Selective Inference for Hierarchical Clustering

Lucy L. Gao[†,*], Jacob Bien[◦], and Daniela Witten [‡]

† Department of Statistics and Actuarial Science, University of Waterloo

◦ Department of Data Sciences and Operations, University of Southern California

‡ Departments of Statistics and Biostatistics, University of Washington

September 23, 2021

Abstract

Classical tests for a difference in means control the type I error rate when the groups are defined *a priori*. However, when the groups are instead defined via clustering, then applying a classical test yields an extremely inflated type I error rate. Notably, this problem persists even if two separate and independent data sets are used to define the groups and to test for a difference in their means. To address this problem, in this paper, we propose a selective inference approach to test for a difference in means between two clusters. Our procedure controls the selective type I error rate by accounting for the fact that the choice of null hypothesis was made based on the data. We describe how to efficiently compute exact p-values for clusters obtained using agglomerative hierarchical clustering with many commonly-used linkages. We apply our method to simulated data and to single-cell RNA-sequencing data.

Keywords: post-selection inference, hypothesis testing, difference in means, type I error

*Corresponding author: lucy.gao@uwaterloo.ca

1 Introduction

Testing for a difference in means between groups is fundamental to answering research questions across virtually every scientific area. Classical tests control the type I error rate when the groups are defined *a priori*. However, it is increasingly common for researchers to instead define the groups via a clustering algorithm. In the context of single-cell RNA-sequencing data, researchers often cluster the cells to identify putative cell types, then test for a difference in mean gene expression between the putative cell types in that same data set (Hwang et al. 2018, Zhang et al. 2019). In fact, the resulting inferential challenges have been described as a "grand challenge" in the field (Lähnemann et al. 2020). Similar issues arise in the field of neuroscience (Kriegeskorte et al. 2009).

In this paper, we develop a valid test for a difference in means between two clusters estimated from the data. We consider the following model for n observations of q features:

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

where $\boldsymbol{\mu} \in \mathbb{R}^{n\times q}$, with rows μ_i, is unknown, and $\sigma^2 > 0$ is known. For $\mathcal{G} \subseteq \{1, 2, \ldots, n\}$, let

$$\bar{\mu}_{\mathcal{G}} \equiv \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \mu_i \quad \text{and} \quad \bar{X}_{\mathcal{G}} \equiv \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} X_i, \tag{2}$$

which we refer to as the mean of $\mathcal{G}$ and the empirical mean of $\mathcal{G}$ in $\mathbf{X}$, respectively. Given a realization $\mathbf{x} \in \mathbb{R}^{n\times q}$ of $\mathbf{X}$, we first apply a clustering algorithm $\mathcal{C}$ to obtain $\mathcal{C}(\mathbf{x})$, a partition of $\{1, 2, \ldots, n\}$. We then use $\mathbf{x}$ to test, for a pair of clusters $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$,

$$H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2} \quad \text{versus} \quad H_1^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} \neq \bar{\mu}_{\hat{\mathcal{C}}_2}. \tag{3}$$

It is tempting to simply apply a Wald test of (3), with p-value given by

$$\mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\left(\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2\right), \tag{4}$$

where (4) is calculated using the $\left(\sigma\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|}+\frac{1}{|\hat{\mathcal{C}}_2|}}\right)\cdot\chi_q$ distribution. However, since we clustered $\mathbf{X}$ to get $\mathcal{C}(\mathbf{X})$, we will observe substantial differences between the empirical means of the estimated clusters in $\mathcal{C}(\mathbf{X})$, even when there is no signal in the data, as is shown in Figure 1(a). In short, we used the data to select the null hypothesis, and so the null distribution of the Wald test statistic is not proportional to a χ_q distribution. Thus, the Wald test is extremely anti-conservative, as illustrated in Figure 1(b).

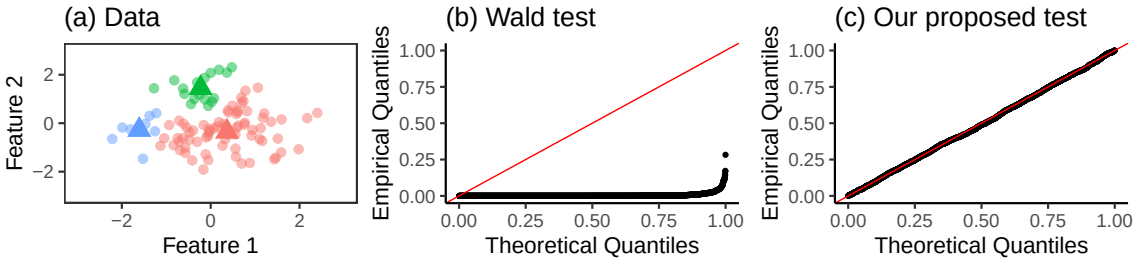

Figure 1: (a) A simulated data set from (1) with $\boldsymbol{\mu}=\mathbf{0}_{100\times 2}$ and $\sigma^2=1$. We apply average linkage hierarchical clustering to get three clusters. The empirical means (defined in (2)) of the three clusters are displayed as triangles. QQ-plots of the Uniform$(0,1)$ distribution against the p-values from (b) the Wald test in (4) and (c) our proposed test, over 2000 simulated data sets from (1) with $\boldsymbol{\mu}=\mathbf{0}_{100\times 2}$ and $\sigma^2=1$. For each simulated data set, a p-value was computed for a randomly chosen pair of estimated clusters.

At first glance, it seems that we might be able to overcome this problem via sample splitting. That is, we divide the observations into a training and a test set, cluster the observations in the training set, and then assign the test set observations to those clusters, as in Figures 2(a)–(c). Then, we apply the Wald test in (4) to the test set. Unfortunately, by assigning test observations to clusters, we have once again used the data to select a null hypothesis to test, in the sense that $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ in (3) is a function of the test observations.

Thus, the Wald test is extremely anti-conservative (Figure 2(d)). In other words, sample splitting does not provide a valid way to test the hypothesis in (3).

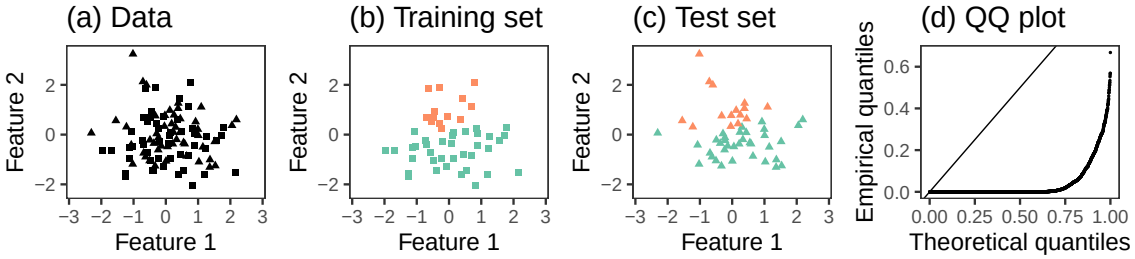

Figure 2: (a) A simulated data set from (1) with $\boldsymbol{\mu} = \mathbf{0}_{100\times 2}$ and $\sigma^2 = 1$. (b) We cluster the training set using average linkage hierarchical clustering. (c) We assign clusters in the test set by training a 3-nearest neighbors classifier on the training set. (d) QQ-plot of the Uniform$(0, 1)$ distribution against the Wald p-values in the test set, over 2000 simulated data sets for which each cluster in the test set was assigned at least one observation.

In this paper, we develop a *selective inference* framework to test for a difference in means after clustering. This framework exploits ideas from the recent literature on selective inference for regression and changepoint detection (Fithian et al. 2014, Loftus & Taylor 2015, Lee et al. 2016, Yang et al. 2016, Hyun et al. 2018, Jewell et al. 2019, Mehrizi & Chenouri 2021). The key idea is as follows: since we chose to test $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ because $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$, we can account for this hypothesis selection procedure by defining a p-value that conditions on the event $\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X})\}$. This yields a correctly-sized test, as seen in Figure 1(c).

A large body of work evaluates the statistical significance of a clustering by testing the goodness-of-fit of models under the misspecification of the number of clusters (Chen et al. 2001, Liu et al. 2008, Chen et al. 2012, Maitra et al. 2012, Kimes et al. 2017) or by assess-

ing the stability of estimated clusters (Suzuki & Shimodaira 2006). Most of these papers conduct bootstrap sampling or asymptotic approximations to the null distribution. Our proposed framework avoids the need for resampling and provides exact finite-sample inference for the difference in means between a pair of estimated clusters, under the assumption that σ in (1) is known. Chapter 3 of Campbell (2018) considers testing for a difference in means after convex clustering (Hocking et al. 2011), a relatively esoteric form of clustering. Our framework is particularly efficient when applied to hierarchical clustering, which is one of the most popular types of clustering across a number of fields. Zhang et al. (2019) proposes splitting the data, clustering the training set, and applying these clusters to the test set as illustrated in Figures 2(a)–(c). They develop a selective inference framework that yields valid p-values for a difference in the mean of a single feature between two clusters in the test set. Our framework avoids the need for sample splitting, and thereby allows inference on the set of clusters obtained from *all* (rather than a subset of) the data.

The rest of the paper is organized as follows. In Section 2, we develop a framework to test for a difference in means after clustering. We apply this framework to compute p-values for hierarchical clustering in Section 3. We describe extensions, simulation results, and applications to real data in Sections 4, 5, and 6. The discussion is in Section 7.

2 Selective inference for clustering

2.1 A test of no difference in means between two clusters

Let $\mathbf{x} \in \mathbb{R}^{n \times q}$ be an arbitrary realization from (1), and let $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ be an arbitrary pair of clusters in $\mathcal{C}(\mathbf{x})$. Since we chose to test $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}$ because $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$, it is natural to

define the p-value as a conditional version of (4),

$$\mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\left(\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 \;\middle|\; \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X})\right). \tag{5}$$

This amounts to asking, "Among all realizations of $\mathbf{X}$ that result in clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$, what proportion have a difference in cluster means at least as large as the difference in cluster means in our observed data set, when in truth $\bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$?" One can show that rejecting $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ when (5) is below α controls the *selective type I error rate* (Fithian et al. 2014) at level α.

Definition 1 (Selective type I error rate for clustering)**.** *For any non-overlapping groups of observations $\mathcal{G}_1, \mathcal{G}_2 \subseteq \{1, 2, \ldots, n\}$, let $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ denote the null hypothesis that $\bar{\mu}_{\mathcal{G}_1} = \bar{\mu}_{\mathcal{G}_2}$. We say that a test of $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ controls the selective type I error rate for clustering if*

$$\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\textit{reject } H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} \textit{ at level } \alpha \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) \leq \alpha, \quad \forall\ 0 \leq \alpha \leq 1. \tag{6}$$

That is, if $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ is true, then the conditional probability of rejecting $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ at level α given that $\mathcal{G}_1$ and $\mathcal{G}_2$ are clusters in $\mathcal{C}(\mathbf{X})$ is bounded by α.

However, (5) is difficult to compute, since the conditional distribution of $\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2$ given $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X})$ involves the unknown nuisance parameters $\boldsymbol{\pi}^{\perp}_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\boldsymbol{\mu}$, where $\boldsymbol{\pi}^{\perp}_{\nu} = \mathbf{I}_n - \frac{\nu\nu^T}{\|\nu\|_2^2}$ projects onto the orthogonal complement of the vector ν, and where

$$[\nu(\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)]_i = \mathbb{1}\{i \in \hat{\mathcal{C}}_1\}/|\hat{\mathcal{C}}_1| - \mathbb{1}\{i \in \hat{\mathcal{C}}_2\}/|\hat{\mathcal{C}}_2|. \tag{7}$$

In other words, it requires knowing aspects of $\boldsymbol{\mu}$ that are not known under the null. Instead, we will define the p-value for $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ in (3) to be

$$p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\Big(\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 \;\Big|\; \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^{\perp}_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\mathbf{x},$$
$$\operatorname{dir}\left(\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\right) = \operatorname{dir}\left(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\right)\Big), \tag{8}$$

where $\text{dir}(w) = \frac{w}{\|w\|_2}\mathbb{1}\{w \neq 0\}$. The following result shows that conditioning on these additional events makes (8) computationally tractable by constraining the randomness in $\mathbf{X}$ to a scalar random variable, while maintaining control of the selective type I error rate.

Theorem 1. *For any realization* $\mathbf{x}$ *from* (1) *and for any non-overlapping groups of observations* $\mathcal{G}_1, \mathcal{G}_2 \subseteq \{1, 2, \ldots, n\}$,

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = 1 - \mathbb{F}\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2; \sigma\sqrt{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}}, \mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})\right), \tag{9}$$

where $p(\cdot;\cdot)$ *is defined in* (8), $\mathbb{F}(t; c, \mathcal{S})$ *denotes the cumulative distribution function of a* $c \cdot \chi_q$ *random variable truncated to the set* $\mathcal{S}$, *and*

$$\mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \left\{\phi \geq 0 : \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1, \mathcal{G}_2)}\mathbf{x} + \left(\frac{\phi}{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}}\right)\nu(\mathcal{G}_1, \mathcal{G}_2)\, dir(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2})^T\right)\right\}. \tag{10}$$

Furthermore, if $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}$ *is true, then* $p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}) \sim \text{Uniform}(0, 1)$, *i.e.*

$$\mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}}\left(p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) = \alpha, \quad \forall\ 0 \leq \alpha \leq 1. \tag{11}$$

That is, rejecting $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}$ *whenever* $p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})$ *is below* α *controls the selective type I error rate (Definition 1) at level* α.

We prove Theorem 1 in Section S1.1 of the supplement. It follows from (9) that to compute the p-value $p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\})$ in (8), it suffices to characterize the one-dimensional set

$$\hat{\mathcal{S}} \equiv \mathcal{S}(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \{\phi \geq 0 : \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\}, \tag{12}$$

where $\mathcal{S}(\mathbf{x};\cdot)$ is defined in (10), and where

$$\mathbf{x}'(\phi) = \boldsymbol{\pi}^{\perp}_{\hat{\nu}}\mathbf{x} + \left(\frac{\phi}{1/|\hat{\mathcal{C}}_1| + 1/|\hat{\mathcal{C}}_2|}\right)\hat{\nu}\ \text{dir}(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2})^T, \quad \hat{\nu} = \nu(\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2), \tag{13}$$

for $\nu(\cdot,\cdot)$ defined in (7).

While the test based on (8) controls the selective type I error rate, the extra conditioning may lead to lower power than a test based on (5)(Lee et al. 2016, Jewell et al. 2019, Mehrizi & Chenouri 2021). However, (8) has a major advantage over (5): Theorem 1 reveals that computing (8) simply requires characterizing $\hat{\mathcal{S}}$ in (12). This is the focus of Section 3.

2.2 Interpreting $\mathbf{x}'(\phi)$ and $\hat{\mathcal{S}}$

Since $\mathbf{x}^T\hat{\nu} = \bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}$, where $\bar{x}_{\hat{\mathcal{C}}_1}$ is defined in (2) and $\hat{\nu}$ is defined in (13), it follows that the ith row of $\mathbf{x}'(\phi)$ in (13) is

$$[\mathbf{x}'(\phi)]_i = \begin{cases} x_i + \left(\frac{|\hat{\mathcal{C}}_2|}{|\hat{\mathcal{C}}_1|+|\hat{\mathcal{C}}_2|}\right)\left(\phi - \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2\right) \text{dir}\left(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\right), & \text{if } i \in \hat{\mathcal{C}}_1, \\ x_i - \left(\frac{|\hat{\mathcal{C}}_1|}{|\hat{\mathcal{C}}_1|+|\hat{\mathcal{C}}_2|}\right)\left(\phi - \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2\right) \text{dir}\left(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\right), & \text{if } i \in \hat{\mathcal{C}}_2, \\ x_i, & \text{if } i \notin \hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2. \end{cases} \tag{14}$$

We can interpret $\mathbf{x}'(\phi)$ as a perturbed version of $\mathbf{x}$, where observations in clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ have been "pulled apart" (if $\phi > \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2$) or "pushed together" (if $0 \leq \phi < \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2$) in the direction of $\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}$. Furthermore, $\hat{\mathcal{S}}$ in (12) describes the set of non-negative ϕ for which applying the clustering algorithm $\mathcal{C}$ to the perturbed data set $\mathbf{x}'(\phi)$ yields $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$. To illustrate this interpretation, we apply average linkage hierarchical clustering to a realization from (1) to obtain three clusters. Figure 3(a)-(c) displays $\mathbf{x} = \mathbf{x}'(\phi)$ for $\phi = \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 = 4$, along with $\mathbf{x}'(\phi)$ for $\phi = 0$ and $\phi = 8$, respectively. The clusters $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$ are shown in blue and orange. In Figure 3(b), since $\phi = 0$, the blue and orange clusters have been "pushed together" so that there is no difference between their empirical means, and average linkage hierarchical clustering no longer estimates these clusters. By

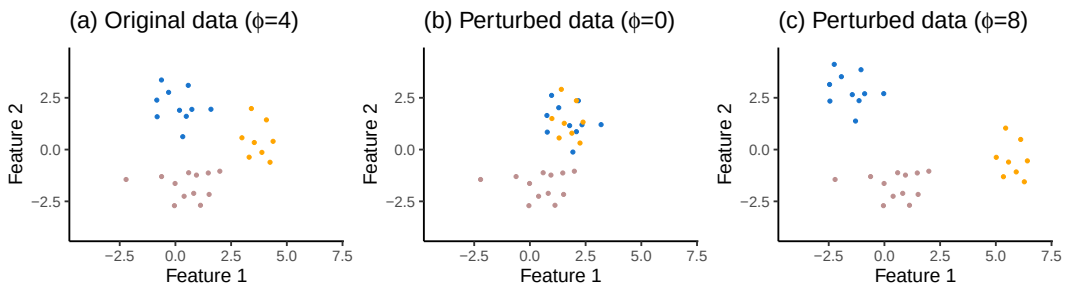

Figure 3: The observations belonging to $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$ are displayed in blue and orange for: (a) the original data set $\mathbf{x} = \mathbf{x}'(\phi)$ with $\phi = \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 = 4$, (b) a perturbed data set $\mathbf{x}'(\phi)$ with $\phi = 0$, and (c) a perturbed data set $\mathbf{x}'(\phi)$ with $\phi = 8$.

contrast, in Figure 3(c), the blue and orange clusters have been "pulled apart", and average linkage hierarchical clustering still estimates these clusters. In this example, $\hat{\mathcal{S}} = [2.8, \infty)$.

3 Computing $\hat{\mathcal{S}}$ for hierarchical clustering

We now consider computing $\hat{\mathcal{S}}$ defined in (12) for clusters defined via hierarchical clustering. After reviewing hierarchical clustering (Section 3.1), we explicitly characterize $\hat{\mathcal{S}}$ (Section 3.2), and then show how specific properties, such as the dissimilarity and linkage used, lead to substantial computational savings in computing $\hat{\mathcal{S}}$ (Sections 3.3–3.4).

3.1 A brief review of agglomerative hierarchical clustering

Agglomerative hierarchical clustering produces a sequence of clusterings. The first clustering, $\mathcal{C}^{(1)}(\mathbf{x})$, contains n clusters, each with a single observation. The $(t+1)$th clustering, $\mathcal{C}^{(t+1)}(\mathbf{x})$, is created by merging the two most similar (or least dissimilar) clusters in the

tth clustering, for $t = 1, \ldots, n-1$. Details are provided in Algorithm 1.

Algorithm 1 Agglomerative hierarchical clustering of a data set $\mathbf{x}$

Let $\mathcal{C}^{(1)}(\mathbf{x}) = \{\{1\}, \{2\}, \ldots, \{n\}\}$. For $t = 1, \ldots, n-1$:

1. Define $\left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\} = \underset{\mathcal{G},\mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \mathcal{G} \neq \mathcal{G}'}{\arg\min} d(\mathcal{G}, \mathcal{G}'; \mathbf{x})$. (We assume throughout this paper that the minimizer is unique.)

2. Merge $\mathcal{W}_1^{(t)}(\mathbf{x})$ and $\mathcal{W}_2^{(t)}(\mathbf{x})$ at the height of $d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right)$ in the dendrogram, and let $\mathcal{C}^{(t+1)}(\mathbf{x}) = \mathcal{C}^{(t)}(\mathbf{x}) \cup \left\{\mathcal{W}_1^{(t)}(\mathbf{x}) \cup \mathcal{W}_2^{(t)}(\mathbf{x})\right\} \setminus \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}$.

Algorithm 1 involves a function $d(\mathcal{G}, \mathcal{G}'; \mathbf{x})$, which quantifies the dissimilarity between two groups of observations. We assume throughout this paper that the dissimilarity between the ith and i'th observations, $d(\{i\}, \{i'\}; \mathbf{x})$, depends on the data through $x_i - x_i'$ only. For example, we could define $d(\{i\}, \{i'\}; \mathbf{x}) = \|x_i - x_{i'}\|_2^2$. When $\max\{|\mathcal{G}|, |\mathcal{G}'|\} > 1$, then we extend the pairwise similarity to the dissimilarity between groups of obervations using the notion of *linakge*, to be discussed further in Section 3.3.

3.2 An explicit characterization of $\hat{\mathcal{S}}$ for hierarchical clustering

We saw in Sections 2.1–2.2 that to compute the p-value $p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\})$ defined in (8), we must characterize the set $\hat{\mathcal{S}} = \{\phi \geq 0 : \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\}$ in (12), where $\mathbf{x}'(\phi)$ in (13) is a perturbed version of $\mathbf{x}$ in which observations in the clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ have been "pulled together" or "pushed apart". We do so now for hierarchical clustering.

Lemma 1. *Suppose that $\mathcal{C} = \mathcal{C}^{(n-K+1)}$, i.e. we perform hierarchical clustering to obtain*

K clusters. Then,

$$d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right) = d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right), \ \forall\ \phi \geq 0, \forall\ t = 1, \ldots, n-K, \quad (15)$$

where $\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right)$ is the "winning pair" of clusters that merged at the t^{th} step of the hierarchical clustering algorithm applied to $\mathbf{x}$. Furthermore, for any $\phi \geq 0$,

$$\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi)) \quad \textit{if and only if} \quad \mathcal{C}^{(t)}(\mathbf{x}'(\phi)) = \mathcal{C}^{(t)}(\mathbf{x})\ \forall\ t = 1, \ldots, n-K+1. \quad (16)$$

We prove Lemma 1 in Section S1.2 of the supplement. The right-hand side of (16) says that the same merges occur in the first $n-K$ steps of the hierarchical clustering algorithm applied to $\mathbf{x}'(\phi)$ and $\mathbf{x}$. To characterize the set of merges that occur in $\mathbf{x}$, consider the set of all "losing pairs", i.e. all cluster pairs that co-exist but are not the "winning pair" in the first $n-K$ steps:

$$\mathcal{L}(\mathbf{x}) = \bigcup_{t=1}^{n-K} \left\{ \{\mathcal{G}, \mathcal{G}'\} : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \mathcal{G} \neq \mathcal{G}', \{\mathcal{G}, \mathcal{G}'\} \neq \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\} \right\}. \quad (17)$$

Each pair $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$ has a "lifetime" that starts at the step where both have been created, $l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \equiv \min\left\{1 \leq t \leq n-K : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\}\right\}$, and ends at step $u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \equiv \max\left\{1 \leq t \leq n-K : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\}\right\}$. By construction, each pair $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$ is never the winning pair at any point in its lifetime, i.e. $d(\mathcal{G}, \mathcal{G}'; \mathbf{x}) > d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right)$ for all $l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t \leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$. Therefore, $\mathbf{x}'(\phi)$ preserves the merges that occur in the first $n-K$ steps in $\mathbf{x}$ if and only if $d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) > d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right)$ for all $l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t \leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ and for all $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$. Furthermore, (15) says that $d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right) = d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right)$ for all $\phi \geq 0$ and $1 \leq t \leq n-K$. This leads to the following result.

Theorem 2. *Suppose that $\mathcal{C} = \mathcal{C}^{(n-K+1)}$, i.e. we perform hierarchical clustering to obtain K clusters. Then, for $\hat{\mathcal{S}}$ defined in* (12)*,*

$$\hat{\mathcal{S}} = \bigcap_{\{\mathcal{G},\mathcal{G}'\}\in\mathcal{L}(\mathbf{x})} \left\{ \phi \geq 0 : d(\mathcal{G},\mathcal{G}';\mathbf{x}'(\phi)) > \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\leq t\leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) \right\}, \quad (18)$$

where $\left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}$ is the pair of clusters that merged at the t^{th} step of the hierarchical clustering algorithm applied to $\mathbf{x}$, $\mathcal{L}(\mathbf{x})$ is defined in (17) *to be the set of "losing pairs" of clusters in $\mathbf{x}$, and $[l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}), u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})]$ is the lifetime of such a pair of clusters in $\mathbf{x}$. Furthermore,* (18) *is the intersection of $\mathcal{O}(n^2)$ sets.*

We prove Theorem 2 in Section S1.3 of the supplement. Theorem 2 expresses $\hat{\mathcal{S}}$ in (12) as the intersection of $\mathcal{O}(n^2)$ sets of the form $\{\phi \geq 0 : d(\mathcal{G},\mathcal{G}';\mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}$, where

$$h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \equiv \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\leq t\leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) \quad (19)$$

is the maximum merge height in the dendrogram of $\mathbf{x}$ over the lifetime of $\{\mathcal{G},\mathcal{G}'\}$. The next subsection is devoted to understanding when and how these sets can be efficiently computed. In particular, by specializing to squared Euclidean distance and a certain class of linkages, we will show that each of these sets is defined by a single quadratic inequality, and that the coefficients of all of these quadratic inequalities can be efficiently computed.

3.3 Squared Euclidean distance and "linear update" linkages

Consider hierarchical clustering with squared Euclidean distance and a linkage that satisfies a linear Lance-Williams update (Lance & Williams 1967) of the form

$$d(\mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{G}_3; \mathbf{x}) = \alpha_1 d(\mathcal{G}_1, \mathcal{G}_3; \mathbf{x}) + \alpha_2 d(\mathcal{G}_2, \mathcal{G}_3; \mathbf{x}) + \beta d(\mathcal{G}_1, \mathcal{G}_2; \mathbf{x}). \quad (20)$$

	Average	Weighted	Ward	Centroid	Median	Single	Complete
Satisfies (20)	✓	✓	✓	✓	✓	✗	✗
Does not produce inversions	✓	✓	✓	✗	✗	✓	✓

Table 1: Properties of seven linkages in the case of squared Euclidean distance (Murtagh & Contreras 2012). Table 1 of Murtagh & Contreras (2012) specifies α_1, α_2, and β in (20).

This includes average, weighted, Ward, centroid, and median linkage (Table 1).

We have seen in Section 3.2 that to evaluate (18), we must evaluate $\mathcal{O}(n^2)$ sets of the form $\{\phi \geq 0 : d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}$ with $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$, where $\mathcal{L}(\mathbf{x})$ in (17) is the set of losing cluster pairs in $\mathbf{x}$. We now present results needed to characterize these sets.

Lemma 2. *Suppose that we define* $d(\{i\}, \{i'\}; \mathbf{x}) = \|x_i - x_{i'}\|_2^2$. *Then, for all* $i \neq i'$, $d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) = a_{ii'}\phi^2 + b_{ii'}\phi + c_{ii'}$, *where for* $\hat{\nu}$ *defined in* (13), $a_{ii'} = \left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)^2$, $b_{ii'} = 2\left(\left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)\langle dir(\mathbf{x}^T\hat{\nu}), x_i - x_{i'}\rangle - a_{ii'}\|\mathbf{x}^T\hat{\nu}\|_2\right)$, *and* $c_{ii'} = \left\|x_i - x_{i'} - \left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)(\mathbf{x}^T\hat{\nu})\right\|_2^2$.

Lemma 2 follows directly from the definition of $\mathbf{x}'(\phi)$ in (13), and does not require (20) to hold. Next, we specialize to squared Euclidean distance *and* linkages satisfying (20), and characterize $d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi))$, the dissimilarity between pairs of clusters in $\mathbf{x}'(\phi)$. The following result follows immediately from Lemma 2 and the fact that linear combinations of quadratic functions of ϕ are also quadratic functions of ϕ.

Proposition 1. *Suppose we define* $d(\{i\}, \{i'\}; \mathbf{x}) = \|x_i - x_{i'}\|_2^2$, *and we define* $d(\mathcal{G}, \mathcal{G}'; \mathbf{x})$ *using a linkage that satisfies* (20). *Then,* $d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi))$ *is a quadratic function of* ϕ *for all* $\mathcal{G} \neq \mathcal{G}'$. *Furthermore, given the coefficients corresponding to the quadratic functions* $d(\mathcal{G}_1, \mathcal{G}_3; \mathbf{x}'(\phi))$, $d(\mathcal{G}_2, \mathcal{G}_3; \mathbf{x}'(\phi))$, *and* $d(\mathcal{G}_1, \mathcal{G}_2; \mathbf{x}'(\phi))$, *we can compute the coefficients cor-*

responding to the quadratic function $d(\mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{G}_3; \mathbf{x}'(\phi))$ *in* $\mathcal{O}(1)$ *time, using* (20).

Lastly, we characterize the cost of computing $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ in (19). Naively, computing $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ could require $\mathcal{O}(n)$ operations. However, if the dendrogram of $\mathbf{x}$ has no inversions below the $(n-K)$th merge, i.e. if $d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) < d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{W}_2^{(t+1)}(\mathbf{x}); \mathbf{x}\right)$ for all $t < n-K$, then $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = d\left(\mathcal{W}_1^{\left(u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\right)}(\mathbf{x}), \mathcal{W}_2^{\left(u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\right)}(\mathbf{x}); \mathbf{x}\right)$. More generally, $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \max_{t \in \mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \cup \{u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right)$, where $\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \Big\{ t : l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t < u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}), d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) > d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{W}_2^{(t+1)}(\mathbf{x}); \mathbf{x}\right) \Big\}$ is the set of steps where inversions occur in the dendrogram of $\mathbf{x}$ during the lifetime of the cluster pair $\{\mathcal{G}, \mathcal{G}'\}$. This leads to the following result.

Proposition 2. *For any* $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$*, given its lifetime* $l_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ *and* $u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$*, and given* $\mathcal{M}(\mathbf{x}) = \Big\{ 1 \leq t \leq n-K : d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) < d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{W}_2^{(t+1)}(\mathbf{x}); \mathbf{x}\right) \Big\}$*, i.e. the set of steps where inversions occur in the dendrogram of* $\mathbf{x}$ *below the* $(n-K)$*th merge, we can compute* $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ *in* $\mathcal{O}(|\mathcal{M}(\mathbf{x})| + 1)$ *time.*

We prove Proposition 2 in Section S1.4 of the supplement. Proposition 2 does not require defining $d(\{i\}, \{i'\}; \mathbf{x}) = \|x_i - x_{i'}\|_2^2$ and does not require (20) to hold. We now characterize the cost of computing $\hat{\mathcal{S}}$ defined in (12), in the case of squared Euclidean distance and linkages that satisfy (20).

Proposition 3. *Suppose we define* $d(\{i\}, \{i'\}; \mathbf{x}) = \|x_i - x_{i'}\|_2^2$*, and we define* $d(\mathcal{G}, \mathcal{G}'; \mathbf{x})$ *using a linkage that satisfies* (20)*. Then, we can compute* $\hat{\mathcal{S}}$ *defined in* (12) *in* $\mathcal{O}\Big((|\mathcal{M}(\mathbf{x})| + \log(n))n^2\Big)$ *time.*

A detailed algorithm for computing $\hat{\mathcal{S}}$ is provided in Section S2 of the supplement. If the linkage we use does not produce inversions, then $|\mathcal{M}(\mathbf{x})| = 0$ for all $\mathbf{x}$. Average, weighted,

and Ward linkage satisfy (20) and are guaranteed not to produce inversions (Table 1), thus $\hat{\mathcal{S}}$ can be computed in $\mathcal{O}(n^2 \log(n))$ time.

3.4 Squared Euclidean distance and single linkage

Single linkage does not satisfy (20) (Table 1), and the inequality that defines the set $\{\phi \geq 0 : d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}$ is not quadratic in ϕ for $|\mathcal{G}| > 1$ or $|\mathcal{G}'| > 1$. Consequently, in the case of single linkage with squared Euclidean distance, we cannot efficiently evaluate the expression of $\hat{\mathcal{S}}$ in (18) using the approach outlined in Section 3.3.

Fortunately, the definition of single linkage leads to an even simpler expression of $\hat{\mathcal{S}}$ than (18). Single linkage defines $d(\mathcal{G}, \mathcal{G}'; \mathbf{x}) = \min_{i \in \mathcal{G}, i' \in \mathcal{G}'} d(\{i\}, \{i'\}; \mathbf{x})$. Applying this definition to (18) yields $\hat{\mathcal{S}} = \bigcap_{\{\mathcal{G},\mathcal{G}'\} \in \mathcal{L}(\mathbf{x})} \bigcap_{i \in \mathcal{G}} \bigcap_{i' \in \mathcal{G}'} \{\phi \geq 0 : d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}$, where $\mathcal{L}(\mathbf{x})$ in (17) is the set of losing cluster pairs in $\mathbf{x}$. Therefore, in the case of single linkage, $\hat{\mathcal{S}} = \bigcap_{\{i,i'\} \in \mathcal{L}'(\mathbf{x})} \hat{\mathcal{S}}_{i,i'}$, where $\mathcal{L}'(\mathbf{x}) = \{\{i, i'\} : i \in \mathcal{G}, i' \in \mathcal{G}', \{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})\}$ and $\hat{\mathcal{S}}_{i,i'} = \left\{\phi \geq 0 : d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > \max_{\{\mathcal{G},\mathcal{G}'\} \in \mathcal{L}(\mathbf{x}) : i \in \mathcal{G}, i' \in \mathcal{G}'} h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\right\}$. The following result characterizes the sets of the form $\hat{\mathcal{S}}_{i,i'}$.

Proposition 4. *Suppose that $\mathcal{C} = \mathcal{C}^{(n-K+1)}$, i.e. we perform hierarchical clustering to obtain K clusters. Let $i \neq i'$. If $i, i' \in \hat{\mathcal{C}}_1$ or $i, i' \in \hat{\mathcal{C}}_2$ or $i, i' \notin \hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2$, then $\hat{\mathcal{S}}_{i,i'} = [0, \infty)$. Otherwise, $\hat{\mathcal{S}}_{i,i'} = \left\{\phi \geq 0 : d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > d\left(\mathcal{W}_1^{(n-K)}(\mathbf{x}), \mathcal{W}_2^{(n-K)}(\mathbf{x}); \mathbf{x}\right)\right\}$.*

We prove Proposition 4 in Section S1.5 of the supplement. Therefore,

$$\begin{aligned}\hat{\mathcal{S}} &= \bigcap_{\{i,i'\} \in \mathcal{L}'(\mathbf{x})} \hat{\mathcal{S}}_{i,i'} \\ &= \bigcap_{\{i,i'\} \in \mathcal{I}(\mathbf{x})} \left\{\phi \geq 0 : \ d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > d\left(\mathcal{W}_1^{(n-K)}(\mathbf{x}), \mathcal{W}_2^{(n-K)}(\mathbf{x}); \mathbf{x}\right)\right\}, \qquad (21)\end{aligned}$$

where $\mathcal{I}(\mathbf{x}) = \mathcal{L}'(\mathbf{x}) \backslash \left[\left\{ \{i, i'\} : i, i' \in \hat{\mathcal{C}}_1 \right\} \cup \left\{ \{i, i'\} : i, i' \in \hat{\mathcal{C}}_2 \right\} \cup \left\{ \{i, i'\} : i, i' \notin \hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2 \right\} \right]$. Recall from Lemma 2 that in the case of squared Euclidean distance, $d(\{i\}, \{i'\}; \mathbf{x}'(\phi))$ is a quadratic function of ϕ whose coefficients can be computed in $\mathcal{O}(1)$ time. Furthermore, $\mathcal{O}(n^2)$ sets are intersected in (21). Therefore, we can evaluate (21) in $\mathcal{O}(n^2 \log(n))$ time by solving $\mathcal{O}(n^2)$ quadratic inequalities and intersecting their solution sets.

4 Extensions

4.1 Monte Carlo approximation to the p-value

We may be interested in computing the p-value $p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\})$ defined in (8) for clustering methods where $\hat{\mathcal{S}}$ in (12) cannot be efficiently computed (e.g. complete linkage hierarchical clustering or non-hierarchical clustering methods). Thus, we develop a Monte Carlo approximation to the p-value that does not require us to compute $\hat{\mathcal{S}}$. Recalling from (12) that $\hat{\mathcal{S}} = \mathcal{S}(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\})$, it follows from Theorem 1 that

$$p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \mathbb{E}\left[\mathbb{1}\left\{\phi \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right\}\right] / \mathbb{E}\left[\mathbb{1}\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right\}\right], \quad (22)$$

for $\phi \sim \sigma\left(\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}}\right) \cdot \chi_q$, and for $\mathbf{x}'(\phi)$ defined in (13). Thus, we could naively sample $\phi_1, \ldots, \phi_N \overset{i.i.d.}{\sim} \sigma\left(\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}}\right) \cdot \chi_q$, and approximate the expectations in (22) with averages over the samples. However, when $\|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2$ is in the tail of the $\sigma\left(\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}}\right) \cdot \chi_q$ distribution, the naive approximation of the expectations in (22) is poor for finite values of N. Instead, we use an importance sampling approach, as in Yang et al. (2016). We sample $\omega_1, \ldots, \omega_N \overset{i.i.d}{\sim} N\left(\|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2, \sigma^2\left(\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}\right)\right)$, and approximate (22) with

$$p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) \approx \left(\sum_{i=1}^{N} \pi_i \mathbb{1}\left\{\omega_i \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\omega_i))\right\}\right) / \left(\sum_{i=1}^{N} \pi_i \mathbb{1}\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\omega_i))\right\}\right),$$

for $\pi_i = \frac{f_1(\omega_i)}{f_2(\omega_i)}$, where f_1 is the density of a $\sigma\left(\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|}+\frac{1}{|\hat{\mathcal{C}}_2|}}\right)\cdot\chi_q$ random variable, and f_2 is the density of a $N\left(\|\bar{x}_{\hat{\mathcal{C}}_1}-\bar{x}_{\hat{\mathcal{C}}_2}\|_2, \sigma^2\left(\frac{1}{|\hat{\mathcal{C}}_1|}+\frac{1}{|\hat{\mathcal{C}}_2|}\right)\right)$ random variable.

4.2 Non-spherical covariance matrix

The selective inference framework in Section 2 assumes that $\mathbf{x}$ is a realization from (1), so that $\mathrm{Cov}(X_i) = \sigma^2\mathbf{I}_q$. In this subsection, we instead assume that $\mathbf{x}$ is a realization from

$$\mathbf{X} \sim \mathcal{MN}_{n\times q}(\boldsymbol{\mu}, \mathbf{I}_n, \boldsymbol{\Sigma}), \tag{23}$$

where $\boldsymbol{\Sigma}$ is a known $q \times q$ positive definite matrix. We define the p-value of interest as

$$p_\Sigma(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\Bigg(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\bar{X}_{\hat{\mathcal{C}}_1}-\bar{X}_{\hat{\mathcal{C}}_2})\|_2^2 \geq \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\bar{x}_{\hat{\mathcal{C}}_1}-\bar{x}_{\hat{\mathcal{C}}_2})\|_2^2 \;\Big|\; \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^\perp_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\mathbf{X} = \boldsymbol{\pi}^\perp_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\mathbf{x},$$
$$\mathrm{dir}\left((\bar{X}_{\hat{\mathcal{C}}_1}-\bar{X}_{\hat{\mathcal{C}}_2})^T\boldsymbol{\Sigma}^{-1}(\bar{X}_{\hat{\mathcal{C}}_1}-\bar{X}_{\hat{\mathcal{C}}_2})\right) = \mathrm{dir}\left((\bar{x}_{\hat{\mathcal{C}}_1}-\bar{x}_{\hat{\mathcal{C}}_2})^T\boldsymbol{\Sigma}^{-1}(\bar{x}_{\hat{\mathcal{C}}_1}-\bar{x}_{\hat{\mathcal{C}}_2})\right)\Bigg).$$

Theorem 3. *For any realization* $\mathbf{x}$ *from* (23)*, and for any non-overlapping groups of observations* $\mathcal{G}_1, \mathcal{G}_2 \subseteq \{1, 2, \ldots, n\}$*,*

$$p_\Sigma(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = 1 - \mathbb{F}\left(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{x}^T\nu(\mathcal{G}_1,\mathcal{G}_2)\|_2; \|\nu(\mathcal{G}_1,\mathcal{G}_2)\|_2, \mathcal{S}_\Sigma(\mathbf{x}; \{\mathcal{G}_1,\mathcal{G}_2\})\right), \tag{24}$$

where $\mathcal{S}_\Sigma(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \left\{\phi \geq 0 : \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\boldsymbol{\pi}^\perp_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{x} + \phi\left(\frac{\nu(\mathcal{G}_1,\mathcal{G}_2)}{\|\nu(\mathcal{G}_1,\mathcal{G}_2)\|_2^2}\right) dir(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{x}^T\nu(\mathcal{G}_1,\mathcal{G}_2))^T\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\right\}$*.*
Furthermore, if $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ *is true, then* $p_\Sigma(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}) \sim \mathrm{Uniform}(0, 1)$*, i.e.*

$$\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}(p_\Sigma(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})) = \alpha, \quad \textit{for all } 0 \leq \alpha \leq 1.$$

That is, rejecting $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ *whenever* $p_\Sigma(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\})$ *is below* α *controls the selective type I error rate (Definition 1).*

We omit the proof of Theorem 3, since it closely follows that of Theorem 1. In the case of hierachical clustering with squared Euclidean distance, we can adapt Sections 3.3–3.4 to compute $\mathcal{S}_\Sigma(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\})$ by replacing $a_{ii'}$, $b_{ii'}$, and $c_{ii'}$ in Lemma 2 with $\tilde{a}_{ii'} = \left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)^2 \left(\frac{\|\mathbf{x}^T\hat{\nu}\|_2}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}^T\hat{\nu}\|_2}\right)^2$, $\tilde{b}_{ii'} = 2\Big(\left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)\left(\frac{\|\mathbf{x}^T\hat{\nu}\|_2}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}^T\hat{\nu}\|_2}\right)\langle \mathrm{dir}(\mathbf{x}^T\hat{\nu}), x_i - x_{i'}\rangle - \tilde{a}_{ii'}\|\mathbf{\Sigma}^{-1/2}\mathbf{x}^T\hat{\nu}\|_2\Big)$, and $\tilde{c}_{ii'} = \left\| x_i - x_{i'} - \left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)(\mathbf{x}^T\hat{\nu})\right\|_2$.

4.3 Unknown variance

The selective inference framework in Section 2 assumes that σ in model (1) is known. If σ is unknown, then we can plug an estimate of σ into (9), as follows:

$$\hat{p}\left(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}\right) = 1 - \mathbb{F}\left(\|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2; \hat{\sigma}(\mathbf{x})\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}},\ \hat{\mathcal{S}}\right). \tag{25}$$

The following result says that if we use an asymptotic over-estimate of σ in (25), then we can asymptotically control the selective type I error rate (Definition 1).

Theorem 4. *For $n = 1, 2, \ldots$, suppose that $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times q}(\boldsymbol{\mu}^{(n)}, \mathbf{I}_n, \sigma^2\mathbf{I}_q)$. Let $\mathbf{x}^{(n)}$ be a realization of $\mathbf{X}^{(n)}$, and $\hat{\mathcal{C}}_1^{(n)}$ and $\hat{\mathcal{C}}_2^{(n)}$ be a pair of clusters estimated from $\mathbf{x}^{(n)}$. Suppose that $\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\}}}\left(\hat{\sigma}(\mathbf{X}^{(n)}) \geq \sigma \;\middle|\; \hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)} \in \mathcal{C}(\mathbf{X}^{(n)})\right) = 1$. Then, for any $\alpha \in [0, 1]$, we have that $\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\}}}\left(\hat{p}(\mathbf{X}^{(n)}; \{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\}) \leq \alpha \;\middle|\; \hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)} \in \mathcal{C}(\mathbf{X}^{(n)})\right) \leq \alpha$.*

We prove Theorem 4 in Section S1.6 of the supplement. In Section S3, we provide an estimator of σ that satisfies the conditions in Theorem 4.

4.4 Consequences of selective type I error rate control

This paper focuses on developing tests of $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} : \bar{\mu}_{\mathcal{G}_1} = \bar{\mu}_{\mathcal{G}_2}$ that control the selective type I error rate (Definition 1). However, it is cumbersome to demonstrate selective type

I error rate control via simulation, as $\mathbb{P}(\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}))$ can be small when $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ holds.

Nevertheless, two related properties can be demonstrated via simulation. Let $\mathcal{H}_0$ denote the set of null hypotheses of the form $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ that are true. The following property holds.

Proposition 5. *When $K = 2$, i.e. the clustering algorithm $\mathcal{C}(\cdot)$ estimates two clusters,*

$$\mathbb{P}\left(p(\mathbf{X}; \mathcal{C}(\mathbf{X})) \leq \alpha \;\middle|\; H_0^{\mathcal{C}(\mathbf{X})} \in \mathcal{H}_0\right) = \alpha, \quad \textit{for all } 0 \leq \alpha \leq 1, \tag{26}$$

where $p(\cdot;\cdot)$ is defined in (8)*. That is, if the two estimated clusters have the same mean, then the probability of falsely declaring otherwise is equal to α.*

We prove Proposition 5 in Section S1.7 of the supplement. What if $K > 2$? Then, given a data set $\mathbf{x}$, we can randomly select (independently of $\mathbf{x}$) a single pair of estimated clusters $\mathcal{G}_1(\mathbf{x}), \mathcal{G}_2(\mathbf{x}) \in \mathcal{C}(\mathbf{x})$. This leads to the following property.

Proposition 6. *Suppose that $K > 2$, i.e. the clustering algorithm $\mathcal{C}(\cdot)$ estimates three or more clusters, and let $\mathcal{G}_1(\mathbf{x}), \mathcal{G}_2(\mathbf{x}) \in \mathcal{C}(\mathbf{x})$ denote a randomly selected pair. If $\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}$ and $\mathbf{X}$ are conditionally independent given $\mathcal{C}(\mathbf{X})$, then for $p(\cdot;\cdot)$ defined in* (8)*,*

$$\mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}) \leq \alpha \;\middle|\; H_0^{\{\mathcal{G}_1(\mathbf{X}),\mathcal{G}_2(\mathbf{X})\}} \in \mathcal{H}_0\right) = \alpha, \quad \textit{for all } 0 \leq \alpha \leq 1. \tag{27}$$

We prove Proposition 6 in Section S1.7 of the supplement. Recall that in Figure 1(c), we simulated data with $\boldsymbol{\mu} = \mathbf{0}_{n\times q}$, so the conditioning event in (27) holds with probability 1. Thus, (27) specializes to $\mathbb{P}(p(\mathbf{X}; \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}) \leq \alpha) = \alpha$, i.e. $p(\mathbf{X}; \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}) \sim \text{Uniform}(0, 1)$. This property is illustrated in Figure 1(c).

5 Simulation results

Throughout this section, we use the efficient implementation of hierarchical clustering in the `fastcluster` package (Müllner et al. 2013) in `R`.

5.1 Uniform p-values under a global null

We generate data from (1) with $\boldsymbol{\mu} = \mathbf{0}_{n\times q}$, so that $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ holds for all pairs of estimated clusters. We simulate 2000 data sets for $n = 150$, $\sigma \in \{1, 2, 10\}$, and $q \in \{2, 10, 100\}$. For each data set, we use average, centroid, single, and complete linkage hierarchical clustering to estimate three clusters, and then test $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ for a randomly-chosen pair of clusters. We compute the p-value defined in (8) as described in Section 3 for average, centroid, and single linkage. For complete linkage, we approximate the p-value as described in Section 4.1 with $N = 2000$. Figure 4 displays QQ plots of the empirical distribution of the p-values against the Uniform$(0, 1)$ distribution. The p-values have a Uniform$(0, 1)$ distribution because our proposed test satisfies (27) and because $\boldsymbol{\mu} = \mathbf{0}_{n\times q}$; see the end of Section 4.4 for a detailed discussion. In Section S4.1 of the supplement, we show that plugging in an over-estimate σ as in (25) yields p-values that are stochastically larger than the Uniform$(0, 1)$ distribution.

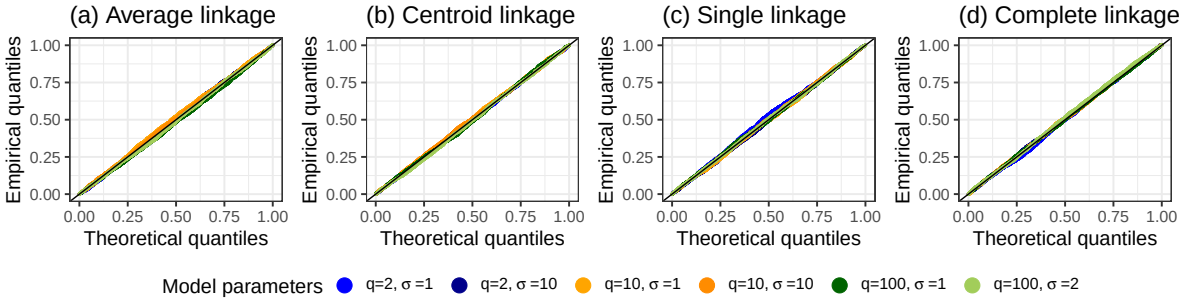

Figure 4: For 2000 draws from (1) with $\boldsymbol{\mu} = \mathbf{0}_{n\times q}$, $n = 150$, $q \in \{2, 10, 100\}$, and $\sigma \in \{1, 2, 10\}$, QQ-plots of the p-values obtained from the test proposed in Section 2.1, using (a) average linkage, (b) centroid linkage, (c) single linkage, and (d) complete linkage.

5.2 Conditional power and detection probability

We generate data from (1) with $n = 30$, and three equidistant clusters,

$$\mu_1 = \cdots = \mu_{\frac{n}{3}} = \begin{bmatrix} -\delta/2 \\ 0_{q-1} \end{bmatrix}, \mu_{\frac{n}{3}+1} = \cdots = \mu_{\frac{2n}{3}} = \begin{bmatrix} 0_{q-1} \\ \sqrt{3}\delta/2 \end{bmatrix}, \mu_{\frac{2n}{3}+1} = \cdots = \mu_n = \begin{bmatrix} \delta/2 \\ 0_{q-1} \end{bmatrix}, \quad (28)$$

for $\delta > 0$. For each simulated data set, we use average, centroid, single, and complete linkage hierarchical clustering to estimate three clusters, and then test $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}$ for a randomly-chosen pair of clusters, with significance level $\alpha = 0.05$. We simulate $300,000$ data sets for $\sigma = 1$, $q = 10$, and seven evenly-spaced values of $\delta \in [4, 7]$. We define the *conditional power* to be the conditional probability of rejecting $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}$, given that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ correspond to true clusters. We estimate the conditional power by

$$\frac{\#\text{ data sets where we reject } H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}\text{, and } \hat{\mathcal{C}}_1 \text{ and } \hat{\mathcal{C}}_2 \text{ are true clusters}}{\#\text{ data sets where } \hat{\mathcal{C}}_1 \text{ and } \hat{\mathcal{C}}_2 \text{ are true clusters}}. \quad (29)$$

Since (29) conditions on the event that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are two true clusters, we are also interested in how often this event occurs. We therefore consider the *detection probability*, the probability that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are true clusters, which we estimate by

$$\frac{\#\text{ data sets where } \hat{\mathcal{C}}_1 \text{ and } \hat{\mathcal{C}}_2 \text{ are true clusters}}{300,000}. \quad (30)$$

Figure 5 displays the conditional power and detection probability as a function of the distance between the true clusters (δ). For all four linkages, the conditional power and detection probability increase as the distance between the true clusters (δ) increases. Average and complete linkage have the highest conditional power, and single linkage has the lowest conditional power. Average, centroid, and complete linkage have substantially higher detection probabilities than single linkage.

We consider an alternative definition of power that does not condition on having correctly estimated the true clusters in Section S4.2 of the supplement.

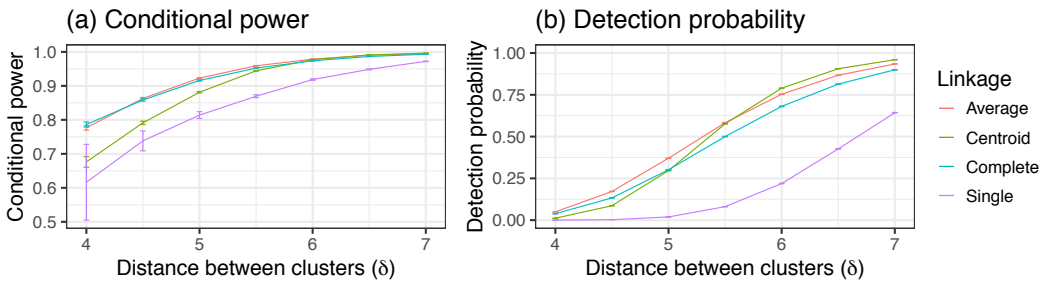

Figure 5: For the simulation study described in Section 5.2, (a) conditional power, defined in (29), of the test proposed in Section 2 versus the difference in means between the true clusters (δ), and (b) detection probability, defined in (30), versus δ.

6 Data applications

6.1 Palmer penguins (Horst et al. 2020)

In this section, we analyze the `penguins` data set from the `palmerpenguins` package in `R` (Horst et al. 2020). We estimate σ with $\hat{\sigma}(\mathbf{x}) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{q}(x_i - \bar{x}_j)^2/(nq-q)}$ for $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, calculated on the bill length and flipper length of 58 female penguins observed in the year 2009. We then consider the 107 female penguins observed in the years 2007–2008 with complete data on species, bill length, and flipper length. Figure 6(a) displays the dendrogram obtained from applying average linkage hierarchical clustering with squared Euclidean distance to the penguins' bill length and flipper length, cut to yield five clusters, and Figure 6(b) displays the data.

We test $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ for all pairs of clusters that contain more than one observation, using the test proposed in Section 2.1, and using the Wald test described in (4). (The latter does not account for the fact that the clusters were estimated from the data, and does

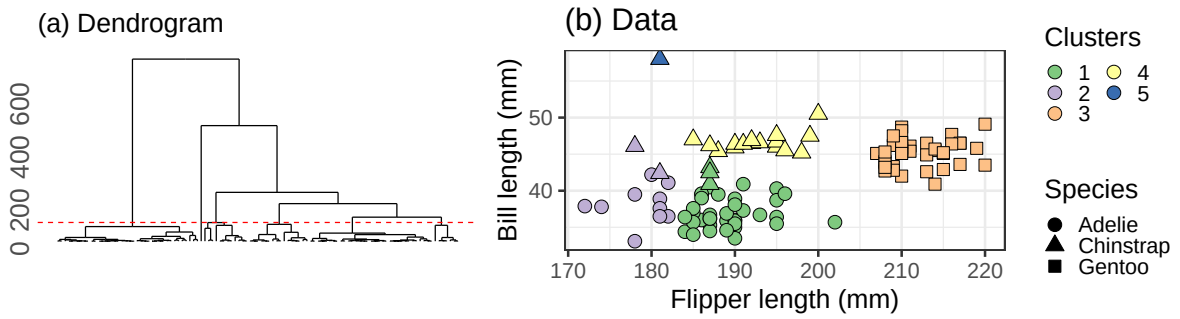

Figure 6: (a) The average-linkage hierarchical clustering dendrogram and (b) the bill lengths and flipper lengths, as well as the true specieis labels and estimated clusters, for the Palmer penguins data described in Section 6.1.

not control the selective type I error rate.) Results are in Table 2. The Wald p-values are small, even when testing for a difference in means between a single species (Clusters 1 and 2). Our proposed test yields a large p-value when testing for a difference in means between a single species (Clusters 1 and 2), and small p-values when the clusters correspond to different species (Clusters 1 and 3, and Clusters 3 and 4). The p-values from our proposed test are large for the remaining three pairs of clusters containing different species, even though visual inspection suggests a large difference between these two clusters. This is because the test statistic is close to the left boundary of $\hat{\mathcal{S}}$ defined in (12), which leads to low power: see the discussion of Figure S9 in Section S4.2 of the supplement.

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

Single-cell RNA sequencing data quantifies the gene expression levels of individual cells. Biologists often cluster the cells to identify putative cell types, and then test for differential

Cluster pairs	$(1,2)$	$(1,3)$	$(1,4)$	$(2,3)$	$(2,4)$	$(3,4)$
Test statistic	10.1	25.0	10.1	33.8	17.1	18.9
Our p-value	0.591	1.70×10^{-14}	0.714	0.070	0.291	2.10×10^{-6}
Wald p-value	0.00383	$< 10^{-307}$	0.00101	$< 10^{-307}$	4.29×10^{-5}	1.58×10^{-11}

Table 2: Results from applying the test of $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_k} = \bar{\mu}_{\hat{\mathcal{C}}_{k'}}$ proposed in Section 2 and the Wald test defined in (4) to the Palmer penguins data set, displayed in Figure 6(b).

gene expression between the clusters, without properly accounting for the fact that the clusters were estimated from the data (Lähnemann et al. 2020). Zheng et al. (2017) classified peripheral blood mononuclear cells prior to sequencing. We will use this data set to demonstrate that testing for differential gene expression after clustering with our proposed selective inference framework yields reasonable results.

6.2.1 Data and pre-processing

We subset the data to the memory T cells, B cells, and monocytes. In line with standard pre-processing protocols (Duò et al. 2018), we remove cells with a high mitochondrial gene percentage, cells with a low or high number of expressed genes, and cells with a low number of total counts. Then, we scale the data so that the total number of counts for each cell equals the average count across all cells. Finally, we apply a $\log_2$ transformation with a pseudo-count of 1, and subset to the 500 genes with the largest pre-normalization average expression levels. We separately apply this pre-processing routine to the memory T cells only, and to all of the cells. After pre-processing, we construct a "no clusters" data set by randomly sampling 600 memory T cells, and a "clusters" data set by randomly sampling 200 each of memory T cells, B cells, and monocytes.

6.2.2 Data analysis

We apply Ward-linkage hierarchical clustering with squared Euclidean distance to the "no clusters" data to get three clusters, containing 64, 428, and 108 cells, respectively. For each pair of clusters, we test $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ using (i) the test proposed in Section 4.2 under model (23) and (ii) using the Wald test under model (23), which has p-value

$$\mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\left((\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2})^T \mathbf{\Sigma}^{-1} (\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}) \geq (\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2})^T \mathbf{\Sigma}^{-1} (\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2})\right). \tag{31}$$

For both tests, we estimate $\mathbf{\Sigma}$ by applying the principal complement thresholding ("POET", Fan et al. 2013) method to the 9,303 memory T cells left out of the "no clusters" data set. Results are in Table 3. The p-values from our test are large, and the Wald p-values are small. Recall that all of the cells are memory T cells, and so (as far as we know) there are no true clusters in the data.

	"No clusters"			"Clusters"l		
Cluster pairs	(1, 2)	(1, 3)	(2, 3)	(1, 2)	(1, 3)	(2, 3)
Test statistic	4.05	4.76	2.96	3.04	4.27	4.38
Our p-value	0.20	0.27	0.70	4.60×10^{-28}	3.20×10^{-82}	1.13×10^{-73}
Wald p-value	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$

Table 3: Results from applying the test of $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ proposed in Section 4.2 and the Wald test in (31) to the "no clusters" and "clusters" data described in Section 6.2.1.

We now apply the same analysis to the "clusters" data. Ward-linkage hierarchical clustering with squared Euclidean distance results in three clusters that almost exactly correspond to memory T cells, B cells, and monocytes. For both tests, we estimate $\mathbf{\Sigma}$ by applying the POET method to the 21,757 memory T cells, B cells, and monocytes left

out of the "clusters" data set. Results are in Table 3. The p-values from both tests are extremely small. This suggests that our proposed approach is able to correctly reject the null hypothesis when it does not hold.

7 Discussion

In this paper, we proposed a selective inference framework for testing the null hypothesis that there is no difference in means between two estimated clusters, under (1). The tests developed in this paper are implemented in the `R` package `clusterpval`. Instructions on how to download and use this package can be found at `http://lucylgao.com/clusterpval`. Links to download the data sets in Section 6 can be found at `https://github.com/lucylgao/clusterpval-experiments`, along with code to reproduce the simulation and real data analysis results from this paper.

Acknowledgments

Lucy L. Gao was supported by the NSERC Discovery Grants program. Daniela Witten and Jacob Bien were supported by NIH Grant R01GM123993. Jacob Bien was supported by NSF CAREER Award DMS-1653017. Daniela Witten was supported by NSF CAREER Award DMS-1252624 and Simons Investigator Award No. 560585.

References

Bourgon, R. (2009), *Overview of the intervals package*. R Vignette, URL `https://cran.r-project.org/web/packages/intervals/vignettes/intervals_overview.pdf`.

Campbell, F. (2018), Statistical Machine Learning Methodology and Inference for Structured Variable Selection, PhD thesis, Rice University.

Chen, H., Chen, J. & Kalbfleisch, J. D. (2001), 'A modified likelihood ratio test for homogeneity in finite mixture models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(1), 19–29.

Chen, J., Li, P. & Fu, Y. (2012), 'Inference on the order of a normal mixture', *Journal of the American Statistical Association* **107**(499), 1096–1105.

Chen, S. & Bien, J. (2020), 'Valid inference corrected for outlier removal', *Journal of Computational and Graphical Statistics* **29**(2), 323–334.

Duò, A., Robinson, M. D. & Soneson, C. (2018), 'A systematic performance evaluation of clustering methods for single-cell RNA-seq data', *F1000Research* **7**.

Fan, J., Liao, Y. & Mincheva, M. (2013), 'Large covariance estimation by thresholding principal orthogonal complements', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4), 603–680.

Fithian, W., Sun, D. & Taylor, J. (2014), 'Optimal inference after model selection', *arXiv preprint arXiv:1410.2597* .

Hocking, T. D., Joulin, A., Bach, F. & Vert, J.-P. (2011), Clusterpath: An algorithm for clustering using convex fusion penalties, *in* 'Proceedings of the 28th International Conference on Machine Learning', pp. 1–8.

Horst, A. M., Hill, A. P. & Gorman, K. B. (2020), *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0, available on CRAN.

Hwang, B., Lee, J. H. & Bang, D. (2018), 'Single-cell RNA sequencing technologies and bioinformatics pipelines', *Experimental & Molecular Medicine* **50**(8), 1–14.

Hyun, S., G'Sell, M., Tibshirani, R. J. et al. (2018), 'Exact post-selection inference for the generalized lasso path', *Electronic Journal of Statistics* **12**(1), 1053–1097.

Jewell, S., Fearnhead, P. & Witten, D. (2019), 'Testing for a change in mean after change-point detection', *arXiv preprint arXiv:1910.04291* .

Kimes, P. K., Liu, Y., Neil Hayes, D. & Marron, J. S. (2017), 'Statistical significance for hierarchical clustering', *Biometrics* **73**(3), 811–821.

Kivaranovic, D. & Leeb, H. (2020), On the length of post-model-selection confidence intervals conditional on polyhedral constraints. To appear in *Journal of the American Statistical Association*.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. & Baker, C. I. (2009), 'Circular analysis in systems neuroscience: the dangers of double dipping', *Nature Neuroscience* **12**(5), 535.

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

Lance, G. N. & Williams, W. T. (1967), 'A general theory of classificatory sorting strategies: 1. hierarchical systems', *The Computer Journal* **9**(4), 373–380.

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

Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. (2008), 'Statistical significance of clustering for high-dimension, low–sample size data', *Journal of the American Statistical Association* **103**(483), 1281–1293.

Loftus, J. R. & Taylor, J. E. (2015), 'Selective inference in regression models with groups of variables', *arXiv preprint arXiv:1511.01478* .

Maitra, R., Melnykov, V. & Lahiri, S. N. (2012), 'Bootstrapping for significance of compact clusters in multi-dimensional datasets', *Journal of the American Statistical Association* **107**(497), 378–392.

Mehrizi, R. V. & Chenouri, S. (2021), 'Valid post-detection inference for change points identified using trend filtering', *arXiv preprint arXiv:2104.12022* .

Müllner, D. et al. (2013), 'fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python', *Journal of Statistical Software* **53**(9), 1–18.

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

Suzuki, R. & Shimodaira, H. (2006), 'Pvclust: an R package for assessing the uncertainty in hierarchical clustering', *Bioinformatics* **22**(12), 1540–1542.

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

Wood, S. (2015), *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.8-31, available on CRAN.

Yang, F., Barber, R. F., Jain, P. & Lafferty, J. (2016), Selective inference for group-sparse linear models, *in* 'Advances in Neural Information Processing Systems', pp. 2469–2477.

Zhang, J. M., Kamath, G. M. & David, N. T. (2019), 'Valid post-clustering differential analysis for single-cell RNA-seq', *Cell Systems* **9**(4), 383–392.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J. et al. (2017), 'Massively parallel digital transcriptional profiling of single cells', *Nature Communications* **8**(1), 1–12.

Supplementary Materials for *Selective Inference for Hierarchical Clustering*

S1 Proofs

S1.1 Proof of Theorem 1

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

For any $\nu \in \mathbb{R}^n$, we have

$$\mathbf{X} = \boldsymbol{\pi}_\nu^\perp \mathbf{X} + (\mathbf{I}_n - \boldsymbol{\pi}_\nu^\perp)\mathbf{X} = \boldsymbol{\pi}_\nu^\perp \mathbf{X} + \left(\frac{\|\mathbf{X}^T \nu\|_2}{\|\nu\|_2^2} \right) \nu \mathrm{dir}(\mathbf{X}^T \nu)^T. \tag{S1}$$

Therefore,

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X} + \left(\frac{\|\mathbf{X}^T\nu(\mathcal{G}_1,\mathcal{G}_2)\|_2}{\|\nu(\mathcal{G}_1,\mathcal{G}_2)\|_2^2}\right)\nu\ \mathrm{dir}(\mathbf{X}^T\nu(\mathcal{G}_1,\mathcal{G}_2))^T \\ &= \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X} + \left(\frac{\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2}{\frac{1}{|\mathcal{G}_1|}+\frac{1}{|\mathcal{G}_2|}}\right)\nu(\mathcal{G}_1,\mathcal{G}_2)\ \mathrm{dir}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2})^T, \qquad \text{(S2)}\end{aligned}$$

where the first equality follows from (S1), and the second equality follows from the definition of $\nu(\mathcal{G}_1,\mathcal{G}_2)$ in (7). Substituting (S2) into the definition of $p(\mathbf{x};\{\mathcal{G}_1,\mathcal{G}_2\})$ given by (8) yields

$$\begin{aligned}&p(\mathbf{x};\{\mathcal{G}_1,\mathcal{G}_2\}) \qquad \text{(S3)}\\ &= \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Bigg(\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2 \ \Bigg| \\ &\qquad \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}\left(\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{x} + \left(\frac{\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2}{\frac{1}{|\mathcal{G}_1|}+\frac{1}{|\mathcal{G}_2|}}\right)\nu(\mathcal{G}_1,\mathcal{G}_2)\mathrm{dir}(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2})^T\right), \\ &\qquad \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{x},\ \mathrm{dir}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}) = \mathrm{dir}(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2})\Bigg).\end{aligned}$$

To simplify (S3), we now show that

$$\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \ \perp\!\!\!\perp\ \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X}, \qquad \text{(S4)}$$

and that under $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}: \bar{\mu}_{\mathcal{G}_1} = \bar{\mu}_{\mathcal{G}_2}$,

$$\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \ \perp\!\!\!\perp\ \mathrm{dir}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}). \qquad \text{(S5)}$$

First, we will show (S4). Recall that $\boldsymbol{\pi}^{\perp}_{\nu}$ is the orthogonal projection matrix onto the subspace orthogonal to ν. Thus, $\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\nu(\mathcal{G}_1,\mathcal{G}_2) = 0_n$. It follows from properties of the matrix normal and multivariate normal distributions that $\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X} \ \perp\!\!\!\perp\ \mathbf{X}^T\nu(\mathcal{G}_1,\mathcal{G}_2)$. This implies (S4), since $\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} = \mathbf{X}^T\nu(\mathcal{G}_1,\mathcal{G}_2)$. Second, we will show (S5). It follows from

(1) that $\mathbf{X}_i \overset{ind}{\sim} N_q(\mu_i, \sigma^2 \mathbf{I}_q)$. Thus, under $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} : \bar{\mu}_{\mathcal{G}_1} = \bar{\mu}_{\mathcal{G}_2}$,

$$\frac{\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}}{\sigma\sqrt{1/|\mathcal{G}_1| + 1/|\mathcal{G}_2|}} = \frac{\mathbf{X}^T \nu(\mathcal{G}_1, \mathcal{G}_2)}{\|\nu(\mathcal{G}_1, \mathcal{G}_2)\|_2} \sim N_q\left(0, \mathbf{I}_q\right), \tag{S6}$$

and (S5) follows from the independence of the length and direction of a standard multivariate normal random vector.

We now apply (S4) and (S5) to (S3). This yields

$$\begin{aligned}
&p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) \\
&= \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2 \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{x} + \left(\frac{\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2}{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}}\right)\nu(\mathcal{G}_1, \mathcal{G}_2)\mathrm{dir}(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2})^T\right)\right) \\
&= \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2 \;\middle|\; \|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \in \mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})\right),
\end{aligned} \tag{S7}$$

where (S7) follows from the definition of $\mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})$ in (10). Under $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} : \bar{\mu}_{\mathcal{G}_1} = \bar{\mu}_{\mathcal{G}_2}$,

$$\frac{\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2^2}{\sigma^2\left(1/|\mathcal{G}_1| + 1/|\mathcal{G}_2|\right)} \sim \chi^2_q,$$

by (S6). That is, under $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$, we have $\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|_2 \sim \left(\sigma\sqrt{1/|\mathcal{G}_1| + 1/|\mathcal{G}_2|}\right) \cdot \chi_q$. Therefore, for $\phi \sim \left(\sigma\sqrt{1/|\mathcal{G}_1| + 1/|\mathcal{G}_2|}\right) \cdot \chi_q$, it follows from (S7) that

$$\begin{aligned}
p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) &= \mathbb{P}\left(\phi \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2 \mid \phi \in \mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})\right) \\
&= 1 - \mathbb{F}\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2; \sigma\sqrt{1/|\mathcal{G}_1| + 1/|\mathcal{G}_2|}, \mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})\right),
\end{aligned} \tag{S8}$$

since $\mathbb{F}(t; c, \mathcal{S})$ denotes the cumulative distribution function of a $c \cdot \chi_q$ random variable truncated to the set $\mathcal{S}$. This is (9).

Finally, we will show (11). It follows from the definition of $p(\cdot; \{\mathcal{G}_1, \mathcal{G}_2\})$ in (8) that for all $\mathbf{x} \in \mathbb{R}^{n \times q}$, we have

$$\begin{aligned}
\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{x}, \\
\mathrm{dir}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}) = \mathrm{dir}(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2})\Big) = \alpha.
\end{aligned} \tag{S9}$$

Therefore, letting $\mathbb{E}_{H_0}$ denote $\mathbb{E}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}$, we have

$$
\begin{aligned}
&\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(p(\mathbf{X};\{\mathcal{G}_1,\mathcal{G}_2\}) \leq \alpha \;\Big|\; \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\Big) \\
&= \mathbb{E}_{H_0}\Big[\mathbb{1}\{p(\mathbf{X};\{\mathcal{G}_1,\mathcal{G}_2\}) \leq \alpha\} \;\Big|\; \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\Big] \\
&= \mathbb{E}_{H_0}\left[\mathbb{E}_{H_0}\Big[\mathbb{1}\{p(\mathbf{X};\{\mathcal{G}_1,\mathcal{G}_2\}) \leq \alpha\} \;\Big|\; \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1,\mathcal{G}_2)}\mathbf{X}, \mathrm{dir}(\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2})\Big] \;\middle|\; \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right] \\
&= \mathbb{E}_{H_0}\Big[\alpha \;\Big|\; \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\Big] \\
&= \alpha,
\end{aligned}
$$

where the second equality follows from the law of total expectation, and the third equality follows from (S9). This is (11).

S1.2 Proof of Lemma 1

First, we state and prove a preliminary result involving the dissimilarities between groups of observations in any perturbed data set $\mathbf{x}'(\phi)$, which holds for any clustering method $\mathcal{C}$.

Lemma S1. *For any $\phi \geq 0$, and for any two sets $\mathcal{G}$ and $\mathcal{G}'$ that are both contained in $\hat{\mathcal{C}}_1$, both contained in $\hat{\mathcal{C}}_2$, or both contained in $\{1,2,\ldots,n\}\backslash[\hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2]$, we have that*

$$d(\mathcal{G},\mathcal{G}';\mathbf{x}'(\phi)) = d(\mathcal{G},\mathcal{G}';\mathbf{x}),$$

where $\mathbf{x}'(\phi)$ is defined in (13).

Proof. By (13), $[\mathbf{x}'(\phi)]_i - [\mathbf{x}'(\phi)]_{i'} = x_i - x_{i'}$ for all $i, i' \in \mathcal{G} \cup \mathcal{G}'$, and therefore all pairwise distances are preserved for all $i, i' \in \mathcal{G} \cup \mathcal{G}'$. The resut follows. □

Suppose that $\mathcal{C} = \mathcal{C}^{(n-K+1)}$. Then, (15) follows immediately from observing that $\mathcal{W}_1^{(t)}(\mathbf{x})$ and $\mathcal{W}_2^{(t)}(\mathbf{x})$ are clusters merged at an earlier stage of the hierarchical clustering algorithm and thus must satisfy the condition in Lemma S1.

We will now show (16). Applying the right-hand side of (16) with $t = n - K + 1$ yields the left-hand side, so the ($\Leftarrow$) direction holds. We will now prove the ($\Rightarrow$) direction by contradiction. Suppose that there exists some $t \in \{1, \ldots, n - K - 1\}$ such that

$$\mathcal{C}^{(t+1)}(\mathbf{x}'(\phi)) \neq \mathcal{C}^{(t+1)}(\mathbf{x}), \tag{S10}$$

and let t specifically denote the smallest such value. Then, $\{\mathcal{W}_1^{(t)}(\mathbf{x}'(\phi)), \mathcal{W}_2^{(t)}(\mathbf{x}'(\phi))\} \neq \{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\}$, which implies that

$$d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right) > d\left(\mathcal{W}_1^{(t)}(\mathbf{x}'(\phi)), \mathcal{W}_2^{(t)}(\mathbf{x}'(\phi)); \mathbf{x}'(\phi)\right). \tag{S11}$$

Since clusters cannot unmerge once they have merged, and $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}^{(n-K+1)}(\mathbf{x})$ by definition, it must be the case that $\mathcal{W}_1^{(t)}(\mathbf{x})$ and $\mathcal{W}_2^{(t)}(\mathbf{x})$ are both in $\hat{\mathcal{C}}_1$, are both in $\hat{\mathcal{C}}_2$, or are both in $[\hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2]^C$. Similarly, since we assumed that $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}^{(n-K+1)}(\mathbf{x}'(\phi))$, it must be the case that $\mathcal{W}_1^{(t)}(\mathbf{x}'(\phi))$ and $\mathcal{W}_2^{(t)}(\mathbf{x}'(\phi))$ are both in $\hat{\mathcal{C}}_1$, are both in $\hat{\mathcal{C}}_2$, or are both in $[\hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2]^C$. Thus, we can apply Lemma S1 to both sides of (S11) to yield

$$d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) > d\left(\mathcal{W}_1^{(t)}(\mathbf{x}'(\phi)), \mathcal{W}_2^{(t)}(\mathbf{x}'(\phi)); \mathbf{x}\right), \tag{S12}$$

which implies that $\left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}$ is not the pair of clusters that merged in the $(t)^{th}$ step of the hierarchical clustering procedure applied to $\mathbf{x}$. This is a contradiction. Therefore, for all $t \in \{1, \ldots, n - K - 1\}$ such that $\mathcal{C}^{(t)}(\mathbf{x}'(\phi)) = \mathcal{C}^{(t)}(\mathbf{x})$, it must be the case that $\mathcal{C}^{(t+1)}(\mathbf{x}'(\phi)) = \mathcal{C}^{(t+1)}(\mathbf{x})$. Observing that $\mathcal{C}^{(1)}(\mathbf{x}'(\phi)) = \mathcal{C}^{(1)}(\mathbf{x}) = \{\{1\}, \{2\}, \ldots, \{n\}\}$ for all $\phi \geq 0$ completes the proof of the ($\Rightarrow$) direction.

S1.3 Proof of Theorem 2

Observe from Algorithm 1 that

$$\mathcal{C}^{(t)}(\mathbf{x}'(\phi)) = \mathcal{C}^{(t)}(\mathbf{x}) \text{ for all } t = 1, 2, \ldots, n-K+1, \tag{S13}$$

if and only if for all $t = 1, 2, \ldots, n-K$,

$$\begin{aligned} d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) &> d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right), \\ \forall\ \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}),\ \mathcal{G} \neq \mathcal{G}',\ \{\mathcal{G}, \mathcal{G}'\} &\neq \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}. \end{aligned} \tag{S14}$$

Equation (15) in Lemma 1 says that

$$d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right) = d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right),\ \forall\ t = 1, 2, \ldots, n-K,\ \forall\ \phi \geq 0.$$

Therefore, (S13) is true if and only if for all $t = 1, 2, \ldots, n-K$,

$$\begin{aligned} d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) &> d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right), \\ \forall\ \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}),\ \mathcal{G} \neq \mathcal{G}',\ \{\mathcal{G}, \mathcal{G}'\} &\neq \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}. \end{aligned} \tag{S15}$$

Recall that $\mathcal{L}(\mathbf{x}) = \bigcup_{t=1}^{n-K} \left\{\{\mathcal{G}, \mathcal{G}'\} : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \mathcal{G} \neq \mathcal{G}', \{\mathcal{G}, \mathcal{G}'\} \neq \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}\right\}$. Observe that (S15) is true for all $t = 1, 2, \ldots, n-K$ if and only if for all $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$,

$$d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) > \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t \leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right), \tag{S16}$$

where $l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \min\left\{1 \leq t \leq n-K : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \{\mathcal{G}, \mathcal{G}'\} \neq \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}\right\}$ and $u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \max\left\{1 \leq t \leq n-K : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\}\right\}$ are the start and end of the lifetime of the cluster pair $\{\mathcal{G}, \mathcal{G}'\}$, respectively. Therefore, it follows from

(16) in Lemma 1 that $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))$ if and only if (S16) is true for all $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$. Recalling from (12) that $\hat{\mathcal{S}} = \{\phi \geq 0 : \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\}$, it follows that

$$\hat{\mathcal{S}} = \bigcap_{\{\mathcal{G},\mathcal{G}'\} \in \mathcal{L}(\mathbf{x})} \left\{ \phi \geq 0 : d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) > \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t \leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) \right\}.$$

This is (18). Furthermore, this intersection involves $\mathcal{O}(n^2)$ sets, because

$$|\mathcal{L}(\mathbf{x})| = \left(\binom{n}{2} - 1 \right) + \sum_{t=1}^{n-K-1} (n - t - 1).$$

S1.4 Proof of Proposition 2

Let $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$, where $\mathcal{L}(\mathbf{x})$ in (17) is the set of losing cluster pairs in $\mathbf{x}$. Suppose that we are given the start and end of the lifetime of this cluster pair in $\mathbf{x}$, $l_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ and $u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$, and we are given the set of steps where inversions occur in the dendrogram of $\mathbf{x}$ below the $(n-K)$th merge,

$$\mathcal{M}(\mathbf{x}) = \left\{ 1 \leq t < n - K : d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) > d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{W}_2^{(t+1)}(\mathbf{x}); \mathbf{x}\right) \right\}.$$

Observe that for $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ in (19),

$$h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t \leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) = \max_{t \in \mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \cup \{u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right),$$

where $\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \left\{ t : l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t < u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}), d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) > d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{G}_2^{(l+1)}(\mathbf{x}); \mathbf{x}\right) \right\}$.
Therefore, we can compute $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ for each cluster pair $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$ as follows:

1. Compute $\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ by checking every step in $\mathcal{M}(\mathbf{x})$ to see if it is in $\left[l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}), u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\right)$.
2. Compute $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ by taking the max over $|\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x})| + 1$ merge heights.

Step 1 requires $2|\mathcal{M}(\mathbf{x})|$ operations, and Step 2 requires $|\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x})| + 1$ operations. Since $|\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x})| < |\mathcal{M}(\mathbf{x})|$, it follows that we can compute $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})$ in $\mathcal{O}(|\mathcal{M}(\mathbf{x})| + 1)$ time.

S1.5 Proof of Proposition 4

Recall that

$$\hat{\mathcal{S}}_{i,i'} = \left\{ \phi \geq 0 : d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > \max_{\{\mathcal{G},\mathcal{G}'\} \in \mathcal{L}(\mathbf{x}): i \in \mathcal{G}, i' \in \mathcal{G}'} h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \right\}. \tag{S17}$$

We will consider two cases: either

$$i, i' \in \hat{\mathcal{C}}_1 \text{ or } i, i' \in \hat{\mathcal{C}}_2 \text{ or } i, i' \notin \hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2, \tag{S18}$$

or (S18) doesn't hold.

Case 1: (S18) **holds**

We first state and prove a preliminary result.

Lemma S2. *Suppose that $\mathcal{C} = \mathcal{C}^{(n-K+1)}$, i.e. we perform hierarchical clustering to obtain K clusters. Let $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$, $i \in \mathcal{G}$, $i' \in \mathcal{G}'$. Suppose that* (S18) *holds. Then, for all $\phi \geq 0$,*

$$d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) = d(\mathcal{G}, \mathcal{G}'; \mathbf{x}). \tag{S19}$$

Proof. Since $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$, it follows that $\mathcal{G}, \mathcal{G}' \in \bigcup_{t=1}^{n-K} \mathcal{C}^{(t)}(\mathbf{x})$. Furthermore, clusters that merge cannot unmerge, and $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}^{(n-K+1)}(\mathbf{x})$. This means that all observations in the set $\mathcal{G}$ must belong to the same cluster in $\mathcal{C}^{(n-K+1)}(\mathbf{x})$, and all observations in the set $\mathcal{G}'$ must belong to the same cluster in $\mathcal{C}^{(n-K+1)}(\mathbf{x})$. Therefore, (S18) implies that either

$$\mathcal{G}, \mathcal{G}' \subseteq \hat{\mathcal{C}}_1 \text{ or } \mathcal{G}, \mathcal{G}' \subseteq \hat{\mathcal{C}}_2 \text{ or } \mathcal{G}, \mathcal{G}' \subseteq \{1, 2, \ldots, n\} \backslash [\hat{\mathcal{C}}_1 \cup \hat{\mathcal{C}}_2].$$

It follows from Lemma S1 that (S19) holds. □

Suppose that i, i' satisfy (S18). Recall from Section 3.2 that

$$d(\mathcal{G}, \mathcal{G}'; \mathbf{x}) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \quad \forall \{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x}). \tag{S20}$$

It follows from (S20) and Lemma S2 that for all $\phi \geq 0$,

$$d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) = d(\mathcal{G}, \mathcal{G}'; \mathbf{x}) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \quad \forall \{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x}) \text{ s.t. } i \in \mathcal{G}, i' \in \mathcal{G}'. \tag{S21}$$

Recall that single linkage defines $d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) = \min_{i \in \mathcal{G}, i' \in \mathcal{G}'} d(\{i\}, \{i'\}; \mathbf{x}'(\phi))$. Thus, it follows from (S21) that for all $\phi \geq 0$,

$$d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \quad \forall \{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x}) \text{ s.t. } i \in \mathcal{G}, i' \in \mathcal{G}'. \tag{S22}$$

Applying (S22) to the definition of $\hat{\mathcal{S}}_{i,i'}$ in (S17) yields $\hat{\mathcal{S}}_{i,i'} = [0, \infty)$.

Case 2: (S18) **does not hold**

Suppose that i, i' do not satisfy (S18). Then, i and i' are assigned to different clusters in $\mathcal{C}(\mathbf{x}) = \mathcal{C}^{(n-K+1)}(\mathbf{x})$. Since $\mathcal{C}^{(n-K+1)}(\mathbf{x})$ is created by merging a pair of clusters in $\mathcal{C}^{(n-K)}(\mathbf{x})$, it follows that i and i' are assigned to different clusters $\mathcal{G}$ and $\mathcal{G}'$ in $\mathcal{C}^{(n-K)}(\mathbf{x})$ such that $\{\mathcal{G}, \mathcal{G}'\}$ is not the winning pair. That is, there exists $\mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(n-K)}(\mathbf{x})$ such that $i \in \mathcal{G}$, and $i' \in \mathcal{G}'$, $\mathcal{G} \neq \mathcal{G}'$, and $\{\mathcal{G}, \mathcal{G}'\} \neq \left\{\mathcal{W}_1^{(n-K)}(\mathbf{x}), \mathcal{W}_2^{(n-K)}(\mathbf{x})\right\}$. This means that $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$, with $u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = n - K$. Furthermore, single linkage cannot produce inversions (Murtagh & Contreras 2012), i.e. $d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) < d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{W}_2^{(t+1)}(\mathbf{x}); \mathbf{x}\right)$ for all $t < n-1$. It follows that

$$\begin{aligned}\max_{\{\mathcal{G},\mathcal{G}'\} \in \mathcal{L}(\mathbf{x}): i \in \mathcal{G}, i' \in \mathcal{G}'} h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) &= \max_{\mathcal{G},\mathcal{G}' \in \mathcal{L}(\mathbf{x}): i \in \mathcal{G}, i' \in \mathcal{G}'} \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t \leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) \\ &= d\left(\mathcal{W}_1^{(n-K)}(\mathbf{x}), \mathcal{W}_2^{(n-K)}(\mathbf{x}); \mathbf{x}\right). \end{aligned} \tag{S23}$$

Applying (S23) to the definition of $\hat{\mathcal{S}}_{i,i'}$ in (S17) yields

$$\hat{\mathcal{S}}_{i,i'} = \left\{\phi \geq 0 : d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > d\left(\mathcal{W}_1^{(n-K)}(\mathbf{x}), \mathcal{W}_2^{(n-K)}(\mathbf{x}); \mathbf{x}\right)\right\}.$$

S1.6 Proof of Theorem 4

The proof of Theorem 4 is adapted from Appendices A.10 and A.11 in Tibshirani et al. (2018). We first state and prove an intermediate result.

Lemma S3. *Recall that $F(t; c, \mathcal{S})$ denotes the CDF of a $c \cdot \chi_q$ random variable truncated to the set $\mathcal{S}$. If $0 < c_1 < c_2$, then*

$$1 - F(t; c_1, \mathcal{S}) < 1 - F(t; c_2, \mathcal{S}), \quad \textit{for all } t \in \mathcal{S}.$$

Proof. Let $f(t; c, \mathcal{S})$ denote the probability density function (pdf) of a $c \cdot \chi_q$ random variable truncated to the set $\mathcal{S}$. Observe that

$$\begin{aligned} f(t; c, \mathcal{S}) &= \frac{\frac{1}{2^{\frac{q}{2}-1}\Gamma(q/2)} c^{-q} t^{q-1} e^{-\frac{t^2}{2c^2}} \mathbb{1}\{t \in \mathcal{S}\}}{\int \frac{1}{2^{\frac{q}{2}-1}\Gamma(q/2)} c^{-q} u^{q-1} e^{-\frac{u^2}{2c^2}} \mathbb{1}\{u \in \mathcal{S}\}\, du} \\ &= \frac{t^{q-1} e^{-\frac{t^2}{2c^2}} \mathbb{1}\{t \in \mathcal{S}\}}{\int u^{q-1} e^{-\frac{u^2}{2c^2}} \mathbb{1}\{u \in \mathcal{S}\}\, du} \\ &= \left(\frac{t^{q-1} \mathbb{1}\{t \in \mathcal{S}\}}{\int u^{q-1} e^{-\frac{u^2}{2c^2}} \mathbb{1}\{u \in \mathcal{S}\}\, du}\right) \left(\exp\left\{-\frac{t^2}{2c^2}\right\}\right). \end{aligned}$$

Thus, $\{f(t; c, \mathcal{S})\}_{c>0}$ is an exponential family with natural parameter $\frac{1}{c^2}$. Therefore, it has a monotone non-increasing likelihood ratio in its sufficient statistic, $-t^2/2$. This means that for any fixed $0 < c_1 < c_2$,

$$\frac{f(u_2; c_1, \mathcal{S})}{f(u_2; c_2, \mathcal{S})} < \frac{f(u_1; c_1, \mathcal{S})}{f(u_1; c_2, \mathcal{S})}, \quad \text{for all } u_1, u_2 \in \mathcal{S},\ u_1 < u_2. \tag{S24}$$

Rearranging terms in (S24) yields:

$$f(u_2; c_1, \mathcal{S}) f(u_1; c_2, \mathcal{S}) < f(u_1; c_1, \mathcal{S}) f(u_2; c_2, \mathcal{S}), \quad \text{for all } u_1, u_2 \in \mathcal{S},\ u_1 < u_2. \tag{S25}$$

Let $t, u_2 \in \mathcal{S}$ with $t < u_2$. Then,

$$\begin{aligned} f(u_2; c_1, \mathcal{S}) F(t; c_2, \mathcal{S}) &= \int_{-\infty}^{t} f(u_2; c_1, \mathcal{S}) f(u_1; c_2, \mathcal{S}) du_1 \\ &< \int_{-\infty}^{t} f(u_1; c_1, \mathcal{S}) f(u_2; c_2, \mathcal{S}) du_1 = F(t; c_1, \mathcal{S}) f(u_2; c_2, \mathcal{S}), \end{aligned}$$

where the inequality follows from (S25). That is, we have shown that

$$f(u_2; c_1, \mathcal{S}) F(u_1; c_2, \mathcal{S}) < F(u_1; c_1, \mathcal{S}) f(u_2; c_2, \mathcal{S}), \quad \text{for all } u_1, u_2 \in \mathcal{S},\ u_1 < u_2. \tag{S26}$$

Let $t \in \mathcal{S}$. Then,

$$\begin{aligned} (1 - F(t; c_1, \mathcal{S})) F(t; c_2, \mathcal{S}) &= \int_{t}^{\infty} f(u_2; c_1, \mathcal{S}) F(t; c_2, \mathcal{S}) du_2 \\ &< \int_{t}^{\infty} F(t; c_1, \mathcal{S}) f(u_2; c_2, \mathcal{S}) du_2 = F(t; c_1, \mathcal{S})(1 - F(t; c_2, \mathcal{S})), \end{aligned} \tag{S27}$$

where the inequality follows from (S26) and the fact that $f(t; c, \mathcal{S}) = 0$ for all $t \notin \mathcal{S}$. Rearranging terms in (S27) yields

$$1 - F(t; c_1, \mathcal{S}) < 1 - F(t; c_2, \mathcal{S}).$$

□

We will now prove Theorem 4. In the following, we will use A_n to denote the event $\left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)} \in \mathcal{C}(\mathbf{X}^{(n)})\right\}$, $\hat{p}_n$ to denote $\hat{p}\left(\mathbf{X}^{(n)}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right)$, and p_n to denote $p\left(\mathbf{X}^{(n)}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right)$.

Observe that

$$\begin{aligned}
&\mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}_n \le \alpha \;\middle|\; A_n\right) \\
&= \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}_n \le \alpha,\ \hat{\sigma}(\mathbf{X}^{(n)}) \ge \sigma \;\middle|\; A_n\right) + \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}_n \le \alpha,\ \hat{\sigma}(\mathbf{X}^{(n)}) < \sigma \;\middle|\; A_n\right) \\
&\le \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}_n \le \alpha,\ \hat{\sigma}(\mathbf{X}^{(n)}) \ge \sigma \;\middle|\; A_n\right) + \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{\sigma}(\mathbf{X}^{(n)}) < \sigma \;\middle|\; A_n\right). \qquad \text{(S28)}
\end{aligned}$$

Recall from (9) that

$$p\left(\mathbf{x}^{(n)}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right) = 1 - \mathbb{F}\left(\left\|\bar{x}_{\hat{\mathcal{C}}_1^{(n)}} - \bar{x}_{\hat{\mathcal{C}}_2^{(n)}}\right\|_2 ; \sigma\sqrt{\frac{1}{|\hat{\mathcal{C}}_1^{(n)}|} + \frac{1}{|\hat{\mathcal{C}}_2^{(n)}|}}, \mathcal{S}\left(\mathbf{x}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right)\right),$$

and recall from (25) that

$$\hat{p}\left(\mathbf{x}^{(n)}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right) = 1 - \mathbb{F}\left(\left\|\bar{x}_{\hat{\mathcal{C}}_1^{(n)}} - \bar{x}_{\hat{\mathcal{C}}_2^{(n)}}\right\|_2 ; \hat{\sigma}(\mathbf{x})\sqrt{\frac{1}{|\hat{\mathcal{C}}_1^{(n)}|} + \frac{1}{|\hat{\mathcal{C}}_2^{(n)}|}}, \mathcal{S}\left(\mathbf{x}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right)\right).$$

It follows from Lemma S3 that for any $\mathbf{x}$ such that $\hat{\sigma}(\mathbf{x}) \ge \sigma$, we have that

$$\hat{p}\left(\mathbf{x}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right) \ge p\left(\mathbf{x}; \left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}\right).$$

Thus, for all $n = 1, 2, \ldots$ and for all $0 \le \alpha \le 1$,

$$\begin{aligned}
\mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}_n \le \alpha, \hat{\sigma}(\mathbf{X}^{(n)}) \ge \sigma \;\middle|\; A_n\right) &\le \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(p_n \le \alpha, \hat{\sigma}(\mathbf{X}^{(n)}) \ge \sigma \;\middle|\; A_n\right) \\
&\le \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(p_n \le \alpha \;\middle|\; A_n\right) \\
&= \alpha, \qquad \text{(S29)}
\end{aligned}$$

where the equality follows from (11). Applying (S29) to (S28) yields

$$\mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}_n \le \alpha \;\middle|\; A_n\right) \le \alpha + \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{\sigma}(\mathbf{X}^{(n)}) < \sigma \;\middle|\; A_n\right). \qquad \text{(S30)}$$

By our assumption in Theorem 4, $\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\}}}\left(\hat{\sigma}(\mathbf{X}^{(n)}) < \sigma \;\middle|\; A_n\right) = 0$. Applying this fact to (S30) yields

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\}}}\left(\hat{p}_n \leq \alpha \;\middle|\; A_n\right) \leq \alpha.$$

This completes the proof of Theorem 4.

S1.7 Proof of Propositions 5 and 6

Recall that $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ denotes the null hypothesis that $\bar{\mu}_{\mathcal{G}_1} = \bar{\mu}_{\mathcal{G}_2}$. Further recall that $\mathcal{H}_0$ denotes the set of hypotheses of the form $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ that are true. We start by rewriting (11) as follows.

$$\text{If } H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} \in \mathcal{H}_0, \text{ then } \mathbb{P}\left(p(\mathbf{X};\{\mathcal{G}_1,\mathcal{G}_2\}) \leq \alpha \;\middle|\; \mathcal{G}_1,\mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) = \alpha. \tag{S31}$$

We will first prove Proposition 5. Let $K = 2$. Then,

$$\begin{aligned}
&\mathbb{P}\left(p(\mathbf{X};\mathcal{C}(\mathbf{X})) \leq \alpha, H_0^{\mathcal{C}(\mathbf{X})} \in \mathcal{H}_0\right) \\
&= \sum_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} \in \mathcal{H}_0} \mathbb{P}\left(p(\mathbf{X};\{\mathcal{G}_1,\mathcal{G}_2\}) \leq \alpha, \mathcal{C}(\mathbf{X}) = \{\mathcal{G}_1,\mathcal{G}_2\}\right) \\
&= \sum_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} \in \mathcal{H}_0} \mathbb{P}\left(p(\mathbf{X};\{\mathcal{G}_1,\mathcal{G}_2\}) \leq \alpha \mid \mathcal{C}(\mathbf{X}) = \{\mathcal{G}_1,\mathcal{G}_2\}\right) \mathbb{P}(\mathcal{C}(\mathbf{X}) = \{\mathcal{G}_1,\mathcal{G}_2\}) \\
&= \sum_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} \in \mathcal{H}_0} \alpha\, \mathbb{P}(\mathcal{C}(\mathbf{X}) = \{\mathcal{G}_1,\mathcal{G}_2\}) \qquad \text{(S32)} \\
&= \alpha\, \mathbb{P}(H_0^{\mathcal{C}(\mathbf{X})} \in \mathcal{H}_0),
\end{aligned}$$

where (S32) follows from (S31) and the fact that $K = 2$. Therefore,

$$\mathbb{P}\left(p(\mathbf{X};\mathcal{C}(\mathbf{X})) \leq \alpha \mid H_0^{\mathcal{C}(\mathbf{X})} \in \mathcal{H}_0\right) = \alpha.$$

This is (26).

We will now prove Proposition 6. Let $K > 2$, and let $\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}$ and $\mathbf{X}$ be conditionally independent given $\mathcal{C}(\mathbf{X})$. Then,

$$\begin{aligned}
&\mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}) \leq \alpha,\ H_0^{\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}} \in \mathcal{H}_0\right) \qquad \text{(S33)}\\
&= \sum_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} \in \mathcal{H}_0} \mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha, \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}\right)\\
&= \sum_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} \in \mathcal{H}_0} \mathbb{P}(\{p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha\} \cap \{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\} \cap \{\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}\})
\end{aligned}$$

where the last equality follows from the fact that $\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}$ implies that $\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})$. Since we assumed that $\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}$ and $\mathbf{X}$ are conditionally independent given $\mathcal{C}(\mathbf{X})$, we have for any $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} \in \mathcal{H}_0$ that

$$\begin{aligned}
&\mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha, \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\} \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right)\\
&= \mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) \mathbb{P}\left(\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\} \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right)\\
&= \alpha\ \mathbb{P}\left(\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\} \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right),
\end{aligned}$$

where the last equality follows from (S31). It follows that for any $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} \in \mathcal{H}_0$, we have that

$$\begin{aligned}
&\mathbb{P}(\{p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha\} \cap \{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\} \cap \{\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}\})\\
&= \alpha\ \mathbb{P}\left(\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\} \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) \mathbb{P}\left(\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right)\\
&= \alpha\ \mathbb{P}\left(\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right). \qquad \text{(S34)}
\end{aligned}$$

Applying (S34) to (S33) yields

$$\begin{aligned}
&\mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}) \leq \alpha,\ H_0^{\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}} \in \mathcal{H}_0\right) \\
&= \sum_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} \in \mathcal{H}_0} \alpha\, \mathbb{P}\left(\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) \\
&= \sum_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} \in \mathcal{H}_0} \alpha\, \mathbb{P}\left(\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}\right) \\
&= \alpha\, \mathbb{P}\left(H_0^{\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}} \in \mathcal{H}_0\right),
\end{aligned}$$

where the second-to-last equality follows from the fact that $\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\} = \{\mathcal{G}_1, \mathcal{G}_2\}$ implies that $\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})$. Therefore,

$$\mathbb{P}\left(p(\mathbf{X}; \{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}) \leq \alpha \mid H_0^{\{\mathcal{G}_1(\mathbf{X}), \mathcal{G}_2(\mathbf{X})\}} \in \mathcal{H}_0\right) = \alpha.$$

This is (27).

S2 Algorithm for computing $\hat{\mathcal{S}}$ in the case of squared Euclidean distance and linkages that satisfy (20)

We begin by defining quantities used in the algorithm. For all $\mathcal{G} \in \bigcup_{t=1}^{n-K} \mathcal{C}^{(t)}(\mathbf{x})$, let

$$l_{\mathcal{G}}(\mathbf{x}) \equiv \min\{1 \leq t \leq n-K : \mathcal{G} \in \mathcal{C}^{(t)}(\mathbf{x})\},\ u_{\mathcal{G}}(\mathbf{x}) \equiv \max\{1 \leq t \leq n-K : \mathcal{G} \in \mathcal{C}^{(t)}(\mathbf{x})\}, \tag{S35}$$

so that for any $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$,

$$l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \max\{l_{\mathcal{G}}(\mathbf{x}), l_{\mathcal{G}'}(\mathbf{x})\}, \tag{S36}$$

$$u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \begin{cases} \min\{u_{\mathcal{G}}(\mathbf{x}), u_{\mathcal{G}'}(\mathbf{x})\} - 1, & \text{if } \{\mathcal{G}, \mathcal{G}'\} \in \left\{\left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\} : 1 \leq t \leq n-K\right\}, \\ \min\{u_{\mathcal{G}}(\mathbf{x}), u_{\mathcal{G}'}(\mathbf{x})\}, & \text{otherwise.} \end{cases}$$

Recall from Section 3.3 that

$$h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \max_{t \in \mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \cup \{u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right), \text{ where} \tag{S37}$$

$$\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \left\{t : t \in \mathcal{M}(\mathbf{x}), l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t < u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\right\}. \tag{S38}$$

Let $\mathcal{G}_{new}^{(t)}(\mathbf{x}) = \mathcal{W}_1^{(t-1)}(\mathbf{x}) \cup \mathcal{W}_2^{(t-1)}(\mathbf{x})$ for all $t = 2, \ldots, n-K$. Define

$$\mathcal{L}_1(\mathbf{x}) = \left\{\{\mathcal{G}, \mathcal{G}'\} : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(1)}(\mathbf{x}), \mathcal{G} \neq \mathcal{G}', \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(1)}(\mathbf{x}), \mathcal{W}_2^{(1)}(\mathbf{x})\}\right\}, \tag{S39}$$

and for $t = 2, \ldots, n-K$, define

$$\mathcal{L}_t(\mathbf{x}) = \left\{\left\{\mathcal{G}_{new}^{(t)}(\mathbf{x}), \mathcal{G}'\right\} : \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}) \backslash \mathcal{G}_{new}^{(t)}(\mathbf{x}), \left\{\mathcal{G}_{new}^{(t)}(\mathbf{x}), \mathcal{G}'\right\} \neq \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}\right\}, \tag{S40}$$

so that for $\mathcal{L}(\mathbf{x})$ defined in (17), $\mathcal{L}(\mathbf{x}) = \bigcup_{t=1}^{n-K} \mathcal{L}_t(\mathbf{x})$. Thus, it follows from Theorem 2 that

$$\hat{\mathcal{S}} = \bigcap_{t=1}^{n-K} \bigcap_{\{\mathcal{G},\mathcal{G}'\} \in \mathcal{L}_t(\mathbf{x})} \hat{\mathcal{S}}_{\mathcal{G},\mathcal{G}'}, \tag{S41}$$

where for all $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$,

$$\hat{\mathcal{S}}_{\mathcal{G},\mathcal{G}'} = \{\phi \geq 0 : d(\mathcal{G}, \mathcal{G}'; \mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}. \tag{S42}$$

The following algorithm computes the set $\hat{\mathcal{S}}$ (defined in (12)) in the case of squared Euclidean distance and linkages that satisfy (20) in $\mathcal{O}\Big((|\mathcal{M}(\mathbf{x})| + \log(n))n^2\Big)$, where $|\mathcal{M}(\mathbf{x})|$ is the number of inversions in the dendrogram of $\mathbf{x}$ below the $(n-K)$th merge:

1. Compute $\mathcal{M}(\mathbf{x}) = \left\{1 \leq t < n-K : d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) > d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}), \mathcal{W}_2^{(t+1)}(\mathbf{x}); \mathbf{x}\right)\right\}$.
2. Compute $l_{\mathcal{G}}(\mathbf{x})$ and $u_{\mathcal{G}}(\mathbf{x})$ defined in (S35) for all $\mathcal{G} \in \bigcup_{t=1}^{n-K} \mathcal{C}^{(t)}(\mathbf{x})$ as follows:

(a) Let $l_{\{i\}}(\mathbf{x}) = 1$ for all i. Let $l_{\mathcal{G}^{(t)}_{new}(\mathbf{x})}(\mathbf{x}) = t$ for all $2 \leq t < n - K$.

(b) For all $\mathcal{G} \in \bigcup_{t=1}^{n-K} \mathcal{C}^{(t)}(\mathbf{x})$, initialize $u_{\mathcal{G}}(\mathbf{x}) = n - K$.

(c) For $t' = 1, 2, \ldots, n - K$, update $u_{\mathcal{W}_1^{(t')}(\mathbf{x})}(\mathbf{x}) = t'$ and $u_{\mathcal{W}_2^{(t')}(\mathbf{x})}(\mathbf{x}) = t'$.

3. For all $\{\{i\}, \{i'\}\} \in \mathcal{L}_1(\mathbf{x})$, compute $\hat{\mathcal{S}}_{\{i\},\{i'\}}$ defined in (S42) as follows:

(a) Use $l_{\{i\}}(\mathbf{x})$ and $l_{\{i'\}}(\mathbf{x})$ to compute $l_{\{i\},\{i'\}}(\mathbf{x})$ and likewise for $u_{\{i\},\{i'\}}(\mathbf{x})$, according to (S36).

(b) Compute $\mathcal{M}_{\{i\},\{i'\}}(\mathbf{x})$ in (S38) by checking which elements of $\mathcal{M}(\mathbf{x})$ are in $[l_{\{i\},\{i'\}}(\mathbf{x}), u_{\{i\},\{i'\}}(\mathbf{x})]$.

(c) Use $\mathcal{M}_{\{i\},\{i'\}}(\mathbf{x})$ and $u_{\{i\},\{i'\}}(\mathbf{x})$ to compute $h_{\{i\},\{i'\}}(\mathbf{x})$ according to (S37).

(d) Compute the coefficients corresponding to the quadratic function $d(\{i\}, \{i'\}; \mathbf{x}'(\phi))$ in constant time, using Lemma 2.

(e) Use these coefficients and $h_{\{i\},\{i'\}}(\mathbf{x})$ to evaluate $\hat{\mathcal{S}}_{\{i\},\{i'\}} = \Big\{\phi \geq 0 : d(\{i\}, \{i'\}; \mathbf{x}'(\phi)) > h_{\{i\},\{i'\}}(\mathbf{x}))\Big\}$. This amounts to solving a quadratic inequality in ϕ.

4. For all $t = 2, \ldots, n - K$ and for all $\{\mathcal{G}^{(t)}_{new}(\mathbf{x}), \mathcal{G}'\} \in \mathcal{L}_t(\mathbf{x})$, compute $\hat{\mathcal{S}}_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}$ defined in (S42) as follows:

(a) Use $l_{\mathcal{G}^{(t)}_{new}(\mathbf{x})}(\mathbf{x})$ and $l_{\mathcal{G}'}(\mathbf{x})$ to compute $l_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ and likewise for $u_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$, according to (S36).

(b) Compute $\mathcal{M}_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ in (S38) by checking which elements of $\mathcal{M}(\mathbf{x})$ are between $l_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ and $u_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$.

(c) Use $\mathcal{M}_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ and $u_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ to compute $h_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ in (S37).

(d) Compute the coefficients corresponding to the quadratic function $d\left(\mathcal{G}^{(t)}_{new}(\mathbf{x}), \mathcal{G}'; \mathbf{x}'(\phi)\right)$ in constant time, using (20) (Proposition 1).

(e) Use these coefficients and $h_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})$ to evaluate $\hat{\mathcal{S}}_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'} = \left\{\phi \geq 0 : d\Big(\mathcal{G}^{(t)}_{new}(\mathbf{x}), \mathcal{G}'; \mathbf{x}'(\phi)\Big) > h_{\mathcal{G}^{(t)}_{new}(\mathbf{x}),\mathcal{G}'}(\mathbf{x})\right\}$. This amounts to solving a quadratic inequality in ϕ.

5. Compute $\hat{\mathcal{S}} = \bigcap_{t=1}^{n-K} \bigcap_{\{\mathcal{G},\mathcal{G}'\}\in\mathcal{L}_t(\mathbf{x})} \hat{\mathcal{S}}_{\mathcal{G},\mathcal{G}'}$.

To see that this algorithm computes $\hat{\mathcal{S}}$ in $\mathcal{O}\Big((|\mathcal{M}(\mathbf{x})| + \log(n))n^2\Big)$ time, observe that:

- Steps 1 and 2 can each be performed in $\mathcal{O}(n)$ time. Note that $\left|\bigcap_{t=1}^{n-K} \mathcal{C}^{(t)}\right| = \mathcal{O}(n)$.
- Each of the $\mathcal{O}(n^2)$ iterations in Step 3 can be performed in $\mathcal{O}(|\mathcal{M}(\mathbf{x})| + 1)$ time.
- Each of the $\mathcal{O}(n^2)$ iterations in Step 4 can be performed in $\mathcal{O}(|\mathcal{M}(\mathbf{x})| + 1)$ time.
- Step 5 can be performed in $\mathcal{O}(n^2 \log(n))$ time. This is because $\hat{\mathcal{S}}_{\mathcal{G},\mathcal{G}'}$ solves a quadratic inequality for each $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(\mathbf{x})$, we can take the intersection over the solution sets of N quadratic inequalities in $\mathcal{O}(N \log N)$ time (Bourgon 2009), and $|\mathcal{L}(\mathbf{x})| = \mathcal{O}(n^2)$.

S3 Supplementary material for Section 4.3

The goal of this section is to establish an estimator of σ that satisfies the condition from Theorem 4. For any data set $\mathbf{x}$, define

$$\hat{\sigma}(\mathbf{x}) = \sqrt{\frac{1}{nq - q} \sum_{i=1}^{n} \sum_{j=1}^{q} (x_{ij} - \bar{x}_j)^2}, \tag{S43}$$

where $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$ for all $j = 1, 2, \ldots, q$. We will first show that under some assumptions, $\hat{\sigma}(\mathbf{X}^{(n)})$ asymptotically over-estimates σ. Then, we will show that if we estimate σ by applying $\hat{\sigma}(\cdot)$ to an independent and identically distributed copy of $\mathbf{X}^{(n)}$, then the condition from Theorem 4 is satisfied.

Let $K^* > 1$ be unknown. We assume that the sequence of mean matrices for $\{\mathbf{X}^{(n)}\}_{n=1}^{\infty}$ in Theorem 4, which we denote as $\{\boldsymbol{\mu}^{(n)}\}_{n=1}^{\infty}$, satisfies the following assumptions.

Assumption 1. *For all* $n = 1, 2, \ldots$, $\left\{\mu_i^{(n)}\right\}_{i=1}^{n} = \{\theta_1, \ldots, \theta_{K^*}\}$, *i.e. there are exactly* K^* *distinct mean vectors among the first* n *observations.*

Assumption 2. *For all* $k = 1, 2, \ldots, K^*$, $\lim_{n\to\infty}\left(\sum_{i=1}^{n} \mathbb{1}\{\mu_i^{(n)} = \theta_k\}/n\right) = \lambda_k$, *where* $\sum_{k=1}^{K^*} \lambda_k = 1$ *and* $\lambda_k > 0$. *That is, the proportion of the first* n *observations that have mean vector* θ_k *converges to* λ_k *for all* k.

This leads to the following result, which says that under Assumptions 1 and 2, $\hat{\sigma}(\mathbf{X}^{(n)})$ asymptotically over-estimates σ.

Lemma S4. *For* $n = 1, 2, \ldots$, *suppose that* $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times q}(\boldsymbol{\mu}^{(n)}, \mathbf{I}_n, \sigma^2\mathbf{I}_q)$, *and that Assumptions 1 and 2 hold for some* $K^* > 1$. *Then,*

$$\lim_{n\to\infty} \mathbb{P}(\hat{\sigma}(\mathbf{X}^{(n)}) \geq \sigma) = 1. \tag{S44}$$

where $\hat{\sigma}(\cdot)$ *is defined in* (S43).

Proof. Observe that

$$\left(\frac{nq-q}{nq}\right)\hat{\sigma}^2(\mathbf{X}^{(n)}) = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{ij}^{(n)}\right)^2 - \frac{1}{q}\sum_{j=1}^{q}\left(\bar{X}_j^{(n)}\right)^2. \tag{S45}$$

We will first show that

$$\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{ij}^{(n)}\right)^2 \xrightarrow{p} \sigma^2 + \frac{1}{q}\sum_{j=1}^{q}\sum_{k=1}^{K^*}\lambda_k\theta_{kj}^2. \tag{S46}$$

Since for any positive integer r and for any $j = 1, 2, \ldots, q$, we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left[\mu_{ij}^{(n)}\right]^r = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K^*}\mathbb{1}\left\{\mu_i^{(n)} = \theta_k\right\}(\theta_{kj})^r = \sum_{k=1}^{K^*}\lambda_k\left(\theta_{kj}\right)^r, \tag{S47}$$

it follows that

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{ij}^{(n)}\right)^2\right] &= \lim_{n\to\infty}\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left[\mathrm{Var}\left[X_{ij}^{(n)}\right] + \left(\mathbb{E}\left[X_{ij}^{(n)}\right]\right)^2\right] \\
&= \lim_{n\to\infty}\left(\sigma^2 + \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left[\mu_{ij}^{(n)}\right]^2\right) \\
&= \sigma^2 + \frac{1}{q}\sum_{j=1}^{q}\sum_{k=1}^{K^*}\lambda_k\left(\theta_{kj}\right)^2. \qquad \text{(S48)}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{ij}^{(n)}\right)^2\right] &\le \frac{1}{(nq)^2}\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[\left(X_{ij}^{(n)}\right)^4\right] \\
&= \frac{1}{(nq)^2}\sum_{i=1}^{n}\sum_{j=1}^{q}\left[\left(\mu_{ij}^{(n)}\right)^4 + 6\left(\mu_{ij}^{(n)}\right)^2\sigma^2 + 3\sigma^4\right] \qquad \text{(S49)} \\
&= \frac{1}{nq}\left[\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(\mu_{ij}^{(n)}\right)^4\right] + \frac{6\sigma^2}{nq}\left[\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left[\mu_{ij}^{(n)}\right]^2\right] + \frac{3\sigma^4}{nq},
\end{aligned}$$

where (S49) follows from the fact that $X_{ij}^{(n)} \sim N(\mu_{ij}^{(n)}, \sigma^2)$. It follows from (S47) that

$$\lim_{n\to\infty}\mathrm{Var}\left[\frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{ij}^{(n)}\right)^2\right] = 0. \tag{S50}$$

Equation (S46) follows from (S48) and (S50). Next, since (S47) holds for $r = 1$ and for all $j = 1, 2, \ldots, q$, we have that

$$\lim_{n\to\infty} \mathbb{E}\left[\bar{X}_j^{(n)}\right] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_{ij}^{(n)} = \sum_{k=1}^{K^*} \lambda_k \theta_{kj}, \qquad \lim_{n\to\infty} \mathrm{Var}\left[\bar{X}_j^{(n)}\right] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Thus, for all $j = 1, 2, \ldots, q$, we have that $\bar{X}_j^{(n)} \xrightarrow{p} \sum_{k=1}^{K^*} \lambda_k \theta_{kj}$. Combining this fact with (S46) and (S45) yields:

$$\left(\frac{nq - q}{nq}\right) \hat{\sigma}^2(\mathbf{X}^{(n)}) \xrightarrow{p} \sigma^2 + \frac{1}{q}\sum_{j=1}^{q}\sum_{k=1}^{K^*} \lambda_k(\theta_{kj} - \tilde{\theta}_j)^2, \tag{S51}$$

where $\tilde{\theta}_j = \sum_{k=1}^{K^*} \lambda_k \theta_{kj}$ for all $j = 1, 2, \ldots, q$. It follows from (S51) that

$$\left(\frac{nq - q}{nq}\right) \hat{\sigma}^2(\mathbf{X}^{(n)}) \xrightarrow{p} \sigma^2 + c, \tag{S52}$$

for $c = \frac{1}{q}\sum_{j=1}^{q}\sum_{k=1}^{K^*} \lambda_k(\theta_{kj} - \tilde{\theta}_j)^2 > 0$. Applying the continuous mapping theorem to (S52) yields $\hat{\sigma}(\mathbf{X}^{(n)}) \xrightarrow{p} \sqrt{\sigma^2 + c}$, where $\sqrt{\sigma^2 + c} > \sigma$. Therefore, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\sigma}(\mathbf{X}^{(n)}) \geq \sigma\right) = 1$. □

The following result establishes an estimator of σ that satisfies the condition in Theorem 4, by making use of an independent and identically distributed copy of $\mathbf{X}^{(n)}$, if such a copy is available. If not, then sample-splitting can be used.

Corollary S1. *For $n = 1, 2, \ldots$, suppose that $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times q}(\boldsymbol{\mu}^{(n)}, \mathbf{I}_n, \sigma^2\mathbf{I}_q)$, and let $\mathbf{x}^{(n)}$ be a realization from $\mathbf{X}^{(n)}$, and that Assumptions 1 and 2 hold for some $K^* > 1$. For all $n = 1, 2, \ldots$, let $\hat{\mathcal{C}}_1^{(n)}$ and $\hat{\mathcal{C}}_2^{(n)}$ be a pair of estimated clusters in $\mathcal{C}(\mathbf{x}^{(n)})$. For all $n = 1, 2, \ldots$, let $\mathbf{Y}^{(n)}$ be an independent and identically distributed copy of $\mathbf{X}^{(n)}$. Then,*

$$\lim_{m,n\to\infty} \mathbb{P}_{H_0\left\{\hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)}\right\}}\left(\hat{p}(\mathbf{X}^{(n)}, \hat{\sigma}(\mathbf{Y}^{(m)})) \leq \alpha \mid \hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)} \in \mathcal{C}(\mathbf{X}^{(n)})\right) \leq \alpha, \text{ for all } 0 \leq \alpha \leq 1. \tag{S53}$$

Proof. By the independence of $\mathbf{X}^{(n)}$ and $\mathbf{Y}^{(n)}$ for all $n = 1, 2, \ldots$, and by Lemma S4,

$$\lim_{m,n\to\infty} \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1^{(n)},\hat{\mathcal{C}}_2^{(n)}\}}} \left(\hat{\sigma}(\mathbf{Y}^{(m)}) > \sigma \mid \hat{\mathcal{C}}_1^{(n)}, \hat{\mathcal{C}}_2^{(n)} \in \mathcal{C}(\mathbf{X}^{(n)}) \right) = 0.$$

Thus, (S53) follows from Theorem 4. □

S4 Additional simulation studies

S4.1 Null p-values when σ is unknown

In Section 4.3, we propose plugging an estimator of σ into the p-value defined in (8) to get (25), and we established conditions under which rejecting $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}$ when (25) is below α asymptotically controls the selective type I error rate (Theorem 4). Furthermore, in Section 4.3, we establish an estimator of σ that satisfisfies the conditions in Theorem 4 (Corollary S1). In the following, we will investigate the finite-sample implications of applying the approach outlined in Section S3.

We generate data from (1) with $n = 200$, $q = 10$, $\sigma = 1$, and two clusters,

$$\mu_1 = \cdots = \mu_{100} = \begin{bmatrix} \delta/2 \\ 0_{q-1} \end{bmatrix}, \mu_{101} = \cdots = \mu_{200} = \begin{bmatrix} 0_{q-1} \\ -\delta/2 \end{bmatrix}, \tag{S54}$$

for $\delta \in \{2, 4, 6\}$. We randomly split each data set into equally sized "training" and "test" sets. For each training set, we use average, centroid, single, and complete linkage hierarchical clustering to estimate three clusters, and then test $H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ for a randomly chosen pair of clusters. We estimate σ by applying $\hat{\sigma}(\cdot)$ in (S43) to the corresponding test set, then use it to compute p-values for average, centroid, and single linkage. In the case of complete linkage, we approximate (25) by replacing σ in the importance sampling procedure described in Section 4.1 with $\hat{\sigma}(\cdot)$ in (S43) applied to the corresponding test

set. Figure S7 displays QQ plots of the empirical distribution of the p-values against the Uniform$(0, 1)$ distribution, over 500 simulated data sets where $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ holds, for $\delta \in \{2, 4, 6\}$.

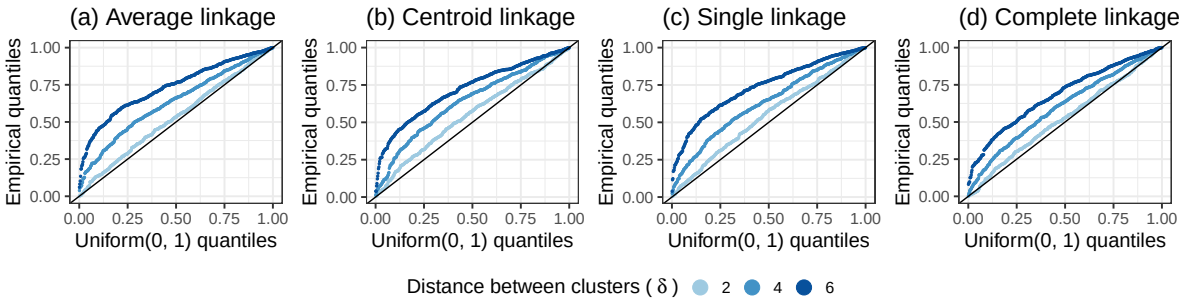

Figure S7: For the simulation study described in Section S4.1, QQ-plots of the p-values, using (a) average linkage, (b) centroid linkage, (c) single linkage, and (d) complete linkage, over 500 simulated data sets where $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ holds, for $\delta \in \{2, 4, 6\}$.

The p-values appear to be stochastically larger than the Uniform$(0, 1)$ distribution, across all linkages and all values of δ in (S54). As δ decreases and the clusters become less separated, $\hat{\sigma}(\cdot)$ in (S43) overestimates σ less severely. Thus, as δ decreases, the distribution of the p-values becomes closer to the Uniform$(0, 1)$ distribution.

S4.2 Power as a function of effect size

Recall that in Section 5.2, we considered the *conditional* power of the test proposed in Section 2.1 to detect a difference in means between two true clusters. We now evaluate the power of the test proposed in Section 2.1 to detect a difference in means between estimated clusters, without conditioning on having correctly estimated the true clusters.

We generate data from (1) with $n = 150$, $\sigma = 1$, $q = 10$, and μ given by (28). We simulate 10,000 data sets for nine evenly-spaced values of $\delta \in [3, 7]$. For each simulated data set, we test $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ for a randomly chosen pair of clusters obtained from single, average, centroid, and complete linkage hierarchical clustering, with significance level $\alpha = 0.05$. We define the *effect size* to be $\Delta = \|\bar{\mu}_{\hat{\mathcal{C}}_1} - \bar{\mu}_{\hat{\mathcal{C}}_2}\|_2/\sigma$ and consider the probability of rejecting $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ as a function of Δ. We smooth our power estimates by fitting a regression spline using the `gam` function in the `R` package `mgcv` (Wood 2015). Figure S8(a)–(d) displays the smoothed estimates when $\min\{|\hat{\mathcal{C}}_1|, |\hat{\mathcal{C}}_2|\} \geq 10$, and Figure S8(e)–(h) displays the smoothed estimates when $\min\{|\hat{\mathcal{C}}_1|, |\hat{\mathcal{C}}_2|\} < 10$. (We stratify our results because the power is much lower when $\min\{|\hat{\mathcal{C}}_1|, |\hat{\mathcal{C}}_2|\}$ is small.)

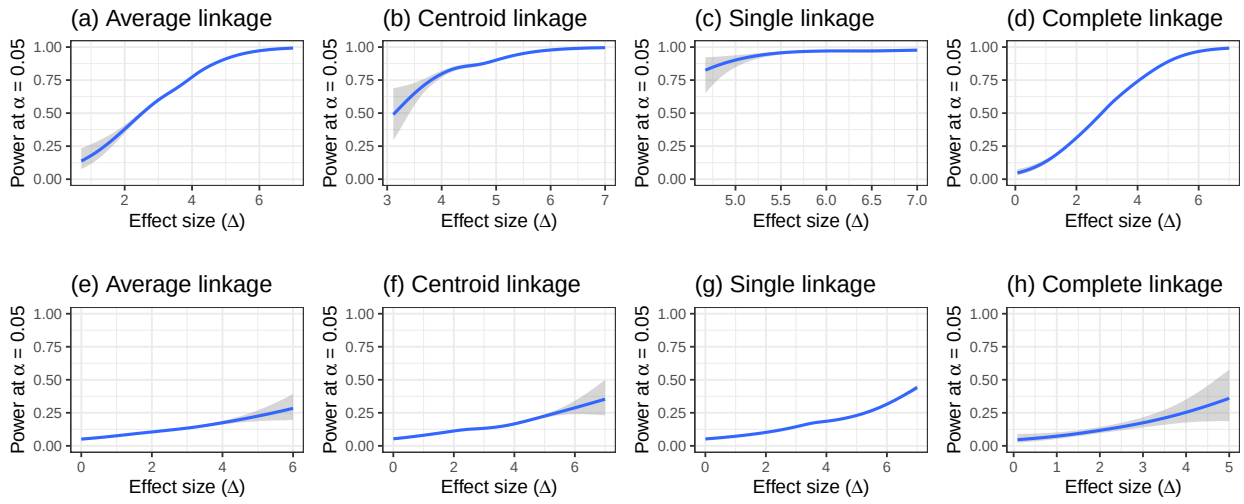

Figure S8: For the simulation study described in Section S4.2, smoothed power estimates as a function of the effect size $\Delta = \|\bar{\mu}_{\hat{\mathcal{C}}_1} - \bar{\mu}_{\hat{\mathcal{C}}_2}\|_2/\sigma$. In (a)–(d), $\min\{|\hat{\mathcal{C}}_1|, |\hat{\mathcal{C}}_2|\} \geq 10$, and in (e)–(h), $\min\{|\hat{\mathcal{C}}_1|, |\hat{\mathcal{C}}_2|\} < 10$.

It may come as a surprise that the x-axis on Figure S8(c) starts at 4.5, while the x-axis on Figure S8(d) starts at 0. This is because single linkage hierarchical clustering produces unbalanced clusters unless it successfully detects the true clusters. Thus, the condition $\min\{|\hat{\mathcal{C}}_1|, |\hat{\mathcal{C}}_2|\} \geq 10$ is only satisfied for single linkage hierarchical clustering when the true clusters are detected, which happens when Δ is greater than 4.5. The x-axis on Figure (b) starts at 3 rather than 0 for a similar reason.

In Figure S8, for all four linkages, the power to reject $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2}$ increases as the effect size (Δ) increases. All four linkages have similar power where the x-axes overlap.

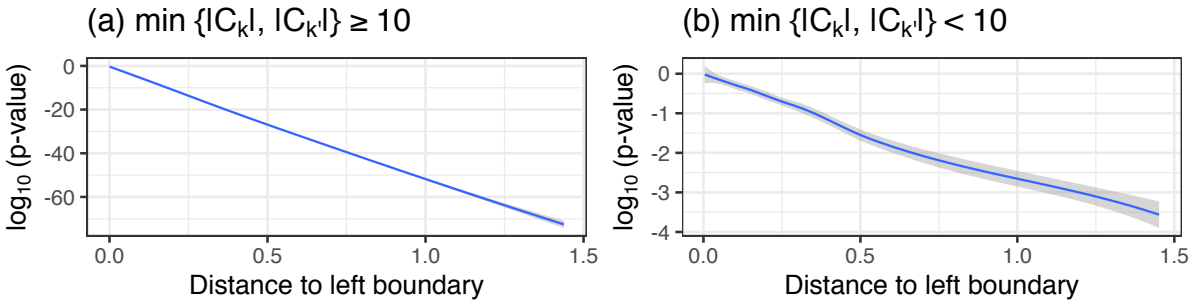

Figure S9: For the simulation study described in Section S4.2 with average linkage and $\Delta = \|\bar{\mu}_{\hat{\mathcal{C}}_1} - \bar{\mu}_{\hat{\mathcal{C}}_2}\|_2/\sigma = 5$, smoothed p-values on the $\log_{10}$ scale as a function of the distance between the test statistic and the left boundary of $\hat{\mathcal{S}}$, defined in (12).

Our test has low power when the test statistic ($\|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2$) is very close to the left boundary of $\hat{\mathcal{S}}$ defined in (12). This is a direct result of the fact that $p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \mathbb{P}(\phi \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 \mid \phi \in \hat{\mathcal{S}})$, for $\phi \sim (\sigma\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}}) \cdot \chi_q$ (Theorem 1). Figure S9 displays the smoothed p-values from average linkage hierarchical clustering with effect size $\Delta = \|\bar{\mu}_{\hat{\mathcal{C}}_1} - \bar{\mu}_{\hat{\mathcal{C}}_2}\|_2/\sigma = 5$ as a function of the distance between the test statistic and the left boundary of $\hat{\mathcal{S}}$. In Figure S9, the p-values increase as the distance increases. Similar results

have been observed in other selective inference frameworks; see Kivaranovic & Leeb (2020) for a detailed discussion.