrt{nK}} \sqrt{x^2 + 2\,\log(d_1)}.$$

Proof. Since $u_j = \frac{1}{\sqrt{nK}} \sum_{k=1}^{K} \sum_{i=1}^{n} \epsilon_i^{(k)} \Psi_i^{(j)} \sim \mathcal{N}(0, \sigma^2)$ we can apply lemma 6.2 of Bühlmann and van de Geer (2011) and result is proved.

For each $j = 1, \dots, d_2$, define the random variables

$$v_j = \frac{\left\| \boldsymbol{\varepsilon}^t \tilde{\boldsymbol{X}}^{(j)} \right\|_2}{\sqrt{nK}} \qquad \text{with} \quad \tilde{\boldsymbol{X}}^{(j)} = \begin{bmatrix} \boldsymbol{\Phi}^{(j)} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Phi}^{(j)} & \cdots & \boldsymbol{0} \\ \cdots & & & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Phi}^{(j)} \end{bmatrix}, \tag{A.12}$$

being a matrix of dimension $nK \times K$ with $\boldsymbol{\Phi}^{(j)}$ the jth column of matrix $\boldsymbol{\Phi}$.

Proposition 2. *For the random variables v_j it holds for any $x > 0$*

$$P\left(\max_{1 \le j \le d_1 2} 2|v_j| < \sqrt{nK}\lambda_0^{\beta} \right) \ge 1 - e^{-x}, \tag{A.13}$$

where

$$\lambda_0^{\beta} = \frac{2\,\sigma}{\sqrt{nK}} \left(1 + \sqrt{(4x + 4\,\log(d_2))/K} + (4x + 4\,\log(d_2))/K \right).$$

Proof. By definition we have that

$$v_j = \frac{1}{\sqrt{K}} \left\| \left(\frac{\sum_{i=1}^{n} \epsilon_i^{(1)} \Phi_i^{(j)}}{\sqrt{n}}, \dots, \frac{\sum_{i=1}^{n} \epsilon_i^{(K)} \Phi_i^{(j)}}{\sqrt{n}} \right) \right\|_2 = \frac{\sigma}{\sqrt{K}} \left(\sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{n} \epsilon_i^{(k)} \Phi_i^{(j)}}{\sigma\sqrt{n}} \right)^2 \right)^{1/2}.$$

Since $\frac{\sum_{i=1}^{n} \epsilon_i^{(k)} \Phi_i^{(j)}}{\sigma\sqrt{n}}$ are K independent normal standard variables, we have that $K v_j^2/\sigma^2 \sim \chi^2(K)$. Finally, applying lemma 8.1 of Bühlmann and van de Geer (2011) result is proved.

Proposition 3. *For all $\boldsymbol{\theta} \in \mathbb{R}^{d_1 + Kd_2 \times 1}$ and for any $x > 0$ it holds*

$$P\left(\frac{2\,\boldsymbol{\varepsilon}^t\, \boldsymbol{X}\, \boldsymbol{\theta}}{nK} \le \lambda_0 \sqrt{G^{\star}} \|\boldsymbol{\theta}\|_{2,1} \right) \ge 1 - 2e^{-x^2/2} - e^{-x},$$

with $\boldsymbol{\varepsilon}$ the concatenation of noise vectors given in Eq. (3.3) and $\lambda_0 = max\{\lambda_0^{\alpha}, \lambda_0^{\beta}/\sqrt{K}\}$.

Proof. By definitions of $\boldsymbol{\theta}$ we can write

$$\frac{2\,\boldsymbol{\varepsilon}^t\, \boldsymbol{X}\, \boldsymbol{\theta}}{nK} = \frac{2\,\boldsymbol{\varepsilon}^t\, \boldsymbol{X}}{nK} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \\ \vdots \\ \boldsymbol{\beta}^{(K)} \end{bmatrix} = \frac{1}{\sqrt{nK}} \left(\sum_{j=1}^{d_1} 2 \frac{\boldsymbol{\varepsilon}^t \boldsymbol{X}^{(j)}}{\sqrt{nK}} \alpha_j + \sum_{j=1}^{d_2} \left(\frac{\boldsymbol{\varepsilon}^t \tilde{\boldsymbol{X}}^{(j)}}{\sqrt{nK}} \right) \boldsymbol{\beta}_j^{(\cdot)} \right),$$

where $\boldsymbol{\beta}_j^{(\cdot)} = \left[\beta_j^{(1)}, \dots, \beta_j^{(K)}\right]^t$, while $\boldsymbol{X}^{(j)}$ and $\tilde{\boldsymbol{X}}^{(j)}$ are given in (A.10) and (A.12). Using Propositions 1 and 2 and the fact that $uv \le |u||v|, \forall\, u, v \in \mathbb{R}$ and $< \boldsymbol{u}, \boldsymbol{v} > \le \|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2, \forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^K$, with probability at least $1 - 2e^{-x^2/2} - e^{-x}$ it follows

$$\begin{aligned}
\frac{2\,\boldsymbol{\varepsilon}^t\,\boldsymbol{X}\,\boldsymbol{\theta}}{nK} &\le \frac{1}{\sqrt{nK}}\left(\sum_{j=1}^{d_1} 2\frac{\left|\boldsymbol{\varepsilon}^t\boldsymbol{X}^{(j)}\right|}{\sqrt{nK}}|\alpha_j| + 2\sum_{j=1}^{d_2}\frac{\left\|\boldsymbol{\varepsilon}^t\tilde{\boldsymbol{X}}^{(j)}\right\|_2}{\sqrt{nK}}\left\|\boldsymbol{\beta}_j^{(\cdot)}\right\|_2\right)\\
&\le \frac{1}{\sqrt{nK}}\left(\sum_{j=1}^{d_1} 2|u_j||\alpha_j| + \sum_{j=1}^{d_2} 2|v_j|\left\|\boldsymbol{\beta}_j^{(\cdot)}\right\|_2\right)\\
&\le \frac{1}{\sqrt{nK}}\left(\max_{1\le j\le d_1} 2|u_j|\,\|\boldsymbol{\alpha}\|_1 + \max_{1\le j\le d_1} 2|v_j|\sum_{j=1}^{d_2}\left\|\boldsymbol{\beta}_j^{(\cdot)}\right\|_2\right)\\
&\le \left(\lambda_0^{\alpha}\,\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_0^{\beta}}{\sqrt{K}}\sqrt{K}\sum_{j=1}^{d_2}\left\|\boldsymbol{\beta}_j^{(\cdot)}\right\|_2\right)\\
&\le \sqrt{G^\star}\lambda_0\left(\frac{1}{\sqrt{G^\star}}\|\boldsymbol{\alpha}\|_1 + \sqrt{\frac{K}{G^\star}}\sum_{j=1}^{d_2}\left\|\boldsymbol{\beta}_j^{(\cdot)}\right\|_2\right) = \sqrt{G^\star}\lambda_0\,\|\boldsymbol{\theta}\|_{2,1},
\end{aligned}$$

where $\lambda_0 = max\{\lambda_0^{\alpha}, \lambda_0^{\beta}/\sqrt{K}\}$.

Proof of Theorem 1. By definition of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ it holds

$$\frac{1}{nK}\left\|\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\right\|_2^2 + \lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}\right\|_{2,1} \le \frac{1}{nK}\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2 + \lambda\sqrt{G^\star}\,\|\boldsymbol{\theta}_0\|_{2,1}\,,$$

then, by using $\mathbf{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}$, it also holds

$$\frac{1}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2^2 + \lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}\right\|_{2,1} \le \frac{2\boldsymbol{\varepsilon}^t\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}{nK} + \lambda\sqrt{G^\star}\,\|\boldsymbol{\theta}_0\|_{2,1}\,.$$

Choose any x, then with probability at least $1 - 2e^{-x^2/2} - e^{-x}$, by Proposition 3, it holds

$$\frac{1}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2^2 + \lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}\right\|_{2,1} \le \sqrt{G^\star}\lambda_0\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_{2,1} + \lambda\sqrt{G^\star}\,\|\boldsymbol{\theta}_0\|_{2,1}\,.$$

Choose $\lambda > 2\lambda_0$, and observe that, whatever $S_0 \subseteq \mathcal{P}$, one has $\|\boldsymbol{\theta}\|_{2,1} = \|\boldsymbol{\theta}(S_0)\|_{2,1} + \left\|\boldsymbol{\theta}(S_0^c)\right\|_{2,1}$ for any $\boldsymbol{\theta}$ and in particular $\|\boldsymbol{\theta}_0\|_{2,1} = \|\boldsymbol{\theta}_0(S_0)\|_{2,1}$, then it holds

$$\begin{aligned}
&\frac{2}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2^2 + 2\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}(S_0^c) - \boldsymbol{\theta}_0(S_0^c)\right\|_{2,1} \le \sqrt{G^\star}\lambda\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{2,1} +\\
&+ 2\lambda\sqrt{G^\star}\left(\|\boldsymbol{\theta}_0(S_0)\|_{2,1} - \left\|\hat{\boldsymbol{\theta}}(S_0)\right\|_{2,1}\right).
\end{aligned}$$

By using the triangle inequality for the l_2/l_1-norm, $\left|\,\|v\|_{2,1} - \|u\|_{2,1}\,\right| \le \|u - v\|_{2,1}$ and rewriting $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{2,1} = \left\|\hat{\boldsymbol{\theta}}(S_0) - \boldsymbol{\theta}_0(S_0)\right\|_{2,1} + \left\|\hat{\boldsymbol{\theta}}(S_0^c) - \boldsymbol{\theta}_0(S_0^c)\right\|_{2,1}$, it holds

$$\frac{2}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2^2 + \lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}(S_0^c) - \boldsymbol{\theta}_0(S_0^c)\right\|_{2,1} \le 3\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}(S_0) - \boldsymbol{\theta}_0(S_0)\right\|_{2,1}. \tag{A.14}$$

Now from Eq. (A.14) we obtain two consequences.

The first is that $\left\|\hat{\boldsymbol{\theta}}(S_0^c) - \boldsymbol{\theta}_0(S_0^c)\right\|_{2,1} \le 3\left\|\hat{\boldsymbol{\theta}}(S_0) - \boldsymbol{\theta}_0(S_0)\right\|_{2,1}$, hence for assumption **(A2)**, it holds

$$G^\star\left\|\hat{\boldsymbol{\theta}}(S_0) - \boldsymbol{\theta}_0(S_0)\right\|_{2,1}^2 \le \frac{\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2^2\, G^\star|S_0|}{nK\phi(S_0)^2}. \tag{A.15}$$

The second is obtained adding $\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}(S_0) - \boldsymbol{\theta}_0(S_0)\right\|_{2,1}$ on both sides of Eq. (A.14), hence

$$\frac{2}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right\|_2^2 + \lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{2,1} \le 4\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}(S_0) - \boldsymbol{\theta}_0(S_0)\right\|_{2,1}. \tag{A.16}$$

Now, substitute Eq. (A.15) into Eq. (A.16) and obtain

$$\frac{2}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\|_2^2+\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right\|_{2,1}\leq 4\lambda\frac{\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\|_2\sqrt{G^\star\,|S_0|}}{\sqrt{nK}\,\phi(S_0)}.$$

Finally, using the inequality $4uv\leq u^2+4v^2$, we obtain

$$\frac{2}{nK}\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\|_2^2+\lambda\sqrt{G^\star}\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right\|_{2,1}\leq\frac{\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right\|_2^2}{nK}+4\frac{\lambda^2\,G^\star\,|S_0|}{\phi(S_0)^2},$$

which gives Eq. (4.9).

References

Argyriou, A., Evgeniou, T., Pontil, M., 2008. Convex multi-task feature learning. Mach. Learn. 73, 243–272. http://dx.doi.org/10.1007/s10994-007-5040-8.

Barros, A.K., Rosipal, R., Girolami, M., Dorffner, G., Ohnishi, N., 2000. Extraction of sleep-spindles from the electroencephalogram (eeg). In: Malmgren, H., Borga, M., Niklasson, L. (Eds.), Perspectives in Neural Computing. In: Artificial Neural Networks Med. Biol., Springer, London, pp. 125–130.

Bayram, I., Selesnick, I.W., 2009. Frequency-domain design of overcomplete rational-dilation wavelet transform. IEEE Trans. Signal Process. 57, 2957–2972. http://dx.doi.org/10.1109/TSP.2009.2020756.

Bobin, J., Moudden, Y., Fadili, J., Starck, J., 2009. Morphological diversity and sparsity for multichannel data restoration. J. Math. Imaging Vision 33, 149–168. http://dx.doi.org/10.1007/s10851-008-0065-6.

Breheny, P., Huang, J., 2009. Penalized methods for bi-level variable selection. Stat. Interface 2, 369–380. http://dx.doi.org/10.4310/SII.2009.v2.n3.a10.

Breheny, P., Huang, J., 2015. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat. Comput. 25, 173–187. http://dx.doi.org/10.1007/s11222-013-9424-2.

Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data. In: Springer Series in Statistics, Springer, Berlin, Heidelberg.

Coppieters, D., Maquet, P., Phillips, C., 2016. Sleep spindles as an electrographic element: Description and automatic detection methods. Neural Plast. 1–19. http://dx.doi.org/10.1155/2016/6783812, Article ID 6783812.

Deun, K.V., Wilderjans, T.F., van den Berg, R.A., Antoniadis, A., Mechelen, I.V., 2011. A flexible framework for sparse simultaneous component based data integration. BMC Bioinformatics 12, 1–17. http://dx.doi.org/10.1186/1471-2105-12-448.

Donoho, D., Elad, M., 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proc. Natl. Acad. Sci. USA 100, 2197–2202.

He, D., Kuhn, D., Parida, L., 2016. Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. Bioinformatics 32, i37–i43. http://dx.doi.org/10.1093/bioinformatics/btw249.

Huang, J., Breheny, P., Ma, S., 2012. A selective review of group selection in high-dimensional models. Statist. Sci. 27, 481–499. http://dx.doi.org/10.1214/12-STS392.

Jenatton, R., Audibert, J., Bach, F., 2011. Structured variable selection with sparsity-inducing norms. J. Mach. Learn. Res. 12, 2777–2824.

Lajnef, T., Chaibi, S., Eichenlaub, J., Ruby, P.M., Aguera, P., Samet, M., Kachouri, A., Jerbi, K., 2015. Sleep spindle and k-complex detection using tunable q-factor wavelet transform and morphological component analysis. Front. Hum. Neurosci. 9, 414. http://dx.doi.org/10.3389/fnhum.2015.00414.

Liu, H., Lafferty, J., Wasserman, L., 2008. Nonparametric regression and classification with joint sparsity constraints. In: Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, NIPS. Curran Associates Inc., Red Hook, New York NY, pp. 969–976.

Lozano, A.C., Swirszcz, G., 2012. Multi-level lasso for sparse multi-task regression. In: Proceedings of the 29th International Coference on International Conference on Machine Learning. Omnipress, USA, pp. 595–602, http://dl.acm.org/citation.cfm?id=30425733042652.

Parekh, A., Selesnick, I., Osorio, R.S., Varga, A., Rapoport, D.M., Ayappa, I., 2017. Multichannel sleep spindle detection using sparse low-rank optimization. J. Neurosci. Methods 288, 1–16. http://dx.doi.org/10.1016/j.jneumeth.2017.06.004.

Parekh, A., Selesnick, I., Rapoport, D.M., Ayappa, I., 2015. Detection of k-complexes and sleep spindles (detoks) using sparse optimization. J. Neurosci. Methods 251, 37–46. http://dx.doi.org/10.1016/j.jneumeth.2015.04.006.

Rakotomamonjy, A., 2011. Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms. Signal Process. 91, 1505–1526. http://dx.doi.org/10.1016/j.sigpro.2011.01.012.

Ruffalo, M., Stojanov, P., Pillutla, V.K., Varma, R., Bar-Joseph, Z., 2017. Reconstructing cancer drug response networks using multitask learning. BMC Syst. Biol. 11, 96. http://dx.doi.org/10.1186/s12918-017-0471-8.

Selesnick, I.W., 2011. Resonance-based signal decomposition: A new sparsity-enabled signal analysis method. Signal Process. 91, 2793–2809. http://dx.doi.org/10.1016/j.sigpro.2010.10.018.

Simon, N., Tibshirani, R., 2012. Standardization and the group lasso penalty. Statist. Sinica 22, 983–1001. http://dx.doi.org/10.5705/ss.2011.075.

Tropp, J.A., 2006. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. Signal Process. 86, 589–602. http://dx.doi.org/10.1016/j.sigpro.2005.05.031.

Tropp, J.A., Gilbert, A., Strauss, M.J., 2006. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. Signal Process. 86, 572–588. http://dx.doi.org/10.1016/j.sigpro.2005.05.030.

Yang, Y., Zou, H., 2015. A fast unified algorithm for solving group-lasso penalize learning problems. Stat. Comput. 25, 1129–1141. http://dx.doi.org/10.1007/s11222-014-9498-5.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 49–67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x.

Yuan, H., Paskov, I., Paskov, H., González, A., Leslie, C., 2016. Multitask learning improves prediction of cancer drug sensitivity. Sci. Rep. 6, 1. http://dx.doi.org/10.1038/srep31619.