Post-clustering difference testing: valid inference and practical considerations

Benjamin Hivert, Denis Agniel, Rodolphe Thiébaut, Boris P. Hejblum

Post-clustering difference testing: valid inference and practical considerations

Benjamin Hivert*[,1,2,3], Denis Agniel[4], Rodolphe Thiébaut[1,2,3,5], and Boris P Hejblum[1,2,3]

[1]Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, Bordeaux F33076, France
[2]INRIA Bordeaux Sud Ouest, SISTM team Talence F-33400, France
[3]Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France
[4]Rand Corporation, Santa Monica, CA 90401, USA
[5]CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France

October 25, 2022

Abstract

Clustering is part of unsupervised analysis methods that consist in grouping samples into homogeneous and separate subgroups of observations also called clusters. To interpret the clusters, statistical hypothesis testing is often used to infer the variables that significantly separate the estimated clusters from each other. However, data-driven hypotheses are considered for the inference process, since the hypotheses are derived from the clustering results. This double use of the data leads traditional hypothesis test to fail to control the Type I error rate particularly because of uncertainty in the clustering process and the potential artificial differences it could create. We propose three novel statistical hypothesis tests which account for the clustering process. Our tests efficiently control the Type I error rate by identifying only variables that contain a true signal separating groups of observations.

Key words: Clustering, hypothesis testing, double-dipping, circular analysis, selective inference, multimodality test, Dip Test

*Corresponding author: benjamin.hivert@u-bordeaux.fr

1 Introduction

Cluster analysis is ubiquitous in medical research (see McLachlan [1992] for a comprehensive overview) to perform data classification, data exploration, and hypothesis generation [Xu and Wunsch, 2008]. Clustering works by grouping homogeneous observations into disjoint subgroups or clusters. When multivariate data are clustered, it is common to seek to identify which variables distinguish two or more of the estimated clusters, in order to interpret the clustering structure and characterise observation groups and how they differ from each other.

Despite the widespread use of clustering, Hennig et al. [2015] state there is no commonly accepted and formal definition of what clusters are. In fact, the definition of what a cluster should be varies depending on the context and the analysis specifics. Here we will use the definition from Everitt and Hothorn [2006], which includes only two criteria: i) homogeneity of observations within a cluster and ii) separability of observations between two different clusters. These two criteria are general enough to encompass the majority of the working definitions of clusters. Both can be quantified using various approaches such as distances or similarity metrics, shape of distribution [Steinbach et al., 2004], multimodality [Kalogeratos and Likas, 2012, Siffer et al., 2018], or distributional assumptions[Liu et al., 2008, Kimes et al., 2017].

While clustering is a multivariate methodology that takes into account all variables, only a set of variables can be expected to differentiate two particular clusters (i.e. separate their observations, according to the second criterion of our definition above). This question of which variable separate clusters of individuals is particularly relevant for high-dimensional data such as omics data [Ntranos et al., 2019, Vandenbon and Diez, 2020]. The current practice to identify such variables is often based on post-clustering hypothesis testing. It leads to a two-step pipeline (a first step of clustering and a second step of inference) that is actually testing data-driven hypotheses in a process sometimes referred to as "double dipping"[Kriegeskorte et al., 2009]. This approach does not efficiently control the type I error rate when testing for differences between clusters. In fact, it is always possible to cluster the data using a clustering method, even if there is no real process separating groups of observations. In this case, the clustering method artificially enforces the differences between the observations by dividing them into different clusters. The significant differences between clusters identified during the inference process could just be an artifact of the previous clustering step. To illustrate this phenomenon, we consider data generated from a univariate Gaussian distribution with mean 0 and variance 1 (Figure 1 **panel A**). Two clusters can be built, e.g., using hierarchical clustering with Ward's method and Euclidean distance (Figure 1 **panel B**). These two estimated clusters are not separated clusters, since all observations come from the same Gaussian distribution. One way to infer their separation is to test for a mean shift between them, for example using the classical t-test. Since there is no real process separating these two clusters, the resulting p-values should be uniformly distributed. However, when we look at the p-values of the t-test for 2000 simulations of the data, the resulting p-values are too small, leading to false positives (Figure 1 **panel B**). This simple example illustrates how it is possible to infer a separation of two clusters, even if this separation is not explained by a real process in the data. Classical inference requires a priori hypothesis. In this toy example, the hypothesis, *i.e* the lack of separation of the two clusters, is based on clusters derived from the data. Moreover, here we force differences between groups of observations by clustering them, so the clustering results do not represent the true structure of the data. Thus, the discoveries are only the results of clustering algorithms and not those of a true biological signal due to this double use of the data and the bad structures forced by clustering. For example, in the context of RNA-seq data analysis, accounting for this clustering step during the inference step is one of the open problems in the eleven grand challenges in single-cell data science mentioned by Lähnemann et al. [2020].

Our goal is to propose new methods for post-clustering inference that take into account the clustering step and the potential artificial differences it may introduce. For any clustering method that can be applied to all features of the data to build clusters, we are interested in testing the null hypothesis that a particular feature does not truly separate two of the estimated clusters. In particular, this null hypothesis allows that the feature: i) is not involved in the separation of the two subgroups and is not affected by the clustering step, and ii) is only involved in this separation because the clustering method applied to the data forced differences.

Recently, some methodological work has been done on post-clustering inference. Since the data is used twice, many of them use selective inference [Tibshirani et al., 2016, Lee et al., 2016] to account for the clustering step. Selective inference aims to control the selective type I error. This is defined as the probability

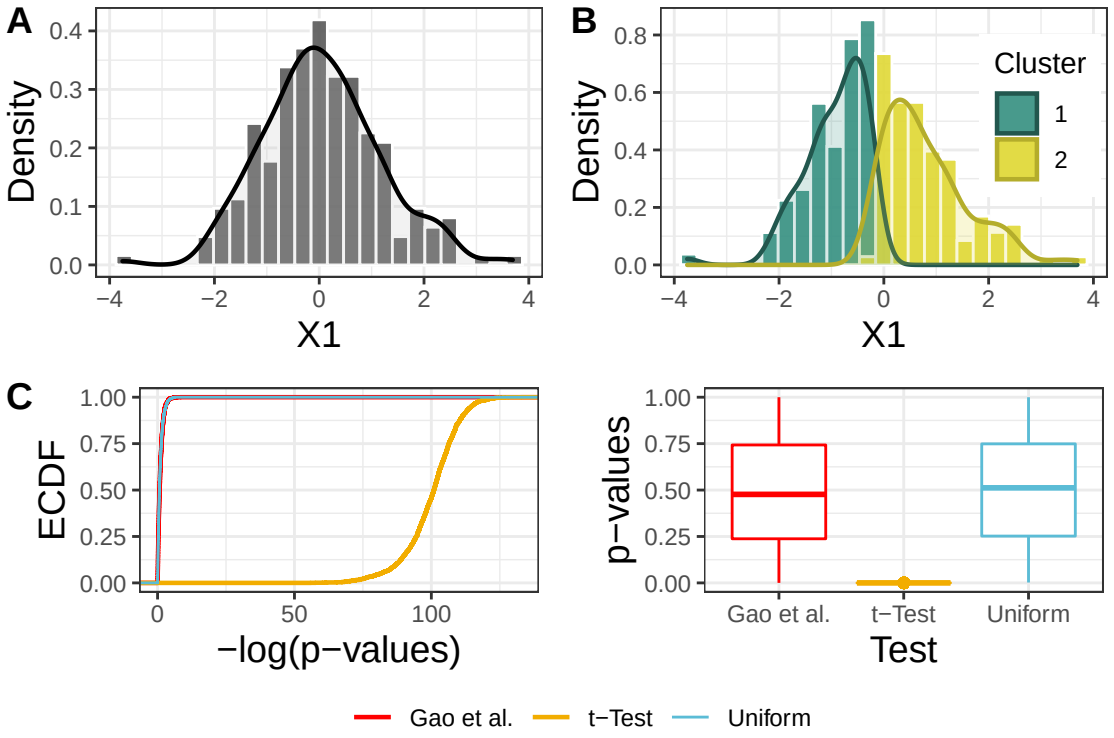

Figure 1: Artificial differences created by clustering. **panel A** Data generated according to 200 realisations of a Gaussian distribution with mean 0 and variance 1. **panel B** Hierarchical clustering with Ward method and Euclidean distance is applied to build two clusters. **panel C** t-test p-values and p-values given by the test proposed by Gao et al. [2022] for separating the two estimated clusters. The uniform distribution is also shown for comparison

under the null of rejecting the null hypothesis, given that the model and the null hypothesis have been selected thanks to the data. When data splitting is not possible, Fithian et al. [2014] has proposed to condition on this selection event during statistical hypothesis testing. In applying this approach, we use two different types of data: the data, to construct the model and the hypothesis, and the data given the fact that it has been used, i.e. data not yet observed, to perform the test. This leads to statistical hypothesis tests that efficiently control the selective type I error. Selective inference was first proposed for linear regression, change point detection [Jewell et al., 2019] and more recently for tree regression [Neufeld et al., 2021]. Clustering is also a framework in which selective inference has been applied recently. For post-clustering inference applied on RNA-seq data, Zhang et al. [2019] have developed a truncated-normal statistic that use selective inference and leads to valid p-values under their null of no differential expression. However, in addition to selective inference, they use data splitting, which is only possible if the number of observations is large enough. They also use a supervised approach to predict the partition formed on half of the data on the remaining half. Instead of conditioning on the clustering event in their statistical hypothesis test, they condition on the fact that in the remaining half of the data, the labels of the observations are predicted thanks to a supervised approach. More recently, Gao et al. [2022] have developed a multivariate selective test to investigate whether two estimated clusters are truly separated or whether the observations they contain come from a single cluster. By using selective inference, they account for the clustering step. Their approach is suitable for cluster validation because their null hypothesis is the equality of two cluster centers. This method also leads to valid p-values under the null hypothesis (Figure 1 **panel c**). However, this method is not suitable for our purpose, since in this particular context the goal is to study the separation of two clusters at the feature level, i.e., in a univariate setting.

In this paper, we introduce three new methods for post-clustering inference. First, we adapt the method proposed by Gao et al. [2022] for univariate hypotheses to investigate whether individual features contain information about group (clustering) structure. In doing so, we use a data-driven and fixed clustering of the data to ensure interpretations. To deal with the multiple clusters case, we also present an extension of this first test based on an aggregation of its p-values. Second, we propose another approach using a test of multimodality that account for the clustering step by investigating the presence of a continuum in the distribution of the variable. The paper proceeds as follows. In the Methods section, we describe the methods we proposed for post-clustering inference. These approaches are then evaluated and compared in the Results section using extensive numerical simulations and a real ecological dataset. Some final comments can be found in the Discussion section.

2 Methods

In the following, let $\boldsymbol{X}$ be a $n \times p$ random variable of n observations of p features, with g^{th} column $\boldsymbol{X}_g$. On $\boldsymbol{X}$ we apply a clustering method $c()$ to create $c(\boldsymbol{X})$, a partition of the n observations into K disjoint clusters $C_1, \ldots, C_K$. We are interested in the ability of a given variable $\boldsymbol{X}_g$ to separate two clusters C_k and C_l estimated using all the information contained in $\boldsymbol{X}$ with the clustering method $c()$.

2.1 Selective test

To develop our statistical hypothesis testing, we first specify a generative model to the observations along $\boldsymbol{X}_g$. We assume that each of the n observations of $\boldsymbol{X}_g$ comes from independent Gaussian distributions with unknown mean μ_{gi} and known variance σ_g^2. Then, for all $i \in \{1, \ldots, n\}$, $X_{gi} \sim \mathcal{N}(\mu_{gi}, \sigma_g^2)$. Because of the independence between each X_{gi}, the multivariate distribution of $\boldsymbol{X}_g$ is a multivariate Gaussian distribution $\mathcal{N}_n\left(\boldsymbol{\mu_g}, \sigma_g^2 \boldsymbol{I_n}\right)$ with mean $\boldsymbol{\mu_g} = (\mu_{g1}, \ldots, \mu_{gn})^t$ and covariance matrix $\boldsymbol{\Sigma} = \sigma_g^2 \boldsymbol{I_n}$. Let $\boldsymbol{x}_g$ be the realisation of $\boldsymbol{X}_g$ observed in $\boldsymbol{X}$. Now, for a cluster C_k, let

$$\overline{\mu}_g^{C_k} = \frac{1}{|C_k|} \sum_{i \in C_k} \mu_{gi} \qquad \text{and} \qquad \overline{X}_g^{C_k} = \frac{1}{|C_k|} \sum_{i \in C_k} X_{gi}$$

be the true mean and empirical mean, respectively, of the variable $\boldsymbol{X}_g$ in cluster C_k. Testing for a mean shift between the two clusters is a straightfoward way to evaluate the separation of two clusters along $\boldsymbol{X}_g$. Thus, we define the two following hypotheses:

$$\mathcal{H}_0 : \overline{\mu}_g^{C_k} = \overline{\mu}_g^{C_l} \qquad \text{vs} \qquad \mathcal{H}_1 : \overline{\mu}_g^{C_k} \neq \overline{\mu}_g^{C_l} \tag{1}$$

By introducing a contrast vector $\boldsymbol{\eta} \in \mathbb{R}^n$ defined by: $\eta_i = \frac{\mathbb{1}_{i \in C_k}}{|C_k|} - \frac{\mathbb{1}_{i \in C_l}}{|C_l|} \forall i = 1, \ldots, n$ following Jewell et al. [2019], Gao et al. [2022], we can rewrite (1) above as:

$$\mathcal{H}_0 : \boldsymbol{\mu}_g^t \boldsymbol{\eta} = 0 \qquad \text{vs} \qquad \mathcal{H}_1 : \boldsymbol{\mu}_g^t \boldsymbol{\eta} \neq 0 \tag{2}$$

$\mathcal{H}_0$ in (2) is actually generated by a function of the data $c(\boldsymbol{X})$, which clearly sets us in the context of selective inference. Conditioning on this clustering event within statistical inference procedures is thus required. In particular, we derive an adaptation of the p-value proposed by Jewell et al. [2019] (originally intended for change point detection) for our purposes of clustering:

$$p_g^{C_k, C_l} \equiv \mathbb{P}_{\mathcal{H}_0} \left(|\boldsymbol{X}_g^t \boldsymbol{\eta}| > |\boldsymbol{x}_g^t \boldsymbol{\eta}| \mid C_k, C_l \in c(\boldsymbol{X}) \right) \tag{3}$$

Here we condition on the estimation of C_k and C_l by $c(\boldsymbol{X})$, which leads to the definition of $\mathcal{H}_0$, and the resulting p-values (3) account for the clustering as well as the uncertainty associated with the estimation of these two clusters. $p_g^{C_k, C_l}$ quantifies the probability that the mean difference between C_k and C_l is as large as the observed difference under $\mathcal{H}_0$ given the observed clustering structure. Its calculation relies on all possible realisations of $\boldsymbol{X}_g$ resulting in the same estimation of C_k and C_l when we apply $c()$ to $\boldsymbol{X}$. Yet, enumerating all such data sets $\boldsymbol{X}$ is hard. To get more tractable p-values, we follow Jewell et al. [2019] and Gao et al. [2022] in constraining the randomness in the random variable $\boldsymbol{X_g}$ and we define our p-value as follows:

$$\tilde{p}_g^{C_k, C_l} \equiv \mathbb{P}_{\mathcal{H}_0} \left(|\boldsymbol{X}_g^t \boldsymbol{\eta}| > |\boldsymbol{x}_g^t \boldsymbol{\eta}| \; | C_k, C_l \in c(\boldsymbol{X}), \boldsymbol{\pi_\eta}^\perp \boldsymbol{X}_g = \boldsymbol{\pi_\eta}^\perp \boldsymbol{x}_g \right) \tag{4}$$

where $\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp} = \boldsymbol{I}_n - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2^2}$ restricts the random variable $\boldsymbol{X}_g$ to a space defined by the scalar $\boldsymbol{\pi_\eta}^\perp \boldsymbol{x}_g$ without losing control of type I error [Gao et al., 2022]. The p-value (4) can be rewritten as (see Supplementary Materials for the proof):

$$\tilde{p}_g^{C_k, C_l} = \mathbb{P}_{\mathcal{H}_0} \left(|\phi_g| > |\boldsymbol{x}_g^t \boldsymbol{\eta}| \mid \phi_g \in S_g \right) \tag{5}$$

where $S_g = \{\phi_g : C_k, C_l \in c\left(\boldsymbol{x}\left(\phi_g\right)\right)\}$ is the set of perturbations of the g^{th} variable from $\boldsymbol{X}$ where both C_k and C_l are conserved by $c()$, and $\phi_g = \boldsymbol{X}_g^t \boldsymbol{\eta} \overset{\mathcal{H}_0}{\sim} \mathcal{N}\left(0, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$. $\boldsymbol{X}(\phi_g)$ thus represents a perturbed version of the data $\boldsymbol{X}$, where only the g^{th} variable is perturbed:

$$\boldsymbol{x_g} - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t \boldsymbol{x}_g}{\|\boldsymbol{\eta}\|_2^2} + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}$$

This perturbation has a clear interpretation: if $|\phi_g| > |\boldsymbol{x}_g^t \boldsymbol{\eta}|$ data from the two clusters are split further apart along $\boldsymbol{X}_g$ than is observed in the data; whereas if $|\phi_g| < |\boldsymbol{x}_g^t \boldsymbol{\eta}|$ instead, they are brought closer together along $\boldsymbol{X}_g$ (and if $\phi_g = \boldsymbol{x}_g^t \boldsymbol{\eta}$ the data are actually not perturbed because in this case $\boldsymbol{X}(\phi_g) = X$). Note that (5) can be rewritten as $\mathbb{P}_{\mathcal{H}_0} \left(|\phi_g| > |\boldsymbol{x}_g^t \boldsymbol{\eta}|, \phi_g \in S_g \right) / \mathbb{P}_{\mathcal{H}_0} \left(\phi_g \in S_g \right)$. So if C_k and C_l can only be preserved when the observation are perturbed further apart, then (5) will be large since $\mathbb{P}_{\mathcal{H}_0} \left(|\phi_g| > |\boldsymbol{x}_g^t \boldsymbol{\eta}|, \phi_g \in S_g \right) \simeq \mathbb{P}_{\mathcal{H}_0} \left(\phi_g \in S_g \right)$. In conclusion, this selective test can be interpreted in terms of separability of the two clusters considered (even though it is based on a difference in means) as it boils down to quantifying the possibility to bring closer together the observations from the two clusters while preserving their separation.

In order to explicitly describe the set S_g while retaining as much generality as possible about $c()$, we follow Gao et al. [2022] and use Monte-Carlo simulations to approximate $\tilde{p}_g^{C_k, C_l}$. This strategy relies on (5) being rewritten as:

$$\tilde{p}_g^{C_k, C_l} = \frac{\mathbb{E}\left[\mathbb{1}\left\{|\phi_g| > |\boldsymbol{x}_g^t \boldsymbol{\eta}|, \phi_g \in S_g\right\}\right]}{\mathbb{E}\left[\mathbb{1}\left\{\phi_g \in S_g\right\}\right]} \tag{6}$$

Namely, we sample $\phi_g^1, \ldots, \phi_g^N \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$ for some large value N, and replace the expectations in (6) with the sums over all samples. This Monte-Carlo procedure avoids the need to formally describe S_g. In order to enhance numerical efficiency, Gao et al. [2022] use an importance sampling approach originally proposed by Yang et al. [2016] to improve the likelihood of preserving the clustering in the perturbed data.

Our proposed estimation of $\tilde{p}_g^{C_k,C_l}$ is thus:

$$\tilde{p}_g^{C_k,C_l} \approx \frac{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{|\omega_g^i| \geq |\boldsymbol{x}_g^t \boldsymbol{\eta}|, C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}} \tag{7}$$

where $\omega_g^1, \ldots, \omega_g^N \sim \mathcal{N}\left(\boldsymbol{x}_g^t \boldsymbol{\eta}, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$, and $\pi^i = \frac{f_1(\omega_g^i)}{f_2(\omega_g^i)}$ the importance sampling probabilities with f_1 the density of a $\mathcal{N}\left(0, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$ distribution and f_2 the distribution of a $\mathcal{N}\left(\boldsymbol{x}_g^t \boldsymbol{\eta}, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$ distribution. Of note, we adapt the method from Phipson and Smyth [2010] to obtain unbiased Monte-Carlo p-values estimations (see Supplementary Materials for details).

At the core of the above test is the scaling variance parameter σ_g^2, which represents the variance of each column of $\boldsymbol{X}_g$. While σ_g^2 is assumed to be known in the test, it is not the case in practice and we propose to use the following plug-in estimate instead:

$$\hat{\sigma}_g^2 = \frac{1}{|C_k| + |C_l| - 1} \sum_{i \in C_k, C_l} \left(X_{gi} - \overline{X}_g^{C_k,C_l}\right)^2 \qquad \text{with} \quad \overline{X}_g^{C_k,C_l} = \frac{1}{|C_k| + |C_l|} \sum_{i \in C_k, C_l} X_{gi}$$

This variance estimate only takes into account observations from the two clusters of interest in the test, in line our null hypothesis of no separation of the two clusters (the variance itself can informs on the separation of the data [Liu et al., 2010]). In some cases, this $\hat{\sigma}_g^2$ could underestimates the variance of $\boldsymbol{X}_g$ (particularly if the clustering induces strong artificial differences). Meanwhile, Gao et al. [2022] rely on a different estimate, that instead overestimate the variance in certain cases. Still they have showed type I error control is guaranteed, even with an overestimated variance, at the cost of being overly conservative (see Supplementary Figure S1 for additional details).

2.2 A more powerful test in the presence of intervening clusters

The above selective test has been designed for comparing a pair of clusters. Yet, in practice there are often more than 2 clusters. In such case, the test could have very limited statistical power even for well separated clusters: if there is one or more additional cluster in-between the two clusters of interest, it quickly becomes impossible to perturbed them closer together without changing the clustering (see Supplementary Figure S2). To overcome this limitation, we extend the above selective test assuming that two estimated clusters C_k and C_l are separated on $\boldsymbol{X}_g$ if and only if at least one of the adjacent cluster pairs in-between them are separated. This means that on the contrary, if there is a continuum on $\boldsymbol{X}_g$ to go from C_k to C_l, then C_k to C_l are not separated. By testing only the separation of pairs of adjacent clusters we retain the statistical power of the selective test. We propose to combine all the in-between adjacent pair selective test p-values into a so called combined selective test to finally assess the separation of C_k and C_l on $\boldsymbol{X}_g$.

To identify the clusters in-between C_k and C_l on $\boldsymbol{X}_g$, we define the set:

$$\mathcal{C}_g^{k:l} := \left\{C_i, i = 1, \ldots, K \;/\; \overline{X}_g^{C_i} \in \left[\min\left(\overline{X}_g^{C_k}, \overline{X}_g^{C_l}\right), \max\left(\overline{X}_g^{C_k}, \overline{X}_g^{C_l}\right)\right]\right\}$$

where $\overline{X}_g^{C_i} = \frac{1}{|C_i|} \sum_{j \in C_i} X_{gj}$, implicitly sorting clusters according to their empirical mean on $\boldsymbol{X}_g$. We define two clusters C_{m_1} and C_{m_2} as adjacent on $\boldsymbol{X}_g$ if $\mathcal{C}_g^{m_1:m_2} = \{C_{m_1}, C_{m_2}\}$. So if $|\mathcal{C}_g^{k:l}| = M$, there are $M-1$ pairs of adjacent clusters in $\mathcal{C}_g^{k:l}$ that draw a path from C_k to C_l. We can then define:

$$p_g^{C_k:C_l} := f\left(p_g^1, \ldots, p_g^{M-1}\right)$$

where f must be a merging function as described in Vovk and Wang [2020]. Based on numerical simulations, we favor the use of the harmonic mean merging function, recommended by Vovk and Wang [2020] for potential dependencies between the p-values – which is our case here since each cluster data contributes to two p-values – and features a good trade-off between type I error and statistical power (see Supplementary Figure S3).

Thus, we use:

$$p_g^{C_k:C_l} = \min\left(e\log(M-1)\frac{M-1}{\sum_{i=1}^{M-1}\frac{1}{p_g^i}}, 1\right)$$

Of note, in order for all $p_g^1, \ldots, p_g^{M-1}$ to be computed using the same variance estimate, we propose this time to plug-in an estimate of σ_g^2 that accounts for all observations belonging to either one the adjacent clusters in $\mathcal{C}_g^{g:l}$:

$$\hat{\sigma}_g^2 = \frac{1}{|C_g^{k:l}| - 1}\sum_{C\in C_g^{k:l}}\sum_{i\in C}\left(X_{gi} - \overline{X}_g^{C_g^{k:l}}\right)^2 \qquad \text{with} \quad \overline{X}_g^{C_g^{k:l}} = \frac{1}{|C_g^{k:l}|}\sum_{C\in C_g^{k:l}}\sum_{i\in C} X_{gi}$$

2.3 Multimodality test

The separation of two clusters according to a given variable is equivalent to this variable's distribution being multimodal. Following Kim et al. [2021], multimodality thus becomes a marker for the separation of clusters: each mode corresponds to a group of homogeneous observations (i.e., a cluster), separated by less dense regions of the distribution. But as with artificial mean differences arising from clustering, multimodality may also be an artefact caused by the clustering method $c()$. We propose to leverage this notion of continuum between two clusters: if C_k and C_l are separated, then there must be dip in the distribution of this variable at some point between the two (i.e. multimodality). On the other hand, if there is a continuum between these two clusters, then they cannot truly be separated (i.e. unimodality). Fortunately, such a continuum cannot be caused by the clustering method.

This second proposal can be seen as a simplified version of our first selective test. Indeed, by perturbing the data in the selective test to see if we can bring the two clusters closer without changing the clustering, we assess how likely it would be to observe a continuum between C_k and C_l. If there is a continuum between C_k and C_l on $\boldsymbol{X}_g$, then its distribution must be unimodal. Thus, to investigate separability of those two clusters on $\boldsymbol{X}_g$, it suffices to apply a unimodality test to its distribution restricted only to the individuals from clusters of the set $C_g^{k:l}$. Indeed, if the $\boldsymbol{X}_g$ separates C_k and C_l, then there are at least two clusters in $C_g^{k:l}$ that are separated from each other, and in particular, since these clusters are between C_k and C_l, there is also a separation between them on $\boldsymbol{X}_g$.

A unimodality test compares the null hypothesis "distribution of $\boldsymbol{X}_g$ is unimodal" to the alternative "distribution of $\boldsymbol{X}_g$ is multimodal". In the context of unsupervised clustering, Kalogeratos and Likas [2012] developed a clustering algorithm based on incremental unimodality testing, and Siffer et al. [2018] developed unimodality a test to assess data clusterability based on their multivariate distribution. Ameijeiras-Alonso et al. [2021] give a recent overview on unimodality testing, but three tests are the most frequent: i) the Silverman test [Silverman, 1981] based on the kernel estimate of the density f of the data, ii) the Dip Test [Hartigan et al., 1985] based on the cumulative distribution function F, and iii) the excess mass test [Müller and Sawitzki, 1991]. The Dip Test avoids the need for estimating of additional parameters or making any distributional assumption and has already been applied to clustering [Kalogeratos and Likas, 2012, Wasserman et al., 2014, Schelling and Plant, 2020]. Furthermore, compared to several multimodality tests available in the R package `multimode` Ameijeiras-Alonso et al. [2021], the Dip Test outperforms its competition both in terms of computation times and performances (see Supplementary Figure S4).

The Dip Test from Hartigan et al. [1985] relies on the dip statistic $\text{dip}(F) = \min_{G\in\mathcal{U}} \rho(F, G)$, where $\rho(F, G) = \sup_x |F(x) - G(x)|$ and $\mathcal{U}$ is the class of unimodal distributions. Thus, the dip statistic is the distance of F to the set of unimodal functions and it measures the deviation of our distribution from unimodality. If F is unimodal, then $\text{dip}(F) = 0$, and conversly if F is multimodal, then $\text{dip}(F) > 0$. In practice p-values can be computed as:

$$p_{\widehat{D}_n} := \mathbb{P}\left(d_{U_n} \geq \widehat{D}_n\right)$$

where d_{U_n} is the dip statistic computed for a n-sample drawn from $\mathcal{U}[0, 1]$ (the standard uniform distribution), $\widehat{D}_n$ is the observed dip statistic, and n is the sample size. Hartigan et al. [1985] showed that the Uniform distribution is the unimodal distribution with the asymptotically largest dip statistic among the unimodal

distributions (intuitively the least favourable candidate for unimodality): a distribution with a dip statistic larger than that of the uniform distribution cannot be unimodal. Thus $p_{\widehat{D}_n}$ is interpreted as the probability under the null case of unimodality that the uniform distribution has a dip statistic greater than the observed dip statistic of $\widehat{F}_n$.

For our purposes, we apply the Dip Test to the distribution of the variable $\boldsymbol{X}_g$ restricted to the individuals that are in the clusters of the set $C_{k:l}$ to test for a continuum between C_k and C_l.

3 Results

3.1 Numerical simulations study

We present here results evaluating the behaviour of our proposed tests in the Methods section both in terms of type-I error control and statistical power.

3.1.1 Behaviour in a two-dimensional setting

We generated two-dimensional data ($p = 2$) under two scenarios: (i) first with no separated clusters from a common standard Gaussian distribution $\mathcal{N}(0, 1)$; and (ii) second with three clusters from Gaussian distributions $\mathcal{N}(\mu^{C_j}, 1)$ and built-in mean differences $\mu^{C_1} = (-5, 0)$, $\mu^{C_2} = (5, 0)$, and $\mu^{C_3} = (0, 10)$ (thus X_1 separated all three clusters while X_2 only separated C_3 from the rest, meaning X_2 was under the null when comparing C_1 and C_2). In both cases, we applied hierarchical clustering with Ward method and Euclidean distance to build three clusters. Figure 2A shows an example realisation for each scenario. In the first scenario, clusters were estimated by forcing differences between groups of observations, creating artificial differences between clusters, while in the second scenario, the estimated clusters represented the true structure of the data (each cluster weas a homogeneous and separate group of observations).

Figure 2B shows the results of the three proposed approaches compared to the p-values from the usual t-test for 2,000 repeated simulations each with a sample size of $n = 200$. For the no cluster scenario, the t-test yielded extremely small p-values which translates into a direct inflation of the type-I error. In fact, the t-test identified the artificial differences created during the clustering process. Taking into account this clustering step, the p-values of the selective test $p_g^{C_k,C_l}$ and the p-values resulting from its merging extension $p_g^{C_k:C_l}$ were fairly uniformly distributed over $[0, 1]$, ensuring a good calibration of the p-values and a control of type-I error. As for the multimodality test, its p-values were overly conservatives but consistent with no real separation of clusters. This was due to its reference being the Uniform distribution (the limit case for a unimodal distribution) while the data were generated from a Gaussian distribution (which has a lower dip statistic than the uniform distribution). Those good results were confirmed under the 3 clusters scenario when comparing C_1 and C_2 along X_2. For all other comparisons under this scenario, all 4 tests correctly detected the separated clusters that are under $\mathcal{H}_1$.

Of note, if clustering does not artificially force differences between groups of observations, e.g by discovering the actual group structure in the data, the t-test also control the type-I error. This illustrates the connections between artificial differences and the estimation of the number of clusters. But, this process is still testing data driven hypothesis, which does not respect the classical inference setting where hypothesis must be specified without using the data.

3.1.2 Statistical power

We now generated data from a univariate mixture of two Gaussian distributions with equal proportions and variance: $0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(\delta, 1)$, where the two components were separated by a mean difference of δ – also called a contamination model [Laurent et al., 2018]. The two components distinguished two different clusters, and the magnitude of δ tuned their separability. We applied hierarchical clustering with Ward method and Euclidean distance to build either 2 or 4 clusters. Figure 3A displays an example realization of this simulation. In the 2-cluster case, the true structure of the data was uncovered, while in the 4-cluster case spurious clusters were introduced. We evaluated the statistical power of the three proposed tests to detect the separation between clusters 1 and 2, the two most extreme clusters in the distribution of the data, according to δ at significance levels $\alpha = 5\%$ using $N = 2000$ Monte-Carlo replicates.

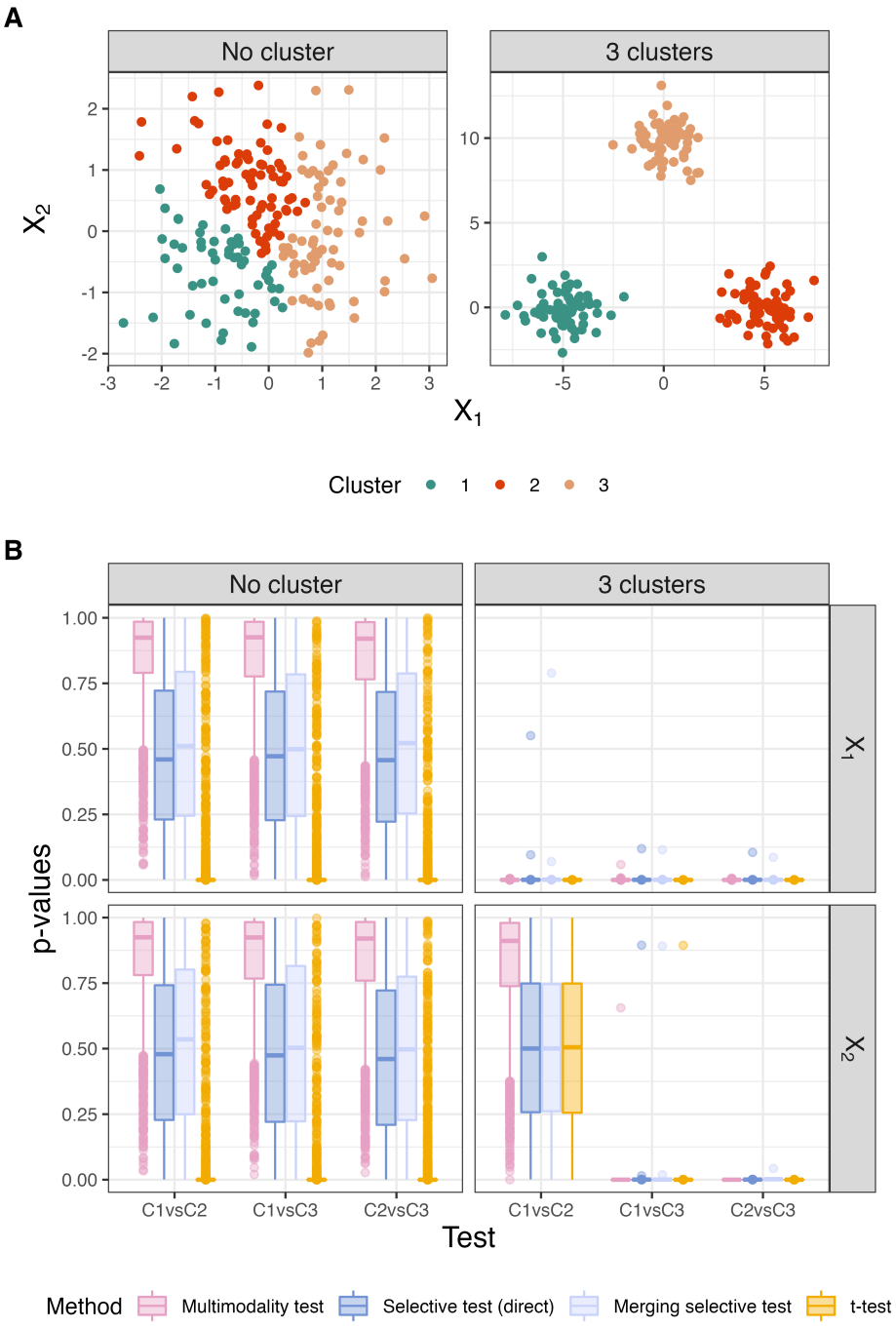

Figure 2: Validity of p-values returned by our proposed tests, comparison with t-test. **panel A** Data generation process. Two cases are studied: a case under a global null of no clusters in the data (No cluster) and a case with three real clusters (3 clusters). In both cases, hierarchical clustering with Ward method and Euclidean distance is used to build 3 clusters. **panel B** Resulting p-values for each possible cluster pair for each variable for 2000 simulations of the data.

Figure 3B displays the results. Intuitively, statistical power increases with δ. The multimodality test appeared the most powerful in this setting , especially when $\delta \geq 3$. Siffer et al. [2018] has shown that $\delta = 3.04$ is a threshold value above which bimodality begins to appear in such two-part Gaussian mixture. Moreover, since all clusters in the 4-cluster case were between cluster 1 and cluster 2, statistical power achieved by the multimodality test is exactly the same as in the 2-clusters case. In the two-clusters case, the direct selective test and the merging selective test had the same statistical power (since only two clusters were estimated, they were necessarily adjacent and therefore the direct selective test was exactly the same as the merging one). Meanwhile, the direct selective test failed in the 4-cluster case, regardless of the value of δ. Indeed, it was impossible to bring clusters 1 and 2 closer without mixing them with clusters 3 and 4. Fortunately, the merging selective test avoided this pitfall, because the direct selective inference test performed favourably on adjacent clusters, and carried over the separation between clusters 3 and 4. Of note, the merging selective test remains valid when the numbers of adjacent p-values increases.

3.2 Application to real ecological data

To further assess our proposed approaches, we also analyzed real data available from the `R` package `palmerpenguins` [Horst et al., 2020]. This benchmark dataset features $p = 4$ measured variables – bill length (mm), bill depth (mm), flipper length (mm), and body mass (g) – for 344 penguins. After removing observations containing missing values for at least one of the 4 variables, $n = 333$ observations were kept in our analysis. The penguins belonged three different species: Adelie, Chinstrap, and Gentoo (with 146, 68 and 119 observations respectively).

3.2.1 Negative control

We initially selected only female Gentoo penguins to create a negative control dataset. Since this dataset contained only observations of the same species and sex, there should be no real differences between observations. We applied hierarchical clustering using Ward method and Euclidean distance on scaled data to build 3 clusters (scaling avoids the variable with the largest variance to dominate the clustering). Since there was no information defining any group structure in this subset, the clustering artificially created differences. Table 1 presents the p-values of each of the 3 proposed test along with the ones from the t-test for all cluster pairs along each of the four measures. Once again, the t-test identified numerous spurious associations. Meanwhile, all 3 proposed tests behaved properly by not identifying any of the four measures to be significantly separating clusters.

Cluster pair tested Variable tested	**Selective test (direct)**	**Merging selective test**	**Multimodality test**	**t-test**
Cluster 1 vs Cluster 2				
bill length	0.4082	0.4110	0.4899	0.0759
bill depth	0.6478	0.6400	0.1478	0.4802
flipper length	0.1160	0.1154	0.0992	0.0017*
body mass	0.3321	0.3425	0.8320	0.0000*
Cluster 1 vs Cluster 3				
bill length	0.1748	0.4995	0.6345	0.0001*
bill depth	0.2914	0.3025	0.5242	0.0000*
flipper length	0.3361	0.3206	0.6146	0.0005*
body mass	0.3404	0.3868	0.2918	0.1190
Cluster 2 vs Cluster 3				
bill length	0.2096	0.2120	0.9140	0.0041*
bill depth	0.1867	0.6618	0.2376	0.0000*
flipper length	0.2101	0.4322	0.1337	0.0000*
body mass	0.1573	0.7967	0.6759	0.0000*

Table 1: P-values for all cluster pair tests along each of the 4 variables from the negative control real data.
* highlights significant p-values at the $\alpha = 5\%$ level

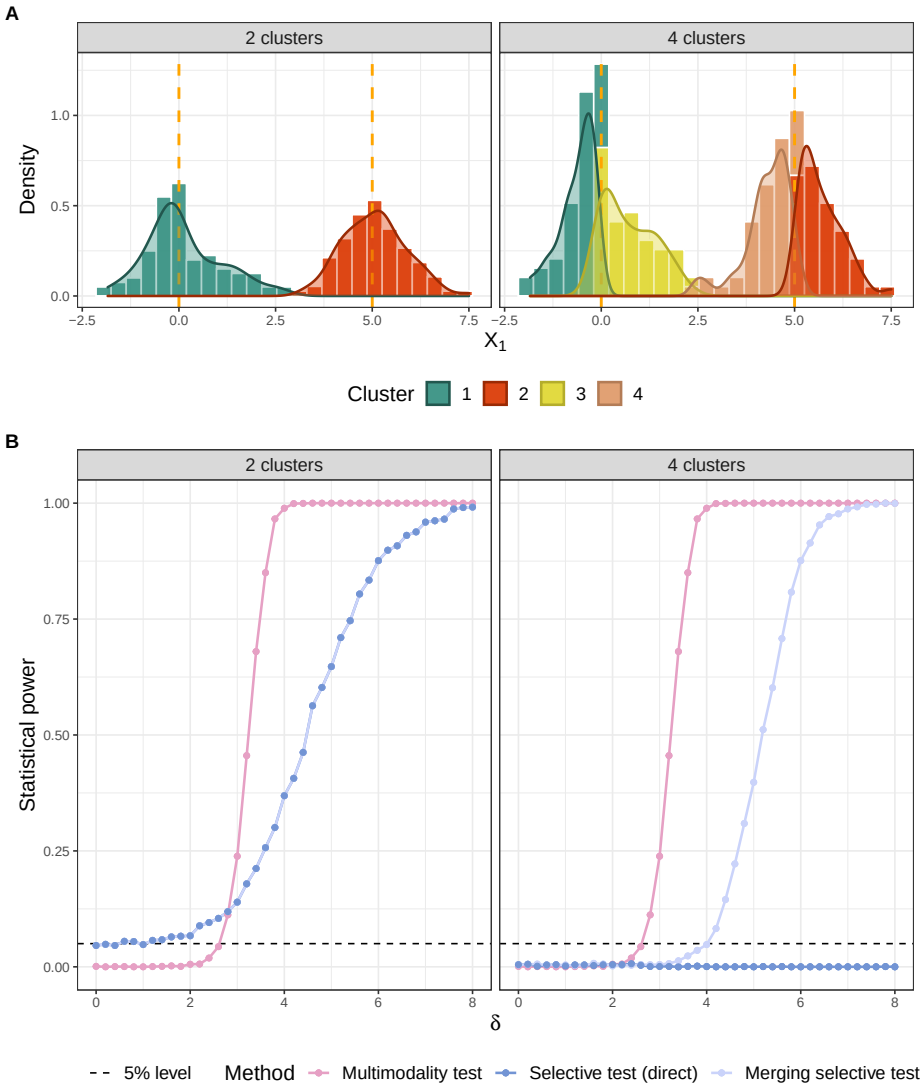

Figure 3: Statistical power of the proposed tests. **panel A** Data generation process: data are generated according to a univariate gaussian mixture with two components (with equal probability and variance) separated by a mean difference δ (contamination model). Two cases are studied: a case where the true numbers of clusters is estimated (2 clusters) and a case where more clusters are estimated (4 clusters). The orange dashed line represents the mean of the component. **panel B** Statistical power (5% level) of the proposed tests for the separation of Cluster 1 and Cluster 2 according the mean difference δ separating the two components of the gaussian mixture.

3.2.2 Full data analysis

We now included all $n = 333$ penguins in our analysis and analyzed the data as if we didn't know the species of the penguins. Since all 3 species were present, we want to identify which features are actually separating them. Figure 4A displays the density distribution for the four (scaled) variables across all 3 species. Adelie and Chinstrap penguins appear harder to distinguish, as they only appear to differ in bill length: Chinstrap penguins have larger bills (comparable to those of Gentoo penguins) than Adelie penguins. The Gentoo penguin species is the easiest to identify, as it is clearly different in the other 3 measures. Once again, we applied hierarchical clustering to scaled data with Euclidean distance and Ward method. Figure 4B displays the results of this clustering where we cut the dendrogram to get three clusters. Those three estimated clusters recovered the true species: cluster 2 and cluster 3 each contained only penguins of the Gentoo and Chinstrap species (respectively) while cluster 1 contained a mixture of two species (100% of the Adelie penguins plus 11 Chinstrap penguins).

Table 2 presents the p-values of each of the 3 proposed test along with the ones from the t-test for all cluster pairs along each of the four measures. Since the identified clusters corresponded to the three real penguin species, the clustering step was not expected to induce any artificial differences, and thus the t-test results can be used as reference. Only one comparison was not significant at the 5% level according to the t-test: bill depth did not separate Adelie (cluster 1) from Chinstrap penguins (cluster 3), which was visually coherent with Figure 4A. The multimodality test seemed to lack statistical power here, but inspection of the measures distribution depicted in Figure 4A showed that only a few comparison exhibited multimodality (namely cluster 1/Adelie compared to either two other clusters along flipper length. Both selective tests identified more significant differences (6/11 for the direct test and 7/11 for the merging test which is more robust when more than 2 clusters are identified). The missed separations can be explained by the lack of statistical power to detect small difference (see Supplementary Table 1 for additional details). By accounting the clustering step, our proposed test could have a reduced power compared to other tests like t-test (which is is the uniformly most powerful test [Lehmann, 2012]). But, this is because they are appropriately accounting for the variability and the uncertainties of the clustering step leading to results that are always valid.

Cluster pair tested Variable tested	**Direct selective test**	**Merging selective test**	**Multimodality test**	**t-test**
Cluster 1 vs Cluster 2 ***(Adelie vs Gentoo)***				
bill length	0.0024*	0.0023*	0.1647	0*
bill depth	0.0015*	0.0017*	0.3687	0*
flipper length	0.0725	0.1832	0.0047*	0*
body mass	0.0439*	0.0008*	0.6402	0*
Cluster 1 vs Cluster 3 ***(Adelie vs Chinstrap)***				
bill length	0.1748	0.0191*	0.0674	0*
bill depth	0.2266	0.2323	0.2373	0.0702
flipper length	0.4318	0.4434	0.0168*	0*
body mass	0.7036	0.7027	0.3311	0.0267*
Cluster 2 vs Cluster 3 ***(Gentoo vs Chinstrap)***				
bill length	0.2263	0.2115	0.0927	0*
bill depth	0.0084*	0.0051*	0.2245	0*
flipper length	0.0186*	0.0205*	0.1585	0*
body mass	0.0002*	0.0002*	0.4174	0*

Table 2: P-values for all cluster pair tests along each of the 4 variables from the positive control real data.
* highlights significant p-values at the $\alpha = 5\%$ level

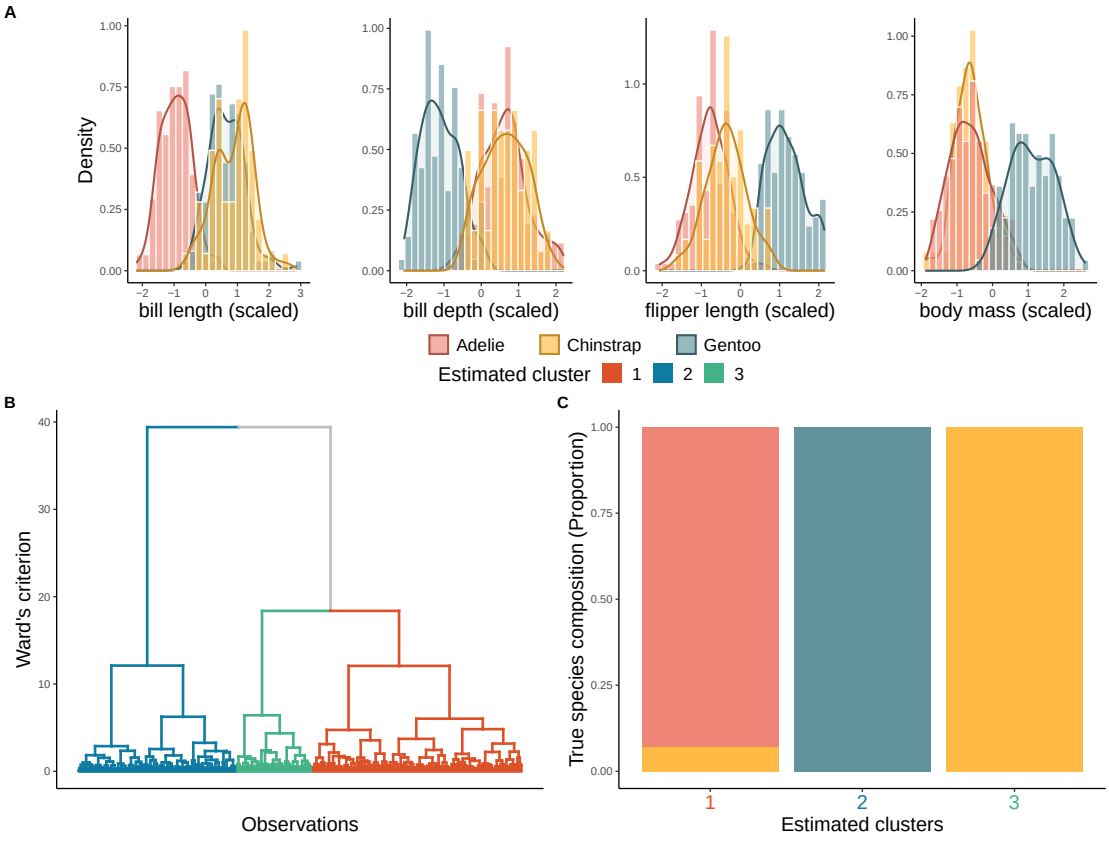

Figure 4: Clustering of the 333 completely observed palmer penguins. **A** Distribution of each scaled variable according to the three true species of penguins. **B** 3 clusters are build thanks to hierarchical clustering. Colors in the dendogram represent the estimated clusters. **C** Clustering results reveals that estimated clusters correspond to true species

4 Discussion

In this paper, we propose three new statistical tests for post-clustering inference that can be used to identify variables that separate two estimated clusters. We show that the double use of data (for clustering and for inference) and the failure to propagate the uncertainty associated with cluster estimation can lead to invalid p-values. This is particularly the case when too many clusters are estimated compared to the true underlying structure of the data. In this case, since there is no true process separating every estimated clusters, the clustering forces artificial differences between observations belonging to a common group of observations. Our three proposed tests, which take into account this clustering step and/or its possible impact on inference, give p-values that indicate a separation not induced by the clustering algorithm but emanating from the underlying data generating process, while controlling the Type-I error rate adequately. Our approaches can be used regardless of the chosen clustering algorithm and take into account many data analysis pipelines where clustering results are used post-hoc to describe and interpret clusters.

All three approaches test whether there is a separation between clusters along a given variable. The selective test is a rigorously defined test based on the concepts of selective inference, adapted from the seminal work of Gao et al. [2022] and makes a Gaussian assumption on the data. Although it tests a univariate mean difference between two clusters, it also exploits the multivariate structure of the data since the (perturbed) clustering uses all variables. The multimodality test, on the other hand, is based on the more intuitive concept of multimodality to characterise the separation of two clusters along a variable. It only relies on univariate considerations, as the separation of clusters is examined based on the distribution of each variable. Thus, unlike the selective test, which has longer computation times (dependent on the number of observations, the number of variables, and the number of Monte-Carlo simulations required to estimate p-values), it is very computationally efficient (see Supplementary Figure S5). However, this simplicity comes at the expense of a larger null hypothesis: the multimodality test requires a clear separation between clusters on the variable to work well, as it only uses the variable-level information and does not consider the entire structure of the data. Finally, since false-negative problems could occur with the selective test (particularly when the two clusters of interest are separated by other clusters), we also propose a merging method based on the aggregation of p-values. This method has the advantage of correcting these false positive problems while guaranteeing good statistical power. However, its computation cost is even greater than the selective test because this approach requires the computation of all the adjacent p-values between C_k and C_l.

The selective test rely on some distributional assumptions. In particular, because it uses the selective inference framework, it assumes Gaussian data to efficiently control the Type I error rate. We show that the selective test remains robust to other distributions (see Supplementary Figure S6). The multimodality test is based on the Dip Test which is a non-parametric test of unimodality. However, in practice, its p-value is computed using the Uniform distribution as the reference distribution under the null of unimodality. It could affect its statistical power but the control of the Type I error is still guaranteed.

The main limitation of our tests lies in the high dimensional setting. Due to the large number of variables and their correlation, perturbation-based approaches can fail. In our case, this result is amplified by the fact that the perturbation is only univariate. Thus, the selective inference test performs poorly in high dimension since it is exclusively based on perturbations. In addition, the calculation of the p-values is done using a Monte-Carlo approach requiring the clustering step to be repeated for each simulation and for each variable. So, if the number of variables is high, the computation time of the selective test can therefore be too long. Another problem of our approaches is related to the "signal vs. noise" ratio of the high dimension. In high dimension, a small signal repeated over a large number of variables is sufficient to create separated clusters in the high dimensional space [Klawonn et al., 2012]. For example in a two-components gaussian mixture, for $n = 100$ observations, a mean difference $\delta = 1$ repeated over $p = 50$ variables is enough to generate separated clusters on the first principal component of a PCA. However, the unimodality test is not powerful enough when the signal is too weak. The problem here is that the existence of clusters is only due to the repetition of the signal on a large number of variables, *i.e.* one has to take into account all the variables and the information they bring to explain the separation between clusters, but the unimodality test is purely univariate, being only interested in the information brought by the tested variable and this is why it lacks power in high dimension. Therefore, all the issues raised by the high-dimension constitute a natural path to apply and extend the results of our work presented here.

References

GJ McLachlan. Cluster analysis and related techniques in medical research. *Statistical Methods in Medical Research*, 1(1):27–48, 1992.

Rui Xu and Don Wunsch. *Clustering*, volume 10. John Wiley & Sons, 2008.

Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci. *Handbook of cluster analysis*. CRC Press, 2015.

Brian S Everitt and Torsten Hothorn. *A handbook ofstatistical analyses using R*. Chapman & Hall, Boca Raton, FL, 2006.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. The challenges of clustering high dimensional data. In *New directions in statistical physics*, pages 273–309. Springer, 2004.

Argyris Kalogeratos and Aristidis Likas. Dip-means: an incremental clustering method for estimating the number of clusters. *Advances in neural information processing systems*, 25:2393–2401, 2012.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Are your data gathered? In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2210–2218, 2018.

Yufeng Liu, David Neil Hayes, Andrew Nobel, and James Stephen Marron. Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008.

Patrick K Kimes, Yufeng Liu, David Neil Hayes, and James Stephen Marron. Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821, 2017.

Vasilis Ntranos, Lynn Yi, Páll Melsted, and Lior Pachter. A discriminative learning approach to differential expression analysis for single-cell rna-seq. *Nature Methods*, 16(2):163–166, 2019.

Alexis Vandenbon and Diego Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature communications*, 11(1):1–10, 2020.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–27, 2022.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after changepoint detection. *arXiv preprint arXiv:1910.04291*, 2019.

Anna C Neufeld, Lucy L Gao, and Daniela M Witten. Tree-values: selective inference for regression trees. *arXiv preprint arXiv:2106.07816*, 2021.

Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4):383–392, 2019.

Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. *Advances in Neural Information Processing Systems*, 29:2469–2477, 2016.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining*, pages 911–916. IEEE, 2010.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Chanwoo Kim, Hanbin Lee, Juhee Jung, Keehoon Jung, and Buhm Han. Marcopolo: a clustering-free approach to the exploration of differentially expressed genes along with group information in single-cell rna-seq data. *bioRxiv*, pages 2020–11, 2021.

Jose Ameijeiras-Alonso, Rosa M Crujeiras, and Alberto Rodriguez-Casal. multimode: An r package for mode assessment. *Journal of Statistical Software*, 97(1):1–32, 2021.

Bernard W Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99, 1981.

John A Hartigan, Pamela M Hartigan, et al. The dip test of unimodality. *Annals of statistics*, 13(1):70–84, 1985.

Dietrich Werner Müller and Günther Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.

Larry Wasserman, Martin Azizyan, and Aarti Singh. Feature selection for high-dimensional clustering. *arXiv preprint arXiv:1406.2240*, 2014.

Benjamin Schelling and Claudia Plant. Dataset-transformation: improving clustering by enhancing the structure with dipscaling and diptransformation. *Knowledge and Information Systems*, 62(2):457–484, 2020.

Béatrice Laurent, Clément Marteau, and Cathy Maugis-Rabusseau. Multidimensional two-component gaussian mixtures detection. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 54, pages 842–865. Institut Henri Poincaré, 2018.

Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. URL https://allisonhorst.github.io/palmerpenguins/. R package version 0.1.0.

Erich L Lehmann. Some history of optimality. In *Selected Works of EL Lehmann*, pages 1033–1039. Springer, 2012.

Frank Klawonn, Frank Höppner, and Balasubramaniam Jayaram. What are clusters in high dimensions and are they difficult to find? In *International workshop on clustering high-dimensional data*, pages 14–33. Springer, 2012.

A Proof: Computation of the selective p-values

We want to compute the selective p-value given in (4):

$$p_g^{C_k,C_l} \equiv \mathbb{P}_{H_0}\left(|\boldsymbol{X}_g{}^t\boldsymbol{\eta}| > |\boldsymbol{x}_g{}^t\boldsymbol{\eta}| \;|C_k, C_l \in c(\boldsymbol{X}), \boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp}\boldsymbol{X}_g = \boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp}\boldsymbol{x}_g\right)$$

To compute (4), we have to write our data matrix $\boldsymbol{X}$ as a function of our statistic $\boldsymbol{X}_g{}^t\boldsymbol{\eta}$ and the residual term $\boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp}\boldsymbol{X}_g$ where $\boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp} = \boldsymbol{I}_n - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2^2}$ Since, $\boldsymbol{X}_g = \boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp}\boldsymbol{X}_g + \left(\boldsymbol{I}_n - \boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp}\right)\boldsymbol{X}_g$, then :

$$\begin{aligned}
c(\boldsymbol{X}) &= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{X}_g|\dots|\boldsymbol{x}_p]\right) \\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + [\boldsymbol{0}_n|\dots|\boldsymbol{X}_g|\dots|\boldsymbol{0}_n]\right) \\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + (\boldsymbol{I}_n - \boldsymbol{\Pi}_{\boldsymbol{\eta}}{}^{\perp})\boldsymbol{X}_g|\dots|\boldsymbol{0}_n\right]\right) \\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \left(\boldsymbol{I}_n - \boldsymbol{I}_n + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2^2}\right)\boldsymbol{X}_g|\dots|\boldsymbol{0}_n\right]\right) \\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2}\boldsymbol{X}_g|\dots|\boldsymbol{0}_n\right]\right) \\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}|\dots|\boldsymbol{0}_n\right]\right) \qquad \text{with } \phi_g = \boldsymbol{X}_g{}^t\boldsymbol{\eta} \\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}|\dots|\boldsymbol{0}_n\right]\right) \\
&= c\left(\left[\boldsymbol{x}_1|\dots|\boldsymbol{x}_g - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t\boldsymbol{x}_g}{\|\boldsymbol{\eta}\|_2^2} + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}|\dots|\boldsymbol{x}_p\right]\right)
\end{aligned}$$

We also have :

$$\boldsymbol{X}_g^t\boldsymbol{\eta} \perp \boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g$$

because $\boldsymbol{\Pi}_{\boldsymbol{\eta}}^{\perp}$ is the orthogonal projection matrix onto the subspace orthogonal to span($\boldsymbol{\eta}$) [Jewell et al., 2019, Gao et al., 2022].

Finally, we have :

$$\begin{aligned}
\boldsymbol{X}_g \sim \mathcal{N}_n(\boldsymbol{\mu_g}, \sigma_g^2\boldsymbol{I}_n) &\Rightarrow \boldsymbol{X}_g^t\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}_g^t\boldsymbol{\eta}, \sigma_g^2\|\boldsymbol{\eta}\|_2^2) \\
&\Rightarrow \phi_g \overset{H_0}{\sim} \mathcal{N}\left(0, \sigma_g^2\|\boldsymbol{\eta}\|_2^2\right)
\end{aligned}$$

Thus, the p-value (4) is equal to :

$$\mathbb{P}_{H_0}\left(|\phi_g| > |\boldsymbol{X}_g{}^t\boldsymbol{\eta}| \mid \phi_g \in S_g\right)$$

with $S_g = \{\phi_g : C_k, C_l \in c\left(\boldsymbol{X}(\phi_g)\right)\}$ and $\phi_g \overset{H_0}{\sim} \mathcal{N}\left(0, \sigma_g^2\|\boldsymbol{\eta}\|_2^2\right)$

B Numerical computation of the selective p-value

Since the selective p-value given in (4) is intractable, in practice it is computed using a Monte-Carlo approach with important sampling resulting in the p-value described in (7). However, Phipson and Smyth [2010] showed that the classical Monte-Carlo estimator of a p-value could be biased for near to zero p-values. In fact, very tiny p-values could be approximated by exactly 0 using the Monte-Carlo approach, leading to statistical hypothesis testing that does not efficiently control the type I error rate. To overcome this problem, they propose to correct the Monte-Carlo p-value by adding 1 in both the numerator and the denominator of the estimated p-value. With this correction, instead of having exactly 0 Monte-Carlo p-value, the near to zero p-values are approximated by $\frac{1}{N+1}$ where N is the number of Monte-Carlo samples.

Unfortunately, we showed using numerical simulation studies that this correction could not work with the selective p-value computed as in (7) for two reasons. The first one is because this p-value originally come from a conditional probability (4), so by definition, there are in fact two probabilities to compute and to correct. The second problem arrives because of the important sample approach. In fact, because under $\mathcal{H}_1$ the π^i in (7) are very small, adding 1 will drastically change the scale of the p-value. So, because of the important sampling approach, we need to correct our Monte-Carlo p-value by adding a constant in the same order of π. We propose to add $\overline{\pi} = \frac{1}{N}\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}$ in both the numerator and the denominator of (7). This correction is reasonable since for small p-value, that is under $\mathcal{H}_1$ we have :

i) $|\boldsymbol{x}_g^t \boldsymbol{\eta}|$ is large because C_k and C_l are truly separated on $\boldsymbol{X}_g$

ii) $\sum_{i=1}^{N} \pi^i \mathbb{1}\{|\omega_g^i| > |\boldsymbol{x}_g^t \boldsymbol{\eta}|, C_k, C_l \in c(\boldsymbol{X}(\omega_g^i)\} \simeq 0$ since each $\pi^i = \frac{f_1(\omega_g^i)}{f_2(\omega_g^i)}$ where f_1 is the density of a gaussian distribution with mean 0. Then, because $\omega_g^i \sim \mathcal{N}(\boldsymbol{x}_g^t \boldsymbol{\eta}, \sigma^2 \|\boldsymbol{\eta}\|_2)$ where $|\boldsymbol{x}_g^t \boldsymbol{\eta}|$ is large, f_1 is evaluated in a point that is far away of the mean, and that is why $f_1(\omega_g^i) \simeq 0$.

So using i) and ii):

$$
\begin{aligned}
\frac{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{|\omega_g^i| \geq |\boldsymbol{x}_g^t \boldsymbol{\eta}|, C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \overline{\pi}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \overline{\pi}} &= \frac{0 + \overline{\pi}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \overline{\pi}} \\
&\simeq \frac{\frac{1}{N}\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \frac{1}{N}\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{C_k, C_l \in c(\boldsymbol{X}(\omega_g^i))\right\}\left[1 + \frac{1}{N}\right]} \\
&= \frac{\frac{1}{N}}{1 + \frac{1}{N}} \\
&= \frac{1}{N+1}
\end{aligned}
$$

So, by correcting our Monte-Carlo p-value by adding $\overline{\pi}$ we obtain the estimator proposed by Phipson and Smyth [2010] for small p-value.

C Supplementary Figure 1

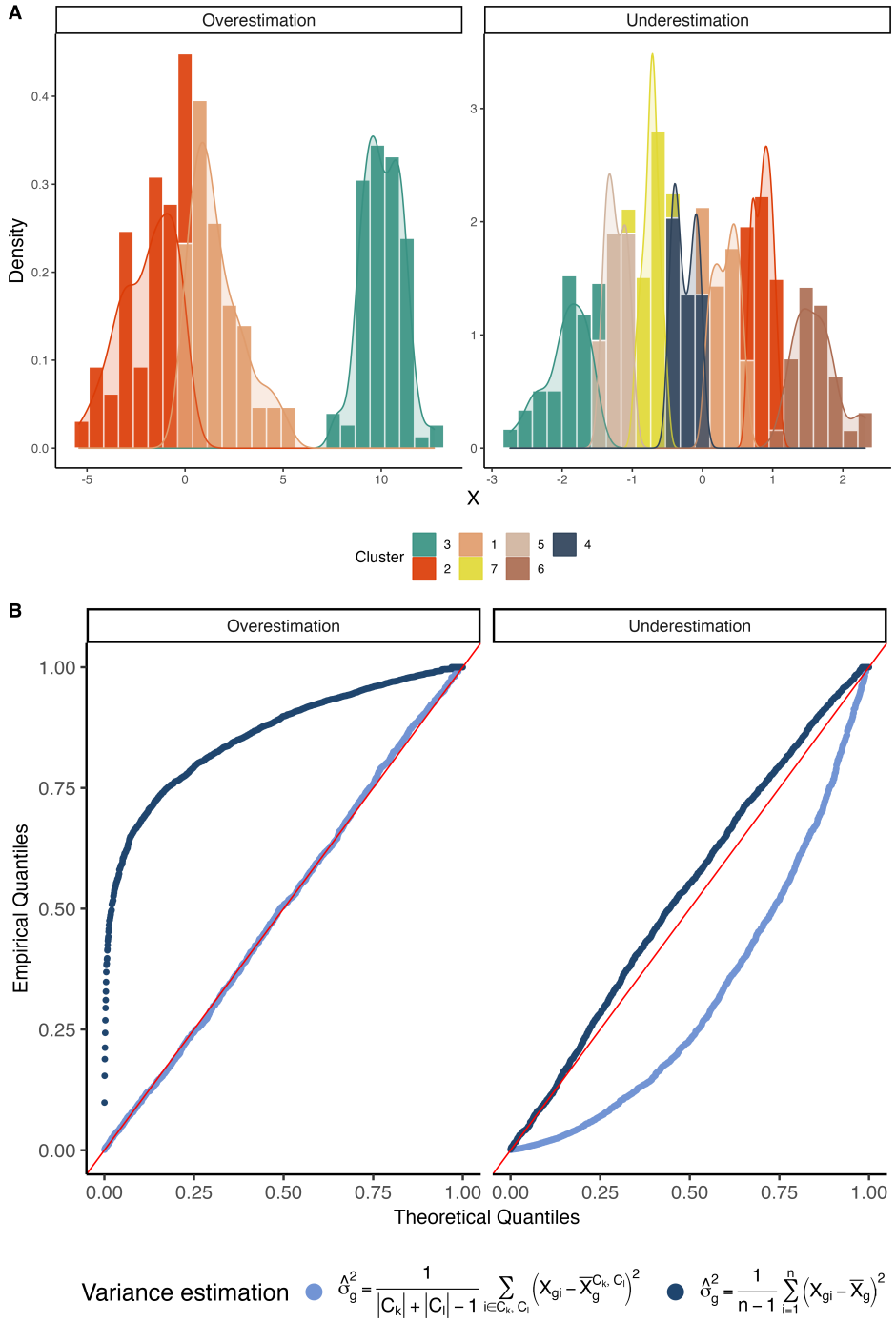

Figure 5: Impact of the variance estimation on the p-values of the selective test. **panel A** The data are simulated according to a two-component Gaussian mixture: $X \sim 0.5\mathcal{N}(0,4) + 0.5\mathcal{N}(10,1)$ for the overestimation panel and according to a standard Gaussian distribution with mean 0 and variance 1 for the underestimation panel. **panel B** QQ-plot of selective p-values against the uniform distribution according to the variance estimator for the test Cluster 1 vs. Cluster 2 for 2000 simulations of the data. Overestimation of variance leads to conservative p-values, while underestimation leads to false positives.

D Supplementary Figure 2

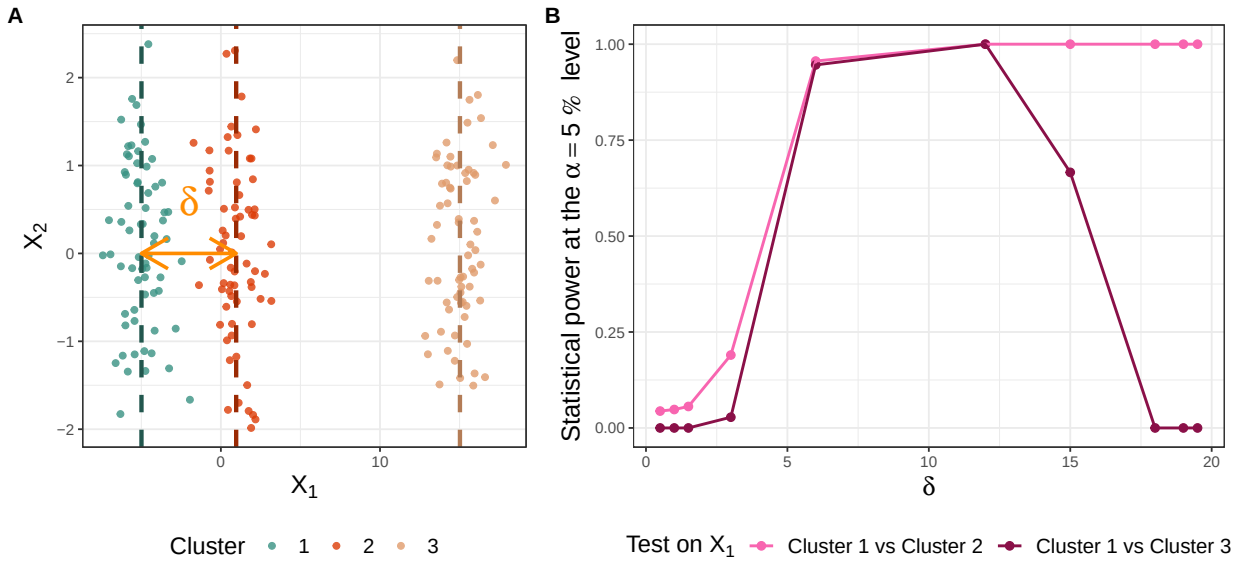

Figure 6: Illustration of the possible loss of statistical power of the selective test in cases where there are more than two estimated clusters. **panel A** Data generation process. A bivariate dataset is simulated such as three clusters are all separated only on X_1. Cluster 1 and Cluster 2 are separated according to a mean difference $\delta \in \{0.5, 1, 1.5, 3, 6, 12, 15, 18, 19, 19.5\}$. **panel B** Statistical power at the $\alpha = 5\%$ level of the selective test computed using 500 simulation of the data as described in **panel A** according to δ, the mean difference between Cluster 1 and Cluster 2. For each simulation, the selective test is applied to test the separation of Cluster 1 vs Cluster 2 and Cluster 1 vs Cluster 3 only on X_1.

E Supplementary Figure 3

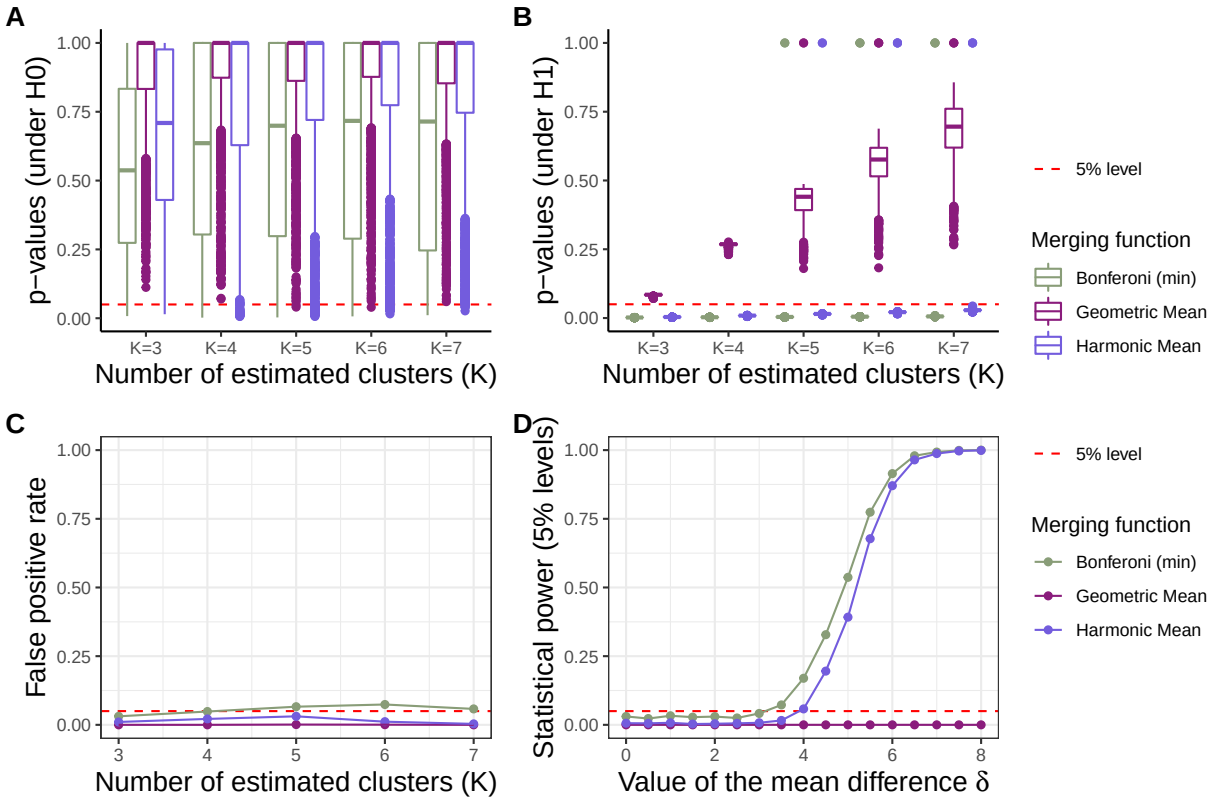

Figure 7: Comparison of three different merging functions presented in Vovk and Wang [2020]. **panel A** Distribution of the resulting merging p-values under H_0 as a function of the number of estimated clusters. **panel B** Distribution of the resulting merging p-values under H_1 (Gaussian mixture with only two components of equal proportion and variance) as a function of the number of estimated clusters. **panel C** False positive rate as a function of the number of estimated clusters. **panel D** Statistical power as a function of the mean difference δ between the two modes of the mixture where $K = 4$ clusters are estimated (the same simulation as in Figure 3. The selective test is always applied to the most extreme clusters, and in such a way that the maximum number of adjacent p-values are merged. 2000 simulations of the data were used.

F Supplementary Figure 4

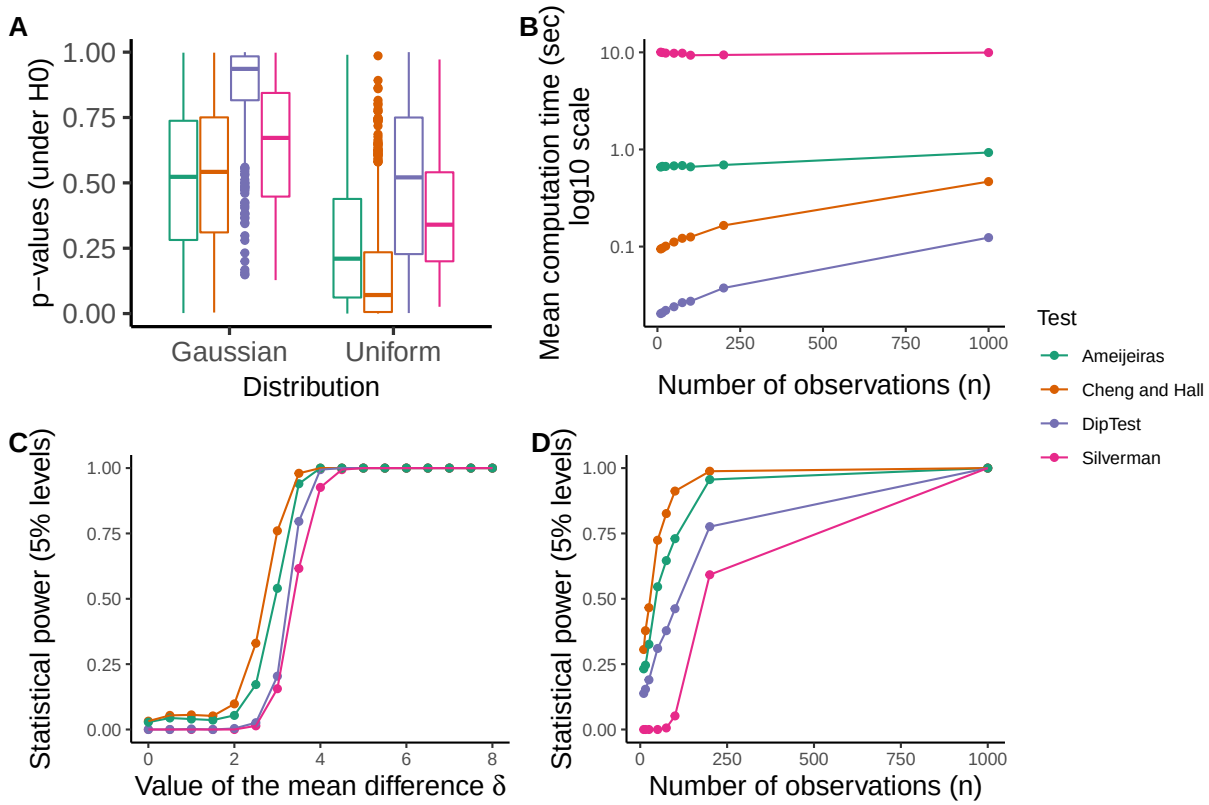

Figure 8: Comparison of different multimodality tests implemented in the `R` package `multimode`[Ameijeiras-Alonso et al., 2021]. **panel A** p-values of each multimodality tests under the null for 500 simulations of 200 realisations of the Gaussian and uniform distributions. **panel B** Mean computation time required by each test as a function of the number of observations n (averaging over the 500 simulations. **panel C** Statistical power (at the $\alpha = 5\%$ level) of each multimodality test as a function of δ, the difference in means between two modes of a two-components Gaussian mixture ($n = 200$ observations). **panel D** Statistical power(at the $\alpha = 5\%$ level) of each multimodality test as a function of the number of observations for $\delta = 3.5$ fixed.

G Supplementary Figure 5

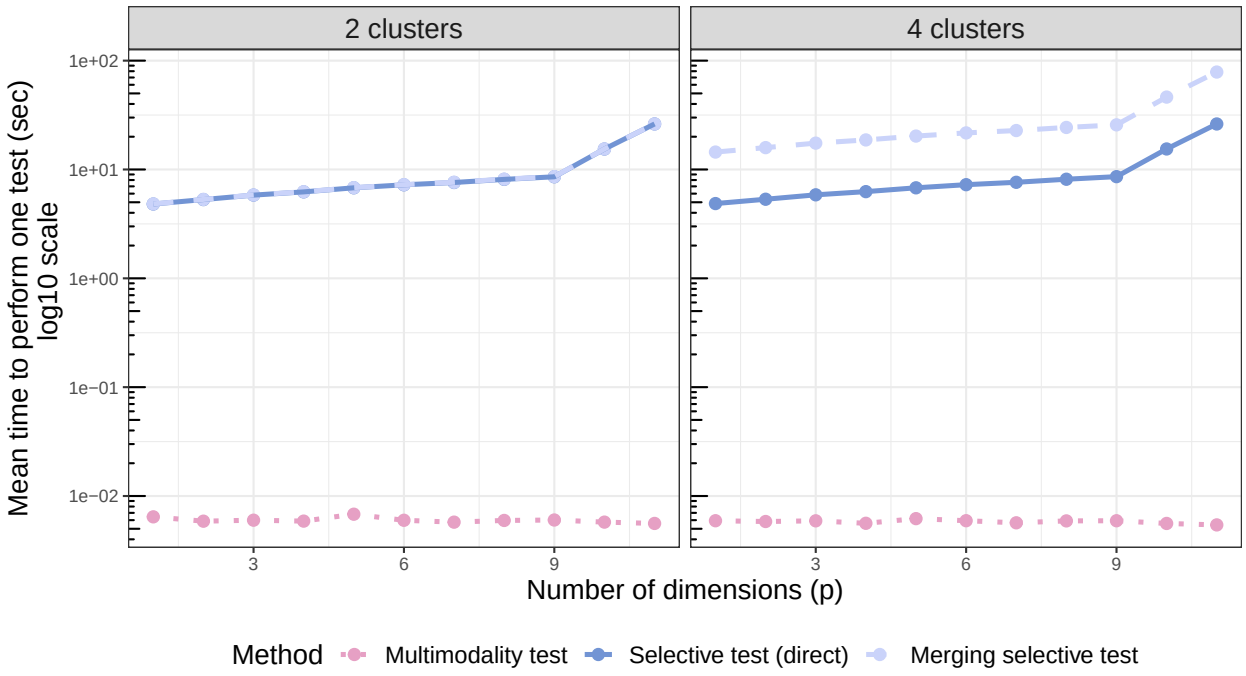

Figure 9: Mean computational time of the three proposed tests (based on 500 simulations of the data) according to the number of dimensions (p) of $\boldsymbol{X}$. The tests are performed only for the first variable, so the dimensionality of the data only affects the computation times of the selective tests since the clustering method must be applied on the data for each Monte-Carlo simulation using all the dimensions of $\boldsymbol{X}$

H Supplementary Figure 6

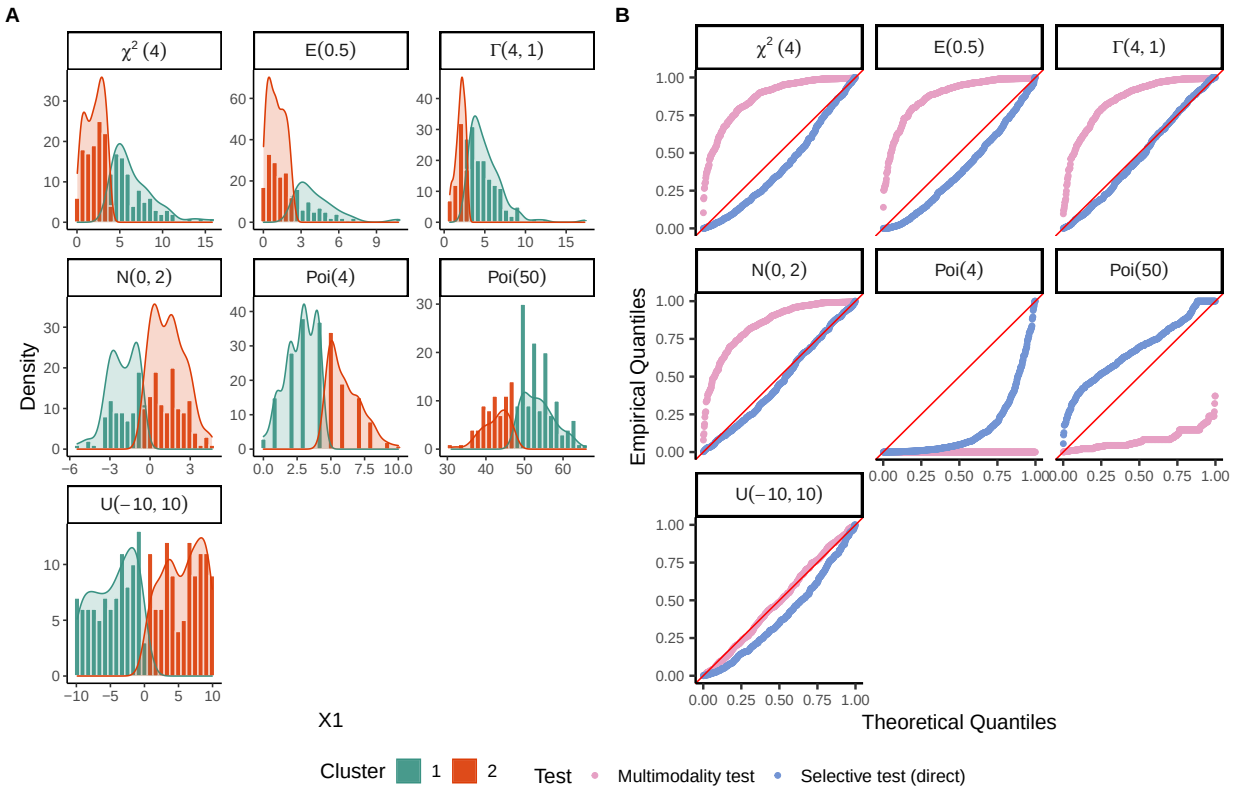

Figure 10: Robustness of our tests for misspecification of the distributional assumption under the null. **panel A** Data generation. The behaviour of our tests was studied on 7 different univariate and unimodal distributions. For each distribution, Ward's clustering on euclidean distance was applied to build 2 clusters. **panel B** QQ-plot against the Uniform distribution of the p-values returned by our two tests for 500 simulations of each distribution.

I Supplementary Table 1

Cluster pair	bill length	bill depth	flipper length	body mass
Cluster 1 vs Cluster 2	1.67	1.53	1.94	1.75
Cluster 1 vs Cluster 3	0.16	1.93	0.50	0.16
Cluster 2 vs Cluster 3	1.83	0.40	1.44	1.59

Table 3: Values of the mean difference (δ) on each (scaled) variable between each estimated pair of clusters