Journal of Machine Learning Research 24 (2023) 1-20 Submitted 4/22; Revised 11/22; Published 5/23

Selective inference for k-means clustering

Yiqun T. Chen YIQUNC@STANFORD.EDU
Data Science Institute and Department of Biomedical Data Science
Stanford University
Stanford, CA 94305, USA

Daniela M. Witten DWITTEN@UW.EDU
Departments of Statistics and Biostatistics
University of Washington
Seattle, WA 98195-4322, USA

Editor: Ji Zhu

Abstract

We consider the problem of testing for a difference in means between clusters of observations identified via k-means clustering. In this setting, classical hypothesis tests lead to an inflated Type I error rate. In recent work, Gao et al. (2022) considered a related problem in the context of hierarchical clustering. Unfortunately, their solution is highly-tailored to the context of hierarchical clustering, and thus cannot be applied in the setting of k-means clustering. In this paper, we propose a p-value that conditions on all of the intermediate clustering assignments in the k-means algorithm. We show that the p-value controls the selective Type I error for a test of the difference in means between a pair of clusters obtained using k-means clustering in finite samples, and can be efficiently computed. We apply our proposal on hand-written digits data and on single-cell RNA-sequencing data.

Keywords: Post-selection inference, Unsupervised learning, Hypothesis testing, Type I error, RNA-sequencing

1. Introduction

Testing for a difference in means between two groups is one of the most fundamental tasks in statistics, with numerous applications. If the groups under investigation are *pre-specified*, i.e., not a function of the observed data, then classical hypothesis tests will control the Type I error rate. However, it is increasingly common to want to test for a difference in means between groups that are *defined through the observed data*, e.g., via the output of a clustering algorithm. For instance, in single-cell RNA-sequencing analysis, researchers often first cluster the cells, and then test for a difference in the expected gene expression levels between the clusters to quantify up- or down-regulation of genes, annotate known cell types, and identify new cell types (Grün et al., 2015; Aizarani et al., 2019; Lähnemann et al., 2020; Zhang et al., 2019; Doughty and Kerkhoven, 2020). In fact, the inferential challenges resulting from testing data-guided hypotheses have been described as a "grand challenge" in the field of genomics (Lähnemann et al., 2020), and papers in the field continue to overlook this issue: as an example, `seurat` (Stuart et al., 2019), the state-of-the-art single-cell RNA sequencing analysis tool, tests for differential gene expression between groups obtained via clustering, with a note that "p-values [from these hypotheses] should be interpreted cautiously, as the genes used for clustering are the same genes tested for differential expression." Testing data-guided hypothesis also arises in the field of neuroscience (Kriegeskorte et al., 2009; Button, 2019), social psychology (Hung and Fithian, 2020), and physical sciences (Friederich et al., 2020; Pollice et al., 2021). When the null hypothesis is a function of the data, classical tests that do not account for this will fail to control the Type I error.

In this paper, we develop a test for a difference in means between two clusters estimated from applying k-means clustering (Lloyd, 1982; MacQueen et al., 1967), an extremely popular clustering algorithm with numerous applications (Xu and Wunsch, 2008). In recent work, Gao et al. (2022) tackled a similar problem for hierarchical clustering. While the two papers share similar notation and setup, our solutions and algorithms are tailored to the iterative and centroid-based nature of k-means clustering, leading to fundamentally different solutions and algorithms than those proposed in Gao et al. (2022). We consider the following simple and well-studied model (Gao et al., 2022; Löffler et al., 2021; Lu and Zhou, 2016) for n observations and q features:

$$X \sim \mathcal{MN}_{n\times q}\left(\mu, \mathbf{I}_n, \sigma^2\mathbf{I}_q\right), \tag{1}$$

where $\mathcal{MN}$ denotes the matrix normal distribution (Bilodeau and Brenner, 1999), $\mu \in \mathbb{R}^{n\times q}$ has unknown rows μ_i, and $\sigma^2 > 0$ is known. Given a realization $x \in \mathbb{R}^{n\times q}$ of X, we first apply the k-means clustering algorithm to obtain $\mathcal{C}(x)$, a partition of the samples $\{1,\ldots,n\}$. We might then consider testing the null hypothesis that the mean is the same across two *estimated* clusters, i.e.,

$$H_0 : \sum_{i\in\hat{\mathcal{C}}_1} \mu_i/|\hat{\mathcal{C}}_1| = \sum_{i\in\hat{\mathcal{C}}_2} \mu_i/|\hat{\mathcal{C}}_2| \text{ versus } H_1 : \sum_{i\in\hat{\mathcal{C}}_1} \mu_i/|\hat{\mathcal{C}}_1| \neq \sum_{i\in\hat{\mathcal{C}}_2} \mu_i/|\hat{\mathcal{C}}_2|, \tag{2}$$

where $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(x)$ are estimated clusters with cardinality $|\hat{\mathcal{C}}_1|$ and $|\hat{\mathcal{C}}_2|$. This is equivalent to testing $H_0 : \mu^\top\nu = 0_q$ versus $H_1 : \mu^\top\nu \neq 0_q$, where

$$\nu_i = 1\Big\{i \in \hat{\mathcal{C}}_1\Big\}/|\hat{\mathcal{C}}_1| - 1\Big\{i \in \hat{\mathcal{C}}_2\Big\}/|\hat{\mathcal{C}}_2|, \quad i = 1,\ldots,n, \tag{3}$$

and $1\{A\}$ equals 1 if the event A holds, and 0 otherwise. Gao et al. (2022) demonstrates that the p-value given by

$$p_{\text{Naive}} = \text{pr}_{H_0}\Big(\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2\Big), \tag{4}$$

where $\|X^\top\nu\|_2 \sim (\sigma\|\nu\|_2)\chi_q$ under H_0, leads to an extremely anti-conservative test. In particular, we constructed the contrast vector in (3) because $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ were obtained by clustering. Therefore, we will observe substantial differences between the cluster centroids $\sum_{i\in\hat{\mathcal{C}}_1} x_i/|\hat{\mathcal{C}}_1|$ and $\sum_{i\in\hat{\mathcal{C}}_2} x_i/|\hat{\mathcal{C}}_2|$, even in the absence of true differences in their population means (left panel Figure 1).

Notably, the problem of testing for a difference in means between two groups obtained via clustering cannot be easily overcome by sample splitting, as pointed out in Gao et al. (2022) and Zhang et al. (2019). To see why, we divide the observations into a training and a test set. We apply k-means clustering on only the training set (left panel of Figure 1), and then assign the test set observations to those clusters (to obtain the center panel of Figure 1,

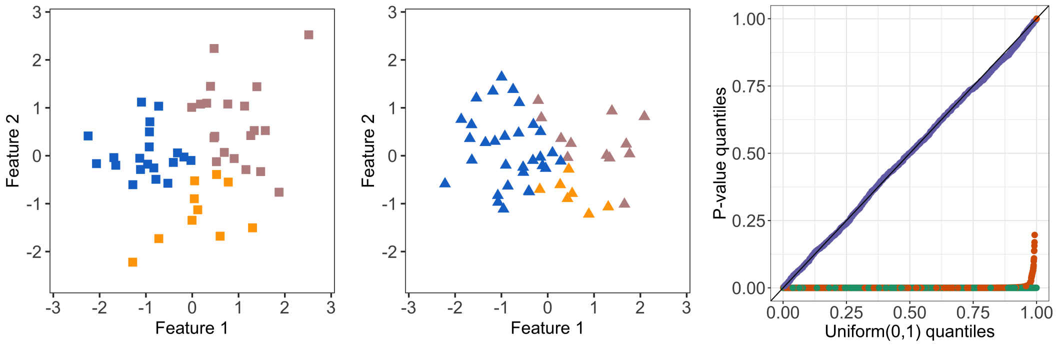


Figure 1: We simulated one dataset from (1) with $\mu = 0_{100\times 2}$ and $\sigma = 1$. We split the data into training (*left*) and test sets (*middle*). *Left:* We apply k-means clustering on the training set to obtain three clusters. *Center:* We apply the training set clusters to the test set using a 3-nearest neighbors classifier. *Right:* Quantile-quantile plot of the naive p-values (4) applied to the training set (green) and the test set (orange), aggregated over 2,000 simulated datasets; as well as our proposed p-values (in (9); displayed in purple) applied to the training set.

we applied a 3-nearest neighbor classifier). Finally, we compute the naive p-values (4) *only* on the test set. Unfortunately, this approach does not work: while we clustered only the training data, we still used the test data to label the test observations, and consequently to construct the contrast vector ν in (3). Therefore, the Wald test based on sample-splitting remains extremely anti-conservative, as shown in the right panel of Figure 1, and does not lead to a valid test of H_0 in (2). We refer the readers to Gao et al. (2022) for further discussion of this point.

In this paper, we develop a test of H_0 that controls the selective Type I error. That is, we wish to ensure that the probability of rejecting H_0 at level α, given that H_0 holds and we decided to test it, is no greater than α:

$$\mathrm{pr}_{H_0}(\text{reject } H_0 \text{ at level } \alpha \mid H_0 \text{ is tested}) \leq \alpha, \ \forall \alpha \in (0,1). \tag{5}$$

To develop the test, we leverage the selective inference framework, which has been applied extensively in high-dimensional linear modeling (Lee et al., 2016; Tibshirani et al., 2016; Fithian et al., 2014; Rügamer et al., 2022; Schultheiss et al., 2021; Taylor and Tibshirani, 2018; Charkhi and Claeskens, 2018; Yang et al., 2016; Loftus and Taylor, 2015), changepoint detection (Jewell et al., 2022; Hyun et al., 2021, 2018; Chen et al., 2021; Le Duy and Takeuchi, 2021; Duy et al., 2020; Benjamini et al., 2019), and clustering (Zhang et al., 2019; Gao et al., 2022; Watanabe and Suzuki, 2021). The key insight behind selective inference is as follows: naive p-values such as (4) lead to anti-conservative tests because the hypothesis H_0 is generated by the same data used for testing. Therefore, to obtain a valid test of H_0, we need to condition on the aspect of the data that led us to test H_0. In our case, we chose to test the null hypothesis in (2) because $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ were obtained via k-means clustering. Therefore, we compute a p-value conditional on the event that k-means clustering yields $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$. This results in selective Type I error control (5), as seen in the right panel of Figure 1.

There is a rich literature on estimating and quantifying the uncertainty in the number of clusters (Li and Chen, 2010; Chen and Li, 2009; Chen et al., 2004; McLachlan et al., 2019; Dobriban, 2020), as well as assessing cluster stability and heterogeneity (Suzuki and Shimodaira, 2006; Kerr and Churchill, 2001; Kimes et al., 2017; Chung, 2020; Jin and Wang, 2016; Aw et al., 2021; Chung and Storey, 2015). Others have examined the asymptotic properties of clustering models from a Bayesian perspective (Guha et al., 2019; Nobile, 2004; Cai et al., 2020). In addition, k-means clustering is a special case of the expectation-maximization algorithm, which allows us to tap into an active line of research on the statistical guarantees of the expectation-maximization algorithm (Zhang and Zhang, 2014; Wang et al., 2015; Cai et al., 2019; Yi and Caramanis, 2015; Balakrishnan et al., 2017). However, most prior work focused the setting with one or more "true" clusters. By contrast, we are interested in a correctly-sized test for the null hypothesis (2), even when $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$ do not correspond to "true" clusters, and even in the absence of "true" clusters in the data. In addition, existing work often relies on asymptotic approximations and bootstrap resampling. Two recent exceptions include Zhang et al. (2019) and Gao et al. (2022), who took a selective inference approach and computed finite-sample p-values for testing the difference in means between estimated clusters obtained via linear classification rules and hierarchical clustering, respectively. Our work is closest to Gao et al. (2022), and extends their framework to k-means clustering. We provide an exact, finite-sample test of the difference in means between a pair of clusters estimated via k-means clustering under model (1), without the need for sample splitting.

The rest of this paper is organized as follows. In Section 2, we briefly review the work of Gao et al. (2022), and outline our proposed test of a difference in means after k-means clustering. It is worth highlighting that while our proposal is inspired by the work of Gao et al. (2022), our solution is *not* simply a minor modification: computing the conditioning set for the p-value in (9) is the key technical challenge of this paper, and the computational insights in Gao et al. (2022) are only applicable to hierarchical clustering. In Section 3, we provide a computationally-efficient approach to compute the p-values corresponding to our proposed test, for a difference in means after k-means clustering. Section 4 outlines some extensions, and we evaluate our proposal in a simulation study in Section 5. We apply our proposal to three real datasets in Section 6, and discuss future work in Section 7. Proofs and additional results are relegated to the Appendix.

Throughout this paper, we will use the following notational conventions. For a matrix A, A_i denotes the ith row and A_{ij} denotes the (i,j)th entry. For a vector $\nu \in \mathbb{R}^n$, $\|\nu\|_2$ denotes its ℓ_2 norm, and $\Pi_\nu^\perp$ is the projection matrix onto the orthogonal complement of ν, i.e., $\Pi_\nu^\perp = \mathbf{I}_n - \nu\nu^\top/\|\nu\|_2^2$, where $\mathbf{I}_n$ is the n-dimensional identity matrix. Moreover, $\text{dir}(\nu) = \nu/\|\nu\|_2$ if $\nu \neq 0_n$ and 0_n otherwise, where 0_n is the n-vector of zeros. We let $\langle \cdot, \cdot \rangle$ and $1\{\cdot\}$ denote the inner product of two vectors and the indicator function, respectively.

2. Selective inference for k-means clustering

2.1 A brief review of k-means clustering

In this section, we review the k-means clustering algorithm. Given samples $x_1, \ldots, x_n \in \mathbb{R}^q$ and a positive integer K, k-means clustering partitions the n samples into disjoint subsets

$\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K$ by solving the following optimization problem:

$$\begin{aligned} &\underset{\mathcal{C}_1,\ldots,\mathcal{C}_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \left\| x_i - \sum_{i \in \mathcal{C}_k} x_i / |\mathcal{C}_k| \right\|_2^2 \right\} \\ &\text{subject to } \bigcup_{k=1}^{K} \mathcal{C}_k = \{1, \ldots, n\}, \mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset, \forall k \neq k'. \end{aligned} \tag{6}$$

It is not typically possible to solve for the global optimum in (6) (Aloise et al., 2009). A number of algorithms are available to find a local optimum (Hartigan and Wong, 1979; Zha et al., 2002; Arthur and Vassilvitskii, 2007); one such approach is Lloyd's algorithm (Lloyd, 1982), given in Algorithm 1. We first sample K out of n observations as initial centroids (step 1 in Algorithm 1). We then assign each observation to the closest centroid (step 2). Next, we iterate between re-computing the centroids and updating the cluster assignments (steps 3a. and 3b.) until the cluster assignments stop changing. The algorithm is guaranteed to converge to a local optimum (Hastie et al., 2001).

In what follows, we will sometimes use $c_i^{(t)}(x)$ and $m_k^{(t)}(x)$ rather than $c_i^{(t)}$ and $m_k^{(t)}$ to emphasize the dependence of the cluster labels and centroids on the data x.

Algorithm 1: Lloyd's algorithm for k-means clustering (Lloyd, 1982)

```
Input: Data $x_1, \ldots, x_n \in \mathbb{R}^q$, number of output clusters $K$, maximum iteration $T$,
       random seed $s$.
Output: Cluster assignments $\left(c_1^{(t)}, \ldots, c_n^{(t)}\right)$.
1. Initialize the centroids $\left(m_1^{(0)}, \ldots, m_K^{(0)}\right)$ by sampling $K$ observations from
   $x_1, \ldots, x_n$ without replacement, using the random seed $s$.
2. Compute assignments $c_i^{(0)} \leftarrow \underset{1 \leq k \leq K}{\text{argmin}} \left\| x_i - m_k^{(0)} \right\|_2^2$, $i = 1, \ldots, n$.
3. Initialize $t = 0$.
while $t \leq T$ do
    a. Update centroids: $m_k^{(t+1)} \leftarrow \left(\sum_{i: c_i^{(t)} = k} x_i\right) / \sum_{i=1}^{n} 1\left\{c_i^{(t)} = k\right\}$, $k = 1, \ldots, K$.
    b. Update assignment: $c_i^{(t+1)} \leftarrow \underset{1 \leq k \leq K}{\text{argmin}} \left\| x_i - m_k^{(t+1)} \right\|_2^2$, $i = 1, \ldots, n$.
    c. if $c_i^{(t+1)} = c_i^{(t)}$ for all $1 \leq i \leq n$
           break
    else
           $t \leftarrow t + 1$.
end
return $\left(c_1^{(t)}, \ldots, c_n^{(t)}\right)$.
```

2.2 A test of (2) for clusters obtained via k-means clustering

Here, we briefly review the proposal of Gao et al. (2022) for selective inference for hierarchical clustering, and outline a selective test for (2) for k-means clustering.

Gao et al. (2022) proposed a selective inference framework for testing hypotheses based on the output of a clustering algorithm. Let $\mathcal{C}(\cdot)$ denote the clustering operator, i.e., a partition of the observations resulting from a clustering algorithm. Since H_0 in (2) is chosen because $\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(x)\right\}$, where $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$ are the two estimated clusters under consideration in (2), Gao et al. (2022) proposed to reject H_0 if

$$\mathrm{pr}_{H_0}\left\{\|X^\top \nu\|_2 \geq \|x^\top \nu\|_2 \;\middle|\; \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(X), \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right\} \tag{7}$$

is small. In (7), conditioning on $\left\{\Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right\}$ eliminates the nuisance parameters $\Pi_\nu^\perp \mu$ and $\mathrm{dir}(\mu^\top \nu)$, where $\Pi_\nu^\perp = \mathbf{I}_n - \nu\nu^\top/\|\nu\|_2$ and $\mathrm{dir}(\mu^\top \nu) = \mu^\top \nu / \|\mu^\top \nu\|_2$ (see, e.g., Section 3.1 of Fithian et al. (2014)). Gao et al. (2022) showed that the test that rejects H_0 when (7) is below α controls the selective Type I error at level α, in the sense of (5). Furthermore, under (1), the conditional distribution of $\|X^\top \nu\|_2$ in (7) is $(\sigma\|\nu\|_2)\chi_q$, truncated to a set. When the operator $\mathcal{C}(\cdot)$ denotes hierarchical clustering, this set can be analytically characterized and efficiently computed, leading to an efficient algorithm for computing (7).

We now extend these ideas to k-means clustering (6). Since the k-means algorithm partitions all n observations, it is natural to condition on the cluster assignments of *all* observations rather than just on $\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(X)\right\}$. This leads to the p-value

$$\mathrm{pr}_{H_0}\left[\|X^\top \nu\|_2 \geq \|x^\top \nu\|_2 \;\middle|\; \bigcap_{i=1}^{n} \left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right], \tag{8}$$

where $c_i^{(T)}(X)$ is the cluster assigned to the ith observation at the final iteration of Algorithm 1. However, computing (8) requires characterizing $\bigcap_{i=1}^n \left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}$, which is not straightforward, and may necessitate enumerating over possibly an exponential number of intermediate cluster assignments $c_i^{(t)}(\cdot)$ for $t = 1, \ldots, T-1$. Hence, we also condition on *all of the intermediate clustering assignments* in Algorithm 1:

$$\begin{aligned} p_{\text{selective}} = \mathrm{pr}_{H_0}\Bigg[\|X^\top \nu\|_2 \geq \|x^\top \nu\|_2 \;\Bigg|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n} \left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \\ \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\Bigg]. \end{aligned} \tag{9}$$

In (9), $c_i^{(t)}(X)$ is the cluster assigned to the ith observation at the tth iteration of Algorithm 1. Roughly speaking, this p-value answers the question:

> Assuming that there is no difference between the population means of $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$, what is the probability of observing such a large difference between their centroids, among all the realizations of X that yield identical results in every iteration of the k-means algorithm?

The p-value in (9) is the focus of this paper. We establish its key properties below.

Proposition 1 *Suppose that x is a realization from* (1), *and let $\phi \sim (\sigma\|\nu\|_2)\chi_q$. Then, under $H_0 : \mu^\top \nu = 0$ with ν defined in* (3),

$$p_{selective} = \text{pr}\left[\phi \geq \|x^\top \nu\|_2 \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right], \tag{10}$$

where $p_{selective}$ is defined in (9), *and*

$$x'(\phi) = x + \left(\phi - \|x^\top \nu\|_2\right)\left(\nu / \|\nu\|_2^2\right) \left\{\text{dir}(x^\top \nu)\right\}^\top . \tag{11}$$

Moreover, the test that rejects $H_0 : \mu^\top \nu = 0$ when $p_{selective} \leq \alpha$ controls the selective Type I error at level α, in the sense of (5).

Proposition 1 states that $p_{\text{selective}}$ can be recast as the survival function of a scaled χ_q random variable, truncated to the set

$$\mathcal{S}_T = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}, \tag{12}$$

where $x'(\phi)$ is defined in (11). Therefore, to compute $p_{\text{selective}}$, it suffices to characterize the set $\mathcal{S}_T$. In (11), $x'(\phi)$ results from applying a perturbation to the observed data x, along the direction of $x^\top \nu$, the difference between the two cluster centroids of interest. Figure 2 illustrates a realization of (1) for k-means clustering with $K = 3$. The left panel displays the observed data x, which corresponds to $x'(\phi)$ with $\phi = \|x^\top \nu\|_2 = 4.37$. Here, ν defined in (3)

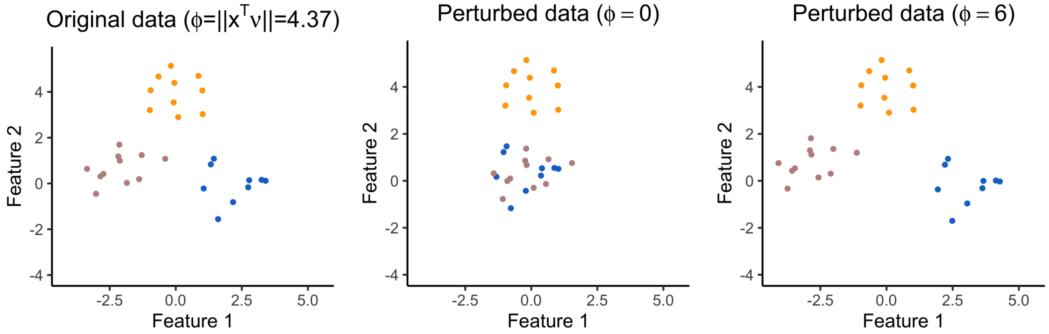


Figure 2: One simulated dataset generated from model (1) with $\mu_i = 1\{1 \leq i \leq 10\}[2.5, 0]^\top + 1\{11 \leq i \leq 20\}[0, -2.5]^\top + 1\{21 \leq i \leq 30\}[\sqrt{18.75}, 0]^\top$ and $\sigma = 1$. *Left:* The original data x corresponds to $\phi = \|x^\top \nu\|_2 = 4.37$. Applying k-means clustering with $K = 3$ yields three clusters, displayed in rosy brown, blue, and orange. Here, ν is chosen to test for a difference in means between $\hat{\mathcal{C}}_1$ (rosy brown) and $\hat{\mathcal{C}}_2$ (blue). *Center:* The perturbed data $x'(\phi)$ with $\phi = 0$. Applying k-means clustering with $K = 3$ does not yield the same set of clusters as in the left panel. *Right:* The perturbed data $x'(\phi)$ with $\phi = 6$. Applying k-means clustering with $K = 3$ yields the same set of clusters as in the left panel.

was chosen to test the difference between $\hat{\mathcal{C}}_1$ (shown in rosy brown) and $\hat{\mathcal{C}}_2$ (shown in blue). The center and right panels of Figure 2 display $x'(\phi)$ with $\phi = 0$ and $\phi = 6$, respectively. In the center panel, with $\phi = 0$, the blue and rosy brown clusters are "pushed together", resulting in $\|x'(\phi)^\top \nu\|_2 = 0$; that is, there is no difference in empirical means between the two clusters under consideration. Applying k-means clustering no longer results in these clusters. By contrast, in the right panel, with $\phi = 6$, the blue and rosy brown clusters are "pulled apart" along the direction of $x^\top \nu$, which results in an increased distance between the centroids of the blue and rosy brown clusters, and k-means clustering does yield the same clusters as on the original data. In this example, $\mathcal{S}_T = (3.59, \infty)$.

3. Computation of the selective p-value

In Section 2, we have shown that the p-value $p_{\text{selective}}$ (9) involves the set $\mathcal{S}_T$ in (12). Indeed, a computationally-efficient characterization of $\mathcal{S}_T$ is the key technical challenge and contribution of our paper. Here, we start with a high-level summary of our approach to characterizing $\mathcal{S}_T$ in (12). First, we rewrite

$$\mathcal{S}_T = \left\{ \phi \in \mathbb{R} : \bigcap_{i=1}^{n} \left\{ c_i^{(0)}\left(x'(\phi)\right) = c_i^{(0)}\left(x\right) \right\} \right\} \cap \left\{ \phi \in \mathbb{R} : \bigcap_{t=1}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}. \tag{13}$$

Next, we consider the first term in (13): according to step 2. of Algorithm 1, for $i = 1, \ldots, n$, $c_i^{(0)}\left(x'(\phi)\right) = c_i^{(0)}\left(x\right)$ if and only if for $i = 1, \ldots, n$, the initial randomly-sampled centroid to which $\left[x'(\phi)\right]_i$ is closest coincides with the initial centroid to which x_i is closest. This condition can be expressed using $K - 1$ inequalities. Furthermore, the same intuition holds for the second term in (13), except that the centroids are a function of the cluster assignments in the previous iteration. We formalize this intuition in Proposition 2, proven in Appendix A.2.

Proposition 2 *Suppose that we apply the k-means clustering algorithm (Algorithm 1) to a matrix $x \in \mathbb{R}^{n \times q}$, to obtain K clusters in at most T steps. Define*

$$w_i^{(t)}(k) = 1\left\{ c_i^{(t)}(x) = k \right\} / \sum_{i'=1}^{n} 1\left\{ c_{i'}^{(t)}(x) = k \right\}. \tag{14}$$

Then, for the set $\mathcal{S}_T$ defined in (12), we have that

$$\mathcal{S}_T = \left(\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[x'(\phi)\right]_i - m^{(0)}_{c_i^{(0)}(x)}\left(x'(\phi)\right) \right\|_2^2 \leq \left\| \left[x'(\phi)\right]_i - m_k^{(0)}\left(x'(\phi)\right) \right\|_2^2 \right\} \right) \cap \tag{15}$$

$$\left(\bigcap_{t=1}^{T} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}\left(c_i^{(t)}(x)\right) \left[x'(\phi)\right]_{i'} \right\|_2^2 \leq \left\| \left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k) \left[x'(\phi)\right]_{i'} \right\|_2^2 \right\} \right). \tag{16}$$

Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned in step 3b. of Algorithm 1 during the tth iteration, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from the data x during step 1 of Algorithm 1. In words, Proposition 2 says that $\mathcal{S}_T$ can be expressed as the intersection of $\mathcal{O}(nKT)$ sets. Therefore, it suffices to characterize the sets in (15) and (16). We now present two lemmas.

Lemma 3 (Lemma 2 in Gao et al. (2022)) *For ν in (3) and $x'(\phi)$ in (11), we have that $\left\|[x'(\phi)]_i - [x'(\phi)]_j\right\|_2^2 = a\phi^2+b\phi+\gamma$, where $a = \left\{(\nu_i - \nu_j)/\|\nu\|_2^2\right\}^2$, $b = 2\left[(\nu_i-\nu_j)/\|\nu\|_2^2 \left\langle x_i - x_j, \mathrm{dir}(x^\top\nu)\right\rangle - \left\{(\nu_i - \nu_j)/\|\nu\|_2^2\right\}^2\|x^\top\nu\|_2\right]$, and $\gamma = \|x_i - x_j - (\nu_i - \nu_j)(x^\top\nu)/\|\nu\|_2^2\|_2^2$.*

Lemma 4 *For ν in (3), $x'(\phi)$ in (11), and $w_i^{(t)}(k)$ in (14), we have that $\left\|[x'(\phi)]_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)[x'(\phi)]_{i'}\right\|_2^2 = \tilde{a}\phi^2 + \tilde{b}\phi + \tilde{\gamma}$, where $\tilde{a} = \left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)^2/\|\nu\|_2^4$, $\tilde{b} = (2/\|\nu\|_2^2)\left\{\left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)\left\langle x_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)x_{i'}, \mathrm{dir}(x^\top\nu)\right\rangle - \left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)^2 (\|x^\top\nu\|_2)/\|\nu\|_2^4\right\}$, $\tilde{\gamma} = \left\|x_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)x_{i'} - \left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)(x^\top\nu)/\|\nu\|_2^2\right\|_2^2$.*

It follows from Lemmas 3 and 4 that all of the inequalities in (15) and (16) are in fact *quadratic* in ϕ, with coefficients that can be analytically computed. Therefore, computing the set $\mathcal{S}_T$ requires solving $\mathcal{O}(nKT)$ quadratic inequalities of ϕ.

Proposition 5 *Suppose that we apply the k-means clustering algorithm (Algorithm 1) to a matrix $x \in \mathbb{R}^{n\times q}$, to obtain K clusters in at most T steps. Then, the set $\mathcal{S}_T$ defined in (12) can be computed in $\mathcal{O}(KT(n+q) + nKT\log(nKT))$ operations.*

4. Extensions

4.1 Non-spherical covariance matrix

Thus far, we have assumed that the observed data x is a realization of (1), which implies that $\mathrm{cov}(X_i) = \sigma^2\mathbf{I}_q$. However, this assumption is often violated in practice. For example, expression levels of genes are highly correlated, and neighbouring pixels in an image tend to be more similar. For a known positive definite matrix Σ, we now let

$$X \sim \mathcal{MN}_{n\times q}\left(\mu, \mathbf{I}_n, \Sigma\right). \tag{17}$$

Under (17), we can whiten the data by applying the transformation $x_i \to \Sigma^{-\frac{1}{2}}x_i$ (Bell and Sejnowski, 1997), where $\Sigma^{-\frac{1}{2}}$ is the unique symmetric positive definite square root of Σ^{-1} (Horn and Johnson, 2012). Note that $\Sigma^{-\frac{1}{2}}X_i \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\mu_i, \mathbf{I}_q)$. Moreover, as $\Sigma^{-\frac{1}{2}} \succ 0$, testing the null hypothesis in (2) is equivalent to testing

$$H_0 : \sum_{i\in\hat{\mathcal{C}}_1} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_1| = \sum_{i\in\hat{\mathcal{C}}_2} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_2| \text{ versus } H_1 : \sum_{i\in\hat{\mathcal{C}}_1} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_1| \neq \sum_{i\in\hat{\mathcal{C}}_2} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_2|. \tag{18}$$

Therefore, to get a correctly-sized test under model (17), we can simply carry out our proposal in Section 2 on the transformed data $\Sigma^{-\frac{1}{2}}x_i$ instead of the original data x_i.

Instead of applying the whitening transformation, we can directly accommodate a known covariance matrix Σ by considering the following extension of $p_{\text{selective}}$ in (9):

$$\begin{aligned} p_{\Sigma,\text{selective}} = \mathrm{pr}_{H_0}\Big[\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \mid \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X\right) = c_i^{(t)}\left(x\right)\right\}, \\ \Pi_\nu^\perp X = \Pi_\nu^\perp x,\ \mathrm{dir}\left(\Sigma^{-\frac{1}{2}}X^\top\nu\right) = \mathrm{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\Big]. \end{aligned} \tag{19}$$

Proposition 6 *Suppose that x is a realization from* (17)*, and let $\phi \sim (\|\nu\|_2)\chi_q$. Then, under $H_0 : \mu^\top \nu = 0$ with ν defined in* (3)*,*

$$p_{\Sigma,selective} = \mathrm{pr}\left[\phi \geq \|\Sigma^{-\frac{1}{2}} x^\top \nu\|_2 \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\phi \frac{\nu}{\|\nu\|_2^2}\right)\left\{\mathrm{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu\right)\right\}^\top \Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x) \right\}\right], \tag{20}$$

where $p_{\Sigma,selective}$ is defined in (19)*. Furthermore, the test that rejects $H_0 : \mu^\top \nu = 0$ when $p_{\Sigma,selective} \leq \alpha$ controls the selective Type I error at level α.*

In addition, we can adapt the results in Section 3 to compute the set $\left\{\phi \in \mathbb{R} : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\phi\nu/\|\nu\|_2^2\right)\left\{\mathrm{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu\right)\right\}^\top \Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x) \right\}\right\}$ by modifying the results in Lemmas 3 and 4. Details are in Section A.5 of the Appendix.

4.2 Unknown variance

When σ is unknown, we can plug in an estimate $\hat{\sigma}$ in (9):

$$\hat{p}_{\text{selective}}(\hat{\sigma}) = \mathrm{pr}\left[\phi(\hat{\sigma}) \geq \|x^\top \nu\|_2 \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi(\hat{\sigma}))\right) = c_i^{(t)}(x) \right\}\right], \tag{21}$$

where $\phi(\hat{\sigma}) \sim (\hat{\sigma}\|\nu\|_2)\chi_q$. If we use a consistent estimator of σ, then a test based on the p-value in (21) provides selective Type I error control (5) asymptotically.

Proposition 7 *For $q = 1, 2, \ldots$, suppose that $X^{(q)} \sim \mathcal{MN}_{n\times q}\big(\mu^{(q)}, \mathbf{I}_n, \sigma^2 \mathbf{I}_q\big)$. Let $x^{(q)}$ be a realization from $X^{(q)}$ and let $c_i^{(t)}(\cdot)$ be the cluster to which the ith observation is assigned during the tth iteration of step 3b. in Algorithm 1. Consider the sequence of null hypotheses $H_0^{(q)} : {\mu^{(q)}}^\top \nu^{(q)} = 0_q$, where $\nu^{(q)}$ defined in* (3) *is the contrast vector resulting from applying k-means clustering on $x^{(q)}$. Suppose that (i) $\hat{\sigma}$ is a consistent estimator of σ, i.e., for all $\epsilon > 0, \lim_{q\to\infty} \mathrm{pr}\big(|\hat{\sigma}(X^{(q)}) - \sigma| \geq \epsilon\big) = 0$; and (ii) there exists $\delta \in (0,1)$ such that $\lim_{q\to\infty} \mathrm{pr}_{H_0^{(q)}}\left[\bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] > \delta$. Then, for all $\alpha \in (0,1)$, we have that $\lim_{q\to\infty} \mathrm{pr}_{H_0^{(q)}}\left[\hat{p}_{selective}(\hat{\sigma}) \leq \alpha \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] = \alpha$.*

In practice, we propose to use the following estimator of σ (Huber, 1981):

$$\hat{\sigma}_{\text{MED}}(x) = \left\{ \underset{1\leq i\leq n, 1\leq j\leq q}{\text{median}} \left(\tilde{x}_{ij}^2\right) / M_{\chi_1^2} \right\}^{1/2}, \tag{22}$$

where $\tilde{x}$ is obtained from subtracting the median of each column in x, and $M_{\chi_1^2}$ is the median of the χ_1^2 distribution. If μ is sparse, i.e., $\sum_{i=1}^{n}\sum_{j=1}^{q} 1\{\mu_{ij} \neq 0\}$ is small, then (22) is consistent with appropriate assumptions; see Appendix A.7.

5. Simulation study

Throughout this section, we consider testing the null hypothesis $H_0 : \mu^\top \nu = 0_q$ versus $H_1 : \mu^\top \nu \neq 0_q$, where, unless otherwise stated, ν defined in (3) is based on a randomly-chosen pair of clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ from k-means clustering. We consider four p-values: p_{Naive} in (4),

$p_{\text{selective}}$ in (9), $\hat{p}_{\text{selective}}$ in (21) with $\hat{\sigma}_{\text{MED}}$ defined in (22), and $\hat{p}_{\text{selective}}$ in (21) with $\hat{\sigma}_{\text{Sample}} = \left\{\sum_{i=1}^{n}\sum_{j=1}^{q}(x_{ij}-\bar{x}_j)^2/(nq-q)\right\}^{1/2}$, where $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$. In the simulations that follow, we compare the selective Type I error (5) and power of the tests that reject H_0 when these p-values are less than $\alpha = 0.05$.

5.1 Selective Type I error under the global null

We generate data from (1) with $\mu = 0_{n\times q}$; therefore, H_0 in (2) holds for any pair of estimated clusters. We simulate 3,000 datasets with $n = 150, \sigma = 1$, and $q = 2, 10, 50, 100$.

For each simulated dataset, we apply k-means clustering with $K = 3$, and then compute p_{Naive}, $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ for a randomly-chosen pair of clusters. Figure 3 displays the observed p-value quantiles versus the Uniform(0,1) quantiles. We see that for all values of q, (i) the naive p-values in (4) are stochastically smaller than a Uniform(0,1) random variable, and the test based on p_{Naive} leads to an inflated Type I error rate; (ii) tests based $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ control the selective Type I error rate in the sense of (5).

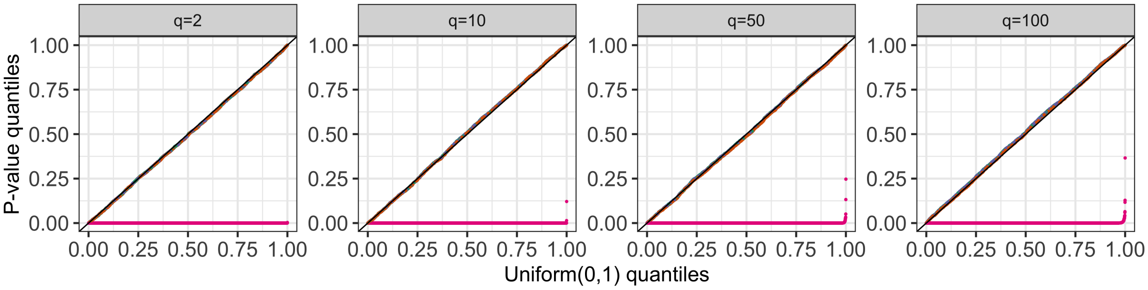


Figure 3: Quantile-quantile plots for p_{Naive} (pink), $p_{\text{selective}}$ (green), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple) under (1) with $\mu = 0_{n\times q}$, stratified by q.

5.2 Conditional power and detection probability

In this section, we show that the tests based on our proposal ($p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$) have substantial power to reject H_0 when it is not true. We generate data from (1) with $n = 150$ and

$$\mu_1 = \ldots = \mu_{\frac{n}{3}} = \begin{bmatrix} -\frac{\delta}{2} \\ 0_{q-1} \end{bmatrix}, \; \mu_{\frac{n}{3}+1} = \ldots = \mu_{\frac{2n}{3}} = \begin{bmatrix} 0_{q-1} \\ \frac{\sqrt{3}\delta}{2} \end{bmatrix}, \; \mu_{\frac{2n}{3}+1} = \ldots = \mu_n = \begin{bmatrix} \frac{\delta}{2} \\ 0_{q-1} \end{bmatrix}. \tag{23}$$

Here, we can think of $\mathcal{C}_1 = \{1,\ldots,n/3\}, \mathcal{C}_2 = \{(n/3)+1,\ldots,(2n/3)\}, \mathcal{C}_3 = \{(2n/3)+1,\ldots,n\}$ as the "true clusters". Moreover, these clusters are equidistant in the sense that the pairwise distance between each pair of population means is $|\delta|$. Recall that we test H_0 in (2) for a pair of estimated clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$, which may not be true clusters. Hence, we will separately consider the *conditional power* and *detection probability* of our proposed tests (Gao et al., 2022; Jewell et al., 2022; Hyun et al., 2021). The conditional power is the probability of rejecting H_0 in (2), given that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are true clusters. Given M simulated datasets with true clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$, we estimate it as

$$\text{Conditional power} = \frac{\sum_{m=1}^{M} 1\left\{\left\{\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}\right\} \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}, p^{(m)} \leq \alpha\right\}}{\sum_{m=1}^{M} 1\left\{\left\{\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}\right\} \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}\right\}}, \tag{24}$$

where , and $p^{(m)}$ and $\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}$ correspond to the p-value and clusters under consideration for the mth simulated dataset. Because the quantity in (24) conditions on the event that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are true clusters, we also estimate how often that event occurs:

$$\text{Detection probability} = \sum_{m=1}^{M} 1\left\{\{\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}\} \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}\right\}/M. \tag{25}$$

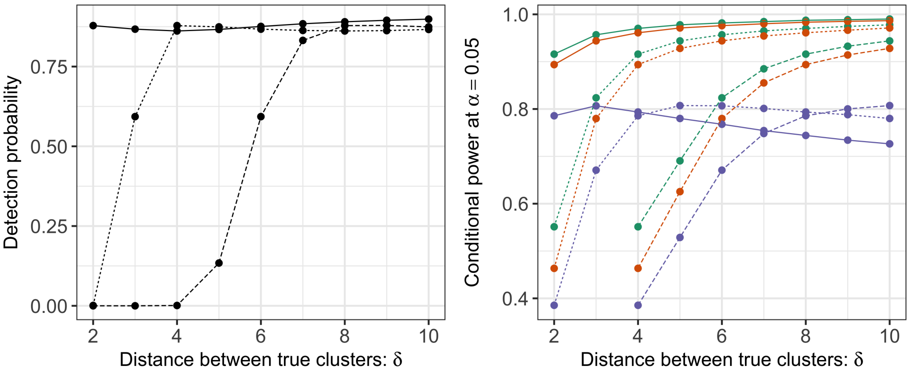


Figure 4: *Left:* The detection probability (25) for k-means clustering with $K = 3$ under model (1) with μ defined in (23), and $\sigma = 0.25$ (solid lines), 0.5 (dashed lines), and 1 (long-dashed lines). *Right:* The conditional power (24) at $\alpha = 0.05$ for the tests based on $p_{\text{selective}}$ (green), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple), under model (1) with μ defined in (23) and $\sigma = 0.25, 0.5, 1$. The conditional power is not displayed for $\delta = 2, 3, \sigma = 1$ because the true clusters were never recovered in simulation.

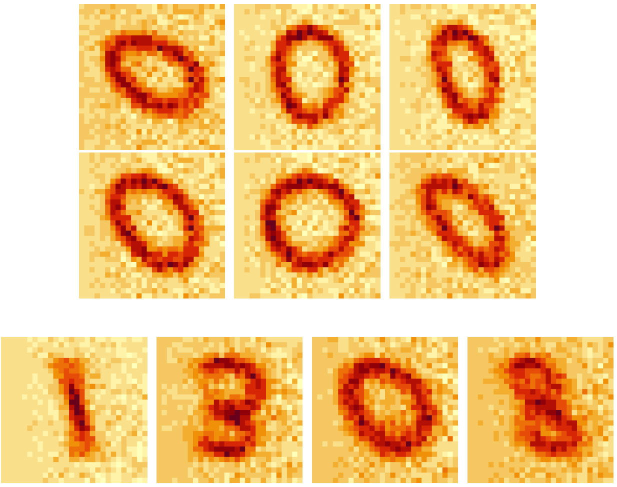
We generate $M = 200,000$ datasets from (23) with $q = 10, \sigma = 0.25, 0.5, 1$, and $\delta = 2, 3, \ldots, 10$. For each simulated dataset, we apply k-means clustering with $K = 3$ and reject $H_0 : \mu^\top \nu = 0_q$ if $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, or $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ is less than $\alpha = 0.05$. In Figure 4, the left panel displays the detection probability (25) of k-means clustering as a function of δ in (23), and the right panel displays the conditional power (24) for the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$. Under model (1), the detection probability and conditional power increase as a function of δ in (23) for all values of σ. For a given value of δ, a larger value of σ leads to lower detection probability and conditional power. The conditional power is not displayed for $\delta = 2, 3, \sigma = 1$ because the true clusters were never recovered in simulation. Moreover, for a given value of δ and σ, the test based on $p_{\text{selective}}$ has the highest conditional power, followed closely by the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. Using $\hat{\sigma}_{\text{Sample}}$ in $\hat{p}_{\text{selective}}$ leads to a less powerful test, especially for large values of δ. This is because $\hat{\sigma}_{\text{Sample}}$ is a conservative estimator of σ in (1), and its bias is an increasing function of δ, the distance between true clusters. By contrast, $\hat{\sigma}_{\text{MED}}$ is a consistent estimator under model (23) (see Appendix A.7).

As an alternative to the conditional power in (24), in Appendix A.8, we consider a notion of power that does not condition on having correctly estimated the true clusters.

6. Real data applications

6.1 MNIST Dataset (Lecun et al., 1998)

Here, we apply our method to the MNIST dataset (Lecun et al., 1998), which consists of 60,000 gray-scale images of handwritten digits. Each image has an accompanying label in $\{0, 1, \ldots, 9\}$, and is stored as a 28×28 matrix that takes on values in $[0, 255]$. We first divide the entries of all the images by 255. Next, since there is no variation in the peripheral pixels of the images (Gallaugher and McNicholas, 2018), which violates model (1), we add an independent perturbation $\mathcal{N}(0, 0.01)$ to each element of the image. Finally, we vectorize each image to obtain a vector $x_i \in \mathbb{R}^{784}$.



	H_0	p_{Naive}	$\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$
"No cluster"			
	$\bar{\mu}_1 = \bar{\mu}_2$	$< 10^{-10}$	$< 10^{-10}$
	$\bar{\mu}_1 = \bar{\mu}_3$	$< 10^{-10}$	0.01
	$\bar{\mu}_1 = \bar{\mu}_4$	$< 10^{-10}$	0.22
	$\bar{\mu}_1 = \bar{\mu}_5$	$< 10^{-10}$	$< 10^{-10}$
	$\bar{\mu}_1 = \bar{\mu}_6$	$< 10^{-10}$	0.53
	$\bar{\mu}_2 = \bar{\mu}_3$	$< 10^{-10}$	$< 10^{-10}$
	$\bar{\mu}_2 = \bar{\mu}_4$	$< 10^{-10}$	1.4×10^{-8}
	$\bar{\mu}_2 = \bar{\mu}_5$	$< 10^{-10}$	$< 10^{-10}$
	$\bar{\mu}_2 = \bar{\mu}_6$	$< 10^{-10}$	0.49
	$\bar{\mu}_3 = \bar{\mu}_4$	$< 10^{-10}$	0.14
	$\bar{\mu}_3 = \bar{\mu}_5$	$< 10^{-10}$	1.1×10^{-6}
	$\bar{\mu}_3 = \bar{\mu}_6$	$< 10^{-10}$	0.45
	$\bar{\mu}_4 = \bar{\mu}_5$	$< 10^{-10}$	2.0×10^{-8}
	$\bar{\mu}_4 = \bar{\mu}_6$	$< 10^{-10}$	0.77
	$\bar{\mu}_5 = \bar{\mu}_6$	$< 10^{-10}$	0.48
"Cluster"			
	$\bar{\mu}_1 = \bar{\mu}_2$	$< 10^{-10}$	$< 10^{-10}$
	$\bar{\mu}_1 = \bar{\mu}_3$	$< 10^{-10}$	8.0×10^{-6}
	$\bar{\mu}_1 = \bar{\mu}_4$	$< 10^{-10}$	6.2×10^{-7}
	$\bar{\mu}_2 = \bar{\mu}_3$	$< 10^{-10}$	6×10^{-3}
	$\bar{\mu}_2 = \bar{\mu}_4$	$< 10^{-10}$	10^{-3}
	$\bar{\mu}_3 = \bar{\mu}_4$	$< 10^{-10}$	4×10^{-4}

Figure 5: *Top left:* Centroids of six clusters from the "no cluster" dataset ($\hat{\mathcal{C}}_1$ to $\hat{\mathcal{C}}_6$ from left to right, top to bottom). *Bottom left:* Same as top left, but for the "cluster" dataset. *Right:* We test the null hypothesis of no difference between each pair of cluster centroids using p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. Here, $\bar{\mu}_i = \sum_{j \in \hat{\mathcal{C}}_i} \mu_j / |\hat{\mathcal{C}}_i|$.

We first construct a "no cluster" dataset by randomly sampling 1,500 images of the 0s; thus, $n = 1,500$ and $q = 784$. To de-correlate the pixels in each image, we whitened the data (see Section 4.1) using $\hat{\Sigma}^{-\frac{1}{2}} = U(\Lambda + 0.01\mathbf{I}_n)^{-\frac{1}{2}}U^\top$ as in prior work (Coates and Ng, 2012), where $U\Lambda U^\top$ is the eigenvalue decomposition of the sample covariance matrix.

We apply k-means clustering with $K = 6$. The centroids are displayed in the top left panel of Figure 5. For each pair of estimated clusters, we compute the p-values p_{Naive}

and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (see Figure 5). The naive p-values are extremely small for all pairs of clusters under consideration, despite the resemblance of the centroids. By contrast, our approach yields modest p-values, congruent with the visual resemblance of the centroids. In addition, for the most part, the pairs for which $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ is small are visually quite different (e.g., clusters 1 and 2, clusters 1 and 5, and clusters 4 and 5).

To demonstrate the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, we also generated a "cluster" dataset by sampling 500 images each from digits $0, 1, 3$, and 8; thus, $n = 2,000$ and $q = 784$. We again whitened the data to obtain uncorrelated features. After applying k-means clustering with $K = 4$, we obtain four clusters that roughly correspond to four digits: cluster 1, 94.0% digit 1; cluster 2, 72.4% digit 3; cluster 3, 83.6% digit 0; cluster 4, 62.4% digit 8 (see the bottom left panel of Figure 5). Results from testing for a difference in means for each pair of clusters using p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ are in Figure 5. Both sets of p-values are small on this "cluster" dataset.

6.2 Single-cell RNA-sequencing data (Zheng et al., 2017)

In this section, we apply our proposal to single-cell RNA-sequencing data collected by Zheng et al. (2017). Single-cell RNA-sequencing quantifies gene expression abundance at the resolution of single cells, thereby revealing cell-to-cell heterogeneity in transcription and allowing for the identification of cell types and marker genes. In practice, biologists often cluster the cells to identify putative cell types, and then perform a differential expression analysis, i.e., they test for a difference in gene expression between two clusters (Stuart et al., 2019; Lähnemann et al., 2020; Grün et al., 2015). Because this approach ignores the fact that the clusters were estimated from the same data used for testing, it does not control the selective Type I error.

Zheng et al. (2017) profiled 68,000 peripheral blood mononuclear cells, and classified them based on their match to the expression profiles of 11 reference transcriptomes from known cell types. We consider the classified cell types to be the "ground truth", and use this information to demonstrate that our proposal in Section 2 yields reasonable results.

As in prior work (Gao et al., 2022; Duò et al., 2018), we first excluded cells with low numbers of expressed genes or total counts, as well as cells in which a large percentage of the expressed genes are mitochondrial. We then divided the counts for each cell by the total sum of counts in that cell. Finally, we applied a $\log_2$ transformation with a pseudo-count of 1 to the expression data, and considered only the subset of 500 genes with the largest average expression levels pre-normalization. We applied the aforementioned pre-processing pipeline separately to memory T cells ($N = 10,224$) and a mixture of five types of cells (memory T cells, B cells, naive T cells, natural killer cells, and monocytes; $N = 43,259$).

To investigate the selective Type I error in the absence of true clusters, we first constructed a "no cluster" dataset by randomly sampling 1,000 out of 10,224 memory T cells after pre-processing (thus, $n = 1,000$ and $q = 500$). Since the gene expression levels are highly correlated, we first whitened the data as described in Section 4.1 by plugging in $\hat{\Sigma}^{-\frac{1}{2}} = U(\Lambda + 0.01\mathbf{I}_n)^{-\frac{1}{2}}U^\top$ (Coates and Ng, 2012), where $U\Lambda U^\top$ is the eigenvalue decomposition of the sample covariance matrix.

We applied k-means clustering to the transformed data with $K = 5$, and obtained five clusters consisting of 97, 223, 172, 165, and 343 cells, respectively (see Figure 8 left panel in Appendix A.9). For each pair of estimated clusters, we computed the p-values p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. The results are displayed in the top panel of Table 1. On this dataset, the naive p-values are extremely small for all pairs of estimated clusters, while our proposed p-values are quite large. In particular, at $\alpha = 0.05$, the test based on p_{Naive} concludes that all five estimated clusters correspond to distinct cell types (even after multiplicity correction). By contrast, our approach does not reject most of the null hypotheses; i.e., it finds no difference between expression levels of the estimated clusters. Because this "no cluster" dataset consists only of memory T cells, we believe that conclusion based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ aligns better with the underlying biology.

Table 1: P-values p_{Naive} in (4) and $\hat{p}_{\text{selective}}$ in (21) with $\hat{\sigma}_{\text{MED}}$ defined in (22) corresponding to the null hypothesis that the means of two estimated clusters are equal, for each pair of estimated clusters in the "no cluster" (top) and the "cluster" datasets (bottom).

H_0	$\bar{\mu}_1 = \bar{\mu}_2$	$\bar{\mu}_1 = \bar{\mu}_3$	$\bar{\mu}_1 = \bar{\mu}_4$	$\bar{\mu}_1 = \bar{\mu}_5$	$\bar{\mu}_2 = \bar{\mu}_3$	$\bar{\mu}_2 = \bar{\mu}_4$	$\bar{\mu}_2 = \bar{\mu}_5$	$\bar{\mu}_3 = \bar{\mu}_4$	$\bar{\mu}_3 = \bar{\mu}_5$	$\bar{\mu}_4 = \bar{\mu}_5$
p_{Naive}	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
$\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$	0.30	0.31	0.43	0.12	0.12	0.002	0.10	0.005	0.04	0.05
H_0	$\bar{\mu}_1 = \bar{\mu}_2$	$\bar{\mu}_1 = \bar{\mu}_3$	$\bar{\mu}_1 = \bar{\mu}_4$	$\bar{\mu}_1 = \bar{\mu}_5$	$\bar{\mu}_2 = \bar{\mu}_3$	$\bar{\mu}_2 = \bar{\mu}_4$	$\bar{\mu}_2 = \bar{\mu}_5$	$\bar{\mu}_3 = \bar{\mu}_4$	$\bar{\mu}_3 = \bar{\mu}_5$	$\bar{\mu}_4 = \bar{\mu}_5$
p_{Naive}	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
$\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$	4.0×10^{-4}	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	5.0×10^{-8}	$< 10^{-10}$

Next, we construct a "cluster" dataset by randomly sampling 400 each of memory T cells, B cells, naive T cells, natural killer cells, and monocytes from the 43,259 cells; thus, $n = 2,000$ and $q = 500$. After whitening the data, we applied k-means clustering to obtain five clusters. We see that these clusters approximately correspond to the five different cell types (cluster 1: 82.5% naive T cells; cluster 2: 95.3% memory T cells; cluster 3: 99.2% B cells; cluster 4: 91.5% nature killer cells; cluster 5: 83.3% monocytes); estimated clusters are visualized in the right panel of Figure 8 in Appendix A.9. We evaluate the p-values p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for all pairs of estimated clusters, and display results in the bottom panel of Table 1. Both sets of p-values are extremely small on this dataset, which suggests that the test based on our p-value has substantial power to reject the null hypothesis when it does not hold.

7. Discussion

We have proposed a test for a difference in means between two clusters estimated from k-means clustering, under (1). Methods developed in this paper are implemented in the `R` package `KmeansInference`, available at `https://github.com/yiqunchen/KmeansInference`. Data and code for reproducing the results in this paper can be found at `https://github.com/yiqunchen/KmeansInference-experiments`. Next, we outline a few directions for future research.

While the p-value in (9) leads to selective Type I error control, it conditions on more information than is used to construct the hypothesis in (2). In practice, data analysts likely only make use of the final cluster assignments (leading to the p-value in (8)), as opposed to all the intermediate assignments (leading to the p-value in (9)). Empirically, conditioning on too much information results in a loss of power (Fithian et al., 2014; Jewell et al., 2022; Liu et al., 2018). In future work, we will investigate the possibility of leveraging recent developments in selective inference (Chen et al., 2022; Le Duy and Takeuchi, 2021; Jewell et al., 2022) to compute the "ideal" p-value (8). Another line of future work is to extend our test for a pairwise difference in means to a difference among multiple groups (Kimes et al., 2017; Suzuki and Shimodaira, 2006). This might further provide a way to determine the number of clusters in k-means clustering.

We could also consider extending our proposal to other data generating models. The normality assumption in (1) is critical to the proof of Proposition 1, because it guarantees that under H_0 in (2), $\|X^\top \nu\|_2$, $\text{dir}(X^\top \nu)$, and $\Pi_\nu^\perp X$ are pairwise independent. However, this normality assumption is often violated in practice; for instance, in single-cell genomics, the data are count-valued and the variance of gene expression levels varies drastically with the mean expression levels of that gene (Stuart et al., 2019; Eling et al., 2018). This has motivated some authors to work with alternative models for gene expression including Poisson (Witten, 2011), negative binomial (Risso et al., 2018), and curved normal (Lin et al., 2021). To extend our framework to other exponential family distributions, we may be able to leverage recent proposals to decompose X into $f(X)$ and $g(X)$ such that both $f(X)$ and $g(X)|f(X)$ have a known, computationally-tractable distribution (Rasines and Alastair Young, 2021; Leiner et al., 2021).

Acknowledgments

This work was partially supported by National Institutes of Health grants (R01 EB026908, R01 GM123993, and R01 DA047869) and a Simons Investigator Award to Daniela Witten; and a Stanford Data Science Fellowship to Yiqun Chen.

References

Nadim Aizarani, Antonio Saviano, Sagar, Laurent Mailly, Sarah Durand, Josip S Herman, Patrick Pessaux, Thomas F Baumert, and Dominic Grün. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, 572(7768):199–204, August 2019.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, May 2009.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium On Discrete Algorithms*, SODA '07, pages 1027–1035, USA, January 2007. Society for Industrial and Applied Mathematics.

Marco Avella-Medina, Heather S Battey, Jianqing Fan, and Quefeng Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, March 2018.

Alan J Aw, Jeffrey P Spence, and Yun S Song. A flexible and robust non-parametric test of exchangeability. *arXiv:2109.15261*, September 2021.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, February 2017.

Anthony J Bell and Terrence J Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, December 1997.

Denis Belomestny, Mathias Trabs, and Alexandre B Tsybakov. Sparse covariance matrix estimation in high-dimensional deconvolution. *Bernoulli*, 25(3):1901 – 1938, 2019. doi: 10.3150/18-BEJ1040A. URL `https://doi.org/10.3150/18-BEJ1040A`.

Yuval Benjamini, Jonathan Taylor, and Rafael A Irizarry. Selection-corrected statistical inference for region detection with high-throughput assays. *Journal of the American Statistical Association*, 114(527):1351–1365, July 2019.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, December 2008.

Martin Bilodeau and David Brenner. *Theory of Multivariate Statistics*. Springer, New York, NY, 1999.

Richard Bourgon. intervals: Tools for working with points and intervals. `https://cran.rstudio.com/web/packages/intervals/index.html`, 2020. Accessed: 2022-2-11.

Katherine S Button. Double-dipping revisited. *Nature Neuroscience*, 22(5):688–690, May 2019.

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models do not reliably learn the number of components. *arXiv:2007.04470*, July 2020.

Tony T Cai, Jing Ma, and Linjun Zhang. CHIME: Clustering of high-dimensional gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, June 2019.

Ali Charkhi and Gerda Claeskens. Asymptotic post-selection inference for the Akaike information criterion. *Biometrika*, 105(3):645–664, June 2018.

Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. Testing for a finite mixture model with two components. *J. R. Stat. Soc. Series B Stat. Methodol.*, 66(1):95–115, February 2004.

Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542, October 2009.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics*, 46(5), October 2018.

Shuxiao Chen and Jacob Bien. Valid inference corrected for outlier removal. *J. Comput. Graph. Stat.*, 29(2):323–334, April 2020.

Yiqun Chen, Sean Jewell, and Daniela Witten. More powerful selective inference for the graph fused lasso. *Journal of Computational and Graphical Statistics*, pages 1–11, 2022.

Yiqun T Chen, Sean W Jewell, and Daniela M Witten. Quantifying uncertainty in spikes estimated from calcium imaging data. *Biostatistics*, October 2021.

Neo Christopher Chung. Statistical significance of cluster membership for unsupervised evaluation of cell identities. *Bioinformatics*, 36(10):3107–3114, May 2020.

Neo Christopher Chung and John D Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, February 2015.

Adam Coates and Andrew Y Ng. Learning feature representations with K-Means. In Grégoire Montavon, Geneviève B Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 561–580. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

L Comminges, O Collier, M Ndaoud, and A B Tsybakov. Adaptive robust estimation in sparse vector model. *Annals of Statistics*, 49(3), June 2021.

Edgar Dobriban. Permutation methods for factor analysis and PCA. *Annals of Statistics*, 48(5), October 2020.

Tyler Doughty and Eduard Kerkhoven. Extracting novel hypotheses and findings from RNA-seq data. *FEMS Yeast Res.*, 20(2), March 2020.

Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, July 2018.

V N L Duy, H Toda, R Sugiyama, and I Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. *arXiv:2002.09132*, 2020.

Nils Eling, Arianne C Richard, Sylvia Richardson, John C Marioni, and Catalina A Vallejos. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Systems*, 7(3):284–294.e12, September 2018.

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

Pascal Friederich, Mario Krenn, Isaac Tamblyn, and Alan Aspuru-Guzik. Scientific intuition inspired by machine learning generated hypotheses. *arXiv:2010.14236*, October 2020.

Michael P B Gallaugher and Paul D McNicholas. Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, 80:83–93, August 2018.

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

Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, September 2015.

Aritra Guha, Nhat Ho, and Xuanlong Nguyen. On posterior contraction of parameters and interpretability in bayesian mixture modeling. *arXiv:1901.05078*, January 2019.

Fang Han and Han Liu. Scale-invariant sparse PCA on high dimensional meta-elliptical data. *Journal of the American Statistical Association*, 109(505):275–287, January 2014.

J A Hartigan and M A Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 28(1):100–108, 1979.

Hastie, Trevor., Hastie, Trevor., Tibshirani, Robert., Friedman, and J H. *The Elements of Statistical Learning : data mining, inference, and prediction.* Springer, New York, 2001.

Roger A Horn and Charles R Johnson. *Matrix Analysis.* Cambridge University Press, 2nd edition edition, October 2012.

Peter J Huber. *Robust Statistics.* John Wiley & Sons, 1981.

Kenneth Hung and William Fithian. Statistical methods for replicability assessment. *The Annals of Applied Statistics*, 14(3):1063–1087, September 2020.

Sangwon Hyun, Max G'Sell, and Ryan J Tibshirani. Exact post-selection inference for the generalized lasso path. *Electron. J. Stat.*, 12(1):1053–1097, 2018.

Sangwon Hyun, Kevin Z Lin, Max G'Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, January 2021.

Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after changepoint detection. *To appear in J. R. Stat. Soc. Series B Stat. Methodol.*, 2022.

Jiashun Jin and Wanjie Wang. Influential features PCA for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359, December 2016.

M K Kerr and G A Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98(16):8961–8965, July 2001.

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

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, May 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, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M Keizer, Indu Khatri, Szymon M Kielbasa, Jan O Korbel, Alexey M Kozlov, Tzu-Hao Kuo, Boudewijn P F Lelieveldt, Ion I Mandoiu, John C Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J Theis, Huan Yang, Alex Zelikovsky, Alice C McHardy, Benjamin J Raphael, Sohrab P Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, February 2020.

Vo Nguyen Le Duy and Ichiro Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *arXiv:2105.04920*, May 2021.

Y Lecun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

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

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data blurring: sample splitting a single sample. *arXiv:2112.11079*, December 2021.

Pengfei Li and Jiahua Chen. Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491):1084–1092, September 2010.

Kevin Z Lin, Jing Lei, and Kathryn Roeder. Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-Seq data. *Journal of the American Statistical Association*, 116(534):457–470, April 2021.

Keli Liu, Jelena Markovic, and Robert Tibshirani. More powerful post-selection inference, with application to the lasso. *arXiv:1801.09037*, January 2018.

S Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, September 1982.

Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, October 2021.

Joshua R Loftus and Jonathan E Taylor. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015.

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv:1612.02099*, December 2016.

MacQueen, J, and author. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, January 1967.

Jelena Markovic, Lucy Xia, and Jonathan Taylor. Unifying approach to selective inference with applications to cross-validation. *arXiv:1703.06559*, 2017.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, February 2018.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, March 2019.

Agostino Nobile. On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, 32(5):2044–2073, October 2004.

Robert Pollice, Gabriel Dos Passos Gomes, Matteo Aldeghi, Riley J Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, Akshatkumar Nigam, Cher Tian Ser, Zhenpeng Yao, and Alán Aspuru-Guzik. Data-Driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860, February 2021.

Daniel G Rasines and G Alastair Young. Splitting strategies for post-selection inference. *arXiv:2102.02159*, February 2021.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17, January 2018.

Peter J Rousseeuw. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

David Rügamer, Philipp F M Baumann, and Sonja Greven. Selective inference for additive and linear mixed models. *Computational Statistics & Data Analysis*, 167:107350, March 2022.

Christoph Schultheiss, Claude Renaux, and Peter Bühlmann. Multicarving for high-dimensional post-selection inference. *Electron. J. Stat.*, 15(1):1695–1742, January 2021.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, June 2006.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for ℓ_1-penalized likelihood models. *The Canadian Journal of Statistics*, 46(1):41–61, March 2018.

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, April 2016.

Ryan J Tibshirani, Alessandro Rinaldo, Rob Tibshirani, and Larry Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287, June 2018.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional EM algorithm: Statistical optimization and asymptotic normality. *Advances in Neural Information Processing Systems*, 28:2512–2520, 2015.

Chihiro Watanabe and Taiji Suzuki. Selective inference for latent block models. *Electron. J. Stat.*, 15(1), January 2021.

Daniela M Witten. Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*, 5(4):2493–2518, December 2011.

S.N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2017.

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

Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2477–2485, Red Hook, NY, USA, December 2016. Curran Associates Inc.

Xinyang Yi and Constantine Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, 28, 2015.

Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, and Horst Simon. Spectral relaxation for k-means clustering. In T Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Series B Stat. Methodol.*, 76(1):217–242, 2014.

Jesse M Zhang, Govinda M Kamath, and David N Tse. Valid post-clustering differential analysis for Single-Cell RNA-Seq. *Cell Systems*, 9(4):383–392.e6, October 2019.

Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, January 2017.

Appendix A. Appendix

A.1 Proof of Proposition 1

The proof of Proposition 1 is similar to the proof of Theorem 1 in Gao et al. (2022), the proof of Theorem 3.1 in Loftus and Taylor (2015), the proof of Lemma 1 in Yang et al. (2016), and the proof of Theorem 3.1 in Chen and Bien (2020).

For any non-zero $\nu \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times q}$, we have that

$$X = \Pi_\nu^\perp X + (\mathbf{I}_n - \Pi_\nu^\perp)X = \Pi_\nu^\perp X + \frac{\nu\nu^\top X}{\|\nu\|_2^2} = \Pi_\nu^\perp X + \left(\frac{\|X^\top \nu\|_2}{\|\nu\|_2^2}\right)\nu\Big\{\mathrm{dir}\Big(X^\top\nu\Big)\Big\}^\top. \tag{A.26}$$

Lemma 8 *Under* (1) *and* $H_0 : \mu^\top\nu = 0_q$*, we have that* $\|X^\top\nu\|_2$, $\Pi_\nu^\perp X$, *and* $\mathrm{dir}(X^\top\nu)$ *are pairwise independent.*

Proof We first prove that $X^\top\nu$ is independent of $\Pi_\nu^\perp X$. The definition of $\Pi_\nu^\perp$ implies that $\Pi_\nu^\perp \nu = 0_n$, and it follows from the properties of the matrix normal distribution that $\Pi_\nu^\perp X$ and $X^\top\nu$ are independent. Therefore, $\|X^\top\nu\|_2$ and $\mathrm{dir}(X^\top\nu)$ are independent of $\Pi_\nu^\perp X$ as well, since both are functions of $X^\top\nu$.

Next, we will show that $\|X^\top\nu\|_2$ and $\mathrm{dir}(X^\top\nu)$ are independent. Under (1) and $H_0 : \mu^\top\nu = 0_q$, we have that $X^\top\nu \sim \mathcal{N}(0_q, \sigma^2\|\nu\|_2^2\mathbf{I}_q)$. It follows that $X^\top\nu$ is rotationally invariant, and therefore $\|X^\top\nu\|_2$ is independent of $\mathrm{dir}(X^\top\nu)$ (see, e.g., Proposition 4.1 and Corollary 4.3 of Bilodeau and Brenner (1999)). ■

We now proceed to prove the statement in (10). Recalling the definition of $p_{\text{selective}}$ in (9), under $H_0 : \mu^\top\nu = 0_q$ with ν defined in (3), we have that

$$\begin{aligned}
p_{\text{selective}} &= \mathrm{pr}_{H_0}\left[\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2 \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top\nu) = \mathrm{dir}(x^\top\nu)\right] \\
&\overset{a.}{=} \mathrm{pr}_{H_0}\Bigg[\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2 \;\Bigg|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp X + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\mathrm{dir}\left(X^\top\nu\right)\right\}^\top\right) = c_i^{(t)}(x)\right\}, \\
&\qquad \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top\nu) = \mathrm{dir}(x^\top\nu)\Bigg] \\
&\overset{b.}{=} \mathrm{pr}_{H_0}\Bigg[\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2 \;\Bigg|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\mathrm{dir}\left(x^\top\nu\right)\right\}^\top\right) = c_i^{(t)}(x)\right\}, \\
&\qquad \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top\nu) = \mathrm{dir}(x^\top\nu)\Bigg] \\
&\overset{c.}{=} \mathrm{pr}_{H_0}\left[\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2 \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\mathrm{dir}\left(x^\top\nu\right)\right\}^\top\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{d.}{=} \mathrm{pr}_{H_0}\left[\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2 \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'(\|X^\top\nu\|_2)\right) = c_i^{(t)}(x)\right\}\right].
\end{aligned} \tag{A.27}$$

Here, step a. follows from substituting X with the expression in (A.26), and step b. follows from replacing $\Pi_\nu^\perp X$ and $\mathrm{dir}(X^\top \nu)$ with $\Pi_\nu^\perp x$ and $\mathrm{dir}(x^\top \nu)$, respectively. Next, in step c., we used Lemma 8. Finally, step d. follows from the definition of $x'(\phi)$ in (11).

Note that under (1) and $H_0 : \mu^\top \nu = 0_q$, we have that $\|X^\top \nu\|_2 \sim \sigma\|\nu\|_2 \chi_q$, which concludes the proof of (10).

It remains to show that the test that rejects $H_0 : \mu^\top \nu = 0$ when $p_{\text{selective}} \leq \alpha$ controls the selective Type I error at level α, in the sense of (5). First of all, recall that we decided to test the null hypothesis in (2) based on the output of Algorithm 1. Therefore, $p_{\text{selective}}$ controls the selective Type I error at level α if, for any $c_i^{(T)}(x)$, $i = 1, \ldots, n$,

$$\mathrm{pr}_{H_0}\left[\text{reject } H_0 \text{ at level } \alpha \,\middle|\, \bigcap_{i=1}^{n} \left\{ c_i^{(T)}(X) = c_i^{(T)}(x) \right\}\right] \leq \alpha, \ \forall \alpha \in (0,1). \tag{A.28}$$

To prove (A.28), we first note that the following holds for any $\alpha \in (0,1)$:

$$\begin{aligned}
&\mathrm{pr}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \leq \alpha \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \, \mathrm{dir}(X^\top\nu) = \mathrm{dir}(x^\top\nu)\right] \\
&\overset{a.}{=} \mathrm{pr}_{H_0}\Bigg[p_{\text{selective}}(\|X^\top\nu\|_2) \leq \alpha \,\Bigg|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp X + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\{\mathrm{dir}(X^\top\nu)\}^\top\right) = c_i^{(t)}(x)\right\}, \\
&\qquad \Pi_\nu^\perp X = \Pi_\nu^\perp x, \, \mathrm{dir}(X^\top\nu) = \mathrm{dir}(x^\top\nu)\Bigg] \\
&\overset{b.}{=} \mathrm{pr}_{H_0}\Bigg[p_{\text{selective}}(\|X^\top\nu\|_2) \leq \alpha \,\Bigg|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\{\mathrm{dir}(x^\top\nu)\}^\top\right) = c_i^{(t)}(x)\right\}, \\
&\qquad \Pi_\nu^\perp X = \Pi_\nu^\perp x, \, \mathrm{dir}(X^\top\nu) = \mathrm{dir}(x^\top\nu)\Bigg] \\
&\overset{c.}{=} \mathrm{pr}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \leq \alpha \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\{\mathrm{dir}(x^\top\nu)\}^\top\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{d.}{=} \mathrm{pr}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \leq \alpha \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'(\|X^\top\nu\|_2)\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{e.}{=} \mathrm{pr}_{H_0}\left[1 - F_q^{\mathcal{S}_T}(\|X^\top\nu\|_2) \leq \alpha \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'(\|X^\top\nu\|_2)\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{f.}{=} \alpha.
\end{aligned} \tag{A.29}$$

Here, steps a. through d. follow from the same line of argument in (A.27). Moreover, (10) implies that, for a given sequence of cluster assignments $c_i^{(T)}(x)$, $i = 1, \ldots, n$, $p_{\text{selective}}$ is the survival function of a χ_q random variable, truncated to the set $\mathcal{S}_T$ defined in (12). Letting $F_q^{\mathcal{S}_T}(\cdot)$ denote the cumulative distribution function of this truncated χ_q random variable, we arrive at step e. Finally, to prove f., we first note that under H_0, the conditional cumulative distribution function of $\|X^\top\nu\|_2$ given $\bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'(\|X^\top\nu\|_2)\right) = c_i^{(t)}(x)\right\}$ is exactly

$F_q^{\mathcal{S}_T}$. The equality, therefore, follows from the probability integral transform, which states that for a continuous random variable Z, $F_Z(Z)$ follows the Uniform(0,1) distribution.

Finally, we have that

$$\begin{aligned}
&\mathrm{pr}_{H_0}\left[p_{\text{selective}}(\|X^\top \nu\|_2) \le \alpha \;\middle|\; \bigcap_{i=1}^{n}\left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}\right] \\
&= \mathrm{E}_{H_0}\left[1\left\{p_{\text{selective}}(\|X^\top \nu\|_2) \le \alpha\right\} \;\middle|\; \bigcap_{i=1}^{n}\left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}\right] \\
&\overset{a.}{=} \mathrm{E}_{H_0}\Bigg(\mathrm{E}_{H_0}\left[1\left\{p_{\text{selective}}(\|X^\top \nu\|_2) \le \alpha\right\} \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\},\right. \\
&\qquad \left.\Pi_\nu^\perp X = \Pi_\nu^\perp x,\, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right] \;\Bigg|\; \bigcap_{i=1}^{n}\left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}\Bigg) \\
&\overset{b.}{=} \mathrm{E}_{H_0}\left[\alpha \;\middle|\; \bigcap_{i=1}^{n}\left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}\right] \\
&= \alpha.
\end{aligned}$$

In the proof above, $a.$ follows from the tower property of conditional expectation, and $b.$ is a direct consequence of (A.29).

Therefore, we conclude that the test based on $p_{\text{selective}}$ controls the selective Type I error in (5), which completes the proof of Proposition 1.

A.2 Proof of Proposition 2

We will derive the expression for $\mathcal{S}_T$ in Proposition 2 using an induction argument. For a positive integer K, we let $[K]$ denote the set $\{1, \ldots, K\}$.

The following two claims (Lemmas 9 and 10) serve as the "base cases" for the proof.

Lemma 9 *Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned during the tth iteration of step 3b. of Algorithm 1 applied to data x, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from x during step 1 of Algorithm 1. For $\mathcal{S}_0$ defined as*

$$\mathcal{S}_0 = \left\{\phi \in \mathbb{R} : \bigcap_{i=1}^{n}\left\{c_i^{(0)}(x'(\phi)) = c_i^{(0)}(x)\right\}\right\}, \tag{A.30}$$

we have that

$$\mathcal{S}_0 = \bigcap_{i=1}^{n}\bigcap_{k=1}^{K}\left\{\phi : \left\|\left[x'(\phi)\right]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi))\right\|_2^2 \le \left\|\left[x'(\phi)\right]_i - m_k^{(0)}(x'(\phi))\right\|_2^2\right\}. \tag{A.31}$$

Proof

We first prove that the set in (A.30) is a subset of the set in (A.31). For an arbitrary $\phi_0 \in$ (A.30) and $1 \le i \le n$, we have that

$$\begin{aligned}
&c_i^{(0)}(x'(\phi_0)) = \underset{1\le k\le K}{\operatorname{argmin}} \left\| \left[x'(\phi_0)\right]_i - m_k^{(0)}(x'(\phi_0))\right\|_2^2 \\
&\overset{a.}{\Longrightarrow} \left\| \left[x'(\phi_0)\right]_i - m_{c_i^{(0)}(x'(\phi_0))}^{(0)}(x'(\phi_0))\right\|_2^2 \le \left\| \left[x'(\phi_0)\right]_i - m_k^{(0)}(x'(\phi_0))\right\|_2^2, \forall k \in [K] \\
&\overset{b.}{\Longrightarrow} \left\| \left[x'(\phi_0)\right]_i - m_{c_i^{(0)}(x)}^{(0)}(x'(\phi_0))\right\|_2^2 \le \left\| \left[x'(\phi_0)\right]_i - m_k^{(0)}(x'(\phi_0))\right\|_2^2, \forall k \in [K].
\end{aligned}$$

Here, the first line follows from the definition of $c_i^{(0)}$ in step 2 of Algorithm 1, and step $a.$ follows from the definition of the argmin function. Step $b.$ follows from the assumption that $\phi_0 \in$ (A.30) satisfies $c_i^{(0)}\left(x'(\phi_0)\right) = c_i^{(0)}\left(x\right)$. Because this holds for an arbitrary $1 \le i \le n$, we have proven that $\phi_0 \in$ (A.30) $\implies$ $\phi_0 \in$ (A.31); or equivalently, (A.30) $\subseteq$ (A.31).

We proceed to prove the other direction. For an arbitrary $\phi_0 \in$ (A.31) and an arbitrary $1 \le i \le n$, we have that

$$\begin{aligned}
&\left\| \left[x'(\phi_0)\right]_i - m_{c_i^{(0)}(x)}^{(0)}(x'(\phi_0))\right\|_2^2 \le \left\| \left[x'(\phi_0)\right]_i - m_k^{(0)}(x'(\phi_0))\right\|_2^2, \forall k \in [K] \\
&\overset{a.}{\Longrightarrow} c_i^{(0)}(x) = \underset{1\le k\le K}{\operatorname{argmin}} \left\| \left[x'(\phi_0)\right]_i - m_k^{(0)}(x'(\phi_0))\right\|_2^2 \\
&\overset{b.}{\Longrightarrow} c_i^{(0)}(x) = c_i^{(0)}(x'(\phi_0)).
\end{aligned}$$

Here, step $a.$ follows from the definition of argmin, and step $b.$ follows from combining the definition of $c_i^{(0)}(x'(\phi))$ in step 2 of Algorithm 1. We conclude that $\phi_0 \in$ (A.31) $\implies$ $\phi_0 \in$ (A.30).

Combining these two directions, we have proven that (A.31) = (A.30). ■

Lemma 10 *Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned in the tth iteration of step 3b. of Algorithm 1 applied to data x, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from x during step 1 of Algorithm 1. For $\mathcal{S}_1$ defined as*

$$\mathcal{S}_1 = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{1} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}, \tag{A.32}$$

and $w_i^{(t)}(k)$ defined in (14)*, we have that*

$$\begin{aligned}
\mathcal{S}_1 = &\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - m_{c_i^{(0)}(x)}^{(0)}(x'(\phi)) \right\|_2^2 \le \left\| [x'(\phi)]_i - m_k^{(0)}(x'(\phi)) \right\|_2^2 \right\} \cap \\
&\left(\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(0)}\left(c_i^{(1)}(x)\right)[x'(\phi)]_{i'} \right\|_2^2 \le \left\| [x'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(0)}(k)[x'(\phi)]_{i'} \right\|_2^2 \right\} \right).
\end{aligned} \tag{A.33}$$

Proof

We first prove that (A.32) $\subseteq$ (A.33). For an arbitrary $\phi_0 \in$ (A.32) and an arbitrary $1 \leq i \leq n$, we have that

$$
\begin{aligned}
& c_i^{(1)}(x'(\phi_0)) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - m_k^{(1)}(x'(\phi_0)) \right\|_2^2 \\
& \overset{a.}{\implies} c_i^{(1)}(x) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - m_k^{(1)}(x'(\phi_0)) \right\|_2^2 \\
& \overset{b.}{\implies} c_i^{(1)}(x) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}} \right\|_2^2 \\
& \overset{c.}{\implies} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = c_i^{(1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = c_i^{(1)}(x)\right\}} \right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}} \right\|_2^2, \forall k \in [K] \\
& \overset{d.}{\implies} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = c_i^{(1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = c_i^{(1)}(x)\right\}} \right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = k\right\}} \right\|_2^2, \forall k \in [K] \\
& \overset{e.}{\implies} \left\| \left[x'(\phi_0)\right]_i - \sum_{i'=1}^n w_{i'}^{(0)}\left(c_i^{(1)}(x)\right)\left[x'(\phi_0)\right]_{i'} \right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \sum_{i'=1}^n w_{i'}^{(0)}(k)\left[x'(\phi_0)\right]_{i'} \right\|_2^2, \forall k \in [K].
\end{aligned}
$$

In the equations above, the first line follows from step 3b. of Algorithm 1 with $t = 0$. Next, step $a.$ follows from the definition of (A.32), which implies that $c_i^{(1)}(x'(\phi_0)) = c_i^{(1)}(x)$. Step $b.$ is a direct consequence of step 3a. of Algorithm 1 with $t = 0$. In steps $c.$ and $d.$, we used the definitions of the argmin function and (A.32). Finally, we apply the definition of $w_i^{(t)}$ in (14) to get e. Because this holds for an arbitrary $1 \leq i \leq n$, $\phi_0 \in$ (A.32) implies that ϕ_0 is an element of the second set in the intersection in (A.33).

Moreover, $\phi_0 \in$ (A.32) implies that $\phi_0 \in$ (A.30), which, according to Lemma 9, further implies that ϕ_0 is an element of the first set in the intersection in (A.33). To summarize, we have proven that $\phi_0 \in$ (A.32) $\implies$ $\phi_0 \in$ (A.33), and as a result, (A.32) $\subseteq$ (A.33).

Next, we prove that the set in (A.33) is a subset of the set in (A.32). For an arbitrary $\phi_0 \in$ (A.33) and an arbitrary $1 \leq i \leq n$, we have that

$$
\begin{aligned}
& \phi_0 \in \text{(A.33)} \overset{a.}{\implies} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = c_i^{(1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = c_i^{(1)}(x)\right\}} \right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = k\right\}} \right\|_2^2, \forall k \in [K] \\
& \overset{b.}{\implies} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = c_i^{(1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = c_i^{(1)}(x)\right\}} \right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}} \right\|_2^2, \forall k \in [K] \\
& \overset{c.}{\implies} c_i^{(1)}(x) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}} \right\|_2^2 \\
& \overset{d.}{\implies} c_i^{(1)}(x) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - m_k^{(1)}(x'(\phi_0)) \right\|_2^2 \\
& \overset{e.}{\implies} c_i^{(1)}(x) = c_i^{(1)}(x'(\phi_0)).
\end{aligned}
$$

Here, step $a.$ follows from the definition of (A.33). In step $b.$, we first apply Lemma 9, which implies that (A.33) $\subseteq$ (A.31). Therefore, $\phi_0 \in$ (A.33) $\implies$ $c_i^{(0)}(x) = c_i^{(0)}(x'(\phi_0))$, for all $i = 1, \ldots, n, k = 1, \ldots, K$, yielding the desired equality. Next, step $c.$ follows from the definition of the argmin function. Finally, steps $d.$ and $e.$ follow directly from the definitions of $m_k^{(t)}$ and $c_i^{(t)}$ in steps 3a. and 3b. of Algorithm 1, respectively.

Because the result above holds for an arbitrary i, we have that $\phi_0 \in$ (A.33) $\implies$ $c_i^{(1)}(x) = c_i^{(1)}(x'(\phi))$, $i = 1, \ldots, n$. Combining this result with the observation that (A.33) $\subseteq$ (A.31), we have that (A.33) $\subseteq$ (A.32), which concludes the proof. ■

Next, we will prove the inductive step in the proof of Proposition 2, which relies on the following claim.

Lemma 11 *Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned in the tth iteration of Algorithm 1 applied to the data x, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from x during initialization. For some $1 \leq \tilde{T} \leq T-1$, define*

$$\mathcal{S}_{\tilde{T}} = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{\tilde{T}} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}. \tag{A.34}$$

Suppose that the following holds for $\tilde{T}$:

$$\begin{aligned} \mathcal{S}_{\tilde{T}} = & \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[x'(\phi)\right]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi)) \right\|_2^2 \leq \left\| \left[x'(\phi)\right]_i - m_k^{(0)}(x'(\phi)) \right\|_2^2 \right\} \\ \cap & \left(\bigcap_{t=1}^{\tilde{T}} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}\left(c_i^{(t)}(x)\right) \left[x'(\phi)\right]_{i'} \right\|_2^2 \leq \left\| \left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k) \left[x'(\phi)\right]_{i'} \right\|_2^2 \right\} \right), \end{aligned} \tag{A.35}$$

where $w_i^{(t)}(\cdot)$ is defined in (14). Then, for $\mathcal{S}_{\tilde{T}+1}$ defined as

$$\mathcal{S}_{\tilde{T}+1} = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{\tilde{T}+1} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}, \tag{A.36}$$

we have that

$$\begin{aligned} \mathcal{S}_{\tilde{T}+1} = & \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[x'(\phi)\right]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi)) \right\|_2^2 \leq \left\| \left[x'(\phi)\right]_i - m_k^{(0)}(x'(\phi)) \right\|_2^2 \right\} \\ \cap & \left(\bigcap_{t=1}^{\tilde{T}+1} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}\left(c_i^{(t)}(x)\right) \left[x'(\phi)\right]_{i'} \right\|_2^2 \leq \left\| \left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k) \left[x'(\phi)\right]_{i'} \right\|_2^2 \right\} \right). \end{aligned} \tag{A.37}$$

Proof Using the definitions in (A.34) and (A.36), we have that

$$\mathcal{S}_{\tilde{T}+1} = \mathcal{S}_{\tilde{T}} \cap \left(\bigcap_{i=1}^{n} \left\{ \phi \in \mathbb{R} : c_i^{(\tilde{T}+1)}\left(x'(\phi)\right) = c_i^{(\tilde{T}+1)}\left(x\right) \right\} \right). \tag{A.38}$$

Therefore, it suffices to prove that (A.38) = (A.37), under the inductive hypothesis (A.35).

We start by proving that (A.38) $\subseteq$ (A.37). For an arbitrary $\phi_0 \in$ (A.38) and an arbitrary $1 \leq i \leq n$, we have that

$$
\begin{aligned}
&c_i^{(\tilde{T}+1)}\left(x'(\phi_0)\right) = c_i^{(\tilde{T}+1)}\left(x\right) \overset{a.}{\Longrightarrow} c_i^{(\tilde{T}+1)}(x) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - m_k^{(\tilde{T}+1)}(x'(\phi_0))\right\|_2^2 \\
&\overset{b.}{\Longrightarrow} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = c_i^{(\tilde{T}+1)}(x)\right\}}\right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = k\right\}}\right\|_2^2, \forall k \in [K] \\
&\overset{c.}{\Longrightarrow} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}}\right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}}\right\|_2^2, \forall k \in [K] \\
&\overset{d.}{\Longrightarrow} \phi_0 \in \text{(A.37)}.
\end{aligned}
$$

Here, the first statement follows from the definition of $\mathcal{S}_{\tilde{T}+1}$. Next, steps $a.$ and $b.$ follow from the definitions of $c_i^{(\tilde{T}+1)}$ and $m_k^{(\tilde{T}+1)}(x'(\phi_0))$ in steps 3b. and 3a. of Algorithm 1, respectively. In step $c.$, we used the fact that $\phi_0 \in$ (A.38) $\Longrightarrow \phi_0 \in \mathcal{S}_{\tilde{T}} \Longrightarrow c_i^{\tilde{T}}(x'(\phi_0)) = c_i^{\tilde{T}}(x)$. Finally, $d.$ follows from the definition of $w_i^{(t)}$ in (14).

We continue with the reverse direction. Applying the inductive hypothesis (A.35), together with the definition of $S_{\tilde{T}+1}$ in (A.37) and the definition of $w_i^{(t)}$ in (14), we have that

$$
\begin{aligned}
\text{(A.37)} = S_{\tilde{T}} \cap \Bigg(\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \Bigg\{ \phi : &\left\| \left[x'(\phi)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}}\right\|_2^2 \leq \\
&\left\| \left[x'(\phi)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}[x'(\phi)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}}\right\|_2^2 \Bigg\}\Bigg).
\end{aligned} \tag{A.39}
$$

For an arbitrary $\phi_0 \in$ (A.37) and any $1 \leq i \leq n$, the following holds:

$$
\begin{aligned}
&\left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}}\right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}}\right\|_2^2, \forall k \in [K] \\
&\overset{a.}{\Longrightarrow} \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = c_i^{(\tilde{T}+1)}(x)\right\}}\right\|_2^2 \leq \left\| \left[x'(\phi_0)\right]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = k\right\}}\right\|_2^2, \forall k \in [K] \\
&\overset{b.}{\Longrightarrow} c_i^{(\tilde{T}+1)}(x) = \operatorname*{argmin}_{1\leq k\leq K} \left\| \left[x'(\phi_0)\right]_i - m_k^{(\tilde{T}+1)}(x'(\phi_0))\right\|_2^2 \\
&\overset{c.}{\Longrightarrow} c_i^{(\tilde{T}+1)}(x) = c_i^{(\tilde{T}+1)}(x'(\phi_0)).
\end{aligned}
$$

Here, to derive step $a.$, we first note that by (A.39), any element ϕ_0 of (A.37) is also an element of $\mathcal{S}_{\tilde{T}}$. Therefore, using the definition of $\mathcal{S}_{\tilde{T}}$ in (A.34), we have that $\bigcap_{t=1}^{\tilde{T}} \left\{ c_i^{(t)}\left(x'(\phi_0)\right) = c_i^{(t)}\left(x\right)\right\}$, and step $a.$ follows directly. Next, steps $b.$ and $c.$ follow directly from steps 3a. and 3b. of Algorithm 1 with $t = \tilde{T}$. By inspecting the form of (A.38), we conclude that $\phi_0 \in$ (A.37) $\Longrightarrow \phi_0 \in$ (A.38).

In conclusion, we have proven that (A.37) = (A.38), which completes the proof. ■

The inductive proof of Proposition 2 follows from combining Lemmas 9, 10 and 11.

A.3 Proof of Lemmas 3 and 4

We first prove Lemma 3, which is also Lemma 2 in Gao et al. (2022).

Proof We first express the inner product $\langle [x'(\phi)]_i, [x'(\phi)]_j \rangle$ as a function of ϕ. From (11), we have that $[x'(\phi)]_i = x_i + \nu_i \left(\frac{\phi - \|x^\top \nu\|_2}{\|\nu\|_2^2} \right) \text{dir}(x^\top \nu) = x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) + \left(\frac{\nu_i}{||\nu||_2^2} \phi \right) \text{dir}(x^\top \nu)$.

Therefore,

$$
\begin{aligned}
\left\langle [x'(\phi)]_i, [x'(\phi)]_j \right\rangle &= \left\langle x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) + \left(\frac{\nu_i}{||\nu||_2^2} \phi \right) \text{dir}(x^\top \nu), x_j - \nu_j \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) + \left(\frac{\nu_j}{||\nu||_2^2} \phi \right) \text{dir}(x^\top \nu) \right\rangle \\
&= \left(\frac{(\nu_i \nu_j)^{1/2}}{||\nu||_2^2} \phi \right)^2 + \left\langle x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu), \left(\frac{\nu_j}{||\nu||_2^2} \right) \text{dir}(x^\top \nu) \right\rangle \cdot \phi \\
&+ \left\langle x_j - \nu_j \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu), \left(\frac{\nu_i}{||\nu||_2^2} \right) \text{dir}(x^\top \nu) \right\rangle \cdot \phi \\
&+ \left\langle x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu), x_j - \nu_j \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) \right\rangle \\
&= \left(\frac{(\nu_i \nu_j)^{1/2}}{||\nu||_2^2} \right)^2 \phi^2 + \left(\frac{\nu_j}{||\nu||_2^2} \left\langle x_i, \text{dir}(x^\top \nu) \right\rangle + \frac{\nu_i}{||\nu||_2^2} \left\langle x_j, \text{dir}(x^\top \nu) \right\rangle - 2 \frac{\nu_i \nu_j ||x^\top \nu||_2}{||\nu||_2^4} \right) \phi \\
&+ \left\langle x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu), x_j - \nu_j \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) \right\rangle .
\end{aligned}
$$

Next, using the expression for $\langle [x'(\phi)]_i, [x'(\phi)]_j \rangle$ above, we have that

$$
\begin{aligned}
\left\| [x'(\phi)]_i - [x'(\phi)]_j \right\|_2^2 &= \left\langle [x'(\phi)]_i - [x'(\phi)]_j, [x'(\phi)]_i - [x'(\phi)]_j \right\rangle \\
&= \left\langle x_i - x_j - (\nu_i - \nu_j) \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) + \left(\frac{(\nu_i - \nu_j)}{||\nu||_2^2} \phi \right) \text{dir}(x^\top \nu), \right. \\
& \left. x_i - x_j - (\nu_i - \nu_j) \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) + \left(\frac{(\nu_i - \nu_j)}{||\nu||_2^2} \phi \right) \text{dir}(x^\top \nu) \right\rangle \\
&= \left(\frac{\nu_i - \nu_j}{||\nu||_2^2} \right)^2 \phi^2 + 2 \left(\frac{\nu_i - \nu_j}{||\nu||_2^2} \left\langle x_i - x_j, \text{dir}(x^\top \nu) \right\rangle - \left(\frac{\nu_i - \nu_j}{||\nu||_2^2} \right)^2 ||x^\top \nu||_2 \right) \phi \\
&+ \left\| x_i - x_j - (\nu_i - \nu_j) \frac{x^\top \nu}{||\nu||_2^2} \right\|_2^2 .
\end{aligned}
$$

■

This completes the proof of Lemma 3.

We continue with the proof of Lemma 4. Using the definition of $w_i^{(t-1)}(k)$ in (14), we have that

$$
\left\| [x'(\phi)]_i - \frac{\sum_{i'=1}^n 1\left\{ c_{i'}^{(t-1)}(x) = k \right\} [x'(\phi)]_{i'}}{\sum_{i'=1}^n 1\left\{ c_{i'}^{(t-1)}(x) = k \right\}} \right\|_2^2 = \left\| [x'(\phi)]_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k) [x'(\phi)]_{i'} \right\|_2^2 ,
$$

where

$$\big[x'(\phi)\big]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\big[x'(\phi)\big]_{i'} = \left(\sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\frac{\nu_i}{||\nu||_2^2}\right)\phi + \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\left(x_i - \nu_i\frac{||x^\top\nu||_2}{||\nu||_2^2}\text{dir}(x^\top\nu)\right)$$

is a linear function of ϕ. The rest of the proof follows directly from the same set of calculations in the proof of Lemma 3.

A.4 Proof of Proposition 5

Recall that n, q, K, T denote the number of samples (see (1)), the number of features (see (1)), the number of clusters (see Algorithm 1), and the maximum number of iterations for which Algorithm 1 is run.

According to Proposition 2, to compute the set $\mathcal{S}_T$ in (12), it suffices to compute the intersection of the two sets in (15) and (16).

We first make the following observations for our timing complexity analysis:

- Observation 1: according to Lemma 3, the set in (15) is an intersection of nK quadratic inequalities.
- Observation 2: according to Lemma 4, the set in (16) is an intersection of nKT quadratic inequalities.
- Observation 3: we can solve a quadratic inequality in $\mathcal{O}(1)$ time using the quadratic formula.
- Observation 4: we can intersect the solution sets of N quadratic inequalities in $\mathcal{O}(N\log N)$ time (Bourgon, 2020).

Equipped with these observations, we will analyze the timing complexity of computing the set (15). Note that the coefficients for each of the nK quadratic inequalities can be computed in $\mathcal{O}(nq)$ operations: first, using the property that $x^\top\nu = \sum_{i\in\hat{\mathcal{C}}_1} x_i/|\hat{\mathcal{C}}_1| - \sum_{i\in\hat{\mathcal{C}}_2} x_i/|\hat{\mathcal{C}}_2|$, we can compute $\|x^\top\nu\|_2$ and $\text{dir}(x^\top\nu)$ in $\mathcal{O}(nq)$ operations. Then, computing the coefficients a, b, and γ in Lemma 3 takes $\mathcal{O}(1)$, $\mathcal{O}(q)$, and $\mathcal{O}(q)$ operations, respectively. For each inequality, obtaining the solution set requires $\mathcal{O}(1)$ operations (see Observation 3). Finally, intersecting the solution sets of the $n(K-1)$ quadratic inequalities incurs another $\mathcal{O}(nK\log(nK))$ operations. Thus, the computational cost for (15) totals to $\mathcal{O}(nKq + nK\log(nK))$ operations.

Next, we analyze the cost of computing the set (16). Note that using Observation 2, we need to solve nKT quadratic inequalities. Here, for each quadratic inequality of the form in Lemma 4, it takes $\mathcal{O}(n), \mathcal{O}(n+q)$, and $\mathcal{O}(n+q)$ operations to compute the coefficients $\tilde{a}, \tilde{b}$, and $\tilde{\gamma}$, respectively. Moreover, for a given iteration t and cluster number k, we only need to compute $\sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\nu_{i'}$ once using $\mathcal{O}(n+q)$ operations once, as opposed to n times, since this formula does not depend on the index i. Therefore, obtaining the nKT solution sets will take $\mathcal{O}((n+q)KT)$ time. Finally, intersecting these sets using Observation 4 adds another $\mathcal{O}(nKT\log(nKT))$ operations.

Combining the costs for computing the set in (15) and the set in (16), we conclude that the cost for computing the set $\mathcal{S}_T$ in (12) is $\mathcal{O}((n+q)KT + nKT\log(nKT))$ operations.

A.5 Proof of Proposition 6 and computation of $p_{\Sigma,\text{selective}}$

The proof of Proposition 6 is similar to that of Proposition 1.

First note that for any non-zero $\nu \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times q}$, we have that

$$X = \Pi_\nu^\perp X + \frac{\nu\nu^\top X \Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}}{\|\nu\|_2^2} = \Pi_\nu^\perp X + \left(\frac{\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\text{dir}\left(\Sigma^{-\frac{1}{2}}X^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}. \quad \text{(A.40)}$$

Lemma 12 *Under* (17) *and* $H_0 : \mu^\top\nu = 0_q$, $\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2$, $\Pi_\nu^\perp X$, *and* $\text{dir}\left(\Sigma^{-\frac{1}{2}}X^\top\nu\right)$ *are pairwise independent.*

Proof As in the proof of Lemma 8, $\Pi_\nu^\perp\nu = 0_n$, and it follows from the property of the matrix normal distribution that $X^\top\nu$ is independent of $\Pi_\nu^\perp X$. Because both $\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2$ and $\text{dir}(\Sigma^{-\frac{1}{2}}X^\top\nu)$ are functions of $X^\top\nu$, both are independent of $\Pi_\nu^\perp X$.

Next, we will show that $\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2$ and $\text{dir}(\Sigma^{-\frac{1}{2}}X^\top\nu)$ are independent. Under (17) and $H_0 : \mu^\top\nu = 0_q$, we have that $\Sigma^{-\frac{1}{2}}X^\top\nu \sim \mathcal{N}(0_q, \|\nu\|_2^2\mathbf{I}_q)$. It then follows that $\Sigma^{-\frac{1}{2}}X^\top\nu$ is rotationally invariant, and therefore $\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2$ is independent of $\text{dir}(\Sigma^{-\frac{1}{2}}X^\top\nu)$ (Bilodeau and Brenner, 1999). ■

Then, recalling the definition of $p_{\Sigma,\text{selective}}$ in (19), we have that

$$p_{\Sigma,\text{selective}} = \text{pr}_{H_0}\left[\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \mid \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \text{dir}\left(\Sigma^{-\frac{1}{2}}X^\top\nu\right) = \text{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\right]$$

$$\overset{a.}{=} \text{pr}_{H_0}\left[\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp X + \frac{\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2}{\|\nu\|_2^2}\nu\left\{\text{dir}\left(\Sigma^{-\frac{1}{2}}X^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x)\right\},\right.$$
$$\left.\Pi_\nu^\perp X = \Pi_\nu^\perp x, \text{dir}(\Sigma^{-\frac{1}{2}}X^\top\nu) = \text{dir}(\Sigma^{-\frac{1}{2}}x^\top\nu)\right]$$

$$\overset{b.}{=} \text{pr}_{H_0}\left[\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \frac{\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2}{\|\nu\|_2^2}\nu\left\{\text{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x)\right\},\right.$$
$$\left.\Pi_\nu^\perp X = \Pi_\nu^\perp x, \text{dir}(\Sigma^{-\frac{1}{2}}X^\top\nu) = \text{dir}(\Sigma^{-\frac{1}{2}}x^\top\nu)\right]$$

$$\overset{c.}{=} \text{pr}_{H_0}\left[\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \frac{\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2}{\|\nu\|_2^2}\nu\left\{\text{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x)\right\}\right].$$

Here, step $a.$ follows from substituting X with the expression in (A.40). Step $b.$ follows from replacing $\Pi_\nu^\perp X$ and $\text{dir}(\Sigma^{-\frac{1}{2}}X^\top\nu)$ with $\Pi_\nu^\perp x$ and $\text{dir}(\Sigma^{-\frac{1}{2}}x^\top\nu)$, respectively. Finally, in step $c.$, we used Lemma 12. Now, under (17) and $H_0 : \mu^\top\nu = 0_q$, we have that $\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \sim \|\nu\|_2\chi_q$, which concludes the proof of (20).

It remains to show that the test that rejects $H_0 : \mu^\top\nu = 0$ when $p_{\Sigma,\text{selective}} \leq \alpha$ controls the selective Type I error, in the sense of (5). We omit the proof here, as it follows directly from the proof of Proposition 1 in Appendix A.1.

Next, we discuss how we could modify the results in Section 3 to compute the p-value $p_{\Sigma,\text{selective}}$. First note that according to Proposition 6, it suffices to compute the set

$$\mathcal{S}_T^{\Sigma} = \left\{ \phi : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)} \left(\Pi_\nu^\perp x + \left(\frac{\phi}{\|\nu\|_2^2} \right) \nu \left\{ \text{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu \right) \right\}^\top \Sigma^{\frac{1}{2}} \right) = c_i^{(t)}(x) \right\} \right\}. \tag{A.41}$$

In addition, letting $\tilde{x}'(\phi)$ denote $\Pi_\nu^\perp x + \left(\frac{\phi}{\|\nu\|_2^2}\right)\nu\left\{\text{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}$, we see that $\tilde{x}'(\phi)$ is in fact a linear function of ϕ with

$$[\tilde{x}'(\phi)]_i = x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \text{dir}(x^\top \nu) + \left(\frac{||x^\top \nu||_2}{||\Sigma^{-\frac{1}{2}} x^\top \nu||_2} \frac{\nu_i}{||\nu||_2^2} \phi \right) \text{dir}(x^\top \nu). \tag{A.42}$$

Therefore, a minor modification of Proposition 2 yields the following corollary.

Corollary 13 *Suppose the k-means clustering algorithm (see Algorithm 1) with K clusters the data x, when applied to the data x, runs for T steps. Then, for the set $\mathcal{S}_T^{\Sigma}$ defined in* (A.41)*, we have that*

$$\begin{aligned} \mathcal{S}_T^{\Sigma} = &\left(\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[\tilde{x}'(\phi)\right]_i - m^{(0)}_{c_i^{(0)}(x)}(\phi) \right\|_2^2 \le \left\| \left[\tilde{x}'(\phi)\right]_i - m_k^{(0)}(\phi) \right\|_2^2 \right\} \right) \cap \\ &\left(\bigcap_{t=1}^{T} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| \left[\tilde{x}'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)} \left(c_{i'}^{(t-1)}(x) \right) \left[\tilde{x}'(\phi)\right]_{i'} \right\|_2^2 \le \left\| \left[\tilde{x}'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k) \left[\tilde{x}'(\phi)\right]_{i'} \right\|_2^2 \right\} \right). \end{aligned} \tag{A.43}$$

We also have the following extensions of Lemmas 3 and 4, which enable efficient computation of the expressions in Corollary 13.

Lemma 14 (Section 4.2 in Gao et al. (2022)) *For $\tilde{x}'(\phi)$ in* (A.42) *and ν in* (3)*, $\left\|[\tilde{x}'(\phi)]_i - [\tilde{x}'(\phi)]_j\right\|_2^2 = a'\phi^2 + b'\phi + \gamma'$, where $a' = \left(\frac{||x^\top\nu||_2}{||\Sigma^{-\frac{1}{2}}x^\top\nu||_2}\right)^2\left(\frac{\nu_i-\nu_j}{\|\nu\|_2^2}\right)^2$, $b' = 2\left(\frac{||x^\top\nu||_2}{||\Sigma^{-\frac{1}{2}}x^\top\nu||_2}\right)\left(\frac{\nu_i-\nu_j}{\|\nu\|_2^2}\langle x_i - x_j, \text{dir}(x^\top\nu)\rangle - \left(\frac{\nu_i-\nu_j}{\|\nu\|_2^2}\right)^2\|x^\top\nu\|_2\right)$, and $\gamma' = \left\|x_i - x_j - (\nu_i - \nu_j)\frac{x^\top\nu}{||\nu||_2^2}\right\|_2^2$.*

Lemma 15 *For $\tilde{x}'(\phi)$ in* (A.42)*, ν in* (3)*, and $w_i^{(t)}(k)$ in* (14)*, $\left\|[\tilde{x}'(\phi)]_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)[\tilde{x}'(\phi)]_{i'}\right\|_2^2 = \tilde{a}'\phi^2 + \tilde{b}'\phi + \tilde{\gamma}'$, where*

$$\tilde{a}' = \frac{1}{\|\nu\|_2^4}\left(\frac{||x^\top\nu||_2}{||\Sigma^{-\frac{1}{2}}x^\top\nu||_2}\right)^2\left(\nu_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\nu_{i'}\right)^2,$$

$$\tilde{b}' = \left(\frac{2||x^\top\nu||_2}{\|\nu\|_2^2||\Sigma^{-\frac{1}{2}}x^\top\nu||_2}\right)\left\{\left(\nu_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\nu_{i'}\right)\left\langle x_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)x_{i'}, \text{dir}(x^\top\nu)\right\rangle - \frac{\|x^\top\nu\|_2}{\|\nu\|_2^4}\left(\nu_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\nu_{i'}\right)^2\right\},$$

and

$$\tilde{\gamma}' = \left\| x_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)x_{i'} - \left(\nu_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)\nu_{i'}\right)\frac{x^\top\nu}{\|\nu\|_2^2}\right\|_2^2.$$

Proofs of Lemmas 14 and 15 follow from the same set of calculations in the proofs of Lemmas 3 and 4 in Appendix A.3.

A.6 Proof of Proposition 7

Proof of Proposition 7 is similar to the proof of Lemma 1 in Markovic et al. (2017) and the proof of Lemma 7 in Tibshirani et al. (2018).

We first present an auxiliary lemma.

Lemma 16 *For any $c_i^{(t)}(x), i = 1, \ldots, n; t = 1, \ldots, T$, $\hat{p}_{selective}(\hat{\sigma})$ defined in* (21) *is a continuous and monotonically increasing function of $\hat{\sigma}$.*

Proof By the definition in (21), we have that

$$\hat{p}_{\text{selective}}(\hat{\sigma}) = \frac{\int_{\|x^\top \nu\|_2}^{\infty} (\frac{1}{2})^{q/2-1} \frac{t^{q-1}}{\Gamma(q/2)} \|\nu\|_2^{-q} \hat{\sigma}^{-q} \exp\left(-\frac{t^2}{2\hat{\sigma}^2 \|\nu\|_2^2}\right) 1\{t \in \mathcal{S}_T\} dt}{\int_0^{\infty} (\frac{1}{2})^{q/2-1} \frac{t^{q-1}}{\Gamma(q/2)} \|\nu\|_2^{-q} \hat{\sigma}^{-q} \exp\left(-\frac{t^2}{2\hat{\sigma}^2 \|\nu\|_2^2}\right) 1\{t \in \mathcal{S}_T\} dt}, \tag{A.44}$$

where $\mathcal{S}_T$ defined in (12) is a function of $c_i^{(t)}(x), i = 1, \ldots, n; t = 1, \ldots, T$. By inspection, (A.44) is a continuous function of $\hat{\sigma}$, because the product or ratio of two continuous functions is still continuous. It remains to show that (A.44) is increasing in $\hat{\sigma}$. This follows directly from Lemma S3. of Gao et al. (2022). ■

Provided that $\hat{\sigma}$ converges to σ in probability, we can combine Lemma 16 and the continuous mapping theorem to see that $\hat{p}_{\text{selective}}(\hat{\sigma})$ converges to $p_{\text{selective}}(\sigma)$ in probability, i.e., for all $\epsilon > 0, \lim_{q\to\infty} \text{pr}(|\hat{p}_{\text{selective}}(\hat{\sigma}) - p_{\text{selective}}(\sigma)| \geq \epsilon) = 0$. Next, letting A_q denote the event $\bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right) \right\}$, we will show that under the assumptions in Proposition 7, $\hat{p}_{\text{selective}}(\hat{\sigma})$ converges to $p_{\text{selective}}(\sigma)$ in probability, *conditional on A_q*. For any $\epsilon > 0$, we have that

$$\begin{aligned}
&\lim_{q\to\infty} \text{pr}_{H_0^{(q)}} \{|\hat{p}_{\text{selective}}(\hat{\sigma}) - p_{\text{selective}}(\sigma)| \leq \epsilon \mid A_q\} \\
&\overset{a.}{=} \lim_{q\to\infty} \frac{\text{pr}_{H_0^{(q)}} \{|\hat{p}_{\text{selective}}(\hat{\sigma}) - p_{\text{selective}}(\sigma)| \leq \epsilon, A_q\}}{\text{pr}_{H_0^{(q)}}(A_q)} \\
&\overset{b.}{\geq} \lim_{q\to\infty} \frac{\text{pr}_{H_0^{(q)}}(A_q) - \text{pr}_{H_0^{(q)}} \{|\hat{p}_{\text{selective}}(\hat{\sigma}) - p_{\text{selective}}(\sigma)| > \epsilon\}}{\text{pr}_{H_0^{(q)}}(A_q)} \\
&\overset{c.}{=} \frac{\lim_{q\to\infty} \text{pr}_{H_0^{(q)}}(A_q) - \lim_{q\to\infty} \text{pr}_{H_0^{(q)}} \{|\hat{p}_{\text{selective}}(\hat{\sigma}) - p_{\text{selective}}(\sigma)| > \epsilon\}}{\lim_{q\to\infty} \text{pr}_{H_0^{(q)}}(A_q)} \\
&\overset{d.}{=} \frac{\delta}{\delta} = 1.
\end{aligned}$$

Here, step a. follows from Bayes rule, and the observation that the denominator is non-zero for finite q. In step b., we used the lower bound that for events A, B defined on the same probability space, $\text{pr}(A \cap B) = \text{pr}(A) - \text{pr}(A \setminus B) \geq \text{pr}(A) - \text{pr}(B^C)$. Next, c. follows from distributing the limit, which is valid because of the assumption that $\lim_{q\to\infty} \text{pr}_{H_0^{(q)}}(A_q) = \delta > 0$; finally, d. follows from the fact that $\hat{p}_{\text{selective}}(\hat{\sigma})$ converges to $p_{\text{selective}}(\sigma)$ in probability for any sequence of $\mu^{(q)}, q = 1, 2, \ldots$, which implies the convergence under $H_0 : {\mu^{(q)}}^\top \nu^{(q)} = 0$ as well.

Finally, we have that

$$\lim_{q\to\infty} \mathrm{pr}_{H_0^{(q)}}\{\hat{p}_{\text{selective}}(\hat{\sigma}) \le \alpha \mid A_q\} \overset{a.}{=} \lim_{q\to\infty} \mathrm{pr}_{H_0^{(q)}}\{p_{\text{selective}}(\sigma) \le \alpha \mid A_q\} \overset{b.}{=} \lim_{q\to\infty} \alpha = \alpha. \tag{A.45}$$

Here, step a. follows from $\hat{p}_{\text{selective}}(\hat{\sigma})$ converging to $p_{\text{selective}}(\sigma)$ in probability, *conditional on A_q*. Step b. follows from the fact that the result of Proposition 1 applies for any positive integer q. This completes the proof of Proposition 7.

Proposition 7 assumes that we have a consistent estimator $\hat{\sigma}$ of σ. In Appendix A.7, we analyze different estimators of σ in (1), and prove that, under appropriate sparsity assumptions on μ in (1), $\hat{\sigma}_{\text{MED}}$ in (22) is a consistent estimator for σ.

As an alternative, we can also use an asymptotically conservative estimator of σ as in Gao et al. (2022). This leads to an asymptotically conservative p-value; details are stated in Corollary 17.

Corollary 17 *For $q = 1, 2, \ldots,$ suppose that $X^{(q)} \sim \mathcal{MN}_{n\times q}\big(\mu^{(q)}, \mathbf{I}_n, \sigma^2\mathbf{I}_q\big)$. Let $x^{(q)}$ be a realization from $X^{(q)}$ and $c_i^{(t)}(\cdot)$ be the cluster to which the ith observation is assigned during the tth iteration of step 3b. of Algorithm 1. Consider the sequence of null hypotheses $H_0^{(q)} : {\mu^{(q)}}^\top \nu^{(q)} = 0_q$, where $\nu^{(q)}$ defined in* (3) *is the contrast vector resulting from applying k-means clustering on $x^{(q)}$. Suppose that (i) $\hat{\sigma}$ is an asymptotically conservative estimator of σ, i.e., $\lim_{q\to\infty} pr\big(\hat{\sigma}(X^{(q)}) \ge \sigma\big) = 1$; and (ii) there exists $\delta \in (0,1)$ such that $\lim_{q\to\infty} pr_{H_0^{(q)}}\left[\bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] > \delta$. Then, $\forall \alpha \in (0,1)$, we have that $\lim_{q\to\infty} pr_{H_0^{(q)}}\left[\hat{p}_{selective}(\hat{\sigma}) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] \le \alpha.$*

We omit the proof of Corollary 17, as it follows directly from combining the proof of Proposition 7 and the fact that $\hat{p}_{\text{selective}}(\hat{\sigma})$ is a monotonically increasing function of $\hat{\sigma}$ (see Lemma 16).

Finally, we remark that, in principle, the result in Proposition 7 can be extended to an unknown covariance matrix Σ. However, estimating Σ is challenging, especially when q is comparable to, or larger than, n (Rousseeuw, 1987; Bickel and Levina, 2008; Avella-Medina et al., 2018). It may be possible to leverage recent advances in robust covariance matrix estimation (e.g., Han and Liu (2014); Chen et al. (2018); Belomestny et al. (2019)) to obtain a consistent estimator of Σ under model (17).

A.7 Estimating σ in (1)

Proposition 7 states that, under appropriate assumptions, a consistent estimator of σ in (1) leads to asymptotic selective Type I error control. In this section, we analyze the asymptotic behavior of the two variance estimators considered in Section 5, $\hat{\sigma}^2_{\text{MED}}$ and $\hat{\sigma}^2_{\text{Sample}}$. In particular, we prove that under model (1) and a sparsity assumption on μ (defined in (1)), a close analog of $\hat{\sigma}^2_{\text{MED}}$ in (22) that does not subtract the column median is a consistent estimator of σ^2. Moreover, we prove that $\hat{\sigma}^2_{\text{Sample}}$ is a conservative estimator of σ^2, and characterize its exact bias.

We first introduce an auxiliary result that specifies the rate of convergence for a median-based estimator of the variance in the sparse vector model (Comminges et al., 2021). For a vector $\theta \in \mathbb{R}^n$, we use $\|\theta\|_0$ to denote its ℓ_0 norm, i.e. $\|\theta\|_0 = \sum_{i=1}^n 1\{\theta_i \neq 0\}$.

Lemma 18 (Proposition 6 in Comminges et al. (2021)) *Consider the model*

$$Y_i = \theta_i + \sigma\xi_i, \quad i = 1, \dots, d, \tag{A.46}$$

where σ is unknown, and the independently and identically distributed noise ξ_i satisfies that (i) $\mathrm{E}(\xi_i) = 0$; (ii) $\mathrm{E}(\xi_i^2) = 1$; and (iii) $\mathrm{E}\big(|\xi_i|^{2+\epsilon}\big) < \infty$ for some $\epsilon > 0$. We further assume that the signal θ is s-sparse, i.e., $\|\theta\|_0 \leq s$. Denoting by $M_{\xi_1^2}$ the median of ξ_1^2, we consider the following estimator of σ^2:

$$\bar{\sigma}^2_{MED} = \mathrm{median}(Y_1^2, \dots, Y_d^2)/M_{\xi_1^2}. \tag{A.47}$$

Then, there exist constants $\gamma \in (0, 1/8)$, $C > 0$ depending only on the cumulative distribution function of ξ_1 such that for all integers s and d satisfying $1 \leq s < \gamma d$,

$$\sup_{\sigma>0} \sup_{\|\theta\|_0 \leq s} \frac{1}{\sigma^2} \mathrm{E}\big\{\big|\bar{\sigma}^2_{MED} - \sigma^2\big|\big\} \leq C \max\left(\frac{1}{d^{1/2}}, \frac{s}{d}\right). \tag{A.48}$$

Building on Lemma 18, in Corollary 19, we analyze the properties of an estimator closely related to $\hat{\sigma}^2_{\mathrm{MED}}$ in (22). In particular, this estimator $\tilde{\sigma}^2_{\mathrm{MED}}$ does not subtract the median of each column in the input data. While $\hat{\sigma}^2_{\mathrm{MED}}$ and $\tilde{\sigma}^2_{\mathrm{MED}}$ are very similar provided that μ is sparse, we expect $\hat{\sigma}^2_{\mathrm{MED}}$ to perform better empirically in scenarios where μ is sparse *up to a constant shift*, i.e., there exists a matrix C such that (i) each column of C takes on the same value; and (ii) $\mu + C$ is sparse.

Corollary 19 *Under model* (1), *consider*

$$\tilde{\sigma}^2_{MED}(X) = \left\{\underset{1\leq i\leq n, 1\leq j\leq q}{\mathrm{median}}\big(X_{ij}^2\big)\right\}/M_{\chi_1^2}, \tag{A.49}$$

where $M_{\chi_1^2}$ is the median of the χ_1^2 distribution. Then, there exist constants $\gamma_0 \in (0, 1/8)$, $c_0 > 0$ such that for all integers s and q satisfying $1 \leq s < \gamma_0 q$,

$$\sup_{\sigma>0} \sup_{\max\limits_{1\leq i\leq n} \|\mu_i\|_0 \leq s} \frac{1}{\sigma^2} \mathrm{E}\big\{\big|\tilde{\sigma}^2_{MED} - \sigma^2\big|\big\} \leq c_0 \max\left\{\frac{1}{(nq)^{1/2}}, \frac{s}{q}\right\}. \tag{A.50}$$

Proof First note that (1) can be re-written into the form of (A.46):

$$X_{ij} = \mu_{ij} + \sigma\xi_{ij}, \quad i = 1, \dots, n,\ j = 1, \dots, q, \tag{A.51}$$

where ξ_{ij} is independently and identically distributed as $\mathcal{N}(0, 1)$. Therefore, the estimator $\tilde{\sigma}^2_{\mathrm{MED}}(X)$ in (A.49) is the estimator (A.47) applied to the model (A.51). Moreover, $\max\limits_{1\leq i\leq n} \|\mu_i\|_0 \leq s$ implies that $\sum_{i=1}^n \sum_{j=1}^q 1\{\mu_{ij} \neq 0\} \leq ns$. Applying Lemma 18, we have that

$$\sup_{\sigma>0} \sup_{\max\limits_{1\leq i\leq n} \|\mu_i\|_0 \leq s} \frac{1}{\sigma^2} \mathrm{E}\big\{\big|\tilde{\sigma}^2_{\mathrm{MED}}(X) - \sigma^2\big|\big\} \leq c_0 \max\left\{\frac{1}{(nq)^{1/2}}, \frac{ns}{nq}\right\} = c_0 \max\left\{\frac{1}{(nq)^{1/2}}, \frac{s}{q}\right\},$$

where c_0 is some universal constant. ■

In words, Corollary 19 states that under model (1), the rate of convergence of $\tilde{\sigma}^2_{\text{MED}}$ in mean (and therefore, in probability) is $\max\left\{1/(nq)^{1/2}, s/q\right\}$. In particular, $\tilde{\sigma}^2_{\text{MED}}$ is a consistent estimator of σ^2 provided that $s/q \to 0$ as $q \to \infty$.

Next, we investigate the property of the sample variance estimator $\hat{\sigma}^2_{\text{Sample}}$.

Proposition 20 *Under model* (1), *for* $\hat{\sigma}^2_{Sample}(X) = \sum_{i=1}^n \sum_{j=1}^q \left(X_{ij} - \bar{X}_j\right)^2/(nq-q)$, *we have that*

$$\mathrm{E}\{\hat{\sigma}^2_{Sample}(X)\} - \sigma^2 = \frac{1}{2n(n-1)q}\sum_{j=1}^q\sum_{i=1}^n\sum_{i'=1}^n(\mu_{ij} - \mu_{i'j})^2. \tag{A.52}$$

Moreover, for any integers s and q such that $ns \leq q$, we have that, for some constant $\tilde{c}_0$,

$$\sup_{\sigma>0}\ \sup_{\max\limits_{1\leq i\leq n}\|\mu_i\|_0\leq s} \frac{1}{\sigma^2}\mathrm{E}\{|\hat{\sigma}^2_{Sample}(X) - \sigma^2|\} \geq \tilde{c}_0\frac{s}{q}. \tag{A.53}$$

Proof We start with the proof of (A.52). Under (1), the following holds:

$$\begin{aligned}
\mathrm{E}\{\hat{\sigma}^2_{\text{Sample}}(X)\} &= \mathrm{E}\left\{\sum_{i=1}^n\sum_{j=1}^q \left(X_{ij} - \bar{X}_j\right)^2/(nq-q)\right\}\\
&= \frac{1}{(n-1)q}\mathrm{E}\left[\sum_{i=1}^n\sum_{j=1}^q\{X_{ij}^2 - (\bar{X}_j)^2\}\right]\\
&= \frac{1}{(n-1)q}\sum_{i=1}^n\sum_{j=1}^q\left[\sigma^2 + \mu_{ij}^2 - \left\{\frac{\sigma^2}{n} + \frac{1}{n^2}\left(\sum_{i'=1}^n \mu_{i'j}\right)^2\right\}\right]\\
&= \sigma^2 + \frac{1}{n^2(n-1)q}\sum_{i=1}^n\sum_{j=1}^q\left\{n^2\mu_{ij}^2 - \left(\sum_{i'=1}^n\mu_{i'j}\right)^2\right\}\\
&= \sigma^2 + \frac{1}{n(n-1)q}\sum_{j=1}^q\left\{\left(\sum_{i=1}^n n\mu_{ij}^2\right) - \left(\sum_{i'=1}^n\mu_{i'j}\right)^2\right\}\\
&= \sigma^2 + \frac{1}{2n(n-1)q}\sum_{j=1}^q\sum_{i=1}^n\sum_{i'=1}^n\left(\mu_{ij} - \mu_{i'j}\right)^2.
\end{aligned}$$

Here, the last equality follows from Langrange's identity, which states that $\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right) - \left(\sum_{i=1}^n a_ib_i\right)^2 = 1/2\sum_{i=1}^n\sum_{i'=1}^n\left(a_ib_{i'} - a_{i'}b_i\right)^2$.

To prove the second statement, we consider a specific matrix $\tilde{\mu} \in \mathbb{R}^{n\times q}$ with exactly $ns \leq q$ non-zero entries. In addition, each column of $\tilde{\mu}$ has at most one non-zero entry and each row of $\tilde{\mu}$ has exactly s non-zero entries. This is possible because ns is assumed to be less than q. Finally, we assume that the square of the minimal non-zero entry of $\tilde{\mu}$, $\min\limits_{i,j:\tilde{\mu}_{ij}\neq 0} \tilde{\mu}_{ij}^2$, is lower bounded by some universal constant M. Then, we have that

$$\begin{aligned}
&\sup_{\sigma>0} \sup_{\max\limits_{1\leq i\leq n} \|\mu_i\|_0 \leq s} \frac{1}{\sigma^2}\mathrm{E}\left\{\left|\hat{\sigma}^2_{\text{Sample}}(X) - \sigma^2\right|\right\} \\
&\qquad \overset{a.}{\geq} \sup_{\sigma>0} \frac{1}{\sigma^2}\mathrm{E}_{X\sim\mathcal{MN}(\tilde{\mu},\mathbf{I}_n,\sigma^2\mathbf{I}_q)}\left\{\left|\hat{\sigma}^2_{\text{Sample}}(X) - \sigma^2\right|\right\} \\
&\qquad \overset{b.}{\geq} \sup_{\sigma>0} \frac{1}{\sigma^2}\mathrm{E}_{X\sim\mathcal{MN}(\tilde{\mu},\mathbf{I}_n,\sigma^2\mathbf{I}_q)}\left\{\hat{\sigma}^2_{\text{Sample}}(X) - \sigma^2\right\} \\
&\qquad \overset{c.}{\geq} \sup_{\sigma>0} \frac{1}{\sigma^2}\frac{1}{2n(n-1)q}\sum_{j=1}^{q}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\tilde{\mu}_{ij} - \tilde{\mu}_{i'j})^2 \\
&\qquad \geq \sup_{\sigma>0} \frac{1}{\sigma^2}\frac{1}{2n(n-1)q}\sum_{j=1}^{q}\sum_{i=1}^{n}\sum_{i'=1}^{n}1\{\tilde{\mu}_{ij}\neq 0\}1\left\{\tilde{\mu}_{i'j} = 0\right\}\left(\tilde{\mu}_{ij} - \tilde{\mu}_{i'j}\right)^2 \\
&\qquad \overset{d.}{\geq} \sup_{\sigma>0} \frac{1}{\sigma^2}\frac{M(n-1)ns}{2n(n-1)q} \\
&\qquad \geq \tilde{c}_0\frac{s}{q}.
\end{aligned}$$

Here, $a.$ follows from picking any $\tilde{\mu}$ satisfying the conditions outlined above, since by construction, $\max\limits_{1\leq i\leq n}\|\tilde{\mu}_i\|_0 = s$. Steps $b.$ and $c.$ follow from the inequality $\mathrm{E}(|X|) \geq \mathrm{E}(X)$ and the expression for $\mathrm{E}\left\{\hat{\sigma}^2_{\text{Sample}}(X)\right\}$ in (A.52), respectively. Finally, to prove $d.$, we note that for each of the ns columns with exactly one non-zero element, there are $n-1$ pairs of $(i,i'), i = 1,\ldots,n; i' = 1,\ldots,n$ such that the product $1\{\tilde{\mu}_{ij}\neq 0\}1\left\{\tilde{\mu}_{i'j} = 0\right\}$ is non-zero. Moreover, each of pair contributes at least M by the assumption that $\min_{i,j:\tilde{\mu}_{ij}\neq 0}\tilde{\mu}_{ij}^2 \geq M$. ■

Contrasting the results in Corollary 19 and Proposition 20, we note that, under (1), the convergence of $\tilde{\sigma}^2_{\text{MED}}$ depends critically on the sparsity parameter s (or, equivalently, the ℓ_0 norm of μ_i), whereas the convergence of $\hat{\sigma}^2_{\text{Sample}}$ is determined by $\sum_{j=1}^{q}\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(\mu_{ij} - \mu_{i'j}\right)^2$. Thus, in scenarios where the underlying means $\mu_i, i = 1,\ldots,n$ are sparse (e.g., (23) in Section 5), we expect $\tilde{\sigma}^2_{\text{MED}}$ (and therefore its "centered" analog $\hat{\sigma}^2_{\text{MED}}$ in (22)) to be a less conservative estimator of σ^2. As a result, we expect the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ to be more powerful than that based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$, as shown in Figure 4 of Section 5.

A.8 Additional power comparisons

In Section 5.2, we compared the conditional power of the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ under (23). Here, we conduct two additional analyses.

In the first analysis, we consider a different notion of power that does not condition on $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ being true clusters. In this case, comparing the power of the tests requires a bit of care, because the effect size $\|\mu^\top\nu\|_2$ may differ across simulated datasets from the same data-generating distribution. As a result, we consider the power of the tests *as a function of* $\|\mu^\top\nu\|_2$. We fit a regression spline using the `gam` function in the `R` package `mgcv` (Wood,

2017) to obtain a smooth estimate of power on the same simulated datasets from Section 5.2. The results are in Figure 6. The power of the tests that reject H_0 if $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, or $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ is less than $\alpha = 0.05$ increases as $\|\mu^\top \nu\|_2$ increases. For a given value of $\|\mu^\top \nu\|_2$ and σ, the test based on $p_{\text{selective}}$ has the highest power, followed by that based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$; the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ has the lowest power.

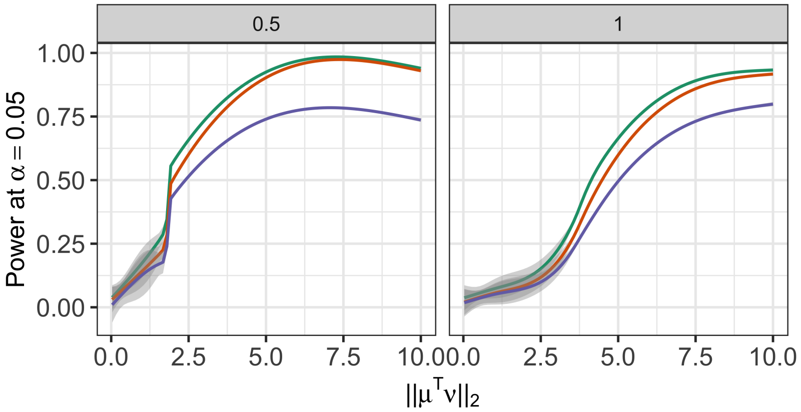


Figure 6: *Left* : Additional analysis of the data in Section 5.2 with $\sigma = 0.5$. We fit a regression spline to display the power of the tests based on $p_{\text{selective}}$ (green line), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange line), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple line) as a function of $\|\mu^\top \nu\|_2$. *Right* : Same as left, but for $\sigma = 1$.

In the second analysis, we consider the conditional power (defined in (24)) of the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ under a different data generating model than (23). We generate data from (1) with $n = 150$ and

$$\mu_1 = \ldots = \mu_{\frac{n}{3}} = \begin{bmatrix} \theta_1 \\ 0_{0.9q} \end{bmatrix}, \ \mu_{\frac{n}{3}+1} = \ldots = \mu_{\frac{2n}{3}} = \begin{bmatrix} \theta_2 \\ 0_{0.9q} \end{bmatrix}, \ \mu_{\frac{2n}{3}+1} = \ldots = \mu_n = \begin{bmatrix} \theta_3 \\ 0_{0.9q} \end{bmatrix}, \tag{A.54}$$

where, q is taken to be a multiple of 10, and for $\delta > 0$, $\theta \in \mathbb{R}^{3\times 0.1q}$ has orthogonal rows, with $\|\theta_i\|_2^2 = \delta/2$ for $i = 1, 2, 3$. As in Section 5.2, we can think of $\mathcal{C}_1 = \{1, \ldots, n/3\}, \mathcal{C}_2 = \{(n/3) + 1, \ldots, (2n/3)\}, \mathcal{C}_3 = \{(2n/3) + 1, \ldots, n\}$ as "true clusters". Under (A.54), the pairwise distance between each pair of true clusters is δ.

We generate $M = 100,000$ datasets from (A.54) with $q = 50$,$\sigma = 0.25, 0.5, 1$, and $\delta = 2, 3, \ldots, 10$. For each simulated dataset, we apply k-means clustering with $K = 3$ and reject $H_0 : \mu^\top \nu = 0_q$ if $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, or $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ is less than $\alpha = 0.05$. Figure 7(a) displays the detection probability (25) of k-means clustering as a function of δ in (A.54). Under model (1), the detection probability increases as a function of δ in (A.54) across all values of σ. For a given value of δ, a larger value of σ leads to lower detection probability. Figure 7(b) displays the conditional power (24) for the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$. For some combinations of δ and σ, the conditional power is not displayed, because the true clusters are never recovered in simulation. For all tests and values of σ under consideration, conditional power is an increasing function of δ. For a given test and a value of δ, smaller σ leads to higher conditional power. Moreover, for

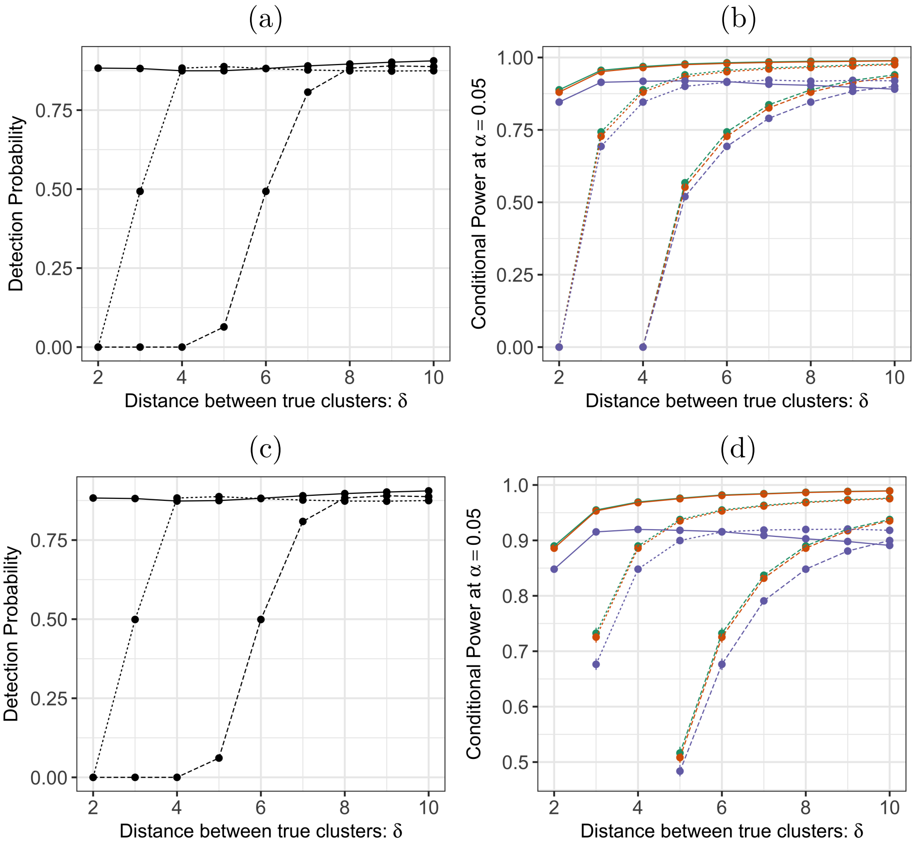


Figure 7: *(a):* Detection probability defined in (25) for k-means clustering with $K = 3$ under model (1) with $n = 150$, $q = 50$, and μ in (A.54), across $\delta = \|\theta_i - \theta_j\|_2$ in (A.54) and $\sigma = 0.25$ (solid lines), 0.5 (dashed lines), and 1 (long-dashed lines). *(b):* The conditional power (24) at $\alpha = 0.05$ for the tests based on $p_{\text{selective}}$ (green), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple), under model (1) with $n = 150$, $q = 50$, and μ in (A.54). *(c):* Same as (a), but for μ in (23). *(d):* Same as (b), but for μ in (23).

the same values of δ and σ, the test based on $p_{\text{selective}}$ has the highest conditional power, followed closely by the test based of $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. Using $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ leads to a less powerful test, especially for larger values of δ. As a comparison, we included the detection probability and conditional power under model (23) with $q = 50$ in panels (c) and (d) of Figure 7. The tests under consideration behave qualitatively similarly as a function of δ and σ. Under (23), we observe an even larger gap between the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ and the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$.

A.9 Additional results for real data applications

In this section, we visualize the estimated clusters for the single cell RNA-sequencing data in Section 6.2.

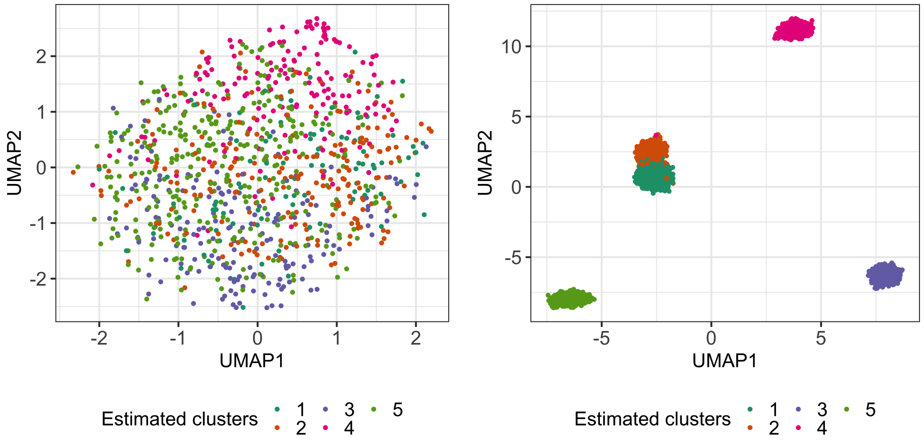


Figure 8: *Left:* The two-dimensional UMAP embedding (McInnes et al., 2018) of the "no cluster" dataset after preprocessing (as described in Section 6.2), colored by the estimated cluster membership via k-means clustering. *Right:* Same as left, but for the "cluster" dataset.