Convex clustering via ℓ_1 fusion penalization

PETER RADCHENKO AND GOURAB MUKHERJEE [*]

Abstract

We study the large sample behavior of a convex clustering framework, which minimizes the sample within cluster sum of squares under an ℓ_1 fusion constraint on the cluster centroids. This recently proposed approach has been gaining in popularity, however, its asymptotic properties have remained mostly unknown. Our analysis is based on a novel representation of the sample clustering procedure as a sequence of cluster splits determined by a sequence of maximization problems. We use this representation to provide a simple and intuitive formulation for the population clustering procedure. We then demonstrate that the sample procedure consistently estimates its population analog, and derive the corresponding rates of convergence. The proof conducts a careful simultaneous analysis of a collection of M-estimation problems, whose cardinality grows together with the sample size. Based on the new perspectives gained from the asymptotic investigation, we propose a key post-processing modification of the original clustering framework. We show, both theoretically and empirically, that the resulting approach can be successfully used to estimate the number of clusters in the population. Using simulated data, we compare the proposed method with existing number of clusters and modality assessment approaches, and obtain encouraging results. We also demonstrate the applicability of our clustering method for the detection of cellular subpopulations in a single-cell virology study.

Some key words: Convex Clustering; Fusion Penalties; Number of Clusters; Rates of Convergence

1 Introduction

Clustering is one of the most popular statistical techniques for unsupervised classification and taxonomy detection (Hartigan, 1975; Kaufman and Rousseeuw, 2009). One serious

[*] University of Southern California

limitation of the traditional methods, such as k-means, is the non-convexity of the corresponding optimization problems. Recently, several convex clustering algorithms have been proposed (Xu *et al.*, 2004; Bach and Harchaoui, 2008; Chi and Lange, 2013). Speed and scalability of these algorithms make them increasingly popular for cluster analysis of massive modern datasets. These approaches use convex relaxations of the traditional non-convex clustering criteria, however, they do not naturally inherit the statistical properties associated with the original methods. Here we study the large sample behavior of a popular convex clustering framework that is based on an ℓ_1 fusion penalty (Hocking *et al.*, 2011).

Consider the problem of clustering n observations, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, which are sampled from a Euclidean space, $\mathbb{R}^d$. The well-studied k-means approach (MacQueen *et al.*, 1967; Hartigan, 1978; Pollard, 1981, 1982; Jain, 2010) is based on minimizing the within cluster sum of squares, $\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\alpha}_i\|_2^2$, with respect to the cluster centroids, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$, under the restriction that the number of distinct cluster centroids is at most k. This restriction can be viewed as an ℓ_0 constraint on the centroids. Motivated by the Lasso and its variants (Tibshirani, 1996; Tibshirani *et al.*, 2005), which successfully use the ℓ_1 constraint as a surrogate for the NP-hard ℓ_0 constraint, Hocking *et al.* (2011) consider the following modification of the k-means clustering criterion:

$$\min_{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\alpha}_i\|_2^2 \text{ subject to } \sum_{1\le i<j\le n} \big\|\boldsymbol{\alpha_i} - \boldsymbol{\alpha_j}\big\|_1 \le t. \tag{1}$$

When $t = 0$, the ℓ_1 penalty fuses all the cluster centroids together. Thus, all the observations are placed in the same cluster. When $t \ge \sum_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1$, we have $\boldsymbol{\alpha}_i = \boldsymbol{x}_i$ for all i, and, thus, each observation forms its own cluster. Varying t between the two extremes creates a path of solutions to the regularized clustering problem. Note that the Lagrangian form of the above criterion, $\min_{\boldsymbol{\alpha}_i} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i<j} \big\|\boldsymbol{\alpha_i} - \boldsymbol{\alpha_j}\big\|_1$, is separable across dimensions. Consequently, the corresponding optimization problem reduces to independently minimizing d univariate convex clustering criteria.

Thus, to understand the large sample behaviour of the multivariate solution, it is sufficient to focus on the analysis of the univariate clustering criterion,

$$\min_{\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{n} (x_i - \alpha_i)^2 + \lambda \sum_{1\le i<j\le n} \big|\alpha_i - \alpha_j\big|. \tag{2}$$

As the penalty parameter λ varies from 0 to ∞, each corresponding solution determines a cluster partition. We are interested in the asymptotics of the entire collection of such partitions, which we view as the outcome of the *sample clustering procedure*.

Summary of the Main Contributions. We analyze the large sample behavior of the sample clustering procedure determined by the solution path for criterion (2). We develop a simple and intuitive formulation for the population clustering procedure, show that under some very mild regularity conditions the sample procedure consistently estimates its population analog, and derive the corresponding rates of convergence.

More specifically, we first demonstrate that the path of solutions to (2) determines a clustering tree, which can be formed by either successive merges of clusters, in a bottom up fashion, or successive splits, in a top down approach. We then study the asymptotic behavior of the full clustering tree by representing each split as a solution to a maximization problem. We define the corresponding population clustering procedure in a similar fashion, but replace sample averages with the corresponding expected values. The asymptotic analysis is significantly complicated by the fact that, unlike in the standard M-estimation setup (e.g. van der Vaart and Wellner 1996; van der Vaart 1998), the number of maximization problems at the sample level tends to infinity together with n, and the number of the corresponding population problems is infinite. We establish consistency and the rates of convergence of the sample clustering procedure through a careful analysis of the population procedure and the corresponding empirical process.

Motivated by the results of our large sample investigation, we introduce a key postprocessing modification to the sample clustering procedure. We show, both theoretically and empirically, that the resulting approach can be successfully used to estimate the number of clusters in the population. We also compare the new methodology with a wide variety of existing modality assessment and number of clusters approaches. Our results provide strong support for the use of fusion penalization in clustering.

Connections to Related Work. Hocking *et al.* (2011), Chi and Lange (2013) and Tan and Witten (2015) have studied modifications of optimization problem (1). These include using ℓ_2 or ℓ_∞ regularization, as well as incorporating weights (Pelckmans *et al.*, 2005; Lindsten *et al.*, 2011; Zhu *et al.*, 2014). The large sample analysis in the papers listed above focusses on showing that if the distance between clusters grows at a sufficiently fast rate, then the corresponding method can separate the groups perfectly. Here we consider a completely different perspective and investigate the asymptotics of a clustering approach in the classical sense of Pollard (1981). We study a clustering procedure that is applied to a random sample, and analyze its convergence to the outcome of the corresponding population procedure, which is based on the underlying probability distribution. As we point out in Section 5, the general framework of our theoretical analysis has the potential to handle the aforementioned modifications of the optimization problem.

The criterion in (1) can be viewed (Hocking *et al.*, 2011) as a convex relaxation of the hierarchical clustering criterion (Hartigan, 1975). However, as clustering is a very mature subject, approaches built on several other philosophies are also widely used in practice. A detailed review of clustering methods can be found in Kaufman and Rousseeuw (2009). One of the most popular methods is the k-means algorithm (MacQueen *et al.*, 1967), which follows a partitioning approach for making clusters. Other popular partitioning methods, such as PAM (Kaufman and Rousseeuw, 1990) and CLARA (Kaufman and Rousseeuw, 1986), are based on the k-medoids algorithm. Density driven approaches, which include mixture model based methods, such as Fraley and Raftery (2002) and Li (2005), as well as non-parametric methods (see Li *et al.* 2007 and the references therein), provide a flexible clustering framework, while spectral clustering methods, such as (Belkin and Niyogi, 2001; Rohe *et al.*, 2011; Shi *et al.*, 2009) perform efficient dimension reduction before segmenting the data. In our empirical analysis we compare the performance of the proposed approach with most of the aforementioned clustering methods.

The ℓ_1 penalty, which is extensively used for variable selection (Tibshirani, 2011), also finds its use in trend filtering (Tibshirani, 2013) and high-dimensional clustering problems (Soltanolkotabi and Candés, 2012; Witten and Tibshirani, 2010). Another related approach, the fused Lasso (Rinaldo *et al.*, 2009; Tibshirani and Walther, 2005; Hoefling, 2010), deals with applications having ordered features and checks for local constancy of their associated coefficients. This approach penalizes the successive differences of the coefficients. Shen and Huang (2010); Shen *et al.* (2012); Ke *et al.* (2013); Bondell and Reich (2008) have proposed methods based on fusion penalties, which apply to all the pairwise differences of coefficients. These approaches can successfully recover the grouping structure of predictors in a high-dimensional regression setup. However, the theory developed for these methods focusses on the homogeneity of regression coefficients and cannot be applied in the unsupervised clustering setup considered in this paper.

Organization of the Paper. In Section 2 we derive two equivalent algorithmic representations of the sample clustering procedure, which we use to formulate the corresponding population procedure. Section 3 contains our main results, in which we establish consistency and the rates of convergence. Our asymptotic analysis reveals that an overwhelming majority of the sample clusters are in some sense negligible. Motivated by this observation, we introduce a key post-processing modification to the clustering procedure. In Section 4 we conduct a detailed empirical analysis of our approach. More specifically, we use simulated data to show its strong performance relative to popular existing approaches for assessing modality and estimating the number of clusters. We also illustrate the use of our method in analysis of single-cell virology datasets. All the proofs, together with additional technical details, are relegated to the Supplementary Material.

2 Sample and Population Clustering Procedures

In this section we derive two equivalent representations of the sample clustering procedure. First, we develop a computationally efficient merging algorithm for producing a path of solutions to the clustering criterion (2). Then, in order to understand the large sample behavior of the solution path, we introduce an equivalent splitting procedure, which can recover all the corresponding cluster splits by solving a sequence of maximization problems. We use the splitting representation to define the population clustering procedure, and describe its basic properties.

2.1 Equivalent Representations for the Sample Solution Path

Note that a solution path for problem (2) could be produced using the highly general fused lasso algorithm in Hoefling (2010). Instead, we obtain a very simple and computationally efficient fitting procedure by analyzing our clustering criterion, (2), directly. The path algorithm we describe here is a bottom up procedure, which starts at $\lambda = 0$, with each observation forming its own cluster, and then gradually merges suitable clusters as λ increases. Fix λ, and suppose that C is one of the clusters identified by the solution to the optimization problem (2). Write α_C for the centroid of cluster C, and denote the corresponding cluster average by $\overline{X}_C$. As pointed out in Hocking *et al.* (2011), the first order conditions for criterion (2) imply

$$\alpha_C = \overline{X}_C + \lambda \sum_{j,\alpha_j \neq \alpha_C} \text{sign}(\alpha_j - \alpha_C). \tag{3}$$

Until the cluster partition or the ordering or the centroids are modified, parameter λ is the only component on the right-hand side of the equation that can change. Thus, equation (3) provides a simple way of tracking the piecewise linear paths of the centroids α_i. Another consequence of the first order conditions is that as λ increases, the only way the clusters get modified is some of them get merged together (Hocking *et al.*, 2011). Hence, we can store the full cluster partition path by keeping track of the merges and the corresponding values of the tuning parameter λ. Algorithm 1 makes this idea precise, and Theorem 1 provides a rigorous justification. Here we use $|\cdot|$ to denote the cardinality of a set.

```
INITIALIZE:
    Sort data in ascending order and store them as $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$.
    Set $K$, the number of clusters, equal to $n$. For each $i$ in $1, ..., n$, set $C_i = \{x_i\}$.
REPEAT:
    Find the adjacent centroid distances standardized by cluster sizes:
        $d(j, j+1) \leftarrow (\overline{X}_{C_{j+1}} - \overline{X}_{C_j}) / (|C_j| + |C_{j+1}|)$.
    Find the clusters that minimize this distance: $j^* \leftarrow \arg\min_j d(j, j+1)$.
    Merge the clusters that were found: $C_{j^*} \leftarrow C_{j^*} \cup C_{j^*+1}$.
    Store the above merge and the corresponding $\lambda$ value: $\lambda = d(j^*, j^*+1)$.
    Relabel the remaining clusters: for $j > j^*$ set $C_j \leftarrow C_{j+1}$.
    Reduce the total number of clusters: $K \leftarrow K - 1$.
UNTIL $K = 1$.
OUTPUT: Sequence of cluster merges and corresponding $\lambda$ values.
```

Algorithm 1: Merging Algorithm

The following result shows that Algorithm 1 reproduces the sequence of cluster partitions and the corresponding λ values from the optimization problem (2). In the proof, which is provided in the Supplementary Material, we also verify that the sequence of λ values, corresponding to successive merges in Algorithm 1, is increasing.

Proposition 1. *Suppose that the observations are generated from a continuous distribution. Then, with probability one, the sequence of merges and λ values produced by the merging algorithm is the same as the sequence corresponding to the optimization criterion* (2).

For the asymptotic analysis, it is helpful to recover the sequence of cluster partitions in a top down approach: we start with everything in one cluster and then split the clusters iteratively. We call a representation of the cluster C as $C = C_1 \cup C_2$ a split if $\max C_1 < \min C_2$. The full collection of splits corresponding to the optimization problem (2) is given by the splitting procedure, described in Algorithm 2 below. Proposition 2, proved in the Supplementary Material, provides theoretical justification. In particular, it shows that each of the cluster splits is chosen to maximize the distance between the two sub-cluster means.

```
INITIALIZE:
    Sort data in ascending order and store them as $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$.
    Set the current partition of $\boldsymbol{x}_n$ to $\boldsymbol{x}_n$.
REPEAT:
    Select one cluster, $C$, with $|C| > 1$, from a current cluster partition of $\boldsymbol{x}_n$.
    Find a split partition $C = C_1 \cup C_2$, that maximizes the distance $\overline{X}_{C_2} - \overline{X}_{C_1}$.
    Store the split $C = C_1 \cup C_2$ and the corresponding value $\lambda = (\overline{X}_{C_2} - \overline{X}_{C_1})/|C|$.
    Replace $C$ with $C_1 \cup C_2$ in the current partition of $\boldsymbol{x}_n$.
UNTIL: All the clusters in the current partition of $\boldsymbol{x}_n$ are of size one.
OUTPUT: Collection of cluster splits and corresponding $\lambda$ values.
```

Algorithm 2: Splitting Procedure

Proposition 2. *Suppose that the observations are generated from a continuous distribution. Then, with probability one, the collection of splits and corresponding λ values produced by the splitting procedure in Algorithm 2 exactly matches the sequence of merges and corresponding λ values produced by the merging algorithm .*

Note that, unlike the merging algorithm, the splitting procedure does not provide a computationally efficient way for producing the clustering tree. Instead, we use the splitting procedure to understand the large sample behavior of the sample clustering procedure. It is reasonable to expect that, as n tends to infinity, the collection of splits in the sample procedure should resemble the collection of splits in an analogous procedure defined on the population. The population procedure can be defined by replacing the averages with the corresponding conditional means. The formal definition is given in the next section.

2.2 Population Clustering Procedure

For the remainder of the paper we assume that the underlying distribution has a finite first moment and a real valued density, f. For concreteness, we focus on the case where the support of the distribution is of the form (L_0, R_0), where $-\infty \leq L_0 < R_0 \leq \infty$. Thus, every open interval in (L_0, R_0) contains positive probability. Given an interval $(l, r) \subseteq (L_0, R_0)$, we write $\mu_{l,r}$ for the population conditional mean on (l, r),

$$\mu_{l,r} = \left(\int_{l}^{r} f(x)dx \right)^{-1} \int_{l}^{r} xf(x)dx. \tag{4}$$

We set $\mu_{r,r} = r$, by continuity. Given an interval $(L, R) \subseteq (L_0, R_0)$, we define

$$G_{L,R}(a) = \mu_{a,R} - \mu_{L,a}, \tag{5}$$

for $a \in [L, R]$. Note that $G_{L,R}(L) = \mu_{L,R} - L$ and $G_{L,R}(R) = R - \mu_{L,R}$.

According to the results in Section 2.1, the sample clustering procedure determines the split partition of a cluster by maximizing the distance between the empirical sub-cluster means. We define the *population clustering procedure* by analogy. Given a cluster (L, R), the population procedure chooses the split that maximizes the distance between the population sub-cluster means. In other words, it finds a point s that maximizes $G_{L,R}$, then partitions (L, R) into subintervals (L, s) and (s, R), on which the procedure is repeated. If s is an interior point of (L, R), we call it a *split point*, and we call the corresponding partition a *split*. Otherwise, the population procedure essentially wants to split off an endpoint, which forces the cluster to be truncated rather than split. More formally, given a cluster (L, R), we distinguish three types of *truncation*, as specified below.

Definition 1. *(i) if* $\arg\max G_{l,R} = \{l\}$ *for all* $l \in [L, L^*)$, *and* $\arg\max G_{L^*,R} \neq \{L^*\}$, *then the interval* (L, R) *is truncated from the* ***left*** *to* (L^*, R^*), *where* $R^* = R$;

(ii) if $\arg\max G_{L,r} = \{r\}$ *for all* $r \in (R^*, R]$, *and* $\arg\max G_{L,R^*} \neq \{R^*\}$, *then* (L, R) *is truncated from the* ***right*** *to* (L^*, R^*), *where* $L^* = L$;

(iii) if there exists a continuous decreasing function $l \mapsto R_l$, satisfying $R_L = R$, for which $\arg\max G_{l,R_l} = \{l, R_l\}$ for all $l \in [L, L^*)$, and $\arg\max G_{L^*,R_{L^*}} \neq \{L^*, R_{L^*}\}$, then (L, R) is truncated, in a ***two-sided*** *fashion, to (L^*, R^*), where $R^* = R_{L^*}$.*

Note that we incorporated a continuity requirement into the definition of a two-sided truncation. In the next subsection we give regularity conditions under which this requirement is satisfied. We are now ready to formulate the full population clustering procedure.

```
INITIALIZE:
    Set the current cluster collection, $\Sigma$, equal to $\{(L_0, R_0)\}$.
REPEAT:
    Select one non-empty cluster, $(L, R)$, from the current cluster collection, $\Sigma$.
    – If the maximum value of $G_{L,R}$ is achieved at a point $s$ in $(L, R)$, then store $s$
      as a split point and replace $(L, R)$ in $\Sigma$ with $(L, s)$ and $(s, R)$.
    – Otherwise, replace $(L, R)$ in $\Sigma$ with the interval $(L^*, R^*)$ from Definition 1.
UNTIL: The current cluster collection, $\Sigma$, consists only of empty clusters.
OUTPUT: Set of split points.
```

Algorithm 3: Population Clustering Procedure

The collection of population split points determines the corresponding clusters. For example, consider the symmetric mixture of two Gaussian distributions examined in Figure 1. The population procedure identifies one split point, located at zero. This specifies the population cluster partition: $(-\infty, 0) \cup (0, \infty)$.

Given an underlying distribution, Algorithm 3 determines the exact behaviour of the population clustering procedure. In Section 4 of the Supplementary Material we document the performance of the population procedure for a variety of Gaussian mixtures. In Section 2.3 we establish an important fact that the population procedure produces no splits for unimodal distributions. We also provide conditions under which the population procedure is *well defined*, by which we mean that it implements finitely many uniquely defined steps.

2.3 Properties of the Population Procedure

The proofs of the results established in this section are provided in the Supplementary Material. We first consider an important special case, where the underlying distribution

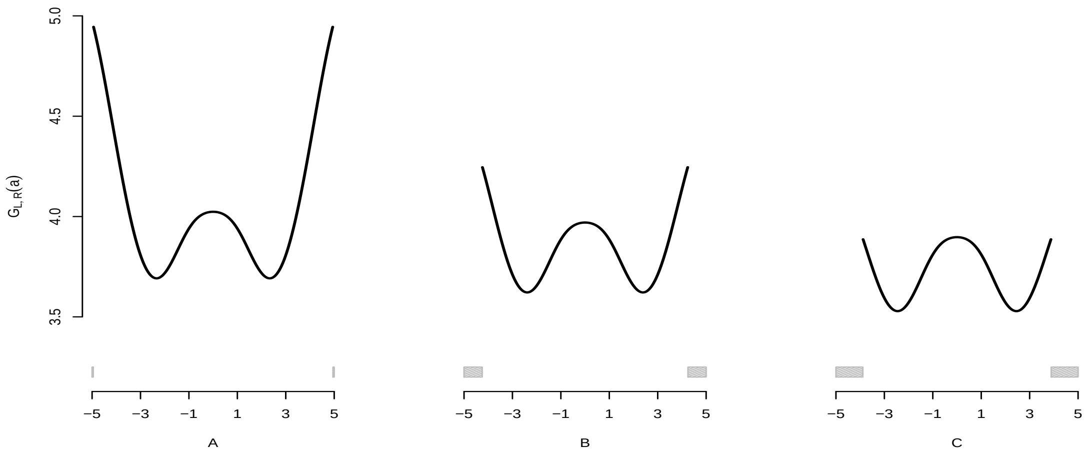

Figure 1: Population criterion function, $G_{L,R}$, corresponding to the Gaussian mixture $0.5\,N(-2,1) + 0.5\,N(2,1)$, is plotted for three choices of (L,R), such that $R = -L$ and $G_{L,R}(L) = G_{L,R}(R) = \max G_{L,R}$. The population clustering procedure continuously truncates the support of the distribution, symmetrically in a two-sided fashion, until $\max G_{L,R}$ can be achieved at an interior point (plot C). Then, the procedure places a split point at zero. Note that the resulting sub-clusters are then truncated down to empty sets according to Proposition 3 in Section 2.3.

is unimodal. We suppose that the density f is either strictly monotone on its support, (L_0, R_0), or there exists a point c for which f is strictly increasing on (L_0, c) and strictly decreasing on (c, R_0). The following result shows that in this setting, under just a continuity assumption on f, the population procedure is unique and does not reveal any clusters.

Proposition 3. *If f is continuous and unimodal, then the population clustering procedure is uniquely defined and produces no splits.*

We now move to the general setting. The following simple regularity condition ensures existence of a population clustering procedure with finitely many steps, as we demonstrate in the proof of Proposition 4 below.

C1. Density f is nonzero and differentiable on (L_0, R_0). It has finitely many modes and, at each of its interior modes, admits a non-constant Taylor approximation.

> **Remark.** The last requirement means the following: for each interior mode c, there exists a positive integer k, such that f is k times differentiable at c with $f^{(k)}(c) \neq 0$.

Note that the differentiability assumption can be slightly relaxed: for example, the results that follow hold for continuous piece-wise linear densities with no constant segments. However, we prefer to keep this assumption, as it simplifies the presentation of the results.

To address the question of uniqueness, consider the following counterexample. Suppose the underlying distribution is uniform on (L, R). Then, the criterion function $G_{L,R}$ is constant on its domain, and the population procedure applied to cluster (L, R) may place a split point anywhere in its interior. Thus, there are infinitely many versions of the population clustering procedure. The following regularity condition explicitly rules out such settings, by requiring that the interior of (L, R) contains at most one maximizer of $G_{L,R}$.

C2. When the population procedure performs a split, the location of the split point is uniquely determined.

Note that condition C2 holds for each distribution with a continuous unimodal density, as a direct consequence of Proposition 3. In the proof of Proposition 3 we also show that C2 holds for all bimodal densities, provided the smoothness condition, C1, is satisfied. The following result establishes existence and uniqueness of the population clustering procedure in the general setting.

Proposition 4. *If regularity conditions C1 and C2 are satisfied, then the population clustering procedure is uniquely defined and implements finitely many steps.*

In Section 3 we show that under the same regularity conditions, C1 and C2, the sample procedure consistently estimates its population counterpart.

3 Main Results

In this section we show that the clustering tree produced by criterion (2) consistently estimates the clustering tree produced by the population procedure defined in Section 2.2. We also derive the corresponding rates of convergence and propose a novel post-processing modification of the sample clustering procedure. Recall that we assume a finite first moment for the underlying distribution.

3.1 Consistency

We start with some useful notation. In both the sample and the population, each split is characterized by a triple (L, s, R), where interval (L, R) is the cluster being split, and s is the split point, located inside (L, R). We write $P_{L,R}$ for the probability assigned to the interval (L, R) by the underlying distribution. For a split $sp = (L, s, R)$ we define its size as $size(sp) = \min\{P_{L,s}, P_{s,R}\}$. When the probabilities in the above definition are replaced with the corresponding sample frequencies, we write $\widehat{size}(sp)$ for the resulting quantity, and refer to it as the empirical size.

The set of all the population splits, denoted by $\mathcal{S}$, defines the population clustering tree. Similarly, the set of all sample splits, $\widehat{\mathcal{S}}$, defines the sample tree. The cardinality of $\widehat{\mathcal{S}}$ tends to infinity as the sample size grows. Alternatively, according to Proposition 4, under mild regularity conditions the population procedure produces finitely many splits, together with some truncations. To establish consistency, we divide the sample splits into "big" and "small", based on their empirical size, then show that the first group converges to the population splits, while the second is asymptotically negligible. The formal definition is given below. We write d_H for the Hausdorff distance between subsets of a Euclidean space.

Definition 2. *Write $\widehat{\mathcal{S}}_\alpha$ for the set $\{sp \in \widehat{\mathcal{S}} : \widehat{size}(sp) > \alpha\}$ and let α^* be the smallest split size in the population procedure. We call the sample clustering procedure* ***strongly consistent*** *if, for each α in $(0, \alpha^*)$, the following statements hold almost surely,*

$$|\widehat{\mathcal{S}}_\alpha| \to |\mathcal{S}| \tag{6}$$

$$d_H(\widehat{\mathcal{S}}_\alpha, \mathcal{S}) \to 0 \qquad \textit{and} \tag{7}$$

$$\max\{size(sp) : sp \in \widehat{\mathcal{S}} \setminus \widehat{\mathcal{S}}_\alpha\} \to 0. \tag{8}$$

If we replace almost sure convergence with convergence in probability, we have a weaker notion of consistency, which holds automatically when the sample procedure is strongly consistent. In particular, displays (6) and (7) imply that, except on a set of probability tending to zero, there is a one to one correspondence between $\widehat{\mathcal{S}}_\alpha$ and the set of all population splits, such that each split in $\widehat{\mathcal{S}}_\alpha$ converges to its population counterpart with respect to the usual Euclidean distance. The next result, which is proved in the Supplementary Material, establishes consistency of the sample procedure.

Theorem 1. *Suppose that regularity conditions C1 and C2, given in Section 2.3, are satisfied. Then, the sample clustering procedure is strongly consistent.*

> **Remark.** It follows from the proof that the result continues to hold if we replace $\widehat{size}$ with $size$ in the definition of $\widehat{\mathcal{S}}_\alpha$ and/or replace $size$ with $\widehat{size}$ in display (8).

If condition C2 is violated, and the locations of the population splits are not uniquely determined, a modification of Theorem 1 continues to hold. More specifically, suppose that the number of versions of the population procedure is finite. Then, the sample procedure converges to the *set* of population versions, rather than a specific one. In other words, the number and the locations of the big sample splits approach the corresponding quantities for an appropriately chosen population version, where the choice depends on the sample.

Consider the important case of a unimodal underlying distribution. Proposition 3 in Section 2.3 and the proof of Theorem 1 imply that in this case condition C1 is not required for consistency. Note also that the population procedure produces no splits, by Proposition 3. It follows that all of the sample splits are uniformly asymptotically negligible.

Corollary 1. *If f is continuous and unimodal, then the maximum size of all the splits in the sample clustering procedure goes to zero almost surely.*

In the next section we extend the results in Theorem 1 by establishing the rates of convergence for the sample clustering procedure.

3.2 Rates of Convergence

To establish the rates of convergence for the sample splitting procedure, we need an additional regularity condition. We use the term *population cluster* to refer to all intervals that appear along the path of the population procedure.

C3. For each population cluster (L, R) and each $t \in \arg\max_{[L,R]} G_{L,R}$, if $t \in (L, R)$, then $G''(t) \neq 0$, otherwise $G'_{L,R}(t) \neq 0$.

The requirement on $G_{L,R}(t)''$, imposed for each population split (L, t, R), is the standard M-estimation assumption that requires the second derivative of the population criterion function to be nonsingular at the population maximum (e.g. van der Vaart and Wellner 1996; van der Vaart 1998). The requirement on $G'_{L,R}$ is the analog of the aforementioned M-estimation assumption in the case where the population criterion function, $G_{L,R}$, is maximized at an endpoint of $[L, R]$, rather than in the interior. In this case, the behaviour of $G_{L,R}$ near its maximum is characterized by the first derivative, rather than by the second.

Let $\widehat{\mathcal{S}}(\tau)$ contain all the sample splits $sp = (L, \widehat{t}, R)$, for which the sample frequencies of (L, R), $(-\infty, \widehat{t})$ and $(\widehat{t}, \infty)$ are greater than or equal to τ. In Theorem 2 we restrict our attention to the sample splits in $\widehat{\mathcal{S}}(\tau)$, for arbitrarily small but positive τ. Without this restriction, the rate of convergence in (10) would change. In particular, split sizes larger than $O_p(\log n/n)$ are produced when the sample procedure is applied to intervals whose widths tend to zero. Also, larger split sizes may appear near the boundary of the support of the distribution. The general approach used in the proof of Theorem 2 would also establish these slower rates of convergence. However, the exact form of the new rates depends on the behavior of the density near the boundary of the support and on the aforementioned intervals of negligible width. Instead, we present a clean result, with just one rate of convergence for all the small sample splits, while only imposing some simple regularity conditions.

Theorem 2. *Suppose that regularity conditions C1-C3 are satisfied. Let α^* be the smallest split size in the population procedure. Then, for each α in $(0, \alpha^*)$,*

$$d_H(\widehat{\mathcal{S}}_\alpha, \mathcal{S}) = O_p\left(n^{-1/3}\right) \qquad \textit{and} \tag{9}$$

$$\max\{size(\boldsymbol{sp}) :\ \boldsymbol{sp} \in (\widehat{\mathcal{S}} \setminus \widehat{\mathcal{S}}_\alpha) \cap \widehat{\mathcal{S}}(\tau)\} = O_p\left(\log n/n\right). \tag{10}$$

> **Remark.** As we point out in the proof, if the domain, (L_0, R_0), of the underlying distribution is bounded, then we can remove the lower bounds on the sample frequencies of $(-\infty, \widehat{t})$ and $(\widehat{t}, \infty)$ from the definition of $\widehat{\mathcal{S}}(\tau)$ by assuming, instead, that $f(L_0)$ and $f(R_0)$ are nonzero. It also follows from the proof that the result continues to hold if we replace $size$ with $\widehat{size}$ in display (10).

The proof of Theorem 2 is provided in the Supplementary Material. The intuition for the presented rates of convergence can be described, informally, as follows. We bound the distance between the sample split point, $\widehat{t}$, and its population counterpart, t, by characterizing the behaviour of the sample criterion function near t. The sample criterion, $\widehat{G}_{L,R}(a)$, is the empirical analog of the population criterion, $G_{L,R}(a)$, and is defined as the difference between the averages of the observations in $[a, R]$ and $[L, a]$, respectively. We examine the decrease in $G_{L,R}(t)$ that occurs when t is perturbed by a small amount, δ. Then, we contrast this decrease with the stochastic term given by the difference between the corresponding deviations in $\widehat{G}_{L,R}$ and $G_{L,R}$. The order of this term is roughly $\sqrt{\delta/n}$. When t is a population split point and, thus, lies in the interior of (L, R), the corresponding decrease in $G_{L,R}$ is quadratic in δ. Balancing out the two terms yields the cube root asymptotic behaviour (cf. Kim and Pollard 1990) for the sample split point. When function $G_{L,R}$ is maximized at an endpoint of the interval $[L, R]$, as in the case of truncation, the decrease in $G_{L,R}$ is linear in δ. Balancing this decrease with the stochastic term of order $\sqrt{\delta/n}$ suggests that the sample split point is a $O_p(n^{-1})$ amount away from the boundary of (L, R). Uniformity of the rate over all the small sample splits requires an additional $\log n$ factor.

In the next section we take advantage of our asymptotic results to propose a key modification to the sample clustering procedure.

3.3 Big Merge Tracker: Post-processing the Sample Procedure

Theorem 1 demonstrates that the sample analog of the truncation operation is peeling a large number of tiny clusters off the ends of a large cluster. It follows that when recording sample splits we should distinguish between those that correspond to splits in the population procedure and those that correspond to truncations. Based on this observation, we propose to post-process the sample clustering procedure by only keeping the splits with significant empirical sizes. More specifically, given a threshold α, if the cardinality of one of the sub-clusters is below αn, the corresponding split is removed from the final output.

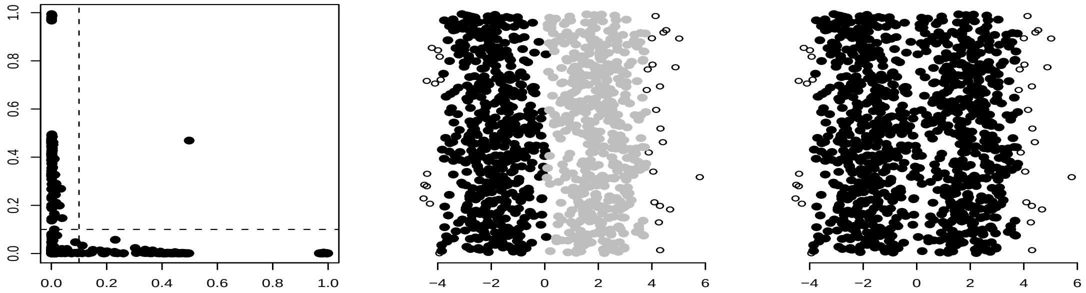

Figure 2: The plots illustrate the path of Algorithm 1 on a sample of 1000 observations from a symmetric Gaussian mixture, $0.5\,N(-2, 1)+0.5\,N(2, 1)$. The scatter plot on the left displays the sample frequencies for each pair of clusters merged along the path. The next two plots show the cluster memberships before and after the *big merge*. For the leftmost plot, the x and y axes denote the proportions of the points in the two merging clusters. For the other two plots, only the x axis, which marks the locations of the points, is informative.

Figure 2 illustrates the path of Algorithm 1 on a sample of 1000 observations, generated independently from the symmetric Gaussian mixture distribution used in Figure 1. The scatter plot on the left displays the sample frequencies for each pair of clusters merged along the path. We found only one merge in which both clusters pass the $\alpha = 0.1$ threshold. The *big merge* occurs at a point where the current number of clusters is 32. The two rightmost plots display cluster memberships before and after the merge. The non-shaded points belong to clusters with non-appreciable size.

The equivalence between the splitting procedure and the merging algorithm implies that in the post-processing step of Algorithm 1 we only keep the merges with the cardinality of each of the merging clusters above αn. For any such merge, we place the split point midway between the two closest representatives of the two clusters being merged. We also replace the stored merges with the corresponding split points. The resulting sequence of split points can then be reinterpreted as a sequence of splits, or a sequence of merges, using the full sample. For example, if the final output contains no split points, then all of the observations in the sample are placed in the same cluster. We call this modified approach the ***Big Merge Tracker*** (BMT) with threshold α. As a direct consequence of Theorems 1 and 2, under the respective regularity conditions, the BMT consistently estimates the number of population clusters, and its split points converge to their population counterparts at the $O_p(n^{-1/3})$ rate. In the next section we analyze the empirical performance of the BMT approach.

4 Simulation Study and Real Data Analysis

In this section we show strong performance of the proposed BMT approach relative to popular existing methods for assessing modality and estimating the number of clusters. We also illustrate the use of our methodology in analysis of single-cell virology datasets. In addition, in Section 5.3 of the Supplementary Material we apply BMT on very large simulated data sets and demonstrate its superior scalability properties.

When the separation between two clusters is very small, the population splitting procedure can still be successful at finding a split point by massively truncating the support. This *zooming-in* effect may result in larger sizes of the *small* sample splits, as we discussed, from a theoretical standpoint, in the paragraph above the statement of Theorem 2. To counteract this phenomenon, we propose an *adjustment* to the Big Merge Tracker. If the sum of the sample frequencies for the two merging clusters in the last big merge is less than 50%, we do not report any merges. Preventing the corresponding splitting procedure from truncating more than 50% of the data, while searching for the first split, slightly reduces its efficiency, but makes it more robust to sampling fluctuations. Throughout this section we use the adjustment described above and set the BMT threshold, α, equal to 0.1 (Algorithm 1 in Section 5 of the Supplementary Material contains the full pseudocode). Note that a large number of additional simulation results, corresponding to a wider range of sample sizes, together with an analysis regarding the choice of α, are provided in Section 5.4 of the Supplementary Material.

4.1 Modality Assessment

Testing for homogeneity of a population is an important statistical problem (Aitkin and Rubin, 1985; Müller and Sawitzki, 1991; Roeder, 1994). Here, we use the BMT to detect the presence of two or more dominant modes in the density. In Table 1 we compare our approach with two popular modality assessment procedures: (i) kernel density estimate based test of Silverman (1981) (ii) histogram based Diptest proposed by Hartigan and Hartigan (1985). P-values of the Silverman test are calculated using the R-package referenced in Vollmer *et al.* (2013). R-packge of Maechler (2013) is used for implementing the Dip test. Detailed descriptions of these procedures are given in Section 5.1 of the Supplementary Material.

We consider 5 different simulation scenarios, in which 100 independent samples of size 10000 were generated and subjected to modality analysis. Table 1 reports the percentage of cases in which multi-modality was detected. P-values for the Dip and Silverman tests were computed based on 1000 MCMC simulations, and decision on the null hypothesis of unimodality was made at the 5% level of significance. The mean and the standard deviation of the p-values from these tests are also reported. In the two unimodal scenarios the BMT is on par with the Silverman and the Dip tests in confirming unimodality of the population distribution with high certainty. In the three non-unimodal cases, which include normal and beta mixtures, the BMT shows better performance in detecting multi-modality.

4.2 Estimating the Number of Clusters

We study the potency of the BMT in detecting the true number of clusters. We compare its performance with the following *number of clusters* estimation methods: (i) the CH index of Caliński and Harabasz (1974) (ii) the KL index of Krzanowski and Lai (1988) (iii) the H measure of Hartigan (1975) (iv) the Silhouette statistic based KR index of Kaufman and Rousseeuw (2009), (v) the Gap statistic of Tibshirani *et al.* (2001) (vi) the Jump

Population Density	Dip Test P-value (D)			Silverman Test P-value (S)			BMT
	Mean (D)	Std(D)	**% multi-mode**	Mean (S)	Std(S)	**% multi-mode**	**% multi-mode**
N(0,1)	0.99	0.04	0.00	0.48	0.25	0.00	0.00
Beta(2,4)	0.98	0.04	0.00	0.54	0.28	2.00	0.00
$\{N(-1.1,1)+N(1.1,1)\}/2$	0.81	0.22	0.00	0.22	0.21	29.00	69.00
$\{Beta(4,6)+Beta(7,3)\}/2$	0.84	0.22	0.00	0.31	0.25	21.00	49.00
$\{N(-2.5,1)+N(0,1)+N(2.5,1)\}/3$	0.10	0.14	52.00	0.03	0.03	79.00	96.00

Table 1: Simulation study to compare multi-modality detection methods

statistic of Sugar and James (2003) (vii) the clustering prediction strength criterion of Tibshirani and Walther (2005), and (viii) the bootstrap based cluster instability minimizing criterion of Fang and Wang (2012), which is inspired by the stable clusters selection approach of Wang (2010). Detailed descriptions of these procedures are provided in Section 5.2 of the Supplementary Material. We consider one multivariate and five univariate regimes. 100 independent replications with the sample size of 5000 are used in each simulation setting, and the distribution of the number of clusters detected by each method is reported. The eight competitor methods are implemented using a number of different clustering approaches via the `NbClust` R-package of Charrad *et al.* (2014) and the `fpc` package of Hennig (2014). More specifically, we use k-means clustering with the Euclidean distance metric (the corresponding results are reported in Table 2), as well as Ward's method (Ward Jr, 1963), Centroid-based clustering (Kaufman and Rousseeuw, 2009), PAM (Kaufman and Rousseeuw, 1990), CLARA (Kaufman and Rousseeuw, 1986), clustering by merging Gaussian mixture components (Hennig, 2010) and hierarchical clustering initialized Gaussian mixture model based clustering method of Fraley and Raftery (2002). All the results for the non-k-means clustering approaches are reported in the tables provided in Section 5.2 of the Supplementary Material.

In our first univariate example, we consider a non-symmetric mixture of two normal densities. Each of them has unit variance and their means are fairly well-separated. We observe that the CH, KL and Hartigan methods struggle to recover the bimodal structure, while the others successfully detect the two clusters (we note that in this setting, the CH index performs better when it uses the centroid based clustering algorithm). The next three simulation scenarios correspond to non-symmetric tri-modal population densities that

True Population Density	Methods	Number of Clusters									
		1	2	3	4	5	6	7	8	9	10+
$0.3\,N(-4,1) + 0.7\,N(4,1)$	CH	0	**29**	12	4	2	2	4	11	14	22
	KL	0	**39**	10	5	5	8	8	8	9	8
	Hartigan	0	**0**	32	16	12	10	11	5	7	7
	Silhouette	0	**100**	0	0	0	0	0	0	0	0
	Gap	0	**100**	0	0	0	0	0	0	0	0
	Jump	0	**99**	0	0	0	0	0	0	0	1
	Pred Str.	0	**100**	0	0	0	0	0	0	0	0
	Stability	0	**100**	0	0	0	0	0	0	0	0
	BMT	0	**93**	7	0	0	0	0	0	0	0
$0.3\,N(-3,1) + 0.35\,N(0,1) + 0.35\,N(3,1)$	CH	0	0	**0**	0	1	0	6	13	28	52
	KL	0	11	**8**	11	9	6	14	12	19	10
	Hartigan	0	0	**69**	13	7	4	2	4	1	0
	Silhouette	0	8	**92**	0	0	0	0	0	0	0
	Gap	0	100	**0**	0	0	0	0	0	0	0
	Jump	0	0	**100**	0	0	0	0	0	0	0
	Pred Str.	0	0	**100**	0	0	0	0	0	0	0
	Stability	0	11	**89**	0	0	0	0	0	0	0
	BMT	0	5	**95**	0	0	0	0	0	0	0
$0.3\,t_1(-3) + 0.35\,t_1(0) + 0.35\,t_1(3)$	CH	0	7	**9**	7	11	7	6	10	14	29
	KL	0	18	**23**	15	14	5	4	9	6	6
	Hartigan	0	0	**46**	24	11	5	4	5	1	4
	Silhouette	0	71	**26**	3	0	0	0	0	0	0
	Gap	0	36	**64**	0	0	0	0	0	0	0
	Jump	17	11	**9**	13	12	5	11	9	7	6
	Pred Str.	0	24	**53**	19	4	0	0	0	0	0
	Stability	0	69	**30**	0	1	0	0	0	0	0
	BMT	0	1	**99**	0	0	0	0	0	0	0
$0.3\,\text{dexp}(-3) + 0.35\,\text{dexp}(0) + 0.35\,\text{dexp}(3)$	CH	0	0	**0**	0	1	0	6	14	23	56
	KL	0	13	**8**	18	9	11	7	20	10	4
	Hartigan	0	0	**55**	19	9	10	3	1	1	2
	Silhouette	0	39	**61**	0	0	0	0	0	0	0
	Gap	0	100	**0**	0	0	0	0	0	0	0
	Jump	0	0	**100**	0	0	0	0	0	0	0
	Pred Str.	0	0	**100**	0	0	0	0	0	0	0
	Stability	0	0	**100**	0	0	0	0	0	0	0
	BMT	0	0	**100**	0	0	0	0	0	0	0
$\big\{Beta(8,2) + Beta(5,5) + Beta(2,8)\big\}/3$	CH	0	0	**2**	0	3	3	6	16	22	48
	KL	0	13	**3**	11	9	10	15	16	9	14
	Hartigan	0	0	**57**	14	9	6	5	5	0	4
	Silhouette	0	0	**100**	0	0	0	0	0	0	0
	Gap	0	100	**0**	0	0	0	0	0	0	0
	Jump	99	0	**1**	0	0	0	0	0	0	0
	Pred Str.	0	78	**17**	4	1	0	0	0	0	0
	Stability	0	0	**25**	60	15	0	0	0	0	0
	BMT	0	0	**100**	0	0	0	0	0	0	0
$\{0.5\,N(-2,1) + 0.5\,N(2,1)\}$ $\otimes\, N(0,1)$ $\otimes\, N(0,1)$ $\otimes\, \chi^2_1$ $\otimes\, \chi^2_1$	CH	0	**0**	1	0	4	5	11	10	23	46
	KL	0	**18**	11	14	7	13	9	11	11	6
	Hartigan	0	**0**	52	16	13	4	7	4	4	0
	Silhouette	0	**0**	87	12	0	1	0	0	0	0
	Gap	0	**100**	0	0	0	0	0	0	0	0
	Jump	0	**14**	0	62	0	0	0	0	8	16
	Pred Str.	0	**100**	0	0	0	0	0	0	0	0
	Stability	0	**85**	0	15	0	0	0	0	0	0
	BMT	0	**96**	4	0	0	0	0	0	0	0

Table 2: Number of clusters detected in 100 trials for six simulation scenarios

are mixtures of standard normals, non-central t-densities with one degree of freedom and double exponential densities with the unit rate parameter, respectively. The medians of the mixing densities and the mixture weights match across the three settings, and the separation between the adjacent medians is not large. In these three simulation settings, the Gap statistic approach, as well as the CH, KL and Hartigan measures, has difficulty detecting the true number of clusters. The Silhouette method performs well for Gaussian mixtures but has difficulties in the other two cases. Jump statistic, prediction strength and bootstrap stability approaches do well in the normal and the double exponential cases, however, they do not show good performance in the considerably thicker-tailed case of the mixture of t-densities. For our fifth simulation setting we consider a bounded population density that is a mixture of three Beta densities. Here, only the BMT and the Silhouette do well in recovering the true number of clusters. In our last example we consider a 5 dimensional data set, which is generated from a product density. The first dimension is generated from a symmetric mixture of two Gaussians; the next two dimensions contain white noise, while the forth and fifth dimensions are generated from a central chi-square distribution with one degree of freedom. We observe that, together with the CH, KL and Hartigan measures, the Silhouette and the Jump approaches do not perform well in detecting the two clusters in this data.

BMT does consistently well across each of the six simulation scenarios, outperforming all the other approaches overall. The prediction strength and the cluster stability methods, which are modern state of the art approaches, deliver the best results among the competitors. However, these two methods have significant trouble in the cases of the beta and the non-central t mixtures. When implemented with various non-k-means clustering approaches (see Section 5.2 of the Supplementary Material), neither of the competitors considerably improves the performance reported in Table 2. Figure 3 in Section 5.2 of the Supplementary Material provides plots of the densities used in the numerical experiments.

4.3 Sub-population Analysis in Single Cell Virology

We demonstrate an application of our clustering approach in a *single-cell Mass Cytometry* (Bendall *et al.*, 2011) based virology study. We analyze the data reported in Sen *et al.* (2014), where the effect of Varicella Zoster Virus (VZV) on human tonsil T cell is studied. VZV is a human herpesvirus and causes varicella and zoster (Zerboni *et al.*, 2014).

We study protein expressions from five independent experiments, each containing an Uninfected (UN) and a Bystander (BY) populations. Bystanders are cells in the VZV infected population, which are not directly infected by the virus, but are influenced by neighboring virus infected cells. Protein expression values are studied on the arcsinh scale. Non-expressed values are uniformly distributed between $[-1, 1]$. Cellular sub-populations are detected by clustering the populations based on the expressions of "core-proteins", which are associated with T cell activation (Newell *et al.*, 2012). Most of the samples have large sizes, usually on the order of $\sim 10^5$. Traditional clustering techniques fail to accommodate such large sample sizes and resort to sub-sampling based approaches (Qiu *et al.*, 2011; Linderman *et al.*, 2012). The BMT, on the other hand, has the advantage of being scalable enough to conduct clustering analysis on the entire sample.

	Experiment I		Experiment II		Experiment III		Experiment IV		Experiment V	
SUB-POPULATIONS	UN	BY	UN	BY	UN	BY	UN	BY	UN	BY
DUAL POSITIVE	8411 (8.8%)	6596 (7.3%)	5253 (5.8%)	4169 5.7%)	4971 (6.0%)	2703 (5.4%)	3795 (5.0%)	1510 (4.6%)	8047 (8.5%)	5225 (8.0%)
DUAL NEGATIVE	2723 (2.8%)	2973 (3.3%)	3537 (3.9%)	2631 (3.6%)	4433 (5.3%)	2935 (5.9%)	4354 (5.8%)	2196 (6.7%)	5012 (5.3%)	2881 (4.4%)
CD4 NON-NAIVE	7993 (8.4%)	10636 (11.8%)	15144 (16.7%)	11556 (15.9%)	21444 (25.9%)	12429 (25.0%)	22149 (29.7%)	8508 (26.1%)	30034 (31.9%)	20469 (31.3%)
CD4 NAIVE	69977 (73.7%)	64119 (71.1%)	57744 (63.7%)	47374 (65.1%)	45764 (55.3%)	27987 (56.3%)	35458 (47.5%)	16390 (50.4%)	40398 (43.0%)	28524 (43.7%)
CD8 NAIVE	5654 (6.0%)	5671 (6.3%)	8599 (9.5%)	6571 (9.0%)	5798 (7.0%)	3490 (7.0%)	8271 (11.1%)	3774 (11.6%)	9869 (10.5%)	7829 (12.0%)
POPULATION SIZE	94837	90157	90641	72699	82637	49672	74540	32497	93878	65244

Table 3: Sizes and Proportions of dominant clusters detected by BMT across 5 independent Virology experiments

We treat three proteins, CD4, CD8 and CD45RA (naive), as core-proteins, as they are typically used to classify T cells. For each of the 10 samples (UN and BY from experiments I-V), based on the expressions of the above three proteins, we performed automated clustering by using BMT in the three dimensional space. Figure 4 and 5 in Section 5.5 of the Supplementary Material show that in all the cases, the BMT detects unimodality for CD4 and CD45RA and bimodality for CD8 expression values. Using the bi-modality of CD8 and the BMT detected splits, we classify cells as CD8-high and CD8-low. Also, considering the expression and non-expressions of the other two markers we simultaneously classify cells into the following clusters, or sub-populations: (i) Dual positive: CD4 expressed and CD8 High (ii) Dual-negative: CD4 non-expressed and CD8 low (iii) CD4 Non-Naive: CD4 expressed and CD45RA non-expressed (iv) CD4 Naive: CD4 expressed and CD45RA non-expressed (v) CD8 Naive: CD8 high and CD45RA expressed.

Table 3 reports the sizes and proportional representations of these sub-populations, across the five experiments. BMT sub-populations resemble the T-cell biology based phenotypic classification in Sen *et al.* (2014). They also revalidate that the sub-population distribution in the Bystander cells is not much different from that of the uninfected, though the UN sub-population distribution varies across experiments. Using this BMT based categorization of the T cells, sub-population level cell-signaling patterns can be subsequently studied. Figure 6 in Section 5.5 of the Supplementary Material shows the heatmap of the protein expressions (core + signaling proteins) of the sub-populations from Experiment I.

5 Discussion

In this paper we focus on the analysis of a popular convex clustering approach that is based on the ℓ_1 fusion regularization. However, we believe that our general theoretical framework can be extended to handle other types of fusion penalties. In particular, based on a preliminary analysis of the corresponding ℓ_2 approach, we conjecture that the asymptotic results in this paper also hold in the ℓ_2 case, under appropriate regularity conditions. The corresponding population procedure can be defined by analogy, as a collection of cluster splits and truncations, where each operation seeks to maximize the Euclidian distance between the corresponding sub-cluster means. The proofs require a rigorous formulation and a thorough analysis of a multivariate analog of the truncation operation.

High computational efficiency of the proposed BMT approach allows it to be applied to massive high-dimensional data sets. In particular, BMT can be used for high-dimensional feature screening, to rule out predictors that do not reveal any clusters in the data. Moreover, we can take further advantage of BMT's computational efficiency and apply the screening procedure to several linear combinations for each pair of variables. This way we can move beyond marginal screening, similarly to how regression models with interaction terms move beyond the simple additive structure.

Acknowledgement

Radchenko's research was partially supported by NSF Grant DMS-1209057. Computing resources were provided by the USC Center for High Performance Computing. The R code and the data sets used for producing the numerical results of Section 4 can be downloaded from `http://www-bcf.usc.edu/~gourab/code-bmt`.

References

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)* 67–75.

Bach, F. R. and Harchaoui, Z. (2008). Diffrac: a discriminative and flexible framework for clustering. In *Advances in Neural Information Processing Systems*, 49–56.

Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, vol. 14, 585–591.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., and Nolan, G. P. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science (New York, N.Y.)* **332**, 6030, 687–696.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics* **64**, 1, 115–123.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* **3**, 1, 1–27.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* **61**, 6, 1–36.

Chi, E. C. and Lange, K. (2013). Splitting methods for convex clustering. *arXiv preprint arXiv:1304.0499* .

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis* **56**, 3, 468–477.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97**, 458, 611–631.

Hartigan, J. (1978). Asymptotic distributions for clustering criteria. *The Annals of Statistics* 117–131.

Hartigan, J. A. (1975). *Clustering algorithms.* Wiley.

Hartigan, J. A. and Hartigan, P. (1985). The dip test of unimodality. *The Annals of Statistics* 70–84.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in data analysis and classification* **4**, 1, 3–34.

Hennig, C. (2014). *fpc: Flexible procedures for clustering.* R package version 2.1-9.

Hocking, T., Vert, J.-P., Bach, F., and Joulin, A. (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In L. Getoor and T. Scheffer, eds., *ICML*, 745–752. Omnipress.

Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics* **19**, 4, 984–1006.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* **31**, 8, 651–666.

Kaufman, L. and Rousseeuw, P. (1986). *Clustering large data sets.* Elsevier.

Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis* 68–125.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons.

Ke, T., Fan, J., and Wu, Y. (2013). Homogeneity in regression. *arXiv preprint arXiv:1303.7409* .

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics* 191–219.

Krzanowski, W. J. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 23–34.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* **14**, 3, 547–568.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* **8**, 8, 1687–1723.

Linderman, M. D., Bjornson, Z., Simonds, E. F., Qiu, P., Bruggner, R. V., Sheode, K., Meng, T. H., Plevritis, S. K., and Nolan, G. P. (2012). Cytospade: high-performance analysis and visualization of high-dimensional cytometry data. *Bioinformatics* **28**, 18, 2400–2401.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, 201–204. IEEE.

MacQueen, J. *et al.* (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, 281–297. California, USA.

Maechler, M. (2013). *diptest: Hartigan's dip test statistic for unimodality - corrected code*. R package version 0.75-5.

Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association* **86**, 415, 738–746.

Newell, E., Sigal, N., Bendall, S., Nolan, G., and Davis, M. (2012). Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of cd8+ t cell phenotypes. *Immunity* **36**, 1, 142 – 152.

Pelckmans, K., De Brabanter, J., Suykens, J., and De Moor, B. (2005). Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics* **9**, 1, 135–140.

Pollard, D. (1982). A central limit theorem for k-means clustering. *The Annals of Probability* 919–926.

Qiu, P., Simonds, E. F., Bendall, S. C., Jr., K. D. G., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature Biotechnology* .

Rinaldo, A. *et al.* (2009). Properties and refinements of the fused lasso. *The Annals of Statistics* **37**, 5B, 2922–2952.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association* **89**, 426, 487–495.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* 1878–1915.

Sen, N., Mukherjee, G., Sen, A., Bendall, S., Sung, P., Nolan, G., and Arvin, A. (2014). Single-cell mass cytometry analysis of human tonsil t cell remodeling by varicella zoster virus. *Cell Reports* **8**, 2, 633 – 645.

Shen, X. and Huang, H.-C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* **105**, 490.

Shen, X., Huang, H.-C., and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika* **99**, 4, 899–914.

Shi, T., Belkin, M., and Yu, B. (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics* 3960–3984.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)* 97–99.

Soltanolkotabi, M. and Candés, E. J. (2012). A geometric analysis of subspace clustering with outliers. *Ann. Statist.* **40**, 4, 2195–2238.

Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset. *Journal of the American Statistical Association* **98**, 463.

Tan, K. M. and Witten, D. (2015). Statistical properties of convex clustering. *arXiv preprint arXiv:1503.08340* .

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 3, 273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 1, 91–108.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics* **14**, 3, 511–528.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 2, 411–423.

Tibshirani, R. J. (2013). Adaptive piecewise polynomial estimation via trend filtering. *arXiv preprint arXiv:1304.2986* .

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag.

Vollmer, S., Holzmann, H., and Schwaiger, F. (2013). Peaks vs components. *Review of Development Economics* **17**, 2, 352–364.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97**, 4, 893–904.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 301, 236–244.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**, 490.

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. (2004). Maximum margin clustering. In *Advances in neural information processing systems*, 1537–1544.

Zerboni, L., Sen, N., Oliver, S. L., and Arvin, A. M. (2014). Molecular mechanisms of varicella zoster virus pathogenesis. *Nature Reviews Microbiology* .

Zhu, C., Xu, H., Leng, C., and Yan, S. (2014). Convex optimization procedure for clustering: Theoretical revisit. In *Advances in Neural Information Processing Systems*, 1619–1627.