KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Dan J., Geirnaert S., Bertrand A. (2022), **Grouped variable selection for generalized eigenvalue problems** *Signal Processing*
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://doi.org/10.1016/j.sigpro.2022.108476
Journal homepage	https://www.sciencedirect.com/journal/signal-processing/
Author contact	jonathan.dan@esat.kuleuven.be
Abstract	
IR	

(article begins on next page)

Grouped Variable Selection for Generalized Eigenvalue Problems

Jonathan Dan[a,b,1,2,*], Simon Geirnaert[a,c,1,2], Alexander Bertrand[a,2]

[a] *KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium*
[b] *Byteflies, Borsbeeksebrug 22, 2600 Berchem, Belgium*
[c] *KU Leuven, Department of Neurosciences, Research Group ExpORL, Herestraat 49 box 721, 3000 Leuven, Belgium*

Abstract

Many problems require the selection of a subset of variables from a full set of optimization variables. The computational complexity of an exhaustive search over all possible subsets of variables is, however, prohibitively expensive, necessitating more efficient but potentially suboptimal search strategies. We focus on sparse variable selection for generalized Rayleigh quotient optimization and generalized eigenvalue problems. Such problems often arise in the signal processing field, e.g., in the design of optimal data-driven filters. We extend and generalize existing work on convex optimization-based variable selection using semidefinite relaxations toward group-sparse variable selection using the $\ell_{1,\infty}$-norm. This group-sparsity allows, for instance, to perform sensor selection for spatio-temporal (instead of purely spatial) filters, and to select variables based on multiple generalized eigenvectors instead of only the dominant one. Furthermore, we extensively compare our method to state-of-the-art methods for sensor selection for spatio-temporal filter design in a simulated sensor network setting. The results show both the proposed algorithm and backward greedy selection method best approximate the exhaustive solution. However, the backward greedy selection has more specific failure cases, in particular for ill-conditioned covariance matrices. As such, the proposed algorithm is the most robust currently available method for group-sparse variable selection in generalized eigenvalue problems.

Keywords: convex optimization, variable selection, sensor selection, generalized Rayleigh quotient, generalized eigenvalue decomposition, group sparsity

[*] Corresponding author
Email addresses: jonathan.dan@esat.kuleuven.be (Jonathan Dan), simon.geirnaert@esat.kuleuven.be (Simon Geirnaert), alexander.bertrand@esat.kuleuven.be (Alexander Bertrand)

[1] These authors contributed equally.

[2] J. D., S. G., and A. B. are also with Leuven.AI - KU Leuven Institute for AI, 3000, Leuven, Belgium.

1. Introduction

Variable selection is an important problem occurring in many mathematical engineering fields. Its goal is to select the subset of variables - often corresponding to specific sensor signals or features thereof - that have the largest impact on the optimization of a specific objective function. Such methods are often used, e.g., to identify the most relevant sensor nodes in a sensor network, or to find the optimal positions to place sensors in a predefined grid [1]. These sensor selection problems arise in many signal processing-related fields, including telecommunication, where antenna placement is critical to the good functioning of a communication network [2, 3, 4, 5, 6, 7, 8, 9], biomedical sensor arrays, e.g., in the context of electro-encephalography (EEG) channel selection or optimal positioning of wearable sensors [10, 11, 12, 13], or wireless acoustic sensor networks, where a microphone subset needs to be selected [14, 15]. The number of sensors is typically constrained by practical factors such as fabrication cost, bandwidth, or physical setup limitations, necessitating an appropriate selection of a limited number of sensors and their location.

In many signal processing applications, the objective function can be written as a generalized Rayleigh quotient (GRQ), which corresponds to solving a generalized eigenvalue decomposition (GEVD). Such GRQ- or GEVD-based objectives are encountered in various beamformer or filter design problems, for example, to maximize the signal-to-noise ratio (SNR) [16, 17, 4, 13], or to maximize discriminative properties of the output signals of a filterbank, e.g., in biomedical sensor arrays [18, 19]. In these contexts, variable/sensor selection helps to reduce the computational complexity and power requirements of processing pipelines, to reduce the risk of overfitting of models, and to improve the overall setup.

In this paper, we focus on grouped variable selection in GRQ/GEVD problems, where the goal is to select a subset of predefined *groups* of variables. For illustrative purposes, but without loss of generality, we will introduce the problem in the context of sensor selection for data-driven spatio-temporal filter design. In sensor networks, an intuitive grouping of the optimization variables is based on the finite impulse response (FIR) filter tap weights in each sensor. However, various (other) types of groupings exist, such as a grouped selection across different filterbands or output filters. Note that all presented methods are besides sensor networks applicable to any other application containing (group-sparse) variable selection for GRQ optimization and GEVD problems.

In sensor selection, the goal is to identify the optimal subset of M out of C available sensors where the choice of M typically leads to a tradeoff between the optimization objective and satisfying some practical constraints. The exhaustive evaluation of all possible sensor combinations is a computationally costly operation. The selection of M out of C sensors is of combinatorial complexity $\frac{C!}{M!\,(C-M)!}$, where each evaluation requires a new GEVD computation, which in itself has a computational cost of $\mathcal{O}(M^3)$. Therefore, computationally efficient methods are required to solve the sensor selection problem. Two popular heuristic methods are found in the greedy forward selection (FS) and backward elimination (BE) algorithms [10, 20], which are easily applied to many selection problems, including the GEVD problem. However, their greedy nature strongly reduces the combinatorial exploration space, which can result in a highly suboptimal selection. Other approaches take the specific problem structure into account and combine optimization of the objective (e.g., the GRQ) with finding a sparse set of sensors. A specific subclass among these optimization-based approaches relaxes the sensor selection problem to a *convex* optimization problem, which can be solved with off-the-shelf convex optimization solvers [21]. More specifically for the GEVD problem, [17] used the sparsity promoting ℓ_1-norm for a purely spatial beamformer (see Section 3.5), which was extended in [2] to a spatio-temporal beamformer using the $\ell_{1,\infty}$-norm as a group-sparse regularizer, albeit in a suboptimal manner (as we will show in Section 3.4). Other approaches in radar beamforming employed the $\ell_{1,2}$-norm as a group-sparse regularizer in combination with successive convex approximation [7, 8]. Furthermore, [2, 7, 8] only cover the case of a single output filter (multiple-input single-output (MISO) filtering), i.e., a single generalized eigenvector is computed, while several GEVD-based signal processing techniques, such as the common spatial patterns (CSP) filterbank, require the extraction of multiple eigenvectors (multiple-input multiple-output (MIMO) filtering).

The main contributions of this work are as follows:

- We extend the GRQ/GEVD sensor selection for purely spatial filtering in [17] to spatio-*temporal* filtering borrowing techniques from [16]. This necessitates the use of a *group-sparse* regularizer. When a sensor is eliminated, all corresponding filter lags should be put to zero.

- We add the possibility to take multiple filters (i.e., multiple generalized eigenvectors) into account (MIMO), whereas previous work only focused on the dominant generalized eigenvector (MISO). This requires consistent removal or zeroing of the filter coefficients corresponding

to an eliminated sensor across *all* filters. This approach can be employed in various other applications, where the notion of a shared selection exists.

- We provide an in-depth and statistical comparison of the proposed method with other state-of-the-art sensor selection methods in GEVD problems, which is largely missing in the aforementioned prior art.

The paper is structured as follows. First, the sensor selection problem for GEVD and GRQ optimization is introduced in Section 2. Next, the convex optimization-based group-sparse sensor selection is explained in Section 3. We then thoroughly compare the proposed method with other (benchmark) sensor selection methods on simulated data in Section 4. In Section 5, we provide an example of the developed method applied on real-world data, in the context of mobile epileptic seizure monitoring. Finally, conclusions are drawn in Section 6.

1.1. Notation

Scalars, vectors, and matrices are denoted by a lowercase (x), bold lowercase ($\mathbf{x}$), and bold uppercase letter ($\mathbf{X}$). The element of matrix $\mathbf{X}$ on the ith row and j^{th} column is given by x_{ij}. $\mathbf{X}^{\text{T}}$ denotes the transpose of a matrix $\mathbf{X}$ and $\text{Tr}\,(\mathbf{X})$ denotes the trace of $\mathbf{X}$. The $N \times N$ identity matrix is denoted by $\mathbf{I}_N$, while $\mathbf{0}_N$ denotes an $N \times N$ matrix with zeros. The ℓ_∞-norm of a vector (i.e., the maximal absolute value) is written as $||\mathbf{x}||_\infty$, the ℓ_1-norm of a vector (i.e., the sum of the absolute values) as $||\mathbf{x}||_1$, the ℓ_2-norm of a vector (i.e., the square root of the sum of squared elements) as $||\mathbf{x}||_2$, and the max-norm of a matrix (i.e., the maximal absolute value across all elements) as $||\mathbf{X}||_{\max}$. $\mathbf{X} \succcurlyeq 0$ denotes that $\mathbf{X}$ is a positive semidefinite matrix. The Kronecker-delta is written as δ_{ij} (i.e., $\delta_{ij} = 0$ if $i \neq j$; $\delta_{ij} = 1$ if $i = j$). Finally, the Kronecker-product of matrices $\mathbf{X} \in \mathbb{R}^{I_x \times J_x}$ and $\mathbf{Y} \in \mathbb{R}^{I_y \times J_y}$ is defined as:

$$\mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & \cdots & x_{1J_x}\mathbf{Y} \\ \vdots & \ddots & \vdots \\ x_{I_x1}\mathbf{Y} & \cdots & x_{I_xJ_x}\mathbf{Y} \end{bmatrix} \in \mathbb{R}^{I_xI_y \times J_xJ_y}.$$

2. Sensor selection for GEVD problems

Consider a setting with C sensors and two stationary zero-mean multi-sensor signals $\mathbf{x}_1(t) \in \mathbb{R}^{CL}$ and $\mathbf{x}_2(t) \in \mathbb{R}^{CL}$, where t denotes the sample (time) index and L denotes the group size as explained

below. $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ could represent the C sensor signals measured during two different states (e.g., EEG during movement of the left arm and movement of the right arm [19]), or they could represent two signal components that are both simultaneously present in the sensor signals (e.g., target signal and noise). We assume that the entries of $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are grouped in blocks of L entries, each group corresponding to a single sensor. For example, a group could consist of L frequency subbands or other features extracted from a single-sensor signal. For illustrative purposes, but without loss of generality, we focus here on spatio-temporal filter design, in which case the L entries of a group correspond to a delay line of length L. In this case, the vector $\mathbf{x}_1(t) \in \mathbb{R}^{CL}$ can be represented as $\mathbf{x}_1(t) = \begin{bmatrix} \underline{\mathbf{x}}_{1,1}(t)^{\mathrm{T}} & \underline{\mathbf{x}}_{1,2}(t)^{\mathrm{T}} & \cdots & \underline{\mathbf{x}}_{1,C}(t)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ where $\underline{\mathbf{x}}_{1,c}(t) = \begin{bmatrix} x_{1,c}(t) & x_{1,c}(t-1) & \cdots & x_{1,c}(t-L+1) \end{bmatrix}^{\mathrm{T}}$ represents the causal FIR filter taps corresponding to the cth sensor (similarly for $\mathbf{x}_2(t)$).

The goal is to find a spatio-temporal filter represented by $\mathbf{w} \in \mathbb{R}^{CL}$ which optimally discriminates between the two signals $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$. Optimal discrimination corresponds to maximizing the energy of the output signal $y_1(t) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_1(t)$, while minimizing the energy of the output signal $y_2(t) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_2(t)$. The optimal $\mathbf{w}$ is thus found by maximizing:

$$\max_{\mathbf{w}\in\mathbb{R}^{CL}} \quad \frac{\mathbb{E}\{(\mathbf{w}^{\mathrm{T}}\mathbf{x}_1(t))^2\}}{\mathbb{E}\{(\mathbf{w}^{\mathrm{T}}\mathbf{x}_2(t))^2\}} = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{R}_1\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{R}_2\mathbf{w}}, \tag{1}$$

where $\mathbf{R}_1 = \mathbb{E}\{\mathbf{x}_1(t)\mathbf{x}_1(t)^{\mathrm{T}}\} \in \mathbb{R}^{CL\times CL}$ and $\mathbf{R}_2 = \mathbb{E}\{\mathbf{x}_2(t)\mathbf{x}_2(t)^{\mathrm{T}}\} \in \mathbb{R}^{CL\times CL}$ are the corresponding covariance matrices. Assuming ergodicity and given T samples of the signals $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$, these covariance matrices can be estimated as $\mathbf{R}_1 = \mathbb{E}\{\mathbf{x}_1(t)\mathbf{x}_1(t)^{\mathrm{T}}\} \approx \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_1(t)\mathbf{x}_1(t)^{\mathrm{T}}$ and similarly for $\mathbf{R}_2$. The problem in (1) is known as a generalized Rayleigh quotient (GRQ) optimization. In the case where $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ represent the target signal and noise components, respectively, Equation (1) implies a maximization of the signal-to-noise ratio, resulting in a so-called max-SNR filter [22]. In max-SNR filtering, the covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ thus correspond to the spatio-temporal covariance matrices related to the target signal and the noise, respectively. In the CSP framework [19], these covariance matrices correspond to the two signal classes that have to be discriminated (e.g., left versus right hand movement).

Because of the scale-invariance of $\mathbf{w}$ in (1), we can arbitrarily set the output power depicted in the denominator to $\mathbf{w}^{\mathrm{T}}\mathbf{R}_2\mathbf{w} = 1$. Using the method of Lagrange multipliers to solve (1) then leads to a GEVD [23]:

$$\mathbf{R}_1\mathbf{w} = \lambda\mathbf{R}_2\mathbf{w}.$$

The optimal filter $\mathbf{w}$ corresponds to the generalized eigenvector (GEVc) corresponding to the largest generalized eigenvalue (GEVl).

In various applications, the GRQ optimization of (1) for MISO filtering is generalized to MIMO filtering, i.e., multiple output filters. In this MIMO case, the goal is to find a filterbank of K spatio-temporal filters $\mathbf{W} \in \mathbb{R}^{CL \times K}$ for which the sum of the energies of the multiple output signals is maximally discriminative:

$$\begin{aligned} \max_{\mathbf{W} \in \mathbb{R}^{CL \times K}} \quad & \frac{\mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}} \mathbf{R}_1 \mathbf{W}\right)}{\mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}} \mathbf{R}_2 \mathbf{W}\right)} \\ \text{s.t.} \quad & \mathbf{W}^{\mathrm{T}} \mathbf{R}_2 \mathbf{W} = \mathbf{I}_K, \end{aligned} \tag{2}$$

with $\mathrm{Tr}\left(\cdot\right)$ denoting the trace operator and $\mathbf{I}_K$ the $K \times K$ identity matrix. The constraint in (2) ensures that the K output channels are orthogonal to each other with respect to the signal component $\mathbf{x}_2(t)$. This constraint is added to obtain K different filters. By plugging this constraint in the cost function in (2), we obtain:

$$\begin{aligned} \max_{\mathbf{W} \in \mathbb{R}^{CL \times K}} \quad & \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}} \mathbf{R}_1 \mathbf{W}\right) \\ \text{s.t.} \quad & \mathbf{W}^{\mathrm{T}} \mathbf{R}_2 \mathbf{W} = \mathbf{I}_K. \end{aligned} \tag{3}$$

The solution of (3) is now given by taking the K GEVcs corresponding to the K largest GEVls:

$$\mathbf{R}_1 \mathbf{W} = \mathbf{R}_2 \mathbf{W} \mathbf{\Lambda}. \tag{4}$$

This generalization to MIMO filters ($K > 1$) is crucial in classification tasks and discriminative analysis as in the CSP framework or in Fisher's discriminant analysis, where the data are projected into a K-dimensional feature space instead of a one-dimensional space. The number of output filters K then introduces a tradeoff between how much information from the original data is preserved and the GRQ (2), which becomes smaller (worse) for larger K.

Our goal is to find the optimal subset of $M \leq C$ out of C sensors, with $M \geq K$, for which the ratio of traces in (2) is maximal. Note that eliminating a sensor means that all time lags corresponding to that sensor need to be zero. Furthermore, the selected sensors must be consistent across all K filters (i.e., columns of $\mathbf{W}$) to be able to physically select only a few sensors. That is why a *group-sparse* sensor selection is required, i.e., the filter weights are grouped per sensor across time lags and filters, and whole groups are put to zero (i.e., eliminated) rather than the individual elements in a group. In the next section, we present a convex optimization-based approach for this sensor selection problem.

3. Optimal group-sparse sensor selection

In this section, we generalize the optimal sensor selection and array design method of [17], which focuses on GRQ optimization and GEVD for purely spatial filtering (i.e., $L = 1$) and for MISO filtering (i.e., $K = 1$), to spatio-temporal filtering and MIMO filtering. Our derivation is based on a similar $\ell_{1,\infty}$-norm regularization technique as proposed in [16] for multicast beamforming and antenna selection. It is noted that during the consolidation of this work, a similar idea to introduce group-sparsity in GEVD problems was published in [2] in the meantime, independently from our work. The work in [2] establishes the $L > 1$ case, yet without generalizing to the $K > 1$ case as also targeted here. Furthermore, our proposed generalization differs from [2] on another crucial aspect, which will be pointed out throughout the derivation (see Section 3.4), and which makes that [2] can not be treated as a special case of our proposed general framework. In Section 4, we will also empirically compare with [2] for the $K = 1$ setting and demonstrate the superiority of our generalization.

Before pursuing group sparsity in (3), let us first vectorize $\mathbf{W} \in \mathbb{R}^{CL\times K}$ as $\mathbf{w} \in \mathbb{R}^{CLK}$, with $\mathbf{w}_k, k \in \{1, \ldots, K\}$, the kth spatio-temporal filter:

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_K \end{bmatrix}, \mathbf{w}_k = \begin{bmatrix} \mathbf{w}_{k,1} \\ \vdots \\ \mathbf{w}_{k,C} \end{bmatrix}, \text{ and } \mathbf{w}_{k,c} = \begin{bmatrix} w_{k,c,1} \\ \vdots \\ w_{k,c,L} \end{bmatrix}. \tag{5}$$

The optimization problem in (3) then becomes:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^{CLK}} \quad & \mathbf{w}^{\mathrm{T}} \left(\mathbf{I}_K \otimes \mathbf{R}_2\right) \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}_k^{\mathrm{T}} \mathbf{R}_1 \mathbf{w}_{k'} = \delta_{kk'}, \quad \forall k, k' \in \{1, \ldots, K\}, \end{aligned} \tag{6}$$

with $\otimes$ the Kronecker-product and $\delta_{kk'}$ the Kronecker-delta (i.e., $\delta_{kk'} = 0$, $\forall k \neq k'$; $\delta_{kk'} = 1$, $\forall k = k'$). Notice that we changed the problem in (3) to a minimization problem to accommodate for an easy introduction of the sparse regularization term. It can be shown that the solution of (3) and (6) are the same up to an arbitrary scaling on each $\mathbf{w}_k$, which is irrelevant as generalized eigenvectors are defined up to a scaling. Using the filter-selector matrix $\mathbf{S}_k \in \mathbb{R}^{CL\times CLK}$, where the subscript indicates the selected coefficients of $\mathbf{w}$:

$$\mathbf{S}_k = \begin{matrix} \begin{bmatrix} \mathbf{0}_{CL} & \ldots & \mathbf{0}_{CL} & \mathbf{I}_{CL} & \mathbf{0}_{CL} & \ldots & \mathbf{0}_{CL} \end{bmatrix} \\ \begin{matrix} 1 & & \text{k-1} & \text{k} & \text{k+1} & & \text{K} \end{matrix} \end{matrix}$$

with an identity matrix on the kth position to select the kth filter $\mathbf{w}_k$ from $\mathbf{w}$, i.e., $\mathbf{S}_k\mathbf{w} = \mathbf{w}_k$, (6) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^{CLK}} \quad & \mathbf{w}^{\mathrm{T}} \left(\mathbf{I}_K \otimes \mathbf{R}_2\right) \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathrm{T}}\mathbf{S}_k^{\mathrm{T}}\mathbf{R}_1\mathbf{S}_{k'}\mathbf{w} = \delta_{kk'}, \quad \forall k, k' \in \{1, \ldots, K\}. \end{aligned} \tag{7}$$

3.1. Group-sparsity promoting regularization

The goal is now to introduce sparsity in (6) on the *sensor* level. This sparsity on the sensor level corresponds to a group-sparse constraint on (6), as all lags of all output filters corresponding to a particular sensor need to be set to zero. Therefore, as in [16], we deploy the convex sparsity-promoting $\ell_{1,\infty}$-norm as a proxy for the optimal but non-convex ℓ_0-norm as a regularization term in (6).

To simplify the notations in the remainder of the derivations, we define the permutation matrix $\mathbf{P} \in \mathbb{R}^{CLK\times CLK}$ that permutes the elements of $\mathbf{w}$ such that they are first ordered by sensor and then by filter and lags (instead of first by filter as in (5)), resulting in $\tilde{\mathbf{w}}$:

$$\tilde{\mathbf{w}} = \mathbf{P}\mathbf{w} = \begin{bmatrix} \tilde{\mathbf{w}}_1 \\ \vdots \\ \tilde{\mathbf{w}}_C \end{bmatrix} \text{ and } \tilde{\mathbf{w}}_c = \begin{bmatrix} \mathbf{w}_{1,c} \\ \vdots \\ \mathbf{w}_{K,c} \end{bmatrix}. \tag{8}$$

Using this notation, the $\ell_{1,\infty}$-norm on the sensor level is defined as:

$$||\mathbf{w}||_{1,\infty} = \sum_{c=1}^{C} ||\tilde{\mathbf{w}}_c||_\infty = \sum_{c=1}^{C} \max_{k=1,\ldots,K} ||\mathbf{w}_{k,c}||_\infty \,, \tag{9}$$

where $||\tilde{\mathbf{w}}_c||_\infty$ corresponds to the maximal absolute value across all lags and filters corresponding to sensor c. As the ℓ_1-norm induces sparsity, while the ℓ_∞-norm is only zero when all elements are zero, the $\ell_{1,\infty}$-norm can be used to put groups of coefficients across lags and filters corresponding to one sensor to zero. Furthermore, in [16], it is also shown that any sparsity-inducing norm can be replaced with the squared norm without changing the regularization properties of the problem. Therefore, the sensor selection problem with the group-sparse regularization term becomes:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^{CLK}} \quad & \mathbf{w}^{\mathrm{T}} \left(\mathbf{I}_K \otimes \mathbf{R}_2\right) \mathbf{w} + \mu \left(\sum_{c=1}^{C} ||\tilde{\mathbf{w}}_c||_\infty\right)^2 \\ \text{s.t.} \quad & \mathbf{w}^{\mathrm{T}}\mathbf{S}_k^{\mathrm{T}}\mathbf{R}_1\mathbf{S}_{k'}\mathbf{w} = \delta_{kk'}, \quad \forall k, k' \in \{1, \ldots, K\}, \end{aligned} \tag{10}$$

where the regularization parameter μ can be used to control the solution's sparsity and thus the number of sensors selected. Note that this is not yet a convex optimization problem due to the quadratic equality constraints.

3.2. Semidefinite formulation and relaxation

To transform (10) into a convex semidefinite problem (SDP), we are using the following trick, as suggested in [24, 16, 17]:

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\left(\mathbf{I}_K \otimes \mathbf{R}_2\right)\mathbf{w} &= \mathrm{Tr}\left(\mathbf{w}^{\mathrm{T}}\left(\mathbf{I}_K \otimes \mathbf{R}_2\right)\mathbf{w}\right)\\ &= \mathrm{Tr}\left(\left(\mathbf{I}_K \otimes \mathbf{R}_2\right)\mathbf{w}\mathbf{w}^{\mathrm{T}}\right)\\ &= \mathrm{Tr}\left(\left(\mathbf{I}_K \otimes \mathbf{R}_2\right)\mathbf{V}\right),\end{aligned}$$

where the second equality holds because of the cyclic property of the trace. Per definition, $\mathbf{V} = \mathbf{w}\mathbf{w}^{\mathrm{T}} \in \mathbb{R}^{CLK\times CLK}$ is thus a rank-1 positive semidefinite matrix. Similarly, the equality constraints can be reformulated as:

$$\mathrm{Tr}\left(\mathbf{R}_1\mathbf{S}_{k'}\mathbf{V}\mathbf{S}_k^{\mathrm{T}}\right) = \delta_{kk'}, \forall k, k' \in \{1,\ldots,K\}.$$

Using the following definition of $\tilde{\mathbf{V}}$:

$$\tilde{\mathbf{V}} = \tilde{\mathbf{w}}\tilde{\mathbf{w}}^{\mathrm{T}} = \mathbf{P}\mathbf{V}\mathbf{P}^{\mathrm{T}} = \begin{bmatrix}\tilde{\mathbf{V}}_{11} & \cdots & \tilde{\mathbf{V}}_{1C}\\ \vdots & \ddots & \vdots\\ \tilde{\mathbf{V}}_{C1} & \cdots & \tilde{\mathbf{V}}_{CC}\end{bmatrix},$$

the group-sparse regularization term in (10) can be reformulated similarly to [16]:

$$\begin{aligned}\left(\sum_{c=1}^{C}||\tilde{\mathbf{w}}_c||_\infty\right)^2 &= \sum_{c_1=1}^{C}\sum_{c_2=1}^{C}||\tilde{\mathbf{w}}_{c_1}||_\infty\,||\tilde{\mathbf{w}}_{c_2}||_\infty\\ &= \sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\left|\left|\tilde{\mathbf{V}}_{c_1c_2}\right|\right|_{\max}\\ &= \mathrm{Tr}\left(\mathbf{1}_C\mathbf{1}_C^{\mathrm{T}}\mathbf{U}\right),\end{aligned} \tag{11}$$

where the max-norm $||\mathbf{A}||_{\max} = \max_{i,j}|a_{ij}|$ is the elementwise maximum over a matrix, $\mathbf{1}_C \in \mathbb{R}^C$ denotes an all-ones vector of length C, and where $\mathbf{U} \in \mathbb{R}^{C\times C}$ is equal to:

$$\mathbf{U} = \begin{bmatrix}\left|\left|\tilde{\mathbf{V}}_{11}\right|\right|_{\max} & \cdots & \left|\left|\tilde{\mathbf{V}}_{1C}\right|\right|_{\max}\\ \vdots & \ddots & \vdots\\ \left|\left|\tilde{\mathbf{V}}_{C1}\right|\right|_{\max} & \cdots & \left|\left|\tilde{\mathbf{V}}_{CC}\right|\right|_{\max}\end{bmatrix}. \tag{12}$$

Using the definition of $\mathbf{U}$ in (12), we finally obtain:

$$\begin{aligned} \min_{\substack{\mathbf{V}\in\mathbb{R}^{CLK\times CLK},\\ \mathbf{U}\in\mathbb{R}^{C\times C}}} \quad & \mathrm{Tr}\left((\mathbf{I}_K \otimes \mathbf{R}_2)\,\mathbf{V}\right) + \mu \mathrm{Tr}\left(\mathbf{1}_C \mathbf{1}_C^{\mathrm{T}} \mathbf{U}\right) \\ \text{s.t.} \quad & \mathrm{Tr}\left(\mathbf{R}_1 \mathbf{S}_{k'} \mathbf{V} \mathbf{S}_k^{\mathrm{T}}\right) = \delta_{kk'}, \forall k, k' \in \{1, \ldots, K\}, \\ & \mathbf{U} \geq |\mathbf{S}_{k,l} \mathbf{V} \mathbf{S}_{k',l'}^{\mathrm{T}}|, \forall k, k' \in \{1, \ldots, K\}, \\ & \qquad\qquad \text{and } \forall l, l' \in \{1, \ldots, L\}, \\ & \mathbf{V} \succcurlyeq 0, \mathrm{rank}(\mathbf{V}) = 1, \end{aligned} \tag{13}$$

with the selector-matrix $\mathbf{S}_{k,l} \in \mathbb{R}^{C\times CLK}$ selecting all coefficients across C sensors for a particular filter k and lag l. The second constraint is an element-wise inequality, which ensures that each element of $\mathbf{U}$ (i.e., for each pair of sensors) is larger than the corresponding element for the corresponding pair of sensors across all filters and lags (expressed by the $\forall$ over the filter and lag indices), and thus implements the max-norm operation. The last two constraints ensure the equivalence between $\mathbf{V}$ and $\mathbf{w}\mathbf{w}^{\mathrm{T}}$.

However, (13) is still not a convex optimization problem due to the rank-1 constraint. Therefore, we approximate (13) by relaxing the rank constraint, which is a technique known as semidefinite relaxation (SDR) and results in an SDP [24]:

$$\begin{aligned} \min_{\substack{\mathbf{V}\in\mathbb{R}^{CLK\times CLK},\\ \mathbf{U}\in\mathbb{R}^{C\times C}}} \quad & \mathrm{Tr}\left((\mathbf{I}_K \otimes \mathbf{R}_2)\,\mathbf{V}\right) + \mu \mathrm{Tr}\left(\mathbf{1}_C \mathbf{1}_C^{\mathrm{T}} \mathbf{U}\right) \\ \text{s.t.} \quad & \mathrm{Tr}\left(\mathbf{R}_1 \mathbf{S}_{k'} \mathbf{V} \mathbf{S}_k^{\mathrm{T}}\right) = \delta_{kk'}, \forall k, k' \in \{1, \ldots, K\}, \\ & \mathbf{U} \geq |\mathbf{S}_{k,l} \mathbf{V} \mathbf{S}_{k',l'}^{\mathrm{T}}|, \forall k, k' \in \{1, \ldots, K\}, \\ & \qquad\qquad \text{and } \forall l, l' \in \{1, \ldots, L\}, \\ & \mathbf{V} \succcurlyeq 0. \end{aligned} \tag{14}$$

This SDR results in practice in a good approximation of the underlying rank-1 solution, potentially using a post-hoc rank-1 approximation of the solution. Note, however, that we are here not interested in the optimal filter coefficients themselves, but only in the selected sensors, which can be retrieved as the non-zero elements of the diagonal of $\mathbf{U}$. Typically, the GEVD problem in (4) is afterward recomputed given the selected sensors from (14).

3.3. Iterative reweighting and algorithm

Similarly to [16, 17], the all-ones matrix $\mathbf{1}_C\mathbf{1}_C^{\mathrm{T}}$ in (14) can be replaced with a reweighting matrix $\mathbf{B}^{(i)} \in \mathbb{R}^{C\times C}$ to implement iteratively reweighted ℓ_1-norm regularization [25]. The optimization problem in (14) can then be iteratively solved by updating $\mathbf{B}^{(i)}$ as:

$$B_{c_1c_2}^{(i+1)} = \frac{1}{U_{c_1c_2}^{(i)} + \epsilon}. \tag{15}$$

This iteratively reweighted ℓ_1-norm regularization procedure compensates for the inherent magnitude-dependency of the ℓ_1-norm. Using the ℓ_1-norm as a proxy for the ℓ_0-norm introduces a too large penalty on the elements that have a large magnitude, while it is only relevant to know whether an element is equal to zero or not [25]. The parameter ϵ, which is set as 10% of the standard deviation of the elements of $\mathbf{U}$ without sensor selection (as suggested in [25] and which can be easily computed using the GEVD in (4)), avoids division by zero. Initially, $\mathbf{B}^{(1)}$ is set to $\mathbf{1}_C\mathbf{1}_C^{\mathrm{T}}$, i.e., (14) is solved. This iterative reweighting procedure generally converges after a few iterations.

To find the optimal set of a specific number M of sensors, a binary search on the hyperparameter μ of (14) can be performed. Once the optimal set of sensors is found, the corresponding spatio-temporal filters $\mathbf{W}$ can be computed by taking the K GEVcs corresponding to the K largest GEVls of the GEVD in (4), using the reduced covariance matrices $\mathbf{R}_{1,2}^{(\mathrm{red})} \in \mathbb{R}^{ML\times ML}$, i.e., by selecting the rows and columns corresponding to the selected sensors. The complete algorithm, which is referred to as GS-$\ell_{1,\infty}$ (GS for group-sparse) in the remainder of the paper, is summarized in Algorithm 1[3]. The convex optimization problem in (14) is solved using the CVX toolbox [26, 27] and MOSEK solver [28].

Remark: It is noted that this algorithm can be easily extended to complex filter coefficients (as often found in beamforming), as the objective function of (14) (with the transpose replaced by Hermitian transpose) is a real-valued function, even though it is function of complex variables, while the inequality constraints are also real. This is due to the use of the trace operator in combination with Hermitian (conjugate symmetric) complex-valued matrices.

[3] An open-source toolbox with the MATLAB implementation of this group-sparse sensor selection algorithm can be found online on `https://github.com/AlexanderBertrandLab/gsl1infSensorSelection`.

Algorithm 1 Group-sparse sensor selection for GEVD (GS-$\ell_{1,\infty}$)

Input:

- $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{R}^{CL \times CL}$: to-be-discriminated covariance matrices
- M: number of sensors to be selected
- K: number of filters/GEVcs to take into account
- $\mu_{\text{LB}}, \mu_{\text{UB}}$: lower and upper bounds of the binary search
- $i_{\max}$: maximal number of reweighting iterations

Output: Optimal subset of M sensors and corresponding filters/GEVCs $\mathbf{W} \in \mathbb{R}^{ML \times K}$

1: Define ϵ as 10% of the standard deviation of the elements of $\mathbf{U}$ corresponding to the solution with all sensors (as can be computed from the GEVD in (4)) and the tolerance τ as 10% of the minimum across the diagonal of $\mathbf{U}$ corresponding to the solution with all sensors

2: **while** Not M sensors selected **do**

3: Initialize $\mathbf{B}^{(1)} = \mathbf{1}_C \mathbf{1}_C^{\mathrm{T}}$

4: $\mu = \mu_{\text{LB}} + \frac{\mu_{\text{UB}} - \mu_{\text{LB}}}{2}$

5: **while** $\mathbf{U}$ changes and $i \leq i_{\max}$ **do**

6: Solve

$$\begin{aligned} \min_{\mathbf{V} \in \mathbb{R}^{CLK \times CLK}, \mathbf{U} \in \mathbb{R}^{C \times C}} \quad & \operatorname{Tr}\left(\left(\mathbf{I}_K \otimes \mathbf{R}_2 \right) \mathbf{V} \right) + \mu \operatorname{Tr}\left(\mathbf{B}^{(i)} \mathbf{U} \right) \\ \text{s.t.} \quad & \operatorname{Tr}\left(\mathbf{R}_1 \mathbf{S}_{k'} \mathbf{V} \mathbf{S}_k^{\mathrm{T}} \right) = \delta_{kk'}, \quad \forall k, k' \in \{1, \ldots, K\}, \\ & \mathbf{U} \geq |\mathbf{S}_{k,l} \mathbf{V} \mathbf{S}_{k',l'}^{\mathrm{T}}|, \qquad \forall k, k' \in \{1, \ldots, K\} \text{ and } \forall l, l' \in \{1, \ldots, L\}, \\ & \mathbf{V} \succcurlyeq 0. \end{aligned}$$

7: Update counter i

8: Update $\mathbf{B}^{(i+1)}$ as:

$$B_{c_1 c_2}^{(i+1)} = \frac{1}{U_{c_1 c_2}^{(i)} + \epsilon}$$

9: **end while**

Determine the (number of) sensors $\hat{M}$ selected by comparing the diagonal of $\mathbf{U}$ with the tolerance τ:
  for $c = 1$ to C **do**
    if $U_{cc} > \tau$ **then** c^{th} sensor selected
    else if $U_{cc} < \tau$ **then** c^{th} sensor eliminated
  end for
  Update the regularization parameter bounds as:
  if $\hat{M} > M$ **then** $\mu_{\text{LB}} = \mu$
  else if $\hat{M} < M$ **then** $\mu_{\text{UB}} = \mu$
end while
Compute optimal filters as the K GEVcs corresponding to the K largest GEVls of the GEVD problem with reduced covariance matrices $\mathbf{R}_{1,2}^{(\text{red})} \in \mathbb{R}^{ML \times ML}$:

$$\mathbf{R}_1^{(\text{red})}\mathbf{W} = \mathbf{R}_2^{(\text{red})}\mathbf{W}\mathbf{\Lambda}$$

3.4. Special case I: MISO filtering

When taking only one filter into account for the sensor selection (i.e., $K = 1$; MISO filtering), the SDR problem in (14) becomes:

$$\begin{aligned} \min_{\substack{\mathbf{V}\in\mathbb{R}^{CL\times CL},\\ \mathbf{U}\in\mathbb{R}^{C\times C}}} \quad & \text{Tr}\left(\mathbf{R}_2\mathbf{V}\right) + \mu\text{Tr}\left(\mathbf{1}_C\mathbf{1}_C^{\text{T}}\mathbf{U}\right) \\ \text{s.t.} \quad & \text{Tr}\left(\mathbf{R}_1\mathbf{V}\right) = 1, \\ & \mathbf{U} \geq |\mathbf{S}_l\mathbf{V}\mathbf{S}_{l'}^{\text{T}}|, \forall l, l' \in \{1, \ldots, L\}, \\ & \mathbf{V} \succcurlyeq 0, \end{aligned} \tag{16}$$

with the selector-matrix $\mathbf{S}_l \in \mathbb{R}^{C\times CL}$ selecting all sensor coefficients corresponding to the lth lag. This simplified problem is very similar to the approach proposed in [2], which was independently published during the consolidation of this work. However, the algorithm derived in [2] has a subtle - yet crucial - difference with (16) in the inequality constraint $\mathbf{U} \geq |\mathbf{S}_l\mathbf{V}\mathbf{S}_{l'}^{\text{T}}|, \forall l, l' \in \{1, \ldots, L\}$. In [2], a different inequality was proposed, which only takes the diagonal elements of the different blocks of $\mathbf{V}$ into account, i.e., $\mathbf{U} \geq |\mathbf{S}_l\mathbf{V}\mathbf{S}_l^{\text{T}}|, \forall l \in \{1, \ldots, L\}$, while we also take the off-diagonal elements of each block of $\mathbf{V}$ into account (remember: the blocks of $\mathbf{V}$ correspond to sensors when $K = 1$, the

elements per block to different combinations of lags (see (5))). While leading to fewer inequality constraints and thus resulting in a decreased computational complexity, this relaxation in [2] alters the solution and leads to a suboptimal sensor selection (as empirically shown in Section 4). The reason is that the off-diagonal blocks also appear in the first constraint of (16), resulting in a mismatch between both constraints. In the remainder of the paper, the variant of [2] is dubbed 'GS-$\ell_{1,\infty}$-[2]'.

3.5. Special case II: purely spatial filtering

In case we do not only constrain to MISO filtering ($K = 1$), but also restrict $\mathbf{w} \in \mathbb{R}^C$ to a purely spatial filter (i.e., $L = 1$), (16) is reduced to:

$$\begin{aligned} \min_{\mathbf{V}\in\mathbb{R}^{C\times C},\mathbf{U}\in\mathbb{R}^{C\times C}} \quad & \mathrm{Tr}\left(\mathbf{R}_2\mathbf{V}\right) + \mu\mathrm{Tr}\left(\mathbf{1}_C\mathbf{1}_C^{\mathrm{T}}\mathbf{U}\right) \\ \text{s.t.} \quad & \mathrm{Tr}\left(\mathbf{R}_1\mathbf{V}\right) = 1, \\ & \mathbf{U} \geq |\mathbf{V}|, \\ & \mathbf{V} \succcurlyeq 0, \end{aligned} \tag{17}$$

which is equivalent to the approach proposed in [17].

3.6. Computational complexity

The computational complexity of the proposed method can be computed from the complexity of interior-point method solvers for quadratic problems with quadratic constraints that are relaxed using semidefinite relaxation. In general, such problems with N^2 variables and T quadratic constraints can be solved to an arbitrary small accuracy ϵ with a complexity of $\mathcal{O}\left(\max(N,T)^4 N^{0.5}\log(\frac{1}{\epsilon})\right)$ [24]. This leads to a complexity of $\mathcal{O}\left((CLK)^{4.5}\log(\frac{1}{\epsilon})\right)$ for our proposed GS-$\ell_{1,\infty}$ algorithm (Algorithm 1).

4. Benchmark study

We compare the proposed GS-$\ell_{1,\infty}$ method with other benchmark sensor selection methods on simulated sensor data with known ground-truth[4]. We use the value of the GRQ (2) as the performance metric (higher is better).

[4] We provide an open-source MATLAB implementation of the benchmark study online on `https://github.com/AlexanderBertrandLab/benchmarkStudySensorSelection`.

Besides the exhaustive search, a random search, and the GS-$\ell_{1,\infty}$-[2] method, the proposed GS-$\ell_{1,\infty}$ method is compared with three other sensor selection methods, which are introduced in Section 4.1. For the random search, the final GRQ is the mean over 1000 random selections of sensors for a given problem. The setup of the benchmark study is described in Section 4.2. The aforementioned methods are compared using only one filter (MISO filtering) in Section 4.3 and using multiple filters (MIMO filtering) in Section 4.4 (for those methods that allow for $K > 1$). Finally, we provide a more in-depth comparison of the two best-performing algorithms, namely the proposed GS-$\ell_{1,\infty}$ method and the backward greedy elimination method (see Section 4.1.1) in Section 4.5.

4.1. Benchmark sensor selection methods

In this section, we briefly introduce other algorithms for sensor selection that will be included in the benchmark study.

4.1.1. Greedy sensor selection methods

Greedy sensor selection methods - also dubbed 'wrapper' methods [10] - sequentially select or eliminate those sensors that maximally increase or minimally decrease the objective, respectively. While these greedy approaches are computationally more efficient than the method proposed in Section 3, due to their sequential nature, the greedy mechanism can result in suboptimal selections, as they are stuck with the selected or eliminated sensors from previous steps. The computational complexity of these greedy methods is dominated by the GEVD computation performed at each iteration, which is $\mathcal{O}\left((CLK)^3\right)$ [29]. The greedy selection can be applied in two directions (forward or backward):

Forward selection (FS). The FS method starts from an empty set of sensors and sequentially adds the sensor (i.e., group of KL variables) that maximally increases the objective (2). New sensors are added until M out of C sensors are selected.

Backward elimination (BE). The BE method starts from the full set of sensors and sequentially removes the sensor that minimally decreases the objective (i.e., the objective in (2)) until M out of C sensors are selected. Many variations on the FS and BE method exist, mostly presented in the context of feature selection for classification [20].

4.1.2. The STECS method

We also compare with the spatio-temporal-filtering-based channel selection (STECS) approach proposed for the GEVD problem in [30]. In the STECS method, initially proposed for $K = 1$, the following optimization problem is solved [30] as a regularized proxy for (1):

$$\min_{\mathbf{w}\in\mathbb{R}^{CL}} \quad \mathbf{w}^{\mathrm{T}}\mathbf{R}_2\mathbf{w} + \frac{1}{\mathbf{w}^{\mathrm{T}}\mathbf{R}_1\mathbf{w}} + \mu\,||\mathbf{w}||_{1,2}\,, \tag{18}$$

with the $\ell_{1,2}$-norm, defined as $||\mathbf{w}||_{1,2} = \sum_{c=1}^{C} ||\mathbf{w}_c||_2$, enforcing the group sparsity over different lags. The GRQ in (1) is split into the first two terms of (18) as it removes the scale-invariance of $\mathbf{w}$, while yielding an equivalent solution [30].

The c-th sensor is then selected if $||\mathbf{w}_c||_2$ is larger than a predefined tolerance τ, which is set as 10% of the minimum $\min_{c\in\{1,\ldots,C\}} ||\mathbf{w}_c||_2$ of the solution with all sensors. The optimization problem in (18) is solved using the line-search method proposed in [30]. The same settings as in [30] have been used. Similarly to the proposed method in Section 3.3, we use a binary search method on the regularization parameter μ to obtain the correct number of sensors M. Furthermore, the authors propose to use the selected sensors with (18) for the first filter $\mathbf{w} \in \mathbb{R}^{CL}$ to recompute the solution of $\mathbf{W} \in \mathbb{R}^{CL\times K}$ for multiple filters with (4), which means the selection does not take the full objective (2) into account. Lastly, note that the optimization problem (18) is non-convex, resulting in potential convergence to non-optimal local minima.

Other sensor selection methods for GEVD problems (and in particular in biomedical applications for CSP problems) have been proposed as well [10], for example, based on filter coefficients magnitude (e.g., [31]), or other variants of ℓ_1-norm regularization (e.g., [7, 8, 32, 33]). However, we do not further consider these methods, as they are not designed for group-sparsity and/or the MIMO case, or have been shown to be outperformed by at least one of the aforementioned methods [30].

4.2. Setup

4.2.1. Simulation model

We assume a $\sqrt{C} \times \sqrt{C}$ square grid of C sensors, each of which are measuring a mixture of N_1 source signals to be maximized (i.e., contributing to $\mathbf{x}_1(t)$ and the numerator of (2)) and N_2 source signals to be minimized (i.e., contributing to $\mathbf{x}_2(t)$ and the denominator of (2)) as well as

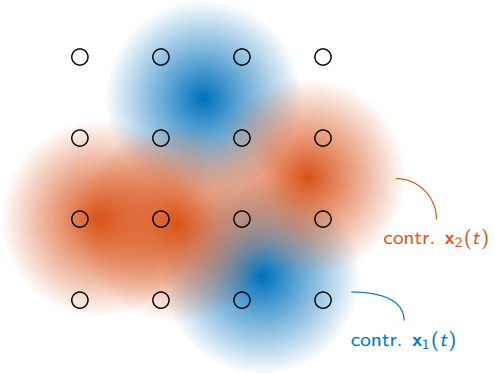

Figure 1: An exemplary generated problem with $C = 16 = 4 \times 4$ sensors, $N_1 = 2$ random signals contributing to $\mathbf{x}_1(t)$, and $N_2 = 3$ random signals contributing to $\mathbf{x}_2(t)$. Each sensor measures a mixture of the underlying sources. The brightness of the color represents the intensity of the signal as perceived by a sensor.

independent sensor noise. This simulated problem resembles point-source models as, for example, found in sensor networks, microphone arrays, neural activity (EEG), and telecommunications. The source signals contributing to $\mathbf{x}_2(t)$ have a power that is approximately 150 times larger than the source signals contributing to $\mathbf{x}_1(t)$. An example is given in Figure 1. In the max-SNR filtering case, one could think of the source signals contributing to $\mathbf{x}_1(t)$ as target signals and source signals contributing to $\mathbf{x}_2(t)$ as noise signals. In that case, the GRQ can be interpreted as an SNR.

Each source signal is a bandpass-filtered white Gaussian signal in a random frequency band between 1 and 9 Hz, sampled at 20 Hz. It originates from a random location within the grid of sensors (drawn from a uniform distribution over the entire area) and propagates with an exponentially decaying amplitude to the sensors. The spread of the exponential decay is set such that the maximal attenuation is equal to a predefined attenuation of 0.5%. Furthermore, a source signal is measured at each sensor with a time delay linear to the distance to that source signal, such that the maximal delay is 100 ms (i.e., 2 samples). The sensor noise at each sensor is white Gaussian noise with twice the maximal attenuation as amplitude.

4.2.2. Monte Carlo runs

For each experiment, i.e., for a given number of sensors C, number of lags L, and number of filters K, 250 (for the $K = 1$ case) and 100 (for the $K > 1$ case) of the random problems in Section 4.2.1 are generated, and the results for each evaluated sensor selection method are averaged over these different problems. For each of these Monte Carlo runs, unless specified otherwise, the number of

	μ_{LB}	μ_{UB}	$i_{\max}$	max. it. binary search
GS-$\ell_{1,\infty}$	10^{-5}	100	15	20
GS-$\ell_{1,\infty}$-[2]	10^{-5}	10^{4}	15	20
STECS	0	10^{16}	/	200

Table 1: The chosen hyperparameters of the different optimization-based sensor selection methods.

signals N_1 and N_2 is randomized between 1 and $2C$.

4.2.3. Hyperparameter choice

Table 1 shows the chosen hyperparameters for the different optimization-based sensor selection methods. The binary search for the GS-$\ell_{1,\infty}$ (Algorithm 1), the GS-$\ell_{1,\infty}$-[2], and the STECS method is aborted if no solution was found after a certain number of iterations. For STECS, this number is taken much larger, which is possible due to its computational efficiency. However, this early stopping criterion leads to a limited number of cases where no solution is found for a certain M. To still produce a meaningful solution in those cases, a random extra sensor is added to the solution obtained for $M-1$ sensors, and the corresponding output GRQ (in dB) is computed. However, when no solution is found for the lowest value of M and the previous solution to still produce a meaningful solution correspondingly fails (because it relies on the solution of the lowest M), the results for all methods for those M for which there is no solution in that particular run are removed.

Furthermore, the hyperparameter μ for the GS-$\ell_{1,\infty}$ (Algorithm 1) and GS-$\ell_{1,\infty}$-[2] algorithm is defined relative to the first target objective part of (16) (i.e., $\text{Tr}\left(\mathbf{R}_2\mathbf{V}\right)$) for the solution with all sensors.

4.2.4. Statistical comparison

To identify statistically significant differences based on hypothesis testing, we use a linear mixed-effect model (LMEM) [34]. Such an LMEM allows the exploitation of all structure in the data by modeling the obtained GRQ as a function of the method while taking the variation due to the different Monte Carlo runs and the effect of a different number of selected sensors M into account as random factors. The following LMEM is chosen based on the Akaike information criterion (AIC),

which takes the model fit and complexity into account:

$$\text{GRQ} \sim 1 + \text{method} + (M|\text{run}).$$

This notation is often used in LMEMs to reflect that the GRQ is modeled with the method as a fixed effect, the number of selected sensors M as random slope (i.e., the GRQ can vary as a function of M independently for each method), and the run as a random intercept. Per fixed term, the estimated regression coefficients (β), standard errors (SE), degrees of freedom (DF), t-value, and p-value are reported. If a significant effect between the different methods is found, we use an additional Tukey-adjusted post-hoc test to assess the pairwise differences between individual methods. The significance level is set to $\alpha = 0.05$. All statistical analyses are performed using the R software package and the `nlme` and `emmeans` packages.

In the statistical hypothesis testing, we limit the number of selected sensors to $\frac{C}{2}$, as we consider this lower half range much more relevant in the context of sensor selection than the upper half range. Typically, one wants to drastically reduce the number of required sensors, i.e., below half of the number of available sensors.

4.3. Comparison in the MISO case ($K = 1$)

First, we evaluate and compare the presented methods in the first special MISO case of Section 3.4, where only one filter (first GEVc) is taken into account, i.e., $K = 1$. In this case, we can also include the comparison with GS-$\ell_{1,\infty}$-[2], which was designed specifically for this case. We choose $C = 25, L = 3$ and look for the optimal sensor selection for M ranging from 2 to 24. Figure 2 shows the output GRQs as a function of M for each separate method (mean over 250 Monte Carlo runs $\pm$ the standard error on the mean). Table 2 shows the outcome of the statistical analysis, for M ranging from 2 to 12 (see Section 4.2.4). All presented methods achieve significantly higher (better) GRQ than random selection but lower (worse) GRQ than the optimal solution obtained through an exhaustive search over all possible combinations (Table 2b).

The greedy sensor selection methods suffer from intrinsic limitations, i.e., they depend on previous choices in their sequential procedure. For example, the FS method starts with a GRQ close to optimal but diverges from the optimal exhaustive solution when M increases, and the other way around for the BE method. However, the FS method achieves overall lower GRQ than the BE method, which is also confirmed by the statistical testing in Table 2b. This could be due

Fixed-effect term	β	**SE**	**DF**	**t-value**	**p-value**
intercept	-5.38	0.79	18729	-6.80	< 0.0001
method = exhaustive$-$BE	2.51	0.07		34.94	< 0.0001
method = exhaustive$-$FS	3.35			46.71	< 0.0001
method = exhaustive$-$STECS	5.48			76.30	< 0.0001
method = exhaustive$-$random	8.73			121.59	< 0.0001
method = exhaustive$-$GS-$\ell_{1,\infty}$-[2]	4.33			60.36	< 0.0001
method = exhaustive$-$GS-$\ell_{1,\infty}$	1.83			25.46	< 0.0001

(a)

	exhaustive	**BE**	**FS**	**STECS**	**random**	**GS-$\ell_{1,\infty}$-[2]**	**GS-$\ell_{1,\infty}$**
exhaustive	/	2.51/$*$	3.53/$*$	5.48/$*$	8.73/$*$	4.33/$*$	1.83/$*$
BE	$-2.51/*$	/	0.85/$*$	2.97/$*$	6.22/$*$	1.83/$*$	$-0.68/*$
FS	$-3.53/*$	$-0.85/*$	/	2.13/$*$	5.38/$*$	0.98/$*$	$-1.53/*$
STECS	$-5.48/*$	$-2.97/*$	$-2.13/*$	/	3.25/$*$	$-1.15/*$	$-3.65/*$
random	$-8.73/*$	$-6.22/*$	$-5.38/*$	$-3.25/*$	/	$-4.40/*$	$-6.90/*$
GS-$\ell_{1,\infty}$-[2]	$-4.33/*$	$-1.83/*$	$-0.98/*$	1.15/$*$	4.40/$*$	/	$-2.51/*$
GS-$\ell_{1,\infty}$	$-1.83/*$	0.68/$*$	1.53/$*$	3.65/$*$	6.90/$*$	2.51/$*$	/

(b)

Table 2: (a) The LMEM fixed-effect outcomes for $M = 2$ to 12 when $C = 25, L = 3, K = 1$. (b) The pairwise differences, showing the estimated difference between average GRQ (method in row $-$ method in column)/p-value per pair of methods (p-values < 0.0001 are indicated with $*$). Statistically significant differences are color coded. Values in green/red indicate that the method in the row outperforms/is outperformed by the method in the column.

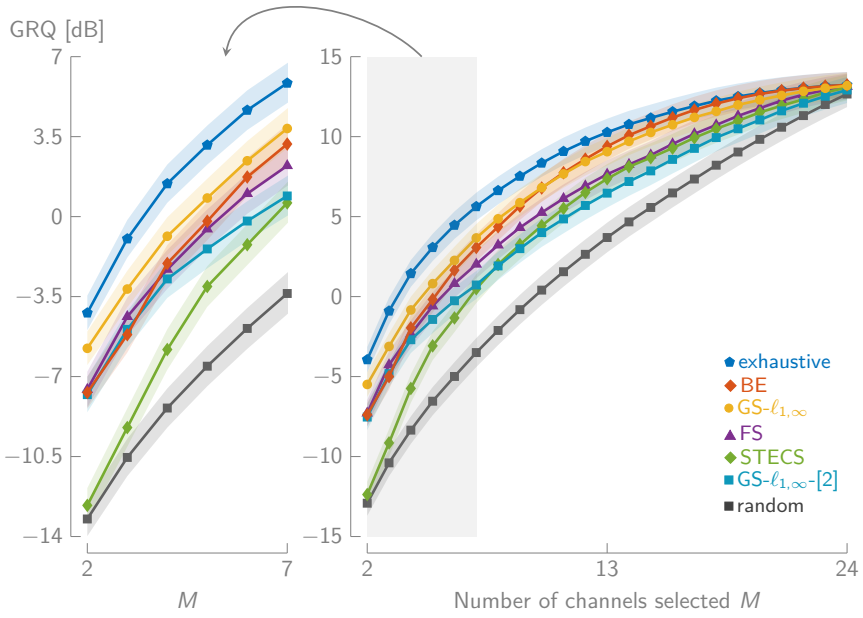

Figure 2: The output GRQ (mean over 250 runs) as a function of M for the different sensor selection methods when $C = 25, L = 3, K = 1$. The shading represents the standard error on the mean.

to the fact the FS method is limited to selecting one sensor at a time, which hampers its capacity to probe combined effects of multiple sensors.

Furthermore, Figure 2 and Table 2b show that the GS-$\ell_{1,\infty}$-[2] method is outperformed by all other methods, except by the STECS method. Our proposed method (significantly) outperforms GS-$\ell_{1,\infty}$-[2] across all M. This is an effect of dropping the off-diagonal blocks of the inequality constraints of (16). However, the gap between both methods becomes smaller for lower M (see also Figure 2). Similarly, the STECS method is outperformed, especially for low M, by all other methods, suffering from its non-convex objective function. This method achieves slightly higher GRQ than the GS-$\ell_{1,\infty}$-[2] method for most larger M, but it achieves much lower GRQ for small M. As a result, there is also a significant difference observed across all M between 2 and 12 between the STECS and GS-$\ell_{1,\infty}$-[2] method (Table 2b).

A remarkable conclusion is that the greedy BE method significantly outperforms almost all other state-of-the-art methods, including GS-$\ell_{1,\infty}$-[2] and STECS, which have not been benchmarked in a group-sparse setting against BE in the corresponding original papers [2] and [30], respectively. The only method that significantly outperforms the BE method is our proposed GS-$\ell_{1,\infty}$ algorithm. Interestingly, although the BE method seems to perform slightly better for larger M, the GS-$\ell_{1,\infty}$

method seems to perform better than the BE method for *small* M especially, explaining the statistically significant difference. The gap between both methods is also larger for these small M than for large M. From an application-based point of view, these smaller M - below half of the total number of sensors C - are often targeted in practice. Indeed, sensor selection is typically performed to substantially decrease the number of required sensors, not to remove only a few sensors. Although the heuristic BE method is computationally much more efficient than the optimization-based GS-$\ell_{1,\infty}$ method, it thus performs worse than the GS-$\ell_{1,\infty}$ method when it most matters. Lastly, it is interesting to identify in how many and in which cases a sensor selection method completely fails. For example, one could define a failure as more than 10 dB difference with the exhaustive method. Using this rule, the fail rate for the BE method across all runs and again for M between 2 and 12 is 3.71%, while this is 0.44% for the GS-$\ell_{1,\infty}$ method. Thus, the BE method has almost 10 times more severe fail cases than the GS-$\ell_{1,\infty}$ method. While these percentages might seem marginal at first sight, it should be taken into account that this percentage is biased by the highly randomized simulated scenarios. After a closer look, these fail cases turn out to mainly occur in cases where the covariance matrix $\mathbf{R}_2$ is ill-conditioned, which is not necessarily a rare case in practical settings, for example, as found in miniaturized EEG sensor networks [11]. In Section 4.5, we further analyze these particular fail cases and provide a more extensive discussion.

4.4. Comparison in the MIMO case ($K > 1$)

Figure 3 and Table 3a show the results of 100 Monte Carlo simulations with $C = 25, L = 2$, and $K = 2$, i.e., the more general case where $K > 1$. As the GS-$\ell_{1,\infty}$-[2] method was only proposed for $K = 1$, it is not included in these simulations.

First of all, the results confirm that the proposed extension to MIMO filtering in Section 3 is valid, as the GS-$\ell_{1,\infty}$ method still obtains GRQs close to those of the optimal exhaustive solution. Furthermore, the results are in line with Section 4.3. The BE and GS-$\ell_{1,\infty}$ method again show a statistically significant difference when evaluated across M between 2 and 12 (Table 3b), confirming that the latter has the advantage for the more relevant low M. Finally, both methods significantly outperform the other benchmark methods (except the exhaustive search).

Fixed-effect term	β	**SE**	**DF**	**t-value**	**p-value**
intercept	-14.64	0.91	6393	-16.01	< 0.0001
method = exhaustive$-$BE	2.10	0.11		18.54	< 0.0001
method = exhaustive$-$FS	3.18			28.04	< 0.0001
method = exhaustive$-$STECS	5.35			47.15	< 0.0001
method = exhaustive$-$random	9.09			80.17	< 0.0001
method = exhaustive$-$GS-$\ell_{1,\infty}$	1.51			13.36	< 0.0001

(a)

	exhaustive	**BE**	**FS**	**STECS**	**random**	**GS-$\ell_{1,\infty}$**
exhaustive	/	2.10/$*$	3.18/$*$	5.35/$*$	9.09/$*$	1.51/$*$
BE	-2.10/$*$	/	1.08/$*$	3.24/$*$	6.99/$*$	-0.59/$*$
FS	-3.18/$*$	-1.08/$*$	/	2.17/$*$	5.91/$*$	-1.66/$*$
STECS	-5.35/$*$	-3.24/$*$	-2.17/$*$	/	3.75/$*$	-3.83/$*$
random	-9.09/$*$	-6.99/$*$	-5.91/$*$	-3.75/$*$	/	-7.58/$*$
GS-$\ell_{1,\infty}$	-1.51/$*$	0.59/$*$	1.66/$*$	3.83/$*$	7.58/$*$	/

(b)

Table 3: (a) The LMEM fixed-effect outcomes for $M = 2$ to 12 when $C = 25, L = 2, K = 2$ (GS-$\ell_{1,\infty}$-[2] is omitted as it is only defined for $K = 1$). (b) The pairwise differences, showing the estimated difference between average GRQ (method in row $-$ method in column)/p-value per pair of methods (p-values < 0.0001 are indicated with $*$). Statistically significant differences are color coded. Values in green/red indicate that the method in the row outperforms/is outperformed by the method in the column.

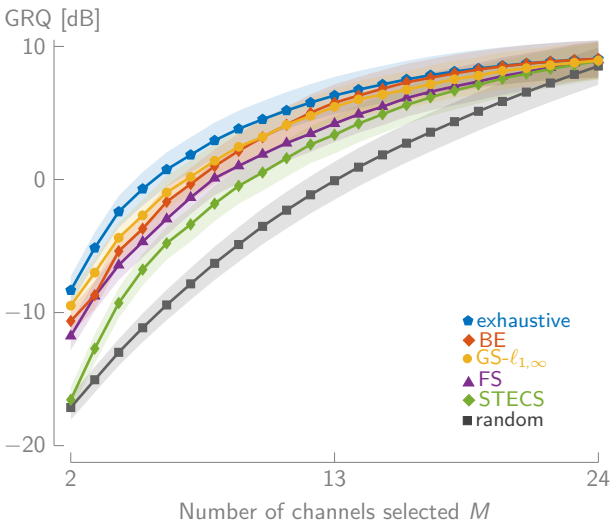

Figure 3: The output GRQ (mean over 100 runs) as a function of M for the different sensor selection methods when $C = 25, L = 2, K = 2$. The shading represents the standard error on the mean.

4.5. Comparison of GS-$\ell_{1,\infty}$ with BE

In this section, we zoom in on the comparison between the BE and GS-$\ell_{1,\infty}$ method, as these two methods achieve the highest GRQ in the previous simulations. As explained in Section 4.1.1, the BE method can suffer from its greedy sequential selection, where previously eliminated sensors can not be recovered when selecting a lower number of sensors. This inherent disadvantage of the BE method can lead to various fail cases (defined here as $> 10\,\text{dB}$ difference with the optimal exhaustive solution). After closer inspection, we identified that the majority of the fail cases (73.53% of all the fail cases in the previous simulations) corresponded to scenarios in which the matrix $\mathbf{R}_2$ was ill-conditioned, i.e., where there was a large difference between the largest and smallest eigenvalue(s).

To thoroughly test this case, we compare the BE method to the GS-$\ell_{1,\infty}$ method on the subset of 60 simulations of Section 4.3 where $2 \leq N_2 \leq 12$, i.e., where the number of signals contributing to $\mathbf{x}_2(t)$ is less than half of the $C = 25$ sensors. These cases correspond to ill-conditioned covariance matrices $\mathbf{R}_2$ in the denominator of the GRQ, where the smallest eigenvalues are determined solely by white Gaussian sensor noise (see Section 4.2.1). The results are shown in Figure 4, where also the performance of the exhaustive solution is shown as a reference. For large M, both methods perform similarly to the exhaustive solution, with very little change in GRQ for increasing values of M. When M decreases, both methods start to diverge from the exhaustive solution. However,

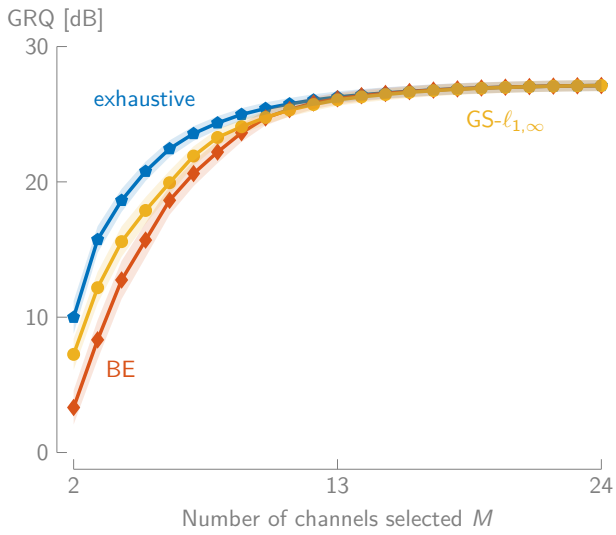

Figure 4: While the BE and GS-$\ell_{1,\infty}$ method performs on par for large M, the GS-$\ell_{1,\infty}$ method starts to outperform the BE method for smaller M in the ill-conditioned $\mathbf{R}_2$ covariance matrix case (mean $\pm$ standard error on the mean).

Fixed-effect term	β	**SE**	**DF**	**t-value**	**p-value**
intercept	31.30	0.44	1259	71.47	< 0.0001
method = GS-$\ell_{1,\infty}$−BE	1.53	0.19	1259	8.20	< 0.0001

Table 4: The LMEM outcomes when including only the BE and GS-$\ell_{1,\infty}$ methods in the case with an ill-conditioned covariance matrix $\mathbf{R}_2$ in the denominator (for $M = 2$ to 12).

the BE method achieves lower GRQ than the GS-$\ell_{1,\infty}$ method for smaller M. This is confirmed by the LMEM including only those two methods, as there is again a significant effect of the method, i.e., the GS-$\ell_{1,\infty}$ method outperforms the BE method (Table 4).

Figure 5 shows the differences in GRQ across all runs and M between 2 and 12 between the BE/GS-$\ell_{1,\infty}$ method and the exhaustive solution when $2 \leq N_2 \leq 12$. The BE method has a heavier tail with more outlying negative differences with the exhaustive solution than the GS-$\ell_{1,\infty}$ method. Of all runs with $2 \leq N_2 \leq 12$, there is a fail rate of 11.52% for the BE method, while this is only 0.91% for the GS-$\ell_{1,\infty}$ method. To summarize, when there is an ill-conditioned covariance matrix $\mathbf{R}_2$ in the denominator of the GRQ, the GS-$\ell_{1,\infty}$ method is more robust than the BE method.

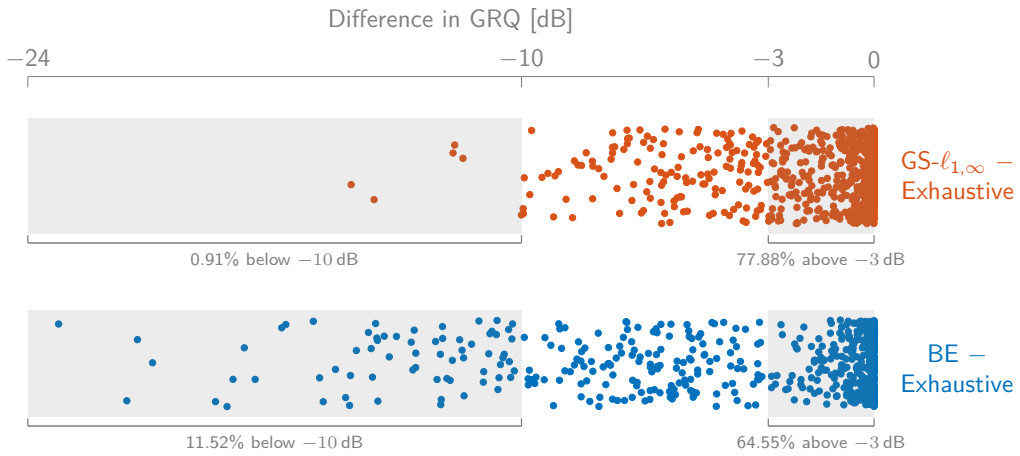

Figure 5: The BE method shows more outlying negative differences in GRQ (across all runs and M between 2 and 12) with the exhaustive solution than the GS-$\ell_{1,\infty}$ method when the covariance matrix in the denominator of the GRQ is ill-conditioned.

5. Example of sensor selection on real-world data

The benchmark study in Section 4 was performed on simulated data, which allowed us to generate a large number of simulations that are generic enough to apply to many sensor selection problems that arise in different signal processing domains. In this section, we show an example of sensor selection in the context of mobile epileptic seizure monitoring. More specifically, the task requires to design a spatiotemporal filter that amplifies multi-channel EEG data during seizures while maximally attenuating peak interferers [13]. The solution is found through max-SNR filtering and can be solved using the GEVD framework described in this paper. More information about the context, problem, and data can be found in [13].

In the following example, we investigate the effect of the reduction of EEG channels on subject three of the study in [13], aiming to design a mobile EEG setup. The data contains 16 channels (i.e., $C = 16$). Five time lags are used per channel (i.e., $L = 5$), while two output filters are computed (i.e., $K = 2$).

Figure 6 shows the GRQ as a function of the number of selected channels for the GS-$\ell_{1,\infty}$ and BE methods, which performed best in the benchmark study (see Section 4.3 to 4.5). The results are in line with the benchmark study. Both methods obtain similar GRQ across the whole range of selected channels, but the GS-$\ell_{1,\infty}$ method outperforms the greedy BE method for a low number of channels M. This confirms that the developed method performs as expected, also on sensor

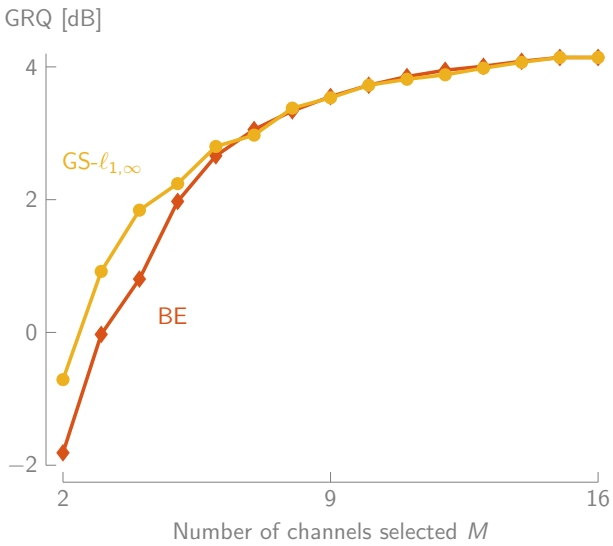

Figure 6: The GS-$\ell_{1,\infty}$ method outperforms the BE method for a low number of channels also on an example with real-world data collected on a patient with epilepsy ($C = 16, L = 5, K = 2$).

selection problems on real-world data.

6. Conclusion

In this paper, we proposed a group-sparse variable selection method using the $\ell_{1,\infty}$-norm for GRQ optimization and GEVD problems applied in the context of sensor selection. This group-sparsity does not only allow to extend spatial to spatio-temporal filtering but also to take multiple filters (eigenvectors) into account and thus extend MISO to MIMO filtering. The latter is essential in various other applications, such as selecting sensors across different filterbands in CSP applications [18, 19].

We have extensively compared the proposed GS-$\ell_{1,\infty}$ method with various other sensor selection methods (greedy, optimization-based, ...). Remarkably, the simple greedy BE method outperformed all methods from the state of the art, except the proposed GS-$\ell_{1,\infty}$ method. While the heuristic BE method is computationally more efficient, it performs worse than the GS-$\ell_{1,\infty}$ method for smaller numbers of selected sensors, and with a higher probability to completely fail. We have shown that one specific fail case of the BE method is when the covariance matrix in the denominator of the GRQ is ill-conditioned.

As the BE method is less robust than the proposed GS-$\ell_{1,\infty}$ method, the latter is the preferred

choice when performing variable selection, in particular if the number of desired variables is small compared to the total number of variables.

Author contributions

Jonathan Dan: Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing. **Simon Geirnaert:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing. **Alexander Bertrand:** Conceptualization, Methodology, Formal Analysis, Writing - Review & Editing, Supervision

Acknowledgements

This work was supported by an Aspirant Grant from the Research Foundation - Flanders (FWO) (for S. Geirnaert - 1136219N), by VLAIO and Byteflies through a Baekeland grant (HBC.2018.0189) (for J. Dan), FWO project nr. G0A4918N, the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No 802895), and the Flemish Government (AI Research Program).

References

[1] S. P. Chepuri, G. Leus, Sparsity-Promoting Sensor Selection for Non-Linear Measurement Models, IEEE Trans. Signal Process. 63 (3) (2015) 684–698. doi:10.1109/TSP.2014.2379662.

[2] S. A. Hamza, M. G. Amin, Sparse Array Beamforming Design for Wideband Signal Models, IEEE Trans. Aerosp. Electron. Syst. 57 (2) (2021) 1211–1226. doi:10.1109/TAES.2020.3037409.

[3] M. Gao, K. F. C. Yiu, S. Nordholm, On the Sparse Beamformer Design, Sensors 18 (10) (2018). doi:10.3390/s18103536.

[4] W. Shi, Y. Li, L. Zhao, X. Liu, Controllable Sparse Antenna Array for Adaptive Beamforming, IEEE Access 7 (2019) 6412–6423. doi:10.1109/ACCESS.2018.2889877.

[5] S. A. Hamza, M. G. Amin, Optimum sparse array receive beamforming for wideband signal model, in: Proc. of the 52nd ACSSC, 2018, pp. 89–93. doi:10.1109/ACSSC.2018.8645552.

[6] S. A. Hamza, M. G. Amin, Sparse Array DFT Beamformers for Wideband Sources, in: Proc. of the IEEE RadarConf19, 2019, pp. 1–5. doi:10.1109/RADAR.2019.8835749.

[7] S. A. Hamza, M. G. Amin, Sparse Array Receiver Beamformer Design for Multi-Functional Antenna, in: Proc. of EUSIPCO 2020, 2021, pp. 1836–1840. doi:10.23919/Eusipco47968.2020.9287795.

[8] S. A. Hamza, W. Zhai, X. Wang, M. G. Amin, Sparse Array Transceiver Design for Enhanced Adaptive Beamforming in MIMO Radar, in: Proc. of ICASSP 2021, 2021, pp. 4410–4414. doi:10.1109/ICASSP39728.2021.9414650.

[9] W. Zhai, X. Wang, S. A. Hamza, M. G. Amin, Cognitive-Driven Optimization of Sparse Array Transceiver for MIMO Radar Beamforming, in: Proc. of the IEEE RadarConf21, 2021, pp. 1–6. doi:10.1109/RadarConf2147009.2021.9455310.

[10] T. Alotaiby, F. E. El-Samie, S. A. Alshebeili, I. Ahmad, A review of channel selection algorithms for EEG signal processing, EURASIP J. Adv. Signal Process. (66) (2015). doi:10.1186/s13634-015-0251-9.

[11] A. M. Narayanan, A. Bertrand, Analysis of Miniaturization Effects and Channel Selection Strategies for EEG Sensor Networks with Application to Auditory Attention Detection, IEEE Trans. Biomed. Eng. 67 (1) (2020) 234–244. doi:10.1109/TBME.2019.2911728.

[12] A. M. Narayanan, P. Patrinos, A. Bertrand, Optimal Versus Approximate Channel Selection Methods for EEG Decoding With Application to Topology-Constrained Neuro-Sensor Networks, IEEE Trans. Neural Syst. Rehabilitation Eng. 29 (2021) 92–102. doi:10.1109/TNSRE.2020.3035499.

[13] J. Dan, B. Vandendriessche, W. V. Paesschen, D. Weckhuysen, A. Bertrand, Computationally-Efficient Algorithm for Real-Time Absence Seizure Detection in Wearable Electroencephalography, Int. J. Neural Syst. 30 (11) (2020) 2050035. doi:10.1142/S0129065720500355.

[14] A. Bertrand, Applications and trends in wireless acoustic sensor networks: A signal processing perspective, in: Proc. 18th IEEE SCVT, 2011, pp. 1–6. doi:10.1109/SCVT.2011.6101302.

[15] J. Zhang, S. P. Chepuri, R. C. Hendriks, R. Heusdens, Microphone Subset Selection for MVDR Beamformer Based Noise Reduction, IEEE/ACM Trans. Audio, Speech, Lang. Process. 26 (3) (2018) 550–563. doi:10.1109/TASLP.2017.2786544.

[16] O. Mehanna, N. D. Sidiropoulos, G. B. Giannakis, Joint Multicast Beamforming and Antenna Selection, IEEE Trans. Signal Process. 61 (10) (2013) 2660–2674. doi:10.1109/tsp.2013.2252167.

[17] S. A. Hamza, M. G. Amin, Hybrid Sparse Array Beamforming Design for General Rank Signal Models, IEEE Trans. Signal Process. 67 (24) (2019) 6215–6226. doi:10.1109/TSP.2019.2952052.

[18] S. Geirnaert, T. Francart, A. Bertrand, Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns, IEEE Trans. Biomed. Eng. 68 (5) (2021) 1557–1568. doi:10.1109/TBME.2020.3033446.

[19] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K.-R. Muller, Optimizing spatial filters for robust EEG single-trial analysis, IEEE Signal Process. Mag. 25 (1) (2007) 41–56. doi:10.1109/MSP.2008.4408441.

[20] M. Dash, H. Liu, Feature Selection for Classification, Intell. Data Anal. 1 (1) (1997) 131–156. doi:10.1016/S1088-467X(97)00008-5.

[21] S. Joshi, S. Boyd, Sensor selection via convex optimization, IEEE Trans. Signal Process. 57 (2) (2009) 451–462. doi:10.1109/TSP.2008.2007095.

[22] B. Van Veen, K. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Mag. 5 (2) (1988) 4–24. doi:10.1109/53.665.

[23] S. Yan, X. Tang, Trace quotient problems revisited, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vi-

sion – ECCV 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 232–244. doi:10.1007/11744047_18.

[24] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, S. Zhang, Semidefinite Relaxation of Quadratic Optimization Problems, IEEE Signal Process. Mag. 27 (3) (2010) 20–34. doi:10.1109/MSP.2010.936019.

[25] E. J. Candès, M. B. Wakin, S. P. Boyd, Enhancing Sparsity by Reweighted ℓ_1 Minimization, J. Fourier Anal. Appl. 14 (5-6) (2008) 877–905. doi:10.1007/s00041-008-9045-x.

[26] M. Grant, S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, version 2.2, http://cvxr.com/cvx (2020).

[27] M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, in: V. Blondel, S. Boyd, H. Kimura (Eds.), Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[28] MOSEK ApS, The MOSEK optimization toolbox for MATLAB manual. Version 9.1.9 (2019). URL http://docs.mosek.com/9.1/toolbox/index.html

[29] G. H. Golub, H. A. Van Der Vorst, Eigenvalue computation in the 20th century, J. Comput. Appl. Math. 123 (1-2) (2000) 35–65. doi:10.1016/S0377-0427(00)00413-1.

[30] F. Qi, W. Wu, Z. L. Yu, Z. Gu, Z. Wen, T. Yu, Y. Li, Spatiotemporal-Filtering-Based Channel Selection for Single-Trial EEG Classification, IEEE Trans. Cybern. 51 (2) (2021) 558–567. doi:10.1109/TCYB.2019.2963709.

[31] J. Meng, G. Liu, G. Huang, X. Zhu, Automated selecting subset of channels based on CSP in motor imagery brain-computer interface system, in: Proc. of IEEE Int. Conf. ROBIO, 2009, pp. 2290–2294. doi:10.1109/ROBIO.2009.5420462.

[32] M. Arvaneh, C. Guan, K. K. Ang, C. Quek, Optimizing the Channel Selection and Classification Accuracy in EEG-Based BCI, IEEE Trans. Biomed. Eng. 58 (6) (2011) 1865–1873. doi:10.1109/TBME.2011.2131142.

[33] I. Onaran, N. F. Ince, A. E. Cetin, Sparse spatial filter via a novel objective function minimization with smooth ℓ_1 regularization, Biomed. Signal Process. Control 8 (3) (2013) 282–288. doi:https://doi.org/10.1016/j.bspc.2012.10.003.

[34] A. Gałecki, T. Burzykowski, Linear Mixed-Effects Models Using R: A Step-by-Step Approach, Springer Texts in Statistics, Springer-Verlag New York, 2013. doi:10.1007/978-1-4614-3900-4.