Testing for a difference in means after selection

Yiqun Chen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Daniela Witten, Chair
Alex Luedtke
Amy Willis

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Testing for a difference in means after selection

Yiqun Chen

Chair of the Supervisory Committee:
Professor Daniela Witten
Departments of Statistics and Biostatistics

In modern data analysis, we often want to test for a difference in means between groups selected based on the observed data. This is a challenging task: when the null hypothesis is selected based on the data, classical tests (e.g., a z-test) that do not account for this will fail to control the Type I error. In this dissertation, we leverage the selective inference framework to develop valid tests for a difference in means when the groups under investigation are selected based on the output of a statistical learning method.

We first consider the task of quantifying the uncertainty of spikes estimated from calcium imaging data. Here, the scientific question can be cast as a test of equality of (weighted) means between groups that are defined through a changepoint detection algorithm. Next, we describe a new test of a null hypothesis that is selected based on the output of the graph fused lasso. Our proposal conditions on less information than existing approaches, thereby leading to higher power while guaranteeing Type I error control. The final chapter is motivated by statistical challenges that arise in single-cell transcriptomics studies, where researchers are interested in ascertaining whether the estimated clusters are truly different from each other. We develop a finite-sample, correctly-sized test for a difference in cluster means when the clusters are obtained via k-means clustering, and demonstrate that our method leads to conclusions that align better with the underlying truth than classical tests.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

ACKNOWLEDGMENTS

I would not be able to finish this dissertation without the help of many people who have supported me over the past five years.

First and foremost, I would like to thank my advisor, Daniela Witten, for her unwavering commitment to my success. Her sharp statistical insights, constant encouragement, and wisdom about life have been truly transformative — I could not have asked for a better mentor, and I inspire to be as good a mentor to others as she has been to me.

I would like to extend my thanks to many mentors and collaborators, who have made the last few years an incredibly intellectually-stimulating experience: Sean Jewell for collaborating on Chapters 2–3 with me and serving as a de facto committee member; Alex Luedtke and Amy Willis for serving on my reading committee and having fun conversations when I first arrived at the University of Washington; Jamie Morgenstern for introducing me to the fascinating field of fairness in machine learning and sharing her insight on my committee; Kelley Harris for serving as the GSR; Lucy Gao for her advice, insights, and keeping me company on Witten group Slack; René Just for our collaboration in rigorous statistical analysis for software engineering and teaching me how to give a stellar conference presentation; Katharina Reinecke, Alexandra To, and Angela DR Smith for the dream team that has supported me in my race and ethnicity research; the SEARCH Collaboration, especially Lillian Brown, Laura Balzer, Carina Marquez, and Maya Petersen, whose dedication to rigorous science and real-world impact will never cease to inspire me; Katie Wilson for serving as a role model for teaching the future biostatisticians and epidemiologists; Katie Kerr for showing me the fun and beauty of statistical consulting and for mentoring me as a teaching assistant; Lin He for teaching me how to succeed and navigate the "tech world" during my summer internship;

my undergraduate mentors, Wenjing Zheng and Will Fithian, for showing me a taste of the research world.

Next round of applauses goes to many others who have made me laugh a little louder and cry a little less, including but certainly not limited to: Yiying Ruan and Christine Liu who have been my ride-or-die since day 1 — we lift each other up when no one else have even noticed that we have fallen; Tina Huang, Cindy Zheng, Tianyi Liu, Yutong Liu for sticking with me since freshman year of college; Tina Zheng, Wan Fung Chui, Forest Yang for serving as my inspiration ever since CS 70; Bill Gao and Ribery Gu for unconditional support since high school; the UW Biostatistics community (a special shout out to Lucy Li, Doris Li, and Eric Morenz); Brandon Wong for great summer memories and fun trips.

A special thank you to Andre Le for always standing by me and believing in us. You told me that education has to be uncomfortable, yet you have held my hand through every struggle.

Lastly, I want to thank my family for their love and support. They have never voiced a single doubt on my pursuit, and their support was a constant that replenishes me during every step of my education, especially the Ph.D. years. Mom and dad (mama and baba), thank you for everything — I love you so so much.

DEDICATION

To my mom, Xianjun He, and grandmother, Xiuqin Wang

Chapter 1

INTRODUCTION

As the availability of large datasets has drastically increased in the past few decades, the goal of data analysis in many applications has shifted from testing *pre-specified* hypotheses to testing *data-driven* hypotheses. In particular, data analysts may want to perform some exploratory data analyses to generate new hypotheses or to select a subset of "interesting" hypotheses, and then test those hypotheses *on the same data*. For instance, in single-cell RNA-sequencing analyses, researchers often first cluster the cells, and then test for a difference in the expected gene expression levels between the clusters to quantify up- or down-regulation of genes, annotate known cell types, and identify new cell types [Aizarani et al., 2019, Doughty and Kerkhoven, 2020, Golub, 2010, Grün et al., 2015, Peters et al., 2014, Zhang et al., 2019]. However, testing hypotheses suggested by the data is a challenging task: once we have used the data to generate the null hypotheses, standard statistical inference tools — e.g., a t-test for a difference in means between two groups — will *fail to control the Type I error* [Fithian et al., 2014]. In fact, the inferential challenges resulting from testing data-guided hypotheses have been described as a "grand challenge" in the field of genomics [Lähnemann et al., 2020]. State-of-the-art toolkits in the field continue to overlook this issue [Stuart et al., 2019, Van den Berge et al., 2020], and one goes so far as to state that "Rather than attaching strong probabilistic interpretations to the p-values, we view the p-values simply as useful numerical summaries for ranking the genes for further inspection." Testing data-guided hypothesis also arises in the field of neuroscience [Button, 2019, Kriegeskorte et al., 2009], social psychology [Hung and Fithian, 2020], and physical sciences [Friederich et al., 2020, Pollice et al., 2021].

This dissertation aims to bridge the gap between the practice of testing data-driven

hypotheses and the existing statistical toolkits for testing such hypotheses. In particular, we are interested in developing a test of the null hypothesis H_0 that controls the selective Type I error [Fithian et al., 2014], when H_0 is chosen based on the output of a statistical learning method applied to the data. That is, we wish to ensure that the probability of rejecting H_0 at level α, given that H_0 holds and we decided to test it, is no greater than α:

$$\mathbb{P}_{H_0}(\text{reject } H_0 \text{ at level } \alpha \mid H_0 \text{ is tested}) \leq \alpha, \quad \text{for all } \alpha \in (0,1). \tag{1.1}$$

To tackle this problem, we take a selective inference (also known as post-selection inference and conditional inference) approach, which has been applied extensively in high-dimensional linear modeling [Charkhi and Claeskens, 2018, Fithian et al., 2014, Lee et al., 2016, Loftus and Taylor, 2014, Rügamer et al., 2022, Schultheiss et al., 2021, Taylor and Tibshirani, 2018, Tibshirani et al., 2016, Yang et al., 2016], changepoint detection [Benjamini et al., 2019, Chen et al., 2021b, Duy et al., 2020, Hyun et al., 2018, 2021, Jewell et al., 2022, Le Duy and Takeuchi, 2021], and correcting selection bias [Hung and Fithian, 2020, Weinstein et al., 2013, Zhong and Prentice, 2008, Zollner and Pritchard, 2007]. In a nutshell, the key insight of selective inference is as follows: to obtain a valid test of H_0, we need to *condition* on the aspect of the data that led us to test it. In Chapters 2–4, we leverage this insight to develop correctly-sized tests for a difference in (weighted) means between two groups that are *defined through the observed data*, e.g., via the output of a changepoint detection or clustering algorithm. In addition to our methodological contributions, we demonstrate the practical use of our tests on calcium imaging data, single-cell RNA-sequencing data, and datasets of drug overdose death rates and teenage birth rates in the contiguous United States.

In Chapter 2, we consider the task of quantifying the uncertainty of spikes estimated from calcium imaging data. Calcium imaging is an increasingly important technology in the field of neuroscience, which allows for simultaneous recording from huge numbers of neurons in behaving animals [Ahrens et al., 2013, Chen et al., 2013, Prevedel et al., 2014]. When a neuron spikes, calcium floods the cell. Due to the use of fluorescent calcium indicators,

this leads to a rapid increase in fluorescence. Thus, for a single neuron, calcium imaging results in a time series of fluorescence intensities that serves as a first-order approximation for the unobserved spike times. Typically, the scientific interest lies in the unobserved spike times, rather than the observed fluorescence traces. In recent years, several algorithms have been developed to estimate a neuron's spike times based on its observed fluorescence traces [Deneux et al., 2016, Friedrich and Paninski, 2016, Friedrich et al., 2017, Jewell and Witten, 2018, Jewell et al., 2019, Pnevmatikakis et al., 2013, Theis et al., 2016, Vogelstein et al., 2010]. However, quantifying the uncertainty associated with these estimated spikes remains an open problem. In particular, after estimating spike times from calcium imaging data, it is natural to consider testing the null hypothesis that the spike did *not* occur — i.e., that there is no increase in calcium — at a timepoint at which we estimated a spike. This is a challenging task, because we will observe a large difference in fluorescence intensities before and after an estimated spike, even in the absence of a true spike. To tackle this problem, we build on a well-studied model for calcium imaging data, which states that calcium decays exponentially in the absence of a spike, and instantaneously increases when a spike occurs [Friedrich and Paninski, 2016, Friedrich et al., 2017, Jewell and Witten, 2018, Jewell et al., 2019, Vogelstein et al., 2010]. We propose a computationally-efficient test for the change in calcium associated with an estimated spike. This test accounts for the fact that the spikes are estimated from the same data used for testing, and as a result, provides selective Type I error control. Moreover, we illustrate the use of this proposed approach to a publicly-available, benchmark calcium imaging dataset [Theis et al., 2016]. The work in Chapter 2 is in press in *Biostatistics* [Chen et al., 2021b].

In Chapter 3, we tackle the problem of testing hypotheses that are a function of the output of the graph fused lasso [Tibshirani and Taylor, 2011], a widely-popular method to reconstruct signals that are *piecewise constant* on a graph, with broad applications in genomics [Gong et al., 2018, Kim and Xing, 2009], imaging analysis [Kang et al., 2018, Xin et al., 2014], and so forth. Its solution can be segmented into *connected components* — that is, elements that share a common value. It's natural to consider testing the null

hypothesis that the means of two connected components estimated from the graph fused lasso are equal. For instance, when the graph fused lasso is used to denoise the chromatin interactions across the entire genome [Gong et al., 2018], testing this hypothesis corresponds to the scientific question of whether two *estimated* regions of the genome have the same rate of chromatin interactions. Since the hypothesis is itself a function of the same data used for testing, naive procedures such as a two-sample z-test fail to control the selective Type I error. To overcome this problem, Hyun et al. [2018] proposed a test that controls the selective Type I error by quantifying the probability of observing the such a large difference in the sample means, conditional on all outputs of the algorithm used to obtain the graph fused lasso solution. However, this proposal conditions on far more information than is needed to determine the null hypothesis under consideration, thereby leading to extremely *low power* in practice [Jewell et al., 2022, Le Duy and Takeuchi, 2021, Liu et al., 2018]. This motivates us to propose a new test for this task that conditions on less information than existing approaches, which leads to substantially higher power while guaranteeing selective Type I error control. In a nutshell, we compute the probability of observing such a large difference in the sample means, conditional only on *the pair of connected components* used to construct the null hypothesis. Our method leads to higher power in simulation, as well as more discoveries on datasets of drug overdose death rates and teenage birth rates in the contiguous United States [Centers for Disease Control and Prevention, 2020a,b]. This chapter is based on Chen et al. [2021a], which has been tentatively accepted by the *Journal of Computational and Graphical Statistics* [Chen et al., 2021a].

In Chapter 4, we investigate the problem of testing for a difference in means between two groups defined via the output of k-means clustering [Lloyd, 1982, MacQueen et al., 1967], an extremely popular method to partition the observations into clusters, each represented by the empirical mean, referred to as the *centroid*, for the observations in that cluster. For instance, in single-cell transcriptomics studies, where the datasets consist of millions of unlabeled cells, estimated clusters via k-means clustering are commonly used to approximate the underlying biological groups (e.g., different cell types or cell states). After obtaining the cluster labels,

researchers are often interested in testing for a difference in means between the centroids to ascertain whether the identified cell types are truly different from each other. This is a challenging task, because the cluster labels, and consequently the null hypothesis, are a function of the data, and classical tests that do not account for this will *fail to control the Type I error*. In this chapter, we propose a finite-sample test that controls the selective Type I error for a difference in means between a pair of clusters obtained using k-means clustering. In addition, we show that the proposed p-value can be efficiently computed, and demonstrate that, when applied to hand-written digits data and single-cell RNA-sequencing data, our method leads to conclusions that align better with the underlying truth than classical tests. The content in Chapter 4 is adapted from Chen and Witten [2022].

We close with a discussion in Chapter 5. Detailed proofs of technical results, as well as some additional information, can be found in the Appendix.

Chapter 2

QUANTIFYING UNCERTAINTY IN SPIKES ESTIMATED FROM CALCIUM IMAGING DATA

2.1 Introduction

In the field of neuroscience, recent advances in calcium imaging have enabled recording from large populations of neurons *in vivo* [Ahrens et al., 2013, Chen et al., 2013, Prevedel et al., 2014]. When a neuron spikes, calcium floods the cell; the presence of fluorescent calcium indicator molecules causes it to fluoresce. Thus, for each neuron, calcium imaging results in a time series of fluorescence intensities that can be seen as a noisy approximation to its unobserved spike times. Typically, the neuron's observed fluorescence trace is not of scientific interest; instead, the interest lies in the unobserved spike times.

A number of methods have been developed to estimate spike times from the fluorescence trace of a neuron [Berens et al., 2018, Jewell and Witten, 2018, Jewell et al., 2019, Pachitariu et al., 2018, Stringer and Pachitariu, 2019, Theis et al., 2016, Vogelstein et al., 2010]. One line of work makes use of a simple model that relates the unobserved calcium c_t and the observed fluorescence Y_t at the tth time step [Friedrich and Paninski, 2016, Jewell and Witten, 2018, Jewell et al., 2019, Vogelstein et al., 2010],

$$\begin{aligned} Y_t &= c_t + \epsilon_t, \quad \epsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad t = 1, \ldots, T, \\ c_t &= \gamma c_{t-1} + z_t, \quad t = 2, \ldots, T, \end{aligned} \tag{2.1}$$

where $z_t \geq 0$ for all t, and $z_t > 0$ indicates the presence of a spike at the tth time step. At most time steps, $z_t = 0$, corresponding to no spike. Between spikes, calcium decays exponentially at a rate $\gamma \in (0, 1)$; γ can be viewed as a property of the calcium indicator, and is taken to be known. Model (2.1) suggests estimating the underlying calcium c_t by

solving the optimization problem

$$\underset{c_1,\ldots,c_T\geq 0;\ z_1,\ldots,z_T}{\text{minimize}} \left\{ \frac{1}{2}\sum_{t=1}^{T}(y_t - c_t)^2 + \lambda \sum_{t=2}^{T} 1\{z_t \neq 0\} \right\} \text{ subject to } z_t = c_t - \gamma c_{t-1} \geq 0 \ , \quad (2.2)$$

where $\lambda \geq 0$ is a tuning parameter that trades off the number of estimated spikes and the fit to the observed fluorescence [Jewell and Witten, 2018], and $1\{A\}$ is the indicator function that equals 1 if the event A occurs, and 0 otherwise. The ℓ_0 penalty $\sum_{t=2}^{T} 1\{z_t \neq 0\}$ is non-convex, which has motivated a number of authors to consider a convex relaxation to (2.2) using an ℓ_1 penalty [Friedrich and Paninski, 2016, Friedrich et al., 2017, Vogelstein et al., 2010]. An efficient dynamic programming algorithm that yields the global optimum to (2.2) has also been proposed [Jewell and Witten, 2018, Jewell et al., 2019].

Despite the extensive literature on estimating a neuron's spike times from its fluorescence intensity [Jewell and Witten, 2018, Jewell et al., 2019, Pachitariu et al., 2018, Theis et al., 2016, Vogelstein et al., 2010], quantifying the uncertainty associated with these estimated spikes remains in large part an open problem. More precisely, suppose we observe a T-vector of fluorescence intensities under model (2.1), and estimate the J spike times $\hat{\tau}_1, \ldots, \hat{\tau}_J$. For fixed $j \in \{1, \ldots, J\}$, consider testing whether there is a spike at $\hat{\tau}_j$, i.e.,

$$H_0 : c_{\hat{\tau}_j+1} - \gamma c_{\hat{\tau}_j} = 0 \quad \text{versus} \quad H_1 : c_{\hat{\tau}_j+1} - \gamma c_{\hat{\tau}_j} > 0, \quad (2.3)$$

where the one-sided alternative reflects the fact that a spike leads to an *increase* (rather than a decrease) in calcium. Despite the apparent simplicity of (2.3), obtaining a test with correct size requires care. For instance, motivated by a Wald test, we can consider the p-value

$$\mathbb{P}_{H_0}\left(Y_{\hat{\tau}_j+1} - \gamma Y_{\hat{\tau}_j} \geq y_{\hat{\tau}_j+1} - \gamma y_{\hat{\tau}_j}\right), \quad (2.4)$$

where $y_1, \ldots, y_T$ is the observed fluorescence, and (2.1) implies that $Y_{\hat{\tau}_j+1} - \gamma Y_{\hat{\tau}_j} \sim \mathcal{N}\left(0, (1+\gamma^2)\sigma^2\right)$ under H_0. But this naive approach ignores the fact that estimation (2.2) and inference (2.3) for $\hat{\tau}_j$ were performed *on the same data* [Button, 2019, Fithian et al., 2014]. Thus, even in the absence of a true spike, we will observe a large value of $y_{\hat{\tau}_j+1} - \gamma y_{\hat{\tau}_j}$; see Figure 2.1(a). Figure 2.1(b) demonstrates that (2.4) does not control the *selective Type I error*: the prob-

ability of a false rejection given that we decided to test this null hypothesis [Fithian et al., 2014].

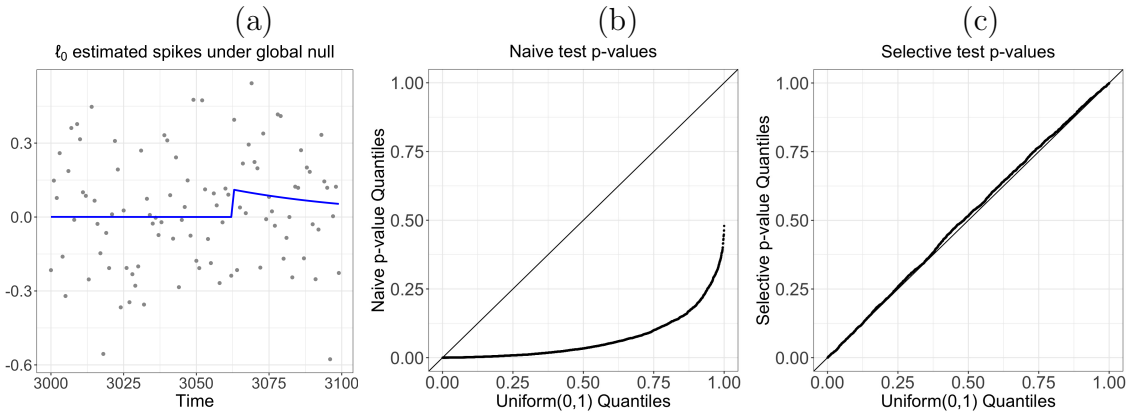


Figure 2.1: *(a):* One simulation with $y_1, \ldots, y_{10,000}$ (grey dots) generated according to model (2.1) with $\gamma = 0.98$, $\sigma = 0.2$, and $z_t = 0$ for all t. The ℓ_0 problem in (2.14) was solved with $\lambda = 0.1$, resulting in 47 estimated spikes with fluorescence increases. Estimated calcium is displayed in blue. We display one estimated spike at time $\hat{\tau} = 3,060$ with $y_{3,000}, \ldots, y_{3,100}$. *(b):* Quantile-quantile plot for the Wald p-values (defined in (2.4)) based on 100 simulations (2,988 hypothesis tests). *(c):* Quantile-quantile plot for the selective p-values (defined in (2.9) with $h = 1$) based on 100 simulations (2,988 hypothesis tests).

In this paper, we leverage the *selective inference* framework, which enables us to test a null hypothesis that was selected using the data, to develop a valid test for (2.3). Related approaches have been developed for a number of problems, including penalized regression [Fithian et al., 2014, Lee et al., 2016, Tibshirani et al., 2016] and changepoint detection [Hyun et al., 2021, Jewell et al., 2022]. In a nutshell, to obtain a test that controls the selective Type I error, we condition on the aspect of the data that led us to test this particular null hypothesis. In particular, since we have chosen to test the null hypothesis $H_0 : c_{\hat{\tau}_j+1} - \gamma c_{\hat{\tau}_j} = 0$ in (2.3) because $\hat{\tau}_j$ is an estimated changepoint, our p-value should be computed *conditional on the event that $\hat{\tau}_j$ is an estimated changepoint.* As seen in Figure 2.1(c), this results in a test that controls the selective Type I error.

Some authors have considered quantifying the uncertainty in the location of an estimated

spike $\hat{\tau}_j$ [Merel et al., 2016, Pnevmatikakis et al., 2016]. Others have applied a Bayesian lens to the uncertainty associated with the magnitude of the change in calcium associated with an estimated spike $\hat{\tau}_j$ [Deneux et al., 2016, Merel et al., 2016, Pnevmatikakis et al., 2016, Soltanian-Zadeh et al., 2018, Theis et al., 2016, Vogelstein et al., 2009]. Despite the flexibility and robustness of Bayesian methods, they do not provide a straightforward way to test (2.3). First, they provide an uncertainty estimate for the change of calcium at *every timepoint*. As a result, we still need to account for selection if we only choose to test the null hypothesis for the estimated spikes [Yekutieli, 2012]. Second, even with appropriate adjustments, Bayesian hypothesis testing typically will not control Type I error [Ghosh, 2011].

The current paper is closely related to the literature on changepoint detection. Jewell and Witten [2018] showed that (2.2) is equivalent to a changepoint detection problem, which allows us to tap into the toolbox of inferential procedures for changepoint detection [Fryzlewicz, 2014, Harchaoui and Lévy-Leduc, 2010, Song et al., 2016, Yao, 1988, Yao and Au, 1989, Zou et al., 2020]. Despite the abundant literature on this topic, a few gaps remain to be filled, as reviewed in Niu et al. [2016]: (i) much of the prior work has focused on quantifying the uncertainty associated with either the number or locations of the estimated changepoints; and (ii) most existing inferential procedures are asymptotic and approximate. Two recent exceptions include Hyun et al. [2021] and Jewell et al. [2022], which took a selective inference approach and computed finite-sample p-values for testing the changes in mean around changepoints estimated using an ℓ_1 and an ℓ_0 penalty, respectively. Our work is closest to Jewell et al. [2022], and extends their proposal to the model (2.1).

In this paper, we propose a general framework to quantify the uncertainty associated with the set of spikes estimated from calcium imaging data, using *any* spike detection algorithm. Our testing framework controls the selective Type I error associated with the null hypothesis (2.3). However, in practice it might be very hard to carry out this framework for an arbitrary spike detection algorithm. Thus, in the special case of spikes estimated by solving a variant of the ℓ_0 optimization problem in (2.2), we provide an algorithm that can be used to efficiently compute p-values and confidence intervals associated with these estimated spikes.

The rest of this paper is organized as follows. In Section 2.2, we detail the null hypothesis of interest, and develop a framework to test it for spikes estimated using *any* spike estimation procedure, under model (2.1). We develop an efficient algorithm to compute the p-values for spikes estimated via a variant of (2.2) in Section 2.3, and develop confidence intervals in Section 2.4. We apply our proposal in a simulation study in Section 2.5, and to calcium imaging data in Section 2.6. The discussion is in Section 2.7. Proofs and other technical details are relegated to the Appendix.

Throughout this paper, upper case Y denotes a random variable, and lower case y denotes a realization of Y. For a vector $\nu \in \mathbb{R}^T$, $\|\nu\|_2$ denotes its ℓ_2 norm, $\nu^\top$ its transpose, and $\Pi_\nu^\perp$ the projection matrix onto its orthogonal complement, i.e., $\Pi_\nu^\perp = I - \frac{\nu\nu^\top}{\|\nu\|_2^2}$ where I denotes the identity matrix. We use $\mathbb{N}$ to denote the natural numbers and $\mathbb{R}$ to denote the real numbers. The notation $1\{\cdot\}$ and $\stackrel{d}{=}$ denote an indicator function and equality in distribution, respectively.

2.2 Selective inference for spike detection

2.2.1 Defining the null hypothesis

We wish to test for an increase in calcium at $\hat{\tau}_j$, an estimated spike time. With $\nu \in \mathbb{R}^T$ defined as

$$\nu_t = \begin{cases} -\gamma, & t = \hat{\tau}_j, \\ 1, & t = \hat{\tau}_j + 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2.5}$$

we can re-write (2.3) as

$$H_0 : \nu^\top c = 0 \text{ versus } H_1 : \nu^\top c > 0. \tag{2.6}$$

However, (2.5) only considers the two timepoints immediately before and after $\hat{\tau}_j$, leaving most data unused. In order to take advantage of a larger data window, we will generalize

the contrast vector ν under a simple assumption.

Assumption 1: There are no spikes within a window of $\pm h$ of $\hat{\tau}_j$. In other words,
$$\gamma^h c_{\hat{\tau}_j-h+1} = \gamma^{h-1} c_{\hat{\tau}_j-h+2} = \ldots = \gamma c_{\hat{\tau}_j} \text{ and } c_{\hat{\tau}_j+1} = c_{\hat{\tau}_j+2}\gamma^{-1} = \ldots = \gamma^{-h+1} c_{\hat{\tau}_j+h}.$$

Under Assumption 1, and treating $\hat{\tau}_j$ as fixed, the log likelihood of $Y_{\hat{\tau}_j-h+1}, \ldots, Y_{\hat{\tau}_j}$ is proportional to $\sum_{t=\hat{\tau}_j-h+1}^{\hat{\tau}_j} \left(Y_t - c_{\hat{\tau}_j}\gamma^{t-\hat{\tau}_j}\right)^2$. Thus, the maximum likelihood estimator for $c_{\hat{\tau}_j}$ is $\hat{c}_{\hat{\tau}_j} = \frac{\gamma^2-1}{\gamma^2-\gamma^{-2h+2}} \sum_{t=\hat{\tau}_j-h+1}^{\hat{\tau}_j} Y_t \gamma^{t-\hat{\tau}_j}$. Similarly, using the h observations $Y_{\hat{\tau}_j+1}, \ldots, Y_{\hat{\tau}_j+h}$, the maximum likelihood estimator for $c_{\hat{\tau}_j+1}$ is $\hat{c}_{\hat{\tau}_j+1} = \frac{\gamma^2-1}{\gamma^{2h}-1} \sum_{t=\hat{\tau}_j+1}^{\hat{\tau}_j+h} Y_t \gamma^{t-(\hat{\tau}_j+1)}$. This suggests that we can test for an increase in calcium at $\hat{\tau}_j$ using (2.6) with ν defined as

$$\nu_t = \begin{cases} -\frac{\gamma(\gamma^2-1)}{\gamma^2-\gamma^{-2h+2}}\gamma^{t-\hat{\tau}_j}, & \hat{\tau}_j - h + 1 \le t \le \hat{\tau}_j, \\ \frac{\gamma^2-1}{\gamma^{2h}-1}\gamma^{t-(\hat{\tau}_j+1)}, & \hat{\tau}_j + 1 \le t \le \hat{\tau}_j + h, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

Details of the form of ν if $\hat{\tau}_j + h > T$ or $\hat{\tau}_j - h + 1 < 1$, as well as a visualization of ν in (2.7), are provided in Appendix A.2.

2.2.2 A selective test for $H_0 : \nu^\top c = 0$ versus $H_1 : \nu^\top c > 0$

Suppose that we test for an increase in calcium only at timepoints at which (i) we estimate a spike; and (ii) there is an increase in fluorescence associated with this estimated spike. This motivates the following p-value to test (2.6):

$$\mathbb{P}_{H_0}\left(\nu^\top Y \ge \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(Y), \nu^\top Y > 0\right), \tag{2.8}$$

where $\mathcal{M}(Y)$ is the set of spikes estimated from Y. Roughly speaking, this p-value answers the question: *Assuming that there is no true spike at $\hat{\tau}_j$, what's the probability of observing such a large increase in fluorescence at $\hat{\tau}_j$, given that we decided to test for a spike at $\hat{\tau}_j$?*

The p-value in (2.8) leads to a test that controls the *selective Type I error* [Fithian et al., 2014], i.e., the probability of *falsely rejecting the null hypothesis, given that the we decided*

to conduct the test. However, computing (2.8) is hard because the conditional distribution of $\nu^\top Y$ given $\hat{\tau}_j(y) \in \mathcal{M}(Y)$ and $\nu^\top Y > 0$ depends on the nuisance parameter $\Pi_\nu^\perp c$, where $\Pi_\nu^\perp = I - \nu\nu^\top/\|\nu\|_2^2$ is the projection matrix onto the orthogonal complement of ν. Therefore, to overcome the computational difficulty, we further condition on $\left\{\Pi_\nu^\perp Y = \Pi_\nu^\perp y\right\}$ to eliminate the dependence on the nuisance parameter, arriving at the p-value:

$$p = \mathbb{P}_{H_0}\left(\nu^\top Y \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(Y), \nu^\top Y > 0, \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right). \tag{2.9}$$

Following arguments in Section 5 of Lee et al. [2016] and Section 3 of Fithian et al. [2014], the test that rejects the null hypothesis in (2.6) when (2.9) is less than α controls the selective Type I error at level α. This p-value is the focus of this paper.

Proposition 1. *Suppose that* $Y \sim \mathcal{N}(c, \sigma^2 I)$. *Then,*

$$\begin{aligned} &\mathbb{P}\left(\nu^\top Y \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(Y), \nu^\top Y > 0, \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\ &= \mathbb{P}\left(\phi \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(y'(\phi)), \phi > 0\right), \end{aligned} \tag{2.10}$$

for $\phi \sim \mathcal{N}\left(\nu^\top c, \sigma^2||\nu||_2^2\right)$, *where*

$$y'(\phi) = \Pi_\nu^\perp y + \phi \cdot \frac{\nu}{||\nu||_2^2} = y + \left(\frac{\phi - \nu^\top y}{||\nu||_2^2}\right)\nu. \tag{2.11}$$

Furthermore, for p *defined in* (2.9), *and* $\phi_0 \sim \mathcal{N}(0, \sigma^2||\nu||_2^2)$,

$$p = \mathbb{P}\left(\phi_0 \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(y'(\phi_0)), \phi_0 > 0\right). \tag{2.12}$$

It follows that to compute the p-value in (2.9), we must characterize the set

$$\mathcal{S} = \{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}. \tag{2.13}$$

Of course, the practical details of computing the set (2.13) will depend on the function $\mathcal{M}(\cdot)$ that yields the estimated spikes. The task of characterizing the set (2.13) is the focus of Section 2.3.

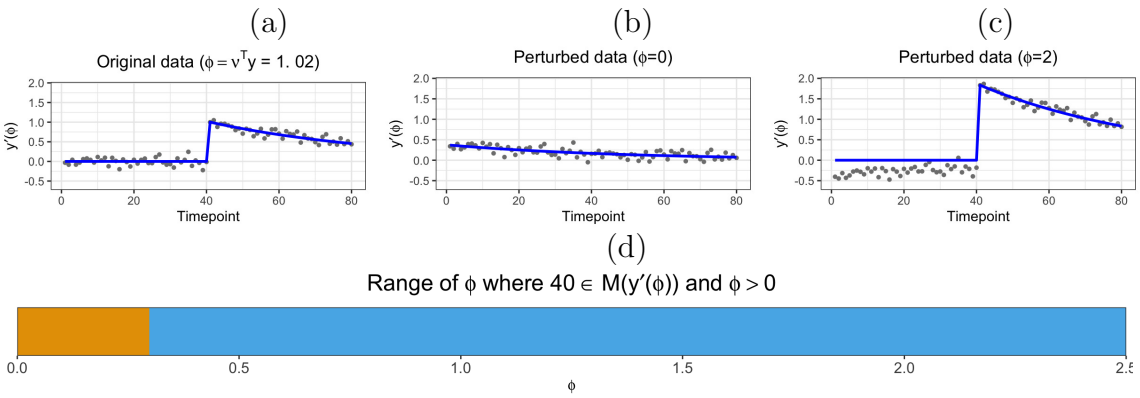


Figure 2.2: Data generated according to (2.1), with $T = 80, \sigma = 0.1, \gamma = 0.98$, and one spike at $t = 40$. Solving the ℓ_0 problem (2.14) with $\lambda = 0.75$ yields a single estimated spike at $t = 40$. *(a):* We plot the original data, which corresponds to $y'(\phi)$ with $\phi = \nu^\top y = 1.02$, where ν is constructed according to (2.7) with $\hat{\tau}_j = 40$ and $h = 40$. The estimated calcium concentration is displayed in blue. *(b):* The perturbed dataset $y'(\phi)$ with $\phi = 0$ is shown. Now there is no increase in calcium at $t = 40$ on $y'(\phi)$, and no spike is estimated. *(c):* The perturbed dataset $y'(\phi)$ with $\phi = 2$ is shown. There is now a very pronounced increase in calcium at $t = 40$, and a spike is estimated. *(d):* The set of ϕ for which $40 \in \mathcal{M}(y'(\phi))$ and $\phi > 0$ is displayed in blue; other values of ϕ are in orange.

In (2.11), $y'(\phi)$ results from perturbing y by a function of ϕ along the direction defined by ν. Elements of y that fall outside of the support of ν are not perturbed. Then, $\mathcal{S}$ in (2.13) is the set of ϕ such that applying $\mathcal{M}(\cdot)$ to the perturbed data $y'(\phi)$ results in an estimated spike at $\hat{\tau}_j$.

As an example, we generate data from (2.1) with $T = 80$, $\sigma = 0.1$, and $\gamma = 0.98$ with a true spike at $t = 40$, and $c_{41} - \gamma c_{40} = 1$. This results in $\phi = \nu^\top y = 1.02$. Solving the optimization problem in (2.2) with $\lambda = 0.75$ results in a single estimated spike at $t = 40$, which means that $\mathcal{S} = \{\phi : 40 \in \mathcal{M}(y'(\phi))\}$. The set-up is displayed in Figure 2.2(a). In panel (b), we perturb the observed data with $\phi = 0$. Now a spike is no longer estimated at $t = 40$, so $0 \notin \mathcal{S}$. In panel (c), we perturb the observed data with $\phi = 2$ to exaggerate the increase in fluorescence; now a spike is estimated at $t = 40$, so $2 \in \mathcal{S}$. In panel (d), we

display the set $\mathcal{S} \cap (0, +\infty) = (0.29, +\infty)$.

2.3 Computation of the selective p-value

Proposition 1 indicates that we can compute the p-value defined in (2.9) provided that we are able to compute the set $\mathcal{S}$ defined in (2.13). In this section, we will show that $\mathcal{S}$ can be efficiently computed for spikes estimated by solving a variant of the ℓ_0 optimization problem in (2.2) that omits the positivity constraint $c_t - \gamma c_{t-1} \geq 0$: namely,

$$\underset{c_1,\ldots,c_T \geq 0}{\text{minimize}} \left\{ \frac{1}{2} \sum_{t=1}^{T} (y_t - c_t)^2 + \lambda \sum_{t=2}^{T} 1\{c_t \neq \gamma c_{t-1}\} \right\}. \tag{2.14}$$

In Section 2.3.1, we briefly review the work of Jewell and Witten [2018] and Jewell et al. [2019], who showed that the solution to (2.14) can be characterized through a recursion involving piecewise quadratic functions. The rest of this section is quite technical. An overview is as follows:

- In Section 2.3.2, we introduce functions $C(\phi)$ and $C'(\phi)$ such that $\mathcal{S} = \{\phi : C(\phi) \leq C'(\phi)\}$.
- Then, in Section 2.3.3, we show that $C(\phi)$ and $C'(\phi)$ are piecewise quadratic in ϕ.
- We can therefore apply approaches from Rigaill [2015] and Maidstone et al. [2017] for efficient manipulation of piecewise quadratic functions, to efficiently compute $\{\phi : C(\phi) \leq C'(\phi)\}$, and in turn, $\mathcal{S}$ in (2.13).

2.3.1 An algorithm to solve (2.14)

Jewell and Witten [2018] noted that (2.14) is equivalent to a changepoint detection problem,

$$\underset{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T,k}{\text{minimize}} \left\{ \sum_{j=0}^{k} \min_{\alpha \geq 0} \left\{ \frac{1}{2} \sum_{t=\tau_j+1}^{\tau_{j+1}} (y_t - \alpha \gamma^{t-\tau_{j+1}})^2 \right\} + \lambda k \right\}, \tag{2.15}$$

in the sense that $\{t : \hat{c}_{t+1} - \gamma\hat{c}_t \neq 0\} = \{\hat{\tau}_1, \ldots, \hat{\tau}_J\}$, where $\hat{c}_1, \ldots, \hat{c}_T$ and $\hat{\tau}_1, \ldots, \hat{\tau}_J, J$ are solutions to (2.14) and (2.15), respectively. Furthermore, let $F(s)$ denote the optimal objective of (2.15) for the first s data points $y_{1:s} = (y_1, \ldots, y_s)$, and define

$$\text{Cost}(y_{1:s}, \alpha; \gamma) = \min_{0 \leq \tau < s} \left\{ F(\tau) + \left\{ \frac{1}{2} \sum_{t=\tau+1}^{s} \left(y_t - \alpha\gamma^{t-s}\right)^2 \right\} + \lambda \right\}. \tag{2.16}$$

In words, $\text{Cost}(y_{1:s}, \alpha; \gamma)$ is the optimal cost of partitioning the data $y_{1:s}$ into exponentially decaying regions with decay parameter γ, given that the calcium at the sth timepoint equals α. It turns out that $\text{Cost}(y_{1:s}, \alpha; \gamma)$ admits a recursion that can be solved efficiently, which provides intuition for characterizing the set $\mathcal{S}$ in the next section.

Proposition 2 (Proposition 1 and Section 2.2.3 in Jewell et al. [2019])**.** *For* $\text{Cost}(y_{1:s}, \alpha; \gamma)$ *defined in* (2.16) *and* $\text{Cost}(y_1, \alpha; \gamma)$ *defined as* $\frac{1}{2}(y_1 - \alpha)^2$*, the following recursion holds:*

$$\text{Cost}(y_{1:s}, \alpha; \gamma) = \min \left\{ \text{Cost}\big(y_{1:(s-1)}, \alpha/\gamma; \gamma\big), \min_{\alpha' \geq 0} \text{Cost}\big(y_{1:(s-1)}, \alpha'; \gamma\big) + \lambda \right\} + \frac{1}{2}(y_s - \alpha)^2, \tag{2.17}$$

. Also, $\text{Cost}(y_{1:s}, \alpha; \gamma)$ *is a piecewise quadratic function of* α*.*

In words, the recursion in (2.17) considers the following two possibilities: (i) there is no spike at the $(s-1)$th time point, in which case the calcium decays exponentially, and the cost equals $\text{Cost}\big(y_{1:(s-1)}, \alpha/\gamma; \gamma\big)$; (ii) there is a spike at the $(s-1)$th time point, and the cost equals the optimal cost up to $s-1$, $\min_{\alpha' \geq 0} \text{Cost}\big(y_{1:(s-1)}, \alpha'; \gamma\big)$, plus the cost of placing a changepoint, λ.

Building on Proposition 2, Jewell et al. [2019] made use of the recent literature on *functional pruning* [Maidstone et al., 2017, Rigaill, 2015] to efficiently compute the cost functions $\text{Cost}(y_{1:s}, \alpha; \gamma)$, as a function of α, using clever manipulations of the piecewise quadratic functions involved in the recursion (2.17). This approach has a worst-case complexity of $O(s^2)$, and is often much faster in practice. Once the cost functions have been computed, it is straightforward to identify the changepoints in (2.15), and, in turn, the spikes in (2.14). Details are provided in Section 2.2 of Jewell et al. [2019].

2.3.2 *Characterizing $\mathcal{S}$ for spikes estimated using (2.14)*

In what follows, we leverage ideas from Jewell et al. [2022] to develop an efficient algorithm to analytically characterize (2.13), i.e., the set of values ϕ such that solving (2.14) on perturbed data $y'(\phi)$ yields an estimated spike $\hat{\tau}_j$. Throughout this section, we define $y_{1:s} = (y_1, \ldots, y_s)$, $y_{T:s} = (y_T, \ldots, y_s)$, $y'_{1:s}(\phi) = ([y'(\phi)]_1, \ldots, [y'(\phi)]_s)$, and $y'_{T:s}(\phi) = ([y'(\phi)]_T, \ldots, [y'(\phi)]_s)$.

Let $\mathcal{M}(y)$ denote the spikes estimated by applying (2.14) to the data y. To begin, we characterize the set $\mathcal{S}$ using the $\text{Cost}(y_{1:s}, \alpha; \gamma)$ function defined in (2.16).

Proposition 3. *Let $\{\hat{\tau}_1, \ldots, \hat{\tau}_J\} = \{t : \hat{c}_{t+1} - \gamma \hat{c}_t \neq 0\}$ be the timesteps of the estimated spikes from (2.14). For $\text{Cost}(y_{1:s}, \alpha; \gamma)$ in (2.16), we have that*

$$C(\phi) = \min_{\alpha \geq 0} \left\{ \text{Cost}\left(y'_{1:\hat{\tau}_j}(\phi), \alpha; \gamma \right) \right\} + \min_{\alpha' \geq 0} \left\{ \text{Cost}\left(y'_{T:(\hat{\tau}_j+1)}(\phi), \alpha'; 1/\gamma \right) \right\} + \lambda \tag{2.18}$$

equals the objective of (2.15) applied to data $y'(\phi)$, subject to the constraint that $\hat{\tau}_j$ is an estimated spike. Furthermore,

$$C'(\phi) = \min_{\alpha \geq 0} \left\{ \text{Cost}\left(y'_{1:\hat{\tau}_j}(\phi), \alpha; \gamma \right) + \text{Cost}\left(y'_{T:(\hat{\tau}_j+1)}(\phi), \gamma\alpha; 1/\gamma \right) \right\} \tag{2.19}$$

equals the objective of (2.15) applied to data $y'(\phi)$, subject to the constraint that $\hat{\tau}_j$ is not an estimated spike. Moreover, for $\mathcal{S}$ defined in (2.13),

$$\mathcal{S} = \{\phi : C(\phi) \leq C'(\phi)\}. \tag{2.20}$$

Therefore, to characterize $\mathcal{S}$ in (2.13), it suffices to characterize $C(\phi)$ in (2.18) and $C'(\phi)$ in (2.19). To do this, we will leverage the toolkit from Jewell et al. [2022] to analytically characterize $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$ as a function of both ϕ and α. While this is related to the task of efficiently characterizing $\text{Cost}(y_{1:s}, \alpha; \gamma)$ in terms of α in Section 2.3.1, it is substantially more challenging, due to the presence of the additional parameter ϕ.

2.3.3 *Efficient computation of* $\mathcal{S}$ *via* $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$

While Proposition 2 cannot be directly applied to $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$, we can arrive at a very similar result by adapting Theorem 2 from Jewell et al. [2022].

Proposition 4. *For* $\hat{\tau}_j - h + 1 \leq s \leq \hat{\tau}_j$ *and* $y'(\phi)$ *defined in* (3.12),

$$\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma) = \min_{f \in \mathcal{C}_s} f(\alpha, \phi), \tag{2.21}$$

where $\mathcal{C}_s$ *is a collection of* $s - \hat{\tau}_j + h + 1$ *piecewise quadratic functions of* α *and* ϕ *constructed with the initialization*

$$\mathcal{C}_{\hat{\tau}_j - h} = \left\{ \text{Cost}(y'_{1:(\hat{\tau}_j - h)}(\phi), \alpha; \gamma) \right\}, \tag{2.22}$$

and the recursion

$$\mathcal{C}_s = \left(\bigcup_{f \in \mathcal{C}_{s-1}} \left\{ f(\alpha/\gamma, \phi) + \frac{1}{2}(y'_s(\phi) - \alpha)^2 \right\} \right) \bigcup \left\{ g_s(\phi) + \frac{1}{2}(y'_s(\phi) - \alpha)^2 \right\}, \tag{2.23}$$

where

$$g_s(\phi) = \min_{f \in \mathcal{C}_{s-1}} \min_{\alpha \geq 0} f(\alpha, \phi) + \lambda. \tag{2.24}$$

Proposition 4 applies when $\hat{\tau}_j - h \geq 1$; Appendix A.6 details the extension for $\hat{\tau}_j - h < 1$. Proposition 4 indicates that $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$ is in fact a bivariate piecewise quadratic function of both ϕ and α (in contrast to a univariate piecewise quadratic function of α, as in $\text{Cost}(y_{1:s}, \alpha; \gamma)$). Moreover, $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$ can be efficiently computed with the recursion in (2.23).

To compute $C(\phi)$ in (2.18), we first use Proposition 4 to compute the collection $\mathcal{C}_{\hat{\tau}_j}$ such that $\text{Cost}\left(y'_{1:\hat{\tau}_j}(\phi), \alpha; \gamma\right) = \min_{f \in \mathcal{C}_{\hat{\tau}_j}} f(\alpha, \phi)$. Using a slight modification of Proposition 4 (see Proposition 18 in Appendix A.5), we also compute the collection $\tilde{\mathcal{C}}_{\hat{\tau}_j + 1}$ such that

$\text{Cost}\left(y'_{T:(\hat{\tau}_j+1)}(\phi), \alpha'; 1/\gamma\right) = \min_{f\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}} f(\alpha', \phi)$. Then, we have that

$$\begin{aligned} C(\phi) &\overset{a.}{=} \min_{\alpha\geq 0}\left\{\min_{f\in\mathcal{C}_{\hat{\tau}_j}} f(\alpha,\phi)\right\} + \min_{\alpha'\geq 0}\left\{\min_{f\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}} f(\alpha',\phi)\right\} + \lambda \\ &\overset{b.}{=} \min_{f\in\mathcal{C}_{\hat{\tau}_j}}\left\{\min_{\alpha\geq 0} f(\alpha,\phi)\right\} + \min_{f\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}}\left\{\min_{\alpha'\geq 0} f(\alpha',\phi)\right\} + \lambda. \end{aligned} \tag{2.25}$$

Here, $a.$ follows from combining the definition of $C(\phi)$ in (2.18) with the expression for $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$ in (2.21) and the expression for $\text{Cost}(y'_{T:s}(\phi), \alpha; 1/\gamma)$ in Appendix A.5; $b.$ follows from changing the order of minimizations. Furthermore, since Proposition 4 states that the functions in $\mathcal{C}_{\hat{\tau}_j}$ are piecewise quadratic in ϕ and α, it follows that $\min_{\alpha\geq 0} f(\alpha, \phi)$ is a piecewise quadratic function of ϕ only. A similar result in Appendix A.5 guarantees that the functions in $\tilde{\mathcal{C}}_{\hat{\tau}_j+1}$ are piecewise quadratic in ϕ and α; therefore, for each $f \in \tilde{\mathcal{C}}_{\hat{\tau}_j+1}$, we have that $\min_{\alpha'\geq 0} f(\alpha', \phi)$ is piecewise quadratic in ϕ. Because minimization and summation over piecewise quadratic functions yields a piecewise quadratic function, it follows that $C(\phi)$ is piecewise quadratic in ϕ.

We now consider computing $C'(\phi)$ in (2.19). Plugging in the expressions for $\text{Cost}(y'_{1:s}(\phi), \alpha; \gamma)$ in (2.21) and $\text{Cost}(y'_{T:s}(\phi), \alpha; 1/\gamma)$ in Appendix A.5 into (2.19), we have

$$\begin{aligned} C'(\phi) &= \min_{\alpha\geq 0}\left\{\min_{f\in\mathcal{C}_{\hat{\tau}_j}} f(\alpha,\phi) + \min_{f\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}} f(\gamma\alpha,\phi)\right\} \\ &= \min_{\alpha\geq 0}\left\{\min_{f\in\mathcal{C}_{\hat{\tau}_j}, \tilde{f}\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}}\left\{f(\alpha,\phi) + \tilde{f}(\gamma\alpha,\phi)\right\}\right\} \\ &= \min_{f\in\mathcal{C}_{\hat{\tau}_j}, \tilde{f}\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}}\left\{\min_{\alpha\geq 0}\left\{f(\alpha,\phi) + \tilde{f}(\gamma\alpha,\phi)\right\}\right\}. \end{aligned} \tag{2.26}$$

By Proposition 4 and Appendix A.5, both $f(\alpha, \phi)$ and $\tilde{f}(\gamma\alpha, \phi)$ are piecewise quadratic in α and ϕ, which implies that $\min_{\alpha\geq 0}\left\{f(\alpha, \phi) + \tilde{f}(\gamma\alpha, \phi)\right\}$ is a piecewise quadratic function of ϕ. Therefore, $C'(\phi)$ is the minimum over a set of piecewise quadratic functions of ϕ, and thus is itself piecewise quadratic in ϕ.

Finally, since both $C(\phi)$ and $C'(\phi)$ are piecewise quadratic in ϕ, we can apply ideas from the functional pruning literature to compute the set $\mathcal{S} = \{\phi : C(\phi) \leq C'(\phi)\}$ efficiently [Maidstone et al., 2017, Rigaill, 2015]. The procedure and computation time are summarized in Algorithm 2 (see Appendix A.7) and Proposition 5, respectively.

Proposition 5. *Once* $\text{Cost}\big(y_{1:(\hat{\tau}_j-h)}, \alpha; \gamma\big)$ *and* $\text{Cost}\big(y_{T:(\hat{\tau}_j+h+1)}, \alpha; 1/\gamma\big)$ *have been computed, Algorithm 2 can be performed in* $O(h^2)$ *operations.*

The worst-case complexity of computing $\text{Cost}\big(y_{1:(\hat{\tau}_j-h)}, \alpha; \gamma\big)$ and $\text{Cost}\big(y_{T:(\hat{\tau}_j+h+1)}, \alpha; 1/\gamma\big)$ is $O(T^2)$, but it is often much faster in practice [Jewell et al., 2019]. Furthermore, $\text{Cost}(y_{1:(\hat{\tau}_j-h)}, \alpha; \gamma)$ was already computed to solve (2.14). Therefore, estimating J changepoints via (2.14) and then computing their corresponding p-values has a worst-case computation time of $O(T^2 + Jh^2)$, and is often much faster in practice. An empirical analysis of the timing complexity of Algorithm 2 can be found in Appendix A.9. We walk through Algorithm 2 on a small example in Appendix A.10.

2.4 Confidence intervals with correct selective coverage

We now construct a $(1-\alpha)$ confidence interval for $\nu^\top c$, the change in calcium associated with an estimated spike $\hat{\tau}_j$.

Proposition 6. *Suppose that* (2.1) *holds, and let* $\hat{\tau}_j$ *denote a spike estimated by solving* (2.14)*. For a given value of* $\alpha \in (0,1)$*, define functions* $\theta_L(t)$ *and* $\theta_U(t)$ *such that*

$$F^{\mathcal{S}\cap(0,\infty)}_{\theta_L(t),\sigma^2||\nu||_2^2}(t) = 1 - \frac{\alpha}{2}, \quad F^{\mathcal{S}\cap(0,\infty)}_{\theta_U(t),\sigma^2||\nu||_2^2}(t) = \frac{\alpha}{2}, \tag{2.27}$$

where $F^{\mathcal{S}\cap(0,\infty)}_{\mu,\sigma^2}(t)$ *is the cumulative distribution function of a normal distribution with mean* μ *and variance* σ^2*, truncated to the set* $\mathcal{S} \cap (0,\infty)$*. Then* $[\theta_L(\nu^\top Y), \theta_U(\nu^\top Y)]$ *is a* $(1-\alpha)$ *confidence interval for* $\nu^\top c$*, in the sense that*

$$\mathbb{P}\Big(\nu^\top c \in \big[\theta_L(\nu^\top Y), \theta_U(\nu^\top Y)\big] \;\Big|\; \hat{\tau}_j(y) \in \mathcal{M}(Y), \nu^\top Y > 0, \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big) = 1 - \alpha. \tag{2.28}$$

Thus, the confidence interval guarantees coverage *conditional on the selection procedure* [Fithian et al., 2014, Lee et al., 2016, Tibshirani et al., 2016]. Computing θ_L (and θ_U) in (2.27) amounts to a root-finding problem, which can be solved, e.g., using bisection [Chen and Bien, 2020].

2.5 Simulation study

Recall that our selective inference framework involves testing the null hypothesis of no increase in calcium at timepoints for which the following two conditions hold: (i) this timepoint was an estimated spike in the solution to (2.14); and (ii) $\nu^\top y > 0$ for this particular timepoint. We let $\{\tilde{\tau}_1, \ldots, \tilde{\tau}_M\}$ denote the set of timepoints satisfying these two conditions, i.e., the set of timepoints to be tested using our selective inference approach. That is,

$$\{\tilde{\tau}_1, \ldots, \tilde{\tau}_M\} = \left\{\hat{\tau}_1, \ldots, \hat{\tau}_J : \nu^\top y > 0\right\}, \tag{2.29}$$

where $\{\hat{\tau}_1, \ldots, \hat{\tau}_J\}$ denotes the set of spikes estimated from (2.14). (2.29) slightly abuses notation, since ν in (2.7) is a function of $\hat{\tau}_j$. Therefore, the right-hand side of (2.29) should be interpreted as the estimated spike times associated with an increase in fluorescence in a window of $\pm h$.

2.5.1 Selective Type I error control under the global null

We simulated $y_1, \ldots, y_{10,000}$ according to (2.1) with $\gamma = 0.98$, $\sigma = 0.2$, and $z_t = 0$ for all $t = 2, \ldots, 10,000$. Thus, the null hypothesis $H_0 : \nu^\top c = 0$ holds for all contrast vectors ν defined in (2.7), regardless of the timepoint being tested, and the value of h in (2.7).

We solved (2.14) with the tuning parameter λ selected to yield $J = 100$ estimated spikes; thus, $J = 100$ in (2.29). Then, for each $\hat{\tau}_j$, we constructed four contrast vectors ν, defined in (2.7), corresponding to $h \in \{1, 2, 10, 20\}$. Then, provided that $\nu^\top y > 0$, we computed the selective p-values in (2.9) and the naive (Wald) p-values defined as

$$\mathbb{P}\big(\nu^\top Y \geq \nu^\top y\big). \tag{2.30}$$

The results, aggregated over 1,000 simulations, are displayed in Figure 2.3. Panels (a) and

(b) display quantile-quantile plots of the naive and selective p-value quantiles versus the Uniform(0,1) quantiles, respectively; we see that for all values of h, (i) the naive procedure in (2.30) is anti-conservative; and (ii) the proposed selective test based on (2.9) controls the selective Type I error.

2.5.2 Power and detection probability

Recall that we test $H_0 : \nu^\top c = 0$ only for timepoints in the set $\{\tilde{\tau}_1, \ldots, \tilde{\tau}_M\}$ defined in (2.29). Therefore, we separately consider the *conditional power* of the proposed test [Hyun et al., 2021, Jewell et al., 2022] and the *detection probability* of the spike estimation procedure.

Given a dataset $y = (y_1, \ldots, y_T)$ with K true spikes $\tau_1, \ldots, \tau_K$, and recalling the definition in (2.29), we define the *conditional power* to be the ratio between (i) the number of true spikes for which the nearest null hypothesis among those tested (i.e., the set $\{\tilde{\tau}_1, \ldots, \tilde{\tau}_M\}$ in (2.29)) is within b timepoints of the true spike *and* has a p-value less than α; and (ii) the number of true spikes for which the nearest tested hypothesis falls within b timepoints:

$$\text{Conditional power} = \frac{\sum_{i=1}^{K} 1\left(p_{m(i)} \leq \alpha, |\tau_i - \tilde{\tau}_{m(i)}| \leq b\right)}{\sum_{i=1}^{K} 1\left(|\tau_i - \tilde{\tau}_{m(i)}| \leq b\right)}, \tag{2.31}$$

where $m(i) = \operatorname{argmin} m|\tau_i - \tilde{\tau}_m|$ indexes the timepoint to be tested that is closest to the ith true spike time, and $p_{m(i)}$ is the corresponding p-value. Since (2.31) conditions on the event that the closest tested timepoint $\tilde{\tau}_{m(i)}$ is within b timepoints of the true spike time τ_i, we also consider the *detection probability*, which tells us how often this event occurs:

$$\text{Detection probability} = \frac{\sum_{i=1}^{K} 1\left(|\tau_i - \tilde{\tau}_{m(i)}| \leq b\right)}{K}. \tag{2.32}$$

We evaluate the detection probability and conditional power on data generated from (2.1) with $T = 10,000$, $\gamma = 0.98$, $z_t \overset{\text{i.i.d.}}{\sim} \text{Poisson}(0.01)$ for all $t = 2, \ldots, T$, and $\sigma \in \{1, 2, \ldots, 10\}$. In (2.14), λ is chosen to yield $J = 100$ estimated spikes, i.e., $J = 100$ in (2.29); this is the expected number of spikes in this simulation. We generate 500 datasets, and consider

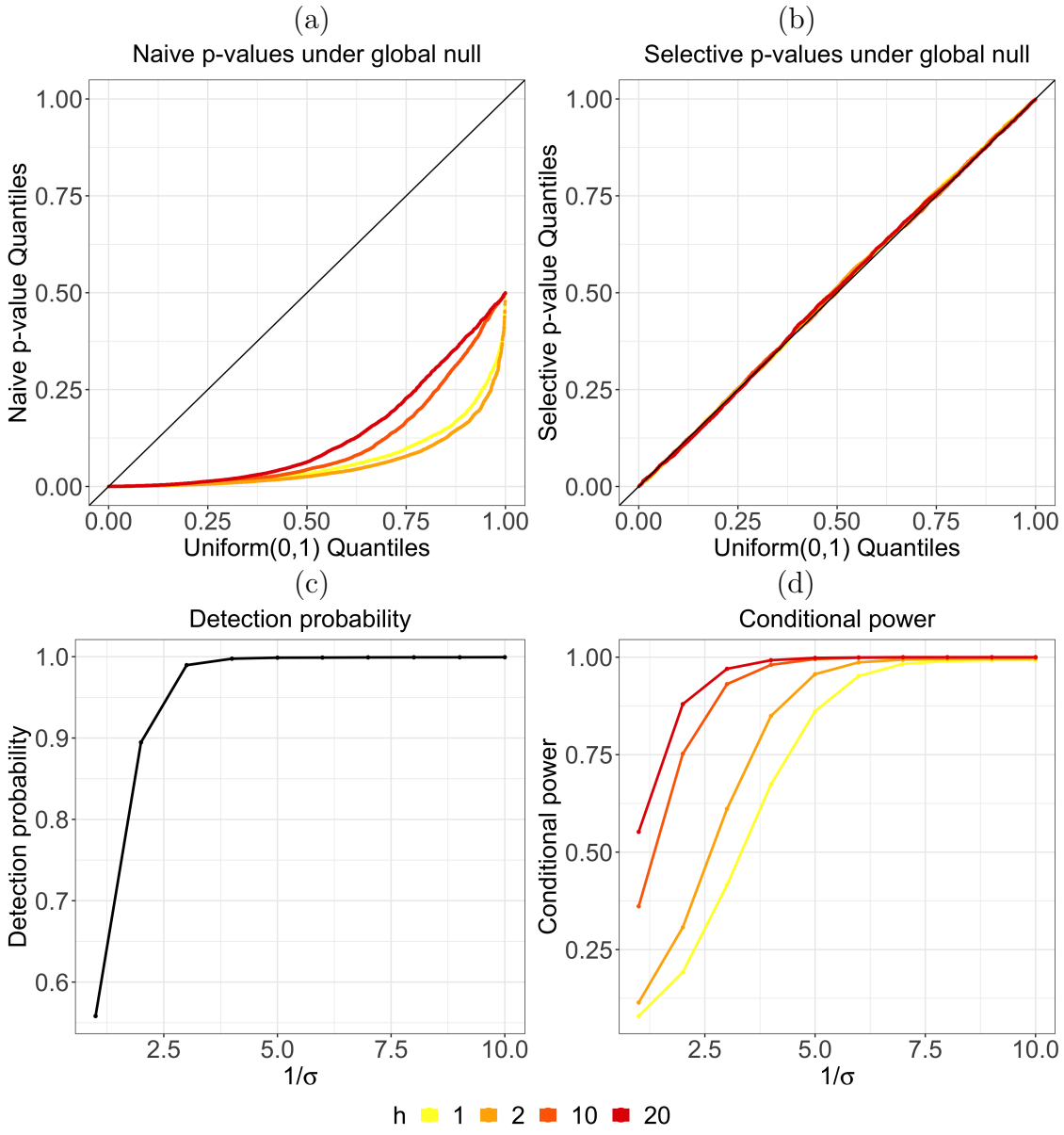


Figure 2.3: *(a):* Quantile-quantile plot for the naive p-values defined in (2.30), which have inflated selective Type I error. *(b):* Quantile-quantile plot for p-values from our proposed selective test in (2.9), which controls selective Type I error. *(c):* Under the model (2.1), detection probability (2.32) is an increasing function of $1/\sigma$. *(d):* Conditional power (2.31) increases as a function of $1/\sigma$ for all h. For a given value of σ, a larger value of h corresponds to higher conditional power, with the caveat that the meaning of the null hypothesis in (2.6) changes as a function of h, and the null hypothesis that holds for a smaller h might not hold for a larger value of h. The constant h appears in the definition of ν in (2.7).

$h \in \{1, 2, 10, 20\}$ in (2.7). Results with $\alpha = 0.05$ and $b = 2$ are displayed in Figure 2.3. Panels (c) and (d) display the detection probability and conditional power, respectively. Both quantities increase as $1/\sigma$ increases. Interpreting the relationship between conditional power and h requires more care: larger values of h typically give rise to higher conditional power for the same value of σ. However, the null hypothesis in (2.6) changes as a function of h, and it may be the case that H_0 holds for a smaller value of h, but not for a larger value.

2.5.3 Confidence interval coverage and width

We now generate data from (2.1) with $T = 10,000$, $\gamma = 0.98$, $z_t \overset{\text{i.i.d.}}{\sim} \text{Poisson}(0.01)$ for $t = 2, \ldots, T$, and $\sigma \in \{1, 2, \ldots, 6\}$. The tuning parameter λ in (2.14) is chosen to yield $J = 100$ estimated spikes, i.e., $J = 100$ in (2.29). For each timepoint $\tilde{\tau}_m$ in (2.29), we construct 95% selective confidence intervals $\left[\theta_L\left(\nu^\top y\right), \theta_U\left(\nu^\top y\right)\right]$ for the parameter $\nu^\top c$, with $h \in \{1, 2, 10, 20\}$. As a comparison, we also construct 95% confidence intervals for $\nu^\top c$ based on the naive p-value (2.30), which do not account for the fact that we decided to test $H_0 : \nu^\top c = 0$ after looking at the data:

$$\left[\nu^\top y - 1.96\sigma||\nu||_2, \nu^\top y + 1.96\sigma||\nu||_2\right]. \tag{2.33}$$

Suppose that we construct M confidence intervals (see (2.29) for the definition of M), we define their coverage, average width, and average midpoint relative to the value of $\nu^\top y$, as follows:

$$\text{Coverage} = \frac{1}{M}\sum_{m=1}^{M} 1\left(\nu^\top c \in \left[\theta_L\left(\nu^\top y\right), \theta_U\left(\nu^\top y\right)\right]\right), \tag{2.34}$$

$$\text{Width} = \frac{1}{M}\sum_{m=1}^{M} \left(\theta_U\left(\nu^\top y\right) - \theta_L\left(\nu^\top y\right)\right), \tag{2.35}$$

$$\text{Midpoint} = \frac{1}{M}\sum_{m=1}^{M} \left(\frac{\theta_L\left(\nu^\top y\right) + \theta_U\left(\nu^\top y\right)}{2} - \nu^\top y\right). \tag{2.36}$$

There is a slight abuse of notation in (2.34)–(2.36), since ν is a function of $\tilde{\tau}_m$ (see (2.7)).

Panels (a) and (b) of Figure 2.4 display the coverage of the selective and naive confidence intervals, respectively. The selective intervals achieve the nominal 95% coverage of the parameter $\nu^\top c$ across all values of σ and h. The naive intervals have poor coverage when $1/\sigma$ is small. As $1/\sigma$ increases, however, the coverage of the naive approach improves. This is because when $1/\sigma$ is very large (and hence σ is very small), the spikes estimated by solving the ℓ_0 problem (2.14) do not change much as a function of ϕ, and thus the truncation set $\{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}$ in (2.13) is very large; this means that ignoring this conditioning set has little effect on the confidence interval computed. A similar observation was made for the lasso in Zhao et al. [2021].

Figure 2.4(c) investigates the average width of the naive and selective confidence intervals as a function of σ, for $h = 1$. Selective intervals are much wider, on average. But the difference in width diminishes as $1/\sigma$ increases. This is congruent with our observations in panel (b): selective intervals can be well-approximated by naive intervals when $1/\sigma$ is large.

To understand how selective intervals achieve the nominal coverage, we plot the average midpoint of the selective intervals, after subtracting out $\nu^\top y$, in panel (d). If a confidence interval is symmetric around $\nu^\top y$ (as is the case for the naive interval in (2.33)), then this value equals zero. A positive value indicates that the interval is shifted upwards relative to $\nu^\top y$, and a negative value indicates the opposite. We see that for all values of h and σ, the selective intervals have a negative value of the midpoint after subtracting out $\nu^\top y$. This indicates that the selective approach provides an interval that is centered below the observed value of $\nu^\top y$.

Throughout this section, we have assumed that σ^2 in (2.1) is known. However, if it is unknown, we propose to use $\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(y_t - \hat{c}_t)^2$ as an estimator for σ^2 in evaluating the p-value in (2.9), where $\hat{c}_t$ is the solution to (2.14). In Appendix A.13, we demonstrate that this estimator leads to adequate selective Type I error control and substantial power in a simulation study.

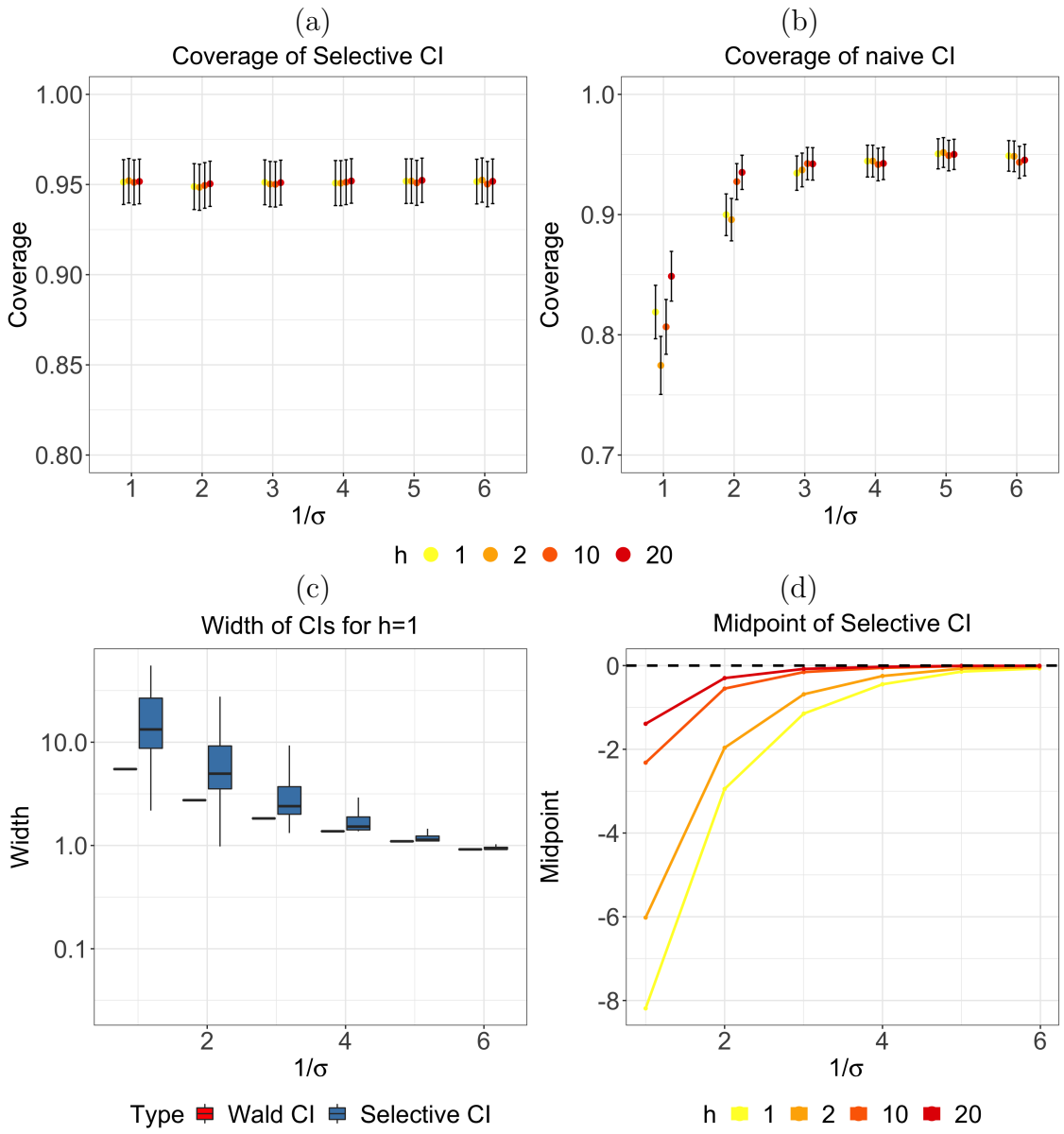


Figure 2.4: *(a):* Selective confidence intervals achieve correct nominal coverage (95% coverage at level $\alpha = 0.05$) across all values of h (defined in (2.7)) and σ (defined in (2.1)). The mean (and standard deviation) over 500 simulated datasets are displayed. *(b):* Naive confidence intervals have poor coverage when $1/\sigma$ is small, for all values of h. *(c):* For $h = 1$, selective confidence intervals are on average wider than naive intervals, but the difference decreases as $1/\sigma$ increases. *(d):* The midpoint of the selective confidence interval is, on average, smaller than $\nu^\top y$.

2.6 Application to calcium imaging data

2.6.1 Overview of data and analysis plan

Here we examine data aggregated as part of the `spikefinder` challenge [Theis et al., 2016]. The data consist of simultaneous electrophysiology and calcium recordings for a number of neurons. We consider the spike times recorded through electrophysiology to be the true, or "ground truth", spike times, against which we assess the accuracy of the spikes estimated via calcium imaging [Berens et al., 2018, Theis et al., 2016]. The calcium recordings have been resampled to 100 Hz, and linear trends removed, as described in Theis et al. [2016].

As in prior work [Jewell et al., 2019, Pachitariu et al., 2018], we set the value of γ in (2.14) based on known properties of the calcium indicators (0.986 for GCamp6f and 0.995 for GCamp6s). In settings where the properties of the calcium indicators are unknown, we can leverage a proposal from Fleming et al. [2021] for estimating γ. Since the calcium has a nonzero baseline, we solve a slight modification of (2.14):

$$\underset{c_1,\ldots,c_T \geq 0, \beta_0}{\text{minimize}} \left\{ \frac{1}{2} \sum_{t=1}^{T} (y_t - c_t - \beta_0)^2 + \lambda \sum_{t=2}^{T} 1\{c_t \neq \gamma c_{t-1}\} \right\}. \tag{2.37}$$

We first computed the average firing rates for data from Chen et al. [2013], which are 0.53 and 0.42 spikes per second for GCamp6f and GCamp6s recordings, respectively. For each recording, we solved (2.37) over a two-dimensional grid of (λ, β_0) values on the first 25% of the recording, and considered only the 20 pairs that yield an estimated average firing rate closest to the average firing rate of the corresponding calcium indicator. Among the 20 pairs, we then chose the (λ, β_0) pair that results in the smallest objective in (2.37) on the first 25% of the recording.

We quantify the accuracy of the estimated spikes resulting from (2.37) by comparing them to the ground truth spikes recorded using electrophysiology on the remaining 75% of each recording, using two widely-used metrics: (i) The *correlation between the true and estimated spikes*, after downsampling to 25 Hz, as described in Theis et al. [2016]. Larger values of

the correlation suggest better agreement between the true and estimated spikes. (ii) The *Victor-Purpura distance between the true and estimated spikes*, with cost parameter 10, as proposed in Victor and Purpura [1996, 1997]. Smaller values of the Victor-Purpura distance suggest better agreement between the true and estimated spikes.

We also quantify the accuracy of the subset of estimated spikes from (2.37) for which the p-values in (2.9) are below 0.05. As before, we computed the p-value (2.9) only on the estimated spikes for which $\nu^\top y > 0$. For each recording, we used $\hat{\sigma}^2 = \sum_{t=1}^{T} (y_t - \hat{c}_t)^2 / (T-1)$ to estimate the variance parameter σ^2, where $\hat{c}_t$ is the solution to (2.37). We used $h = 20$ in (2.7); this choice is motivated by the half decay times of the calcium indicators used in Chen et al. [2013], which are approximately 150 ms and 250 ms for GCamp6f and GCamp6s, respectively. Results for other values of h, as well as diagnostics to model (2.1), are in Section A.12 of the Appendix.

2.6.2 Results for a single cell

In Figure 2.5, we display results for a single cell: recording 29 of dataset 7 from the `spikefinder` challenge. Each panel displays the following quantities, at varying levels of zoom: (i) the fluorescence trace (grey dots); (ii) the estimated spikes from (2.37) (orange ticks); (iii) the estimated spikes from (2.37) for which the p-values from (2.9) with $h = 20$ are below 0.05 (blue ticks); and (iv) the true spikes (black ticks).

We see that the estimated spikes with p-values less than 0.05 match very closely with the true spikes. For example, (2.37) estimates spikes near 79.3, 83.0, 89.1, and 92.9 seconds. None of these correspond to a true spike, and none have a p-value less than 0.05. Thus, the spikes with p-values above 0.05 appear to be false positives. By contrast, those with p-values below 0.05 are mostly true positives. The quantitative measures defined in Section 2.6.1 further indicate that considering only spikes with p-values below 0.05 increases accuracy: the correlations between the true spikes and the estimated spikes including and excluding p-values below 0.05 are 0.54 and 0.62, respectively, and the Victor-Purpura distances are 278 and 244, respectively.

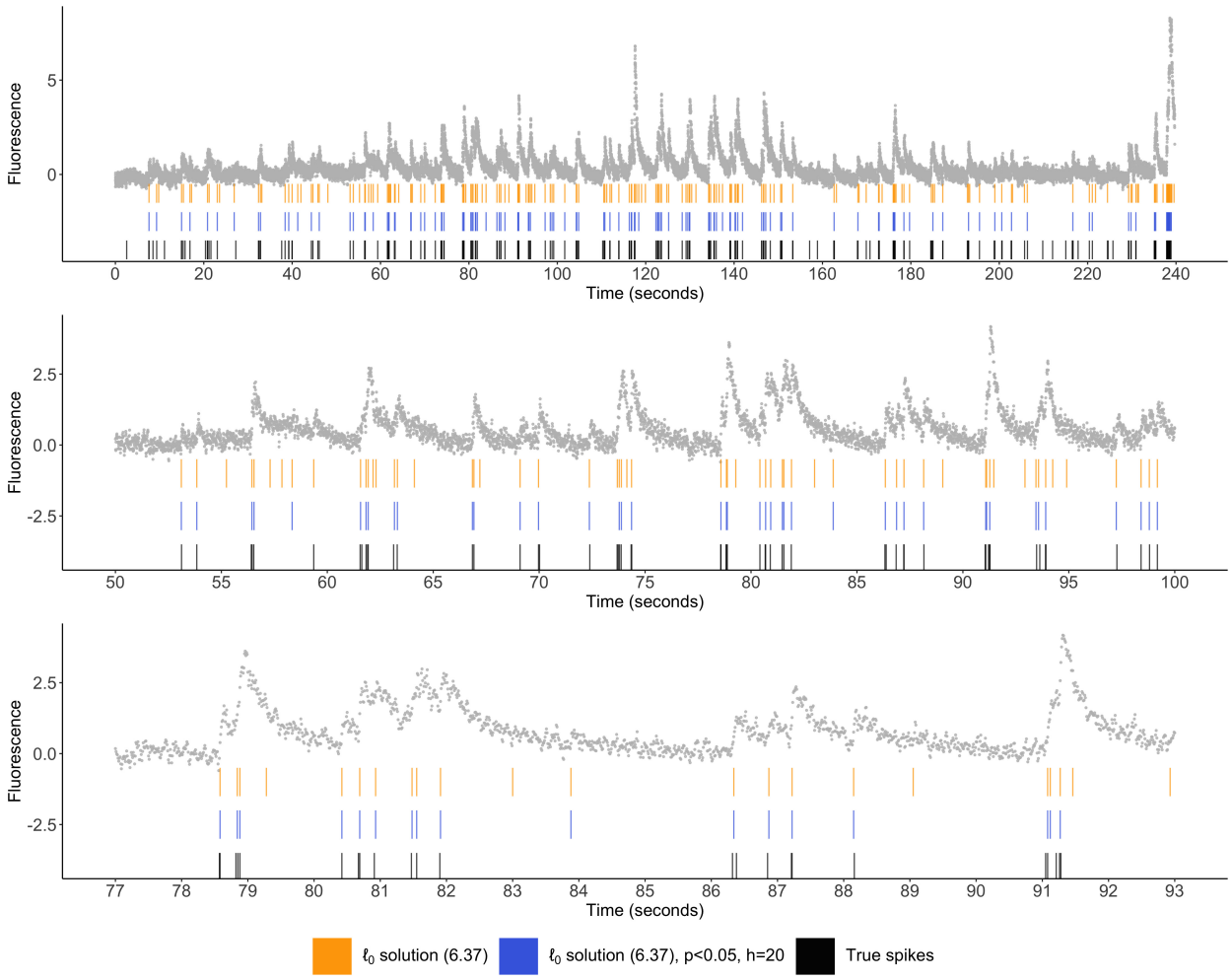


Figure 2.5: Illustrative example for recording 29 from Chen et al. [2013], which uses the GCaMP6f indicator, after preprocessing as described in Theis et al. [2016]. The cell's fluorescence trace is displayed in grey. Estimated spikes from (2.37) are displayed in orange; the spikes with p-values from (2.9) below 0.05 (with $h = 20$) are displayed in blue; and the true spike times are shown in black.

2.6.3 Results for recordings in Chen et al. [2013]

We now examine datasets 7 and 8 of the `spikefinder` challenge. Their original source is Chen et al. [2013]. The data consist of 58 recordings; each is approximately 230 seconds long.

Figure 2.6 displays the accuracy — relative to the ground truth spikes obtained via electrophysiology — of the spikes estimated via (2.37) (in orange), along with the subset of those spikes for which the p-value is below 0.05 (in blue). Accuracy is measured using Victor-Purpura distance and correlation. We find that the spikes from (2.37) with p-values below 0.05 are more accurate than the full set of spikes from (2.37). These results are based on $h = 20$ in (2.9). Results for $h = 5$ and $h = 50$ are similar; see Figures A.3 and A.4 in Section A.12 of the Appendix.

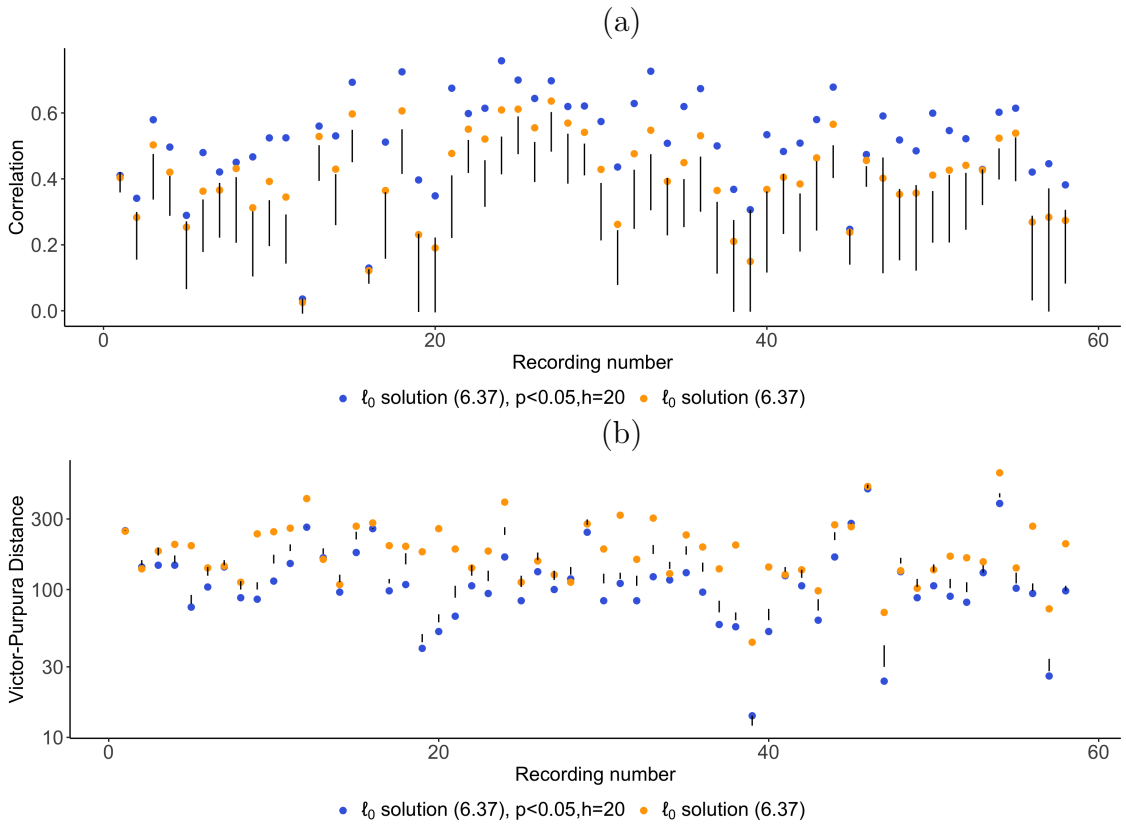


Figure 2.6: Result for recordings from the Chen et al. [2013] dataset. *(a):* The correlations between the true spike times and the spikes estimated from (2.37) are plotted in orange. The correlations between the true spike times and the subset of the spikes from (2.37) with p-value (2.9) below 0.05 are plotted in blue. For each recording, the black line represents the 2.5% and 97.5% quantiles of the resampling distribution with 1,000 samples. *(b):* As in (a), but Victor-Purpura distance is displayed instead of correlation.

It is natural to wonder whether retaining only estimated spikes with p-values below 0.05 improves the correlation and Victor-Purpura distance merely as a byproduct of reducing the number of estimated spikes, rather than due to the high quality of the estimated spikes with p-values below 0.05. We assess this using a resampling approach. Let M denote the number of spikes for which p-values are computed, and let $\tilde{M}$ denote the number that are below 0.05. We sample $\tilde{M}$ out of M estimated spike times for which p-values are computed without replacement, and compute the correlation and Victor-Purpura distance between the true spike times and the sampled subset. We do this 1,000 times, and record the 2.5% and 97.5% quantiles of the accuracy measures obtained. These are shown as the endpoints of the black lines displayed in Figure 2.6. We see that even after taking into account the effect of a smaller number of estimated spikes, excluding spikes with p-values greater than 0.05 still provides improved accuracy, measured using either correlation (56 out of 58 recordings) or Victor-Purpura distance (51 out of 58 recordings).

2.7 Discussion

Methods developed in this paper are implemented in the `R` package `SpikeInference`, available at https://github.com/yiqunchen/SpikeInference. We provide a tutorial for the package at https://yiqunchen.github.io/SpikeInference/. Code for reproducing the results in this paper can be found at https://github.com/yiqunchen/SpikeInference-experiments.

Our work leads to a few future directions of research.

2.7.1 Alternative conditioning sets and contrast vectors for testing (2.6)

Instead of conditioning on the jth estimated spike $\hat{\tau}_j$ to obtain the p-value in (2.9), we could instead condition on $\hat{\tau}_j$ and its immediate neighbors, $\hat{\tau}_{j-1}$ and $\hat{\tau}_{j+1}$. This would allows us to

define the contrast vector ν as

$$\nu_t = \begin{cases} -\frac{\gamma(\gamma^2-1)}{\gamma^2-\gamma^{2(\hat{\tau}_{j-1}-\hat{\tau}_j)}} \cdot \gamma^{t-\hat{\tau}_j}, & \hat{\tau}_{j-1}+1 \le t \le \hat{\tau}_j \\ \frac{\gamma^2-1}{\gamma^{2(\hat{\tau}_{j+1}-\hat{\tau}_j)}-1} \cdot \gamma^{t-(\hat{\tau}_j+1)}, & \hat{\tau}_j+1 \le t \le \hat{\tau}_{j+1} \\ 0, & \text{otherwise,} \end{cases}$$

leading to a p-value given by $\mathbb{P}\left(\phi \ge \nu^\top y \;\middle|\; \{\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}\} \subseteq \mathcal{M}(y'(\phi)), \phi > 0\right)$, where $\phi \sim \mathcal{N}(0, \sigma^2||\nu||_2^2)$. This approach eliminates the need to specify a window size h, and instead chooses the window size adaptively. Computing this new p-value requires only minor modifications of the results in Section 2.3, using ideas from Jewell et al. [2022]; we leave the details to future work.

As an alternative, we could keep the conditioning set in (2.9), but define a contrast vector ν that uses different numbers of timepoints to the left and right of $\hat{\tau}_j$ (in contrast to (2.7)).

2.7.2 Selective inference for other spike detection methods

In this paper, we considered selective inference on spikes estimated via the ℓ_0 problem in (2.14). However, another line of research [Friedrich and Paninski, 2016, Friedrich et al., 2017, Vogelstein et al., 2010] involves estimating spikes via an ℓ_1-penalized approach:

$$\underset{c_1,\ldots,c_T \ge 0; z_1,\ldots,z_T}{\text{minimize}} \left\{ \frac{1}{2}\sum_{t=1}^{T}(y_t - c_t)^2 + \lambda \sum_{t=2}^{T} |z_t| \right\} \text{ subject to } z_t = c_t - \gamma c_{t-1} \ge 0. \tag{2.38}$$

Spikes are estimated at timepoints for which $\hat{c}_t \neq \gamma \hat{c}_{t-1}$. To conduct inference on these estimated spikes, we could leverage the framework in Section 2.2.2, along with recent developments in selective inference for the lasso and related problems [Hyun et al., 2021, Lee et al., 2016].

2.7.3 Propagating uncertainty to downstream data analysis

This article focused on quantifying the uncertainty associated with $\nu^\top c$, the change in calcium associated with an estimated spike. It is also of interest to propagate this uncertainty to downstream analyses, such as the neural decoding model [Pillow et al., 2011, Ventura, 2008]. This model is similar to (2.1) with $z_t \overset{\text{i.i.d.}}{\sim} \text{Poisson}(f(\theta_t))$ for a function f; the goal is to estimate the coefficients θ_t. We could leverage the framework proposed in Wei et al. [2019] to propagate uncertainty of estimating $\nu^\top c$ to θ_t.

Chapter 3

MORE POWERFUL SELECTIVE INFERENCE FOR THE GRAPH FUSED LASSO

3.1 Introduction

We consider a vector $Y \in \mathbb{R}^n$, assumed to be a noisy realization of a signal $\beta \in \mathbb{R}^n$,

$$Y_j = \beta_j + \epsilon_j, \quad \epsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad j = 1, \ldots, n, \tag{3.1}$$

with known variance σ^2. We assume that β has some underlying structure of interest. For instance, β might be *sparse*, with few non-zero elements, or *piecewise constant*, meaning that the elements of β are ordered, and adjacent elements tend to take on equal values.

It is natural to estimate β by solving the optimization problem

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \beta\|_2^2 + \lambda \|D\beta\|_1 \right\}, \tag{3.2}$$

where D is an $m \times n$ penalty matrix that encodes the structure of β. Problem (3.2) is a special case of the generalized lasso with an identity design matrix [Arnold and Tibshirani, 2016, Hastie et al., 2015, Tibshirani and Taylor, 2011]. While the ideas in this chapter apply for a general design matrix, we make use of an identity design matrix to simplify the discussion. Many well-known regression problems involving ℓ_1 penalties can be viewed as special cases of the generalized lasso; examples include the lasso [Tibshirani, 1996], the fused lasso signal approximator [Friedman et al., 2007, Rinaldo, 2009, Tibshirani et al., 2005], the graph fused lasso [Hastie et al., 2015, Tibshirani and Taylor, 2011], and trend filtering [Kim et al., 2009, Tibshirani, 2014].

Despite the abundant literature on algorithms for computing $\hat{\beta}$ in (3.2) [Arnold and Tib-

shirani, 2016, Friedman et al., 2007, Johnson, 2013, Ramdas and Tibshirani, 2016, Tibshirani and Taylor, 2011, Xin et al., 2014, Zhu, 2017] and on its theoretical properties [Harchaoui and Lévy-Leduc, 2010, Rinaldo, 2009, Sadhanala et al., 2016, Tibshirani and Taylor, 2011], the topic of *inference* for the generalized lasso remains less developed. In this work, we focus on testing a null hypothesis that was determined after observing $\hat{\beta}$ in (3.2).

More precisely, suppose that we perform the graph fused lasso, a special case of (3.2),

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \sum_{(j,j') \in E} |\beta_j - \beta_{j'}| \right\}, \tag{3.3}$$

where $G = (V, E)$ is an undirected graph, $V = \{1, \ldots, n\}$, and $(j, j') \in E$ indicates that the jth and j'th vertices in the graph are connected by an edge [Tibshirani and Taylor, 2011]. For sufficiently large values of the non-negative tuning parameter λ, we will have $\hat{\beta}_j = \hat{\beta}_{j'}$ for some $(j, j') \in E$. We can segment $\hat{\beta}$ into *connected components* — that is, sets of elements of $\hat{\beta}$ that are connected in the original graph and share a common value. We might then consider testing the null hypothesis that the true mean of β is the same across two *estimated* connected components, i.e.,

$$H_0 : \sum_{j \in \hat{C}_1} \beta_j / |\hat{C}_1| = \sum_{j' \in \hat{C}_2} \beta_{j'} / |\hat{C}_2| \text{ versus } H_1 : \sum_{j \in \hat{C}_1} \beta_j / |\hat{C}_1| \neq \sum_{j' \in \hat{C}_2} \beta_{j'} / |\hat{C}_2|, \tag{3.4}$$

where $\hat{C}_1 \subseteq V$ and $\hat{C}_2 \subseteq V$ are connected components of $\hat{\beta}$, with cardinality $|\hat{C}_1|$ and $|\hat{C}_2|$, and $\hat{C}_1 \cap \hat{C}_2 = \emptyset$. This is equivalent to testing $H_0 : \nu^\top \beta = 0$ versus $H_1 : \nu^\top \beta \neq 0$, where

$$\nu_j = 1\Big\{j \in \hat{C}_1\Big\}/|\hat{C}_1| - 1\Big\{j \in \hat{C}_2\Big\}/|\hat{C}_2|, \quad j = 1, \ldots, n. \tag{3.5}$$

Here, H_0 is chosen based on the data, i.e., we selected the contrast vector ν in (3.5) because $\hat{C}_1$ and $\hat{C}_2$ are estimated connected components. We focus on developing a test of H_0 that controls the *selective Type I error* [Fithian et al., 2014], i.e., one for which the probability of

rejecting H_0 at level α, given that H_0 holds *and we decided to test it*, is no greater than α:

$$\mathbb{P}_{H_0}(\text{reject } H_0 \text{ at level } \alpha \mid H_0 \text{ is tested}) \leq \alpha, \ \forall \alpha \in (0,1). \tag{3.6}$$

It is not hard to see that a standard two-sample z-test of $H_0 : \nu^\top \beta = 0$, with p-value $\mathbb{P}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y|\right)$, fails to account for the fact that we decided to test H_0 after looking at the data, and therefore does not control the selective Type I error rate (3.6). To address this problem, Hyun et al. [2018] propose an elegant approach for testing $H_0 : \nu^\top \beta = 0$ that makes use of the selective inference framework developed by Lee et al. [2016], Fithian et al. [2014], and Tibshirani et al. [2016]. Their key insight is as follows: the set of Y that yields a particular output for the first K steps of the dual path algorithm for solving (3.2) is a polyhedron, of the form $\{Y : AY \geq 0\}$, for a matrix A that can be explicitly computed. Thus, *conditional on Y belonging to this polyhedral set*, the linear contrast $\nu^\top Y$ follows a truncated normal distribution, with parameters that are a function of A, σ^2, and ν, for any ν that is based on the output of (3.2). It is thus possible to compute valid p-values for the null hypothesis in (3.4) in the sense of (3.6), *by conditioning on the outputs from the first K steps of the dual path algorithm.*

Our work relies on a simple observation: the proposal considered in Hyun et al. [2018] involves conditioning on much more information than is used to construct the contrast vector ν in (3.5). As pointed out by Fithian et al. [2014] and Liu et al. [2018], conditioning on unnecessary information leads to reduced power. In this chapter, we make use of recent ideas from Jewell et al. [2022] to develop a computationally-efficient test of $H_0 : \nu^\top \beta = 0$ that conditions on substantially less information than Hyun et al. [2018], thereby obtaining much higher power while still guaranteeing valid inference in the sense of (4.5).

While this chapter was in preparation, Le Duy and Takeuchi [2021] independently developed a test of $H_0 : \nu^\top \beta = 0$ that has higher power than Hyun et al. [2018]. Compared to that paper, our proposal (i) conditions on less unnecessary information; (ii) enjoys better numerical stability; and (iii) leads to more interpretable p-values. Detailed discussion and

experimental results are provided in Appendix B.13.

The rest of this chapter is organized as follows. In Section 3.2, we briefly review the dual path algorithm for solving (3.2), and the existing proposals for selective inference for this problem. In Section 3.3, we introduce our selective inference procedure, which provides a computationally-efficient approach to condition on less information than Hyun et al. [2018]. Section 3.4 outlines some extensions, and Section 3.5 compares the performance of our proposal to that of Hyun et al. [2018] in simulation. A real data application is in Section 3.6, and a discussion of future work is in Section 3.7. Proofs and other technical details are relegated to the Appendix.

Throughout this chapter, we will use the following notational conventions. The ith row of a matrix A is denoted A_i. Given a set S of positive integers, A_S is the submatrix with rows in S, A_{-S} is the submatrix with rows not in S, and $|S|$ is the cardinality of the set S. For a vector $\nu \in \mathbb{R}^n$, $||\nu||_1, ||\nu||_2$, and $||\nu||_\infty$ denote its ℓ_1, ℓ_2, and ℓ_∞ norms, respectively. In addition, $\Pi_\nu^\perp$ denotes the projection matrix onto the orthogonal complement of ν, i.e., $\Pi_\nu^\perp = I_n - \frac{\nu\nu^\top}{||\nu||_2^2}$, where I_n is the n-dimensional identity matrix. We use $1_{(\cdot)}$ to denote the indicator function. For a positive integer m, we define $[m] = \{1, 2, \ldots, m\}$.

3.2 Background on the generalized lasso

In this section, we review the selective inference framework of Hyun et al. [2018] for testing hypotheses based upon the generalized lasso estimator (3.2), which includes the graph fused lasso as a special case. Their framework relies on the dual path algorithm of Tibshirani and Taylor [2011] for solving (3.2). Thus, we begin with a very brief overview of that algorithm.

3.2.1 The dual problem, and the dual path algorithm

Tibshirani and Taylor [2011] develop an efficient path algorithm for solving the dual problem for (3.2), which takes the form

$$\hat{u}(\lambda) = \underset{u \in \mathbb{R}^m}{\operatorname{argmin}} \quad \left\| y - D^\top u \right\|_2^2 \quad \text{subject to} \quad \|u\|_\infty \leq \lambda, \tag{3.7}$$

and is related to (3.2) through the identity $\hat{\beta}(\lambda) = y - D^\top \hat{u}(\lambda)$, where the notation $\hat{\beta}(\lambda)$ and $\hat{u}(\lambda)$ makes explicit that $\hat{\beta}$ and $\hat{u}$ are functions of λ. This dual path algorithm is detailed in Appendix B.1. While the details of the algorithm are not important for the current paper, we briefly summarize the main idea. The algorithm begins with $\lambda = \infty$, and then proceeds through a series of steps, corresponding to decreasing values of λ. The kth step involves computing a boundary set $B_k \subseteq [m]$, which consists of the subset of indices of the vector u for which the inequality constraint in (3.7) is tight. The signs of the elements of u associated with this boundary set, s_{B_k}, are also computed. These quantities satisfy

$$\hat{\beta}(\lambda) = P_{\mathrm{Null}(D_{-B_k})}\big(y - \lambda \cdot D_{B_k}^\top s_{B_k}\big) \tag{3.8}$$

for a range of λ values corresponding to the kth step [Tibshirani and Taylor, 2011]. In (3.8), D_{B_k} and D_{-B_k} correspond to the submatrices of D with rows in B_k and not in B_k, respectively, and $P_{\mathrm{Null}(D_{-B_k})}$ is the projection matrix onto the null space of D_{-B_k}. To summarize, (3.8) indicates that $\hat{\beta}(\lambda)$ can be computed from (B_k, s_{B_k}), for an appropriate range of λ values.

The next proposition considers the special case of the graph fused lasso problem (3.3).

Proposition 7. *Let B_k denote the boundary set that results from the kth step of the dual path algorithm for (3.3), and let $\hat{\beta}$ denote the solution to (3.3). Let G_{-B_k} denote the subgraph of G with edges in the boundary set B_k removed, and let $C_1, \ldots, C_L$ denote the L connected components of G_{-B_k}. Then, under (3.1), with probability 1, $\hat{\beta}_j = \hat{\beta}_{j'}$ if and only if $j, j' \in C_l$ for some $l \in [L]$.*

We close with a brief summary of the main points in this section:

- For the generalized lasso (3.2), $\hat{\beta}$ can be computed from λ and (B_k, s_{B_k}) from the dual path algorithm.

- In the special case of the graph fused lasso (3.3), the connected components of $\hat{\beta}$ can be computed from B_k from the dual path algorithm.

3.2.2 Existing work on selective inference for the generalized lasso

The main idea behind selective inference is as follows: when testing a null hypothesis that is a function of the data, to control the selective Type I error in the sense of (3.6), we must condition on the information used to construct that null hypothesis [Fithian et al., 2014, Lee et al., 2016, Tibshirani et al., 2016]. In particular, to test a null hypothesis of the form $H_0 : \nu^\top \beta = 0$ where ν is a function of the data, we must condition on the information used to construct ν.

In a recent elegant line of work, a number of authors have shown that the model selection events of several well-known model selection procedures, including the lasso [Lee et al., 2016], stepwise regression [Loftus and Taylor, 2014, Tibshirani et al., 2016], and marginal screening [Reid et al., 2017], can be written as polyhedral constraints on Y. More precisely, conditioning on the selected model (and in some cases, additional information) is equivalent to conditioning on a polyhedral set $\{Y : AY \leq b\}$, where the matrix A and the vector b can be explicitly computed. Thus, we can test null hypotheses that are a function of the selected model by considering the null distribution of Y truncated to a polyhedral set.

Recently, Hyun et al. [2018] extended this line of work to develop an approach for selective inference for the generalized lasso (3.2). Their key insight is as follows:

> *The set of Y that leads to a specified output for the first K steps of the dual path algorithm for (3.2) is a polyhedron, i.e., $\{Y : AY \leq 0\}$, for a matrix A that can be explicitly computed.*

Proposition 8 details this result.

Proposition 8 (Proposition 3.1 in Hyun et al. [2018])**.** *Consider solving (3.2) using the dual path algorithm for $y \in \mathbb{R}^n$. For the kth step, $k = 1, \ldots, K$, define*

$$M_k(y) \equiv (B_k(y), s_{B_k}(y), R_k(y), L_k(y)) \tag{3.9}$$

where the boundary set $B_k(y)$ and the sign vector of the boundary set $s_{B_k}(y)$ are defined in

Algorithm 3 (see Appendix B.1), and $R_{k+1}(y) = (\text{sign}(a_i) : i \notin B_k(y))$ *and* $L_{k+1}(y) = \{i : i \in B_k(y), c_i < 0, d_i < 0\}$, *for* a_i, c_i, *and* d_i *specified in Algorithm 3.*

Then the set $\left\{Y \in \mathbb{R}^n : \bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}\right\}$ *is of the form* $\{Y : AY \leq 0\}$ *for some matrix* A *that can be constructed explicitly based on* $M_1(y), \ldots, M_K(y)$.

Motivated by this result, Hyun et al. [2018] proposed to test $H_0 : \nu^\top \beta = 0$, where ν is a function of the generalized lasso estimator, via a p-value of the form

$$p_{\text{Hyun}} \equiv \mathbb{P}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}, \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right). \tag{3.10}$$

In words, p_{Hyun} answers the following question:

> *Assuming that there is no difference between the population means of* $\hat{C}_1$ *and* $\hat{C}_2$, *then what's the probability of observing such a large difference in the sample means of* $\hat{C}_1$ *and* $\hat{C}_2$, *given that the first* K *steps of the dual path algorithm yield the same results as on the observed data?*

In (3.10), conditioning on $\Pi_\nu^\perp Y$ eliminates the nuisance parameter $\Pi_\nu^\perp \beta$; see Section 3.1 of Fithian et al. [2014]. Now, under (3.1), the conditional distribution of $\nu^\top Y$ is normal with mean zero and variance $\sigma^2 ||\nu||_2^2$, truncated to a set that can be characterized and efficiently computed using Proposition 8. This yields the p-value in (3.10). Furthermore, unlike the z-test based on the naive p-value $\mathbb{P}_{H_0}\big(|\nu^\top Y| \geq |\nu^\top y|\big)$, a test that rejects H_0 when the p-value in (3.10) is less than some level α controls the selective Type I error rate, as in (3.6).

We emphasize that the p-value in (3.10) conditions on the event $\bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}$; that is, on all of the outputs of the first K steps of the dual path algorithm (rather than simply the Kth step). However, typically the contrast vector ν in $H_0 : \nu^\top \beta = 0$ is constructed using only (at most) the output of the Kth step in the dual path algorithm. In what follows, we will consider conditioning on much less information than (3.10). This will result in a test that controls the selective Type I error in the sense of (3.6), and that has substantially higher power under the alternative.

3.3 Proposed approach

3.3.1 *What* should *we condition on?*

To control the selective Type I error in (3.6), we must condition on the aspect of the data that led us to test the specific null hypothesis $H_0 : \nu^\top \beta = 0$ [Fithian et al., 2014, Hyun et al., 2018].

If a data analyst wishes to choose the contrast vector ν in the null hypothesis $H_0 : \nu^\top \beta = 0$ by inspecting the elements of $\hat{\beta}$ resulting from the Kth step of the dual algorithm of the generalized lasso problem (3.2), then there is no reason to condition on $\bigcap_{k=1}^{K-1} \{M_k(Y) = M_k(y)\}$ (as was done by Hyun et al. [2018]), since the outputs of the first $K-1$ steps of the dual path algorithm are not considered in constructing ν. In fact, according to (3.8), which states that B_K and s_{B_K} uniquely determine $\hat{\beta}$, the data analyst need only condition on B_K and s_{B_K}, rather than on $M_K = (B_K, s_{B_K}, R_K, L_K)$.

Furthermore, the data analyst might construct the contrast vector ν in $H_0 : \nu^\top \beta = 0$ to take on a constant value within each connected component of $\hat{\beta}$, as in (3.4) and (3.5). Recall from Proposition 7 that the connected components of $\hat{\beta}$ are equivalent to the connected components of the subgraph G_{-B_K}. Therefore, it suffices to condition only on the connected components of the subgraph G_{-B_K}, or even on just the pair of connected components under investigation in (3.4).

What is the disadvantage of conditioning on $\bigcap_{k=1}^{K} \{M_k(Y) = M_k(y)\}$, as in Hyun et al. [2018]? Conditioning on too much information leads to a loss of power [Fithian et al., 2014, Jewell et al., 2022, Lee et al., 2016, Liu et al., 2018]. We wish to condition on less information to achieve higher power than Hyun et al. [2018], while controlling the selective Type I error (3.6). Of course, this could result in computational challenges, as the conditioning sets described in the use cases above are not polyhedral, so the conditional distribution of $\nu^\top Y$ is no longer a normal truncated to an easily-characterized set.

In what follows, we focus on the case where the data analyst constructs the contrast vector ν to take on a constant value in each connected component of $\hat{\beta}$, and consider a

p-value of the form

$$p_{\hat{C}_1,\hat{C}_2} \equiv \mathbb{P}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right). \tag{3.11}$$

In (3.11), $\hat{C}_1(y)$ and $\hat{C}_2(y)$ are two connected components estimated from the data realization y and used to construct the contrast vector ν in (3.5), and $\mathcal{CC}_K(Y)$ is the set of connected components obtained from applying K steps of the dual path algorithm for (3.3) to the random variable Y. Roughly speaking, this p-value answers the following question:

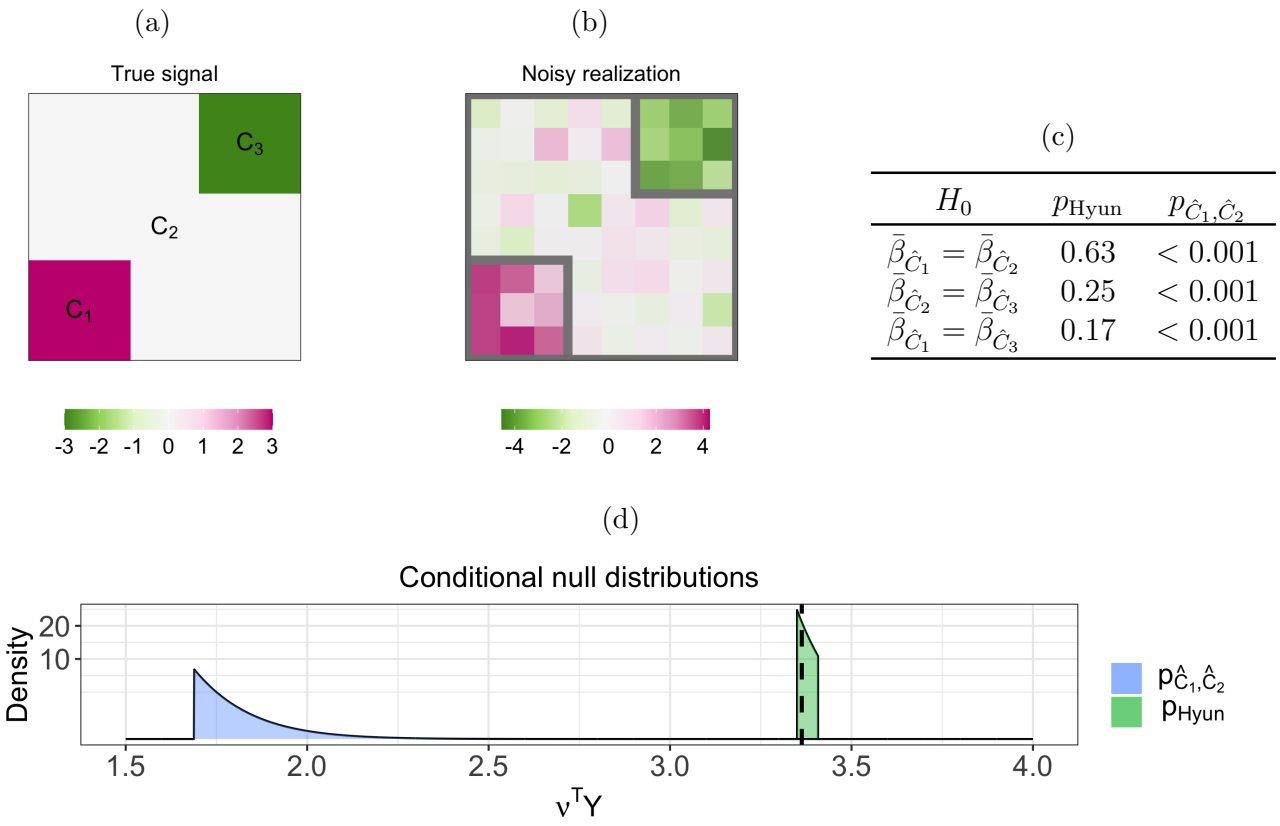
> *Assuming that there is no difference between the population means of $\hat{C}_1$ and $\hat{C}_2$, then what's the probability of observing such a large difference in the sample means of $\hat{C}_1$ and $\hat{C}_2$, given that these two connected components were estimated from the data?*

While our proposed p-value $p_{\hat{C}_1,\hat{C}_2}$ conditions on far less information than Hyun et al. [2018], the recent proposal of Le Duy and Takeuchi [2021] takes an intermediate approach. They condition on the full boundary set B_K at the Kth step of the dual path algorithm (and thus, implicitly, on *all* of the connected components in $\mathcal{CC}_K(Y)$), whereas we condition only on the two connected components of interest. See Appendix B.13 for further discussion and power comparison.

3.3.2 Illustrative example

We now demonstrate that conditioning on less information leads to increased power, using an example involving the graph fused lasso applied to a two-dimensional grid graph.

To begin, we constructed a graph composed of 64 nodes arranged in an 8×8 grid, such that each node is connected to its four closest (up, down, left, right) neighbors. We generated data on this grid according to (3.1), where β has three piecewise constant segments, C_1, C_2, and C_3, with means of 3, 0, and -3, respectively. The true values of β, as well as the data generated from this model, are shown in Figures 3.1(a)–(b). On this particular data



(c)

H_0	p_{Hyun}	$p_{\hat{C}_1,\hat{C}_2}$
$\bar{\beta}_{\hat{C}_1} = \bar{\beta}_{\hat{C}_2}$	0.63	< 0.001
$\bar{\beta}_{\hat{C}_2} = \bar{\beta}_{\hat{C}_3}$	0.25	< 0.001
$\bar{\beta}_{\hat{C}_1} = \bar{\beta}_{\hat{C}_3}$	0.17	< 0.001

Figure 3.1: *(a):* We generated β on an 8×8 grid. There are three true connected components, which take on values of -3, 0, and 3. *(b):* A noisy realization from the model $Y \sim \mathcal{N}(\beta, I_{64})$. In this particular example, running 13 steps of the dual path algorithm for the graph fused lasso results in perfect recovery of the true connected components of β (displayed in grey). *(c):* For each pair of estimated connected components, we tested the null hypothesis of equality in means using p_{Hyun} in (3.10) and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11). *(d):* The conditional null distributions of $\nu^\top Y$, where ν is chosen to test for a difference in means between $\hat{C}_1$ and $\hat{C}_2$, conditional on the conditioning sets in the definitions of p_{Hyun} in (3.10) and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11). In *(d)*, the test statistic $|\nu^\top y| = 3.36$ is displayed as a dashed black line; this value is quite large relative to the null distribution of $p_{\hat{C}_1,\hat{C}_2}$, but modest relative to that of p_{Hyun}.

set, $K = 13$ steps of the dual path algorithm for the graph fused lasso recovered the true connected components exactly.

For each pair of connected components, we then constructed a contrast vector ν as in

(3.4), so that $H_0 : \nu^\top \beta = 0$ posits that the two components being tested have the same mean. We tested H_0 using the p-values p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ given in (3.10) and (3.11), respectively. The p-values for all pairs of connected components are displayed in Figure 3.1(c). Because p_{Hyun} conditions on unnecessary information, the test based on p_{Hyun} has extremely low power and it cannot reject any H_0. By contrast, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher power. In Figure 3.1(d), we display the null distribution of $\nu^\top Y$, conditional on the conditioning sets in (3.10) and (3.11).

3.3.3 Properties of $p_{\hat{C}_1,\hat{C}_2}$

The following result establishes key properties of $p_{\hat{C}_1,\hat{C}_2}$ in (3.11).

Proposition 9. *Suppose that $Y \sim \mathcal{N}(\beta, \sigma^2 I_n)$. Define*

$$y'(\phi) = \Pi_\nu^\perp y + \phi \cdot \frac{\nu}{||\nu||_2^2} = y + \left(\frac{\phi - \nu^\top y}{||\nu||_2^2} \right) \nu. \tag{3.12}$$

Let $\phi \sim \mathcal{N}(0, \sigma^2 ||\nu||_2^2)$. Then, under $H_0 : \nu^\top \beta = 0$,

$$p_{\hat{C}_1,\hat{C}_2} = \mathbb{P}\Big(|\phi| \geq |\nu^\top y| \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(y'(\phi)) \Big). \tag{3.13}$$

Moreover, the test that rejects H_0 if $p_{\hat{C}_1,\hat{C}_2} \leq \alpha$ controls the selective Type I error at level α.

Therefore, to compute the p-value in (3.11), it suffices to characterize the set

$$\mathcal{S}_{\hat{C}_1,\hat{C}_2} \equiv \left\{ \phi \in \mathbb{R} : \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(y'(\phi)) \right\}. \tag{3.14}$$

We can think of $y'(\phi)$ in (3.12) as a perturbation of the data by a function of ϕ along the direction defined by ν. Figure 3.2 illustrates this intuition in the toy example from Figure 3.1, in the context of a test for the difference in the means of $\hat{C}_1$ and $\hat{C}_2$ (see Figure 3.2(a)). Panel (a) displays the observed data, for which $\phi = \nu^\top y = 3.36$, where ν is defined in (3.5). In panel (b), we perturb the observed data to $\phi = 0$. Now the graph fused lasso with $K = 13$ no longer

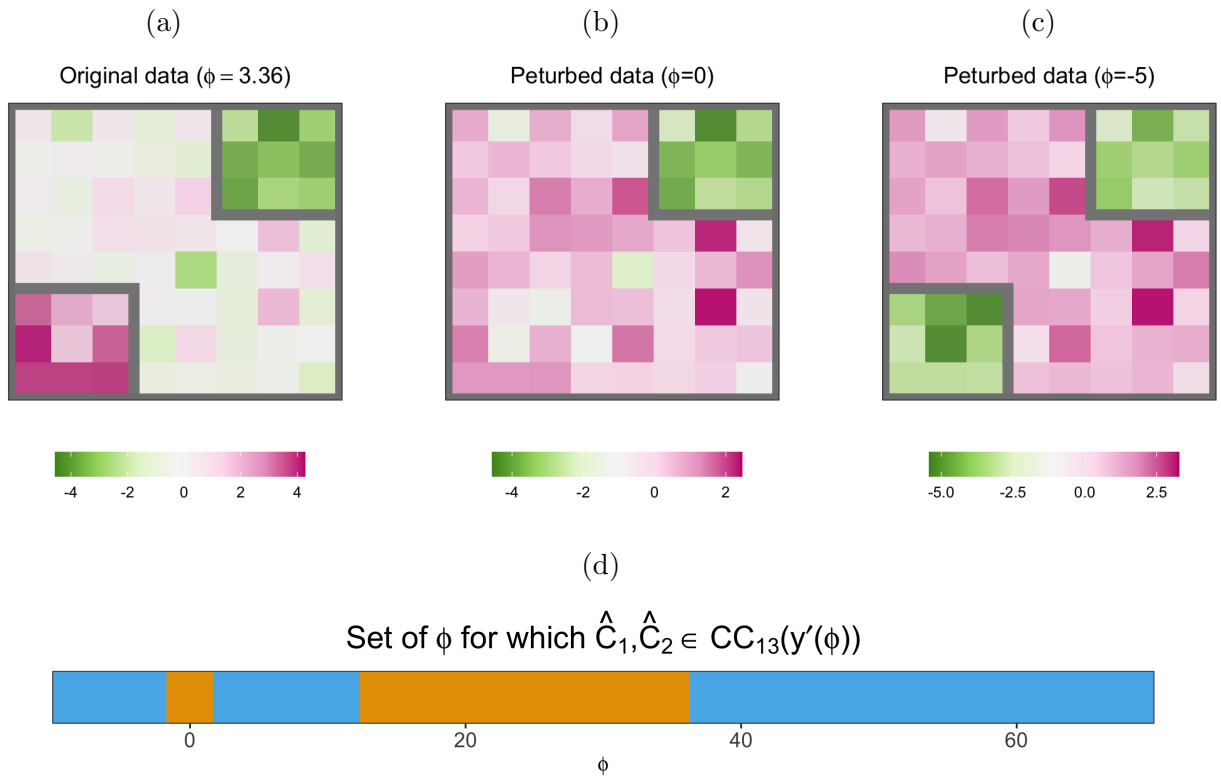


Figure 3.2: Data generated according to the model in Figure 3.1. *(a):* The data y in Figure 3.1(b) corresponds to $y'(\phi)$ with $\phi = \nu^\top y = 3.36$. Applying the graph fused lasso with $K = 13$ steps in the dual path algorithm results in three estimated connected components, displayed in grey boxes. Here, ν is chosen to test for a difference between the means of $\hat{C}_1(y)$ (lower left) and $\hat{C}_2(y)$ (middle). *(b):* The perturbed dataset $y'(\phi)$ with $\phi = 0$. Applying the graph fused lasso with $K = 13$ results in two connected components, displayed in grey boxes. *(c):* The perturbed dataset $y'(\phi)$ with $\phi = -5$. Applying the graph fused lasso with $K = 13$ results in $\hat{C}_1(y)$ and $\hat{C}_2(y)$. *(d):* The set of ϕ for which $\hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_{13}(y'(\phi))$ is displayed in blue; other values are in orange.

detects the three connected components. In panel (c), we perturb the observed data to $\phi = -5$; in this case, the graph fused lasso with $K = 13$ estimates all three connected components. Therefore, $\phi = 3.36$ and -5 are in the set $\left\{\phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_{13}(y'(\phi))\right\}$, but $\phi = 0$ is not. Panel (d) displays $\left\{\phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_{13}(y'(\phi))\right\} = (-\infty, -1.71) \cup (1.69, 12.3) \cup (36.3, \infty)$.

We now leverage ideas from Jewell et al. [2022] to develop an efficient approach to compute

the set (3.14). First, we characterize the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ in (3.14) in Proposition 10. Recall that $M_k(y) = (B_k(y), s_{B_k}(y), R_k(y), L_k(y))$ is the output of the kth step of Algorithm 3. We first present a corollary of Proposition 8.

Corollary 1. *The set* $\left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y)\}\right\}$ *is an interval.*

Proposition 10. *Let* $\mathcal{I}$ *be the set of possible outputs of Algorithm 3 that yield* $\hat{C}_1$ *and* $\hat{C}_2$ *and can be obtained via a perturbation of* y *defined in* (3.12)*, i.e.,*

$$\mathcal{I} \equiv \left\{(m_1, \ldots, m_K) : \exists \alpha \in \mathbb{R} \text{ such that } \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(y'(\alpha)), \bigcap_{k=1}^{K} \{M_k(y'(\alpha)) = m_k\}\right\}. \tag{3.15}$$

Then, there exists an index set $\mathcal{J}$ *and scalars* $\ldots < a_{-2} < a_{-1} < a_0 < a_1 < a_2 < \ldots$ *such that*

1. *the set* $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ *in* (3.14) *is the union of* $|\mathcal{J}|$ *intervals:*

$$\mathcal{S}_{\hat{C}_1,\hat{C}_2} = \left\{\phi \in \mathbb{R} : \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(y'(\phi))\right\} = \bigcup_{i \in \mathcal{J}} [a_i, a_{i+1}]; \tag{3.16}$$

2. $|\mathcal{I}| = |\mathcal{J}|$ *(i.e., the sets* $\mathcal{I}$ *and* $\mathcal{J}$ *have the same cardinality); and*

3. $\forall i \in \mathcal{J}$, $\exists (m_1, \ldots, m_K) \in \mathcal{I}$ *such that* $[a_i, a_{i+1}] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = (m_1, \ldots, m_K)\}\right\}$.

In words, Proposition 10 states that the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ in (3.14) can be expressed as a union of intervals, each of which can be computed by applying Corollary 1 on a perturbation of y. Next, we use Proposition 10 to develop an efficient recipe to compute $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ by constructing the index set $\mathcal{J}$ and scalars $\ldots < a_{-2} < a_{-1} < a_0 < a_1 < a_2 < \ldots$. To begin, we run the first K steps of the dual path algorithm on the data y. We then apply Corollary 1 to obtain the set $[a_0, a_1] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y)\}\right\}$. By construction, $[a_0, a_1] \subset \mathcal{S}_{\hat{C}_1,\hat{C}_2}$, because $\hat{C}_1$ and $\hat{C}_2$ are connected components estimated from the data y. Therefore, we initialize the

index set $\mathcal{J}$ as $\{0\}$. Then, for a small $\eta > 0$, we apply Corollary 1 to obtain the interval $\left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_1 + \eta))\}\right\}$. If the left endpoint of this interval does not equal a_1, then we must repeat with a smaller value of η until we obtain an interval of the form $[a_1, a_2]$. We can then check whether $\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(a_1+\eta))$: if so, then $[a_1, a_2] \subset \mathcal{S}_{\hat{C}_1,\hat{C}_2}$ and we update $\mathcal{J}$ to include $\{1\}$. Otherwise, $\mathcal{J}$ remains unchanged. We continue in this vein, along the positive real line, until we reach an interval for which the right endpoint equals ∞.

Finally, we proceed along the negative real line: we apply Corollary 1 to compute the interval $[a_{-1}, a_0] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_0 - \eta))\}\right\}$. If $\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(a_0 - \eta))$, then $\mathcal{J}$ is set to $\mathcal{J} \cup \{-1\}$; otherwise, $\mathcal{J}$ remains unchanged. We iterate until the algorithm outputs an interval for which the left endpoint equals $-\infty$. Finally, $\mathcal{S}_{\hat{C}_1,\hat{C}_2} = \bigcup_{i\in\mathcal{J}} [a_i, a_{i+1}]$. The procedure is summarized in Algorithm 4 of Appendix B.3.

In our implementation, we initialize with $\eta = 10^{-4}$, which proves to be an efficient choice in experiments in Section 3.5 (see details in Appendix B.3). In principle, the running time of Algorithm 4 can be quite slow, and potentially even exponential in K. However, in practice, the runtime of Algorithm 4 is nowhere near the worst-case upper bound (see Appendix B.11 for a detailed empirical study of the timing complexity of Algorithm 4). In addition, in Proposition 11, we describe an "early stopping" rule that guarantees a conservative p-value and only requires running Algorithm 4 until we reach intervals containing $|\nu^\top y| + \delta$ and $-|\nu^\top y| - \delta$ for some $\delta > 0$, as opposed to ∞ and $-\infty$. Then, the set is appended with $(-\infty, -|\nu^\top y| - \delta]$ and $[|\nu^\top y| + \delta, \infty)$. This "early stopping" rule also applies to the extensions in Section 3.4.

Proposition 11. *Provided that* $\mathbb{P}\left(\phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right) > 0$, *for any* $\delta > 0$, *we have that*

$$\mathbb{P}\left(|\phi| \geq |\nu^\top y| \;\middle|\; \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\right) \geq \mathbb{P}\left(|\phi| \geq |\nu^\top y| \;\middle|\; \phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right), \tag{3.17}$$

where $\tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \equiv (-\infty, -|\nu^\top y| - \delta] \cup \mathcal{S}_{\hat{C}_1,\hat{C}_2} \cup [|\nu^\top y| + \delta, \infty)$.

3.4 Extensions

3.4.1 Confidence intervals for $\nu^\top \beta$

We now construct a $(1-\alpha)$ confidence interval for $\nu^\top \beta$, the difference between the population means of two connected components $\hat{C}_1$ and $\hat{C}_2$ resulting from the graph fused lasso.

Proposition 12. *Suppose that (3.1) holds, and let $\hat{C}_1$ and $\hat{C}_2$ be two connected components obtained from performing K steps of the dual path algorithm for the graph fused lasso (3.3). For a given value of $\alpha \in (0,1)$, define functions $\theta_l(t)$ and $\theta_u(t)$ such that*

$$F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\theta_l(t),\sigma^2||\nu||_2^2}(t) = 1 - \frac{\alpha}{2}, \quad F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\theta_u(t),\sigma^2||\nu||_2^2}(t) = \frac{\alpha}{2}, \tag{3.18}$$

where $F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\mu,\sigma^2}(t)$ is the cumulative distribution function of a $\mathcal{N}(\mu,\sigma^2)$ random variable, truncated to the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ defined in (3.14). Then $[\theta_l(\nu^\top Y), \theta_u(\nu^\top Y)]$ has $(1-\alpha)$ selective coverage [Fithian et al., 2014, Lee et al., 2016, Tibshirani et al., 2016] for $\nu^\top \beta$, i.e.,

$$\mathbb{P}\Big(\nu^\top \beta \in \big[\theta_l(\nu^\top Y), \theta_u(\nu^\top Y)\big] \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big) = 1-\alpha. \tag{3.19}$$

Computing θ_l and θ_u in (3.18) amounts to a root-finding problem, which can be solved using bisection [Chen and Bien, 2020]. A similar result is used to construct confidence intervals corresponding to p_{Hyun} in Hyun et al. [2018].

3.4.2 An alternative conditioning set

The conditioning set for $p_{\hat{C}_1,\hat{C}_2}$ involves the connected components of the graph fused lasso solution after K steps of the dual path algorithm. However, in practice, a data analyst might prefer a more "user-facing" choice of K, such as the value that yields L connected components in the solution $\hat{\beta}$.

For this reason, we now consider a slight modification of $p_{\hat{C}_1,\hat{C}_2}$,

$$p^*_{\hat{C}_1,\hat{C}_2} = \mathbb{P}_{H_0}\Big(|\nu^\top Y| \geq |\nu^\top y| \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big), \tag{3.20}$$

where the subscript K on $\mathcal{CC}$ has been dropped, indicating that the number of steps of the graph fused lasso algorithm is no longer fixed; instead, the function $\mathcal{CC}$ now represents the graph fused lasso estimator tuned to yield exactly L connected components. Thus, in $p^*_{\hat{C}_1,\hat{C}_2}$, we condition on datasets for which $\hat{C}_1(y), \hat{C}_2(y)$ are among L connected components estimated using the graph fused lasso. It is not hard to show that Proposition 10 and Algorithm 4 require only minor modifications to enable the computation of the p-values $p^*_{\hat{C}_1,\hat{C}_2}$; details are provided in Section B.8 of the Appendix.

3.5 Simulation study

We consider testing the null hypothesis $H_0 : \nu^\top \beta = 0$ versus $H_1 : \nu^\top \beta \neq 0$, where, unless otherwise stated, ν is defined in (3.5) for a randomly-chosen pair of estimated connected components $\hat{C}_1$, $\hat{C}_2$ of the solution to (3.3). We consider three p-values: p_{Hyun} in (3.10), $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), and the naive p-value

$$p_{\text{Naive}} \equiv \mathbb{P}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y|\right), \tag{3.21}$$

and compare the selective Type I error (3.6) and power of the tests that reject H_0 when these p-values are less than $\alpha = 0.05$.

In the simulations that follow, comparing the power of the tests requires a bit of care. Because the null hypothesis $H_0 : \nu^\top \beta = 0$ involves the contrast vector ν, which is a function of the data, the effect size $|\nu^\top \beta|$ may differ across simulated datasets from the same data-generating distribution. Therefore, in what follows, we consider the power as a function of $|\nu^\top \beta|$. Alternatively, we can separately assess the detection probability (i.e., the probability that $\hat{C}_1$ and $\hat{C}_2$ are true piecewise constant segments) and the "conditional power" [Gao et al., 2020, Hyun et al., 2021] (i.e., the probability of rejecting H_0, given that $\hat{C}_1$, $\hat{C}_2$ are true piecewise constant segments). Details are in Appendix B.9.

3.5.1 One-dimensional fused lasso

We first consider the special case of the graph fused lasso on a chain graph, in which the observations are ordered, and there is an edge between each pair of adjacent observations. This leads to the one-dimensional fused lasso problem [Tibshirani and Taylor, 2011]. We simulated from the "middle mutation" model of Hyun et al. [2021], where the signal contains two true changepoints of size δ, and in turn, three connected components:

$$Y_j \overset{ind.}{\sim} \mathcal{N}(\beta_j, \sigma^2), \quad \beta_j = \delta \times 1\{101 \leq j \leq 140\}, \quad j = 1, \ldots, 200. \tag{3.22}$$

Figure 3.3(a) displays an example of this synthetic data with $\delta = 3$ and $\sigma = 1$.

Selective Type I error control under the global null

We simulated $y_1, \ldots, y_{200}$ according to (3.22) with $\delta = 0$ and $\sigma = 1$. Therefore, the null hypothesis $H_0 : \nu^\top \beta = 0$ holds for all contrast vectors ν in (3.5), regardless of the pair of estimated connected components under consideration.

We solved (3.3) with $K = 2$ steps in the dual path algorithm, which yields exactly three estimated connected components by the properties of the one-dimensional fused lasso. Then, for each simulated dataset, we computed $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), p_{Hyun} in (3.10), and the naive p-value in (3.21).

Figure 3.3(b) displays the observed p-value quantiles versus Uniform$(0, 1)$ quantiles, aggregated over 1,000 simulated datasets. We see that (i) the test based on the naive p-value in (3.21), which does not account for the fact that the connected components were estimated from the data, is anti-conservative; and (ii) tests based on p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ control the selective Type I error (3.6).

Power as a function of effect size

Next, we show that the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher power than that based on p_{Hyun}. We generated 1,500 datasets from (3.22) with $\sigma \in \{0.5, 1, 2\}$, for each of ten evenly-spaced

values of $\delta \in [0.5, 5]$. For every simulated dataset, we solved (3.3) with $K = 2$. We then rejected $H_0 : \nu^\top \beta = 0$ if p_{Hyun} or $p_{\hat{C}_1,\hat{C}_2}$ was less than $\alpha = 0.05$. Recalling that ν in (3.5) is a function of the data, and the effect size $|\nu^\top \beta|$ will vary across simulated datasets drawn from an identical distribution, we created seven evenly-spaced bins of the observed values of $|\nu^\top \beta|$, and then computed the proportion of simulated datasets for which we rejected H_0 within each bin.

Results are in Figure 3.3(c). The power of each test increases as the value of $|\nu^\top \beta|$ increases. For a given bin of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher power than the test based on p_{Hyun}. For a given test and bin of $|\nu^\top \beta|$, a smaller value of σ results in higher power. As an alternative to binning, we can use regression splines to estimate the power as a smooth function of the effect size; see Appendix B.9.

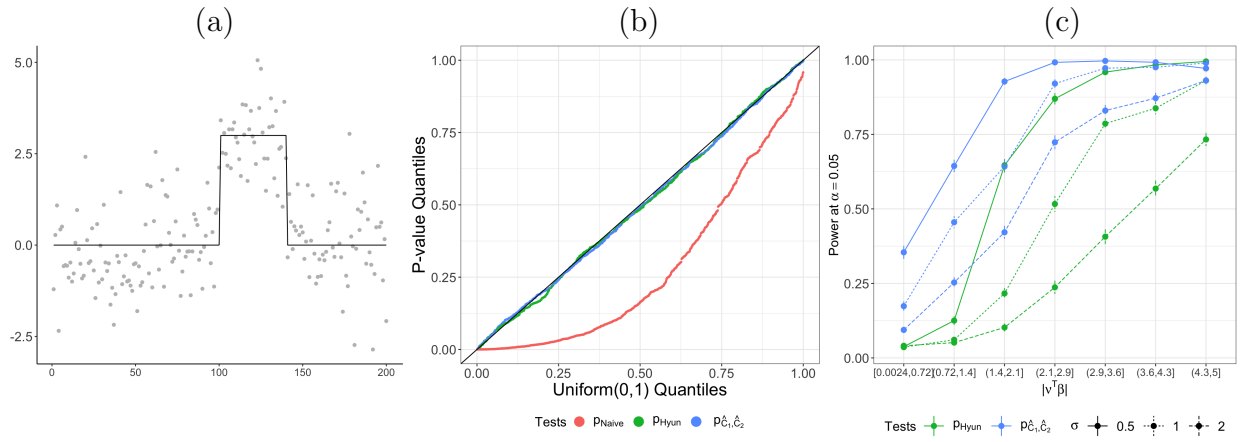


Figure 3.3: *(a):* One realization of y generated according to (3.22) with $\delta = 3$ and $\sigma = 1$ (grey dots), along with the true signal β (black curve). *(b):* When $\delta = 0$, tests based on both p_{Hyun} in (3.10) and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11) control the selective Type I error in the sense of (3.6). By contrast, the naive p-value in (3.21) leads to a test with inflated selective Type I error. *(c):* The power of the tests based on both p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ increases as a function of $|\nu^\top \beta|$. For a given bin of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher power than the test based on p_{Hyun}; the power of each test increases as σ decreases.

3.5.2 Two-dimensional fused lasso

We consider the graph fused lasso on a grid graph, constructed by connecting each node to its four closest neighbors (up, down, left, right). This leads to the two-dimensional fused lasso problem, also known as total-variation denoising [Rudin et al., 1992, Tibshirani and Taylor, 2011].

The signal β consists of with 64 observations arranged in an 8×8 grid. It has three piecewise constant segments with means δ, 0, and $-\delta$, displayed in Figure 3.4(a):

$$Y_j \overset{ind.}{\sim} \mathcal{N}(\beta_j, \sigma^2), \quad \beta_j = \delta \times 1\{j \in C_1\} + (-\delta) \times 1\{j \in C_3\}, \quad j = 1, \ldots, 64. \tag{3.23}$$

Selective Type I error control under the global null

We simulated $y_1, \ldots, y_{64}$ according to (3.23) with $\delta = 0$ and $\sigma = 1$. Thus, the null hypothesis $H_0 : \nu^\top \beta = 0$ holds for any contrast vector ν under consideration.

For each simulated dataset, we solved (3.3) with $K = 15$ steps in the dual path algorithm, which typically yields between 2 and 4 estimated connected components. Then, provided that there was more than one connected component in the solution $\hat{\beta}$, we computed p_{Naive} in (3.21), p_{Hyun} in (3.10), and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11). We rejected H_0 if the p-values are less than $\alpha = 0.05$.

Panel (b) of Figure 3.3 displays the observed p-values quantiles versus the Uniform$(0, 1)$ quantiles, over 1,000 simulated datasets. As in Section 3.5.1, the tests based on both p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ control the selective Type I error in (3.6), whereas the test based on p_{Naive} is anti-conservative.

Power as a function of effect size

We generated data according to (3.23) with each of eight evenly-spaced values of $\delta \in [0.5, 4]$ and $\sigma \in \{0.5, 1, 2\}$. For each simulated dataset, we solved (3.3) with $K = 15$ steps in the dual path algorithm. Provided that there were at least two estimated connected components,

we then computed p_{Hyun} in (3.10) and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), and rejected H_0 if the p-values were less than 0.05.

In Figure 3.4(c), we display the proportion of simulated datasets for which we rejected H_0 using the two tests, over seven evenly-spaced bins of $|\nu^\top \beta|$. For a given bin, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has substantially higher power than that based on p_{Hyun}; the power of each test increases as σ decreases.

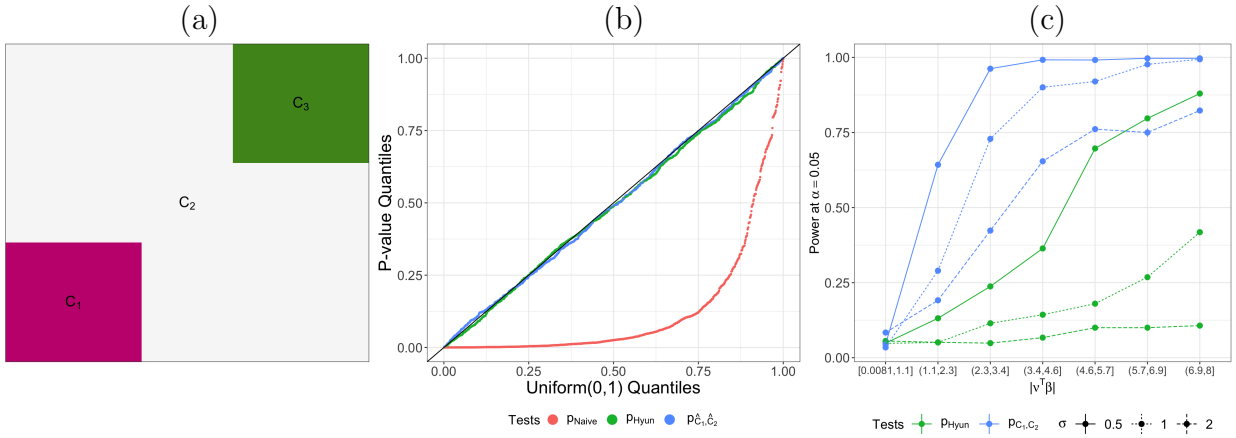


Figure 3.4: *(a):* The piecewise constant segments of β in (3.23). *(b):* When $\delta = 0$, tests based on both p_{Hyun} in (3.10) and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11) control the selective Type I error. By contrast, the test based on p_{Naive} in (3.21) has an inflated selective Type I error. *(c):* The power of the tests based on p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ increases as a function of $|\nu^\top \beta|$. For a given bin of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has substantially higher power than that based on p_{Hyun}. The power of both tests increases as σ decreases.

3.5.3 Allowing for unknown variance

Throughout this section, we have assumed that σ^2 in (3.1) is known. In Appendix B.10, we investigate the Type I error control and power of several variance estimators in simulations.

3.6 Data applications

In this section, we apply our proposed p-value $p_{\hat{C}_1,\hat{C}_2}$ to a dataset consisting of two measures: (i) drug overdose death rates (deaths per 100,000 persons), and (ii) teenage birth rates

(births per 1,000 females aged 15–19), in each of the 48 contiguous states in the United States [Centers for Disease Control and Prevention, 2020a,b]. In what follows, we consider the two measures after applying a log transformation. We can think of the data as noisy measurements of the true drug overdose death and teenage birth rate rates in each state, which are known to exhibit geographic trends [Amin et al., 2017, Schieber et al., 2019, Ventura et al., 2014]. Therefore, we solve the graph fused lasso in (3.3) with a custom graph that encodes the geography of the 48 states: each state is a node, and there is an edge between each contiguous pair of states. We then consider testing the equality of measures for pairs of estimated connected components.

For each pair of connected components, we computed three p-values: $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), p_{Hyun} in (3.10), and p_{Naive} in (3.21). We also computed confidence intervals for $\nu^\top\beta$, the difference between population means of a pair of estimated connected components, using $p_{\hat{C}_1,\hat{C}_2}$ and p_{Hyun}, as described in Section 2.4, along with the naive confidence interval $\left[\nu^\top y - z_{1-\alpha/2}\cdot\sigma\|\nu\|_2, \nu^\top y + z_{1-\alpha/2}\cdot\sigma\|\nu\|_2\right]$, where z_α is the αth quantile of the standard normal distribution. For each p-value and confidence interval, we plugged in $\hat{\sigma}^2 = \sum_{l=1}^{L}\sum_{j\in\hat{C}_l}\left\{y_j - \left(\sum_{j'\in\hat{C}_l} y_{j'}\right)/|\hat{C}_l|\right\}^2/(48-L)$ as an estimate for σ^2 in (3.1), where $\hat{C}_1,\ldots,\hat{C}_L$ are the estimated connected components.

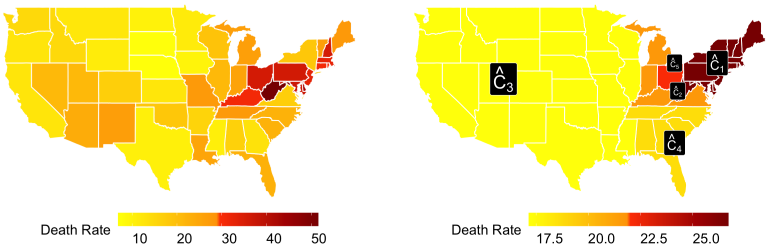
3.6.1 Drug overdose death rates in the contiguous U.S. in 2018

Figure 3.5(a) displays the drug overdose death rate in a color map. We solved (3.3) with $K = 30$ steps in the dual path algorithm, which resulted in five connected components (see Appendix B.12 for results with other choices of K); the results are displayed in Figure 3.5(b). We have estimated a constant drug overdose death rate in five geographical regions, which we refer to as the Northeast ($\hat{C}_1$), Ohio ($\hat{C}_2$), the West and Mountain region ($\hat{C}_3$), the Southeast ($\hat{C}_4$), and the Midwest ($\hat{C}_5$). Among these regions, the Northeast and Midwest have the highest estimated drug overdose rates.

We assess the equality of the means of each pair of connected components using p_{Naive} in (3.21), p_{Hyun} in (3.10), and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11). The results are in Figure 3.5(c). The subset

of pairs for which $p_{\hat{C}_1,\hat{C}_2}$ is below 0.05 and p_{Hyun} is not is displayed in bold. For instance, the Northeast ($\hat{C}_1$) and the Southeast ($\hat{C}_4$) have a statistically significant difference in mean drug overdose death rates using the test based on $p_{\hat{C}_1,\hat{C}_2}$, but not using the test based on p_{Hyun}, at level $\alpha = 0.05$. Confidence intervals corresponding to these p-values are displayed in Figure 3.5(d). Intervals based on p_{Hyun} are much wider than those based on $p_{\hat{C}_1,\hat{C}_2}$ across all ten pairs of connected components. In addition, the confidence intervals based on $p_{\hat{C}_1,\hat{C}_2}$ are not much wider than those based on p_{Naive}, even though the latter do not have correct coverage for the true parameter $\nu^\top \beta$.

(a) (b) (c)



H_0	p_{Naive}	$p_{\hat{C}_1,\hat{C}_2}$	p_{Hyun}
$\bar{\beta}_1 = \bar{\beta}_2$	0.78	0.78	0.45
$\mathbf{\bar{\beta}_1 = \bar{\beta}_3}$	**< 0.001**	**< 0.001**	**0.890**
$\mathbf{\bar{\beta}_1 = \bar{\beta}_4}$	**0.003**	**0.024**	**0.820**
$\bar{\beta}_1 = \bar{\beta}_5$	0.12	0.49	0.68
$\bar{\beta}_2 = \bar{\beta}_3$	0.007	0.64	0.58
$\bar{\beta}_2 = \bar{\beta}_4$	0.10	0.70	0.65
$\bar{\beta}_2 = \bar{\beta}_5$	0.29	0.78	0.52
$\bar{\beta}_3 = \bar{\beta}_4$	0.03	0.21	0.36
$\mathbf{\bar{\beta}_3 = \bar{\beta}_5}$	**0.003**	**0.039**	**0.650**
$\bar{\beta}_4 = \bar{\beta}_5$	0.36	0.24	0.60

(d)

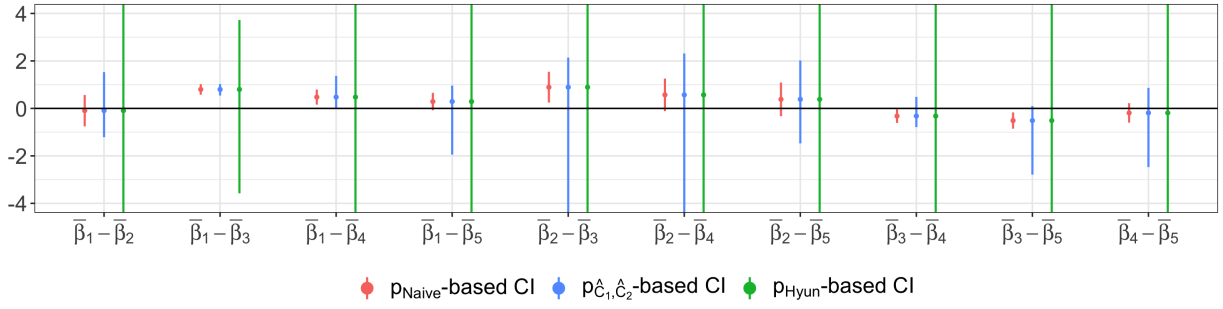


Figure 3.5: *(a):* The observed drug overdose death rates (deaths per 100,000 persons) for the 48 contiguous U.S. states in the year 2018. *(b):* Applying the graph fused lasso to the drug overdose data results in five estimated connected components. *(c):* For each pair of estimated connected components, we computed p_{Naive} in (3.21), p_{Hyun} in (3.10), and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11). For brevity, we use the notation $\bar{\beta}_l = \sum_{j \in \hat{C}_l} \beta_j / |\hat{C}_l|$. *(d):* For each pair of estimated connected components, we constructed confidence intervals for the difference in means, corresponding to p_{Naive}, p_{Hyun}, and $p_{\hat{C}_1,\hat{C}_2}$.

3.6.2 Teenage birth rates in the contiguous U.S. in 2018

Figure 3.6(a) displays the teenage birth rate in each of the 48 states. We solved the graph fused lasso with $K = 30$ steps of the dual path algorithm, which results in five estimated connected components displayed in Figure 3.6(b); Appendix B.12 contains additional results for $K = 20$. For each pair of estimated connected components, we computed the p-values p_{Naive}, p_{Hyun}, and $p_{\hat{C}_1,\hat{C}_2}$, along with the corresponding confidence intervals for the difference in means. The results are displayed in Figures 3.6(c) and (d).

As in Section 3.6.1, at level $\alpha = 0.05$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ makes more rejections than that based on p_{Hyun}. Additionally, the confidence intervals based on $p_{\hat{C}_1,\hat{C}_2}$ are much narrower than those based on p_{Hyun}; in some cases, the former are of comparable length to those based on p_{Naive}.

3.7 Discussion

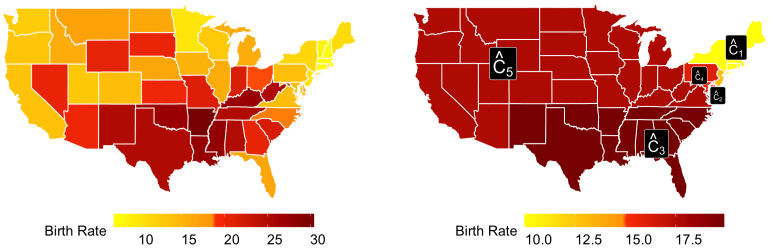
We have proposed a new procedure for testing for the difference in the means of two connected components resulting from the graph fused lasso. Our approach conditions on less information than existing approaches, leading to substantially higher power while still controlling the selective Type I error.

Methods developed in this paper are implemented in the R package `GFLassoInference`. Instructions on how to download and use this package can be found at https://yiqunchen.github.io/GFLassoInference. Code and files to reproduce the results in the paper can be found at https://github.com/yiqunchen/GFLassoInference-experiment.

3.7.1 Incorporating the selection of the tuning parameter

Throughout this paper, we have chosen K, the number of steps in the dual path algorithm for (3.3), without making use of the data. However, in practice, the tuning parameter K is often selected based on the data. For instance, we could choose the value of K that minimizes the modified Bayesian information criterion [Hyun et al., 2018, Zhang and Siegmund, 2007].

(a) (b) (c)



H_0	p_{Naive}	$p_{\hat{C}_1,\hat{C}_2}$	p_{Hyun}
$\bar{\beta}_1 = \bar{\beta}_2$	0.67	0.99	0.74
$\boldsymbol{\bar{\beta}_1 = \bar{\beta}_3}$	**< 0.001**	**< 0.001**	**0.57**
$\boldsymbol{\bar{\beta}_1 = \bar{\beta}_4}$	**0.001**	**0.005**	**0.26**
$\boldsymbol{\bar{\beta}_1 = \bar{\beta}_5}$	**< 0.001**	**< 0.001**	**0.30**
$\boldsymbol{\bar{\beta}_2 = \bar{\beta}_3}$	**< 0.001**	**0.02**	**0.260**
$\bar{\beta}_2 = \bar{\beta}_4$	0.13	0.24	0.26
$\bar{\beta}_2 = \bar{\beta}_5$	0.02	0.10	0.24
$\boldsymbol{\bar{\beta}_3 = \bar{\beta}_4}$	**< 0.001**	**0.04**	**0.71**
$\boldsymbol{\bar{\beta}_3 = \bar{\beta}_5}$	**< 0.001**	**0.002**	**0.61**
$\bar{\beta}_4 = \bar{\beta}_5$	0.34	0.38	0.71

(d)

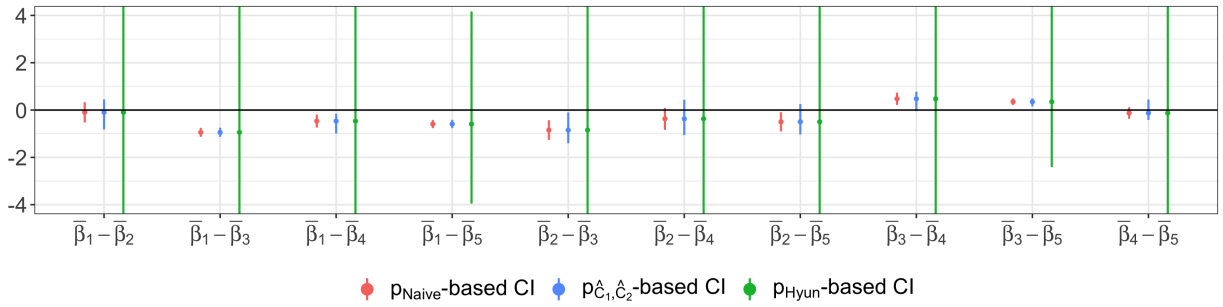


Figure 3.6: *(a):* The observed teenage birth rates (births per 1,000 females aged 15–19) for the 48 contiguous U.S. states in 2018. *(b):* The graph fused lasso solution with $K = 30$ results in five connected components, displayed in distinct colors. *(c):* For each pair of estimated connected components, we computed p_{Naive} in (3.21), p_{Hyun} in (3.10), and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11). Pairs for which the test based on $p_{\hat{C}_1,\hat{C}_2}$ results in a rejection at $\alpha = 0.05$, but not for the test based on p_{Hyun}, are in bold. *(d):* Confidence intervals for the differences in means for each pair of connected components.

We leave details to future work.

3.7.2 Extension to other generalized lasso problems

Ideas in this paper apply beyond the setting of the piecewise constant model in (3.1) and the graph fused lasso estimator in (3.3). For instance, we can consider extending our proposal to the trend filtering problem, which postulates that the underlying signal is ordered and piecewise polynomial [Kim et al., 2009, Tibshirani, 2014]. Because trend filtering is a special case of (3.2) and can be solved using the dual path algorithm, an extension of the approach in Section 3.3 can be applied.

In addition, we can extend our proposal from an identity matrix in (3.2) to any design matrix $X \in \mathbb{R}^{n \times q}$ with full column rank, i.e., $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \right\}$. Hyun et al. [2018] showed that a p-value similar to (3.10) can be used in this case to test the hypothesis (3.4). Therefore, we can directly apply the computational insights in Section 3.3 to obtain a more powerful test.

We leave the details of outlined extensions, as well as comparisons to recent selective inference tools for trend filtering (e.g., Leiner et al. [2021], Mehrizi and Chenouri [2021]), to future work.

3.7.3 Relaxing assumptions in (3.1)

While the idea of conditioning on less information to improve the power of a selective inference procedure applies regardless of the distributions of the observations, the assumptions in model (3.1) are critical to the proof of Proposition 9, and therefore, the efficient computation of $p_{\hat{C}_1, \hat{C}_2}$. A line of recent work in selective inference has focused on relaxing these assumptions in high-dimensional linear modeling [Charkhi and Claeskens, 2018, Tian and Taylor, 2018, Tibshirani et al., 2018a], and may be applicable to the generalized lasso. Alternatively, we can extend (3.1) to other exponential family distributions by leveraging the recent developments in generalized data carving [Leiner et al., 2021, Rasines and Alastair Young, 2021, Schultheiss et al., 2021].

Chapter 4

SELECTIVE INFERENCE FOR K-MEANS CLUSTERING

4.1 Introduction

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

In this paper, we develop a test for a difference in means between two clusters estimated

from applying k-means clustering [Lloyd, 1982, MacQueen et al., 1967], an extremely popular clustering algorithm with numerous applications [Hand and Adams, 2015, Xu and Wunsch, 2008]. We consider the following simple and well-studied model [Gao et al., 2020, Löffler et al., 2021, Lu and Zhou, 2016] for n observations and q features:

$$X \sim \mathcal{MN}_{n\times q}\left(\mu, \mathbf{I}_n, \sigma^2\mathbf{I}_q\right), \tag{4.1}$$

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

$$H_0 : \sum_{i\in\hat{\mathcal{C}}_1} \mu_i/|\hat{\mathcal{C}}_1| = \sum_{i\in\hat{\mathcal{C}}_2} \mu_i/|\hat{\mathcal{C}}_2| \text{ versus } H_1 : \sum_{i\in\hat{\mathcal{C}}_1} \mu_i/|\hat{\mathcal{C}}_1| \neq \sum_{i\in\hat{\mathcal{C}}_2} \mu_i/|\hat{\mathcal{C}}_2|, \tag{4.2}$$

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

$$\nu_i = 1\left\{i \in \hat{\mathcal{C}}_1\right\}/|\hat{\mathcal{C}}_1| - 1\left\{i \in \hat{\mathcal{C}}_2\right\}/|\hat{\mathcal{C}}_2|, \quad i = 1, \ldots, n, \tag{4.3}$$

and $1\{A\}$ equals 1 if the event A holds, and 0 otherwise. At first glance, we can test the hypothesis in (4.2) by applying a Wald test, with p-value given by

$$p_{\text{Naive}} = \mathbb{P}_{H_0}\left(\|X^\top\nu\|_2 \geq \|x^\top\nu\|_2\right), \tag{4.4}$$

where $\|X^\top\nu\|_2 \sim (\sigma\|\nu\|_2)\chi_q$ under H_0. But this "naive" approach ignores the fact that H_0 is chosen based on the data, i.e., we constructed the contrast vector in (4.3) because $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ were obtained by clustering. Therefore, we will observe substantial differences between the cluster centroids $\sum_{i\in\hat{\mathcal{C}}_1} x_i/|\hat{\mathcal{C}}_1|$ and $\sum_{i\in\hat{\mathcal{C}}_2} x_i/|\hat{\mathcal{C}}_2|$, even in the absence of true differences in their population means (left panel Figure 4.1). The center panel of Figure 4.1 illustrates that

the test based on (4.4) does not control the selective Type I error: that is, the probability of falsely rejecting a null hypothesis, given that we decided to test it [Fithian et al., 2014].

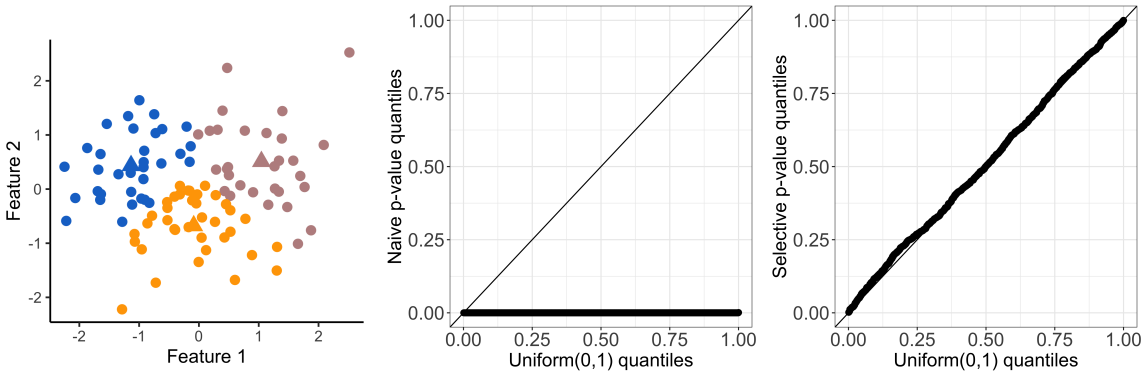


Figure 4.1: *Left:* One simulated dataset generated from (4.1) with $\mu = 0_{100\times 2}$ and $\sigma = 1$. We apply k-means clustering to obtain three clusters. The cluster centroids are displayed as triangles. *Center:* Quantile-quantile plot of the naive p-values (defined in (4.4)) applied to 2,000 simulated datasets from (4.1) with $\mu = 0_{100\times 2}$ and $\sigma = 1$. *Right:* Quantile-quantile plot of our proposed p-values (defined in (4.9)) applied to the same simulated datasets.

Notably, the problem of testing for a difference in means between two groups obtained via clustering cannot be easily overcome by sample splitting, as pointed out in Gao et al. [2020] and Zhang et al. [2019]. To see why, we divide the observations into a training and a test set. We apply k-means clustering on only the training set (left panel of Figure 4.2), and then assign the test set observations to those clusters (to obtain the center panel of Figure 4.2, we applied a 3-nearest neighbor classifier). Finally, we compute the naive p-values (4.4) *only* on the test set. Unfortunately, this approach does not work: while we clustered only the training data, we still used the test data to label the test observations, and consequently to construct the contrast vector ν in (4.3). Therefore, the Wald test based on sample-splitting remains extremely anti-conservative, as shown in the right panel of Figure 4.2, and does not lead to a valid test of H_0 in (4.2).

In this paper, we develop a test of H_0 that controls the selective Type I error. That is, we wish to ensure that the probability of rejecting H_0 at level α, given that H_0 holds and

we decided to test it, is no greater than α:

$$\mathbb{P}_{H_0}(\text{reject } H_0 \text{ at level } \alpha \mid H_0 \text{ is tested}) \leq \alpha, \ \forall \alpha \in (0,1). \tag{4.5}$$

To develop the test, we leverage the selective inference framework, which has been applied extensively in high-dimensional linear modeling [Charkhi and Claeskens, 2018, Fithian et al., 2014, Lee et al., 2016, Loftus and Taylor, 2014, Rügamer et al., 2022, Schultheiss et al., 2021, Taylor and Tibshirani, 2018, Tibshirani et al., 2016, Yang et al., 2016], changepoint detection [Benjamini et al., 2019, Chen et al., 2021b, Duy et al., 2020, Hyun et al., 2018, 2021, Jewell et al., 2022, Le Duy and Takeuchi, 2021], and clustering [Gao et al., 2020, Watanabe and Suzuki, 2021, Zhang et al., 2019]. The key insight behind selective inference is as follows: to obtain a valid test of H_0, we need to *condition* on the aspect of the data that led us to test it. In our case, we chose to test the null hypothesis in (4.2) because $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ were obtained via k-means clustering. Therefore, we compute a p-value conditional on the event that k-means clustering yields $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$. This results in selective Type I error control (4.5), as seen in the right panel of Figure 4.1.

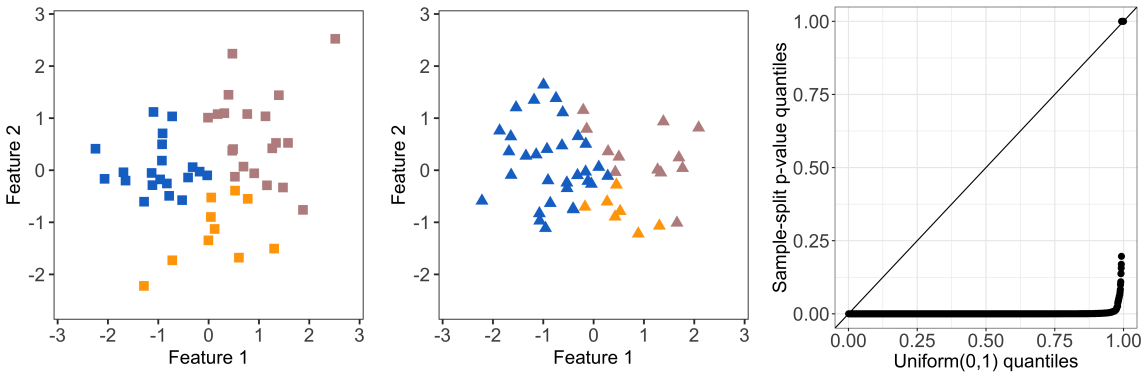


Figure 4.2: *Left:* One simulated dataset generated from (4.1) with $\mu = 0_{100\times 2}$ and $\sigma = 1$. We apply k-means clustering on the training set to obtain three clusters. *Center:* We apply the training set clusters to the test set using a 3-nearest neighbors classifier. *Right:* Quantile-quantile plot of the naive p-values (4.4) applied to the test set, aggregated over 2,000 simulated datasets.

There is a rich literature on estimating and quantifying the uncertainty in the number of clusters [Chen et al., 2004, Chen and Li, 2009, Dobriban, 2020, Li and Chen, 2010, McLachlan et al., 2019], as well as assessing cluster stability and heterogeneity [Aw et al., 2021, Chung, 2020, Chung and Storey, 2015, Jin and Wang, 2016, Kerr and Churchill, 2001, Kimes et al., 2017, Suzuki and Shimodaira, 2006]. Others have examined the asymptotic properties of clustering models from a Bayesian perspective [Cai et al., 2020, Guha et al., 2019, Nobile, 2004]. In addition, k-means clustering is a special case of the expectation-maximization algorithm, which allows us to tap into the active line of research on the statistical guarantees of the expectation-maximization algorithm [Balakrishnan et al., 2017, Cai et al., 2019, Wang et al., 2015, Yi and Caramanis, 2015, Zhang and Zhang, 2014]. However, most prior work focused on scenarios where the number of clusters is correctly specified, and the estimated clusters memberships are close to the truth. By contrast, we are interested in a correctly-sized test for the null hypothesis (4.2), even when $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$ do not correspond to true clusters. In addition, existing work often relies on asymptotic approximations and bootstrap resampling. Two recent exceptions include Zhang et al. [2019] and Gao et al. [2020], who took a selective inference approach and computed finite-sample p-values for testing the difference in means between estimated clusters obtained via linear classification rules and hierarchical clustering, respectively. Our work is closest to Gao et al. [2020], and extends their framework to k-means clustering. We provide an exact, finite-sample test of the difference in means between a pair of clusters estimated via k-means clustering under model (4.1), without the need for sample splitting.

The rest of this paper is organized as follows. In Section 4.2, we briefly review the work of Gao et al. [2020], and outline our proposed test of a difference in means after k-means clustering. In Section 4.3, we provide a computationally-efficient approach to compute the p-values corresponding to our proposed test. Section 4.4 outlines some extensions, and we evaluate our proposal in a simulation study in Section 4.5. We apply our proposal to three real datasets in Section 4.6, and discuss future work in Section 4.7. Proofs and additional results are relegated to the Appendix.

Throughout this paper, we will use the following notational conventions. For a matrix

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

4.2 Selective inference for k-means clustering

4.2.1 A brief review of k-means clustering

In this section, we review the k-means clustering algorithm. Given samples $x_1, \ldots, x_n \in \mathbb{R}^q$ and a positive integer K, k-means clustering partitions the n samples into disjoint subsets $\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K$ by solving the following optimization problem:

$$\begin{aligned} &\underset{\mathcal{C}_1,\ldots,\mathcal{C}_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \left\| x_i - \sum_{i \in \mathcal{C}_k} x_i/|\mathcal{C}_k| \right\|_2^2 \right\} \\ &\text{subject to } \bigcup_{k=1}^{K} \mathcal{C}_k = \{1, \ldots, n\}, \mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset, \forall k \neq k'. \end{aligned} \tag{4.6}$$

It is not typically possible to solve for the global optimum in (4.6) [Aloise et al., 2009]. A number of algorithms are available to find a local optimum [Arthur and Vassilvitskii, 2007, Hartigan and Wong, 1979, Zha et al., 2002]; one such approach is Lloyd's algorithm [Lloyd, 1982], given in Algorithm 1. We first sample K out of n observations as initial centroids (step 1 in Algorithm 1). We then assign each observation to the closest centroid (step 2). Next, we iterate between re-computing the centroids and updating the cluster assignments (steps 3a. and 3b.) until the cluster assignments stop changing. The algorithm is guaranteed to converge to a local optimum [Hastie et al., 2001].

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

Algorithm 1: Lloyd's algorithm for k-means clustering [Lloyd, 1982]

Input: Data $x_1, \ldots, x_n \in \mathbb{R}^q$, number of output clusters K, maximum iteration T, random seed s.

Output: Cluster assignments $\left(c_1^{(t)}, \ldots, c_n^{(t)}\right)$.

1. Initialize the centroids $\left(m_1^{(0)}, \ldots, m_K^{(0)}\right)$ by sampling K observations from $x_1, \ldots, x_n$ without replacement, using the random seed s.
2. Compute assignments $c_i^{(0)} \leftarrow \underset{1 \le k \le K}{\operatorname{argmin}} \left\| x_i - m_k^{(0)} \right\|_2^2$, $i = 1, \ldots, n$.
3. Initialize $t = 0$.

while $t \le T$ **do**

a. Update centroids: $m_k^{(t+1)} \leftarrow \left(\sum_{i:c_i^{(t)}=k} x_i\right) / \sum_{i=1}^n 1\left\{c_i^{(t)} = k\right\}$, $k = 1, \ldots, K$.

b. Update assignment: $c_i^{(t+1)} \leftarrow \underset{1 \le k \le K}{\operatorname{argmin}} \left\| x_i - m_k^{(t+1)} \right\|_2^2$, $i = 1, \ldots, n$.

c. **if** $c_i^{(t+1)} = c_i^{(t)}$ *for all* $1 \le i \le n$
    break
else
    $t \leftarrow t + 1$.

end

return $\left(c_1^{(t)}, \ldots, c_n^{(t)}\right)$.

4.2.2 A test of (4.2) for clusters obtained via k-means clustering

Here, we briefly review the proposal of Gao et al. [2020] for selective inference for hierarchical clustering, and outline a selective test for (4.2) for k-means clustering.

Gao et al. [2020] proposed a selective inference framework for testing hypotheses based on the output of a clustering algorithm. Let $\mathcal{C}(\cdot)$ denote the clustering operator, i.e., a partition of the observations resulting from a clustering algorithm. Since H_0 in (4.2) is chosen because $\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(x)\right\}$, where $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$ are the two estimated clusters under consideration in (4.2), Gao et al. [2020] proposed to reject H_0 if

$$\mathbb{P}_{H_0}\left\{\|X^\top \nu\|_2 \geq \|x^\top \nu\|_2 \;\middle|\; \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(X), \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right\} \tag{4.7}$$

is small. In (4.7), conditioning on $\left\{\Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right\}$ eliminates the nuisance parameters $\Pi_\nu^\perp \mu$ and $\mathrm{dir}(\mu^\top \nu)$, where $\Pi_\nu^\perp = \mathbf{I}_n - \nu\nu^\top/\|\nu\|_2$ and $\mathrm{dir}(\mu^\top \nu) = \mu^\top \nu / \|\mu^\top \nu\|_2$ (see, e.g., Section 3.1 of Fithian et al. [2014]). Gao et al. [2020] showed that the test that rejects H_0 when (4.7) is below α controls the selective Type I error at level α, in the sense of (4.5). Furthermore, under (4.1), the conditional distribution of $\|X^\top \nu\|_2$ in (4.7) is $(\sigma\|\nu\|_2)\chi_q$, truncated to a set. When the operator $\mathcal{C}(\cdot)$ denotes hierarchical clustering, this set can be analytically characterized and efficiently computed, leading to an efficient algorithm for computing (4.7).

We now extend these ideas to k-means clustering (4.6). Since the k-means algorithm partitions all n observations, it is natural to condition on the cluster assignments of *all* observations rather than just on $\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(X)\right\}$. This leads to the p-value

$$\mathbb{P}_{H_0}\left[\|X^\top \nu\|_2 \geq \|x^\top \nu\|_2 \;\middle|\; \bigcap_{i=1}^{n}\left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \mathrm{dir}(X^\top \nu) = \mathrm{dir}(x^\top \nu)\right], \tag{4.8}$$

where $c_i^{(T)}(X)$ is the cluster assigned to the ith observation at the final iteration of Algorithm 1. However, computing (4.8) requires characterizing $\bigcap_{i=1}^n \left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}$, which is not straightforward, and may necessitate enumerating over possibly an exponential num-

ber of intermediate cluster assignments $c_i^{(t)}(\cdot)$ for $t = 1, \ldots, T-1$. Hence, we also condition on *all of the intermediate clustering assignments* in Algorithm 1:

$$p_{\text{selective}} = \mathbb{P}_{H_0}\Bigg[\|X^\top \nu\|_2 \geq \|x^\top \nu\|_2 \;\Bigg|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \\ \text{dir}(X^\top \nu) = \text{dir}(x^\top \nu)\Bigg]. \tag{4.9}$$

In (4.9), $c_i^{(t)}(X)$ is the cluster assigned to the ith observation at the tth iteration of Algorithm 1. Roughly speaking, this p-value answers the question:

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

The p-value in (4.9) is the focus of this paper. We establish its key properties below.

Proposition 13. *Suppose that x is a realization from (4.1), and let $\phi \sim (\sigma\|\nu\|_2)\chi_q$. Then, under $H_0 : \mu^\top \nu = 0$ with ν defined in (4.3),*

$$p_{selective} = \mathbb{P}\left[\phi \geq \|x^\top \nu\|_2 \;\Bigg|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(x'(\phi)) = c_i^{(t)}(x)\right\}\right], \tag{4.10}$$

where $p_{selective}$ is defined in (4.9), and

$$x'(\phi) = x + \left(\phi - \|x^\top \nu\|_2\right)\left(\nu/\|\nu\|_2^2\right)\left\{\text{dir}(x^\top \nu)\right\}^\top. \tag{4.11}$$

Moreover, the test that rejects $H_0 : \mu^\top \nu = 0$ when $p_{selective} \leq \alpha$ controls the selective Type I error at level α, in the sense of (4.5).

Proposition 13 states that $p_{\text{selective}}$ can be recast as the survival function of a scaled χ_q

random variable, truncated to the set

$$\mathcal{S}_T = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}, \tag{4.12}$$

where $x'(\phi)$ is defined in (4.11). Therefore, to compute $p_{\text{selective}}$, it suffices to characterize the set $\mathcal{S}_T$. In (4.11), $x'(\phi)$ results from applying a perturbation to the observed data x, along

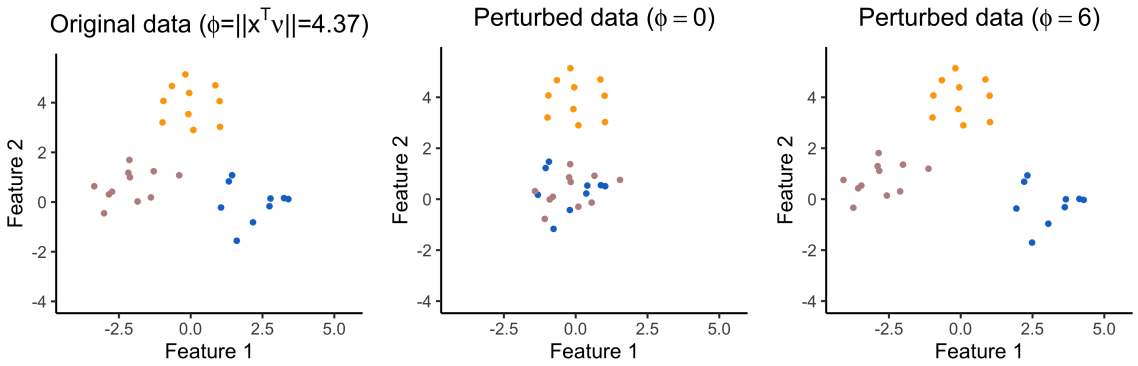


Figure 4.3: One simulated dataset generated from model (4.1) with $\mu_i = 1\{1 \leq i \leq 10\}\left[\begin{smallmatrix}2.5\\0\end{smallmatrix}\right] + 1\{11 \leq i \leq 20\}\left[\begin{smallmatrix}0\\-2.5\end{smallmatrix}\right] + 1\{21 \leq i \leq 30\}\left[\begin{smallmatrix}\sqrt{18.75}\\0\end{smallmatrix}\right]$ and $\sigma = 1$. *Left:* The original data x corresponds to $\phi = \|x^\top \nu\|_2 = 4.37$. Applying k-means clustering with $K = 3$ yields three clusters, displayed in pink, blue, and orange. Here, ν is chosen to test for a difference in means between $\hat{\mathcal{C}}_1$ (pink) and $\hat{\mathcal{C}}_2$ (blue). *Center:* The perturbed data $x'(\phi)$ with $\phi = 0$. Applying k-means clustering with $K = 3$ does not yield the same set of clusters as in the left panel. *Right:* The perturbed data $x'(\phi)$ with $\phi = 6$. Applying k-means clustering with $K = 3$ yields the same set of clusters as in the left panel.

the direction of $x^\top \nu$, the difference between the two cluster centroids of interest. Figure 4.3 illustrates a realization of (4.1) for k-means clustering with $K = 3$. The left panel displays the observed data x, which corresponds to $x'(\phi)$ with $\phi = \|x^\top \nu\|_2 = 4.37$. Here, ν defined in (4.3) was chosen to test the difference between $\hat{\mathcal{C}}_1$ (shown in pink) and $\hat{\mathcal{C}}_2$ (shown in blue). The center and right panels of Figure 4.3 display $x'(\phi)$ with $\phi = 0$ and $\phi = 6$, respectively. In the center panel, with $\phi = 0$, the blue and pink clusters are "pushed together", resulting in $\|x'(\phi)^\top \nu\|_2 = 0$; that is, there is no difference in empirical means between the two clusters under consideration. Applying k-means clustering no longer results in these clusters. By

contrast, in the right panel, with $\phi = 6$, the blue and pink clusters are "pulled apart" along the direction of $x^\top \nu$, which results in an increased distance between the centroids of the blue and pink clusters, and k-means clustering does yield the same clusters as on the original data. In this example, $\mathcal{S}_T = (3.59, \infty)$.

4.3 Computation of the selective p-value

In Section 4.2, we have shown that the p-value $p_{\text{selective}}$ (4.9) involves the set $\mathcal{S}_T$ in (4.12). Here, we start with a high-level summary of our approach to characterizing $\mathcal{S}_T$. We rewrite $\mathcal{S}_T$ as $\left\{\phi \in \mathbb{R} : \bigcap_{i=1}^n \left\{c_i^{(0)}\left(x'(\phi)\right) = c_i^{(0)}\left(x\right)\right\}\right\} \cap \left\{\phi \in \mathbb{R} : \bigcap_{t=1}^T \bigcap_{i=1}^n \left\{c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right)\right\}\right\}$. Next, we consider the first term in the intersection: according to step 2. of Algorithm 1, for $i = 1, \ldots, n$, $c_i^{(0)}\left(x'(\phi)\right) = c_i^{(0)}\left(x\right)$ if and only if for $i = 1, \ldots, n$, the initial randomly-sampled centroid to which $[x'(\phi)]_i$ is closest coincides with the initial centroid to which x_i is closest. This condition can be expressed using $K - 1$ inequalities. Furthermore, the same intuition holds for the second term in the intersection, except that the centroids are a function of the cluster assignments in the previous iteration. We formalize this intuition in Proposition 14, proven in Appendix C.2.

Proposition 14. *Suppose that we apply the k-means clustering algorithm (Algorithm 1) to a matrix $x \in \mathbb{R}^{n \times q}$, to obtain K clusters in at most T steps. Define*

$$w_i^{(t)}(k) = 1\left\{c_i^{(t)}(x) = k\right\} / \sum_{i'=1}^n 1\left\{c_{i'}^{(t)}(x) = k\right\}. \tag{4.13}$$

Then, for the set $\mathcal{S}_T$ defined in (4.12), we have that

$$\mathcal{S}_T = \left(\bigcap_{i=1}^n \bigcap_{k=1}^K \left\{\phi : \left\|[x'(\phi)]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi))\right\|_2^2 \leq \left\|[x'(\phi)]_i - m_k^{(0)}(x'(\phi))\right\|_2^2\right\}\right) \cap \tag{4.14}$$

$$\left(\bigcap_{t=1}^T \bigcap_{i=1}^n \bigcap_{k=1}^K \left\{\phi : \left\|[x'(\phi)]_i - \sum_{i'=1}^n w_{i'}^{(t-1)}\left(c_i^{(t)}(x)\right)[x'(\phi)]_{i'}\right\|_2^2 \leq \left\|[x'(\phi)]_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)[x'(\phi)]_{i'}\right\|_2^2\right\}\right). \tag{4.15}$$

Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned in step 3b.

of Algorithm 1 during the tth iteration, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from the data x during step 1 of Algorithm 1. In words, Proposition 14 says that $\mathcal{S}_T$ can be expressed as the intersection of $\mathcal{O}(nKT)$ sets. Therefore, it suffices to characterize the sets in (4.14) and (4.15). We now present two lemmas.

Lemma 1 (Lemma 2 in Gao et al. [2020]). *For ν in (4.3) and $x'(\phi)$ in (4.11), we have that $\left\|[x'(\phi)]_i - [x'(\phi)]_j\right\|_2^2 = a\phi^2+b\phi+\gamma$, where $a = \left\{(\nu_i-\nu_j)/\|\nu\|_2^2\right\}^2$, $b = 2\left[(\nu_i-\nu_j)/\|\nu\|_2^2 \left\langle x_i - x_j, \mathrm{dir}(x^\top\nu)\right\rangle - \left\{(\nu_i-\nu_j)/\|\nu\|_2^2\right\}^2\|x^\top\nu\|_2\right]$, and $\gamma = \left\|x_i - x_j - (\nu_i-\nu_j)(x^\top\nu)/||\nu||_2^2\right\|_2^2$.*

Lemma 2. *For ν in (4.3), $x'(\phi)$ in (4.11), and $w_i^{(t)}(k)$ in (4.13), we have that $\left\|[x'(\phi)]_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)[x'(\phi)]_{i'}\right\|_2^2 = \tilde{a}\phi^2 + \tilde{b}\phi + \tilde{\gamma}$, where $\tilde{a} = \left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)^2/\|\nu\|_2^4$, $\tilde{b} = (2/\|\nu\|_2^2)\left\{\left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)\left\langle x_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)x_{i'}, \mathrm{dir}(x^\top\nu)\right\rangle - \left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)^2(\|x^\top\nu\|_2)/\|\nu\|_2^4\right\}$ and $\tilde{\gamma} = \left\|x_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)x_{i'} - \left(\nu_i - \sum_{i'=1}^n w_{i'}^{(t-1)}(k)\nu_{i'}\right)(x^\top\nu)/\|\nu\|_2^2\right\|_2^2$.*

It follows from Lemmas 1 and 2 that all of the inequalities in (4.14) and (4.15) are in fact *quadratic* in ϕ, with coefficients that can be analytically computed. Therefore, computing the set $\mathcal{S}_T$ requires solving $\mathcal{O}(nKT)$ quadratic inequalities of ϕ.

Proposition 15. *Suppose that we apply the k-means clustering algorithm (Algorithm 1) to a matrix $x \in \mathbb{R}^{n\times q}$, to obtain K clusters in at most T steps. Then, the set $\mathcal{S}_T$ defined in (4.12) can be computed in $\mathcal{O}(nKT(n+q) + nKT\log(nKT))$ operations.*

4.4 Extensions

4.4.1 Non-spherical covariance matrix

Thus far, we have assumed that the observed data x is a realization of (4.1), which implies that $\mathrm{cov}(X_i) = \sigma^2\mathbf{I}_q$. However, this assumption is often violated in practice. For example, expression levels of genes are highly correlated, and neighbouring pixels in an image tend to be more similar. For a known positive definite matrix Σ, we now let

$$X \sim \mathcal{MN}_{n\times q}(\mu, \mathbf{I}_n, \Sigma). \tag{4.16}$$

Under (4.16), we can whiten the data by applying the transformation $x_i \to \Sigma^{-\frac{1}{2}} x_i$ [Bell and Sejnowski, 1997], where $\Sigma^{-\frac{1}{2}}$ is the unique symmetric positive definite square root of Σ^{-1} [Horn and Johnson, 2012]. Note that $\Sigma^{-\frac{1}{2}} X_i \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\mu_i, \mathbf{I}_q)$. Moreover, as $\Sigma^{-\frac{1}{2}} \succ 0$, testing the null hypothesis in (4.2) is equivalent to testing

$$H_0 : \sum_{i\in\hat{\mathcal{C}}_1} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_1| = \sum_{i\in\hat{\mathcal{C}}_2} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_2| \text{ versus } H_1 : \sum_{i\in\hat{\mathcal{C}}_1} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_1| \neq \sum_{i\in\hat{\mathcal{C}}_2} \Sigma^{-\frac{1}{2}}\mu_i/|\hat{\mathcal{C}}_2|. \quad (4.17)$$

Therefore, to get a correctly-sized test under model (4.16), we can simply carry out our proposal in Section 4.2 on the transformed data $\Sigma^{-\frac{1}{2}} x_i$ instead of the original data x_i.

Instead of applying the whitening transformation, we can directly accommodate a known covariance matrix Σ by considering the following extension of $p_{\text{selective}}$ in (4.9):

$$\begin{aligned} p_{\Sigma,\text{selective}} = \mathbb{P}_{H_0}\Big[\|\Sigma^{-\frac{1}{2}}X^\top\nu\|_2 \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \mid \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \\ \Pi_\nu^\perp X = \Pi_\nu^\perp x,\ \operatorname{dir}\left(\Sigma^{-\frac{1}{2}}X^\top\nu\right) = \operatorname{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\Big]. \end{aligned} \quad (4.18)$$

Proposition 16. *Suppose that x is a realization from (4.16), and let $\phi \sim (\|\nu\|_2)\chi_q$. Then, under $H_0 : \mu^\top\nu = 0$ with ν defined in (4.3),*

$$p_{\Sigma,selective} = \mathbb{P}\left[\phi \geq \|\Sigma^{-\frac{1}{2}}x^\top\nu\|_2 \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\phi\frac{\nu}{\|\nu\|_2^2}\right)\left\{\operatorname{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x)\right\}\right], \quad (4.19)$$

where $p_{\Sigma,selective}$ is defined in (4.18). Furthermore, the test that rejects $H_0 : \mu^\top\nu = 0$ when $p_{\Sigma,selective} \leq \alpha$ controls the selective Type I error at level α.

In addition, we can adapt the results in Section 4.3 to compute the set $\left\{\phi \in \mathbb{R} : \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + (\phi\nu/\|\nu\|_2^2)\left\{\operatorname{dir}\left(\Sigma^{-\frac{1}{2}}x^\top\nu\right)\right\}^\top\Sigma^{\frac{1}{2}}\right) = c_i^{(t)}(x)\right\}\right\}$ by modifying the results in Lemmas 1 and 2. Details are in Section C.5 of the Appendix.

4.4.2 Unknown variance

When σ is unknown, we can plug in an estimate $\hat{\sigma}$ in (4.9):

$$\hat{p}_{\text{selective}}(\hat{\sigma}) = \mathbb{P}\left[\phi(\hat{\sigma}) \geq \|x^\top \nu\|_2 \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'(\phi(\hat{\sigma}))\right) = c_i^{(t)}\left(x\right)\right\}\right], \tag{4.20}$$

where $\phi(\hat{\sigma}) \sim (\hat{\sigma}\|\nu\|_2)\chi_q$. If we use a consistent estimator of σ, then a test based on the p-value in (4.20) provides selective Type I error control (4.5) asymptotically.

Proposition 17. *For $q = 1, 2, \ldots$, suppose that $X^{(q)} \sim \mathcal{MN}_{n\times q}\big(\mu^{(q)}, \mathbf{I}_n, \sigma^2\mathbf{I}_q\big)$. Let $x^{(q)}$ be a realization from $X^{(q)}$ and let $c_i^{(t)}(\cdot)$ be the cluster to which the ith observation is assigned during the tth iteration of step 3b. in Algorithm 1. Consider the sequence of null hypotheses $H_0^{(q)} : {\mu^{(q)}}^\top \nu^{(q)} = 0_q$, where $\nu^{(q)}$ defined in (4.3) is the contrast vector resulting from applying k-means clustering on $x^{(q)}$. Suppose that (i) $\hat{\sigma}$ is a consistent estimator of σ, i.e., for all $\epsilon > 0, \lim_{q\to\infty}\mathbb{P}\big(|\hat{\sigma}(X^{(q)}) - \sigma| \geq \epsilon\big) = 0$; and (ii) there exists $\delta \in (0,1)$ such that $\lim_{q\to\infty}\mathbb{P}_{H_0^{(q)}}\left[\bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] > \delta$. Then, for all $\alpha \in (0,1)$, we have that $\lim_{q\to\infty}\mathbb{P}_{H_0^{(q)}}\left[\hat{p}_{selective}(\hat{\sigma}) \leq \alpha \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] = \alpha$.*

In practice, we propose to use the following estimator of σ [Huber, 1981]:

$$\hat{\sigma}_{\text{MED}}(x) = \left\{\underset{1\leq i\leq n, 1\leq j\leq q}{\text{median}}\left(\tilde{x}_{ij}^2\right)/M_{\chi_1^2}\right\}^{1/2}, \tag{4.21}$$

where $\tilde{x}$ is obtained from subtracting the median of each column in x, and $M_{\chi_1^2}$ is the median of the χ_1^2 distribution. If μ is sparse, i.e., $\sum_{i=1}^{n}\sum_{j=1}^{q} 1\{\mu_{ij} \neq 0\}$ is small, then (4.21) is consistent with appropriate assumptions; see Appendix C.7.

4.5 Simulation study

Throughout this section, we consider testing the null hypothesis $H_0 : \mu^\top\nu = 0_q$ versus $H_1 : \mu^\top\nu \neq 0_q$, where, unless otherwise stated, ν defined in (4.3) is based on a randomly-chosen pair of clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ from k-means clustering. We consider four p-values: p_{Naive} in (4.4),

$p_{\text{selective}}$ in (4.9), $\hat{p}_{\text{selective}}$ in (4.20) with $\hat{\sigma}_{\text{MED}}$ defined in (4.21), and $\hat{p}_{\text{selective}}$ in (4.20) with $\hat{\sigma}_{\text{Sample}} = \left\{\sum_{i=1}^{n}\sum_{j=1}^{q}(x_{ij}-\bar{x}_j)^2/(nq-q)\right\}^{1/2}$, where $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$. In the simulations that follow, we compare the selective Type I error (4.5) and power of the tests that reject H_0 when these p-values are less than $\alpha = 0.05$.

4.5.1 Selective Type I error under the global null

We generate data from (4.1) with $\mu = 0_{n\times q}$; therefore, H_0 in (4.2) holds for any pair of estimated clusters. We simulate 3,000 datasets with $n = 150, \sigma = 1$, and $q = 2, 10, 50, 100$.

For each simulated dataset, we apply k-means clustering with $K = 3$, and then compute p_{Naive}, $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ for a randomly-chosen pair of clusters. Figure 4.4 displays the observed p-value quantiles versus the Uniform(0,1) quantiles. We see that for all values of q, (i) the naive p-values in (4.4) are stochastically smaller than a Uniform(0,1) random variable, and the test based on p_{Naive} leads to an inflated Type I error rate; (ii) tests based $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ control the selective Type I error rate in the sense of (4.5).

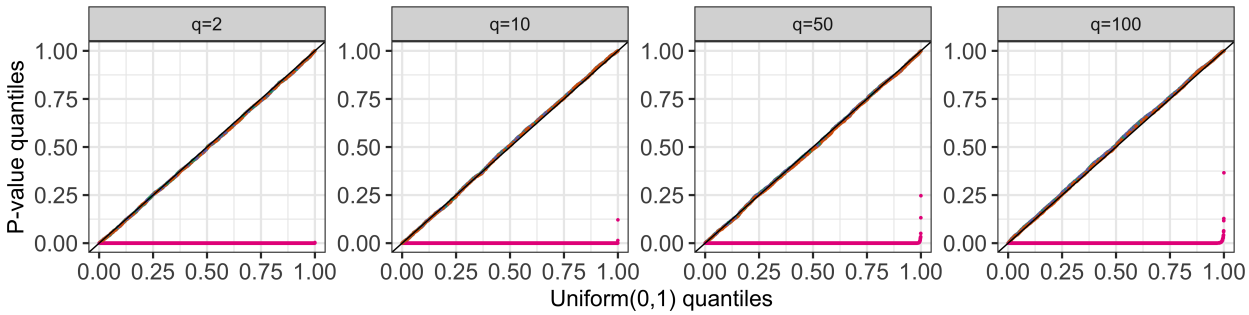


Figure 4.4: Quantile-quantile plots for p_{Naive} (pink), $p_{\text{selective}}$ (green), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple) under (4.1) with $\mu = 0_{n\times q}$, stratified by q.

4.5.2 Conditional power and detection probability

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

from (4.1) with $n = 150$ and

$$\mu_1 = \ldots = \mu_{\frac{n}{3}} = \begin{bmatrix} -\frac{\delta}{2} \\ 0_{q-1} \end{bmatrix}, \; \mu_{\frac{n}{3}+1} = \ldots = \mu_{\frac{2n}{3}} = \begin{bmatrix} 0_{q-1} \\ \frac{\sqrt{3}\delta}{2} \end{bmatrix}, \; \mu_{\frac{2n}{3}+1} = \ldots = \mu_n = \begin{bmatrix} \frac{\delta}{2} \\ 0_{q-1} \end{bmatrix}. \quad (4.22)$$

Here, we can think of $\mathcal{C}_1 = \{1, \ldots, n/3\}, \mathcal{C}_2 = \{(n/3)+1, \ldots, (2n/3)\}, \mathcal{C}_3 = \{(2n/3)+1, \ldots, n\}$ as the "true clusters". Moreover, these clusters are equidistant in the sense that the pairwise distance between each pair of population means is $|\delta|$. Recall that we test H_0 in (4.2) for a pair of estimated clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$, which may not be true clusters. Hence, we will separately consider the *conditional power* and *detection probability* of our proposed tests [Gao et al., 2020, Hyun et al., 2021, Jewell et al., 2022]. The conditional power is the probability of rejecting H_0 in (4.2), given that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are true clusters. Given M simulated datasets with true clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$, we estimate it as

$$\text{Conditional power} = \frac{\sum_{m=1}^{M} 1\Big\{\Big\{\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}\Big\} \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}, p^{(m)} \leq \alpha\Big\}}{\sum_{m=1}^{M} 1\Big\{\Big\{\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}\Big\} \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}\Big\}}, \quad (4.23)$$

where $p^{(m)}$ and $\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}$ correspond to the p-value and clusters under consideration for the mth simulated dataset. Because the quantity in (4.23) conditions on the event that $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are true clusters, we also estimate how often that event occurs:

$$\text{Detection probability} = \sum_{m=1}^{M} 1\Big\{\{\hat{\mathcal{C}}_1^{(m)}, \hat{\mathcal{C}}_2^{(m)}\} \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}\Big\}/M. \quad (4.24)$$

We generate $M = 200,000$ datasets from (4.22) with $q = 10, \sigma = 0.25, 0.5, 1$, and $\delta = 2, 3, \ldots, 10$. For each simulated dataset, we apply k-means clustering with $K = 3$ and reject $H_0 : \mu^\top \nu = 0_q$ if $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, or $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ is less than $\alpha = 0.05$. In Figure 4.5, the left panel displays the detection probability (4.24) of k-means clustering as a function of δ in (4.22), and the right panel displays the conditional power (4.23) for the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$. Under model (4.1), the

detection probability and conditional power increase as a function of δ in (4.22) for all values of σ. For a given value of δ, a larger value of σ leads to lower detection probability and conditional power. The conditional power is not displayed for $\delta = 2, 3, \sigma = 1$ because the true clusters were never recovered in simulation. Moreover, for a given value of δ and σ, the test based on $p_{\text{selective}}$ has the highest conditional power, followed closely by the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. Using $\hat{\sigma}_{\text{Sample}}$ in $\hat{p}_{\text{selective}}$ leads to a less powerful test, especially for large values of δ. This is because $\hat{\sigma}_{\text{Sample}}$ is a conservative estimator of σ in (4.1), and its bias is an increasing function of δ, the distance between true clusters. By contrast, $\hat{\sigma}_{\text{MED}}$ is a consistent estimator under model (4.22) (see Appendix C.7).

As an alternative to the conditional power in (4.23), in Appendix C.8, we consider a notion of power that does not condition on having correctly estimated the true clusters.

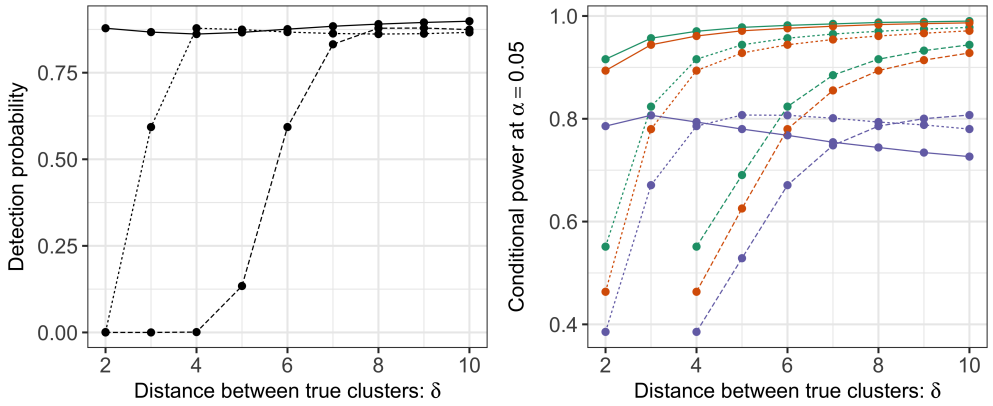
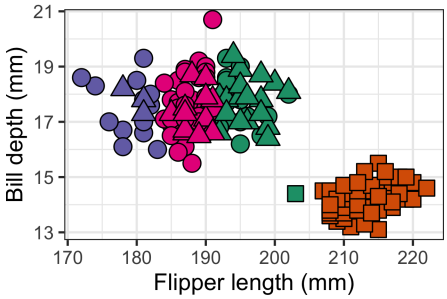


Figure 4.5: *Left:* The detection probability (4.24) for k-means clustering with $K = 3$ under model (4.1) with μ defined in (4.22), and $\sigma = 0.25$ (solid lines), 0.5 (dashed lines), and 1 (long-dashed lines). *Right:* The conditional power (4.23) at $\alpha = 0.05$ for the tests based on $p_{\text{selective}}$ (green), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple), under model (4.1) with μ defined in (4.22) and $\sigma = 0.25, 0.5, 1$. The conditional power is not displayed for $\delta = 2, 3, \sigma = 1$ because the true clusters were never recovered in simulation.

4.6 Real data applications

4.6.1 Palmer Penguins [Horst et al., 2020]

Here we analyze the Palmer penguins dataset from the `palmerpenguins` package in `R` [Horst et al., 2020]. We consider the 165 female penguins with complete observations, and apply k-means clustering with $K = 4$ to two of the collected features: bill depth and flipper length. Figure 4.6 displays the estimated clusters. We assess the equality of the means of



H_0	p_{Naive}	$\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$
$\bar{\mu}_1 = \bar{\mu}_2$	$< 10^{-10}$	3.3×10^{-5}
$\bar{\mu}_1 = \bar{\mu}_3$	$< 10^{-10}$	0.007
$\bar{\mu}_1 = \bar{\mu}_4$	$< 10^{-10}$	0.24
$\bar{\mu}_2 = \bar{\mu}_3$	$< 10^{-10}$	2.2×10^{-6}
$\bar{\mu}_2 = \bar{\mu}_4$	$< 10^{-10}$	7.8×10^{-6}
$\bar{\mu}_3 = \bar{\mu}_4$	$< 10^{-10}$	0.10

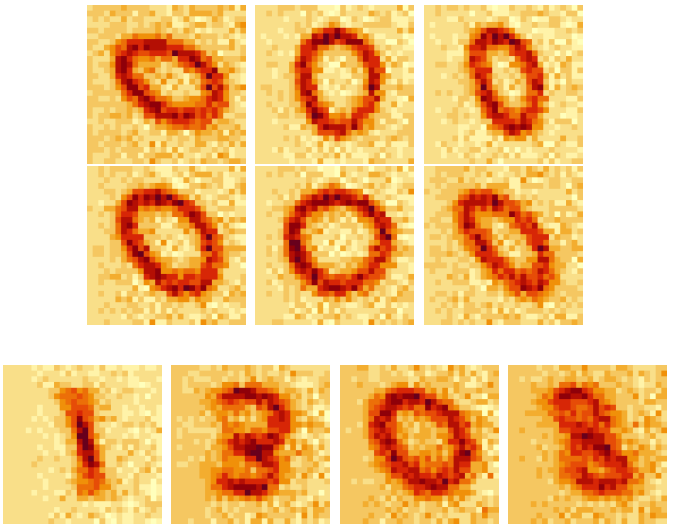
Figure 4.6: *Left:* The bill depths and flipper lengths of female Palmer penguins, along with true species labels (Adelie: circle; Gentoo: square; Chinstrap: triangle) and clusters estimated using k-means clustering (cluster 1: green; cluster 2: orange; cluster 3: purple; cluster 4: pink). *Right:* We test the null hypothesis that the means of two estimated clusters are equal, for each pair of clusters estimated via k-means clustering, using p_{Naive} in (4.4) and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ in (4.20) with $\hat{\sigma}_{\text{MED}}$ defined in (4.21). Here, $\bar{\mu}_i = \sum_{j \in \hat{\mathcal{C}}_i} \mu_j / |\hat{\mathcal{C}}_i|$.

each pair of estimated clusters using p_{Naive} in (4.4) and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ in (4.20) with $\hat{\sigma}_{\text{MED}}$ defined in (4.21). The results are in Figure 4.6. The naive p-values are small for all pairs of estimated clusters, even when the underlying species distributions are nearly identical (e.g., both clusters 1 and 4 are a mix of Chinstrap and Adelie penguins). By contrast, our proposal results in large p-values when testing for a difference in means between clusters composed of the same species (clusters 1 and 4, clusters 3 and 4), and small p-values when the clusters correspond to different species (e.g., clusters 1 and 2, clusters 2 and 3).

4.6.2 MNIST Dataset [Lecun et al., 1998]

In this section, we apply our method to the MNIST dataset [Lecun et al., 1998], which consists of 60,000 gray-scale images of handwritten digits. Each image has an accompanying

label in $\{0, 1, \ldots, 9\}$, and is stored as a 28×28 matrix that takes on values in $[0, 255]$. We first divide the entries of all the images by 255. Next, since there is no variation in the peripheral pixels of the images [Gallaugher and McNicholas, 2018], which violates model (4.1), we add an independent perturbation $\mathcal{N}(0, 0.01)$ to each element of the image. Finally, we vectorize each image to obtain a vector $x_i \in \mathbb{R}^{784}$. We first construct a "no cluster" dataset by randomly



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

Figure 4.7: *Top left:* Centroids of six clusters from the "no cluster" dataset ($\hat{\mathcal{C}}_1$ to $\hat{\mathcal{C}}_6$ from left to right, top to bottom). *Bottom left:* Same as top left, but for the "cluster" dataset. *Right:* We test the null hypothesis of no difference between each pair of cluster centroids using p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. Here, $\bar{\mu}_i = \sum_{j \in \hat{\mathcal{C}}_i} \mu_j / |\hat{\mathcal{C}}_i|$.

sampling 1,500 images of the 0s; thus, $n = 1,500$ and $q = 784$. To de-correlate the pixels in each image, we whitened the data (see Section 4.4.1) using $\hat{\Sigma}^{-\frac{1}{2}} = U(\Lambda + 0.01\mathbf{I}_n)^{-\frac{1}{2}}U^\top$ as in prior work [Coates and Ng, 2012], where $U\Lambda U^\top$ is the eigenvalue decomposition of the sample covariance matrix.

We apply k-means clustering with $K = 6$. The centroids are displayed in the top left panel of Figure 4.7. For each pair of estimated clusters, we compute the p-values p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (see Figure 4.7). The naive p-values are extremely small for all pairs of clusters under consideration, despite the resemblance of the centroids. By contrast, our

approach yields modest p-values, congruent with the visual resemblance of the centroids. In addition, for the most part, the pairs for which $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ is small are visually quite different (e.g., clusters 1 and 2, clusters 1 and 5, and clusters 4 and 5). In other words, even though all estimated clusters are composed of digit 0s, the pairs for which $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ is small usually correspond to 0s with different stroke patterns.

To demonstrate the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, we also generated a "cluster" dataset by sampling 500 images each from digits $0, 1, 3$, and 8; thus, $n = 2,000$ and $q = 784$. We again whitened the data to obtain uncorrelated features. After applying k-means clustering with $K = 4$, we obtain four clusters that roughly correspond to four digits: cluster 1, 94.0% digit 1; cluster 2, 72.4% digit 3; cluster 3, 83.6% digit 0; cluster 4, 62.4% digit 8 (see the bottom left panel of Figure 4.7). Results from testing for a difference in means for each pair of clusters using p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ are in Figure 4.7. Both sets of p-values are small on this "cluster" dataset.

4.6.3 Single-cell RNA-sequencing data [Zheng et al., 2017]

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

Zheng et al. [2017] profiled 68,000 peripheral blood mononuclear cells, and classified them based on their match to the expression profiles of 11 reference transcriptomes from known cell types. We consider the classified cell types to be the "ground truth", and use this information to demonstrate that our proposal in Section 4.2 yields reasonable results.

As in prior work [Duò et al., 2018, Gao et al., 2020], we first excluded cells with low numbers of expressed genes or total counts, as well as cells in which a large percentage of the expressed genes are mitochondrial. We then divided the counts for each cell by the total sum of counts in that cell. Finally, we applied a $\log_2$ transformation with a pseudo-count of 1 to the expression data, and considered only the subset of 500 genes with the largest average expression levels pre-normalization. We applied the aforementioned pre-processing pipeline separately to memory T cells ($N = 10,224$) and a mixture of five types of cells (memory T cells, B cells, naive T cells, natural killer cells, and monocytes; $N = 43,259$).

To investigate the selective Type I error in the absence of true clusters, we first constructed a "no cluster" dataset by randomly sampling 1,000 out of 10,224 memory T cells after pre-processing (thus, $n = 1,000$ and $q = 500$). Since the gene expression levels are highly correlated, we first whitened the data as described in Section 4.4.1 by plugging in $\hat{\Sigma}^{-\frac{1}{2}} = U(\Lambda + 0.01\mathbf{I}_n)^{-\frac{1}{2}}U^\top$ [Coates and Ng, 2012], where $U\Lambda U^\top$ is the eigenvalue decomposition of the sample covariance matrix.

We applied k-means clustering to the transformed data with $K = 5$, and obtained five clusters consisting of 97, 223, 172, 165, and 343 cells, respectively (see Figure C.3 left panel in Appendix C.9). For each pair of estimated clusters, we computed the p-values p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. The results are displayed in the top panel of Table 4.1. On this dataset, the naive p-values are extremely small for all pairs of estimated clusters, while our proposed p-values are quite large. In particular, at $\alpha = 0.05$, the test based on p_{Naive} concludes that all five estimated clusters correspond to distinct cell types (even after multiplicity correction), whereas our approach does not reject the null hypothesis that the expression levels (and thus cell types) are, in fact, the same between estimated clusters. Because this "no cluster" dataset consists only of memory T cells, we believe that conclusion based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ aligns better with the underlying biology.

Next, we construct a "cluster" dataset by randomly sampling 400 each of memory T cells, B cells, naive T cells, natural killer cells, and monocytes from the $43,259$ cells; thus, $n = 2,000$ and $q = 500$. After whitening the data, we applied k-means clustering to obtain

Table 4.1: P-values p_{Naive} in (4.4) and $\hat{p}_{\text{selective}}$ in (4.20) with $\hat{\sigma}_{\text{MED}}$ defined in (4.21) corresponding to the null hypothesis that the means of two estimated clusters are equal, for each pair of estimated clusters in the "no cluster" (top) and the "cluster" datasets (bottom).

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

five clusters. We see that these clusters approximately correspond to the five different cell types (cluster 1: 82.5% naive T cells; cluster 2: 95.3% memory T cells; cluster 3: 99.2% B cells; cluster 4: 91.5% nature killer cells; cluster 5: 83.3% monocytes); estimated clusters are visualized in the right panel of Figure C.3 in Appendix C.9. We evaluate the p-values p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for all pairs of estimated clusters, and display results in the bottom panel of Table 4.1. Both sets of p-values are extremely small on this dataset, which suggests that the test based on our p-value has substantial power to reject the null hypothesis when it does not hold.

4.7 Discussion

We have proposed a test for a difference in means between two clusters estimated from k-means clustering, under (4.1). Methods developed in this paper are implemented in the `R` package `KmeansInference`, available at https://github.com/yiqunchen/KmeansInference. Data and code for reproducing the results in this paper can be found at https://github.com/yiqunchen/KmeansInference-experiments. Next, we outline a few directions for future research.

While the p-value in (4.9) leads to selective Type I error control, it conditions on more information than is used to construct the hypothesis in (4.2). In practice, data analysts likely only make use of the final cluster assignments (leading to the p-value in (4.8)), as opposed to all the intermediate assignments (leading to the p-value in (4.9)). Empirically,

conditioning on too much information results in a loss of power [Fithian et al., 2014, Jewell et al., 2022, Liu et al., 2018]. In future work, we will investigate the possibility of leveraging recent developments in selective inference [Chen et al., 2021a, Jewell et al., 2022, Le Duy and Takeuchi, 2021] to compute the "ideal" p-value (4.8).

We could also consider extending our proposal to other data generating models. The normality assumption in (4.1) is critical to the proof of Proposition 13, because it guarantees that under H_0 in (4.2), $\|X^\top \nu\|_2$, $\text{dir}(X^\top \nu)$, and $\Pi_\nu^\perp X$ are pairwise independent. However, this normality assumption is often violated in practice, which might result in inflated Type I error rate in real data analysis (e.g., a few pairs of visually-similar centroids in the top left panel of Figure 4.7 correspond to small p-values). For instance, in single-cell genomics, the data are count-valued and the variance of gene expression levels varies drastically with the mean expression levels of that gene [Eling et al., 2018, Stuart et al., 2019]. This has motivated some authors to work with alternative models for gene expression including Poisson [Witten, 2011], negative binomial [Risso et al., 2018], and curved normal [Lin et al., 2021]. To extend our framework to other exponential family distributions, we may be able to leverage recent proposals to decompose X into $f(X)$ and $g(X)$ such that both $f(X)$ and $g(X)|f(X)$ have a known, computationally-tractable distribution [Leiner et al., 2021, Rasines and Alastair Young, 2021].

Chapter 5

DISCUSSION

5.1 Summary

In this dissertation, we proposed new tests for a difference in means between two groups after selection; that is, when the groups under investigation are *defined through the observed data*. In Chapter 2, we tackle the problem of quantifying the uncertainty of spikes estimated from calcium imaging. Building on a well-studied model for calcium imaging data [Friedrich and Paninski, 2016, Friedrich et al., 2017, Vogelstein et al., 2010], we formulate the problem of interest as testing for a change in weighted means of the calcium concentration around an estimated spike. Our proposal accounts for the fact that the spikes are estimated from the data, thereby providing selective Type I error control and correct selective coverage.

In Chapter 3, we consider testing for a difference in means between two groups defined via the output of the graph fused lasso [Tibshirani and Taylor, 2011], a popular method to reconstruct *piecewise constant* signals. The solution to the graph fused lasso can be segmented into *connected components*, which are groups of elements that share a common value. It's natural to wish to test whether the means are equal across two *estimated* connected components. We propose a new test for this task that conditions on less information than existing approaches [Hyun et al., 2018], which leads to substantially higher power while guaranteeing selective Type I error control.

We pivot to the multivariate setting in Chapter 4, where we consider testing for a difference in cluster means after applying k-means clustering to the data. Our work is motivated by the field of genomics, where datasets in single-cell transcriptomics consist of millions of unlabeled cells, to which researchers routinely apply clustering algorithms. Under a simple and well-studied matrix-variate normal model, we propose a finite-sample, computationally-

efficient test that controls the selective Type I error for this task. Our proposal leads to conclusions that align better with the underlying truth than classical tests on both handwritten digits data and single-cell RNA-sequencing data.

Methods developed in Chapters 2–4 are implemented in open source software available online (links can be found in Sections 2.7, 3.7, and 4.7); we have extensively documented our software, with vignettes demonstrating usage of our methods on real data applications.

5.2 Future work

The tests proposed in Chapters 2–4 were developed under the assumption that the observations are normally distributed. The normality assumption is critical, as it guarantees the efficient computation of our proposed p-values. For instance, in Chapter 4, the normality assumption in (4.1) is critical to the proof of Proposition 13, because it guarantees that under H_0 in (4.2), $\|X^\top \nu\|_2$, $\text{dir}(X^\top \nu)$, and $\Pi_\nu^\perp X$ are pairwise independent. Similarly, in Chapter 3, we rely on the property of a multivariate normal distribution to prove Proposition 9, and consequently, to compute $p_{\hat{C}_1,\hat{C}_2}$.

However, this normality assumption might be violated in practice. For instance, in single-cell genomics, the data are count-valued, which has motivated several authors to work with alternative models for gene expression such as Poisson [Witten, 2011] and negative binomial [Risso et al., 2018]. A line of recent work in selective inference has focused on relaxing these assumptions in high-dimensional linear modeling [Charkhi and Claeskens, 2018, Tian and Taylor, 2018, Tibshirani et al., 2018a] by carefully analyzing the asymptotics of the selection events, and may be applicable to the settings in Chapters 2–4. Alternatively, we may be able to leverage the recent developments in generalized data splitting and carving to extend our selective inference approach to other exponential family distributions [Leiner et al., 2021, Rasines and Alastair Young, 2021, Schultheiss et al., 2021]. As an example, in the context of testing for equality in cluster means in Chapter 4, this amounts to decomposing the data X into $f(X)$ and $g(X)$ for some functions f and g such that both $f(X)$ and $g(X)|f(X)$ have a known, computationally-tractable distribution [Leiner et al., 2021,

Rasines and Alastair Young, 2021].

In the future, it may be worthwhile to investigate the connections between ideas in this dissertation and related developments in machine learning. Recent work has recast parameter inference after selection as a binary classification problem [Liu et al., 2022, Markovic et al., 2019]. Building on these observations, we could leverage toolkits from adaptive data analysis such as disagreement-based active classification [Balcan et al., 2010, Settles, 2011] to provide computationally-efficient selective inference tools for any black-box selection procedures. As another example, *multiverse analysis*, an approach that outlines and executes all "reasonable" data analysis plans, and then interprets the full range of possible outcomes collectively, has emerged as a popular tool for data analysis [Dragicevic et al., 2019, Simonsohn et al., 2019, Steegen et al., 2016]. Examining the role of rigorous statistical procedures, such as those developed in this dissertation, in empirical practice such as multiverse analysis will be an important direction of future work.

BIBLIOGRAPHY

Ahrens, M. B., Orger, M. B., Robson, D. N., Li, J. M., and Keller, P. J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods*, 10(5):413–420.

Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Herman, J. S., Pessaux, P., Baumert, T. F., and Grün, D. (2019). A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, 572(7768):199–204.

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248.

Amin, R., Decesare, J. Z., Hans, J., and Roussos-Ross, K. (2017). Epidemiologic surveillance of teenage birth rates in the United States, 2006-2012. *Obstetrics and Gynecology*, 129(6):1068–1077.

Anscombe, F. J. (1948). The transformation of poisson, binomial and Negative-Binomial data. *Biometrika*, 35(3/4):246–254.

Arnold, T. B. and Tibshirani, R. J. (2016). Efficient implementations of the generalized lasso dual path algorithm. *J. Comput. Graph. Stat.*, 25(1):1–27.

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

Avella-Medina, M., Battey, H. S., Fan, J., and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284.

Aw, A. J., Spence, J. P., and Song, Y. S. (2021). A flexible and robust non-parametric test of exchangeability. *arXiv:2109.15261*.

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

Balcan, M.-F., Hanneke, S., and Vaughan, J. W. (2010). The true sample complexity of active learning. *Machine Learning*, 80(2):111–139.

Bar-Lev, S. K. and Enis, P. (1988). On the classical choice of variance stabilizing transformations and an application for a poisson variate. *Biometrika*, 75(4):803–804.

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

Belomestny, D., Trabs, M., and Tsybakov, A. B. (2019). Sparse covariance matrix estimation in high-dimensional deconvolution. *Bernoulli*, 25(3):1901–1938.

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

Berens, P., Freeman, J., Deneux, T., Chenkov, N., McColgan, T., Speiser, A., Macke, J. H., Turaga, S. C., Mineault, P., Rupprecht, P., Gerhard, S., Friedrich, R. W., Friedrich, J., Paninski, L., Pachitariu, M., Harris, K. D., Bolte, B., Machado, T. A., Ringach, D., Stone, J., Rogerson, L. E., Sofroniew, N. J., Reimer, J., Froudarakis, E., Euler, T., Román Rosón, M., Theis, L., Tolias, A. S., and Bethge, M. (2018). Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.*, 14(5):e1006157.

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

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*. Springer, New York, NY.

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

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

Cai, D., Campbell, T., and Broderick, T. (2020). Finite mixture models do not reliably learn the number of components. *arXiv:2007.04470*.

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

Centers for Disease Control and Prevention (2020a). 2018 Drug Overdose Death Rates. https://www.cdc.gov/drugoverdose/data/statedeaths/drug-overdose-death-2018.html.

Centers for Disease Control and Prevention (2020b). 2018 Teenage Birth Rates. https://www.cdc.gov/nchs/pressroom/sosmap/teen-births/teenbirths.htm.

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

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

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

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics*, 46(5).

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

Chen, T.-W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., and Kim, D. S. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300.

Chen, Y., Jewell, S., and Witten, D. (2021a). More powerful selective inference for the graph fused lasso. *arXiv:2109.10451*.

Chen, Y. T., Jewell, S. W., and Witten, D. M. (2021b). Quantifying uncertainty in spikes estimated from calcium imaging data. *To appear in Biostatistics*.

Chen, Y. T. and Witten, D. M. (2022). Selective inference for k-means clustering. *arXiv:2203.15267*.

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

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

Coates, A. and Ng, A. Y. (2012). Learning feature representations with K-Means. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 561–580. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Deneux, T., Kaszas, A., Szalay, G., Katona, G., Lakner, T., Grinvald, A., Rózsa, B., and Vanzetta, I. (2016). Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nat. Commun.*, 7:12190.

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

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

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., and Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, number Paper 65 in CHI '19, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Duò, A., Robinson, M. D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141.

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

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

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

Fleming, W., Jewell, S., Engelhard, B., Witten, D., and Witten, I. (2021). Inferring spikes from calcium imaging in dopamine neurons. *PloS one*, 16(6):e0252345.

Friederich, P., Krenn, M., Tamblyn, I., and Aspuru-Guzik, A. (2020). Scientific intuition inspired by machine learning generated hypotheses. *arXiv:2010.14236*.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

Friedrich, J. and Paninski, L. (2016). Fast active set methods for online spike inference from calcium imaging. In *Advances In Neural Information Processing Systems*, pages 1984–1992.

Friedrich, J., Zhou, P., and Paninski, L. (2017). Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.*, 13(3):e1005423.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

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

Gao, L. L., Bien, J., and Witten, D. (2020). Selective inference for hierarchical clustering. *arXiv:2012.02936*.

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Stat. Sci.*, 26(2):187–202.

Golub, T. (2010). Counterpoint: Data first. *Nature*, 464(7289):679.

Gong, Y., Lazaris, C., Sakellaropoulos, T., Lozano, A., Kambadur, P., Ntziachristos, P., Aifantis, I., and Tsirigos, A. (2018). Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature Communications*, 9(1):542.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.

Guha, A., Ho, N., and Nguyen, X. (2019). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *arXiv:1901.05078*.

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

Hand, D. J. and Adams, N. M. (2015). *Data Mining*, pages 1–7. John Wiley & Sons, Ltd, Chichester, UK.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple Change-Point estimation with a total variation penalty. *J. Am. Stat. Assoc.*, 105(492):1480–1493.

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

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

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, 2nd edition edition.

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

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

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

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

Hyun, S., Lin, K. Z., G'Sell, M., and Tibshirani, R. J. (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*.

Jewell, S., Fearnhead, P., and Witten, D. (2022). Testing for a change in mean after change-point detection. *To appear in J. R. Stat. Soc. Series B Stat. Methodol.*

Jewell, S. and Witten, D. (2018). Exact spike train inference via ℓ_0 optimization. *Ann. Appl. Stat.*, 12(4):2457–2482.

Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M. (2019). Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*.

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

Johnson, N. A. (2013). A dynamic programming algorithm for the fused lasso and L0-segmentation. *J. Comput. Graph. Stat.*, 22(2):246–260.

Kang, J., Reich, B. J., and Staicu, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105(1):165–184.

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

Kim, S. and Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLOS Genetics*, 5(8):e1000587.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Rev.*, 51(2):339–360.

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

Kivaranovic, D. and Leeb, H. (2020). On the length of Post-Model-Selection confidence intervals conditional on polyhedral constraints. *J. Am. Stat. Assoc.*, pages 1–13.

Kovács, S., Li, H., and Bühlmann, P. (2020). Seeded intervals and noise level estimation in change point detection: a discussion of fryzlewicz (2020). *Journal of the Korean Statistical Society*, 49(4):1081–1089.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540.

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., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P. F., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31.

Le Duy, V. N. and Takeuchi, I. (2021). More powerful conditional selective inference for generalized lasso by parametric programming. *arXiv:2105.04920*.

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

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

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2021). Data blurring: sample splitting a single sample. *arXiv:2112.11079*.

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

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

Liu, K., Markovic, J., and Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv:1801.09037*.

Liu, S., Markovic, J., and Taylor, J. (2022). Black-box selective inference via bootstrapping. *arXiv:2203.14504*.

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

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

Loftus, J. R. and Taylor, J. E. (2014). A significance test for forward stepwise model selection. *arXiv:1405.3920*.

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv:1612.02099*.

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

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Stat. Comput.*, 27(2):519–533.

Markovic, J., Taylor, J., and Taylor, J. (2019). Inference after black box selection. *arXiv:1901.09973*.

Markovic, J., Xia, L., and Taylor, J. (2017). Unifying approach to selective inference with applications to cross-validation. *arXiv:1703.06559*.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378.

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

Merel, J., Shababo, B., Naka, A., Adesnik, H., and Paninski, L. (2016). Bayesian methods for event analysis of intracellular currents. *J. Neurosci. Methods*, 269:21–32.

Niu, Y. S., Hao, N., and Zhang, H. (2016). Multiple Change-Point detection: A selective overview. *Stat. Sci.*, 31(4):611–623.

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

Pachitariu, M., Stringer, C., and Harris, K. D. (2018). Robustness of spike deconvolution for neuronal calcium imaging. *J. Neurosci.*, 38(37):7976–7985.

Peters, D. P. C., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014). Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere*, 5(6):1–15.

Pillow, J. W., Ahmadian, Y., and Paninski, L. (2011). Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains. *Neural Comput.*, 23(1):1–45.

Pnevmatikakis, E. A., Merel, J., Pakman, A., and Paninski, L. (2013). Bayesian spike inference from calcium imaging data. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 349–353. IEEE.

Pnevmatikakis, E. A., Soudry, D., Gao, Y., Machado, T. A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., Ahrens, M., Bruno, R., Jessell, T. M., Peterka, D. S., Yuste, R., and Paninski, L. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299.

Pollice, R., Dos Passos Gomes, G., Aldeghi, M., Hickman, R. J., Krenn, M., Lavigne, C., Lindner-D'Addario, M., Nigam, A., Ser, C. T., Yao, Z., and Aspuru-Guzik, A. (2021). Data-Driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860.

Prevedel, R., Yoon, Y.-G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M., Boyden, E. S., and Vaziri, A. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods*, 11(7):727–730.

Ramdas, A. and Tibshirani, R. J. (2016). Fast and flexible ADMM algorithms for trend filtering. *J. Comput. Graph. Stat.*, 25(3):839–858.

Rasines, D. G. and Alastair Young, G. (2021). Splitting strategies for post-selection inference. *arXiv:2102.02159*.

Reid, S., Taylor, J., and Tibshirani, R. (2017). Post-selection point and interval estimation of signal sizes in Gaussian samples. *Can. J. Stat.*, 45(2):128–148.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to *kmax* change-points. *Journal de la Société Française de Statistique*, 156(4):180–205.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Ann. Stat.*, 37(5B):2922–2952.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17.

Rousseeuw, P. J. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D.: Nonlinear Phenomena*, 60(1):259–268.

Rügamer, D., Baumann, P. F. M., and Greven, S. (2022). Selective inference for additive and linear mixed models. *Computational Statistics & Data Analysis*, 167:107350.

Rügamer, D. and Greven, S. (2020). Inference for L_2-boosting. *Statistics and computing*, 30(2):279–289.

Sadhanala, V., Wang, Y.-X., and Tibshirani, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. 29:3513–3521.

Schieber, L. Z., Guy, G. P., Seth, P., Young, R., Mattson, C. L., Mikosz, C. A., and Schieber, R. A. (2019). Trends and patterns of geographic variation in opioid prescribing practices by state, United States, 2006-2017. *JAMA Network Open*, 2(3).

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

Settles, B. (2011). From theories to queries: Active learning in practice. volume 16 of *Proceedings of Machine Learning Research*, pages 1–18. PMLR.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2019). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *Available at SSRN 2694998*.

Soltanian-Zadeh, S., Gong, Y., and Farsiu, S. (2018). Information-theoretic approach and fundamental limits of resolving two closely timed neuronal spikes in mouse brain calcium imaging. *IEEE Trans. Biomed. Eng.*, 65(11):2428–2439.

Song, R., Banerjee, M., and Kosorok, M. R. (2016). Asymptotics for change-point models under varying degrees of mis-specification. *Ann. Stat.*, 44(1):153–182.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 11(5):702–712.

Stringer, C. and Pachitariu, M. (2019). Computational processing of neural recordings from calcium imaging data. *Curr. Opin. Neurobiol.*, 55:22–31.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, 3rd, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21.

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

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

Theis, L., Berens, P., Froudarakis, E., Reimer, J., Román Rosón, M., Baden, T., Euler, T., Tolias, A. S., and Bethge, M. (2016). Benchmarking spike rate inference in population calcium imaging. *Neuron*, 90(3):471–482.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(1):91–108.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.*, 42(1):285–323.

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

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

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Stat.*, 39(3):1335–1371.

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

Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1):1201.

Ventura, S. J., Hamilton, B. E., and Matthews, T. J. (2014). National and state patterns of teen births in the United States, 1940-2013. *National Vital Statistics Reports*, 63(4):1–34.

Ventura, V. (2008). Spike train decoding without spike sorting. *Neural Computation*, 20(4):923–963.

Victor, J. D. and Purpura, K. P. (1996). Nature and precision of temporal coding in visual cortex: a metric-space analysis. *Journal of Neurophysiology*, 76(2):1310–1326.

Victor, J. D. and Purpura, K. P. (1997). Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164.

Vogelstein, J., Watson, B., Packer, A., Yuste, R., Jedynak, B., and Paninski, L. (2009). Spike inference from calcium imaging using sequential monte carlo methods. *Biophysical journal*, 97(2):636–655.

Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology*, 104(6):3691–3704.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality. *Advances in Neural Information Processing Systems*, 28:2512–2520.

Warton, D. I. (2018). Why you cannot transform your way out of trouble for small counts. *Biometrics*, 74(1):362–368.

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

Wei, X.-X., Zhou, D., Grosmark, A., Ajabi, Z., Sparks, F., Zhou, P., Brandon, M., Losonczy, A., and Paninski, L. (2019). A zero-inflated gamma model for post-deconvolved calcium imaging traces.

Weinstein, A., Fithian, W., and Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108(501):165–176.

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

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

Xin, B., Kawahara, Y., Wang, Y., and Gao, W. (2014). Efficient generalized fused lasso and its application to the diagnosis of alzheimer's disease. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Xu, R. and Wunsch, D. (2008). *Clustering*. John Wiley & Sons.

Yang, F., Barber, R. F., Jain, P., and Lafferty, J. (2016). Selective inference for group-sparse linear models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2477–2485.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Stat. Probab. Lett.*, 6(3):181–189.

Yao, Y.-C. and Au, S. T. (1989). Least-Squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 51(3):370–381.

Yekutieli, D. (2012). Adjusted Bayesian inference for selected parameters. *J. R. Stat. Soc. Series B Stat. Methodol.*, 74(3):515–541.

Yi, X. and Caramanis, C. (2015). Regularized EM algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*.

Zha, H., He, X., Ding, C., Gu, M., and Simon, H. (2002). Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems*, volume 14.

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

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

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zhao, S., Witten, D., and Shojaie, A. (2021). In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science*.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049.

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9(4):621–634.

Zhu, Y. (2017). An augmented ADMM algorithm with application to the generalized lasso problem. *J. Comput. Graph. Stat.*, 26(1):195–204.

Zollner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American Journal of Human Genetics*, 80(4):605–615.

Zou, C., Wang, G., and Li, R. (2020). Consistent selection of the number of change-points via sample-splitting. *Ann. Stat.*, 48(1):413–439.

Appendix A

A.1 Proof of Proposition 1

We first prove the statement (2.10). The following equalities hold:

$$\begin{aligned}
&\mathbb{P}\left(\nu^\top Y \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y, \nu^\top Y > 0\right) \\
&\overset{a.}{=} \mathbb{P}\left(\nu^\top Y \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}(\Pi_\nu^\perp y + \Pi_\nu Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y, \nu^\top Y > 0\right) \\
&\overset{b.}{=} \mathbb{P}\left(\phi \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}\left(y'(\phi)\right), \Pi_\nu^\perp Y = \Pi_\nu^\perp y, \nu^\top Y > 0\right) \\
&\overset{c.}{=} \mathbb{P}\left(\phi \geq \nu^\top y \;\middle|\; \hat{\tau}_j(y) \in \mathcal{M}\left(y'(\phi)\right), \phi > 0\right).
\end{aligned}$$

Here, $a.$ follows from the fact that $Y = \Pi_\nu^\perp Y + \Pi_\nu Y$, and the fact that we have conditioned on the event $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$. To prove $b.$, we first note that $I = \Pi_\nu + \Pi_\nu^\perp$ and $\Pi_\nu Y = \frac{\nu\nu^\top}{||\nu||_2^2} Y$, which implies

$$\Pi_\nu^\perp y + \Pi_\nu Y = y - \Pi_\nu y + \Pi_\nu Y = y - \frac{\nu^\top y}{||\nu||_2^2}\nu + \frac{\nu^\top Y}{||\nu||_2^2}\nu = y'(\phi),$$

where we define $\phi = \nu^\top Y \sim \mathcal{N}(\nu^\top c, \sigma^2||\nu||_2^2)$. Finally, $c.$ follows from the fact that $Y \sim \mathcal{N}(c, \sigma^2 I)$ implies independence of $\phi = \nu^\top Y$ and $\Pi_\nu^\perp Y$.

Now to prove (2.12), we note that under H_0 in (2.6), $\nu^\top Y \sim \mathcal{N}(0, \sigma^2||\nu||_2^2)$. Therefore, applying the result above with $\nu^\top c = 0$ completes the proof.

A.2 General case for the contrast vector ν

The definition (2.7) only applies when $\hat{\tau}_j - h + 1 \geq 1$ and $\hat{\tau}_j + h \leq T$. In the case that $\hat{\tau}_j - h + 1 < 1$ or $\hat{\tau}_j + h > T$, we define the contrast vector ν as follows:

$$\nu_t = \begin{cases} -\gamma \frac{\gamma^2 - 1}{\gamma^2 - \gamma^{2(\hat{\tau}_L - \hat{\tau}_j)}} \cdot \gamma^{t - \hat{\tau}_j}, & \hat{\tau}_L \leq t \leq \hat{\tau}_j \\ \frac{\gamma^2 - 1}{\gamma^{2(\hat{\tau}_R - \hat{\tau}_j)} - 1} \cdot \gamma^{t - (\hat{\tau}_j + 1)}, & \hat{\tau}_j + 1 \leq t \leq \hat{\tau}_R \\ 0, & \text{otherwise} \end{cases}, \tag{A.1}$$

where $\hat{\tau}_L = \max(1, \hat{\tau}_j - h + 1)$, and $\hat{\tau}_R = \min(T, \hat{\tau}_j + h)$.

In Figure A.1, we plot the contrast vector in (2.7), generated with $T = 50$, $\gamma = 0.98$, $\hat{\tau}_j = 20$, and $h = 5$.

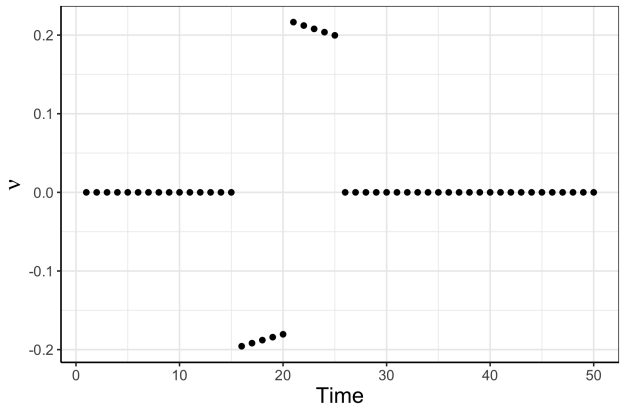


Figure A.1: Plot of the contrast ν generated according to (2.7), with $T = 50$, $\gamma = 0.98$, $\hat{\tau}_j = 20$, and $h = 5$.

A.3 Proof of Proposition 3

Recall that the ℓ_0 problem (2.14) is equivalent to the changepoint detection problem (2.15), in the sense that (2.14) results in an estimated changepoint at $\hat{\tau}_j$ if and only if $\hat{\tau}_j$ is in the

solution to (2.15).

We first prove that $C(\phi)$ defined in (2.18) equals the objective of (2.15) applied to data $y'(\phi)$, subject to the constraint that $\hat{\tau}_j$ is in the solution.

$$
\begin{aligned}
C(\phi) &= \min_{\alpha\geq 0}\left\{\text{Cost}\left(y'_{1:\hat{\tau}_j}(\phi),\alpha;\gamma\right)\right\} + \min_{\alpha\geq 0}\left\{\text{Cost}\left(y'_{T:(\hat{\tau}_j+1)}(\phi),\alpha;1/\gamma\right)\right\} + \lambda \\
&\overset{a.}{=} \min_{\alpha\geq 0}\left\{\text{Cost}\left(y'_{1:\hat{\tau}_j}(\phi),\alpha;\gamma\right)\right\} + \min_{\alpha\geq 0}\left\{\text{Cost}\left(y'_{(\hat{\tau}_j+1):T}(\phi),\alpha;\gamma\right)\right\} + \lambda \\
&\overset{b.}{=} \min_{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\hat{\tau}_j,k}\left\{\sum_{j=0}^{k}\min_{\alpha\geq 0}\left(\frac{1}{2}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y'_t(\phi)-\alpha\gamma^{t-\tau_{j+1}}\right)^2\right)+\lambda k\right\} \\
&+ \min_{\hat{\tau}_j=\tilde{\tau}_0<\tilde{\tau}_1<\ldots<\tilde{\tau}_l<\tilde{\tau}_{l+1}=T,l}\left\{\sum_{j=0}^{l}\min_{\alpha\geq 0}\left(\frac{1}{2}\sum_{t=\tilde{\tau}_j+1}^{\tilde{\tau}_{j+1}}\left(y'_t(\phi)-\alpha\gamma^{t-\tilde{\tau}_{j+1}}\right)^2\right)+\lambda l\right\} + \lambda \\
&\overset{c.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\hat{\tau}_j,k,\\ \hat{\tau}_j=\tilde{\tau}_0<\tilde{\tau}_1<\ldots<\tilde{\tau}_l<\tilde{\tau}_{l+1}=T,l}}\left\{\sum_{j=0}^{k}\min_{\alpha\geq 0}\left(\frac{1}{2}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y'_t(\phi)-\alpha\gamma^{t-\tau_{j+1}}\right)^2\right)\right. \\
&\left.+\sum_{j=0}^{l}\min_{\alpha\geq 0}\left(\frac{1}{2}\sum_{t=\tilde{\tau}_j+1}^{\tilde{\tau}_{j+1}}\left(y'_t(\phi)-\alpha\gamma^{t-\tilde{\tau}_{j+1}}\right)^2\right)+\lambda(k+l)\right\}+\lambda \\
&\overset{d.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T,\\ k,\hat{\tau}_j\in\{\tau_1,\ldots,\tau_k\}}}\left\{\sum_{j=0}^{k}\min_{\alpha\geq 0}\left(\frac{1}{2}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y'_t(\phi)-\alpha\gamma^{t-\tau_{j+1}}\right)^2\right)+\lambda k\right\}.
\end{aligned}
$$

Here, *a.* follows from Lemma 5 and *b.* follows from Lemma 4. Part *c.* follows from combining the two minimization problems, and finally part *d.* follows from treating $(k+l)$ as a new variable in the optimization problem.

Next, we show that $C'(\phi)$ defined in (2.19) equals the objective of (2.15) applied to data $y'(\phi)$, subject to the constraint that $\hat{\tau}_j$ is *not* in the solution.

$$
\begin{aligned}
C'(\phi) &= \min_{\alpha \geq 0} \left\{ \text{Cost}\Big(y'_{1:\hat{\tau}_j}(\phi), \alpha; \gamma\Big) + \text{Cost}\Big(y'_{T:(\hat{\tau}_j+1)}(\phi), \gamma\alpha; 1/\gamma\Big) \right\} \\
&\overset{a.}{=} \min_{\alpha \geq 0} \left\{ \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\hat{\tau}_j, \\ \alpha_0,\ldots,\alpha_{k-1}\geq 0, \alpha_k=\alpha, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha_j \gamma^{t-\tau_{j+1}}\right)^2 + \lambda k \right\} \right. \\
&\quad \left. + \min_{\substack{\hat{\tau}_j=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T, \\ \alpha_0,\ldots,\alpha_k\geq 0, \alpha_k=\gamma\alpha, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_{T+\hat{\tau}_j+1-t}(\phi) - \alpha_j (1/\gamma)^{t-\tau_{j+1}}\right)^2 + \lambda k \right\} \right\} \\
&\overset{b.}{=} \min_{\alpha \geq 0} \left\{ \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\hat{\tau}_j, \\ \alpha_0,\ldots,\alpha_{k-1}\geq 0, \alpha_k=\alpha, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha_j \gamma^{t-\tau_{j+1}}\right)^2 + \lambda k \right\} \right. \\
&\quad \left. + \min_{\substack{T=\tilde{\tau}_0>\tilde{\tau}_1>\ldots>\tilde{\tau}_k>\tilde{\tau}_{k+1}=\hat{\tau}_j, \\ \alpha_0,\ldots,\alpha_k\geq 0, \alpha_k=\gamma\alpha, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tilde{\tau}_{j+1}}^{\tilde{\tau}_j+1} \left(y'_t(\phi) - \alpha_j \gamma^{t-\tilde{\tau}_{j+1}}\right)^2 + \lambda k \right\} \right\} \\
&\overset{c.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\hat{\tau}_j, \\ \alpha_0,\ldots,\alpha_{k-1}\geq 0, \alpha_k=\alpha, k, \\ \hat{\tau}_j=\tilde{\tau}_{\tilde{k}+1}<\tilde{\tau}_{\tilde{k}}<\ldots<\tilde{\tau}_1<\tilde{\tau}_0=T \\ \tilde{\alpha}_0,\ldots,\tilde{\alpha}_{\tilde{k}}\geq 0, \tilde{\alpha}_{\tilde{k}}=\gamma\alpha, \tilde{k}}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha_j \gamma^{t-\tau_{j+1}}\right)^2 + \lambda k + \frac{1}{2} \sum_{j=0}^{\tilde{k}} \sum_{t=\tilde{\tau}_j+1}^{\tilde{\tau}_{j+1}} \left(y'_t(\phi) - \tilde{\alpha}_j \gamma^{t-\tilde{\tau}_{j+1}}\right)^2 + \lambda \tilde{k} \right\} \\
&\overset{d.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T, \\ \alpha_0,\ldots,\alpha_k\geq 0, k, \\ \hat{\tau}_j\neq\tau_j, \forall j=1,\ldots,k.}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha_j \gamma^{t-\tau_{j+1}}\right)^2 + \lambda k \right\} \\
&\overset{e.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T, \\ k, \hat{\tau}_j\notin\{\tau_1,\ldots,\tau_k\}}} \left\{ \sum_{j=0}^{k} \min_{\alpha\geq 0} \left(\frac{1}{2} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha \gamma^{t-\tau_{j+1}}\right)^2 \right) + \lambda k \right\}.
\end{aligned}
$$

Part $a.$ follows from expanding $\text{Cost}(\cdot)$ using Lemma 3. We then change the optimization variable in the second term from τ_j to $\tilde{\tau}_j = T + \hat{\tau}_j - \tau_j$, which does not change the optimization problem because the mapping between $\tilde{\tau}_j$ and τ_j is invertible; re-indexing the summation completes part $b.$ Next, $c.$ follows from combining the two optimization problems. In step $d.$, we observe that the two constraints $\alpha_k = \alpha$ (i.e., fitted value at timepoint $\hat{\tau}_j$ is α) and $\tilde{\alpha}_{\tilde{k}} = \gamma\alpha$ (i.e., fitted value at timepoint $\hat{\tau}_j + 1$ is $\gamma\alpha$) are equivalent to a single constraint that

$\hat{\tau}_j$ is not a changepoint. Finally, step $e.$ follows from pulling the optimization over α_j inside the summation.

To summarize, we have proven that

$$C(\phi) = \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T,\\ k,\hat{\tau}_j\in\{\tau_1,\ldots,\tau_k\}}} \left\{ \sum_{j=0}^{k} \min_{\alpha\geq 0} \left(\frac{1}{2} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha\gamma^{t-\tau_{j+1}} \right)^2 \right) + \lambda k \right\}, \tag{A.2}$$

and

$$C'(\phi) = \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=T,\\ k,\hat{\tau}_j\notin\{\tau_1,\ldots,\tau_k\}}} \left\{ \sum_{j=0}^{k} \min_{\alpha\geq 0} \left(\frac{1}{2} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y'_t(\phi) - \alpha\gamma^{t-\tau_{j+1}} \right)^2 \right) + \lambda k \right\}. \tag{A.3}$$

By inspection of (A.2) and (A.3), we conclude that $\{\phi : C(\phi) \leq C'(\phi)\} = \{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}$, which completes the proof.

We present the technical lemmas used in the proof below.

Lemma 3. *For* $\text{Cost}(y_{1:s}, \alpha; \gamma)$ *defined in* (2.16)*, we have*

$$\text{Cost}(y_{1:s}, \alpha; \gamma) = \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s,\\ \alpha_0,\ldots,\alpha_k\geq 0, \alpha_k=\alpha, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y_t - \alpha_j\gamma^{t-\tau_{j+1}} \right)^2 + \lambda k \right\}. \tag{A.4}$$

Proof.

$$\begin{aligned}
\text{Cost}(y_{1:s},\alpha;\gamma) &\overset{a.}{=} \min_{0\le\tau<s}\left\{F(\tau)+\frac{1}{2}\left(\sum_{t=\tau+1}^{s}(y_t-\alpha\gamma^{t-s})^2\right)+\lambda\right\}\\
&\overset{b.}{=} \min_{0\le\tau<s}\left\{\min_{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\tau,k}\frac{1}{2}\left(\sum_{j=0}^{k}\min_{\alpha\ge0}\left\{\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha\gamma^{t-\tau_{j+1}}\right)^2\right\}+\lambda k\right)\right.\\
&\left.+\frac{1}{2}\left(\sum_{t=\tau+1}^{s}\left(y_t-\alpha\gamma^{t-s}\right)^2\right)+\lambda\right\}\\
&\overset{c.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=\tau<s,\\ \alpha_0,\ldots,\alpha_k\ge0,k,\tau}}\left\{\frac{1}{2}\sum_{j=0}^{k}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha_j\gamma^{t-\tau_{j+1}}\right)^2+\lambda k+\frac{1}{2}\sum_{t=\tau_{k+1}+1}^{s}\left(y_t-\alpha\gamma^{t-s}\right)^2+\lambda\right\}\\
&\overset{d.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s,\\ \alpha_0,\ldots,\alpha_k\ge0,\alpha_k=\alpha,k}}\left\{\frac{1}{2}\sum_{j=0}^{k}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha_j\gamma^{t-\tau_{j+1}}\right)^2+\lambda k\right\}.
\end{aligned}$$

Here, *a.* follows from the definition in (2.16) and *b.* follows from the definition of $F(\tau)$, the optimal cost of segmenting the first τ data points. Part *c.* follows from pulling the $\min_{\alpha\ge0}$ operation out of the summation, which is performed separately for each data segment $y_{(\tau_j+1):\tau_{j+1}}$. Finally, part *d.* follows by inspection. □

Lemma 4. *For* $\text{Cost}(y_{1:s},\alpha;\gamma)$ *defined in* (2.16)*, we have*

$$\min_{\alpha\ge0}\left\{\text{Cost}(y_{1:s},\alpha;\gamma)\right\}=\min_{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s,k}\left\{\sum_{j=0}^{k}\min_{\alpha\ge0}\left\{\frac{1}{2}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha\gamma^{t-\tau_{j+1}}\right)^2\right\}+\lambda k\right\}.$$

Proof.

$$\begin{aligned}
\min_{\alpha\ge0}\left\{\text{Cost}(y_{1:s},\alpha;\gamma)\right\} &\overset{a.}{=} \min_{\alpha\ge0}\left\{\min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s,\\ \alpha_0,\ldots,\alpha_k\ge0,\alpha_k=\alpha,k}}\left\{\frac{1}{2}\sum_{j=0}^{k}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha_j\gamma^{t-\tau_{j+1}}\right)^2+\lambda k\right\}\right\}\\
&= \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s,\\ \alpha_0,\ldots,\alpha_k\ge0,k}}\left\{\frac{1}{2}\sum_{j=0}^{k}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha_j\gamma^{t-\tau_{j+1}}\right)^2+\lambda k\right\}\\
&\overset{b.}{=} \min_{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s,k}\left\{\frac{1}{2}\sum_{j=0}^{k}\min_{\alpha\ge0}\left\{\frac{1}{2}\sum_{t=\tau_j+1}^{\tau_{j+1}}\left(y_t-\alpha\gamma^{t-\tau_{j+1}}\right)^2\right\}+\lambda k\right\}.
\end{aligned}$$

Here, *a.* follows from Lemma 3. *b.* follows from noting that α_j can be minimized independently for each data segment $y_{(\hat{\tau}_j+1):\hat{\tau}_{j+1}}$. □

Lemma 5. *For* $\text{Cost}(y_{1:s}, \alpha; \gamma)$ *defined in* (2.16)*, we have*

$$\min_{\alpha \geq 0} \{\text{Cost}(y_{1:s}, \alpha; \gamma)\} = \min_{\alpha \geq 0} \{\text{Cost}(y_{s:1}, \alpha; 1/\gamma)\}.$$

Proof.

$$\begin{aligned}
\min_{\alpha \geq 0} \{\text{Cost}(y_{1:s}, \alpha; \gamma)\} &\overset{a.}{=} \min_{\substack{0=\tau_0<\tau_1<\ldots<\tau_k<\tau_{k+1}=s, \\ \alpha_0,\ldots,\alpha_k \geq 0, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tau_j+1}^{\tau_{j+1}} \left(y_t - \alpha_j \gamma^{t-\tau_{j+1}}\right)^2 + \lambda k \right\} \\
&\overset{b.}{=} \min_{\substack{s=\tilde{\tau}_0>\tilde{\tau}_1>\ldots>\tilde{\tau}_k>\tilde{\tau}_{k+1}=0, \\ \alpha_0,\ldots,\alpha_k \geq 0, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=s-\tilde{\tau}_j+1}^{s-\tilde{\tau}_{j+1}} \left(y_t - \alpha_j \gamma^{t-(s-\tilde{\tau}_{j+1})}\right)^2 + \lambda k \right\} \\
&\overset{c.}{=} \min_{\substack{s=\tilde{\tau}_0>\tilde{\tau}_1>\ldots>\tilde{\tau}_k>\tilde{\tau}_{k+1}=0, \\ \alpha_0,\ldots,\alpha_k \geq 0, k}} \left\{ \frac{1}{2} \sum_{j=0}^{k} \sum_{t=\tilde{\tau}_{j+1}}^{\tilde{\tau}_j} \left(y_{s-t} - \alpha_j (1/\gamma)^{t-\tilde{\tau}_{j+1}}\right)^2 + \lambda k \right\} \\
&\overset{d.}{=} \min_{\alpha \geq 0} \{\text{Cost}(y_{s:1}, \alpha; 1/\gamma)\}.
\end{aligned}$$

Part *a.* follows from Lemma 3. In step *b.*, we change the optimization variable from τ_j to $\tilde{\tau}_j = s - \tau_j$, which does not change the optimization problem because the mapping between $\tilde{\tau}_j$ and τ_j is invertible. Step *c.* follows from re-indexing the summation, and finally *d.* follows from Lemma 3 again. □

A.4 Proof of Proposition 4

To begin, we will prove (2.21) using an induction argument. The following claim serves as the "base case" for the recursion.

Lemma 6.

$$\text{Cost}\left(y'_{1:(\hat{\tau}_j-h+1)}(\phi), \alpha; \gamma\right) = \min_{f \in \mathcal{C}_{\hat{\tau}_j-h+1}} f(\alpha, \phi), \tag{A.5}$$

where

$$\mathcal{C}_{\hat{\tau}_j-h+1} = \Bigg\{ \text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi), \alpha/\gamma; \gamma\right) + \frac{1}{2}\left(y'_{\hat{\tau}_j-h+1}(\phi) - \alpha\right)^2, \\ \min_{\alpha' \geq 0}\left\{\text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi), \alpha'; \gamma\right)\right\} + \lambda + \frac{1}{2}\left(y'_{\hat{\tau}_j-h+1}(\phi) - \alpha\right)^2 \Bigg\}. \tag{A.6}$$

Proof. To prove Lemma 6, we will first compute Cost $\left(y'_{1:(\hat{\tau}_j-h+1)}(\phi), \alpha; \gamma\right)$ using the definition in (2.16); we will then show that this equals $\min_{f \in \mathcal{C}_{\hat{\tau}_j-h+1}} f(\alpha, \phi)$, with $\mathcal{C}_{\hat{\tau}_j-h+1}$ in (A.6).

Per the definition of ν in (4.3), $y'_{1:(\hat{\tau}_j-h)}(\phi) = y_{1:(\hat{\tau}_j-h)}$; therefore, $\mathcal{C}_{\hat{\tau}_j-h} = \text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi), \alpha; \gamma\right) = \text{Cost}\left(y_{1:(\hat{\tau}_j-h)}, \alpha; \gamma\right)$. From Proposition 2, this means that $\mathcal{C}_{\hat{\tau}_j-h}$ is a piecewise quadratic function of α only.

Now we consider the function Cost $\left(y'_{1:(\hat{\tau}_j-h+1)}(\phi), \alpha; \gamma\right)$. There are two possibilities:

1. *There is no changepoint at the* $(\hat{\tau}_j - h)$*th time step.* In this case, Cost $\left(y'_{1:(\hat{\tau}_j-h+1)}(\phi), \alpha; \gamma\right)$ equals

   $$\text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi), \alpha/\gamma; \gamma\right) + \frac{1}{2}\left(y'_{\hat{\tau}_j-h+1}(\phi) - \alpha\right)^2,$$

   where α/γ accounts for the exponential calcium decay.

2. *There is a changepoint at the* $(\hat{\tau}_j - h)$*th time step.* In this case, Cost $\left(y'_{1:(\hat{\tau}_j-h+1)}(\phi), \alpha; \gamma\right)$ equals

   $$\min_{\alpha' \geq 0}\left\{\text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi), \alpha'; \gamma\right)\right\} + \lambda + \frac{1}{2}\left(y'_{\hat{\tau}_j-h+1}(\phi) - \alpha\right)^2,$$

   where the changepoint incurs a penalty of λ, and there can be an arbitrary change in the calcium from timepoint $\hat{\tau}_j - h$ to $\hat{\tau}_j - h + 1$.

Therefore,

$$\begin{aligned}\text{Cost}\left(y'_{1:(\hat{\tau}_j-h+1)}(\phi),\alpha;\gamma\right) &= \min\Bigg\{\text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi),\alpha/\gamma;\gamma\right)+\frac{1}{2}\left(y'_{\hat{\tau}_j-h+1}(\phi)-\alpha\right)^2, \\ &\qquad \min_{\alpha'\geq 0}\left\{\text{Cost}\left(y'_{1:(\hat{\tau}_j-h)}(\phi),\alpha';\gamma\right)\right\}+\lambda+\frac{1}{2}\left(y'_{\hat{\tau}_j-h+1}(\phi)-\alpha\right)^2\Bigg\} \\ &= \min_{f\in\mathcal{C}_{\hat{\tau}_j-h+1}} f(\alpha,\phi), \end{aligned} \tag{A.7}$$

where the last equality follows from (A.6). This completes the proof. □

We will now prove the inductive step for the recursion, which relies on the following claim.

Lemma 7. *Suppose that for some* $s\in\{\hat{\tau}_j-h+1,\ldots,\hat{\tau}_j-1\}$,

$$\text{Cost}\left(y'_{1:s}(\phi),\alpha;\gamma\right)=\min_{f\in\mathcal{C}_s} f(\alpha,\phi). \tag{A.8}$$

Then,

$$\text{Cost}\left(y'_{1:(s+1)}(\phi),\alpha;\gamma\right)=\min_{f\in\mathcal{C}_{s+1}} f(\alpha,\phi), \tag{A.9}$$

where $\mathcal{C}_{s+1}$ *is defined recursively according to* (2.23).

Proof. To begin, we apply Proposition 2 with $y'(\phi)$ instead of y and get

$$\text{Cost}\left(y'_{1:(s+1)}(\phi),\alpha;\gamma\right)=\min\left\{\text{Cost}\left(y'_{1:s}(\phi),\alpha/\gamma;\gamma\right),\min_{\alpha'\geq 0}\text{Cost}\left(y'_{1:s}(\phi),\alpha';\gamma\right)+\lambda\right\}+\frac{1}{2}\left(y'_{s+1}(\phi)-\alpha\right)^2. \tag{A.10}$$

Applying the inductive hypothesis in (A.8) with α/γ instead of α, we have that

$$\text{Cost}\left(y'_{1:s}(\phi),\alpha/\gamma;\gamma\right)=\min_{f\in\mathcal{C}_s} f(\alpha/\gamma,\phi), \tag{A.11}$$

and

$$\min_{\alpha' \geq 0} \left\{ \text{Cost}\left(y'_{1:s}(\phi), \alpha'; \gamma\right) \right\} = \min_{\alpha' \geq 0} \left\{ \min_{f \in \mathcal{C}_s} f(\alpha', \phi) \right\}. \tag{A.12}$$

Therefore,

$$\text{Cost}\left(y'_{1:(s+1)}(\phi), \alpha; \gamma\right) \stackrel{a.}{=} \min \left\{ \min_{f \in \mathcal{C}_s} f(\alpha/\gamma, \phi), \min_{\alpha' \geq 0} \left\{ \min_{f \in \mathcal{C}_s} f(\alpha', \phi) \right\} + \lambda \right\} + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \tag{A.13}$$

$$\stackrel{b.}{=} \min \left\{ \min_{f \in \mathcal{C}_s} f(\alpha/\gamma, \phi) + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2, \min_{f \in \mathcal{C}_s} \left\{ \min_{\alpha' \geq 0} \left\{ f(\alpha', \phi) \right\} \right\} + \lambda + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \right\}, \tag{A.14}$$

where $a.$ follows from (A.8) and (A.10), and $b.$ follows from exchanging the order of minimization and distributing the $\frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2$ term inside.

Furthermore,

$$\min_{f \in \mathcal{C}_{s+1}} f(\alpha, \phi) \stackrel{a.}{=} \min_{f \in \left\{ \left(\bigcup_{f \in \mathcal{C}_s} \left\{ f(\alpha/\gamma, \phi) + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \right\} \right) \bigcup \left\{ g_{s+1}(\phi) + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \right\} \right\}} f(\alpha, \phi) \tag{A.15}$$

$$\stackrel{b.}{=} \min \left\{ \min_{f \in \mathcal{C}_s} \left\{ f(\alpha/\gamma, \phi) + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \right\}, g_{s+1}(\phi) + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \right\} \tag{A.16}$$

$$\stackrel{c.}{=} \min \left\{ \min_{f \in \mathcal{C}_s} \left\{ f(\alpha/\gamma, \phi) \right\} + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2, \min_{f \in \mathcal{C}_s} \left\{ \min_{\alpha \geq 0} \left\{ f(\alpha, \phi) \right\} \right\} + \lambda + \frac{1}{2}\left(y'_{s+1}(\phi) - \alpha\right)^2 \right\}, \tag{A.17}$$

where $a.$ follows from the definition of $\mathcal{C}_{s+1}$ in (2.23); $b.$ follows from noting that $\min_{f \in A \bigcup B} f = \min\left\{\min_{f \in A} f, \min_{f \in B} f\right\}$; and $c.$ follows from the definition of $g_{s+1}(\phi)$ in (2.24).

Now by inspection, (A.14) is equal to (A.17); this completes the proof. □

The inductive proof of (2.21) follows directly from combining Lemmas 6 and 7.

We will now show that for $\hat{\tau}_j - h + 1 \leq s \leq \hat{\tau}_j$, $\mathcal{C}_s$ is a collection of piecewise quadratic functions of α and ϕ. We will show this by induction. We first make the following observations, which follow from simple algebra:

- *Observation 1:* For $\hat{\tau}_j - h + 1 \leq s \leq \hat{\tau}_j$, $\frac{1}{2}(y'_s(\phi) - \alpha)^2$ is a quadratic function of ϕ and α, where $y'(\phi)$ is defined in (3.12).

- *Observation 2:* If both $f_1(\alpha, \phi)$ and $f_2(\alpha, \phi)$ are piecewise quadratic functions of α and ϕ, then $f_1 + f_2$ is also a piecewise quadratic function of α and ϕ.

- *Observation 3:* If $f(\alpha, \phi)$ is a piecewise quadratic function of ϕ and α, then $\min_{\alpha \geq 0} f(\alpha, \phi)$ is a piecewise quadratic function of only ϕ.

- *Observation 4:* If $\mathcal{C}_s$ is a finite set of piecewise quadratic functions of ϕ and α, then $\min_{f \in \mathcal{C}_s} f(\alpha, \phi)$ is a piecewise quadratic function of ϕ and α.

In our induction, Lemma 8 serves as our "base case". The induction step is presented in Lemma 9.

Lemma 8. *$\mathcal{C}_{\hat{\tau}_j - h + 1}$ is a collection of piecewise quadratic functions of α and ϕ.*

Proof. Applying the recursion in (2.23), we see that

$$\mathcal{C}_{\hat{\tau}_j - h + 1} = \Bigg\{ \text{Cost}\Big(y'_{1:(\hat{\tau}_j - h)}(\phi), \alpha; \gamma\Big) + \frac{1}{2}\Big(y'_{\hat{\tau}_j - h + 1}(\phi) - \alpha\Big)^2, \\ \min_{\alpha \geq 0} \Big\{ \text{Cost}\Big(y'_{1:(\hat{\tau}_j - h)}(\phi), \alpha; \gamma\Big) \Big\} + \lambda + \frac{1}{2}\Big(y'_{\hat{\tau}_j - h + 1}(\phi) - \alpha\Big)^2 \Bigg\}.$$

By Proposition 2, $\text{Cost}(y'_{1:(\hat{\tau}_j - h)}(\phi), \alpha; \gamma) = \text{Cost}(y_{1:(\hat{\tau}_j - h)}, \alpha; \gamma)$ is a piecewise quadratic function of α. Furthermore, $\frac{1}{2}(y'_{\hat{\tau}_j - h + 1}(\phi) - \alpha)^2$ is a quadratic function of ϕ and α, according to Observation 1. Therefore, the first term in $\mathcal{C}_{\hat{\tau}_j - h + 1}$ is a piecewise quadratic function of ϕ and α according to Observation 2. As for the second term, we note that $\min_{\alpha \geq 0} \Big\{ \text{Cost}(y'_{1:(\hat{\tau}_j - h)}(\phi), \alpha; \gamma) \Big\}$ is a piecewise quadratic function of ϕ according to Observation 3, so its sum with $\lambda + \frac{1}{2}(y'_{\hat{\tau}_j - h + 1}(\phi) - \alpha)^2$ is piecewise quadratic in ϕ and α. □

Lemma 9. *Suppose that for some $s \in \{\hat{\tau}_j - h + 1, \ldots, \hat{\tau}_j - 1\}$, $\mathcal{C}_s$ is a collection of piecewise quadratic functions of α and ϕ. Then,*

$$\mathcal{C}_{s+1} = \left(\bigcup_{f \in \mathcal{C}_s} \left\{ f(\alpha/\gamma, \phi) + \frac{1}{2}\big(y'_{s+1}(\phi) - \alpha\big)^2 \right\} \right) \bigcup \left\{ g_{s+1}(\phi) + \frac{1}{2}\big(y'_{s+1}(\phi) - \alpha\big)^2 \right\} \quad \text{(A.18)}$$

is also a collection of piecewise quadratic functions of α and ϕ, where g_{s+1} is defined in (2.24).

Proof. According to the induction hypothesis, each $f \in \mathcal{C}_s$ is a piecewise quadratic function of α and ϕ. Therefore, $f(\alpha/\gamma, \phi) + \frac{1}{2}(y'_{s+1}(\phi) - \alpha)^2$ is a piecewise quadratic function of α and ϕ for all $f \in \mathcal{C}_s$, according to Observation 2. Furthermore, from Observations 3 and 4, we can see that

$$g_{s+1}(\phi) = \min_{f \in \mathcal{C}_s} \min_{\alpha \geq 0} f(\alpha, \phi) + \lambda$$

is a piecewise quadratic function of ϕ. □

Combining Lemmas 8 and 9 completes the argument that for $s \in \{\hat{\tau}_j - h, \ldots, \hat{\tau}_j\}$, $\mathcal{C}_s$ is a collection of piecewise quadratic functions.

To complete the proof of Proposition 4, it remains to show that for $s \in \{\hat{\tau}_j - h, \ldots, \hat{\tau}_j\}$, $|\mathcal{C}_s| = s - \hat{\tau}_j + h + 1$, where $|\mathcal{C}_s|$ is the cardinality of the set $\mathcal{C}_s$. According to (2.22), $\mathcal{C}_{\hat{\tau}_j - h}$ consists of a single function. At each iteration of the recursion in (2.23), only one additional function is added; therefore, $\mathcal{C}_s$ consists of $1 + s - (\hat{\tau}_j - h) = s - \hat{\tau}_j + h + 1$ functions.

A.5 Extension of Proposition 4 to $y'_{T:(\hat{\tau}_j+1)}(\phi)$

The following proposition is a straightforward extension of Proposition 4 to the sequence $y'_{T:(\hat{\tau}_j+1)}(\phi)$ with decay parameter $1/\gamma$ to account for the time reversal.

Proposition 18. *For $\hat{\tau}_j + 1 \leq s \leq \hat{\tau}_j + h$,*

$$\text{Cost}(y'_{T:s}(\phi), \alpha; 1/\gamma) = \min_{f \in \tilde{\mathcal{C}}_s} f(\alpha, \phi), \tag{A.19}$$

where $\tilde{\mathcal{C}}_s$ is a collection of $\hat{\tau}_j + h + 2 - s$ piecewise quadratic functions of α and ϕ, $f(\alpha, \phi)$,

constructed with the initialization

$$\tilde{\mathcal{C}}_{\hat{\tau}_j+h+1} = \left\{ \text{Cost}\left(y'_{T:(\hat{\tau}_j+h+1)}(\phi), \alpha; 1/\gamma \right) \right\}, \tag{A.20}$$

and the recursion

$$\tilde{\mathcal{C}}_s = \left(\bigcup_{f \in \tilde{\mathcal{C}}_{s+1}} \left\{ f(\alpha\gamma, \phi) + \frac{1}{2}(y'_s(\phi) - \alpha)^2 \right\} \right) \bigcup \left\{ g_s(\phi) + \frac{1}{2}(y'_s(\phi) - \alpha)^2 \right\}, \tag{A.21}$$

where

$$g_s(\phi) = \min_{f \in \tilde{\mathcal{C}}_{s+1}} \min_{\alpha \geq 0} f(\alpha, \phi) + \lambda \tag{A.22}$$

and $y'(\phi)$ is defined in (3.12).

A.6 General case for Propositions 4 and 18

Propositions 4 and 18 assumed that $\hat{\tau}_j - h \geq 1$ and $\hat{\tau}_j + h + 1 \leq T$ (where T is the length of the observed data), respectively. We now provide details for the cases where $\hat{\tau}_j - h < 1$ and $\hat{\tau}_j + h + 1 > T$.

- *Case 1:* $\hat{\tau}_j - h < 1$. Define $\hat{\tau}_L = \max\{1, \hat{\tau}_j - h\}$ and initialize with

$$\mathcal{C}_{\hat{\tau}_L} = \left\{ \text{Cost}\left(y'_{1:\hat{\tau}_L}(\phi), \alpha; \gamma \right) \right\} \tag{A.23}$$

  in Proposition 4 instead of (2.22), with the convention $y'_{1:1}(\phi) = y'_1(\phi)$.

- *Case 2:* $\hat{\tau}_j + h + 1 > T$. Define $\hat{\tau}_R = \min\{T, \hat{\tau}_j + h + 1\}$ and initialize with

$$\mathcal{C}_{\hat{\tau}_R} = \left\{ \text{Cost}\left(y'_{T:\hat{\tau}_R}(\phi), \alpha; 1/\gamma \right) \right\} \tag{A.24}$$

  in Proposition 18 instead of (A.20), with the convention $y'_{T:T}(\phi) = y'_T(\phi)$.

A.7 Algorithm for computing $\mathcal{S}$ in (2.20)

Algorithm 2: Computing $\mathcal{S}$ in (2.20) for a spike $\hat{\tau}_j$ resulting from (2.14)

Input : Data $y_{1:T}$, spike location $\hat{\tau}_j$, exponential decay parameter γ
Output: Set $\mathcal{S}$

1. Compute the collection of functions $\mathcal{C}_{\hat{\tau}_j}$ using Proposition 4.
2. Compute the collection of functions $\tilde{\mathcal{C}}_{\hat{\tau}_j+1}$ using Proposition 18.
3. Compute $C(\phi)$ using (2.25).
4. Compute $C'(\phi)$ using (2.26).
5. Compute $\mathcal{S} = \{\phi : C(\phi) \leq C'(\phi)\}$.

A.8 Proof of Proposition 5

Throughout the proof, we assume that the number of pieces in the piecewise quadratic functions under consideration is a constant that does not depend on h and T. Moreover, we will leverage the toolkit from Jewell et al. [2022], Maidstone et al. [2017], Rigaill [2015], which allows for efficient manipulation of both univariate and bivariate piecewise quadratic functions. Provided with an efficient implementation of the toolkit, we make the following two observations for our timing complexity analysis:

- *Observation 1:* $\min_{f \in \mathcal{C}} f(\phi)$ can be computed in $O(|\mathcal{C}|)$ operations, provided that $f(\phi)$ is a piecewise quadratic function of ϕ;
- *Observation 2:* $\forall f_1, f_2 \in \mathcal{C}$, $f_1(\alpha, \phi) + f_2(\alpha, \phi)$ can be computed in $O(1)$ operations, provided that $f_1(\alpha, \phi)$ and $f_2(\alpha, \phi)$ are piecewise quadratic functions of α and ϕ with $O(1)$ pieces.

Finally, we recall that if $f(\alpha, \phi)$ is a piecewise quadratic function of α and ϕ, then

$\min_{\alpha\geq 0}\{f(\alpha,\phi)\}$ is a piecewise quadratic function of ϕ only and can be computed analytically.

Now we will characterize the computational complexity of Algorithm 2:

1. Step 1: We first consider the time to compute $\mathcal{C}_s$ for some $s \in \{\hat{\tau}_j - h + 1, \ldots, \hat{\tau}_j\}$, assuming that we have computed $\mathcal{C}_{s-1}$.

   (a) We first compute $\bigcup_{f\in\mathcal{C}_{s-1}}\{f(\alpha/\gamma,\phi) + \frac{1}{2}(y'_s(\phi) - \alpha)^2\}$, which takes $O(|\mathcal{C}_{s-1}|) = O(s - \hat{\tau}_j + h)$ operations.

   (b) We then compute $g_s(\phi)$ using (2.24): the inner minimization over $\alpha \geq 0$ takes $O(1)$ operations for each $f \in \mathcal{C}_{s-1}$ since it admits an analytical solution; the outer minimization over $\mathcal{C}_{s-1}$ takes $O(|\mathcal{C}_{s-1}|) = O(s - \hat{\tau}_j + h)$ operations according to Observation 1.

   In summary, computing $\mathcal{C}_s$ from $\mathcal{C}_{s-1}$ takes $O(s - \hat{\tau}_j + h)$ operations for any $s \in \{\hat{\tau}_j - h + 1, \ldots, \hat{\tau}_j\}$. The first step of Algorithm 2 requires computing $\mathcal{C}_s$ for all $s \in \{\hat{\tau}_j - h + 1, \ldots, \hat{\tau}_j\}$, a total of $O\Big(\sum_{t=\hat{\tau}_j-h+1}^{\hat{\tau}_j}(t - \hat{\tau}_j + h)\Big) = O(h^2)$ operations.

2. Step 2: Applying the same logic used in analyzing Step 1 to the second step of Algorithm 2, we conclude that computing $\tilde{\mathcal{C}}_{\hat{\tau}_j+1}$ takes $O(h^2)$ operations using Proposition 18.

3. According to (2.25), computing $C(\phi)$ requires $\min_{f\in\mathcal{C}_{\hat{\tau}_j}}\{\min_{\alpha\geq 0} f(\alpha,\phi)\}$ and $\min_{f\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}}\{\min_{\alpha'\geq 0} f(\alpha',\phi)\}$. Both terms can be computed in $O(|\mathcal{C}_{\hat{\tau}_j}|) = O(h)$ operations using Observation 1; moreover, the summation will take $O(1)$ operations according to Observation 2. Hence Step 3 takes $O(h)$ operations in total.

4. According to (2.26),

$$C'(\phi) = \min_{f\in\mathcal{C}_{\hat{\tau}_j}, \tilde{f}\in\tilde{\mathcal{C}}_{\hat{\tau}_j+1}} \left\{ \min_{\alpha\geq 0} \left\{ f(\alpha,\phi) + \tilde{f}(\gamma\alpha,\phi) \right\} \right\}.$$

(a) Computing the set $\left\{ f(\alpha,\phi) + \tilde{f}(\gamma\alpha,\phi) \;\middle|\; f \in \mathcal{C}_{\hat{\tau}_j}, \tilde{f} \in \tilde{\mathcal{C}}_{\hat{\tau}_j+1} \right\}$ takes $O(|\mathcal{C}_{\hat{\tau}_j}| \cdot |\tilde{\mathcal{C}}_{\hat{\tau}_j+1}|) = O(h^2)$ operations, since each addition takes $O(1)$ operations (Observation 2) and there are $|\mathcal{C}_{\hat{\tau}_j}| \cdot |\tilde{\mathcal{C}}_{\hat{\tau}_j+1}|$ such sums.

(b) Minimizing over $\alpha \geq 0$ for each $f(\alpha,\phi) + \tilde{f}(\gamma\alpha,\phi)$ takes $O(1)$ operations, so the cost of minimization over the entire collection is $O(h^2)$.

(c) Computing $C'(\phi)$ as the minimum of $O(h^2)$ piecewise quadratic functions of ϕ requires $O(h^2)$ operations by Observation 1.

To summarize, we need $O(h^2)$ operations to compute $C'(\phi)$.

5. To carry out Step 5, we first compute $\min\{C(\phi), C'(\phi)\}$, the minimum of two piecewise quadratic functions of ϕ only, which takes $O(1)$ operations by Observation 1. In $O(1)$ operations, we can obtain $\mathcal{S}$ in (2.13) by computing the set of ϕ such that $\min\{C(\phi), C'(\phi)\} = C(\phi)$.

To summarize, computing $\mathcal{S}$ defined in (2.13) using Algorithm 2 takes $O(h^2)$ operations.

A.9 Empirical timing results for Proposition 5

In this section, we investigate the claim from Proposition 5 that computing the set $\mathcal{S}$ defined in (2.13) requires $O(h^2)$ operations, where h is the window size that appears in (2.7).

Figure A.2 displays the running time, computed on a MacBook Pro with a 1.4 GHz Intel Core i5 processor, as a function of the window size, h, over 50 replicate datasets simulated according to (2.1) with $T = 10,000$, $\gamma = 0.98$, and $z_t \overset{\text{i.i.d.}}{\sim} \text{Poisson}(0.01)$; the tuning parameter λ for the ℓ_0 problem in (2.14) is set to 0.3, which yields between 50 and 100 spikes. With $h = 20$, the average running time is 2.1 seconds for each dataset. In addition, a quadratic fit is plotted for reference. We see that the running time is indeed approximately quadratic in the window size h.

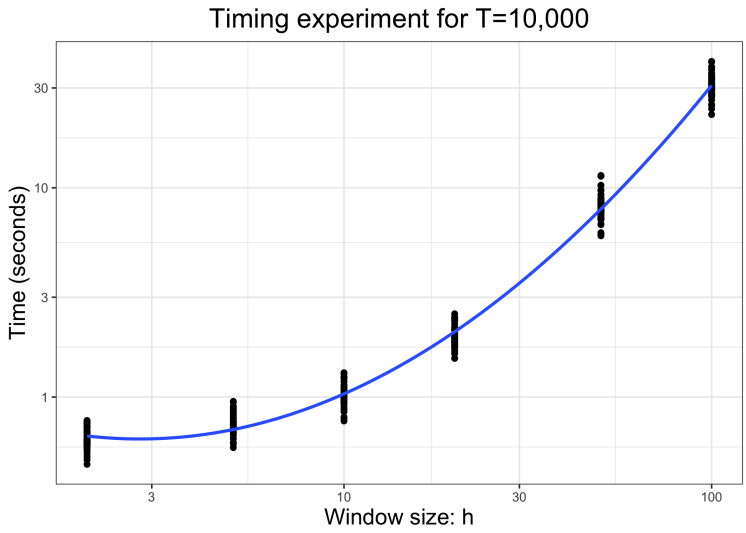


Figure A.2: Running time of Algorithm 2 over 50 replicate datasets, as a function of the window size, h. Each point represents a separate dataset. Each dataset is simulated according to (2.1), and the ℓ_0 problem is solved with $\lambda = 0.3$. A quadratic equation (Time $= 0.003h^2 - 0.002h + 0.695$) is plotted for reference.

A.10 An illustrative example for Propositions 4 and 18

In this section, we walk through a very simple example of characterizing the set $\mathcal{S} = \{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}$ in (2.13) using Proposition 3.

Suppose $y_{1:4} = (8, 4, 6, 3)$, and we want to compute $\mathcal{S}$ for $\hat{\tau}_j = 2$ with $h = 1$ (i.e., $\hat{\tau}_j - h = 1, \hat{\tau}_j + h = 3$), $\gamma = \frac{1}{2}$, and $\lambda = 1$. We first compute ν according to (2.7) and $y'(\phi)$ according to (2.11):

$$\nu = \begin{pmatrix} 0 \\ -0.5 \\ 1 \\ 0 \end{pmatrix}, \quad y'(\phi) = \begin{pmatrix} 8 \\ 5.6 - 0.4\phi \\ 2.8 + 0.8\phi \\ 3 \end{pmatrix}.$$

According to (2.20), to compute $\mathcal{S}$, it suffices to compute $C(\phi)$ in (2.18) and $C'(\phi)$ in (2.19).

We first compute $C(\phi)$ using Proposition 4. We start with $\mathcal{C}_{\hat{\tau}_j - h} = \mathcal{C}_1$.

1. $\mathcal{C}_1$ has only one function

$$\mathcal{C}_1 = \text{Cost}(y_1'(\phi), \alpha; \gamma) = \frac{1}{2}(8 - \alpha)^2.$$

2. To compute $\mathcal{C}_2$, we apply (2.23):

$$\mathcal{C}_2 = \Big\{\frac{1}{2}(8 - \alpha/0.5)^2 + \frac{1}{2}(5.6 - 0.4\phi - \alpha)^2, \\ \frac{1}{2}(5.6 - 0.4\phi - \alpha)^2 + g_2(\phi)\Big\},$$

where

$$g_2(\phi) = \min_{\alpha \geq 0} \text{Cost}(y_1, \alpha; \gamma) + \lambda = 0 + \lambda = 1.$$

This completes the calculation

$$\text{Cost}\Big(y'_{1:\hat{\tau}_j}(\phi), \alpha; \gamma\Big) = \text{Cost}(y'_{1:2}(\phi), \alpha; \gamma) = \min_{f \in \mathcal{C}_2} f(\alpha, \phi).$$

For the reverse direction, we will apply Proposition 18 to compute sets $\mathcal{C}_4$ and $\mathcal{C}_3$.

1. $\mathcal{C}_4$ consists of a single function:

$$\mathcal{C}_4 = \text{Cost}(y_4'(\phi), \alpha; 1/\gamma) = \frac{1}{2}(3 - \alpha)^2.$$

2. Applying (A.21), we get

$$\mathcal{C}_3 = \min\Big\{\frac{1}{2}(3 - \alpha/2)^2 + \frac{1}{2}(2.8 + 0.8\phi - \alpha)^2, \\ \min_{\alpha' \geq 0}\Big\{\frac{1}{2}(3 - \alpha'/2)^2\Big\} + \lambda + \frac{1}{2}(2.8 + 0.8\phi - \alpha)^2\Big\},$$

which yields

$$\text{Cost}\Big(y'_{T:\hat{\tau}_j+1}(\phi),\alpha;1/\gamma\Big)=\text{Cost}(y'_{4:3}(\phi),\alpha;1/\gamma)=\min_{f\in\mathcal{C}_3}f(\alpha,\phi).$$

According to (2.18),

$$C(\phi)=\min_{\alpha\geq 0}\big\{\text{Cost}(y'_{1:2}(\phi),\alpha;\gamma)\big\}+\min_{\alpha\geq 0}\big\{\text{Cost}(y'_{4:3}(\phi),\alpha;1/\gamma)\big\}+\lambda,$$

where

$$\min_{\alpha\geq 0}\big\{\text{Cost}\big(y'_{1:2}(\phi),\alpha;\gamma\big)\big\}=\min\left\{\min_{\alpha\geq 0}\left\{\frac{1}{2}(8-\alpha/(0.5))^2+\frac{1}{2}(5.6-0.4\phi-\alpha)^2\right\},\min_{\alpha\geq 0}\left\{\frac{1}{2}(5.6-0.4\phi-\alpha)^2+1\right\}\right\}$$

$$=\min\left\{\tilde{f}_1(\phi)=\begin{cases}0.064\phi^2-0.512\phi+1.024 & \phi\leq 54\\ 0.08\phi^2-2.24\phi+47.68 & \phi>54\end{cases},\tilde{f}_2(\phi)=\begin{cases}1 & \phi\leq 14\\ 0.08\phi^2-2.24\phi+16.68 & \phi>14\end{cases}\right\}$$

$$=\begin{cases}1 & \phi\leq 0.047\\ 0.064\phi^2-0.512\phi+1.024 & 0.047<\phi\leq 7.953\\ 1 & 7.953<\phi\leq 14\\ 0.08\phi^2-2.24\phi+16.68 & \phi>14\end{cases},$$

and

$$\min_{\alpha\geq 0}\left\{\text{Cost}\left(y'_{4:3}(\phi),\alpha;1/\gamma\right)\right\}=\min\left\{\min_{\alpha\geq 0}\left\{\frac{1}{2}(3-\alpha/2)^2+\frac{1}{2}(2.8+0.8\phi-\alpha)^2\right\},\min_{\alpha\geq 0}\left\{\frac{1}{2}(2.8+0.8\phi-\alpha)^2+1\right\}\right\}$$

$$=\min\left\{\tilde{f}_1(\phi)=\begin{cases}0.32\phi^2+2.24\phi+8.42 & \phi<-5.375\\ 0.064\phi^2-0.512\phi+1.024 & \phi\geq -5.375\end{cases},\tilde{f}_2(\phi)=\begin{cases}0.32\phi^2+2.24\phi+4.92 & \phi<-3.5\\ 1 & \phi\geq -3.5\end{cases}\right\}$$

$$=\begin{cases}0.32\phi^2+2.24\phi+4.92 & \phi\leq -3.5\\ 1 & -3.5<\phi\leq 0.047\\ 0.064\phi^2-0.512\phi+1.024 & 0.047<\phi\leq 7.953\\ 1 & \phi>7.953\end{cases}.$$

Therefore,

$$C(\phi)=\begin{cases}0.32\phi^2+2.24\phi+5.92 & \phi\leq -3.5\\ 2 & -3.5<\phi\leq 0.047\\ 0.064\phi^2-0.512\phi+2.024 & 0.047<\phi\leq 7.953\\ 2 & 7.953<\phi\leq 14\\ 0.08\phi^2-2.24\phi+17.68 & \phi>14\end{cases}\cdot$$

Moreover, according to (2.19),

$$
\begin{aligned}
C'(\phi) &= \min_{\alpha\geq 0}\left\{\text{Cost}(y'_{1:2}(\phi), u/0.5; \gamma) + \text{Cost}(y'_{4:3}(\phi), \alpha; 1/\gamma)\right\} \\
&= \min\Bigg\{ \min_{\alpha\geq 0}\left\{\frac{1}{2}(8-\alpha/0.25)^2 + \frac{1}{2}(5.6-0.4\phi-\alpha/0.5)^2 + \frac{1}{2}(3-\alpha/2)^2 + \frac{1}{2}(2.8+0.8\phi-\alpha)^2\right\}, \\
&\quad \min_{\alpha\geq 0}\left\{\frac{1}{2}(8-\alpha/0.25)^2 + \frac{1}{2}(5.6-0.4\phi-\alpha/0.5)^2 + 1 + \frac{1}{2}(2.8+0.8\phi-\alpha)^2\right\}, \\
&\quad \min_{\alpha\geq 0}\left\{\frac{1}{2}(5.6-0.4\phi-\alpha/0.25)^2 + 1 + \frac{1}{2}(3-\alpha)^2 + \frac{1}{2}(2.8+0.8\phi-\alpha)^2\right\}, \\
&\quad \min_{\alpha\geq 0}\left\{\frac{1}{2}(5.6-0.4\phi-\alpha/0.25)^2 + 1 + 1 + \frac{1}{2}(2.8+0.8\phi-\alpha)^2\right\}\Bigg\} \\
&= 0.4\phi^2 + 2.
\end{aligned}
$$

Finally, to determine $\mathcal{S}$, we take the minimum of these two functions:

$$
\min\left\{C(\phi), C'(\phi)\right\} = \begin{cases}
0.32\phi^2 + 2.24\phi + 5.92 & \phi \leq -3.5 \;\text{ Minimizer: } C(\phi) \\
3 & -3.5 < \phi \leq -1.581 \;\text{ Minimizer: } C(\phi) \\
0.4\phi^2 + 2 & -1.581 \leq \phi < 0.837 \;\text{ Minimizer: } C'(\phi) \\
0.064\phi^2 - 0.512\phi + 3.024 & 0.837 < \phi \leq 7.953 \;\text{ Minimizer: } C(\phi) \\
3 & \phi \geq 7.953 < \phi \leq 14 \;\text{ Minimizer: } C(\phi) \\
30.08\phi^2 - 2.24\phi + 17.68 & \phi > 14 \;\text{ Minimizer: } C(\phi)
\end{cases}.
$$

According to (2.20), $\mathcal{S} = (-\infty, -1.581) \cup [0.837, \infty)$ for this example.

A.11 Proof of Proposition 6

We first present an auxiliary result.

Lemma 10 (Lemma A.2. in Kivaranovic and Leeb [2020]). *Let $F^{\mathcal{S}}_{\mu,\sigma^2}$ denote the cumulative distribution function for a normal random variable with mean μ and variance σ^2, truncated*

to the set $\mathcal{S} \subseteq \mathbb{R}$. *For each* $t \in \mathcal{S}$, $F^{\mathcal{S}}_{\mu,\sigma^2}(t)$ *is continuous and monotonically decreasing in* μ .

We now present the proof of Proposition 6.

According to Lemma 10, $F^{\mathcal{S}\cap(0,\infty)}_{\mu,\sigma^2\|\nu\|_2^2}(t)$ is a monotonically decreasing function of μ for each $t \in \mathcal{S} \cap (0,\infty)$. Since $\frac{\alpha}{2} < 1 - \frac{\alpha}{2}$, it follows that $\theta_L(t)$ and $\theta_U(t)$ defined in (2.27) are unique, and that $\theta_L(t) < \theta_U(t)$.

In addition, monotonicity implies that $\forall t \in \mathcal{S} \cap (0,\infty)$, (i) $\nu^\top c > \theta_L(t)$ if and only if $F^{\mathcal{S}\cap(0,\infty)}_{\nu^\top c,\sigma^2\|\nu\|_2^2}(t) < 1 - \alpha/2$; and (ii) $\nu^\top c < \theta_U(t)$ if and only if $F^{\mathcal{S}\cap(0,\infty)}_{\nu^\top c,\sigma^2\|\nu\|_2^2}(t) > \alpha/2$.

These two observations imply that

$$\left\{\nu^\top c : \nu^\top c \in [\theta_L(t), \theta_U(t)]\right\} = \left\{\nu^\top c : \frac{\alpha}{2} \leq F^{\mathcal{S}\cap(0,\infty)}_{\nu^\top c,\sigma^2\|\nu\|_2^2}(t) \leq 1 - \frac{\alpha}{2}\right\}, \forall t \in \mathcal{S} \cap (0,\infty). \tag{A.25}$$

Recall that $Y \sim \mathcal{N}(c, \sigma^2 I)$. This implies that

$$\begin{aligned}
&\mathbb{P}\Big(\nu^\top c \in \left[\theta_L(\nu^\top Y), \theta_U(\nu^\top Y)\right] \;\Big|\; \hat{\tau}_j \in \mathcal{M}(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y, \nu^\top Y > 0\Big) \\
&\overset{a.}{=} \mathbb{P}\Big(\frac{\alpha}{2} \leq F^{\mathcal{S}\cap(0,\infty)}_{\nu^\top c,\sigma^2\|\nu\|_2^2}(\nu^\top Y) \leq 1 - \frac{\alpha}{2} \;\Big|\; \hat{\tau}_j \in \mathcal{M}(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y, \nu^\top Y > 0\Big) \\
&\overset{b.}{=} \mathbb{P}\Big(F^{\mathcal{S}\cap(0,\infty)}_{\nu^\top c,\sigma^2\|\nu\|_2^2}(Z) \in \Big[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\Big]\Big) \\
&\overset{c.}{=} 1 - \alpha.
\end{aligned}$$

To prove *a.*, we note that (A.25) holds for all $t \in \mathcal{S} \cap (0,\infty)$; therefore it holds for $\nu^\top Y$ conditioning on $\left\{\hat{\tau}_j \in \mathcal{M}(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y, \nu^\top Y > 0\right\}$ as well. Step *b.* follows from Proposition 9 and letting Z denote a normal random variable with mean $\nu^\top c$ and variance $\sigma^2||\nu||_2^2$, truncated to the set $\mathcal{S} \cap (0,\infty)$. The last step follows from the probability integral transform, which states that for a continuous random variable X, $F_X(X)$ is distributed as the Uniform(0,1) distribution.

A.12 Additional information for data analysis in Section 2.6

Data for the `spikefinder` challenge are available for download at
`https://s3.amazonaws.com/neuro.datasets/challenges/spikefinder/spikefinder.train.zip`.

In what follows, we reproduce Figure 2.6 with different choices of h (defined in (2.7)). In Figures A.3 and A.4, we compare the accuracy — as measured by the Victor-Purpura distance and correlation — of the spikes estimated via (2.37) (in orange), as well as the subset of spikes estimated via (2.37) for which the p-value is below 0.05 (in blue). The black lines indicate the 2.5% and 97.5% quantiles of the accuracy measures obtained over 1,000 resampled datasets, where each resampled dataset contains a subset of the estimated spikes from (2.37); details are as in Section 2.6.3. The results using $h = 5$ and $h = 50$ are quite similar to those with $h = 20$ (see Figure 2.6): the subset of spikes estimated via (2.37) for which the p-value is below 0.05 is the most accurate in almost every recording.

In addition, we performed simple diagnostics of the normality assumption of the error terms in (2.1). In Figure A.5, we plot the residuals $(y_t - \hat{c}_t)$ for recordings from the Chen et al. [2013] dataset; for most recordings, residuals appear approximately normal.

A.13 Estimation of the error variance σ^2 in (2.1)

Throughout the paper, we have assumed that σ^2 in (2.1) is known. However, if it is unknown, we propose to use $\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(y_t - \hat{c}_t)^2$ as an estimator for σ^2 in evaluating the p-value in (2.9). In Figure A.6, we present the results of a simulation study using the estimator $\hat{\sigma}^2$ and demonstrate that it leads to (i) adequate selective Type I error control under the global null (see Figure A.6(a)), (ii) substantial power under the alternative (see Figure A.6(b)), and (iii) correct selective coverage of the parameter $\nu^\top c$ (see Figure A.6(c)).

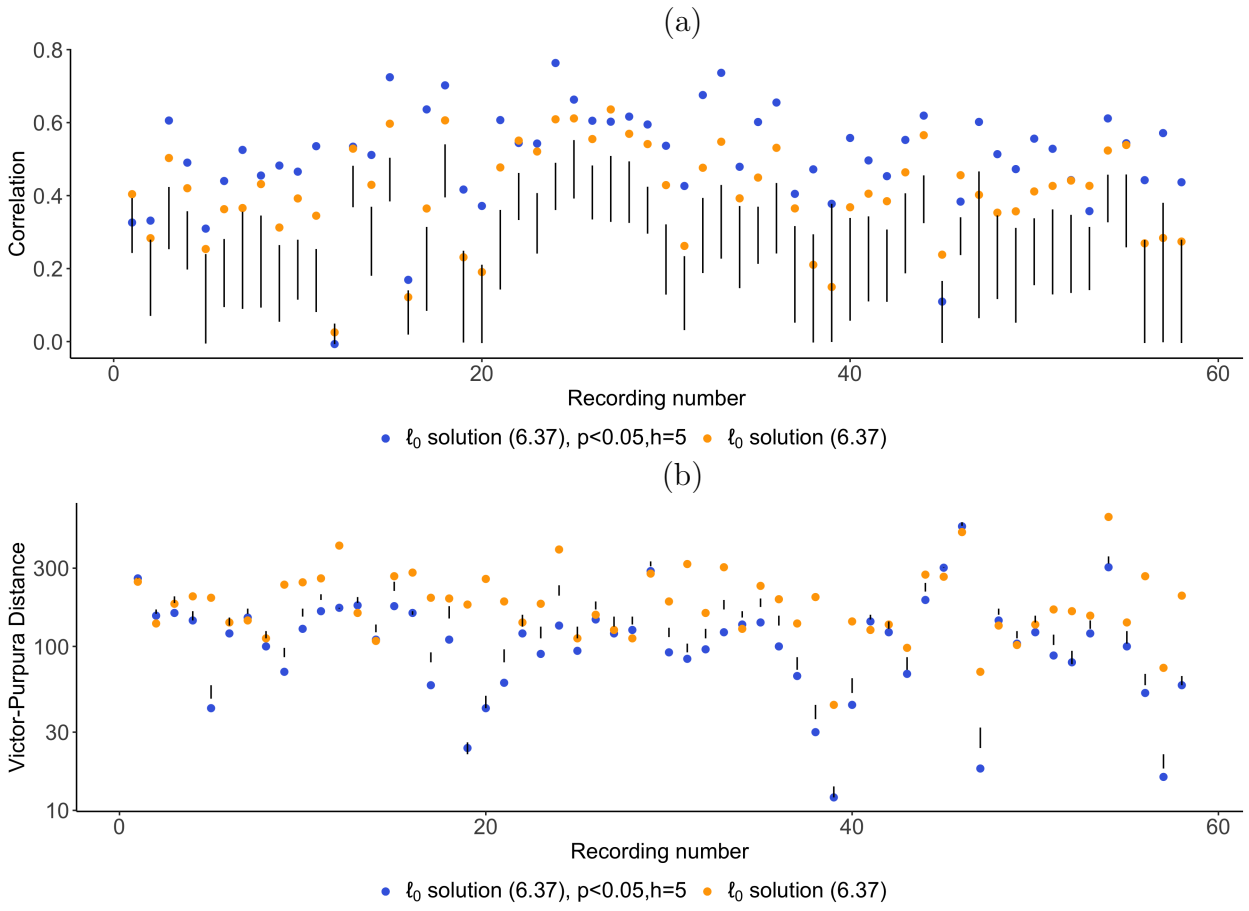


Figure A.3: Results for the Chen et al. [2013] dataset. Details are as in Figure 2.6 but with $h = 5$.

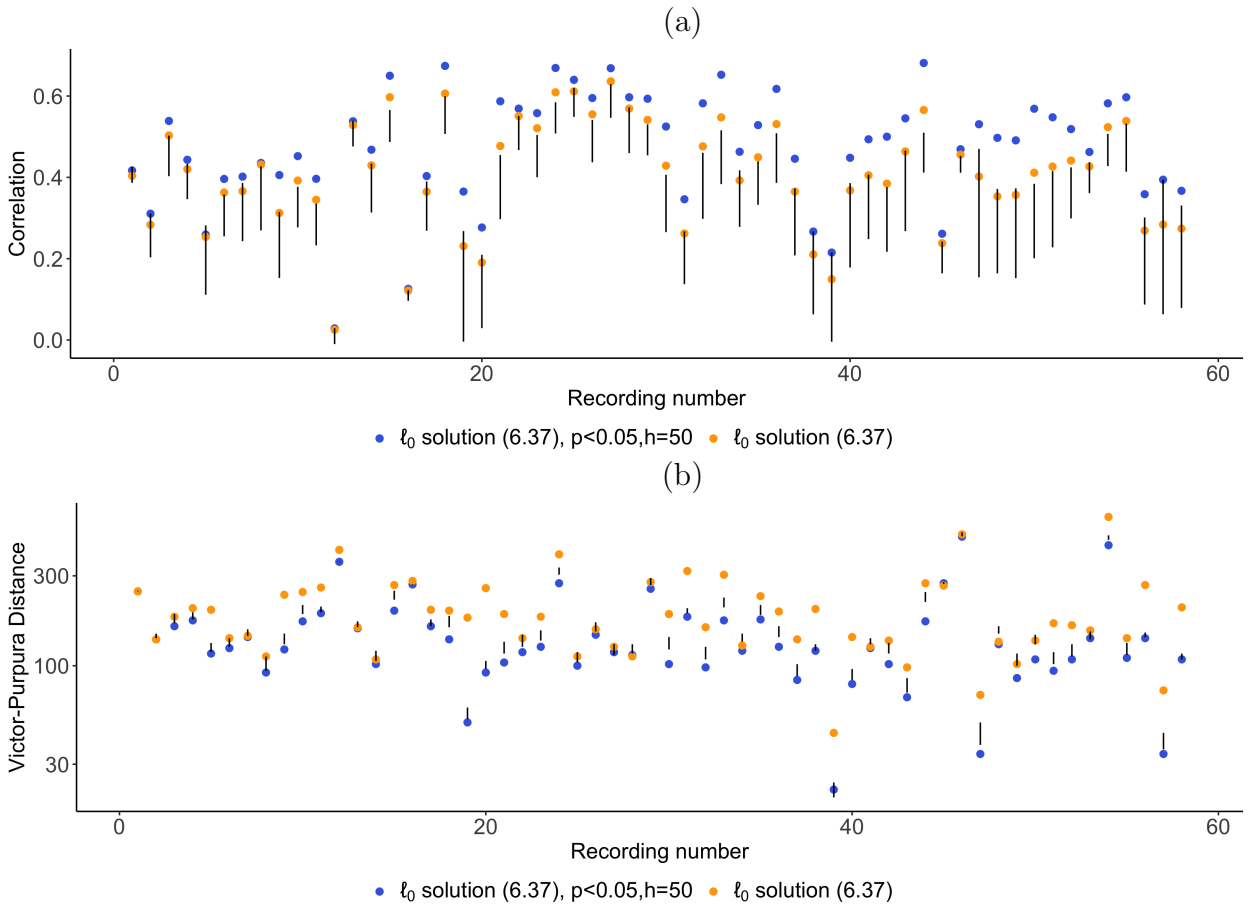


Figure A.4: Results for the Chen et al. [2013] dataset. Details are as in Figure 2.6 but with $h = 50$.

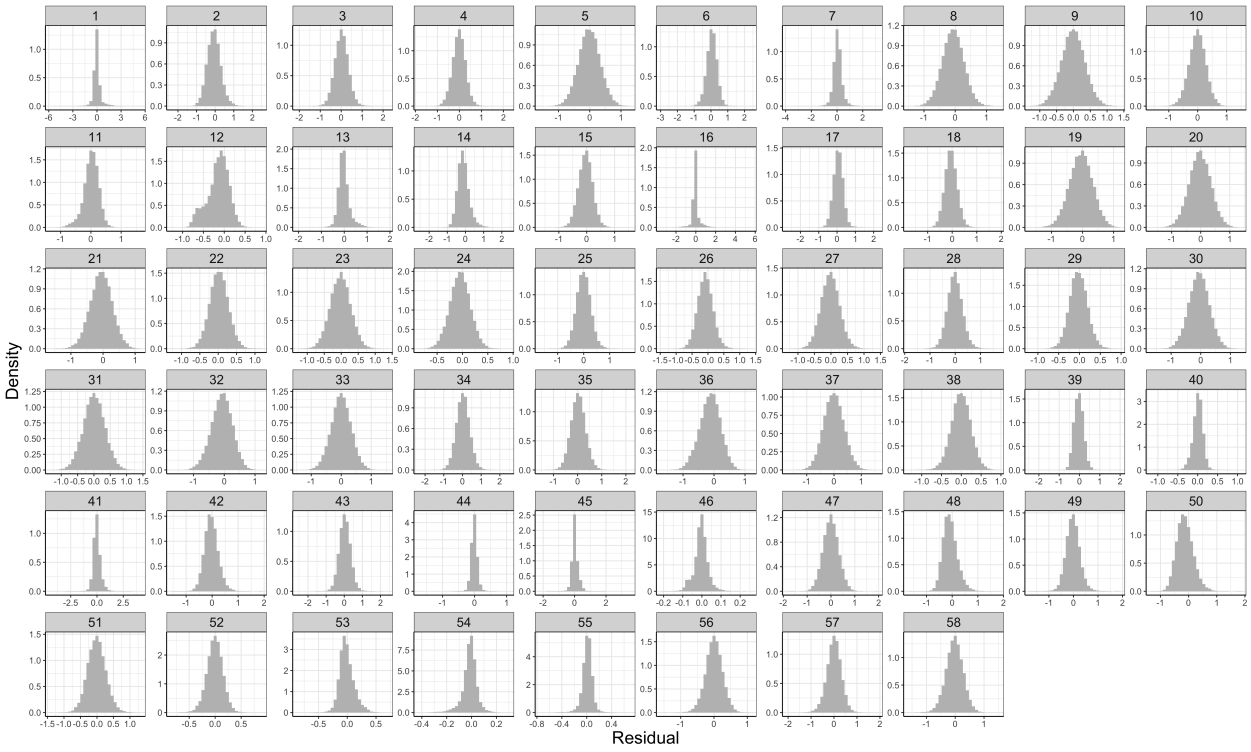


Figure A.5: Residuals, $y_t - \hat{c}_t$, for recordings from the Chen et al. [2013] dataset, where $\hat{c}_t$ is the solution to (2.37).

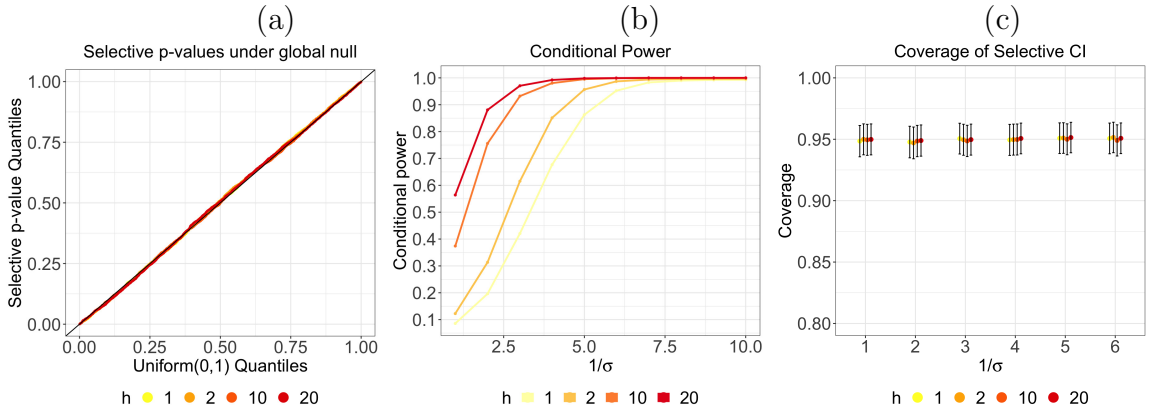


Figure A.6: *(a):* Quantile-quantile plot for selective p-values computed using estimated variance $\hat{\sigma}^2$ based on 100 simulations (2,988 hypothesis tests) under the global null. *(b):* Conditional power for selective p-values with estimated variance $\hat{\sigma}^2$. *(c):* Selective confidence intervals computed using estimated variance $\hat{\sigma}^2$ achieve correct nominal coverage (95% coverage at level $\alpha = 0.05$) across all values of h and σ.

Appendix B

B.1 Dual path algorithm for (3.2) with $X = I$ [Tibshirani and Taylor, 2011]

Algorithm 3 details the dual path algorithm for computing the generalized lasso problem (3.2) with an identity design matrix [Tibshirani and Taylor, 2011]. For a matrix $A \in \mathbb{R}^{m \times m}$, we use $A^{\dagger}$ to denote its Moore-Penrose inverse; for a vector $a \in \mathbb{R}^n$, we use $\text{diag}(a)$ to denote the $n \times n$ diagonal matrix with elements of a on the diagonal.

Algorithm 3: Dual path algorithm for the generalized lasso problem (3.2) when $X = I$.

Input: Input data $y \in \mathbb{R}^n$, Penalty matrix $D \in \mathbb{R}^{m\times n}$, Number of steps K

Output: Solution $\hat{\beta}$

Initialize $k = 0$, $\lambda_0 = \infty$, $B_0 = \emptyset$, $s_{B_0} = \emptyset$;

1. Compute $\hat{u} = \big(DD^\top\big)^\dagger y$.
2. Compute $\lambda_1 = \max_{i\in[m]} |\hat{u}_i|, i^* = \operatorname{argmax}_{i\in[m]} |\hat{u}_i|$.
3. Update $B_{k+1} \leftarrow B_k \cup i^*, s_{B_{k+1}} \leftarrow s_{B_k} \cup \operatorname{sign}(\hat{u}_{i^*})$.
4. Record the solution $\forall \lambda \in [\lambda_1, \infty]$:
$$\hat{u}(\lambda) = \hat{u}, \quad \hat{\beta}(\lambda) = y - D^\top \hat{u}(\lambda).$$
5. $k \leftarrow k + 1$

while $\lambda_k > 0$ *and* $k \le K$ **do**

1. Compute
$$a = (D_{-B_k} D_{-B_k}^\top)^\dagger D_{-B_k} y, \quad b = (D_{-B_k} D_{-B_k}^\top)^\dagger D_{-B_k} D_{B_k}^\top s_{B_k}.$$
2. Compute hitting times $t_i^{(\text{hit})} = \max\left\{\frac{a_i}{b_i+1}, \frac{a_i}{b_i-1}\right\}, \forall i \in [m] \setminus B_k$.
3. Set $h_{k+1} = \max_{i\notin B_k} t_i^{(\text{hit})}$ and $i^* = \operatorname{argmax}_{i\notin B_k} t_i^{(\text{hit})}$.
4. Compute
$$c = \operatorname{diag}(s_{B_k}) D_{B_k}\big(y - D_{-B_k}^\top a\big), \quad d = \operatorname{diag}(s_{B_k}) D_{B_k}\big(D_{B_k}^\top s_{B_k} - D_{-B_k}^\top b\big).$$
5. Compute leaving times $t_i^{(\text{leave})} = \frac{c_i}{d_i} \cdot 1\{c_i < 0\} \cdot 1\{d_i < 0\}, \forall i \in B_k$.
6. Set $l_{k+1} = \max_{i\in B_k} t_i^{(\text{leave})}$and $i^\diamond = \operatorname{argmax}_{i\in B_k} t_i^{(\text{leave})}$.
7. Set $\lambda_{k+1} = \max\{h_{k+1}, l_{k+1}\}$.
8. **if** $h_{k+1} \ge l_{k+1}$ **then**
$$B_{k+1} \leftarrow B_k \cup i^*, \quad s_{B_{k+1}} \leftarrow s_{B_k} \cup \operatorname{sign}(t_{i^*}^{(\text{hit})})$$
else
$$B_{k+1} \leftarrow B_k \setminus i^\diamond, \quad s_{B_{k+1}} \leftarrow s_{B_k} \setminus \operatorname{sign}(t_{i^\diamond}^{(\text{leave})})$$
end
9. Record the solution: $\forall \lambda \in [\lambda_{k+1}, \lambda_k]$:
$$\hat{u}(\lambda) = a - \lambda b, \quad \hat{\beta}(\lambda) = y - D^\top \hat{u}(\lambda).$$
10. $k \leftarrow k + 1$

end

B.2 Proof of Proposition 7

The proof of Proposition 7 is similar to the arguments in Section 6.2 of Tibshirani and Taylor [2011].

We first prove the "if" direction:

$$j, j' \in C_l \text{ for some } l \in [L] \implies \hat{\beta}_j = \hat{\beta}_{j'}. \tag{B.1}$$

First recall the result in (3.8), which states that $\hat{\beta} = P_{\text{Null}(D_{-B_k})}\big(y - \lambda D_{B_k}^\top s_{B_k}\big)$. It follows that $\hat{\beta}_j = \left(\left[P_{\text{Null}(D_{-B_k})}\right]_j\right)^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big)$ and $\hat{\beta}_{j'} = \left(\left[P_{\text{Null}(D_{-B_k})}\right]_{j'}\right)^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big)$, where $\left[P_{\text{Null}(D_{-B_k})}\right]_j$ denotes the jth row of the matrix $P_{\text{Null}(D_{-B_k})}$, written as a column vector. Therefore, to prove (B.1), it suffices to prove that

$$j, j' \in C_l \text{ for some } l \in [L] \implies \left[P_{\text{Null}(D_{-B_k})}\right]_j = \left[P_{\text{Null}(D_{-B_k})}\right]_{j'}. \tag{B.2}$$

To prove (B.2), we first compute $P_{\text{Null}(D_{-B_k})}$. The null space $\text{Null}(D_{-B_k})$ has dimension L and admits the basis $\{\mathbf{1}_{C_1}, \ldots, \mathbf{1}_{C_L}\}$ [Tibshirani and Taylor, 2011], where the jth element of $\mathbf{1}_{C_l}$ equals $1\{\text{Node } j \in C_l\}, \quad j = 1, \ldots, n$. Therefore, denoting $|C_l|$ the cardinality of the set C_l, we have that

$$P_{\text{Null}(D_{-B_k})} = \sum_{l=1}^{L} \frac{\mathbf{1}_{C_l}\mathbf{1}_{C_l}^\top}{|C_l|}.$$

It follows from algebra that

$$\left[P_{\text{Null}(D_{-B_k})}\right]_j = \sum_{l=1}^{L} \frac{[\mathbf{1}_{C_l}]_j}{|C_l|}\mathbf{1}_{C_l}, \quad \left[P_{\text{Null}(D_{-B_k})}\right]_{j'} = \sum_{l=1}^{L} \frac{[\mathbf{1}_{C_l}]_{j'}}{|C_l|}\mathbf{1}_{C_l}. \tag{B.3}$$

Now, according to the definition of $\mathbf{1}_{C_l}$ and the assumption that j and j' are in the same connected component, we have $[\mathbf{1}_{C_l}]_j = [\mathbf{1}_{C_l}]_{j'}, \forall l \in [L]$, which implies that $\left[P_{\text{Null}(D_{-B_k})}\right]_j = \left[P_{\text{Null}(D_{-B_k})}\right]_{j'}$ and therefore completes the proof for (B.2).

Next, we will prove the "only if" direction: that with probability one,

$$\hat{\beta}_j = \hat{\beta}_{j'} \implies j, j' \in C_l \text{ for some } l \in [L]. \tag{B.4}$$

Combining the results in (3.8) and (B.3), we have that

$$\hat{\beta}_j = \sum_{l=1}^{L} \frac{[\mathbf{1}_{C_l}]_j}{|C_l|} (\mathbf{1}_{C_l})^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big), \quad \hat{\beta}_{j'} = \sum_{l=1}^{L} \frac{[\mathbf{1}_{C_l}]_{j'}}{|C_l|} (\mathbf{1}_{C_l})^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big). \tag{B.5}$$

We proceed to prove (B.4) by contradiction. Suppose without loss of generality that $j \in C_1, j' \in C_2$ and $C_1 \cap C_2 = \emptyset$, we will show that, with probability one, $\hat{\beta}_j \neq \hat{\beta}_{j'}$. Combining our assumption and (B.5), we have that

$$\begin{aligned} \hat{\beta}_j - \hat{\beta}_{j'} &= \frac{1}{|C_1|} (\mathbf{1}_{C_1})^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big) - \frac{1}{|C_2|} (\mathbf{1}_{C_2})^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big) \\ &= \left(\frac{\mathbf{1}_{C_1}}{|C_1|} - \frac{\mathbf{1}_{C_2}}{|C_2|}\right)^\top \big(y - \lambda D_{B_k}^\top s_{B_k}\big). \end{aligned} \tag{B.6}$$

By our assumption, $\mathbf{1}_{C_1}/|C_1| - \mathbf{1}_{C_2}/|C_2|$ is a non-zero vector. In addition, it follows from algebra that entries of the vector $\lambda D_{B_k}^\top s_{B_k}$ can only take values that are integer multiples of λ. Under model (3.1), $y \sim \mathcal{N}(0, \sigma^2 I_n)$, which implies that the inner product in (B.6) will be non-zero with probability one. Therefore, we have proven that with probability one, (B.4) holds.

B.3 Algorithm for computing $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ in (3.14)

We now present an algorithm to compute $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ (3.14).

In Algorithm 4, we initialize with $\eta = 10^{-4}$, and apply Corollary 1 to obtain the set $[\tilde{a}_1, a_2] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_1 + \eta))\}\right\}$ (see step 4(a) in Algorithm 4 with $i = 1$). If the left endpoint $\tilde{a}_1 > a_1 + \epsilon$ (where ϵ is a small constant set to 1.5×10^{-8} by default), we repeat with a smaller value of η (see step 4(b)i of Algorithm 4).

In a simulation study, we investigate how often, using the initialization $\eta = 10^{-4}$, we need

Algorithm 4: Characterizing the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ in (3.14)

Input : Data y, $\hat{C}_1$, $\hat{C}_2$, number of steps K for Algorithm 3, $\eta = 10^{-4}$
Output: $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ in (3.14)

1. $\mathcal{J} \leftarrow \{0\}$, $i \leftarrow 1$, $j \leftarrow 0$.
2. $[a_0, a_1] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y)\}\right\}$, where M_k was defined in (3.9).
3. **while** $a_i < \infty$ **do**
   (a) Compute $[\tilde{a}_i, a_{i+1}] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_i + \eta))\}\right\}$.
   (b) **while** $\tilde{a}_i \neq a_i$ **do**
      i. $\eta \leftarrow \frac{1}{2} \cdot \eta$,
      ii. $[\tilde{a}_i, a_{i+1}] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_i + \eta))\}\right\}$.
   end
   (c) **if** $\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(a_i + \eta))$ **then**
      $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$.
   end
   (d) $i \leftarrow i + 1$.
   end
4. **while** $a_j > -\infty$ **do**
   (a) $[a_{j-1}, \tilde{a}_j] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_j - \eta))\}\right\}$.
   (b) **while** $\tilde{a}_j \neq a_j$ **do**
      i. $\eta \leftarrow \frac{1}{2} \cdot \eta$,
      ii. $[a_{j-1}, \tilde{a}_j] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_j - \eta))\}\right\}$.
   end
   (c) **if** $\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(a_j - \eta))$ **then**
      $\mathcal{J} \leftarrow \mathcal{J} \cup \{j - 1\}$.
   end
   (d) $j \leftarrow j - 1$.
   end

$\mathcal{S}_{\hat{C}_1,\hat{C}_2} \leftarrow \bigcup_{i \in \mathcal{J}} [a_i, a_{i+1}]$.

to perform the halving operation in step 4(b)i of Algorithm 4. Results are aggregated in Figure B.1. We almost never have to halve the initial value 10^{-4}, and the number of halving operations never exceeds seven.

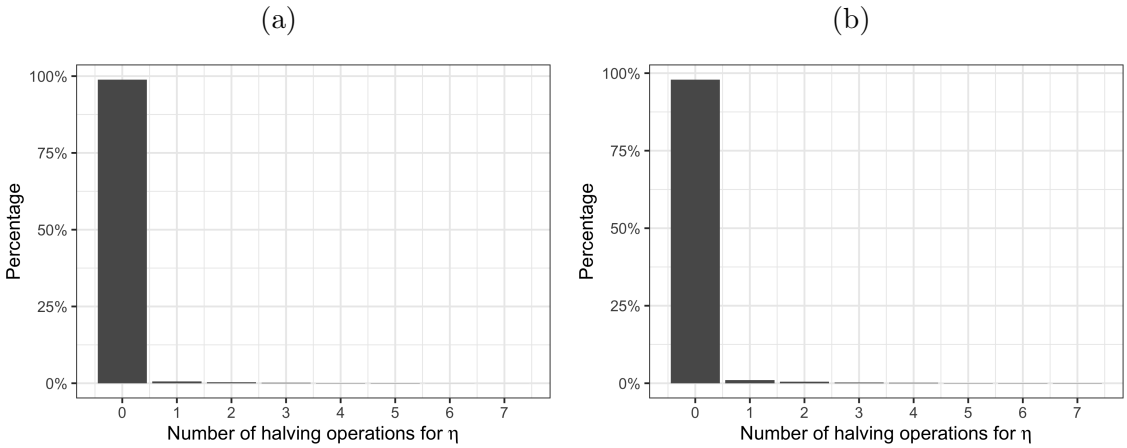


Figure B.1: *(a):* Number of halving operations for η (defined in Section 3.3.3 and Algorithm 4 of Appendix B.3) in the one-dimensional fused lasso simulations described in Section 3.5.1 with $\sigma = 1$. *(b):* Same as (a), but for the two-dimensional fused lasso simulations described in Section 3.5.2 with $\sigma = 1$.

B.4 Proof of Proposition 9

We first prove the statement (3.13). The following equalities hold for an arbitrary vector $\nu \in \mathbb{R}^n$:

$$\begin{aligned}
&\mathbb{P}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\
&\overset{a.}{=} \mathbb{P}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(\Pi_\nu^\perp y + \Pi_\nu Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\
&\overset{b.}{=} \mathbb{P}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\
&\overset{c.}{=} \mathbb{P}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right)\right).
\end{aligned} \tag{B.7}$$

Here, $a.$ follows from the fact that $Y = \Pi_\nu^\perp Y + \Pi_\nu Y$, and the fact that we have conditioned on the event $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$. To prove $b.$, we first note that $I = \Pi_\nu + \Pi_\nu^\perp$ and $\Pi_\nu Y = \frac{\nu\nu^\top}{||\nu||_2^2} Y$, which implies

$$\Pi_\nu^\perp y + \Pi_\nu Y = y - \Pi_\nu y + \Pi_\nu Y = y - \frac{\nu^\top y}{||\nu||_2^2}\nu + \frac{\nu^\top Y}{||\nu||_2^2}\nu = y'(\nu^\top Y),$$

where the notation $y'(\cdot)$ is defined in (3.12). Finally, $c.$ follows from the fact that $Y \sim \mathcal{N}(\beta, \sigma^2 I)$ implies independence of $\nu^\top Y$ and $\Pi_\nu^\perp Y$.

Now under H_0 in (3.4), we have that

$$\begin{aligned} p_{\hat{C}_1,\hat{C}_2} &\overset{a.}{=} \mathbb{P}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\ &\overset{b.}{=} \mathbb{P}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right)\right) \\ &\overset{c.}{=} \mathbb{P}\left(|\phi| \geq |\nu^\top y| \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'(\phi)\right)\right). \end{aligned}$$

Here, step $a.$ is the definition of $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), and step $b.$ follows from applying the result in (B.7). In $c.$, we used the fact that under H_0, $\nu^\top Y \sim \mathcal{N}(0, \sigma^2||\nu||_2^2) \overset{d}{=} \phi$, which completes the proof.

Next, we will prove that the test that rejects $H_0 : \nu^\top \beta = 0$ when $p_{\hat{C}_1,\hat{C}_2} \leq \alpha$ controls the selective Type I error as in (3.6). First of all, a test for the null hypothesis in (3.4) controls the selective Type I error in (3.6) if

$$\mathbb{P}_{H_0}\left(\text{reject } H_0 \text{ at level } \alpha \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(Y\right)\right) \leq \alpha, \;\; \forall \alpha \in (0,1). \tag{B.8}$$

In what follows, we will write $p_{\hat{C}_1,\hat{C}_2}$ as $p_{\hat{C}_1,\hat{C}_2}\left(\nu^\top y\right)$ to emphasize that $p_{\hat{C}_1,\hat{C}_2}$ is a function of the observed difference in means $\nu^\top y$. Moreover, using the result in (3.13), we have that

$$p_{\hat{C}_1,\hat{C}_2}\left(\nu^\top y\right) = 1 - \tilde{F}^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{0,\sigma^2||\nu||_2^2}\left(|\nu^\top y|\right), \tag{B.9}$$

where $\tilde{F}^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{0,\sigma^2||\nu||_2^2}(\cdot)$ is the cumulative distribution function of *the magnitude* of a $\mathcal{N}(0, \sigma^2||\nu||_2^2)$

random variable, truncated to the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$. For all $\alpha \in (0,1)$, we have that

$$
\begin{aligned}
&\mathbb{P}_{H_0}\Big(p_{\hat{C}_1,\hat{C}_2}(\nu^\top Y) \leq \alpha \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\Big) \\
&\quad \overset{a.}{=} \mathbb{P}\left(p_{\hat{C}_1,\hat{C}_2}(\nu^\top Y) \leq \alpha \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(\Pi_\nu^\perp y + \Pi_\nu Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\
&\quad \overset{b.}{=} \mathbb{P}_{H_0}\left(p_{\hat{C}_1,\hat{C}_2}(\nu^\top Y) \leq \alpha \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right) \\
&\quad \overset{c.}{=} \mathbb{P}_{H_0}\left(p_{\hat{C}_1,\hat{C}_2}(\nu^\top Y) \leq \alpha \;\middle|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right)\right) \\
&\quad \overset{d.}{=} \mathbb{P}_{H_0}\Big(1 - \tilde{F}^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{0,\sigma^2||\nu||_2^2}\left(|\nu^\top Y|\right) \leq \alpha \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right)\Big) \\
&\quad \overset{e.}{=} \mathbb{P}_Z\Big(\tilde{F}^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{0,\sigma^2||\nu||_2^2}(Z) \geq 1-\alpha\Big) \\
&\quad \overset{f.}{=} \alpha.
\end{aligned}
\tag{B.10}
$$

Steps $a.$ through $c.$ follow from the same line of argument in (B.7). In step $d.$, we use the identity in (B.9). To prove step $e.$, we first note that, under H_0, the conditional cumulative distribution function of $|\nu^\top Y|$ given $\hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K\left(y'\left(\nu^\top Y\right)\right)$, is exactly $\tilde{F}^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{0,\sigma^2||\nu||_2^2}$; the equality follows by denoting Z the corresponding random variable. Finally, $f.$ is a direct consequence of the probability integral transform, which states that for a continuous random variable Z, $F_Z(Z)$ is distributed as the Uniform(0,1) distribution.

Now for the test that rejects H_0 if $p_{\hat{C}_1,\hat{C}_2} \leq \alpha$, we have that

$$
\begin{aligned}
&\mathbb{P}_{H_0}\Big(p_{\hat{C}_1,\hat{C}_2} \leq \alpha \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y)\Big) \\
&\quad = \mathbb{E}_{H_0}\Big[1_{(p_{\hat{C}_1,\hat{C}_2} \leq \alpha)} \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y)\Big] \\
&\quad \overset{a.}{=} \mathbb{E}_{H_0}\Big[\mathbb{E}_{H_0}\Big[1_{(p_{\hat{C}_1,\hat{C}_2} \leq \alpha)} \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\Big] \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y)\Big] \\
&\quad \overset{b.}{=} \mathbb{E}_{H_0}\Big[\alpha \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(Y)\Big] \\
&\quad = \alpha.
\end{aligned}
\tag{B.11}
$$

In the proof above, step $a.$ follows from the tower property of conditional expectation. To prove $b.$, we apply the results from (B.10).

By inspection of (B.8) and (B.11), we conclude that the test based on $p_{\hat{C}_1,\hat{C}_2}$ controls the selective Type I error, which completes the proof.

B.5 Proof of Corollary 1 and Proposition 10

We first prove Corollary 1: that is, $\left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y)\}\right\}$ is an interval, where M_k is defined in (3.9).

Proof.

$$\begin{aligned}
\left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y)\}\right\} &\overset{a.}{=} \{\phi \in \mathbb{R} : Ay'(\phi) \le 0\} \\
&\overset{b.}{=} \left\{\phi \in \mathbb{R} : A\left(y - \frac{\nu^\top y}{||\nu||_2^2}\nu + \frac{\phi}{||\nu||_2^2}\nu\right) \le 0\right\} \\
&\overset{c.}{=} \left\{\phi \in \mathbb{R} : \phi \cdot (A\nu) \le (\nu^\top y)A\nu - ||\nu||_2^2 Ay\right\} \\
&\overset{d.}{=} \left[\max_{i:(A\nu)_i<0} \frac{(\nu^\top y)(A\nu)_i - ||\nu||_2^2(Ay)_i}{(A\nu)_i}, \min_{i:(A\nu)_i>0} \frac{(\nu^\top y)(A\nu)_i - ||\nu||_2^2(Ay)_i}{(A\nu)_i}\right].
\end{aligned}$$

Here, $a.$ follows from Proposition 8, which states that the set $\left\{Y \in \mathbb{R}^n : \bigcap_{k=1}^{K} M_k(Y) = M_k(y)\right\} = \{Y \in \mathbb{R}^n : AY \le 0\}$ for some matrix A, where $\le$ is interpreted as the component-wise inequality. Next, $b.$ follows from the definition of $y'(\phi)$ in (3.12). Finally, $c.$ and $d.$ follow from solving the linear inequality in ϕ. Note that in $d.$, we implicitly assumed that $\exists i$ (or i') such that $(A\nu)_i < 0$ (or $(A\nu)_{i'} > 0$); if that doesn't hold, the corresponding expression in $d.$ is replaced by $-\infty$ (or $+\infty$).

□

We proceed to prove Proposition 10.

$$
\begin{aligned}
\mathcal{S}_{\hat{C}_1,\hat{C}_2} &\overset{a.}{=} \left\{\phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(\phi))\right\} \\
&\overset{b.}{=} \bigcup_{(m_1,\dots m_K)\in\mathcal{M}_K} \left\{\phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(\phi)), \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = m_k\}\right\} \\
&\overset{c.}{=} \bigcup_{(\tilde{m}_1,\dots \tilde{m}_K)\in\mathcal{I}} \left\{\phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(\phi)), \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = \tilde{m}_k\}\right\} \\
&\overset{d.}{=} \bigcup_{(\tilde{m}_1,\dots \tilde{m}_K)\in\mathcal{I}} \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = \tilde{m}_k\}\right\} \\
&\overset{e.}{=} \bigcup_{i\in\mathcal{J}} [a_i, a_{i+1}],
\end{aligned}
$$

where in $b.$, $\mathcal{M}_K$ is the set of all possible outputs of the first K steps of the dual path algorithm. In the proof above, $a.$ is the definition of $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$, and $b.$ follows from the definition of $\mathcal{M}_K$. Furthermore, steps $c.$ and $d.$ follows from the definition of the index set $\mathcal{I}$ (see (3.15)). Finally, to prove $e.$, we apply Corollary 1, which implies that for each $\tilde{m}_k$, $\left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = \tilde{m}_k\}\right\}$ is an interval. This concludes the proof of Proposition 10.

B.6 Proof and an empirical analysis of Proposition 11

The proof of Proposition 11 is similar to the proof of Proposition 4 in Jewell et al. [2022]. First, note that by the definition of $\tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}$ in Proposition 11, $\mathbb{P}\left(\phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right) > 0 \implies$

$\mathbb{P}\Big(\phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\Big) > 0$. Next, we have that

$$\begin{aligned}
\mathbb{P}\left(|\phi| \geq |\nu^\top y| \,\middle|\, \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\right) &\overset{a.}{=} \frac{\mathbb{P}\left(|\phi| \geq |\nu^\top y|, \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\right)}{\mathbb{P}\Big(\phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\Big)} \\
&\overset{b.}{=} \frac{\mathbb{P}\left(|\phi| \geq |\nu^\top y|, \phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right) + \mathbb{P}\left(|\phi| \geq |\nu^\top y|, \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \setminus \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right)}{\mathbb{P}\Big(\phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big) + \mathbb{P}\Big(\phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \setminus \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big)} \\
&\overset{c.}{=} \frac{\mathbb{P}\left(|\phi| \geq |\nu^\top y|, \phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right) + \mathbb{P}\left(\phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \setminus \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right)}{\mathbb{P}\Big(\phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big) + \mathbb{P}\Big(\phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \setminus \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big)} \\
&\geq \frac{\mathbb{P}\left(|\phi| \geq |\nu^\top y|, \phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right)}{\mathbb{P}\Big(\phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big)} \\
&= \mathbb{P}\left(|\phi| \geq |\nu^\top y| \,\middle|\, \phi \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}\right).
\end{aligned}$$

Here, $a.$ follows from Bayes' rule and the fact that $\mathbb{P}\Big(\phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\Big) > 0$, and $b.$ is a direct consequence of the law of total probability. Step $c.$ follows from the definition of $\tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}$, which implies that $\big\{\phi : |\phi| \geq |\nu^\top y|\big\} \cap \Big\{\phi : \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \setminus \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big\} = \Big\{\phi : \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2} \setminus \mathcal{S}_{\hat{C}_1,\hat{C}_2}\Big\}$.

Next, we investigate the approximation error of using $\tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}$ to compute the p-value instead of $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$. In what follows, we denote $\mathbb{P}\left(|\phi| \geq |\nu^\top y| \,\middle|\, \phi \in \tilde{\mathcal{S}}_{\hat{C}_1,\hat{C}_2}\right)$ as $p_{\hat{C}_1,\hat{C}_2}(\delta)$ for brevity. In prior work, several choices of δ have been proposed (e.g., $\max\big\{0, 10\sigma\|\nu\|_2 - |\nu^\top y|\big\}$ [Jewell et al., 2022] and $\max\big\{0, 20\sigma - |\nu^\top y|\big\}$ [Liu et al., 2018]). In this section, we computed $p_{\hat{C}_1,\hat{C}_2}(\delta)$ with $\delta = \max\{0, 10\sigma\|\nu|_2 - |\nu^\top y|\}$ for the experiments in the one-dimensional fused lasso case (Section 3.5.1 of the main text). Results are aggregated and displayed in Figure B.2. Panel (a) displays the p-values $p_{\hat{C}_1,\hat{C}_2}$ versus $p_{\hat{C}_1,\hat{C}_2}(\delta)$ with $\delta = \max\{0, 10\sigma\|\nu\|_2 - |\nu^\top y|\}$ on the $-\log_{10}$ scale under the global null. We see that the two set of p-values are nearly identical; the same holds for the datasets simulated with true changepoints (see Figure B.2(b)). Regarding computational efficiency, this choice of δ reduces the overall running time of Algorithm 2 by 15–20%, depending on the specific simulation parameters. In conclusion, we suggest using $\delta = \max\{0, 10\sigma\|\nu\|_2 - |\nu^\top y|\}$ to balance

the inferential accuracy and computational efficiency.

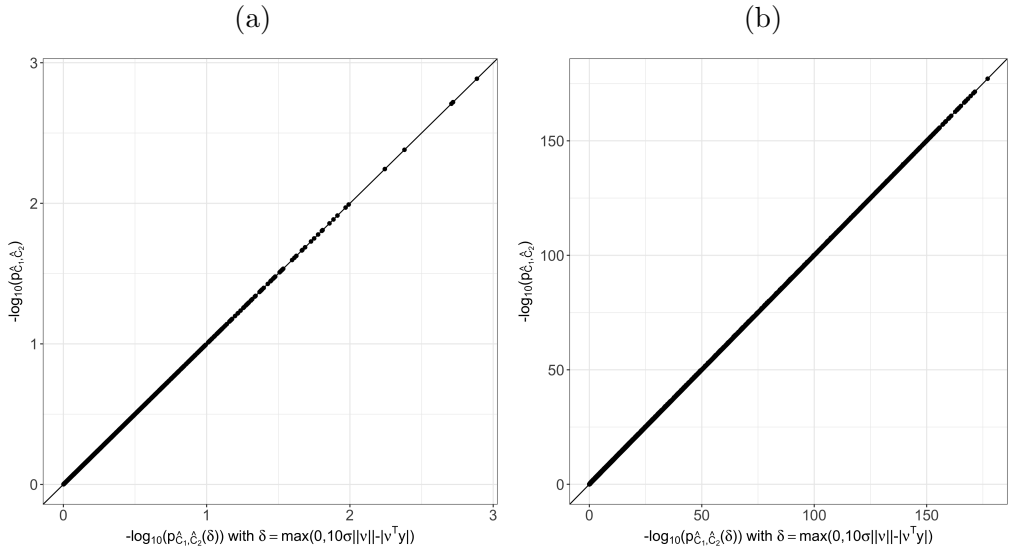


Figure B.2: *(a):* For the one-dimensional fused lasso simulations described in Section 3.5.1, the p-values $p_{\hat{C}_1,\hat{C}_2}$ and $p_{\hat{C}_1,\hat{C}_2}(\delta)$ with $\delta = \max\{0, 10\sigma\|\nu\|_2 - |\nu^\top y|\}$ on the $-\log_{10}$ scale. *(b):* Same as (a), but for the simulations described in Section 3.5.1.

B.7 *Proof of Proposition 12*

The proof of Proposition 12 is similar to the proof of Theorem 6.1 of Lee et al. [2016], the proof of Corollary 3.1 of Chen and Bien [2020], and the proof of Proposition B.7 in this dissertation.

Recalling the definition that $F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\mu,\sigma^2||\nu||_2^2}(t)$ is the cumulative distribution function of a $\mathcal{N}(\mu, \sigma^2||\nu||_2^2)$ random variable, truncated to the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ defined in (3.14), we have that $F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\mu,\sigma^2||\nu||_2^2}(t)$ is a monotonically decreasing function of μ for each $t \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}$ (see, e.g., Lemma A.2. of Kivaranovic and Leeb 2020). Since $\frac{\alpha}{2} < 1 - \frac{\alpha}{2}$, it follows that $\theta_l(t)$ and $\theta_u(t)$ defined in (3.18) are unique, and that $\theta_l(t) < \theta_u(t)$.

In addition, monotonicity implies that for all $t \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}$, (i) $\nu^\top\beta > \theta_l(t)$ if and only if $F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2||\nu||_2^2}(t) < 1-\alpha/2$; and (ii) $\nu^\top\beta < \theta_u(t)$ if and only if $F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2||\nu||_2^2}(t) > \alpha/2$. In other words, we have that

$$\left\{\nu^\top\beta : \nu^\top\beta \in [\theta_l(t), \theta_u(t)]\right\} = \left\{\nu^\top\beta : \frac{\alpha}{2} \leq F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2||\nu||_2^2}(t) \leq 1 - \frac{\alpha}{2}\right\}, \ \forall t \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}. \quad \text{(B.12)}$$

Now, under (3.1), $Y \sim \mathcal{N}(\beta, \sigma^2 I_n)$, which implies that

$$\begin{aligned}
&\mathbb{P}\Big(\nu^\top\beta \in \big[\theta_l(\nu^\top Y), \theta_u(\nu^\top Y)\big] \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big) \\
&\overset{a.}{=} \mathbb{P}\Big(\frac{\alpha}{2} \leq F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2\|\nu\|_2^2}(\nu^\top Y) \leq 1 - \frac{\alpha}{2} \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big) \\
&\overset{b.}{=} \mathbb{P}\Big(\frac{\alpha}{2} \leq F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2\|\nu\|_2^2}(\nu^\top Y) \leq 1 - \frac{\alpha}{2} \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(\nu^\top Y))\Big) \\
&\overset{c.}{=} \mathbb{P}\Big(F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2\|\nu\|_2^2}(Z) \in \Big[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\Big]\Big) \\
&\overset{d.}{=} 1 - \alpha.
\end{aligned}$$

In step *a.*, we use the fact that (B.12) holds for all $t \in \mathcal{S}_{\hat{C}_1,\hat{C}_2}$; therefore it holds for $\nu^\top Y$ conditioning on the event $\left\{\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y\right\}$ as well. Next, step *b.* follows from the definition of $y'(\phi)$ in (3.12), and the fact that under H_0 and (3.1), $\nu^\top Y$ is independent of $\Pi^\perp_\nu Y$. To prove *c.*, we first recall that $\nu^\top Y$, conditional on $\left\{\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_K(y'(\nu^\top Y))\right\}$, is a

$\mathcal{N}(\nu^\top\beta, \sigma^2||\nu||_2^2)$ random variable, truncated to the set $\mathcal{S}_{\hat{C}_1,\hat{C}_2}$ (see Appendix B.4). Moreover, letting Z denote that random variable completes the proof for c. The last step follows from combining the fact that the cumulative distribution of Z is $F^{\mathcal{S}_{\hat{C}_1,\hat{C}_2}}_{\nu^\top\beta,\sigma^2\|\nu\|_2^2}$ and the probability integral transform.

B.8 *Modification of Algorithm 4 to compute* $p^*_{\hat{C}_1,\hat{C}_2}$

For the graph fused lasso with an arbitrary graph, it may be the case that for two integers K, K', $|\mathcal{CC}_K(Y)| = |\mathcal{CC}_{K'}(Y)|$ and $\mathcal{CC}_K(Y) \neq \mathcal{CC}_{K'}(Y)$ [Tibshirani and Taylor, 2011]. In other words, $\mathcal{CC}(Y)$ in (3.20) need not to be unique. Therefore, we cannot directly apply the recipes in Section 3.3 to characterize $p^*_{\hat{C}_1,\hat{C}_2}$ in (3.20).

In what follows, we propose to characterize a variant of $p^*_{\hat{C}_1,\hat{C}_2}$ that makes use of the smallest value of K to yield exactly L connected components:

$$\tilde{p}^*_{\hat{C}_1,\hat{C}_2} \equiv \mathbb{P}_{H_0}\Big(|\nu^\top Y| \geq |\nu^\top y| \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_{\text{argmin}_k\{k:|\mathcal{CC}_k(Y)|=L\}}(Y), \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big). \tag{B.13}$$

We remark that the definitions of $\tilde{p}^*_{\hat{C}_1,\hat{C}_2}$ in (B.13) and $p^*_{\hat{C}_1,\hat{C}_2}$ in (3.20) coincide in the special case of the one-dimensional fused lasso, when the number of estimated connected components L is uniquely determined by the number of steps in the dual path algorithm K, and, as a result, $\mathcal{CC}(Y)$ is uniquely defined.

Using a similar argument to Proposition 9, we have that for $\phi \sim \mathcal{N}(0, \sigma^2||\nu||_2^2)$,

$$\tilde{p}^*_{\hat{C}_1,\hat{C}_2} = \mathbb{P}\Big(|\phi| \geq |\nu^\top y| \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_{\text{argmin}_k\{k:|\mathcal{CC}_k(y'(\phi))|=L\}}(y'(\phi))\Big). \tag{B.14}$$

Therefore, the key to computing $p^*_{\hat{C}_1,\hat{C}_2}$ in (3.20) is to characterize the set

$$\mathcal{S}^*_{\hat{C}_1,\hat{C}_2} \equiv \Big\{\phi : \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_{\text{argmin}_k\{k:|\mathcal{CC}_k(y'(\phi))|=L\}}(y'(\phi))\Big\}. \tag{B.15}$$

The algorithm for characterizing the set $\mathcal{S}^*_{\hat{C}_1,\hat{C}_2}$ in (B.15) requires only minor modifications to Algorithm 4. Details are provided in Algorithm 5.

Algorithm 5: Characterizing the set $\mathcal{S}^*_{\hat{C}_1,\hat{C}_2}$ in (B.15)

Input: Data y, $\hat{C}_1$, $\hat{C}_2$, number of connected components L, $\eta = 10^{-4}$

Output: $\mathcal{S}^*_{\hat{C}_1,\hat{C}_2}$ in (B.15)

1. $\mathcal{J} \leftarrow \{0\}$, $i \leftarrow 1$, $j \leftarrow 0$.
2. $[a_0, a_1] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K^*} \{M_k(y'(\phi)) = M_k(y)\}\right\}$, where $K^* = \text{argmin}_k \{|\mathcal{CC}_k(y)| = L\}$.
3. **while** $a_i < \infty$ **do**
   (a) Compute $[\tilde{a}_i, a_{i+1}] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K^*} \{M_k(y'(\phi)) = M_k(y'(a_i + \eta))\}\right\}$, where $K^* = \text{argmin}_k \{|\mathcal{CC}_k(y'(a_i + \eta))| = L\}$.
   (b) **while** $\tilde{a}_i \neq a_i$ **do**
      i. $\eta \leftarrow \frac{1}{2} \cdot \eta$,
      ii. $[\tilde{a}_i, a_{i+1}] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K^*} \{M_k(y'(\phi)) = M_k(y'(a_i + \eta))\}\right\}$, where $K^* = \text{argmin}_k \{|\mathcal{CC}_k(y'(a_i + \eta))| = L\}$
   end
   (c) **if** $\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_{K^*}(y'(a_i + \eta))$, **then** $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$.
   (d) $i \leftarrow i + 1$.
end
4. **while** $a_j > -\infty$ **do**
   (a) $[a_{j-1}, \tilde{a}_j] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K^*} \{M_k(y'(\phi)) = M_k(y'(a_j - \eta))\}\right\}$, where $K^* = \text{argmin}_k \{|\mathcal{CC}_k(y'(a_j - \eta))| = L\}$.
   (b) **while** $\tilde{a}_j \neq a_j$ **do**
      i. $\eta \leftarrow \frac{1}{2} \cdot \eta$,
      ii. $[a_{j-1}, \tilde{a}_j] = \left\{\phi \in \mathbb{R} : \bigcap_{k=1}^{K^*} \{M_k(y'(\phi)) = M_k(y'(a_j - \eta))\}\right\}$, where $K^* = \text{argmin}_k \{|\mathcal{CC}_k(y'(a_j - \eta))| = L\}$
   end
   (c) **if** $\hat{C}_1, \hat{C}_2 \in \mathcal{CC}_{K^*}(y'(a_j - \eta))$, **then** $\mathcal{J} \leftarrow \mathcal{J} \cup \{j - 1\}$.
   (d) $j \leftarrow j - 1$.
end
5. $\mathcal{S}^*_{\hat{C}_1,\hat{C}_2} \leftarrow \bigcup_{i \in \mathcal{J}} [a_i, a_{i+1}]$

Figure B.3 displays the results for the test based on $p^*_{\hat{C}_1,\hat{C}_2}$ with $L = 3$ for the two-dimensional fused lasso simulations described in Section 3.5.2. Panel (a) demonstrates that the test based on $p^*_{\hat{C}_1,\hat{C}_2}$ controls the selective Type I error. In panel (b) of Figure B.3, we see that the test based on $p^*_{\hat{C}_1,\hat{C}_2}$ has substantially higher power than the test based on p_{Hyun}. Figures B.3(c) and (d) investigate the detection probability (defined in (B.17)) and conditional power (defined in (B.16)) of $p^*_{\hat{C}_1,\hat{C}_2}$.

B.9 Additional power comparisons

In Section 3.5, we considered the power of the tests based on $p_{\hat{C}_1,\hat{C}_2}$ and p_{Hyun} as a function of the binned effect size $|\nu^\top \beta|$. Here, we conduct three additional analyses on the same simulated datasets from Section 3.5.

In the first analysis, we separately consider two quantities considered in prior work [Gao et al., 2020, Hyun et al., 2021, Jewell et al., 2022]: (i) the *detection probability* (i.e., the probability that $\hat{C}_1$ and $\hat{C}_2$ are true piecewise constant regions in the original signal β) of the graph fused lasso estimator in Eq. (3) of the main manuscript, and (ii) the *conditional power* of the tests based on $p_{\hat{C}_1,\hat{C}_2}$ and p_{Hyun} (i.e., the probability of rejecting $H_0 : \nu^\top \beta = 0$, *given that* $\hat{C}_1$ and $\hat{C}_2$ are true piecewise constant regions).

Given M simulated datasets, we can estimate the conditional power as

$$\text{Conditional power} = \frac{\sum_{m=1}^{M} 1\Big\{\exists j, j' \text{ s.t. } \hat{C}_1^{(m)} = C_j, \hat{C}_2^{(m)} = C_{j'}, p^{(m)} \leq \alpha\Big\}}{\sum_{m=1}^{M} 1\Big\{\exists j, j' \text{ s.t. } \hat{C}_1^{(m)} = C_j, \hat{C}_2^{(m)} = C_{j'}\Big\}}, \tag{B.16}$$

where $p^{(m)}$ and $\hat{C}_1^{(m)}, \hat{C}_2^{(m)}$ correspond to the p-values and estimated connected components under consideration for the mth simulated dataset. Here, C_j is the jth true piecewise constant segment in β. Because the quantity in (B.16) conditions on the event that $\hat{C}_1^{(m)}$ and $\hat{C}_2^{(m)}$ correspond to true piecewise constant segments, we also estimate how often this occurs:

$$\text{Detection probability} = \frac{\sum_{m=1}^{M} 1\Big\{\exists j, j' \text{ s.t. } \hat{C}_1^{(m)} = C_j, \hat{C}_2^{(m)} = C_{j'}\Big\}}{M}. \tag{B.17}$$

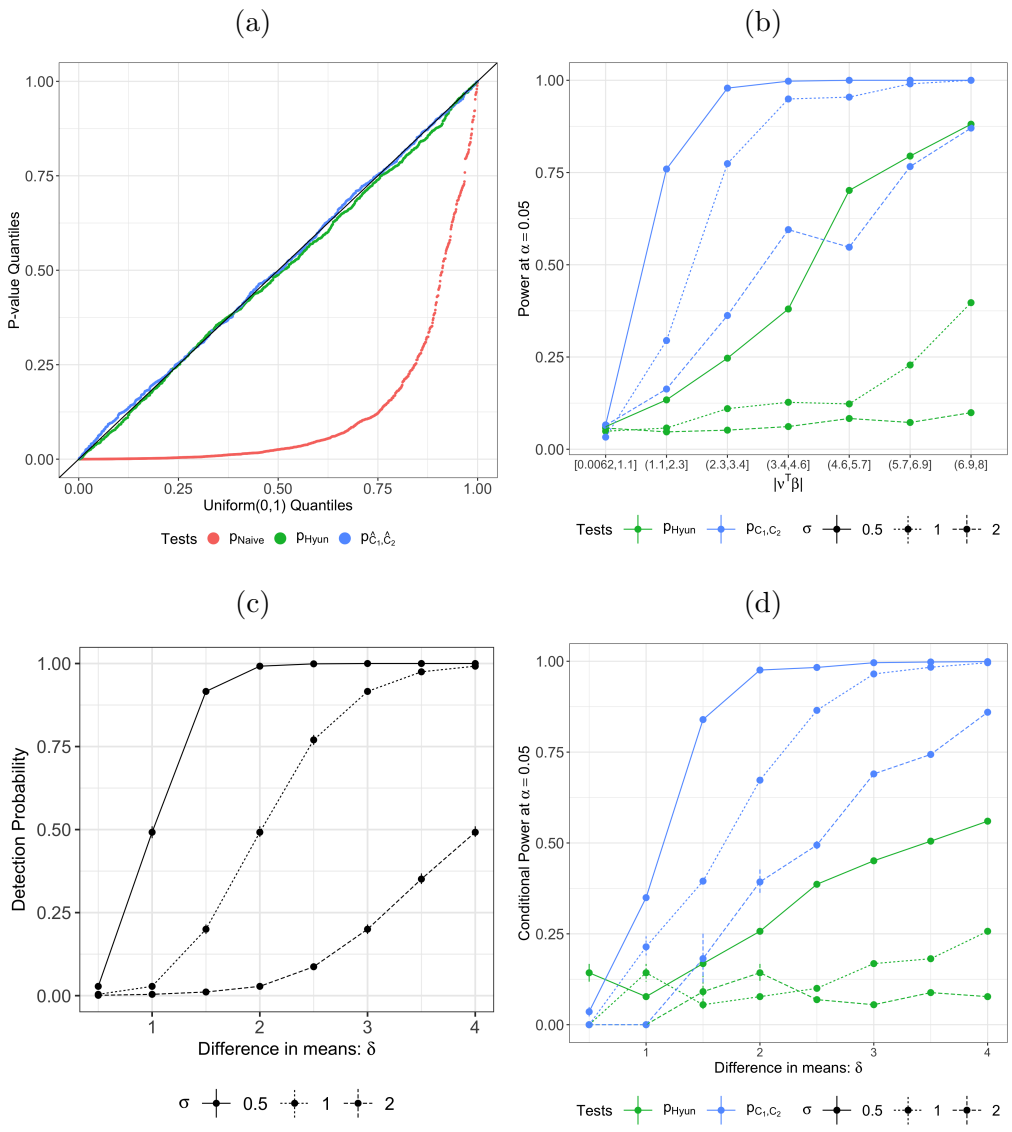


Figure B.3: *(a):* For the two-dimensional fused lasso model in (3.23), when $\delta = 0$, tests based on p_{Hyun} and $p^*_{\hat{C}_1,\hat{C}_2}$ control the selective Type I error. By contrast, the naive p-value leads to an inflated selective Type I error. *(b):* Under the simulation setup in Section 3.5.2, the power of tests based on p_{Hyun} and $p^*_{\hat{C}_1,\hat{C}_2}$ increases as a function of $|\nu^\top \beta|$. For a given bin of $|\nu^\top \beta|$, the test based on $p^*_{\hat{C}_1,\hat{C}_2}$ has substantially higher power than that based on p_{Hyun}. *(c):* Under the same setup as (b), the detection probability defined in (B.17) increases as a function of the difference in means between two piecewise constant segments (δ in (3.23)). Moreover, a larger value of noise variance σ^2 leads to a smaller detection probability. *(d):* Under the same setup as (b), the test based on $p^*_{\hat{C}_1,\hat{C}_2}$ has substantially higher conditional power (defined in (B.16)) than that based on p_{Hyun}. For both tests, power increases as a function of δ (defined in (B.17)).

We evaluated detection probability and conditional power on data generated from the one-dimensional and two-dimensional fused lasso models, with the same simulation setup as in Sections 3.5.1 and 3.5.2, respectively. Results aggregated over 1,500 simulations are displayed in Figure B.4. Panels (a) and (b) display the detection probability and conditional power (with $\alpha = 0.05$) for the one-dimensional fused lasso, respectively. Both quantities increase as the difference in means between the two piecewise constant segments (δ in (3.22)) increases. By contrast, both quantities decrease as the variance σ^2 increases. In addition, for a given value of σ and δ, the conditional power of the test based on $p_{\hat{C}_1,\hat{C}_2}$ is higher than that based on p_{Hyun}. We observe similar trends in the two-dimensional fused lasso case; see Figure B.4(c)–(d). In the second analysis, instead of binning $|\nu^\top \beta|$, we fit a regression spline using the `gam` function in the `R` package `mgcv` [Wood, 2017] to obtain a smooth estimate for the one-dimensional fused lasso simulations in Section 3.5.1. The results are in Figure B.5. As in Figure 3.3(c), the power of the tests that reject H_0 if $p_{\hat{C}_1,\hat{C}_2}$ or p_{Hyun} is below $\alpha = 0.05$ increases as $|\nu^\top \beta|$ increases. For a given value of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher power than that based on p_{Hyun}.

Finally, in the third analysis, we assess the sensitivity of our conclusions to the choice of K in the dual path algorithm, using the one-dimensional fused lasso model in (3.22). Recall that in Section 3.5.1, we choose $K = 2$ so the number of estimated connected components resulting from the one-dimensional fused lasso equals the true number of connected components in (3.22).

Here, we repeat the experiments in Section 3.5.1 with $K = 4$, which yields five estimated connected components. Results are displayed in Figure B.6. Panel (a) displays the observed p-value quantiles versus Uniform(0,1) quantiles, aggregated over 1,000 simulated datasets. As in the case of $K = 2$, only tests based on p_{Hyun} or $p_{\hat{C}_1,\hat{C}_2}$ control the selective Type I error. In Figure B.6(b), we see that the power of the tests based on p_{Hyun} or $p_{\hat{C}_1,\hat{C}_2}$ increases as $|\nu^\top \beta|$ increases. For a given value of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has substantially higher power than the test based on p_{Hyun}. In other words, the substantial increase in power, as well as the selective Type I error control, of the test based on $p_{\hat{C}_1,\hat{C}_2}$, *does not* depend on

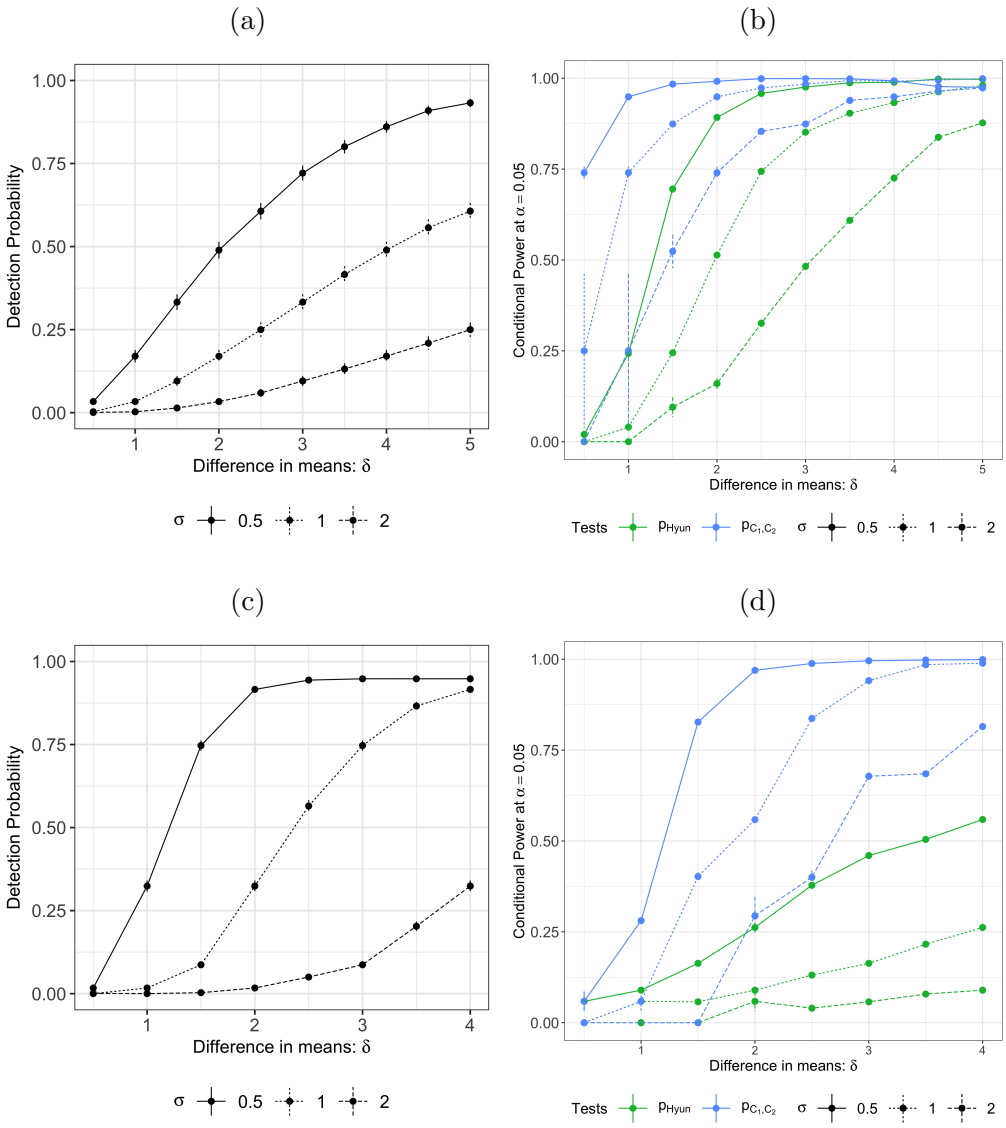


Figure B.4: *(a):* For the one-dimensional fused lasso simulations described in Section 3.5.1, the detection probability of the tests based on both p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ increases as a function of the difference in means between two piecewise constant segments, δ, across all values of σ. *(b):* For the one-dimensional fused lasso simulations described in Section 3.5.1, the conditional power of the tests based on both p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ increases as a function of $|\nu^\top \beta|$, across all values of σ. For a given bin of $|\nu^\top \beta|$ and a given value of σ, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher conditional power than the test based on p_{Hyun}. *(c):* Same as (a), but for the two-dimensional fused lasso simulations described in Section 3.5.2 *(d):* Same as (b), but for the two-dimensional fused lasso simulations described in Section 3.5.2.

correctly specifying K.

B.10 Estimation of the error variance σ^2 in (3.1)

Throughout this section, we have assumed that σ^2 in (3.1) is known. In practice, we can plug in an estimate $\hat{\sigma}$ when computing the p-values $p_{\hat{C}_1,\hat{C}_2}$ and p_{Hyun}. That is, we use

$$p_{\hat{C}_1,\hat{C}_2}(\hat{\sigma}) = \mathbb{P}\Big(|\phi(\hat{\sigma})| \geq |\nu^\top y| \;\Big|\; \hat{C}_1(y), \hat{C}_2(y) \in \mathcal{CC}_K(y'(\phi(\hat{\sigma})))\Big), \tag{B.18}$$

where $\phi(\hat{\sigma}) \sim \hat{\sigma} \cdot \mathcal{N}(0, ||\nu||_2^2)$.

In this section, we investigate the empirical selective Type I error control and power of the following estimators of σ^2:

- $\hat{\sigma}^2_{\text{Residual}} = \frac{1}{n-L}\sum_{l=1}^{L}\sum_{j\in\hat{C}_l}\left(y_j - \left(\sum_{j'\in\hat{C}_l} y_{j'}\right)/|\hat{C}_l|\right)^2$, where $\hat{C}_1, \ldots, \hat{C}_L$ are the L estimated connected components of the graph fused lasso solution $\hat{\beta}$;

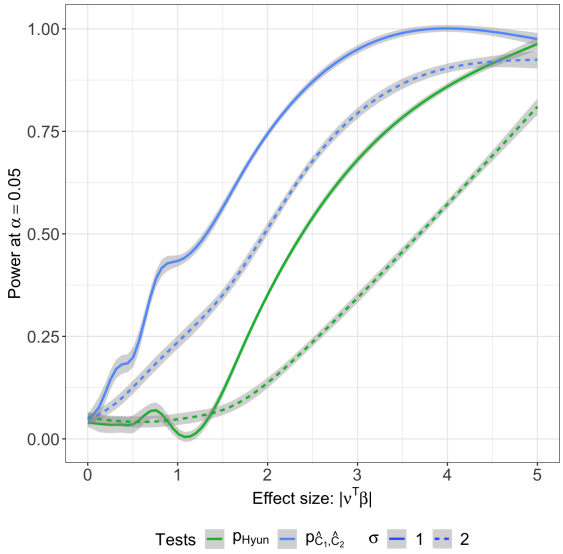


Figure B.5: Additional analysis of the data in Section 3.5.1. We used a generalized additive model to obtain the power of the tests based on p_{Hyun} in (3.10) and $p_{\hat{C}_1,\hat{C}_2}$ in (3.11) as a smooth function of $|\nu^\top\beta|$.

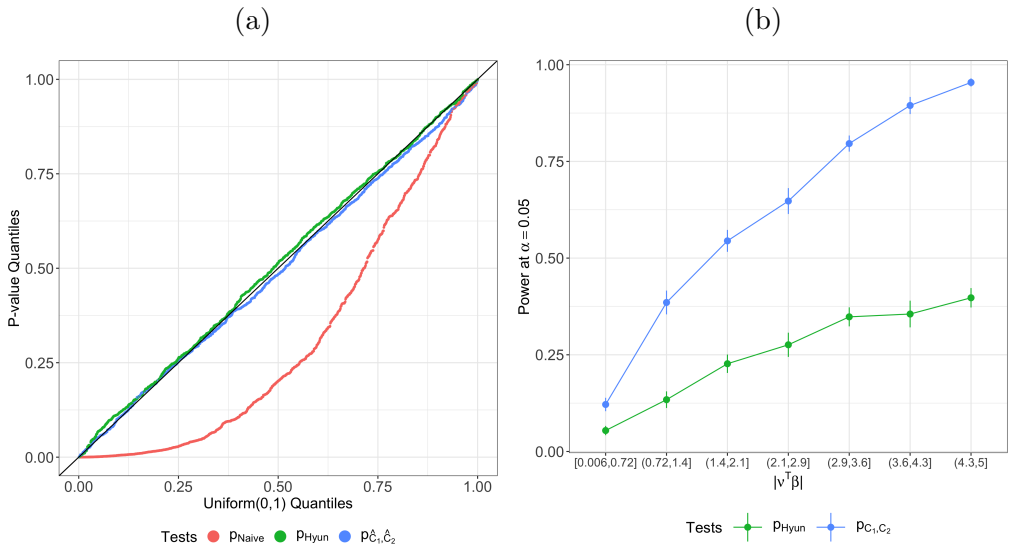


Figure B.6: *(a):* For the one-dimensional fused lasso simulations in Section 3.5.1, when $\delta = 0$ and the graph fused lasso is solved using the dual path algorithm with $K = 4$, tests based on p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ control the selective Type I error. By contrast, the naive p-value leads to an inflated selective Type I error. *(b):* For the one-dimensional fused lasso simulations in Section 3.5.1 with $\sigma = 1$, the power of tests based on p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ increases as a function of $|\nu^\top \beta|$. For a given bin of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has substantially higher power than that based on p_{Hyun}.

- $\hat{\sigma}^2_{\text{Sample}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_j - \bar{y})^2$, where $\bar{y}$ is the mean of the data y; and
- $\hat{\sigma}^2_{\text{MAD}} = \frac{\text{median}_{i=2,\ldots,n}(|z_i - \tilde{z}|)}{\sqrt{2}\Phi^{-1}(3/4)}$, where $\tilde{z} = \text{median}_{i=2,\ldots,n}(z_i)$, and $z_i = y_i - y_{i-1}$.

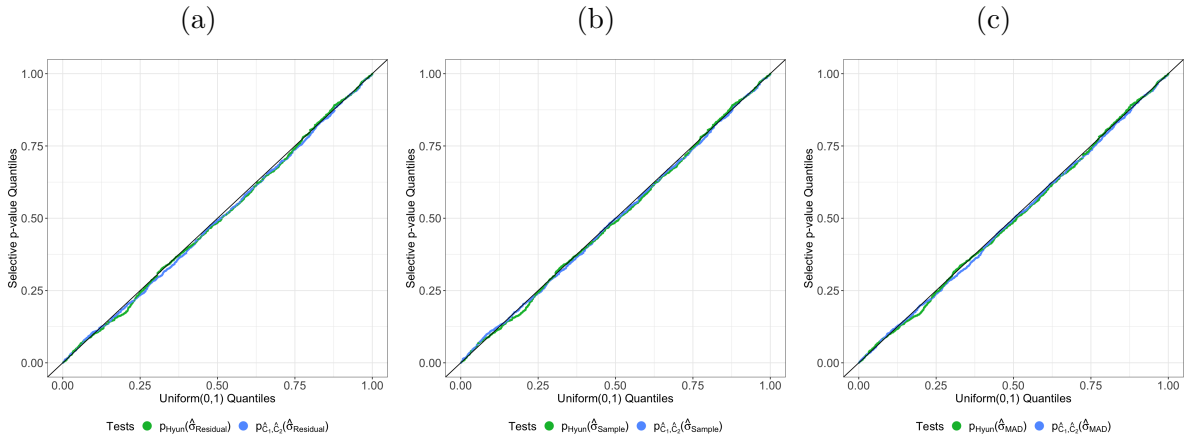


Figure B.7: *(a):* Quantile-quantile plot for p-values p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ computed using the estimated variances $\hat{\sigma}^2_{\text{Residual}}$ for the one-dimensional fused lasso simulations described in Section 3.5.1 of the manuscript. Results are based on 1,000 simulated datasets under the global null. *(b):* Same as (a), but for the variance estimator $\hat{\sigma}^2_{\text{Sample}}$. *(c):* Same as (a), but for the variance estimator $\hat{\sigma}^2_{\text{MAD}}$.

Figure B.7 displays the quantiles of the p-values p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ computed using the estimated variances with the same simulation setup as in Section 3.5.1. All three estimators ($\hat{\sigma}^2_{\text{Residual}}$ with $L = 3$, $\hat{\sigma}^2_{\text{Sample}}$, and $\hat{\sigma}^2_{\text{MAD}}$) lead to selective Type I error control under the global null.

In addition, we compared the power of the tests based on estimated variance with that obtained using the true variance. Results from a simulation study with the same setup as in Section 3.5.1 are aggregated in Figure B.8. We see that the tests based on $\hat{\sigma}^2_{\text{Residual}}$ with $L = 3$ or $\hat{\sigma}^2_{\text{MAD}}$ result in nearly identical power to the test based on the true variance σ^2. By contrast, using $\hat{\sigma}^2_{\text{Sample}}$ leads to a less powerful test, especially for larger values of $|\nu^\top\beta|$. Moreover, the test based on $p_{\hat{C}_1,\hat{C}_2}$ is more powerful than the counterpart based on p_{Hyun}, *regardless of the chosen variance estimator.*

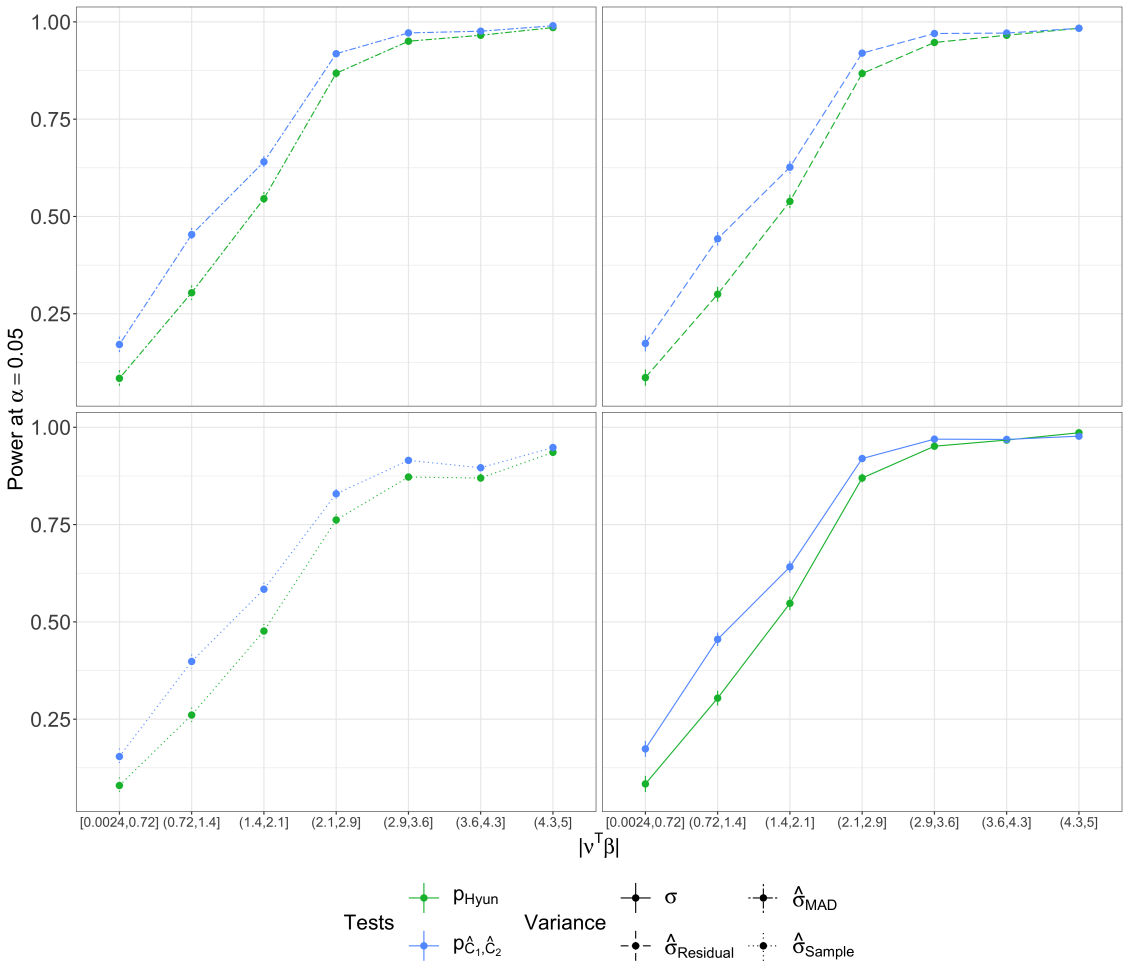


Figure B.8: The power of the tests based on both p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$ computed using the true variance and three different variance estimators at level $\alpha = 0.05$, for the one-dimensional fused lasso simulations described in Section 3.5.1 of the manuscript. Tests based on the true variance have the highest power, followed closely by the ones based on $\hat{\sigma}^2_{\text{MAD}}$ and $\hat{\sigma}^2_{\text{Residual}}$. Tests based on $\hat{\sigma}^2_{\text{Sample}}$ have the lowest power.

Our empirical results agree with observations made in related problems for selective inference: (i) when the global null does not hold, $\hat{\sigma}^2_{\text{Sample}}$ is a conservative estimator of the true variance σ^2 [Gao et al., 2020, Hyun et al., 2021, Rügamer and Greven, 2020, Tibshirani et al., 2018a]; (ii) $\hat{\sigma}^2_{\text{Residual}}$ has good empirical performance when L is correctly specified; and (iii) in the case of the one-dimensional fused lasso, $\hat{\sigma}^2_{\text{MAD}}$ is an asymptotically consistent estimator under appropriate assumptions [Jewell et al., 2022, Kovács et al., 2020].

B.11 Timing complexity for Algorithm 4

In this section, we first characterize the computational complexity of Algorithm 4 using the following Proposition.

Proposition 19. *Recall that $M_k(y) = (B_k(y), s_{B_k}(y), R_k(y), L_k(y))$ is the output of the kth step of the dual path algorithm (see Algorithm 3 in Appendix B.1). Define*

$$\tilde{\mathcal{I}} \equiv \left\{ (m_1, \ldots, m_K) : \exists \alpha \in \mathbb{R} \text{ such that } \bigcap_{k=1}^{K} \{M_k(y'(\alpha)) = m_k\} \right\}. \tag{B.19}$$

Then, computing the p-value $p_{\hat{C}_1, \hat{C}_2}$ using Algorithm 4 in Appendix B.3 requires running the dual path algorithm $\mathcal{O}(|\tilde{\mathcal{I}}|)$ times, where $|\cdot|$ denotes the cardinality of a set.

We omit the proof of Proposition 19, as it directly follows from Algorithm 4 and the definition of $\tilde{\mathcal{I}}$. Proposition 19 implies that the time complexity for Algorithm 4 is instance-dependent, and can in principle be prohibitively large. For instance, in the one-dimensional fused lasso case, $|\tilde{\mathcal{I}}|$ is upper bounded by $2^K \cdot \frac{n!}{(n-K)!}$, where n is the number of observations, and K is the number of steps in the dual path algorithm. This upper bound comes from the observation that in the one-dimensional fused lasso problem, $M_k(y)$ in (3.9) is equivalent to the *location and sign* of the estimated changepoint during the kth step of the dual path algorithm (see Algorithm 3 in Appendix B.1) [Hyun et al., 2018, Tibshirani and Taylor, 2011]. *However, in practice, we are nowhere near this worst case scenario:* in the experiments in Section 4.5, $|\tilde{\mathcal{I}}|$ is of reasonable size. In particular, for the one-dimensional fused lasso

simulations described in Section 3.5.1, the upper bound postulates that $|\tilde{\mathcal{I}}|$ can be as large as $158,400$ ($n = 200, K = 2$). However, as displayed in Figure B.9(a), empirically, $|\tilde{\mathcal{I}}|$ falls under 500 in all instances.

In addition, because we re-implemented the polyhedron approach in Hyun et al. [2018] using the ideas from Arnold and Tibshirani [2016], each graph fused lasso instance and its corresponding intervals of the form $[a_i, a_{i+1}]$ (see Section 3.3 for more details) can be computed efficiently.

Figure B.9(b) displays the running time of Algorithm 2, computed on a MacBook Pro with a 1.4 GHz Intel Core i5 processor, over 1,000 replicate datasets simulated according to the one-dimensional fused lasso model in Section 3.5.1 with $n = 200, \delta = 0, \sigma = 1$. The graph fused lasso problem is solved using the the dual path algorithm with $K = 2$. The average running time for running Algorithm 4 to test each hypothesis $H_0 : \nu^\top \beta = 0$ is 2.7 seconds.

When the empirical size of $\tilde{\mathcal{I}}$ defined in (B.19) is large, we could alternatively use an importance sampling approach to obtain an approximate p-value with ideas from, e.g., Rügamer et al. [2022], Rügamer and Greven [2020], Yang et al. [2016].

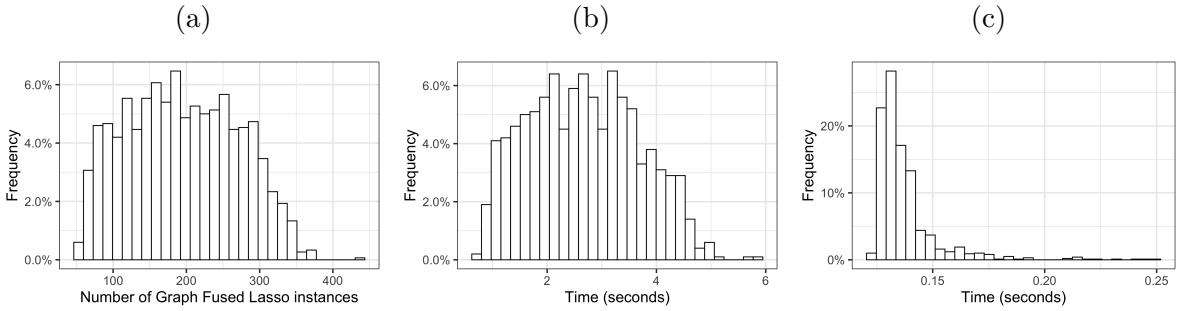
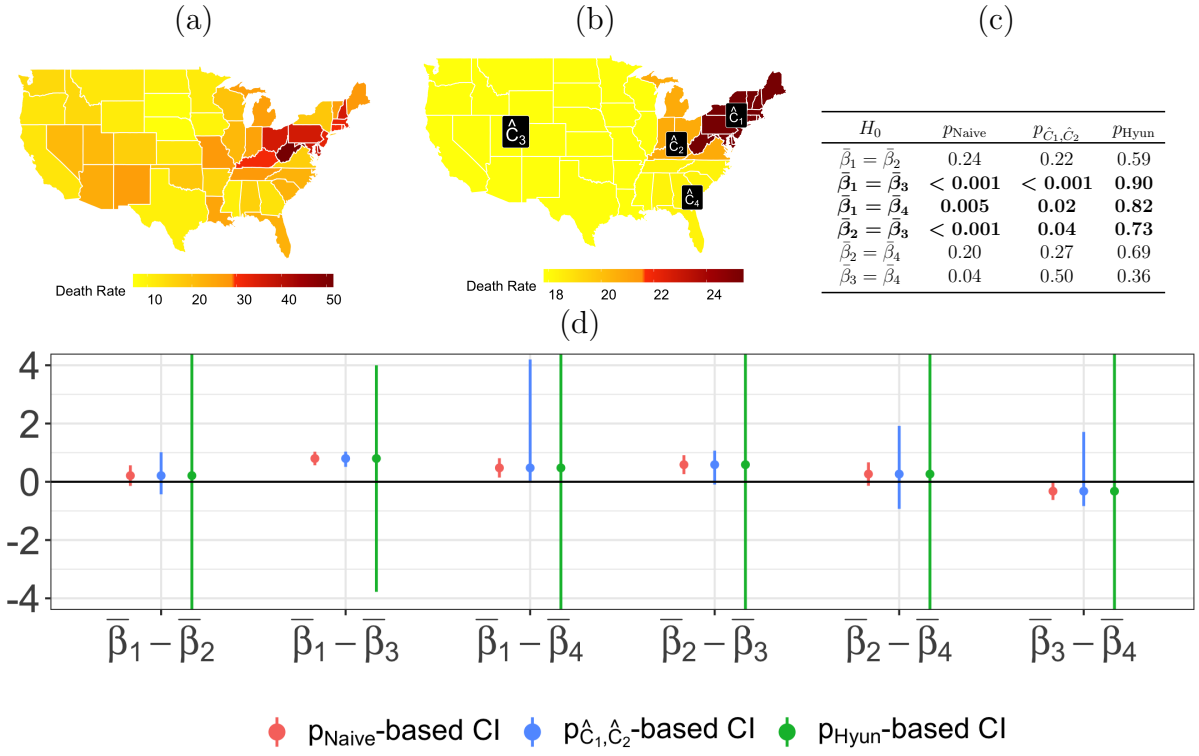


Figure B.9: *(a):* Empirical distribution of $|\tilde{\mathcal{I}}|$ over 1,000 replicate datasets. Each dataset is simulated according to the one-dimensional fused lasso model described in Section 3.5.1. We solved the graph fused lasso problem with $K = 2$ steps in the dual path algorithm. *(b):* Same as (a), but for the running time of Algorithm 4. *(c):* Same as (b), but for the running time of computing p_{Hyun}.

B.12 Additional results for data applications

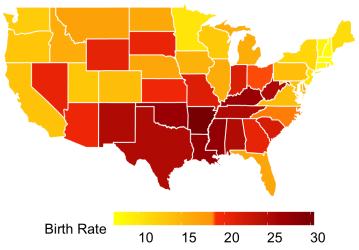
Here, we repeat the analysis in Sections 3.6.1 and 3.6.2 with $K = 27$ and $K = 20$, respectively. Results are displayed in Figures B.10 and B.11. Similar to the case of $K = 30$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ leads to more rejections than the test based on p_{Hyun} at $\alpha = 0.05$. Furthermore, the confidence intervals based on $p_{\hat{C}_1,\hat{C}_2}$ are considerably shorter than those based on p_{Hyun}, and in some cases, even comparable to the naive confidence intervals that do not have proper coverage for the true parameter $\nu^\top \beta$.



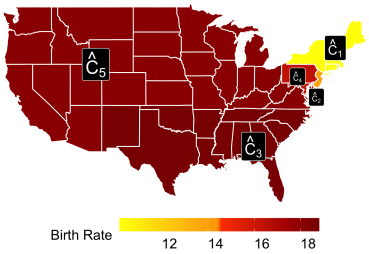
H_0	p_{Naive}	$p_{\hat{C}_1,\hat{C}_2}$	p_{Hyun}
$\bar{\beta}_1 = \bar{\beta}_2$	0.24	0.22	0.59
$\bar{\beta}_1 = \bar{\beta}_3$	**< 0.001**	**< 0.001**	**0.90**
$\bar{\beta}_1 = \bar{\beta}_4$	**0.005**	**0.02**	**0.82**
$\bar{\beta}_2 = \bar{\beta}_3$	**< 0.001**	**0.04**	**0.73**
$\bar{\beta}_2 = \bar{\beta}_4$	0.20	0.27	0.69
$\bar{\beta}_3 = \bar{\beta}_4$	0.04	0.50	0.36

Figure B.10: *(a):* The observed drug overdose death rates (deaths per 100,000 persons) for the 48 contiguous U.S. states in the year 2018. *(b):* Applying the graph fused lasso to the drug overdose data with $K = 27$ results in four estimated connected components. *(c):* For each pair of estimated connected components, we computed p_{Naive}, p_{Hyun}, and $p_{\hat{C}_1,\hat{C}_2}$. For brevity, we use the notation $\bar{\beta}_l = \sum_{j \in \hat{C}_l} \beta_j / |\hat{C}_l|$. *(d):* For each pair of estimated connected components, we constructed confidence intervals for the difference in means, corresponding to p_{Naive}, p_{Hyun}, and $p_{\hat{C}_1,\hat{C}_2}$.

(a)



(b)



(c)

H_0	p_{Naive}	$p_{\hat{C}_1,\hat{C}_2}$	p_{Hyun}
$\bar{\beta}_1 = \bar{\beta}_2$	0.67	0.99	0.74
$\bar{\beta}_1 = \bar{\beta}_3$	< 0.001	0.07	0.57
$\boldsymbol{\bar{\beta}_1 = \bar{\beta}_4}$	**0.001**	**0.01**	**0.26**
$\bar{\beta}_1 = \bar{\beta}_5$	< 0.001	0.63	0.30
$\bar{\beta}_2 = \bar{\beta}_3$	< 0.001	0.08	0.27
$\bar{\beta}_2 = \bar{\beta}_4$	0.13	0.34	0.26
$\bar{\beta}_2 = \bar{\beta}_5$	0.02	0.65	0.24
$\bar{\beta}_3 = \bar{\beta}_4$	< 0.001	0.12	0.71
$\bar{\beta}_3 = \bar{\beta}_5$	< 0.001	0.14	0.61
$\bar{\beta}_4 = \bar{\beta}_5$	0.34	0.78	0.71

(d)

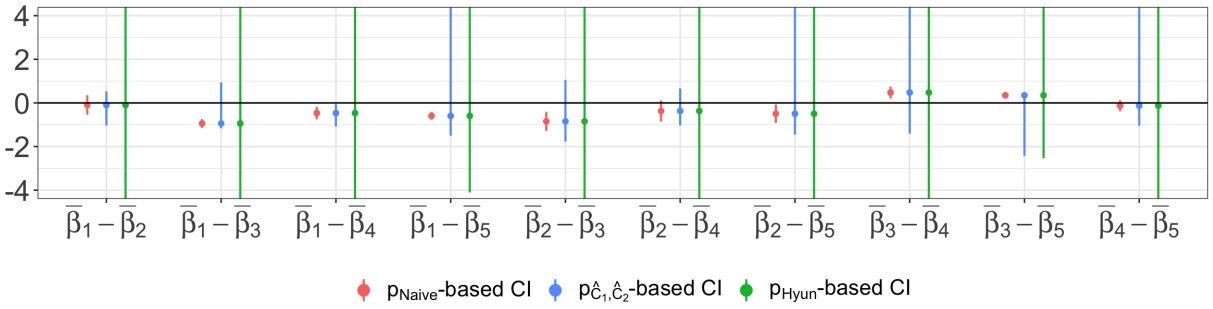


Figure B.11: *(a):* The observed teenage birth rates (births per 1,000 females aged 15–19) for the 48 contiguous U.S. states in 2018. *(b):* The graph fused lasso solution with $K = 20$ results in five estimated connected components, displayed in distinct colors. *(c):* For each pair of estimated connected components, we computed p_{Naive}, p_{Hyun}, and $p_{\hat{C}_1,\hat{C}_2}$. Pairs for which the test based on $p_{\hat{C}_1,\hat{C}_2}$ results in a rejection at $\alpha = 0.05$, but not for the test based on p_{Hyun}, are in bold. *(d):* Confidence intervals for the differences in means for each pair of connected components.

B.13 A comparison of $p_{\hat{C}_1,\hat{C}_2}$ and p_{LeDuy}

In this section, we first briefly review the p-value proposal of Le Duy and Takeuchi [2021] (henceforth referred to as p_{LeDuy}), and elaborate on the conceptual differences between p_{LeDuy} and $p_{\hat{C}_1,\hat{C}_2}$ (3.11). Next, we provide results from a simulation study that compares the selective Type I error (3.6) and the power of the tests based on $p_{\hat{C}_1,\hat{C}_2}$, p_{Hyun}, and p_{LeDuy}. In the simulations that follow, we will only consider the one-dimensional fused lasso problem, since the extension to a non-chain graph has not been implemented by Le Duy and Takeuchi [2021] at the time of writing. We used the software for computing p_{LeDuy} provided by the authors at https://github.com/vonguyenleduy/parametric_generalized_lasso_selective_inference.

B.13.1 A conceptual comparison of $p_{\hat{C}_1,\hat{C}_2}$ and p_{LeDuy}

Le Duy and Takeuchi [2021] consider the p-value p_{LeDuy} defined as

$$p_{\text{LeDuy}} \equiv \mathbb{P}_{H_0}\Big(|\nu^\top Y| \geq |\nu^\top y| \;\Big|\; \Big\{i : \big(D\hat{\beta}(Y)\big)_i \neq 0\Big\} = \Big\{i : \big(D\hat{\beta}(y)\big)_i \neq 0\Big\}, \Pi^\perp_\nu Y = \Pi^\perp_\nu y\Big), \tag{B.20}$$

where, with a slight abuse of notation, $\hat{\beta}(Y)$ is the solution to (3.2) with data Y. Using a similar argument to Proposition 9, they showed that (B.20) can be recast as the cumulative distribution function of a $\mathcal{N}(0, \sigma^2||\nu^2||)$ random variable, truncated to a set that can be efficiently computed. We note that in the case of the graph fused lasso (3.3), the set $\Big\{i : \big(D\hat{\beta}(Y)\big)_i \neq 0\Big\}$ is equivalent to the set of edges used to determine the connected components of $\hat{\beta}$; in other words, conditioning on the event $\Big\{i : \big(D\hat{\beta}(Y)\big)_i \neq 0\Big\} = \Big\{i : \big(D\hat{\beta}(y)\big)_i \neq 0\Big\}$ is equivalent to conditioning on $\{B_K(Y) = B_K(y)\}$, for an appropriate choice of K [Tibshirani and Taylor, 2011].

What advantages, then, does our proposal $p_{\hat{C}_1,\hat{C}_2}$ in (3.11) offer when compared to p_{LeDuy}?

- *Smaller conditioning set:* first of all, we condition on *even less* information when constructing $p_{\hat{C}_1,\hat{C}_2}$: p_{LeDuy} conditions on the set of edges that are used to determine the

connected components of $\hat{\beta}$, and therefore implicitly, on *all* of the connected components in $\hat{\beta}$ (see, e.g., Proposition 7). By contrast, in (3.11), we condition only on the pair of connected components under investigation, thereby obtaining higher power.

- *Interpretability:* the conditioning set for $p_{\hat{C}_1,\hat{C}_2}$ is based on the connected components of $\hat{\beta}$ after running the dual path algorithm for K steps. By contrast, p_{LeDuy} conditions on the output of (3.3) with a specific λ. We argue that, as a result, $p_{\hat{C}_1,\hat{C}_2}$ is more interpretable. Consider the widely-popular one-dimensional fused lasso problem and a pair of estimated piecewise constant segments $\hat{C}_1, \hat{C}_2$. $p_{\hat{C}_1,\hat{C}_2}$ answers the question:

  > *Assuming that there is no difference between the population means of $\hat{C}_1$ and $\hat{C}_2$, then what's the probability of observing such a large difference in the sample means of $\hat{C}_1$ and $\hat{C}_2$, given that $\hat{C}_1$ and $\hat{C}_2$ are among the $K+1$ piecewise constant segments estimated from the data?*

  On the other hand, p_{LeDuy} answers the question:

  > *Assuming that there is no difference between the population means of $\hat{C}_1$ and $\hat{C}_2$, then what's the probability of observing such a large difference in the sample means of $\hat{C}_1$ and $\hat{C}_2$, given that $\hat{C}_1$ and $\hat{C}_2$ are among the piecewise constant segments estimated from the data with a specific λ?*

  Here, $p_{\hat{C}_1,\hat{C}_2}$ is answering the question about $K+1$ piecewise constant segments, where K is a very interpretable quantity (i.e., the number of estimated changepoints). In contrast, for p_{LeDuy}, the meaning of λ could vary greatly across different datasets — the value of λ that yields K estimated changepoints on one dataset could yield far more or fewer estimated changepoints on another dataset.

- *Numerical stability:* Le Duy and Takeuchi [2021] solved the primal problem (3.3) using an iterative solver, which in practice leads to numerical issues when identifying the

set $\left\{i : \left(D\hat{\beta}(Y)\right)_i \neq 0\right\}$ [Arnold and Tibshirani, 2016], and consequently, in computing p_{LeDuy}. By contrast, we chose to work with the dual problem (3.7), which avoids these numerical issues and yields the *exact* connected components of $\hat{\beta}$ when computing $p_{\hat{C}_1,\hat{C}_2}$.

B.13.2 A simulation study comparing p_{Hyun}, $p_{\hat{C}_1,\hat{C}_2}$, and p_{LeDuy}

Next, we conducted a simulation study to compare the selective Type I error and power of the tests based on the following p-values: p_{Hyun} in (3.10), $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), and p_{LeDuy} in (B.20). We tested the null hypothesis $H_0 : \nu^\top \beta = 0$, where ν is defined in (3.5) for a randomly-chosen pair of *adjacent* piecewise constant segments $\hat{C}_1$, $\hat{C}_2$ in the solution to the one-dimensional fused lasso problem.

The signal $\beta \in \mathbb{R}^{500}$ is piecewise constant with 10 changepoints $\{\tau_1, \ldots, \tau_{10}\}$ (or equivalently, 11 piecewise constant segments), and the values of β alternate between 0 and δ after each changepoint:

$$Y_j \overset{ind.}{\sim} \mathcal{N}(\beta_j, \sigma^2), \quad \beta_j = \sum_{i=0}^{11} \delta \times 1\{i \text{ odd}, \tau_i \leq j \leq \tau_{i+1}\}, \quad j = 1, \ldots, 500, \tag{B.21}$$

where $\tau_0 \equiv 0$ and $\tau_{11} \equiv 500$. Figure B.12(a) displays an instance of (B.21) with $\delta = 3$ and $\sigma = 1$.

In the simulations that follow, the software accompanying Le Duy and Takeuchi [2021] returned an empty string for p_{LeDuy} for around 27% of the hypotheses. Upon inquiring, the authors of Le Duy and Takeuchi [2021] said that this is due to numerical stability issues in identifying the conditioning set in p_{LeDuy}. Consequently, the displayed results for p_{LeDuy} are based on the subset of the hypotheses for which the authors' software successfully returned a p-value.

We first investigate the selective Type I error control by simulating $y_1, \ldots, y_{500}$ from (B.21) with $\delta = 0$ and $\sigma = 1$. Therefore, the null hypothesis $H_0 : \nu^\top \beta = 0$ holds for all contrast vectors ν, regardless of the pair of piecewise constant segments under investigation.

For p_{Hyun} and $p_{\hat{C}_1,\hat{C}_2}$, we solved (3.3) with $K = 10$ steps in the dual path algorithm, which yields exactly 11 piecewise constant segments by the properties of the one-dimensional fused lasso. For p_{LeDuy}, we selected the tuning parameter λ so that (3.3) yields exactly 11 piecewise constant segments on the data.

Figure B.12(b) displays the observed p-value quantiles versus Uniform$(0, 1)$ quantiles, aggregated over 1,000 hypothesis tests. We see that all three tests based on p_{Hyun} (3.10), $p_{\hat{C}_1,\hat{C}_2}$ (3.11), and p_{LeDuy} (B.20) control the selective Type I error as in (3.6).

Next, we show that the test based on $p_{\hat{C}_1,\hat{C}_2}$ has higher power than the test based on p_{LeDuy}, and both tests have higher power than the test based on p_{Hyun}. We generated 500 datasets from (B.21) for each of ten evenly-spaced values of $\delta \in [0.5, 5]$. For each simulated dataset, we solved (3.3) with $K = 10$ for p_{Hyun}, and $p_{\hat{C}_1,\hat{C}_2}$, and chose the tuning parameter λ so that (3.3) yields 10 estimated changepoints for p_{LeDuy}. We rejected the null hypothesis $H_0 : \nu^\top \beta = 0$ if the p-value was less than $\alpha = 0.05$. As in Section 3.5.1, we consider the power as a function of $|\nu^\top \beta|$.

Figure B.12(c) displays the power estimated by first creating six evenly-spaced bins of the observed values of $|\nu^\top \beta|$, and then computing the proportion of simulated datasets for which we rejected H_0 within each bin. Alternatively, we could estimate the power as a smooth function of $|\nu^\top \beta|$ using a regression spline (see Figure B.12(d)). In both cases, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has 10–15% higher power than the test based p_{LeDuy}, and both have substantially higher power than the test based on p_{Hyun}.

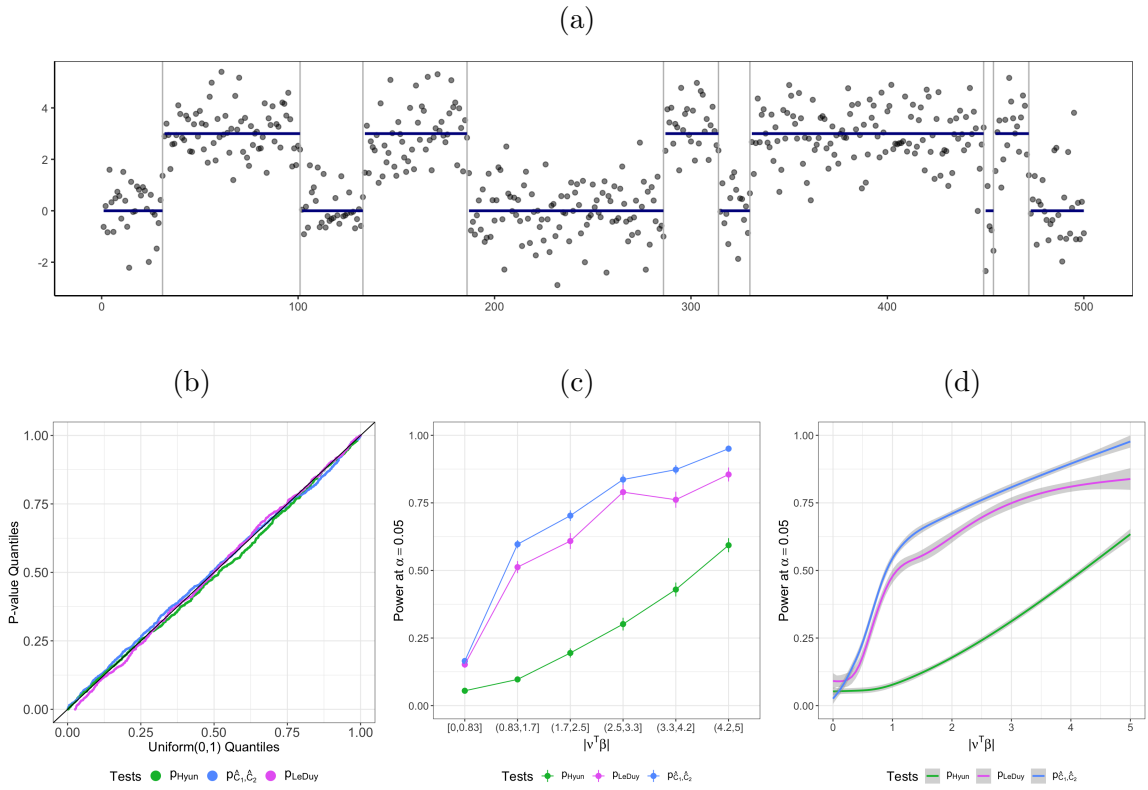


Figure B.12: *(a):* One realization of y generated according to (B.21) with $\delta = 3$ and $\sigma = 1$ (grey dots), along with the true signal β (blue curve). *(b):* When $\delta = 0$, tests based on p_{Hyun} in (3.10), $p_{\hat{C}_1,\hat{C}_2}$ in (3.11), and p_{LeDuy} in (B.20) control the selective Type I error in (3.6). *(c):* The power of the tests based on p_{Hyun}, $p_{\hat{C}_1,\hat{C}_2}$, and p_{LeDuy} increases as a function of the effect size $|\nu^\top \beta|$. For a given bin of $|\nu^\top \beta|$, the test based on $p_{\hat{C}_1,\hat{C}_2}$ has the highest power, followed by the test based on p_{LeDuy}, and finally the test based on p_{Hyun}. *(d):* Same as (c), but the power of the three tests are estimated using the `gam` function in the `R` package `mgcv` [Wood, 2017] instead of binning.

Appendix C

C.1 Proof of Proposition 13

The proof of Proposition 13 is similar to the proof of Theorem 1 in Gao et al. [2020], the proof of Theorem 3.1 in Loftus and Taylor [2014], the proof of Lemma 1 in Yang et al. [2016], and the proof of Theorem 3.1 in Chen and Bien [2020].

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

$$X = \Pi_\nu^\perp X + (\mathbf{I}_n - \Pi_\nu^\perp)X = \Pi_\nu^\perp X + \frac{\nu\nu^\top X}{\|\nu\|_2^2} = \Pi_\nu^\perp X + \left(\frac{\|X^\top \nu\|_2}{\|\nu\|_2^2}\right)\nu\big\{\mathrm{dir}\big(X^\top\nu\big)\big\}^\top. \quad \text{(C.1)}$$

Lemma 11. *Under (4.1) and $H_0 : \mu^\top \nu = 0_q$, we have that $\|X^\top\nu\|_2$, $\Pi_\nu^\perp X$, and $\mathrm{dir}(X^\top\nu)$ are pairwise independent.*

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

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

We now proceed to prove the statement in (4.10). Recalling the definition of $p_{\text{selective}}$ in

(4.9), under $H_0 : \mu^\top \nu = 0_q$ with ν defined in (4.3), we have that

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

Here, step *a.* follows from substituting X with the expression in (C.1), and step *b.* follows from replacing $\Pi_\nu^\perp X$ and $\text{dir}(X^\top\nu)$ with $\Pi_\nu^\perp x$ and $\text{dir}(x^\top\nu)$, respectively. Next, in step *c.*, we used Lemma 11. Finally, step *d.* follows from the definition of $x'(\phi)$ in (4.11).

Note that under (4.1) and $H_0 : \mu^\top\nu = 0_q$, we have that $\|X^\top\nu\|_2 \sim \sigma\|\nu\|_2\chi_q$, which concludes the proof of (4.10).

It remains to show that the test that rejects $H_0 : \mu^\top\nu = 0$ when $p_{\text{selective}} \leq \alpha$ controls the selective Type I error at level α, in the sense of (4.5). First of all, recall that we decided to test the null hypothesis in (4.2) based on the output of Algorithm 1. Therefore, $p_{\text{selective}}$ controls the selective Type I error at level α if, for any $c_i^{(T)}(x)$, $i = 1, \ldots, n$,

$$
\mathbb{P}_{H_0}\left[\text{reject } H_0 \text{ at level } \alpha \;\middle|\; \bigcap_{i=1}^{n}\left\{c_i^{(T)}(X) = c_i^{(T)}(x)\right\}\right] \leq \alpha,\ \forall\alpha \in (0,1). \tag{C.3}
$$

To prove (C.3), we first note that the following holds for any $\alpha \in (0,1)$:

$$
\begin{aligned}
&\mathbb{P}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}(X) = c_i^{(t)}(x)\right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x,\, \text{dir}(X^\top\nu) = \text{dir}(x^\top\nu)\right] \\
&\overset{a.}{=} \mathbb{P}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp X + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\text{dir}\left(X^\top\nu\right)\right\}^\top\right) = c_i^{(t)}(x)\right\},\right. \\
&\quad \left.\Pi_\nu^\perp X = \Pi_\nu^\perp x,\, \text{dir}(X^\top\nu) = \text{dir}(x^\top\nu)\right] \\
&\overset{b.}{=} \mathbb{P}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\text{dir}\left(x^\top\nu\right)\right\}^\top\right) = c_i^{(t)}(x)\right\},\right. \\
&\quad \left.\Pi_\nu^\perp X = \Pi_\nu^\perp x,\, \text{dir}(X^\top\nu) = \text{dir}(x^\top\nu)\right] \\
&\overset{c.}{=} \mathbb{P}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(\Pi_\nu^\perp x + \left(\frac{\|X^\top\nu\|_2}{\|\nu\|_2^2}\right)\nu\left\{\text{dir}\left(x^\top\nu\right)\right\}^\top\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{d.}{=} \mathbb{P}_{H_0}\left[p_{\text{selective}}(\|X^\top\nu\|_2) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'\left(\|X^\top\nu\|_2\right)\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{e.}{=} \mathbb{P}_{H_0}\left[1 - F_q^{\mathcal{S}_T}(\|X^\top\nu\|_2) \le \alpha \;\middle|\; \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'\left(\|X^\top\nu\|_2\right)\right) = c_i^{(t)}(x)\right\}\right] \\
&\overset{f.}{=} \alpha.
\end{aligned}
\tag{C.4}
$$

Here, steps $a.$ through $d.$ follow from the same line of argument in (C.2). Moreover, (4.10) implies that, for a given sequence of cluster assignments $c_i^{(T)}(x)$, $i = 1, \ldots, n$, $p_{\text{selective}}$ is the survival function of a χ_q random variable, truncated to the set $\mathcal{S}_T$ defined in (4.12). Letting $F_q^{\mathcal{S}_T}(\cdot)$ denote the cumulative distribution function of this truncated χ_q random variable, we arrive at step $e.$ Finally, to prove $f.$, we first note that under H_0, the conditional cumulative distribution function of $\|X^\top\nu\|_2$ given $\bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(x'\left(\|X^\top\nu\|_2\right)\right) = c_i^{(t)}(x)\right\}$ is exactly $F_q^{\mathcal{S}_T}$. The equality, therefore, follows from the probability integral transform, which states that for a continuous random variable Z, $F_Z(Z)$ follows the Uniform(0,1) distribution.

Finally, we have that

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

In the proof above, $a.$ follows from the tower property of conditional expectation, and $b.$ is a direct consequence of (C.4).

Therefore, we conclude that the test based on $p_{\text{selective}}$ controls the selective Type I error in (4.5), which completes the proof of Proposition 13.

C.2 Proof of Proposition 14

We will derive the expression for $\mathcal{S}_T$ in Proposition 14 using an induction argument. For a positive integer K, we let $[K]$ denote the set $\{1, \ldots, K\}$.

The following two claims (Lemmas 12 and 13) serve as the "base cases" for the proof.

Lemma 12. *Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned during the tth iteration of step 3b. of Algorithm 1 applied to data x, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from x during step 1 of Algorithm 1. For $\mathcal{S}_0$ defined as*

$$
\mathcal{S}_0 = \left\{\phi \in \mathbb{R} : \bigcap_{i=1}^{n}\left\{c_i^{(0)}(x'(\phi)) = c_i^{(0)}(x)\right\}\right\}, \tag{C.5}
$$

we have that

$$\mathcal{S}_0 = \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi)) \right\|_2^2 \leq \left\| [x'(\phi)]_i - m^{(0)}_{k}(x'(\phi)) \right\|_2^2 \right\}. \tag{C.6}$$

Proof. We first prove that the set in (C.5) is a subset of the set in (C.6). For an arbitrary $\phi_0 \in$ (C.5) and $1 \leq i \leq n$, we have that

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

Here, the first line follows from the definition of $c_i^{(0)}$ in step 2 of Algorithm 1, and step *a.* follows from the definition of the argmin function. Step *b.* follows from the assumption that $\phi_0 \in$ (C.5) satisfies $c_i^{(0)}(x'(\phi_0)) = c_i^{(0)}(x)$. Because this holds for an arbitrary $1 \leq i \leq n$, we have proven that $\phi_0 \in$ (C.5) $\implies$ $\phi_0 \in$ (C.6); or equivalently, (C.5) $\subseteq$ (C.6).

We proceed to prove the other direction. For an arbitrary $\phi_0 \in$ (C.6) and an arbitrary $1 \leq i \leq n$, we have that

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

Here, step *a.* follows from the definition of argmin, and step *b.* follows from combining the definition of $c_i^{(0)}(x'(\phi))$ in step 2 of Algorithm 1. We conclude that $\phi_0 \in$ (C.6) $\implies$ $\phi_0 \in$ (C.5).

Combining these two directions, we have proven that (C.6) = (C.5). □

Lemma 13. *Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned in the tth iteration of step 3b. of Algorithm 1 applied to data x, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from x during step 1 of Algorithm 1. For $\mathcal{S}_1$ defined as*

$$\mathcal{S}_1 = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{1} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}, \tag{C.7}$$

and $w_i^{(t)}(k)$ defined in (4.13), we have that

$$\begin{aligned}\mathcal{S}_1 = &\bigcap_{i=1}^{n}\bigcap_{k=1}^{K}\left\{\phi : \left\|\left[x'(\phi)\right]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi))\right\|_2^2 \le \left\|\left[x'(\phi)\right]_i - m_k^{(0)}(x'(\phi))\right\|_2^2\right\} \cap \\ &\left(\bigcap_{i=1}^{n}\bigcap_{k=1}^{K}\left\{\phi : \left\|\left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(0)}\left(c_i^{(1)}(x)\right)\left[x'(\phi)\right]_{i'}\right\|_2^2 \le \left\|\left[x'(\phi)\right]_i - \sum_{i'=1}^{n} w_{i'}^{(0)}(k)\left[x'(\phi)\right]_{i'}\right\|_2^2\right\}\right).\end{aligned} \tag{C.8}$$

Proof. We first prove that (C.7) $\subseteq$ (C.8). For an arbitrary $\phi_0 \in$ (C.7) and an arbitrary $1 \le i \le n$, we have that

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

In the equations above, the first line follows from step 3b. of Algorithm 1 with $t = 0$. Next, step *a.* follows from the definition of (C.7), which implies that $c_i^{(1)}(x'(\phi_0)) = c_i^{(1)}(x)$. Step *b.*

is a direct consequence of step 3a. of Algorithm 1 with $t = 0$. In steps $c.$ and $d.$, we used the definitions of the argmin function and (C.7). Finally, we apply the definition of $w_i^{(t)}$ in (4.13) to get $e.$ Because this holds for an arbitrary $1 \leq i \leq n$, $\phi_0 \in$ (C.7) implies that ϕ_0 is an element of the second set in the intersection in (C.8).

Moreover, $\phi_0 \in$ (C.7) implies that $\phi_0 \in$ (C.5), which, according to Lemma 12, further implies that ϕ_0 is an element of the first set in the intersection in (C.8). To summarize, we have proven that $\phi_0 \in$ (C.7) $\implies$ $\phi_0 \in$ (C.8), and as a result, (C.7) $\subseteq$ (C.8).

Next, we prove that the set in (C.8) is a subset of the set in (C.7). For an arbitrary $\phi_0 \in$ (C.8) and an arbitrary $1 \leq i \leq n$, we have that

$$
\begin{aligned}
\phi_0 \in \text{(C.8)} &\overset{a.}{\implies} \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = c_i^{(1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = c_i^{(1)}(x)\right\}} \right\|_2^2 \leq \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x) = k\right\}} \right\|_2^2, \forall k \in [K] \\
&\overset{b.}{\implies} \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = c_i^{(1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = c_i^{(1)}(x)\right\}} \right\|_2^2 \leq \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}} \right\|_2^2, \forall k \in [K] \\
&\overset{c.}{\implies} c_i^{(1)}(x) = \underset{1\leq k\leq K}{\operatorname{argmin}} \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(0)}(x'(\phi_0)) = k\right\}} \right\|_2^2 \\
&\overset{d.}{\implies} c_i^{(1)}(x) = \underset{1\leq k\leq K}{\operatorname{argmin}} \left\| [x'(\phi_0)]_i - m_k^{(1)}(x'(\phi_0)) \right\|_2^2 \\
&\overset{e.}{\implies} c_i^{(1)}(x) = c_i^{(1)}(x'(\phi_0)).
\end{aligned}
$$

Here, step $a.$ follows from the definition of (C.8). In step $b.$, we first apply Lemma 12, which implies that (C.8) $\subseteq$ (C.6). Therefore, $\phi_0 \in$ (C.8) $\implies$ $c_i^{(0)}(x) = c_i^{(0)}(x'(\phi_0))$, for all $i = 1, \ldots, n, k = 1, \ldots, K$, yielding the desired equality. Next, step $c.$ follows from the definition of the argmin function. Finally, steps $d.$ and $e.$ follow directly from the definitions of $m_k^{(t)}$ and $c_i^{(t)}$ in steps 3a. and 3b. of Algorithm 1, respectively.

Because the result above holds for an arbitrary i, we have that $\phi_0 \in$ (C.8) $\implies$ $c_i^{(1)}(x) = c_i^{(1)}(x'(\phi))$, $i = 1, \ldots, n$. Combining this result with the observation that (C.8) $\subseteq$ (C.6), we have that (C.8) $\subseteq$ (C.7), which concludes the proof. □

Next, we will prove the inductive step in the proof of Proposition 14, which relies on the

following claim.

Lemma 14. *Recall that $c_i^{(t)}(x)$ denotes the cluster to which the ith observation is assigned in the tth iteration of Algorithm 1 applied to the data x, and that $m_k^{(0)}(x)$ denotes the kth centroid sampled from x during initialization. For some $1 \leq \tilde{T} \leq T-1$, define*

$$\mathcal{S}_{\tilde{T}} = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{\tilde{T}} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}. \tag{C.9}$$

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

$$\begin{aligned} \mathcal{S}_{\tilde{T}} = &\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi)) \right\|_2^2 \leq \left\| [x'(\phi)]_i - m_k^{(0)}(x'(\phi)) \right\|_2^2 \right\} \\ \cap &\left(\bigcap_{t=1}^{\tilde{T}} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}\left(c_i^{(t)}(x)\right)[x'(\phi)]_{i'} \right\|_2^2 \leq \left\| [x'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)[x'(\phi)]_{i'} \right\|_2^2 \right\} \right), \end{aligned} \tag{C.10}$$

where $w_i^{(t)}(\cdot)$ is defined in (4.13). Then, for $\mathcal{S}_{\tilde{T}+1}$ defined as

$$\mathcal{S}_{\tilde{T}+1} = \left\{ \phi \in \mathbb{R} : \bigcap_{t=0}^{\tilde{T}+1} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(x'(\phi)\right) = c_i^{(t)}\left(x\right) \right\} \right\}, \tag{C.11}$$

we have that

$$\begin{aligned} \mathcal{S}_{\tilde{T}+1} = &\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - m^{(0)}_{c_i^{(0)}(x)}(x'(\phi)) \right\|_2^2 \leq \left\| [x'(\phi)]_i - m_k^{(0)}(x'(\phi)) \right\|_2^2 \right\} \\ &\cap \left(\bigcap_{t=1}^{\tilde{T}+1} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [x'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}\left(c_i^{(t)}(x)\right)[x'(\phi)]_{i'} \right\|_2^2 \leq \left\| [x'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k)[x'(\phi)]_{i'} \right\|_2^2 \right\} \right). \end{aligned} \tag{C.12}$$

Proof. Using the definitions in (C.9) and (C.11), we have that

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

Therefore, it suffices to prove that (C.13) = (C.12), under the inductive hypothesis

(C.10).

We start by proving that (C.13) $\subseteq$ (C.12). For an arbitrary $\phi_0 \in$ (C.13) and an arbitrary $1 \leq i \leq n$, we have that

$$
\begin{aligned}
& c_i^{(\tilde{T}+1)}\left(x'(\phi_0)\right) = c_i^{(\tilde{T}+1)}\left(x\right) \overset{a.}{\implies} c_i^{(\tilde{T}+1)}(x) = \underset{1\leq k\leq K}{\operatorname{argmin}} \left\| [x'(\phi_0)]_i - m_k^{(\tilde{T}+1)}(x'(\phi_0)) \right\|_2^2 \\
& \overset{b.}{\implies} \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = c_i^{(\tilde{T}+1)}(x)\right\}} \right\|_2^2 \leq \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x'(\phi_0)) = k\right\}} \right\|_2^2 \\
& \overset{c.}{\implies} \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}} \right\|_2^2 \leq \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}} \right\|_2^2 \\
& \overset{d.}{\implies} \phi_0 \in \text{(C.12)}.
\end{aligned}
$$

Here, the first statement follows from the definition of $\mathcal{S}_{\tilde{T}+1}$. Next, steps *a.* and *b.* follow from the definitions of $c_i^{(\tilde{T}+1)}$ and $m_k^{(\tilde{T}+1)}(x'(\phi_0))$ in steps 3b. and 3a. of Algorithm 1, respectively. In step *c.*, we used the fact that $\phi_0 \in$ (C.13) $\implies \phi_0 \in \mathcal{S}_{\tilde{T}} \implies c_i^{\tilde{T}}(x'(\phi_0)) = c_i^{\tilde{T}}(x)$. Finally, *d.* follows from the definition of $w_i^{(t)}$ in (4.13).

We continue with the reverse direction. Applying the inductive hypothesis (C.10), together with the definition of $S_{\tilde{T}+1}$ in (C.12) and the definition of $w_i^{(t)}$ in (4.13), we have that

$$
\begin{aligned}
\text{(C.12)} = S_{\tilde{T}} \cap \Bigg(\bigcap_{i=1}^n \bigcap_{k=1}^K \Bigg\{ \phi : \Bigg\| [x'(\phi)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}} \Bigg\|_2^2 \leq \\
\Bigg\| [x'(\phi)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}[x'(\phi)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}} \Bigg\|_2^2 \Bigg\} \Bigg).
\end{aligned}
\tag{C.14}
$$

For an arbitrary $\phi_0 \in$ (C.12) and any $1 \leq i \leq n$, the following holds:

$$\left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = c_i^{(\tilde{T}+1)}(x)\right\}} \right\|_2^2 \leq \left\| [x'(\phi_0)]_i - \frac{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}[x'(\phi_0)]_{i'}}{\sum_{i'=1}^n 1\left\{c_{i'}^{(\tilde{T})}(x) = k\right\}} \right\|_2^2, \forall k \in [K]$$

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

$$\overset{b.}{\Longrightarrow} c_i^{(\tilde{T}+1)}(x) = \underset{1\leq k\leq K}{\operatorname{argmin}} \left\| [x'(\phi_0)]_i - m_k^{(\tilde{T}+1)}(x'(\phi_0)) \right\|_2^2$$

$$\overset{c.}{\Longrightarrow} c_i^{(\tilde{T}+1)}(x) = c_i^{(\tilde{T}+1)}(x'(\phi_0)).$$

Here, to derive step *a.*, we first note that by (C.14), any element ϕ_0 of (C.12) is also an element of $\mathcal{S}_{\tilde{T}}$. Therefore, using the definition of $\mathcal{S}_{\tilde{T}}$ in (C.9), we have that $\bigcap_{t=1}^{\tilde{T}} \left\{c_i^{(t)}\left(x'(\phi_0)\right) = c_i^{(t)}\left(x\right)\right\}$, and step *a.* follows directly. Next, steps *b.* and *c.* follow directly from steps 3a. and 3b. of Algorithm 1 with $t = \tilde{T}$. By inspecting the form of (C.13), we conclude that $\phi_0 \in$ (C.12) $\implies$ $\phi_0 \in$ (C.13).

In conclusion, we have proven that (C.12) = (C.13), which completes the proof. □

The inductive proof of Proposition 14 follows from combining Lemmas 12, 13 and 14.

C.3 Proof of Lemmas 1 and 2

We first prove Lemma 1, which is also Lemma 2 in Gao et al. [2020].

Proof. We first express the inner product $\langle [x'(\phi)]_i, [x'(\phi)]_j \rangle$ as a function of ϕ. From (4.11), we have that $[x'(\phi)]_i = x_i + \nu_i \left(\frac{\phi - \|x^\top \nu\|_2}{\|\nu\|_2^2}\right) \operatorname{dir}(x^\top \nu) = x_i - \nu_i \frac{\|x^\top \nu\|_2}{\|\nu\|_2^2} \operatorname{dir}(x^\top \nu) + \left(\frac{\nu_i}{\|\nu\|_2^2}\phi\right) \operatorname{dir}(x^\top \nu)$.

Therefore,

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

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

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

□

This completes the proof of Lemma 1.

We continue with the proof of Lemma 2. Using the definition of $w_i^{(t-1)}(k)$ in (4.13), we

have that

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

where

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

is a linear function of ϕ. The rest of the proof follows directly from the same set of calculations in the proof of Lemma 1.

C.4 Proof of Proposition 15

Recall that n, q, K, T denote the number of samples (see (4.1)), the number of features (see (4.1)), the number of clusters (see Algorithm 1), and the maximum number of iterations for which Algorithm 1 is run.

According to Proposition 14, to compute the set $\mathcal{S}_T$ in (4.12), it suffices to compute the intersection of the two sets in (4.14) and (4.15).

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

- Observation 1: according to Lemma 1, the set in (4.14) is an intersection of nK quadratic inequalities.
- Observation 2: according to Lemma 2, the set in (4.15) is an intersection of nKT quadratic inequalities.
- Observation 3: we can solve a quadratic inequality in $\mathcal{O}(1)$ time using the quadratic formula.
- Observation 4: we can intersect the solution sets of N quadratic inequalities in $\mathcal{O}(N \log N)$ time [Bourgon, 2020].

Equipped with these observations, we will analyze the timing complexity of computing the set (4.14). Note that the coefficients for each of the nK quadratic inequalities can be computed in $\mathcal{O}(nq)$ operations: first, using the property that $x^\top \nu = \sum_{i \in \hat{\mathcal{C}}_1} x_i/|\hat{\mathcal{C}}_1| - \sum_{i \in \hat{\mathcal{C}}_2} x_i/|\hat{\mathcal{C}}_2|$, we can compute $\|x^\top \nu\|_2$ and $\mathrm{dir}(x^\top \nu)$ in $\mathcal{O}(nq)$ operations. Then, computing the coefficients a, b, and γ in Lemma 1 takes $\mathcal{O}(1)$, $\mathcal{O}(q)$, and $\mathcal{O}(q)$ operations, respectively. For each inequality, obtaining the solution set requires $\mathcal{O}(1)$ operations (see Observation 3). Finally, intersecting the solution sets of the $n(K-1)$ quadratic inequalities incurs another $\mathcal{O}(nK \log(nK))$ operations. Thus, the computational cost for (4.14) totals to $\mathcal{O}(nKq + nK \log(nK))$ operations.

Next, we analyze the cost of computing the set (4.15). Note that using Observation 2, we need to solve nKT quadratic inequalities. Here, for each quadratic inequality of the form in Lemma 2, it takes $\mathcal{O}(n), \mathcal{O}(n+q)$, and $\mathcal{O}(n+q)$ operations to compute the coefficients $\tilde{a}, \tilde{b}$, and $\tilde{\gamma}$, respectively. Therefore, obtaining the nKT solution sets will take $\mathcal{O}(nKT(n+q))$ time. Finally, intersecting these sets using Observation 4 adds another $\mathcal{O}(nKT \log(nKT))$ operations.

Combining the costs for computing the set in (4.14) and the set in (4.15), we conclude that the cost for computing the set $\mathcal{S}_T$ in (4.12) is $\mathcal{O}(nKT(n+q) + nKT \log(nKT))$ operations.

C.5 Proof of Proposition 16 and computation of $p_{\Sigma,\text{selective}}$

The proof of Proposition 16 is similar to that of Proposition 13.

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

$$X = \Pi_\nu^\perp X + \frac{\nu\nu^\top X \Sigma^{-\frac{1}{2}} \Sigma^{\frac{1}{2}}}{\|\nu\|_2^2} = \Pi_\nu^\perp X + \left(\frac{\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2}{\|\nu\|_2^2} \right) \nu \left\{ \mathrm{dir}\left(\Sigma^{-\frac{1}{2}} X^\top \nu \right) \right\}^\top \Sigma^{\frac{1}{2}}. \tag{C.15}$$

Lemma 15. *Under (4.16) and $H_0 : \mu^\top \nu = 0_q$, $\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2$, $\Pi_\nu^\perp X$, and $\mathrm{dir}\left(\Sigma^{-\frac{1}{2}} X^\top \nu\right)$ are pairwise independent.*

Proof. As in the proof of Lemma 11, $\Pi_\nu^\perp \nu = 0_n$, and it follows from the property of the

matrix normal distribution that $X^\top \nu$ is independent of $\Pi_\nu^\perp X$. Because both $\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2$ and $\mathrm{dir}(\Sigma^{-\frac{1}{2}} X^\top \nu)$ are functions of $X^\top \nu$, both are independent of $\Pi_\nu^\perp X$.

Next, we will show that $\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2$ and $\mathrm{dir}(\Sigma^{-\frac{1}{2}} X^\top \nu)$ are independent. Under (4.16) and $H_0 : \mu^\top \nu = 0_q$, we have that $\Sigma^{-\frac{1}{2}} X^\top \nu \sim \mathcal{N}(0_q, \|\nu\|_2^2 \mathbf{I}_q)$. It then follows that $\Sigma^{-\frac{1}{2}} X^\top \nu$ is rotationally invariant, and therefore $\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2$ is independent of $\mathrm{dir}(\Sigma^{-\frac{1}{2}} X^\top \nu)$ [Bilodeau and Brenner, 1999]. □

Then, recalling the definition of $p_{\Sigma,\text{selective}}$ in (4.18), we have that

$$
\begin{aligned}
p_{\Sigma,\text{selective}} &= \mathbb{P}_{H_0}\Big[\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2 \geq \|\Sigma^{-\frac{1}{2}} x^\top \nu\|_2 \mid \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}(X) = c_i^{(t)}(x) \right\}, \Pi_\nu^\perp X = \Pi_\nu^\perp x, \\
&\qquad \mathrm{dir}\left(\Sigma^{-\frac{1}{2}} X^\top \nu\right) = \mathrm{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu\right)\Big] \\
&\overset{a.}{=} \mathbb{P}_{H_0}\left[\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2 \geq \|\Sigma^{-\frac{1}{2}} x^\top \nu\|_2 \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(\Pi_\nu^\perp X + \frac{\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2}{\|\nu\|_2^2} \nu \left\{ \mathrm{dir}\left(\Sigma^{-\frac{1}{2}} X^\top \nu\right) \right\}^\top \Sigma^{\frac{1}{2}} \right) = c_i^{(t)}(x) \right\}, \right. \\
&\qquad \left. \Pi_\nu^\perp X = \Pi_\nu^\perp x,\, \mathrm{dir}(\Sigma^{-\frac{1}{2}} X^\top \nu) = \mathrm{dir}(\Sigma^{-\frac{1}{2}} x^\top \nu) \right] \\
&\overset{b.}{=} \mathbb{P}_{H_0}\left[\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2 \geq \|\Sigma^{-\frac{1}{2}} x^\top \nu\|_2 \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(\Pi_\nu^\perp x + \frac{\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2}{\|\nu\|_2^2} \nu \left\{ \mathrm{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu\right) \right\}^\top \Sigma^{\frac{1}{2}} \right) = c_i^{(t)}(x) \right\}, \right. \\
&\qquad \left. \Pi_\nu^\perp X = \Pi_\nu^\perp x,\, \mathrm{dir}(\Sigma^{-\frac{1}{2}} X^\top \nu) = \mathrm{dir}(\Sigma^{-\frac{1}{2}} x^\top \nu) \right] \\
&\overset{c.}{=} \mathbb{P}_{H_0}\left[\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2 \geq \|\Sigma^{-\frac{1}{2}} x^\top \nu\|_2 \;\middle|\; \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)}\left(\Pi_\nu^\perp x + \frac{\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2}{\|\nu\|_2^2} \nu \left\{ \mathrm{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu\right) \right\}^\top \Sigma^{\frac{1}{2}} \right) = c_i^{(t)}(x) \right\} \right].
\end{aligned}
$$

Here, step *a.* follows from substituting X with the expression in (C.15). Step *b.* follows from replacing $\Pi_\nu^\perp X$ and $\mathrm{dir}(\Sigma^{-\frac{1}{2}} X^\top \nu)$ with $\Pi_\nu^\perp x$ and $\mathrm{dir}(\Sigma^{-\frac{1}{2}} x^\top \nu)$, respectively. Finally, in step *c.*, we used Lemma 15. Now, under (4.16) and $H_0 : \mu^\top \nu = 0_q$, we have that $\|\Sigma^{-\frac{1}{2}} X^\top \nu\|_2 \sim \|\nu\|_2 \chi_q$, which concludes the proof of (4.19).

It remains to show that the test that rejects $H_0 : \mu^\top \nu = 0$ when $p_{\Sigma,\text{selective}} \leq \alpha$ controls the selective Type I error, in the sense of (4.5). We omit the proof here, as it follows directly from the proof of Proposition 13 in Appendix C.1.

Next, we discuss how we could modify the results in Section 4.3 to compute the p-value

$p_{\Sigma,\text{selective}}$. First note that according to Proposition 16, it suffices to compute the set

$$\mathcal{S}_T^{\Sigma} = \left\{ \phi : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ c_i^{(t)} \left(\Pi_\nu^{\perp} x + \left(\frac{\phi}{\|\nu\|_2^2} \right) \nu \left\{ \operatorname{dir}\left(\Sigma^{-\frac{1}{2}} x^\top \nu \right) \right\}^\top \Sigma^{\frac{1}{2}} \right) = c_i^{(t)}(x) \right\} \right\}. \tag{C.16}$$

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

$$[\tilde{x}'(\phi)]_i = x_i - \nu_i \frac{||x^\top \nu||_2}{||\nu||_2^2} \operatorname{dir}(x^\top \nu) + \left(\frac{||x^\top \nu||_2}{||\Sigma^{-\frac{1}{2}} x^\top \nu||_2} \frac{\nu_i}{||\nu||_2^2} \phi \right) \operatorname{dir}(x^\top \nu). \tag{C.17}$$

Therefore, a minor modification of Proposition 14 yields the following corollary.

Corollary 2. *Suppose the k-means clustering algorithm (see Algorithm 1) with K clusters the data x, when applied to the data x, runs for T steps. Then, for the set $\mathcal{S}_T^{\Sigma}$ defined in (C.16), we have that*

$$\begin{aligned} \mathcal{S}_T^{\Sigma} = & \left(\bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [\tilde{x}'(\phi)]_i - m^{(0)}_{c_i^{(0)}(x)}(\phi) \right\|_2^2 \le \left\| [\tilde{x}'(\phi)]_i - m_k^{(0)}(\phi) \right\|_2^2 \right\} \right) \cap \\ & \left(\bigcap_{t=1}^{T} \bigcap_{i=1}^{n} \bigcap_{k=1}^{K} \left\{ \phi : \left\| [\tilde{x}'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)} \left(c_{i'}^{(t-1)}(x) \right) [\tilde{x}'(\phi)]_{i'} \right\|_2^2 \le \left\| [\tilde{x}'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k) [\tilde{x}'(\phi)]_{i'} \right\|_2^2 \right\} \right). \end{aligned} \tag{C.18}$$

We also have the following extensions of Lemmas 1 and 2, which enable efficient computation of the expressions in Corollary 2.

Lemma 16 (Section 4.2 in Gao et al. [2020])**.** *For $\tilde{x}'(\phi)$ in (C.17) and ν in (4.3), $\left\| [\tilde{x}'(\phi)]_i - [\tilde{x}'(\phi)]_j \right\|_2^2 = a'\phi^2 + b'\phi + \gamma'$, where $a' = \left(\frac{||x^\top \nu||_2}{||\Sigma^{-\frac{1}{2}} x^\top \nu||_2} \right)^2 \left(\frac{\nu_i - \nu_j}{\|\nu\|_2^2} \right)^2$, $b' = 2 \left(\frac{||x^\top \nu||_2}{||\Sigma^{-\frac{1}{2}} x^\top \nu||_2} \right) \left(\frac{\nu_i - \nu_j}{\|\nu\|_2^2} \langle x_i - x_j, \operatorname{dir}(x^\top \nu) \rangle - \left(\frac{\nu_i - \nu_j}{\|\nu\|_2^2} \right)^2 \|x^\top \nu\|_2 \right)$, and $\gamma' = \left\| x_i - x_j - (\nu_i - \nu_j) \frac{x^\top \nu}{\|\nu\|_2^2} \right\|_2^2$.*

Lemma 17. *For $\tilde{x}'(\phi)$ in (C.17), ν in (4.3), and $w_i^{(t)}(k)$ in (4.13), $\left\| [\tilde{x}'(\phi)]_i - \sum_{i'=1}^{n} w_{i'}^{(t-1)}(k) [\tilde{x}'(\phi)]_{i'} \right\|_2^2 = \tilde{a}'\phi^2 + \tilde{b}'\phi + \tilde{\gamma}'$, where*

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

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

and

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

Proofs of Lemmas 16 and 17 follow from the same set of calculations in the proofs of Lemmas 1 and 2 in Appendix C.3.

C.6 Proof of Proposition 17

Proof of Proposition 17 is similar to the proof of Lemma 1 in Markovic et al. [2017] and the proof of Lemma 7 in Tibshirani et al. [2018b].

We first present an auxiliary lemma.

Lemma 18. *For any $c_i^{(t)}(x), i = 1, \ldots, n; t = 1, \ldots, T$, $\hat{p}_{selective}(\hat{\sigma})$ defined in (4.20) is a continuous and monotonically increasing function of $\hat{\sigma}$.*

Proof. By the definition in (4.20), we have that

$$\hat{p}_{\text{selective}}(\hat{\sigma}) = \frac{\int_{\|x^\top \nu\|_2}^\infty (\frac{1}{2})^{q/2-1} \frac{t^{q-1}}{\Gamma(q/2)} \|\nu\|_2^{-q} \hat{\sigma}^{-q} \exp\left(-\frac{t^2}{2\hat{\sigma}^2\|\nu\|_2^2}\right) 1\{t \in \mathcal{S}_T\} dt}{\int_0^\infty (\frac{1}{2})^{q/2-1} \frac{t^{q-1}}{\Gamma(q/2)} \|\nu\|_2^{-q} \hat{\sigma}^{-q} \exp\left(-\frac{t^2}{2\hat{\sigma}^2\|\nu\|_2^2}\right) 1\{t \in \mathcal{S}_T\} dt}, \tag{C.19}$$

where $\mathcal{S}_T$ defined in (4.12) is a function of $c_i^{(t)}(x), i = 1, \ldots, n; t = 1, \ldots, T$. By inspection, (C.19) is a continuous function of $\hat{\sigma}$, because the product or ratio of two continuous functions is still continuous. It remains to show that (C.19) is increasing in $\hat{\sigma}$. This follows directly from Lemma S3. of Gao et al. [2020]. □

Provided that $\hat{\sigma}$ converges to σ in probability, we can combine Lemma 18 and the continuous mapping theorem to see that $\hat{p}_{\text{selective}}(\hat{\sigma})$ converges to $p_{\text{selective}}(\sigma)$ in probability, i.e., for all $\epsilon > 0, \lim_{q\to\infty} \mathbb{P}(|\hat{p}_{\text{selective}}(\hat{\sigma}) - p_{\text{selective}}(\sigma)| \geq \epsilon) = 0$. Next, letting A_q denote the event $\bigcap_{t=0}^T \bigcap_{i=1}^n \left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}$, we will show that under the assumptions in Proposition 17, $\hat{p}_{\text{selective}}(\hat{\sigma})$ converges to $p_{\text{selective}}(\sigma)$ in probability, *conditional on* A_q. For any $\epsilon > 0$,

we have that

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

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

Finally, we have that

$$\lim_{q\to\infty} \mathbb{P}_{H_0^{(q)}}\{\hat{p}_{\text{selective}}(\hat{\sigma}) \le \alpha \mid A_q\} \overset{a.}{=} \lim_{q\to\infty} \mathbb{P}_{H_0^{(q)}}\{p_{\text{selective}}(\sigma) \le \alpha \mid A_q\} \overset{b.}{=} \lim_{q\to\infty} \alpha = \alpha. \quad \text{(C.20)}$$

Here, step $a.$ follows from $\hat{p}_{\text{selective}}(\hat{\sigma})$ converging to $p_{\text{selective}}(\sigma)$ in probability, *conditional on* A_q. Step $b.$ follows from the fact that the result of Proposition 13 applies for any positive integer q. This completes the proof of Proposition 17.

Proposition 17 assumes that we have a consistent estimator $\hat{\sigma}$ of σ. In Appendix C.7, we analyze different estimators of σ in (4.1), and prove that, under appropriate sparsity assumptions on μ in (4.1), $\hat{\sigma}_{\text{MED}}$ in (4.21) is a consistent estimator for σ.

As an alternative, we can also use an asymptotically conservative estimator of σ as in Gao et al. [2020]. This leads to an asymptotically conservative p-value; details are stated in Corollary 3.

Corollary 3. *For $q = 1, 2, \ldots$, suppose that $X^{(q)} \sim \mathcal{MN}_{n\times q}\left(\mu^{(q)}, \mathbf{I}_n, \sigma^2\mathbf{I}_q\right)$. Let $x^{(q)}$ be a realization from $X^{(q)}$ and $c_i^{(t)}(\cdot)$ be the cluster to which the ith observation is assigned during the tth iteration of step 3b. of Algorithm 1. Consider the sequence of null hypotheses $H_0^{(q)} : {\mu^{(q)}}^\top \nu^{(q)} = 0_q$, where $\nu^{(q)}$ defined in (4.3) is the contrast vector resulting from applying k-means clustering on $x^{(q)}$. Suppose that (i) $\hat{\sigma}$ is an asymptotically conservative estimator of σ, i.e., $\lim_{q\to\infty} \mathbb{P}\left(\hat{\sigma}(X^{(q)}) \geq \sigma\right) = 1$; and (ii) there exists $\delta \in (0,1)$ such that $\lim_{q\to\infty} \mathbb{P}_{H_0^{(q)}}\left[\bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] > \delta$. Then, $\forall \alpha \in (0,1)$, we have that $\lim_{q\to\infty} \mathbb{P}_{H_0^{(q)}}\left[\hat{p}_{selective}(\hat{\sigma}) \leq \alpha \,\middle|\, \bigcap_{t=0}^{T}\bigcap_{i=1}^{n}\left\{c_i^{(t)}\left(X^{(q)}\right) = c_i^{(t)}\left(x^{(q)}\right)\right\}\right] \leq \alpha$.*

We omit the proof of Corollary 3, as it follows directly from combining the proof of Proposition 17 and the fact that $\hat{p}_{\text{selective}}(\hat{\sigma})$ is a monotonically increasing function of $\hat{\sigma}$ (see Lemma 18).

Finally, we remark that, in principle, the result in Proposition 17 can be extended to an unknown covariance matrix Σ. However, estimating Σ is challenging, especially when q is comparable to, or larger than, n [Avella-Medina et al., 2018, Bickel and Levina, 2008, Rousseeuw, 1987]. It may be possible to leverage recent advances in robust covariance matrix estimation (e.g., Belomestny et al. [2019], Chen et al. [2018], Han and Liu [2014]) to obtain a consistent estimator of Σ under model (4.16).

C.7 Estimating σ in (4.1)

Proposition 17 states that, under appropriate assumptions, a consistent estimator of σ in (4.1) leads to asymptotic selective Type I error control. In this section, we analyze the asymptotic behavior of the two variance estimators considered in Section 4.5, $\hat{\sigma}^2_{\text{MED}}$ and $\hat{\sigma}^2_{\text{Sample}}$. In particular, we prove that under model (4.1) and a sparsity assumption on μ (defined in (4.1)), a close analog of $\hat{\sigma}^2_{\text{MED}}$ in (4.21) that does not subtract the column median

is a consistent estimator of σ^2. Moreover, we prove that $\hat{\sigma}^2_{\text{Sample}}$ is a conservative estimator of σ^2, and characterize its exact bias.

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

Lemma 19 (Proposition 6 in Comminges et al. [2021])**.** *Consider the model*

$$Y_i = \theta_i + \sigma\xi_i, \quad i = 1, \ldots, d, \tag{C.21}$$

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

$$\bar{\sigma}^2_{MED} = \text{median}(Y_1^2, \ldots, Y_d^2)/M_{\xi_1^2}. \tag{C.22}$$

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

$$\sup_{\sigma>0} \sup_{\|\theta\|_0 \leq s} \frac{1}{\sigma^2} \mathrm{E}\left\{\left|\bar{\sigma}^2_{MED} - \sigma^2\right|\right\} \leq C \max\left(\frac{1}{d^{1/2}}, \frac{s}{d}\right). \tag{C.23}$$

Building on Lemma 19, in Corollary 4, we analyze the properties of an estimator closely related to $\hat{\sigma}^2_{\text{MED}}$ in (4.21). In particular, this estimator $\tilde{\sigma}^2_{\text{MED}}$ does not subtract the median of each column in the input data. While $\hat{\sigma}^2_{\text{MED}}$ and $\tilde{\sigma}^2_{\text{MED}}$ are very similar provided that μ is sparse, we expect $\hat{\sigma}^2_{\text{MED}}$ to perform better empirically in scenarios where μ is sparse *up to a constant shift*, i.e., there exists a matrix C such that (i) each column of C takes on the same value; and (ii) $\mu + C$ is sparse.

Corollary 4. *Under model (4.1), consider*

$$\tilde{\sigma}^2_{MED}(X) = \left\{ \underset{1\le i\le n, 1\le j\le q}{\text{median}} \left(X_{ij}^2\right) \right\} / M_{\chi_1^2}, \tag{C.24}$$

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

$$\sup_{\sigma>0} \sup_{\max_{1\le i\le n} \|\mu_i\|_0 \le s} \frac{1}{\sigma^2} \mathrm{E}\left\{ \left| \tilde{\sigma}^2_{MED} - \sigma^2 \right| \right\} \le c_0 \max\left\{ \frac{1}{(nq)^{1/2}}, \frac{s}{q} \right\}. \tag{C.25}$$

Proof. First note that (4.1) can be re-written into the form of (C.21):

$$X_{ij} = \mu_{ij} + \sigma \xi_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, q, \tag{C.26}$$

where ξ_{ij} is independently and identically distributed as $\mathcal{N}(0, 1)$. Therefore, the estimator $\tilde{\sigma}^2_{\mathrm{MED}}(X)$ in (C.24) is the estimator (C.22) applied to the model (C.26). Moreover, $\max_{1\le i\le n} \|\mu_i\|_0 \le s$ implies that $\sum_{i=1}^n \sum_{j=1}^q 1\{\mu_{ij} \ne 0\} \le ns$. Applying Lemma 19, we have that

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

where c_0 is some universal constant. □

In words, Corollary 4 states that under model (4.1), the rate of convergence of $\tilde{\sigma}^2_{\mathrm{MED}}$ in mean (and therefore, in probability) is $\max\left\{1/(nq)^{1/2}, s/q\right\}$. In particular, $\tilde{\sigma}^2_{\mathrm{MED}}$ is a consistent estimator of σ^2 provided that $s/q \to 0$ as $q \to \infty$.

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

Proposition 20. *Under model (4.1), for $\hat{\sigma}^2_{Sample}(X) = \sum_{i=1}^n \sum_{j=1}^q \left(X_{ij} - \bar{X}_j\right)^2/(nq-q)$, we*

have that

$$\mathrm{E}\big\{\hat{\sigma}^2_{Sample}(X)\big\} - \sigma^2 = \frac{1}{2n(n-1)q}\sum_{j=1}^{q}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\mu_{ij}-\mu_{i'j})^2. \tag{C.27}$$

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

$$\sup_{\sigma>0}\sup_{\max_{1\le i\le n}\|\mu_i\|_0\le s}\frac{1}{\sigma^2}\mathrm{E}\big\{\big|\hat{\sigma}^2_{Sample}(X)-\sigma^2\big|\big\} \ge \tilde{c}_0\frac{s}{q}. \tag{C.28}$$

Proof. We start with the proof of (C.27). Under (4.1), the following holds:

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

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

To prove the second statement, we consider a specific matrix $\tilde{\mu}\in\mathbb{R}^{n\times q}$ with exactly $ns\le q$ non-zero entries. In addition, each column of $\tilde{\mu}$ has at most one non-zero entry and each row of $\tilde{\mu}$ has exactly s non-zero entries. This is possible because ns is assumed to be less

than q. Finally, we assume that the square of the minimal non-zero entry of $\tilde{\mu}$, $\min_{i,j:\tilde{\mu}_{ij}\neq 0}\tilde{\mu}_{ij}^2$, is lower bounded by some universal constant M. Then, we have that

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

Here, $a.$ follows from picking any $\tilde{\mu}$ satisfying the conditions outlined above, since by construction, $\max_{1\leq i\leq n}\|\tilde{\mu}_i\|_0=s$. Steps $b.$ and $c.$ follow from the inequality $\mathrm{E}(|X|)\geq\mathrm{E}(X)$ and the expression for $\mathrm{E}\left\{\hat{\sigma}^2_{\text{Sample}}(X)\right\}$ in (C.27), respectively. Finally, to prove $d.$, we note that for each of the ns columns with exactly one non-zero element, there are $n-1$ pairs of $(i,i'), i=1,\ldots,n; i'=1,\ldots,n$ such that the product $1\{\tilde{\mu}_{ij}\neq 0\}1\{\tilde{\mu}_{i'j}=0\}$ is non-zero. Moreover, each of pair contributes at least M by the assumption that $\min_{i,j:\tilde{\mu}_{ij}\neq 0}\tilde{\mu}_{ij}^2\geq M$. □

Contrasting the results in Corollary 4 and Proposition 20, we note that, under (4.1), the convergence of $\tilde{\sigma}^2_{\text{MED}}$ depends critically on the sparsity parameter s (or, equivalently, the ℓ_0 norm of μ_i), whereas the convergence of $\hat{\sigma}^2_{\text{Sample}}$ is determined by $\sum_{j=1}^{q}\sum_{i=1}^{n}\sum_{i'=1}^{n}(\mu_{ij}-\mu_{i'j})^2$. Thus, in scenarios where the underlying means $\mu_i, i=1,\ldots,n$ are sparse (e.g., (4.22) in Section 4.5), we expect $\tilde{\sigma}^2_{\text{MED}}$ (and therefore its "centered" analog $\hat{\sigma}^2_{\text{MED}}$ in (4.21)) to be a less conservative estimator of σ^2. As a result, we expect the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ to be more powerful than that based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$, as shown

in Figure 4.5 of Section 4.5.

C.8 Additional power comparisons

In Section 4.5.2, we compared the conditional power of the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ under (4.22). Here, we conduct two additional analyses.

In the first analysis, we consider a different notion of power that does not condition on $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ being true clusters. In this case, comparing the power of the tests requires a bit of care, because the effect size $\|\mu^\top \nu\|_2$ may differ across simulated datasets from the same data-generating distribution. As a result, we consider the power of the tests *as a function of* $\|\mu^\top \nu\|_2$. We fit a regression spline using the `gam` function in the `R` package `mgcv` [Wood, 2017] to obtain a smooth estimate of power on the same simulated datasets from Section 4.5.2. The results are in Figure C.1. The power of the tests that reject H_0 if $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, or $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ is less than $\alpha = 0.05$ increases as $\|\mu^\top \nu\|_2$ increases. For a given value of $\|\mu^\top \nu\|_2$ and σ, the test based on $p_{\text{selective}}$ has the highest power, followed by that based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$; the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ has the lowest power.

In the second analysis, we consider the conditional power (defined in (4.23)) of the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ under a different data generating model than (4.22). We generate data from (4.1) with $n = 150$ and

$$\mu_1 = \ldots = \mu_{\frac{n}{3}} = \begin{bmatrix} \theta_1 \\ 0_{0.9q} \end{bmatrix}, \; \mu_{\frac{n}{3}+1} = \ldots = \mu_{\frac{2n}{3}} = \begin{bmatrix} \theta_2 \\ 0_{0.9q} \end{bmatrix}, \; \mu_{\frac{2n}{3}+1} = \ldots = \mu_n = \begin{bmatrix} \theta_3 \\ 0_{0.9q} \end{bmatrix}, \quad \text{(C.29)}$$

where, q is taken to be a multiple of 10, and for $\delta > 0$, $\theta \in \mathbb{R}^{3\times 0.1q}$ has orthogonal rows, with $\|\theta_i\|_2^2 = \delta/2$ for $i = 1, 2, 3$. As in Section 4.5.2, we can think of $\mathcal{C}_1 = \{1, \ldots, n/3\}, \mathcal{C}_2 = \{(n/3) + 1, \ldots, (2n/3)\}, \mathcal{C}_3 = \{(2n/3) + 1, \ldots, n\}$ as "true clusters". Under (C.29), the pairwise distance between each pair of true clusters is δ.

We generate $M = 100,000$ datasets from (C.29) with $q = 50$,$\sigma = 0.25, 0.5, 1$, and $\delta =$

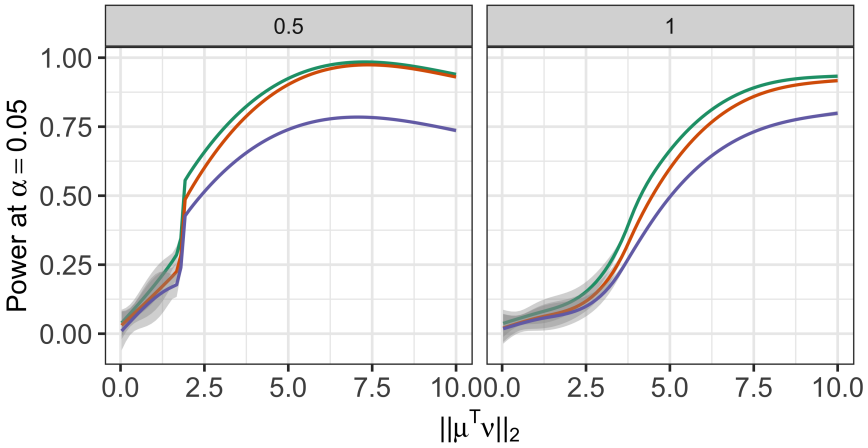


Figure C.1: *Left* : Additional analysis of the data in Section 4.5.2 with $\sigma = 0.5$. We fit a regression spline to display the power of the tests based on $p_{\text{selective}}$ (green line), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange line), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple line) as a function of $\|\mu^\top \nu\|_2$. *Right* : Same as left, but for $\sigma = 1$.

$2, 3, \ldots, 10$. For each simulated dataset, we apply k-means clustering with $K = 3$ and reject $H_0 : \mu^\top \nu = 0_q$ if $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, or $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ is less than $\alpha = 0.05$. Figure C.2(a) displays the detection probability (4.24) of k-means clustering as a function of δ in (C.29). Under model (4.1), the detection probability increases as a function of δ in (C.29) across all values of σ. For a given value of δ, a larger value of σ leads to lower detection probability. Figure C.2(b) displays the conditional power (4.23) for the tests based on $p_{\text{selective}}$, $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$, and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$. For some combinations of δ and σ, the conditional power is not displayed, because the true clusters are never recovered in simulation. For all tests and values of σ under consideration, conditional power is an increasing function of δ. For a given test and a value of δ, smaller σ leads to higher conditional power. Moreover, for the same values of δ and σ, the test based on $p_{\text{selective}}$ has the highest conditional power, followed closely by the test based of $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$. Using $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ leads to a less powerful test, especially for larger values of δ. As a comparison, we included the detection probability and conditional power under model (4.22) with $q = 50$ in panels (c) and (d) of Figure C.2. The tests under consideration behave qualitatively similarly as a function of δ

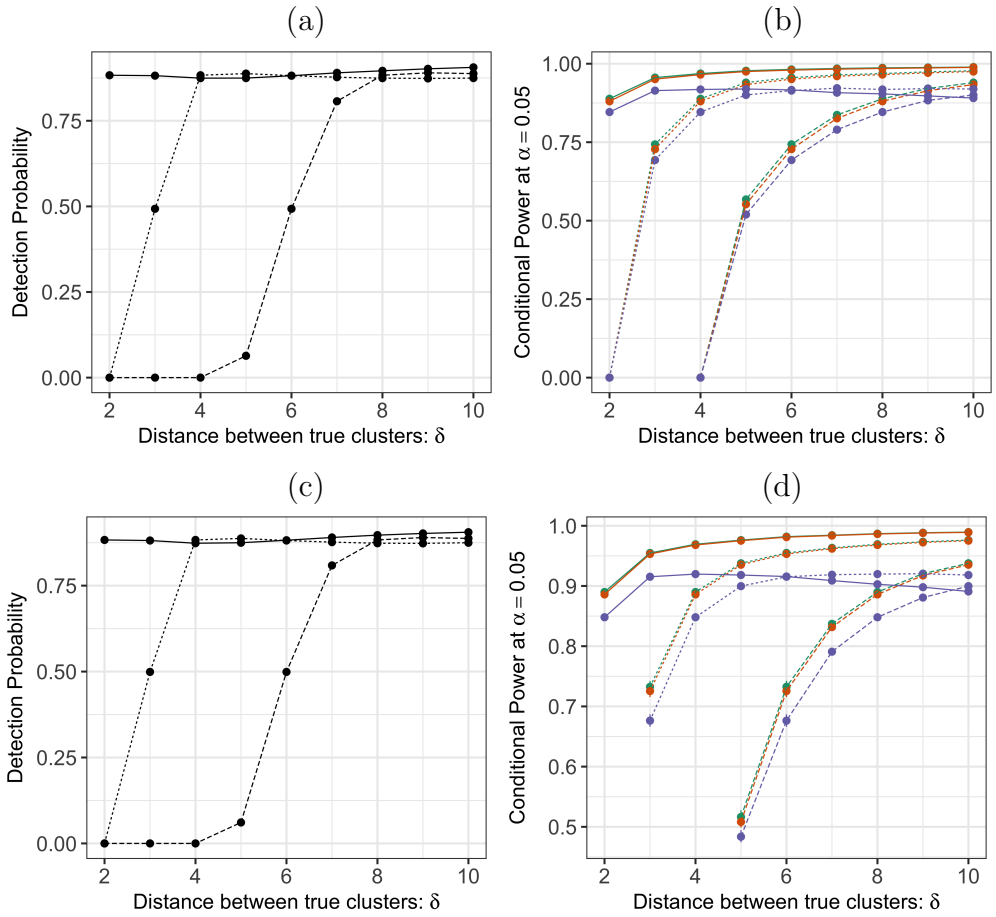


Figure C.2: *(a):* Detection probability defined in (4.24) for k-means clustering with $K = 3$ under model (4.1) with $n = 150$, $q = 50$, and μ in (C.29), across $\delta = \|\theta_i - \theta_j\|_2$ in (C.29) and $\sigma = 0.25$ (solid lines), 0.5 (dashed lines), and 1 (long-dashed lines). *(b):* The conditional power (4.23) at $\alpha = 0.05$ for the tests based on $p_{\text{selective}}$ (green), $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ (orange), and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ (purple), under model (4.1) with $n = 150$, $q = 50$, and μ in (C.29). *(c):* Same as (a), but for μ in (4.22). *(d):* Same as (b), but for μ in (4.22).

and σ. Under (4.22), we observe an even larger gap between the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{Sample}})$ and the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$.

C.9 Additional results for real data applications

In this section, we visualize the estimated clusters for the single cell RNA-sequencing data in Section 4.6.3.

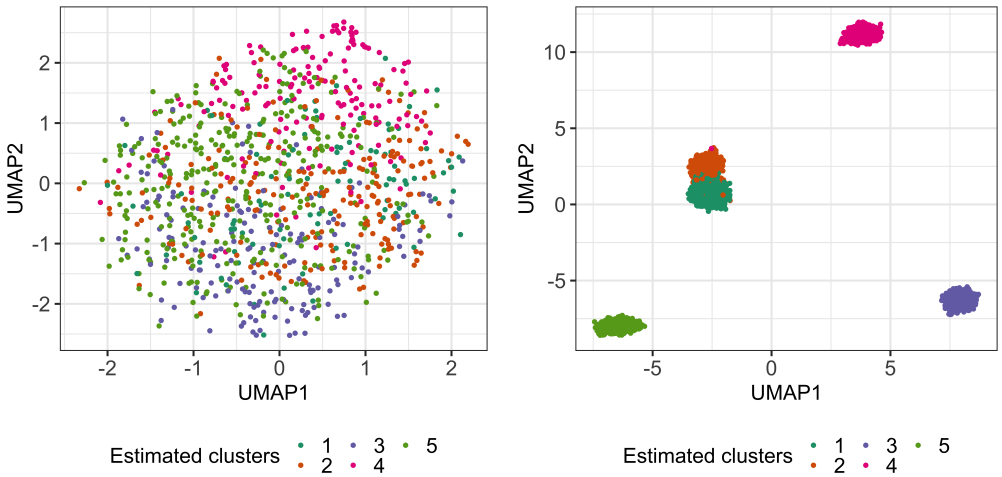


Figure C.3: *Left:* The two-dimensional UMAP embedding [McInnes et al., 2018] of the "no cluster" dataset after preprocessing (as described in Section 4.6.3), colored by the estimated cluster membership via k-means clustering. *Right:* Same as left, but for the "cluster" dataset.

C.10 Applying $\hat{p}_{selective}(\hat{\sigma}_{MED})$ to other data-generating models

In this section, we investigate the empirical performance of our method on non-normal data. We repeat the experiments in Section 4.5 with data generated from Poisson and negative binomial models after applying variance-stabilizing transformations [Anscombe, 1948, Bar-Lev and Enis, 1988].

- **Poisson:** To investigate the empirical Type I error control for Poisson data, we first

generate data from

$$X_{ij} \overset{ind}{\sim} \text{Poisson}(5), \quad i = 1, \ldots, 300\,; j = 1, \ldots, 50, \tag{C.30}$$

and apply the Anscombe transformation [Anscombe, 1948] $x \to 2\sqrt{x + 3/8}$ to the data. Next, we perform k-means clustering and computed $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ on the transformed data. Results aggregated from 1,000 simulated data sets are displayed in Figure C.4(a). We see that the naive p-values lead to an anti-conservative test, whereas the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ appears to control the selective Type I error.

Next, we show that the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ has power to reject H_0 when it is not true. We generate data from

$$X_{ij} \overset{ind}{\sim} \text{Poisson}\Big(\big(\mu_{ij}^2 - 3/8\big)^2\Big), \quad i = 1, \ldots, 300\,; j = 1, \ldots, 50, \tag{C.31}$$

where

$$\mu_1 = \ldots = \mu_{100} = \begin{bmatrix} 5 - \frac{\delta}{2} \\ \mathbf{5}_{49} \end{bmatrix}, \; \mu_{101} = \ldots = \mu_{200} = \begin{bmatrix} \mathbf{5}_{49} \\ 5 + \frac{\sqrt{3}\delta}{2} \end{bmatrix}, \; \mu_{201} = \ldots = \mu_{300} = \begin{bmatrix} 5 + \frac{\delta}{2} \\ \mathbf{5}_{49} \end{bmatrix}, \tag{C.32}$$

and $\mathbf{5}_q$ is the q-dimensional vector of all fives. Here, we can think of $\mathcal{C}_1 = \{1, \ldots, 100\}, \mathcal{C}_2 = \{101, \ldots, 200\}, \mathcal{C}_3 = \{201, \ldots, 300\}$ as the "true clusters". Moreover, these clusters are equidistant in the sense that the pairwise distance between each pair of population means is δ. We generate $M = 2,000$ datasets from (C.31) with $\delta = 4, 5, \ldots, 9$. For each simulated dataset, we apply k-means clustering with $K = 3$ on the Anscombe transformed data and reject $H_0 : \mu^\top \nu = 0_{50}$ if $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ is less than $\alpha = 0.05$. The conditional power (defined in (4.23)) increases as a function of δ in (C.32) (see Figure C.4(c)).

- **Negative Binomial:** To investigate the empirical Type I error control for negative binomial data, we generate data from

$$X_{ij} \overset{ind}{\sim} \text{Negative Binomial}(5; 1), \quad i = 1, \ldots, 300\,; j = 1, \ldots, 50, \tag{C.33}$$

and apply the log transformation $x \to \log(x+1)$ to the data. Next, we perform k-means clustering and computed $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ on the log-transformed data. We display results aggregated from 1,000 simulated data sets in Figure C.4(b); the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ appears to control the selective Type I error rate.

Next, we investigate the power of the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for negative binomial data after log transformation. We generate data from

$$X_{ij} \overset{ind}{\sim} \text{Negative Binomial}(\exp(\mu_{ij} - 1); 1), \quad i = 1, \ldots, 300\,; j = 1, \ldots, 50, \tag{C.34}$$

where

$$\mu_1 = \ldots = \mu_{100} = \begin{bmatrix} 10 - \frac{\delta}{2} \\ \mathbf{10}_{49} \end{bmatrix}, \; \mu_{101} = \ldots = \mu_{200} = \begin{bmatrix} \mathbf{10}_{49} \\ 10 + \frac{\sqrt{3}\delta}{2} \end{bmatrix}, \; \mu_{201} = \ldots = \mu_{300} = \begin{bmatrix} 10 + \frac{\delta}{2} \\ \mathbf{10}_{49} \end{bmatrix}, \tag{C.35}$$

and $\mathbf{10}_q$ is the q-dimensional vector of all tens. As in the Poisson case, we can think of $\mathcal{C}_1 = \{1, \ldots, 100\}, \mathcal{C}_2 = \{101, \ldots, 200\}, \mathcal{C}_3 = \{201, \ldots, 300\}$ as the "true clusters", and the pairwise distance between each pair of population means is δ. We generate $M = 2,000$ datasets from (C.34) with $\delta = 7, 8, \ldots, 15$. For each simulated dataset, we apply k-means clustering with $K = 3$ on the log transformed data and reject $H_0 : \mu^\top \nu = 0_{50}$ if $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ is less than $\alpha = 0.05$. The conditional power (defined in (4.23)) increases as a function of δ in (C.35) (see Figure C.4(d)).

To sum up, the test based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ leads to adequate Type I error rate control and substantial power on Poisson and negative binomial data with variance-stabilizing trans-

formations. However, prior work has suggested that if the mean counts are small, then no data transformation can be expected to stabilize variances [Warton, 2018]. In future work, we will investigate other strategies to extend our framework to non-normal data (see, e.g., Section 4.4 of the dissertation).

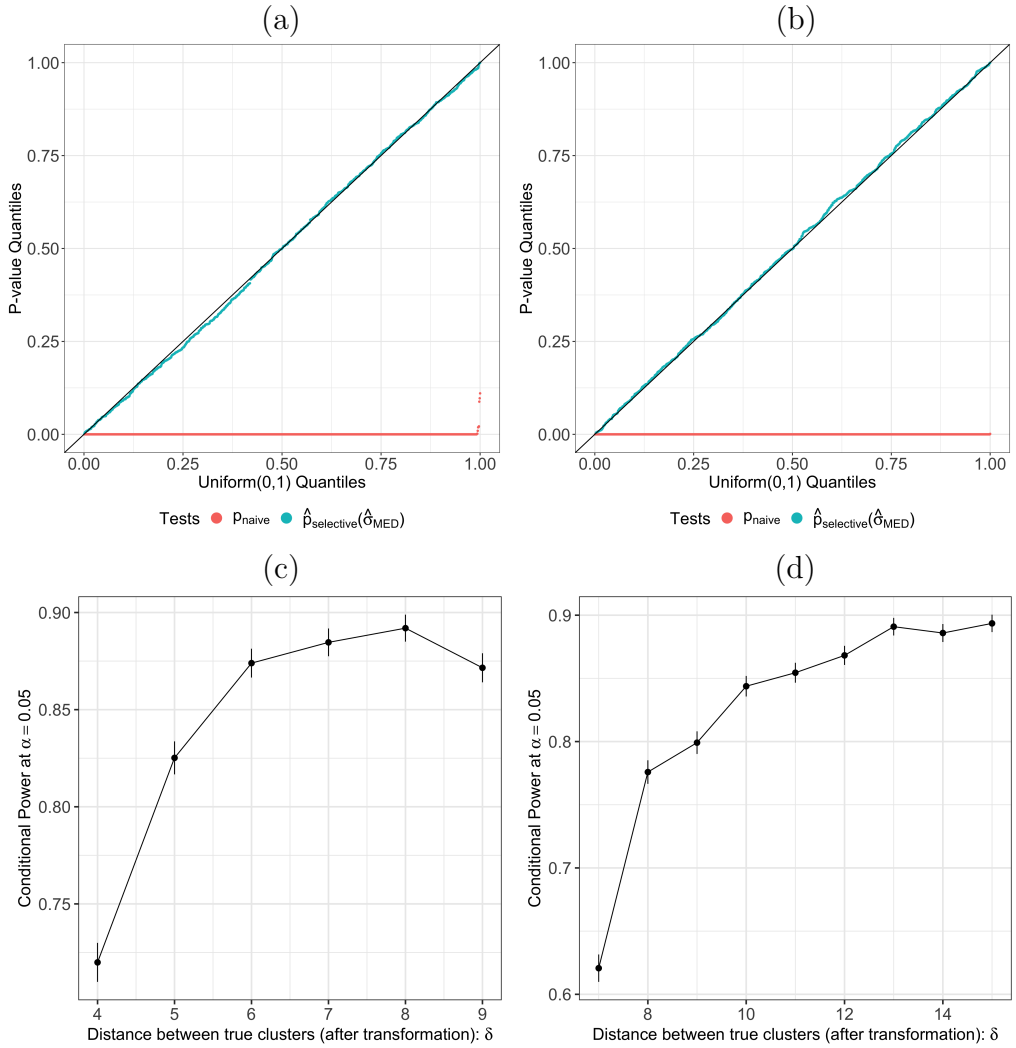


Figure C.4: *(a):* Quantile-quantile plots for p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for Anscombe-transformed Poisson data. *(b):* Quantile-quantile plots for p_{Naive} and $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for log-transformed negative binomial data. *(c):* Conditional power (defined in (4.23)) at $\alpha = 0.05$ for the tests based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for Anscombe-transformed Poisson data. *(d):* Conditional power (defined in (4.23)) at $\alpha = 0.05$ for the tests based on $\hat{p}_{\text{selective}}(\hat{\sigma}_{\text{MED}})$ for log-transformed negative binomial data.