Spatially relaxed inference on high-dimensional linear models

Jérôme-Alexis Chevalier¹ · Tuan-Binh Nguyen² · Bertrand Thirion¹ · Joseph Salmon³

Received: 1 June 2021 / Accepted: 26 August 2022 / Published online: 30 September 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
We consider the inference problem for high-dimensional linear models, when covariates have an underlying spatial organization reflected in their correlation. A typical example of such a setting is high-resolution imaging, in which neighboring pixels are usually very similar. Accurate point and confidence intervals estimation is not possible in this context with many more covariates than samples, furthermore with high correlation between covariates. This calls for a reformulation of the statistical inference problem that takes into account the underlying spatial structure: if covariates are locally correlated, it is acceptable to detect them up to a given spatial uncertainty. We thus propose to rely on the $\delta$-FWER, that is, the probability of making a false discovery at a distance greater than $\delta$ from any true positive. With this target measure in mind, we study the properties of ensembled clustered inference algorithms which combine three techniques: spatially constrained clustering, statistical inference, and ensembling to aggregate several clustered inference solutions. We show that ensembled clustered inference algorithms control the $\delta$-FWER under standard assumptions, for $\delta$ equal to the largest cluster diameter. We complement the theoretical analysis with empirical results, demonstrating accurate $\delta$-FWER control and decent power achieved by such inference algorithms.

Keywords · Clustering · High-dimension · Linear model · Spatial tolerance · Statistical inference · Structured data · Support recovery

1 Introduction
High-dimensional regression corresponds to a setting where the number of covariates (or features) $p$ exceeds the number of samples $n$. It notably occurs when searching for conditional associations among some high-dimensional observations and some outcome of interest: the target. Typical examples of the high-dimensional setting include inference problems on high-resolution images, where one aims at pixel-or voxel-level analysis, e.g., in neuroimaging (Norman et al. 2006), astronomy (Richards et al. 2009), but also in other fields where covariates display a spatial structure, e.g., in genomics (Dehman et al. 2015). In all these examples, it actually turns out that not only $n < p$, but also $n \ll p$ and the covariates are spatially structured because of the physics of the problem or the measurements process. Because such high-dimensional data lead to high-variance results, probing statistical significance is important to give a level of confidence in the reported association. For this reason, the present analysis departs from traditional sparse modeling methods such as the Lasso (Tibshirani 1996) that simply aim at selecting a good set of predictive covariates without considering statistical significance. In this challenging context, we consider the multivariate linear model $y = X\beta^* + \varepsilon$, where the target $y \in \mathbb{R}^n$ is continuous, the random design matrix $X \in \mathbb{R}^{n \times p}$ is Gaussian, the parameter vector $\beta^* \in \mathbb{R}^p$ is sparse and spatially smooth, and the noise vector $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$ with standard deviation $\sigma_\varepsilon > 0$, is independent from $X$. The aim is to infer $\beta^*$, with statistical guarantees on the estimate, in particular regarding the support, i.e., the set of covariates with nonzero importance.

In high-dimensional settings, standard statistical inference methodology does not apply, but numerous methods have recently been proposed to recover the nonzero parameters of $\beta^*$ with statistical guarantees. Many methods rely on resampling: bootstrap procedures (Bach 2008; Chatterjee and Lahiri 2011), perturbation resampling-based proce-
dure, if not misleading. Contrarily to the screening/inference procedure, post-selection inference procedures generally merge the screening and inference steps into one and then use all the samples (Berk et al. 2013; Lockhart et al. 2014), resulting in potentially more powerful tests than sample splitting. Yet, these approaches do not scale well with large $p$. Another family of methods rely on debiasing procedures: the most prominent examples are corrected ridge (Bühlmann 2013) and desparsified Lasso (Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014) which is an active area of research (Javanmard and Montanari 2018; Bellec and Zhang 2022). Additionally, knockoff filters (Barber and Candès 2015; Candès et al. 2018) consist in creating noisy copies of the original variables, and checking which original variables are selected prior to the fake ones. Finally, a general framework for statistical inference in sparse high-dimensional models has been proposed recently (Ning and Liu 2017).

In practice, in the $n \ll p$ setting, we consider the previous methods are not well-adapted as they are often powerless or computationally intractable. In particular, the number of predictive parameters (i.e., the support size) denoted $s(\hat{\beta}^*)$ can be greater than the number of samples even in the sparse setting, where $s(\hat{\beta}^*) \ll p$. There is an underlying identifiability problem: in general, one cannot retrieve all predictive parameters, as highlighted, e.g., in Wainwright (2009). Beyond the fact that statistical inference is impossible when $p \gg n$, the problem is aggravated by the following three effects. First, as outlined above, dense covariate sampling leads to high values for $p$ and induces high correlation among covariates, further challenging the conditions for recovery, as shown in Wainwright (2009). Second, when testing multiple hypotheses, the correction cost is heavy; for example with Bonferroni correction (Dunn 1961), $p$ values are corrected by a factor $p$ when testing every covariate. At least for this reason, these inference methods lack power in the high-dimensional setting. Third, the above approaches are at least quadratic or cubic in the support size, hence become prohibitive whenever both $p$ and $n$ are large.

Nevertheless, in these settings, variables often reflect some underlying spatial structure, such as smoothness. For example, in medical imaging, an image has a 3D structure and a given voxel is highly correlated with neighboring voxels; in genomics, there exist blocks of single nucleotide polymorphisms (SNPs) that tend to be jointly predictive or not. Hence, $\hat{\beta}^*$ can in general be assumed to share the same structure: among several highly correlated covariates, asserting that only one is important to predict the target seems meaningless, if not misleading.

A computationally attractive solution that alleviates high dimensionality is to group correlated neighboring covariates. This step can be understood as a design compression: it produces a closely related, yet reduced version of the original problem (see e.g., Varoquaux et al. 2012). Inference combined with a fixed clustering has been proposed by Bühlmann et al. (2013) and can overcome the dimensionality issue, yet this study does not provide procedures that derive cluster-wise confidence intervals or $p$ values. Moreover, in most cases groups (or clusters) are not pre-determined nor easily identifiable from data, and their estimation simply represents a local optimum among a huge, non-convex space of solutions. It is thus problematic to base inference upon such an arbitrary data representation. Inspired by this dimension reduction approach, we have proposed (Chevalier et al. 2018) the ensemble of clustered desparsified Lasso (EnCluDL) procedure that exhibits strong empirical performances (Chevalier et al. 2021) in terms of support recovery even when $p \gg n$. EnCluDL is an ensemble cluster-based inference algorithm, i.e., it combines a spatially constrained clustering procedure that reduces the problem dimension, an inference procedure that performs statistical inference at the cluster level, and an ensembling method that aggregates several cluster-level solutions. Concerning the inference step, the desparsified Lasso (Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014) was preferred over other high-dimensional statistical inference procedures based on the comparative study of Dezeure et al. (2015) and on the research activity around it Javanmard and Montanari (2018), Bellec and Zhang (2022); however, it is possible to use another inference procedure to compute $p$ values. By contrast, we did not consider the popular knockoff procedure (Barber and Candès 2015; Candès et al. 2018) since it does not produce $p$ values and does not control the family-wise error rate (FWER). Nevertheless, the algorithm of Janson and Su (2016), an extension of the knockoffs, leverages on the knockoff procedure to estimate a support while ensuring an FWER-type control: it does not control the standard FWER but another relaxed version of the FWER called $k$-FWER. As it is a relevant alternative to ensemble clustered inference algorithms, we have included this algorithm in the empirical comparison. In Nguyen et al. (2020), a variant of the knockoffs is proposed to control the FWER, but it does not handle large-$p$ problems. Another extension that produces $p$ value, called conditional randomization test, has been presented in Candès et al. (2018), but its computational cost is prohibitive. Additionally, Meinshausen (2015) provides “group-bound” confidence intervals, corresponding to confidence intervals on the $\ell_1$-norm of several parameters, without making further assumptions on the design matrix. However, this method is known to be conservative in practice (Mitra and Zhang 2016; Javanmard and Montanari 2018). Finally, hierarchical testing (Blanchard and Geman 2005; Meinshausen 2008) also lever-
ages this clustering/inference combination but in a different way. Their approach consists in performing significance tests along the tree of a hierarchical clustering algorithm starting from the root node and descending subsequently into children of rejected nodes. This procedure has the drawback of being constrained by the clustering tree, which is often not known, thus replaced by some noisy estimate.

Producing a cluster-wise inference is not completely satisfactory as it relies on an arbitrary clustering choice. Instead, we look for methods that derive covariate-wise statistics enabling support identification with a spatially relaxed false detection control.

In Sect. 2, we describe the model and the key data assumptions. In Sect. 3, we present a generalization of the FWER, called $\delta$-FWER that takes into account a spatial tolerance of magnitude $\delta$ for the false discoveries. Then, in Sect. 4, we prove that ensembled clustered inference algorithms control the $\delta$-FWER under reasonable assumptions for a given tolerance parameter $\delta$. In Sect. 5, we apply the ensembled clustered inference scheme to the desparsified Lasso leading to the EnCluDL algorithm and conduct two simulations: we show that EnCluDL exhibits a good statistical power in comparison with alternative procedures, and we verify that it displays the expected $\delta$-FWER control. In Sect. 6, we describe an application of the proposed methods to neuroimaging and provide comparative results on a large-scale dataset. In Sect. 7, we discuss the key features of the ensembled clustered inference algorithms, such as the clustering step.

2 Model and data assumptions

2.1 Notation

Throughout the remainder of this article, for any $p \in \mathbb{N}^*$, we write $[p]$ for the set $\{1, \ldots, p\}$. For a vector $\beta$, $\beta_j$ refers to its $j$-th coordinate. For a matrix $X$, $X_{i,j}$ refers to the $i$-th row and $X_{i,j}$ to the $j$-th column, and $X_{i,j}$ refers to the element in the $i$-th row and $j$-th column.

2.2 Generative models of high-dimensional data: random fields

In the setting that we consider, we assume that the covariates come with a natural representation in a discretized metric space, generally the discretized 2D or 3D Euclidean space. In such settings, discrete random fields are convenient to model the random variables representing the covariates. Indeed, denoting by $X = (X_{i,j})_{i \in [n], j \in [p]}$ the random design matrix, where $n$ is the number of samples and $p$ the number of covariates, the rows $(X_{i,j})_{i \in [n]}$ are sampled from a random field defined on a discrete domain.

2.3 Gaussian random design model and high-dimensional settings

We assume that the covariate samples are independent and identically distributed and follow a centered Gaussian distribution, i.e., for all $i \in [n]$, $X_{i,j} \sim \mathcal{N}(0,\Sigma)$ where $\Sigma$ is the covariance matrix of the covariates and $X_{i,j}$ is independent from $X_{j,i}$, for all $j \in [n]\backslash \{i\}$. Our aim is to derive confidence bounds or $p$ values on the coefficients of the parameter vector denoted by $\beta^*$, under the Gaussian linear model:

$$y = X\beta^* + \varepsilon,$$

where $y \in \mathbb{R}^n$ is the target, $X \in \mathbb{R}^{n \times p}$ is the (random) design matrix, $\beta^* \in \mathbb{R}^p$ is the vector or parameters, and $\varepsilon \sim \mathcal{N}(0,\sigma^2_nI_n)$ is the noise vector with standard deviation $\sigma_n > 0$. We make the assumption that $\varepsilon$ is independent of $X$.

2.4 Data structure

Since the covariates have a natural representation in a metric space, we assume that the spatial distances between covariates are known. With a slight abuse of notation, the distance between covariates $j$ and $k$ is denoted by $d(j,k)$ for $(j,k) \in [p] \times [p]$ and the correlation between covariates $j$ and $k$ is given by $\text{Cor}(X_{j,:},X_{k,:}) = \Sigma_{j,k}/\sqrt{\Sigma_{j,j}\Sigma_{k,k}}$. We now introduce a key structural assumption: two covariates at a spatial distance smaller than $\delta$ are positively correlated.

Assumption 2.1 The covariates verify the spatial homogeneity assumption with distance parameter $\delta > 0$ if, for all $(j,k) \in [p] \times [p]$, $d(j,k) \leq \delta$ implies that $\Sigma_{j,k} \geq 0$.

Under model (1), each coordinate of the parameter vector $\beta^*$ links one covariate to the target. Then, $\beta^*$ has the same underlying organization as the covariates and is also called weight map in these settings. Defining its support as $S(\beta^*) = \{j \in [p]: \beta^*_j \neq 0\}$ and its cardinal as $s(\beta^*) = |S(\beta^*)|$, we assume that the true model is sparse, meaning that $\beta^*$ has a small number of nonzero entries, i.e., $s(\beta^*) \ll p$. The complement of $S(\beta^*)$ in $[p]$ is called the null region and is denoted by $N(\beta^*)$, i.e., $N(\beta^*) = \{j \in [p]: \beta^*_j = 0\}$. Additionally to the sparse assumption, we assume that $\beta^*$ is (spatially) smooth. To reflect sparsity and smoothness, we introduce another key assumption: weights associated with close enough covariates share the same sign, zero being both positive and negative.

Assumption 2.2 The weight vector $\beta^*$ verifies the sparse-smooth assumption with distance parameter $\delta > 0$ if, for all $(j,k) \in [p] \times [p]$, $d(j,k) \leq \delta$ implies that $\text{sign}(\beta^*_j) = \text{sign}(\beta^*_k)$.

Equivalently, the sparse-smooth assumption with parameter $\delta$ holds if the distance between the two closest weights...
of opposite sign is larger than δ. In Fig. 2a, we give an example of a weight map verifying the sparse-smooth assumption with δ = 2.

3 Statistical control with spatial tolerance

Under the spatial assumption we have discussed, discoveries that are closer than δ from the true support are not considered as false discoveries: inference at a resolution finer than δ might be unrealistic. This means that δ can be interpreted as a tolerance parameter on the (spatial) support we aim at recovering. Then, we introduce a new metric closely related to the FWER that takes into account spatial tolerance, and we call it δ-family-wise error rate (δ-FWER). A similar extension of the false discovery rate (FDR) has been introduced by Cheng et al. (2020), Nguyen et al. (2019), Gimenez and Zou (2019), but, to the best of our knowledge, this has not been considered yet for the FWER. In the following, we consider a general estimator $\hat{\beta}$ that comes with $p$ values, testing the nullity of the corresponding parameters, denoted by $\hat{p} = (\hat{p}_j)_{j \in [p]}$. Also, we denote by $S(\hat{\beta}) \subset [p]$ a general estimate of the support $S(\beta^*)$ derived from the estimator $\hat{\beta}$.

Definition 3.1 (δ-null hypothesis) For all $j \in [p]$, the δ-null hypothesis for the $j$-th covariate, $H_0^\delta(j)$, states that all other covariates at distance less than δ have a zero weight in the true model (1); the alternative hypothesis is denoted $H_1^\delta(j)$:

$H_0^\delta(j)$: “for all $k \in [p]$ such that $d(j, k) \leq \delta$, $\beta_k^* = 0$”,

$H_1^\delta(j)$: “there exists $k \in [p]$ such that $d(j, k) \leq \delta$ and $\beta_k^* \neq 0$”.

Thus, we say that a δ-Type 1 error is made if a null covariate $j \in [p]$ is selected, i.e., $j \in S(\hat{\beta})$, while $H_0^\delta(j)$ holds true. Taking δ = 0 recovers the usual null-hypothesis $H_0(j)$: “$\beta_j^* = 0$” and usual Type 1 error.

Definition 3.2 (Control of the δ-Type 1 error) The $p$ value related to the $j$-th covariate denoted by $\hat{p}_j$ controls the δ-Type 1 error if, under $H_0^\delta(j)$, for all $\alpha \in (0, 1)$, we have:

$$P(\hat{p}_j \leq \alpha) \leq \alpha,$$

where $P$ is the probability distribution with respect to the random dataset of observations $(X, y)$.

Definition 3.3 (δ-null region) The set of indexes of covariates verifying the δ-null hypothesis is called the δ-null region and is denoted by $N^\delta(\beta^*)$ (or simply $N^\delta$):

$$N^\delta(\beta^*) = \{ j \in [p] : \text{ for all } k \in [p], d(j, k) \leq \delta \text{ implies that } \beta_k^* = 0 \}.$$

When δ = 0 the δ-null region is simply the null region: $N^0(\beta^*) = N(\beta^*)$. We also point out the nested property of δ-null regions with respect to δ: for $0 \leq \delta_1 \leq \delta_2$ we have $N^{\delta_2}(\beta^*) \subseteq N^{\delta_1}(\beta^*) \subseteq N(\beta^*)$ (see Fig. 2d for an example of δ-null region).

Definition 3.4 (Rejection region) Given a family of $p$ values $\hat{p} = (\hat{p}_j)_{j \in [p]}$ and a threshold $\alpha \in (0, 1)$, the rejection region, $R_\alpha(\hat{p})$, is the set of indexes having a $p$ value lower than $\alpha$:

$$R_\alpha(\hat{p}) = \{ j \in [p] : \hat{p}_j \leq \alpha \}.$$

Definition 3.5 (δ-Type 1 error region) Given a family of $p$ values $\hat{p} = (\hat{p}_j)_{j \in [p]}$ and a threshold $\alpha \in (0, 1)$, the δ-Type 1 error region at level $\alpha$ is $\delta_\alpha^\delta$, the set of indexes belonging both to the δ-null region and to the rejection region at level $\alpha$.

We also refer to this region as the erroneous rejection region at level $\alpha$ with tolerance $\delta$:

$$\delta_\alpha^\delta(\hat{p}) = N^\delta \cap R_\alpha(\hat{p}).$$

When δ = 0, the δ-Type 1 error region recovers the Type 1 error region which is denoted by $\delta_\alpha^0(\hat{p})$. Again, one can verify a nested property: for $0 \leq \delta_1 \leq \delta_2$ we have $\delta_{\alpha}^{\delta_2}(\hat{p}) \subseteq \delta_{\alpha}^{\delta_1}(\hat{p})$.

Definition 3.6 (δ-family wise error rate) Given a family of $p$ values $\hat{p} = (\hat{p}_j)_{j \in [p]}$ and a threshold $\alpha \in (0, 1)$, the δ-FWER at level $\alpha$ with respect to the family $\hat{p}$, denoted $\delta_{\text{FWER}}(\hat{p})$, is the probability that the δ-Type 1 error region at level $\alpha$ is not empty:

$$\delta_{\text{FWER}}(\hat{p}) = P(|\delta_\alpha^\delta(\hat{p})| \geq 1) = P(\min_{j \in N^\delta} \hat{p}_j \leq \alpha).$$

Definition 3.7 (δ-FWER control) We say that the family of $p$ values $\hat{p} = (\hat{p}_j)_{j \in [p]}$ controls the δ-FWER if, for all $\alpha \in (0, 1)$:

$$\delta_{\text{FWER}}(\hat{p}) \leq \alpha.$$

When δ = 0 the δ-FWER is the usual FWER. Additionally, for $0 \leq \delta_1 \leq \delta_2$, one can verify that $\delta_{2\text{-FWER}}(\hat{p}) \leq \delta_{1\text{-FWER}}(\hat{p}) \leq \text{FWER}(\hat{p})$. Thus, δ-FWER control is a weaker property than usual FWER control.

4 δ-FWER control with clustered inference algorithms

4.1 Clustered inference algorithms

A clustered inference algorithm consists in partitioning the covariates into groups (or clusters) before applying a sta-
tistical inference procedure. In Algorithm 1 of Sect. 4.1, we describe a standard clustered inference algorithm that produces a (corrected) \( p \) value family on the parameters of the model (1). In this algorithm, in addition to the observations \((X, y)\), we take as input the transformation matrix \( A \in \mathbb{R}^{p \times C} \) which maps and averages covariates into \( C \) clusters. The statistical_inference function corresponds to a given statistical inference procedure that takes as inputs the clustered data \( Z \) and the target \( y \) and produces valid \( p \) values for every cluster. If \( C < n \), least squares are suitable, otherwise, procedures such as multi-sample split (Wasserman and Roeder 2009; Meinshausen et al. 2009), corrected ridge (Bühlmann 2013) or desparsiﬁed Lasso (Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014) might be relevant whenever their assumptions are veriﬁed. Then, the computed \( p \) values are corrected for multiple testing by multiplying by a factor \( C \). Finally, covariate-wise \( p \) values are inherited from the corresponding cluster-wise \( p \) values.

**Algorithm 1** Clustering inference

```plaintext```
input : \( X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n, A \in \mathbb{R}^{p \times C} \)

\( Z = XA \); // compressed design matrix
\( \hat{p}^0 = \text{statistical}_\text{inference}(Z, y) \); // uncorrected cluster-wise \( p \)-values
\( \hat{q}^0 = C \times \hat{p}^0 \); // corrected cluster-wise \( p \)-values

for \( j = 1, \ldots, p \) do
    \( \hat{q}_j = \hat{q}^0_j \) if \( j \) in cluster \( c \); // corrected covariate-wise \( p \)-values

return \( \hat{q} = (\hat{q}_j)_{j \in [p]} \); // family of corrected covariate-wise \( p \)-values
```

**Algorithm 2** Ensembled clustered inference

```plaintext```
input : \( X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n \)

param: \( C, B \)

for \( b = 1, \ldots, B \) do
    \( X^{(b)} = \text{sampling}(X) \); // sampling rows of \( X \)
    \( A^{(b)} = \text{clustering}(C, X^{(b)}) \); // transformation matrix
    \( \hat{q}^{(b)} = \text{clustered}_\text{inference}(X, y, A^{(b)}) \); // families of corr. covariate-wise \( p \)-val.

for \( j = 1, \ldots, p \) do
    \( \hat{q}_j = \text{ensembling}((\hat{q}^{(b)}_j, b \in [B])) \); // aggregated corrected covariate-wise \( p \)-values

return \( \hat{q} = (\hat{q}_j)_{j \in [p]} \); // family of aggregated corrected covariate-wise \( p \)-values
```

Ensembled clustered inference algorithms correspond to the ensembling of several clustered inference solutions for different choice of clusterings using the \( p \) value aggregation proposed by Meinshausen et al. (2009). In Algorithm 2 of Sect. 4.1, we give a standard ensembled clustered inference algorithm that produces a (corrected) \( p \) value family on the parameters of the model (1). In this algorithm, the sampling function corresponds to a subsampling of the data, i.e., a subsampling of the rows of \( X \). The clustering function derives a choice of clustering in \( C \) clusters, it produces a transformation matrix \( A^{(b)} \in \mathbb{R}^{p \times C} \) that should vary for each bootstrap \( b \in [B] \) since the subsampled data \( X^{(b)} \) varies. Once the clustering inference steps are completed, the ensembling function aggregates the \( B \) (corrected) \( p \) value families into a single one.

Figure 1 can help the reader to better understand the organization of the next sections, aiming eventually at establishing the \( \delta \)-FWER control property of the clustered inference and ensembled clustered inference algorithms.

### 4.2 Compressed representation

The motivation for using groups of covariates that are spatially concentrated is to reduce the dimension while preserving large-scale data structure. The number of groups is denoted by \( C < p \) and, for \( c \in [C] \), we denote by \( G_c \) the \( c \)-th group. The collection of all the groups is denoted by \( \mathcal{G} = \{G_1, G_2, \ldots, G_C\} \) and forms a partition of \( [p] \). Every group representative variable is defined by the average of the covariates it contains. Then, denoting by \( Z \in \mathbb{R}^{n \times C} \) the compressed random design matrix that contains the group representative variables in columns and, without loss of generality, assuming a suitable ordering of the columns of \( X \), dimension reduction can be written:

\[
Z = XA ,
\]

where \( A \in \mathbb{R}^{p \times C} \) is the transformation matrix defined by:

\[
A = \begin{bmatrix}
\alpha_1 & \cdots & 0 & \cdots & 0 & \cdots \\
0 & \cdots & \alpha_2 & \cdots & 0 & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & \cdots & \alpha_C & \cdots \\
\end{bmatrix} ,
\]

where \( \alpha_c = 1/|G_c| \) for all \( c \in [C] \). More clearly, for all \( c \in [C] \), \( A_{j,c} = \alpha_c \) if \( j \in G_c \) and \( A_{j,c} = 0 \) otherwise. Consequently, the distribution of the \( i \)-th row of \( Z \) is given by \( Z_{ic} \sim \mathcal{N}_C(0, \Sigma) \), where \( \Sigma = A \Sigma A^T \). The correlation between the groups \( c \in [C] \) and \( c' \in [C] \) is given by \( \text{Cor}(Z_{ic}, Z_{ic'}) = \Sigma_{cc'}/\sqrt{\Sigma_{cc'} \Sigma_{c'c'}} \). As mentioned in Bühlmann et al. (2013), because of the Gaussian assumption in (1), we have the following compressed representation:

\[
y = Z\theta^* + \eta ,
\]
where \( \theta^* \in \mathbb{R}^C, \eta \sim \mathcal{N}(0, \sigma^2 \eta_n), \sigma_\eta \geq \sigma_\varepsilon > 0 \) and \( \eta \) is independent of \( Z \).

**Remark 4.1** The \((n, C)\) regime is obviously milder than the original \((n, p)\) regime. This is not the unique desirable effect produced by clustering to improve the conditions for statistical inference.

Indeed, clustering-based design compression also generally improves the conditioning of the problem.

Assumptions needed for valid statistical inference are thus more likely to be met.

For more details about this conditioning enhancement, the reader may refer to Bühlmann et al. (2013).

### 4.3 Properties of the compressed model weights

We now give a property of the weights of the compressed model, which is a consequence of (Proposition 4.3 Bühlmann et al. 2013).

**Proposition 4.1** Considering the Gaussian linear model in (1) and assuming:

(i) for all \( c \in [C] \), for all \((j, k) \in (G_c)^2\), \( \Sigma_{j,k} \geq 0 \),

(ii) for all \( c \in [C] \), for all \( c' \in [C] \setminus \{c\} \), \( \Upsilon_{c,c'} = 0 \),

(iii) for all \( c \in [C] \), \( \{\beta^*_j \geq 0 \text{ for all } j \in G_c\} \) or \( \{\beta^*_j \leq 0 \text{ for all } j \in G_c\} \),

then, in the compressed representation (3), for \( c \in [C] \), \( \theta^*_c \neq 0 \) if and only if there exists \( j \in G_c \) such that \( \beta^*_j \neq 0 \). If such an index \( j \) exists, then \( \text{sign}(\theta^*_c) = \text{sign}(\beta^*_j) \).

**Proof** See Supplement E.1.

Assumption (i) states that the covariates in a group are all positively correlated. Let us define the group diameter (or cluster diameter) of \( G_c \) by the distance that separates its two most distant covariates, i.e., \( \text{Diam}(G_c) = \max \{d(j, k) : (j, k) \in (G_c)^2\} \) and the clustering diameter of \( G \) by the largest group diameter, i.e., \( \text{Diam}(G) = \max \{\text{Diam}(G_c) : c \in [C]\} \).

In Fig. 2b, we propose a clustering of the initial weight map in Fig. 2a for which the clustering diameter is equal to 2 for the \( \ell_1 \) distance. Assumption (i) notably holds when \( \text{Diam}(G) \leq \delta \) under the spatial homogeneity assumption (Assumption 2.1) with parameter \( \delta \). Assumption (ii) assumes independence of the groups. A sufficient condition is when the covariates covariance matrix \( \Sigma \) is block diagonal, with blocks coinciding with the group structure; i.e., assumption (ii) holds when covariates of different groups are independent. In practice, this assumption may be unmet, and we relax it in Supplement B. Assumption (iii) states that all the weights in a group share the same sign. This is notably the case when the clustering diameter is smaller than \( \delta \) and the weight map satisfies the sparse-smooth assumption (Assumption 2.2) with parameter \( \delta \). For instance, a clustering-based compressed representation of the weight map in Fig. 2a is given in Fig. 2c.

### 4.4 Statistical inference on the compressed model

To perform the statistical inference on the compressed problem (3), one might consider any statistical inference procedure that produces cluster-wise \( p \) values \( \hat{P}^G = (\hat{P}^G_c)_{c \in [C]} \), given a choice of clustering \( G \), that control the Type I error. More precisely, for any \( c \in [C] \), under \( H_0(G_c) \), i.e., the null hypothesis which states that \( \theta^*_c \) is equal to zero in the compressed model, we assume that the \( p \) value associated with the \( c \)-th cluster verifies:

\[
\mathbb{P}(\hat{P}^G_c \leq \alpha) \leq \alpha . \tag{4}
\]

To correct for multiple comparisons, we consider Bonferroni correction (Dunn 1961) which is a conservative procedure but has the advantage of being valid without any additional assumptions. Furthermore, here the correction factor is only equal to the number of groups, not the number of covariates. Then, the family of corrected cluster-wise \( p \) values \( \hat{q}^G = (\hat{q}^G_c)_{c \in [C]} \) is defined by:

\[
\hat{q}^G_c = \min \{1, C \times \hat{P}^G_c\} . \tag{5}
\]

Let us denote by \( N_G(\theta^*) \) (or simply \( N_G \)) the null region in the compressed problem for a given choice of clustering \( G \), i.e., \( N_G(\theta^*) = \{c \in [C] : \theta^*_c = 0\} \). Then, for all \( \alpha \in (0, 1) \):

\[
\text{FWER}_G(\hat{q}^G) = \mathbb{P}(\min_{c \in N^G} \hat{q}^G_c \leq \alpha) \leq \alpha . \tag{6}
\]

This means that the cluster-wise \( p \) value family \( \hat{q}^G \) controls FWER.
4.5 De-grouping

Given the families of cluster-wise \( p \) values \( \hat{p}_G \) and corrected \( p \) values \( \hat{q}_G \) as defined in (9) and (5), our next aim is to derive families of \( p \) values and corrected p-values related to the covariates of the original problem. To construct these families, we simply set the (corrected) \( p \) value of the \( j \)-th covariate to be equal to the (corrected) \( p \) value of its corresponding group:

\[
\text{for all } j \in [p], \quad \hat{p}_j = \sum_{c \in [C]} \mathbb{1}_{\{j \in G_c\}} \hat{p}_c^G ,
\]

\[
\text{for all } j \in [p], \quad \hat{q}_j = \sum_{c \in [C]} \mathbb{1}_{\{j \in G_c\}} \hat{q}_c^G .
\]

Proposition 4.2 Under the assumptions of Proposition 4.1 and assuming that the clustering diameter is smaller than \( \delta \), then:
The family 
lies 
and ensembled clustered inference algorithms control the 
δ
4.7
Let us consider 
B
4.6 Ensembling

The previous de-grouping properties is shown in Fig. 2d. Roughly, since all the clusters that intersect the δ-null region have low p value with low probability, one can conclude that all the covariates of the δ-null region also have low p value with low probability.

4.6 Ensembling

Let us consider B families of corrected p values that control the δ-FWER. For any b ∈ [B], we denote by 
q(b)
the b-th family of corrected p values. Then, we show that the ensembling method proposed in Meinshausen et al. (2009) yields a family that also enforces δ-FWER control.

Proposition 4.3 Assume that for b ∈ [B], the p value families 
q(b)
control the δ-FWER. Then, for any γ ∈ (0, 1), the ensembled p value family 
q(γ)
is defined by:

for all j ∈ [p], 
q_j(γ)

= \min \left\{ 1, \gamma\text{-quantile} \left( \frac{\hat{q}_j(b)}{\gamma} : b \in [B] \right) \right\},

controls the δ-FWER.

Proof See Supplement E.3. □

Remark 4.2 In practice, given that we advise to take B = 25 (see Sect. 5.2), we recommend to take γ ∈ [0.2, 0.5]. Indeed, γ ≥ 0.2 ensures that a covariate is assigned a low ensembled p value only if at least 5 of its bootstrap p values are low. Then, the ensembling effect has good chances to be effective. Additionally, we do not recommend to take γ > 0.5 since the ensembling procedure becomes quite conservative and the resulting inference loses sensitivity.

4.7 δ-FWER control

We can now state our main result: the clustered inference and ensembled clustered inference algorithms control the δ-FWER.

Theorem 4.1 Assume the model given in (1) and that the data structure assumptions, Assumptions 2.1 and 2.2, are satisfied for a distance parameter larger than δ. Assume that all the clusterings considered have a diameter smaller than δ. Assume that the uncorrelated cluster assumption, i.e., assumption (ii) of Proposition 4.1, is verified for each clustering and further assume that the statistical inference performed on the compressed model (3) is valid, i.e., (4) holds. Then, the p value family obtained from the clustered inference algorithm controls the δ-FWER. Additionally, the p value family derived by the ensembled clustered inference algorithm controls the δ-FWER.

Proof See Supplement E.4. □

Remark 4.3 When the Type 1 error control offered by the statistical inference procedure is only asymptotic, the result stated by Proposition 4.1 remains true asymptotically.

This is notably the case when using desparsified Lasso: under the assumptions of Proposition 4.1 and the assumptions specific to desparsified Lasso (cf. Supplement A), ensemble of clustered desparsified Lasso (EnCluDL) controls the δ-FWER asymptotically.

5 Numerical simulations

5.1 CluDL and EnCluDL

To test the (ensembled) clustered inference algorithms, we use the desparsified Lasso for statistical inference (Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014) leading to the clustered desparsified Lasso (CluDL) and the ensemble of clustered desparsified Lasso (EnCluDL) algorithms that were first presented in Chevalier et al. (2018).

In Supplement A, we detail the assumptions and refinements that occur when choosing the desparsified Lasso for the statistical inference step. An important aspect is the fact that all the results are asymptotic. In Supplement C, we present a diagram illustrating the mechanism of EnCluDL and analyse its numerical complexity.

5.2 2D simulation with regular support

We run a series of simulations on 2D (image) data in order to give empirical evidence of the theoretical properties of CluDL and EnCluDL and compare their recovery properties with two other procedures. For an easier visualization of the results, we consider one central scenario, whose parameters are written in bold in the following of this section, with several variations, changing only one parameter at a time.

In all these simulations, the feature space considered is a 2D square with edge length \( H = 40 \) leading to \( p = H^2 = \)
1600 covariates, with a sample size \( n \in \{50, 100, 200, 400\} \). To construct \( \beta^* \), we define a 2D weight map \( \hat{\beta}^* \) with four active regions, as shown in Fig. 3, and then flatten \( \hat{\beta}^* \) to a vector \( \beta^* \) of size \( p \). Each active region is a square of width \( h \in \{2, 4, 6, 8\} \), leading to a size of support of 1%, 4%, 9% or 16%. To construct the design matrix, we first build a 2D data matrix \( \tilde{X} \) by drawing \( p \) random normal vectors of size \( n \) that are spatially smoothed with a 2D Gaussian filter to create a correlation structure related to the covariates’ spatial organization. The same flattening process as before is used to get the design matrix \( X \in \mathbb{R}^{n \times p} \). The spatial smoothing parameter is adjusted to achieve an average correlation between two adjacent covariates (local correlation) of \( \rho \in \{0.5, 0.75, 0.9, 0.95\} \). We also set the noise standard deviation \( \sigma_n \in \{1, 2, 3, 4\} \), which corresponds to a signal-to-noise ratio (SNR) \( \text{SNR}_Y = \|X\beta^*\|_2/\|\varepsilon\|_2 \). For each scenario, we run 100 simulations to derive meaningful statistics. A Python implementation of the simulations and procedures presented in this paper is available on https://github.com/ja-che/hidimstat. Regarding the clustering step in CluDL and EnCluDL, we used a spatially constrained agglomerative clustering algorithm with Ward criterion. This algorithm is popular in many applications (Varoquaux et al. 2012; Dehman et al. 2015), as it tends to create compact, balanced clusters. Since the optimal number of clusters \( C \) is unknown a priori, we have tested several values \( C \in \{50; 800\} \). A smaller \( C \) generally improves recovery, but entails a higher spatial tolerance. Following theoretical considerations, we compute the largest cluster diameter for every value of \( C \) and set \( \delta \) to this value. We obtained the couples \((C, \delta) \in \{(50, 11), (100, 8), (200, 6), (400, 4), (800, 2)\}\). In Fig. 3, we represent the tolerance region for \( \delta = 6 \). It corresponds to the null-weight covariates that surround the square-shaped active regions. Concerning EnCluDL, we took a number of bootstraps \( B \) equal to 25, as we observed that it was sufficient to benefit from most of the effect of randomized clustering.

5.3 Alternative methods

We compare the recovering properties of CluDL and EnCluDL with two other procedures: desparsified Lasso and knockoffs. Contrarily to CluDL and EnCluDL, the latter two do not include a compression step. The version of the desparsified Lasso we have tested is the one presented in van de Geer et al. (2014) that outputs \( p \) values. Using Bonferroni correction, it controls the classical FWER at any desired rate. The original version of knockoffs (Barber and Candès 2015; Candès et al. 2018) only controls the false discovery rate (FDR) which is a weaker control than the classical FWER. Yet Janson and Su (2016) modifies the covariate selection process leading to a procedure that controls the k-FWER, i.e., the probability of making at least \( k \) false discoveries. We tested this extension of knockoffs. Note that the choice of \( k \) depends on the rate at which we want to control the k-FWER. More precisely, if we want a k-FWER control at 10%, we need to tolerate \( k = 4 \) at least, otherwise the estimated support would always be empty.

Since k-FWER and \( \delta \)-FWER controls are both weaker than the usual FWER control whenever \( k > 1 \) and \( \delta > 0 \), one can expect desparsified Lasso to be less powerful than knockoffs, CluDL and EnCluDL. Besides, there is no relation between k-FWER and \( \delta \)-FWER controls when \( k > 1 \) and \( \delta > 0 \), hence it is not possible to establish which one is less prohibitive for support recovery. However, when data are spatially structured, \( \delta \)-FWER control might be more relevant since it controls the very undesirable far-from-support false discoveries.

5.4 Results of the simulation with square-shaped support

In Fig. 3, we plot the maps estimated by knockoffs, desparsified Lasso, CluDL and EnCluDL for \( C = 200 \), when solving one instance of the central scenario simulation. Regarding knockoffs and desparsified Lasso solutions, we notice that the power is low and the methods select few covariates in each predictive region. The CluDL method is more powerful and recovers groups of covariates that correspond more closely to the true weights. However, the shape of the CluDL solution depends on the clustering choice. The EnCluDL solution seems even more powerful than CluDL and recovers groups of covariates that correspond almost perfectly to the true weights. Both CluDL and EnCluDL are only accurate up to the \( \delta = 6 \) pixels spatial tolerance, but EnCluDL fits the ground truth more tightly.

In Fig. 4, we focus on the central scenario to get more insight about the statistical properties of the methods and the influence of the \( C \) hyper-parameter for CluDL and EnCluDL. First, we observe that all methods reach the targeted control: desparsified Lasso controls the FWER, knockoffs control the k-FWER and, CluDL and EnCluDL control the \( \delta \)-FWER. Desparsified Lasso, Knockoffs and EnCluDL are conservative since their empirical FWER remains equal to 0%, while the nominal rate is 10%. CluDL is the only method which saturates the nominal rate, since it benefits from the clustering to improve the well-posedness of the estimation problem, and avoids the conservative aggregation step. Second, considering the true positive rates (TPR), we notice that the methods that do not integrate a compression step, i.e., knockoffs and desparsified Lasso, have a limited statistical power in the \( n \ll p \) regime. However, CluDL has decent power and EnCluDL improves over CluDL thanks to clustering randomization. Finally, CluDL and EnCluDL are reasonably flexible.
Fig. 3 True support and estimated supports for one instance of the central scenario of the simulation described in Sect. 5.2. Left The support in yellow is composed of four squared regions of width \( h = 4 \) pixels. The tolerance region (set of null-weight covariates close to the support) in green surrounds the support, its width is \( \delta = 6 \) pixels. The remaining covariates in blue form the \( \delta \)-null region (set of null-weight covariates far from the support). Others The yellow squares are the covariates selected by each method. Knockoffs selects few covariates when controlling the \( k \)-FWER at 10% for \( k = 4 \). Desparsified Lasso only retrieves 3 covariates when controlling the FWER at 10%. For \( C = 200 \), CluDL and EnCluDL have good power and control the \( \delta \)-FWER at 10% for \( \delta = 6 \). (Color figure online)

Fig. 4 Results for fixed simulation parameters corresponding to the central scenario simulation. The green line with circles correspond to knockoffs, the brown line with triangles is the desparsified Lasso, the purple squared line correspond to CluDL and the blue plain line is EnCluDL. Left Empirical FWER for desparsified Lasso, \( k \)-FWER for knockoffs and \( \delta \)-FWER for CluDL and EnCluDL. The 80% confidence intervals are obtained by Binomial approximation. Right Median true positive rate (TPR) for all the procedures, together with 80% confidence interval obtained by taking the first decile and last decile TPR. (Color figure online)

Fig. 5 True weights, true support and supports estimated by knockoffs, desparsified Lasso, CluDL and EnCluDL for \( C = 200 \) when solving the first seed of the simulation described in Sect. 5.5. As in Sect. 5.2, the power of knockoffs and desparsified Lasso is low since the methods select few covariates. Again, CluDL and, in a greater extent, EnCluDL improve upon the the other two.

5.5 2D simulation with variable support shape

In this section, we describe a second set of simulations to study the impact of the shape of the support. The setting is similar to the one of Sect. 5.2, except the support which is less regular, notably regarding its shape and weights distribution. We take the parameters of the central scenario \( \sigma_x = 2 \), \( \text{SNR}_y = 3.5 \) and \( \rho = 0.75 \), except for the number samples which is taken equal to \( n = 400 \) since this scenario is more challenging. In this simulation, \( h \) is no longer a simulation parameter since the support, displayed in Fig. 5, is fixed. Indeed, the support is derived from the compression of an image composed of the letters—“S,” “T,” “c” and “o”—with various thickness: “T” and “c” are wider, “o” is thinner. After compression, we discarded the weights smaller than 10% of the largest weight and normalized the weights such that the largest weight is equal to 1.0, i.e., weights of the support are between 0.1 and 1.0. The recovery of this support is more challenging for four main reasons: the shape of the support is less compact, the thickness of the support varies, the proximity between predictive regions can be high, the weights have different magnitude. Using this support, the aforementioned simulation parameters and the same generative model as in Sect. 5.2, we build 100 simulations to derive meaningful statistics.

5.6 Results of simulations with variable support shape

In Fig. 5, we plot true weights, true support and supports estimated by knockoffs, desparsified Lasso, CluDL and EnCluDL for \( C = 200 \) when solving the first seed of the simulation described in Sect. 5.5. As in Sect. 5.2, the power of knockoffs and desparsified Lasso is low since the methods select few covariates. Again, CluDL and, in a greater extent, EnCluDL improve upon the other two.

with respect to the choice of \( C \) since the TPR does not vary dramatically with \( C \).

We have also studied the influence of the simulation parameters by varying one parameter of the central scenario. The corresponding results are available in Supplement D.
As discussed in Sect. 5.5, the support of this simulation is more challenging to estimate than the support of Sect. 5.2. Notably, if we focus on the letter “o,” we notice that both CLuDL and EnCLuDL are struggling to recover this thin letter. On the opposite, EnCLuDL recovers quite accurately the wider letters “S,” “T,” and “c.” The fact that the letters “S” and “T” are close did not lead to a concatenation when using CLuDL or EnCLuDL. Nevertheless, since the tolerance region fills the gap between two letters, support junction may occur. Depending on the application, this might be a problem or not. To avoid such junction, one might want reduce the size of the clusters, reducing the size of the tolerance region and increasing the spatial accuracy of the solution.

However, as shown in Fig. 6, increase in C, hence in spatial accuracy, likely comes at the cost of power loss. In Fig. 6, we observe similar results as in Fig. 4: FWER-type controls holds and clustering-based inference methods have more power when the compression is greater. More insights regarding the choice of clustering and notably the choice of C are provided in Sect. 5.7.

5.7 Insights regarding the choice of clustering

Since clustered inference methods require to make one or several clustering, it is important to give some insights regarding the choice of clustering and eventually the choice of C, the number of clusters. However, in preamble, we insist on the fact that in general, there is not ideal choice of clustering since there is no true clustering of the data. This explains why we have proposed ensemble of clustered inference methods that mitigate the choice of clustering with randomization. Moreover, since the data are actually not structured in clusters, it is not straightforward to define what a good clustering is.

From Assumptions (i) and (ii) of Proposition 4.1 (see also assumptions (i), (ii.a) and (ii.b) of Proposition B.1), a given clustering is good if it results in low inter-cluster correlation and high intra-cluster feature correlation. We have also seen in Sects. 5.4 and 5.6 that a strong compression of the problem might be necessary for the sake of power. Moreover, from Sect. 4.2 and Supplement A, another desirable property of the clustering is that the variance of the noise in the compressed model defined in Eq. (3) is not much higher than the variance of the noise in the original model, i.e., the compression is accurate. This is the case if the assumptions of Proposition 4.1 are verified. Overall, a good criterion to decide whether or not a given clustering is admissible is to check if assumptions (i) and (ii) of Proposition 4.1 are (nearly) verified.

In the simulations of Sect. 5 and in the real data experiments of Sect. 6, we have used a spatially constrained agglomerative clustering with Ward criterion since it is a computationally attractive technique that groups correlated covariates. It has been shown to have good performance in neuroimaging experiments in Varoquaux et al. (2012). Alternatively, other fast agglomerative clustering techniques such as Hoyos-Idrobo et al. (2018) reduce computation time when the data have very high dimensionality, e.g., when $p \geq 10^5$.

Given a choice of clustering algorithm and an admissible range for C—with respect to assumptions (i) and (ii) of Proposition 4.1—we recommend to take the smallest C that leads to a clustering diameter smaller than the desired tolerance parameter $\delta$. Indeed, decreasing C generally increases the power of clustering-based inference, see the simulation results in Sects. 5.4, 5.6 and in Supplement D.

6 Real data experiments

In this section, we present two experiments derived from an actual functional magnetic resonance imaging dataset (fMRI) to illustrate the power gains brought by compressed
Fig. 6 Results for the simulation described in Sect. 5.5. These plots display the same FWER and TPR statistics as in Fig. 4 for the four inference methods considered: knockoffs, desparsified Lasso, CluDL and EnCluDL.

Fig. 7 Conditional inference on a large-scale functional Magnetic Resonance Imaging dataset. Four approaches are compared to infer what region’s activity predicts hand motor action across participants in a large-scale dataset of 400 subjects leading to $n = 800$. For each participant, we have one brain activity image displaying left-hand activity and another one displaying right-hand activity, and want to infer which brain regions predict whether the subject was moving the left or right hand. Knockoffs and desparsified Lasso work on an image representation with large supervoxels (10k) for the sake of computational tractability, while CluDL and EnCluDL rely on further reduction to $C = 500$ brain regions. The results are displayed on axial views of standard brain template, cut at $z = 50$ mm and $z = 60$ mm. This analysis reveals expected contralateral motor regions involved in prediction for CluDL and EnCluDL, a subset of those with desparsified Lasso and no detection with knockoffs.

The Human Connectome Project (HCP900) dataset is a collection of neuroimaging and behavioral data on healthy young adults aged 22–35. Participants were asked to perform different tasks inside an MRI scanner, while blood oxygenation level-dependent (BOLD) signals of the brain were recorded. The analysis investigates what brain regions are predictive of the subtle variations of cognitive activity across participants, conditionally to other brain regions. The setting is high-dimensional with $n = 800$ brain images taken from 400 participants (two images per participant) and $p = 156,437$ brain voxels.

In Fig. 7, we present conditional association tests that show regions involved in predicting whether the subject use their left or right hand in a hand motion experiment. We use knockoffs, desparsified Lasso, CluDL and EnCluDL. For knockoffs and desparsified Lasso, we had to reduce the dimension from $p = 156,437$ to 10k using a supervoxel approach with Ward algorithm (Varoquaux et al. 2012) for the sake of computational tractability. CluDL and EnCluDL involve further reductions to $C = 500$ regions.

We observe that knockoffs are powerless in this setting, and that desparsified Lasso makes very few detections. By contrast, CluDL and EnCluDL uncover expected contralateral motor regions. This power gain is a clear benefit of the
Fig. 8 Conditional inference on a second large-scale functional magnetic resonance imaging datasets. The figure presents four attempts to infer what region’s activity predicts foot motor activity across participants in a large-scale dataset of 400 subjects (leading to \( n = 800 \)). For each participant, we have one brain activity image displaying left-foot activity and another one displaying right-foot activity, and want to infer which brain regions predict whether the subject was moving the left or right foot. Knockoffs and desparsified Lasso work on an image representation with large supervoxels (10k) for the sake of computational tractability, while CluDL and EnCluDL rely on further reduction to \( C = 500 \) brain regions.

aggressive dimension reduction approach involved in CluDL and EnCluDL.

In Fig. 8, we present the results for another HCP900 task, i.e., we consider another brain activity contrast. In this experiment, we assess regions predicting whether participants are moving the left or right foot. Again, EnCluDL and to some extent CluDL display good face validity, outlining expected cortical and cerebellar motor regions, while desparsified Lasso and knockoffs are powerless. Additional experiments on functional neuroimaging datasets have been published in Chevalier et al. (2021).

7 Discussion

When \( n \ll p \), statistical inference on predictive model parameters is a hard problem. However, when the data are spatially structured, we have shown that ensembled clustered inference procedures are attractive, as they exhibit statistical guarantees and good power. The price to pay is to accept that inference is only accurate up to spatial distance \( \delta \) corresponding to the clustering diameter, thus replacing FWER with \( \delta \)-FWER control guarantees.

One of the most obvious field of application of this class of algorithms is neuroscience where it can be used to solve source localization problems. In that regards, a wide empirical validation of EnCluDL has been conducted in Chevalier et al. (2021) including fMRI data experiments. Also, an extension of EnCluDL was proposed in Chevalier et al. (2020) to address the magneto/electroencephalography source localization problem which involves spatio-temporal data.

With EnCluDL, the statistical inference step is performed by the desparsified Lasso. In Nguyen et al. (2019), another ensembled clustered inference method that leverages the knockoff technique (Barber and Candès 2015) leading to a procedure called ECKO has been tested. However, formal \( \delta \)-FDR control guarantees have not been established yet for this model. It would be also quite natural to try other inference techniques such as the (distilled) conditional randomization test (Candès et al. 2018).

In the present work, we have only considered the linear regression setup. However, combining the same algorithmic scheme with statistical inference solutions for generalized linear models, we could extend this work to the logistic regression setup. This would extend the usability of ensembled clustered inference to many more application settings.

8 Supplementary material

Supplementary material available online includes an analysis of the technical assumptions required when choosing the desparsified Lasso to perform the statistical inference step.
in Supplement A, a proposition for relaxing assumption (ii) of Proposition 4.1 in Supplement B, a diagram summarizing EnCluDL and a study of the complexity of EnCluDL in Supplement C, complementary results for studying the influence of the simulation parameters in Supplement D and the proofs in Supplement E.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11222-022-10139-6.

**Acknowledgements** This study has been funded by Labex DigiCosme (ANR-11-LABEX-0045-DIGICOSME) as part of the program “Investissement d’Avenir” (ANR-11-IDEX-0003-02), by the Fast-Big Project (ANR-17-CE23-0011) and the KARAIB AI Chair (ANR-20-CHIA-0025-01). This study has also been supported by the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 945539, Human Brain Project SGA3).

**References**

Bach, F.R.: Bolasso: model consistent Lasso estimation through the bootstrap. In: Proceedings of the 25th International Conference on Machine Learning, pp. 33–40 (2008)

Barber, R.F., Candès, E.: Controlling the false discovery rate via knockoffs. Ann. Stat. 43(5), 2055–2085 (2015)

Belloc, P.C., Zhang, C.-H.: De-biasing the Lasso with degrees-of-freedom adjustment. Bernoulli 28(2), 713–743 (2022)

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L.: Valid post-selection inference. Ann. Stat. 41(2), 802–837 (2013)

Blanchard, G., Geman, D.: Hierarchical testing designs for pattern recognition. Ann. Stat. 33(3), 1155–1202 (2005)

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

Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.-H.: Correlated variables in regression: clustering and sparse estimation. J. Stat. Plan. Infer. 143(11), 1835–1858 (2013)

Candès, E., Fan, Y., Janson, L., Lv, J.: Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 80(3), 551–577 (2018)

Chatterjee, A., Lahiri, S.N.: Bootstrapping Lasso estimators. J. Am. Stat. Assoc. 106(494), 608–625 (2011)

Cheng, D., He, Z., Schwartzman, A., et al.: Multiple testing of local extrema for detection of change points. Electron. J. Stat. 14(2), 3705–3729 (2020)

Chevalier, J.-A., Salmon, J., Thirion, B.: Statistical inference with ensemble of clustered desparsified Lasso. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 638–646. Springer (2018)

Chevalier, J.-A., Gramfort, A., Salmon, J., Thirion, B.: Statistical control for spatio-temporal MEG/EEG source imaging with desparsified multi-task Lasso. Adv. Neural Inf. Process. Syst. 33, 1759–1770 (2020)

Chevalier, J.-A., Nguyen, T.-B., Salmon, J., Varoquaux, G., Thirion, B.: Decoding with confidence: statistical control on decoder maps. Neuroimage 117921 (2021)

Dehman, A., Ambroise, C., Neuviel, P.: Performance of a blockwise approach in variable selection using linkage disequilibrium information. BMC Bioinform. 16(1), 148 (2015)

Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N.: High-dimensional inference: confidence intervals, p-values and R-software hdci. Stat. Sci. 30(4), 533–558 (2015)

Dunn, O.J.: Multiple comparisons among means. J. Am. Stat. Assoc. 56(293), 52–64 (1961)

Gimenez, J.R., Zou, J.: Discovering conditionally salient features with statistical guarantees. In: International Conference on Machine Learning, pp. 2290–2298 (2019)

Hoyos-Idrobo, A., Varoquaux, G., Kahn, J., Thirion, B.: Recursive nearest aggregate glomeration (ReNa): fast clustering for approximation of structured signals. IEEE Trans. Pattern Anal. Mach. Intell. 41(3), 669–681 (2018)

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

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

Javanmard, A., Montanari, A.: Debiasing the Lasso: optimal sample size for Gaussian designs. Ann. Stat. 46(6A), 2593–2622 (2018)

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R.: A significance test for the Lasso. Ann. Stat. 42(2), 413 (2014)

Meinshausen, N.: Hierarchical testing of variable importance. Biometrika 95(2), 265–278 (2008)

Meinshausen, N.: Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. J. R. Stat. Soc. Ser. B Stat. Methodol. 923–945 (2015)

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

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

Minnier, J., Tian, L., Cai, T.: A perturbation method for inference on regularized regression estimates. J. Am. Stat. Assoc. 106(496), 1371–1382 (2011)

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

Nguyen, T.-B., Chevalier, J.-A., Thirion, B.: Ecko: ensemble of clustered knockoffs for robust multivariate inference on fMRI data. In: International Conference on Information Processing in Medical Imaging, pp. 454–466. Springer (2019)

Nguyen, T.-B., Chevalier, J.-A., Thirion, B., Arlot, S.: Aggregation of multiple knockoffs. In: International Conference on Machine Learning, pp. 7283–7293. PMLR (2020)

Ning, Y., Liu, H.: A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann. Stat. 45(1), 158–195 (2017)

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10(9), 424–430 (2006)

Richards, J.W., Freeman, P.E., Lee, A.B., Scharf, C.M.: Exploiting low-dimensional structure in astronomical spectra. Astrophysics. J. 691(1), 32 (2009)
Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58(1), 267–288 (1996)

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

Varoquaux, G., Gramfort, A., Thirion, B.: Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In: International Conference on Machine Learning (2012)

Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using \( \ell_1 \)-constrained quadratic programming (Lasso). IEEE Trans. Image Process. 55(5), 2183–2202 (2009)

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

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

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.