Statistical control for spatio-temporal MEG/EEG source imaging with desparsified multi-task Lasso

Jerome-Alexis Chevalier
Inria Saclay
Paris-Saclay, France
jerome-alexis.chevalier@inria.fr

Alexandre, Gramfort
Inria Saclay
Paris-Saclay, France
alexandre.gramfort@inria.fr

Joseph Salmon
IMAG, Université de Montpellier
Montpellier, France
joseph.salmon@umontpellier.fr

Bertrand, Thirion
Inria Saclay, CEA
Paris-Saclay, France
bertrand.thirion@inria.fr

Abstract

Detecting where and when brain regions activate in a cognitive task or in a given clinical condition is the promise of non-invasive techniques like magnetoencephalography (MEG) or electroencephalography (EEG). This problem, referred to as source localization, or source imaging, poses however a high-dimensional statistical inference challenge. While sparsity promoting regularizations have been proposed to address the regression problem, it remains unclear how to ensure statistical control of false detections. Moreover, M/EEG source imaging requires to work with spatio-temporal data and autocorrelated noise. To deal with this, we adapt the desparsified Lasso estimator —an estimator tailored for high dimensional linear model that asymptotically follows a Gaussian distribution under sparsity and moderate feature correlation assumptions— to temporal data corrupted with autocorrelated noise. We call it the desparsified multi-task Lasso (d-MTLasso). We combine d-MTLasso with spatially constrained clustering to reduce data dimension and with ensembling to mitigate the arbitrary choice of clustering; the resulting estimator is called ensemble of clustered desparsified multi-task Lasso (ecd-MTLasso). With respect to the current procedures, the two advantages of ecd-MTLasso are that *i)*it offers statistical guarantees and *ii)*it allows to trade spatial specificity for sensitivity, leading to a powerful adaptive method. Extensive simulations on realistic head geometries, as well as empirical results on various MEG datasets, demonstrate the high recovery performance of ecd-MTLasso and its primary practical benefit: offer a statistically principled way to threshold MEG/EEG source maps.

1 Introduction

Source imaging with magnetoencephalography (MEG) and electroencephalography (EEG) delivers insights into brain activity with high temporal and good spatial resolution in a non-invasive way (Baillet et al., 2001). It however requires to solve the bioelectromagnetic inverse problem, which is a high-dimensional ill-posed regression problem. Various approaches have been proposed to regularize the estimation of the regression coefficients that map activity to brain locations. Historically, ℓ_2 regularization was considered first (Hämäläinen and Ilmoniemi, 1994), with successive improvements known as dSPM (Dale et al., 2000) and sLORETA (Pascual-Marqui, 2002) that are referred to as "noise normalized" solutions. The reason is that the coefficients are standardized with an estimate

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

of the noise standard deviation, producing outputs that are comparable to T or F statistics, yet not statistically calibrated. These latter techniques have since become standard when using ℓ_2 approaches.

More recently, alternative approaches based on sparsity assumptions have been proposed with the ambition to improve the spatial specificity of M/EEG source imaging (Matsuura and Okabe, 1995; Haufe et al., 2009; Gramfort et al., 2012; Lucka et al., 2012; Wipf and Nagarajan, 2009). The output of such methods consists of focal sources as opposed to blurred images obtained with ℓ_2 regularization. However, obtaining statistics ("noise normalized") from sparse or non-linear estimators seems challenging, especially since M/EEG data are spatio-temporal data with complex noise structure. A natural way to deal with the temporal dimension is to consider a multi-task estimator and structured sparse priors based on ℓ_1/ℓ_2 mixed norms (Ou et al., 2009; Gramfort et al., 2012).

In the statistical literature, some attempts to obtain an estimate of both regression coefficients and their variance have been proposed for linear models in high dimension (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Bühlmann, 2013). These estimates can then be translated to p-value maps, *i.e.,* maps of p-values associated with each covariate. Some methods adapted for sparse scenarios have then proposed to debias the Lasso to obtain p-values or confidence intervals (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014). We refer to such variants as desparsified Lasso. Recently, desparsified extensions of group Lasso have also been considered (Mitra and Zhang, 2016; Stucky and van de Geer, 2018). However, all these previous methods generally lack of power when $p \gg n$. Here, we propose to address a multi-task setting in the presence of correlated noise, and to deal with high-dimensional when $p \gg n$ leveraging on data structure as done by Chevalier et al. (2018). All these challenges need to be considered for M/EEG source imaging.

Our first contribution is to propose the desparsified multi-task Lasso (d-MTLasso), an extension of the desparsified Lasso (d-Lasso) (Zhang and Zhang, 2014; van de Geer et al., 2014) to multi-task setting (Obozinski et al., 2010). More precisely, we adapt the group formulation by Mitra and Zhang (2016) to the multi-task setting that enjoys *i)*a simple statistic test formula with *ii)*a natural integration of auto-correlated noise and *iii)*a simplification of the assumptions. Our second contribution is to introduce ensemble of clustered desparsified multi-task Lasso (ecd-MTLasso), which has two advantages compared to current methods: *i)*it offers statistical guarantees and *ii)*it allows to trade spatial specificity for sensitivity, leading to a powerful adaptive method. Our third contribution is an empirical validation of the theoretical claims. In particular, we run extensive simulations on realistic head geometries, as well as empirical results on various MEG datasets to demonstrate the high recovery performance of ecd-MTLasso and its primary practical benefit: offer a statistically principled way to threshold MEG/EEG source maps.

2 Theoretical Background

In this section, we give the noise model, we provide standard tools for solving the source localization problem and, mainly, we present three new methods with their assumptions and statistical guarantees.

2.1 Model and notation

For clarity, we use bold lowercase for vectors and bold uppercase for matrices. For any positive integer $p \in \mathbb{N}^*$, we write $[p]$ for the set $\{1, \dots, p\}$. For a vector $\boldsymbol{\beta}$, $\boldsymbol{\beta}_j$ refers to its j-th coordinate. For a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{X}^{(-j)}$ refers to matrix $\mathbf{X}$ without the j-th column, $\mathbf{X}_{i,\cdot}$ refers to the i-th row and $\mathbf{X}_{\cdot,j}$ to the j-th column and $\mathbf{X}_{i,j}$ refers to the element in the i-th row and j-th column. The notation $\|\cdot\|$ refers to the Frobenius norm for matrices and to the standard Euclidean norm for vectors. For a covariance matrix $\mathbf{M}$, the Mahalanobis norm is denoted by $\|\cdot\|_{\mathbf{M}^{-1}}$ and for a given vector $\mathbf{a}$ we have $\|\mathbf{a}\|^2_{\mathbf{M}^{-1}} \triangleq \mathrm{Tr}(\mathbf{a}^\top \mathbf{M}^{-1} \mathbf{a})$. For $\mathbf{B} \in \mathbb{R}^{p \times T}$, $\|\mathbf{B}\|_{2,1} = \sum_{j=1}^{p} \|\mathbf{B}_{j,\cdot}\|$, and its (row) support is $\mathrm{Supp}(\mathbf{B}) = \{j \in [p] : \mathbf{B}_{j,\cdot} \neq 0\}$. We assume that the underlying model is linear:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \ , \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times T}$ is the signal observed on M/EEG sensors, $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix representing the M/EEG forward model, $\mathbf{B} \in \mathbb{R}^{p \times T}$ the underlying signal in source space and $\mathbf{E} \in \mathbb{R}^{n \times T}$ the noise. We assume that there exist $\rho \in [0, 1)$ and $\sigma > 0$ such that all $t \in [T]$, $\mathbf{E}_{\cdot,t} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and that for all $i \in [n]$ and all $t \in [T-1]$, $\mathrm{Cor}(\mathbf{E}_{i,t}, \mathbf{E}_{i,t+1}) = \rho$. For all $i \in [n]$, $\mathbf{E}_{i,\cdot}$ is Gaussian with

Toeplitz covariance, *i.e.,* defining $\mathbf{M} \in \mathbb{R}^{T \times T}$ by $\mathbf{M}_{t,u} = \sigma^2 \rho^{|t-u|}$ for all $(t, u) \in [T]^2$, we have:

$$\mathbf{E}_{i,\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{M}) \ . \tag{2}$$

We further assume that $\mathbf{X}$ has been column-wise standardized and denote by $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{p \times p}$ the empirical covariance matrix of $\mathbf{X}$, *i.e.,* $\hat{\mathbf{\Sigma}} = \mathbf{X}^\top \mathbf{X}/n$ with $\hat{\mathbf{\Sigma}}_{j,j} = 1$. All proofs are given in Appendix D.

2.2 Metrics for statistical inference in M/EEG

To quantify the ability of a M/EEG source imaging technique to obtain a good estimated $\hat{\mathbf{B}}$, a commonly reported quantity is the Peak Localization Error (PLE) (Hauk et al., 2011). It consists in measuring the distance (in mm) along the cortical surface between the true simulated source and the location with maximum amplitude in the estimator. By contrast, spatial dispersion (SD) measures how much the activity is spread out by the inverse method (Molins et al., 2008).

To quantify the control of statistical errors, we consider a generalization of the Family Wise Error Rate (FWER) (Hochberg and Tamhane, 1987): the δ-FWER. As illustrated in Figure 5 in appendix, it is the FWER taken with respect to a ground truth dilated spatially by an amount δ —in the present study a distance in mm. A rigorous definition of δ-FWER is given in Appendix A. The rationale is that detections made outside of the support, but less than δ away from the support should count as slight inaccuracies of the methods, not as false positives. In an analogous manner, δ-FDR $=$ $(1 - \delta$-precision$)$ has been proposed recently as an extension of the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) to include a spatial tolerance (Nguyen et al., 2019; Gimenez and Zou, 2019). We thus characterize the selection capabilities of the methods through a δ-precision/recall curve.

2.3 Classical Solutions

The sLORETA and dSPM estimators are derived from the ridge estimator (Hoerl and Kennard, 1970):

$$\hat{\mathbf{B}}^{\text{Ridge}} = \mathbf{K}\mathbf{Y} \quad \text{where} \quad \mathbf{K} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1} \ . \tag{3}$$

They are obtained by scaling each row j in $\hat{\mathbf{B}}^{\text{Ridge}}$ by an estimate of the noise level at location j. It reads (Lin et al., 2006) $\hat{\mathbf{B}}^{\text{dSPM}}_{j,t} = \hat{\mathbf{B}}^{\text{Ridge}}_{j,t} / \sigma_j^{\text{dSPM}}$ and $\hat{\mathbf{B}}^{\text{sLORETA}}_{j,t} = \hat{\mathbf{B}}^{\text{Ridge}}_{j,t} / \sigma_j^{\text{sLORETA}}$, where $\sigma_j^{\text{dSPM}} = \sqrt{\sigma^2 [\mathbf{K}\mathbf{K}^\top]_{j,j}}$ and $\sigma_j^{\text{sLORETA}} = \sqrt{[\mathbf{K}(\sigma^2 \mathbf{I} + \mathbf{X}\mathbf{X}^\top)\mathbf{K}^\top]_{j,j}}$. Interestingly, it can be proved that in the absence of noise and when only a single coefficient is non-zero, the sLORETA estimate has its maximum at the correct location (Pascual-Marqui, 2002). Assuming $\mathbf{B}_{\cdot,t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the covariance of $\mathbf{Y}$ reads $\sigma^2 \mathbf{I} + \mathbf{X}\mathbf{X}^\top$. Hence, one can consider that sLORETA adds to dSPM an extra term in the sensor covariance matrix that comes from the sources. Note that these methods treat each time instant independently, hence ignoring source and noise temporal autocorrelations.

2.4 Desparsified multi-task Lasso (d-MTLasso)

Let us first recall the definition of the multi-task Lasso (MTLasso) estimator (Obozinski et al., 2010) in our setting. For a tuning parameter[1] $\lambda > 0$, it is defined as

$$\hat{\mathbf{B}}^{\text{MTL}} \in \underset{\mathbf{B} \in \mathbb{R}^{p \times T}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 + \lambda \|\mathbf{B}\|_{2,1} \right\} \ . \tag{4}$$

It is well known that similarly to the Lasso, MTLasso is biased: it tends to shrink rows with large amplitude towards zero. Below, we provide an adaptation of the Desparsified Lasso following the approach by Zhang and Zhang (2014), see also Mitra and Zhang (2016), to ensure statistical control. The approach relies on the introduction of score vectors $\mathbf{z}_1, \ldots, \mathbf{z}_p$ in $\mathbb{R}^n$ defined by

$$\mathbf{z}_j = \mathbf{X}_{\cdot,j} - \mathbf{X}^{(-j)} \hat{\boldsymbol{\beta}}^{(-j)}_{\boldsymbol{\alpha}_j} \ , \tag{5}$$

where, for $j \in [p]$, $\hat{\boldsymbol{\beta}}^{(-j)}_{\boldsymbol{\alpha}_j}$ is the Lasso solution (Tibshirani (1996); Chen and Donoho (1994)) of the regression of $\mathbf{X}_{\cdot,j}$ against $\mathbf{X}^{(-j)}$ with regularization parameter[2] $\boldsymbol{\alpha}_j$. Note that these score vectors

[1] λ is set by cross-validation on a logarithmic grid going from $\frac{\lambda_{\max}}{100}$ to $\lambda_{\max}$, where $\lambda_{\max} = \left\|\mathbf{X}^\top \mathbf{Y}\right\|_{2,\infty}$.

[2] In (Zhang and Zhang, 2014, Table 1) an algorithm for choosing $\boldsymbol{\alpha}_j$ is proposed. We noticed that taking for all $j \in [p]$, $\boldsymbol{\alpha}_j = c\boldsymbol{\alpha}_{\max,j} := c\|\mathbf{X}^{(-j)}\mathbf{X}_{\cdot,j}\|_\infty / n$ with $c = 0.5\%$ for M/EEG data allows to make a significant computation gain and yields adequate residuals for $C = 1000$ (see Sec. 2.6).

are independent of $\mathbf{Y}$ and their computation is then equivalent to solving the node-wise Lasso (Meinshausen and Bühlmann, 2006). For such vectors, the noise model in (1) yields

$$\frac{\mathbf{z}_j^\top \mathbf{Y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} = \mathbf{B}_{j,\cdot} + \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} + \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \mathbf{B}_{k,\cdot}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \quad . \tag{6}$$

Discarding the noise term and plugging $\hat{\mathbf{B}}^{\mathrm{MTL}}_{k,\cdot}$ as a preliminary estimator of $\mathbf{B}_{k,\cdot}$ in (6), we coin the desparsified multi-task Lasso (d-MTLasso), a debiased estimator of $\hat{\mathbf{B}}^{\mathrm{MTL}}$ defined for all $j \in [p]$ by

$$\hat{\mathbf{B}}^{(\mathrm{d-MTLasso})}_{j,\cdot} = \frac{\mathbf{z}_j^\top \mathbf{Y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \hat{\mathbf{B}}^{\mathrm{MTL}}_{k,\cdot}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \quad . \tag{7}$$

To derive d-MTLasso statistical properties, we need the extended Restricted Eigenvalue (RE) property (Lounici et al., 2011, Assumption 3.1), detailed in Appendix B. More precisely, we assume that

(A1) RE($\mathbf{X}$, s) is verified on $\mathbf{X}$ for a sparsity parameter $s \geq |\mathrm{Supp}(\mathbf{B})|$ and a constant $\kappa = \kappa(s) > 0$.

Roughly, A1 can be seen as a combination of sparsity and "moderate" feature correlation assumptions.

Proposition 2.1. *Considering the model in Equation (1), assuming A1 and for a choice of λ large enough*[3] *in Equation (4), then with high probability:*

$$\sqrt{n}(\hat{\mathbf{B}}^{(\mathrm{d-MTLasso})} - \mathbf{B}) = \mathbf{\Lambda} + \mathbf{\Delta} \quad , \tag{8}$$

$$\mathbf{\Lambda}_{j,\cdot} \sim \mathcal{N}_p(\mathbf{0}, \hat{\mathbf{\Omega}}_{j,j}\mathbf{M}), \text{ for all } j \in [p], \quad \text{where } \hat{\mathbf{\Omega}}_{j,k} = \frac{n \mathbf{z}_j^\top \mathbf{z}_k}{|\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}||\mathbf{z}_k^\top \mathbf{X}_{\cdot,k}|}$$

$$\|\mathbf{\Delta}\|_{2,1} = \mathrm{O}\left(\frac{s\lambda\sqrt{\log(p)}}{\kappa^2}\right) \tag{9}$$

Then, under the j-th null hypothesis $H_0^{(j)}$: "$\mathbf{B}_{j,\cdot} = 0$" and neglecting the term $\mathbf{\Delta}$ (see Appendix D.2 for more details) in (8) as done by van de Geer et al. (2014), $\hat{\mathbf{B}}^{(\mathrm{d-MTLasso})}_{j,\cdot}$ is Gaussian with zero-mean. Finally, using standard results on χ^2 distributions (see Appendix D.1), we obtain

$$n \left\| \hat{\mathbf{B}}^{(\mathrm{d-MTLasso})}_{j,\cdot} \right\|^2_{\mathbf{M}^{-1}} \sim \hat{\mathbf{\Omega}}_{j,j} \chi^2_T \quad .$$

If $\mathbf{M}$ is known, the quantity $n\|\hat{\mathbf{B}}^{(\mathrm{d-MTLasso})}_{j,\cdot}\|^2_{\mathbf{M}^{-1}}/\hat{\mathbf{\Omega}}_{j,j}$ can be used as a decision statistic to obtain a p-value testing the importance of source j by comparison with the χ^2_T distribution. In practice we need to estimate $\mathbf{M}$ by $\hat{\mathbf{M}}$. Notably, assuming that we have an estimator $\hat{\sigma}$ of σ that verifies approximately $(n - \hat{s})\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-\hat{s}}$, where $\hat{s} = |\mathrm{Supp}(\hat{\mathbf{B}}^{\mathrm{MTL}})|$ (see Sec. 2.5), we take

$$\hat{f}_j := \frac{n\|\hat{\mathbf{B}}^{(\mathrm{d-MTLasso})}_{j,\cdot}\|^2_{\hat{\mathbf{M}}^{-1}}}{T\,\hat{\mathbf{\Omega}}_{j,j}} \quad , \tag{10}$$

as statistic to compare with a Fisher distribution with parameters T and $n - \hat{s}$, to compute the p-values. The full d-MTLasso algorithm is given in Algorithm 1. Note that, a Python implementation of the procedures presented in this paper is available on `https://github.com/ja-che/hidimstat` along with some examples.

2.5 Noise parameters estimation

In Sec. 2.1 noise is assumed homogeneous across sensors, allowing to obtain a robust estimator. Extending Reid et al. (2016) to multi-task regression, we consider the residuals $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{\mathrm{MTL}}$, and the estimated support size $\hat{s}$. Defining, for $t \in [T]$, $\hat{\sigma}_t^2 = \|\hat{\mathbf{E}}_{\cdot,t}\|^2/(n - \hat{s})$, an estimate of σ^2 is:

$$\hat{\sigma}^2 = \mathrm{median}(\{\hat{\sigma}_t^2, t \in [T]\}) \quad .$$

Taking the median instead of the mean avoids depending on prospective under-fitted time steps and turns out to be more robust empirically. Similarly, defining for all $t \in [T-1]$, $\hat{\rho}_t = \mathrm{cor}_n(\hat{\mathbf{E}}_{\cdot,t}, \hat{\mathbf{E}}_{\cdot,t+1})$ (where $\mathrm{cor}_n(.,.)$ is the empirical correlation), ρ is estimated by taking $\hat{\rho} = \mathrm{median}(\{\hat{\rho}_t, t \in [T-1]\})$. Then, an estimator $\hat{\mathbf{M}}$ of $\mathbf{M}$ is given by $\hat{\mathbf{M}}_{t,u} = \hat{\sigma}^2\hat{\rho}^{|t-u|}$.

[3] See the proof of (Lounici et al., 2011, Theorem3.1).

Algorithm 1 d-MTLasso

input : $\mathbf{X} \in \mathbb{R}^{n \times p}, \mathbf{Y}$

$\hat{\mathbf{B}}^{\mathrm{MTL}} \leftarrow \mathrm{MTL}(\mathbf{X}, \mathbf{Y})$ // cross-validated multi-task Lasso

$\hat{\mathbf{E}} \leftarrow \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{\mathrm{MTL}}$ // Residuals

$\hat{s} \leftarrow |\mathrm{Supp}(\hat{\mathbf{B}}^{\mathrm{MTL}})|$

for $t \in [T]$ **do** // Noise level estimation

  $\hat{\sigma}_t^2 = \|\hat{\mathbf{E}}_{.,t}\|^2/(n - \hat{s})$

$\hat{\sigma}^2 = \mathrm{median}(\{\hat{\sigma}_t^2, t \in [T]\})$

Get $\hat{\mathbf{M}}$ thanks to Sec. 2.5

for $j \in [p]$ **do**

  $\mathbf{z}_j \leftarrow \mathrm{Lasso}(\mathbf{X}^{(-j)}, \mathbf{X}_{.,j})$ // cross-validated Lasso

  $\hat{\mathbf{\Omega}}_{j,j} \leftarrow \frac{n\mathbf{z}_j^\top \mathbf{z}_j}{|\mathbf{z}_j^\top \mathbf{X}_{.,j}||\mathbf{z}_j^\top \mathbf{X}_{.,j}|}$

  $\hat{\mathbf{B}}_{j,.}^{(\mathrm{d-MTLasso})} \leftarrow \frac{\mathbf{z}_j^\top \mathbf{Y}}{\mathbf{z}_j^\top \mathbf{X}_{.,j}} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{.,k} \hat{\mathbf{B}}_{k,.}^{\mathrm{MTL}}}{\mathbf{z}_j^\top \mathbf{X}_{.,j}}$ // Desparsified multi-task Lasso

  $\hat{f}_j \leftarrow \frac{n\|\hat{\mathbf{B}}_{j,.}^{(\mathrm{d-MTLasso})}\|_{\hat{\mathbf{M}}^{-1}}^2}{T\hat{\mathbf{\Omega}}_{j,j}}$ // Inference statistics

return $\hat{f}_1, \ldots, \hat{f}_p$

2.6 Clustering to handle spatially structured high-dimensional data

In the high-dimensional inference scenario considered, the number of sensors is more than one order of magnitude smaller than the number of sources, $n \ll p$. Therefore, estimators of conditional association between sources and observations struggle to identify the solution. The setting is even more difficult due to the presence of high correlation between sources (see Figure 6 in appendix). Further gains can however come from a compression of the design matrix (Bühlmann et al., 2013; Mandozzi and Bühlmann, 2016). For this we introduce a clustering step that reduces data dimensionality while leveraging spatial structure. We consider a spatially-constrained hierarchical clustering algorithm described by Varoquaux et al. (2012) that uses Ward criterion[4]. Other clustering schemes might be considered, as long as they yield spatially contiguous regions of the cortical surface. The combination of this clustering algorithm with the d-Lasso or d-MTLasso algorithms will be respectively referred to as clustered desparsified Lasso (cd-Lasso) and clustered desparsified multi-task Lasso (cd-MTLasso).

The number of clusters is denoted by C and, for $r \in [C]$, we denote by G_r the r-th group. Every cluster representative variable is given by the average of the covariates it contains. Then, reordering conveniently the columns of $\mathbf{X}$, the compressed design matrix $\mathbf{Z} \in \mathbb{R}^{n \times C}$ is given by:

$$\mathbf{Z} = \mathbf{X}\mathbf{A} \ , \quad \mathbf{A} = \begin{bmatrix} \frac{1}{|G_1|} & \cdots & \frac{1}{|G_1|} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{1}{|G_2|} & \cdots & \frac{1}{|G_2|} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \frac{1}{|G_r|} & \cdots & \frac{1}{|G_r|} \end{bmatrix} , \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{p \times C}$. We say that the compression of $\mathbf{X}$ is of good quality if:

(A2) there exists $\mathbf{\Gamma} \in \mathbb{R}^{C \times T}$ such that $\mathbf{\Gamma}_{r,.} = \sum_{j \in G_r} w_j \mathbf{B}_{j,.}$ with $w_j \geq 0$ for all $j \in [p]$, and the associated compression loss $\mathbf{X}\mathbf{B} - \mathbf{Z}\mathbf{\Gamma}$ is "small enough" with respect to the model noise (see Appendix D.3 for more details).

(A3)[5] $\mathrm{RE}(\mathbf{Z}, s')$ is verified on $\mathbf{Z}$ for sparsity parameter $s' \geq |\mathrm{Supp}(\mathbf{\Gamma})|$ and constant $\kappa' = \kappa'(s') > 0$.

Proposition 2.2. *Assume Equation (1), A2, A3, a choice of regularization parameter in the MTLasso regression of* $\mathbf{Z}$ *against* $\mathbf{Y}$ *that is large enough, and that the largest cluster of the compression is of size* δ*, then cd-MTLasso controls the* δ*-FWER.*

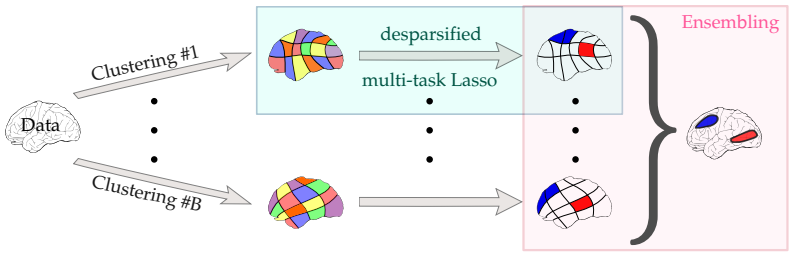

Figure 1: **ecd-MTL overview diagram.** While cd-MTLasso applies d-MTLasso to clustered data, ecd-MTLasso aggregates several cd-MTLasso solutions.

2.7 Ensemble of clustered desparsified multi-task Lasso (ecd-MTLasso)

To reduce the sensitivity of cd-MTLasso to small data perturbations, we propose to randomize over the clustering. We build several clustering solution, considering $B = 100$ different random subsamples of size 10% of the full sample; then we aggregate the p-value maps output by cd-MTLasso. To aggregate the B cd-MTLasso solutions, we use the adaptive quantile aggregation proposed by Meinshausen et al. (2009) detailed in Appendix C. The full procedure of ensembling B cd-MTLasso (resp. cd-Lasso), solutions is called ecd-MTLasso for ensemble of clustered desparsified multi-task Lasso (resp. ecd-Lasso). Algorithm of ecd-MTLasso is given in Algorithm 2 in appendix. Also, we give an overview diagram to clarify the nesting structure of the proposed solutions in Figure 1.

Proposition 2.3. *Assume that for each of the B compressions the hypotheses of Prop. 2.2 are verified, then ecd-MTLasso controls the δ-FWER.*

This result is conservative and mixing several cd-MTLasso usually reduces the spatial tolerance δ. Additional details on the procedure and computational complexity are deferred to Appendix E.

3 Experiments

In this section, we give empirical evidence of the advantages of ecd-MTLasso for source localization. First, in a typical point source simulation, we compare the methods with respect to the standard PLE metric; notably, we study the effect of i/clustering and ii/integrating time dimension. In a second simulation with more realistic features, we examine the δ-FWER control property and compare the support recovery properties of all methods. Lastly, working on real MEG data, we show that, contrary to sLORETA, ecd-MTLasso retrieves expected patterns using a universal threshold.

3.1 Simulation study

Here, we study how the proposed estimators perform compared to standard ℓ_2 regularized approaches, and assess whether time-aware statistical analysis improves upon static d-Lasso as it is essential for M/EEG source imaging. We use the head anatomy and the recording setup from the *sample* dataset publicly available from the MNE software (Gramfort et al., 2014). The design matrix $\mathbf{X}$ is computed with a three-shell boundary element model with $p = 7498$ candidate cortical locations, and a 306-channels Elekta Neuromag Vectorview system with 102 magnetometers and 204 gradiometers. We only keep the gradiometers and remove one defective sensor leading to $n = 203$. When considering multiple consecutive time instants to demonstrate the ability of the solver to leverage spatio-temporal data, the source is fixed and the temporal noise autocorrelation is set to $\rho = 0.3$.

Figure 2 reports the normalized histograms of PLE for the 7498 locations for the different methods investigated; results on spatial dispersion (SD) are available in Figure 7 in appendix. While it might seem simplistic to consider a single source, this experiment allows to demonstrate that d-Lasso improves over sLORETA in the presence of noise (see Figure 2, left). In the same figure, one can observe that clustering degrades this performance, as it carries an intrinsic spatial blur. However, even in this adversarial scenario (Dirac-like source location), cd-Lasso and ecd-Lasso remain competitive *w.r.t.* sLORETA, avoiding extreme PLE values. Note that, here, a single time point was used (T=1).

[4] A typical choice is $C = 1000$ clusters for M/EEG data.

[5] $|\mathrm{Supp}(\mathbf{\Gamma})| \leq |\mathrm{Supp}(\mathbf{B})|$ and $\mathbf{Z}$ is generally better conditioned than $\mathbf{X}$ making A3 more plausible than A1.

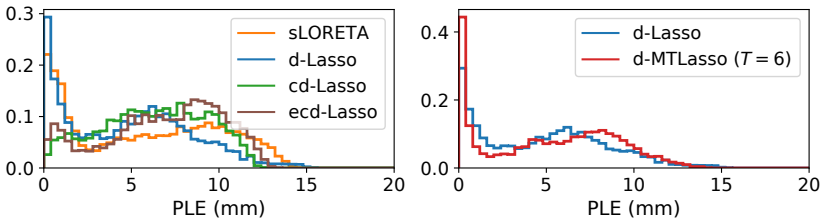

Figure 2: **Peak Localization Error (PLE) histograms.** (left): PLE on a fixed time point (T=1), sLORETA is outperformed by desparsified Lasso; cd-Lasso and ecd-Lasso are more concentrated and exhibit a smaller number of very low PLE but also a smaller number of extreme PLE values. (right): PLE for desparsified multi-task Lasso (d-MTLasso) with T=6 compared to d-Lasso (T=1). More time points improve the results by reducing the PLE.

The right panel in Figure 2 shows that d-MTLasso (T=6) significantly outperforms d-Lasso (T=1) in terms of PLE. Leveraging spatio-temporal data indeed increases the signal-to-noise ratio, which enhances spatial specificity. Effects in terms of SD are minor (see appendix, Figure 7).

3.2 Experiments on FWER control

We now investigate whether the different versions of d-MTLasso control the δ-FWER on a realistic simulation, and compare their support recovery properties. The data are the same as in Sec. 3.1. To simulate the sources, we randomly draw 3 active regions by selecting parcels from a subdivided cortical Freesurfer parcellation with 448 parcels (Khan et al., 2018). For each selected parcel we take as sources all the dipoles at a 10-mm geodesic distance from the center of the parcel (around 10 dipoles per region), fixing the amplitude at 10 nAm. To evaluate how the methods control the δ-FWER, we perform 100 simulations and count how often active sources are found outside the δ-dilated ground truth. In the left panel of Figure 3, we see that d-MTLasso does not control the δ-FWER, due to the violation of some hypotheses of proposition 1, in particular those regarding source correlation. However, we notice that handling noise autocorrelation reduces the empirical δ-FWER. Using clustering, assumptions of Prop. 2.2 are more easily met, in particular the conditioning of the problem is improved (Mattout et al., 2005). Yet cd-MTLasso does not control the δ-FWER for $\delta = 40$ mm, because the δ-FWER is controlled if δ is smaller than the largest cluster diameter, which may not hold. Finally, randomization via ecd-MTLasso further improves FWER control. Empirically, we observe that the δ-FWER is controlled for δ around twice the average cluster diameter. Then, with the limitation of having a compressed design matrix well conditioned (C not too large), we can reduce the tolerance δ by increasing C (empirical support of this claim in appendix in Figure 9). We have excluded sLORETA from this study since it does not provide guarantees on the false discoveries.

The right panel of Figure 3 shows the δ-precision recall curve of the different methods. We first notice that d-MTLasso cannot compete with sLORETA, because the high dimensionality of the problem makes the computation of the source importance overly ill-posed. cd-MTLasso improves detection accuracy, but still does not perform as well as sLORETA. However, adding the ensembling step, the δ-precision improves strongly, making ecd-MTLasso much better than sLORETA. In Figure 8 in appendix, we obtain similar results when considering the standard precision-recall curve.

3.3 Results on three MEG datasets

We now report results on three MEG datasets spanning three types of sensory stimuli: auditory, visual and somatosensory. Additional results on EEG datasets are presented in Appendix H. The auditory evoked fields (AEF) and visual evoked field (VEF) are obtained using stimuli in the left ear and left visual hemifield. The somatosensory evoked fields (SEF) are obtained following electrical stimulation of the left median nerve on the wrist. The detailed description of the data is provided in Appendix F.

Experimental results are presented in Figure 4 and Figure 10 (cf. Appendix H). Among the many methods for M/EEG source imaging present in the literature, the methods that are compared here have in common to output a statistical map. The ℓ_2 regularized sLORETA method is compared to the

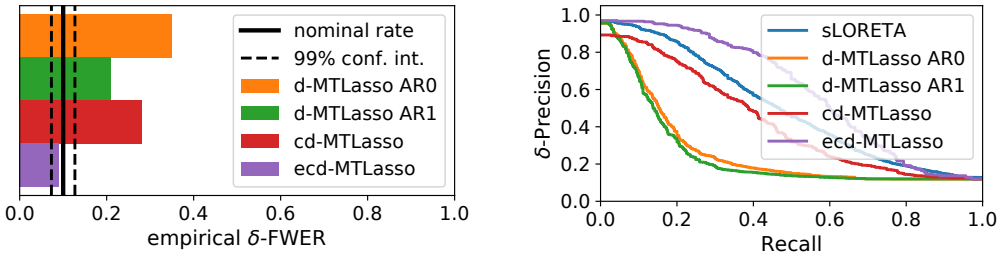

Figure 3: **δ-FWER and δ-Precision-Recall.** (left): δ-FWER control of the different d-MTLasso methods. δ-FWER control is hard for d-MTLasso and cd-MTLasso, as some detections are made far from the true sources, due to remote correlations. Ensembles of clusters allow to limit these false detections. (right): δ-Precision-Recall curves: sLORETA outperforms d-MTLasso AR0 and AR1, because the problem is too high dimensional for the d-MTLasso to work properly. Clustering improves the outcome, and ensembling brings further benefits: ecd-MTLasso outperforms sLORETA.

debiased sparse estimators presented and evaluated above. The input for all solvers is a time window of data: from $t = 50$ to $t = 100$ ms for AEF and VEF, and from $t = 30$ to $t = 40$ ms for SEF. During such time intervals one can expect the sources to originate primarily from the early sensory cortices whose locations are anatomically known for normal subjects.

First one can observe that all methods manage to highlight the proper functional sensory units (planum temporale for AEF, calcarine region for VEF and central sulcus for SEF). Considering sLORETA results, one can observe that at a common threshold of 3.0 on the Student statistic, the estimator is quite spatially specific for VEF, but is overly conservative for AEF and clearly leading to many false positives for SEF. By inspection of the d-MTLasso solution, one can observe that taking into account the autocorrelation of the noise leads to a better calibrated noise variance, and therefore fewer dubious detection. Considering ecd-MTLasso results, while all maps are also thresholded with a single level, one can see that it retrieves expected patterns without making dubious discoveries.

3.4 Summary, guidelines and limitations

Summary of experiments. In Sec. 3.1, we have shown that taking into account the time dimension improve the results in terms of PLE. Also, we have seen that even in this adversarial point source scenario (cf. Sec. 3.1), clustered methods remain competitive. In Sec. 3.2, while no control of false discoveries is proposed by sLORETA, ecd-MTL is the only method that offers statistical control in practice. Namely, it controls the δ-FWER for δ equals to twice the average cluster diameter. Additionally, in this realistic simulation, ecd-MTL exhibits the best support recovery properties. In Sec. 3.3, working on real MEG data, we show that, contrary to sLORETA, ecd-MTLasso produces calibrated statistics with universal threshold and retrieves expected patterns without making dubious discoveries. Overall, ecd-MTL offers statistical guarantees and is our privileged method.

Guidelines for statistical inference with ecd-MTLasso on temporal M/EEG data. First, we try to give guidelines concerning the number of clusters C. Hoyos-Idrobo et al. (2015) exhibit that clustering improves problem conditioning, this means that the Restricted Eigenvalue (RE) property (see assumptions A1 and A3) is more likely to be verified. Complementary, we argue that, keeping C over a hundred (limiting compression loss), the fewer clusters, the more A3 is likely to be verified for Prop. 2.2 and Prop. 2.3 to hold but also the better the sensitivity of ecd-MTL. However, small C also requires a higher spatial tolerance. We then hit a fundamental trade-off for statistical inference between sensitivity and spatial specificity. Then, C can be chosen depending on the problem setting: if it is difficult (noisy), it seems natural to lower spatial tolerance expectations (diminish C); in that sense ecd-MTL is an adaptive method (cf. Figure 9). For the present use case, taking $C = 1000$ seems an adequate trade-off to ensure δ-FWER control with reasonable spatial tolerance.

Now, we give recommendation for time sampling and window size. Choosing too short windows complicate AR model estimation due to the lack of data, while choosing too large windows may lead to non stationary support. We recommend taking windows of 20 to 50ms with a time sampling at 5 to 10ms as keeping $T < 10$ reduces computation time and should not decrease sensitivity significantly.

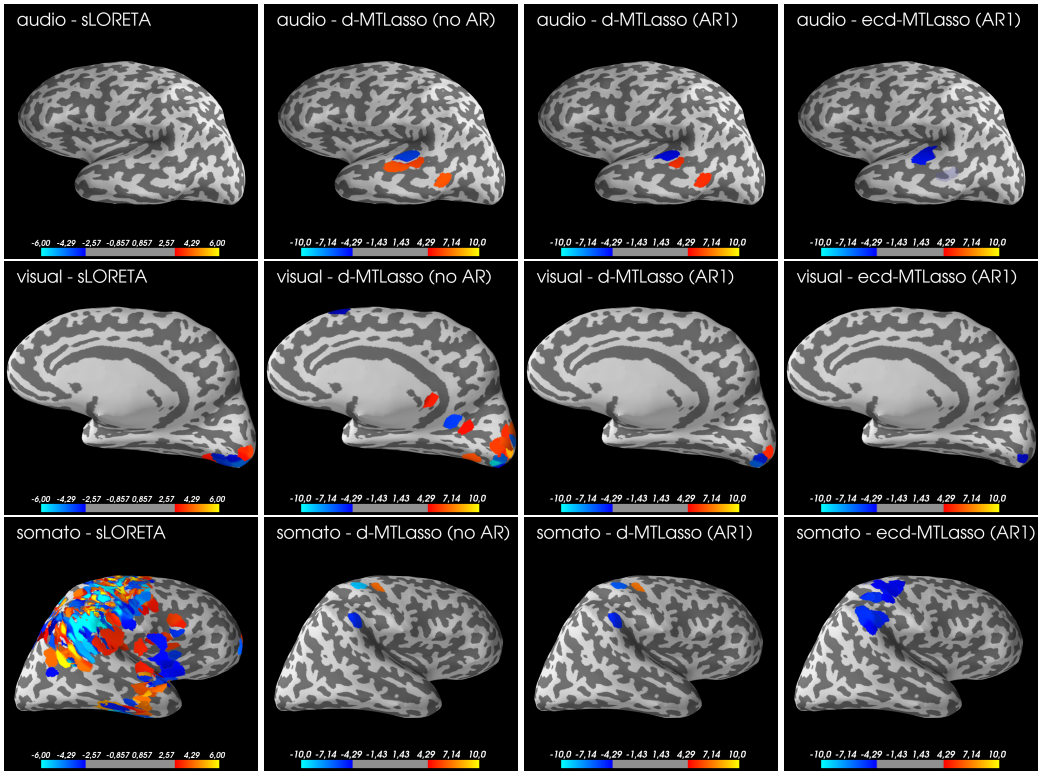

Figure 4: **Empirical comparison on 3 MEG datasets.** From left to right one can see sLORETA, d-MTLasso without AR modeling (assuming non-autocorrelated noise), d-MTLasso with an AR1 noise model and the ecd-MTLasso using also an AR1. Results correspond to auditory (top), visual (middle) and somatosensory (bottom) evoked fields. Colormaps are fixed across datasets and adjusted based on meaningful statistical thresholds in order to qualitatively illustrate FWER control issues.

Finally, when working with M/EEG data, we recommend to use only 10% of the full data to compute several clustering solutions with spatial constraint and Ward criteria to ensure enough diversity.

Limitations. The main limitation is the fact that mixing different types of sensors violates modeling assumptions both on temporal correlations and on spatial correlations, that is why we had to treat MEG and EEG sensors separately. A possibility to handle heterogeneous sensors is to follow Massias et al. (2018b), but for the temporal part further developments are required and left for future work.

Also left for future work, is the possibility of studying windows larger than 50ms. A simple solution is to slide a window of 20 to 50ms over the considered period of time.

Finally, a more common limitation is the fact that assumptions are hard to test in practice.

4 Conclusion

The MEG source imaging problem poses a hard statistical inference challenge: namely that of high-dimensional statistical analysis, furthermore with high correlations in the design. We have proposed an estimator that calibrates correctly the effects size and variance, up to a number of hypotheses, that are not easily met: some level of sparsity, mild correlation across sensors, homogeneity and heteroscedasticity of the noise. Up to these hypotheses, and up to a spatial tolerance on the exact location of the sources, we provide the first method with statistical guarantees for source imaging. This is made possible by bringing several improvements to the original desparsified Lasso solution: a multi-task formulation that increases power by basing inference on multiple time steps, a clustering step that renders the design less ill-posed and an ensembling step that mitigates the (hard) choice of clusters. Finally, our privileged method, ecd-MTLasso, runs in less than 10 mn on a real dataset on non-specialized hardware, making it usable by practitioners.

5 Statement of broader impact

Magnetoencephalography (MEG) and electroencephalography (EEG) offer a unique opportunity to image brain activity non-invasively with a temporal resolution in the order of milliseconds. This is relevant for cognitive neuroscience to describe the sequence of active areas during certain cognitive tasks, but also for clinical neuroscience, where electrophysiology is used for diagnosis (*e.g.,* sleep medicine, epilepsy presurgical mapping). Yet, doing brain imaging with M/EEG requires to solve a challenging high-dimensional inverse problem for which statistical guarantees are crucially important. In this work, we address this statistical challenge when using sparsity promoting regularization and when considering the specificity of M/EEG signals: data are spatio-temporal and the noise is temporally autocorrelated. The proposed algorithm is built on very recent work in optimization to speed up Lasso-type solvers, as well as work in mathematical statistics on desparsified Lasso estimators. We believe that this work, whose contribution is both on the modeling side and on the inference aspects, brings sparse estimators close to a wide adoptions in the neuroscience community.

We also would like to emphasize that the inference framework can be adapted to many other high-dimensional problems where data structure can be leveraged: biomedical data and physical observations (cardiac or brain monitoring, genomics, seismology, etc.), especially those that involve severely ill-posed inverse problems.

Acknowledgements This research is supported under funding of French ANR project FastBig (ANR-17- CE23-0011), the KARAIB AI chair (ANR-20-CHIA-0025-01), the European Research Council Starting Grant SLAB ERC-StG-676943 and Labex DigiCosme (ANR-11-LABEX-0045-DIGICOSME).

References

S. Baillet, J. C. Mosher, and R. M. Leahy. Electromagnetic brain mapping. *IEEE Signal Proc. Mag.*, 18(6):14–30, Nov. 2001.

R. F. Barber and E. J. Candès. A knockoff filter for high-dimensional selective inference. *Ann. Statist.*, 47(5):2504–2537, 2019.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1):289–300, 1995.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 09 2013.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, Nov 2013.

S. Chen and D. L. Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

J.-A. Chevalier, J. Salmon, and B. Thirion. Statistical inference with ensemble of clustered desparsified lasso. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 638–646, 2018.

A. M. Dale, A. K. Liu, B. R. Fischl, R. L. Buckner, J. W. Belliveau, J. D. Lewine, and E. Halgren. Dynamic statistical parametric mapping. *Neuron*, 26(1):55–67, 2000.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statist. Sci.*, 30(4):533–558, 2015.

O. J. Dunn. Multiple comparisons among means. *J. Amer. Statist. Assoc.*, 56(293):52–64, 1961.

J. R. Gimenez and J. Zou. Discovering conditionally salient features with statistical guarantees. In *ICML*, pages 2290–2298, 2019.

A. Gramfort, M. Kowalski, and M. Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Phys. Med. Biol.*, 57(7):1937–1961, 2012.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, 2014.

M. S. Hämäläinen and R. J. Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, 32(1):35–42, Jan 1994.

S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and Guido Nolte. Estimating vector fields using sparse basis field expansions. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NeurIPS*, pages 617–624. Curran Associates, Inc., 2009.

O. Hauk, D. G. Wakeman, and R. Henson. Comparison of noise-normalized minimum norm estimates for meg analysis using multiple resolution metrics. *NeuroImage*, 54(3):1966 – 1974, 2011.

Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1987.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Andrés Hoyos-Idrobo, Yannick Schwartz, Gaël Varoquaux, and Bertrand Thirion. Improving sparse recovery on structured images with bagged clustering. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 73–76. IEEE, 2015.

L. Janson and W. Su. Familywise error rate control via knockoffs. *Electron. J. Stat.*, 10(1):960–975, 2016.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014.

S. Khan, J. A. Hashmi, F. Mamashli, K. Michmizos, M. G. Kitzbichler, H. Bharadwaj, Y. Bekhti, S. Ganesan, K.-L. A. Garel, S. Whitfield-Gabrieli, R. L. Gollub, J. Kong, L. M. Vaina, K. D. Rana, S. M. Stufflebeam, M. S. Hämäläinen, and T. Kenet. Maturation trajectories of cortical resting-state networks depend on the mediating frequency band. *NeuroImage*, 174:57 – 68, 2018.

F. H. Lin, T. Witzel, S. P. Ahlfors, S. M. Stufflebeam, J. W. Belliveau, and M. S. Hämäläinen. Assessing and improving the spatial accuracy in meg source localization by depth-weighted minimum-norm estimates. *NeuroImage*, 31(1):160–71, 2006.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

F. Lucka, S. Pursiainen, M. Burger, and C. Wolters. Hierarchical bayesian inference for the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents. *NeuroImage*, 61(4):1364–1382, Apr. 2012.

J. Mandozzi and P. Bühlmann. Hierarchical testing in the high-dimensional setting with correlated variables. *J. Amer. Statist. Assoc.*, 111(513):331–343, 2016.

M. Massias, A. Gramfort, and J. Salmon. Celer: a Fast Solver for the Lasso with Dual Extrapolation. In *ICML*, volume 80, pages 3315–3324, 2018a.

M. Massias, S. Vaiter, A. Gramfort, and J. Salmon. Dual extrapolation for sparse generalized linear models. *arXiv preprint arXiv:1907.05830*, 2019.

Mathurin Massias, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. In *International Conference on Artificial Intelligence and Statistics*, pages 998–1007, 2018b.

K. Matsuura and Y. Okabe. Selective minimum-norm solution of the biomagnetic inverse problem. *IEEE Trans. Biomed. Eng.*, 42(6):608–615, June 1995. ISSN 0018-9294.

J. Mattout, M. Pélégrini-Issac, L. Garnero, and H. Benali. Multivariate source prelocalization (MSP): Use of functionally informed basis functions for better conditioning the MEG inverse problem. *NeuroImage*, 26(2):356 – 373, 2005.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72: 417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009.

R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electron. J. Stat.*, 10(2):1829–1873, 2016.

A. Molins, S. M. Stufflebeam, E. N. Brown, and M. S. Hämäläinen. Quantification of the benefit from integrating MEG and EEG data in minimum L2-norm estimation. *NeuroImage*, 42(3):1069 – 1077, 2008.

T.-B. Nguyen, J.-A. Chevalier, and B. Thirion. Ecko: Ensemble of clustered knockoffs for robust multivariate inference on fMRI data. In *International Conference on Information Processing in Medical Imaging*, pages 454–466, 2019.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

W. Ou, M. S. Hämaläinen, and P. Golland. A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946, Feb. 2009.

R. Pascual-Marqui. Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find. Exp. Clin. Pharmacology*, 24(D):5–12, 2002.

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Stat. Sin.*, 26(1):35–67, 2016.

B. Stucky and S. van de Geer. Asymptotic confidence regions for high-dimensional structured sparsity. *IEEE Trans. Signal Process.*, 66(8):2178–2190, 2018.

S. Taulu. Spatiotemporal Signal Space Separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, 51(7):1759–1769, 2006.

J. Taylor and R. J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *ICML*, pages 1375–1382, 2012.

L. Wasserman and K. Roeder. High-dimensional variable selection. *Ann. Statist.*, 37(5A):2178–2201, 2009.

D. Wipf and S. Nagarajan. A unified bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966, Feb. 2009.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(1):217–242, 2014.

APPENDIX

A Formal definition of δ-FWER control

Now we give a more formal definition of the δ-FWER.

Definition A.1 (δ-family wise error rate)**.** *Given a family of (corrected) p-values $\hat{p} = (\hat{p}_j)_{j\in[p]}$ and a threshold $x \in (0,1)$, the δ-FWER, also denoted $\mathrm{FWER}_x^\delta(\hat{p})$, is the probability to make at least one false discovery at a distance at least δ from the true support:*

$$\mathrm{FWER}_x^\delta(\hat{p}) = \mathbb{P}(\min_{j\in N^\delta} \hat{p}_j \leq x) \ , \tag{12}$$

with $N^\delta = \{j \in [p] : \forall k \in \mathrm{Supp}(\mathbf{B}), d(j,k) \geq \delta\}$ and $d(j,k)$ is the distance between source j and k.

Definition A.2 (δ-FWER control)**.** *We say that the family of (corrected) p-values $\hat{p} = (\hat{p}_j)_{j\in[p]}$ controls the δ-FWER if, for all $x \in (0,1)$:*

$$\mathrm{FWER}_x^\delta(\hat{p}) \leq x \ . \tag{13}$$

B Extended Restricted Eigenvalue assumption

Here, we rewrite (Lounici et al., 2011, Assumption 3.1), adjusting it for the multi-task Lasso case (particular case of the more general group Lasso). Notice that for a given value of T, the assumption is equivalent to (Lounici et al., 2011, Assumption 4.1). Let $1 \leq s \leq p$ be an integer that gives an upper bound on the sparsity $|\mathrm{Supp}(\mathbf{B})|$. The extended Restricted Eigenvalue assumption RE($\mathbf{X}, s$) is verified on $\mathbf{X}$ for sparsity parameter s and constant $\kappa = \kappa(s) > 0$, if:

$$\min \left\{ \frac{\|\mathbf{X}\mathbf{\Theta}\|}{\sqrt{nT}\,\|\mathbf{\Theta}_J\|} : |J| \leq s, \mathbf{\Theta} \in \mathbb{R}^{p\times T} \setminus \{\mathbf{0}\}, \|\mathbf{\Theta}_{J^C}\|_{2,1} \leq 3\,\|\mathbf{\Theta}\|_{2,1} \right\} \geq \kappa \ , \tag{14}$$

where $J \subset [p]$ and J^C denotes its complementary *i.e.,* $J^C = [p] \setminus J$, and $\mathbf{\Theta}_J$ refers to the matrix $\mathbf{\Theta}$ without the rows J^C.

C Adaptive quantile aggregation of p-values and ecd-MTLasso algorithm

In this section, we provide some more details on the way we perform aggregation of p-values across the p-values maps created through the clustering randomization, then we give the full ecd-MTLasso algorithm.

For the j-th features (or source) we have a vector $(p_j^{(b)})_{b\in[B]}$ of p-values, with one p-value computed for each of the B clusterings. Then, the final p-value of the j-th feature is given by the adaptive quantile aggregation, as proposed by Meinshausen et al. (2009):

$$p_j = \min \left\{ (1 - \log(\gamma_{\min})) \inf_{\gamma\in(\gamma_{\min},1)} \left(\gamma\text{-quantile}\left\{ \frac{p_j^{(b)}}{\gamma}; b \in [B] \right\} \right) \ , \ 1 \right\} \ ,$$

where we have taken $\gamma_{\min} = 0.25$ in our experiments. Taking a value of $\gamma_{\min}$ not too small (*e.g.,* $\gamma_{\min} \geq 0.25$) allows to recover sources that have received small p-values several times (*e.g.,* at least for $B/4$ different choices of clustering).

We give the full algorithm of ecd-MTLasso in Algorithm 2.

D Proofs

D.1 Probability lemma

Lemma D.1. *Let $\boldsymbol{\varepsilon} \in \mathbb{R}^T$ be a centered Gaussian random vector with (symmetric positive definite) covariance $\mathbf{M} \in \mathbb{R}^{T\times T}$. Then, the random variable $\boldsymbol{\varepsilon}^\top \mathbf{M}^{-1} \boldsymbol{\varepsilon}$ follows a χ^2_T distribution.*

Algorithm 2 ecd-MTLasso

input : $\mathbf{X} \in \mathbb{R}^{n \times p}, \mathbf{Y}$

param : $C = 1000, B = 100$

for $b = 1, \ldots, B$ **do**

$\mathbf{X}^{(b)} = \texttt{sample}(\mathbf{X})$

$\mathbf{A}^{(b)} = \texttt{Ward}(C, \mathbf{X}^{(b)})$

$\mathbf{Z}^{(b)} = \mathbf{X}\mathbf{A}^{(b)}$

$q^{(b)} = \frac{\texttt{d-MTLasso}(\mathbf{Z}^{(b)}, \mathbf{Y})}{C}$ // corrected cluster-wise p-values at bootstrap b

for $j = 1, \ldots, p$ **do**

$p_j^{(b)} = q_r^{(b)}$ if $j \in G_r$ // corrected feature-wise p-values at bootstrap b

for $j = 1, \ldots, p$ **do**

$p_j = \texttt{aggregation}(p_j^{(b)}, b \in [B])$ // aggregated corrected feature-wise p-values

return p_j for $j \in [p]$

Proof. Note first that since $\mathbf{M}$ is symmetric positive definite, its square-root $\mathbf{N} \in \mathbb{R}^{T \times T}$ exists and is a symmetric positive definite matrix satisfying $\mathbf{N}^2 = \mathbf{M}$. Hence, this leads to the following displays

$$\boldsymbol{\varepsilon}^\top \mathbf{M}^{-1} \boldsymbol{\varepsilon} = (\mathbf{N}^{-1}\boldsymbol{\varepsilon})^\top (\mathbf{N}^{-1}\boldsymbol{\varepsilon}).$$

We have that $\mathbf{N}^{-1}\boldsymbol{\varepsilon}$ is a centered Gaussian random vector, and its covariance matrix reads:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{N}^{-1}\boldsymbol{\varepsilon})(\mathbf{N}^{-1}\boldsymbol{\varepsilon})^\top\right] &= \mathbb{E}\left[\mathbf{N}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{N}^{-1}\right] \\
&= \mathbb{E}\left[\mathbf{N}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{N}^{-1}\right] \\
&= \mathbf{N}^{-1}\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\mathbf{N}^{-1} \\
&= \mathbf{N}^{-1}\mathbf{M}\mathbf{N}^{-1} \\
&= \mathbf{N}^{-1}\mathbf{N}^2\mathbf{N}^{-1} \\
&= \mathrm{Id}_T \quad .
\end{aligned}$$

To conclude $\mathbf{N}^{-1}\boldsymbol{\varepsilon} \in \mathbb{R}^T$ is a centered Gaussian vector with covariance Id_T, hence its squared Euclidean norm $\left\|\mathbf{N}^{-1}\boldsymbol{\varepsilon}\right\|^2 = (\mathbf{N}^{-1}\boldsymbol{\varepsilon})^\top(\mathbf{N}^{-1}\boldsymbol{\varepsilon})$ follows a χ^2_T distribution. □

D.2 Proof of Prop. 2.1

Now, we give a proof of Prop. 2.1:

Proof. First, let us fix an index $j \in [p]$. Then, using Equation (7) we have:

$$\begin{aligned}
\sqrt{n}(\hat{\mathbf{B}}^{(\mathrm{d-MTLasso})}_{j,\cdot} - \mathbf{B}_{j,\cdot}) &= \sqrt{n}\frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} - \sum_{k \neq j} \frac{\sqrt{n}\,\mathbf{z}_j^\top \mathbf{X}_{\cdot,k}(\hat{\mathbf{B}}^{\mathrm{MTL}}_{k,\cdot} - \mathbf{B}_{k,\cdot})}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \\
&= \boldsymbol{\Lambda}_{j,\cdot} + \boldsymbol{\Delta}_{j,\cdot} \quad ,
\end{aligned} \tag{15}$$

where $\boldsymbol{\Lambda}_{j,\cdot} = \sqrt{n}\frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}$ and $\boldsymbol{\Delta}_{j,\cdot} = \sqrt{n}\sum_{k \neq j} \mathbf{P}_{j,k}(\mathbf{B}_{k,\cdot} - \hat{\mathbf{B}}^{\mathrm{MTL}}_{k,\cdot})$ with

$$\mathbf{P}_{j,k} = \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} \quad .$$

Now, we show that $\boldsymbol{\Lambda}_{j,\cdot} \sim \mathcal{N}_p(\mathbf{0},\, \hat{\boldsymbol{\Omega}}_{j,j}\mathbf{M})$, or equivalently we show that $\mathbf{E}^\top \mathbf{z}_j \sim \mathcal{N}(0, n\left\|\mathbf{z}_j\right\|^2 \mathbf{M})$. It is clear that $\mathbf{E}^\top \mathbf{z}_j$ is a centered Gaussian vector. Then, its covariance denoted by $\mathbf{V}^{(j)}$, can be computed as follows:

$$\mathbf{V}^{(j)} = \mathbb{E}(\mathbf{E}^\top \mathbf{z}_j \mathbf{z}_j^\top \mathbf{E}) \in \mathbb{R}^{T \times T} \quad ,$$

whose general term is given for $t, t' \in [T]$ by

$$\begin{aligned}
\mathbf{V}^{(j)}_{t,t'} &= \mathbb{E}(\mathbf{E}^\top_{\cdot,t}\mathbf{z}_j\mathbf{z}_j^\top\mathbf{E}_{\cdot,t'}) \\
&= \mathbb{E}(\mathbf{z}_j^\top\mathbf{E}_{\cdot,t'}\mathbf{E}^\top_{\cdot,t}\mathbf{z}_j) && \text{(scalar values commute)} \\
&= \mathbf{z}_j^\top\mathbb{E}(\mathbf{E}_{\cdot,t'}\mathbf{E}^\top_{\cdot,t})\mathbf{z}_j \\
&= \mathbf{z}_j^\top\mathbb{E}(\sum_{i=1}^{n}\mathbf{E}_{i,t'}\mathbf{E}^\top_{i,t})\mathbf{z}_j \\
&= \mathbf{z}_j^\top\sum_{i=1}^{n}\mathbb{E}(\mathbf{E}_{i,t'}\mathbf{E}^\top_{i,t})\mathbf{z}_j \ .
\end{aligned}$$

Then, the noise structure in Equation (2) yields $\mathbf{V}^{(j)}_{t,t'} = \mathbf{z}_j^\top n\mathbf{M}_{t,t'}\mathbf{z}_j = n\left\|\mathbf{z}_j\right\|^2\mathbf{M}_{t,t'}$.

Now, we show that with high probability $\|\boldsymbol{\Delta}\|_{2,1} = \mathrm{O}\left(\frac{s\lambda\sqrt{\log(p)}}{\kappa^2}\right)$. First, notice that:

$$\|\boldsymbol{\Delta}\|_{2,1} \leq \sqrt{n}\max_{k\neq j}|\mathbf{P}_{j,k}|\left\|\hat{\mathbf{B}}^{\mathrm{MTL}} - \mathbf{B}\right\|_{2,1} \ .$$

For a convenient choice of the regularization parameters $\boldsymbol{\alpha}$, using Bühlmann and van de Geer (2011, Lemma 2.1) and following the same approach as Dezeure et al. (2015, Appendix A.1), we obtain, with high probability:

$$\sqrt{n}\max_{k\neq j}|\mathbf{P}_{j,k}| = \mathrm{O}\left(\sqrt{\log(p)}\right) \ .$$

Bounds on $\|\hat{\mathbf{B}}^{\mathrm{MTL}} - \mathbf{B}\|_{2,1}$ are also available in the literature (Lounici et al., 2011) for $\rho = 0$ and can be extended to $\rho > 0$ similarly. Notably, provided $\rho = 0$, assuming A1 for a sparsity parameter $|\mathrm{Supp}(\mathbf{B}^*)| \leq s$, a given constant $\kappa = \kappa(s) > 0$, and a choice of λ large enough in Equation (4), (Lounici et al., 2011, Theorem 3.1) gives directly the following bound, with high probability:

$$\left\|\hat{\mathbf{B}}^{\mathrm{MTL}} - \mathbf{B}\right\|_{2,1} = \mathrm{O}\left(\frac{s\lambda}{\kappa^2}\right) \ .$$

□

Remark D.1. *Following van de Geer et al. (2014), to neglect $\boldsymbol{\Delta}$ we need to have $\|\boldsymbol{\Delta}\|_\infty = \mathrm{o}(1)$. This condition is verified if $s = \mathrm{o}\left(\frac{\kappa^2}{\lambda\sqrt{\log(p)}}\right)$.*

D.3 Proof of Prop. 2.2

Before starting the proof, let us give more precision on assumption A2, the complete assumption is the following:

(A2) there exists $\boldsymbol{\Gamma} \in \mathbb{R}^{C\times T}$ such that $\boldsymbol{\Gamma}_{r,\cdot} = \sum_{j\in G_r} w_j\mathbf{B}_{j,\cdot}$ with $w_j \geq 0$ for all $j \in [p]$, so that the associated compression loss $\mathbf{XB} - \mathbf{Z\Gamma}$ is bounded as follows:

$$\|\mathbf{XB} - \mathbf{Z\Gamma}\|^2_{2,2} \leq \xi\frac{T\phi^2_{\min}(\mathbf{M})}{n} = \xi\frac{T\phi^2_{\min}(\mathbf{R})\sigma^2}{n} \ , \tag{16}$$

where $\xi > 0$ is an arbitrary small constant, $\phi^2_{\min}(\mathbf{M}) > 0$ is the smallest eigenvalue of $\mathbf{M}$ and $\phi^2_{\min}(\mathbf{R}) > 0$ is the smallest eigenvalue of $\mathbf{R}$, the temporal correlation matrix of the noise defined by $\mathbf{R} = \mathbf{M}/\sigma^2$. The hypothesis plainly means that the noise induced by design matrix compression is small enough with respect to the model noise.

Now we give a proof of Prop. 2.2:

Proof. First, we derive the d-MTLasso for the compressed problem, for $r \in [C]$:

$$\hat{\mathbf{\Gamma}}^{(\mathrm{d-MTLasso})}_{r,.} = \frac{\mathbf{a}_r^\top \mathbf{Y}}{\mathbf{a}_r^\top \mathbf{Z}_{.,r}} - \sum_{l \neq r} \frac{\mathbf{a}_r^\top \mathbf{Z}_{.,l} \hat{\mathbf{\Gamma}}^{\mathrm{MTL}}_{r,.}}{\mathbf{a}_r^\top \mathbf{Z}_{.,r}} \quad , \tag{17}$$

where a_r's are the residuals obtained by nodewise Lasso on $\mathbf{Z}$ playing the same role as the z_j's in Equation (7). Then, as done in Appendix D.2, we derive:

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{\Gamma}}^{(\mathrm{d-MTLasso})}_{r,.} - \mathbf{\Gamma}_{r,.}) &= \sqrt{n}\frac{\mathbf{a}_r^\top \mathbf{E}}{\mathbf{a}_r^\top \mathbf{Z}_{.,r}} - \sum_{l \neq r} \frac{\sqrt{n}\, \mathbf{a}_r^\top \mathbf{Z}_{.,l}(\hat{\mathbf{\Gamma}}^{\mathrm{MTL}}_{l,.} - \mathbf{\Gamma}_{l,.})}{\mathbf{a}_r^\top \mathbf{Z}_{.,r}} + \frac{\sqrt{n}\, \mathbf{a}_r^\top (\mathbf{XB} - \mathbf{Z\Gamma})}{\mathbf{a}_r^\top \mathbf{Z}_{.,r}} \\ &= \mathbf{\Lambda}'_{r,.} + \mathbf{\Delta}'_{r,.} + \mathbf{\Pi}_{r,.} \quad ,\end{aligned} \tag{18}$$

We treat $\mathbf{\Lambda}'$ and $\mathbf{\Delta}'$ as in Appendix D.2, assuming that the hypotheses that are used to bound (hence, neglect) $\mathbf{\Delta}'$ are verified (notably A3).

Next, for $r \in [C]$, we want to establish that $\frac{n\|\mathbf{\Pi}_{r,.}\|^2_{\mathbf{M}^{-1}}}{T\hat{\mathbf{\Omega}}'_{r,r}}$ is negligible, *i.e.,* that $\mathbf{\Pi}$ has a negligible effect on all decision statistics, where the covariance $\hat{\mathbf{\Omega}}'$ has the following generic diagonal term:

$$\hat{\mathbf{\Omega}}'_{r,r} = \frac{n\,\|\mathbf{a}_r\|^2}{|\mathbf{a}_r^\top \mathbf{Z}_{.,r}|^2} \quad .$$

Given that

$$\|\mathbf{\Pi}_{r,.}\|^2_{\mathbf{M}^{-1}} = \frac{n\left\|\mathbf{a}_r^\top (\mathbf{XB} - \mathbf{Z\Gamma})\right\|^2_{\mathbf{M}^{-1}}}{|\mathbf{a}_r^\top \mathbf{Z}_{.,r}|^2} \tag{19}$$

$$\leq n\frac{\left\|\mathbf{a}_r^\top\right\|^2}{|\mathbf{a}_r^\top \mathbf{Z}_{.,r}|^2}\frac{\|\mathbf{XB} - \mathbf{Z\Gamma}\|^2_{2,2}}{\phi^2_{\min}(\mathbf{M})} \quad , \tag{20}$$

where $\|\cdot\|_{2,2}$ denotes the spectral norm. Then, we obtain that

$$\frac{n\,\|\mathbf{\Pi}_{r,.}\|^2_{\mathbf{M}^{-1}}}{T\hat{\mathbf{\Omega}}'_{r,r}} \leq \frac{n}{T}\frac{\|\mathbf{XB} - \mathbf{Z\Gamma}\|^2_{2,2}}{\phi^2_{\min}(\mathbf{M})} \leq \xi \quad . \tag{21}$$

Then, if A2 is verified for ξ small enough, we can also neglect $\mathbf{\Pi}$ in front of $\mathbf{\Lambda}'$.

Then, by neglecting $\mathbf{\Pi}$ and $\mathbf{\Delta}'$, we have:

$$\sqrt{n}(\hat{\mathbf{\Gamma}}^{(\mathrm{d-MTLasso})} - \mathbf{\Gamma}) \sim \mathcal{N}_C(\mathbf{0},\, \hat{\mathbf{\Omega}}'_{r,r}\mathbf{M}) \quad . \tag{22}$$

Then we can construct p-values that test the r-th null hypothesis $H_0^{(r)}$: "$\mathbf{\Gamma}_{j,.} = 0$", applying the same technique as in Sec. 2.4. By correcting these p-values —*e.g.,* using the Bonferroni correction (Dunn, 1961), we multiply by C the initial p-values—, we obtain cluster-wise corrected p-values that control the FWER.

Since, for all $r \in [C]$, $\mathbf{\Gamma}_{r,.}$ is a linear combination of $\mathbf{B}_{j,.}$ for $j \in G_r$, then $\mathbf{\Gamma}_{r,.} \neq 0$ if at least there exist $j \in G_r$ such that $\mathbf{B}_{j,.} \neq 0$.

Then, defining the feature-wise corrected p-values by the corrected p-values of the corresponding cluster, and assuming that clusters are at most of size δ, such corrected p-values control the δ-FWER.

□

Remark D.2. *In assumption A2, having a positive linear combination is not necessary, a simple linear combination is sufficient.*

However, we assumed that $\mathbf{\Gamma}_{r,.}$ was a positive linear combination of $\mathbf{B}_{j,.}$ for $j \in G_r$, to get the following desired properties:

"If additionally for $r \in [C]$, for all $j \in G_r$ and all $k \in G_r$, we have $\mathrm{sign}(\mathbf{B}_{j,.}) = \mathrm{sign}(\mathbf{B}_{k,.})$, then $\mathrm{sign}(\mathbf{\Gamma}_{r,.}) = \mathrm{sign}(\mathbf{B}_{j,.})$ (zero being booth positive and negative)."

This means that if all the features' weights in a cluster have the same sign, there exists a compression verifying A2 such that the cluster weight preserves the sign.

D.4 Proof of Prop. 2.3

Proof. Assuming the hypotheses of Prop. 2.3 and applying Prop. 2.2, we can, for each of the B compression of the problem in Equation (1), construct a corrected p-value family that control the δ-FWER. Applying the quantile aggregate method in Equation (15), we derive a corrected p-value family taking into account for each compression choice. Applying Meinshausen et al. (2009, Theorem 3.2), this aggregated corrected p-value family also controls the δ-FWER. □

E Computational aspects

Here we give some elements about the computational aspect of the algorithms we propose.

For solving Lasso or multi-task Lasso problems, we rely for additional speed-up on `celer`[6] (Massias et al., 2018a, 2019), a solver which is much more efficient than the standard coordinate descent (speed up by more than 10x on our experiments).

To compute d-MTLasso, we must solve p Lasso of size $(n, (p-1))$, and 1 multi-task Lasso with cross-validation on a dataset of size (n, p, T). For $n = 200$, $p = 7500$ and $T = 10$, the algorithms can be run on a standard laptop in around 10 hours (using only 1 CPU). However, the algorithm is embarrassingly parallel and requires around 15 minutes if run on a machine with 50 CPUs. To compute cd-MTLasso, we must solve C Lasso of size $(n, (C-1))$. and 1 multi-task Lasso with cross-validation on a dataset of size (n, C, T). For $n = 200$, $C = 1000$ and $T = 10$, it can be run on a standard local device in less than 1 minute (using only 1 CPU). Finally, to compute ecd-MTLasso, we must solve B cd-MTLasso. For $B = 100$ (25 is already a good value to get most of the advantages of ensembling), $n = 200$, $C = 1000$ and $T = 10$, it can be run on a standard laptop in around 1 hour (using only 1 CPU) and around 1 minute on a machine with 50 CPUs.

Although, when using coordinate-descent-like algorithms, the complexity depends on solver parameters such as tolerance on stopping criteria, the complexity in C (or p) appears empirically to be cubic, while it is linear in n and T. It is also linear in B.

F Detailed data description

For AEF and VEF, data contained one artifactual channel leading to $n = 203$, while for SEF data were preprocessed for removal of environmental noise leading to an effective number of samples of $n = 64$ (Taulu, 2006). For the AEF dataset, we report results for AEFs evoked by left auditory stimulation with pure tones of 500 Hz. The analysis window for source estimation was chosen from 50 ms to 200 ms based on visual inspection of the evoked data to capture the dominant N100m component, leading to $T = 6$. For the SEF dataset, we analyzed SEFs evoked by bipolar electrical stimulation (0.2 ms in duration) of the left median nerve. To capture the main peaks of the evoked response and exclude the strong stimulus artifact, the analysis window was chosen from 18 ms to 200 ms based on visual inspection of the sensor signal.

Preprocessing was done following the standard pipeline from the MNE software (Gramfort et al., 2014). Baseline correction using pre-stimulus data (from -200 ms to 0 ms) was used. Epochs with peak-to-peak amplitudes exceeding predefined rejection parameters (3 pT for magnetometers and 400 pT/m for gradiometers, and 150 μV for EOG on AEF and VEF and 350 μV for SEF) were assumed to be affected by artifacts and discarded. This resulted in 55 (AEF), 67 (SEF) and 111 (SEF) artifact-free measurements which were average to produce the target matrix $\mathbf{Y}$. The gain matrix was computed using a set of $p = 7498$ cortical locations, and a three-layer boundary element model.

G Related Work

The topic of high-dimensional inference has been addressed in many recent works. Yet, to the best of our knowledge, none of this literature has been applied to the source localization problem we consider here.

[6] https://github.com/mathurinm/CELER

- The idea of associating clustering with high-dimensional inference can also be found in recent works with application to genetic data: Bühlmann et al. (2013) has used a fixed clustering step, which is made adaptive in Mandozzi and Bühlmann (2016). Our contribution deviates from these works in two regards: unlike Bühlmann et al. (2013), we do not consider that a fixed clustering, however good it is, indeed captures the essence of the problem: this is why we resort to an ensemble of different clustering solutions. Unlike Mandozzi and Bühlmann (2016), we do not try to narrow down the inference in a hierarchical fashion, because we do not consider that source imaging can in effect be traced down to the vertex level: given the difficulty of the source imaging problem, we find it more satisfactory to outline a region of putative activity.
- Another family of inference methods based on sample splits has been introduced by Meinshausen et al. (2009): train data are used to select regions, test data to assess their statistical significance. The choice of splits can be varied and aggregated upon to mitigate the impact of arbitrary splits selection. However, data splitting has a high cost in terms of statistical power, making these approaches weakly sensitive Taylor and Tibshirani (2015).
- An alternative method yielding family-wise error rate (FWER) control is the stability selection method, that builds on bootstrapped randomized sparse regression Meinshausen and Bühlmann (2010). Yet, this approach has been found too weakly sensitive and it has not been considered in further statistical inference works, see *e.g.,* Dezeure et al. (2015).
- Post-selection inference Taylor and Tibshirani (2015) is an approach that typically relies on a sparse estimator (such as Lasso) and then assesses the significance of the selected variables. It accounts for the selection in the inference process, avoiding the undesirable bias of selecting and testing on the same data. However, we have not not found an implementation that scales in a numerically sound way to the problem size that we are considering here: thousand features, even after clustering.
- Knockoff inference (with or without clustering) is probably the most recent alternative developed for high-dimensional inference (Barber and Candès, 2019): it consists in appending noisy copy of the problem features and selecting only variables that are much more significantly associated than their noisy copy. While this approach is computationally relevant for the problem at hand, it suffers from the arbitrary knockoff variable set used; it yields a control of the false discovery rate of the detection problem, that is not directly comparable with the family-wise error rate (FWER) considered here. FWER control is possible with knockoff Janson and Su (2016), yet very weakly sensitive.

H Supplementary figures

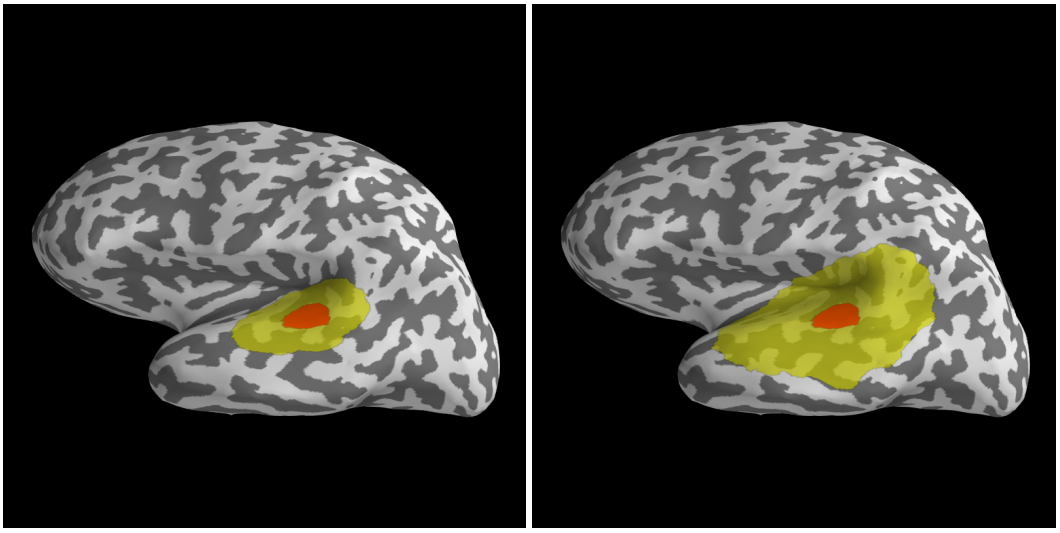

Figure 5: **Illustrating spatial tolerance of size $\delta = 20$ mm and $\delta = 40$ mm.** The true source in red has a 10 mm radius (distance measured on the cortical surface) and the spatial tolerance extend this region by 20 mm on the left side and 40 mm on the right side in yellow. The δ-FWER is the probability of making false discoveries outside of the extended region. Then, a false discovery made in the yellow region is not counted neither as an error nor a true positive.

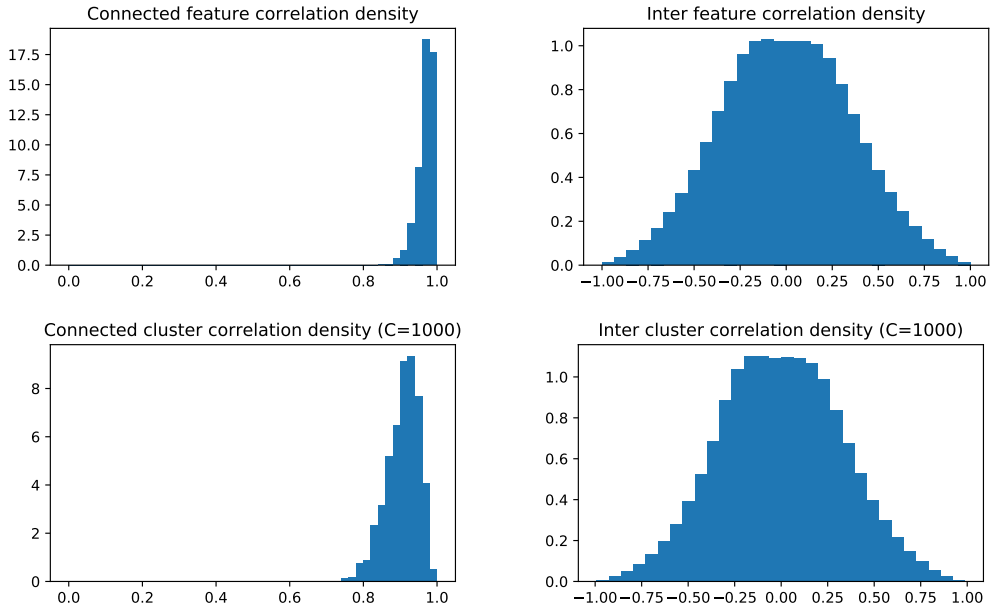

Figure 6: **Illustrating correlation in MNE sample MEG data.** (left): Distribution of the maximum correlation between a feature (resp. cluster) and another connected feature (resp. cluster). (Top) the maximum connected feature correlation is close to 0.98 in average. (Bottom) the maximum connected cluster correlation is lower, close to 0.9 on average. Clustering improves conditioning significantly. (right): The density of the inter feature correlation (top) looks similar to the density of the inter cluster correlation (bottom). By focusing the extreme values of correlation, we see a little decrease of extreme values for the clustered data.

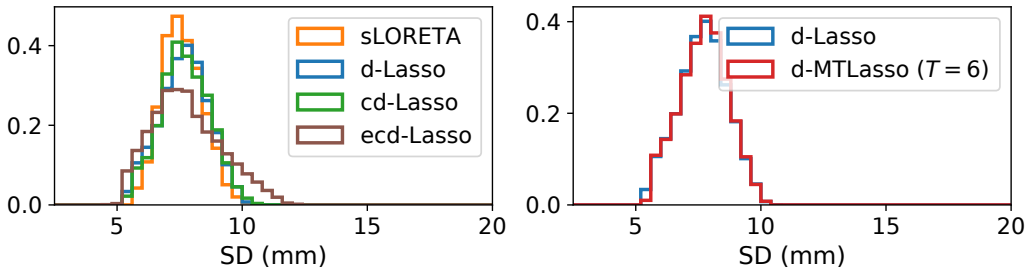

Figure 7: **Spatial Dispersion (SD) histograms.** (left): SD on a fixed time point (Hauk et al., 2011). All methods lead to comparable spatial dispersion. (right): SD for desparsified multi-task Lasso (d-MTLasso) with increasing time points. See Figure 2 for PLE histograms on the same experiments.

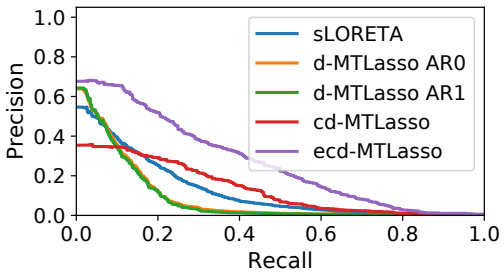

Figure 8: **Precision-Recall.** See Figure 3 for $\delta-$Precision-Recall curves computed on the same data.

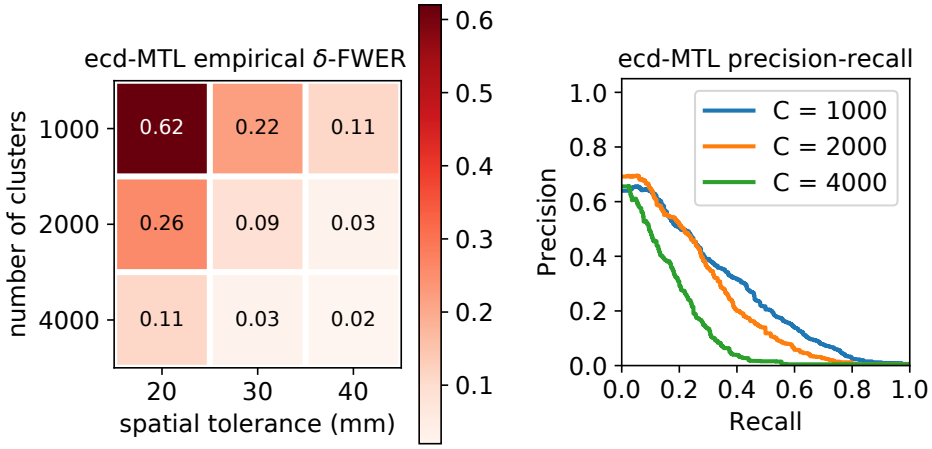

Figure 9: **ecd-MTLasso empirical δ-FWER and precision recall for different choice of cluster sizes.** (left): Running the same simulation as in Sec. 3.2, we observe that the spatial tolerance δ can be reduced to 20 mm by increasing the number of clusters up to 4000. With $C = 1000$ clusters (resp. $C = 2000$, $C = 4000$), the average cluster diameter is around 18 mm (resp. 13 mm and 9 mm). It turns out that the δ-FWER is controlled for around twice the diameter (if the compressed design matrix verifies assumption A1). (right): We see that this decrease in spatial tolerance comes with a price regarding support recovery: the precision-recall curve declines with when C is increased. (both): Note that we need to set the hyper-parameter c that is used to compute the regularization parameters $\boldsymbol{\alpha}$ (see note coming with Equation (5)). We found empirically that it should be inversely proportional to C: for $C = 1000$, $c = 0.5\%$; for $C = 2000$, $c = 0.25\%$; for $C = 4000$, $c = 0.15\%$.

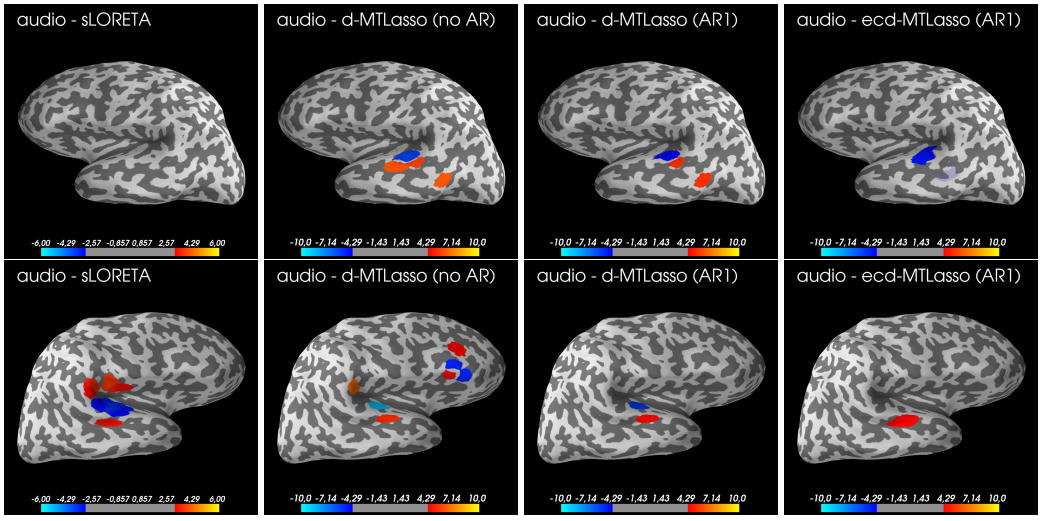

Figure 10: **Comparison on audio dataset on both hemispheres.** From left to right are compared sLORETA, d-MTLasso without AR modeling (noise is assumed non-autocorrelated), d-MTLasso with an AR1 noise model and the ecd-MTLasso using also an AR1. The results correspond to auditory (top) evoked fields. Colormaps are fixed across datasets and adjusted based on meaningful statistical thresholds in order to outline FWER control issues.

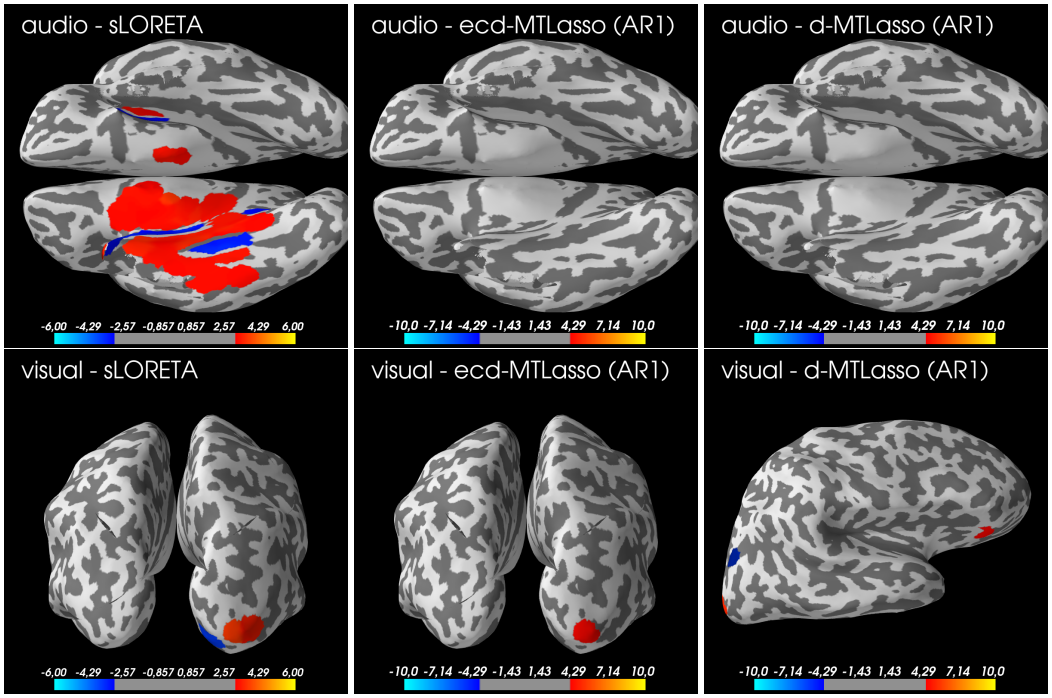

Figure 11: **Results on real data keeping only EEG sensors.** Auditory activations (top) have historically been hard to infer with EEG sensors: sLORETA produces only false discoveries while ecd-MTL and d-MTL make no discoveries. In the visual experiment (bottom): sLORETA and ecd-MTL produce expected patterns, d-MTL produces expected patterns plus one false discovery in the frontal lobe. In our work, we have emphasized MEG experiments: they offer more sensors compared to EEG leading to improved statistical power.