Spatially relaxed inference on high-dimensional linear models

Jérôme-Alexis CHEVALIER
Inria Paris-Saclay, CEA, Université Paris-Saclay
and
Tuan-Binh NGUYEN
Inria Paris-Saclay, CEA, Université Paris-Saclay, LMO
and
Bertrand THIRION
Inria Paris-Saclay, CEA, Université Paris-Saclay
and
Joseph SALMON
IMAG, Université de Montpellier, CNRS

jerome-alexis.chevalier@inria.fr

June 1, 2021

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 δ-FWER, that is the probability of making a false discovery at a distance greater than δ 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 δ-FWER under standard assumptions for δ equal to the largest cluster diameter. We complement the theoretical analysis with empirical results, demonstrating accurate δ-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 setting. 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, De Martino et al., 2008], astronomy [Richards et al., 2009], but also in other fields where covariates display a spatial structure *e.g.,* in genomics [Balding, 2006, Dehman et al., 2015]. In all these examples, it actually turns out that not only $n < p$ but even $n \ll p$ 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 context, a first approach is to consider the multivariate linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \ ,$$

where the target is denoted by $\mathbf{y} \in \mathbb{R}^n$, the design matrix by $\mathbf{X} \in \mathbb{R}^{n \times p}$, the parameter vector by $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and the random error vector by $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. The aim is to infer $\boldsymbol{\beta}^*$, with statistical guarantees on the estimate, in particular regarding the support, *i.e.,* the set of covariates with non-zero importance.

Statistical inference on individual parameters. In high-dimensional settings, standard statistical inference methodology does not apply, but numerous methods have recently been proposed to recover the non-zero parameters of $\boldsymbol{\beta}^*$ with statistical guarantees. Many methods rely on resampling: bootstrap procedures [Bach, 2008, Chatterjee and Lahiri, 2011, Liu and Yu, 2013], perturbation resampling-based procedures [Minnier et al., 2011], stability selection procedures [Meinshausen and Bühlmann, 2010] and randomized sample splitting [Wasserman and Roeder, 2009, Meinshausen et al., 2009]. All of these approaches suffer from limited power. 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, Lee et al., 2016, Tibshirani et al., 2016], 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, 2019, Celentano et al., 2020]. Additionally, knockoff filters [Barber and Candès, 2015, Candès et al., 2018] consist in creating noisy "fake" 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].

Failure of existing statistical inference methods. 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 $s(\boldsymbol{\beta}^*)$ can be greater than the number of samples even in the sparse setting, where $s(\boldsymbol{\beta}^*) \ll p$. 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 for several multiple hypothesis, the correction cost is heavy [Dunn, 1961, Westfall and Young, 1993, Benjamini and Hochberg, 1995]; for example with Bonferroni correction [Dunn, 1961], p-values are corrected by a factor p when testing every covariate. This make this type of inference methods powerless in our settings (see Fig. 3 for instance). Third, the above approaches are at least quadratic or cubic in the support size, hence become prohibitive whenever both p and n are large.

Combining clustering and inference. 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, $\boldsymbol{\beta}^*$ 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.,* Park et al. [2006], Varoquaux et al. [2012], Hoyos-Idrobo et al. [2018]). 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 ensembled clustered 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 ag-

gregates 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 [Dezeure et al., 2017, Javanmard and Montanari, 2018, Bellec and Zhang, 2019, Celentano et al., 2020]; however, it is be possible to use another inference procedure that produces a p-value family controlling the classical FWER. By contrast, we did not consider the popular knockoff procedure [Barber and Candès, 2015, Candès et al., 2018], that does not produce p-values and does not control the family-wise error rate (FWER). However, an extension of the knockoffs to FWER-type control was proposed by Janson and Su [2016]. It does not control the standard FWER but another relaxed version of the FWER called k-FWER. As it is a relevant alternative to ensembled clustered inference algorithms, we have included it in our empirical comparison (see Section 5). 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 ℓ_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 [Mandozzi and Bühlmann, 2016, Blanchard and Geman, 2005, Meinshausen, 2008] also leverages 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 available, thus replaced by some noisy estimate.

Contributions. 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 that regard, our first contribution is to present a generalization of the FWER called δ-FWER, that takes into account a spatial tolerance of magnitude δ for the false discoveries. Then, our main contribution is to prove that ensembled clustered inference algorithms control the δ-FWER under reasonable assumptions for a given tolerance parameter δ. Finally, we apply the ensembled clustered inference scheme to the desparsified Lasso leading to the EnCluDL algorithm and conduct an empirical study: we show that EnCluDL exhibits a good statistical power in comparison with alternative procedures and we verify that it displays the expected δ-FWER control.

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

i-th row and j-th column.

2 Model and data assumptions

2.1 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 $\mathbf{X} = (\mathbf{X}_{i,j})_{i\in[n],j\in[p]}$ the random design matrix, where n is the number of samples and p the number of covariates, the rows $(\mathbf{X}_{i,.})_{i\in[n]}$ are sampled from a random field defined on a discrete domain.

2.2 Gaussian random design model and high dimensional settings

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \ , \tag{1}$$

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

2.3 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 $\mathrm{Cor}(\mathbf{X}_{.,j}, \mathbf{X}_{.,k}) = \mathbf{\Sigma}_{j,k}/\sqrt{\mathbf{\Sigma}_{j,j}\mathbf{\Sigma}_{k,k}}$. We now introduce a key structural assumption: two covariates at a spatial distance smaller than δ 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 $\mathbf{\Sigma}_{j,k} \geq 0$.*

Under model (1), each coordinate of the parameter vector $\boldsymbol{\beta}^*$ links one covariate to the target. Then, $\boldsymbol{\beta}^*$ has the same underlying organization as the covariates and is also called weight map in these settings. Defining its *support* as $S(\boldsymbol{\beta}^*) = \{j \in [p] : \beta_j^* \neq 0\}$ and its cardinal as $s(\boldsymbol{\beta}^*) = |S(\boldsymbol{\beta}^*)|$, we assume that the true model is sparse, meaning that $\boldsymbol{\beta}^*$ has a small number of non-zero entries, *i.e.,* $s(\boldsymbol{\beta}^*) \ll p$. The complementary of $S(\boldsymbol{\beta}^*)$ in $[p]$ is called the *null region* and is denoted by $N(\boldsymbol{\beta}^*)$, *i.e.,* $N(\boldsymbol{\beta}^*) = \{j \in$

$[p] : \boldsymbol{\beta}^*_j = 0\}$. Additionally to the sparse assumption, we assume that $\boldsymbol{\beta}^*$ 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 $\boldsymbol{\beta}^*$ 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* $\text{sign}(\boldsymbol{\beta}^*_j) = \text{sign}(\boldsymbol{\beta}^*_k)$.

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

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 $\hat{\boldsymbol{\beta}}$ 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 $S(\hat{\boldsymbol{\beta}}) \subset [p]$ a general estimate of the support $S(\boldsymbol{\beta}^*)$ derived from the estimator $\hat{\boldsymbol{\beta}}$.

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

$$H_0^\delta(j) : \text{“for all } k \in [p] \text{ such that } d(j,k) \leq \delta,\ \boldsymbol{\beta}^*_k = 0\text{”} ,$$
$$H_1^\delta(j) : \text{“there exists } k \in [p] \text{ such that } d(j,k) \leq \delta \text{ and } \boldsymbol{\beta}^*_k \neq 0\text{”} .$$

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

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

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

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 $N^\delta(\boldsymbol{\beta}^*)$ (or simply N^δ):*

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

When $\delta = 0$ the δ-null region is simply the null region : $N^0(\boldsymbol{\beta}^*) = N(\boldsymbol{\beta}^*)$. 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 $N^{\delta_2}(\boldsymbol{\beta}^*) \subseteq N^{\delta_1}(\boldsymbol{\beta}^*) \subseteq N(\boldsymbol{\beta}^*)$ (see Fig. 2-(d) 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 α:*

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

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 α is $\mathscr{E}^\delta_\alpha$, the set of indexes belonging both to the δ-null region and to the rejection region at level α. We also refer to this region as the erroneous rejection region at level α with tolerance δ:*

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

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

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 α with respect to the family $\hat{p}$, denoted δ-FWER$_\alpha(\hat{p})$, is the probability that the δ-type 1 error region at level α is not empty:*

$$\delta\text{-}FWER_\alpha(\hat{p}) = \mathbb{P}(|\mathscr{E}^\delta_\alpha(\hat{p})| \geq 1) = \mathbb{P}(\min_{j\in N^\delta} \hat{p}_j \leq \alpha) \quad .$$

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 $\alpha \in (0,1)$:*

$$\delta\text{-}FWER_\alpha(\hat{p}) \leq \alpha \quad .$$

When $\delta = 0$ the δ-FWER is the usual FWER. Additionally, for $0 \leq \delta_1 \leq \delta_2$, one can verify that δ_2-FWER$_\alpha(\hat{p}) \leq \delta_1$-FWER$_\alpha(\hat{p}) \leq$ FWER$_\alpha(\hat{p})$. 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 statistical inference procedure. In Sec. 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 $(\mathbf{X}, \mathbf{y})$, we take as input the transformation matrix $\mathbf{A} \in \mathbb{R}^{p \times C}$ 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 $\mathbf{Z}$ and the target $\mathbf{y}$ 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 Clustered inference

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

$\mathbf{Z} = \mathbf{X}\mathbf{A}$ `// compressed design matrix`

$\hat{p}^{\mathcal{G}} = \texttt{statistical\_inference}(\mathbf{Z}, \mathbf{y})$ `// uncorrected cluster-wise p-values`

$\hat{q}^{\mathcal{G}} = C \times \hat{p}^{\mathcal{G}}$ `// corrected cluster-wise p-values`

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

  $\hat{q}_j = \hat{q}^{\mathcal{G}}_c$ 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

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

param : C, B

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

  $\mathbf{X}^{(b)} = \texttt{sampling}(\mathbf{X})$ `// sampling rows of` $\mathbf{X}$

  $\mathbf{A}^{(b)} = \texttt{clustering}(q, \mathbf{X}^{(b)})$ `// transformation matrix`

  $\hat{q}^{(b)} = \texttt{clustered\_inference}(\mathbf{X}, \mathbf{y}, \mathbf{A}^{(b)})$ `// families of corr. covariate-wise p-val.`

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

  $\hat{q}_j = \texttt{ensembling}(\{\hat{q}_j^{(b)}, b \in [B]\})$ `// 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 Sec. 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 $\mathbf{X}$. The `clustering` function derives a choice of clustering in C clusters, it produces a transformation matrix $\mathbf{A}^{(b)} \in \mathbb{R}^{p \times C}$

that should vary for each bootstrap $b \in [B]$ since the subsampled data $\mathbf{X}^{(b)}$ varies. Once the clustering inference steps are completed, the `ensembling` function aggregates the B (corrected) p-value families into a single one.

Fig. 1 can help the reader to better understand the organization of the next sections, aiming eventually at establishing the δ-FWER control property of the clustered inference and ensembled clustered inference algorithms.

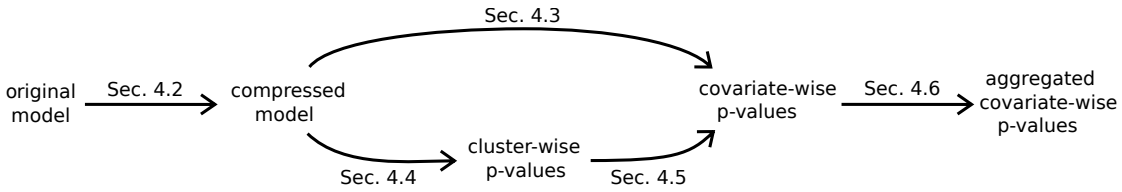

Figure 1: Organization of Section 4.

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 $r \in [q]$, we denote by G_r the r-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 $\mathbf{Z} \in \mathbb{R}^{n \times C}$ 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 $\mathbf{X}$, dimension reduction can be written:

$$\mathbf{Z} = \mathbf{X}\mathbf{A} \ , \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{p \times q}$ is the transformation matrix defined by:

$$\mathbf{A} = \begin{bmatrix} \alpha_1 \text{—} \alpha_1 & 0 \text{—} 0 & \ldots & 0 \text{—} 0 \\ 0 \text{—} 0 & \alpha_2 \text{—} \alpha_2 & \ldots & 0 \text{—} 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 \text{—} 0 & 0 \text{—} 0 & \ldots & \alpha_C \text{—} \alpha_C \end{bmatrix} \ ,$$

where $\alpha_c = 1/|G_c|$ for all $c \in [C]$. Consequently, the distribution of the i-th row of $\mathbf{Z}$ is given by $\mathbf{Z}_{i,\cdot} \sim \mathcal{N}_q(0, \boldsymbol{\Upsilon})$, where $\boldsymbol{\Upsilon} = \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A}$. The correlation between the groups $r \in [q]$ and $l \in [q]$ is given by $\mathrm{Cor}(\mathbf{Z}_{\cdot,r}, \mathbf{Z}_{\cdot,l}) = \boldsymbol{\Upsilon}_{r,l} / \sqrt{\boldsymbol{\Upsilon}_{r,r} \boldsymbol{\Upsilon}_{l,l}}$. As mentioned in Bühlmann et al. [2013], because of the Gaussian assumption in (1), we have the following compressed representation:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta}^* + \boldsymbol{\eta} \ , \tag{3}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^q$, $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma_\eta^2 \mathbf{I}_n)$, $\sigma_\eta \geq \sigma_\varepsilon > 0$ and $\boldsymbol{\eta}$ is independent of $\mathbf{Z}$.

Remark 4.1. *Dimension reduction is not the unique desirable effect of clustering with regards to statistical inference. Indeed, this 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 problem which is a consequence of Bühlmann et al. [2013, Proposition 4.3].

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

$$
\begin{aligned}
&(i)\ \textit{for all } c \in [C],\ \textit{for all } (j,k) \in (G_c)^2,\ \boldsymbol{\Sigma}_{j,k} \geq 0\ ,\\
&(ii)\ \textit{for all } c \in [C],\ \textit{for all } c' \in [C] \setminus \{c\},\ \boldsymbol{\Upsilon}_{c,c'} = 0\ ,\\
&(iii)\ \textit{for all } c \in [C], \left(\boldsymbol{\beta}^*_j \geq 0 \textit{ for all } j \in G_c\right) \text{ or } \left(\boldsymbol{\beta}^*_j \leq 0 \textit{ for all } j \in G_c\right)\ ,
\end{aligned}
$$

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

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.,* $\operatorname{Diam}(G_c) = \max\{d(j,k) : (j,k) \in (G_c)^2\}$ and the clustering diameter of $\mathcal{G}$ by the largest group diameter, *i.e.,* $\operatorname{Diam}(\mathcal{G}) = \max\{\operatorname{Diam}(G_c) : c \in [C]\}$. In Fig. 2-(b), we propose a clustering of the initial weight map in Fig. 2-(a) for which the clustering diameter is equal to 2 for the ℓ_1 distance. Assumption (i) notably holds when $\operatorname{Diam}(\mathcal{G}) \leq \delta$ under the spatial homogeneity assumption (Ass. 2.1) with parameter δ. Assumption (ii) assumes independence of the groups. A sufficient condition is when the covariates covariance matrix $\boldsymbol{\Sigma}$ 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 δ and the weight map satisfies the sparse-smooth assumption (Ass. 2.2) with parameter δ. For instance, a clustering-based compressed representation of the weight map in Fig. 2-(a) is given in Fig. 2-(c).

4.4 Statistical inference on the compressed model

To perform the statistical inference on the compressed problem (3), we could consider any statistical inference procedure that produces cluster-wise p-values $\hat{p}^{\mathcal{G}} = (\hat{p}^{\mathcal{G}}_c)_{c \in [C]}$, given a choice of clustering $\mathcal{G}$, that control the type 1 error. More precisely, for any $c \in [C]$, under $H_0(G_c)$, *i.e.,* the null hypothesis which states that θ^*_c is equal to zero

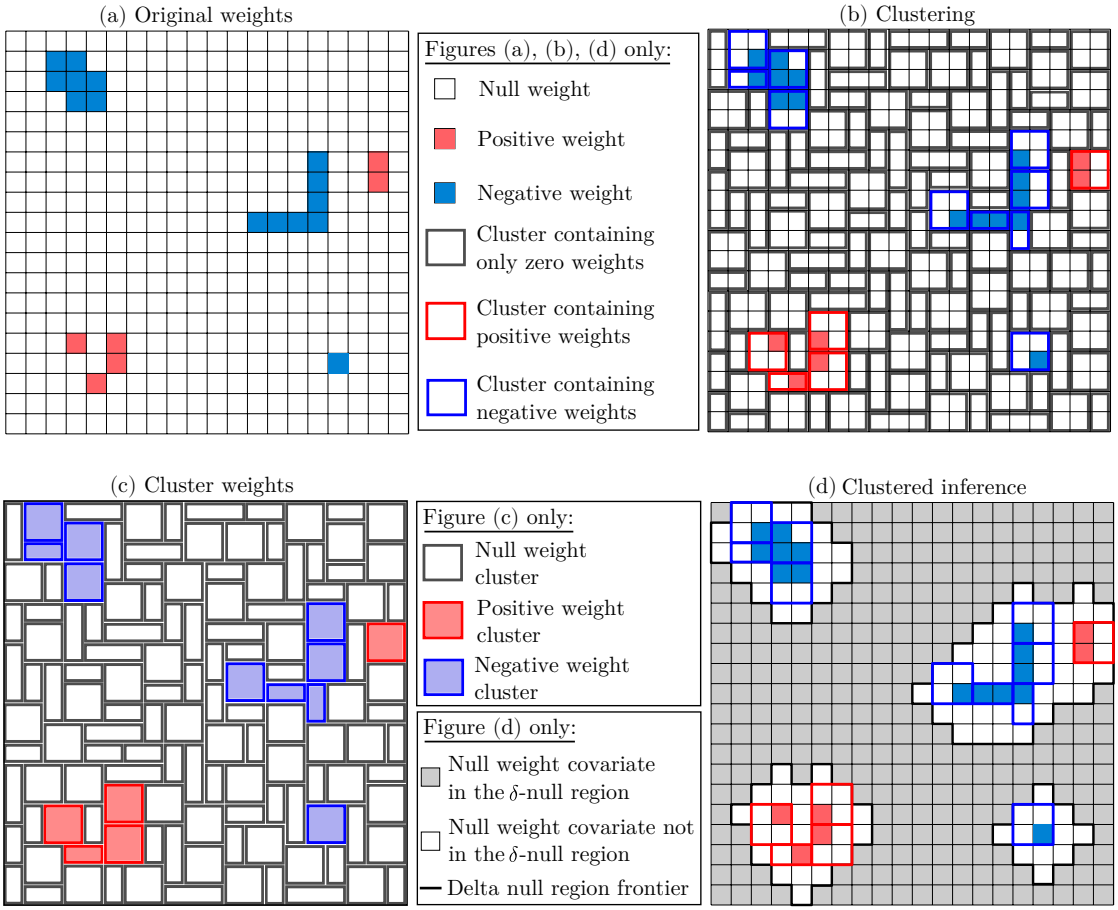

Figure 2: Clustered inference mechanism on 2D-spatially structured data. Item a: Example of weight map with a 2D-structure. Voxels represent covariates, with blue (resp. red) corresponding to negative (resp. positive) weights; others are null weights. Item b: Arbitrary choice of spatially constrained clustering with a diameter of $\delta = 2$ units for the ℓ_1-distance. Rectangles delimited by black lines represent clusters that contain only zero-weight covariates. Blue (resp. red) rectangles refer to clusters that contain negative-weight (resp. positive) covariates. Item c: Compressed model weights: under the assumptions of Prop. 4.1, the cluster weights share the same signs as the covariate weights they contain. Blue (resp. red) rectangles correspond to negative-weight (resp. positive-weights) clusters. Item d: The grey area corresponds to the δ-null region ($\delta = 2$). Under the same assumptions, the non-zero weight groups have no intersection with the δ-null region.

in the compressed model, we assume that the p-value associated with the c-th cluster verifies:

$$\mathbb{P}(\hat{p}_c^{\mathcal{G}} \leq \alpha) \leq \alpha \ . \tag{4}$$

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 $\hat{q}^{\mathcal{G}} = (\hat{q}^{\mathcal{G}}_c)_{c\in[C]}$ is defined by:

$$\hat{q}^{\mathcal{G}}_c = \min\{1, C \times \hat{p}^{\mathcal{G}}_c\} \ . \tag{5}$$

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

$$\mathrm{FWER}_\alpha(\hat{q}^{\mathcal{G}}) = \mathbb{P}(\min_{c\in N_{\mathcal{G}}} \hat{q}^{\mathcal{G}}_c \leq \alpha) \leq \alpha \ . \tag{6}$$

This means that the cluster-wise p-value family $\hat{q}^{\mathcal{G}}$ controls FWER.

4.5 De-grouping

Given the families of cluster-wise p-values $\hat{p}^{\mathcal{G}}$ and corrected p-values $\hat{q}^{\mathcal{G}}$ as defined in (10) 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:

$$\begin{aligned} &\text{for all } j \in [p], \quad \hat{p}_j = \sum_{c\in[C]} \mathbb{1}_{\{j\in G_c\}}\ \hat{p}^{\mathcal{G}}_c \ , \\ &\text{for all } j \in [p], \quad \hat{q}_j = \sum_{c\in[C]} \mathbb{1}_{\{j\in G_c\}}\ \hat{q}^{\mathcal{G}}_c \ . \end{aligned} \tag{7}$$

Proposition 4.2. *Under the assumptions of Prop. 4.1 and assuming that the clustering diameter is smaller than δ, then:*

(i) elements of the family $\hat{p}$ defined in (7) control the δ-type 1 error:

$$\textit{for all } j \in N^\delta, \ \textit{for all } \alpha \in (0, 1), \ \mathbb{P}(\hat{p}_j \leq \alpha) \leq \alpha \ ,$$

(ii) the family $\hat{q}$ defined in (7) controls the δ-FWER:

$$\textit{for all } \alpha \in (0, 1), \ \mathbb{P}(\min_{j\in N^\delta}(\hat{q}_j) \leq \alpha) \leq \alpha \ .$$

Proof. See Supplement E.2. □

The previous de-grouping properties can be seen in Fig. 2-(d). 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 $b \in [B]$, we denote by $\hat{q}^{(b)}$ 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 $b \in [B]$, the p-value families $\hat{q}^{(b)}$ control the δ-FWER. Then, for any $\gamma \in (0,1)$, the ensembled p-value family $\tilde{q}(\gamma)$ defined by:*

$$\text{for all } j \in [p],\ \tilde{q}_j(\gamma) = \min\left\{1, \gamma\text{-quantile}\left(\left\{\frac{\hat{q}_j^{(b)}}{\gamma} : b \in [B]\right\}\right)\right\}\ , \tag{8}$$

controls the δ-FWER.

Proof. See Supplement E.3. □

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, Ass. 2.1 and Ass. 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 Prop. 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 4.2. *When the type 1 error control offered by the statistical inference procedure is only asymptotic, the result stated by Theorem 4.1 remains true asymptotically. This is notably the case when using desparsified Lasso: under the assumptions of Theorem 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

For testing the (ensembled) clustered inference algorithms, we have decided to make the inference step using the desparsified Lasso [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 to perform the statistical inference step. A notable difference is the fact that all the results becomes asymptotic. In Supplement C, we present a diagram illustrating the mechanism of EnCluDL and analyse its numerical complexity.

5.2 2D Simulation

We run a series of simulations on 2D 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 $p = H^2 = 1\,600$ covariates, with a sample size $n \in \{50, \mathbf{100}, 200, 400\}$. To construct $\boldsymbol{\beta}^*$, we define a 2D weight map $\tilde{\boldsymbol{\beta}}^*$ with four active regions (as illustrated in Fig. 3) and then flatten $\tilde{\boldsymbol{\beta}}^*$ to a vector $\boldsymbol{\beta}^*$ of size p. Each active region is a square of width $h \in \{2, \mathbf{4}, 6, 8\}$, leading to a size of support of 1%, **4**%, 9% or 16%. To construct the design matrix, we first build a 2D data matrix $\tilde{\mathbf{X}}$ 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 $\mathbf{X} \in \mathbb{R}^{n \times p}$. The intensity of the spatial smoothing is adjusted to achieve a correlation between two adjacent covariates (local correlation) of $\rho \in \{0.5, \mathbf{0.75}, 0.9, 0.95\}$. We also set the noise standard deviation $\sigma_\varepsilon \in \{1, \mathbf{2}, 3, 4\}$, which corresponds to a signal to noise ratio (SNR) $\mathrm{SNR}_y \in \{6.5, \mathbf{3.5}, 2.2, 1.5\}$, where the SNR is defined by $\mathrm{SNR}_y = \|\mathbf{X}\boldsymbol{\beta}^*\|_2 / \|\boldsymbol{\varepsilon}\|_2$. 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 $C \in [100; 400]$. 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 δ to this value. We obtained the couples $(C, \delta) \in \{(100, 8), (200, 6), (300, 5), (400, 4)\}$. The tolerance region is represented in Fig. 3 for $\delta = 6$. 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 clustering randomization.

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, none of these includes 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 last extension of knockoffs. Depending on the nominal rate at which we want to control the k-FWER, the choice of k is not arbitrary. 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 δ-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 δ-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, δ-FWER control might be more relevant since it controls the very undesirable far-from-support false discoveries.

5.4 Results

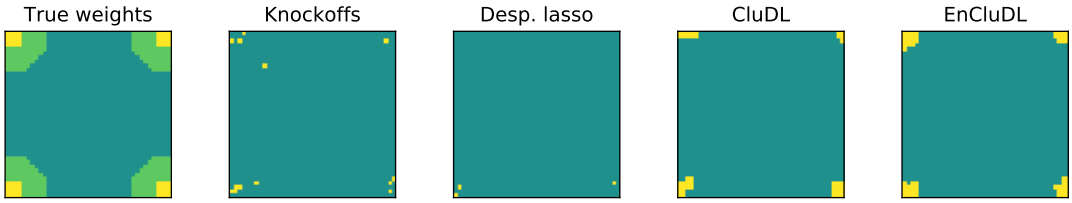

Figure 3: True support and estimated support for the first seed of the central scenario. Left: The support in yellow is composed of four regions of width $h = 4$ covariates. The tolerance region in green surrounds the support, its width is $\delta = 6$ covariates. The remaining covariates in blue form the δ-null region. 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 δ-FWER at 10% for $\delta = 6$.

In Fig. 3, we plot the maps estimated by knockoffs, desparsified Lasso, CluDL and EnCluDL for $C = 200$ when solving the first seed 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 the CluDL one and recovers groups of covariates that correspond almost perfectly to the true weights. Both CluDL and EnCluDL are only accurate up to the spatial tolerance which is $\delta = 6$, but EnCluDL fits the ground truth more tightly.

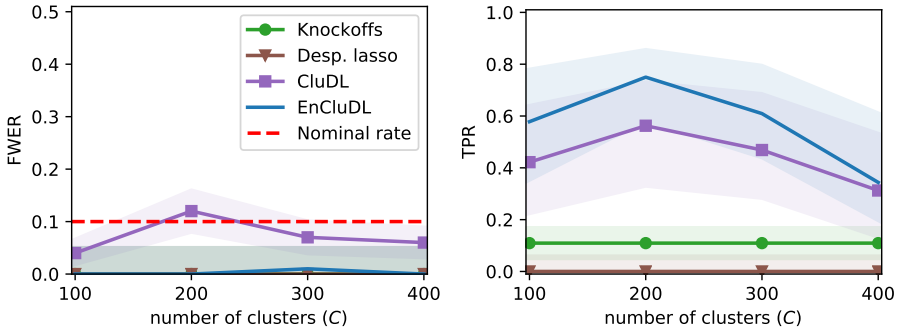

Figure 4: Results for fixed simulation parameters corresponding to the central scenario simulation. The green line with circles correspond to knockoffs, the brown line with triangles is the desparsified Lasso, the purple squared line correspond to CluDL and the blue plain line is EnCluDL. Left: Empirical FWER for desparsified Lasso, k-FWER for knockoffs and δ-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.

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 δ-FWER. 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 due to $n \ll p$. However, CluDL has decent power and EnCluDL improves over CluDL thanks to clustering randomization. Finally, CluDL and EnCluDL are flexible with respect to the choice of C since the TPR varies quite slowly 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. The main conclusion gained from these complementary results is the fact that, up to the limit given by the desired spatial tolerance δ, the choice of C should be made in function of the data structure. More precisely, good clustering creates clusters that are weakly correlated and contains covariates that are highly correlated. This observation is linked to assumption (ii) of Prop. 4.1.

6 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 δ corresponding to the clustering diameter, thus replacing FWER with δ-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 δ-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, Liu and Janson, 2020].

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.

Acknowledgement

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 FastBig 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).

Supplementary material

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

References

F. R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40, 2008. 2

D. J. Balding. A tutorial on statistical methods for population association studies. *Nature reviews genetics*, 7(10):781–791, 2006. 2

R. F. Barber and E. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 10 2015. 2, 4, 15, 17

P. C. Bellec and C.-H. Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *arXiv preprint arXiv:1902.08885*, 2019. 2, 4

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57 (1):289–300, 1995. 3

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

G. Blanchard and D. Geman. Hierarchical testing designs for pattern recognition. *The Annals of Statistics*, 33(3):1155–1202, 2005. 4

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19 (4):1212–1242, 09 2013. 2, 8

P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835 – 1858, 2013. 3, 9, 10, 24, 28, 29

E. Candès, Y. Fan, L. Janson, and J. Lv. 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. 2, 4, 15, 17

M. Celentano, A. Montanari, and Y. Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020. 2, 4

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.*, 106(494):608–625, 2011. 2

D. Cheng, Z. He, A. Schwartzman, et al. Multiple testing of local extrema for detection of change points. *Electron. J. Stat.*, 14(2):3705–3729, 2020. 6

J.-A. Chevalier, J. Salmon, and B. Thirion. Statistical inference with ensemble of clustered desparsified lasso. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 638–646. Springer, 2018. 3, 14

J.-A. Chevalier, A. Gramfort, J. Salmon, and B. Thirion. Statistical control for spatio-temporal meg/eeg source imaging with desparsified multi-task lasso. In *Thirty-fourth Conference on Neural Information Processing Systems*, 2020. 17

J.-A. Chevalier, T.-B. Nguyen, J. Salmon, G. Varoquaux, and B. Thirion. Decoding with confidence: Statistical control on decoder maps. *NeuroImage*, page 117921, 2021. 3, 17

F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, 43(1):44–58, 2008. 2

A. Dehman, C. Ambroise, and P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC bioinformatics*, 16 (1):148, 2015. 2, 14

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

R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017. 4

O. J. Dunn. Multiple comparisons among means. *J. Amer. Statist. Assoc.*, 56(293): 52–64, 1961. 3, 12

J. R. Gimenez and J. Zou. Discovering conditionally salient features with statistical guarantees. *International Conference on Machine Learning*, pages 2290–2298, 2019. 6

A. Hoyos-Idrobo, G. Varoquaux, J. Kahn, and B. Thirion. Recursive nearest agglomeration (rena): fast clustering for approximation of structured signals. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):669–681, 2018. 3

L. Janson and W. Su. Familywise error rate control via knockoffs. *Electron. J. Stat.*, 10 (1):960–975, 2016. 4, 15

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014. 2, 4, 8, 13, 22

A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *Ann. Statist.*, 46(6A):2593–2622, 2018. 2, 4

J. Lee, D. Sun, Y. Sun, and J. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 2016. 2

H. Liu and B. Yu. Asymptotic properties of lasso+ mls and lasso+ ridge in sparse high-dimensional linear regression. *Electron. J. Stat.*, 7:3124–3169, 2013. 2

M. Liu and L. Janson. Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*, 2020. 17

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42(2):413, 2014. 2

J. Mandozzi and P. Bühlmann. Hierarchical testing in the high-dimensional setting with correlated variables. *J. Amer. Statist. Assoc.*, 111(513):331–343, 2016. 4

N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008. 4

N. Meinshausen. 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.*, pages 923–945, 2015. 4

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72:417–473, 2010. 2

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009. 2, 8, 13, 30

J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. *J. Amer. Statist. Assoc.*, 106(496):1371–1382, 2011. 2

R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with debiased scaled group lasso. *Electron. J. Stat.*, 10(2):1829–1873, 2016. 4

E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. In *Journal of Physics: Conference Series*, volume 904, page 012006. IOP Publishing, 2017. 22

T.-B. Nguyen, J.-A. Chevalier, and B. Thirion. Ecko: Ensemble of clustered knockoffs for robust multivariate inference on fMRI data. In *International Conference on Information Processing in Medical Imaging*, pages 454–466. Springer, 2019. 6, 17

T.-B. Nguyen, J.-A. Chevalier, B. Thirion, and S. Arlot. Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR, 2020. 4

Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, 45(1):158–195, 2017. 3

K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430, 2006. 2

M. Y. Park, T. Hastie, and R. Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, 05 2006. 3

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016. 22

J.W. Richards, P.E. Freeman, A.B. Lee, and C.M. Schafer. Exploiting low-dimensional structure in astronomical spectra. *The Astrophysical Journal*, 691(1):32, 2009. 2

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996. 2

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.*, 111(514):600–620, 2016. 2

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014. 2, 4, 8, 13, 15, 22

G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *International Conference on Machine Learning*, 2012. 3, 14

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Image Process.*, 55 (5):2183–2202, 2009. 3

L. Wasserman and K. Roeder. High-dimensional variable selection. *Ann. Statist.*, 37 (5A):2178–2201, 2009. 2, 8

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993. 3

G. Yu and J. Bien. Estimating the error variance in a high-dimensional linear model. *Biometrika*, 106(3):533–546, 2019. 22

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(1):217–242, 2014. 2, 4, 8, 13, 22

Supplementary material for "Spatially relaxed inference on high-dimensional linear models"

A Desparsified Lasso on the compressed model

Here, we clarify the assumptions and refinements that occur when chosing the desparsified Lasso as the procedure that performs the statistical inference on the compressed model. The desparsified Lasso was first developed in Zhang and Zhang [2014] and Javanmard and Montanari [2014], and thoroughly analyzed in van de Geer et al. [2014]. Following notation in Eq. (3), the true support in the compressed model is denoted by $S(\boldsymbol{\theta}^*) = \{c \in [C] : \boldsymbol{\theta}^*_c \neq 0\}$ and its cardinality by $s(\boldsymbol{\theta}^*) = |S(\boldsymbol{\theta}^*)|$. We also denote by $\boldsymbol{\Omega} \in \mathbb{R}^{C\times C}$ the inverse of the population covariance matrix of the groups, *i.e.*, $\boldsymbol{\Omega} = \boldsymbol{\Upsilon}^{-1}$. Then, for $c \in [C]$, the sparsity of the c-th row of $\boldsymbol{\Omega}$ (or c-th column) is $s(\boldsymbol{\Omega}_{c,.}) = |S(\boldsymbol{\Omega}_{c,.})|$, where $S(\boldsymbol{\Omega}_{c,.}) = \{c' \in [C] : \boldsymbol{\Omega}_{c,c'} \neq 0\}$. We also denote the smallest eigenvalue of $\boldsymbol{\Upsilon}$ by $\phi_{\min}(\boldsymbol{\Upsilon}) > 0$. We can now state the assumptions required for probabilistic inference with desparsified Lasso [van de Geer et al., 2014]:

Theorem A.1 (Theorem 2.2 of van de Geer et al. [2014])**.** *Considering the model in Eq. (3) and assuming:*

$$
\begin{aligned}
&(i)\ 1/\phi_{\min}(\boldsymbol{\Upsilon}) = \mathcal{O}(1)\ , \\
&(ii)\ \max_{c\in[C]}(\boldsymbol{\Upsilon}_{c,c}) = \mathcal{O}(1)\ , \\
&(iii)\ s(\theta^*) = o(\sqrt{n}/\log(C))\ , \\
&(iv)\ \max_{c\in[C]}(s(\boldsymbol{\Omega}_{c,.})) = o(n/\log(C))\ ,
\end{aligned}
$$

then, denoting by $\hat{\boldsymbol{\theta}}$ the desparsified Lasso estimator derived from the inference procedure described in van de Geer et al. [2014], the following holds:

$$
\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \boldsymbol{\xi} + \boldsymbol{\zeta}\ , \\
&\boldsymbol{\xi}|\mathbf{Z} \sim \mathcal{N}(0_C, \sigma^2_{\boldsymbol{\eta}}\hat{\boldsymbol{\Omega}})\ , \\
&\|\boldsymbol{\zeta}\|_\infty = o_{\mathbb{P}}(1)\ ,
\end{aligned}
$$

where $\hat{\boldsymbol{\Omega}}$ is such that $\left\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\right\|_\infty = o_{\mathbb{P}}(1)$.

Remark A.1. *In Theorem A.1, to compute confidence intervals, the noise standard deviation σ_η in the compressed problem has to be estimated. We refer the reader to the surveys that are dedicated to this subject such as Reid et al. [2016], Ndiaye et al. [2017], Yu and Bien [2019].*

As argued in van de Geer et al. [2014], from Theorem A.1 we obtain asymptotic confidence intervals for the r-th element of $\boldsymbol{\theta}^*$ from the following equations, for all

$z_1 \in \mathbb{R}$ and $z_2 \in \mathbb{R}^+$:

$$\begin{aligned}
&\mathbb{P}\left[\frac{\sqrt{n}(\hat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}_c^*)}{\sigma_\eta\sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \leq z_1 \;\middle|\; \mathbf{Z}\right] - \Phi(z_1) = o_{\mathbb{P}}(1) \;\;, \\
&\mathbb{P}\left[\frac{\sqrt{n}|\hat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}_c^*|}{\sigma_\eta\sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \leq z_2 \;\middle|\; \mathbf{Z}\right] - (2\Phi(z_2) - 1) = o_{\mathbb{P}}(1) \;\;,
\end{aligned} \tag{9}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Thus, for each $c \in [C]$ one can provide a p-value that assesses whether or not $\boldsymbol{\theta}_c^*$ is equal to zero. In the case of a two-sided single test, for each $c \in [C]$, the p-value denoted by $\hat{p}_c^{\mathcal{G}}$ is:

$$\hat{p}_c^{\mathcal{G}} = 2\left(1 - \Phi\left(\frac{\sqrt{n}|\hat{\boldsymbol{\theta}}_c|}{\sigma_\eta\sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}}\right)\right) \;\;. \tag{10}$$

Under $H_0(G_c)$, from (9), we have, for any $\alpha \in (0,1)$:

$$\begin{aligned}
\mathbb{P}(\hat{p}_c^{\mathcal{G}} \leq \alpha \mid \mathbf{Z}) &= 1 - \mathbb{P}\left[\frac{\sqrt{n}|\hat{\boldsymbol{\theta}}_c|}{\sigma_\eta\sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \;\middle|\; \mathbf{Z}\right] \\
&= \alpha + o_{\mathbb{P}}(1) \;\;.
\end{aligned} \tag{11}$$

Then, (11) shows that the p-values $\hat{p}_c^{\mathcal{G}}$ asymptotically control type 1 errors. Using the Bonferroni correction, the family of corrected p-values $\hat{q}^{\mathcal{G}} = (\hat{q}_c^{\mathcal{G}})_{c\in[C]}$ remains defined by:

$$\hat{q}_c^{\mathcal{G}} = \min\{1, C \times \hat{p}_c^{\mathcal{G}}\} \;\;. \tag{12}$$

Then, for all $\alpha \in (0,1)$:

$$\mathrm{FWER}_\alpha(\hat{q}^{\mathcal{G}}) = \mathbb{P}(\min_{c\in N_{\mathcal{G}}} \hat{q}_c^{\mathcal{G}} \leq \alpha \mid \mathbf{Z}) \leq \alpha + o_{\mathbb{P}}(1) \;\;. \tag{13}$$

Then, (13) shows that the p-value family $\hat{q}^{\mathcal{G}}$ asymptotically control FWER. Finally, we have shown that desparsified Lasso applied to a compressed version of the original problem provides cluster-wise p-value families $\hat{p}^{\mathcal{G}}$ and $\hat{q}^{\mathcal{G}}$ that control respectively the type 1 error and the FWER in the compressed model only asymptotically.

B Relaxing the uncorrelated clusters assumption

As noted in Sec. 4.3, assumption (ii) of Prop. 4.1 is often unmet in practice. Here, taking the particular case in which the inference step is performed by desparsified Lasso, we relax the assumption and show that it is still possible to compute an adjusted corrected p-value that asymptotically controls the δ-FWER. Hopefully, the technique used to derive

this relaxation would also applicable to other parametric statistical inference methods such as corrected ridge. To better understand the development made in this section, the adjusted p-values of this section should be compared with the original p-values of Supplement A. Note that, this extension is easy to integrate in the proof of the main results Theorem 4.1 as it just requires to use the adjusted corrected p-value instead of the original corrected p-value. Also, it does not provide much more insight about clustered inference algorithms. This is why we have decided to keep this extension for Supplementary Materials.

First, we replace Prop. 4.1 by the next proposition that is a consequence of Bühlmann et al. [2013, Proposition 4.4].

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

(i) for all $c \in [C]$, *for all* $j, k \in G_c^2$, $\operatorname{Cov}(\mathbf{X}_{.,j}, \mathbf{X}_{.,k} \mid \{\mathbf{Z}_{.,c'} : c' \neq c\}) \geq 0$,

(ii.a) for all $c \in [C]$, *there exists* $\boldsymbol{\nu}_c \in \mathbb{R}^+$ *s.t. for all* $j \in G_c$, *for all* $k \notin G_c$,

$$|\operatorname{Cov}(\mathbf{X}_{.,j}, \mathbf{X}_{.,k} \mid \{\mathbf{Z}_{.,c'} : c' \neq c\})| \leq \boldsymbol{\nu}_c \ ,$$

(ii.b) for all $c \in [C]$, *there exists* $\boldsymbol{\tau}_c > 0$ *s.t.* $\operatorname{Var}(\mathbf{Z}_{.,c} \mid \{\mathbf{Z}_{.,c'} : c' \neq c\}) \geq \boldsymbol{\tau}_c$,

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

then, in the compressed representation (3), $\boldsymbol{\theta}^*$ *admits the following decomposition:*

$$\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}} + \boldsymbol{\kappa} \ , \tag{14}$$

where, for all $c \in [C]$, $|\boldsymbol{\kappa}_c| \leq (\boldsymbol{\nu}_c \ / \ \boldsymbol{\tau}_c)\|\boldsymbol{\beta}^*\|_1$ *and* $\tilde{\boldsymbol{\theta}}_c \neq 0$ *if and only if there exists* $j \in G_c$ *such that* $\boldsymbol{\beta}_j^* \neq 0$. *If such an index* j *exists then* $\operatorname{sign}(\tilde{\boldsymbol{\theta}}_c) = \operatorname{sign}(\boldsymbol{\beta}_j^*)$.

Proof. See Supplement E.1. □

The assumptions (i) and (ii) in Prop. 4.1 are replaced by (i), $(ii.a)$ and $(ii.b)$ in Prop. B.1. More precisely, instead of assuming that the covariates inside a group are positively correlated, we assume that they are positively correlated conditionally to all other groups. Also, we relax the more questionable assumption of groups independence; we assume instead that the conditional covariance of two covariates of different groups is bounded above $(ii.a)$ and that the conditional variance of the group representative variable is non-zero $(ii.b)$. In practice, except when group representative variables are linearly dependent, we can always find values for which $(ii.a)$ and $(ii.b)$ are verified, but we would like the upper bound of $(ii.a)$ as low as possible and the lower bound of $(ii.b)$ as high as possible. Finally, assumption (iii) remains unchanged.

Then, as done in Supplement A, we can build $\hat{\boldsymbol{\theta}}$. Under the same assumptions, Theorem A.1 is still valid and $\hat{\boldsymbol{\theta}}$ still verifies (9). However, here we want to estimate $\tilde{\boldsymbol{\theta}}$, not $\boldsymbol{\theta}^*$. Combining Theorem A.1 and Prop. B.1, we can see $\hat{\boldsymbol{\theta}}$ as a biased estimator of

$\tilde{\boldsymbol{\theta}}$. To take this bias into account, we need to adjust the definition of the p-values given by (10). Let us assume that, for a given $a \in \mathbb{R}^+$,

$$\max_{c \in [C]} \left(\frac{\boldsymbol{\nu}_c}{\boldsymbol{\tau}_c \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \right) \leq \frac{a\, \sigma_\varepsilon}{\|\boldsymbol{\beta}^*\|_1} \ . \tag{15}$$

And, for all $c \in [C]$, let us define the adjusted p-values:

$$\hat{p}_c^{\mathcal{G}} = 2 \left(1 - \Phi \left(\sqrt{n} \left[\frac{|\hat{\boldsymbol{\theta}}_c|}{\sigma_\eta \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} - a \right]_+ \right) \right) \ . \tag{16}$$

Let us denote by $q_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ the $1-\frac{\alpha}{2}$ quantile of the standard Gaussian distribution. Then, under $H_0(G_c)$, the hypothesis which states that $\boldsymbol{\beta}^*_j = 0$ for $j \in G_c$ implying that $\tilde{\boldsymbol{\theta}}_c = 0$, we have, for any $\alpha \in (0,1)$:

$$\begin{aligned}
\mathbb{P}(\hat{p}_c^{\mathcal{G}} \leq \alpha \mid \mathbf{Z}) &= 1 - \mathbb{P}\left[\sqrt{n} \left[\frac{|\hat{\boldsymbol{\theta}}_c|}{\sigma_\eta \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} - a \right]_+ \leq q_{1-\frac{\alpha}{2}} \,\middle|\, \mathbf{Z} \right] \\
&\leq 1 - \mathbb{P}\left[\sqrt{n} \left[\frac{|\hat{\boldsymbol{\theta}}_c|}{\sigma_\eta \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} - \frac{\boldsymbol{\nu}_c \|\boldsymbol{\beta}^*\|_1}{\sigma_\varepsilon \boldsymbol{\tau}_c \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \right]_+ \leq q_{1-\frac{\alpha}{2}} \,\middle|\, \mathbf{Z} \right] \\
&\leq 1 - \mathbb{P}\left[\sqrt{n} \left[\frac{|\hat{\boldsymbol{\theta}}_c| - |\boldsymbol{\kappa}_c|}{\sigma_\eta \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \right]_+ \leq q_{1-\frac{\alpha}{2}} \,\middle|\, \mathbf{Z} \right] \\
&= 1 - \mathbb{P}\left[\sqrt{n} \left[\frac{|\hat{\boldsymbol{\theta}}_c| - |\boldsymbol{\theta}^*_c|}{\sigma_\eta \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \right]_+ \leq q_{1-\frac{\alpha}{2}} \,\middle|\, \mathbf{Z} \right] \\
&\leq 1 - \mathbb{P}\left[\sqrt{n} \frac{|\hat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}^*_c|}{\sigma_\eta \sqrt{\hat{\boldsymbol{\Omega}}_{c,c}}} \leq q_{1-\frac{\alpha}{2}} \,\middle|\, \mathbf{Z} \right] \\
&= \alpha + o_{\mathbb{P}}(1) \ .
\end{aligned} \tag{17}$$

Finally, we have built a cluster-wise adjusted p-value family that asymptotically exhibits, with low probability ($< \alpha$), low value ($< \alpha$) for the clusters which contain only zero weight covariates. To complete the proof in the case of correlated clusters, one can proceed as in uncorrelated cluster case taking (16) instead of (10).

Now, let us come back to the interpretation and choice for the constant a. In Prop. B.1, we have shown that, when groups are not independent, a group weight in the compressed model can be non-zero even if the group only contains zero weight covariates. However, the absolute value of the weight of such a group is necessarily upper bounded. We thus introduce $a \in \mathbb{R}^+$ in (16) to increase the p-values by a relevant amount and

keep statistical guarantees concerning the non-discovery of a such group. The value of a depends on the physics of the problem and on the choice of clustering. While the physics of the problem is fixed, the choice of clustering has a strong impact on the left term of (15) and a "good" choice of clustering results in a lower a (less correction). To estimate a, we need to find an upper bound of $\|\beta^*\|_1$, a lower bound of σ_ε and to estimate the left term of (15). In practice, to compute p-values, we took $a = 0$ since the formula in (10) was already conservative for all the problems we considered.

C EnCluDL

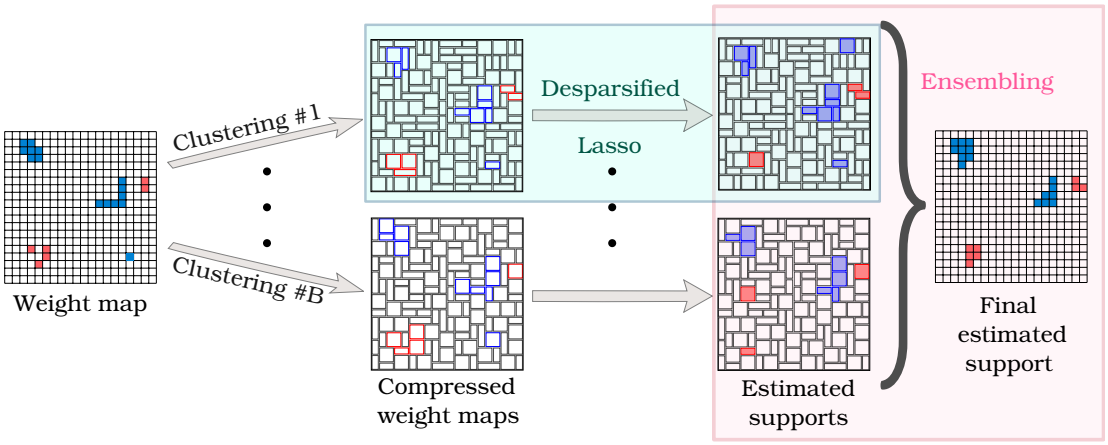

Figure 5: Summary of the mechanism of ensemble of clustered desparsified Lasso (EnCluDL). EnCluDL combines three algorithmic steps: a clustering procedure, the desparsified Lasso statistical inference procedure to derive p-value maps, and an ensembling method that synthesizes several p-value maps into one.

Computationally, to derive the EnCluDL solution we must solve B independent CluDL problems, making the global problem embarrassingly parallel; nevertheless, we could run the CluDL algorithm on standard desktop stations without parallelization with $n = 400$, $p \approx 10^5$, $C = 500$ and $B = 25$ in less than 10 minutes. Note that, the clustering step being much quicker than the inference step, p has a very limited impact on the total computation time.

The complexity for solving the Lasso depends significantly on the choice of solver, we then give the complexity in numbers of Lasso. The complexity for solving EnCluDL is given by the complexity of the resolution of $\mathcal{O}(B \times C)$ Lasso problems with n samples and C covariates, *i.e.,* with clustering. It is noteworthy that the complexity for solving the desparsified Lasso on the original problem is given by the complexity of the resolution of $\mathcal{O}(p)$ Lasso problems with n samples and p covariates, *i.e.,* without clustering. Then, EnCluDL should be much faster than the desparsified Lasso whenever $p \gg C$.

D Complementary simulation results

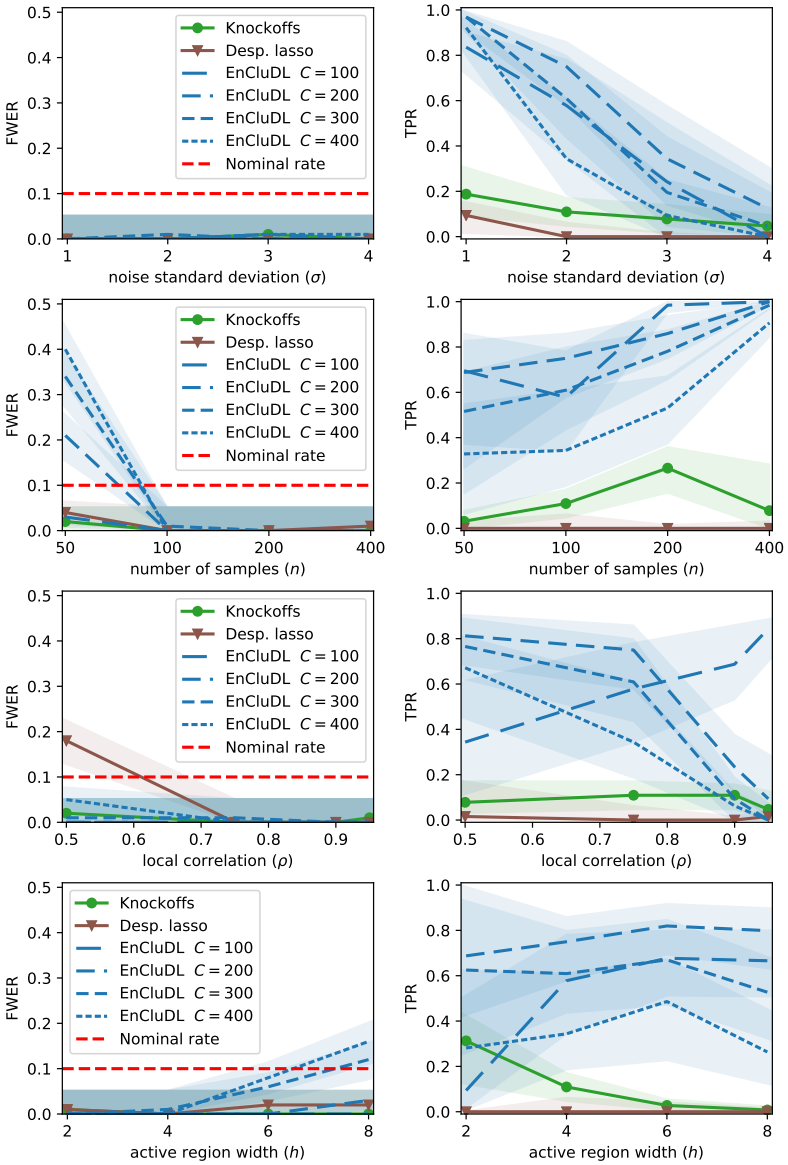

Figure 6: Results for various simulation parameters. The green line with circles correspond to the knockoffs, the brown line with triangle is the desparsified lasso, the dashed blue lines are for EnCluDL with length of the dashes increasing when C diminishes: large dashes are for $C = 100$, medium for $C = 200$, small for $C = 300$, tiny for $C = 400$. We compute the same FWER and TPR quantities as in Fig. 4, and the same 80% confidence intervals: by Binomial approximation for the FWER and taking first and last deciles for the TPR.

In Fig. 6, we study the influence of the simulation parameters by varying one parameter of the central scenario at a time. We vary the noise standard deviation, the number of samples, the local correlation and the size of the support. For a better readability of the figures, we do not analyze the results of CLuDL since it is expected to be always a bit less powerful than EnCluDL while showing a similar behavior. First, we look at the plots where we vary the noise standard deviation σ. We observe that the methods reach the targeted FWER control and notice that EnCluDL benefits more strongly from the decrease of σ regarding support recovery. Second, we analyze the results for various sample sizes (n) values. Concerning EnCluDL, we notice that the δ-FWER is not controlled when $n = 50$ except for $C = 100$. This is not surprising since the δ-FWER control is asymptotic and $n = 50$ is not sufficient. In terms of support recovery, the problem gets easier with larger n, but only EnCluDL benefits strongly from an increase of n. Third, we investigate the influence of the level of correlation between neighboring covariates (ρ). Regarding FWER control, desparsified lasso does not control the FWER when $\rho = 0.5$. Regarding the statistical power of EnCluDL, as one would expect, when the spatial structure is strong *i.e.,* $\rho > 0.9$, it is relevant to pick larger clusters, *i.e.,* to take a smaller C. Indeed, to make a relevant choice for C, data structure has to be taken into account to derive good covariates' clustering; this is true up to the limit given by the desired spatial tolerance. A good clustering creates clusters that are weakly correlated and contains covariates that are highly correlated. This observation is linked to assumption (ii) of Prop. 4.1 or to assumption $(ii.a)$ and $(ii.b)$ of Prop. B.1. Finally, we consider the results for different support sizes coded by the active region width h. Sparsity is a crucial assumption for desparsified lasso and then for EnCluDL. Also, when p (or C) increases the required sparsity is greater. This explains why when $h = 8$ and $C \geq 300$, the empirical δ-FWER is slightly above the expected nominal rate. Regarding the statistical power of EnCluDL, as one could expect, when the active regions are large, it is relevant to use large clusters. However, it can be difficult to estimate this parameter in advance, thus we prefer to consider desired spatial tolerance parameter δ and data structure to set C.

E Proofs

E.1 Proof of Prop. 4.1 and Prop. B.1

First, we start by the proof of Prop. 4.1 which is derived from Bühlmann et al. [2013, Proposition 4.3]:

Proof. With assumption (ii) and Bühlmann et al. [2013, Proposition 4.3], we have, for all $c \in [C]$:

$$\boldsymbol{\theta}_c^* = |G_c| \sum_{j \in G_c} w_j \boldsymbol{\beta}_j^* \ ,$$

where, for all $j \in G_c$:

$$w_j = \frac{\sum_{k \in G_c} \boldsymbol{\Sigma}_{j,k}}{\sum_{k \in G_c} \sum_{k' \in G_c} \boldsymbol{\Sigma}_{k,k'}} \ .$$

From assumption (i), we have $w_j > 0$ for all $j \in G_c$. Assumption (iii) ensures that, for all $j \in G_c$, the $\boldsymbol{\beta}^*_j$ have the same sign. Then, $\boldsymbol{\theta}^*_c$ is of the same sign as the $\boldsymbol{\beta}^*_j$ and is non-zero only if there exists $j \in G_c$ such that $\boldsymbol{\beta}^*_j \neq 0$. □

Now, we give the proof of Prop. B.1 which is mainly derived from Bühlmann et al. [2013, Proposition 4.4]:

Proof. With assumption $(ii.a)$ and $(ii.b)$ and Bühlmann et al. [2013, Proposition 4.4], we have, for all $c \in [C]$:

$$\boldsymbol{\theta}^*_c = |G_c| \sum_{j \in G_c} w'_j \boldsymbol{\beta}^*_j + \boldsymbol{\kappa}_c \ ,$$

where

$$w'_j = \frac{\sum_{k \in G_c} \mathrm{Cov}(\mathbf{X}_{.,j}, \mathbf{X}_{.,k} \mid \{\mathbf{Z}_{.,c'} : c' \neq c\})}{\sum_{k \in G_c} \sum_{k' \in G_c} \mathrm{Cov}(\mathbf{X}_{.,k}, \mathbf{X}_{.,k'} \mid \{\mathbf{Z}_{.,c'} : c' \neq c\})} \ ,$$

and, for all $c \in [C]$

$$|\boldsymbol{\kappa}_c| \leq (\boldsymbol{\nu}_c \ / \ \boldsymbol{\tau}_c) \|\boldsymbol{\beta}^*\|_1 \ .$$

Let us define $\tilde{\boldsymbol{\theta}}$ by

$$\tilde{\boldsymbol{\theta}}_c = |G_c| \sum_{j \in G_c} w'_j \boldsymbol{\beta}^*_j \ .$$

Then,

$$\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}} + \boldsymbol{\kappa} \ ,$$

And, similarly as in the proof of Prop. 4.1, from assumption (i) and (iii), $\tilde{\boldsymbol{\theta}}_c$ is of the same sign as the $\boldsymbol{\beta}^*_j$ for $j \in G_c$ and is non-zero only if there exists $j \in G_c$ such that $\boldsymbol{\beta}^*_j \neq 0$. □

E.2 Proof of Prop. 4.2

Before going trough the proof of Prop. 4.2, we introduce the grouping function g that matches the covariate index to its corresponding group index:

$$\begin{aligned} g : [p] &\to [C] \\ j &\mapsto c \quad \text{if } j \in G_c \ . \end{aligned}$$

Then, (7) can be rewritten as follows:

$$\begin{aligned} &\text{for all } j \in [p], \quad \hat{p}_j = \hat{p}^{\mathcal{G}}_{g(j)} \ , \\ &\text{for all } j \in [p], \quad \hat{q}_j = \hat{q}^{\mathcal{G}}_{g(j)} \ . \end{aligned} \tag{18}$$

Proof. *(i)* Suppose that we are under $H_0^\delta(j)$. Since the cluster diameters are all smaller than δ, all the covariates in $G_{g(j)}$ have a corresponding weight equal to zero. Thus, using Prop. 4.1, we have $\boldsymbol{\theta}^*_{g(j)} = 0$, *i.e.,* we are under $H_0(G_{g(j)})$. Under this last null-hypothesis, using (11) and (18), we have:

$$\text{for all } \alpha \in (0,1),\ \mathbb{P}(\hat{p}^{\mathcal{G}}_{g(j)} \leq \alpha) = \mathbb{P}(\hat{p}_j \leq \alpha) = \alpha \ .$$

This last result being true for any $j \in N^\delta$, we have shown that the elements of the family $\hat{p}$ control the δ-type 1 error.

(ii) As mentioned in Sec. 4.4, we know that, the family $\hat{q}^{\mathcal{G}}$ controls the FWER, *i.e.,* for $\alpha \in (0,1)$ we have $\mathbb{P}(\min_{c \in N_{\mathcal{G}}} \hat{q}^{\mathcal{G}}_c \leq \alpha) \leq \alpha$. Let us denote by $g^{-1}(N_{\mathcal{G}})$ the set of indexes of covariates that belong to the groups of $N_{\mathcal{G}}$, *i.e.,* $g^{-1}(N_{\mathcal{G}}) = \{j \in [p] : g(j) \in N_{\mathcal{G}}\}$. Again, given that all the cluster diameters are smaller than δ and using Prop. 4.1, if $j \in N^\delta$ then $g(j) \in N_{\mathcal{G}}$. That is to say $N^\delta \subset g^{-1}(N_{\mathcal{G}})$. Then, we have:

$$\min_{j \in N^\delta}(\hat{q}_j) \geq \min_{j \in g^{-1}(N_{\mathcal{G}})}(\hat{q}_j) \ .$$

We can also notice that:

$$\begin{aligned}\min_{j \in g^{-1}(N_{\mathcal{G}})}(\hat{q}_j) &= \min_{j \in g^{-1}(N_{\mathcal{G}})}(\hat{q}^{\mathcal{G}}_{g(j)}) \\ &= \min_{g(j) \in N_{\mathcal{G}}}(\hat{q}^{\mathcal{G}}_{g(j)}) \ .\end{aligned}$$

Replacing $g(j) \in [C]$ by $c \in [C]$, and using (6), we obtain:

$$\text{for all } \alpha \in (0,1),\ \mathbb{P}(\min_{j \in N^\delta}(\hat{q}_j) \leq \alpha) \leq \mathbb{P}(\min_{c \in N_{\mathcal{G}}} \hat{q}^{\mathcal{G}}_c \leq \alpha) \leq \alpha \ .$$

This last result states that the family $(\hat{q}_j)_{j \in [p]}$ controls the δ-FWER. □

E.3 Proof of Prop. 4.3

The proof of Prop. 4.3 is inspired by the one proposed by Meinshausen et al. [2009]. However, it is subtly different since we can not remove the term $\min_{j \in N^\delta}$ and have to work with it to obtained the desired inequality. First, we start by making a short remark about the γ-quantile quantity.

Definition E.1 (empirical γ-quantile)**.** *For a set V of real numbers and $\gamma \in (0,1)$, let*

$$\gamma\text{-}quantile(V) = \min\left\{v \in V : \frac{1}{|V|}\sum_{w \in V} \mathbb{1}_{w \leq v} \geq \gamma\right\} \ . \tag{19}$$

Remark E.1. *For a set of real number V and for $a \in \mathbb{R}$, let us define the quantity $\pi(a, V)$ by the following:*

$$\pi(a, V) = \frac{1}{|V|}\sum_{v \in V} \mathbb{1}\left(v \leq a\right) \tag{20}$$

Then, for $\gamma \in (0,1)$, the two events $E_1 = \{\pi(a, V) \geq \gamma\}$ and $E_2 = \{\gamma\text{-}quantile(V) \leq a\}$ are identical.

Now, we give the proof of Prop. 4.3.

Proof. First, one can notice that, from (8), we have:

$$\min_{j\in N^\delta}(\tilde{q}_j(\gamma)) \geq \min\left\{1, \gamma\text{-quantile}\left(\left\{\min_{j\in N^\delta}\left(\frac{\hat{q}_j^{(b)}}{\gamma}\right) : b\in[B]\right\}\right)\right\} \ .$$

Then, for $\alpha \in (0,1)$:

$$\begin{aligned}\mathbb{P}\left(\min_{j\in N^\delta}(\tilde{q}_j(\gamma)) \leq \alpha\right) &\leq \mathbb{P}\left(\min\left\{1, \gamma\text{-quantile}\left(\left\{\min_{j\in N^\delta}\left(\frac{\hat{q}_j^{(b)}}{\gamma}\right) : b\in[B]\right\}\right)\right\} \leq \alpha\right) \\ &= \mathbb{P}\left(\gamma\text{-quantile}\left(\left\{\min_{j\in N^\delta}\left(\frac{\hat{q}_j^{(b)}}{\gamma}\right) : b\in[B]\right\}\right) \leq \alpha\right) \ .\end{aligned}$$

Using Rem. E.1, for $\gamma \in (0,1)$, with:

$$V = \left\{\min_{j\in N^\delta}\left(\frac{\hat{q}_j^{(b)}}{\gamma}\right) : b\in[B]\right\} \quad \text{and} \quad a = \alpha \ ,$$

and noticing that:

$$\pi\left(\alpha, \left\{\min_{j\in N^\delta}\left(\frac{\hat{q}_j^{(b)}}{\gamma}\right) : b\in[B]\right\}\right) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{\min_{j\in N^\delta}(\hat{q}_j^{(b)}) \leq \alpha\gamma\right\} \ ,$$

then, we have:

$$\mathbb{P}\left(\gamma\text{-quantile}\left(\left\{\min_{j\in N^\delta}\left(\frac{\hat{q}_j^{(b)}}{\gamma}\right) : b\in[B]\right\}\right) \leq \alpha\right) = \mathbb{P}\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{\min_{j\in N^\delta}(\hat{q}_j^{(b)}) \leq \alpha\gamma\right\} \geq \gamma\right).$$

Then, the Markov inequality gives:

$$\mathbb{P}\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{\min_{j\in N^\delta}(\hat{q}_j^{(b)}) \leq \alpha\gamma\right\} \geq \gamma\right) \leq \frac{1}{\gamma}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{\min_{j\in N^\delta}(\hat{q}_j^{(b)}) \leq \alpha\gamma\right\}\right] \ .$$

Then, using the assumption that the B families $(\hat{q}_j^{(b)})_{j\in[p]}$ control of the δ-FWER (last inequality), we have:

$$\frac{1}{\gamma}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{\min_{j\in N^\delta}(\hat{q}_j^{(b)}) \leq \alpha\gamma\right\}\right] = \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}\left(\min_{j\in N^\delta}(\hat{q}_j^{(b)}) \leq \alpha\gamma\right) \leq \alpha \ .$$

Finally, we have shown that, for $\alpha \in (0,1)$:

$$\mathbb{P}\left(\min_{j\in N^\delta}(\tilde{q}_j(\gamma)) \leq \alpha\right) \leq \alpha \ .$$

This establishes that the family $(\tilde{q}_j(\gamma))_{j\in[p]}$ controls the δ-FWER. □

E.4 Proof of Theorem 4.1

To show Theorem 4.1, we connect the previous results: Prop. 4.1, Prop. 4.2 and Prop. 4.3. First, we prove that clustered inference algorithms produce a p-value family that controls the δ-FWER.

Proof. Assuming the noise model (1), assuming that Ass. 2.1 and Ass. 2.2 are verified for a distance parameter larger than δ and that the clustering diameter is smaller than δ, then we directly obtain the assumption (i) and (iii) of Prop. 4.1. This means that the compressed representation has the correct pattern of non-zero coefficients, in particular it does not include in the support clusters of null-only covariates. Additionally, if one is able to perform a valid statistical inference on the compressed model (3), *i.e.,* to produce cluster-wise p-values such that (4) holds, then Prop. 4.2 ensures that the p-value family constructed using the de-grouping method presented in (7) controls the δ-FWER. □

Now, we prove that ensembled clustered inference algorithms produce a p-value family that controls the δ-FWER.

Proof. Given the above arguments, the p-value families produced by clustered inference algorithms subject to all clusterings fulfilling the theorem hypotheses control the δ-FWER. Then, using the aggregation method given by (8), we know from Prop. 4.3 that the aggregated p-value family also controls the δ-FWER. □