



## Decoding with Confidence: Statistical Control on Decoder Maps

Jérôme-Alexis Chevalier, Tuan-Binh Nguyen, Joseph Salmon, Gaël Varoquaux, Bertrand Thirion

### ► To cite this version:

Jérôme-Alexis Chevalier, Tuan-Binh Nguyen, Joseph Salmon, Gaël Varoquaux, Bertrand Thirion. Decoding with Confidence: Statistical Control on Decoder Maps. NeuroImage, Elsevier, 2021, pp.117921. 10.1016/j.neuroimage.2021.117921 . hal-03179728

HAL Id: hal-03179728

<https://hal.archives-ouvertes.fr/hal-03179728>

Submitted on 24 Mar 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

<sup>1</sup> Decoding with Confidence:  
<sup>2</sup> Statistical Control on Decoder Maps

<sup>3</sup> Jérôme-Alexis Chevalier<sup>1,2,3</sup>, Tuan-Binh Nguyen<sup>1,2,3,4</sup>, Joseph Salmon<sup>5</sup>,  
Gaël Varoquaux<sup>1,2,3</sup>, Bertrand Thirion<sup>1,2,3</sup>

[jerome-alexis.chevalier@inria.fr](mailto:jerome-alexis.chevalier@inria.fr)

<sup>4</sup> <sup>1</sup> Parietal project-team, Inria Saclay-Ile de France, Palaiseau, France

<sup>5</sup> <sup>2</sup> CEA/Neurospin bat 145, Gif-Sur-Yvette, France

<sup>6</sup> <sup>3</sup> Université Paris-Saclay, Gif-Sur-Yvette, France

<sup>7</sup> <sup>4</sup> LMO, Université Paris-Saclay, Orsay, France

<sup>8</sup> <sup>5</sup> IMAG, Université de Montpellier, CNRS, Montpellier, France

<sup>9</sup> <sup>10</sup> Keywords: fMRI, Decoding, Statistical Methods, Multivariate Model, Inference, Statistical Control, Support Recovery, High Dimension.

<sup>11</sup> **Abstract**

<sup>12</sup> In brain imaging, decoding is widely used to infer relationships between brain  
<sup>13</sup> and cognition, or to craft brain-imaging biomarkers of pathologies. Yet, standard  
<sup>14</sup> decoding procedures do not come with statistical guarantees, and thus do not give  
<sup>15</sup> confidence bounds to interpret the pattern maps that they produce. Indeed, in  
<sup>16</sup> whole-brain decoding settings, the number of explanatory variables is much greater  
<sup>17</sup> than the number of samples, hence classical statistical inference methodology cannot  
<sup>18</sup> be applied. Specifically, the standard practice that consists in thresholding decoding  
<sup>19</sup> maps is not a correct inference procedure. We contribute a new statistical-testing  
<sup>20</sup> framework for this type of inference. To overcome the statistical inefficiency of  
<sup>21</sup> voxel-level control, we generalize the Family Wise Error Rate (FWER) to account  
<sup>22</sup> for a spatial tolerance  $\delta$ , introducing the  $\delta$ -Family Wise Error Rate ( $\delta$ -FWER).  
<sup>23</sup> Then, we present a decoding procedure that can control the  $\delta$ -FWER: the Ensemble  
<sup>24</sup> of Clustered Desparsified Lasso (EnCluDL), a procedure for multivariate statistical  
<sup>25</sup> inference on high-dimensional structured data. We evaluate the statistical properties  
<sup>26</sup> of EnCluDL with a thorough empirical study, along with three alternative procedures  
<sup>27</sup> including decoder map thresholding. We show that EnCluDL exhibits the best  
<sup>28</sup> recovery properties while ensuring the expected statistical control.

<sup>29</sup> **1 Introduction**

<sup>30</sup> Predicting behavior or diseases status from brain images is an important analytical  
<sup>31</sup> approach for imaging neurosciences, as it provides an effective evaluation of the infor-  
<sup>32</sup> mation carried by brain images. Machine learning tools, mostly supervised learning, are

indeed used on brain images to infer cognitive states [Haynes and Rees, 2006, Norman et al., 2006] or to perform diagnosis or prognosis [Demirci et al., 2008, Fan et al., 2008]. Brain images are obtained from MRI or PET imaging, or even EEG- or MEG-based volume-based activity reconstruction. They are used to predict a *target* outcome: binary (*e.g.*, two-condition tasks), discrete (*e.g.*, multiple-condition tasks) or continuous (*e.g.*, age). The *decoding models* used for such predictions are most often linear models, characterized by a weight map that can be represented as a brain image [Mourao-Miranda et al., 2005, Varoquaux and Thirion, 2014].

Besides the prediction accuracy achieved, this estimated weight map is crucial to assess the information captured by the model. Typically, the produced weight maps are used to identify discriminative patterns [Haxby et al., 2001, Mourao-Miranda et al., 2005, Gramfort et al., 2013] and support reverse inferences [Poldrack, 2011, Schwartz et al., 2013, Varoquaux et al., 2018], *i.e.*, conclude on the implication of brain regions in the studied process.

Unlike in standard analysis —statistical parametric mapping [Poldrack et al., 2011, chap 7]—, in decoding the feature importance is tested *conditional on other brain features*, *i.e.*, it assesses whether each feature *adds* to information conveyed by other features. Weichwald et al. [2015] highlight the fact that decoding, *i.e.*, multivariate or conditional analysis, and encoding, *i.e.*, univariate or marginal analysis, are complementary. They notably argue that taking the two perspectives is essential for causal interpretation regarding the implication of brain regions in the target outcome (see also Haufe et al. [2014]).

While decoding optimizes the prediction of a target outcome, little or nothing can be concluded about the significant features of weight maps. Indeed, those maps do not come with well-controlled statistical properties, making decoding models hard to interpret. For instance, considering linear Support Vector Machines (SVM) [Cortes and Vapnik, 1995] or linear Support Vector Regression (SVR) [Smola and Schölkopf, 2004], that are popular in neuroimaging [Pereira et al., 2009, Rizk-Jackson et al., 2011], a natural way to recover predictive regions from their weight maps is to threshold these maps (*e.g.* Mourao-Miranda et al. [2005], Rehme et al. [2015], Sato et al. [2013], Lee et al. [2010]). However, this approach is problematic for two reasons: there exists no clear way to choose the threshold as a function of a desired significance, and it is unclear whether such a thresholded map is still an accurate predictor of the outcome. Solutions that bypass the arbitrary threshold choice have been proposed, such as Recursive Feature Elimination (RFE) [De Martino et al., 2008], but the produced maps still lack statistical guarantees.

In this work, we show that the natural procedure that consists in thresholding standard decoders, such as SVR, is not a relevant solution. In this respect, we consider two thresholding strategies: one that keeps extreme weights, and another one that computes the threshold by performing a permutation test. Unlike RFE, these two thresholding strategies can be derived from statistical testing considerations —yet, these statistical properties are not assumption free. We also consider decoders that provide confidence intervals around the estimated weight map. As detailed in the next section, these ap-

76 proaches also face severe challenges in terms of statistical power and computational  
77 tractability. They have to rely on algorithmic shortcuts, approximations and hypotheses  
78 that are more or less problematic in practice.

79 Hence, for all methods considered, the control of false detections is only achieved  
80 within a certain theoretical framework, and given a series of assumptions that are not  
81 easily checked. It is thus fundamental to analyze their statistical behavior with an  
82 extensive empirical study. We present here a set of experiments assessing the accuracy  
83 of the error rate control and support recovery on real and semi-synthetic brain-imaging  
84 data.

85 Additionally, to achieve a reasonable compromise between error control and power,  
86 we introduce a new type of error control adapted to imaging problems. The proposed  
87 quantity is a generalization of the Family Wise Error Rate (FWER) [Hochberg and  
88 Tamhane, 1987] including a spatial tolerance parametrized by a distance  $\delta$ . We call it  
89  $\delta$ -FWER.

90 In Section 2, we bring useful background, discuss the statistical guarantees that we  
91 aim at for pattern maps, and make the theoretical and practical inference challenges  
92 explicit. In Section 3 we provide a definition of the  $\delta$ -FWER along with a geometrical  
93 interpretation of this quantity. We also describe several statistical inference methods  
94 producing statistical maps reflecting the significance of conditional association of brain  
95 regions with a target, while controlling the FWER or  $\delta$ -FWER. Section 4 and Section 5  
96 follow with extensive experiments on simulations and large-scale fMRI datasets that  
97 study the behavior of the benchmarked solutions regarding false positive control and  
98 recovery.

## 99 2 Context: decoding-map recovery

100 In this section, we first review a result due to Weichwald et al. [2015] about the com-  
101plementarity of univariate and multivariate inference, then we present the statistical  
102 guarantees that we aim at for on brain-wide decoding maps, lastly we formalize the  
103 problem of statistical inference on such maps.

### 104 2.1 Complementarity of univariate and multivariate inference

105 Statistical inference in neuroimaging can be performed using a mass univariate model-  
106 ing, *i.e.*, fitting brain activity maps from an outcome —leading to *encoding models*—  
107 or by predicting an outcome from brain maps using multivariate modeling —leading to  
108 *decoding models*. The complementarity of univariate and multivariate analyses has been  
109 demonstrated in Weichwald et al. [2015]. Specifically, they argued: “We showed that only  
110 encoding models in a stimulus-based setting support unambiguous causal statements.  
111 This result appears to imply that decoding models, despite their gaining popularity in  
112 neuroimaging, are of little value for investigating the neural causes of cognition. In the  
113 following, we argue that this is not the case. Specifically, we show that by combining  
114 encoding and decoding models, we gain insights into causal structure that are not pos-

115 sible by investigating each type of model individually.” This statement clearly implies  
 116 that inference tools are needed for multivariate analysis. The present work is thus fully  
 117 dedicated to multivariate inference. We simply provide some univariate inference results  
 118 for reference, given that they address different yet complementary questions.

## 119 2.2 Statistical control with spatial tolerance

120 In decoding, the signals from voxels are used concurrently to predict an outcome. Given  
 121 that they display high correlations, trying to identify the effect of each covariate (voxel)  
 122 is not possible. Precise voxel-level control may not be necessary: current brain models  
 123 are rather specified at a regional scale, see *e.g.*, [Glasser et al., 2016]. Additionally, to  
 124 control a statistical error, detecting a voxel adjacent to a truly predictive region is less  
 125 problematic than detecting a false positive far from such a predictive region. These  
 126 two facts argue in favor of incorporating a spatial tolerance in the sought statistical  
 127 control, as with efforts in standard analysis [Smith and Nichols, 2009, Da Mota et al.,  
 128 2014, Bowring et al., 2019]. Hence, we introduce a generalization of the Family Wise  
 129 Error Rate (FWER) [Hochberg and Tamhane, 1987]: the  $\delta$ -FWER. This generalization  
 130 is related to the extension of the False Discovery Rate (FDR) [Benjamini and Hochberg,  
 131 1995] proposed by Nguyen et al. [2019] and Gimenez and Zou [2019], called  $\delta$ -FDR and  
 132 local-FDR, respectively.

## 133 2.3 Formal problem setting

134 **Notation.** For clarity, we use bold lowercase for vectors and bold uppercase for  
 135 matrices. For  $p \in \mathbb{N}$ , we write  $[p]$  for the set  $\{1, \dots, p\}$ . For a vector  $\mathbf{w}$ ,  $\mathbf{w}_j$  refers to its  $j$ -th  
 136 coordinate. For a matrix  $\mathbf{X}$ ,  $\mathbf{X}_{i,j}$  refers to the element in the  $i$ -th row and  $j$ -th column.

137 **Formalizing the decoding problem.** The target (outcome to decode) is observed in  
 138  $n$  samples and denoted by  $\mathbf{y} \in \mathbb{R}^n$  ( $\mathbf{y}$  can be binary, discrete or continuous). The brain  
 139 volume is discretized into  $p$  voxels. The corresponding  $p$  voxel signals are also referred  
 140 to as explanatory variables, covariates or features. We denote by  $\mathbf{X} \in \mathbb{R}^{n \times p}$  the matrix  
 141 containing (column-wise) the  $p$  covariates  $\{\mathbf{X}_1, \dots, \mathbf{X}_p\}$ . We assume that, for all  $i \in [n]$ ,  
 142 the samples  $(\mathbf{y}_i, \mathbf{X}_{i,.})$  are i.i.d. Then, further assuming a linear dependency between the  
 143 covariates and the response, the generative model is as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\varepsilon}, \quad (1)$$

144 where  $\mathbf{w}^* \in \mathbb{R}^p$  is the true weight map and  $\boldsymbol{\varepsilon}$  is the noise vector. In the present study,  
 145 we assume for simplicity that the noise is Gaussian, *i.e.*,  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$ , but extension  
 146 to sub-Gaussian noise is possible.

147 **High dimensionality and structure of the data.** Given  $\mathbf{X}$  and  $\mathbf{y}$ , a standard  
 148 procedure computes an estimate  $\hat{\mathbf{w}}$  of  $\mathbf{w}^*$ . Getting statistical guarantees on  $\mathbf{w}_j^*$ ,  $j \in [p]$ ,  
 149 means assessing with some degree of uncertainty that  $\mathbf{w}_j^*$  is non-zero, or equivalently,

150 giving a confidence interval for  $\mathbf{w}_j^*$ . This is hard in high dimension and when short-  
151 and long-range correlations are present in the data. Indeed, for brain imaging data,  $n$   
152 is typically hundreds (or less), whereas  $p$  may amount to hundreds of thousands. In  
153 addition, voxel signals are highly correlated, which makes model identification harder  
154 due to multicollinearity and ill-posedness. Theoretical studies, *e.g.*, Wainwright [2009],  
155 have revealed that in such settings there is no hope to *recover* completely and accurately  
156 the predictive regions.

## 157 2.4 Current practices: thresholding decoding maps

158 **Uniform threshold.** Probably the most natural procedure used to recover discriminative  
159 patterns is to threshold decoders with high prediction performance —a popular  
160 choice is the linear SVM/SVR decoder [Pereira et al., 2009, Rizk-Jackson et al., 2011].  
161 Thresholding decoder maps at a uniform value —*i.e.*, the threshold is the same for all  
162 weights— is probably the most common practice in neuroimaging; threshold value being  
163 generally arbitrary: "naked-eye criteria". It is not thought of as a statistical operation,  
164 and is sometimes left to the reader, who is presented unthresholded maps and yet told  
165 to interpret only the salient features of these maps.

166 Permutation testing can also be used to derive a uniform threshold with explicit guarantees.  
167 The classical Westfall-Young permutation test procedure [Westfall and Young,  
168 1993] is well-known in the univariate context to control the FWER [Anderson, 2001],  
169 but its application to multivariate testing is not as straightforward. Then, instead of  
170 considering the usual  $t$ -statistics, a permutation test can use the linear SVR weights.  
171 An estimated weight map must be computed for the original problem and for several  
172 permuted problems before performing the Westfall-Young procedure; this method is  
173 detailed in Sec. 3.3.

174 Under some assumptions (see Sec. 3.2 and Sec. 3.3) that are more or less problematic  
175 in practice, the uniform thresholding strategies might recover the predictive patterns  
176 with FWER control. However, we will see that these naive strategies are not satisfactory  
177 in practice.

178 **Non-uniform threshold.** Another method proposed by Gaonkar and Davatzikos  
179 [2012], specifically designed for neuroimaging settings, relies on the analytic approxima-  
180 tion of a permutation test performed over a linear SVM/SVR estimator. This method  
181 computes confidence intervals around the weights of the proposed estimator. Then, un-  
182 der some assumptions (see Sec. 3.4) that are not always met in practice, this procedure  
183 controls the FWER. It is almost equivalent to thresholding the SVR weights with a  
184 non-uniform threshold —*i.e.*, the threshold is specific to each weight. We refer to it as  
185 Adaptive Permutation Threshold SVR (Ada-SVR) from now on.

## 186 2.5 Building decoders designed for statistical control

187 **Dimension reduction by voxel grouping.** A computationally attractive solution to  
188 alleviate high dimensionality is to leverage the data structure and group adjacent —and

correlated— voxels, producing a closely related, yet compressed version of the original problem. In decoding, the grouping of voxels via spatially-constrained clustering algorithms has already been used to reduce the problem dimension [Gramfort et al., 2012, Varoquaux et al., 2012, Wang et al., 2015]. Specifically, groups of contiguous voxels can be replaced by the average signal they carry, reducing the dimensionality while improving the conditioning of the estimation problem. However, such a compression introduces a bias, as the patterns are constrained by the clusters shape. This bias is problematic as there is no unique grouping or clustering of the voxels [Thirion et al., 2014]: many different groupings capture the signal as accurately. One way to mitigate this bias is to use aggregation of models [Breiman, 1996, Zhou, 2012] obtained from several voxel groupings. Varoquaux et al. [2012] implemented this idea by computing different groupings from different random subsamples of the full data sample. The corresponding procedure yields decoders with more stable maps as well as a better prediction accuracy. In this subsampling spirit, random subspace methods [Ho, 1998, Kuncheva et al., 2010, Kuncheva and Rodríguez, 2010] also improve the prediction accuracy with more stable solutions—but in this case the subsampling is performed on the raw features. More recently, a procedure, *Fast Regularized Ensembles of Models* (FReM) [Hoyos-Idrobo et al., 2018], has combined clustering and ensembling to reduce the variance of the weight map, while ensuring high prediction accuracy. Yet, FReM weight maps do not enjoy statistical guarantees.

**High-dimensional statistics tools.** There have been a variety of procedures to produce p-value maps (map of p-values associated to every covariate) for linear models in high dimension [Wasserman and Roeder, 2009, Meinshausen et al., 2009, Bühlmann, 2013, Zhang and Zhang, 2014, Javanmard and Montanari, 2014]. Yet, they are not directly applicable to brain-imaging settings, as the dimensionality is too high. Based on a comparative review of those procedures [Dezeure et al., 2015], we have focused on the so-called Desparsified Lasso (DL), introduced in Zhang and Zhang [2014] and thoroughly analyzed by van de Geer et al. [2014]. Roughly, Desparsified Lasso can be seen as a Lasso-type [Tibshirani, 1996] extension of the least-squares to high dimensional settings, producing weight maps with well-controlled statistical distribution.

However, when the number  $p$  of features is much greater than the number  $n$  of samples, Desparsified Lasso lacks statistical power [Chevalier et al., 2018] and the computational cost becomes prohibitive. Indeed, solving Desparsified Lasso entails solving  $p$  Lasso problems with a design matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$ . Using the standard coordinate descent implementation [Friedman et al., 2007] the computation time is  $\mathcal{O}(Tnp^2)$ , with  $T$  the number of epochs used to solve the Lasso. However, when  $p$  is of order of few thousands and  $n$  few hundreds, Desparsified Lasso remains feasible with modest computer resources. In this context, the recently proposed Ensemble of Clustered Desparsified Lasso (EnCluDL) [Chevalier et al., 2018] combines three steps: a clustering procedure that reduces the problem dimension but preserves data structure, the Desparsified Lasso procedure that is tractable on the compressed problem, and an ensembling method introduced by Meinshausen et al. [2009] that aggregates several solutions of the compressed

problem. This method, summarized in Sec. 3.5, follows a scheme similar to FReM but the inference and ensembling procedures are different since it aims at producing p-value maps with statistical properties. Indeed, under some assumptions (see Sec. 3.5), it can be shown that EnCluDL controls the  $\delta$ -FWER at the desired nominal level.

Finally, Knockoff filters [Barber and Candès, 2015, Candès et al., 2018], extended to work on images by Nguyen et al. [2019], are also an appealing procedure, though they can only control the FDR [Barber and Candès, 2015] or a relaxed version of the FWER [Janson and Su, 2016] incompatible with our spatial control, the  $\delta$ -FWER detailed below. In this study, following the previous work of Chevalier et al. [2018], we focus on FWER or  $\delta$ -FWER control. We then defer the extension of EnCluDL to FDR-controlling procedures and the benchmarking with alternatives to future work.

### 3 Materials and methods

#### 3.1 $\delta$ -Family Wise Error Rate ( $\delta$ -FWER)

In this section, we introduce a new way of controlling false detections that is well suited for neuroimaging settings as it incorporates spatial tolerance.

**True support under linear model assumption.** When considering multivariate inference, the *support*  $S \subset [p]$  is the set of covariates that are non-independent of  $\mathbf{y}$  conditionally to the other covariates. The rest of the voxels form the *null region*  $N$  i.e.,  $N = [p] \setminus S$ . Formally,  $S$  is the unique set that verifies:

$$\begin{aligned} \forall j \in S, \quad & \mathbf{X}_j \not\perp\!\!\!\perp \mathbf{y} \mid \{\mathbf{X}_k, k \in [p] \setminus \{j\}\} , \\ \forall j \in N, \quad & \mathbf{X}_j \perp\!\!\!\perp \mathbf{y} \mid \{\mathbf{X}_k, k \in S\} , \end{aligned} \tag{2}$$

where the sign  $\perp\!\!\!\perp$  denotes independence. Under the linear assumption made in (1),  $S$  becomes simply the set of non zero weights and  $N$  the set of zero weights:

$$\begin{aligned} S &= \{j \in [p] : w_j^* \neq 0\} , \\ N &= \{j \in [p] : w_j^* = 0\} . \end{aligned} \tag{3}$$

**$\delta$ -neighborhood.** The variables  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$  can also be characterized by the spatial proximity of their underlying voxels in brain space: given  $\delta \geq 0$ , a voxel  $k \in [p]$  is in the  $\delta$ -neighborhood of a voxel (or a set of voxels) if their distance is less than  $\delta$ .

**$\delta$ -null region.** For  $\delta \geq 0$ , we denote by  $S^{(\delta)}$  the  $\delta$ -dilation of the support  $S$ , i.e., the set of voxels in  $S$  or in its  $\delta$ -neighborhood. By definition,  $S \subset S^{(\delta)}$ . We denote by  $N^{(-\delta)}$  the  $\delta$ -erosion (inverse operation of a  $\delta$ -dilation) of the null region  $N$ , implying that  $N^{(-\delta)} \subset N$ . From the definition of  $N$  we have immediately:

$$N^{(-\delta)} = [p] \setminus S^{(\delta)} , \tag{4}$$

259 We refer to  $N^{(-\delta)}$  as the  $\delta$ -null region. As shown in Fig. 1, we interpret the  $\delta$ -null region  
 260 as the subset of the covariates which are at a distance less than  $\delta$  from the support  
 261 covariates. We also give a practical example of the  $\delta$ -null region in the case of real fMRI  
 262 data in appendix in Fig. 14.

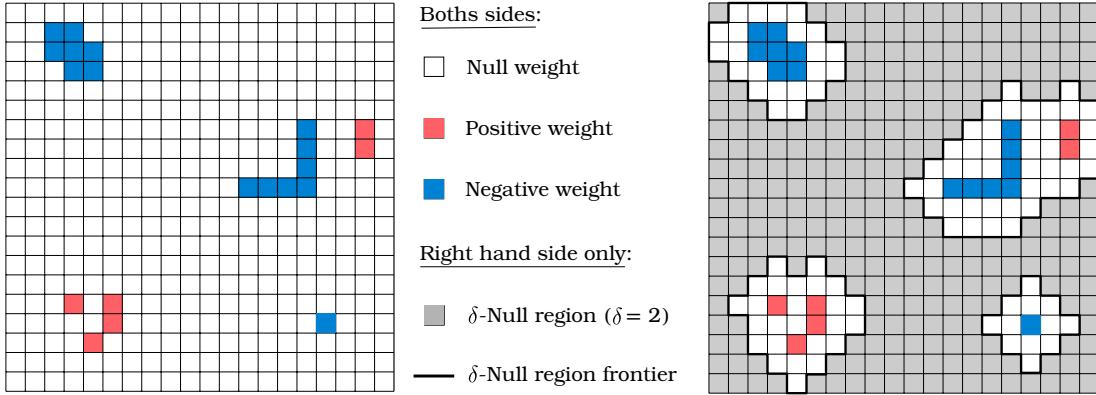


Figure 1: **Spatial tolerance to false discoveries.** Left: example of 2D-weight map, small squares represent voxels. The map is sparse. Right: representation of the  $\delta$ -null region for the associated map with  $\delta = 2$ . The covariates in the  $\delta$ -null region are "far" from non-null covariates, discoveries in this area are highly undesired. Discovering a null covariate "close" to a non-null covariate is tolerated.

263  **$\delta$ -Family Wise Error Rate ( $\delta$ -FWER).** If we have an estimate of the support  
 264  $\hat{S} \subset [p]$ , we recall that the Family Wise Error Rate (FWER) is defined as the probability  
 265 of making a false detection [Hochberg and Tamhane, 1987]:

$$\text{FWER}(\hat{S}) = \mathbb{P}(\hat{S} \cap N \neq \emptyset) . \quad (5)$$

266 Similarly, given  $\delta \geq 0$ , we defined the  $\delta$ -FWER to be

$$\delta\text{-FWER}(\hat{S}) = \mathbb{P}(\hat{S} \cap N^{(-\delta)} \neq \emptyset) , \quad (6)$$

267 *i.e.*, the probability of making a detection at distance more than  $\delta$  from the true support.  
 268 The  $\delta$ -FWER control is thus weaker than the FWER control, except when  $\delta = 0$  and  
 269 when the true support is empty (*i.e.*,  $N = [p]$ ), in which case the  $\delta$ -FWER coincides  
 270 with the classical FWER.

### 271 3.2 Thresholded SVR (Thr-SVR)

272 In this section, we introduce Thresholded SVR (Thr-SVR), a procedure that thresholds  
 273 uniformly the estimated SVR weight map, keeping extreme weights; this method corre-  
 274 sponds to the most standard and simple approach to recover predictive patterns. The  
 275 first step is to derive the SVR weights  $\hat{\mathbf{w}}^{\text{SVR}}$ . Then, assuming that the estimated weights

276 of the null region are sampled from a given distribution centered on 0, the corresponding  
 277 standard deviation  $\sigma_{\text{SVR}}$  can be approximated with the following estimator:

$$\hat{\sigma}_{\text{SVR}} = \sqrt{\frac{1}{p} \sum_{j=1}^p (\hat{\mathbf{w}}_j^{\text{SVR}})^2} . \quad (7)$$

278 We could also consider other estimators to approximate this quantity (*e.g.*, Schwartzman  
 279 et al. [2009]) but the former is simple and at worst biased upward when the support is  
 280 not empty. Now, assuming a Gaussian distribution for the SVR weights in the null  
 281 region, *i.e.*, for  $j \in N$ :

$$\hat{\mathbf{w}}_j^{\text{SVR}} \sim \mathcal{N}(0, \sigma_{\text{SVR}}^2) , \quad (8)$$

282 we can produce (corrected) p-values by applying a Bonferroni correction. The produced  
 283 p-values are at worst conservative under the two assumptions discussed in Section 6.  
 284 In this procedure, the regression method considered is a linear SVR but similar results  
 285 were obtained with other procedures (*e.g.*, Ridge regression).

### 286 3.3 Permutation Test SVR (Perm-SVR)

287 Now, we introduce another uniform thresholding strategy of SVR weights based upon a  
 288 permutation test procedure. To derive corrected p-values from a permutation test, we  
 289 first regress the design matrix against the response vector using a linear SVR to obtain  
 290 an estimate  $\hat{\mathbf{w}}^{\text{SVR}}$  of the weights map similarly as made in the Thr-SVR procedure.  
 291 Then, permuting randomly  $R$  times the response vector and regressing the design matrix  
 292 against the permuted response by a linear SVR, we obtain  $R$  maps  $(\hat{\mathbf{w}}^{\text{SVR},(r)})_{r \in [R]}$ .  
 293 We can now apply the Westfall-Young step-down maxT adjusted p-values algorithm  
 294 [Westfall and Young, 1993, p. 116-117] taking the raw SVR weights instead of the usual  
 295  $t$ -statistics to derive the corrected p-values. A sufficient assumption to ensure the validity  
 296 of the p-values is the pivotality of the SVR weights. Keeping the corrected p-values that  
 297 are less than a given significance level —equal to 10% in this study— this procedure is  
 298 equivalent to thresholding the SVR weight map. We call this procedure Permutation Test  
 299 SVR (Perm-SVR). The only difference between Perm-SVR and the Thr-SVR procedure  
 300 is the way of computing the threshold. To perform the permutation test procedure, we  
 301 took  $R = 1000$  permutations.

### 302 3.4 Adaptive Permutation Threshold SVR (Ada-SVR)

303 Here, we introduce Adaptive Permutation Threshold SVR (Ada-SVR), a statistical in-  
 304 ference procedure that produces a weight map and confidence intervals around it; it is  
 305 also almost equivalent to thresholding the SVR weights non-uniformly. Ada-SVR was  
 306 first presented by Gaonkar and Davatzikos [2012]. First, the authors derived an esti-  
 307 mated weight  $\hat{\mathbf{w}}^{\text{APT}}$  linearly related to the target by approximating the hard margin  
 308 SVM formulation, their estimator is given by the following equation:

$$\hat{\mathbf{w}}^{\text{APT}} = \mathbf{L} \mathbf{y} , \quad (9)$$

309 where  $\mathbf{y}$  is the target variable and  $\mathbf{L} \in \mathbb{R}^{p \times n}$  only depends on the design matrix  $\mathbf{X}$ :

$$\mathbf{L} = \mathbf{X}^\top \left[ (\mathbf{X}\mathbf{X}^\top)^{-1} - (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{1} (\mathbf{1}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{1})^{-1} \mathbf{1}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \right], \quad (10)$$

310 where  $\mathbf{1} \in \mathbb{R}^n$  is a vector of ones. The approximation made by (9) is notably valid under  
311 the assumption that all the data samples are support vectors, which might hold at least  
312 if  $n \ll p$ . Then, if  $\mathbf{y}$  is standardized and if  $n$  is large enough (so that the central limit  
313 theorem holds), one expects that under the null hypothesis for the  $j$ -th covariate:

$$\hat{\mathbf{w}}_j^{\text{APT}} \sim \mathcal{N}(0, \sum_{k=1}^n \mathbf{L}_{j,k}^2). \quad (11)$$

314 From (11), p-values can be computed and corrected by applying a Bonferroni correction  
315 (multiplying the raw p-values by a factor  $p$ ).

### 316 3.5 Ensemble of Clustered Desparsified Lasso Algorithm (EnCluDL)

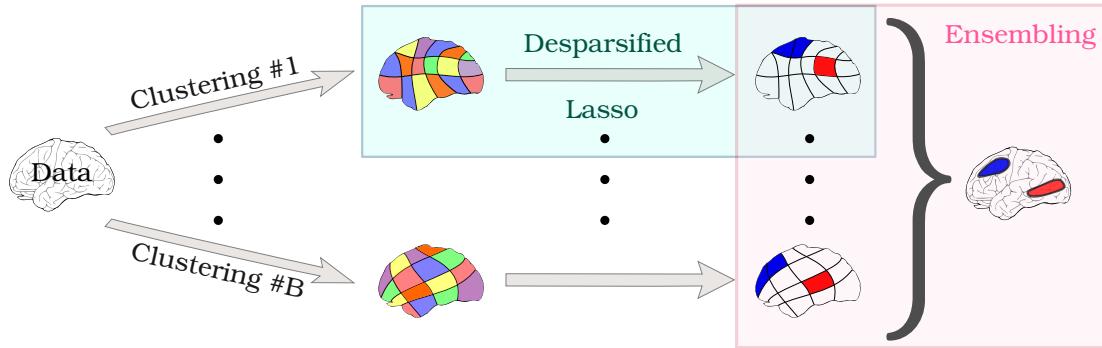


Figure 2: **Ensemble of Clustered Desparsified Lasso (EnCluDL) algorithm.** The EnCluDL algorithm combines three algorithmic steps: a clustering (or parcellation) procedure applied to images, the Desparsified Lasso procedure (statistical inference) to derive statistical maps, and an ensembling method that synthesizes several statistical maps. In the first step,  $B$  clusterings of voxels are generated using  $B$  random subsamples of the original sample. Then, for each grouping-based data reduction, a statistical inference procedure is run resulting in  $B$  z-score maps (or p-value maps). Finally, these maps are ensembled into a final z-score map using an aggregation method that preserves statistical properties.

317 Ensemble of Clustered Desparsified Lasso (EnCluDL) is a multivariate statistical  
318 inference procedure designed for spatial data; it was first introduced by Chevalier et al.  
319 [2018]. EnCluDL relies on three steps: a spatially-constrained clustering algorithm for  
320 reducing the problem dimension, a statistical inference procedure for deriving statistical  
321 maps, and an ensembling method for aggregating the statistical maps.

322 **Statistical inference with Desparsified Lasso.** Desparsified Lasso (DL) is a statistical  
323 inference procedure that can be viewed as a generalization of the least-squares-based  
324 inference in high dimension under sparsity assumptions. It was proposed and thoroughly  
325 analyzed by [Zhang and Zhang \[2014\]](#) and [van de Geer et al. \[2014\]](#). This estimator pro-  
326 duces p-values on linear model parameters even when the number of parameters  $p$  is  
327 (reasonably) greater than the number of samples  $n$ . A technical description of Despar-  
328 sified Lasso is available in [Sec. 7.1](#). In the neuroimaging context, the initial parameters are  
329 related to the voxels, which are of the order of one hundred thousand while the number  
330 of samples is almost always lower than one thousand. In such settings Desparsified Lasso  
331 is inefficient due to a lack of statistical power, hence dimension reduction is required.

332 **Clustering.** As argued in [Section 1](#), while performing dimension reduction, we aim at  
333 keeping the spatial structure of the data and avoid mixing voxels "far" from each other.  
334 This is achieved with data-driven parcellation along with a spatially constrained clus-  
335 tering algorithm following the conclusions by [Varoquaux et al. \[2012\]](#) and [Thirion et al.](#)  
336 [\[2014\]](#). Another interesting aspect of this dimension reduction method is its denoising  
337 property [[Hoyos-Idrobo et al., 2018](#)] since it produces averages from groups of noisy vox-  
338 els. Note that this choice ultimately calls for a spatial tolerance on the statistical control,  
339 *i.e.*, considering the  $\delta$ -FWER instead of the standard FWER. Through the clustering,  
340 the  $p$  voxels are grouped into  $C$  clusters, where  $C \ll p$ . Then, Desparsified Lasso is  
341 directly applied to the compressed problem in order to produce corrected p-values. Not-  
342 ably, corrected p-values are obtained from the initial p-values by applying Bonferroni  
343 correction [[Dunn, 1961](#)] with a factor  $C \ll p$ . Following the terminology in [[Chevalier](#)  
344 [et al., 2018](#)], we refer to this procedure as Clustered Desparsified Lasso (CluDL). CluDL  
345 however suffers from high variance [[Chevalier et al., 2018](#)] as it depends on an arbitrary  
346 grouping choice. This can be alleviated by ensembling techniques, as described next.

347 **Ensembling.** [Varoquaux et al. \[2012\]](#), [Hoyos-Idrobo et al. \[2018\]](#) have shown that  
348 randomizing the grouping choice and adding an ensembling step to aggregate several  
349 solutions can stabilize the overall procedure. Additionally, [Chevalier et al. \[2018\]](#) have  
350 highlighted that the ensembling step is also beneficial in terms of support recovery. To  
351 perform  $B$  groupings of the covariates, we train the parcellations algorithm with  $B$   
352 different random subsamples of the original data sample. Then, thanks to the CluDL  
353 procedure, we obtain  $B$  statistical maps that are aggregated into one through an en-  
354 sembling procedure. The ensembling procedure we considered in the statistical inference  
355 procedure is adapted from [Meinshausen et al. \[2009\]](#) that is described in appendix in  
356 [Sec. 7.2](#). We refer to the full inference algorithm as Ensemble of Clustered Desparsified  
357 Lasso (EnCluDL). Under hypothesis ensuring Desparsified Lasso statistical properties  
358 —notably sparsity and smoothness of the true weight map and i.i.d. data samples—  
359 EnCluDL gives statistical guarantees, namely it controls the  $\delta$ -FWER.

360 **Choosing  $\delta$  for  $\delta$ -FWER control** Theoretically, the minimal spatial tolerance  $\delta$   
361 that guarantees a control of the  $\delta$ -FWER with EnCluDL is the largest parcel diameter.

362 However, in practice, we aggregate many statistical maps obtained from different choices  
 363 of voxel grouping; then the required spatial tolerance is reduced to the average radius.  
 364 Then, the value of  $\delta$  for which we observe the  $\delta$ -FWER control varies approximately  
 365 linearly with the cubic root of the average number of voxels per cluster. In standard  
 366 fMRI settings, we propose the following formula for  $\delta$ :

$$\delta_0 = \left( \frac{p}{2C} \right)^{1/3}, \quad (12)$$

367 the ratio  $p/C$  being the average number of voxels per cluster,  $\delta_0$  is a distance in voxel  
 368 size unit.

369 Note that the previous formula is an estimate of the average cluster radius that  
 370 assumes that the shape of the clusters have identical cubic shape. In practice, this  
 371 formula tends to underestimate the average cluster radius but was suitable in all our  
 372 experiments. In Sec. 7.6, we study empirically the distribution of the cluster radius  
 373 distribution as a function of the number of clusters, and compare it with  $\delta_0$ .

374 Additionally, note that when the setting is particularly favorable for inference, *e.g.*, if  
 375  $\log(n)/C$  is large, the choice of  $\delta$  given by (12) might be slightly too liberal. To address  
 376 these specific cases, we propose a more refined formula to estimate  $\delta$  in appendix in  
 377 Sec. 7.5.

378 **EnCluDL hyper parameters.** The number of clusters  $C$  is a crucial hyperparameter  
 379 of EnCluDL. Generally, a suitable  $C$  depends on intrinsic physical properties of the  
 380 problem and on the targeted spatial tolerance  $\delta$ . Decreasing  $C$  increases the statistical  
 381 power while reducing the spatial precision. In the neuroimaging context, taking  $C =$   
 382 500 is a fair default value achieving a suitable trade-off between spatial precision and  
 383 statistical power when the number of samples is a few hundreds. With this choice, the  
 384 spatial tolerance should be close to  $\delta = 10$  mm when working with masked fMRI data.

385 As a more adaptive approach, we recommend tuning  $C$  according to  $n$  *e.g.*,  $C \in$   
 386  $[n/2, n]$ . This choice should still ensure the  $\delta$ -FWER control with  $\delta$  given by (12) (or its  
 387 corrected version, see appendix Sec. 7.5) and is justified in Sec. 4.5.

388 The parameter  $B$ , the number of CluDL solutions to be aggregated, is discussed in  
 389 Sec. 3.5. The larger  $B$  the more stable the solution, yet the heavier the computational  
 390 cost. In our experiments, we have set  $B = 25$  (see Hoyos-Idrobo et al. [2018] for a more  
 391 complete discussion on this parameter).

392 **Empirical analysis of data structure assumptions for EnCluDL.** The core part  
 393 of EnCluDL consists in applying Desparsified Lasso to a clustered version of the original  
 394 problem. As disclaimed in van de Geer et al. [2014], some technical hypotheses on the  
 395 structure of the design matrix  $\mathbf{X}$  —*i.e.*, of the reduced data— are necessary to produce  
 396 valid confidence intervals on the parameters with Desparsified Lasso. Roughly, it is  
 397 necessary that the features are "not too much correlated". In appendix in Sec. 7.3, we  
 398 show in a simple setting that as long as the correlation between two predictive features

399 is less than 0.8, it is possible to recover both features. However when the correlation  
400 between features is more than 0.9, only one of the two features can be identified.

401 In Sec. 7.4, we show that in standard fMRI datasets neighboring voxels can have  
402 a correlation greater than 0.9. Thus applying Desparsified Lasso at the voxel level  
403 certainly leads to many false negatives. However, since Desparsified Lasso is applied  
404 to the clustered problem, we have to consider correlation between clusters instead. In  
405 Sec. 7.4, we show on HCP data that such inter-cluster correlation is almost always  
406 lower than 0.8 and always lower than 0.85. This means that data structure assumptions  
407 for EnCluDL are sustainable. Additionally, the fact that EnCluDL aggregates several  
408 CluDL solutions increases the tolerance to inter-cluster correlation.

### 409 3.6 A complementary univariate solution

410 Given the complementarity of univariate and multivariate inference noted previously, we  
411 add to our study a univariate inference method, namely *univariate permuted OLS* (Univ-  
412 OLS). This method does not test the same null hypothesis as the other methods: it tests  
413 whether or not a voxel is marginally associated with the target. Then, while it should  
414 not be benchmarked with the other methods, we propose to consider jointly the results  
415 obtained by the marginal and the conditional analyses, as advocated by Weichwald et al.  
416 [2015].

417 The Univ-OLS method is based on the generalized linear model (GLM) [Friston et al.,  
418 1994]. For every voxel we compute a t-statistic by applying the OLS procedure on the  
419 linear model that associates each voxel with the target. Subsequently, we also derive the  
420 permuted t-statistic distribution by performing the OLS on permuted data. Finally, to  
421 obtain corrected p-values, we use the standard maxT procedure [Westfall and Young,  
422 1993]. Note that, for this method, we have used the `permuted_ols` function implemented  
423 in the Nilearn python package [Abraham et al., 2014] with 1000 permutations.

### 424 3.7 Implementation

425 The Python code that implements Thr-SVR, Perm-SVR, Ada-SVR and EnCluDL can be  
426 found on <https://github.com/ja-che/hidimstat>. Our algorithms are implemented  
427 with Python = 3.6.8 and need the following packages Numpy = 1.16.2 [Van der Walt  
428 et al., 2011], Scipy = 1.2.1 [Virtanen et al., 2020], Scikit-Learn = 0.21 [Pedregosa et al.,  
429 2011], Joblib = 0.11 and Nilearn = 0.6.0 [Abraham et al., 2014].

## 430 4 Experimental procedures

### 431 4.1 Data

432 To validate empirically the statistical guarantees of the four algorithms —Thr-SVR,  
433 Perm-SVR, Ada-SVR and EnCluDL— described in Section 3, we perform several ex-  
434 periments on resting-state fMRI and task fMRI data. We also show some results for  
435 Univ-OLS to highlight the complementarity of univariate and multivariate analyses, in

436 particular when studying predictive patterns on real data. We focus on three datasets:  
437 HCP900 resting-state fMRI, HCP900 task fMRI and RSVP task fMRI.

438 **HCP900 resting-state fMRI data.** HCP900 resting-state fMRI dataset [Van Essen  
439 et al., 2012] contains 4 runs of 15 minutes resting-state recordings with a 0.76s-repetition  
440 time (corresponding to 1200 frames per run) for 796 subjects. We use the MNI-resampled  
441 images provided in the HCP900 release. For this dataset the number of samples is equal  
442 to 1200 (only one run is used) and the number of voxels is 156 374 after gray-matter  
443 masking (the spatial resolution being 2 mm isotropic).

444 **HCP900 task fMRI data.** We also use the HCP900 task-evoked fMRI dataset [Van  
445 Essen et al., 2012], in which we take the masked 2 mm z-maps of the 796 subjects  
446 from 6 tasks to solve 7 binary classification problems: emotion (*emotional face* vs *shape*  
447 *outline*), gambling (*reward* vs *loss*), language (*story* vs *math*), motor hand (*left* vs *right*  
448 hand), motor foot (*left* vs *right* foot), relational (*relational* vs *match*) and social (*mental*  
449 *interaction* vs *random interaction*). We consider the fixed-effect maps for each outcome  
450 (or condition), yielding one image per subject per condition (which corresponds to two  
451 images per subject for each classification problem). Then, for each problem, the number  
452 of samples available is 1592 ( $= 2 \times 796$ ) and the number of voxels is 156 374 after  
453 gray-matter masking.

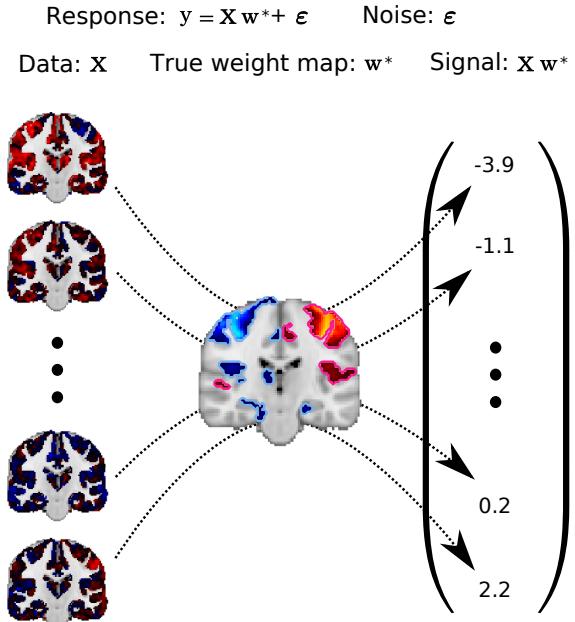
454 **Unmasked RSVP task fMRI data.** We also use activation maps obtained from  
455 a rapid serial visual presentation (RSVP) task of the individual brain charting dataset  
456 [Pinho et al., 2018], augmented with 9 additional subjects performing the same task,  
457 under the same experimental procedures and scanning parameters. No masking is used  
458 for this dataset, so that out-of-brain voxels are not withdrawn from preprocessing. We  
459 consider the unmasked 3 mm-resolution statistical z-maps of the 6 sessions of the 21 sub-  
460 jects for a reading task with 6 different contrasts that have been grouped into 2 classes:  
461 language (words, simple sentences, complex sentences) vs pseudo-language (consonant  
462 strings, pseudo-word lists, jabberwocky). The images are all registered to MNI space and  
463 per-condition effects are estimated with Nistats v0.0.1 library [Abraham et al., 2014].  
464 For this dataset the number of samples available is equal to 756 (21 subjects  $\times$  6 runs  $\times$  6  
465 images per run) and the number of voxels is 173 628 (unmasked images resampled at  
466 3-mm resolution). We run the inter-subject experiment described in Sec. 4.4 with this  
467 dataset.

## 468 4.2 Statistical control on semi-simulated data

469 A first series of experiments study whether the four different methods exhibit the ex-  
470 pected  $\delta$ -FWER control and are competitive in terms of support recovery, as measured  
471 with the precision-recall curve. To do so, we have to construct the true weight map  $\mathbf{w}^*$ .  
472 We generate “semi-simulated” data: generating signals from estimates on real data. To

473 avoid circularity in the definition of the ground truth, we used two different tasks: one  
 474 to build  $\mathbf{w}^*$  and another one to define  $\mathbf{X}$ .

475 **Building a reference weight map from HCP900 motor hand dataset.** To construct an underlying weight map, we use the motor hand (MH) task of the HCP900 task  
 476 fMRI dataset described in Sec. 4.1. Specifically, we build a design matrix  $\mathbf{X}_{\text{MH}} \in \mathbb{R}^{n \times p}$   
 477 from the motor hand task z-maps of all subjects associated with a binary target index  
 478  $\mathbf{y}_{\text{MH}}$ . To obtain an initial weight map  $\mathbf{w}_{\text{MH}}^{\text{SVC}}$  we regress  $\mathbf{X}_{\text{MH}}$  against  $\mathbf{y}_{\text{MH}}$  by fitting  
 479 a linear Support Vector Classifier (SVC) [Cortes and Vapnik, 1995]. From  $\mathbf{w}_{\text{MH}}^{\text{SVC}}$  we  
 480 only kept the 10% most extreme values ensuring that the connected groups of non zero-  
 481 weight voxels have a minimal size of  $1 \text{ cm}^3$  by removing small clusters. We chose this  
 482 map (represented in Fig. 3 and Fig. 4) to be the true weight map  $\mathbf{w}^* \in \mathbb{R}^p$  for the whole  
 483 simulated experiments.



**Figure 3: Generating a hybrid dataset with known ground truth and actual fMRI data.** To generate the response for a given sample we multiply the corresponding brain activation map by the true weight map and add a Gaussian noise with fixed variance. To highlight the predictive regions, we circle them in pink for positive coefficients and in light blue for negative coefficients. As an illustration, we take four different data samples with negative or positive output value.

485 **Simulating responses with HCP900 emotion dataset.** We then take  $\mathbf{X}$  to be  
 486 the set of z-maps from the emotion task of the HCP900 task fMRI dataset described in  
 487 Sec. 4.1. To generate a continuous response vector  $\mathbf{y}$ , we draw a Gaussian random noise  
 488 vector  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\epsilon}^2 \mathbf{I}_n)$  and use the linear model introduced in (1), where  $\sigma_{\epsilon} = 0.2$  to  
 489 reach  $\text{SNR}_y = 10$ , where  $\text{SNR}_y$  is given by:

$$\text{SNR}_y = \frac{\|\mathbf{X} \mathbf{w}^*\|^2}{n \sigma_{\epsilon}^2} . \quad (13)$$

490 The way we simulate  $\mathbf{y}$  is summarized in Fig. 3.

491 **Quantification of error control and detection accuracy.** To obtain representa-  
492 tive results, we then run the procedures described in [Section 3](#) for 100 different response  
493 vectors  $\mathbf{y}$  generated from different random samples of subjects and different draws of  $\varepsilon$ .  
494 We let the number of samples vary from  $n = 50$  (25 random subjects taken among the  
495 796) to  $n = 1200$  (600 subjects), the number of voxels being  $p = 156\,374$ . For each sim-  
496 ulation, we record the empirical  $\delta$ -FWER and the precision-recall curves. Importantly,  
497 we do not recommend running such analysis with  $n < 100$ , since the estimation problem  
498 is hard and statistical guarantees are only asymptotic.

499 **Heavy-tailed version of the semi-simulated experiment.** In the above experi-  
500 ment the noise is Gaussian, hence we also benchmark the inference procedures for Laplace  
501 and Student noise to assess the impact of noise distribution.

502 **Binary version of the semi-simulated experiment.** In the main experiment the  
503 response vector  $\mathbf{y}$  is continuous, hence we also benchmark the inference procedures for a  
504 binary response. For that, we simply take as response vector the signs of the continuous  
505  $\mathbf{y}$  generated as in the previous paragraph.

506 **Univ-OLS solves another inference problem.** Univariate methods do not compete  
507 with multivariate methods, as they do not test the same null hypotheses. However, for  
508 pedagogical purpose, we show that Univ-OLS based FWER control is not valid in the  
509 multivariate analysis setup.

### 510 **4.3 Statistical control under the global null with i.i.d. data**

511 In this experiment, we test whether the procedures control the FWER under a global  
512 null model. EnCluDL only controls the  $\delta$ -FWER theoretically but, when the true weight  
513 vector  $\mathbf{w}^*$  is null, the  $\delta$ -FWER and the classical FWER are identical. Then, all pro-  
514 cedures should control the FWER. Here, we considered the tasks of the HCP900 task  
515 fMRI dataset described in [Sec. 4.1](#) keeping all the subjects ( $n = 1592$ ). Then, to get  
516 a noise-only response, we (uniformly) randomly permute the original response vector.  
517 Similarly as in [Sec. 4.2](#), the i.i.d. hypothesis is legitimate, since the data correspond to  
518 z-maps of different subjects. For each task, we draw 100 different permutations of the  
519 response and check if the different methods enforce the chosen nominal FWER of 10%.

520 to illustrate the importance of checking the underlying assumptions, in appendix in  
521 [Sec. 7.8](#), we describe an additional experiment to show that FWER (or  $\delta$ -FWER) is not  
522 controlled anymore when working with an autocorrelated response vector, breaking the  
523 i.i.d hypothesis. This experiment is adapted from [Eklund et al. \[2016\]](#).

### 524 **4.4 Statistical control of out-of-brain detections**

525 In this experiment we test the four procedures on an unmasked task fMRI dataset to  
526 verify that no spurious detection is made outside of the brain —up to the allowed error  
527 rate. Indeed, the non-null coefficients of the weight vector  $\mathbf{w}^*$  should all be contained

528 in the brain since there is no informative signal in out-of-brain voxels. To do so, we  
529 take the unmasked RSVP task fMRI dataset, described in Sec. 4.1 (with design matrix  
530  $\mathbf{X}$  containing  $n = 756$  unmasked z-maps). Then, we report how frequently some voxels  
531 are detected outside the brain volume. For the sake of completeness, we also check the  
532 non-occurrence of out-of-brain detections with Univ-OLS.

### 533 4.5 Insights on the choice of number of clusters

534 In this experiment, we assess empirically the impact of  $C$ , the number of clusters used in  
535 the EnCluDL algorithm. We use the same generative method as in Sec. 4.2 to produce  
536 an experiment with known ground truth. Then, we run the EnCluDL algorithm varying  
537 the numbers of clusters  $C$  from  $C = 200$  to  $C = 1000$ . We also vary the number of  
538 samples  $n$  from 100 to 1200. As in Sec. 4.2, we run the experiment for 100 different  
539 response vectors and report aggregated results. We report two statistics: the empirical  
540  $\delta$ -FWER and the AUC of the precision-recall curve for every value of  $C$  and  $n$ .

### 541 4.6 Face validity on HCP dataset

542 In this experiment, we consider the output of the procedures in terms of brain regions  
543 that are conditionally associated with the task performed by the subjects. Similarly as  
544 in Sec. 4.3, we consider the tasks of the HCP900 task fMRI dataset described in Sec. 4.1,  
545 keeping this time the true response vector. We run all the procedures on every task and  
546 report the statistical maps thresholded such that the FWER < 10% or the  $\delta$ -FWER <  
547 10% (for EnCluDL). For this, we use all the available samples ( $n = 1592$ ). We also  
548 include Univ-OLS to compare the discriminative patterns obtained with a univariate  
549 inference.

### 550 4.7 Prediction performance

551 Even if it is not the purpose of this study, we also checked the prediction performance of  
552 the decoders produced by each method. Since Thr-SVR and Perm-SVR rely on the same  
553 predictive function, there are three different decoders: SVR, Ada-SVR and EnCluDL.  
554 To perform this experiment, we consider the tasks of the HCP900 task fMRI dataset  
555 described in Sec. 4.1. We run all the procedures on every task using a sample size  
556  $n = 400$ , keeping the rest of the samples to test the trained model. For each task and  
557 each method, we take 100 different random subsamples to produce the results. This  
558 experiment being a side study, we give the results in appendix in Sec. 7.12.

## 559 5 Results

560 In this section, after setting the value of the tolerance parameter  $\delta$  in the different  
561 datasets, we present the experimental results.

562 **5.1 Estimating  $\delta$  in HCP and RSVP datasets**

563 In all the experiments, unless specified otherwise, we run EnCluDL with the default  
 564 choice  $C = 500$ . Reversing (12), we obtain a tolerance parameter of  $\delta_{\text{HCP}} = 5.4$  voxels  
 565 for HCP900 and  $\delta_{\text{RSVP}} = 5.6$  voxels for RSVP, corresponding to  $\delta_{\text{HCP}} = 12 \text{ mm}$  and  
 566  $\delta_{\text{RSVP}} = 18 \text{ mm}$  respectively after rounding up. In Fig. 14 in appendix, we display the  
 567 spatial tolerance of 6 voxels in the case of HCP data.

568 **5.2 Statistical control with known ground truth**

569 Here, we describe the results obtained from the experiment described in Sec. 4.2.

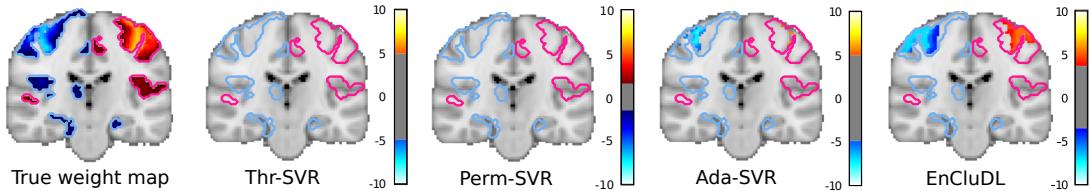
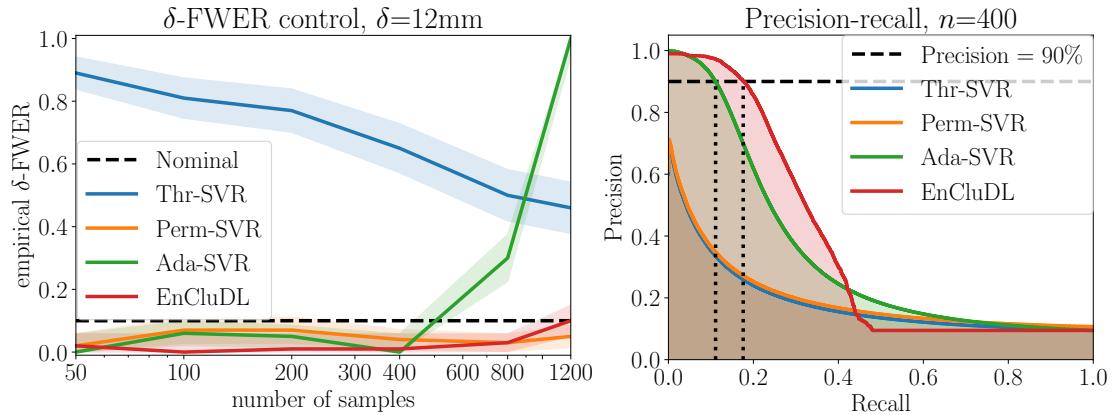


Figure 4: **Qualitative comparison of the model solutions.** Here, we show the solutions (z-maps) given by the four inference procedures, for a single random draw of the noise vector in the experiment described in Sec. 4.2. The weight maps are thresholded such that  $\delta\text{-FWER} < 10\%$  theoretically. We can observe that none of the methods yield false discoveries but the Ensemble of Clustered Desparsified Lasso (EnCluDL) procedure is the most powerful followed by Adaptive Permutation Threshold SVR (Ada-SVR).

570 **Qualitative comparison of the model solutions.** In Fig. 4, we present a qualitative  
 571 comparison of the model solutions when  $n = 400$ . None of the methods yields false  
 572 discoveries for the chosen threshold —taken such that  $\delta\text{-FWER} < 10\%$ . EnCluDL  
 573 recovers more active regions than the other procedures, which makes it the most powerful  
 574 procedure, followed by Ada-SVR. The other two procedures do not discover the expected  
 575 patterns. These results displayed are obtained for a single random draw of the noise  
 576 vector, but similar results holds for different draws.

577  **$\delta$ -FWER control.** In this experiment, we check if Thr-SVR, Perm-SVR, Ada-SVR  
 578 and EnCluDL control the  $\delta$ -FWER at the targeted nominal level (here being 10%). Fig. 5  
 579 shows that Perm-SVR and EnCluDL procedures control the  $\delta$ -FWER for all sample sizes  
 580 since their empirical  $\delta$ -FWER remain below the targeted nominal level, whereas Thr-  
 581 SVR and Ada-SVR fail to control the  $\delta$ -FWER in every setting. In particular, the  
 582 empirical  $\delta$ -FWER for Ada-SVR is above the targeted nominal level for  $n \geq 800$ . This  
 583 might occur since the approximation made by (9) is valid only if  $n$  remains “sufficiently  
 584 low” [Gaonkar and Davatzikos, 2012]. Thr-SVR fails to control empirically the  $\delta$ -FWER  
 585 for any value of  $n$ . This might be due to the two assumptions made in Sec. 3.2 not being  
 586 satisfied—it is indeed unlikely that the SVR weights of the null region follow the same

587 distribution. We further discuss this point in [Section 6](#). Concerning EnCluDL, one can  
 588 notice that the empirical  $\delta$ -FWER is slightly larger for  $n = 1200$ , this effect is explained  
 589 in appendix in [Sec. 7.5](#) and [Sec. 7.6](#). We report additional results, notably heavy-tailed  
 590 and binary version of the experiment, in appendix in [Sec. 7.10](#). These lead to the same  
 591 statistical behavior as observed here.



**Figure 5:  $\delta$ -FWER control and precision-recall curve on semi-simulated data (known ground truth).** Left: The results of the experiment described in [Sec. 4.2](#) show that the permutation test (Perm-SVR) and Ensemble of Clustered Desparsified Lasso (EnCluDL) are the only procedures that correctly control the  $\delta$ -FWER at the nominal level (10%). This is not the case for Adaptive Permutation Threshold SVR (Ada-SVR) and Thresholded SVR (Thr-SVR) procedures. Right: For the same experiment, EnCluDL has the best performance in terms of precision-recall curve. For  $n = 400$ , and ensuring 90% precision, EnCluDL obtains a recall of 23% and Ada-SVR a recall of 16%. Thr-SVR and Perm-SVR share the same precision-recall curve and were not able to reach 90% precision.

592 **Precision-recall.** In this experiment, we also evaluate the recovery properties of the  
 593 four methods by comparing the precision-recall curve for different value of  $n$ . [Fig. 5](#)  
 594 shows that EnCluDL has the best precision-recall curve for  $n = 400$ . We recall that  
 595 the perfect precision-recall curve is reached if the precision is equal to 1 for any value  
 596 of recall between 0 and 1. Similar results were obtained for the other sample sizes  
 597 tested (appendix [Fig. 17](#)). Indeed, when  $n = 400$ , for a 90% precision, EnCluDL gives  
 598 a recall of 23% and Ada-SVR a recall of 16%. Thr-SVR and Perm-SVR share the same  
 599 precision-recall curve since they both produce p-values arranged in the reverse order of  
 600 the absolute SVR weights. These thresholding methods were not able to reach the 90%  
 601 precision; their recovery properties are much weaker.

602 We report additional results in [Sec. 7.10](#).

### 603 5.3 Statistical control under the global null with i.i.d. data

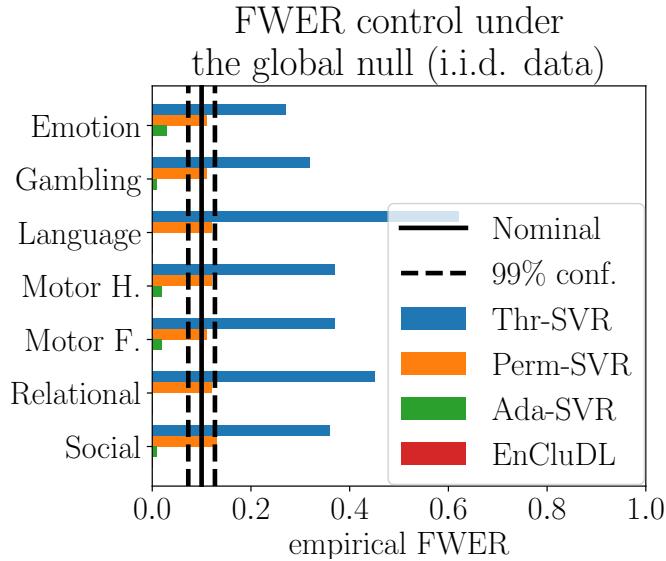
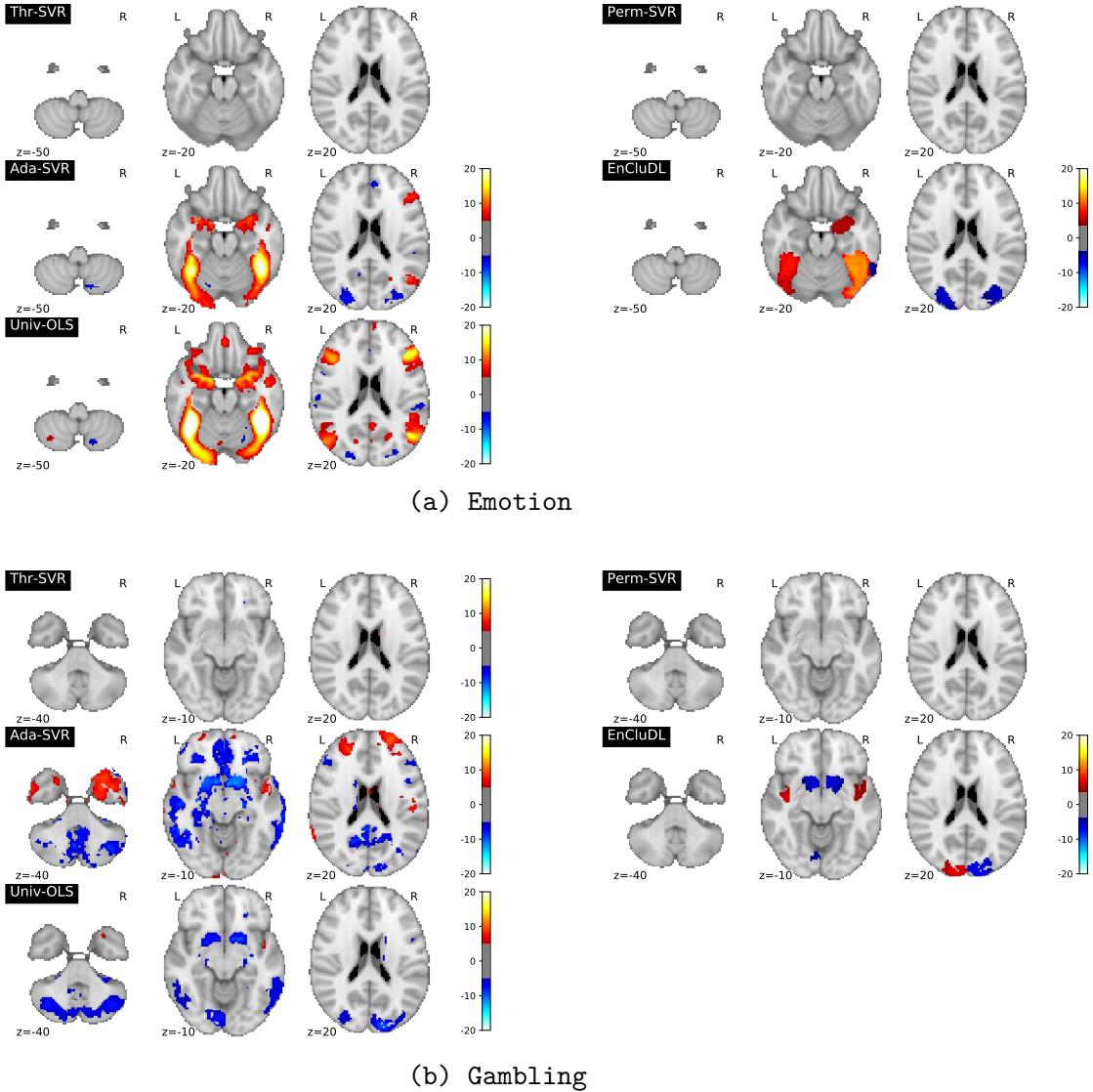


Figure 6: **FWER control under the global null with i.i.d. data** The results of the experiment with i.i.d. data under the global null, described in Sec. 4.3, show that, only the Thresholded SVR (Thr-SVR) fails to control the FWER empirically in this context. EnCluDL makes no detection: it is a conservative approach, as one could expect from theory.

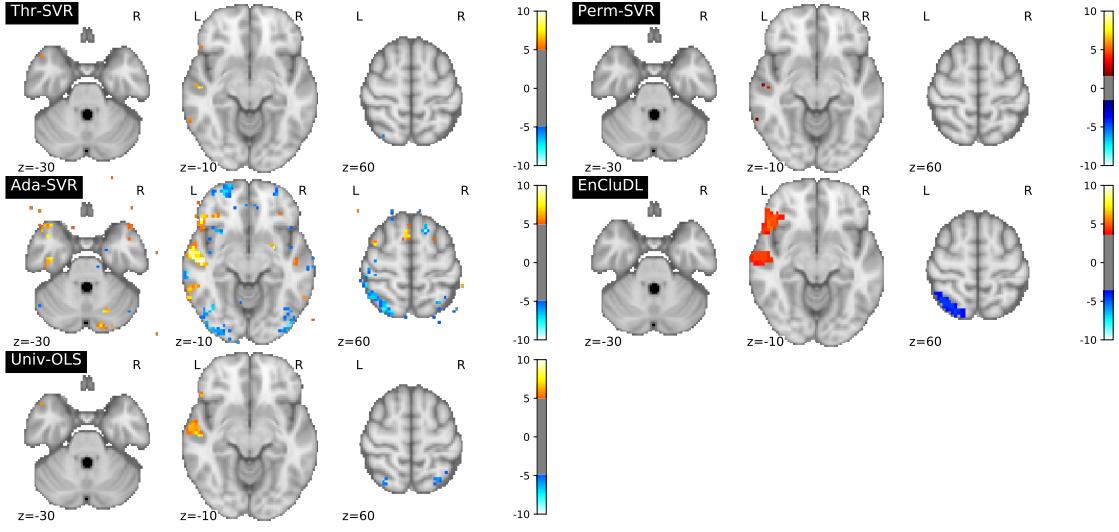
604 **FWER control under the global null (permuted response).** Here, we summarize  
 605 the results of the experiment testing control of the FWER in a global null setting  
 606 (Sec. 4.3). Fig. 6 shows that, when samples are i.i.d., all the procedures control the  
 607 FWER, except Thr-SVR. EnCluDL is even conservative since the empirical FWER re-  
 608 mains at 0 for all the different tasks tested. This result is not surprising since at least  
 609 two steps of the EnCluDL procedure are conservative: the Bonferroni correction and the  
 610 ensembling of the p-values maps.

611 **Face validity (original response).** Additionally, we run the procedures with the  
 612 original (not permuted) response vector to check whether the methods can recover pre-  
 613 dictive patterns; this corresponds to the experiment described Sec. 4.6. We plot the  
 614 results for the two first tasks (emotion and gambling) in Fig. 7; see appendix Fig. 23  
 615 for the five other tasks. Qualitatively, EnCluDL recovers the most plausible predictive  
 616 patterns, Ada-SVR sometimes makes dubious discoveries: patterns are too wide and  
 617 implausible. The two other methods exhibit a very weak statistical power.

618 Comparing EnCluDL and Univ-OLS solutions, we see that the discovered patterns  
 619 are not a subset of each other. This result was expected given the arguments in Weich-  
 620 wald et al. [2015]: the advantage of combining the two paradigms is to get more insight  
 621 on the causal nature of the relation between the voxel signals and the target.



**Figure 7: Estimated predictive patterns on standard task fMRI dataset.** Here, we plot the results for the emotion and gambling tasks of the experiment described in Sec. 4.6 thresholding the statistical maps such that the  $\delta$ -FWER stays lower than 10% for  $\delta = 12$  mm. Qualitatively, EnCluDL discovers the most plausible patterns, Ada-SVR sometimes makes dubious discoveries, patterns are too wide and implausible, while the two other methods exhibit a very weak statistical power. Univariate analysis results obtain with Univ-OLS clearly provide distinct information about the relationship between the voxel signals and the outcome. The results of the five other tasks are available in Fig. 23.



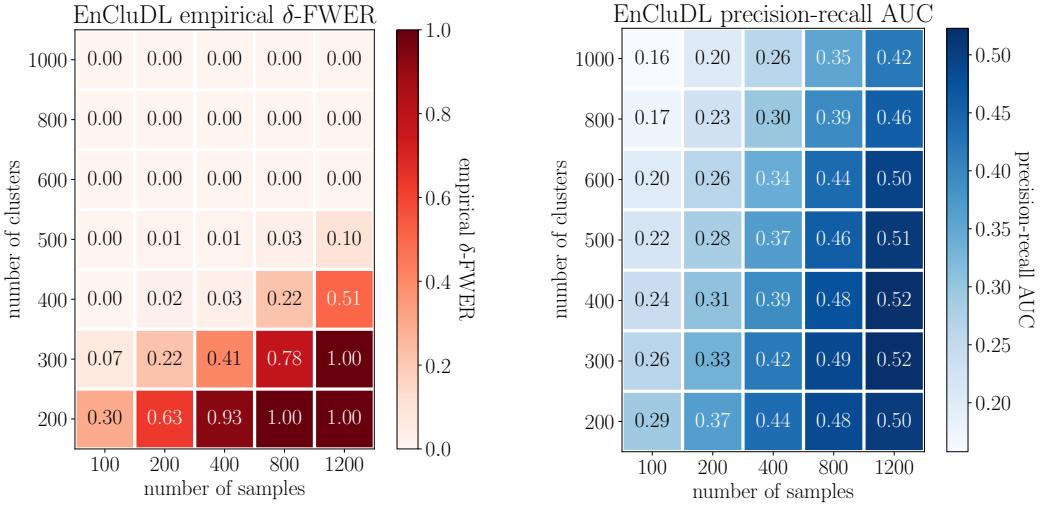
**Figure 8: Statistical maps for unmasked RVSP data.** The results of the unmasked task-fMRI experiment, described in Sec. 4.4, show that EnCluDL, Thresholded SVR (Thr-SVR) and the permutation test (Perm-SVR) do not return out-of-brain discoveries, while the Adaptive Permutation Threshold SVR (Ada-SVR) does. Here z-score maps are thresholded such that the  $\delta$ -FWER is at most 10% for  $\delta = 6$  voxels (or 18 mm). Thr-SVR and the Perm-SVR do not yield spurious detections but very few detections are made, hence these method have low statistical power. EnCluDL does not make any spurious detection; rather it makes detections in the temporal lobe and Broca’s area, which are expected for a reading task. Univ-OLS does not make any out-of-brain detection either but returns significant associations in the temporal lobe.

#### 5.4 Statistical control of out-of-brain discoveries

We now report the results from the unmasked RSVP task data experiment (Sec. 4.4). Here, we check whether out-of-brain detections are made. In Fig. 8, the z-score maps are thresholded such that the FWER (for Perm-SVR, Thr-SVR, and Ada-SVR) or the  $\delta$ -FWER (for EnCluDL) are at most 10% for  $\delta = 6$  voxels (or 18 mm). We observe that Ada-SVR makes some out-of-brain discoveries, and it does not control the FWER empirically. Thr-SVR and Perm-SVR do not yield spurious detections but very few detections are made, hence these methods have low statistical power. EnCluDL does not make any out-of-brain detections and it outlines predictive regions in the temporal lobe and Broca’s area, expected for a reading task. Finally, Univ-OLS does not make any spurious detection either; it only makes detections in the temporal lobe.

#### 5.5 Insights on choosing the number of clusters

Here, we report the results obtained of the experiment task-fMRI data (Sec. 4.5) studying the impact of  $C$  (number of clusters) on the  $\delta$ -FWER control and the recovery



**Figure 9: Influence of the number  $C$  of clusters on  $\delta$ -FWER control and the recovery properties of EnCluDL.** The results of the experiment described in Sec. 4.5 show the impact of  $C$  on the  $\delta$ -FWER control and the recovery score of EnCluDL. When  $C \geq 500$ , clusters are smaller, hence the  $\delta$ -FWER is controlled for  $\delta = 12$  mm (and potentially lower values of  $\delta$ ) since all the empirical  $\delta$ -FWER's are lower than the 10% nominal rate. Conversely, when  $C < 500$ , clusters are wider and the spatial tolerance is overcome by the model inaccuracy, hence the  $\delta$ -FWER is not controlled for  $\delta = 12$  mm. However, it remains controlled for higher values of  $\delta$ . Concerning the recovery properties we see that reducing the number of clusters improves the precision-recall curves. Thus, the more spatial uncertainty is tolerated, the best recovery properties EnCluDL offers.

properties of EnCluDL for various sample sizes. These results are obtained with 100 repetitions for every sample and cluster sizes. In Fig. 9, we notice that a lower  $C$  leads to improved recovery, according to the area under the precision-recall curves, for  $\delta = 6$  voxels (or 12 mm). However, when the number of cluster is lower, the average cluster radius increases and overcomes the spatial tolerance of  $\delta$ , leading to inflated error rates (cf. Sec. 7.6). More precisely, the  $\delta$ -FWER is controlled when  $C \geq 500$ . Note that for  $C < 500$ , it is possible to control the  $\delta$ -FWER, even when  $n$  is small, provided a larger spatial tolerance  $\delta > 6$  voxels. To compute the requested  $\delta$ , one can use (12). Besides, we observe that the recovery score of EnCluDL improves when  $n$  increases, as expected. We also notice that the empirical  $\delta$ -FWER increases with  $n$ . To explain this effect, we first recall that theoretically the  $\delta$ -FWER is controlled for  $\delta$  equal to the largest cluster diameter, likely to be too large in practice. In this study, we have taken  $\delta$  equal to  $\delta_0$ , which is slightly smaller than the average radius of the clusters (cf. Sec. 7.6), since in practice this choice ensures the  $\delta$ -FWER control. However, when the setting is particularly favorable for inference (*e.g.*, if  $\log(n)/C > 1.5 \times 10^{-2}$ ), some false discoveries can be made at a distance greater than the average radius from the support. The choice of  $\delta$

is further discussed in Sec. 3.5 and in appendix in Sec. 7.5. Additionally, we can notice from Fig. 9 that for a fixed  $C/n$  ratio the recovery capability is stable (see also appendix Sec. 7.9). Then, as discussed in Sec. 3.5, we advise taking  $C$  of the same order as  $n$  (e.g.,  $C \in [n/2, n]$ ) when the goal is to recover most of the predictive regions without strict requirements on the accuracy of their shapes —since the value of  $\delta$  given by (12) might be not small with regards to the predictive region itself.

## 6 Discussion

Decoding models are fundamental for causal interpretation of the implication of brain regions for an outcome of interest, mental process or disease status [Weichwald et al., 2015]. They produce weight maps that are needed to support this type of inference [Poldrack, 2011, Varoquaux et al., 2018]. These weight maps capture how brain regions relate to the outcome, *conditional on* the other regions, which is a key difference with respect to standard brain mapping based on mass univariate models. However, the weight maps produced by the common decoders come without statistical guarantees. Indeed, decoders optimize the quality of their prediction, but give no control on conditional feature importance. This is difficult due to the large number of covariates —voxels— as well as the severe multi-collinearity: voxel-level inference is untenable. On the other hand, given the spatial structure of the data, a spatial tolerance in the statistical control is natural, as in Gaussian random field theory used in standard analysis [Nichols, 2012].

Our first contribution is to formalize this spatial statistical control by introducing the  $\delta$ -FWER, a control of false discoveries up to a spatial slack  $\delta$ . This definition uncovers a fundamental trade-off between accuracy in the localization of the brain structures involved and statistical power: here we deliberately degrade spatial accuracy, acknowledging current concerns on statistical power in neuroimaging studies [Button et al., 2013, Noble et al., 2019].

Our second contribution is to study empirically the statistical control of four procedures computing decoding maps, ranging from thresholding procedures applied to SVR weights, to a dedicated decoding procedure, EnCluDL. Experiments show that the Thr-SVR procedure, thresholding SVR weights, fails to achieve useful statistical control. Exact permutation testing yields the expected statistical control but with very poor statistical power for all experimental settings we have studied. On the other hand, Adaptive Permutation Threshold SVR (Ada-SVR) [Gaonkar and Davatzikos, 2012], does not control the FWER as it should, though it exhibits a fair precision-recall curve in our semi-simulated experiments. This shows how difficult it is to identify a statistically valid threshold for SVR weight maps. This is due to the fact that under the null hypothesis, estimated weights are not distributed according to a fixed distribution —notably because of the dependency structure of the data— and more precisely, the variance of these distributions differs. Then, thresholding linear decoders (SVR, logistic regression) based on their estimated weights amplitudes is not a principled approach to control false discoveries.

EnCluDL uses a different decoding procedure to estimate the weight maps [Chevalier

693 et al., 2018], and as a result comes with theoretical statistical guarantees: it controls  
694 the  $\delta$ -FWER for a predetermined tolerance parameter  $\delta$  equal to the largest diameter  
695 of the clusters, assuming that the observed samples are i.i.d. and that the weight maps  
696 are homogeneous and sparse. The experiments show that, indeed, for i.i.d. scenarios,  
697 EnCluDL controls the  $\delta$ -FWER for  $\delta$  equal to the average radius of the clusters. Though,  
698 in some very high SNR or high sample size regimes, it might be necessary to take  $\delta$  larger  
699 than the average radius (see Sec. 7.5). In practice, our choice of  $\delta$  is conservative, and  
700 with current fMRI datasets,  $\delta$ -FWER control holds for smaller  $\delta$ , even in relatively large  
701 cohorts ( $n = 1200$ ).

702 In our experiments, the spatial tolerance is around 1cm. Given that the definition  
703 of spatial location is blurred by inter-subject variability in group studies, this tolerance  
704 does not seem problematic. The method can thus be used for inference in cognitive  
705 neuroscience and population studies in psychiatry, neurology or epidemiology.

706 In addition, EnCluDL exhibits the best support recovery performance in the pro-  
707 posed semi-simulated experiments with fMRI data but also finds patterns with good  
708 face validity in more qualitative experiments plotted in Fig. 7. On the other hand, we  
709 also notice that EnCluDL tends to be over-conservative. Taking into account the diffi-  
710 culty of the problem and the fact that the convergence results are only asymptotic, we  
711 do not recommend using EnCluDL with  $n < 100$ .

712 In the present study, we have considered that the confounding variable effects have  
713 been removed during fMRI data preprocessing. However, it is still possible to include an  
714 additional confounding variable to the covariates before performing the inference. With  
715 regards to EnCluDL, we note that confounding variables should be handled separately  
716 from the clustered brain features.

717 Although it is not the main purpose of this study, we also checked the prediction  
718 performance of the decoders produced by each method. It is important to note that  
719 EnCluDL has been designed for the recovery of conditional statistical associations, not  
720 for prediction. In practice, the prediction performance is almost the same for SVR and  
721 Ada-SVR, and is slightly better than the one of EnCluDL (see Fig. 24). For prediction  
722 purpose, we recommend using *Fast Regularized Ensembles of Models* (FReM) [Hoyos-  
723 Idrobo et al., 2018], which is a stable and computationally efficient decoder with state-  
724 of-the-art prediction performance.

725 For pedagogical purpose, we have also considered a dataset where cross-sample in-  
726 dependence is violated due to serial correlation, reproducing an experiment of Eklund  
727 et al. [2016]. The ensuing loss of statistical control underlines the importance of the  
728 i.i.d. hypothesis. Hence, EnCluDL should not be used to make inference from intra-  
729 subject dataset recorded over one session. With these warnings in mind, we think that  
730 EnCluDL can be used safely in neuroimaging context. Our code, implemented with  
731 Python 3, can be found on <https://github.com/ja-che/hidimstat> along with some  
732 examples.

733 We have not considered the method proposed by Nguyen et al. [2019] based on the  
734 Knockoff filters [Barber and Candès, 2015, Candès et al., 2018] that yet appear to be an  
735 appealing procedure, as it can only control the FDR. In this study we have focused on

736  $\delta$ -FWER control, and hence defer the analysis of FDR-controlling procedures to future  
737 work. Also, we have not benchmarked post-selection inference procedures [Lee et al.,  
738 2016, Berk et al., 2013], as we found them challenging to run in high dimensional settings  
739 and prone to numerical underflows.

740 Our empirical results clearly show that standard thresholding procedures, including  
741 classical permutation tests, are not reliable to infer regions importance on decoder maps,  
742 due to the high number of covariates. Since, in neuroimaging studies, these maps are  
743 used to give evidence on the brain regions that supports an outcome, it is crucial to use  
744 a procedure with statistical control on the brain maps. Our study shows that EnCluDL  
745 provides such a control.

746 **Acknowledgements.** Data were provided in part by the Human Connectome Project,  
747 WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil;  
748 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH  
749 Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neu-  
750 roscience at Washington University. This project has received funding from the Euro-  
751 pean Union's Horizon 2020 research and innovation program under grant agreement No  
752 826421, and has been supported by Labex DigiCosme (project ANR-11-LABEX-0045-  
753 DIGICOSME) operated by ANR as part of the program "Investissement d'Avenir" Idex  
754 Paris Saclay (ANR-11-IDEX-0003-02), as well as by the ANR-17-CE23-0011 Fast-Big  
755 project .

756 **Ethics statement.** No experiments on living beings were performed for this study.  
757 Hence, IRB approval was not necessary. The data that we used were acquired in original  
758 studies that had received approval by the original institution's IRB. All data were used  
759 accordingly to respective usage guidelines.

## 760 References

- 761 A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gram-  
762 fort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-  
763 learn. *Frontiers in neuroinformatics*, 8:14, 2014. 13, 14
- 764 M. J. Anderson. Permutation tests for univariate or multivariate analysis of variance and  
765 regression. *Canadian journal of fisheries and aquatic sciences*, 58(3):626–639, 2001. 5
- 766 R. F. Barber and E. Candès. Controlling the false discovery rate via knockoffs. *Ann.*  
767 *Statist.*, 43(5):2055–2085, 10 2015. 7, 25
- 768 Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and  
769 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57  
770 (1):289–300, 1995. 4

- 771 R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann.*  
772 *Statist.*, 41(2):802–837, 2013. 26
- 773 A. Bowring, F. Telschow, A. Schwartzman, and T. E. Nichols. Spatial confidence sets  
774 for raw effect size images. *NeuroImage*, 203:116187, 2019. 4
- 775 L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 6
- 776 P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19  
777 (4):1212–1242, 09 2013. 6
- 778 K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson,  
779 and M. R. Munafò. Power failure: why small sample size undermines the reliability  
780 of neuroscience. *Nature Reviews Neuroscience*, 14:365, 2013. 24
- 781 E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold:‘model-X’ knockoffs for high  
782 dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 80  
783 (3):551–577, 2018. 7, 25
- 784 J.-A. Chevalier, J. Salmon, and B. Thirion. Statistical inference with ensemble of clus-  
785 tered desparsified lasso. In *International Conference on Medical Image Computing*  
786 and *Computer-Assisted Intervention*, pages 638–646. Springer, 2018. 6, 7, 10, 11, 24
- 787 C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297,  
788 1995. 2, 15
- 789 B. Da Mota, V. Fritsch, G. Varoquaux, T. Banaschewski, G. J. Barker, A. L. W. Bokde,  
790 U. Bromberg, P. J. Conrod, J. Gallinat, H. Garavan, J.-L. Martinot, F. Nees, T. Paus,  
791 Z. Pausova, M. Rietschel, M. N. Smolka, A. Ströhle, V. Frouin, J.-B. Poline, and  
792 B. Thirion. Randomized parcellation based inference. *NeuroImage*, 89:203–215, 2014.  
793 4
- 794 F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano.  
795 Combining multivariate voxel selection and support vector machines for mapping and  
796 classification of fMRI spatial patterns. *Neuroimage*, 43(1):44–58, 2008. 2
- 797 O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl,  
798 G. D. Pearlson, and V. D. Calhoun. A review of challenges in the use of fMRI for dis-  
799 ease classification/characterization and a projection pursuit application from a multi-  
800 site fMRI schizophrenia study. *Brain imaging and behavior*, 2(3):207–226, 2008. 2
- 801 R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference:  
802 Confidence intervals,  $p$ -values and R-Software hdi. *Statist. Sci.*, 30(4):533–558, 2015.  
803 6, 33
- 804 O. J. Dunn. Multiple comparisons among means. *J. Amer. Statist. Assoc.*, 56(293):  
805 52–64, 1961. 11

- 806 A. Eklund, T. Nichols, and H. Knutsson. Cluster failure: Why fMRI inferences for  
807 spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.*, 113  
808 (28):7900–7905, 2016. [16](#), [25](#), [40](#)
- 809 Y. Fan, N. Batmanghelich, C. M. Clark, C. Davatzikos, and Alzheimer’s Disease Neu-  
810 roimaging Initiative. Spatial patterns of brain atrophy in mci patients, identified via  
811 high-dimensional pattern classification, predict subsequent cognitive decline. *Neu-*  
812 *roimage*, 39(4):1731–1743, 2008. [2](#)
- 813 J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization.  
814 *Ann. Appl. Stat.*, 1(2):302–332, 2007. [6](#)
- 815 K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frack-  
816 owiak. Statistical parametric maps in functional imaging: a general linear approach.  
817 *Human brain mapping*, 2(4):189–210, 1994. [13](#)
- 818 B. Gaonkar and C. Davatzikos. Deriving statistical significance maps for svm based  
819 image classification and group comparisons. In *International Conference on Medical*  
820 *Image Computing and Computer-Assisted Intervention*, pages 723–730. Springer, 2012.  
821 [5](#), [9](#), [18](#), [24](#)
- 822 J. R. Gimenez and J. Zou. Discovering conditionally salient features with statistical  
823 guarantees. *International Conference on Machine Learning*, pages 2290–2298, 2019. [4](#)
- 824 M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub,  
825 K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van  
826 Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536:171–178,  
827 2016. [4](#)
- 828 A. Gramfort, G. Varoquaux, and B. Thirion. Beyond brain reading: randomized spar-  
829 sity and clustering to simultaneously predict and identify. In *Machine Learning and*  
830 *Interpretation in Neuroimaging*, pages 9–16. Springer, 2012. [6](#)
- 831 A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fMRI  
832 with TV-L1 prior. In *2013 International Workshop on Pattern Recognition in Neu-*  
833 *roimaging*, pages 17–20. IEEE, 2013. [2](#)
- 834 S. Haufe, Frank Meinecke, Kai G., S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bieß-  
835 mann. On the interpretation of weight vectors of linear models in multivariate neu-  
836 roimaging. *Neuroimage*, 87:96–110, 2014. [2](#)
- 837 J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini.  
838 Distributed and overlapping representations of faces and objects in ventral temporal  
839 cortex. *Science*, 293(5539):2425–2430, 2001. [2](#)
- 840 J.-D. Haynes and G. Rees. Neuroimaging: decoding mental states from brain activity  
841 in humans. *Nature Reviews Neuroscience*, 7(7):523, 2006. [2](#)

- 842 T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans.*  
843 *Pattern Anal. Mach. Intell.*, 20(8):832–844, 1998. 6
- 844 Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. Wiley Series in  
845 Probability and Statistics. John Wiley & Sons, Inc., 1987. 3, 4, 8
- 846 A. Hoyos-Idrobo, G. Varoquaux, Y. Schwartz, and B. Thirion. Frem—scalable and stable  
847 decoding with fast regularized ensemble of models. *NeuroImage*, 180:160–172, 2018.  
848 6, 11, 12, 25
- 849 L. Janson and W. Su. Familywise error rate control via knockoffs. *Electron. J. Stat.*, 10  
850 (1):960–975, 2016. 7
- 851 A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-  
852 dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014. 6, 33
- 853 L. I. Kuncheva and J. J. Rodríguez. Classifier ensembles for fMRI data analysis: an  
854 experiment. *Magnetic resonance imaging*, 28(4):583–593, 2010. 6
- 855 L. I. Kuncheva, J. J. Rodríguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston.  
856 Random subspace ensembles for fMRI classification. *IEEE Trans. Med. Imaging*, 29  
857 (2):531–542, 2010. 6
- 858 J. Lee, D. Sun, Y. Sun, and J. Taylor. Exact post-selection inference, with application  
859 to the lasso. *Ann. Statist.*, 44(3):907–927, 2016. 26
- 860 S. Lee, S. Halder, A. Kübler, N. Birbaumer, and R. Sitaram. Effective functional map-  
861 ping of fmri data with support-vector machines. *Human brain mapping*, 31(10):1502–  
862 1511, 2010. 2
- 863 N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression.  
864 *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009. 6, 11, 34
- 865 J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying  
866 brain states and determining the discriminating activation patterns: support vector  
867 machine on functional MRI data. *NeuroImage*, 28(4):980–995, 2005. 2
- 868 T.-B. Nguyen, J.-A. Chevalier, and B. Thirion. Ecko: Ensemble of clustered knock-  
869 offs for robust multivariate inference on fMRI data. In *International Conference on*  
870 *Information Processing in Medical Imaging*, pages 454–466. Springer, 2019. 4, 7, 25
- 871 T. E. Nichols. Multiple testing corrections, nonparametric methods, and random field  
872 theory. *Neuroimage*, 62:811, 2012. 24
- 873 S. Noble, D. Scheinost, and R. Constable. Cluster failure or power failure? evaluating  
874 sensitivity in cluster-level inference. *NeuroImage*, 209:116468, 12 2019. 24

- 875 K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-  
876 voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430, 2006.  
877 2
- 878 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel,  
879 P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,  
880 M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python.  
881 *J. Mach. Learn. Res.*, 12:2825–2830, 2011. 13
- 882 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a  
883 tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009. 2, 5
- 884 A. L. Pinho, A. Amadon, T. Ruest, M. Fabre, E. Dohmatob, I. Denghien, C. Ginisty,  
885 S. Becuwe-Desmidt, S. Roger, L. Laurier, et al. Individual brain charting, a high-  
886 resolution fMRI dataset for cognitive mapping. *Scientific data*, 5:180105, 2018. 14
- 887 R. A. Poldrack. Inferring mental states from neuroimaging data: from reverse inference  
888 to large-scale decoding. *Neuron*, 72(5):692–697, 2011. 2, 24
- 889 R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data  
890 analysis*. Cambridge University Press, 2011. 2
- 891 A. K. Rehme, L. J. Volz, D.-L. Feis, I. Bomilcar-Focke, T. Liebig, S. B. Eickhoff, G. R.  
892 Fink, and C. Grefkes. Identifying neuroimaging markers of motor disability in acute  
893 stroke by machine learning techniques. *Cerebral cortex*, 25(9):3046–3056, 2015. 2
- 894 A. Rizk-Jackson, D. Stoffers, S. Sheldon, J. Kuperman, A. Dale, J. Goldstein, J. Corey-  
895 Bloom, R. A. Poldrack, and A. R. Aron. Evaluating imaging biomarkers for neu-  
896 rodegeneration in pre-symptomatic Huntington’s disease using machine learning tech-  
897 niques. *Neuroimage*, 56(2):788–796, 2011. 2, 5
- 898 J. R. Sato, R. Basilio, F. F. Paiva, G. J. Garrido, I. E. Bramati, P. Bado, F. Tovar-  
899 Moll, R. Zahn, and J. Moll. Real-time fmri pattern decoding and neurofeedback using  
900 friend: an fsl-integrated bci toolbox. *PLoS One*, 8(12):e81658, 2013. 2
- 901 Y. Schwartz, B. Thirion, and G. Varoquaux. Mapping cognitive ontologies to and from  
902 the brain. In *Advances in neural information processing systems*, pages 1673–1681,  
903 2013. 2
- 904 A. Schwartzman, R. F. Dougherty, J. Lee, D. Ghahremani, and J. E. Taylor. Empirical  
905 null and false discovery rate analysis in neuroimaging. *Neuroimage*, 44(1):71–82, 2009.  
906 9
- 907 S. M. Smith and T. E. Nichols. Threshold-free cluster enhancement: addressing problems  
908 of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*,  
909 44(1):83–98, 2009. 4

- 910 A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Stat. Comput.*,  
911 14(3):199–222, 2004. 2
- 912 B. Thirion, G. Varoquaux, E. Dohmatob, and J.-B. Poline. Which fMRI clustering gives  
913 good brain parcellations? *Frontiers in Neuroscience*, 8:167, 2014. 6, 11
- 914 R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B  
915 Stat. Methodol.*, 58(1):267–288, 1996. 6
- 916 S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal  
917 confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–  
918 1202, 2014. 6, 11, 12, 40
- 919 S. Van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for  
920 efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30,  
921 2011. 13
- 922 D. C. Van Essen, K. Ugurbil, E. J. Auerbach, D. M. Barch, T. E. J. Behrens, R. Bu-  
923 cholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. A. Feinberg,  
924 M. F. Glasser, N. Harel, A. C. Heath, L. J. Larson-Prior, D. S. Marcus, G. Michalar-  
925 eas, S. Moeller, R. Oostenveld, S. E. Petersen, F. W. Prior, B. L. Schlaggar, S. M.  
926 Smith, A. Z. Snyder, J. Xu, and E. Yacoub. The Human Connectome Project: a data  
927 acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012. 14
- 928 G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging.  
929 *GigaScience*, 3(1):28, 2014. 2
- 930 G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recov-  
931 ery on spatially correlated designs with randomization and clustering. In *International  
932 Conference on Machine Learning*, 2012. 6, 11
- 933 G. Varoquaux, Y. Schwartz, R. A. Poldrack, B. Gauthier, D. Bzdok, J.-B. Poline, and  
934 B. Thirion. Atlases of cognition with large-scale human brain mapping. *PLoS com-  
935 putational biology*, 14(11):e1006565, 2018. 2, 24
- 936 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau,  
937 E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett,  
938 J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern,  
939 E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde,  
940 J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald,  
941 A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0:  
942 Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.  
943 13
- 944 M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery  
945 using  $\ell_1$ -constrained quadratic programming (lasso). *IEEE Trans. Image Process.*, 55  
946 (5):2183–2202, 2009. 5

- 947 Y. Wang, J. Zheng, S. Zhang, X. Duan, and H. Chen. Randomized structural sparsity  
 948 via constrained block subsampling for improved sensitivity of discriminative voxel  
 949 identification. *Neuroimage*, 117:170–183, 2015. 6
- 950 L. Wasserman and K. Roeder. High-dimensional variable selection. *Ann. Statist.*, 37  
 951 (5A):2178–2201, 2009. 6
- 952 S. Weichwald, T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, and M. Grosse-Wentrup.  
 953 Causal interpretation rules for encoding and decoding models in neuroimaging. *Neu-  
 954 roimage*, 110:48–59, 2015. 2, 3, 13, 20, 24
- 955 P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and meth-  
 956 ods for p-value adjustment*, volume 279. John Wiley & Sons, 1993. 5, 9, 13
- 957 C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in  
 958 high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(1):217–242,  
 959 2014. 6, 11, 32, 33, 40
- 960 Z.-H. Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC,  
 961 2012. 6

## 962 7 Appendix

### 963 7.1 Desparsified Lasso

964 **Additional notation.** For a matrix  $\mathbf{X}$ ,  $\mathbf{X}_{i,\cdot}$  refers to the  $i$ -th row and  $\mathbf{X}_{\cdot,j}$  to the  $j$ -th  
 965 column,  $\mathbf{X}_{i,j}$  refers to the element  $(i, j)$ , and  $\mathbf{X}^{(-j)}$  refers to the matrix  $\mathbf{X}$  without the  
 966  $j$ -th column.  $\mathbf{X}^\dagger$  denotes the Moore-Penrose pseudo-inverse of  $\mathbf{X}$ .

967 **Small-dimension insight.** The Desparsified Lasso procedure, introduced by [Zhang](#)  
 968 and [Zhang \[2014\]](#) extends the Ordinary Least Squares (OLS) procedure to  $n < p$  cases.  
 969 Let us first recall the standard OLS framework ( $n > p$ ). Starting from model (1), let us  
 970 define  $\mathbf{z}_j \in \mathbb{R}^n$  the residual of the OLS regression of  $\mathbf{X}_{\cdot,j}$  versus  $\mathbf{X}^{(-j)}$  given by:

$$\mathbf{z}_j = \mathbf{X}_{\cdot,j} - \mathbf{X}^{(-j)} \hat{\mathbf{w}}^{(-j)}, \quad (14)$$

971 where  $\hat{\mathbf{w}}^{(-j)}$  refers to the estimator of the OLS regression of  $\mathbf{X}_{\cdot,j}$  versus  $\mathbf{X}^{(-j)}$ . In  
 972 particular,  $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$  for all  $k \in [p] \setminus \{j\}$ . In this setting, we also have the following  
 973 result:

974 **Proposition 7.1.** *If  $n > p$  and  $\text{rank}(\mathbf{X}) = p$ , then, for all  $j \in [p]$ :*

$$\hat{\mathbf{w}}_j^{\text{OLS}} = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}, \quad (15)$$

975 where  $\hat{\mathbf{w}}^{\text{OLS}}$  is the parameter vector estimates obtained from the OLS regression of  $\mathbf{y}$   
 976 against  $\mathbf{X}$ .

977 **Desparsified Lasso.** In this setting, it is not possible to construct a non-zero vector  
978 family  $\{\mathbf{z}_j, j \in [p]\}$  (*i.e.*, a family verifying  $\mathbf{z}_j \neq \mathbf{0}$  for all  $j \in [p]$ ), such that  $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$   
979 for all  $k \neq j$ . The idea proposed by [Zhang and Zhang \[2014\]](#) is to construct a family  
980  $\{\mathbf{z}_j, j \in [p]\}$  which would play the same role as the residual of the OLS regression of  $\mathbf{X}_{\cdot,j}$   
981 versus  $\mathbf{X}^{(-j)}$  in (14) but relaxing (slightly) the constraint  $\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} = 0$ . To do so, instead  
982 of computing  $\{\mathbf{z}_j, j \in [p]\}$  by OLS regression, they proposed to take the residual of the  
983 Lasso regressions<sup>1</sup> of  $\mathbf{X}_{\cdot,j}$  against  $\mathbf{X}^{(-j)}$ . Then, from (1), one can derive the following:

$$\frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} = \mathbf{w}_j^* + \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} + \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \mathbf{w}_k^*}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} . \quad (16)$$

984 Noticing that the second term in (16) is a noise term and plugging in an initial estimator  
985  $\hat{\mathbf{w}}^{(\text{init})}$  of  $\mathbf{w}^*$  in the third term —a standard choice being the Lasso— they propose the  
986 following estimator:

$$\hat{\mathbf{w}}_j = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} \hat{\mathbf{w}}_k^{(\text{init})}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}} . \quad (17)$$

987 Here, one can notice that (17) generalizes (15) to  $n < p$ . Then, from (16) and (17) one  
988 can derive:

$$\sigma_\varepsilon^{-1} (\hat{\mathbf{w}}_j - \mathbf{w}_j^*) = \underbrace{\sigma_\varepsilon^{-1} \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}}_{\eta_j} + \underbrace{\sigma_\varepsilon^{-1} \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{X}_{\cdot,k} (\mathbf{w}_k^* - \hat{\mathbf{w}}_k^{(\text{init})})}{\mathbf{z}_j^\top \mathbf{X}_{\cdot,j}}}_{\mu_j} . \quad (18)$$

989 This yields:

$$\sigma_\varepsilon^{-1} (\hat{\mathbf{w}} - \mathbf{w}^*) = \boldsymbol{\eta} + \boldsymbol{\mu}, \quad \boldsymbol{\eta} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Omega}) , \quad (19)$$

990 where:

$$\boldsymbol{\Omega}_{jk} = \frac{\mathbf{z}_j^\top \mathbf{z}_k}{(\mathbf{z}_j^\top \mathbf{X}_{\cdot,j})(\mathbf{z}_k^\top \mathbf{X}_{\cdot,k})} . \quad (20)$$

991 Asymptotically and under some sparsity assumptions (one can refer to [Dezeure et al.,  
992 2015](#) for more details), one can neglect the last term  $\boldsymbol{\mu}$  and obtain:

$$\sigma_\varepsilon^{-1} (\boldsymbol{\Omega}_{jj})^{-1/2} (\hat{\mathbf{w}}_j - \mathbf{w}_j^*) \sim \mathcal{N}(0, 1) . \quad (21)$$

993 From (21), one can compute the confidence intervals and p-values of the coefficients of  
994 the estimated weight map. Note that similar estimators have been derived in parallel in  
995 [Javanmard and Montanari \[2014\]](#).

---

<sup>1</sup>From our analysis, taking  $\lambda_j$ , the regularization parameter used in the Lasso regression of  $\mathbf{X}_{\cdot,j}$  against  $\mathbf{X}^{(-j)}$ , equal to  $0.01 \times \max_{k \in [p] \setminus \{j\}} |\mathbf{X}_{\cdot,j}^\top \mathbf{X}_{\cdot,k}| / n$  is appropriate to compute  $\mathbf{z}_j$ . Empirically, it results in a more conservative solution than the one proposed by [Zhang and Zhang \[2014\]](#) but it avoids doing computationally expensive grid-search.

996 **7.2 Adaptive quantile aggregation of p-values**

For the  $j$ -th voxel, 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]:

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

997 where we have taken  $\gamma_{\min} = 0.20$  in our experiments. Taking a value of  $\gamma_{\min}$  not too  
 998 small (*e.g.*,  $\gamma_{\min} \geq 0.20$ ) ensures that the discovered sources have received small p-values  
 999 many times (*e.g.*, at least for  $B/5$  different choices of clustering).

1000 **7.3 Empirical analysis of data structure impact**

1001 In this section, we propose two simulations to gain more insight concerning the assump-  
 1002 tions about data structure that are necessary for Desparsified Lasso and EnCluDL to  
 1003 have power. More precisely, we investigate up to which level of correlation two corre-  
 1004 lated predictive features (having non-zero weight) are both identified. Indeed, when two  
 1005 predictive features are highly correlated, there is a risk that the inference procedure only  
 1006 detects one of the two.

1007 The first simulation has modest data dimension, which corresponds to that of data  
 1008 after clustering. We use it to analyze the behavior of Desparsified Lasso. The second  
 1009 simulation has a 2D structure with larger data dimension, it introduces short- and long-  
 1010 range correlation structure, it is used to study EnCluDL.

1011 **First simulation: approximating the clustered data setting.** In this simulation  
 1012 we set  $n = 100$  and  $p = 500$ . We construct the design matrix  $\mathbf{X}$  such that features are  
 1013 normally distributed and the first two features have a correlation equal to parameter  $\rho$ ,  
 1014 while all the other features are independent. The weight  $\mathbf{w}^*$  is such that  $\mathbf{w}_j^* = 1$  for  
 1015  $1 \leq j \leq 10$  and  $\mathbf{w}_j^* = 0$  otherwise. We also set  $\sigma_\varepsilon = 1$  giving approximately  $\text{SNR}_y = 12$   
 1016 close to the SNR estimated in real fMRI datasets.

1017 To check the ability of Desparsified Lasso to identify two correlated features, we  
 1018 compare the smallest z-score of the first two first features (“correlated features”) with the  
 1019 smallest z-score of the two following features (“control features”) for different value of  $\rho \in$   
 1020  $(0, 1)$ . While the minimum z-score of the control features should not vary significantly  
 1021 and corresponds to a control value, the minimum z-score of the two correlated features  
 1022 should decrease towards 0 when  $\rho$  increases to 1. Also, we look at the z-score of a random  
 1023 non-predictive feature (“random null feature”) to get insight about the z-score threshold  
 1024 value to declare a feature significant.

1025 **First simulation results.** In Fig. 10, we give the results for the first simulation.  
 1026 When the correlation of the two correlated features increases, their identification using  
 1027 the Desparsified Lasso procedure becomes harder. In this experiment, we observe that

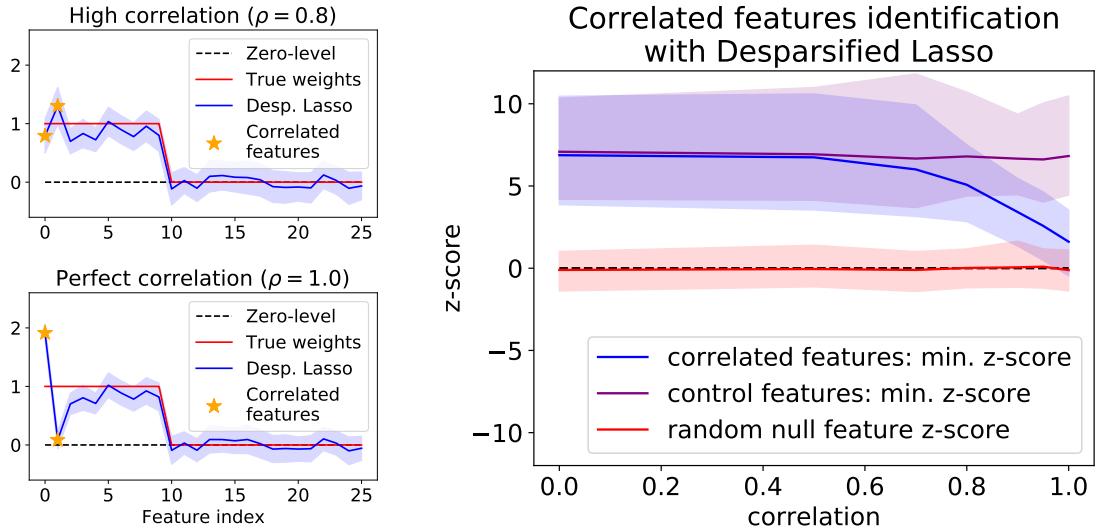
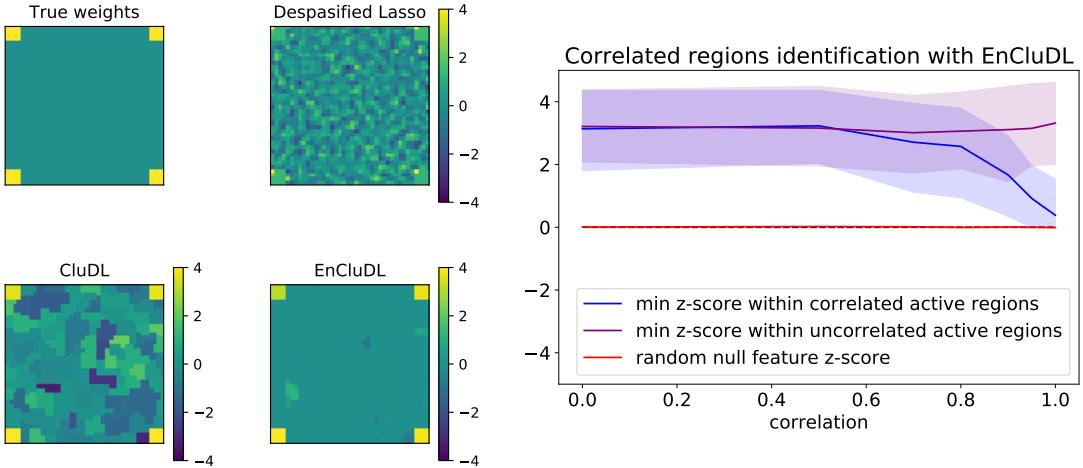


Figure 10: **Impact of correlation when trying to identify two correlated features.** Left: We plot the Desparsified Lasso estimator and its 95% confidence intervals. The correlation between the first two features is set to  $\rho$ , while the other features are uncorrelated. The higher  $\rho$  the harder it is to identify each of the two correlated features. For  $\rho = 1.0$ , it is impossible, while for  $\rho = 0.8$ , the identification of both features is successful. Right: Quantitative summary of the simulations. When the correlation increases the minimum z-score of the two first features (“correlated features”) decreases (90% confidence intervals also displayed). The correlation between the two following features (“control features”) remains equal to zero, thus the minimum z-score of these features is used as a control value that should not vary significantly. Also we plot the z-score of a random non-predictive feature (“random null feature”). We observe that for a correlation lower than 0.8 the deviation is limited and it is possible to identify the two correlated variables. For a correlation larger than 0.9 the deviation is massive and it becomes impossible to recover the two correlated variables.

1028 below a correlation of 0.8, Desparsified Lasso can identify accurately the two correlated  
 1029 variables. However, above a correlation of 0.9, Desparsified Lasso might fail to recover  
 1030 the both correlated variables.

1031 **Second simulation: 2D data structure.** The simulation we consider has a 2D data  
 1032 structure. It aims at approximating the short- and long-range correlation structure that  
 1033 can be observed in fMRI data (see Sec. 7.4). The feature space considered is a square  
 1034 with edge length  $H = 40$ , then  $p = H^2 = 1600$  features and we took  $n = 100$  samples.  
 1035 To construct  $\mathbf{w}^*$ , we define a 2D weight map  $\tilde{\mathbf{w}}^*$  of size  $H \times H$  with four active regions  
 1036 then we flatten  $\tilde{\mathbf{w}}^*$  in a vector  $\mathbf{w}^*$  of size  $p$ . Each active region is a small square of width  
 1037  $h = 4$ , leading to support of size  $4 \times h^2 = 64$ . The four active regions are located in the  
 1038 corners of the weight map. The true weight map is represented in Fig. 11. To construct



**Figure 11: Impact of correlation when trying to identify two correlated regions.** Left: True weight map, and z-scores estimated by Desparsified Lasso, CluDL and EnCluDL, obtained for  $\rho = 0.9$ . Desparsified Lasso cannot handle the extreme short-range correlation that occurs within each predictive region and only identifies one feature in each. CluDL and EnCluCL benefit from the clustering, as they identify all the features for every predictive regions. We can also observe that EnCluDL improves upon CluDL thanks to the smoothing effect produced by ensembling. Focusing on the EnCluDL solution, we can see that the z-score of the upper left active region is a bit lower than for the other active regions. This is due to the high correlation between the upper left and bottom right regions. Right: Summary of the results of the second simulation. When the correlation increases the minimum z-score within the correlated active regions decreases. The minimum z-score between the two uncorrelated regions is used as a control. We also plot the z-score of a random non-predictive feature, we notice that due to the ensembling step of EnCluDL, the empirical confidence intervals are much thinner than in Fig. 10. We observe that for a correlation lower than 0.8 the deviation is limited and it is possible to identify the two correlated predictive regions. For a correlation larger than 0.9 the deviation becomes large and recovering the two correlated regions becomes impossible.

the design matrix  $\mathbf{X}$ , we first construct a 2D matrix  $\tilde{\mathbf{M}}$  by drawing  $p$  random normal vectors of size  $n$  that are spatially smoothed with a 2D Gaussian filter (the smoothing is only made in the feature space for each sample independently, the samples are not mixed and remain independent). We flatten the vectors to go from  $\tilde{\mathbf{M}}$  of size  $n \times H \times H$  to  $\mathbf{M}$  of size  $n \times p$ . The spatial smoothing enforces a 2D structure on the data. Then, we further modify  $\mathbf{M}$  such that (i) all the features of an active region are perfectly correlated and (ii) two of the four active regions are correlated at a given value  $\rho \in (0, 1)$ , the two other active regions being unmodified (hence uncorrelated). The first transformation aims at showing that the clustering is useful to handle the short-range correlation that might be very high for fMRI data (see Sec. 7.4). The second transformation aims at

1049 testing whether EnCluDL can recover two correlated predictive regions; this is notably  
 1050 desirable in the case of long-range correlation (*e.g.*, two contralateral brain regions). The  
 1051 two uncorrelated regions are used to provide control values. With these transformations  
 1052 we obtain the design matrix  $\mathbf{X}$ . In Sec. 7.4, the two active regions that are correlated  
 1053 are located in the upper left corner and in the bottom right corner while the other two  
 1054 are uncorrelated. Finally, we also set  $\sigma_\varepsilon = 10$ , to approximately get  $\text{SNR}_y = 4$ .

1055 To check the ability of EnCluDL to identify two correlated regions, we compare  
 1056 the smallest z-score of the features that belong to one of the correlated regions with the  
 1057 smallest z-score of the features that belong to the uncorrelated active regions; we analyze  
 1058 the results for several values of  $\rho \in (0, 1)$ . To understand the effect of the clustering and  
 1059 ensembling, we compare Desparsified Lasso, CluDL and EnCluDL solutions qualitatively.  
 1060 Since the features that belong to the same active region are perfectly correlated, we  
 1061 expect that Desparsified Lasso identifies only one feature per region at best. We also  
 1062 report the z-score of a random non-predictive feature.

1063 **Second simulation results.** In Fig. 11, we give the results for the second simulation.  
 1064 Clustering turns out to be crucial to produce valid statistical inference solution in the  
 1065 presence of extreme short-range correlation. Additionally, we show that when the corre-  
 1066 lation of the two correlated active regions increases, their identification using EnCluDL  
 1067 becomes harder. In this experiment, we observe that below a correlation of 0.8, En-  
 1068 CluDL can identify accurately the two correlated regions. However, above a correlation  
 1069 of 0.9, EnCluDL generally fails to recover the two correlated regions.

## 1070 7.4 fMRI data structure

1071 In Sec. 7.3, we have shown that one may encounter multicollinearity issues. It is thus  
 1072 necessary to analyze the correlation structure of actual fMRI data.

1073 In Fig. 12, we study the correlation observed in the HCP900 Emotion task data.  
 1074 Considering correlation between random voxels, then neighboring voxels, we can see  
 1075 that the correlation is much higher in the case of neighboring voxel. Notably, the median  
 1076 correlation between two random voxels is 0.1 while the median correlation between two  
 1077 neighboring voxels is above 0.8, and often larger than 0.9. We have shown in Sec. 7.3, that  
 1078 Desparsified Lasso may fail to detect two features when they are so strongly correlated.

1079 Correlation histograms after clustering the data as shown in Fig. 12. For example,  
 1080 taking  $C = 500$  clusters, the median correlation between two random clusters is 0.3 while  
 1081 it is 0.7 for two neighboring clusters. Inter-cluster correlation always remains below 0.85  
 1082 and almost always below 0.8. In practice, we have shown in Sec. 7.3 that Desparsified  
 1083 Lasso can handle scenarios where features have correlation lower than 0.8.

## 1084 7.5 Estimating $\delta$ for which EnCluDL controls the $\delta$ -FWER

In Sec. 3.5, we recommend using  $\delta$ , in regular brain imaging settings with (12):

$$\delta_0 = \left( \frac{p}{2C} \right)^{1/3},$$

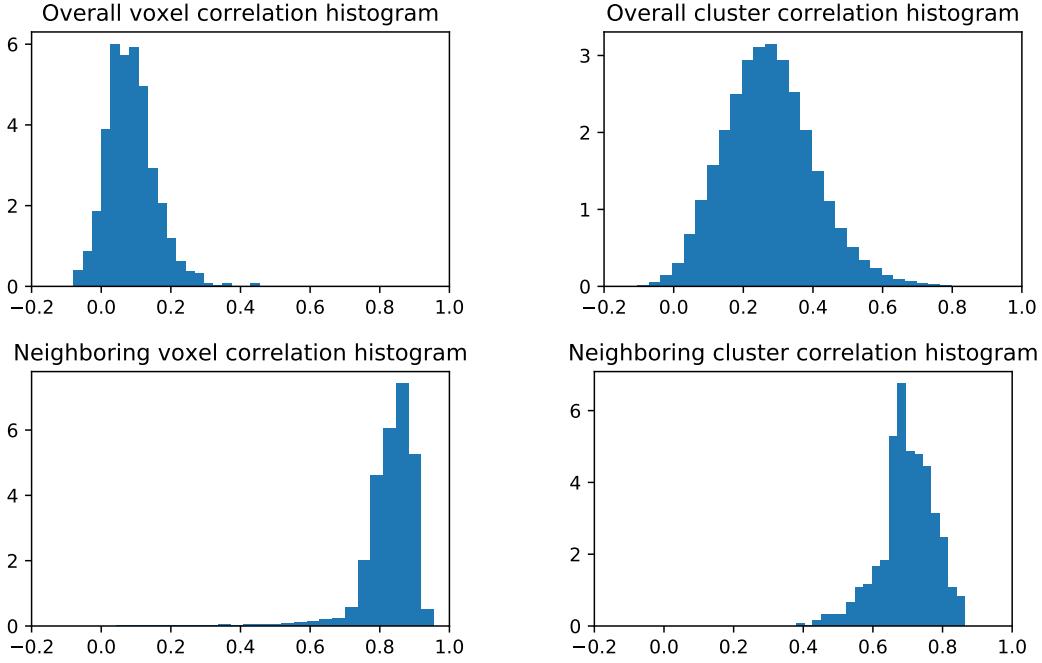


Figure 12: **Data structure in HCP900 emotion task.** Left: Correlation histogram of the fMRI data at voxel level. The correlation between two random voxels is quite low, a typical value being around 0.1. However, when looking at neighboring voxels, we observe that the correlation is often higher than 0.9. This exhibits the short- and long-range correlation structure but also suggests that raw Desparsified Lasso would not be adapted to this setting. Right: Correlation histogram of the clustered data for  $C = 500$ . The correlation between two random clusters is around 0.3, while the correlation between two neighboring clusters is around 0.7 and almost always below 0.8. Then, thanks to clustering, highly correlated voxels are aggregated into groups and Desparsified Lasso is adapted to this setting.

1085  $\delta_0$  being a distance in voxel unit close to the average radius of the clusters used in En-  
 1086 CluDL. However, when the setting is particularly favorable for inference, *i.e.*, if  $\log(n)/C$   
 1087 is large or  $\sigma_\varepsilon$  is small, the choice of  $\delta$  given by (12) may be over-optimistic and we might  
 1088 need to correct this formula. We have found empirically that a suitable multiplicative  
 1089 factor, denoted by  $\tau > 0$ , that could be used to correct  $\delta_0$  is given by:

$$\tau = -45 \log\left(\frac{\sigma_\varepsilon}{\text{std}(\mathbf{y})}\right) \frac{\log(n)}{C}, \quad (22)$$

1090 where  $\sigma_\varepsilon$  is the standard deviation of the noise  $\varepsilon$ . In practice  $\sigma_\varepsilon$  has to be estimated; in  
 1091 the fMRI datasets we studied, estimates of  $\frac{\sigma_\varepsilon}{\text{std}(\mathbf{y})}$  were close to 0.1. However, given the  
 1092 heuristic derivation of this quantity and the uncertainty about the value of  $\tau$ , we do not  
 1093 recommend correcting  $\delta_0$  with a factor lower than 1 as it could lead to a dramatic under-  
 1094 estimation of the valid  $\delta$ . Then, the final formula to compute the  $\delta$  such that  $\delta$ -FWER

1095 control is ensured, is:

$$\delta^* = \max(1, \tau) \delta_0 . \quad (23)$$

1096 Note that the formula given by (12) and even (23) are not bullet proof but rather give  
1097 reasonable estimates of  $\delta$ .

1098 **7.6 Cluster size analysis**

1099 In Sec. 3.5, we have proposed a formula to compute a valid spatial tolerance parameter  
1100  $\delta_0$ . In Fig. 13, we show that  $\delta_0$  is close but slightly lower than the average cluster radius.  
1101 Also, one can notice that taking a larger number of clusters, the size of the clusters is  
1102 smaller. As a consequence, the statistical control is valid for a lower spatial tolerance.  
1103 Finally, by looking at the shape of the distribution of the cluster radius, we observe that  
1104 there are only few large clusters.

1105 In general  $\delta_0$  is a suitable choice, however when the setting is particularly favorable  
1106 for inference, the mixing effect produced by ensembling might not be sufficient and voxels  
1107 far (further than  $\delta_0$ ) from the support might be discovered. This effect can be explained  
by the detection of large clusters that are overlapping the support and the null region.

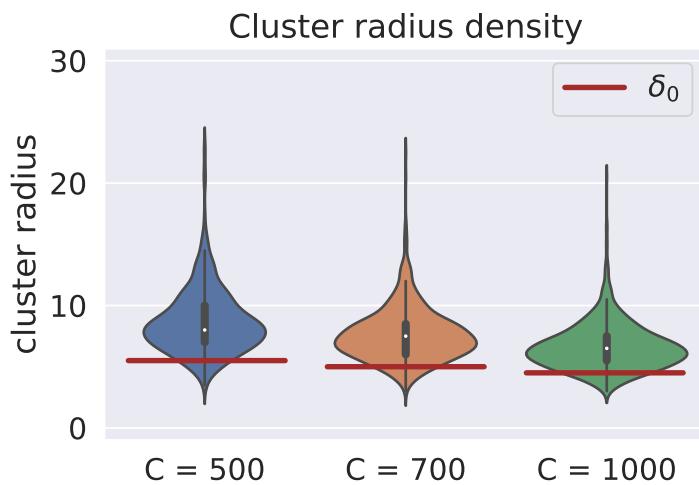


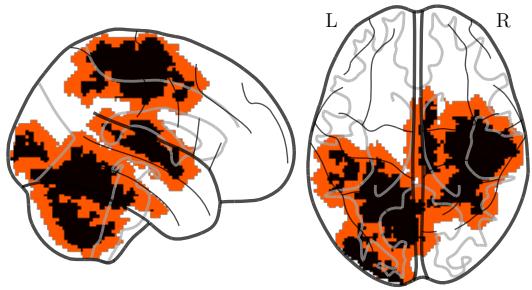
Figure 13: **Comparing  $\delta_0$  with the distribution of the cluster radius as a function of  $C$ .**  
By taking a larger number of clusters, we decrease the size of the clusters. The statistical  
control is thus valid for a smaller spatial tolerance. Comparing the distribution of the cluster  
radius with the recommended choice of spatial tolerance parameter  $\delta_0$ , we observe that  $\delta_0$   
is a bit lower than the empirical average cluster radius. Finally, we observe that few clusters  
are much wider than the others, this may occasionally lead to false discoveries far from the  
support in high SNR scenarios.

1108

1109 **7.7 Illustrating spatial tolerance on real brain geometry**

1110 In Fig. 14, we display a brain pattern with spatial tolerance in the case of the HCP data.

Figure 14: **Expanding HCP maps by 6 voxels.** The black-colored voxels represent the positive weights of the reference map constructed in Sec. 4.2. The red-colored voxels are the  $\delta$ -dilation of the previous map where  $\delta = 6$  voxels, i.e., the tolerance we have taken in all experiments. Then,  $\delta$ -FWER controls the false discoveries made outside of the colored voxels (see also Sec. 3.1).



1111

1112 **7.8 Statistical control under the global null with autocorrelated data**

1113 **Experiment.** In this experiment, we study how the different procedures control the  
1114 FWER when the data are temporally autocorrelated; hence violating the i.i.d. assumption.  
1115 Notably, this is the case if the data correspond to fMRI signal recordings of one  
1116 given subject during an acquisition. We consider data from the HCP900 resting-state  
1117 fMRI dataset described in Sec. 4.1 with full samples ( $n = 1200$ ). The design matrix  
1118  $\mathbf{X}$  contains the 15-minutes fMRI signal records. As in Eklund et al. [2016], we con-  
1119 struct  $\mathbf{y}$  such that it corresponds to two activity paradigms: block or event responses,  
1120 with several frequencies: 10s on/off, 20s on/off, 30s on/off, 2s-activation/6s-rest, 4s-  
1121 activation/8s-rest. Thus,  $\mathbf{y}$  is temporally autocorrelated. In these simulations  $\mathbf{w}^* = \mathbf{0}$   
1122 so the  $\delta$ -FWER and the classical FWER are identical. To better assess the impact of  
1123 correlation, we also generate  $\mathbf{y}$  as an i.i.d. —uncorrelated— Bernoulli or standard Gaus-  
1124 sian random variable (here again  $\mathbf{w}^* = \mathbf{0}$ ), breaking spurious correlations between  $\mathbf{X}$   
1125 and  $\mathbf{y}$ . These two cases enable to check if the procedures still control the FWER at  
1126 the targeted nominal level on this dataset under the i.i.d. hypothesis. For each kind of  
1127 response, we repeat the experiment 100 times, using data from 100 different subjects.

1128 **Results.** we now report the results of the experiment. In Fig. 15, we observe that for  
1129 all the fictitious block response paradigms, for every procedure, the empirical FWER ex-  
1130 ceeds the targeted nominal level (10%), as one would expect. This result is not surprising  
1131 since independence across samples is a key assumption for a valid statistical inference  
1132 with any of the four procedures. Notably, concerning EnCluDL, Desparsified Lasso needs  
1133 the i.i.d. hypothesis [Zhang and Zhang, 2014, van de Geer et al., 2014] to produce valid  
1134 confidence intervals or p-values. This assumption is not verified for the block or event

1135 response paradigms due to the temporal dependency in the data. However, when the  
 1136 target  $\mathbf{y}$  is i.i.d. —*i.e.*, without temporal dependency (Bernoulli or Gaussian random  
 1137 responses)— the FWER is controlled (except for Thr-SVR). Indeed, the model is no  
 longer confounded by the correlation structure underlying the data.

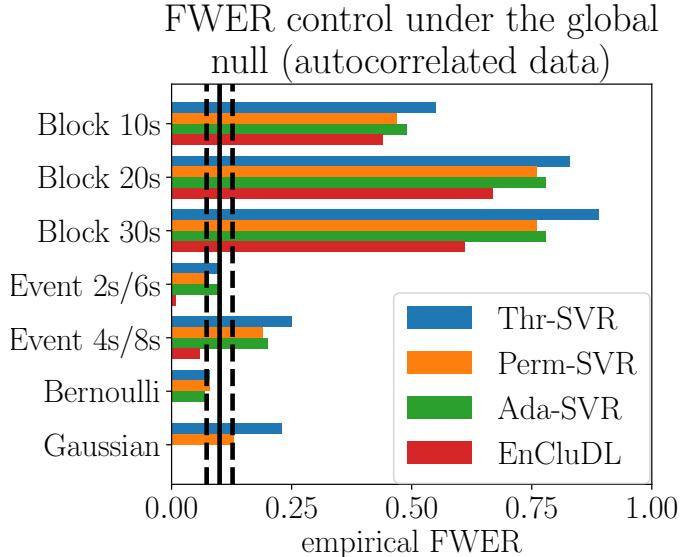


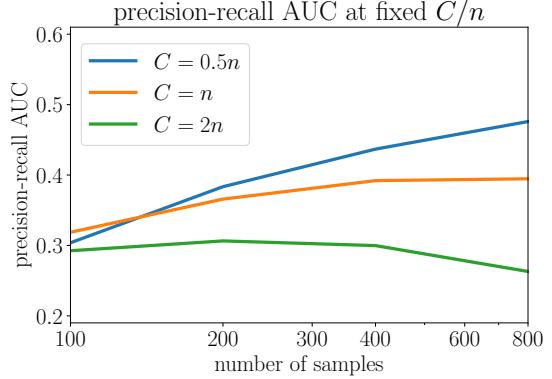
Figure 15: **FWER control under the global null with autocorrelated data.** The results of the experiment with correlated data under the global null, described in Sec. 7.8, show that, when the data are temporally autocorrelated, all the procedures fail to control the FWER. Indeed, for all the fictitious block response paradigms, the empirical FWER exceeds the targeted nominal level of 10% for every procedure. This result is not surprising as the procedures control the  $\delta$ -FWER under the hypothesis that the samples are i.i.d.; this is not the case for the block or event response paradigms. However, when the fictitious response breaks the temporal dependency (binary or Gaussian random responses), the i.i.d. hypothesis is met and the FWER is empirically well controlled except for the Thr-SVR procedure.

1138

## 1139 7.9 Influence of the $C/n$ ratio on the recovery property of EnCluDL

1140 When using EnCluDL, the number  $C$  of clusters is an arbitrary parameter. We proposed  
 1141 some default choice in Sec. 4.5, yet intuitively,  $C$  should adapt to the amount of data  
 1142 available: larger samples size lead to better estimation, allowing refined localization,  
 1143 hence higher  $C$ . In Fig. 16, we show on semi-simulated data that for  $C \in [n/2, n]$ ,  $C/n$   
 1144 being fixed, the precision-recall AUC on real data does not depend on  $n$ , suggesting to  
 1145 chose  $C$  proportional to  $n$ .

**Figure 16: Influence of the  $C/n$  ratio on the precision-recall AUC.** The results of the experiment described in Sec. 4.5 show that the precision-recall AUC depends almost linearly on  $\log(C/n)$  except when  $C$  is critically low creating very wide clusters and deteriorating the precision-recall curve. This limit depends on the physical properties of the problem; here,  $C$  should not be lower than 100. Keeping this limit in mind, we advise taking  $C \in [n/2, n]$  to recover most of the predictive regions.

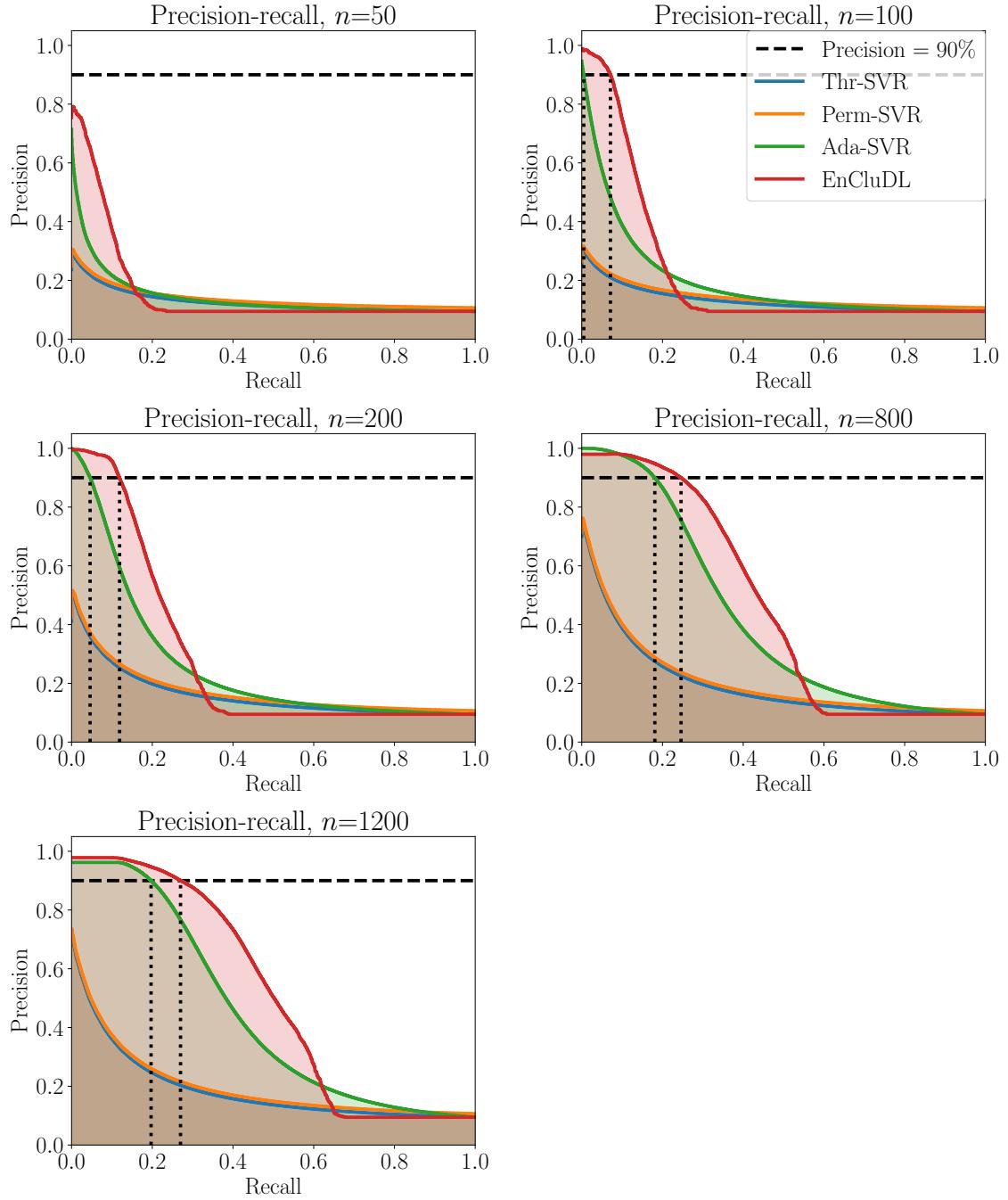


## 1146 7.10 Statistical control with known ground truth: additional plots

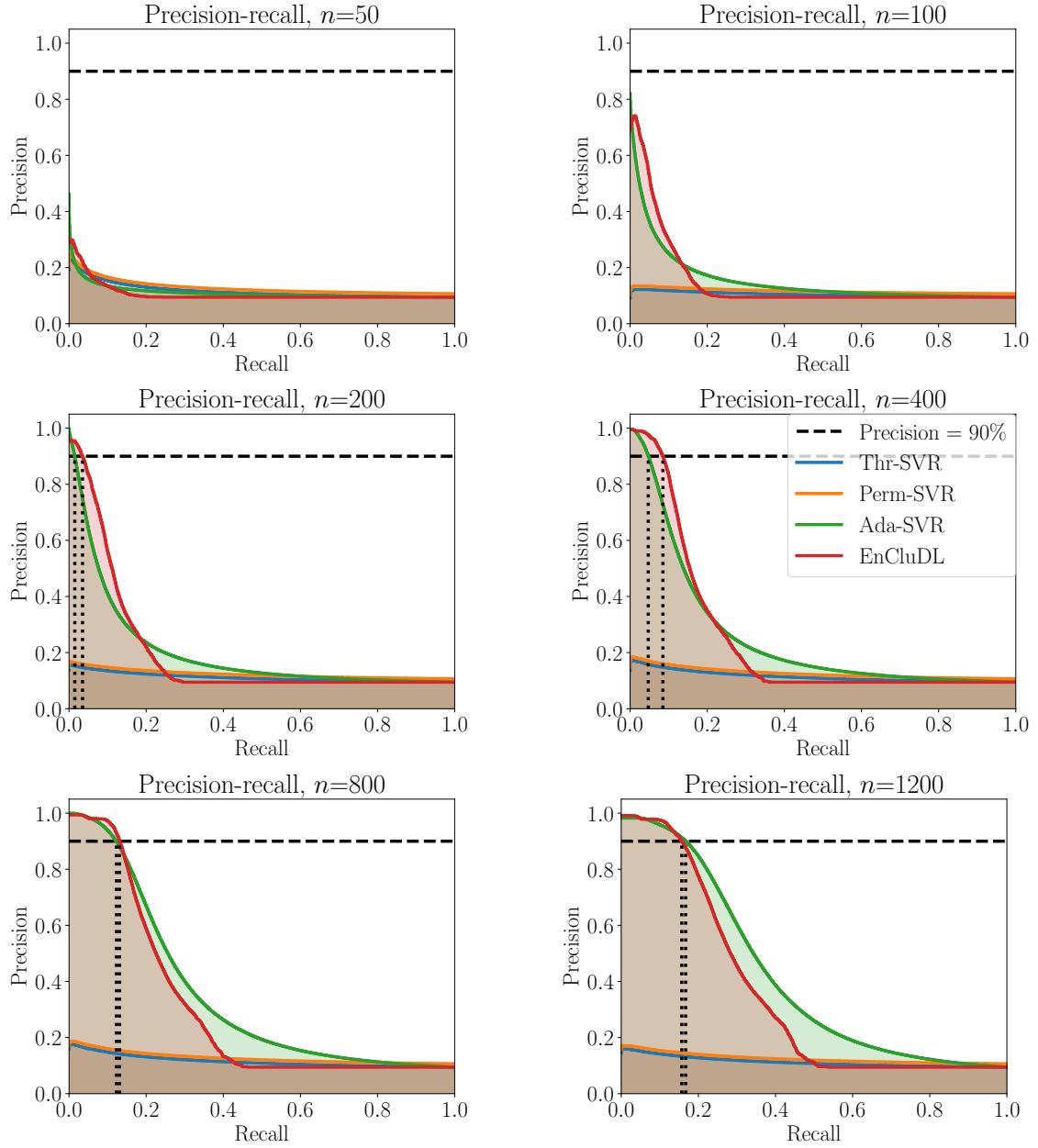
1147 In this section, we provide additional experimental results to assess the detection accuracy  
 1148 of the multivariate estimators, to complement the results in Sec. 4.2. Fig. 17 shows  
 1149 additional precision-recall curves, obtained for different values of  $n$ : these different settings  
 1150 preserve the relative performance of the methods, while larger  $n$  results in better  
 1151 curves. However, we do not recommend running such analysis with  $n < 100$ , since the  
 1152 estimation problem is hard and statistical guarantees only hold in asymptotic regime.  
 1153 Fig. 18 and Fig. 19 display the performance of the methods in terms of  $\delta$ -FWER control  
 1154 and precision-recall curves on semi-simulated data where  $\mathbf{y}$  is binary. This induces a violation  
 1155 of the EnCluDL model that reduces its performance in terms of  $\delta$  precision-recall.  
 1156 Yet, unlike Ada-SVR, it still controls the  $\delta$ -FWER accurately.

## 1157 7.11 Face validity on HCP dataset

1158 In Fig. 23, we plot the results for five tasks taken from the HCP dataset, besides of the  
 1159 two described in Sec. 4.6. For all methods, the statistical maps are thresholded such that  
 1160 the  $\delta$ -FWER stays lower than 10% for  $\delta = 12$  mm. Qualitatively, EnCluDL discovers  
 1161 the most plausible patterns, Ada-SVR often makes dubious discoveries, patterns are  
 1162 too wide and implausible, while the two other methods exhibit a very weak statistical  
 1163 power. As discussed in the main person, Univ-OLS provides complementary results that  
 1164 highlight marginal association between the data and the target.



**Figure 17: Precision-recall curves on semi-simulated data with continuous response vector.** The results of the experiment described in Sec. 4.2 show that EnCluDL has the best performance in terms of precision-recall curve.



**Figure 18: Precision-recall curves on semi-simulated data with binary response vector.** The results of the experiment described in Sec. 4.2 with binary response show that Ada-SVR and EncluDL outperform alternatives in terms of feature recovery. These results are quite similar to the one presented in Fig. 6.

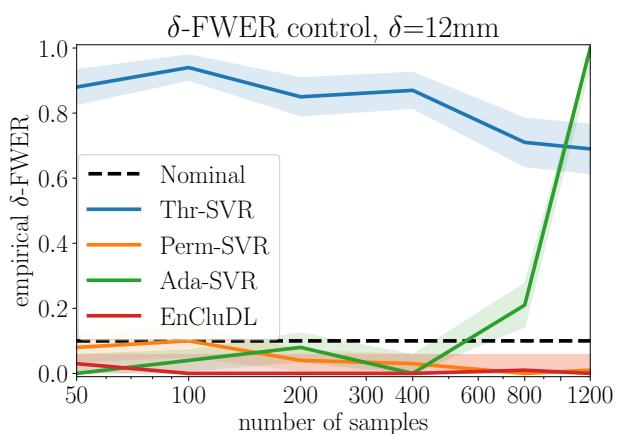
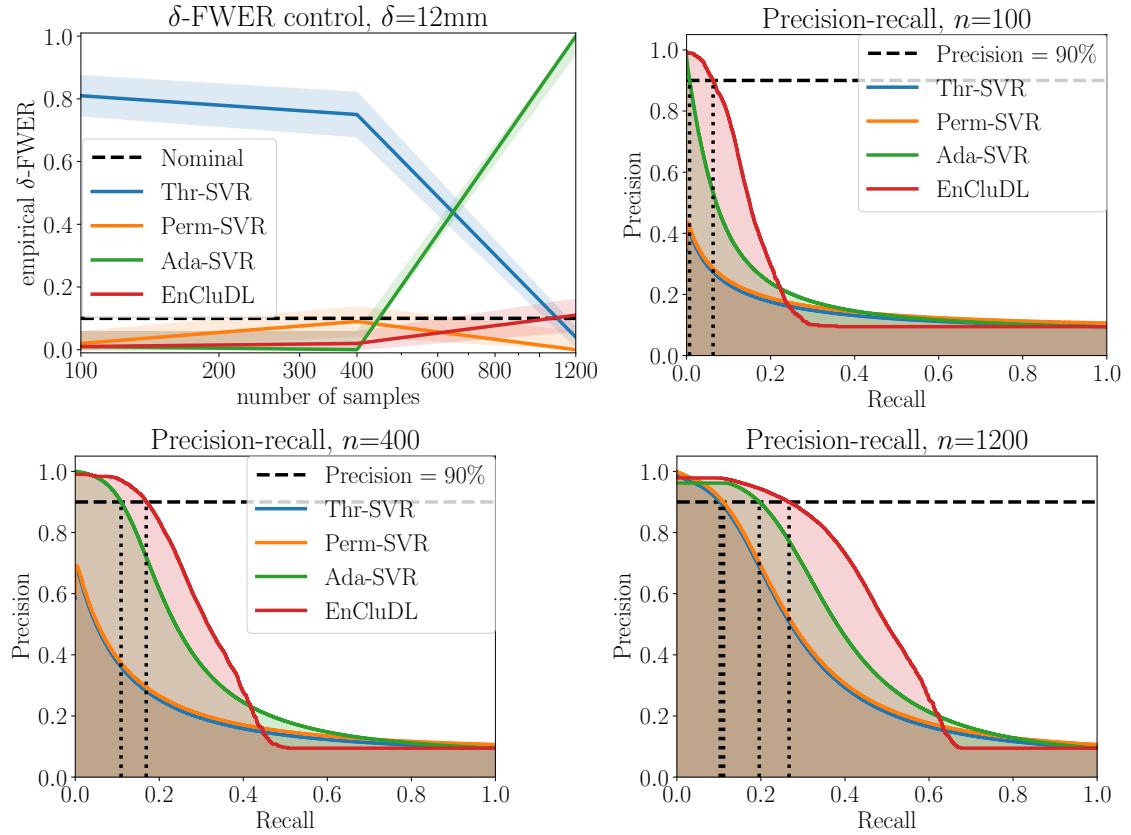
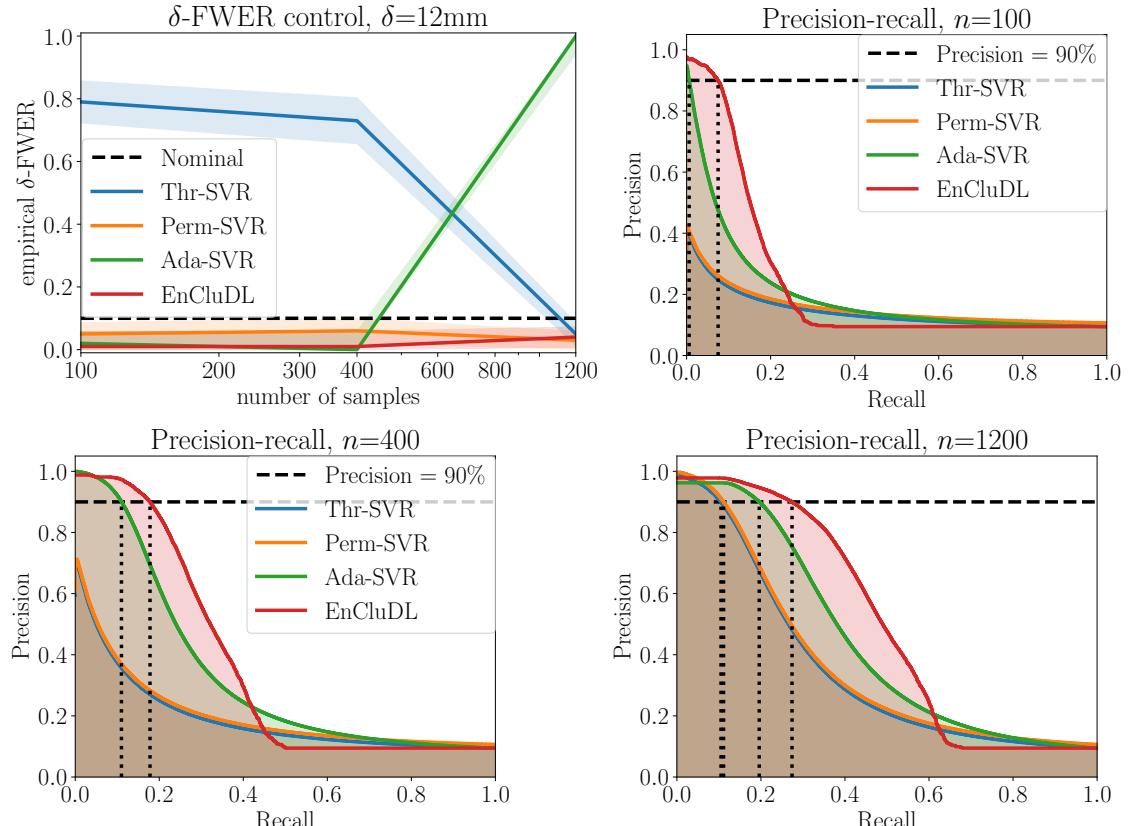


Figure 19:  **$\delta$ -FWER control on semi-simulated data with binary response vector.** The results of the experiment described in Sec. 4.2 with binary response show that only Perm-SVR and EnCluDL actually control the  $\delta$ -FWER.

45



**Figure 20:  $\delta$ -FWER control and precision-recall curves on semi-simulated data with continuous response vector with Laplace noise.** The results of the experiment described in Sec. 4.2 with Laplace noise are similar to the one presented in Fig. 6 for Gaussian noise.



**Figure 21:  $\delta$ -FWER control and precision-recall curves on semi-simulated data with continuous response vector with Student noise.** The results of the experiment described in Sec. 4.2 with Student (with 5 degrees of freedom) noise are similar to the one presented in Fig. 6 for Gaussian noise.

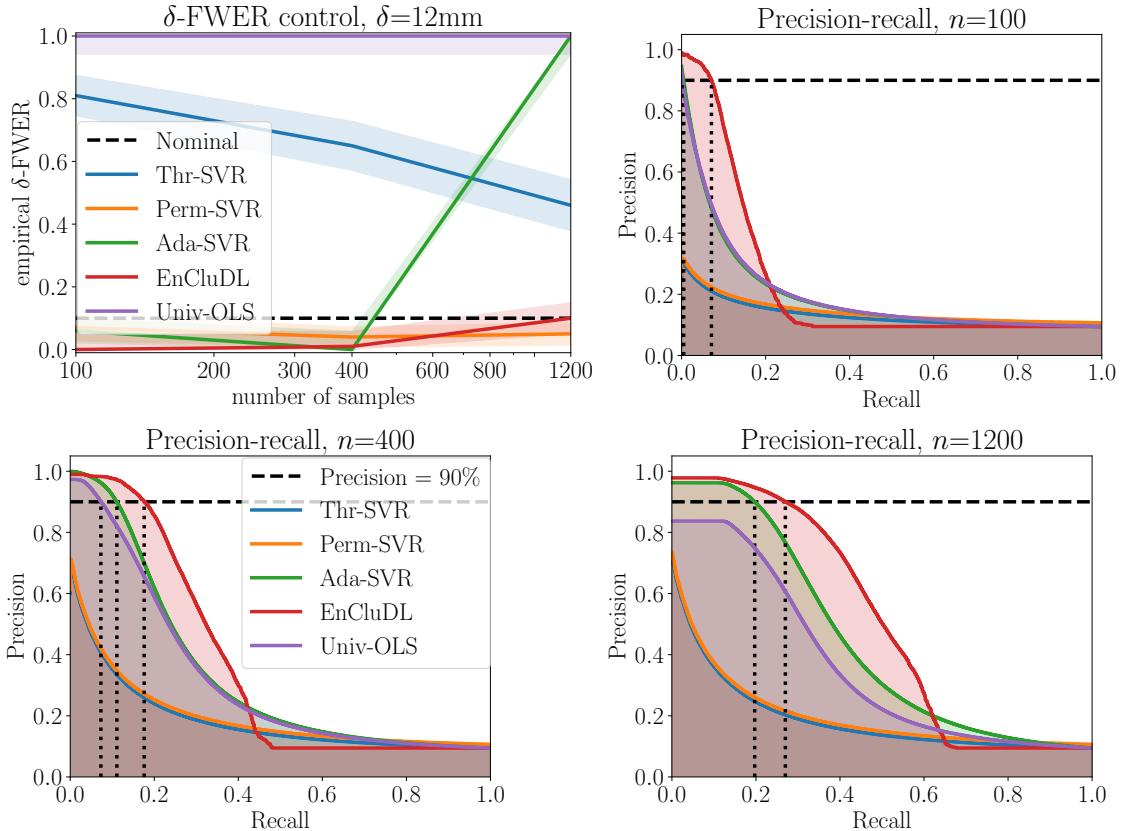
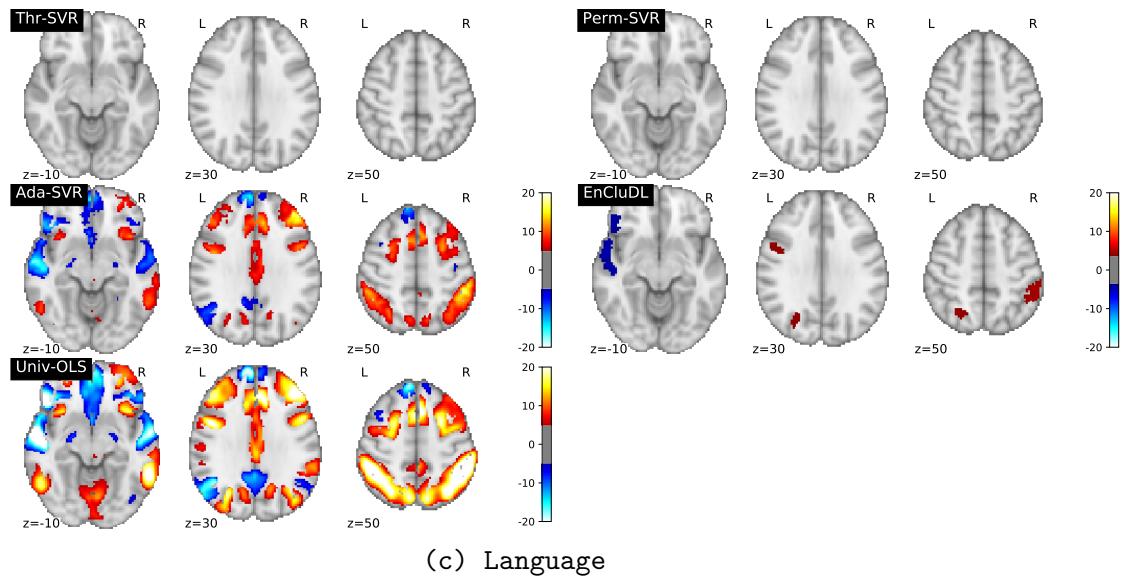
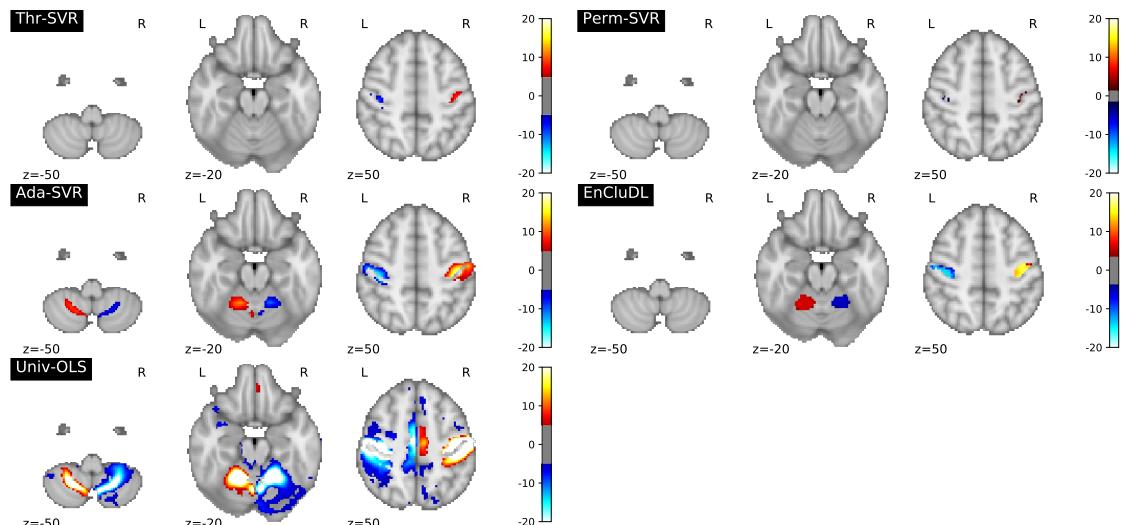


Figure 22:  **$\delta$ -FWER control and precision-recall curves on semi-simulated data with continuous response vector including a univariate method.** These results show that the FWER control guaranteed by Univ-OLS for univariate inference does not match the control granted by EncluDL in the conditional paradigm. This is due the fact that the null hypotheses being tested are not the same.

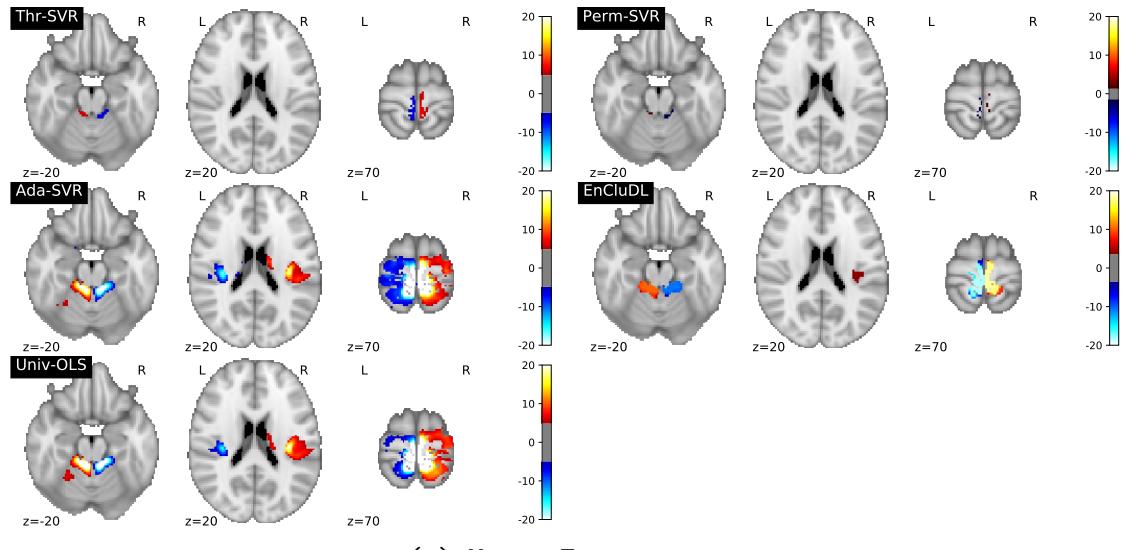


(c) Language

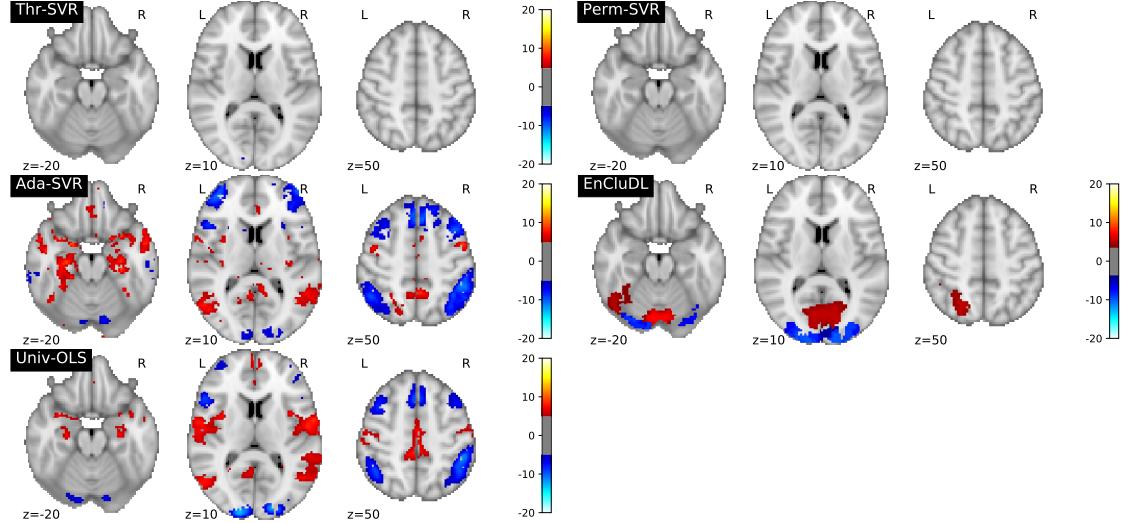


(d) Motor Hand

Figure 18a: cf. Fig. 23 for description.



(e) Motor Foot



(f) Relational

Figure 18b: cf. Fig. 23 for description.

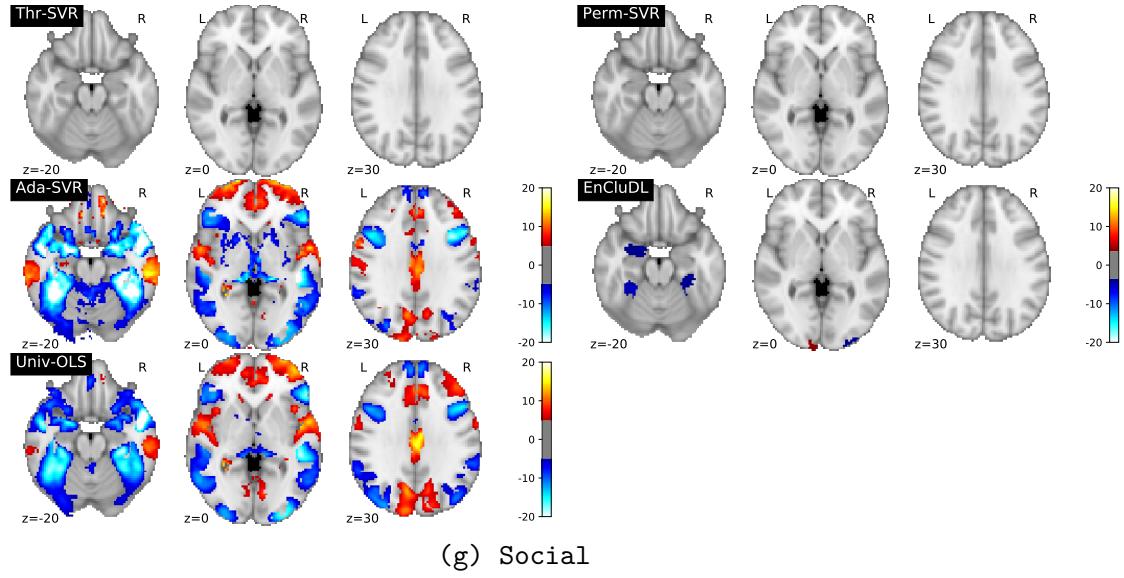


Figure 23: **Estimated predictive patterns on standard task fMRI dataset.** Here, we plot the results for five tasks of the experiment described in Sec. 4.6 thresholding the statistical maps such that the  $\delta$ -FWER stays lower than 10% for  $\delta = 12$  mm. Qualitatively, EnCluDL discovers the most plausible patterns, Ada-SVR often makes dubious discoveries, patterns are too wide and implausible, while the two other methods exhibit a very weak statistical power. As discussed before, Univ-OLS provides complementary results that display marginal associations between voxel signals and the target. The results of emotion and gambling tasks are available in Fig. 7.

1165 **7.12 Prediction performance**

1166 In this section, we give results on the prediction performance of the methods. In Fig. 24,  
 1167 we plot the results of the experiment described in Sec. 4.7. We notice that the classifica-  
 1168 tion error rate is almost the same for SVR (the weight map of Thr-SVR and Perm-SVR)  
 1169 and Ada-SVR, their prediction performance is slightly better than the one of EnCluDL.  
 1170 Hence, we do not recommend using EncluDL to achieve state-of-the art prediction ac-  
 curacy, but only for statistical inference purpose.

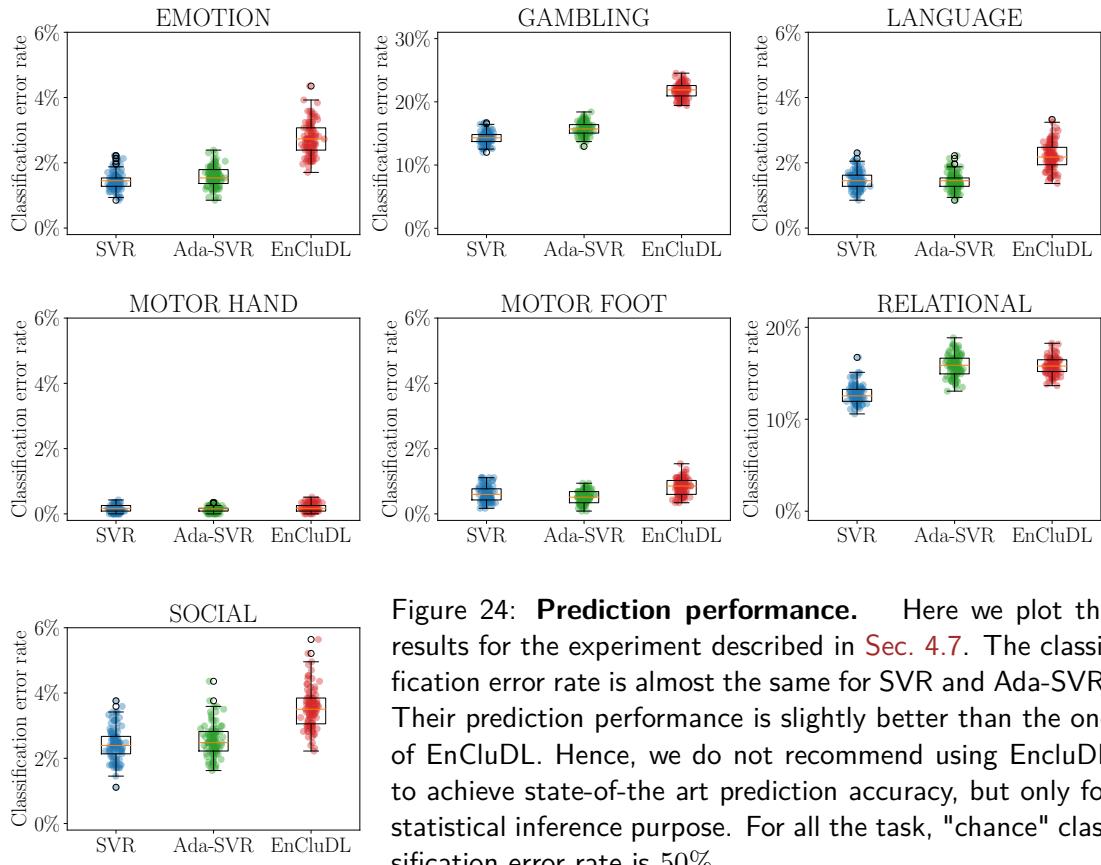


Figure 24: **Prediction performance.** Here we plot the results for the experiment described in Sec. 4.7. The classification error rate is almost the same for SVR and Ada-SVR. Their prediction performance is slightly better than the one of EnCluDL. Hence, we do not recommend using EncluDL to achieve state-of-the art prediction accuracy, but only for statistical inference purpose. For all the task, "chance" clas-  
 sification error rate is 50%.

1171