Selective inference after convex clustering with ℓ_1 penalization

François Bachoc, Cathy Maugis, Pierre Neuvial

Selective inference after convex clustering with ℓ_1 penalization

François Bachoc[⋆], **Cathy Maugis-Rabusseau**[◦], **and Pierre Neuvial**[•]

[⋆] *Institut de Mathématiques de Toulouse; UMR5219*
Université de Toulouse; CNRS
UT3, F-31062 Toulouse

[◦] *Institut de Mathématiques de Toulouse; UMR5219*
Université de Toulouse; CNRS
INSA, F-31077 Toulouse, France

[•] *Institut de Mathématiques de Toulouse; UMR5219*
Université de Toulouse; CNRS

Abstract: Classical inference methods notoriously fail when applied to data-driven test hypotheses or inference targets. Instead, dedicated methodologies are required to obtain statistical guarantees for these selective inference problems. Selective inference is particularly relevant post-clustering, typically when testing a difference in mean between two clusters. In this paper, we address convex clustering with ℓ_1 penalization, by leveraging related selective inference tools for regression, based on Gaussian vectors conditioned to polyhedral sets. In the one-dimensional case, we prove a polyhedral characterization of obtaining given clusters, than enables us to suggest a test procedure with statistical guarantees. This characterization also allows us to provide a computationally efficient regularization path algorithm. Then, we extend the above test procedure and guarantees to multi-dimensional clustering with ℓ_1 penalization, and also to more general multi-dimensional clusterings that aggregate one-dimensional ones. With various numerical experiments, we validate our statistical guarantees and we demonstrate the power of our methods to detect differences in mean between clusters. Our methods are implemented in the `R` package `poclin`.

MSC 2010 subject classifications: Primary: 62F03, 62H30.
Keywords and phrases: Selective inference, clustering, regularization path, hypothesis test, truncated Gaussian.

1. Context and objectives

The problem of **selective inference** occurs when the same dataset is used (i) to detect a statistical signal and (ii) to evaluate the strength of this signal [27]. In this article, we focus on the problem of post-clustering testing, where step (i) corresponds to a clustering of the input data, and step (ii) to an hypothesis test stemming from the clustering step. In such a situation, the naive application of a test that does not account for the data-driven clustering step is bound to violate type I error control [6].

This problem occurs in several applications. For instance, it is well-identified in the analysis of single-cell RNA-seq data (see [12]) where the genes expression is measured for several cells: we want to test if each gene has a differential expression between two cells clusters, which are determined beforehand with a clustering procedure on the same expression matrix. This practical question has motivated numerous recent statistical developments to address this post-clustering testing problem.

A data splitting strategy has been studied by [36], but the assignment of labels (from the clustering of the first sample) to the second sample before the test procedure is not taken into

account in the correction. A conditional testing approach has been proposed by [6] for the problem of the difference in mean between two clusters. The authors condition by the event "the two compared clusters are obtained by the random clustering" and by an additional one, allowing p-values to be exactly computed in the case of agglomerative hierarchical clustering. This approach has been extended to the test of the difference in mean between two clusters for each fixed variable [9]. A strategy to aggregate these p-values, and another approach using tests of multimodality (without statistical guarantees) are also suggested in [9]. In the context of single-cell data analysis, a count splitting approach under a Poisson assumption [19] and a more flexible Negative Binomial assumption [20] have recently been proposed. In the same line of work, a data thinning strategy is explored in [5, 18], that consists in generating two (or more) independent random matrices that sum to the initial data matrix. This idea can be applied to various distributions belonging to the exponential family.

The present contribution takes a different route from the above references and builds on [14], where a Gaussian linear model is considered, and test procedures are provided, together with associated guarantees post-selection of variables based on the Lasso. The nature of the Lasso optimization problem is carefully analyzed in [14], and conditionally valid test procedures are obtained, based on properties of Gaussian vectors conditioned to polyhedral sets.

We will extend this approach and its statistical guarantees to clustering procedures based on solving a convex optimization problem with ℓ_1 penalization.

Let us now describe the setting of the paper in more details. We observe, for n observations of p variables (or features), a matrix $\boldsymbol{Y} = (Y_{ij})_{i\in[|n|],j\in[|p|]}$, where $[|u|] := \{1,\dots,u\}$ for any positive integer u. We assume that $\mathrm{vec}(\boldsymbol{Y})$ is a np-dimensional Gaussian vector with mean vector $\boldsymbol{\beta}$ and $np \times np$ covariance matrix $\boldsymbol{\Gamma}$, where vec(.) denotes the vectorization by column of a matrix. The vector $\boldsymbol{\beta}$ is unknown but the matrix $\boldsymbol{\Gamma}$ is assumed to be known (as in several of the articles cited above, we will discuss this hypothesis in Section 4.3). Note that this setup covers in particular the case considered e.g. in [6], where $\boldsymbol{Y}$ follows the matrix normal distribution $\mathcal{MN}_{n\times p}(\mathbf{u}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ where $\mathbf{u}$ is the $n \times p$ mean matrix, $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix among rows and $\boldsymbol{\Delta}$ is the $p \times p$ covariance matrix among variables. Indeed, this matrix normal setup is equivalent (by definition) to that $\mathrm{vec}(\boldsymbol{Y})$ is a np-dimensional Gaussian vector with mean vector $\boldsymbol{\beta} := \mathrm{vec}(\mathbf{u})$ and $np \times np$ covariance matrix $\boldsymbol{\Gamma} = \boldsymbol{\Delta} \otimes \boldsymbol{\Sigma}$, where $\otimes$ denotes the Kronecker product.

Under this framework, as announced, we will develop test procedures that extend the line of analysis of [14] to a clustering counterpart of the Lasso in linear models. Thus we consider the convex clustering problem [10, 15, 24] which consists in solving the following optimization problem

$$\widehat{\boldsymbol{B}}(\boldsymbol{Y}) \in \underset{\boldsymbol{B}=(\boldsymbol{B}_{1.}^\top,\dots,\boldsymbol{B}_{n.}^\top)^\top\in\mathbb{R}^{n\times p}}{\operatorname{argmin}} \ \frac{1}{2}||\boldsymbol{B}-\boldsymbol{Y}||_F^2 + \lambda \sum_{\substack{i,i'=1\\ i<i'}}^{n} ||\boldsymbol{B}_{i'.} - \boldsymbol{B}_{i.}||_1 \tag{1}$$

where $||\cdot||_F$ is the Frobenius norm and $\boldsymbol{B}_{i.}$ denotes the i-th row of $\boldsymbol{B}$. The quantity $\lambda > 0$ is a tuning parameter that we consider fixed here (as for the covariance matrix $\boldsymbol{\Gamma}$, this assumption is further discussed in Section 4.3). We can immediately notice that Problem (1) is separable, and can be solved by addressing, for $j \in [|p|]$, the one-dimensional problem

$$\widehat{\boldsymbol{B}}_{.j}(\boldsymbol{Y}_{.j}) \in \underset{\boldsymbol{B}_{.j}=(B_{1j},\dots,B_{nj})^\top\in\mathbb{R}^{n}}{\operatorname{argmin}} \ \frac{1}{2}||\boldsymbol{B}_{.j}-\boldsymbol{Y}_{.j}||_2^2 + \lambda \sum_{\substack{i,i'=1\\ i<i'}}^{n} |B_{i'j} - B_{ij}|, \tag{2}$$

where $\boldsymbol{B}_{\cdot j}$ is the j-th column of $\boldsymbol{B}$. It is worth pointing out that if the norm $\|\cdot\|_1$ is replaced by another norm $\|\cdot\|_q$, $q \in (0,\infty)\backslash\{1\}$ in (1), then the optimization problem is no longer separable. Hence, it becomes more challenging from a computational perspective. This topic has been the object of a fair amount of recent work, see [4, 25, 31, 33, 37] and our discussions at the end of Section 2.4 and in Section 4.4.

The solution $\widehat{\boldsymbol{B}}_{\cdot j}(\boldsymbol{Y}_{\cdot j})$ of (2) naturally provides a one-dimensional clustering $\mathcal{C}^{(j)}$ of the observations for the variable j, by affecting i and i' to the same cluster if and only if $\widehat{\boldsymbol{B}}_{ij} = \widehat{\boldsymbol{B}}_{i'j}$. Similarly, the solution of (1) provided by the matrix $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{B}}_{\cdot 1}, \dots, \hat{\boldsymbol{B}}_{\cdot p})$ naturally yields a multi-dimensional clustering of the observations, by affecting i and i' to the same cluster if and only if $\widehat{\boldsymbol{B}}_{i\cdot} = \widehat{\boldsymbol{B}}_{i'\cdot}$. In this article, we will consider more general multi-dimensional clusterings that can be obtained by aggregation of the one-dimensional clusterings $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(p)}$ (see Section 3.1). A clustering of the rows of $\boldsymbol{Y}$ in K clusters will be denoted by $\mathcal{C} = \mathcal{C}(\boldsymbol{Y}) = (\mathcal{C}_1(\boldsymbol{Y}), \dots, \mathcal{C}_K(\boldsymbol{Y}))$. Of course these clusters and the number of clusters K are random (depending on $\boldsymbol{Y}$).

Our goal is to provide test procedures for a (data-dependent) hypothesis of the form

$$\boldsymbol{\kappa}^\top \boldsymbol{\beta} = 0,$$

where $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\mathcal{C}(\boldsymbol{Y}))$ is a deterministic function of the clustering $\mathcal{C}(\boldsymbol{Y})$ and where we recall that $\boldsymbol{\beta}$ is the $np \times 1$ mean vector of $\mathrm{vec}(\boldsymbol{Y})$. We refer to Section 4.2 for further discussions on the merits and interpretations of the tests considered in this paper.

Example 1 (feature-level two-group test)**.** *The following typical example of a choice of $\boldsymbol{\kappa}$ enables to compare, for a variable $j_0 \in [|p|]$, the average signal difference between two clusters $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$, $k_1, k_2 \in [|K|]$, $k_1 \neq k_2$. We write, for $i \in [|n|]$ and $j \in [|p|]$,*

$$\boldsymbol{\kappa}_{i+(j-1)n} = \left(\frac{\mathbb{1}_{i \in \mathcal{C}_{k_1}}}{|\mathcal{C}_{k_1}|} - \frac{\mathbb{1}_{i \in \mathcal{C}_{k_2}}}{|\mathcal{C}_{k_2}|} \right) \mathbb{1}_{j=j_0}, \tag{3}$$

where $|A|$ denotes the cardinality of any finite set A. This yields

$$\boldsymbol{\kappa}^\top \boldsymbol{\beta} = \frac{1}{|\mathcal{C}_{k_1}|} \sum_{i \in \mathcal{C}_{k_1}} \beta_{i+(j_0-1)n} - \frac{1}{|\mathcal{C}_{k_2}|} \sum_{i \in \mathcal{C}_{k_2}} \beta_{i+(j_0-1)n}. \tag{4}$$

In the particular matrix normal setup discussed above,

$$\boldsymbol{\kappa}^\top \boldsymbol{\beta} = \frac{1}{|\mathcal{C}_{k_1}|} \sum_{i \in \mathcal{C}_{k_1}} \mathbf{u}_{i,j_0} - \frac{1}{|\mathcal{C}_{k_2}|} \sum_{i \in \mathcal{C}_{k_2}} \mathbf{u}_{i,j_0}.$$

Rejecting this hypothesis corresponds to deciding that the clusters $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$ have a discriminative power for the variable j_0, since their average signal indeed differs.

The separation of Problem (1) into p one-dimensional optimization problems in (2) will be key for the testing procedures we develop in this paper. In Section 2, we will thus develop our methodology and theory related to the one-dimensional Problem (2). A test procedure is proposed and its statistical guarantees are established in Section 2.3. In Section 2.4, a discussion of the existing optimization procedures to solve Problem (2) is given and an original regularization path algorithm is also provided, specifically for this problem (obtained by leveraging our theoretical results in Section 2.2). In Section 3, the proposed test procedure

and its guarantees are extended to the p-dimensional framework. Numerical experiments are presented in Sections 2.5 for $p = 1$ and 3.3 for $p > 1$. In Section 4, we provide a detailed overview of our contributions, together with various conclusive discussions regarding them and remaining open problems. The proofs are postponed to Appendices A to C. Appendix D contains additional material regarding the computational aspects of convex clustering, in particular with our suggested regularization path. Appendix E contains additional numerical illustrations.

2. The one-dimensional case

2.1. Setting and notation

In this section, for notational simplification, we consider a single Gaussian vector $\boldsymbol{X}$ of size $n \times 1$, with unknown mean vector $\boldsymbol{\mu}$ and known covariance matrix $\boldsymbol{\Sigma}$. This vector $\boldsymbol{X}$ should be thought of as an instance of $\boldsymbol{Y}_{\cdot j}$ in (2) for some fixed $j \in [|p|]$.

We consider the convex clustering procedure (as Problem (2)) obtained for a given $\lambda > 0$ by

$$\widehat{\boldsymbol{B}}(\boldsymbol{X}) \in \underset{\boldsymbol{B}=(B_1,\dots,B_n)\in\mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}||\boldsymbol{B}-\boldsymbol{X}||_2^2 + \lambda \sum_{\substack{i,i'=1 \\ i<i'}}^{n} |B_{i'} - B_i|. \tag{5}$$

Solving this optimization problem defines a clustering of the n observations, each cluster corresponding to a distinct value of $\widehat{\boldsymbol{B}}(\boldsymbol{X})$. This mapping is formalized by the following definition.

Definition 1. *For $\boldsymbol{B} = (B_1, \dots, B_n) \in \mathbb{R}^n$, let $b_1 > b_2 > \cdots > b_K$ be the sorted distinct values of the set $\{B_i : i \in [|n|]\}$. The clustering associated to $\boldsymbol{B}$ is $\mathcal{C} = (\mathcal{C}_k)_{k\in[|K|]}$, where $\mathcal{C}_k = \{i : B_i = b_k\}$ for $k \in [|K|]$.*

Note that, indifferently, we address clusterings of a set of elements $(x_1, \dots, x_n)$ (for instance scalars or vectors) either with clusters that are subsets of $(x_1, \dots, x_n)$ or subsets of $[|n|]$. It is convenient to point out the following basic property of the optimization of Problem (5), implying in particular that the clusters are composed by successive scalar observed values, which is very natural.

Lemma 1. *Consider a fixed $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$. Consider $\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{B}}(\boldsymbol{x})$ given by Problem (5). Then, for $i, i' \in [|n|], i \neq i'$,*

1. *$x_i = x_{i'}$ implies $\widehat{B}_i = \widehat{B}_{i'}$*
2. *$x_i \geq x_{i'}$ implies $\widehat{B}_i \geq \widehat{B}_{i'}$.*

Similarly as discussed in Section 1, for the clustering $\mathcal{C} = \mathcal{C}(\boldsymbol{X}) = (\mathcal{C}_1(\boldsymbol{X}), \dots, \mathcal{C}_K(\boldsymbol{X}))$ obtained from (5), we will provide a valid test procedure for an hypothesis of the form $\boldsymbol{\eta}^\top \boldsymbol{\mu} = 0$, where $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathcal{C}(\boldsymbol{X}))$.

2.2. Polyhedral characterization of convex clustering in dimension one

As in [14], we will suggest a test procedure (see Section 2.3) based on analyzing Gaussian vectors conditioned to polyhedral sets. At first sight, one could thus aim at showing that the

observation vector $\boldsymbol{X}$ yields a given clustering with (5) if and only if it belongs to a corresponding polyhedral set. However, this does not hold in general. Hence, we will characterize a more restricted event with a polyhedral set. This event is that (i) a given clustering is obtained and (ii) the scalar observations are in a given order. The same phenomenon occurs in [14], where variables are selected in a linear model. There, it does not hold that a given set of variables is selected by the Lasso if and only if the observation vector belongs to a given polyhedral set. Nevertheless, the event that can be characterized with a polyhedral set is that (i) a given set of variables is selected and (ii) the signs of the estimated coefficients for these variables take a given sequence of values. We refer to Section 4.6 for further discussion on conditioning also by the observations' order.

Before stating the polyhedral characterization, let us provide some notation. We let $\mathfrak{S}_n$ be the set of permutations of $[|n|]$. Consider observations $x_1, \ldots, x_n$, ordered as $x_{\sigma(1)} \geq \cdots \geq x_{\sigma(n)}$ for $\sigma \in \mathfrak{S}_n$. When these observations are clustered into K clusters of successive values, the clustering is in one-to-one correspondence with the positions of the cluster right-limits $t_1, \ldots, t_K$, where $0 = t_0 < t_1 \ldots < t_K = n$, and where for $k \in [|K|]$, cluster $\mathcal{C}_k$ is composed by the indices $\sigma(t_{k-1}+1), \ldots, \sigma(t_k)$, for $k \in [|K|]$. This corresponds to the following definition.

Definition 2. *For $n \in \mathbb{N}$ and $K \in [|n|]$, let*

$$\mathcal{T}_{K,n} := \{(t_k)_{0 \leq k \leq K};\ 0 = t_0 < t_1 < \cdots < t_K = n\}\,.$$

For any $\sigma \in \mathfrak{S}_n$ and any vector $\mathbf{t} \in \mathcal{T}_{K,n}$, the clustering associated to $(\mathbf{t}, \sigma)$ is defined as $\mathcal{C}(\mathbf{t}, \sigma) = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, where for $k \in [|K|]$, $n_k = t_k - t_{k-1}$ and $\mathcal{C}_k = \{\sigma(t_{k-1}+i)\}_{i \in [|n_k|]}$.

In particular, let us consider the clustering $\mathcal{C} = (\mathcal{C}_k)_{k \in [|K|]}$ obtained from Definition 1 by solving Problem (5) for a given $\boldsymbol{x} \in \mathbb{R}^n$. This clustering can be written as $\mathcal{C}(\mathbf{t}, \sigma)$, for any σ such that $x_{\sigma(1)} \geq \cdots \geq x_{\sigma(n)}$, $t_0 = 0$ and $t_k = \sum_{j \in [|k|]} |\mathcal{C}_j|$ for $k \in [|K|]$.

Example 2. *To illustrate Definition 2 and Lemma 1, let $\boldsymbol{x} = (2, 6, 11, 10, 7, 1, 6.5, 7)$ be observed data. A permutation reordering the values of $\boldsymbol{x}$ by decreasing order is*

$$\sigma : (1, \ldots, n = 8) \mapsto (3, 4, 5, 8, 7, 2, 1, 6).$$

For the clustering $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3)$ with $\mathcal{C}_1 = \{11, 10\}$, $\mathcal{C}_2 = \{7, 7, 6.5, 6\}$ and $\mathcal{C}_3 = \{2, 1\}$, the associated vector $\mathbf{t}$ is $t_0 = 0$, $t_1 = 2$, $t_2 = 6$ and $t_3 = 8$, as shown in Figure 1. Note that the clustering $\mathcal{C}$ of observations is equivalent to the clustering of indices $\mathcal{C}_1 = \{\sigma(1), \sigma(2)\} = \{3, 4\}$, $\mathcal{C}_2 = \{\sigma(3), \sigma(4), \sigma(5), \sigma(6)\} = \{5, 8, 7, 2\}$ and $\mathcal{C}_3 = \{\sigma(7), \sigma(8)\} = \{1, 6\}$. The regularization path (see Section 2.4) associated to the convex clustering problem on the observed values $\boldsymbol{x}$ is represented in Figure 2. The vertical line at $x = \lambda$ intersects the regularization path at $y = \hat{B}_i$. The order property between x_i and $\hat{B}_i$ stated in Lemma 1 is observed all along the regularization path. For $\lambda = 0.5$, we find the clustering in three clusters where the $\hat{B}_i$ values take three distinct values $\hat{b}_k$ ($\hat{b}_1 = 7.5$, $\hat{b}_2 = 6.625$ and $\hat{b}_3 = 4.5$).

Next, we can provide the announced polyhedral characterization of obtaining a given clustering, together with a given order of the observations.

Theorem 2. *Let $\mathbf{t}$ be a fixed vector in $\mathcal{T}_{K,n}$ with $K \in [|n|]$, and let $\sigma \in \mathfrak{S}_n$ be a fixed permutation of $[|n|]$. Let $\mathcal{C} = \mathcal{C}(\mathbf{t}, \sigma)$ be the clustering obtained from Definition 2, with cluster cardinalities $n_1, \ldots, n_K$. Consider a fixed $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Let $\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{B}}(\boldsymbol{x})$ be the solution of Problem (5) for some fixed $\lambda > 0$, with $\boldsymbol{X}$ replaced by $\boldsymbol{x}$. From Definition 1, $\widehat{\boldsymbol{B}}$*

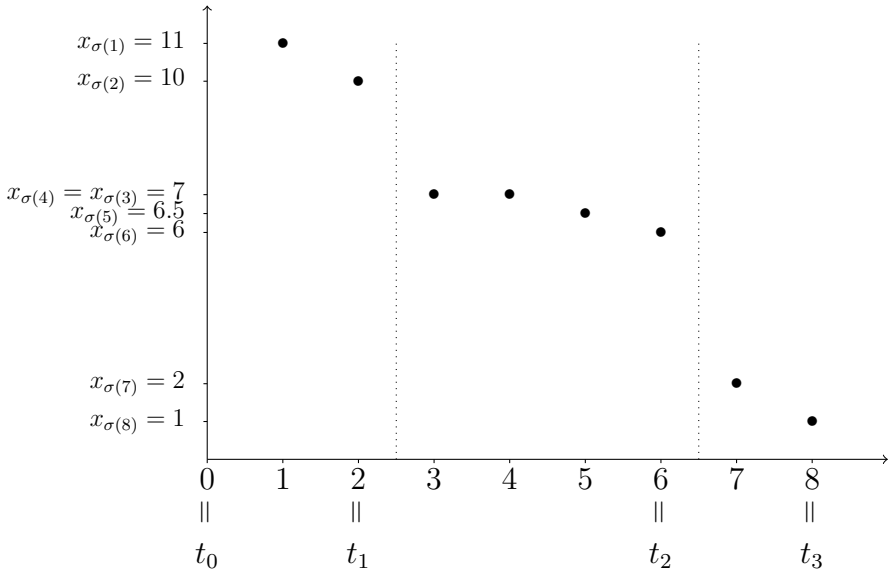

Figure 1: Illustration of Definition 2 for one clustering with $K = 3$ clusters of the observed values $\boldsymbol{x} = (2, 6, 11, 10, 7, 1, 6.5, 7)$

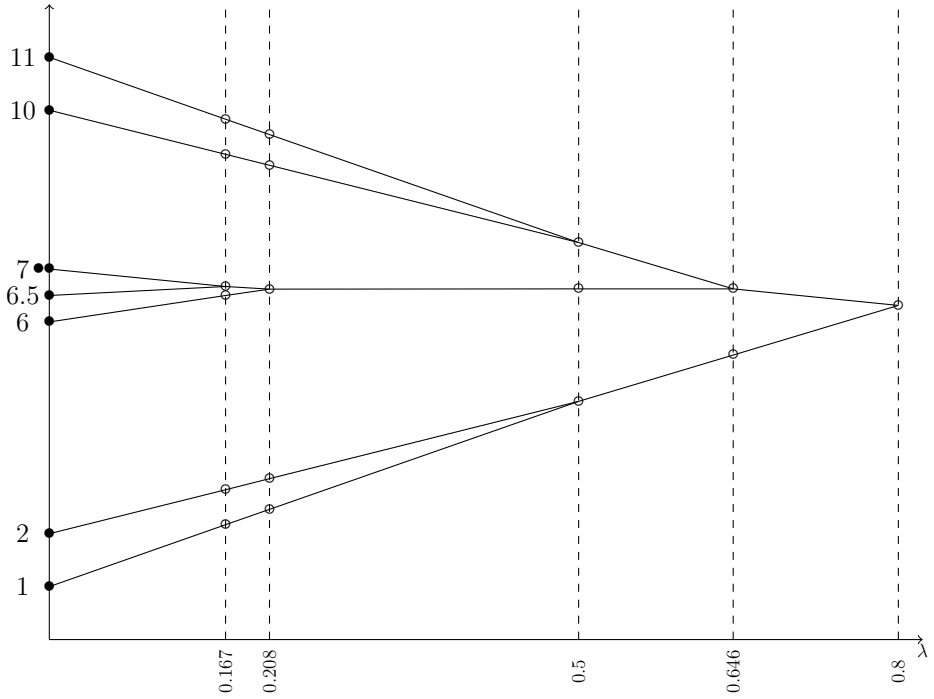

Figure 2: Regularization path (see Section 2.4) associated to the convex clustering problem for the observed values $\boldsymbol{x} = (2, 6, 11, 10, 7, 1, 6.5, 7)$.

yields a clustering. Then the set of conditions

$$\mathcal{C}(\mathbf{t}, \sigma) \text{ is the clustering given by } \widehat{\boldsymbol{B}}, \tag{6}$$
$$x_{\sigma(1)} \geq x_{\sigma(2)} \geq \cdots \geq x_{\sigma(n)} \tag{7}$$

is equivalent to the set of the three following conditions

$$\text{for } k \in [|K-1|]:$$
$$\frac{1}{n_k} \sum_{i=1}^{n_k} x_{\sigma(t_{k-1}+i)} - \frac{1}{n_{k+1}} \sum_{i=1}^{n_{k+1}} x_{\sigma(t_k+i)} > \lambda(t_{k+1} - t_{k-1}), \tag{8}$$
$$\text{for } k \in [|K|] \text{ such that } n_k \geq 2, \text{ for } \ell \in [|n_k - 1|]:$$
$$\frac{1}{n_k} \sum_{i=1}^{n_k} x_{\sigma(t_{k-1}+i)} - \frac{1}{\ell} \sum_{i=1}^{\ell} x_{\sigma(t_{k-1}+i)} \geq \lambda(\ell - n_k), \tag{9}$$
$$x_{\sigma(1)} \geq x_{\sigma(2)} \geq \cdots \geq x_{\sigma(n)}. \tag{10}$$

Finally, when (6) and (7) hold, then for $i \in [|n|]$, for $k \in [|K|]$ with $i \in \mathcal{C}_k$, we have

$$\hat{B}_i = \frac{1}{n_k} \sum_{i' \in \mathcal{C}_k} x_{i'} + \lambda \sum_{k'=1}^{k-1} n_{k'} - \lambda \sum_{k'=k+1}^{K} n_{k'}. \tag{11}$$

In (11), note that by convention $\sum_{k'=a}^{b} \cdots = 0$ for $a, b \in \mathbb{Z}$, $a > b$. We will use this convention in the rest of the paper. Note also that, apart from the polyhedral characterization given by (8) to (10), Theorem 2 also provides the explicit expression of the optimal $\widehat{\boldsymbol{B}}$, solution of Problem (5). This expression depends of the optimal clustering, so it cannot be directly computed to optimize (5) in practice. Nevertheless, Theorem 2 is the basis of a regularization path algorithm provided in Section 2.4.

Next, the following lemma provides a formulation of (8) to (10) in Theorem 2 as an explicit polyhedral set. In this lemma and in the rest of the paper, for $a \in \mathbb{N}$, we let $\mathbf{0}_a$ be the $a \times 1$ vector composed of zeros.

Lemma 3. *Consider the setting of Theorem 2. Let $\mathbf{P}_\sigma$ be the $n \times n$ permutation matrix associated to $\sigma \in \mathfrak{S}_n$: $\mathbf{P}_\sigma \boldsymbol{x} = (\boldsymbol{x}_{\sigma(1)}, \ldots, \boldsymbol{x}_{\sigma(n)})^\top$, for a $n \times 1$ vector $\boldsymbol{x}$. Then, Conditions (8), (9) and (10) can be written as*

$$\{\mathbf{M}(\mathbf{t})\mathbf{P}_\sigma \boldsymbol{x} \leq \lambda\ \mathbf{m}(\mathbf{t})\} \tag{12}$$

where $\mathbf{M}(\mathbf{t}) \in \mathbb{R}^{2(n-1)\times n}$ and $\mathbf{m}(\mathbf{t}) \in \mathbb{R}^{2(n-1)}$ are given by:

$$\mathbf{M}(\mathbf{t}) = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2(\mathbf{t}) \\ \mathbf{M}_3(\mathbf{t}) \end{pmatrix} \text{ and } \mathbf{m}(\mathbf{t}) = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2(\mathbf{t}) \\ \mathbf{m}_3(\mathbf{t}) \end{pmatrix},$$

with $\mathbf{M}_1 \in \mathbb{R}^{n-1\times n}$, $\mathbf{M}_2(\mathbf{t}) \in \mathbb{R}^{K-1\times n}$ and $\mathbf{M}_3(\mathbf{t}) \in \mathbb{R}^{n-K\times n}$, explicitly expressed in Appendix B (Equations (25), (27) and (29) respectively); $\mathbf{m}_1 = \mathbf{0}_{n-1}$, $\mathbf{m}_2(\mathbf{t}) \in \mathbb{R}^{K-1}$ and $\mathbf{m}_3(\mathbf{t}) \in \mathbb{R}^{n-K}$, explicitly expressed in Appendix B (Equations (26) and (28) respectively). Furthermore, the inequality $\mathbf{M}_2(\mathbf{t})\mathbf{P}_\sigma \boldsymbol{x} \leq \lambda \mathbf{m}_2(\mathbf{t})$ is strict in (12).

2.3. Test procedure and its guarantees

In this section, we construct the test procedure and provide its theoretical guarantees, based on Theorem 2 and Lemma 3. Since the polyhedral characterization has been shown from these two results, the construction and guarantees here are obtained similarly as in [14]. We nevertheless provide the full details, for the sake of self-completeness.

2.3.1. Construction of the test procedure

We want to test

$$\boldsymbol{\eta}^\top \boldsymbol{\mu} = 0,$$

where $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathcal{C}(\boldsymbol{X}))$ and $\mathcal{C}(\boldsymbol{X})$ is obtained from Problem (5) and Definition 1. The test statistic is naturally

$$\boldsymbol{\eta}^\top \boldsymbol{X},$$

and we will construct an invariant statistic from it, based on the polyhedral lemma (Lemma 5.1) of [14], that we restate in our setting for convenience. In the next statement, $\boldsymbol{I}_a$ is the identity matrix in dimension $a \in \mathbb{N}$ and we use the conventions that the minimum over an empty set is $+\infty$ and the maximum over an empty set is $-\infty$.

Proposition 4 (Polyhedral lemma, adapted from [14])**.** *Let $\mathbf{t}$ be a fixed vector in $\mathcal{T}_{K,n}$ with $K \in [|n|]$. Let $\sigma \in \mathfrak{S}_n$ be a fixed permutation of $[|n|]$, and $\mathbf{P}_\sigma$ be the $n \times n$ associated permutation matrix.*

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ invertible and let $\boldsymbol{\eta}$ be a fixed non-zero $n \times 1$ vector (allowed to depend on $\mathbf{t}$ and σ).
Let $\boldsymbol{Z} := \boldsymbol{Z}(\boldsymbol{X}) := [\boldsymbol{I}_n - \boldsymbol{c}\boldsymbol{\eta}^\top]\boldsymbol{X}$ with $\boldsymbol{c} = \boldsymbol{\Sigma}\boldsymbol{\eta}(\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-1}$. Let $\mathbf{M} := \mathbf{M}(\mathbf{t})$ and $\lambda\mathbf{m} := \lambda\mathbf{m}(\mathbf{t})$ defined in (12). Then, for any fixed $\lambda > 0$, we have the following properties:

- *$\boldsymbol{Z}$ is uncorrelated with, and hence independent of, $\boldsymbol{\eta}^\top\boldsymbol{X}$.*
- *The conditioning set can be written as follows*

$$\{\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\, \mathbf{m}\} = \{\mathcal{V}^-(\boldsymbol{Z}) \leq \boldsymbol{\eta}^\top \boldsymbol{X} \leq \mathcal{V}^+(\boldsymbol{Z}), \mathcal{V}^0(\boldsymbol{Z}) \geq 0\} \tag{13}$$

  where
  - $\mathcal{V}^-(\boldsymbol{Z}) := \max_{l:(\mathbf{M}\mathbf{P}_\sigma \boldsymbol{c})_l < 0} \frac{\lambda m_l - (\mathbf{M}\mathbf{P}_\sigma \boldsymbol{Z})_l}{(\mathbf{M}\mathbf{P}_\sigma \boldsymbol{c})_l}$
  - $\mathcal{V}^+(\boldsymbol{Z}) := \min_{l:(\mathbf{M}\mathbf{P}_\sigma \boldsymbol{c})_l > 0} \frac{\lambda m_l - (\mathbf{M}\mathbf{P}_\sigma \boldsymbol{Z})_l}{(\mathbf{M}\mathbf{P}_\sigma \boldsymbol{c})_l}$
  - $\mathcal{V}^0(\boldsymbol{Z}) := \min_{l:(\mathbf{M}\mathbf{P}_\sigma \boldsymbol{c})_l = 0} \lambda m_l - (\mathbf{M}\mathbf{P}_\sigma \boldsymbol{Z})_l.$

Note that $\mathcal{V}^-(\boldsymbol{Z})$, $\mathcal{V}^+(\boldsymbol{Z})$ and $\mathcal{V}^0(\boldsymbol{Z})$ are independent of $\boldsymbol{\eta}^\top\boldsymbol{X}$. Finally, when the event in (13) has non-zero probability, conditionally to this event, the probability that $\mathcal{V}^-(\boldsymbol{Z}) = \mathcal{V}^+(\boldsymbol{Z})$ is zero.

From Proposition 4, it is shown in [14] that, for any fixed $\mathbf{z}_0$ with $\mathcal{V}^-(\mathbf{z}_0) < \mathcal{V}^+(\mathbf{z}_0)$, under the null hypothesis $\boldsymbol{\eta}^\top\boldsymbol{\mu} = 0$, conditionally to $\{\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\, \mathbf{m}, \boldsymbol{Z} = \mathbf{z}_0\}$, the following invariant statistic based on the test statistic $\boldsymbol{\eta}^\top\boldsymbol{X}$ fulfills

$$T(\boldsymbol{X}, \mathbf{t}, \sigma) := F^{[\mathcal{V}^-(\mathbf{z}_0), \mathcal{V}^+(\mathbf{z}_0)]}_{0, \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}}(\boldsymbol{\eta}^\top \boldsymbol{X}) \sim \mathcal{U}[0,1], \tag{14}$$

where $\mathcal{U}[0,1]$ denotes the uniform distribution and $F_{\nu,\tau^2}^{[a,b]}(.)$ is the cumulative distribution function (cdf) of a Gaussian distribution $\mathcal{N}(\nu,\tau^2)$ truncated on the interval $[a,b]$.

The p-value, corresponding to considering two-sided alternative hypotheses to $\boldsymbol{\eta}^\top\boldsymbol{\mu}=0$, is then

$$\text{pval}(\boldsymbol{x},\mathbf{t},\sigma) = 2\min\left[T(\boldsymbol{x},\mathbf{t},\sigma), 1-T(\boldsymbol{x},\mathbf{t},\sigma)\right] \tag{15}$$

for a $n\times 1$ observation vector $\boldsymbol{x}$. Note that the two definitions (14) and (15) require $\mathcal{V}^-(\mathbf{z}_0) < \mathcal{V}^+(\mathbf{z}_0)$, which holds almost surely conditionally to $\mathbf{MP}_\sigma\boldsymbol{X} \leq \lambda\,\mathbf{m}$, as stated in Proposition 4.

2.3.2. Conditional level

Next, we show that the suggested test is conditionally valid. That is, conditionally to a clustering and a data order, when the null hypothesis (that is fixed by the clustering) holds, the p-value is uniformly distributed. In particular, the probability of rejection is equal to the prescribed level. Conditional validity naturally yields unconditional validity, as shown in Section 2.3.3. Hence conditional validity is mathematically a stronger property than unconditional validity. A statistical benefit of conditional validity is that the null hypothesis is fixed after conditioning; in particular $\boldsymbol{\eta}^\top\boldsymbol{\mu}$ becomes a fixed target of interest, which is beneficial for interpretability. In the related context of linear models, for instance, the tests obtained from the confidence intervals of [2, 3] are unconditionally valid while the tests provided in [14, 29] are conditionally (and unconditionally) valid. The interpretability benefit we discuss above is also discussed in [14].

Proposition 5. *Let $\mathbf{t}$ be a fixed vector in $\mathcal{T}_{K,n}$ with $K \in [|n|]$. Let $\sigma \in \mathfrak{S}_n$ be a fixed permutation of $[|n|]$, and $\mathbf{P}_\sigma$ be the $n\times n$ associated permutation matrix. Let $\mathcal{C} = \mathcal{C}(\mathbf{t},\sigma)$ be the clustering obtained from Definition 2, with cluster cardinalities $n_1,\dots,n_K$.*

Let $\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ invertible. Consider a fixed $n\times 1$ non-zero vector $\boldsymbol{\eta}\in\mathbb{R}^n$ (that is only allowed to depend on $(\mathbf{t},\sigma)$). Assume that

$$\boldsymbol{\eta}^\top\boldsymbol{\mu} = 0.$$

Let $\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{B}}(\boldsymbol{X})$ from Problem (5) for some fixed $\lambda>0$. Assume that with non-zero probability, the event

$$E_{\mathbf{t},\sigma} \quad := \quad \left\{\mathcal{C}(\mathbf{t},\sigma) \text{ is the clustering given by } \widehat{\boldsymbol{B}},\quad X_{\sigma(1)} \geq X_{\sigma(2)} \geq \cdots \geq X_{\sigma(n)}\right\}$$

holds. Then, conditionally to $E_{\mathbf{t},\sigma}$, $\text{pval}(\boldsymbol{X},\mathbf{t},\sigma)$ is uniformly distributed on $[0,1]$:

$$\mathbb{P}_{\boldsymbol{\eta}^\top\boldsymbol{\mu}=0}\left(\text{pval}(\boldsymbol{X},\mathbf{t},\sigma)\leq t \middle| E_{\mathbf{t},\sigma}\right) = t \qquad \forall t\in[0,1].$$

2.3.3. Unconditional level

We now show that $\text{pval}(\boldsymbol{X},\mathbf{t},\sigma)$ is unconditionally uniformly distributed, which we call unconditional validity. Here, "unconditionally" means that the clustering is not fixed, but it is still necessary to condition by the fact that the null hypothesis $\boldsymbol{\eta}^\top\boldsymbol{\mu}=0$ is well-defined and true. Regarding well-definiteness, the vector $\boldsymbol{\eta}=\boldsymbol{\eta}(\mathcal{C}(\boldsymbol{X}))$ may indeed not be well-defined for all clusterings $\mathcal{C}(\boldsymbol{X})$. In the next proposition, we thus introduce the set $\mathcal{E}$ of clusterings,

indexed by an ordering σ and a sequence of right-limits $\mathbf{t}$ as in Definition 2, that make $\boldsymbol{\eta}$ well-defined.

For instance, in the case of the two-group test of Example 1, $\boldsymbol{\eta}$ can be defined similarly as in (4), with

$$\boldsymbol{\eta}^\top \boldsymbol{\mu} = \frac{1}{|\mathcal{C}_{k_1}(\boldsymbol{X})|} \sum_{i \in \mathcal{C}_{k_1}(\boldsymbol{X})} \mu_i - \frac{1}{|\mathcal{C}_{k_2}(\boldsymbol{X})|} \sum_{i \in \mathcal{C}_{k_2}(\boldsymbol{X})} \mu_i. \tag{16}$$

In this case, $\mathcal{E}$ is the set of clusterings for which the number of clusters is larger than or equal to $\max(k_1, k_2)$, enabling $\boldsymbol{\eta}$ to be well-defined. When $k_1 = 1$ and $k_2 = 2$, this definition is possible for all clusterings, except the one with only one cluster. In this case, $\mathcal{E}$ should thus be defined as restricting $\mathbf{t}$ to have at least 3 elements $0 = t_0 < t_1 < t_2 = n$, that is to correspond to a clustering with at least two clusters.

Then, Proposition 6 shows that conditionally to $\mathcal{E}$ and to $\boldsymbol{\eta}^\top \boldsymbol{\mu} = 0$, the p-value is uniformly distributed, which we call unconditional validity, in the sense that we do not condition by a single clustering, as commented above.

Proposition 6. *Let $\mathcal{E}$ be a subset of the set of all possible values of $(\mathbf{t}, \sigma)$ in Proposition 5. Consider a deterministic function $\boldsymbol{\eta} : \mathcal{E} \to \mathbb{R}^n$, outputing a non-zero column vector. Assume that $\boldsymbol{\Sigma}$ is invertible. Let $\widehat{\boldsymbol{B}}$ as in (5). Let $S = S(\boldsymbol{X})$ be a random permutation obtained by reordering $\boldsymbol{X}$ as: $X_{S(1)} \geq \cdots \geq X_{S(n)}$ (uniquely defined with probability one). Let $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}$ be the random clustering given by $\widehat{\boldsymbol{B}}$ (Definition 1), of random dimension $K(\boldsymbol{X}) = K$. Let $\mathbf{T}(\boldsymbol{X}) = \mathbf{T} \in \mathcal{T}_{K,n}$ be the random vector, such that $\mathbf{T}$ and S yield $\mathcal{C}$ as in Definition 2.*

Assume that

$$\mathbb{P}\Big((\mathbf{T}, S) \in \mathcal{E}, \boldsymbol{\eta}(\mathbf{T}, S)^\top \boldsymbol{\mu} = 0\Big) > 0.$$

Then, conditionally to the above event, $\mathrm{pval}(\boldsymbol{X}, \mathbf{T}, S)$ *is uniformly distributed on* $[0, 1]$*:*

$$\mathbb{P}\Big(\mathrm{pval}(\boldsymbol{X}, \mathbf{T}, S) \leq t \Big| (\mathbf{T}, S) \in \mathcal{E}, \boldsymbol{\eta}(\mathbf{T}, S)^\top \boldsymbol{\mu} = 0\Big) = t \qquad \forall t \in [0, 1].$$

2.4. Regularization path

At first sight, (5) is a convex optimization problem, whose (unique) minimizer does not have any explicit expression, and thus (5) requires numerical optimization to approximate its solution. Furthermore, this numerical optimization would be repeated for different values of λ. However, thanks to the polyhedral characterization of Theorem 2, we can provide a regularization path for solving (5). This regularization path is an algorithm, only performing elementary operations, that provides the entire sequence of exact solutions to (5), for all values of λ. This algorithm is exposed in Algorithm 1. Then, Theorem 7 shows that this algorithm is well-defined and indeed provides the set of solutions to Problem (5).

Theorem 7. *Algorithm 1 stops at a final value of r that we write $r_{\max}$, such that $r_{\max} \leq n - 1$ and we have $K^{(0)} > \cdots > K^{(r_{\max})} = 1$. Let $\lambda^{(r_{\max}+1)} = +\infty$ by convention. For $r \in \{0, \ldots, r_{\max}\}$ and $\lambda \in [\lambda^{(r)}, \lambda^{(r+1)})$, $(\hat{B}_i^{(r)}(\lambda))_{i \in [|n|]}$ minimizes Problem (5)*[1].

[1]Even if Algorithm 1 stops at $r = r_{\max}$, we can still define $(\hat{B}_i^{(r_{\max})}(\lambda))_{i \in [|n|]}$ there, with (17), with the convention that $\sum_{k'=1}^{0} n_{k'}^{(r_{\max})} = 0$ and $\sum_{k'=2}^{1} n_{k'}^{(r_{\max})} = 0$. This vector has all its components equal to $\sum_{i=1}^{n} x_i / n$.

Algorithm 1: Regularization path for one-dimensional convex clustering

Input: $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$

Initialization

$r \leftarrow 0$; $\lambda^{(0)} \leftarrow 0$;

$\tilde{x}_1 > \cdots > \tilde{x}_{K^{(0)}}$: the $K^{(0)}$ distinct values in $\boldsymbol{x}$;

$\mathcal{C}^{(0)} = (\mathcal{C}_1^{(0)}, \ldots, \mathcal{C}_{K^{(0)}}^{(0)}) \leftarrow$ clustering of $[|n|]$ where $\mathcal{C}_k^{(0)} = \{i \in [|n|] : x_i = \tilde{x}_k\}$;

$n_k^{(0)} \leftarrow |\mathcal{C}_k^{(0)}|$ for $k \in [|K^{(0)}|]$;

$\hat{b}_k^{(0)}(\lambda^{(0)}) \leftarrow \tilde{x}_k$ for $k \in [|K^{(0)}|]$;

$\hat{B}_i^{(0)}(\lambda^{(0)}) \leftarrow \hat{b}_k^{(0)}(\lambda^{(0)})$ if $i \in \mathcal{C}_k^{(0)}$ (k is unique) for $i \in [|n|]$;

while $K^{(r)} \geq 2$ **do**

For all $\lambda \geq \lambda^{(r)}$ *we define*

$$\hat{b}_k^{(r)}(\lambda) := \hat{b}_k^{(r)}(\lambda^{(r)}) + \left(\lambda - \lambda^{(r)}\right) \left(\sum_{k'=1}^{k-1} n_{k'}^{(r)} - \sum_{k'=k+1}^{K^{(r)}} n_{k'}^{(r)} \right) \quad \forall k \in [|K^{(r)}|] \tag{17}$$

$\hat{B}_i^{(r)}(\lambda) := \hat{b}_k^{(r)}(\lambda)$ if $i \in \mathcal{C}_k^{(r)}$ (k is unique) for $i \in [|n|]$;

$$\lambda^{(r+1)} \leftarrow \lambda^{(r)} + \min_{k \in [|K^{(r)}-1|]} \frac{\hat{b}_k^{(r)}(\lambda^{(r)}) - \hat{b}_{k+1}^{(r)}(\lambda^{(r)})}{n_k^{(r)} + n_{k+1}^{(r)}}; \tag{18}$$

$(\hat{b}_k^{(r+1)}(\lambda^{(r+1)}))_{k \in [|K^{(r+1)}|]} \leftarrow$ distinct values of $(\hat{b}_k^{(r)}(\lambda^{(r+1)}))_{k \in [|K^{(r)}|]}$, sorted decreasingly;

$\mathcal{C}^{(r+1)} \leftarrow$ clustering of $[|n|]$ obtained from $\left(\hat{B}_i^{(r)}(\lambda^{(r+1)})\right)_{i \in [|n|]}$ by Definition 1;

$n_k^{(r+1)} \leftarrow |\mathcal{C}_k^{(r+1)}|$ for $k \in [|K^{(r+1)}|]$;

$r \leftarrow r + 1$;

end

By way of illustration, Algorithm 1 was applied to the observations of Example 2, and the resulting regularization path is shown in Figure 2. In Algorithm 1, since $r \mapsto K^{(r)}$ is strictly decreasing during the execution, there are at most $n-1$ induction steps. A straightforward implementation of (18) can lead to a time complexity of order $\mathcal{O}(K^{(r)})$ for each step, and thus a total time complexity of order $\mathcal{O}(n^2)$ in the worst case. The space complexity is linear ($\mathcal{O}(n)$). Indeed, in order to recover the entire regularization path, it is sufficient to record at each step r the labels of the clusters merged at this step. We have implemented this algorithm in the open source `R` package `poclin` (which stands for "post convex clustering inference"), which is available from `https://plmlab.math.cnrs.fr/pneuvial/poclin`. The empirical time complexity of our implementation is substantially below $\mathcal{O}(n^2)$ for $n \leq 10^5$, as illustrated in Appendix D. In this appendix, we also explain that the time complexity of Algorithm 1 could be further decreased to $\mathcal{O}(n \log(n))$ without compromising the linear space complexity by storing merge candidates more efficiently using a min heap.

Remark 1 (Final value of the regularization parameter)**.** *As a consequence of Theorem 2 (see in particular (9)), the final value of λ in Algorithm 1 is obtained analytically as:*

$$\lambda^{(r_{\max})} = \max_{i \in [\![n-1]\!]} \frac{\frac{1}{i}\sum_{i'=1}^{i} x_{(i')} - \frac{1}{n}\sum_{i'=1}^{n} x_{(i')}}{n-i}. \tag{19}$$

It corresponds to the smallest value of λ for which the convex clustering yields exactly one cluster. The range of values for which there are two or more clusters has also been studied by [26] for convex clustering procedures that include Problem (5). We note that in the specific case of Problem (5), $\lambda^{(r_{\max})}$ can be computed using (19) in linear time after an initial sorting of the input vector. Our numerical experiments below make use of (19) to choose λ in a non data-driven way, see also Appendix E.1.

Relation to other existing regularization path algorithms. Algorithm 1 has similarities with the following two more general regularization path algorithms, that can be applied to Problem (5). First, for the generalized lasso, a penalization term of the form $\|\boldsymbol{DB}\|_1$ is studied in [30], for a general matrix $\boldsymbol{D}$. It is then simple to find a $n(n-1)/2 \times n$ (sparse) matrix $\boldsymbol{D}$ leading to the penalization term $\lambda \sum_{i,i'=1, i<i'}^{n} |B_{i'} - B_i|$ of (5). The benchmarks that we have conducted in Appendix D show that the procedure based on the generalized lasso has a very large memory footprint and is very slow (more than 10 seconds for $n = 50$), as it relies on the matrix $\boldsymbol{D}$, whose total number of entries is $\mathcal{O}(n^3)$. Second, the fused lasso signal approximator (FLSA) suggested by [11] can handle a penalization term of the form $\lambda \sum_{i,i'=1,(i,i')\in E}^{n} |B_{i'} - B_i|$, where E is a set of pairs of indices. Similarly as before, taking E as the complete set of pairs recovers the penalization term of (5). The theoretical time complexity of the regularization path for FLSA has been shown in [10] to be $\mathcal{O}(n \log(n))$ in this case. The benchmarks that we have conducted in Appendix D show that the procedure based on FLSA is much more efficient than the one based on the generalized lasso. Nevertheless, our implementation of Algorithm 1 remains preferable, as it can address larger dataset sizes (see Figure 6).

On top of these numerical performances, the benefit of Algorithm 1, relatively to these two general procedures, is that its description and proof of validity (Theorem 7) are self-contained and specific to the one-dimensional convex clustering problem (5). Furthermore, the proof of validity exploits the specific analysis of (5) given by Theorem 2.

2.5. Numerical experiments

In order to illustrate the behaviour of our post-clustering testing procedure, we have performed the following numerical experiments in the one-dimensional framework. The code to reproduce these numerical experiments and the associated figures is available from https://plmlab.math.cnrs.fr/pneuvial/poclin-paper.

We consider a Gaussian sample $\boldsymbol{X} = (X_1, \dots, X_n)$ with mean vector $\boldsymbol{\mu} = (\nu \mathbf{1}_{n/2}^\top, \mathbf{0}_{n/2}^\top)^\top$ and known covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_n$. Here and in the rest of the paper, for $a \in \mathbb{N}$, we let $\mathbf{1}_a$ be the $a \times 1$ vector composed of ones.

We set $n = 1000$ and $\lambda = 0.0025$. This value of λ has been chosen to ensure that with high probability, the convex clustering finds at least two clusters under the null hypothesis. The procedure that we have used in our numerical experiments to achieve this property relies on (19) and is described in Appendix E.1. Let $\mathcal{C} = (\mathcal{C}_k)_{k \in [|K|]}$ be the result of the one-dimensional convex clustering obtained from Algorithm 1 with $\lambda = 0.0025$. If $K > 2$, we merge adjacent clusters in to obtain a 2-class clustering of the form $\overline{\mathcal{C}}_1 := \mathcal{C}_1 \cup \dots \cup \mathcal{C}_q, \overline{\mathcal{C}}_2 := \mathcal{C}_{q+1} \cup \dots \cup \mathcal{C}_K$, where q is chosen so that the sizes of $\overline{\mathcal{C}}_1$ and $\overline{\mathcal{C}}_2$ are as balanced as possible. We then the test procedure introduced in Section 2.3.1 to compare the means of $\overline{\mathcal{C}}_1$ and $\overline{\mathcal{C}}_2$, as in Example 1. Note that this yields $\eta_i = \mathbb{1}_{i \in \overline{\mathcal{C}}_1}/|\overline{\mathcal{C}}_1| - \mathbb{1}_{i \in \overline{\mathcal{C}}_2}/|\overline{\mathcal{C}}_2|$ for $i \in [|n|]$, which is indeed a deterministic function of $\mathcal{C}_1, \dots, \mathcal{C}_K$ and thus in the scope of the guarantees obtained in Section 2.3. For each signal value $\nu \in \{0, 1, 2, 3, 4, 5\}$, we retain $N = 1000$ numerical experiments for which $K \geq 2$. Note that the event $K \geq 2$ corresponds to the set $\mathcal{E}$ in Proposition 6.

Figure 3 (left) gives the empirical density of $\boldsymbol{\eta}^\top \boldsymbol{\mu}$, the difference between the true means of the estimated clusters, for each value of ν considered. This plot quantifies the performance of the clustering step: for a perfect clustering, we would have $\boldsymbol{\eta}^\top \boldsymbol{\mu} = \nu$, corresponding to the diagonal line. As expected, the larger the signal (ν increases), the easier the clustering step.

Figure 3 (right) shows the empirical p-value distribution of the proposed test (see (15)). For $\nu = 0$ (no signal), the curve illustrates the uniformity of the distribution of the p-values: it shows that the level of the test is appropriately controlled. Another simulation to control the level of the test is available in Appendix E.2. As expected, the power of the test is an increasing function of the distance between the null and the alternative hypotheses (as encoded by the parameter ν). Our conditional test is able to detect the signal only for $\nu > 1$.

3. The p-dimensional case

3.1. Aggregating one-dimensional clusterings

Consider the p-dimensional setting of Section 1. For $j \in [|p|]$, consider the one-dimensional clustering $\mathcal{C}^{(j)} = \mathcal{C}^{(j)}(\boldsymbol{Y}_{\cdot j}) = (\mathcal{C}_1^{(j)}(\boldsymbol{Y}_{\cdot j}), \dots, \mathcal{C}_{K^{(j)}}^{(j)}(\boldsymbol{Y}_{\cdot j}))$ obtained by computing $\hat{\boldsymbol{B}}_{\cdot j}$ by solving (2) and with Definition 1. We consider a p-dimensional clustering $\mathcal{C}$ obtained by aggregation of the one-dimensional clusterings $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(p)}$ as follows.

For $i \in [|n|]$ and $j \in [|p|]$, let $\tilde{Y}_{ij}$ be the class index of Y_{ij} in the clustering $\mathcal{C}^{(j)}$, rescaled from $\{1, 2, \dots, K^{(j)}\}$ to $\{0, 1/(K^{(j)} - 1), \dots, 1\}$. We obtain a p-dimensional clustering $\mathcal{C}$ by applying a clustering procedure to the rows of the $n \times p$ matrix $\tilde{\boldsymbol{Y}}$, for instance a hierarchical clustering [17] with the Euclidean distance. We are then in a position to test an hypothesis $\boldsymbol{\kappa}^\top \boldsymbol{\beta} = 0$, where $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\mathcal{C})$, as motivated in Section 1. In particular, we can test the signal difference for the column j_0 between two clusters $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$ in the multi-dimensional clustering $\mathcal{C}$, as in (4).

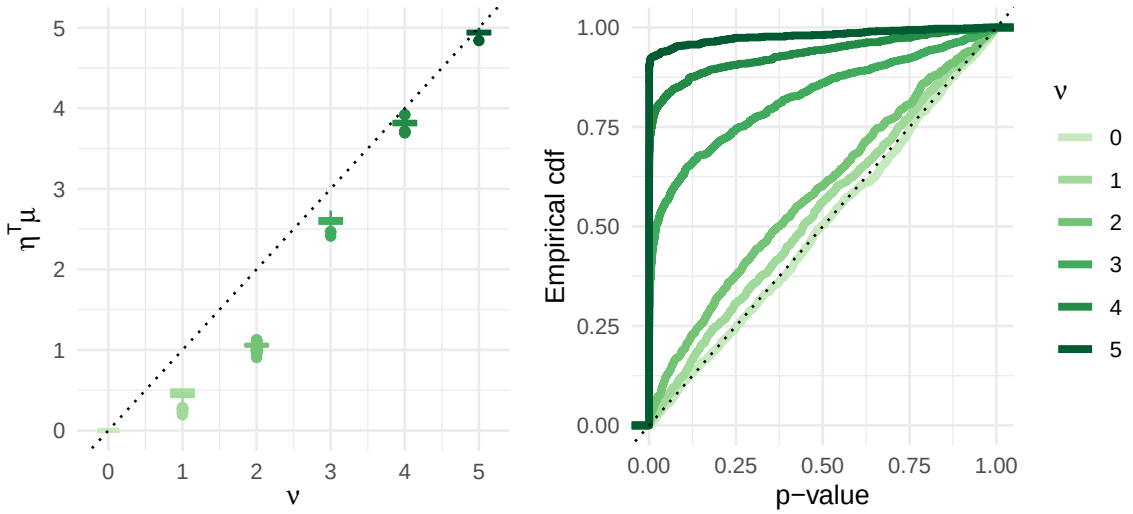

Figure 3: Left: empirical density of $\boldsymbol{\eta}^\top \boldsymbol{\mu}$ for each ν. Right: empirical cumulative distribution functions of the p-value of the test of equality between the means of two clusters.

Remark 2. *Above, we focus on a specific aggregation using the hierarchical clustering with the Euclidean distance for simplicity. However, we can construct more general p-dimensional clusterings $\mathcal{C}$ by more general aggregations of $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(p)}$. Indeed, our statistical framework (see Section 3.2) encompasses any case where $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\mathcal{C})$, as long as $\mathcal{C}$ is a function of the one-dimensional clusterings and orderings. In particular, one could also consider the hierarchical clustering with the Hamming distance, or the "unanimity" clustering, (i and i' are in the same cluster of $\mathcal{C}$ if and only if they are in the same cluster for each $\mathcal{C}^{(j)}$). This latter clustering is actually the one provided by Problem (1). For more background on clustering aggregation, we refer for instance to [7, 21, 32] and references therein.*

3.2. Test procedure and its guarantees

3.2.1. Construction of the test procedure

The test procedure for the hypothesis $\boldsymbol{\kappa}^\top \boldsymbol{\beta} = 0$ is constructed similarly as in Section 2.3.1. We consider p permutations $\sigma^{(1)}, \dots, \sigma^{(p)}$ that provide the orderings of the columns of the $n \times p$ observation matrix $\boldsymbol{Y}$. As in Definition 2, we identify the p clusterings $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(p)}$ by their numbers of classes $K^{(1)}, \dots, K^{(p)} \in [|n|]$ and by the right-limit sequences $\mathbf{t}^{(j)} \in \mathcal{T}_{K^{(j)},n}$ for $j \in [|p|]$.

For $j \in [|p|]$, we consider the matrix $\mathbf{M}(\mathbf{t}^{(j)})\mathbf{P}_{\sigma^{(j)}}$ of size $2(n-1) \times n$ and the vector $\lambda \mathbf{m}(\mathbf{t}^{(j)})$ of size n, defined in Lemma 3. Recall from Section 2 that, if only the variable j and its clustering $\mathcal{C}^{(j)}$ and order $\sigma^{(j)}$ were considered, then the conditioning event would be $\{\mathbf{M}(\mathbf{t}^{(j)})\mathbf{P}_{\sigma^{(j)}} \boldsymbol{Y}_{\cdot j} \leq \lambda \mathbf{m}(\mathbf{t}^{(j)})\}$.

We then explicit the conditioning constraints in dimension p, corresponding to all the clusterings $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(p)}$ and orders $\sigma^{(1)}, \dots, \sigma^{(p)}$. We define the matrix $\boldsymbol{\mathscr{M}}$ of size $2(n-1)p \times np$ in the following block-wise fashion. There are p^2 rectangular blocks (corresponding to dividing the rows into p groups and the columns into p groups). The block indexed by row-group j and column-group j' has size $2(n-1) \times n$. It is zero if $j \neq j'$ and it is equal to $\mathbf{M}(\mathbf{t}^{(j)})$ if

$p=1$	$\boldsymbol{x}$	$\boldsymbol{X}$	$\boldsymbol{\Sigma}$	$\boldsymbol{\eta}$	$\mathbf{P}_\sigma$	$\mathbf{M}$	$\lambda\mathbf{m}$	$\boldsymbol{Z}$	$\boldsymbol{c}$
size	n	n	$n\times n$	n	$n\times n$	$2(n-1)\times n$	n	n	n
$p>1$	$\mathrm{vec}(\mathbf{y})$	$\mathrm{vec}(\boldsymbol{Y})$	$\boldsymbol{\Gamma}$	$\boldsymbol{\kappa}$	$\mathbf{D}_\sigma$	$\boldsymbol{\mathscr{M}}$	$\lambda\boldsymbol{m}$	$\mathrm{vec}(\overline{\boldsymbol{Z}})$	$\mathrm{vec}(\overline{\boldsymbol{c}})$
size	np	np	$np\times np$	np	$np\times np$	$2(n-1)p\times np$	np	np	np

TABLE 1

Correspondence between the notation of Section 2.3.1 (dimension one) and the notation of Sections 3.2.1 and 3.2.2 (dimension p).

$j=j'$. Define also $\mathbf{D}_\sigma$ as the $np\times np$ block diagonal matrix with p diagonal blocks and block j equal to $\mathbf{P}_{\sigma^{(j)}}$, for $j\in[|p|]$. With these definitions, we have

$$\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma\mathrm{vec}(\boldsymbol{Y})=\begin{pmatrix}\mathbf{M}(\mathbf{t}^{(1)})\mathbf{P}_{\sigma^{(1)}}\boldsymbol{Y}_{.1}\\ \vdots\\ \mathbf{M}(\mathbf{t}^{(p)})\mathbf{P}_{\sigma^{(p)}}\boldsymbol{Y}_{.p}\end{pmatrix}.$$

We let $\lambda\boldsymbol{m}$ be the vector obtained by stacking the column vectors $\lambda\mathbf{m}(\mathbf{t}^{(j)})$, $j\in[|p|]$, one above the other. The conditioning constraints in dimension p are then $\{\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma\mathrm{vec}(\boldsymbol{Y})\le\lambda\boldsymbol{m}\}$.

Consider a column vector $\boldsymbol{\kappa}$ of size np, that is allowed to depend on $(\mathbf{t}^{(j)},\sigma^{(j)}),j\in[|p|]$. This includes the setting $\boldsymbol{\kappa}=\boldsymbol{\kappa}(\mathcal{C}^{(1)},\dots,\mathcal{C}^{(p)})$ of Section 3.1, with the additional mathematical flexibility that $\boldsymbol{\kappa}$ is allowed to depend on the orderings of the columns, besides their clusterings.

Recall that $\boldsymbol{\Gamma}$ is the $np\times np$ covariance matrix of $\mathrm{vec}(\boldsymbol{Y})$. Note that in the definition of $\mathrm{pval}(\boldsymbol{x},\mathbf{t},\sigma)$ in Section 2.3.1 (one-dimensional case), the values of $\boldsymbol{x}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\eta}$, $\mathbf{MP}_\sigma$ and $\lambda\mathbf{m}$ are sufficient to determine the invariant statistic $T(\boldsymbol{X},\mathbf{t},\sigma)$ in (14) and the p-value $\mathrm{pval}(\boldsymbol{x},\mathbf{t},\sigma)$ in (15). Thus we can define the test statistic $\boldsymbol{\kappa}^\top\mathrm{vec}(\boldsymbol{Y})$, then the invariant statistic $T(\boldsymbol{Y})=T(\boldsymbol{Y},\mathbf{t}^{(1)},\dots,\mathbf{t}^{(p)},\sigma^{(1)},\dots,\sigma^{(p)})$ in the same way as $T(\boldsymbol{X},\mathbf{t},\sigma)$ in (14) and consequently the p-value $\mathrm{pval}(\mathbf{y})$, for a $n\times p$ realization $\mathbf{y}$ of $\boldsymbol{Y}$, in the same way as $\mathrm{pval}(\boldsymbol{x},\mathbf{t},\sigma)$ in (15). The explicit correspondence between the notation of the one-dimensional case and the present notation is given in Table 1. The next section provides additional explanations on the computation of the invariant statistic $T(\boldsymbol{Y})$, in the special case of independent variables, for the sake of exposition.

3.2.2. A detailed example: testing the signal difference along a variable j_0 with independent variables

Consider testing the signal difference for the column j_0 between two clusters $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$ in the multi-dimensional clustering $\mathcal{C}$, as in Example 1. It is interesting to explicit the construction of the invariant statistic in the special case of the matrix normal distribution (see Section 1) where $\boldsymbol{\Delta}$ is diagonal, that is the p n-dimensional observation vectors corresponding to the p variables are independent. For the sake of simplicity, let us even consider that $\boldsymbol{\Delta}=\boldsymbol{I}_p$.

Observe first that the test statistic satisfies $\boldsymbol{\kappa}^\top\mathrm{vec}(\boldsymbol{Y})=\boldsymbol{\eta}^\top\boldsymbol{Y}_{.j_0}$, where $\eta_i=\mathbb{1}_{i\in\mathcal{C}_{k_1}}/|\mathcal{C}_{k_1}|-\mathbb{1}_{i\in\mathcal{C}_{k_2}}/|\mathcal{C}_{k_2}|$. That is, the test statistic is constructed as it would be in the one-dimensional case (Section 2.3.1), except that the one-dimensional clustering $\mathcal{C}^{(j_0)}$ is replaced by the aggregated one $\mathcal{C}$. The variance of the test statistic (unconditional to the clusterings and orders of observations) is thus $\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}$ and is as in the one-dimensional case (up to the distinction between $\mathcal{C}^{(j_0)}$ and $\mathcal{C}$). Then, the next proposition specifies the computation of the invariant statistic.

Proposition 8. *In the context of Section 3.2.2, computing the invariant statistic as described in Section 3.2.1 is equivalent to proceed as described in Section 2.3.1 (one-dimensional case), with* $\boldsymbol{\eta}$ *defined by* $\eta_i = \mathbb{1}_{i\in\mathcal{C}_{k_1}}/|\mathcal{C}_{k_1}| - \mathbb{1}_{i\in\mathcal{C}_{k_2}}/|\mathcal{C}_{k_2}|$ *for* $i \in [|n|]$*, with* $\boldsymbol{X}$ *replaced by* $\boldsymbol{Y}_{\cdot j_0}$ *and with the conditioning set* $\{\mathbf{MP}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}\}$ *replaced by* $\left\{\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{Y}_{\cdot j_0} \leq \lambda \mathbf{m}(\mathbf{t}^{(j_0)})\right\}$.

In Proposition 8, the observations corresponding to the variables $j \neq j_0$, for which the average signal difference is not tested, have an impact on the clusterings $\mathcal{C}^{(j)}$, $j \neq j_0$, and thus have an impact on the multi-dimensional clustering $\mathcal{C}$ and thus on $\boldsymbol{\eta}$. Besides $\boldsymbol{\eta}$, these observations have no other influence on the construction of the invariant statistic, which is computed only from $\boldsymbol{Y}_{\cdot j_0}$ and its conditioning set $\{\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{Y}_{\cdot j_0} \leq \lambda \mathbf{m}(\mathbf{t}^{(j_0)})\}$ as in the one-dimensional case. This fact can be interpreted in light of the general properties of conditioning and independence. Indeed, we are studying events of the form E_j on $\boldsymbol{Y}_{\cdot j}$, $j \in [|p|]$ and we are studying a test statistic $\boldsymbol{\eta}(E_1, \dots, E_p)^\top \boldsymbol{Y}_{\cdot j_0}$ conditionally to these events. Here E_j encodes the event corresponding to (6) and (7) in Theorem 2 for variable j. By independence of $\boldsymbol{Y}_{\cdot j}$, $j \in [|p|]$, the events E_j, $j \neq j_0$ simply have an influence on $\boldsymbol{\eta}$, while the event E_{j_0} also has an impact on the conditional distribution of $\boldsymbol{Y}_{\cdot j_0}$ given E_{j_0}.

3.2.3. Conditional level

For $j \in [|p|]$, let $\widehat{\boldsymbol{B}}_{\cdot j}$ be obtained from (2). The next proposition is similar to Proposition 5 and proves that the p-value $\text{pval}(\boldsymbol{Y})$ in Section 3.2.1 is uniformly distributed, conditionally to the one-dimensional clusterings and orders, when the null hypothesis is true. We remark that in the context of Section 3.1, this implies that the p-value is also uniformly distributed conditionally to the p-dimensional clustering obtained by aggregation, when the null hypothesis is true.

Proposition 9. *Consider* p *fixed permutations* $\sigma^{(1)}, \dots, \sigma^{(p)}$ *of* $[|n|]$*. Let* $K^{(1)}, \dots, K^{(p)} \in [|n|]$*. For* $j \in [|p|]$*, let* $\mathbf{t}^{(j)} \in \mathcal{T}_{K^{(j)},n}$ *and consider the clustering* $\mathcal{C}^{(j)}$ *associated to* $(\mathbf{t}^{(j)}, \sigma^{(j)})$ *by Definition 2.*

Consider a fixed non-zero vector $\boldsymbol{\kappa} \in \mathbb{R}^{np}$ *(that is only allowed to depend on* $(\mathbf{t}^{(j)}, \sigma^{(j)}), j \in [|p|]$*). Assume that*

$$\boldsymbol{\kappa}^\top \boldsymbol{\beta} = 0.$$

Assume that with non-zero probability, the event

$$E \quad := \quad \left\{\text{for } j \in [|p|],\ \mathcal{C}^{(j)} \text{ is the clustering given by } \widehat{\boldsymbol{B}}_{\cdot j} \text{ and } Y_{\sigma^{(j)}(1)j} \geq \cdots \geq Y_{\sigma^{(j)}(n)j}\right\}$$

holds. Assume also that the $np \times np$ *matrix* $\boldsymbol{\Gamma}$ *is invertible. Then, conditionally to* E, $\text{pval}(\boldsymbol{Y})$ *is uniformly distributed on* $[0, 1]$ *under the null hypothesis.*

3.2.4. Unconditional level

The unconditional guarantee is similar to that of Proposition 6 for the one-dimensional case. In particular, here we also introduce the subset $\mathcal{E}$ on which the null hypothesis is well-defined.

Proposition 10. *Let* $\mathcal{E}$ *be a subset of the set of all possible values of* $(\mathbf{t}^{(j)}, \sigma^{(j)})_{j\in[|p|]}$ *in Proposition 9. Consider a deterministic function* $\boldsymbol{\kappa} : \mathcal{E} \to \mathbb{R}^{np}$*, outputing a non-zero column vector. Assume that* $\boldsymbol{\Gamma}$ *is invertible. For* $j \in [|p|]$*, let* $\widehat{\boldsymbol{B}}_{\cdot j}$ *be obtained from* (2)*. Let also* $S^{(j)} = S^{(j)}(\boldsymbol{Y}_{\cdot j})$ *be the random permutation obtained by the order of* $\boldsymbol{Y}_{\cdot j}$*:* $Y_{S^{(j)}(1)j} \geq \cdots \geq Y_{S^{(j)}(n)j}$

(uniquely defined with probability one). Let $\mathcal{C}^{(j)}(\boldsymbol{Y}_{\cdot j}) = \mathcal{C}^{(j)}$ *be the random clustering given by* $\widehat{\boldsymbol{B}}_{\cdot j}$ *(Definition 1). Let* $\mathbf{T}^{(j)}(\boldsymbol{Y}_{\cdot j}) = \mathbf{T}^{(j)} \in \mathcal{T}_{K^{(j)},n}$ *be the random vector (with random* $K^{(j)}(\boldsymbol{Y}_{\cdot j}) = K^{(j)}$*), such that* $(\mathbf{T}^{(j)}, S^{(j)})$ *yields* $\mathcal{C}^{(j)}$ *as in Definition 2.*

Assume that

$$\mathbb{P}\left((\mathbf{T}^{(j)}, S^{(j)})_{j\in[|p|]} \in \mathcal{E}, \boldsymbol{\kappa}((\mathbf{T}^{(j)}, S^{(j)})_{j\in[|p|]})^\top \boldsymbol{\beta} = 0\right) > 0.$$

Then, conditionally to the above event, $\mathrm{pval}(\boldsymbol{Y})$ *is uniformly distributed on* $[0, 1]$*.*

3.3. Numerical experiments

In this section, we describe the numerical experiments that we have performed in order to illustrate the behaviour of our post-clustering testing procedure for $p > 1$. The code to reproduce these numerical experiments and the associated figures is available from `https://plmlab.math.cnrs.fr/pneuvial/poclin-paper`.

We consider the specific case where $\boldsymbol{Y}$ is distributed from a matrix normal distribution $\mathcal{MN}_{n\times p}(\mathbf{u}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ (see Section 1) with $p = 3$, $\mathbf{u} = \begin{pmatrix} \nu\mathbf{1}_{n/2} & \mathbf{0}_{n/2} & \mathbf{0}_{n/2} \\ -\nu\mathbf{1}_{n/2} & \mathbf{0}_{n/2} & \mathbf{0}_{n/2} \end{pmatrix}$ with $\nu \in \{0, 1, 2, 5\}$, $\boldsymbol{\Sigma} = \boldsymbol{I}_n$, and $\boldsymbol{\Delta} = \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix}$ with $\rho \in \{0, 0.3, 0.5\}$.

We obtain $K = 2$ clusters by aggregating one-dimensional convex clusterings obtained for a given value of λ, as explained in Section 3.1. For each variable $j \in \{1, 2, 3\}$, we want to compare the means of the two clusters. This corresponds to the test of the null hypothesis $\boldsymbol{\kappa}^\top\boldsymbol{\beta} = 0$, where $\boldsymbol{\kappa}$ is defined by (3) (see Example 1). We compare our procedure with $\lambda = 0.016$ (resp. $\lambda = 0.0025$) for $n = 100$ (resp. $n = 1000$) and the two-group Wilcoxon rank sum test as implemented in the `R` function `wilcox.test`. This choice of λ ensures to have at least two clusters under the null hypothesis with high probability, as explained in Section 2.5 and Appendix E.1. The empirical cumulative distribution function of the p-values $\mathrm{pval}(\mathbf{y})$ across 500 experiments is represented for different values of the simulation parameters in Figures 4 and 5 for $n = 100$ and $n = 1000$, respectively. For each parameter combination, the p-value distribution of the proposed method (in green) is compared to that of the two-group Wilcoxon rank sum test (in orange) for all three variables $\boldsymbol{Y}_{\cdot j}$, for $j = 1, 2, 3$ (in columns). Each row corresponds to a value of ν and each line type corresponds to a value of ρ.

First, the clustering procedure described in Section 3.1 works reasonably well in this setting. Indeed, for the variable $\boldsymbol{Y}_{\cdot 1}$, the absolute value of the difference between the true means of the estimated clusters (obtained as $\boldsymbol{\kappa}^\top\boldsymbol{\beta}$) is generally close to the true value of the signal (that is 2ν), see Figure 9 in Appendix E.3.

The proposed test controls the type I error rate: in all situations where there is no signal (that is, for $\nu = 0$ or $j \in \{2, 3\}$), the empirical p-value distribution is close to the uniform distribution on $[0, 1]$ ($y = x$). Under the alternative hypothesis (i.e. for $j = 1$ and $\nu > 0$), our proposed test is able to detect some signal for $\nu \geq 2$. For $\nu = 1$ the signal is too small to be detected.

In contrast, the naive Wilcoxon test yields severely anti-conservative p-values in absence of signal. This test is naturally much more sensitive than our proposed test. However, it should

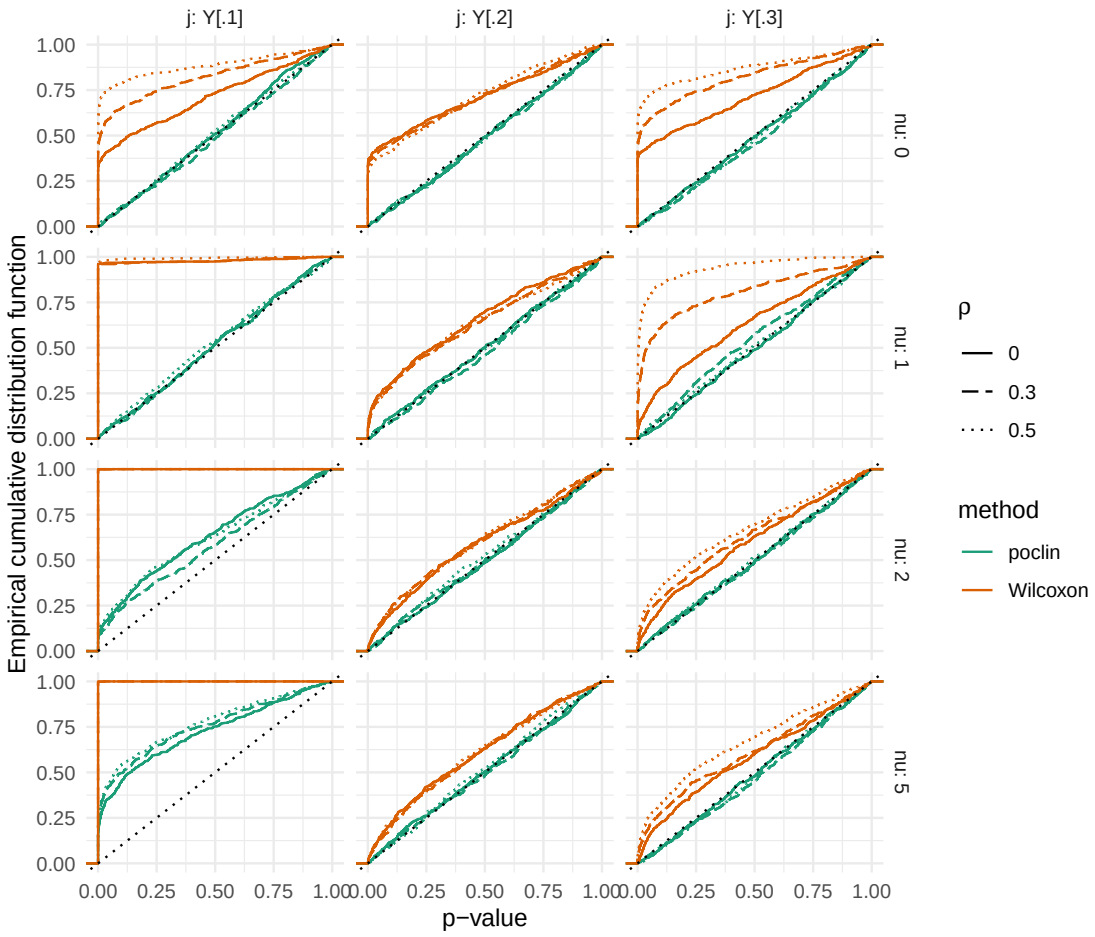

Figure 4: The empirical cumulative distribution function of the p-values across 500 experiments for $n = 100$ with our method `poclin` (in green) and the Wilcoxon test (in orange). Each column corresponds to a variable j, each row to a value of ν and each line type to a value of ρ.

be noted that one cannot compare the power of the two tests, since the Wilcoxon test fails to control type I error.

Regarding the influence of n: our proposed method does not gain much power as n increases from 100 to 1000. This is consistent with the fact that the signal is not different across values of n, see Figure 9. For $n = 1000$, the Wilcoxon test is able to distinguish the signal from the noise when $\rho = 0$ and actually becomes well-calibrated for $\boldsymbol{Y}_{\cdot 2}$ when $\nu \neq 0$. However, due to the correlation between $\boldsymbol{Y}_{\cdot 1}$ and $\boldsymbol{Y}_{\cdot 3}$, the Wilcoxon test is anti-conservative for $\boldsymbol{Y}_{\cdot 3}$.

4. Discussion

We first provide an overview of our contributions, and then we discuss various specific aspects of them and various remaining open questions.

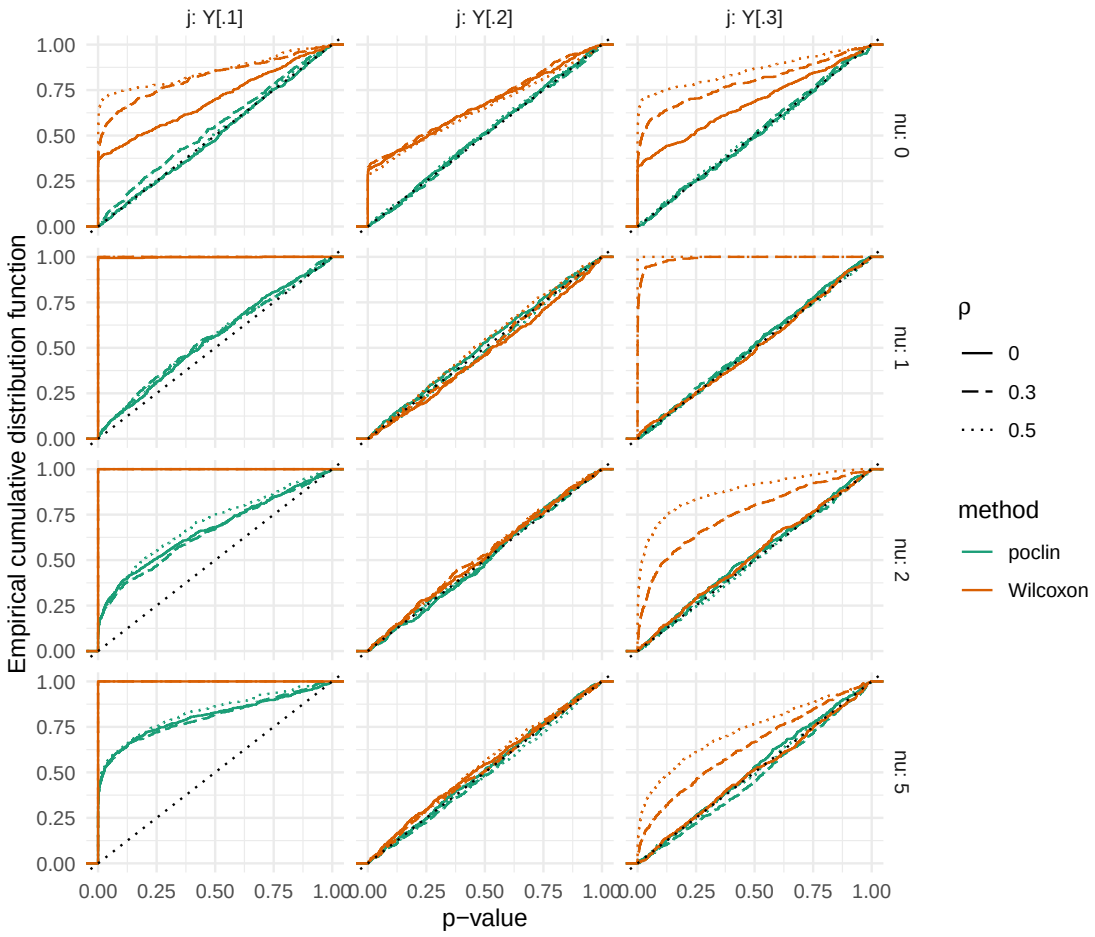

Figure 5: The empirical cumulative distribution function of the p-values across 500 experiments for $n = 1000$ with our method `poclin` (in green) and the Wilcoxon test (in orange). Each column corresponds to a variable j, each row to a value of ν and each line type to a value of ρ.

4.1. Overview of the contributions

Selective inference, in the post-clustering context, is a challenging problem and statistical guarantees could be obtained for it only in the recent years, see the references provided in Section 1. In this paper, we suggest a solution based on exhibiting polyhedral conditioning sets for Gaussian vectors, extending a line of work that has proved to be very successful in other statistical contexts, especially for regression models. This line of work was pioneered by [14] and then developed by [23, 29], among others.

Nevertheless, extending the existing approaches from regression models to clustering models is challenging. As such, the proofs we provide require innovations (for instance for Theorems 2 and 7). Furthermore, obtaining polyhedral conditioning sets is made possible by focusing on intermediate one-dimensional convex clustering optimization problems based on ℓ_1 penalties (see (2)). In the end, we provide the following workflow for selective inference post-clustering.

(1) We characterize a one-dimensional clustering by polyhedral constraints on the obser-

vation vector (Section 2.2).

(2) As a by by-product, we provide a regularization path algorithm to implement this clustering (Section 2.4). The computational efficiency of this algorithm is demonstrated numerically, also in comparison with other existing procedures.

(3) Following [14], from the polyhedral constraints, we obtain a test procedure which is conditionally and unconditionally valid post-clustering (Section 2.3). The procedure enables to test the nullity of any linear combination of the unknown mean vector, provided this combination only depends on the clustering (and on the order of the observations). In particular, it is possible to test for the significance of the signal difference between two clusters as in Example 1 (see Equation (16)). Although we do not develop it in this paper, confidence intervals for the above linear combination can be constructed from our test procedure, similarly as in [14]. Numerical experiments (Section 2.5) confirm the validity of the test procedure, and indicate that it has power to detect cases where the clustering on the observation vector was able to cluster the unknown mean vector as well into inhomogeneous groups.

(4) We suggest to aggregate one-dimensional clusterings to form a single multi-dimensional clustering for the data matrix. Our above contributions can thus be naturally leveraged to obtain a valid test procedure, posterior to this multi-dimensional clustering (Section 3.2). In particular, we can test the significance of the signal difference between clusters along a specific variable, as in Example 1. This feature could be beneficial in potential applications to single-cell RNA-seq data, since in this context, testing along a specific variable enables to study genes expressions individually. It is also a welcome complement to related references, in particular [6], that focuses on testing the global nullity of the signal mean difference vector across two clusters, rather than considering individual components (i.e. variables).
This workflow (1)-(4) depends on a regularization parameter λ that should not be data-driven (see Section 4.3 below). From a practical point of view, we provide a procedure to choose λ in a non data-driven way, from a choice of the covariance matrix, see Sections 2.5 and 3.3, and Appendix E.1.

Similarly as in the one-dimensional case, we provide numerical experiments (Section 3.3) that both confirm the validity of the test procedure and demonstrate its power to validate when the clustering procedure successfully yields clusters with significant signal difference for individual variables. These numerical experiments (as well as those in Section 2.5) also indicate that inference post-clustering is challenging, in that statistical procedures that do not account for the data-driven nature of the clustering are strongly anti-conservative. Indeed, the standard Wilcoxon test wrongly indicates signal differences across clusters in many cases where there is actually no difference. Note that the numerical experiments are focused on the hierarchical-clustering-based aggregation of one-dimensional clusterings, as described in Section 3.1. In future investigations, it would be relevant to quantify the benefit brought by alternative aggregation methods. Indeed, a flexibility of our framework is that our statistical guarantees hold for any aggregation procedure.

4.2. Benefits of the test procedure in well- and misspecified clustering problems

For simplicity, let us focus on the one-dimensional case of Section 2, with the observation vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The discussion of the multi-dimensional case of Section 3 would be similar. The clustering problem can be considered as well-specified if there are clusters of indices for the mean vector $\boldsymbol{\mu}$ with equal values, corresponding to a Gaussian mixture setting (see for instance [8, 13, 16, 22] for expositions and recent contributions on mixture models).

In the well-specified case, there are thus intrinsic classes of the observations and it is natural to aim at recovering them.

Consider for the sake of discussion that $n/2$ components of $\boldsymbol{\mu}$ are zero and the other $n/2$ components are one (there are two intrinsic classes) and that the clustering procedure yields two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ of equal size. Then if the null hypothesis $(2/n)\sum_{i\in\mathcal{C}_1}\mu_i = (2/n)\sum_{i\in\mathcal{C}_2}\mu_i$ is rejected by our test procedure, it means that one empirical cluster contains a strict majority of individuals from one intrinsic class, and vice versa for the second cluster. If our test procedure is extended to yield a confidence interval on $(2/n)\sum_{i\in\mathcal{C}_1}\mu_i - (2/n)\sum_{i\in\mathcal{C}_2}\mu_i$ showing that with high probability this quantity is larger than some $\delta \in (0,1)$, then one can see that the first empirical cluster contains at least $n(\delta+1)/4$ observations from an intrinsic class (corresponding to mean one; and conversely for the second cluster). Hence, generally speaking, for a well-specified clustering problem with intrinsic classes, our test procedure is relevant to recover these classes, similarly as statistical procedures that are dedicated to finite mixture problems, see the references given above.

On the other hand, the clustering problem can be considered as misspecified when the n components of $\boldsymbol{\mu}$ are two-by-two distinct. In this case one can consider that there are no intrinsic classes. Nevertheless, providing tests or confidence intervals on the same quantity $(2/n)\sum_{i\in\mathcal{C}_1}\mu_i - (2/n)\sum_{i\in\mathcal{C}_2}\mu_i$ as before enables to assess if the clustering procedure was able to cluster the unknown mean vector, besides the random/noisy observations. Hence, a benefit of the post-clustering framework considered here is that it is meaningful both in well- and misspecified settings. A similar discussion can be made in the related context of selective inference in regression settings, see in particular [1, 3].

4.3. Known covariance matrix and fixed λ

As pointed out above, we assume the covariance matrix ($\boldsymbol{\Sigma}$ in Section 2 and $\boldsymbol{\Gamma}$ in Section 3) to be known and the tuning parameter λ to be fixed. These two assumptions are necessary for our statistical guarantees in Sections 2 and 3. Indeed, the obtention of these guarantees relies first on exhibiting a Gaussian vector constrained to a polyhedron. Then, the Gaussian vector is decomposed into a linear combination (corresponding to the statistical hypothesis to test) and an independent remainder. This two-step strategy corresponds in particular to Lemma 3 and Proposition 4 in the one-dimensional case. It was previously suggested by [14] in the related context of post-selection inference for the lasso model selector, with Gaussian linear models.

Obtaining a polyhedron in the first step relies on λ not depending on the data, and computing the decomposition in the second step relies on knowing the covariance matrix. Broadly speaking, in the selective inference context, it is relatively common to assume known covariance matrices, or fixed tuning parameters, in order to obtain rigorous mathematical guarantees. This is indeed the case in [14] mentioned above, but also for instance in [6]. In this latter reference, the covariance matrix is assumed to be proportional to the identity, with a known variance for most of the theoretical results. Asymptotic results are given there in Section 4.3 for the case of a conservative variance estimator. Also, data thinning procedures, for instance in [18], usually require knowledge of the data distribution in order to produce independent parts, where the independence property enables valid statistical inference.

In our setting, obtaining theoretical guarantees (finite-sample or asymptotic) with an estimated covariance matrix or a data-dependent tuning parameter is of course an important problem for future work. In other contexts, successes have been obtained in this direction,

see in particular [28, 35]. Note that relaxing the assumption of known covariance matrix can yield identifiability issues, because the mean vector $\boldsymbol{\beta}$ is unrestricted (see also the discussion of misspecified clustering problems in Section 4.2). These identifiability issues boil down to the fact that multiple pairs of mean vector and covariance matrix can "explain" the same dataset. Studying which minimal assumptions circumvent these identifiability issues is thus an important problem in the prospect of extending this work to an estimated covariance matrix.

4.4. Choice of the ℓ_1 norm in the multi-dimensional convex clustering problem (1)

Our test procedure and its statistical guarantees for the multi-dimensional case rely on aggregating one-dimensional clusterings. As discussed in Remark 2, solving Problem (1) with the multi-dimensional ℓ_1 norm penalization boils down to one such aggregation. Hence, our procedure and guarantees apply to multi-dimensional convex clustering with ℓ_1 penalization.

One can see that our arguments, and crucially the proof of Theorem 2, cannot be applied directly to convex clusterings obtained by replacing the ℓ_1 penalization by a more general ℓ_q one, $q > 0$, and especially by the ℓ_2 one. In fact, we view the following question as an important open problem: is it possible to characterize the set of observation matrices $\boldsymbol{Y}$, such that Problem (1), with the ℓ_1 penalization replaced by the ℓ_q one, yields a given clustering, with polyhedral sets or other tractable sets?

Nevertheless, we note that the ℓ_1 penalization in Problem (1) has computational benefits. Indeed, the problem is separable, and for each subproblem, we have obtained an exact regularization path in Section 2.4 that stops after a maximal number of iterations known in advance. To our knowledge, such a favorable regularization path is not available for a general ℓ_q penalization. In agreement with this, the reference [10] (from 2011) concludes that Problem (1) can be readily solved for thousands of data points, while if the ℓ_1 penalization is replaced by the ℓ_q one, this is the case for (only) hundreds of data points.

4.5. Comparison with data splitting strategies

For the problem of post-clustering inference, data splitting (or data fission, data thinning) strategies [5, 18, 36] consist in separating the dataset into two stochastically independent ones, keeping the same indexing of individuals as the original dataset. Then, a clustering can be computed from the first dataset and then applied to the second data set. By independence, the distribution of a post-clustering statistic of interest (for instance the difference of average between two classes for a variable, in view of studying (4)) on the second data set remains simple. For instance if the original dataset is Gaussian, this distribution remains Gaussian conditionally to the clustering. Hence, a benefit of data splitting compared to our approach is a simplicity of implementation. Furthermore, any clustering procedure can be used.

On the other hand, with data splitting, conclusions are provided for a clustering computed on a dataset that differs from the original one. Hence, the conclusions of data splitting approaches might be more difficult to interpret for practitioners, compared to those of the present work, since these conclusions do not apply to the clustering that they would compute on the original data set.

Note also that data splitting and our approach share two similar difficulties. First, they share hyperparameters that should not be data-driven for the statistical guarantees. Indeed,

with data splitting we need to fix the splitting mechanism to general the two datasets above. Similarly, we fix the regularization parameter λ in (1). Second, considering Gaussian data, the covariance matrix should be known for data splitting and our approach, as already discussed in Section 4.3.

4.6. On conditioning by the orders

Let us consider the one-dimensional setting (Section 2) for simplicity of exposition. A similar discussion could be made for the multi-dimensional case as well. Our test procedure is valid conditionally to both the clustering and the order of observations, see Proposition 5, and our discussion at the beginning of Section 2.2. Being valid conditionally to the clustering can be considered as a desirable statistical feature, since the clustering is an object of interest in itself (see also the discussion before Proposition 5). However, being valid conditionally to the order is more a by-product of our approach than a desirable statistical feature. Indeed, in order to obtain a polyhedral set with a tractable number of linear pieces ($2(n-1)$) in Theorem 2, it was necessary in the proof to condition by the observation order. Importantly, the constraint (9) is not a linear constraint on the observation vector if the order is not fixed.

It could be the case that, if a test procedure could be derived by only conditioning by the clustering, this test could have more power than the one we obtain in Section 2.3, which is an interesting perspective for future work. In other words, it is possible that we pay a price when conditioning by the order of observations? In the related regression context, a similar phenomenon occurs in [14]. There, a first test procedure is obtained by conditioning by the selected variables and a second one is obtained by conditioning by the selected variables and the signs of the coefficients. The first procedure has a computational cost that is exponential in the number of variables, but is more powerful. The second procedure has a small computational cost. In Section 6 of [14], it is written on this point that "one may be willing to sacrifice statistical efficiency for computational efficiency".

Acknowledgements

This work was supported by the Project GAP (ANR-21-CE40-0007) of the French National Research Agency (ANR), and by the MITI at CNRS through the DDisc project.

References

[1] F. Bachoc, H. Leeb, and B. M. Pötscher. Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics*, 47(3):1475–1504, 2019.

[2] F. Bachoc, D. Preinerstorfer, and L. Steinberger. Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48(1):440–463, 2020.

[3] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.

[4] E. C. Chi and K. Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

[5] A. Dharamshi, A. Neufeld, K. Motwani, L. L. Gao, D. Witten, and J. Bien. Generalized data thinning using sufficient statistics. *arXiv preprint arXiv:2303.12931*, 2023.

[6] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, 2022.

[7] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1):4–es, mar 2007.

[8] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46(6A):2844–2870, 2018.

[9] B. Hivert, D. Agniel, R. Thiébaut, and B. P. Hejblum. Post-clustering difference testing: valid inference and practical considerations. *arXiv preprint arXiv:2210.13172*, 2022.

[10] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th International Conference on Machine Learning*, page 1, 2011.

[11] H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[12] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[13] B. Laurent, C. Marteau, and C. Maugis-Rabusseau. Non-asymptotic detection of two-component mixtures with unknown means. *Bernoulli*, 22(1):242–274, 2016.

[14] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[15] F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 201–204. IEEE, 2011.

[16] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6(1):355–378, 2019.

[17] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[18] A. Neufeld, A. Dharamshi, L. L. Gao, and D. Witten. Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*, 2023.

[19] A. Neufeld, L. L. Gao, J. Popp, A. Battle, and D. Witten. Inference after latent variable estimation for single-cell RNA sequencing data. *arXiv preprint arXiv:2207.00554*, 2022.

[20] A. Neufeld, J. Popp, L. L. Gao, A. Battle, and D. Witten. Negative binomial count splitting for single-cell rna sequencing data. *arXiv preprint arXiv:2307.12985*, 2023.

[21] N. Nguyen and R. Caruana. Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 607–612, 2007.

[22] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370, 2013.

[23] S. Panigrahi and J. Taylor. Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association*, pages 1–11, 2022.

[24] K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor. Convex clustering shrinkage. In *PASCAL workshop on statistics and optimization of clustering workshop*, 2005.

[25] D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *Journal of Machine Learning Research*, 22:9–1, 2021.

[26] K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electronic journal of statistics*, 9(2):2324, 2015.

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

[28] X. Tian, J. R. Loftus, and J. E. Taylor. Selective inference with unknown variance via

the square-root lasso. *Biometrika*, 105(4):755–768, 2018.

[29] R. J. Tibshirani, A. Rinaldo, R. Tibshirani, and L. Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics*, 46(3):1255 – 1287, 2018.

[30] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011.

[31] B. Wang, Y. Zhang, W. W. Sun, and Y. Fang. Sparse convex clustering. *Journal of Computational and Graphical Statistics*, 27(2):393–403, 2018.

[32] X. Wang, C. Yang, and J. Zhou. Clustering aggregation by probability accumulation. *Pattern Recognition*, 42(5):668–675, 2009.

[33] M. Weylandt, J. Nagorski, and G. I. Allen. Dynamic visualization and fast computation for convex clustering via algorithmic regularization. *Journal of Computational and Graphical Statistics*, 29(1):87–96, 2020.

[34] J. W. J. Williams. Algorithm 232: heapsort. *Commun. ACM*, 7:347–348, 1964.

[35] Y. Yun and R. F. Barber. Selective inference for clustering with unknown variance. *arXiv preprint arXiv:2301.12999*, 2023.

[36] J. M. Zhang, G. M. Kamath, and N. T. David. Valid post-clustering differential analysis for single-cell RNA-seq. *Cell systems*, 9(4):383–392, 2019.

[37] X. Zhou, C. Du, and X. Cai. An efficient smoothing proximal gradient algorithm for convex clustering. *arXiv preprint arXiv:2006.12592*, 2020.

Appendix A: Technical lemmas and their proofs

Lemma 11. *Consider a fixed* $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$. *Let, for* $\boldsymbol{B} = (B_1, \dots, B_n) \in \mathbb{R}^n$,

$$R(\boldsymbol{B}) = ||\boldsymbol{B} - \boldsymbol{x}||_2^2.$$

Then, for $i, i' \in [|n|], i \neq i'$ *such that* $x_i = x_{i'}$, *if* $\boldsymbol{B}$ *is such that* $B_i \neq B_{i'}$, *replacing* B_i *and* $B_{i'}$ *by* $(B_i + B_{i'})/2$ *strictly decreases* $R(\boldsymbol{B})$. *Furthermore, for* $i, i' \in [|n|], i \neq i'$ *such that* $x_i < x_{i'}$, *if* $B_i > B_{i'}$, *exchanging* B_i *and* $B_{i'}$ *in* $\boldsymbol{B}$ *strictly decreases* $R(\boldsymbol{B})$.

Proof of Lemma 11. In the first case, we compute the change of $R(\boldsymbol{B})$,

$$\text{before} - \text{after} = (B_i - x_i)^2 + (B_{i'} - x_i)^2 - 2\left(\frac{B_i + B_{i'}}{2} - x_i\right)^2$$

which is strictly positive by strict convexity and because $B_i \neq B_{i'}$. In the second case,

$$\begin{aligned}
\text{before} - \text{after} &= (B_i - x_i)^2 + (B_{i'} - x_{i'})^2 - (B_i - x_{i'})^2 - (B_{i'} - x_i)^2 \\
&= -2B_i x_i - 2B_{i'} x_{i'} + 2B_i x_{i'} + 2B_{i'} x_i \\
&= 2(B_i - B_{i'})(x_{i'} - x_i) \\
&> 0.
\end{aligned}$$

□

Lemma 12. *Let* $k \in \mathbb{N}$. *Let* $f : \mathbb{R}^k \to \mathbb{R}$ *be convex and continuously differentiable. Let* $g : \mathbb{R}^k \to \mathbb{R}$ *be convex and continuous. Let* $\boldsymbol{x} \in \mathbb{R}^k$. *For a continuously differentiable function* $\psi : \mathbb{R}^k \to \mathbb{R}$, *we let* $\mathrm{Lin}_{\boldsymbol{x}}(\psi)$ *be the function* $\mathbf{t} \mapsto \nabla_\psi(\boldsymbol{x})^\top(\mathbf{t} - \boldsymbol{x})$, *letting* $\nabla_\psi(\boldsymbol{x})$ *be the gradient of* ψ *at* $\boldsymbol{x}$. *Then* $\boldsymbol{x}$ *is a minimizer of* $\mathrm{Lin}_{\boldsymbol{x}}(f) + g$ *if and only if* $\boldsymbol{x}$ *is a minimizer of* $f + g$.

Proof of Lemma 12. For a convex function ϕ, $\boldsymbol{x}$ is a minimizer of ϕ if and only if, for any $\boldsymbol{v} \in \mathbb{R}^k$,

$$\lim_{\substack{u \to 0 \\ u > 0}} \frac{\phi(\boldsymbol{x} + u\boldsymbol{v}) - \phi(\boldsymbol{x})}{u} \geq 0.$$

For any $\boldsymbol{v} \in \mathbb{R}^k$, the above limit is identical when $\phi = f + g$ and when $\phi = \mathrm{Lin}_{\boldsymbol{x}}(f) + g$. Hence the above limit is non-negative when $\phi = f + g$ if and only if it is non-negative when $\phi = \mathrm{Lin}_{\boldsymbol{x}}(f) + g$. □

Lemma 13. *Let* $n \in \mathbb{N}$ *and* $(a_1, \dots, a_n) \in \mathbb{R}^n$ *with* $\sum_{i=1}^n a_i = 0$. *Then the function*

$$\begin{aligned}
g : \mathbb{R}^n &\to \mathbb{R} \\
(u_1, \dots, u_n) &\mapsto \sum_{i=1}^n a_i u_i + \sum_{\substack{i,i'=1 \\ i<i'}}^n |u_i - u_{i'}|
\end{aligned}$$

is minimal at 0 *if and only if, with* $a_{[1]} \leq \cdots \leq a_{[n]}$ *the ordered values of* $a_1, \dots, a_n$, *for* $\ell \in [|n-1|]$,

$$\sum_{i=1}^{\ell} a_{[i]} + \ell(n - \ell) \geq 0.$$

Proof of Lemma 13. We write $b_1 \leq \cdots \leq b_n$ for the ordered values of $a_1, \ldots, a_n$. Then g is minimal at 0 if and only if h is minimal at 0 with

$$\begin{aligned} h : \mathbb{R}^n &\to \mathbb{R} \\ (u_1, \ldots, u_n) &\mapsto \sum_{i=1}^n b_i u_i + \sum_{\substack{i,i'=1 \\ i<i'}}^n |u_i - u_{i'}|. \end{aligned}$$

The minimum of h is reached when $u_1, \ldots, u_n$ satisfy $u_1 \geq \cdots \geq u_n$. Indeed if there is $i < i'$ with $u_i < u_{i'}$, we can swap u_i and $u_{i'}$ which lets the sum of absolute values unchanged and changes the linear combination as

$$\text{before} - \text{after} = b_i u_i + b_{i'} u_{i'} - b_i u_{i'} - b_{i'} u_i = (b_i - b_{i'})(u_i - u_{i'}) \geq 0.$$

We can do this swap each time there is $i < i'$ with $u_i < u_{i'}$, until we have $u_1 \geq \cdots \geq u_n$ and g has not been increased. Hence, to minimize h it is sufficient to consider $u_1 \geq \cdots \geq u_n$.

Let $v_\ell = u_\ell - u_{\ell+1} \geq 0$ for $\ell \in [|n-1|]$. We have

$$\begin{aligned} \sum_{i=1}^n b_i u_i &= u_n \sum_{i=1}^n b_i + \sum_{\ell=1}^{n-1} v_\ell \left(\sum_{i=1}^\ell b_i \right) \\ &= \sum_{\ell=1}^{n-1} v_\ell \left(\sum_{i=1}^\ell b_i \right) \end{aligned}$$

since by assumption $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = 0$. We also have

$$\begin{aligned} \sum_{\substack{i,i'=1 \\ i<i'}}^n |u_i - u_{i'}| &= \sum_{\substack{i,i'=1 \\ i<i'}}^n \sum_{\ell=i}^{i'-1} v_\ell \\ &= \sum_{\ell=1}^{n-1} v_\ell \ell (n-\ell). \end{aligned}$$

Therefore,

$$\sum_{i=1}^n b_i u_i + \sum_{\substack{i,i'=1 \\ i<i'}}^n |u_i - u_{i'}| = \sum_{\ell=1}^{n-1} v_\ell \left(\sum_{i=1}^\ell b_i + \ell(n-\ell) \right).$$

Hence h is minimal at 0 if and only if, for $\ell \in [|n-1|]$, $\sum_{i=1}^\ell b_i + \ell(n-\ell) \geq 0$. □

Appendix B: Proofs for Section 2

Proof of Lemma 1. For the first part, let $i, i' \in [|n|], i \neq i'$ such that $x_i = x_{i'}$ and assume that $\widehat{B}_i \neq \widehat{B}_{i'}$. Let us consider the increment of the criterion in (5) when replacing $\widehat{B}_i$ and $\widehat{B}_{i'}$ by $(\widehat{B}_i + \widehat{B}_{i'})/2$. From Lemma 11, the quadratic part is strictly decreased. Let us show that

the absolute value part is decreased. This will lead to a contradiction since there is a unique minimizer in (5) by strict convexity. The increment of the absolute value part is given by

$$\text{before} - \text{after} = |\widehat{B}_i - \widehat{B}_{i'}| + \sum_{\substack{\iota=1 \\ \iota \notin \{i,i'\}}}^{n} \left(\left|\widehat{B}_\iota - \widehat{B}_i\right| + \left|\widehat{B}_\iota - \widehat{B}_{i'}\right| - 2\left|\widehat{B}_\iota - \frac{\widehat{B}_i + \widehat{B}_{i'}}{2}\right| \right).$$

In the right-hand side above, $|\widehat{B}_i - \widehat{B}_{i'}| > 0$ and the second sum is non-negative by convexity. This conclude the proof of the first part.

For the second part, let $i, i' \in [|n|], i \neq i'$ such that $x_i > x_{i'}$ and assume that $\widehat{B}_i < \widehat{B}_{i'}$. Let us consider again the increment of the criterion in (5) obtained by exchanging $\widehat{B}_i$ and $\widehat{B}_{i'}$. From Lemma 11, the quadratic part is strictly decreased. The absolute value part is left unchanged and thus the criterion in (5) is strictly decreased which is a contradiction as before. □

Proof of Theorem 2.

Proof that (6) and (7) imply (8),(9),(10)

By (6), for any $k \in [|K|]$, all the $\widehat{B}_i$ for $i \in \mathcal{C}_k$ are identical to a value that we denote by $\widehat{b}_k$, with $\widehat{b}_1, \dots, \widehat{b}_K$ two-by-two distinct. By Definition 1, Lemma 1 and (7), we have $\widehat{b}_1 > \cdots > \widehat{b}_K$. With this notation, the vector $(\widehat{b}_k)_{k \in [|K|]}$ is locally solution of

$$\min_{(b_k)_k} \quad \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (b_k - x_i)^2 + \lambda \sum_{\substack{k,k'=1 \\ k'>k}}^{K} n_k n_{k'} (b_k - b_{k'}).$$

Indeed, in (5) we can assign to all the $(B_i)_{i \in \mathcal{C}_k}$ the same new value b_k close to $\widehat{b}_k$, and we have $|B_i - B_{i'}| = b_k - b_{k'}$ for all $i \in \mathcal{C}_k, i' \in \mathcal{C}_{k'}$, $k < k'$. Canceling the gradient with respect to $b_1, \dots, b_K$ at $\widehat{b}_1, \dots, \widehat{b}_K$ then provides, for $k \in [|K|]$,

$$\sum_{i \in \mathcal{C}_k} \left(\widehat{b}_k - x_i \right) - \lambda \sum_{k'=1}^{k-1} n_k n_{k'} + \lambda \sum_{k'=k+1}^{K} n_k n_{k'} = 0.$$

This provides

$$\widehat{b}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} x_i + \lambda \sum_{k'=1}^{k-1} n_{k'} - \lambda \sum_{k'=k+1}^{K} n_{k'} \tag{20}$$

$$= \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} x_i + \lambda t_{k-1} - \lambda (t_K - t_k). \tag{21}$$

Hence, we have for $k, k' \in [|K|], k < k'$:

$$\widehat{b}_k - \widehat{b}_{k'} = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} x_i - \frac{1}{n_{k'}} \sum_{i \in \mathcal{C}_{k'}} x_i + \lambda (t_{k-1} - t_{k'-1}) + \lambda (t_k - t_{k'}),$$

so that (8) holds by the previous observation that $\widehat{b}_1 > \cdots > \widehat{b}_K$ and taking $k' = k + 1$.

Now we fix $k \in [|K|]$. If we replace $\widehat{B}_i = \widehat{b}_k$ by $\widehat{b}_k + U_i$ for $i \in \mathcal{C}_k$ and we keep the $\widehat{B}_i$, $i \notin \mathcal{C}_k$ unchanged, we increase the cost function in Problem (5). Hence the following function of $(U_i)_{i\in\mathcal{C}_k}$

$$\frac{1}{2}\sum_{i\in\mathcal{C}_k}\left(\widehat{b}_k + U_i - x_i\right)^2 + \lambda\sum_{\substack{k'=1\\k'\neq k}}^{K}\sum_{i\in\mathcal{C}_k} n_{k'}\ \mathrm{sign}(k'-k)\left(\widehat{b}_k - \widehat{b}_{k'} + U_i\right)$$
$$+\lambda\sum_{\substack{i,i'\in\mathcal{C}_k\\ i<i'}}|U_i - U_{i'}|$$

is minimal locally around 0. Above, we let $\mathrm{sign}(t) = 1$ if $t > 0$, $\mathrm{sign}(0) = 0$ and $\mathrm{sign}(t) = -1$ if $t < 0$. From Lemma 12, this implies that the function

$$\sum_{i\in\mathcal{C}_k}\left(\widehat{b}_k - x_i\right)U_i + \lambda\sum_{\substack{k'=1\\k'\neq k}}^{K}\sum_{i\in\mathcal{C}_k} n_{k'}\ \mathrm{sign}(k'-k)U_i + \lambda\sum_{\substack{i,i'\in\mathcal{C}_k\\ i<i'}}|U_i - U_{i'}| \tag{22}$$

of $(U_i)_{i\in\mathcal{C}_k}$ has a local minimum at zero. From (20), this function is

$$\sum_{i\in\mathcal{C}_k}\left(\left(\frac{1}{n_k}\sum_{i'\in\mathcal{C}_k}x_{i'}\right) - x_i\right)U_i + \lambda\sum_{\substack{i,i'\in\mathcal{C}_k\\ i<i'}}|U_i - U_{i'}|. \tag{23}$$

If $n_k = 1$ this function is 0. Otherwise, because this function has a local minimum at zero, and because $a_i := \frac{1}{n_k}\left(\sum_{i'\in\mathcal{C}_k} x_{i'}\right) - x_i$ satisfies $a_{\sigma(t_{k-1}+1)} \leq \cdots \leq a_{\sigma(t_k)}$ by (7) , Lemma 13 implies that for all $\ell \in [|n_k|]$,

$$\sum_{i=1}^{\ell}\left[\left(\frac{1}{n_k}\sum_{i'\in\mathcal{C}_k}x_{i'}\right) - x_{\sigma(t_{k-1}+i)}\right] + \lambda\ell(n_k - \ell) \geq 0 \tag{24}$$

so that (9) holds. Note that (24) also holds trivially for $\ell = n_k$. Finally, (10) holds, being identical to (7) .

Proof that (8),(9),(10) imply (6) and (7)

Let $\tilde{b}_k$ be given by the right hand side of (20) for $k \in [|K|]$. Let $\tilde{B}_i = \tilde{b}_k$ for $k \in [|K|]$ and $i \in \mathcal{C}_k$. Let us show that $\widetilde{\boldsymbol{B}} = (\tilde{B}_1, \ldots, \tilde{B}_n)$ provides a minimum of (5) (that is $\widetilde{\boldsymbol{B}} = \widehat{\boldsymbol{B}}$). Note that (8) and (21) provide $\tilde{b}_k > \tilde{b}_{k'}$ for $k < k'$. Then we can write the cost function at $\widetilde{B}_i + U_i, i \in [|n|]$, locally around 0 for $U = (U_1, \ldots, U_n) \in \mathbb{R}^n$, using :

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k}\left(\widetilde{b}_k - x_i + U_i\right)^2 + \lambda\sum_{\substack{k,k'=1\\k'>k}}^{K}\sum_{\substack{i\in\mathcal{C}_k\\ i'\in\mathcal{C}_{k'}}}\left(\widetilde{b}_k - \widetilde{b}_{k'} + U_i - U_{i'}\right) + \lambda\sum_{k=1}^{K}\sum_{\substack{i,i'\in\mathcal{C}_k\\ i<i'}}|U_i - U_{i'}|.$$

From Lemma 12 a sufficient condition to have a local minimum at 0 is to have a local minimum at 0 of

$$\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k}\left(\widetilde{b}_k - x_i\right)U_i + \lambda\sum_{\substack{k,k'=1\\k'>k}}^{K}\sum_{\substack{i\in\mathcal{C}_k\\ i'\in\mathcal{C}_{k'}}}(U_i - U_{i'}) + \lambda\sum_{k=1}^{K}\sum_{\substack{i,i'\in\mathcal{C}_k\\ i<i'}}|U_i - U_{i'}|.$$

This is a sum of functions of $(U_i)_{i\in\mathcal{C}_k}$, the sum being over k. Hence, it is enough that the following function of $(U_i)_{i\in\mathcal{C}_k}$ is locally minimal at 0, for $k \in [|K|]$,

$$\sum_{i\in\mathcal{C}_k}\left(\widetilde{b}_k - x_i\right)U_i + \lambda\sum_{i\in\mathcal{C}_k}\sum_{\substack{k'=1\\k'\neq k}}^{K} n_{k'}\ \mathrm{sign}(k'-k)U_i + \lambda\sum_{\substack{i,i'\in\mathcal{C}_k\\ i<i'}}|U_i - U_{i'}|.$$

This is the same function as in (22) and (23). It is 0 when $n_k = 1$. Otherwise, since the weights of the linear combination of $(U_i)_{i\in\mathcal{C}_k}$ have sum zero and with Condition (9) we indeed have a minimum at $U_i = 0, i \in \mathcal{C}_k$ from Lemma 13. Hence $\widetilde{\boldsymbol{B}}$ as defined above is the global minimizer of (5) (since it is a local minimizer). Hence since we have seen before that $\tilde{b}_k \neq \tilde{b}_{k'}$ for $k \neq k'$, then (6) is satisfied. Finally, (7) holds, being identical to (10).
Proof of (11)**:** This equation was established in (20). □

Proof and full expressions for Lemma 3.

Condition (10): $x_{\sigma(1)} \geq x_{\sigma(2)} \geq \cdots \geq x_{\sigma(n)}$ is equivalent to $\mathbf{M}_1\mathbf{P}_\sigma\boldsymbol{x} \leq \lambda\mathbf{m}_1$ with

$$\mathbf{M}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0\\ 0 & -1 & 1 & 0 & \dots & 0\\ 0 & 0 & -1 & 1 & \dots & 0\\ & \ddots & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{n-1\times n} \tag{25}$$

and $\mathbf{m}_1 = \mathbf{0}_{n-1}$.
Condition (8): For $k \in [|K-1|]$,

$$\frac{1}{n_k}\sum_{i=1}^{n_k} x_{\sigma(t_{k-1}+i)} - \frac{1}{n_{k+1}}\sum_{i=1}^{n_{k+1}} x_{\sigma(t_k+i)} > \lambda(t_{k+1} - t_{k-1})$$

is equivalent to $\mathbf{M}_2(\mathbf{t})\mathbf{P}_\sigma\boldsymbol{x} < \lambda\mathbf{m}_2(\mathbf{t})$, where $\mathbf{M}_2(\mathbf{t}) \in \mathbb{R}^{K-1\times n}$ and $\mathbf{m}_2(\mathbf{t}) \in \mathbb{R}^{K-1}$ are defined by

$$\mathbf{m}_2(\mathbf{t}) = -\left(t_2 - t_0, t_3 - t_1, \dots, t_K - t_{K-2}\right)^\top \tag{26}$$

and

$$\mathbf{M}_2(\mathbf{t}) = \begin{pmatrix} -\frac{1}{n_1} & \dots & -\frac{1}{n_1} & \frac{1}{n_2} & \dots & \frac{1}{n_2} & 0 & \dots & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0\\ 0 & \dots & 0 & -\frac{1}{n_2} & \dots & -\frac{1}{n_2} & \frac{1}{n_3} & \dots & \frac{1}{n_3} & \dots & 0 & \dots & 0 & 0 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & -\frac{1}{n_{K-1}} & \dots & -\frac{1}{n_{K-1}} & \frac{1}{n_K} & \dots & \frac{1}{n_K} \end{pmatrix}, \tag{27}$$

where the number of repetitions of each $\pm 1/n_k$ in each line is n_k.
Condition (9) : For $k \in [|K|]$ such that $n_k \geq 2$, for $\ell \in [|n_k - 1|]$,

$$\frac{1}{n_k}\sum_{i=1}^{n_k} x_{\sigma(t_{k-1}+i)} - \frac{1}{\ell}\sum_{i=1}^{\ell} x_{\sigma(t_{k-1}+i)} \geq \lambda(\ell - n_k)$$

is equivalent to $\mathbf{M}_3(\mathbf{t})\mathbf{P}_\sigma \boldsymbol{x} \leq \lambda \mathbf{m}_3(\mathbf{t})$ where $\mathbf{M}_3(\mathbf{t}) \in \mathbb{R}^{n-K\times n}$ and $\mathbf{m}_3(\mathbf{t}) \in \mathbb{R}^{n-K}$ are as follows. We have

$$\mathbf{m}_3(\mathbf{t}) = (n_1 - 1, n_1 - 2, \ldots, 1, n_2 - 1, \ldots, 1, \ldots, n_K - 1, \ldots, 1)^\top \tag{28}$$

with the convention that $(n_k - 1, n_k - 2, \ldots, 1)$ is empty when $n_k = 1$, and $\mathbf{M}_3(\mathbf{t}) = \mathrm{diag}(\mathbf{M}_3^{(1)}, \ldots, \mathbf{M}_3^{(K)})$ with

$$\mathbf{M}_3^{(k)} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & \ldots & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{n_k-1} & \frac{1}{n_k-1} & \frac{1}{n_k-1} & \ldots & \frac{1}{n_k-1} & 0 \end{pmatrix} - \frac{1}{n_k}\mathbf{1}_{n_k-1\times n_k}, \tag{29}$$

with the convention that $\mathbf{M}_3^{(k)}$ is 0×0 when $n_k = 1$ and where $\mathbf{1}_{n_k-1\times n_k}$ is the $n_k - 1 \times n_k$ matrix composed of ones. □

Proof of Proposition 4. The proposition is obtained from Lemma 5.1 in [14]. We will only show the last claim that the probability that $\mathcal{V}^-(\boldsymbol{Z}) = \mathcal{V}^+(\boldsymbol{Z})$ is zero, conditionally to the event in (13). We have, letting $\mathbb{1}_{(13)}$ denote the indicator function that the event in (13) holds,

$$\mathbb{E}\left[\mathbb{1}_{(13)}\mathbb{1}_{\mathcal{V}^-(\boldsymbol{Z})=\mathcal{V}^+(\boldsymbol{Z})}\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{V}^-(\boldsymbol{Z})=\mathcal{V}^+(\boldsymbol{Z})}\mathbb{E}\left[\mathbb{1}_{(13)}\middle|\, \mathcal{V}^-(\boldsymbol{Z}) = \mathcal{V}^+(\boldsymbol{Z})\right]\right].$$

The above conditional expectation is zero from (13), because $\boldsymbol{\eta}^\top \boldsymbol{X}$ is independent from $\boldsymbol{Z}$ and has non-zero variance because $\boldsymbol{\Sigma}$ is invertible and $\boldsymbol{\eta}$ is non-zero. □

Proof of Proposition 5. The proof follows closely that of Theorem 5.2 in [14]. Fix $t \in [0, 1]$. Remark that in Lemma 3, all the lines of $\mathbf{M}_2(\mathbf{t})$ are non-zero. Furthermore, $\boldsymbol{\Sigma}$ is invertible. This provides, from Theorem 2 and Lemma 3 that the events $E_{\mathbf{t},\sigma}$ and $\{\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}\}$ have their symmetric difference of probability zero. Hence we have

$$\mathbb{P}\left(\mathrm{pval}(\boldsymbol{X}, \mathbf{t}, \sigma) \leq t \middle|\, E_{\mathbf{t},\sigma}\right) = \mathbb{P}\left(\mathrm{pval}(\boldsymbol{X}, \mathbf{t}, \sigma) \leq t \middle|\, \mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}\right).$$

We then have

$$\begin{aligned} &\mathbb{P}\left(\mathrm{pval}(\boldsymbol{X}, \mathbf{t}, \sigma) \leq t \middle|\, E_{\mathbf{t},\sigma}\right) \\ &= \int_{\mathbb{R}^n} \mathbb{P}\left(\mathrm{pval}(\boldsymbol{X}, \mathbf{t}, \sigma) \leq t \middle|\, \mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}, \boldsymbol{Z} = \mathbf{z}_0\right) d\mathbb{P}_{|\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}}(\mathbf{z}_0), \end{aligned}$$

where $d\mathbb{P}_{|\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}}(\mathbf{z}_0)$ denotes the law of $\boldsymbol{Z}$ conditionally to $\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}$.

Consider $\mathbf{z}_0$ in the support of $\mathbb{P}_{|\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}}$ such that $\mathcal{V}^-(\mathbf{z}_0) < \mathcal{V}^+(\mathbf{z}_0)$, which holds with $\mathbb{P}_{|\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}}$-probability one from Proposition 4. Then, as discussed in Section 2.3.1 and shown in [14], conditionally to $\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}$ and $\boldsymbol{Z} = \mathbf{z}_0$, under the null hypothesis $\boldsymbol{\eta}^\top \boldsymbol{\mu} = 0$, $T(\boldsymbol{X}, \mathbf{t}, \sigma)$ is uniformly distributed on $[0, 1]$ and thus so is $\mathrm{pval}(\boldsymbol{X}, \mathbf{t}, \sigma)$. We thus obtain,

$$\begin{aligned} \mathbb{P}\left(\mathrm{pval}(\boldsymbol{X}, \mathbf{t}, \sigma) \leq t \middle|\, E_{\mathbf{t},\sigma}\right) &= \int_{\mathbb{R}^n} t\ d\mathbb{P}_{|\mathbf{M}\mathbf{P}_\sigma \boldsymbol{X} \leq \lambda\ \mathbf{m}}(\mathbf{z}_0) \\ &= t. \end{aligned}$$

This concludes the proof. □

Proof of Proposition 6. Fix $t \in [0,1]$. We have

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{pval}(\boldsymbol{X},\mathbf{T},S) \leq t \middle| (\mathbf{T},S) \in \mathcal{E}, \boldsymbol{\eta}(\mathbf{T},S)^\top \boldsymbol{\mu} = 0\right) \\
&= \sum_{\substack{(\mathbf{t},\sigma)\in\mathcal{E} \\ \boldsymbol{\eta}(\mathbf{t},\sigma)^\top\boldsymbol{\mu}=0 \\ \mathbb{P}((\mathbf{T},S)=(\mathbf{t},\sigma))>0}} \\
&\quad \mathbb{P}\left((\mathbf{T},S) = (\mathbf{t},\sigma) \middle| (\mathbf{T},S) \in \mathcal{E}, \boldsymbol{\eta}(\mathbf{T},S)^\top \boldsymbol{\mu} = 0\right) \mathbb{P}\left(\mathrm{pval}(\boldsymbol{X},\mathbf{t},\sigma) \leq t \middle| (\mathbf{T},S) = (\mathbf{t},\sigma)\right).
\end{aligned}$$

In the conditional probability $\mathbb{P}\left(\mathrm{pval}(\boldsymbol{X},\mathbf{t},\sigma) \leq t \middle| (\mathbf{T},S) = (\mathbf{t},\sigma)\right)$ of the above sum, conditionally to $(\mathbf{T},S) = (\mathbf{t},\sigma)$, one can check that all the conditions of Proposition 5 hold. Hence, from this proposition, we have

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{pval}(\boldsymbol{X},\mathbf{T},S) \leq t \middle| (\mathbf{T},S) \in \mathcal{E}, \boldsymbol{\eta}(\mathbf{T},S)^\top \boldsymbol{\mu} = 0\right) \\
&= \sum_{\substack{(\mathbf{t},\sigma)\in\mathcal{E} \\ \boldsymbol{\eta}(\mathbf{t},\sigma)^\top\boldsymbol{\mu}=0 \\ \mathbb{P}((\mathbf{T},S)=(\mathbf{t},\sigma))>0}} \mathbb{P}\left((\mathbf{T},S) = (\mathbf{t},\sigma) \middle| (\mathbf{T},S) \in \mathcal{E}, \boldsymbol{\eta}(\mathbf{T},S)^\top \boldsymbol{\mu} = 0\right) \times t \\
&= t.
\end{aligned}$$

This concludes the proof. □

Proof of Theorem 7.
We will show that, for the successive values of r, for $k \in [|K^{(r)}|]$, for $\lambda \geq \lambda^{(r)}$,

$$\hat{b}_k^{(r)}(\lambda) = \frac{1}{n_k^{(r)}} \sum_{i \in \mathcal{C}_k^{(r)}} x_i + \lambda \sum_{k'=1}^{k-1} n_{k'}^{(r)} - \lambda \sum_{k'=k+1}^{K^{(r)}} n_{k'}^{(r)}. \tag{30}$$

We will also show that for the successive values of r,

$$\lambda^{(r+1)} = \inf\left\{\lambda \geq \lambda^{(r)};\ \text{there exists } k \in [|K^{(r)} - 1|] \text{ such that } \hat{b}_k^{(r)}(\lambda) = \hat{b}_{k+1}^{(r)}(\lambda)\right\}. \tag{31}$$

We will prove by induction that the following properties $\mathcal{O}^{(r)}$, $\mathcal{P}^{(r)}$, $\mathcal{Q}^{(r)}$ and $\mathcal{R}^{(r)}$ hold for $r = 0, 1, \ldots$ and as long as $K^{(r)} \geq 2$:

$$\begin{aligned}
\mathcal{O}^{(r)} &= \text{“(30) holds for } k \in [|K^{(r)}|] \text{ and } \lambda \geq \lambda^{(r)}\text{”}, \\
\mathcal{P}^{(r)} &= \text{“the set in (31) is non-empty and (31) holds”}, \\
\mathcal{Q}^{(r)} &= \text{“for } \lambda \in [\lambda^{(r)}, \lambda^{(r+1)}), \text{ we have } \hat{b}_1^{(r)}(\lambda) > \cdots > \hat{b}_{K^{(r)}}^{(r)}(\lambda)\text{”}, \\
\mathcal{R}^{(r)} &= \text{“for } \lambda \in [\lambda^{(r)}, \lambda^{(r+1)}), (\hat{B}_i^{(r)}(\lambda))_{i\in[|n|]} \text{ minimizes Problem (5)”}.
\end{aligned}$$

Along proving these properties by induction, we will show that $K^{(r)} > K^{(r+1)}$. Doing this, and discussing the case $r = r_{\max}$ at the end, will conclude the proof.

Initialization: $r = 0$. When $r = 0$, we have $\lambda^{(0)} = 0$ and, for $k \in [|K^{(0)}|]$,

$$\hat{b}_k^{(0)}(\lambda^{(0)}) = \tilde{x}_k = \frac{1}{n_k^{(0)}} \sum_{i \in \mathcal{C}_k^{(0)}} x_i \tag{32}$$

and thus the right-hand sides of (17) and (30) are equal and so $\mathcal{O}^{(0)}$ holds. Furthermore, from (32), $\hat{b}_1^{(0)}(0) > \cdots > \hat{b}_{K^{(0)}}^{(0)}(0)$. We have, for $k \in [|K^{(0)} - 1|]$, using (30),

$$\begin{aligned}\hat{b}_k^{(0)}(\lambda) - \hat{b}_{k+1}^{(0)}(\lambda) &= \tilde{x}_k + \lambda \sum_{k'=1}^{k-1} n_{k'}^{(0)} - \lambda \sum_{k'=k+1}^{K^{(0)}} n_{k'}^{(0)} - \tilde{x}_{k+1} - \lambda \sum_{k'=1}^{k} n_{k'}^{(0)} + \lambda \sum_{k'=k+2}^{K^{(0)}} n_{k'}^{(0)} \\ &= \underbrace{\tilde{x}_k - \tilde{x}_{k+1}}_{>0} - \lambda \underbrace{\left(n_k^{(0)} + n_{k+1}^{(0)}\right)}_{>0}. \end{aligned} \tag{33}$$

Hence, we see that indeed the set in (31) is non-empty.

Let $\tilde{\lambda}^{(1)}$ be given by the right-hand side of (31). The values of $\hat{b}_k^{(0)}(\lambda)$, $k \in [|K^{(0)}|]$, are continuous in λ and thus by definition of $\tilde{\lambda}^{(1)}$ they remain two-by-two distinct and in the same order on $[0, \tilde{\lambda}^{(1)})$. Furthermore , from (33),

$$\tilde{\lambda}^{(1)} = \min_{k \in [|K^{(0)}-1|]} \frac{\tilde{x}_k - \tilde{x}_{k+1}}{n_k^{(0)} + n_{k+1}^{(0)}} = \lambda^{(0)} + \min_{k \in [|K^{(0)}-1|]} \frac{\hat{b}_k^{(0)}(\lambda^{(0)}) - \hat{b}_{k+1}^{(0)}(\lambda^{(0)})}{n_k^{(0)} + n_{k+1}^{(0)}} = \lambda^{(1)}.$$

Hence indeed (31) holds and thus $\mathcal{P}^{(0)}$ holds. Since $\lambda^{(1)}$ is given by (31), then also $\mathcal{Q}^{(0)}$ holds.

Let us now show $\mathcal{R}^{(0)}$. Let $\lambda \in [\lambda^{(0)}, \lambda^{(1)})$. We will apply Theorem 2, with σ there being a permutation such that $x_{\sigma(1)} \geq \cdots \geq x_{\sigma(n)}$ and $\mathcal{C}$ being the clustering $\mathcal{C}^{(0)}$. For $k \in [|K^{(0)} - 1|]$, since $\hat{b}_k^{(0)}(\lambda) - \hat{b}_{k+1}^{(0)}(\lambda) > 0$ as seen above, we obtain from (30) that (8) holds, using (20) and (21). It is immediate that (9) holds because the left-hand term is zero and the right-hand term is non-positive. Hence from (11) in Theorem 2, $\mathcal{R}^{(0)}$ indeed holds, also from (30). Also, $K^{(1)} < K^{(0)}$ because, by definition of $\lambda^{(1)}$ in (31), the values $\hat{b}_1^{(0)}(\lambda^{(1)}), \ldots, \hat{b}_{K^{(0)}}^{(0)}(\lambda^{(1)})$ are not two-by-two distinct.

Induction: from r to $r+1$. Let now $r \in \mathbb{N}$ such that $K^{(r+1)} \geq 2$. Assume that $\mathcal{O}^{(r)}$, $\mathcal{P}^{(r)}$, $\mathcal{Q}^{(r)}$ and $\mathcal{R}^{(r)}$ hold. For any $B \in \mathbb{R}^n$, from $\mathcal{R}^{(r)}$, for $\lambda \in [\lambda^{(r)}, \lambda^{(r+1)})$,

$$\frac{1}{2} \sum_{i=1}^{n} \left(\hat{B}_i^{(r)}(\lambda) - x_i\right)^2 + \lambda \sum_{\substack{i,i'=1 \\ i<i'}}^{n} \left|\hat{B}_i^{(r)}(\lambda) - \hat{B}_{i'}^{(r)}(\lambda)\right| \leq \frac{1}{2} \sum_{i=1}^{n} (B_i - x_i)^2 + \lambda \sum_{\substack{i,i'=1 \\ i<i'}}^{n} |B_i - B_{i'}|.$$

As $\lambda \to \lambda^{(r+1)}$, this yields

$$\begin{aligned}&\frac{1}{2} \sum_{i=1}^{n} \left(\hat{B}_i^{(r)}(\lambda^{(r+1)}) - x_i\right)^2 + \lambda^{(r+1)} \sum_{\substack{i,i'=1 \\ i<i'}}^{n} \left|\hat{B}_i^{(r)}(\lambda^{(r+1)}) - \hat{B}_{i'}^{(r)}(\lambda^{(r+1)})\right| \\ &\leq \frac{1}{2} \sum_{i=1}^{n} (B_i - x_i)^2 + \lambda^{(r+1)} \sum_{\substack{i,i'=1 \\ i<i'}}^{n} |B_i - B_{i'}|.\end{aligned}$$

Hence, the minimizer of (5) for $\lambda = \lambda^{(r+1)}$ is equal to $\hat{B}_i^{(r)}(\lambda^{(r+1)})_{i \in [|n|]}$. Hence, $\mathcal{C}^{(r+1)}$ is the clustering obtained by minimizing (5). We can see from the successive definitions of $(\hat{b}_k^{(r)}(\lambda^{(r+1)}))_{k \in [|K^{(r)}|]}$, $(\hat{b}_k^{(r+1)}(\lambda^{(r+1)}))_{k \in [|K^{(r+1)}|]}$ and $(\hat{B}_i^{(r+1)}(\lambda^{(r+1)}))_{i \in [|n|]}$ that $(\hat{B}_i^{(r)}(\lambda^{(r+1)}))_{i \in [|n|]} = (\hat{B}_i^{(r+1)}(\lambda^{(r+1)}))_{i \in [|n|]}$.

Hence, from (11) in Theorem 2, the right-hand side of (30), with (r, λ) there replaced by $(r+1, \lambda^{(r+1)})$ and for $k \in [|K^{(r+1)}|]$, is equal to $\hat{b}_k^{(r+1)}(\lambda^{(r+1)})$. Hence (30) holds at step $r+1$ for $\lambda = \lambda^{(r+1)}$. Hence (30) holds for $\lambda \geq \lambda^{(r+1)}$ since the right-hand-sides of (17) and (30) have the same slope w.r.t. λ. Thus $\mathcal{O}^{(r+1)}$ is proved. The properties $\mathcal{P}^{(r+1)}$ and $\mathcal{Q}^{(r+1)}$ are shown similarly as in the initialization step.

Let us finally show $\mathcal{R}^{(r+1)}$. Let $\lambda \in [\lambda^{(r+1)}, \lambda^{(r+2)})$. Similarly as for the initialization step, we will apply Theorem 2, with the same permutation σ and with $\mathcal{C}$ being the clustering $\mathcal{C}^{(r+1)}$. Equation (8) is shown to hold similarly as before, using (30). From $\mathcal{R}^{(r)}$ and with the above, we obtain that $\left(\hat{B}_i^{(r)}(\lambda^{(r+1)})\right)_{i\in[|n|]} = \left(\hat{B}_i^{(r+1)}(\lambda^{(r+1)})\right)_{i\in[|n|]}$ minimizes (5) when $\lambda = \lambda^{(r+1)}$. Hence Equation (9) holds when $\lambda = \lambda^{(r+1)}$ from Theorem 2. For $\lambda \in [\lambda^{(r+1)}, \lambda^{(r+2)})$, the clustering is the same as when $\lambda = \lambda^{(r+1)}$ so the left-hand-side of (9) is unchanged compared to when $\lambda = \lambda^{(r+1)}$. On the other hand, the right-hand side is decreased. Hence, (9) also holds for $\lambda \in [\lambda^{(r+1)}, \lambda^{(r+2)})$. Hence from (11) in Theorem 2, and (30), $\mathcal{R}^{(r+1)}$ indeed holds.

When $r = r_{\max}$. As before, we show that $\mathcal{O}^{(r_{\max}-1)}$ implies $\mathcal{O}^{(r_{\max})}$. Then using (30) for $r_{\max}$ we obtain, by the same arguments as when showing $\mathcal{R}^{(r+1)}$ above, that for $\lambda \geq \lambda^{(r_{\max})}$, $(\hat{B}_i^{(r_{\max})}(\lambda))_{i\in[|n|]}$ minimizes (5). Note that the right-hand-side of (30) is constant in λ now and there is a single class. Hence the common value of $(\hat{B}_i^{(r_{\max})})_{i\in[|n|]}$ minimizes (5) as $\lambda \to \infty$, so this value is $\sum_{i=1}^n x_i/n$. □

Appendix C: Proofs for Section 3

Proof of Proposition 8. Computing the invariant statistic as described in Section 3.2.1 requires computing $\mathrm{vec}(\overline{\boldsymbol{Z}}) := [\boldsymbol{I}_{np} - \mathrm{vec}(\overline{\boldsymbol{c}})\boldsymbol{\kappa}^\top]\mathrm{vec}(\boldsymbol{Y})$ with $\mathrm{vec}(\overline{\boldsymbol{c}}) = (\boldsymbol{I}_p \otimes \boldsymbol{\Sigma})\boldsymbol{\kappa}(\boldsymbol{\kappa}^\top(\boldsymbol{I}_p \otimes \boldsymbol{\Sigma})\boldsymbol{\kappa})^{-1}$, similarly as $\boldsymbol{Z}$ and $\boldsymbol{c}$ in Proposition 4. Using the properties of Kronecker products, we have, letting $\boldsymbol{e}_{j_0}$ be the j_0th base column vector in $\mathbb{R}^p$,

$$\begin{aligned}
\mathrm{vec}(\overline{\boldsymbol{c}}) =& (\boldsymbol{I}_p \otimes \boldsymbol{\Sigma})(\boldsymbol{e}_{j_0} \otimes \boldsymbol{\eta})(\boldsymbol{\kappa}^\top(\boldsymbol{I}_p \otimes \boldsymbol{\Sigma})\boldsymbol{\kappa})^{-1} \\
=& (\boldsymbol{e}_{j_0} \otimes \boldsymbol{\Sigma}\boldsymbol{\eta})(\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-1} \\
=& \boldsymbol{e}_{j_0} \otimes \boldsymbol{c},
\end{aligned}$$

where $\boldsymbol{c} = \boldsymbol{\Sigma}\boldsymbol{\eta}(\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-1}$ is as defined for the one-dimensional case in Proposition 4. Hence, $\mathrm{vec}(\overline{\boldsymbol{c}})$ is a $np \times 1$ vector where the subvector corresponding to the variable j_0 is equal to $\boldsymbol{c}$ and the subvectors corresponding to the other variables are zero. Then,

$$\begin{aligned}
\mathrm{vec}(\overline{\boldsymbol{Z}}) =& \mathrm{vec}(\boldsymbol{Y}) - \mathrm{vec}(\overline{\boldsymbol{c}})\boldsymbol{\kappa}^\top\mathrm{vec}(\boldsymbol{Y}) \\
=& \mathrm{vec}(\boldsymbol{Y}) - (\boldsymbol{e}_{j_0} \otimes \boldsymbol{c})(\boldsymbol{e}_{j_0}^\top \otimes \boldsymbol{\eta}^\top)\mathrm{vec}(\boldsymbol{Y}) \\
=& \mathrm{vec}(\boldsymbol{Y}) - \left((\boldsymbol{e}_{j_0}\boldsymbol{e}_{j_0}^\top) \otimes (\boldsymbol{c}\boldsymbol{\eta}^\top)\right)\mathrm{vec}(\boldsymbol{Y}) \\
=& \mathrm{vec}(\boldsymbol{Y}) - \boldsymbol{e}_{j_0} \otimes \left(\boldsymbol{c}\boldsymbol{\eta}^\top\boldsymbol{Y}_{\cdot j_0}\right) \\
=& \boldsymbol{e}_{j_0} \otimes \boldsymbol{Z} + \mathrm{vec}(\boldsymbol{Y}_{\cdot -j_0}),
\end{aligned}$$

where $\boldsymbol{Z} = [\boldsymbol{I}_n - \boldsymbol{c}\boldsymbol{\eta}^\top]\boldsymbol{Y}_{\cdot j_0}$ is defined as in Proposition 4 and $\boldsymbol{Y}_{\cdot -j_0}$ is defined by replacing the column j_0 of $\boldsymbol{Y}$ by zero. Hence, $\mathrm{vec}(\overline{\boldsymbol{Z}})$ is a $np \times 1$ vector which subvector corresponding to the variable j_0 is $\boldsymbol{Z}$ which is computed as in the one-dimensional case.

The next step for obtaining the invariant statistic is to compute

$$\mathcal{V}^-(\mathrm{vec}(\overline{\boldsymbol{Z}})) := \max_{l:(\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{c}}))_l<0} \frac{\lambda m_l - (\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{Z}}))_l}{(\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{c}}))_l}$$

and

$$\mathcal{V}^+(\mathrm{vec}(\overline{\boldsymbol{Z}})) := \min_{l:(\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{c}}))_l>0} \frac{\lambda m_l - (\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{Z}}))_l}{(\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{c}}))_l}.$$

In $\mathcal{V}^-(\mathrm{vec}(\overline{\boldsymbol{Z}}))$ the set of indices l is the disjoint union of p sets of cardinality $2(n-1)$ each, corresponding to the p variables. Consider l in the set corresponding to a variable $j \neq j_0$. Then in $(\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{c}}))_l$, the row l of $\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma$, of size np has non-zero components only for the indices corresponding to the variable j. On the other hand, as seen above, $\mathrm{vec}(\overline{\boldsymbol{c}})$ has non-zero components only for the indices corresponding to the variable j_0. Hence, taking the inner product, $(\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\overline{\boldsymbol{c}}))_l = 0$. Hence the maximum in $\mathcal{V}^-(\mathrm{vec}(\overline{\boldsymbol{Z}}))$ can simply be taken with the indices l corresponding to the variable j_0. This, together with the expressions of $\mathrm{vec}(\overline{\boldsymbol{c}})$ and $\mathrm{vec}(\overline{\boldsymbol{Z}})$ above yields

$$\mathcal{V}^-(\mathrm{vec}(\overline{\boldsymbol{Z}})) = \max_{l:(\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{c})_l<0} \frac{\lambda \mathbf{m}(\mathbf{t}^{(j_0)})_l - (\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{Z})_l}{(\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{c})_l} := \mathcal{V}^-(\boldsymbol{Z}),$$

where $\mathcal{V}^-(\boldsymbol{Z})$ has the same expression as in Proposition 4 for the one-dimensional case. We obtain similarly

$$\mathcal{V}^+(\mathrm{vec}(\overline{\boldsymbol{Z}})) = \min_{l:(\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{c})_l>0} \frac{\lambda \mathbf{m}(\mathbf{t}^{(j_0)})_l - (\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{Z})_l}{(\mathbf{M}(\mathbf{t}^{(j_0)})\mathbf{P}_{\sigma^{(j_0)}}\boldsymbol{c})_l} := \mathcal{V}^+(\boldsymbol{Z}).$$

The invariant statistic is thus, similarly as in Section 2.3.1,

$$T(\boldsymbol{Y}) = F_{0,\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta}}^{[\mathcal{V}^-(\boldsymbol{Z}),\mathcal{V}^+(\boldsymbol{Z})]}(\boldsymbol{\eta}^\top \boldsymbol{Y}_{\cdot j_0}).$$

This concludes the proof. □

Proof of Proposition 9. From Theorem 2 and Lemma 3, for $j \in [|p|]$, the event

$$\left\{ \mathcal{C}^{(j)} \text{ is the clustering given by } \widehat{\boldsymbol{B}}_{\cdot j} \text{ and } Y_{\sigma^{(j)}(1)j} \geq \cdots \geq Y_{\sigma^{(j)}(n)j} \right\}$$

is equal to the event $\left\{\mathbf{M}(\mathbf{t}^{(j)})\mathbf{P}_{\sigma^{(j)}}\boldsymbol{Y}_{\cdot j} \leq \lambda \mathbf{m}(\mathbf{t}^{(j)})\right\}$, up to a symmetric difference of $\boldsymbol{Y}$-probability 0 (because the rows of $\mathbf{M}(\mathbf{t}^{(j)})$ are non-zero and the covariance matrix of $\boldsymbol{Y}_{\cdot j}$ is invertible). Hence, up to a symmetric difference of $\boldsymbol{Y}$-probability 0, the event E is equal to the event $\{\boldsymbol{\mathscr{M}}\mathbf{D}_\sigma \mathrm{vec}(\boldsymbol{Y}) \leq \lambda \boldsymbol{m}\}$, with the construction of Section 3.2.1. The rest of the proof is the same as the proof of Proposition 5. □

Proof of Proposition 10. The proof is the same as for Proposition 6. □

Appendix D: Time complexity of convex clustering

D.1. Benchmarking existing implementations of convex clustering

In this section we compare the observed time complexities of existing implementations of one-dimensional convex clustering in the `R` language:

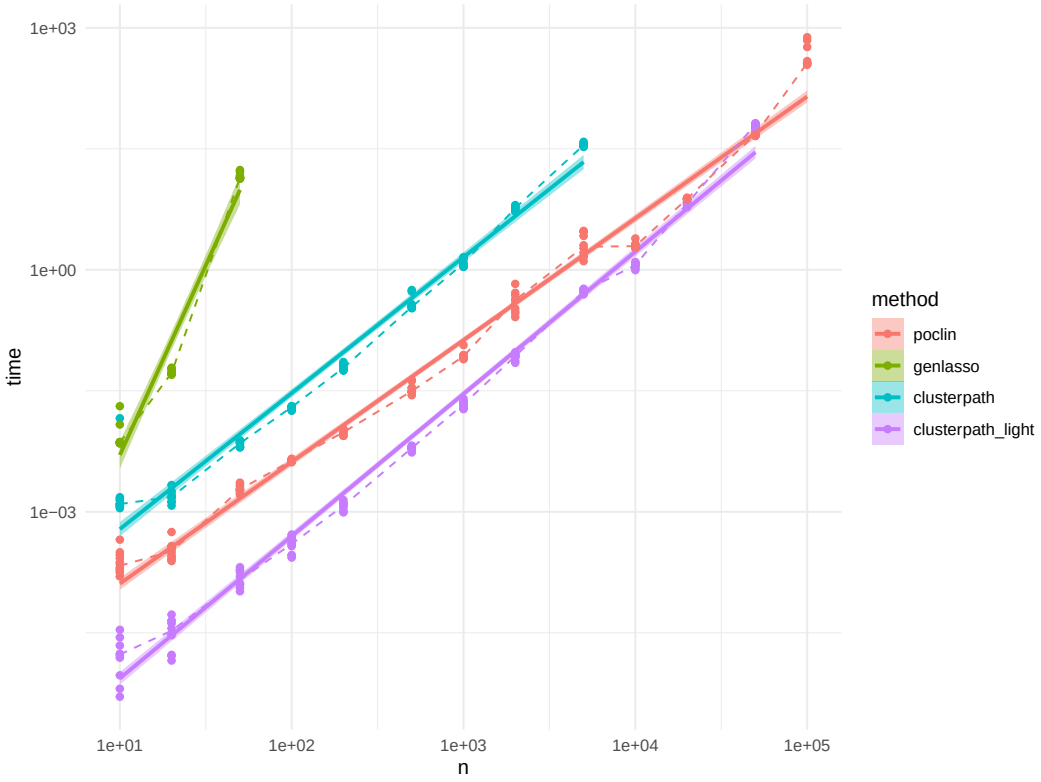

Figure 6: Comparison of the empirical complexities of existing `R` implementations of convex clustering. The axes are on a logarithmic scale. Each dot represents one experimental run. For a given method, the median computation times for a given problem size are connected by dashed lines. The solid lines have been obtained by a linear regression of time against problem size (on the log scale).

- the `convex_clustering_1d` method in our `R` package `poclin`, which is available from https://plmlab.math.cnrs.fr/pneuvial/poclin;
- the `genlasso` function in the `R` package `genlasso`, which is available from CRAN at https://CRAN.R-project.org/package=genlasso;
- the `clusterpath.l1.id` function in the `R` package `clusterpath`, which is available from `R`-forge at https://clusterpath.r-forge.r-project.org/. The core functions of this package are implemented in `C`.

We have used the `R` package `microbenchmark` to compare the execution time of these implementations on standard Gaussian signal of size $n \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000\}$. The results are displayed in Figure 6 on the log-log scale. Each dot represents one experimental run. For a given method, the median computation times for a given problem size are connected by dashed lines. The solid lines have been obtained by a linear regression of time against problem size (on the log scale).

The computation time of `genlasso` is much larger than for the other implementations; we were not able to get results for $n \geq 100$ with this method. This is explained by the fact that `genlasso` is a generic implementation where the constraints are stored in a $n(n-1)/2 \times n$

matrix. In contrast, the other `clusterpath` and `poclin` implementations are quite efficient. For `clusterpath`, we report two computational times, which are labeled as 'clusterpath' and 'clusterpath_light' in Figure 6, respectively:

- 'clusterpath' corresponds to the direct application of the `clusterpath.l1.id` function. We were not able to include 'clusterpath' for $n \geq 10000$ due to memory issues – one run of this function for $n = 5000$ takes up 6.6 Gb of RAM;
- 'clusterpath_light' directly calls the underlying `C` function `join_clusters_convert` in the `clusterpath` package, thereby avoiding some computational overhead. Thanks to this modification, we were able to run 'clusterpath_light' for $n \leq 50000$. However, it was not possible to run it for $n \geq 100000$ because of memory issues.

In contrast, since the space complexity of `poclin` is linear, we were able to run `poclin` without any memory issue for $n \leq 100000$. The computational time of `poclin` is slightly higher than that of 'clusterpath_light' for $n \leq 20000$. However, we note that the implementation in `poclin` only uses `R` code, while 'clusterpath_light' only uses `C` code: we expect that a `C` implementation of `poclin` would lead to improve computational times. More interestingly, a linear regression in the log/log space showed that the slope of the `poclin` curve is approximately 1.5, while that of 'clusterpath_light' and 'clusterpath' are approximately 1.9. This implies that the empirical complexity of `poclin` is of the order of $\mathcal{O}(n^{1.5})$, while that of `clusterpath` is of the order of $\mathcal{O}(n^{1.9})$.

D.2. Further reducing the time complexity of Algorithm 1

The complexity of Algorithm 1 can be reduced to the order $\mathcal{O}(n \log(n))$ without compromising the linear memory complexity. This section gives an informal description of the main idea for this reduction. We consider the pairs of consecutive clusters, associated to consecutive values of $\hat{b}$ in Algorithm 1. Let us define the "merging distance" of each of these pairs as the value of λ for which the corresponding value of $\hat{b}$ become equal, that is, where this pair of clusters should be merged into one. If two clusters are merged, the merging distances are updated *only for these two clusters and the one or two adjacent ones*. This property could be exploited in the implementation of Algorithm 1, by storing these merging distances in a min heap binary tree [34]. Indeed, the minimal element of a min heap (here, corresponding to the next merge), is obtained in constant time ($\mathcal{O}(1)$) as the root of the tree, while the cost of inserting an element in the heap is logarithmic ($\mathcal{O}(\log(n))$), corresponding to the depth of the binary heap. Exploiting the binary min-heap tree, we can keep a $\mathcal{O}(\log(n))$ cost at each step when two clusters are merged. This yields a total computational complexity of $\mathcal{O}(n \log(n))$ for $\mathcal{O}(n)$ steps.

Note also that if more than two clusters are merged, then the computational cost of the corresponding step can be higher, but the total number of steps is more reduced. We eschew a full description of an implementation of Algorithm 1 with a binary min-heap tree for the sake of concision and to promote explicit formulas such as (17) and (18).

Appendix E: Additional illustrations

E.1. Calibration of the regularization parameter

We describe the procedure used in the numerical experiments to calibrate the value of the regularization parameter λ. As explained in the main text, the goal of this procedure is to

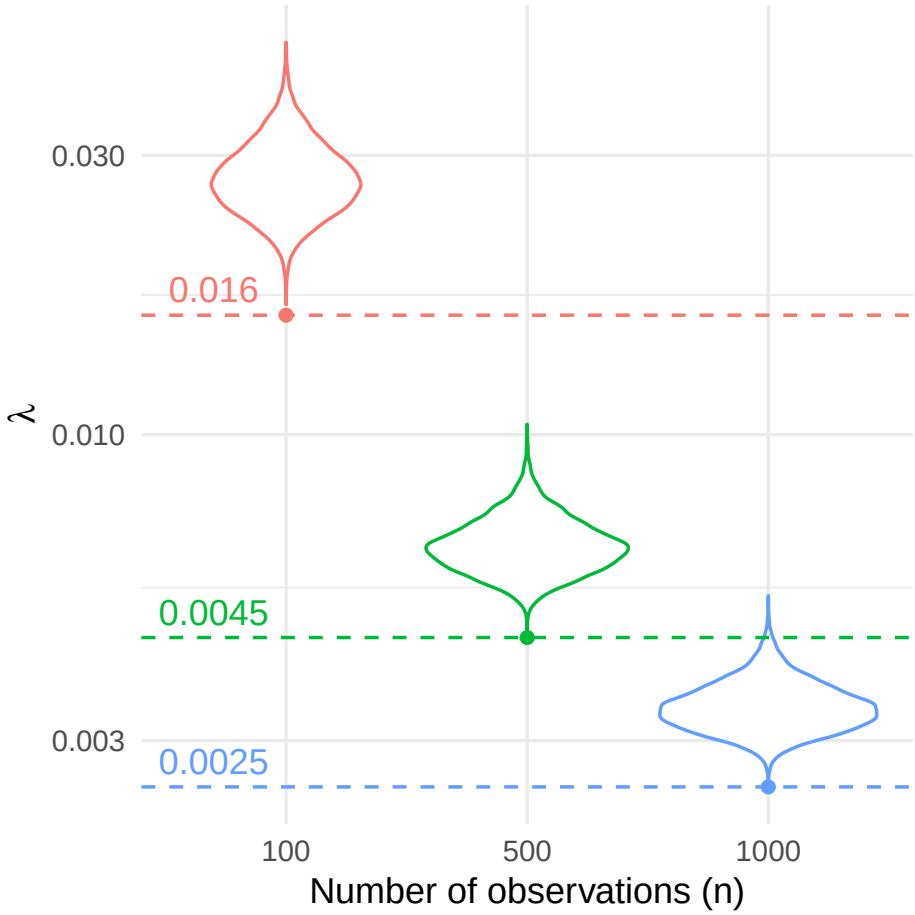

Figure 7: Empirical distribution of $\lambda^{(r_{\max})}$ across 10000 replications of the procedure to choose λ (see main text for details). The dots correspond to the chosen values.

ensure that with high probability, the one-dimensional convex clustering finds at least two clusters under the null hypothesis.

We generate $B = 10000$ replicated null data sets $\boldsymbol{Z}_b \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}), b \in [|B|]$. For each of these data sets, we calculate $\lambda_b^{(r_{\max})}$, the smallest value of the regularization parameter λ for which the convex clustering yields a single cluster. For a given input data set, this value is obtained analytically in linear time using (19). Finally, we set λ to $q_{0.01}(\boldsymbol{\lambda}) - \mathrm{sd}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_b^{(r_{\max})})_{b\in[|B|]}$, $q_{0.01}(\boldsymbol{z})$ is the first percentile of the vector $\boldsymbol{z}$ and $\mathrm{sd}(\boldsymbol{z})$ is the standard deviation of the vector $\boldsymbol{z}$. The result of this procedure for $\boldsymbol{\Sigma} = \boldsymbol{I}_n$ is also illustrated by Figure 7. For $n \in \{100, 500, 1000\}$, the empirical distribution of $\lambda^{(r_{\max})}$ is summarized by a violin plot (mirrored kernel density estimate) and the obtained value of λ is represented by a dot and a dashed horizontal line.

E.2. Level of the test in the one dimensional case

We set a Gaussian sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ with mean vector $\boldsymbol{\mu} = \mathbf{0}_n$ and known covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_n$. For the given fixed value $\lambda = 0.0025$, we use our test in order to compare

the means of cluster $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ for $1 \leq k < k' \leq K_0 = 10$. This corresponds to the test of the $K_0(K_0-1)/2$ null hypotheses ${\boldsymbol{\eta}^{[kk']}}^\top \boldsymbol{\mu} = 0$, where $\boldsymbol{\eta}^{[kk']} \in \mathbb{R}^n$ is defined by $\boldsymbol{\eta}_i^{[kk']} = 1/n_k \mathbb{1}_{i \in \mathcal{C}_k} - 1/n_{k'} \mathbb{1}_{i \in \mathcal{C}_{k'}}$. We retain $N = 1000$ numerical experiments such that the clustering $\mathcal{C}(\boldsymbol{X})$ associated to λ verifies $K(\boldsymbol{X}) \geq K_0$. The result is summarized in Figure 8 by the empirical cumulative distribution of the conditional p-value (15) for each pair of clusters. Figure 8 illustrates the uniformity of the distribution of each of these p-values, for $n = 1000$.

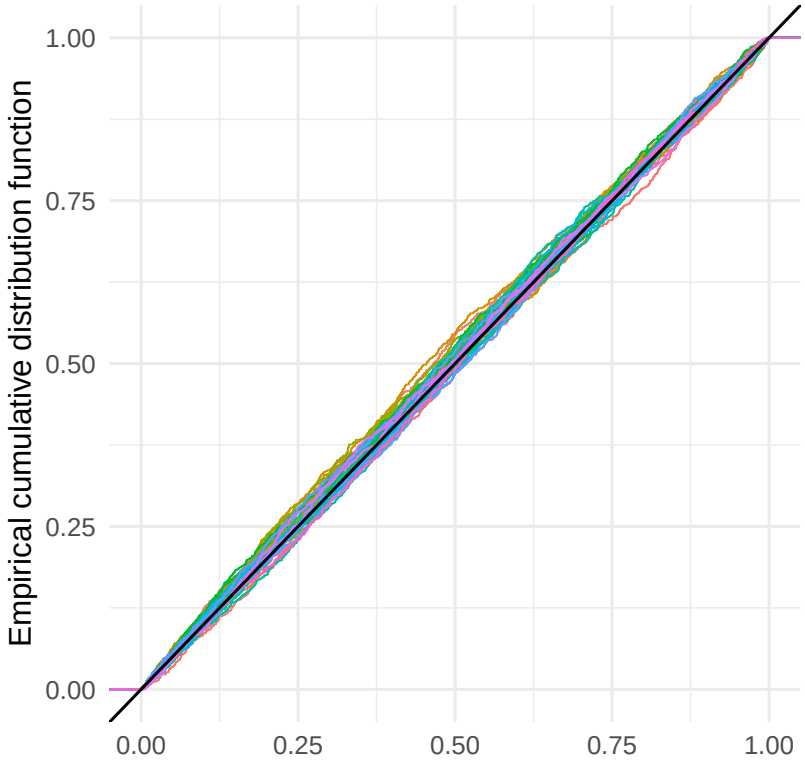

Figure 8: Experiments under the null hypothesis: empirical cumulative distribution functions of the p-value of the test of equality of means of all pairs of clusters. Each curve corresponds to a specific pair of clusters.

E.3. The p-dimensional case

In Figure 9 we plot the empirical distribution (across 500 numerical experiments described in Section 3.3) of the absolute value of the difference between the true means of the estimated clusters for $n = 100$ and $n = 1000$, for the variable $\boldsymbol{Y}_{\cdot 1}$. By construction, under this simulation scenario, this quantity is bounded by 2ν, as indicated by a dashed horizontal line. The other variables, $\boldsymbol{Y}_{\cdot 2}$ and $\boldsymbol{Y}_{\cdot 3}$, are not displayed because they do not carry any signal. This indicates that the convex clustering procedure works reasonably well: indeed, the difference between the true means of the estimated clusters is close to the difference between the true means of the true clusters. Moreover, the variability is greater for $n = 100$ than $n = 1000$, consistently with the available information to solve the clustering problem.

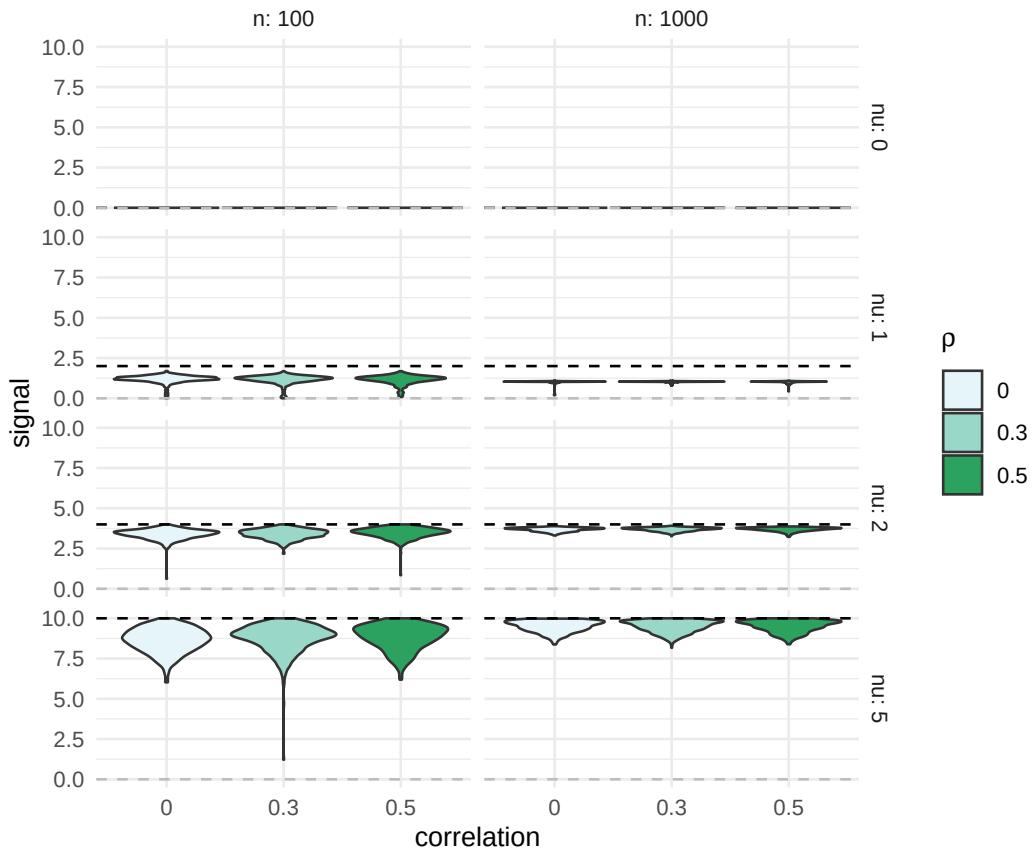

Figure 9: Empirical distribution (across 500 experiments) of the absolute value of the difference between the true means of the estimated clusters. Dashed horizontal lines are drawn for reference at $y = 2\nu$ (in black) and $y = 0$ (in gray).