Selective Inference for Hierarchical Clustering

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

† Department of Statistics, University of British Columbia

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

‡ Departments of Statistics and Biostatistics, University of Washington

August 18, 2022

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@stat.ubc.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 means between the putative cell types in that same data set; see e.g. Hwang et al. (2018). Unfortunately, available tests do not properly account for the double-use of data, which invalidates the resulting inference. One example of this problematic issue can be seen in the `FindMarkers` function in the popular and highly-cited `R` package `Seurat` (Satija et al. 2015, Butler et al. 2018, Stuart et al. 2019, Hao et al. 2021). Many recent papers have described the issues associated with the use of data for both clustering and downstream testing, without proposing suitable solutions (Luecken & Theis 2019, Lähnemann et al. 2020, Deconinck et al. 2021). In fact, testing for a difference in means between a pair of estimated clusters while controlling the type I error rate has even been described as one of eleven "grand challenges" in the entire field of single-cell data science (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. (We discuss the case where σ^2 is unknown in Section 4.3.) 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 $\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \overset{H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}}{\sim} \left(\sigma \sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}} \right) \cdot \chi_q$. However, since we clustered $\mathbf{x}$ to get $\mathcal{C}(\mathbf{x}) = \{\hat{\mathcal{C}}_k\}_{k=1}^K$, we will observe substantial differences between $\{\bar{x}_{\hat{\mathcal{C}}_k}\}_{k=1}^K$ even when there is no signal in the data, as is shown in Figure 1(a). That is, in (4), the random variable on the left-hand side of the inequality follows a scaled χ_q distribution, but the right-hand side of the inequality is *not* drawn from a scaled χ_q distribution, because $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are functions of $\mathbf{x}$. In short, the problem is that we used the data to select a null hypothesis to test. Since the Wald test does not account for this hypothesis selection procedure, it is extremely anti-conservative, as is shown in Figure 1(b).

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

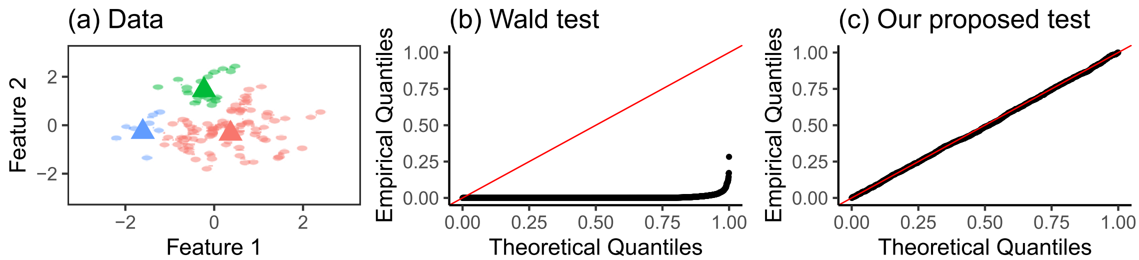

Figure 1: (a) A simulated data set from (1) with $\boldsymbol{\mu} = \mathbf{0}_{100\times 2}$ and $\sigma^2 = 1$. We apply average linkage hierarchical clustering to get three clusters. The empirical means (defined in (2)) of the three clusters are displayed as triangles. QQ-plots of the Uniform$(0,1)$ 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.

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).

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

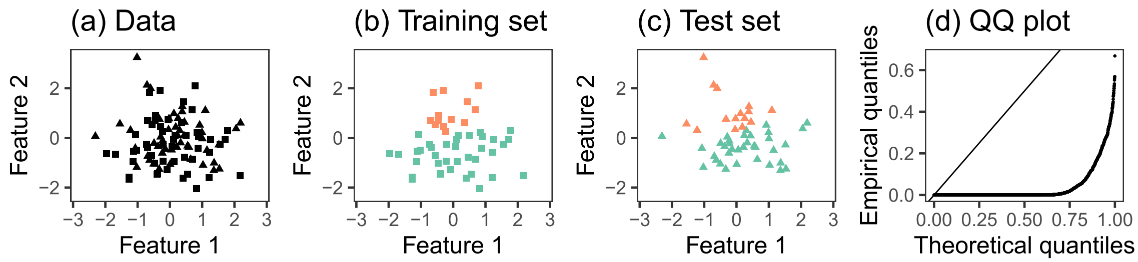

Figure 2: (a) A simulated data set from (1) with $\boldsymbol{\mu} = \mathbf{0}_{100\times 2}$ and $\sigma^2 = 1$. (b) We cluster the training set using average linkage hierarchical clustering. (c) We assign clusters in the test set by training a 3-nearest neighbors classifier on the training set. (d) 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.

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 assessing 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 $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}$ based on $\mathbf{X}$ controls the selective type I error rate for clustering* *at level* α *if*

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

That is, if $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ *is true, then the conditional probability of rejecting* $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ *based on* $\mathbf{X}$ *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) cannot be calculated, 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

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

where $\mathrm{dir}(w) = \frac{w}{\|w\|_2}\mathbb{1}\{w \neq 0\}$. 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

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

Furthermore, if $H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}$ is true, then

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

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

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

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

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

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

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

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 $0 \le \phi < \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2$) 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 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

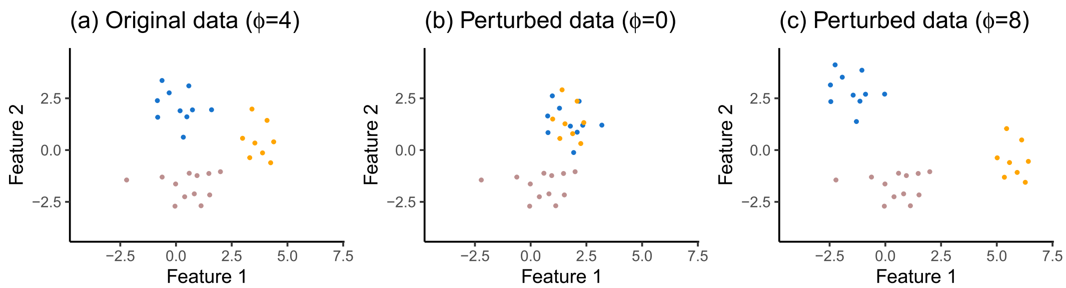

Figure 3: The observations belonging to $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$ are displayed in blue and orange for: (a) the original data set $\mathbf{x} = \mathbf{x}'(\phi)$ with $\phi = \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 = 4$, (b) a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

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 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.

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 $\mathcal{C}^{(t+1)}(\mathbf{x}) = \mathcal{C}^{(t)}(\mathbf{x}) \cup \left\{\mathcal{W}_1^{(t)}(\mathbf{x}) \cup \mathcal{W}_2^{(t)}(\mathbf{x})\right\} \setminus \left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}$.

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

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

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

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

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

	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).

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 corresponding 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 $h_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = d\left(\mathcal{W}_1^{(u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}))}(\mathbf{x}),\mathcal{W}_2^{(u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}))}(\mathbf{x});\mathbf{x}\right)$. 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 $\mathcal{M}_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) = \left\{t : l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \leq t < u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}), d\left(\mathcal{W}_1^{(t)}(\mathbf{x}),\mathcal{W}_2^{(t)}(\mathbf{x});\mathbf{x}\right) > d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}),\mathcal{W}_2^{(t+1)}(\mathbf{x});\mathbf{x}\right)\right\}$ 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* $\mathcal{M}(\mathbf{x}) = \left\{1 \leq t \leq n-K : d\left(\mathcal{W}_1^{(t)}(\mathbf{x}),\mathcal{W}_2^{(t)}(\mathbf{x});\mathbf{x}\right) < d\left(\mathcal{W}_1^{(t+1)}(\mathbf{x}),\mathcal{W}_2^{(t+1)}(\mathbf{x});\mathbf{x}\right)\right\}$*,* *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 $d(\mathcal{G},\mathcal{G}';\mathbf{x}) = \min\limits_{i\in\mathcal{G},i'\in\mathcal{G}'} d(\{i\},\{i'\};\mathbf{x})$. Applying this definition to (18) yields $\hat{\mathcal{S}} = \bigcap\limits_{\{\mathcal{G},\mathcal{G}'\}\in\mathcal{L}(\mathbf{x})} \bigcap\limits_{i\in\mathcal{G}} \bigcap\limits_{i'\in\mathcal{G}'} \{\phi \geq 0 : d(\{i\},\{i'\};\mathbf{x}'(\phi)) > h_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\}$, 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 $\text{Cov}(X_i) = \sigma^2 \mathbf{I}_q$. In this subsection, we instead assume that $\mathbf{x}$ is a realization from

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

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

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

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

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

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

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

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 $\tilde{a}_{ii'} = \left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)^2\left(\frac{\|\mathbf{x}^T\hat{\nu}\|_2}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}^T\hat{\nu}\|_2}\right)^2$, $\tilde{b}_{ii'} = 2\left(\left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)\left(\frac{\|\mathbf{x}^T\hat{\nu}\|_2}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}^T\hat{\nu}\|_2}\right)\langle \mathrm{dir}(\mathbf{x}^T\hat{\nu}), x_i - x_{i'}\rangle - \tilde{a}_{ii'}\|\mathbf{\Sigma}^{-1/2}\mathbf{x}^T\hat{\nu}\|_2\right)$, and $\tilde{c}_{ii'} = \left\|x_i - x_{i'} - \left(\frac{\hat{\nu}_i - \hat{\nu}_{i'}}{\|\hat{\nu}\|_2^2}\right)(\mathbf{x}^T\hat{\nu})\right\|_2^2$.

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:

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

Intuitively, if we plug in a consistent estimate of σ, we might expect to asymptotically control the selective type I error rate. For example, this is true in the context of selective inference for low-dimensional linear regression (Tian & Taylor 2017), where the error variance can be consistently estimated under mild assumptions.

Unfortunately, consistently estimating σ in (1) presents a major challenge. Similar issues arise when estimating the error variance in high-dimensional linear regression; see e.g. Reid et al. (2016). Thus, we adopt a similar approach to Tibshirani et al. (2018), which studied the theoretical implications of plugging in a simple over-estimate of the error variance in the context of selective inference for high-dimensional linear regression. In the following result, we show that plugging in an asymptotic over-estimate of σ in (25) leads to asymptotic control of 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.

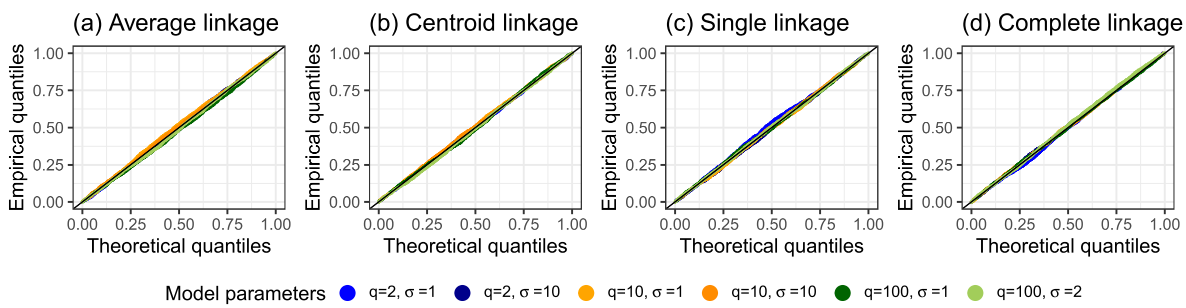

Figure 4: For 2000 draws from (1) with $\boldsymbol{\mu} = \mathbf{0}_{n\times q}$, $n = 150$, $q \in \{2, 10, 100\}$, and $\sigma \in \{1, 2, 10\}$, QQ-plots of the p-values obtained from the test proposed in Section 2.1, using (a) average linkage, (b) centroid linkage, (c) single linkage, and (d) complete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

5.2 Conditional power and recovery 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 $H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}$ for a randomly-chosen pair of estimated clusters, given that the randomly-chosen pair of estimated clusters correspond to two true clusters. Since this conditions on the event that the randomly-chosen pair of estimated clusters correspond to two true clusters, we are also interested in how often this event occurs. We therefore consider the *recovery probability*, the probability that the randomly-chosen pair of estimated clusters correspond to two true clusters. Figure 5 displays the conditional power and recovery probability as a function of the distance between the true clusters (δ).

Figure 5 displays conditional power and recovery probability as a function of the distance between the true clusters (δ). For all four linkages, the conditional power and recovery

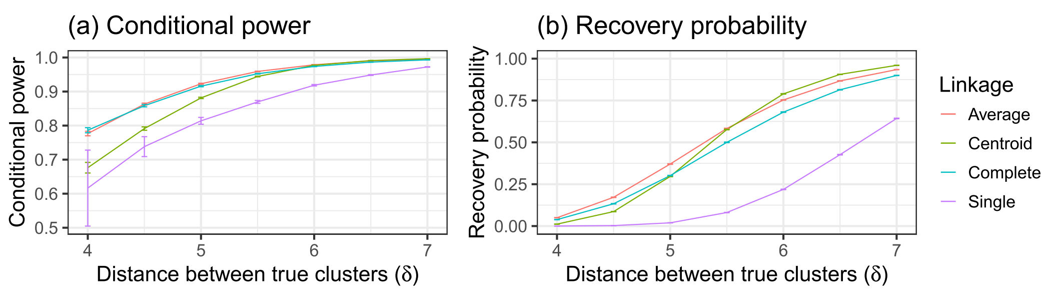

Figure 5: For the simulation study described in Section 5.2, (a) conditional power of the test proposed in Section 2 versus the difference in means between the true clusters (δ), and (b) recovery probability versus δ. Conditional power and recovery probability are defined in Section 5.2.

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 recovery 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.

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 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

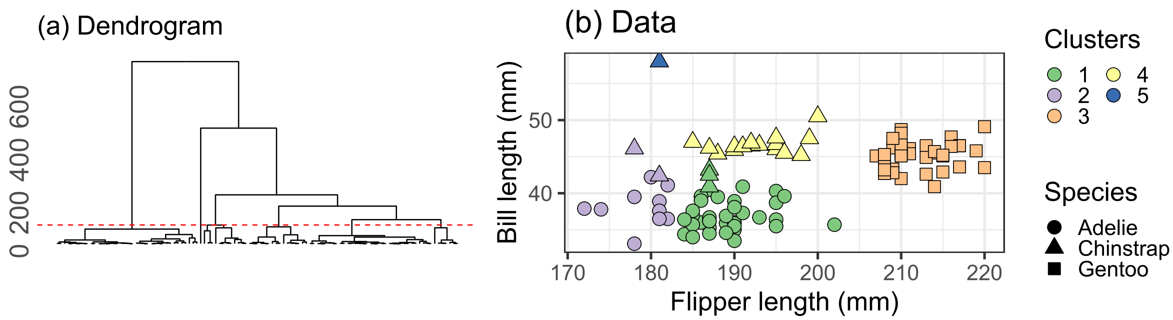

Figure 6: (a) The average-linkage hierarchical clustering dendrogram and (b) the bill lengths and flipper lengths, as well as the true species labels and estimated clusters, for the Palmer penguins data described in Section 6.1.

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 S3 in Section S4.2 of the supplement.

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).

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 gene expression between the clusters, without properly accounting for the fact that the clusters were estimated from the data (Luecken & Theis 2019, Lähnemann et al. 2020, Deconinck et al. 2021). 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

$$\mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\left((\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2})^T \boldsymbol{\Sigma}^{-1}(\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}) \geq (\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2})^T \boldsymbol{\Sigma}^{-1}(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2})\right). \tag{29}$$

For both tests, we estimate $\boldsymbol{\Sigma}$ 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"1		
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 (29) 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 $\boldsymbol{\Sigma}$ 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). This directly solves a problem that routinely occurs when biologists analyze data, e.g. when testing for differential expression between clusters estimated from single-cell RNA-sequencing data (Luecken & Theis 2019, Lähnemann et al. 2020, Deconinck et al. 2021).

The framework proposed in Section 2 assumed that σ in (1) was known. Similarly, the extended framework in Section 4.2 assumed that $\mathbf{\Sigma}$ in (23) was known. These assumptions are unlikely to hold in practice. Thus, in the data applications of Sections 6.1–6.2, we replaced σ in (1) and $\mathbf{\Sigma}$ in (23) with estimates. In Section 4.3, we explored the theoretical implications of this replacement, under model (1). Future work could include investigating how best to estimate σ in (1) and $\mathbf{\Sigma}$ in (23). Another potential avenue for future work would be to develop an analogue of Theorem 4 in Section 4.3 under model (23).

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 R01-GM123993. Jacob Bien was supported by NSF CAREER Award DMS-1653017. Daniela Witten was supported by NSF CAREER Award DMS-1252624, NIH Grants R01-EB026908 and R01-DA047869, and an Simons Investigator Award in Mathematical Modeling of Living Systems (No. 560585).

Conflict of interest statement

The authors have no relevant financial or non-financial competing interests to report.

References

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. (2018), 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology* **36**(5), 411–420.

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.

Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B. & Saeys, Y. (2021), 'Recent advances in trajectory inference from single-cell 'omics data', *Current Opinion in Systems Biology* **27**, 100344.

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* .

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M. et al. (2021), 'Integrated analysis of multimodal single-cell data', *Cell* **184**(13), 3573–3587.

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 changepoint 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.

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* .

Luecken, M. D. & Theis, F. J. (2019), 'Current best practices in single-cell RNA-seq analysis: a tutorial', *Molecular Systems Biology* **15**(6), e8746.

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.

Reid, S., Tibshirani, R. & Friedman, J. (2016), 'A study of error variance estimation in lasso regression', *Statistica Sinica* pp. 35–67.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. (2015), 'Spatial reconstruction of single-cell gene expression data', *Nature Biotechnology* **33**(5), 495–502.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. (2019), 'Comprehensive integration of single-cell data', *Cell* **177**(7), 1888–1902.

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

Tian, X. & Taylor, J. (2017), ‘Asymptotics of selective inference’, *Scandinavian Journal of Statistics* **44**(2), 480–499.

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.

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.