**Selective inference for** k**-means clustering**

**Yiqun T. Chen** yiqunc@stanford.edu

Data Science Institute and Department of Biomedical Data Science Stanford University

Stanford, CA 94305, USA

**Daniela M. Witten** dwitten@uw.edu

Departments of Statistics and Biostatistics University of Washington

Seattle, WA 98195-4322, USA

**Editor:** Ji Zhu

**Abstract**

We consider the problem of testing for adiference in means between clusters of observations identiﬁed via k-means clustering. In this setting, classical hypothesis tests lead to an in且ated Type I error rate. In recent work, [Gao et al.](#bookmark2) [(2022) considered a related problem](#bookmark2) in the context of hierarchical clustering. Unfortunately, their solution is highly-tailored to the context of hierarchical clustering, and thus cannot be applied in the setting of k-means clustering. In this paper, we propose a p-value that conditions on all of the intermediate clustering assignments in the k-means algorithm. We show that the p-value controls the selective Type I error for a test of the diference in means between a pair of clusters obtained using k-means clustering in ﬁnite samples, and can be e伍ciently computed. We apply our proposal on hand-written digits data and on single-cell RNA-sequencing data.

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

**1. Introduction**

Testing for a diference 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 diference in means be- tween 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 clus- ter the cells, and then test for adiference in the expected gene expression levels between the clusters to quantify up- or down-regulation of genes, annotate known cell types, and iden- [tify new cell types (Gru…n et al.,](#bookmark3) [2015;](#bookmark3) [Aizarani et al.,](#bookmark4) [2019;](#bookmark4) [La…hnemann et al.,](#bookmark5) [2020;](#bookmark5) [Zhang](#bookmark6) [et al.,2019;](#bookmark6)[Doughty and Kerkhoven,2020)](#bookmark7). In fact, the inferential challenges resulting from testing data-guided hypotheses have been described as a “grand challenge” in the ﬁeld of [genomics (La…hnemann et al.,](#bookmark5) [2020), and papers in the ﬁeld continue to overlook this issue:](#bookmark5) as an example, seurat [(Stuart et al.,](#bookmark8) [2019), the state-of-the-art single-cell RNA sequencing](#bookmark8) analysis tool, tests for diferential gene expression between groups obtained via clustering, with a note that “p-values [from these hypotheses] should be interpreted cautiously, as the

oc2023 Yiqun T. Chen and Daniela M. Witten.

License: CC-BY 4.0, see [https://creativecommons.org/licenses/by/4.0/.](https://creativecommons.org/licenses/by/4.0/) Attribution requirements are provided at [http://jmlr.org/papers/v24/22-0371.html.](http://jmlr.org/papers/v24/22-0371.html)

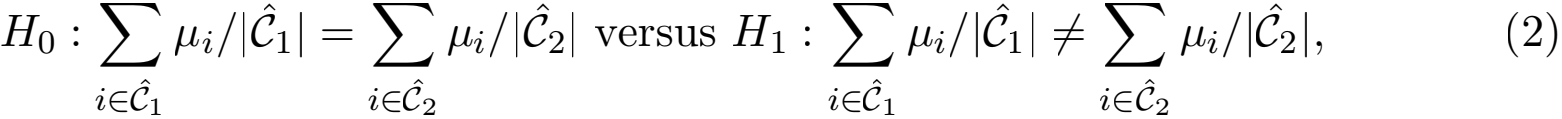
Chen and Witten

genes used for clustering are the same genes tested for diferential expression.” Testing data- [guided hypothesis also arises in the ﬁeld of neuroscience (Kriegeskorte et al.,](#bookmark9) [2009;](#bookmark9) [Button,](#bookmark10) [2019), social psychology (Hung and Fithian,](#bookmark10) [2020), and physical sciences (Friederich et al.,](#bookmark11) [2020;](#bookmark12) [Pollice et al.,](#bookmark13) [2021)](#bookmark13). 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.

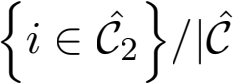
In this paper, we develop a test for a diference in means between two clusters estimated from applying [k-means clustering (Lloyd,](#bookmark14) [1982;](#bookmark14) [MacQueen et al.,](#bookmark15) [1967), an extremely pop](#bookmark15)- [ular clustering algorithm with numerous applications (Xu and Wunsch,](#bookmark16) [2008)](#bookmark16). In recent work, [Gao et al.](#bookmark2) [(2022) tackled a similar problem for hierarchical clustering](#bookmark2). While the two papers share similar notation and setup, our solutions and algorithms are tailored to the iterative and centroid-based nature of k-means clustering, leading to fundamentally difer- ent solutions and algorithms than those proposed in [Gao et al.](#bookmark2) [(2022)](#bookmark2). We consider the [following simple and well-studied model (Gao et al.,](#bookmark2) [2022;](#bookmark2) [Lo… er et al.,](#bookmark17) [2021;](#bookmark17) [Lu and Zhou,](#bookmark18) [2016) for](#bookmark18) n observations and q features:

X ~ MN n ×q (μ,**I**n, σ2 **I**q) , (1)

where MN [denotes the matrix normal distribution (Bilodeau and Brenner,1999),](#bookmark20) μ ∈ Rn ×q has unknown rows μi, and σ 2 > 0 is known. Given a realization x ∈ Rn ×q of X, we ﬁrst apply the k-means clustering algorithm to obtain C(x), a partition of the samples {1, . . . , n}. We might then consider testing the null hypothesis that the mean is the same across two estimated clusters, i.e.,



where 1 , 2 ∈ C(x) are estimated clusters with cardinality | 1 | and | 2 | . This is equivalent to testing H0 : μTV = 0q versus H1 : μTV  0q, where

Vi = 1 1 | - 1 2 | , i = 1, . . . , n, 

and 1{A} equals 1 if the event A holds, and 0 otherwise. [Gao et al.](#bookmark2) [(2022) demonstrates](#bookmark2) that the p-value given by

pNaive = prH0 ( ⅡXTVⅡ2 ≥ ⅡxTVⅡ2 ), (4)

where ⅡXTVⅡ2 ~ (σⅡVⅡ2 )χq under H0, leads to an extremely anti-conservative test. In

[particular, we constructed the contrast vector in (3) because](#bookmark22) 1 and 2 were obtained by

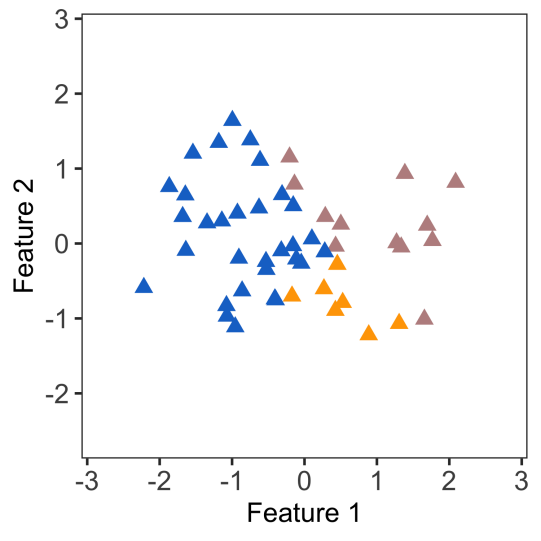
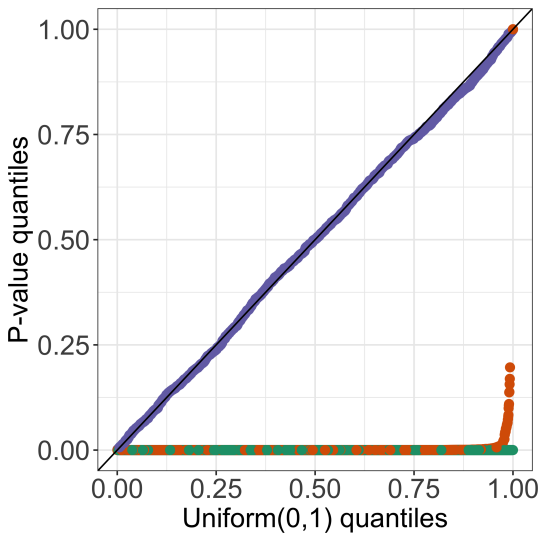
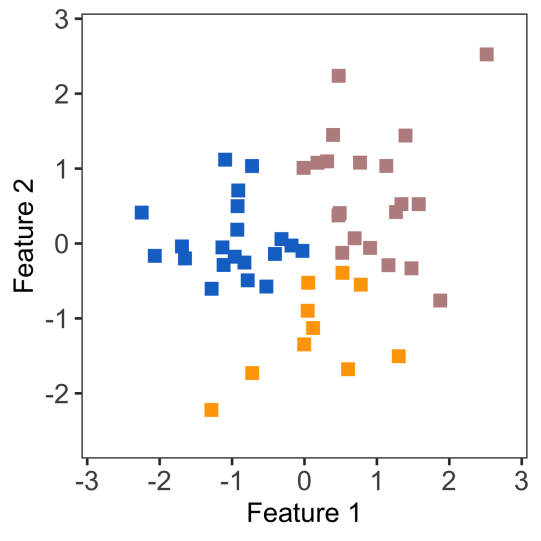
clustering. Therefore, we will observe substantial diferences between the cluster centroids

Σi∈1 xi/|1 | andΣi∈2 xi/|2 |, even in the absence of true diferences in their population

means (left panel Figure [1)](#bookmark24).

Notably, the problem of testing for a diference in means between two groups obtained via clustering cannot be easily overcome by sample splitting, as pointed out in [Gao et al.](#bookmark2) [(2022) and](#bookmark2)[Zhang et al.](#bookmark6) [(2019)](#bookmark6). To see why, we divide the observations into a training and a test set. We apply k-means clustering on only the training set (left panel of Figure [1), and](#bookmark24) then assign the test set observations to those clusters (to obtain the center panel of Figure [1,](#bookmark24)

Selective inference for k-means clustering



[Figure 1: We simulated one dataset from (1) with](#bookmark19) μ = 0100×2 and σ = 1. We split the data into training (left ) and test sets (middle). Left: We apply k-means clustering on the training set to obtain three clusters. Center: We apply the training set clusters to the test set using a 3-nearest neighbors classiﬁer. Right: Quantile- quantile plot of the naive [p-values (4) applied to the training set (green) and](#bookmark23) the test set (orange), aggregated over 2,000 simulated datasets; as well as our proposed [p-values (in (9); displayed in purple) applied to the training set](#bookmark25).

we applied a 3-nearest neighbor classiﬁer). [Finally, we compute the naive p-values (4)](#bookmark23) 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 v [in (3)](#bookmark22). Therefore, the Wald test based on sample-splitting remains extremely anti-conservative, as shown in the right panel of Figure [1, and does not](#bookmark24) lead to a valid test of H0 [in (2)](#bookmark21). We refer the readers to [Gao et al.](#bookmark2) [(2022) for further](#bookmark2) discussion of this point.

In this paper, we develop a test of H0 that controls the selective Type I error. That is, we wish to ensure that the probability of rejecting H0 at level Q, given that H0 holds and we decided to test it, is no greater than Q:

prH0 (reject H0 at level Q j H0 is tested) ≤ Q, 8Q ∈ (0, 1). (5)

To develop the test, we leverage the selective inference framework, which has been applied [extensively in high-dimensional linear modeling (Lee et al.,](#bookmark27) [2016;](#bookmark27) [Tibshirani et al.,](#bookmark28) [2016;](#bookmark28) [Fithian et al.,](#bookmark29) [2014;](#bookmark29) [Ru…gamer et al.,](#bookmark30) [2022;](#bookmark30) [Schultheiss et al.,](#bookmark31) [2021;](#bookmark31) [Taylor and Tibshirani,](#bookmark32) [2018;](#bookmark32)[Charkhi and Claeskens,2018;](#bookmark33)[Yang et al.,2016;](#bookmark34)[Loftus and Taylor,2015), changepoint](#bookmark35) [detection (Jewell et al.,](#bookmark36) [2022;](#bookmark36) [Hyun et al.,](#bookmark37) [2021,](#bookmark37) [2018;](#bookmark38) [Chen et al.,](#bookmark39) [2021;](#bookmark39) [Le Duy and](#bookmark40) [Takeuchi,](#bookmark40) [2021;](#bookmark40) [Duy et al.,](#bookmark41) [2020;](#bookmark41) [Benjamini et al.,](#bookmark42) [2019), and clustering (Zhang et al.,](#bookmark42) [2019;](#bookmark6) [Gao et al.,](#bookmark2) [2022;](#bookmark2) [Watanabe and Suzuki,](#bookmark43) [2021)](#bookmark43). The key insight behind selective inference is as follows: naive [p-values such as (4) lead to anti-conservative tests because](#bookmark23) the hypothesis H0 is generated by the same data used for testing. Therefore, to obtain a valid test of H0, we need to condition on the aspect of the data that led us to test H0 . In

[our case, we chose to test the null hypothesis in (2) because](#bookmark21) 1 and 2 were obtained via

k-means clustering. Therefore, we compute ap-value conditional on the event that k-means

clustering yields 1 and 2 . [This results in selective Type I error control (5), as seen in the](#bookmark26)

right panel of Figure [1.](#bookmark24)

Chen and Witten

There is a rich literature on estimating and quantifying the uncertainty in the number of [clusters (Li and Chen,](#bookmark44) [2010;](#bookmark44) [Chen and Li,](#bookmark45) [2009;](#bookmark45) [Chen et al.,](#bookmark46) [2004;](#bookmark46) [McLachlan et al.,](#bookmark47) [2019;](#bookmark47) [Dobriban,](#bookmark48) [2020), as well as assessing cluster stability and heterogeneity (Suzuki and Shi-](#bookmark48) [modaira,](#bookmark49) [2006;](#bookmark49) [Kerr and Churchill,](#bookmark50) [2001;](#bookmark50) [Kimes et al.,](#bookmark51) [2017;](#bookmark51) [Chung,](#bookmark52) [2020;](#bookmark52) [Jin and Wang,](#bookmark53) [2016;](#bookmark53) [Aw et al.,](#bookmark54) [2021;](#bookmark54) [Chung and Storey,](#bookmark55) [2015)](#bookmark55). Others have examined the asymptotic [properties of clustering models from a Bayesian perspective (Guha et al.,](#bookmark56) [2019;](#bookmark56) [Nobile,](#bookmark57) [2004;](#bookmark57) [Cai et al.,](#bookmark58) [2020)](#bookmark58). In addition, k-means clustering is a special case of the expectation- maximization algorithm, which allows us to tap into an active line of research on the statis- [tical guarantees of the expectation-maximization algorithm (Zhang and Zhang,](#bookmark59) [2014;](#bookmark59) [Wang](#bookmark60) [et al.,](#bookmark60) [2015;](#bookmark60) [Cai et al.,](#bookmark61) [2019;](#bookmark61) [Yi and Caramanis,](#bookmark62) [2015;](#bookmark62) [Balakrishnan et al.,](#bookmark63) [2017)](#bookmark63). However, most prior work focused the setting with one or more “true” clusters. By contrast, we are

[interested in a correctly-sized test for the null hypothesis (2), even when](#bookmark21) 1 , 2 do not corre-

spond to “true” clusters, and even in the absence of “true” clusters in the data. In addition, existing work often relies on asymptotic approximations and bootstrap resampling. Two recent exceptions include [Zhang et al.](#bookmark6) [(2019) and](#bookmark6) [Gao et al.](#bookmark2) [(2022), who took a selective](#bookmark2) inference approach and computed ﬁnite-sample p-values for testing the diference in means between estimated clusters obtained via linear classiﬁcation rules and hierarchical cluster- ing, respectively. Our work is closest to [Gao et al.](#bookmark2) [(2022), and extends their framework](#bookmark2) to k-means clustering. We provide an exact, ﬁnite-sample test of the diference in means between a pair of clusters estimated via [k-means clustering under model (1), without the](#bookmark19) need for sample splitting.

The rest of this paper is organized as follows. In Section [2, we brie且y review the work](#bookmark64) of [Gao et al.](#bookmark2) [(2022), and outline our proposed test of a diference in means after](#bookmark2) k-means clustering. It is worth highlighting that while our proposal is inspired by the work of [Gao](#bookmark2) [et al.](#bookmark2) [(2022), our solution is](#bookmark2) not simply a minor modiﬁcation: computing the conditioning [set for the p-value in (9) is the key technical challenge of this paper, and the computational](#bookmark25) insights in [Gao et al.](#bookmark2) [(2022) are only applicable to hierarchical clustering](#bookmark2). In Section [3, we](#bookmark65) provide a computationally-e伍cient approach to compute the p-values corresponding to our proposed test, for a diference in means after k-means clustering. Section [4](#bookmark66) outlines some extensions, and we evaluate our proposal in a simulation study in Section [5.](#bookmark67) We apply our proposal to three real datasets in Section [6, and discuss future work in Section](#bookmark68) [7.](#bookmark69) Proofs and additional results are relegated to the Appendix.

Throughout this paper, we will use the following notational conventions. For a matrix A, Ai denotes the ith row and Aij denotes the (i, j)th entry. For a vector v ∈ Rn , ⅡvⅡ2

denotes its l2 norm, and Π is the projection matrix onto the orthogonal complement of

v, i.e., Π = **I**n — vvT /ⅡvⅡ, where **I**n is the n-dimensional identity matrix. Moreover,

dir(v) = v/ⅡvⅡ2 if v  0n and 0n otherwise, where 0n is the n-vector of zeros. We let 〈· , ·〉 and 1{·} denote the inner product of two vectors and the indicator function, respectively.

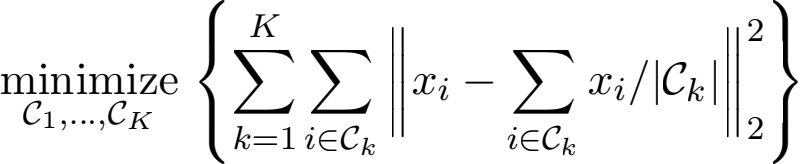
**2. Selective inference for k-means clustering**

**2.1 A brief review of** k**-means clustering**

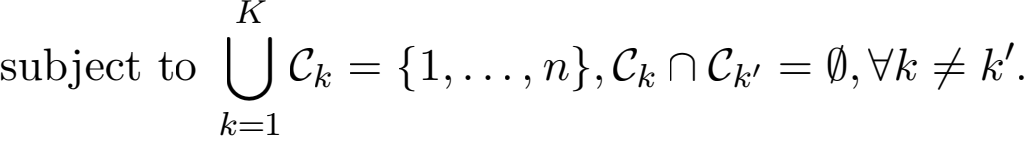
In this section,wereview the k-means clustering algorithm. Given samples x1 , . . . , xn ∈ Rq and a positive integer K , k-means clustering partitions the n samples into disjoint subsets

Selective inference for k-means clustering

1 , . . . , K by solving the following optimization problem:



(6)



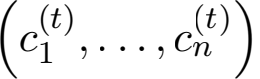
[It is not typically possible to solve for the global optimum in (6) (Aloise et al.,](#bookmark70) [2009)](#bookmark71). A [number of algorithms are available to ﬁnd a local optimum (Hartigan and Wong,](#bookmark72) [1979;](#bookmark72) [Zha](#bookmark73) [et al.,](#bookmark73) [2002;](#bookmark73) [Arthur and Vassilvitskii,](#bookmark74) [2007); one such approach is Lloyd's algorithm (Lloyd,](#bookmark74) [1982), given in Algorithm](#bookmark14) [1.](#bookmark75) We ﬁrst sample K out of n observations as initial centroids (step 1 in Algorithm [1)](#bookmark75). 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.,](#bookmark76) [2001)](#bookmark76).

In what follows, we will sometimes use ct)(x) and mt)(x) rather than ci(t) and mk(t) to emphasize the dependence of the cluster labels and centroids on the data x.

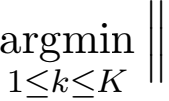
**Algorithm 1:** Lloyd's algorithm for [k-means clustering (Lloyd,](#bookmark14) [1982)](#bookmark14)

**Input:** Data x1 , . . . , xn ∈ Rq, number of output clusters K, maximum iteration T, random seed s.

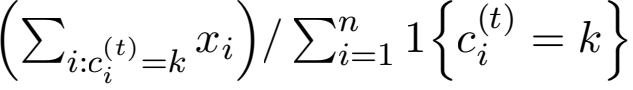
**Output:** Cluster assignments  .

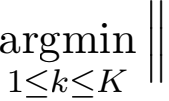
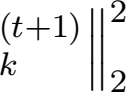
1. Initialize the centroids (m  , . . . , m) by sampling K observations from

x1 , . . . , xn without replacement, using the random seed s.

2. Compute assignments c ← xi — m , i = 1, . . . , n.

3. Initialize t = 0. **while** t ≤ T **do**

a. Update centroids: m ← , k = 1, . . . , K.

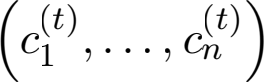
b. Update assignment: c ← xi — m , i = 1, . . . , n.

c. **if** ci(t+1) = ci(t) for all 1 ≤ i ≤ n **break**

**else**

t ← t + 1.

**end**

**return** .

**2.2 A test of** [(2)](#bookmark21) **for clusters obtained via** k**-means clustering**

Here, we brie且y review the proposal of [Gao et al.](#bookmark2) [(2022) for selective inference for hierar](#bookmark2)- [chical clustering, and outline a selective test for (2) for](#bookmark21) k-means clustering.

Chen and Witten

[Gao et al.](#bookmark2) [(2022) proposed a selective inference framework for testing hypotheses based](#bookmark2) on the output of a clustering algorithm. Let C(·) denote the clustering operator, i.e., a partition of the observations resulting from a clustering algorithm. Since H0 [in (2) is chosen](#bookmark21)

because {1 , 2 ∈ C(x)}, where 1 , 2 are the two estimated clusters under consideration in

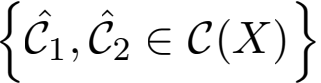
[(2),](#bookmark21) [Gao et al.](#bookmark2) [(2022) proposed to reject](#bookmark2) H0 if

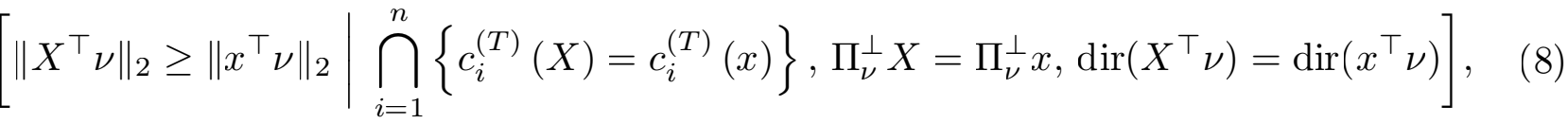
prH0 { ⅡXTνⅡ2 ≥ ⅡxT νⅡ2 **I** 1 , 2 ∈ C(X), ΠX = Πx, dir(XTν) = dir(xT ν)} (7)

is small. [In (7), conditioning on](#bookmark77) {ΠX = Πx, dir(XTν) = dir(xT ν)} eliminates the nui-

sance parameters Πμ and dir(μT ν), where Π = **I**n—ννT /ⅡνⅡ2 and dir(μT ν) = μT ν/ⅡμT νⅡ2

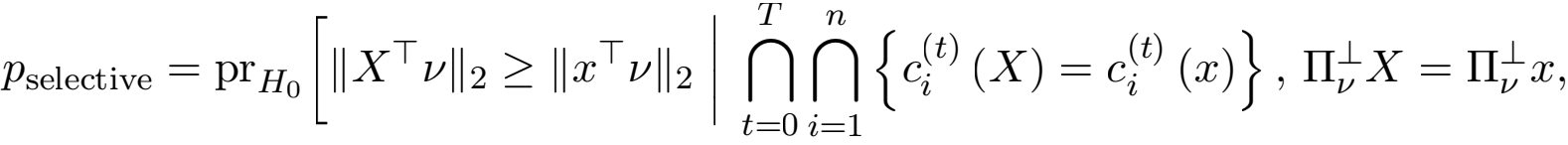
(see, e.g., Section 3.1 of [Fithian et al.](#bookmark29) [(2014))](#bookmark29). [Gao et al.](#bookmark2) [(2022) showed that the test that](#bookmark2) rejects H0 [when (7) is below](#bookmark77) Q controls the selective Type I error at level Q, in the sense [of (5)](#bookmark26). [Furthermore, under (1), the conditional distribution of](#bookmark19) ⅡXTνⅡ2 [in (7) is (σⅡνⅡ](#bookmark77)2 )χq , truncated to a set. When the operator C(·) denotes hierarchical clustering, this set can be analytically characterized and e伍ciently computed, leading to an e伍cient algorithm for [computing (7)](#bookmark77).

We now extend these ideas to [k-means clustering (6)](#bookmark70). 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  . This leads to the p-value

prH0 

where cT)(X) is the cluster assigned to the ith observation at the ﬁnal iteration of Algo-

rithm [1.](#bookmark75) [However, computing (8) requires characterizing](#bookmark78) ∩cT)(X) = cT)(x)}, which is not straightforward, and may necessitate enumerating over possibly an exponential num- ber of intermediate cluster assignments ct)(·) fort = 1, . . . , T — 1. Hence, we also condition on all of the intermediate clustering assignments in Algorithm [1:](#bookmark75)



(9)

dir(XTν) = dir(xT ν)].

[In (9),](#bookmark25) ct)(X) is the cluster assigned to the ith observation at the tth iteration of Algorithm [1.](#bookmark75) Roughly speaking, this p-value answers the question:

Assuming that there is no diference between the population means of 1 and

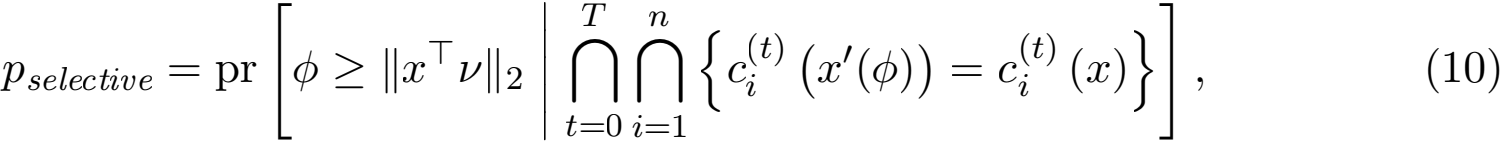
2 , what is the probability of observing such a large diference between their

centroids, among all the realizations of X that yield identical results in every iteration of the k-means algorithm?

The [p-value in (9) is the focus of this paper](#bookmark25). We establish its key properties below.

Selective inference for k-means clustering

**Proposition 1** Suppose that 儿 is a realization from [(1), and let](#bookmark19) φ ~ (σⅡνⅡ2 )χq . Then, under H0 : µT ν = 0 with ν deﬁned in [(3)](#bookmark22),

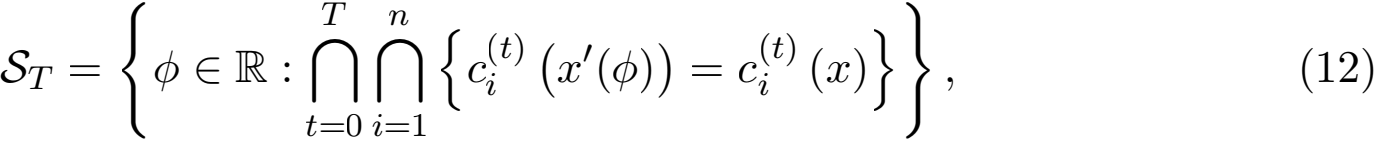


where p selective is deﬁned in [(9), and](#bookmark25)

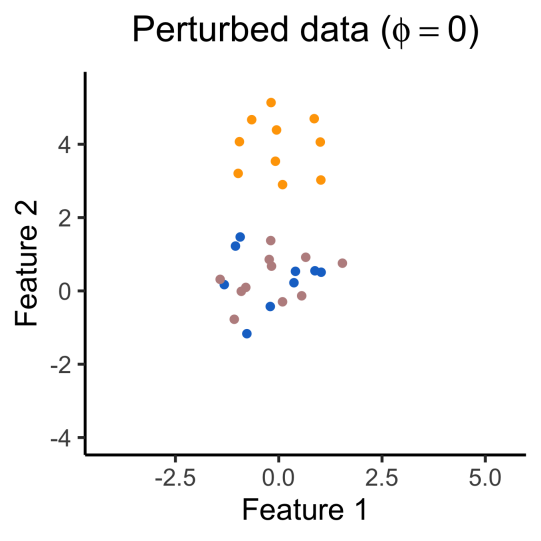
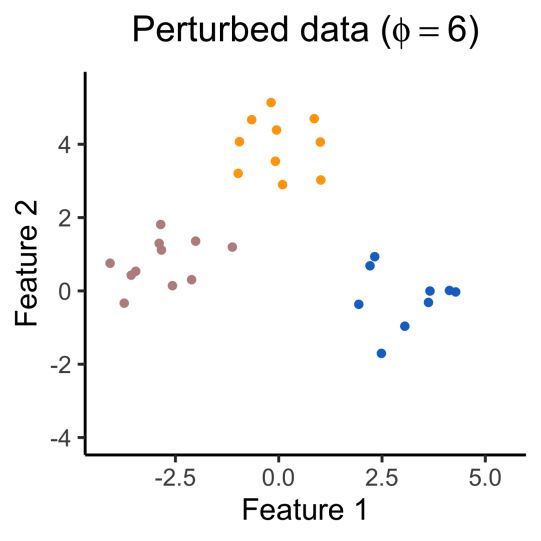
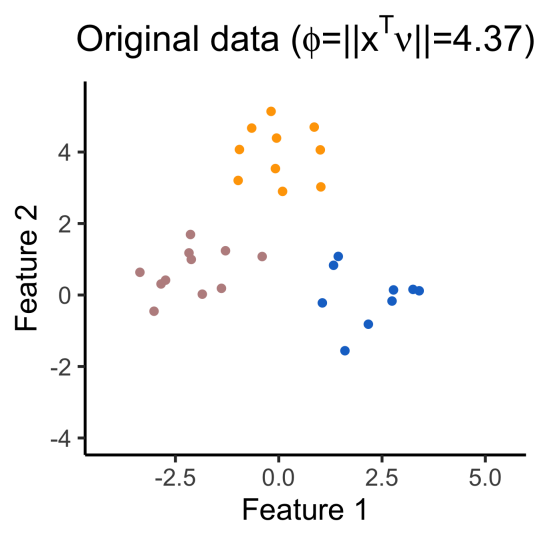
儿9 (φ) = 儿 + (φ — Ⅱ儿T νⅡ2 ) (ν/ⅡνⅡ){dir(儿T ν)}T . (11)

Moreover, the test that rejects H0 : µT ν = 0 when p selective ≤ α controls the selective Type I error at level α, in the sense of [(5)](#bookmark26).

Proposition [1](#bookmark79) states that p selective can be recast as the survival function of a scaled χq random variable, truncated to the set



where 儿9 [(φ) is deﬁned in (11)](#bookmark81). Therefore, to compute p selective, it su伍ces to characterize the set ST . [In (11),](#bookmark81) 儿9 (φ) results from applying a perturbation to the observed data 儿, along



[Figure 2: One simulated dataset generated from model (1) with](#bookmark19) µi = 1{1 ≤ i ≤ 10}[2.5, 0]T + 1{11 ≤ i ≤ 20}[0, —2.5]T + 1{21 ≤ i ≤ 30}[√18.75, 0]T and σ = 1. Left: The original data 儿 corresponds to φ = Ⅱ儿T νⅡ2 = 4.37. Applying k-means clustering with K = 3 yields three clusters, displayed in rosy brown, blue, and or-

ange. Here, ν is chosen to test for adiference in means between 1 (rosy brown)

and 2 (blue). Center: The perturbed data 儿9 (φ) with φ = 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 儿9 (φ) with φ = 6. Applying k-means clustering with K = 3 yields the same set of clusters as in the left panel.

the direction of 儿T ν, the diference between the two cluster centroids of interest. Figure [2](#bookmark83) [illustrates a realization of (1) for](#bookmark19) k-means clustering with K = 3. The left panel displays the observed data 儿, which corresponds to 儿9 (φ) with φ = Ⅱ儿T νⅡ2 = 4.37. Here, [ν deﬁned in (3)](#bookmark22)

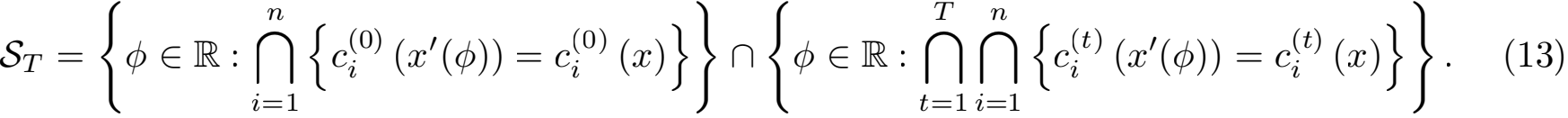
Chen and Witten

was chosen to test the diference between 1 (shown in rosy brown) and 2 (shown in blue).

The center and right panels of Figure [2](#bookmark83) display xI (φ) with φ = 0 and φ = 6, respectively. In the center panel, with φ = 0, the blue and rosy brown clusters are “pushed together”, resulting in ⅡxI (φ)TVⅡ2 = 0; that is, there is no diference 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 φ = 6, the blue and rosy brown clusters are “pulled apart” along the direction of xTV, which results in an increased distance between the centroids of the blue and rosy brown clusters, and k-means clustering does yield the same clusters as on the original data. In this example, ST = (3.59, ∞ ).

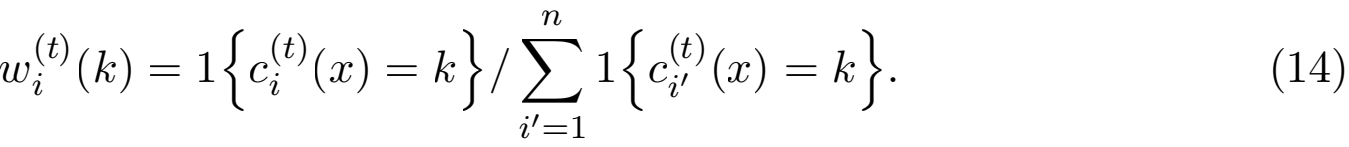
**3. Computation of the selective p-value**

In Section [2, we have shown that the](#bookmark64) p-value p selective [(9) involves the set](#bookmark25) ST [in (12)](#bookmark82). Indeed, a computationally-e伍cient characterization of ST is the key technical challenge and contribution of our paper. Here, we start with a high-level summary of our approach to characterizing ST [in (12)](#bookmark82). First, we rewrite

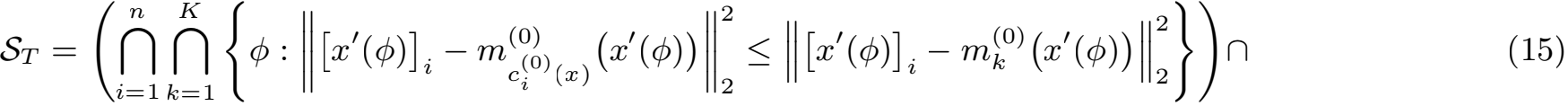


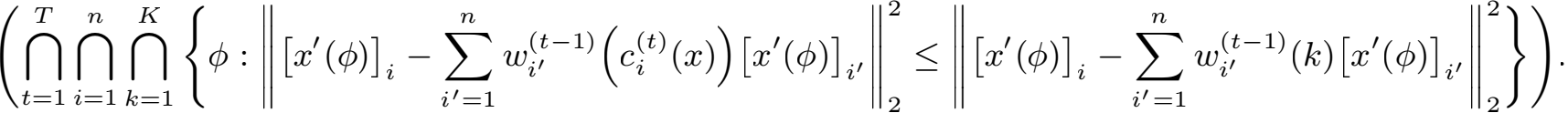
[Next, we consider the ﬁrst term in (13): according to step 2](#bookmark84). of Algorithm [1, for](#bookmark75) i = 1, . . . , n, c0)(xI (φ)) = c0)(x) if and only if for i = 1, . . . , n, the initial randomly-sampled centroid to which [xI (φ)]i is closest coincides with the initial centroid to which xi is closest. This condition can be expressed using K — 1 inequalities. Furthermore, the same intuition [holds for the second term in (13), except that the centroids are a function of the cluster](#bookmark84) assignments in the previous iteration. We formalize this intuition in Proposition [2, proven](#bookmark85) in Appendix [A.2.](#bookmark86)

**Proposition 2** Suppose that we apply the k-means clustering algorithm (Algorithm [1) to](#bookmark75) a matrix x ∈ Rn ×q , to obtain K clusters in at most T steps. Deﬁne



Then,for the set ST deﬁned in [(12), we have that](#bookmark82)





(16)

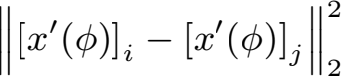
Recall that ct)(x) denotes the cluster to which the ith observation is assigned instep 3b.

of Algorithm [1](#bookmark75) during the tth iteration, and that m(x) denotes the kth centroid sampled

from the data x during step 1 of Algorithm [1.](#bookmark75) In words, Proposition [2](#bookmark85) says that ST can be expressed as the intersection of O(nKT) sets. Therefore, it su伍ces to characterize the sets [in (15) and (16)](#bookmark88). We now present two lemmas.

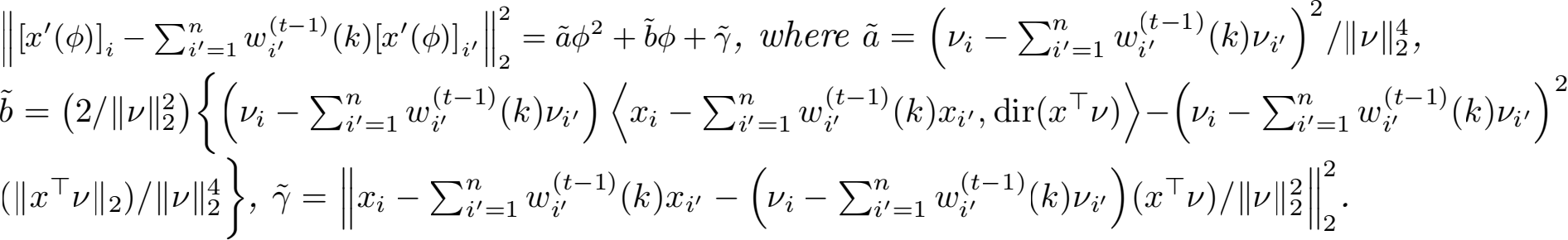
Selective inference for k-means clustering

**Lemma 3 (Lemma 2 in** [**Gao et al.**](#bookmark2)[**(2022))**](#bookmark2)For ν in [(3)](#bookmark22) and x/ (φ) in [(11), we have](#bookmark81)

that  = aφ2 +bφ+√ , where a = 2 , b = 2 [ /ⅡνⅡ

{(νi - νj )/ⅡνⅡ}2 ⅡxT νⅡ2], and √ = **Ⅱ**xi - xj - (νi - νj )(xT ν)/jjνjj **Ⅱ** .

**Lemma 4** For ν in [(3)](#bookmark22), x/ (φ) [in (11), and w i(t)(k)](#bookmark81) in [(14), we have that](#bookmark87)



It follows from Lemmas [3](#bookmark90) and [4](#bookmark91) [that all of the inequalities in (15) and (16) are in fact](#bookmark88) quadratic in φ, with coe伍cients that can be analytically computed. Therefore, computing the set ST requires solving O(nKT) quadratic inequalities of φ .

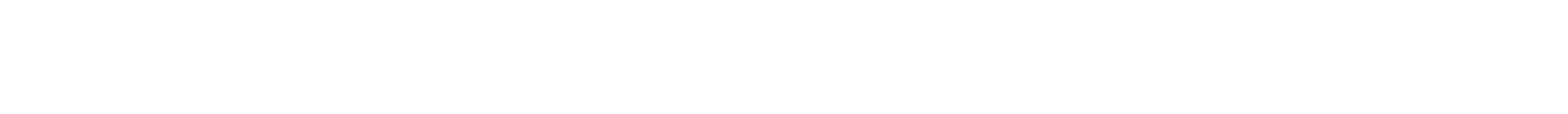
**Proposition 5** Suppose that we apply the k-means clustering algorithm (Algorithm [1) to](#bookmark75) a matrix x ∈ Rn ×q , to obtain K clusters in at most T steps. Then, the set ST deﬁned in [(12)](#bookmark82) can be computed in O(KT (n + q) + nKT log (nKT)) operations.

**4. Extensions**

**4.1 Non-spherical covariance matrix**

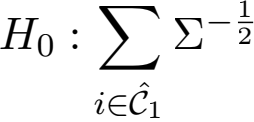
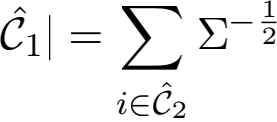
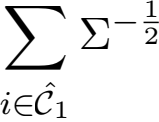
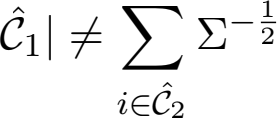
Thus far, we have assumed that the observed data x [is a realization of (1), which implies](#bookmark19) that cov(Xi) = σ 2 **I**q . However, this assumption is often violated in practice. For example, expression levels of genes are highly correlated, and neighbouring pixels in an image tend to be more similar. For a known positive deﬁnite matrix Σ, we now let

X ~ MN n ×q(µ,**I**n, Σ) . (17)

[Under (17), we can whiten the data by applying the transformation](#bookmark93) xi → Σ— 2 xi (Bell

1

and Sejnowski, 1997), where Σ —  is the unique symmetric positive deﬁnite square root of Σ—1 [(Horn and Johnson,](#bookmark94) [2012)](#bookmark94). Note that Σ —  ~ N(Σ—  **I**q). Moreover, as Σ —  > 0, [testing the null hypothesis in (2) is equivalent to testing](#bookmark21)

 µi/| µi/|2 | versus H1 :  µi/| µi/| (18)

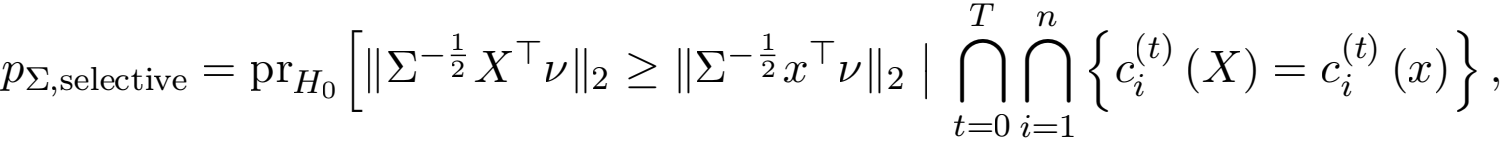
Therefore, to get a correctly-sized test under mo[del (17), we can simply carry out our](#bookmark93)

1

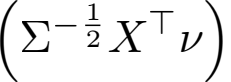
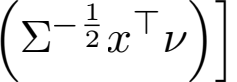
[](#bookmark64)proposal in Section 2 on the transformed data Σ — 2 xi instead of the original data xi.

Instead of applying the whitening transformation, we can directly accommodate a known

covariance matrix Σ by considering the following extension of p selective [in (9):](#bookmark25)

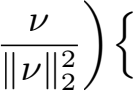


(19)

x, dir  = dir  .

Chen and Witten

**Proposition 6** Suppose that x is a realization from [(17), and let](#bookmark93) φ ∼ (ⅡνⅡ2 )χq . Then, under H0 : µT ν = 0 with ν deﬁned in [(3)](#bookmark22),

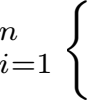
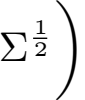
pΣ ,selective = pr [φ ≥ ⅡΣ -  儿T νⅡ2 ''''' t0 i1 {ci(t) (Π儿 + (φ dir (Σ -  儿T

ν)}T Σ  = ci(t) (儿) }] ;

(20)

where pΣ ,selective is deﬁned in [(19)](#bookmark96). Furthermore, the test that rejects H0 : µT ν = 0 when pΣ ,selective ≤ α controls the selective Type I error at level α .

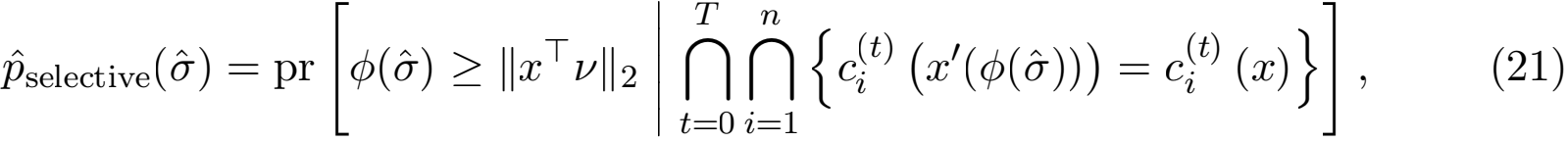
In addition, we can adapt the results in Section [3](#bookmark65) to compute the set

{φ ∈ R : ∩ ∩ci(t) (Πx +(φν/ⅡνⅡ){dir (Σ- T ν)}T  = ci(t) (x)}} by modifying the

results in Lemmas [3](#bookmark90) and [4.](#bookmark91) Details are in Section [A.5](#bookmark98) of the Appendix.

**4.2 Unknown variance**

When σ is unknown, we can plug in an estimate [in (9):](#bookmark25)



where φ() ∼ (ⅡνⅡ2 )χq . If we use a consistent estimator of σ, then a test based on the [p-value in (21) provides selective Type I error control (5) asymptotically.](#bookmark99)

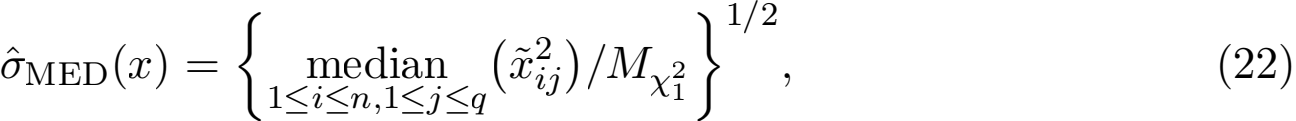
**Proposition 7** For q = 1, 2, . . . , suppose that X(q) ∼ MN n ×q (µ(q) , **I**n, σ2 **I**q) . Let x(q) be a realization from X(q) and let ct)(·) be the cluster to which the ith observation is assigned during the tth iteration of step 3b. in Algorithm [1](#bookmark75). Consider the sequence of null hypotheses H0(q) : µ(q)Tν(q) = 0q, where ν(q) deﬁned in [(3)](#bookmark22) is the contrast vector resulting from applying

k-means clustering on x(q) . Suppose that (i) is a consistent estimator of σ, i.e., for

all ∈ > 0, lim q→∞ pr (|(X(q)) − σ| ≥ ∈ ) = 0; and (ii) there exists δ ∈ (0, 1) such that

lim q→∞ prH0(q) [∩ ∩ct) (X(q)) = ct) (x(q))}] > δ . Then, for all α ∈ (0, 1), we have that lim q→∞ prH0(q) [selective() ≤ α ''' ∩ ∩ct) (X(q)) = ct) (x(q))}] = α .

In practice, we propose to use the following estimator of σ [(Huber,](#bookmark101) [1981):](#bookmark101)



where is obtained from subtracting the median of each column in x, and Mχ is the

median of the χ distribution. If µ is sparse, i.e., Σ Σ=1 1{µij  0} [is small, then (22)](#bookmark102)

is consistent with appropriate assumptions; see Appendix [A.7.](#bookmark103)

**5. Simulation study**

Throughout this section, we consider testing the null hypothesis H0 : µT ν = 0q versus H1 : µT ν  0q , where, unless otherwise stated, ν [deﬁned in (3) is based on a randomly-chosen](#bookmark22)

pair of clusters 1 and 2 from k-means clustering. We consider four p-values: pNaive [in (4),](#bookmark23)

Selective inference for k-means clustering

p selective [in (9),](#bookmark25) selective [in (21) with](#bookmark99) MED [deﬁned in (22), and](#bookmark102) selective [in (21) with](#bookmark99) Sample =

{Σ Σ=1 (xij - x-j)2 /(nq - q)}1/2, where j = Σ xij/n. In the simulations that

[follow, we compare the selective Type I error (5) and power of the tests that reject](#bookmark26) H0 when these p-values are less than Q = 0.05.

**5.1 Selective Type I error under the global null**

[We generate data from (1) with μ = 0n](#bookmark19) ×q; therefore, H0 [in (2) holds for any pair of estimated](#bookmark21) clusters. We simulate 3,000 datasets with n = 150, σ = 1, and q = 2, 10, 50, 100.

For each simulated dataset, we apply k-means clustering with K = 3, and then compute pNaive , p selective , selective(MED ), and selective(Sample) for a randomly-chosen pair of clusters. Figure [3](#bookmark104) displays the observed p-value quantiles versus the Uniform(0,1) quantiles. We see that for all values of q, (i) the naive [p-values in (4) are stochastically smaller than a](#bookmark23) Uniform(0,1) random variable, and the test based on pNaive leads to an in且ated Type I error rate; (ii) tests based p selective , selective(MED ), and selective(Sample) control the selective [Type I error rate in the sense of (5)](#bookmark26).

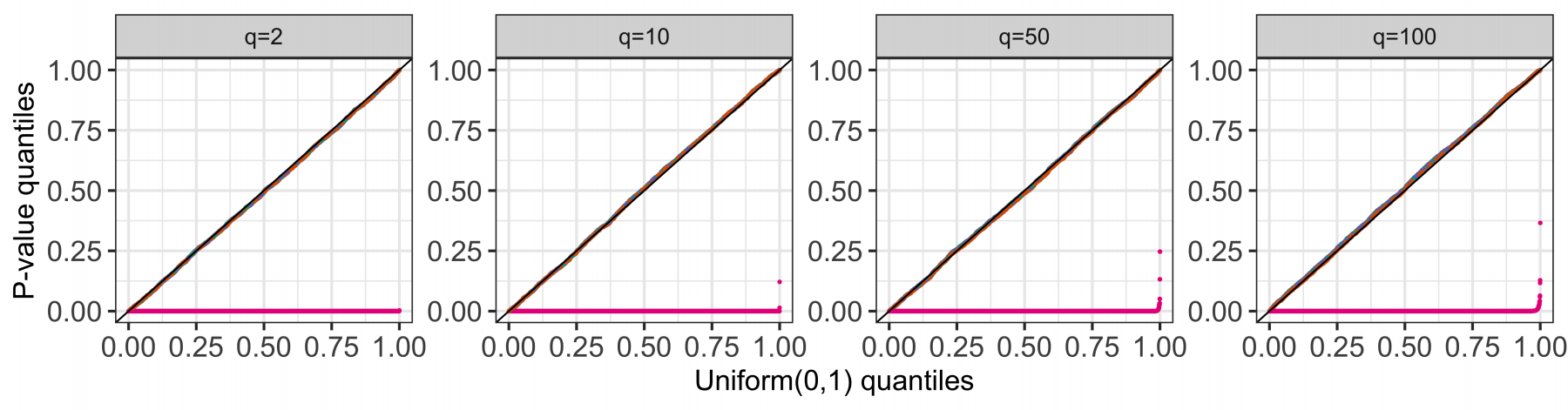
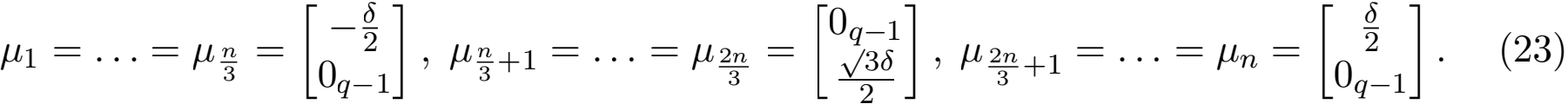


Figure 3: Quantile-quantile plots for pNaive (pink), p selective (green), selective(MED ) (or- ange), and selective([Sample) (purple) under (1) with](#bookmark19) μ = 0n ×q, stratiﬁed by q.

**5.2 Conditional power and detection probability**

In this section, we show that the tests based on our proposal (pselective , selective(MED ), and selective(Sample)) have substantial power to reject H0 when it is not true. We generate data [from (1) with](#bookmark19) n = 150 and



Here, we can think of C1 = {1, . . . , n/3}, C2 = {(n/3) + 1, . . . , (2n/3)}, C3 = {(2n/3) + 1, . . . , n} as the “true clusters” . Moreover, these clusters are equidistant in the sense that the pairwise distance between each pair of population means is jδj . Recall that we test H0

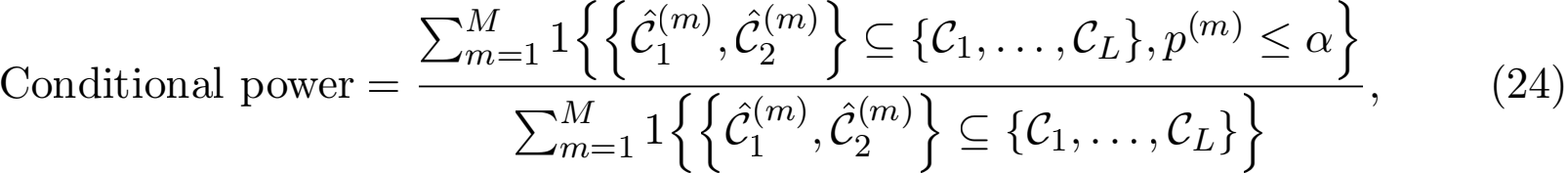
[in (2) for a pair of estimated clusters](#bookmark21) 1 and 2 , which may not be true clusters. Hence,

we will separately consider the conditional power and detection probability of our proposed [tests (Gao et al.,](#bookmark2) [2022;](#bookmark2) [Jewell et al.,](#bookmark36) [2022;](#bookmark36) [Hyun et al.,](#bookmark37) [2021)](#bookmark37). The conditional power is the

Chen and Witten

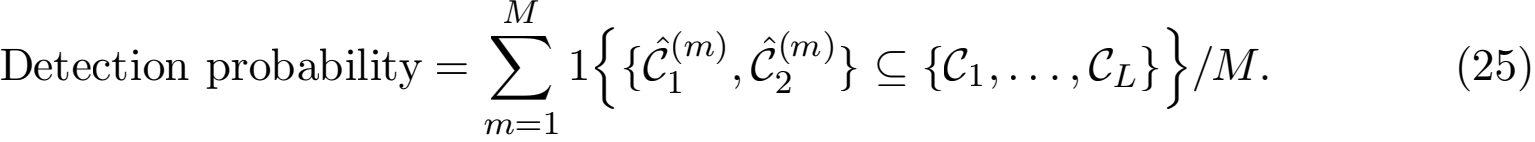
probability of rejecting H0 [in (2), given that](#bookmark21) 1 and 2 are true clusters. Given M simulated

datasets with true clusters {C1 , . . . , CL}, we estimate it as



where , and p(m) and 1(m) , 2(m) correspond to the p-value and clusters under consideration

for the mth simulated dataset. [Because the quantity in (24) conditions on the event that](#bookmark107) 1 and 2 are true clusters, we also estimate how often that event occurs:



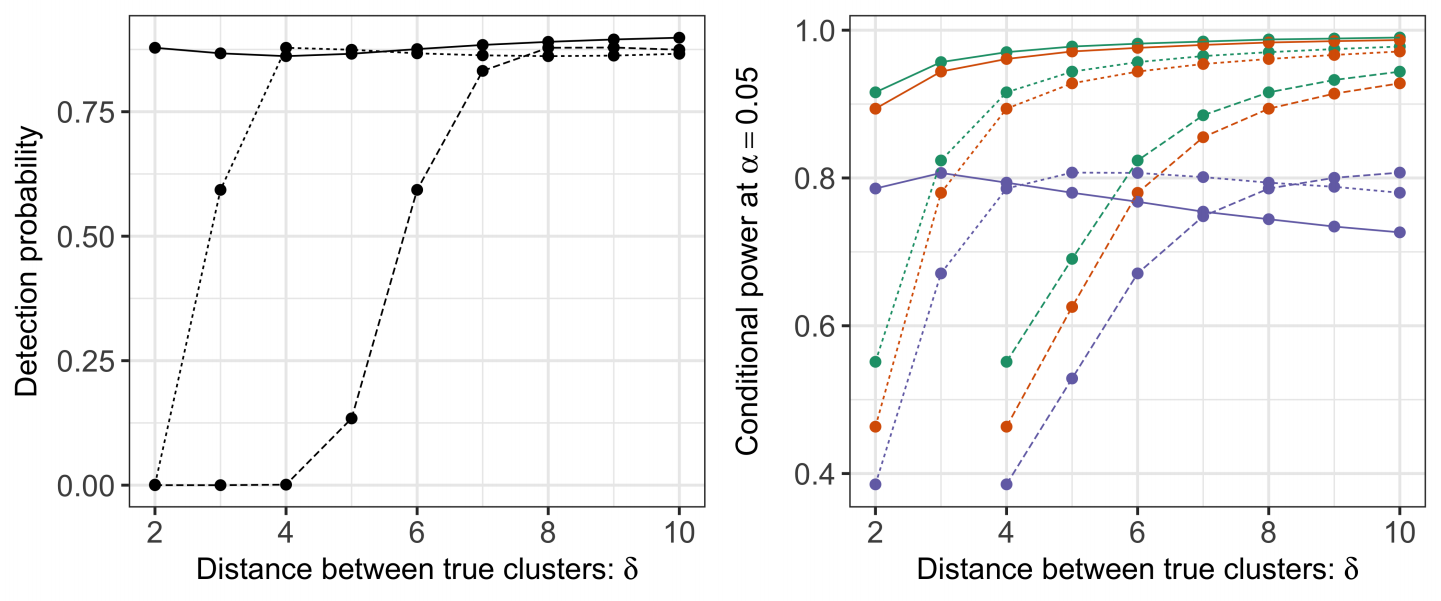


Figure 4: Left: [The detection probability (25) for](#bookmark108) k-means clustering with K = 3 under [model (1) with](#bookmark19) µ [deﬁned in (23), and](#bookmark106) σ = 0.25 (solid lines), 0.5 (dashed lines), and 1 (long-dashed lines). Right: [The conditional power (24) at](#bookmark107) α = 0.05 for the tests based on p selective (green), selective(MED ) (orange), and selective(Sample) [(purple), under model (1) with](#bookmark19) µ [deﬁned in (23) and](#bookmark106) σ = 0.25, 0.5, 1. The con- ditional power is not displayed for δ = 2, 3, σ = 1 because the true clusters were never recovered in simulation.

We generate M = 200, [000 datasets from (23) with](#bookmark106) q = 10, σ = 0.25, 0.5, 1, and δ = 2, 3, . . . , 10. For each simulated dataset, we apply k-means clustering with K = 3 and

reject H0 : µT ν = 0q if p selective , selective(MED ), or selective(Sample) is less than α = 0.05.

In Figure [4, the left panel displays the detection probability (25) of](#bookmark108) k-means clustering as a function of δ [in (23), and the right panel displays the conditional power (24) for](#bookmark106)

the tests based on p selective , selective(MED ), and selective(Sample). [Under model (1), the](#bookmark19)

detection probability and conditional power increase as a function of δ [in (23) for all values](#bookmark106) 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 δ = 2, 3, σ = 1 because the true clusters were never recovered in simulation. Moreover, for a given value of δ and σ , the test based on pselective has the highest conditional power, followed closely by the test

Selective inference for k-means clustering

based on selective(MED ). Using Sample in selective leads to a less powerful test, especially

for large values of δ . This is because Sample is a conservative estimator of σ [in (1), and its](#bookmark19)

bias is an increasing function of δ, the distance between true clusters. By contrast, MED

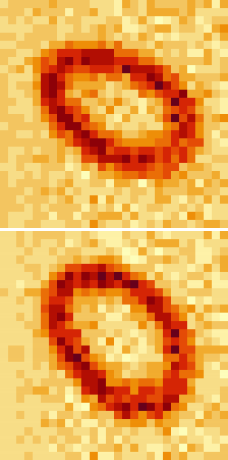
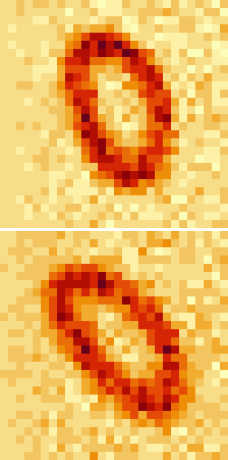
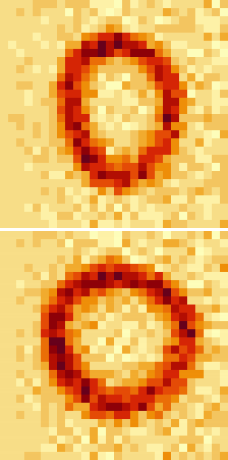
[is a consistent estimator under model (23) (see Appendix](#bookmark106) [A.7)](#bookmark103).

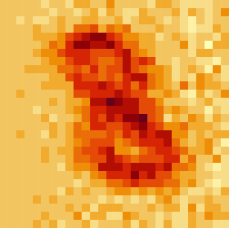
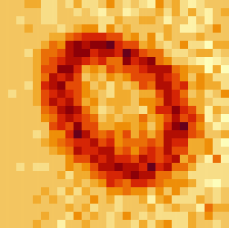
[As an alternative to the conditional power in (24), in Appendix](#bookmark107) [A.8, we consider a](#bookmark109) notion of power that does not condition on having correctly estimated the true clusters.

**6. Real data applications**

[**6.1 MNIST Dataset (Lecun et al.,**](#bookmark110)[**1998)**](#bookmark110)

[Here, we apply our method to the MNIST dataset (Lecun et al.,](#bookmark110) [1998), which consists of](#bookmark110) 60,000 gray-scale images of handwritten digits. Each image has an accompanying label in {0, 1, . . . , 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,](#bookmark111) [2018), which violates model (1), we add](#bookmark111) an independent perturbation N(0, 0.01) to each element of the image. Finally, we vectorize each image to obtain a vector xi ∈ R784 .





H0

“No cluster”

1 = 2 1 = 3 1 = 4 1 = 5 1 = 6 2 = 3 2 = 4 2 = 5 2 = 6 3 = 4 3 = 5 3 = 6 4 = 5 4 = 6 5 = 6

“Cluster”

1 = 2 1 = 3 1 = 4 2 = 3 2 = 4 3 = 4

pNaive

< 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10

< 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10

selective(MED )

< 10-10

0.01

0.22

< 10-10

0.53

< 10-10

1.4 × 10-8

< 10-10

0.49

0.14

1.1 × 10-6

0.45

2.0 × 10-8

0.77

0.48

< 10-10

8.0 × 10-6

6.2 × 10-7

6 × 10-3

10-3

4 × 10-4

Figure 5: Top left: Centroids of six clusters from the “no cluster” dataset (1 to 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 pNaive and selective(MED ). Here, i = Σj∈i μj/|i| .

We ﬁrst construct a “no cluster” dataset by randomly sampling 1,500 images of the 0s; thus, n = 1, 500 and q = 784. To de-correlate the pixels in each image, we whitened the

data (see Section [4.1) using](#bookmark66) -  = U (Λ + 0.01**I**n)-  [as in prior work (Coates and Ng,](#bookmark113)

[2012), where](#bookmark113) UΛUT is the eigenvalue decomposition of the sample covariance matrix.

We apply k-means clustering with K = 6. The centroids are displayed in the top left panel of Figure [5.](#bookmark112) For each pair of estimated clusters, we compute the p-values pNaive

Chen and Witten

and selective(MED ) (see Figure [5)](#bookmark112). 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 selective(MED ) is small are visually quite diﬀerent (e.g., clusters 1 and 2, clusters 1 and 5, and clusters 4 and 5).

To demonstrate the power of the test based on selective(MED ), we also generated a “cluster” dataset by sampling 500 images each from digits 0 , 1, 3, and 8; thus, n = 2, 000 and q = 784. We again whitened the data to obtain uncorrelated features. After applying k-means clustering with K = 4, we obtain four clusters that roughly correspond to four digits: cluster 1, 94.0% digit 1; cluster 2, 72.4% digit 3; cluster 3, 83.6% digit 0; cluster 4, 62.4% digit 8 (see the bottom left panel of Figure [5)](#bookmark112). Results from testing for a diﬀerence

in means for each pair of clusters using pNaive and selective(MED ) are in Figure [5.](#bookmark112) Both

sets of p-values are small on this “cluster” dataset.

[**6.2 Single-cell RNA-sequencing data (Zheng et al.,**](#bookmark115)[**2017)**](#bookmark115)

In this section, we apply our proposal to single-cell RNA-sequencing data collected by [Zheng](#bookmark115) [et al.](#bookmark115) [(2017)](#bookmark115). Single-cell RNA-sequencing quantiﬁes gene expression abundance at the resolution of single cells, thereby revealing cell-to-cell heterogeneity in transcription and allowing for the identiﬁcation of cell types and marker genes. In practice, biologists often cluster the cells to identify putative cell types, and then perform a diﬀerential expression [analysis, i.e., they test for a diﬀerence in gene expression between two clusters (Stuart et al.,](#bookmark8) [2019;](#bookmark8) [L¨ahnemann et al.,](#bookmark5) [2020;](#bookmark5) [Gr¨un et al.,](#bookmark3) [2015)](#bookmark3). 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.](#bookmark115) [(2017) proﬁled 68,000 peripheral blood mononuclear cells, and classiﬁed](#bookmark115) them based on their match to the expression proﬁles of 11 reference transcriptomes from known cell types. We consider the classiﬁed cell types to be the “ground truth”, and use this information to demonstrate that our proposal in Section [2](#bookmark64) yields reasonable results.

[As in prior work (Gao et al.,](#bookmark2) [2022;](#bookmark2) [Du`o et al.,](#bookmark116) [2018), we ﬁrst excluded cells with low](#bookmark116) 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 log2 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 con- structed a “no cluster” dataset by randomly sampling 1,000 out of 10,224 memory T cells after pre-processing (thus, n = 1, 000 and q = 500). Since the gene expression levels are highly correlated, we ﬁrst whitened the data as described in Section [4.1](#bookmark66) by plugging in

-  = U (Λ + 0.01**I**n)-  [(Coates and Ng,](#bookmark113) [2012), where](#bookmark113) UΛUT is the eigenvalue decom-

position of the sample covariance matrix.

We applied k-means clustering to the transformed data with K = 5, and obtained ﬁve clusters consisting of 97, 223, 172, 165, and 343 cells, respectively (see Figure [8](#bookmark117) left panel

Selective inference for k-means clustering

in Appendix [A.9)](#bookmark118). For each pair of estimated clusters, we computed the p-values pNaive and

selective(MED ). The results are displayed in the top panel of Table [1.](#bookmark119) 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 Q = 0.05, the test based on pNaive concludes that all ﬁve estimated clusters correspond to distinct cell types (even after multiplicity correction). By contrast, our approach does not reject most of the null hypotheses; i.e., it ﬁnds no diference between expression levels of the estimated clusters. Because this “no cluster” dataset consists only of memory T cells, we believe that conclusion based on selective(MED ) aligns better with the underlying biology.

Table 1: P-values pNaive [in (4) and](#bookmark23) selective [in (21) with](#bookmark99) MED [deﬁned in (22) corresponding](#bookmark102)

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).

H0 μ1 = μ2 μ1 = μ3 μ1 = μ4 μ1 = μ5 μ2 = μ3 μ2 = μ4 μ2 = μ5 μ3 = μ4 μ3 = μ5 μ4 = μ5

- - - - - - - - - - - - - - - - - - - -

pNaive < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10

selective(MED ) 0.30 0.31 0.43 0.12 0.12 0.002 0.[10 0.005 0.04 0.05](100.0050.040.05)

H0 μ1 = μ2 μ1 = μ3 μ1 = μ4 μ1 = μ5 μ2 = μ3 μ2 = μ4 μ2 = μ5 μ3 = μ4 μ3 = μ5 μ4 = μ5

- - - - - - - - - - - - - - - - - - - -

pNaive < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10

selective(MED ) 4.0 × 10-4 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 < 10-10 5.0 × 10-8 < 10-10

Next, we construct a “cluster” dataset by randomly sampling 400 each of memory T cells, B cells, naive T cells, natural killer cells, and monocytes from the 43, 259 cells; thus, n = 2, 000 and q = 500. After whitening the data, we applied k-means clustering to obtain ﬁve clusters. We see that these clusters approximately correspond to the ﬁve diferent cell types (cluster 1: 82.5% naive T cells; cluster 2: 95.3% memory T cells; cluster 3: 99.2% B cells; cluster 4: 91.5% nature killer cells; cluster 5: 83.3% monocytes); estimated clusters are visualized in the right panel of Figure [8](#bookmark117) in Appendix [A.9.](#bookmark118) We evaluate the p-values pNaive and selective(MED ) for all pairs of estimated clusters, and display results in the bottom panel of Table [1.](#bookmark119) Both sets of p-values are extremely small on this dataset, which suggests that the test based on our p-value has substantial power to reject the null hypothesis when it does not hold.

**7. Discussion**

We have proposed a test for a diference in means between two clusters estimated from

[k-means clustering, under (1)](#bookmark19). 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](https://github.com/yiqunchen/KmeansInference-experiments.Next), we outline a few directions for future research.

While the [p-value in (9) leads to selective Type I error control, it conditions on more](#bookmark25) [information than is used to construct the hypothesis in (2)](#bookmark21). In practice, data analysts likely only make use of the ﬁnal cluster assignments (leading to the [p-value in (8)), as opposed to](#bookmark78) [all the intermediate assignments (leading to the p-value in (9))](#bookmark25). Empirically, conditioning

Chen and Witten

[on too much information results in a loss of power (Fithian et al.,](#bookmark29) [2014;](#bookmark29) [Jewell et al.,](#bookmark36) [2022;](#bookmark36) [Liu et al.,](#bookmark120) [2018)](#bookmark120). In future work, we will investigate the possibility of leveraging recent [developments in selective inference (Chen et al.,](#bookmark121) [2022;](#bookmark121) [Le Duy and Takeuchi,](#bookmark40) [2021;](#bookmark40) [Jewell](#bookmark36) [et al.,](#bookmark36) [2022) to compute the “ideal”](#bookmark36) [p-value (8)](#bookmark78). Another line of future work is to extend [our test for a pairwise diference in means to a diference among multiple groups (Kimes](#bookmark51) [et al.,](#bookmark51) [2017;](#bookmark51) [Suzuki and Shimodaira,](#bookmark49) [2006)](#bookmark49). This might further provide away to determine the number of clusters in k-means clustering.

We could also consider extending our proposal to other data generating models. The [normality assumption in (1) is critical to the proof of Proposition](#bookmark19) [1, because it guarantees](#bookmark79)

that under H0 [in (2),](#bookmark21) ⅡXTVⅡ2 , dir(XTV), and ΠX are pairwise independent. However,

this normality assumption is often violated in practice; for instance, in single-cell genomics, the data are count-valued and the variance of gene expression levels varies drastically with [the mean expression levels of that gene (Stuart et al.,](#bookmark8) [2019;](#bookmark8) [Eling et al.,](#bookmark122) [2018)](#bookmark122). This has motivated some authors to work with alternative models for gene expression including [Poisson (Witten,](#bookmark123) [2011), negative binomial (Risso et al.,](#bookmark123) [2018), and curved normal (Lin](#bookmark124) [et al.,](#bookmark125) [2021)](#bookmark125). 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)jf (X) have a known, computationally-tractable distribution (Rasines and](#bookmark126) [Alastair Young,](#bookmark126) [2021;](#bookmark126) [Leiner et al.,](#bookmark127) [2021)](#bookmark127).

**Acknowledgments**

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

**References**

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

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

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

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

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

Selective inference for k-means clustering

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chen and Witten

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

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

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

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

Adam Coates and Andrew Y Ng. Learning feature representations with K-Means. In Gr/egoire Montavon, Genevi、eve B Orr, and Klaus-Robert Mu…ller, editors, Neural Net- works: Tricks of the Trade: Second Edition, pages 561–580. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

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

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

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

Angelo Du、o, Mark D Robinson, and Charlotte Soneson. A systematic performance eval- uation of clustering methods for single-cell RNA-seq data. F1000Research, 7:1141, July 2018.

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

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

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

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

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

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

Selective inference for k-means clustering

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chen and Witten

David La…hnemann, Johannes Ko…ster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto At- tolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, An- tonio Cappuccio, Giacomo Corleone, Bas E Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M Keizer, Indu Kha- tri, Szymon M Kielbasa, Jan O Korbel, Alexey M Kozlov, Tzu-Hao Kuo, Boudewijn P F Lelieveldt, Ion I Mandoiu, John C Marioni, Tobias Marschall, Felix Mo…lder, Amir Nikne- jad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J Theis, Huan Yang, Alex Zelikovsky, Alice C McHardy, Benjamin J Raphael, Sohrab P Shah, and Alexander Scho…nhuth. Eleven grand challenges in single-cell data science. Genome Biology, 21(1):31, February 2020.

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

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

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

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

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

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

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

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

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

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

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

Selective inference for k-means clustering

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

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

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

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

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

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

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

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

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

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

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

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

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

Jonathan Taylor and Robert Tibshirani. Post-selection inference for l1-penalized likelihood models. The Canadian Journal of Statistics, 46(1):41–61, March 2018.

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

Selective inference for k-means clustering

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

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

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

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

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

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

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

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

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

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

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

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

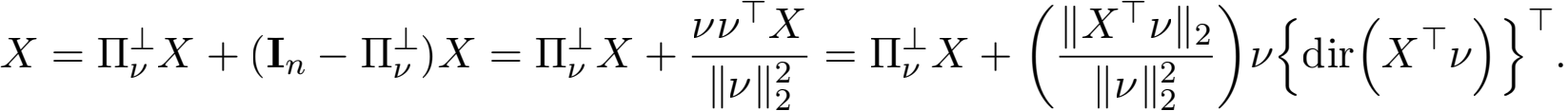
Chen and Witten

**Appendix A. Appendix**

**A.1 Proof of Proposition** [**1**](#bookmark79)

The proof of Proposition [1](#bookmark79) is similar to the proof of Theorem 1 in [Gao et al.](#bookmark2) [(2022), the](#bookmark2) proof of Theorem 3.1 in [Loftus and Taylor](#bookmark35) [(2015), the proof of Lemma 1 in](#bookmark35) [Yang et al.](#bookmark34) [(2016), and the proof of Theorem 3.1 in](#bookmark34) [Chen and Bien](#bookmark134) [(2020)](#bookmark134).

For any non-zero V ∈ Rn and X ∈ Rn ×q , we have that



(A.26)

**Lemma 8** Under [(1)](#bookmark19) and H0 : μTV = 0q, we have that ⅡXTVⅡ2 , ΠX, and dir(XTV) are

pairwise independent.

**Proof** We ﬁrst prove that XTV is independent of ΠX . The deﬁnition of Π implies that

ΠV = 0n, and it follows from the properties of the matrix normal distribution that ΠX

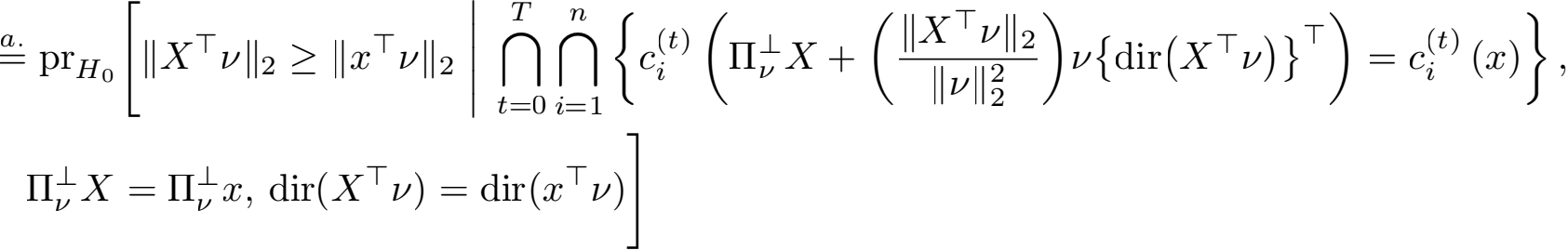
and XTV are independent. Therefore, ⅡXTVⅡ2 and dir(XTV) are independent of ΠX as

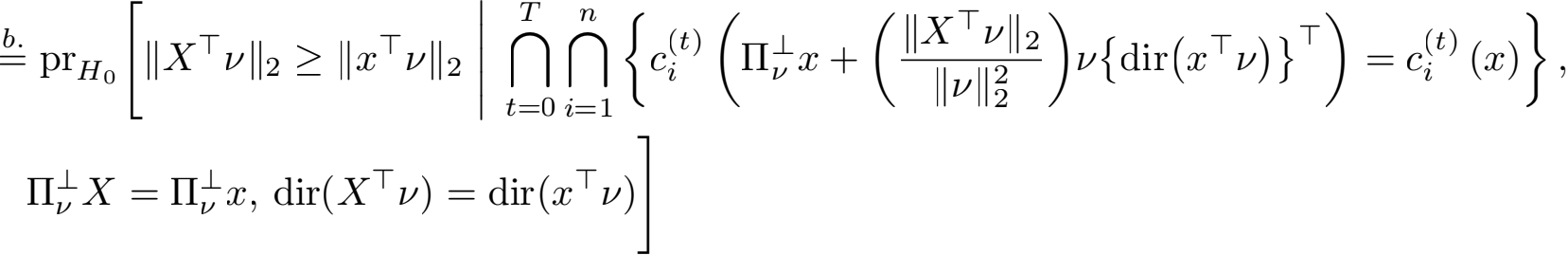
well, since both are functions of XTV .

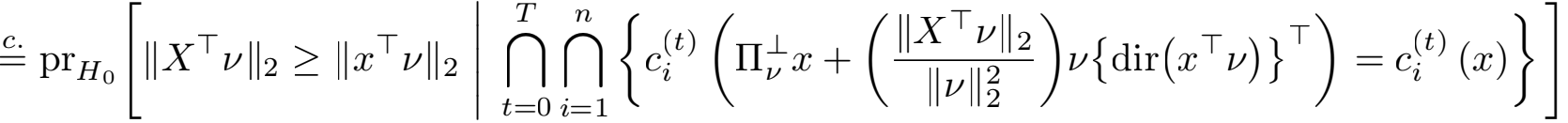
Next, we will show that ⅡXTVⅡ2 and dir(XTV) are independent. [Under (1) and](#bookmark19) H0 : μTV = 0q, we have that XTV ~ N(0q, σ2 ⅡVⅡ**I**q). It follows that XTV is rotationally in- variant, and therefore ⅡXTVⅡ2 is independent of dir(XTV) (see, e.g., Proposition 4.1 and Corollary 4.3 of [Bilodeau and Brenner](#bookmark20) [(1999))](#bookmark20). 

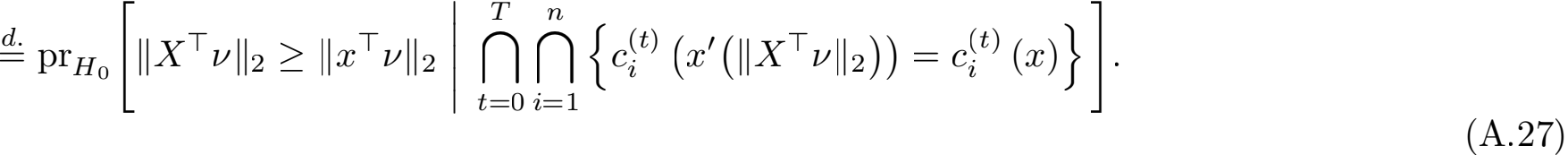
[We now proceed to prove the statement in (10)](#bookmark80). Recalling the deﬁnition of p selective in [](#bookmark22)









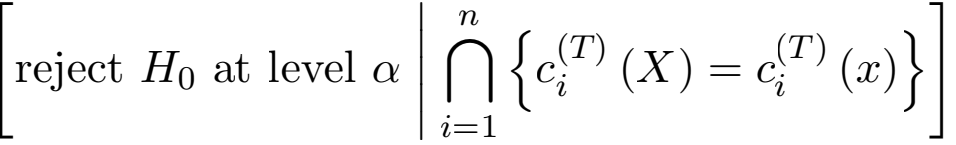


Selective inference for k-means clustering

Here, step a. follows from substituting X [with the expression in (A.26), and step](#bookmark143) b. follows from replacing ΠX and dir(XTν) with Πx and dir(xT ν), respectively. Next, in step c., we used Lemma [8.](#bookmark144) Finally, step d. follows from the deﬁnition of x/ [(φ) in (11)](#bookmark81).

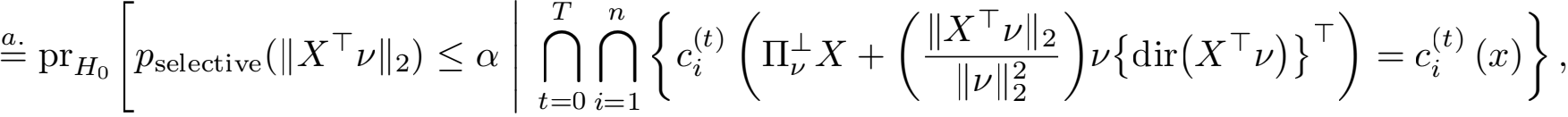
[Note that under (1) and](#bookmark19) H0 : µT ν = 0q, we have that ⅡXTνⅡ2 ~ σⅡνⅡ2 χq, which [concludes the proof of (10)](#bookmark80).

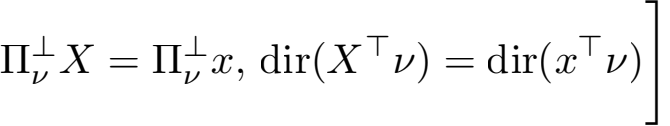
It remains to show that the test that rejects H0 : µT ν = 0 when p selective ≤ α controls the selective Type I error at level [α, in the sense of (5)](#bookmark26). First of all, recall that we decided [to test the null hypothesis in (2) based on the output of Algorithm](#bookmark21) [1.](#bookmark75) Therefore, p selective controls the selective Type I error at level α if, for any cT)(x), i = 1, . . . , n,

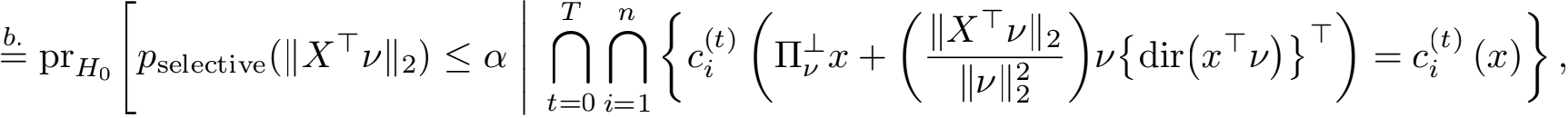
prH0  ≤ α , 8α ∈  (A.28)

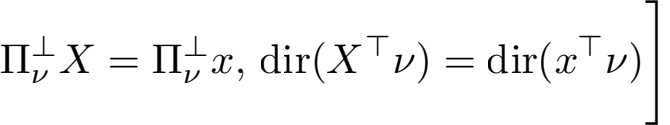
[](#bookmark146)

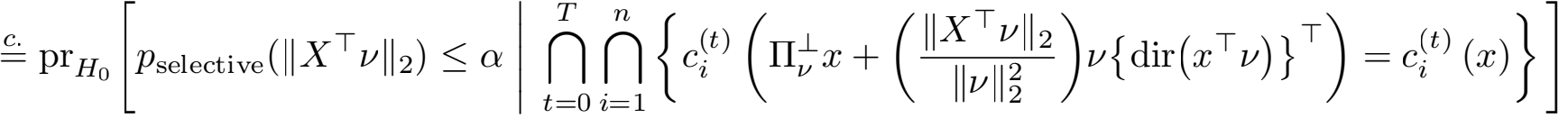


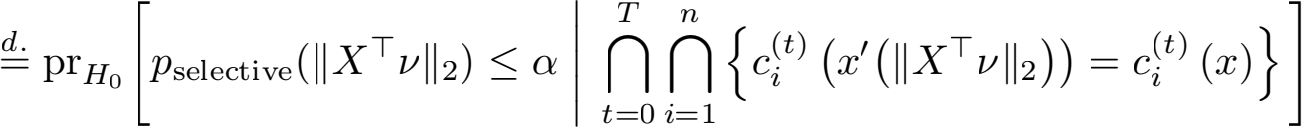


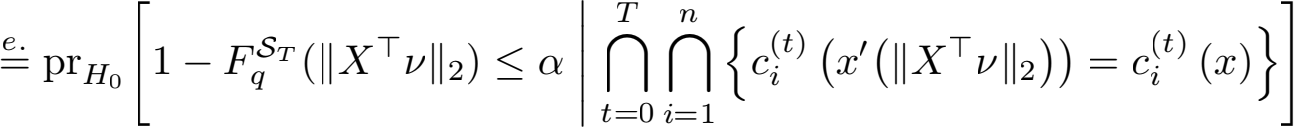














(A.29)

Here, steps a. through d. [follow from the same line of argument in (A.27)](#bookmark145). [Moreover, (10)](#bookmark80) implies that, for a given sequence of cluster assignments cT)(x), i = 1, . . . , n, p selective is the survival function of a χq random variable, truncated to the set ST [deﬁned in (12)](#bookmark82). Letting

FT (·) denote the cumulative distribution function of this truncated χq random variable, we

arrive at step e. Finally, to prove f., we ﬁrst note that under H0, the conditional cumulative

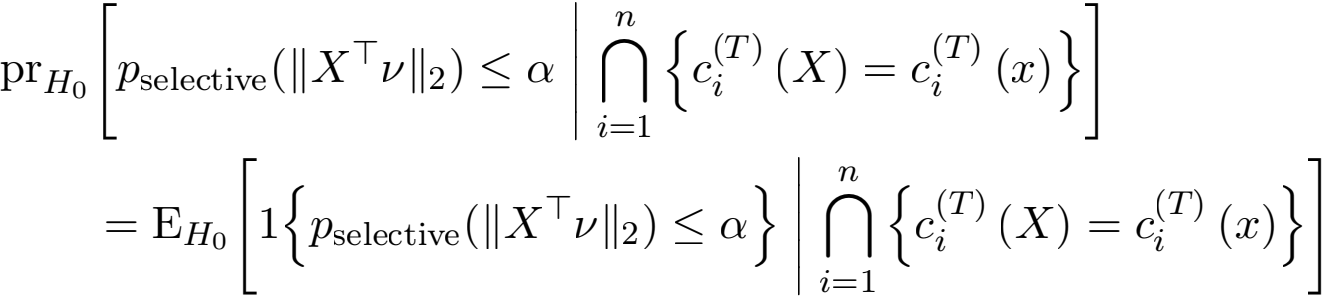
distribution function of ⅡXTνⅡ2 given ∩ ∩ct) (x/ (ⅡXTνⅡ2 )) = ct)(x)} is exactly

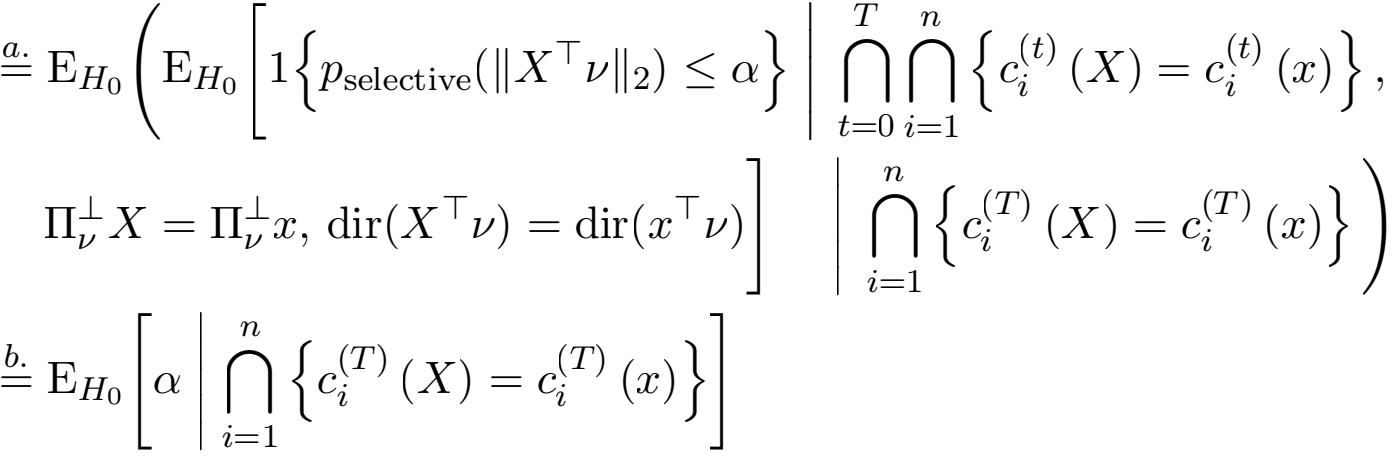
Chen and Witten

FT . The equality, therefore, follows from the probability integral transform, which states

that for a continuous random variable Z , FZ(Z) follows the Uniform(0,1) distribution.







= Q.

In the proof above, a. follows from the tower property of conditional expectation, and b. is [a direct consequence of (A.29)](#bookmark147).

Therefore, we conclude that the test based on p selective controls the selective Type I error [in (5), which completes the proof of Proposition](#bookmark26) [1.](#bookmark79)

**A.2 Proof of Proposition** [**2**](#bookmark85)

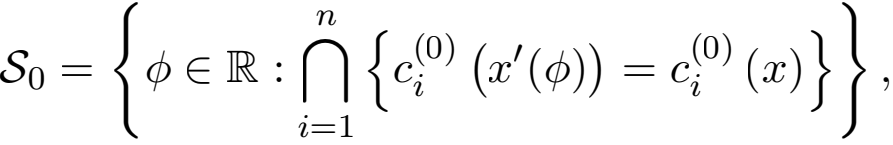
We will derive the expression for ST in Proposition [2](#bookmark85) using an induction argument. For a positive integer K, we let [K] denote the set {1, . . . , K}.

The following two claims (Lemmas [9](#bookmark86) and [10) serve as the “base cases” for the proof](#bookmark148).

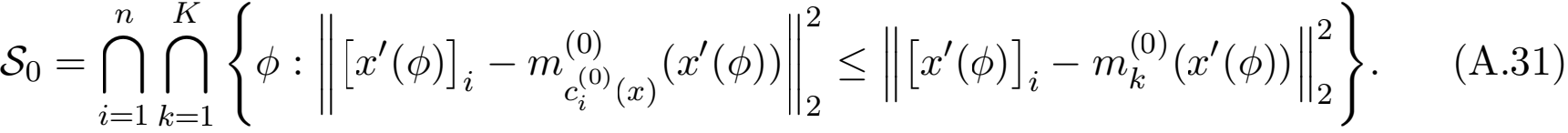
**Lemma 9** Recall that ct)(x) denotes the cluster to which the ith observation is assigned

during the tth iteration of step 3b. of Algorithm [1](#bookmark75) applied to data x, and that m(x)

denotes the kth centroid sampled from x during step 1 of Algorithm [1](#bookmark75). For S0 deﬁned as

 (A.30)

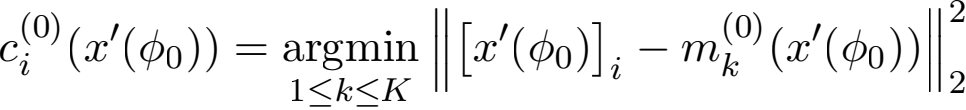
we have that

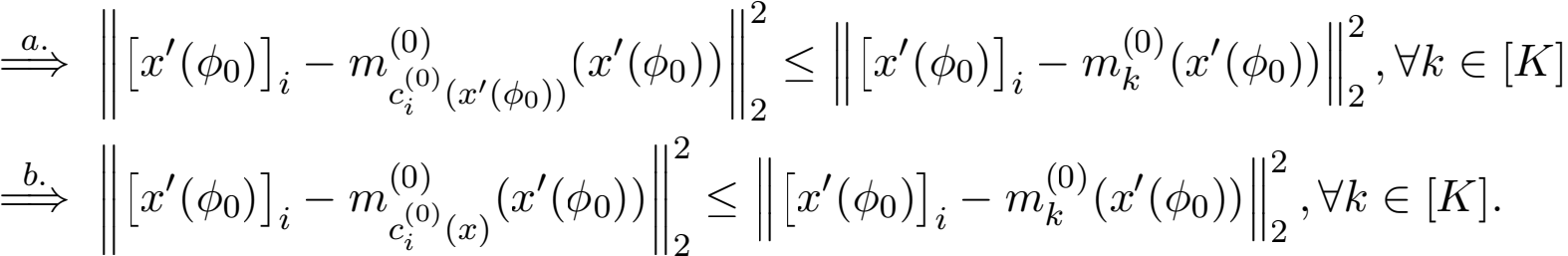


**Proof**

Selective inference for k-means clustering

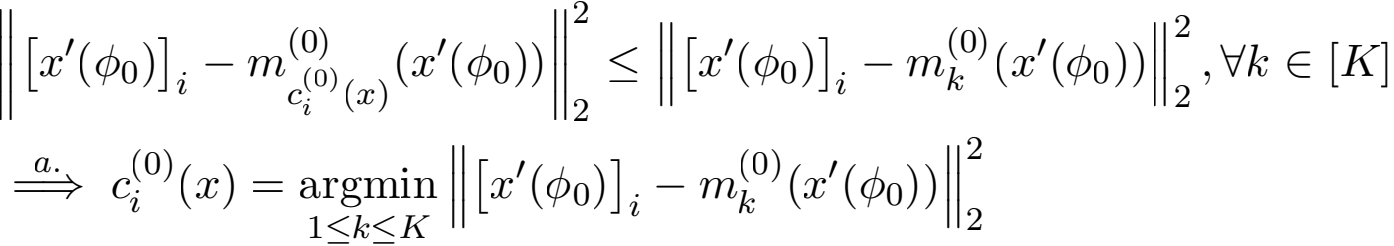
[We ﬁrst prove that the set in (A.30) is a subset of the set in (A.31)](#bookmark149). For an arbitrary φ0 ∈ [(A.30) and 1](#bookmark149) ≤ i ≤ n, we have that

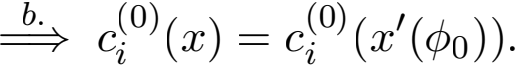




Here, the ﬁrst line follows from the deﬁnition of ci(0) in step 2 of Algorithm [1, and step](#bookmark75) a: follows from the deﬁnition of the argmin function. Step b: follows from the assumption that φ0 ∈ [(A.30) satisﬁes c](#bookmark149)0)(x, (φ0 )) = c0)(x). Because this holds for an arbitrary 1 ≤ i ≤ n, we have proven that φ0 ∈ [(A.30) =](#bookmark149)⇒ φ0 ∈ [(A.31); or equivalently, (A.30)](#bookmark150) ≤ [(A.31)](#bookmark150).

We proceed to prove the other direction. For an arbitrary φ0 ∈ [(A.31) and an arbitrary](#bookmark150) 1 ≤ i ≤ n, we have that





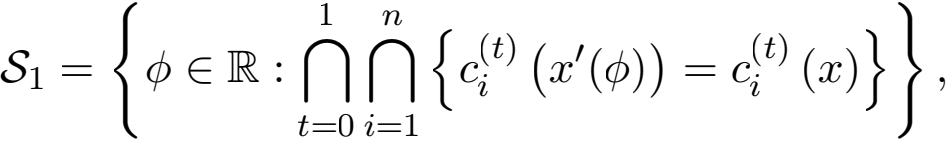
Here, step a: follows from the deﬁnition of argmin, and step b: follows from combining the deﬁnition of c0)(x, (φ)) in step 2 of Algorithm [1.](#bookmark75) We conclude that φ0 ∈ [(A.31) =](#bookmark150)⇒ φ0 ∈ [(A.30)](#bookmark149).

[Combining these two directions, we have proven that (A.31) = (A.30)](#bookmark150). 

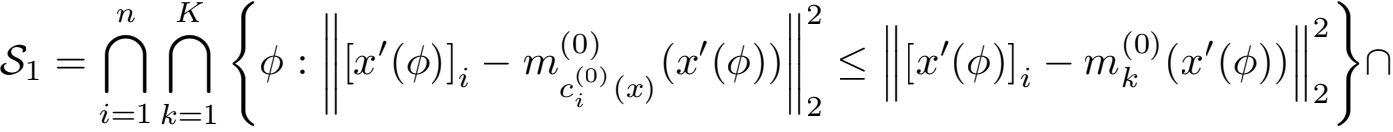
**Lemma 10** Recall that ct)(x) denotes the cluster to which the ith observation is assigned

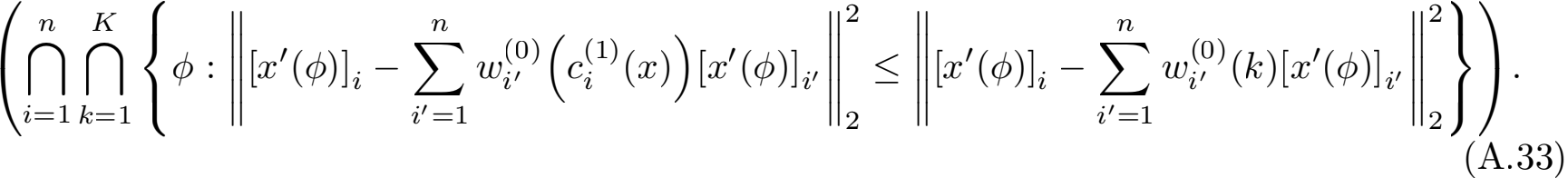
in the tth iteration of step 3b. of Algorithm [1](#bookmark75) applied to data x, and that m(x) denotes

the kth centroid sampled from x during step 1 of Algorithm [1](#bookmark75). For S1 deﬁned as

 (A.32)

and wi(t)(k) deﬁned in [(14), we have that](#bookmark87)

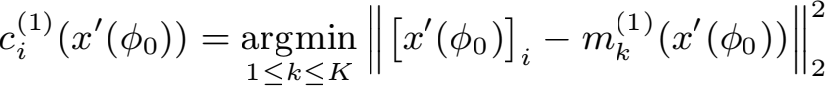


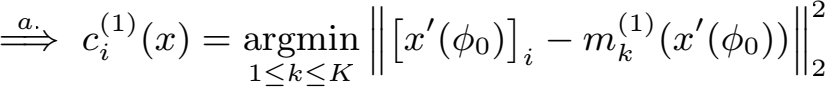


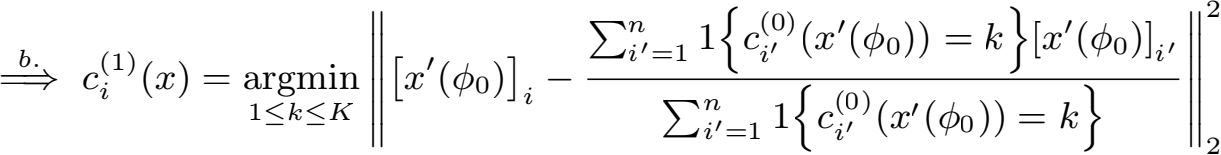
Chen and Witten

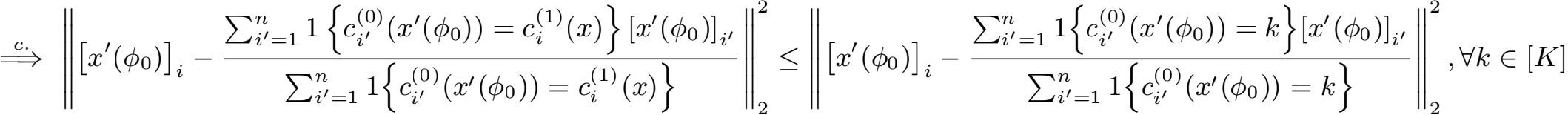
**Proof**

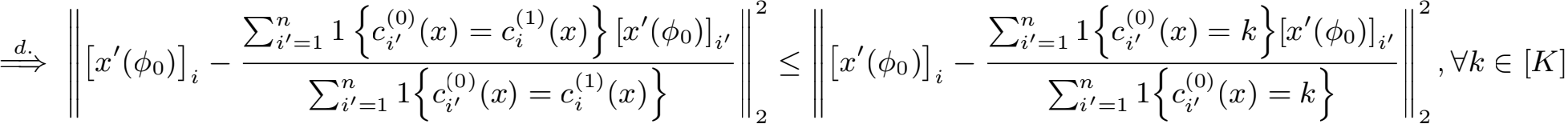
[We ﬁrst prove that (A.32)](#bookmark151) ≤ [(A.33)](#bookmark152). For an arbitrary φ0 ∈ [(A.32) and an arbitrary](#bookmark151) 1 ≤ i ≤ n, we have that

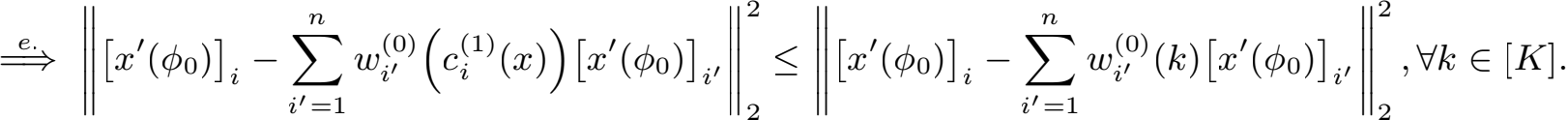










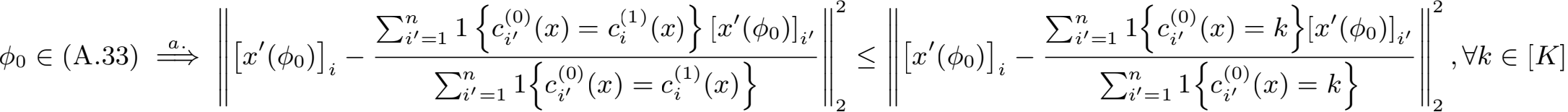


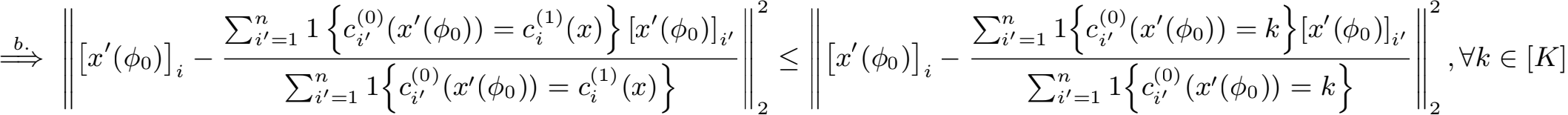
In the equations above, the ﬁrst line follows from step 3b. of Algorithm [1](#bookmark75)with t = 0. Next, step a. [follows from the deﬁnition of (A.32), which implies that c](#bookmark151)1)(x/ (φ0 )) = c1)(x). Step

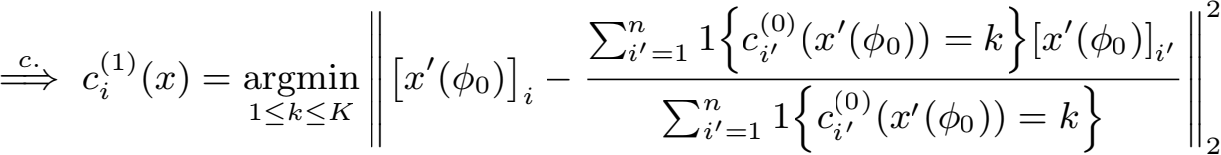
b. is a direct consequence of step 3a. of Algorithm [1](#bookmark75) with t = 0. In steps c. and d., we used [the deﬁnitions of the argmin function and (A.32)](#bookmark151). Finally, we apply the deﬁnition of wi(t) [in (14) to get](#bookmark87) e. Because this holds for an arbitrary 1 ≤ i ≤ n, φ0 ∈ [(A.32) implies that](#bookmark151) φ0 [is an element of the second set in the intersection in (A.33)](#bookmark152).

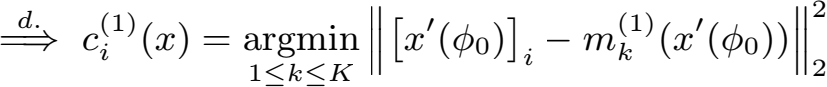
Moreover, φ0 ∈ [(A.32) implies that](#bookmark151) φ0 ∈ [(A.30), which, according to Lemma](#bookmark149) [9, further](#bookmark86) implies that φ0 [is an element of the ﬁrst set in the intersection in (A.33)](#bookmark152). To summarize, we have proven that φ0 ∈ [(A.32) =](#bookmark151)) φ0 ∈ [(A.33), and as a result, (A.32)](#bookmark152) ≤ [(A.33)](#bookmark152).

[Next, we prove that the set in (A.33) is a subset of the set in (A.32)](#bookmark152). For an arbitrary φ0 ∈ [(A.33) and an arbitrary 1](#bookmark152) ≤ i ≤ n, we have that

[](#bookmark152)







 ci(1) (x) = ci(1) (x, (φ0 )).

Here, step a. [follows from the deﬁnition of (A.33)](#bookmark152). In step b., we ﬁrst apply Lemma [9,](#bookmark86) [which implies that (A.33)](#bookmark152) ≤ [(A.31)](#bookmark150). Therefore, φ0 ∈ [(A.33) =](#bookmark152)) c0)(x) = c0)(x/ (φ0 )), for all i = 1, . . . , n, k = 1, . . . , K, yielding the desired equality. Next, step c. follows from the

Selective inference for k-means clustering

deﬁnition of the argmin function. Finally, steps d. and e. follow directly from the deﬁnitions

of mk(t) and ci(t) in steps 3a. and 3b. of Algorithm [1, respectively.](#bookmark75)

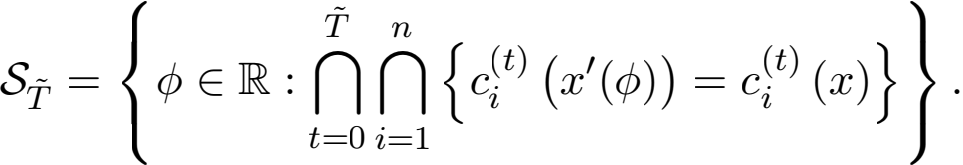
Because the result above holds for an arbitrary i, we have that φ0 ∈ [(A.33) =](#bookmark152)) c1)(x) = c1)(x/ (φ)), i = 1, . . . , n. [Combining this result with the observation that (A.33)](#bookmark152) ≤ [(A.31), we have that (A.33)](#bookmark150) ≤ [(A.32), which concludes the proof](#bookmark151). 

Next, we will prove the inductive step in the proof of Proposition [2, which relies on the](#bookmark85) following claim.

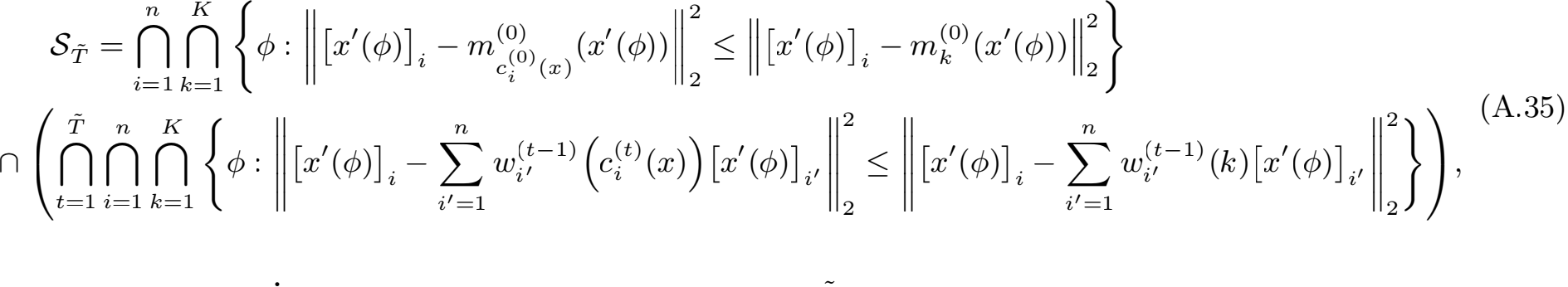
**Lemma 11** Recall that ct)(x) denotes the cluster to which the ith observation is assigned

in the tth iteration of Algorithm [1](#bookmark75) applied to the data x, and˜that m(x) denotes the kth

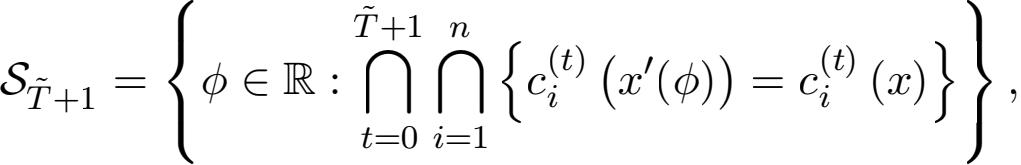
centroid sampled from x during initialization. For some 1 ≤ T ≤ T — 1, deﬁne

 (A.34)

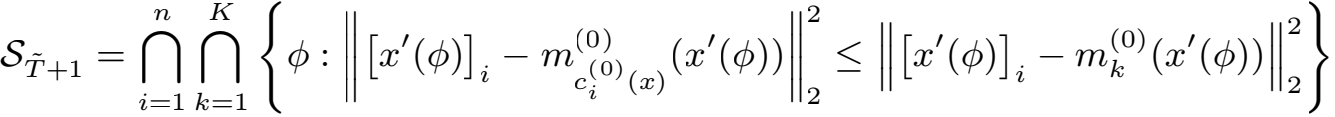
Suppose that the following holds for :

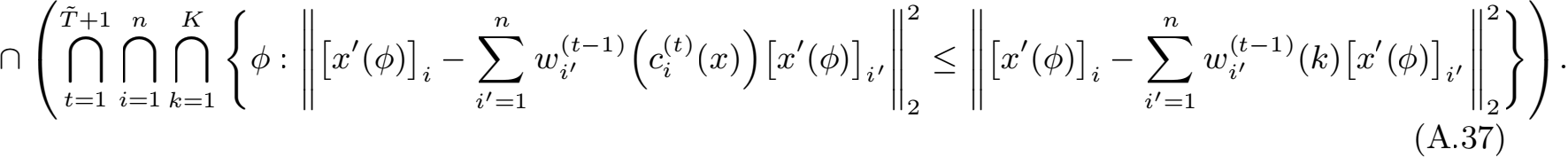


where wi(t)( ) is deﬁned in [(14)](#bookmark87). Then, for ST+1 deﬁned as

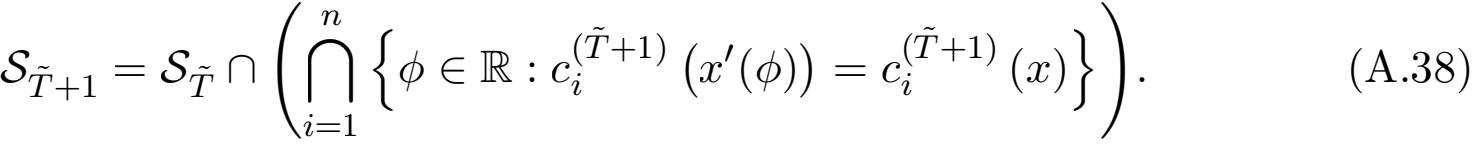
 (A.36)

we have that





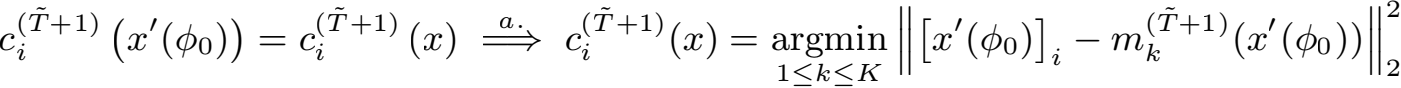
**Proof** [Using the deﬁnitions in (A.34) and (A.36), we have that](#bookmark154)

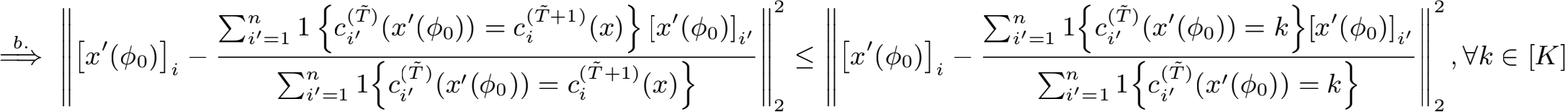


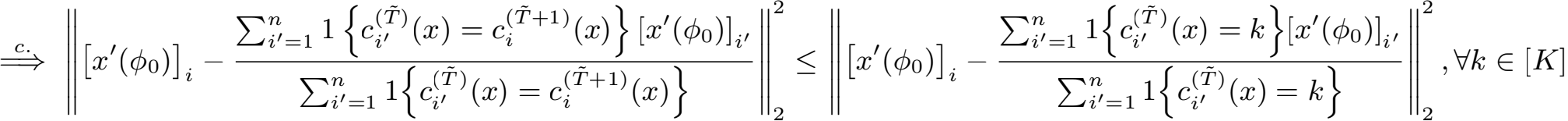
[Therefore, it su伍ces to prove that (A.38) = (A.37), under the inductive hypothesis](#bookmark158) [(A.35)](#bookmark155).

Chen and Witten

[We start by proving that (A.38)](#bookmark158) ≤ [(A.37)](#bookmark157). For an arbitrary φ0 ∈ [(A.38) and an arbitrary](#bookmark158) 1 ≤ i ≤ n, we have that







 φ0 ∈ [(A.37)](#bookmark157).

Here, the ﬁrst statement ~follows from ~the deﬁnition of S+1 . Next, steps a. and b. follow

from the deﬁnitions of ci(T+1) and mT+1)(x/ (φ0 )) in steps 3b. and 3a. of Al~gorithm [1,](#bookmark75)

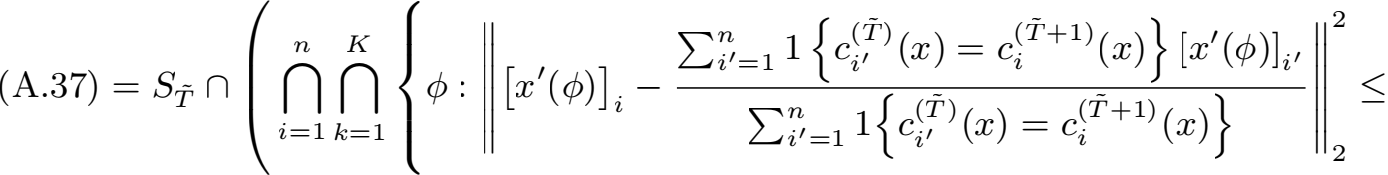
re~spectively. In step c., we used the fact that φ0 ∈ [(A.38) =](#bookmark158)) φ0 ∈ S =) c(x/ (φ0 )) =

c(x). Finally, d. follows from the deﬁnition of wi(t) [in (14)](#bookmark87).

We continue with the reverse direction. [Applying the inductive hypothesis (A.35), to](#bookmark155)-

gether with the deﬁnition of S+1 [in (A.37) and the deﬁnition of](#bookmark157) wi(t) [in (14), we have](#bookmark87)

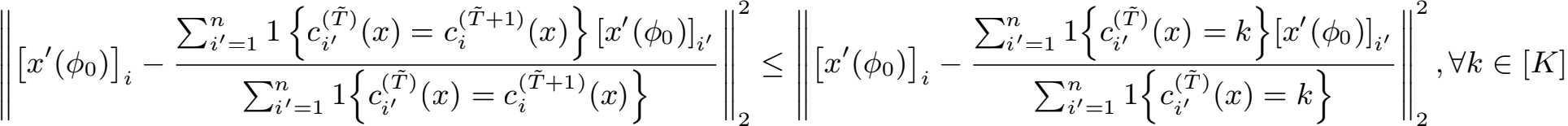
that

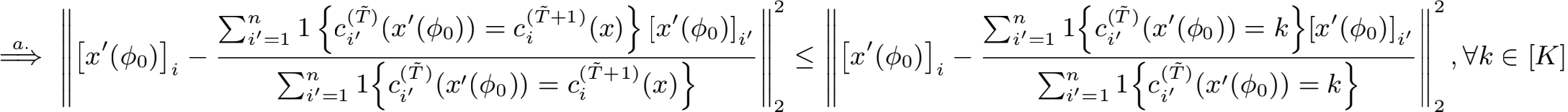
[](#bookmark157)

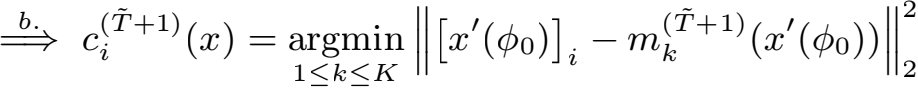
(A.39)



For an arbitrary φ0 ∈ [(A.37) and any 1](#bookmark157) ≤ i ≤ n, the following holds:









Here, to derive step a.[, we ﬁrst note that by (A.39), any element](#bookmark159) φ0 [of (A.37) is](#bookmark157)

als~o an element of S . Therefore, using the deﬁnition of S [in (A.34), we have that](#bookmark154)

∩ct)(x/ (φ0 )) = ct)(x)}, and step a. follows directly. ~Next, steps b. and c. follow

directly from steps 3a. and 3b. of Algorithm [1](#bookmark75) with t = T. By inspecting the form of [(A.38), we conclude that](#bookmark158) φ0 ∈ [(A.37) =](#bookmark157)) φ0 ∈ [(A.38)](#bookmark158).

[In conclusion, we have proven that (A.37) = (A.38), which completes the proof](#bookmark157). 

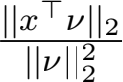
Selective inference for k-means clustering

The inductive proof of Proposition [2](#bookmark85) follows from combining Lemmas [9,](#bookmark86) [10](#bookmark148) and [11.](#bookmark153)

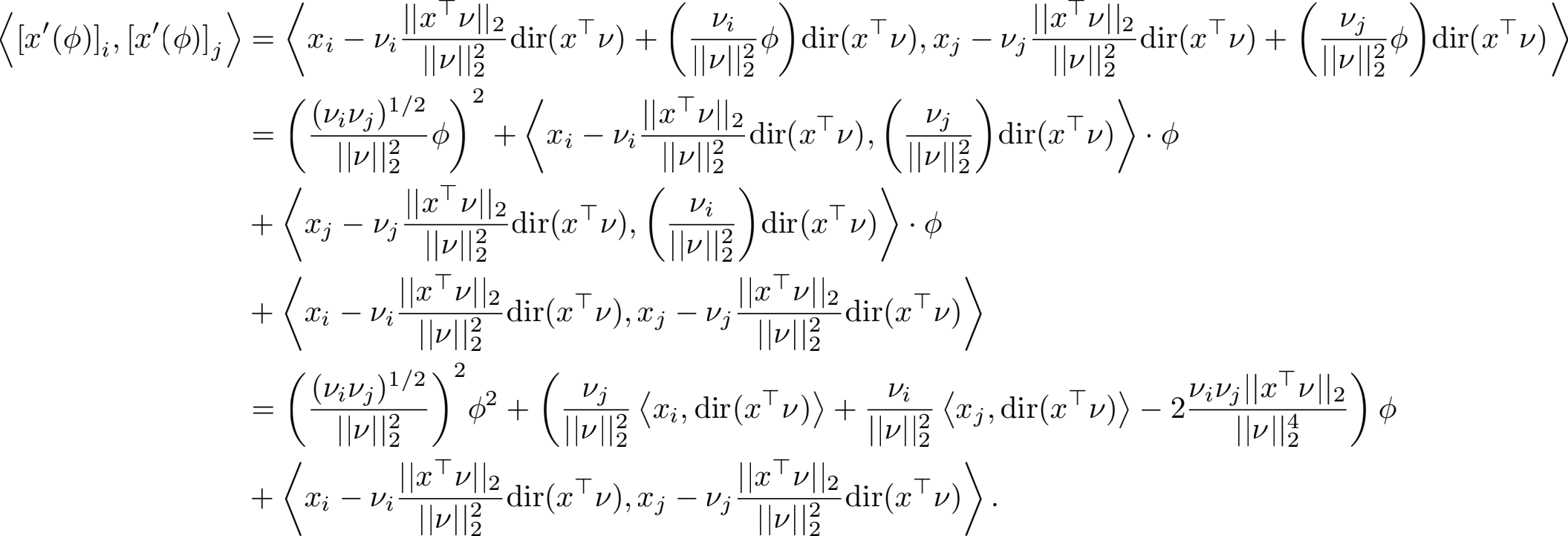
**A.3 Proof of Lemmas** [**3**](#bookmark90) **and** [**4**](#bookmark91)

We ﬁrst prove Lemma [3, which is also Lemma 2 in](#bookmark90) [Gao et al.](#bookmark2) [(2022)](#bookmark2).

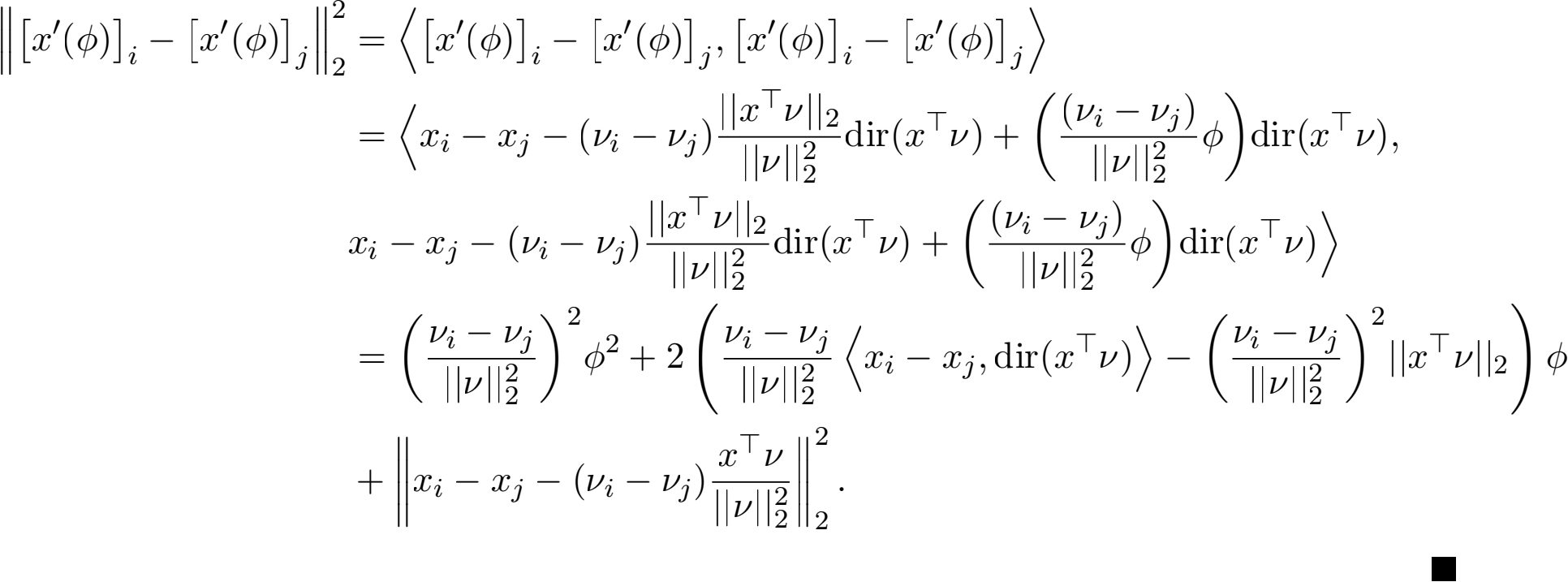
**Proof** We ﬁrst express the inner product 〈[xI (φ)]i, [xI (φ)]j〉 as a function of φ . [From (11), we](#bookmark81)

have that [xI (φ)]i = xi+Vi ( φ-νⅡ2 ) dir(xTV) = xi —Vi dir(xTV)+ ( jj φ)dir(xTV).

Therefore,

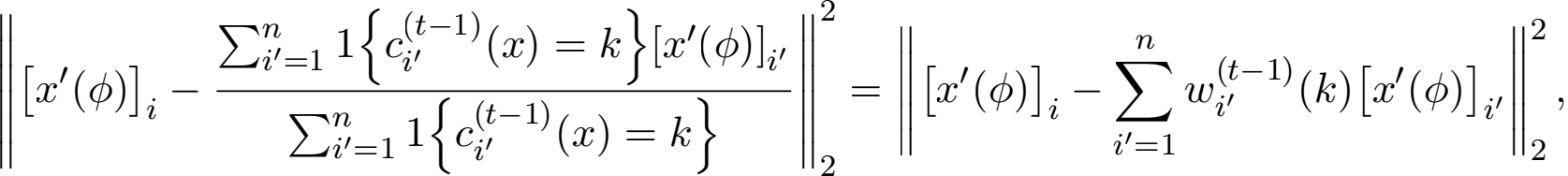


Next, using the expression for 〈[xI (φ)]i, [xI (φ)]j〉 above, we have that



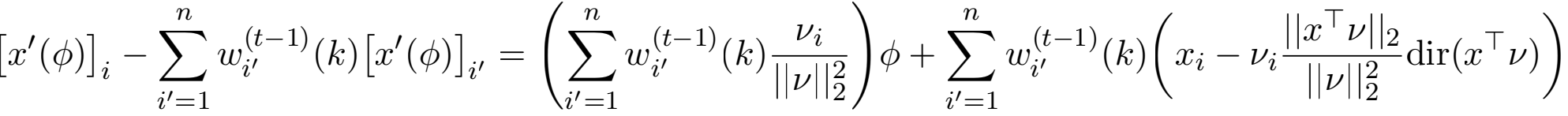
This completes the proof of Lemma [3.](#bookmark90)

We continue with the proof of Lemma [4.](#bookmark91) [Using the deﬁnition of w i(t-1)(k) in (14), we](#bookmark87) have that



Chen and Witten

where



is a linear function of φ . The rest of the proof follows directly from the same set of calculations in the proof of Lemma [3.](#bookmark90)

**A.4 Proof of Proposition** [**5**](#bookmark92)

Recall that n,q, K, T [denote the number of samples (see (1)), the number of features (see](#bookmark19) [(1)), the number of clusters (see Algorithm](#bookmark19) [1), and the maximum number of iterations for](#bookmark75) which Algorithm [1](#bookmark75) is run.

According to Proposition [2, to compute the set](#bookmark85) ST [in (12), it su伍ces to compute the](#bookmark82) [intersection of the two sets in (15) and (16)](#bookmark88).

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

• Observation 1: according to Lemma [3, the set in (15) is an intersection of](#bookmark90) nK quadratic inequalities.

• Observation 2: according to Lemma [4, the set in (16) is an intersection of](#bookmark91) nKT quadratic inequalities.

• Observation 3: we can solve a quadratic inequality in O(1) time using the quadratic formula.

• Observation 4: we can intersect the solution sets of N quadratic inequalities in [O(N log N) time (Bourgon,](#bookmark132) [2020)](#bookmark132).

Equipped with these observations, we will analyze the timing complexity of comput- [ing the set (15)](#bookmark88). Note that the coe伍cients for each of the nK quadratic inequalities can

be computed in O(nq) operations: ﬁrst, using the property that xTV = Σi∈1 xi/j1 j —

Σi∈2 xi/j2 j, we can compute ⅡxTVⅡ2 and dir(xTV) in O(nq) operations. Then, com-

puting the coe伍cients a,b, and √ in Lemma [3](#bookmark90) takes O(1), O(q), and O(q) operations, respectively. For each inequality, obtaining the solution set requires O(1) operations (see Observation 3). Finally, intersecting the solution sets of the n(K — 1) quadratic inequalities incurs another O(nK log(nK)) operations. [Thus, the computational cost for (15) totals to](#bookmark88) O(nKq + nK log(nK)) operations.

[Next, we analyze the cost of computing the set (16)](#bookmark89). Note that using Observation 2, we

need to solve nKT quadratic inequalities. Here, for each quadratic inequality of the form i˜n

Lemma [4, it takes](#bookmark91) O(n), O(n + q), and O(n + q) operations to compute the coe伍cients , b,



to compute Σ, =1 wi(—1)(k)Vi, once using O(n + q) operations once, as opposed to n times,

since this formula does not depend on the index i. Therefore, obtaining the nKT solution sets will take O((n + q)KT) time. Finally, intersecting these sets using Observation 4 adds another O(nKT log(nKT)) operations.

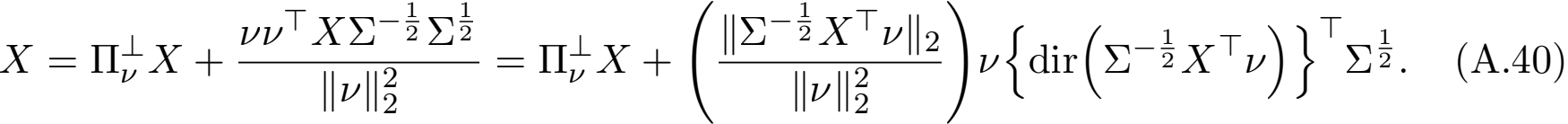
[Combining the costs for computing the set in (15) and the set in (16), we conclude that](#bookmark88) the cost for computing the set ST [in (12) is](#bookmark82) O((n + q)KT + nKT log(nKT)) operations.

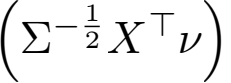
Selective inference for k-means clustering

**A.5 Proof of Proposition** [**6**](#bookmark95) **and computation of** pΣ ;**selective**

The proof of Proposition [6](#bookmark95) is similar to that of Proposition [1.](#bookmark79)

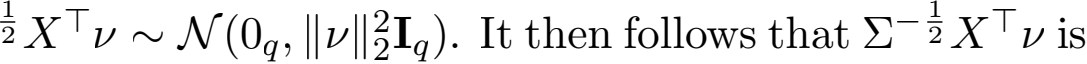
First note that for any non-zero V ∈ Rn and X ∈ Rn ×q , we have that



**Lemma 12** Under [(17)](#bookmark93) and H0 : μTV = 0q , ⅡΣ — XTVⅡ2 , X , and dir  are pairwise independent.

**Proof** As in the proof of Lemma [8, Π](#bookmark144)V = 0n, and it follows from the property of the

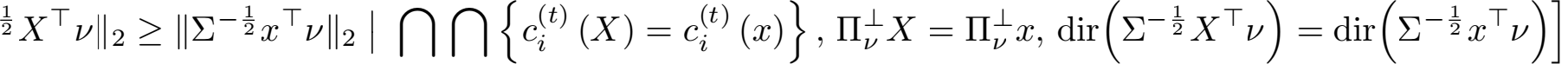
matrix normal distribution that XTV is independent of ΠX . Because both ⅡΣ — VⅡ2 and dir(Σ — V) are functions of XTV, both are independent of ΠX .

Next, we will show that ⅡΣ — [](#bookmark93) and H0 : μTV = 0q, we have that Σ — 

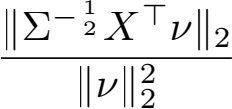
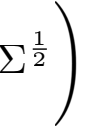
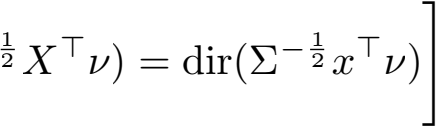
rotationally invariant, and therefore ⅡΣ — VⅡ2 is independent of dir(Σ — [V) (Bilodeau](#bookmark20) [and Brenner,](#bookmark20) [1999)](#bookmark20). 

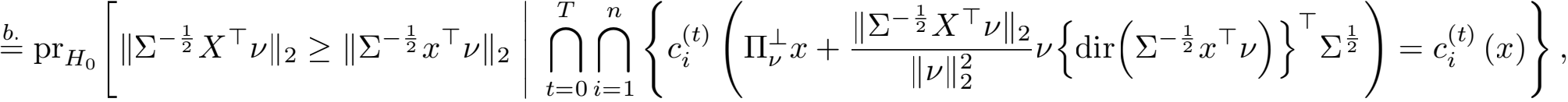
Then, recalling the deﬁnition of pΣ ;selective [in (19), we have that](#bookmark96)

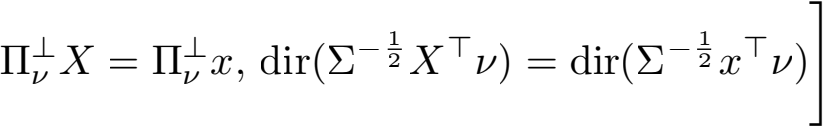
T n

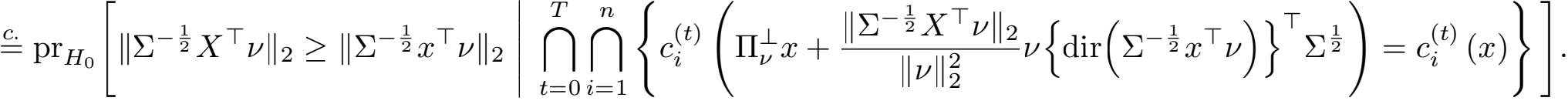
pΣ ,selective = prH0 [ ⅡΣ- 

t=0 i=1

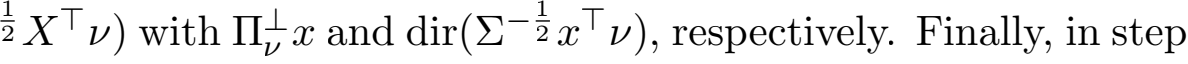
 prH0 [ ⅡΣ- TVⅡ2 ≥ ⅡΣ- TVⅡ2  t0 i1 {ci(t) (ΠX +  V {dir (Σ- TV)}T  = ci(t) (x)} , ΠX = Πx, dir(Σ- 







Here, step a. [follows from substituting X with the expression in (A.40)](#bookmark161). Step b. follows from

replacing ΠX and dir(Σ — 

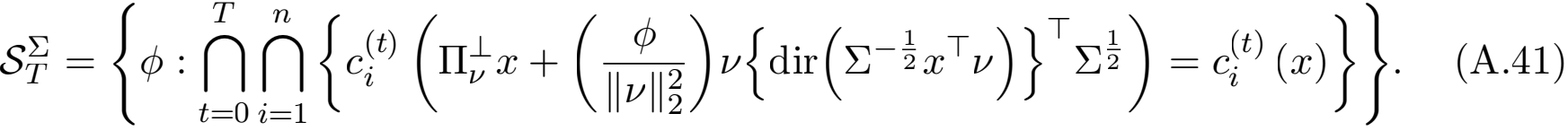
c., we used Lemma [12.](#bookmark161) [Now, under (17) and](#bookmark93) H0 : μTV = 0q, we have that ⅡΣ — VⅡ2 ~

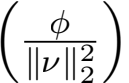
ⅡVⅡ2 [χq, which concludes the proof of (20)](#bookmark97).

It remains to show that the test that rejects H0 : μTV = 0 when pΣ ;selective ≤ Q controls [the selective Type I error, in the sense of (5)](#bookmark26). We omit the proof here, as it follows directly from the proof of Proposition [1](#bookmark79) in Appendix [A.1.](#bookmark142)

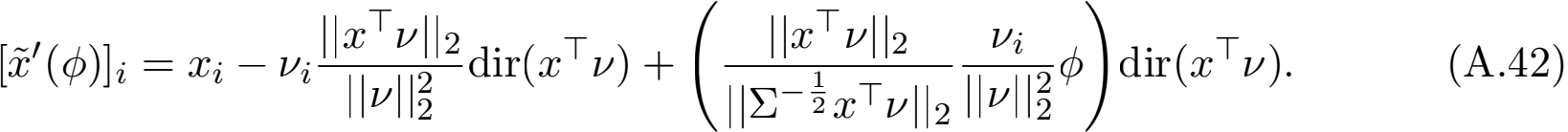
Chen and Witten

Next, we discuss how we could modify the results in Section [3](#bookmark65) to compute the p-value pΣ ,selective. First note that according to Proposition [6, it su伍ces to compute the set](#bookmark95)



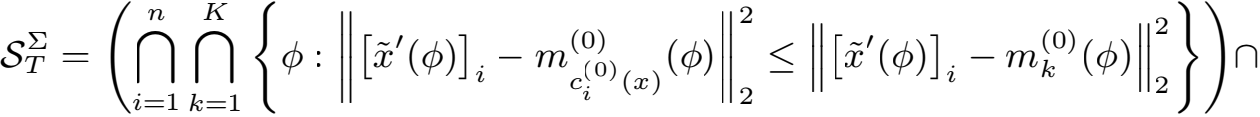
In addition, letting I (φ) denote Πx + v {dir (Σ- Tv)}T Σ  I (φ) is

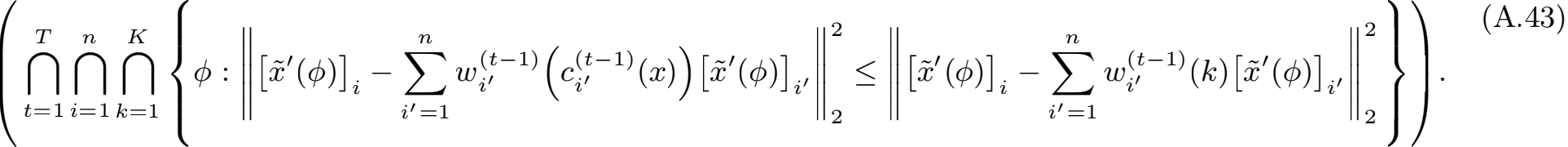
in fact a linear function of φ with



Therefore, a minor modiﬁcation of Proposition [2](#bookmark85) yields the following corollary.

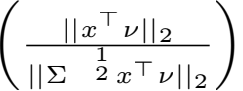
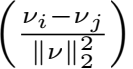
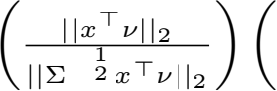
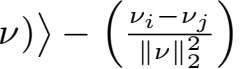
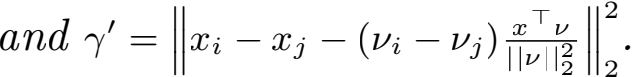
**Corollary 13** Suppose the k-means clustering algorithm (see Algorithm [1) with](#bookmark75) K clusters the data x, when applied to the data x, runs for T steps. Then, for the set S deﬁned in [(A.41), we have that](#bookmark162)





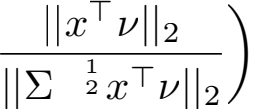
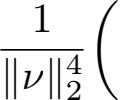
We also have the following extensions of Lemmas [3](#bookmark90) and [4, which enable e伍cient com](#bookmark91)- putation of the expressions in Corollary [13.](#bookmark164)

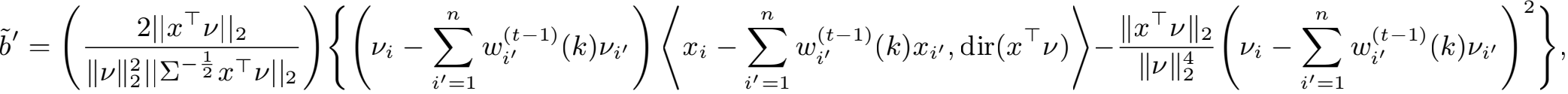
**Lemma 14 (Section 4.2 in** [**Gao et al.**](#bookmark2)[**(2022))**](#bookmark2)For I (φ) in [(A.42)](#bookmark163) and v in [(3)](#bookmark22), **"**/ (φ)]i - / (φ)]j **"** =

a/ φ2 +b/ φ+√/ , where a/ = - 2 2 , b/ = 2 -  〈xi - xj, dir(xT 2 ⅡxT νⅡ2 ), 

**L**~**emm**~**a 15** For I [(φ) in (A.42)](#bookmark163), [v in (3), and w i(t)(k)](#bookmark22) in [(14)](#bookmark87), **"**/ (φ)]i -Σ, =1 wi(—1)/ (φ)]i, **"** =

a/ φ2 + b/ φ 

I = - 2 (vi - iΣ, wi(-1)(k)vi, )2 ;





Proofs of Lemmas [14](#bookmark165) and [15](#bookmark166) follow from the same set of calculations in the proofs of Lemmas [3](#bookmark90) and [4](#bookmark91) in Appendix [A.3.](#bookmark160)

Selective inference for k-means clustering

**A.6 Proof of Proposition** [**7**](#bookmark100)

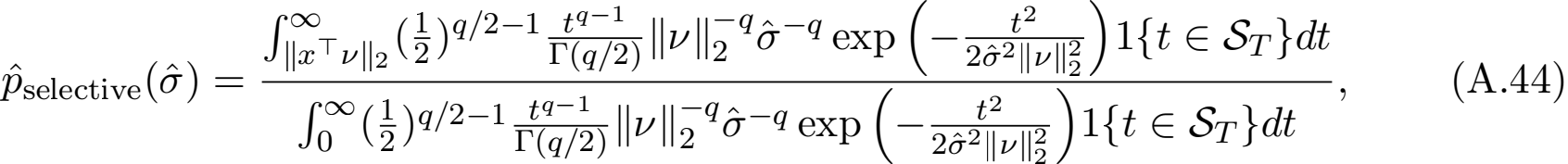
Proof of Proposition [7](#bookmark100) is similar to the proof of Lemma 1 in [Markovic et al.](#bookmark137) [(2017) and the](#bookmark137) proof of Lemma 7 in [Tibshirani et al.](#bookmark140) [(2018)](#bookmark140).

We ﬁrst present an auxiliary lemma.

**Lemma 16** For any ct)(x), i = 1, . . . , n; t = 1, . . . , T, selective() deﬁned in [(21)](#bookmark99) is a

continuous and monotonically increasing function of .

**Proof** [By the deﬁnition in (21), we have that](#bookmark99)



where ST [deﬁned in (12) is a function of c](#bookmark82)t)(x), i = 1, . . . , n; t = 1, . . . , T. By inspection,

[(A.44) is a continuous function of](#bookmark168) , because the product or ratio of two continuous func-

tions is still continuous. [It remains to show that (A.44) is increasing in](#bookmark168) . This follows

directly from Lemma S3. of [Gao et al.](#bookmark2) [(2022)](#bookmark2). 

Provided that converges to σ in probability, we can combine Lemma [16](#bookmark167) and the

continuous mapping theorem to see that selective() converges top selective(σ) in probability,

i.e., for all ∈ > 0, lim q→∞ pr(|selective() - p selective(σ)| ≥ ∈) = 0. Next, letting Aq denote

the event ∩ ∩ct) (X(q)) = ct) (x(q))}, we will show that under the assumptions in

Proposition [7,](#bookmark100) selective() converges to p selective(σ) in probability, conditional on Aq . For

any ∈ > 0, we have that

qlprH0(q) {| selective() - p selective(σ)| ≤ ∈ | Aq}

a.

=

b. ≥

c.

=

prH (q) {| selective() - p selective(σ)| ≤ ∈, Aq}

lim 0

q→∞ prH0(q) (Aq)

prH (q) (Aq) - prH (q) {| selective() - p selective(σ)| > ∈}

lim 0 0

q→∞ prH0(q) (Aq)

lim q→∞ prH (q) (Aq) - lim q→∞ prH (q) {| selective() - p selective(σ)| > ∈}

0 0

limq→∞ prH0(q) (Aq)



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, pr(A ∩ B) = pr(A) - pr(A \ B) ≥ pr(A) - pr(BC). Next, c. follows from distributing the limit, which is valid because of the assumption that lim q→∞ prH (q) (Aq) =

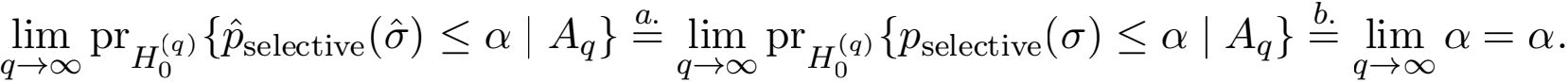
0

δ > 0; ﬁnally, d. follows from the fact that selective() converges top selective(σ) in probability

for any sequence of μ(q), q = 1, 2, . . ., which implies the convergence under H0 : μ(q)TV(q) = 0 as well.

Chen and Witten

Finally, we have that



(A.45)

Here, step a. follows from selective() converging to p selective(σ) in probability, conditional on Aq . Step b. follows from the fact that the result of Proposition [1](#bookmark79) applies for any positive integer q. This completes the proof of Proposition [7.](#bookmark100)

Proposition [7](#bookmark100) assumes that we have a consistent estimator of σ . In Appendix [A.7,](#bookmark103)

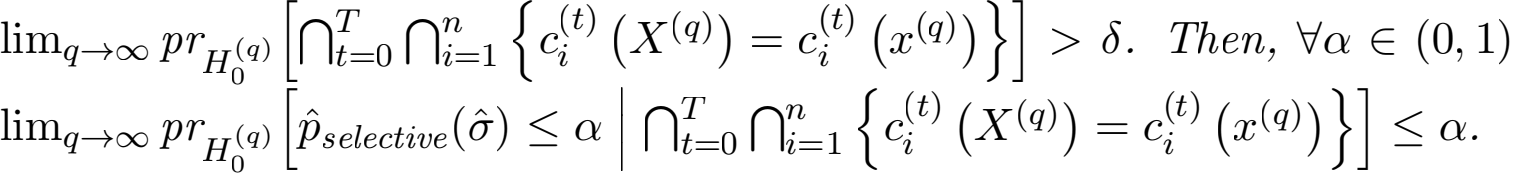
we analyze diferent estimators of σ [in (1), and prove that, under appropriate sparsity](#bookmark19) assumptions on µ [in (1),](#bookmark19) MED [in (22) is a consistent estimator for](#bookmark102) σ .

As an alternative, we can also use an asymptotically conservative estimator of σ as in [Gao et al.](#bookmark2) [(2022)](#bookmark2). This leads to an asymptotically conservative p-value; details are stated in Corollary [17.](#bookmark169)

**Corollary 17** For q = 1, 2, . . . , suppose that X(q) ~ MN n ×q (µ(q) , **I**n, σ2 **I**q) . Let x(q) be a realization from X(q) and ct)(·) be the cluster to which the ith observation is assigned during the tth iteration of step 3b. of Algorithm [1](#bookmark75). Consider the sequence of null hypothe- ses H0(q) : µ(q)Tν(q) = 0q, where ν(q) deﬁned in [(3)](#bookmark22) is the contrast vector resulting from

applying k-means clustering on x(q) . Suppose that (i) is an asymptotically conservative

estimator of σ, i.e., lim q→∞ pr((X(q)) ≥ σ ) = 1; and (ii) there exists δ ∈ (0, 1) such that

, we have that

We omit the proof of Corollary [17, as it follows directly from combining the proof of](#bookmark169) Proposition [7](#bookmark100) and the fact that selective() is a monotonically increasing function of (see Lemma [16)](#bookmark167).

Finally, we remark that, in principle, the result in Proposition [7](#bookmark100) can be extended to an unknown covariance matrix Σ . However, estimating Σ is challenging, especially when q is comparable to, or larger than, n [(Rousseeuw,](#bookmark139) [1987;](#bookmark139) [Bickel and Levina,](#bookmark131) [2008;](#bookmark131) [Avella-Medina](#bookmark128) [et al.,](#bookmark128) [2018)](#bookmark128). It may be possible to leverage recent advances in robust covariance matrix estimation (e.g., [Han and Liu](#bookmark136) [(2014);](#bookmark136)[Chen et al.](#bookmark133) [(2018);](#bookmark133) [Belomestny et al.](#bookmark130) [(2019)) to obtain](#bookmark130) [a consistent estimator of Σ under model (17)](#bookmark93).

**A.7 Estimating** σ **in** [(1)](#bookmark19)

Proposition [7](#bookmark100) states that, under appropriate assumptions, a consistent estimator of σ in [(1) leads to asymptotic selective Type I error control](#bookmark19). In this section, we analyze the

asymptotic behavior of the two variance estimators considered in Section [5,](#bookmark67) ED and

ample. [In particular, we prove that under model (1) and a sparsity assumption on](#bookmark19) µ

[(deﬁned in (1)), a close analog of](#bookmark19) ED [in (22) that does not subtract the column median](#bookmark102)

is aconsistent estimator of σ 2 . Moreover, we prove that ample 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.,](#bookmark135) [2021)](#bookmark135). For a vector θ ∈ Rn, we use ⅡθⅡ0 to denote its l0 norm, i.e. ⅡθⅡ0 = Σ 1{θi  0}.

Selective inference for k-means clustering

**Lemma 18 (Proposition 6 in** [**Comminges et al.**](#bookmark135)[**(2021))**](#bookmark135)Consider the model

Yi = θi + σξi, i = 1, . . . , d, (A.46)

where σ is unknown, and the independently and identically distributed noise ξi satisﬁes that

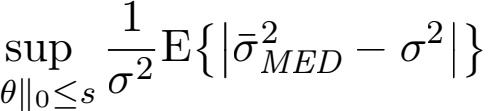
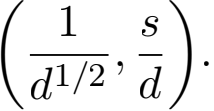
(i) E(ξi) = 0; (ii) E(ξ) = 1; and (iii) E(|ξi| 2+∈) < ∞ for some ∈ > 0. We further assume

that the signal θ is s-sparse, i.e., ⅡθⅡ0 ≤ s. Denoting by Mξ the median of ξ, we consider

the following estimator of σ2 :

ED = median(Y12 , . . . , Yd2 )/Mξ . (A.47)

Then, there exist constants γ ∈ (0, 1/8), C > 0 depending only on the cumulative distribu- tion function of ξ1 such that for all integers s and d satisfying 1 ≤ s < γd,

 Ⅱ ≤ C max  (A.48)

Building on Lemma [18, in Corollary](#bookmark170)[19, we analyze the properties of an estimator closely](#bookmark173)

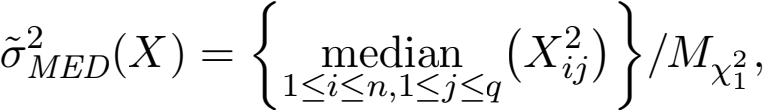
related to ED [in (22)](#bookmark102). In particular, this estimator ED does not subtract the median

of each column in the input data. While ED and ED are very similar provided that µ

is sparse, we expect ED to perform better empirically in scenarios where µ is sparse up

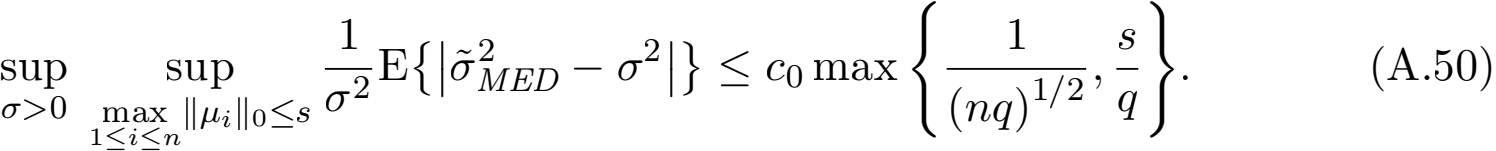
to a constant shift, i.e., there exists a matrix C such that (i) each column of C takes on the same value; and (ii) µ + C is sparse.

**Corollary 19** Under model [(1), consider](#bookmark19)

 (A.49)

where Mχ is the median of the χ distribution. Then, there exist constants γ0 ∈ (0, 1/8),

c0 > 0 such that for all integers s and q satisfying 1 ≤ s < γ0 q,



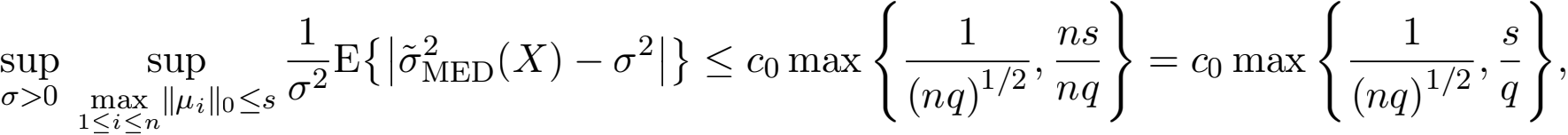
**Proof** [First note that (1) can be re-written into the form of (A.46):](#bookmark19)

Xij = µij + σξij, i = 1, . . . , n, j = 1, . . . , q, (A.51)

where ξij is independently and identically distributed as N(0, 1). Therefore, the estima- tor ED [(X) in (A.49) is the estimator (A.47) applied to the model (A.51)](#bookmark174). Moreover,

1n ⅡµiⅡ0 ≤ s implies that Σ Σ=1 1{µij  0} ≤ ns. Applying Lemma [18, we have](#bookmark170)

that



Chen and Witten

where c0 is some universal constant. 

In words, Corollary [19](#bookmark173)[states that under model (1](#bookmark19)ED in mean

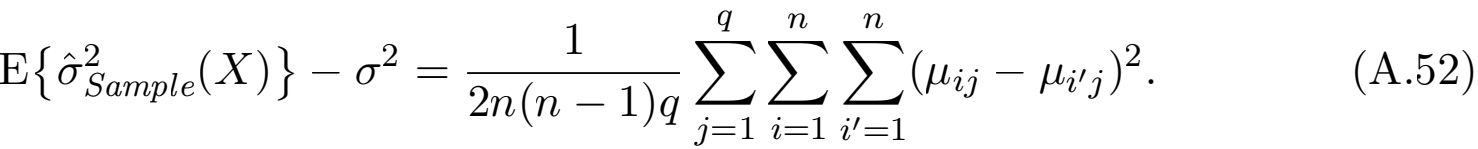
(and therefore, in probability) is max {1/(nq)1/2, s/q }. 

estimator of σ 2 provided that s/q → 0 as q → ∞ .

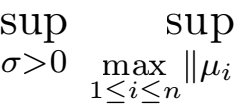
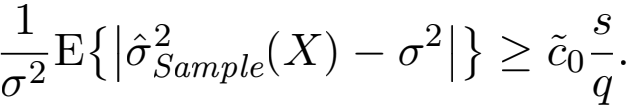
Next, we investigate the property of the sample variance estimator ample.

**Proposition 20** Under model [(1), for](#bookmark19) ample(X) = Σ Σ=1 (Xij - X-j )2 /(nq - q), we

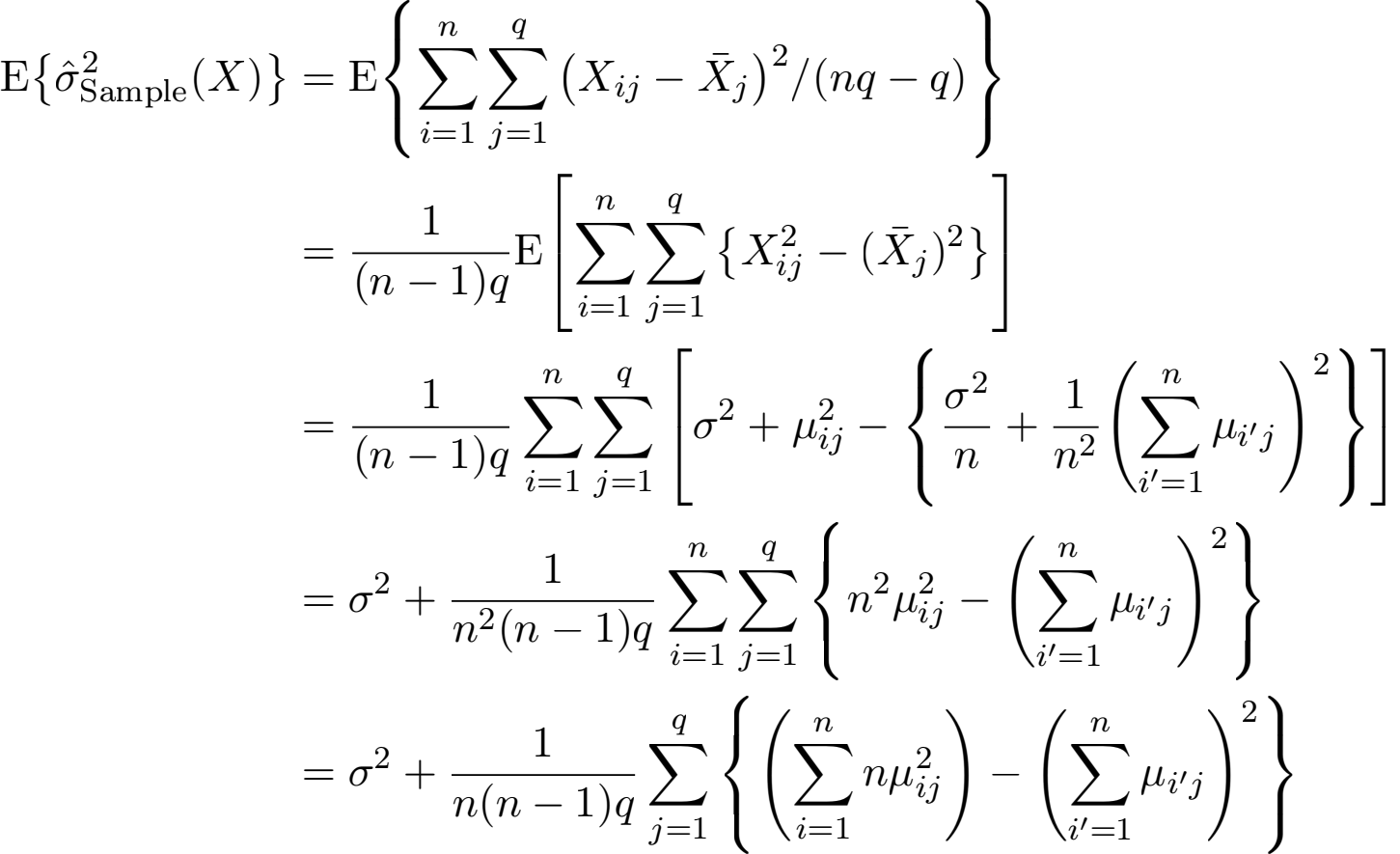
have that

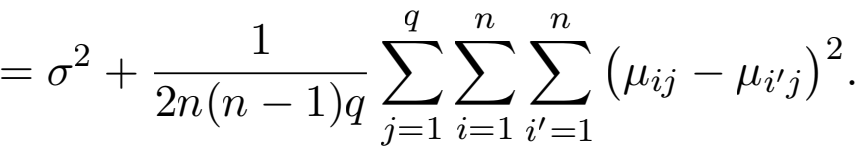


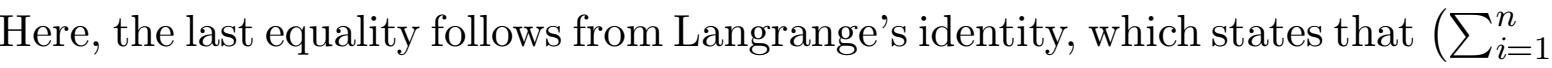
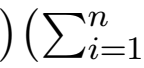
Moreover, for any integers s and q such that ns ≤ 0 ,

Ⅱ0 ≤s  (A.53)

**Proof** [We start with the proof of (A.52)](#bookmark177). [Under (1), the following holds:](#bookmark19)



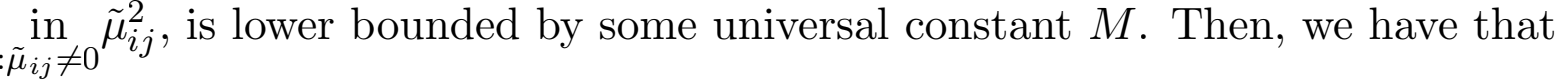


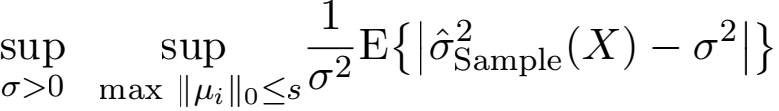
 a b)- (Σ a ibi)2 = 1/2 Σ Σ, =1 (a ibi, - ai, bi)2 .



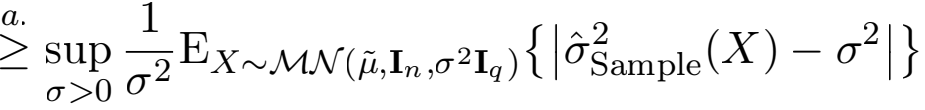
ns ≤ q non-zero entries.  

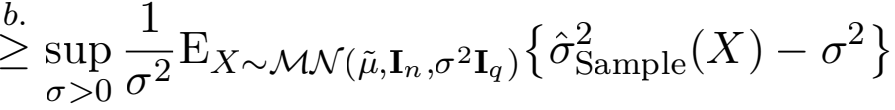
Selective inference for k-means clustering

to be less than q.  , i,jm

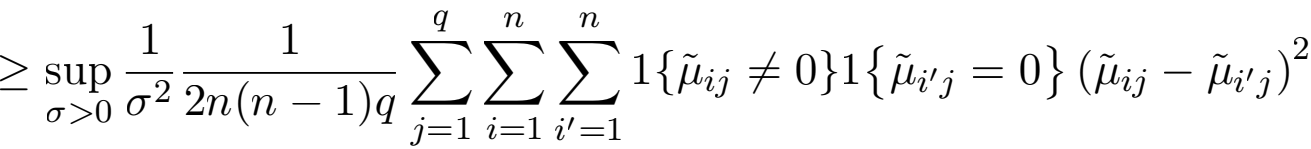


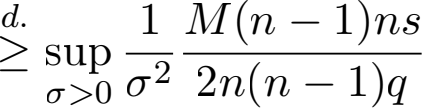
1≤i≤n

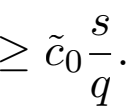












Here, a. -



1≤i≤n

the expression for E{ample(X)} [in (A.52), respectively.](#bookmark177) 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, iI ), i = 1, . . . , n; iI = 1, . . . , n ij  0}1} i~s non-zero.

0 μ ≥ M. 

Contrasting the results in Corollary [19](#bookmark173) and Proposition [20, we note that, under (1), the](#bookmark176)

ED depends critically on the sparsity parameter s (or, equivalently, the l0

norm of μi), whereas the convergence of ample is determined byΣ=1 Σ Σ, =1 (μij - μi,j)2 .

Thus, in scenarios where the underlying means μi, i = 1, . . . , n [are sparse (e.g., (23) in Sec](#bookmark106)- tion [5](#bookmark67)ED (and therefore its \centered" analog ED [in (22)) to be a less](#bookmark102) conservative estimator of σ 2 . As a result, we expect the test based on selective(MED ) to be more powerful than that based on selective(Sample), as shown in Figure [4](#bookmark108) of Section [5.](#bookmark67)

**A.8 Additional power comparisons**

In Section [5.2, we compared the conditional power of the tests based on p selective](#bookmark105) , selective(MED ), and selective([Sample) under (23)](#bookmark106). Here, we conduct two additional analyses.

In the ﬁrst analysis, we consider a diferent notion of power that does not condition on

1 and 2 being true clusters. In this case, comparing the power of the tests requires a bit

of care, because the efect size ⅡμTVⅡ2 may difer across simulated datasets from the same data-generating distribution. As a result, we consider the power of the tests as a function of ⅡμTVⅡ2 . We ﬁt a regression spline using the gam function in the R package mgcv [(Wood,](#bookmark141)

Chen and Witten

[2017) to obtain a smooth estimate of power on the same simulated datasets from Section](#bookmark141)[5.2.](#bookmark105) The results are in Figure [6.](#bookmark178) The power of the tests that reject H0 if p selective , selective(MED ), or selective(Sample) is less than α = 0.05 increases as ⅡµT νⅡ2 increases. For a given value of ⅡµT νⅡ2 and σ, the test based on p selective has the highest power, followed by that based

on selective(MED ); the test based on selective(Sample) has the lowest power.

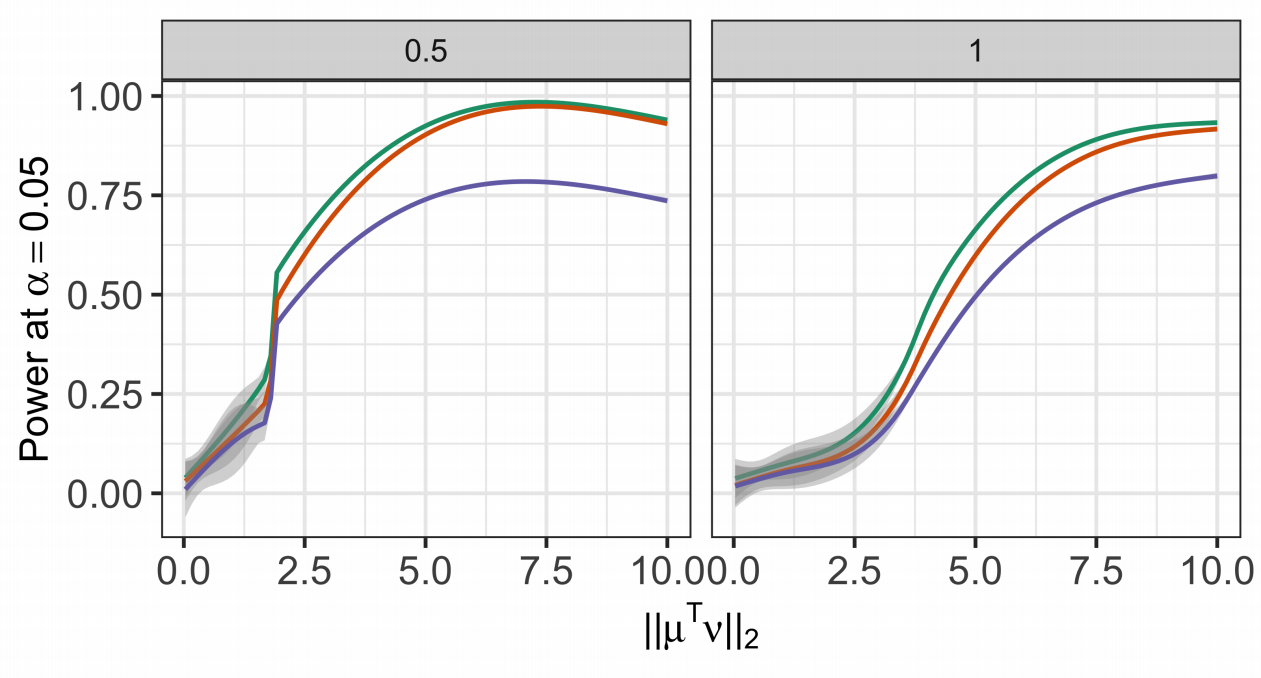
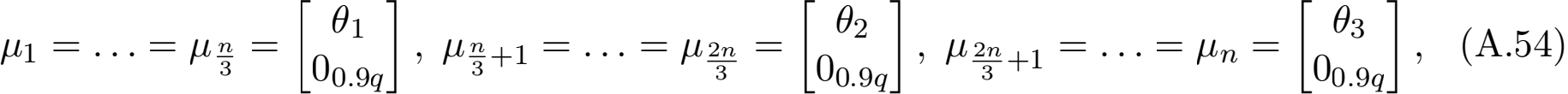


Figure 6: Left : Additional analysis of the data in Section [5.2](#bookmark105) with σ = 0.5. We ﬁt a regression spline to display the power of the tests based on p selective (green line), selective(MED ) (orange line), and selective(Sample) (purple line) as a function of ⅡµT νⅡ2 . Right : Same as left, but for σ = 1.

[In the second analysis, we consider the conditional power (deﬁned in (24)) of the tests](#bookmark107) based on p selective , selective(MED ), and selective(Sample) under a diferent data generating [model than (23)](#bookmark106). [We generate data from (1) with](#bookmark19) n = 150 and



where, q is taken to be a multiple of 10, and for δ > 0, θ ∈ R3×0.1q has orthogonal rows,

with ⅡθiⅡ = δ/2 for i = 1, 2, 3. As in Section [5.2, we can think of](#bookmark105) C1 = {1, . . . , n/3}, C2 =

{(n/3) + 1, . . . , (2n/3)}, C3 = {(2n/3) + 1, . . . , n} as \true clusters". [Under (A.54), the](#bookmark179) pairwise distance between each pair of true clusters is δ .

We generate M = 100, [000 datasets from (A.54) with](#bookmark179) q = 50,σ = 0.25, 0.5, 1, and δ = 2, 3, . . . , 10. For each simulated dataset, we apply k-means clustering with K = 3 and

reject H0 : µT ν = 0q if p selective , selective(MED ), or selective(Sample) is less than α = 0.05.

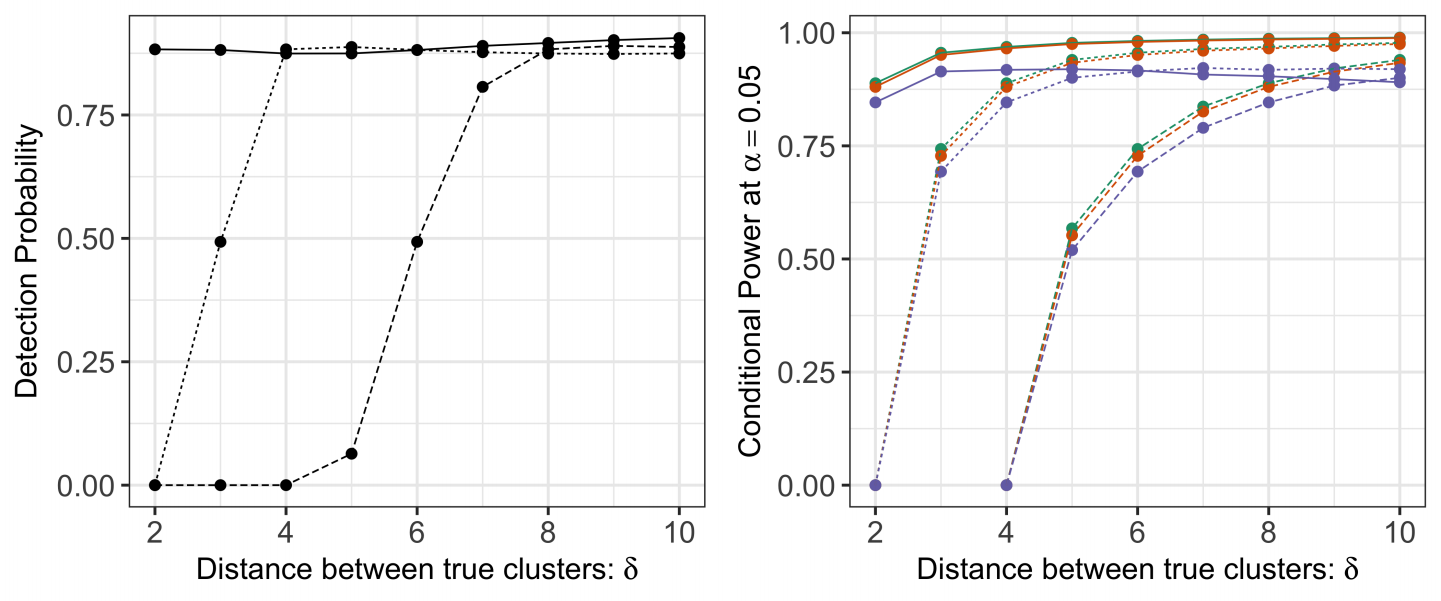
Figure [7](#bookmark180)[(a) displays the detection probability (25) of](#bookmark108) k-means clustering as a function of δ [in (A.54)](#bookmark179). [Under model (1), the detection probability increases as a function of](#bookmark19) [δ in (A.54)](#bookmark179) across all values of σ . For a given value of δ, a larger value of σ leads to lower detection probability. Figure [7(b) displays the conditional power (24) for the tests based on](#bookmark180) pselective ,

selective(MED ), and selective(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

Selective inference for k-means clustering

(b)



( ) (d)

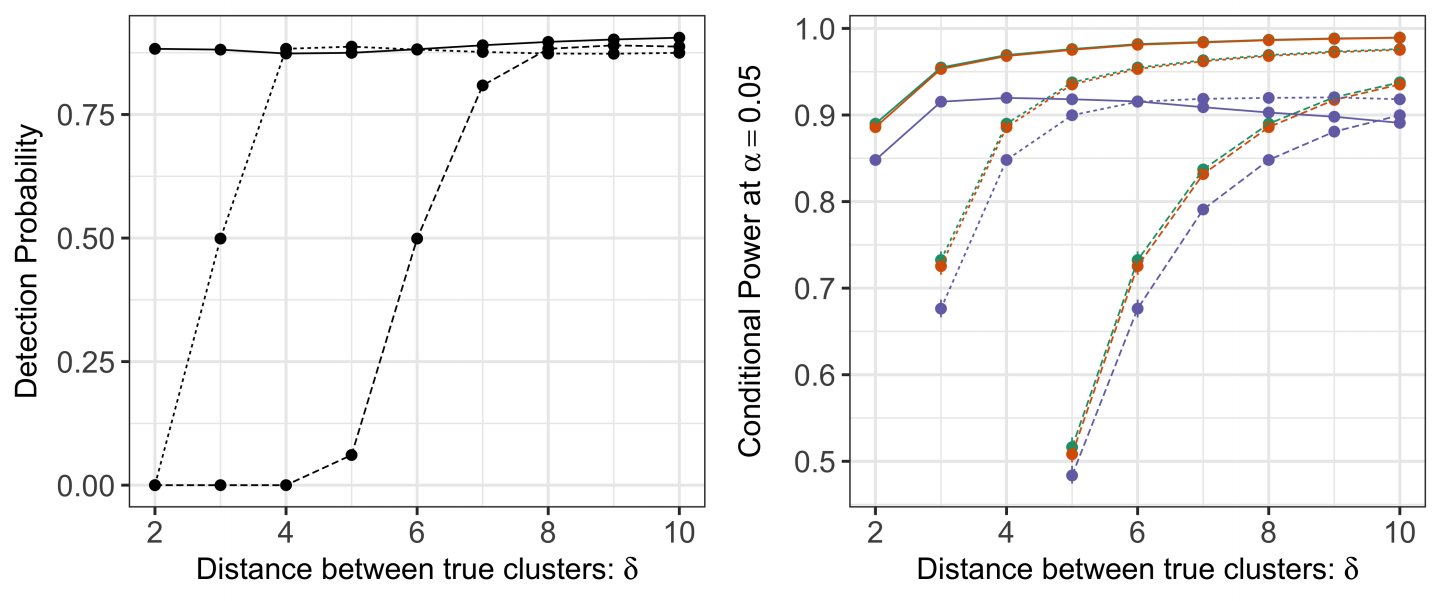


Figure 7: (a): [Detection probability deﬁned in (25) for](#bookmark108) k-means clustering with K = 3 [under model (1) with](#bookmark19) n = 150, q = 50, and µ [in (A.54), across](#bookmark179) δ = Ⅱθi - θjⅡ2 in [(A.54) and](#bookmark179) σ = 0.25 (solid lines), 0.5 (dashed lines), and 1 (long-dashed lines). (b): [The conditional power (24) at](#bookmark107) α = 0.05 for the tests based on p selective (green), selective(MED ) (orange), and selective(Sample) (purple), under model

[(1) with](#bookmark19) n = 150, q = 50, and µ [in (A.54)](#bookmark179). (c): Same as (a), but for µ [in (23)](#bookmark106).

(d): Same as (b), but for µ [in (23)](#bookmark106).

Chen and Witten

the same values of δ and σ, the test based on p selective has the highest conditional power, followed closely by the test based of selective(MED ). Using selective(Sample) leads to a less powerful test, especially for larger values of δ . As a comparison, we included the detection

[probability and conditional power under model (23) with](#bookmark106) q = 50 in panels (c) and (d) of Figure [7.](#bookmark180) The tests under consideration behave qualitatively similarly as a function of δ and σ . [Under (23), we observe an even larger gap between the power of the test based on](#bookmark106) selective(Sample) and the power of the test based on selective(MED ).

**A.9 Additional results for real data applications**

In this section, we visualize the estimated clusters for the single cell RNA-sequencing data in Section [6.2.](#bookmark114)

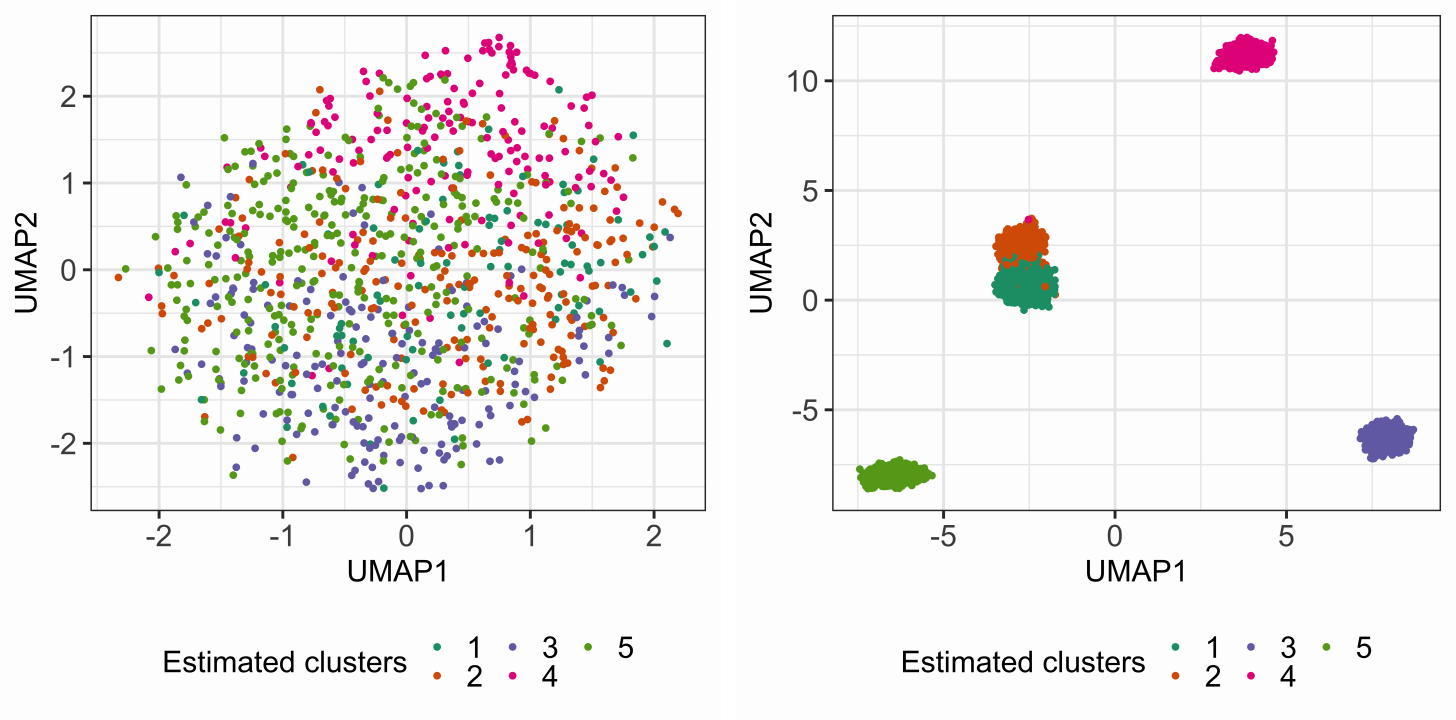


Figure 8: Left: [The two-dimensional UMAP embedding (McInnes et al.,](#bookmark138) [2018) of the “no](#bookmark138) cluster” dataset after preprocessing (as described in Section [6.2), colored by the](#bookmark114)

estimated cluster membership via k-means clustering. Right: Same as left, but

for the “cluster” dataset.