Initializing Partition-Optimization Algorithms

Ranjan Maitra

***Abstract*— Clustering datasets is a challenging problem needed in a wide array of applications. Partition-optimization approaches, such as k-means or expectation-maximization (EM) algorithms, are sub-optimal and find solutions in the vicinity of their initialization. This paper proposes a staged approach to specifying initial values by finding a large number of local modes and then obtaining representatives from the most separated ones. Results on test experiments are excellent. We also provide a detailed comparative assessment of the suggested algorithm with many commonly-used initialization approaches in the literature. Finally, the methodology is applied to two datasets on diurnal microarray gene expressions and industrial releases of mercury.**

***Index Terms*— Toxic Release Inventory, methylmercury, multi-Gaussian mixtures, protein localization, singular value decomposition**

I. Introduction

There is a substantial body of literature devoted to the issue of grouping data into an unknown number of clusters [1]–[13]. Indeed, the numerous methods proposed in the literature reflect both the challenges and the wide applicability of the problem. Most approaches involve a certain degree of empiricism but broadly fall into either the hierarchical clustering or the partition-optimization categories. The former provide a tree-like structure for demarcating groups, with the property that all observations in a group at some branch node are also together higher up the tree. Both agglomerative and divisive approaches exist, with groups merged or split at a node according to a previously defined between-groups similarity measure.

An entirely different class of clustering algorithms divides the dataset into a number of homogeneous clusters based on some optimality criterion such as the minimization of some aspect (commonly the trace or determinant) of the within-sums-of-squares-and-products (SSP_W) matrix [14], [15], or the maximization of likelihood and estimation of parameters in a model-based setting, followed by an assignment of each observation to the class with maximum posterior probability. In a one-dimensional framework, an optimal partition is found, following Fisher's [16] algorithm which is computationally intractable in a multivariate setting. Available algorithms are sub-optimal in all but the most trivial of cases: two common methodologies are k-means and the expectation-maximization (EM) approach to estimation in an appropriately specified mixture model. Several generalizations of k-means exist: the k-medoids algorithm (*cf.* Chapter 2 of Kaufman and Rousseeuw [17]) is perhaps the one most familiar to statisticians. These are all iterative approaches, finding optima in a neighborhood of their initialization, with identified groups heavily dependent on these starting values [18]. Thus, it is imperative to have methodology that provides good initializers for these algorithms, two applications of which are presented next.

Manuscript received October 24, 2006. This material is based, in part, upon work supported by the National Science Foundation (NSF) under its CAREER Grant No. DMS-0437555. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

R. Maitra is with the Department of Statistics at Iowa State University, Ames, IA 50011-1210, USA. E-mail: maitra@iastate.edu.

A. Two Case Studies in the Life and Health Sciences

This section illustrates two scenarios in the public health and biological sciences that would benefit from improved initialization methods for partition-optimization algorithms. The first is in the context of analyzing microarray gene expression data on the diurnal starch content of *Arabidopsis L. Heynth* leaves. A second application profiles mercury releases reported by different industrial facilities in the United States in the year 2000, in a first step towards understanding and addressing factors behind this important public health concern.

1) Identifying Similar-Acting Genes in a Plant's Diurnal Cycle: The processes and pathways involved in the synthesis and degradation of starch – the most abundantly stored form of carbon in plants, a major source of calories in the human diet, and an important industrial commodity – is not fully understood. In Arabidopsis, the two are integrated processes, occurring at rates that relate to the duration of day and night. Further, the chemical structure and composition of starch in Arabidopsis leaves is similar to that in crops, so a detailed investigation of its synthesis and degradation in the former can provide clarification and deeper understanding of the processes in the plastids of living cells [19], [20].

According to Smith *et al* [21], the Arabidopsis genome sequence reveals many genes encoding enzymes potentially involved in starch synthesis and degradation. While the functionality of some of these genes may be considered to be well-established, having been extensively studied in other species, that of several other genes is uncertain. Determining the latter is an important step in determining the processes of starch breakdown and synthesis. One approach is to correlate changes in their abundance levels to those of the known genes. Microarray technology readily permits such measurement: the European Arabidopsis Stock Centre website at $http://nasc.nott.ac.uk$ provides Affymetrix ATH1 microarray data on 22,810 genes from plants exposed to equal periods of light and darkness in the diurnal cycle. Leaves were harvested at eleven time-points, at the start of the experiment (end of the light period) and subsequently after 1, 2, 4, 8 and 12 hours of darkness and light each. The whole experiment was repeated on plants in the same growth chamber, resulting in data from a randomized complete block design (RCBD).

Some of the genes encoding enzymes in the dataset are believed to be involved in starch synthesis – *eg* PGI1, PGM1,

APLs, AGPase and STS are surmised to encode predicted plastidial transit properties [21]. The pathway for effecting starch breakdown is even less understood, with the presumed involvement, with varying degrees of certainty, of GWDs and their derivatives, AMYs, BAMs, DPEs and so on. Functionality can be identified by grouping the transcriptomes into homogeneous clusters of similar-acting genes. A distance measure of choice here is correlation because it can be used to group genes that act together, regardless of the exact value of the gene expressions. This can be used in conjunction with k-means, for instance, to derive clusters of similar-acting genes in the starch diurnal cycle. An added incentive for the choice of k-means is that standard software can be used because the Euclidean distance applied to observations transformed to be on the zero-centered unit sphere orthogonal to the unit vector, is the square root of an affine transform of correlation. Having thus arrived at a grouping, one could investigate membership and potentially draw conclusions on the roles of different enzymes in starch breakdown and synthesis. Here again, the potential use of k-means draws attention to the need for an effective initialization strategy.

2) Profiling Industrial Facilities that Release Mercury: Prenatal exposure to mercury (and the more toxic methylmercury) has been a long-standing concern for public health officials because it can cause adverse developmental and cognitive effects in children, even at low doses [22], [23]. Such children are at elevated risk of performing neurobehavioral tasks poorly, and possible adverse effects on their cardiovascular, immune, and reproductive systems [24], [25]. Despite the US government's decades-long efforts to curtail home uses of mercury, the 2003 Environmental Protection Agency (EPA) [26] report on "America's Children and the Environment" found that about 8 percent of women of child-bearing age in 1999-2000 (defined to be in the age group of 16–49 years) had at least 5.8 parts per billion (ppb) mercury in their blood. There is no safe limit for methylmercury, which is more readily absorbed in the body than inorganic mercury and therefore, more potent. It enters the food chain following its conversion from elemental mercury in the environment by bacteria, and is then transferred to humans through eating contaminated fish. The elemental form of mercury entering the food chain is believed to originate as emissions and releases from industrial facilities, sometimes carried over long distances on global air currents far away from their source [27]. Devising effective policies for limiting industrial releases of mercury is essential, and a first step towards this goal is to understand the characteristics of its many different aspects.

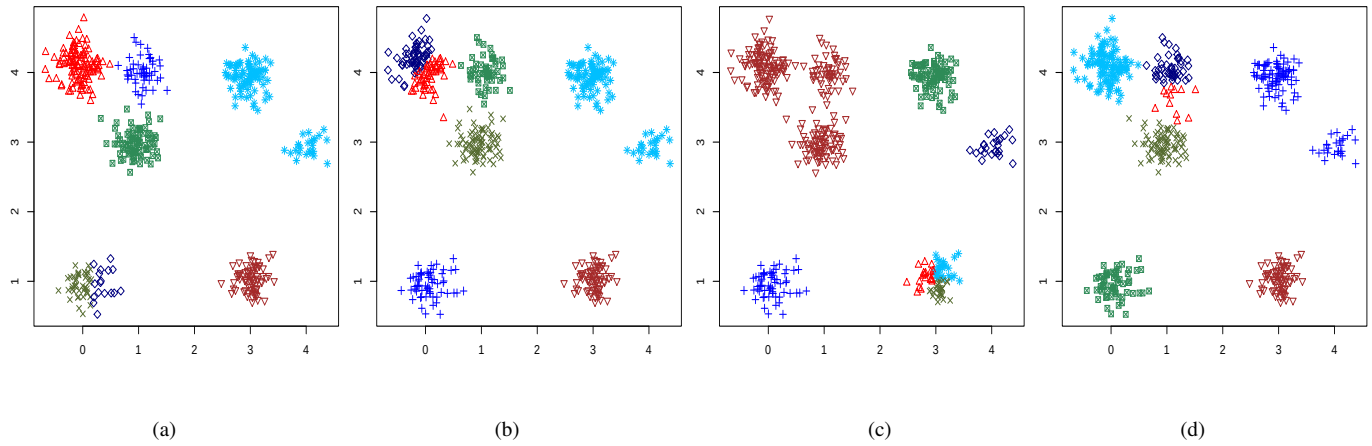
The EPA's Toxic Release Inventory (TRI) database contains annual release data on a list of chemicals as reported by eligible industrial facilities. Only 1,596 of 91,513 reports submitted in 2000 concerned mercury and its compounds. Combining multiple reports from the same facility resulted in 1,409 separate facility reports for releases of mercury (in pounds) into air (whether as fugitive or stack air emissions), water, land, underground injection into wells, or as off-site disposal. Electric, gas and sanitary services facilities accounted for 539 of these reports (with coal- or gas-combusting electricity-generating plants submitting 464 reports) followed by facilities involved with chemicals manufacture or processing (162 reports), stone, clay, glass and concrete products (149 reports), primary metals (115 reports), petroleum refining and related products (110 reports) and paper and allied products (100 reports). In all, 24 different broad classes of industrial facilities, as identified by the 2-digit Standard Industry Classification (SIC) codes, reported mercury releases in the year 2000.

An unsupervised learning approach is an invaluable tool in understanding the different characteristics of industrial mercury releases because the conclusions flow directly from the reported data. Finding facilities that are similar in the context of their reported releases, and hence their impact on public health, would make it possible to draw profiles for each category. Groups of particular interest could be analyzed further in terms of industry and regional compositions, resulting in the formulation of more targeted policies designed to have maximum impact on reducing mercury pollution. Clustering thus becomes an effective tool in framing public policy in this environmental health scenario. There are no natural hierarchies expected in such groupings, which means that we are led to use an partition-optimization algorithm, and inherently to the need for methods to effectively initialize them.

B. Background and Significance

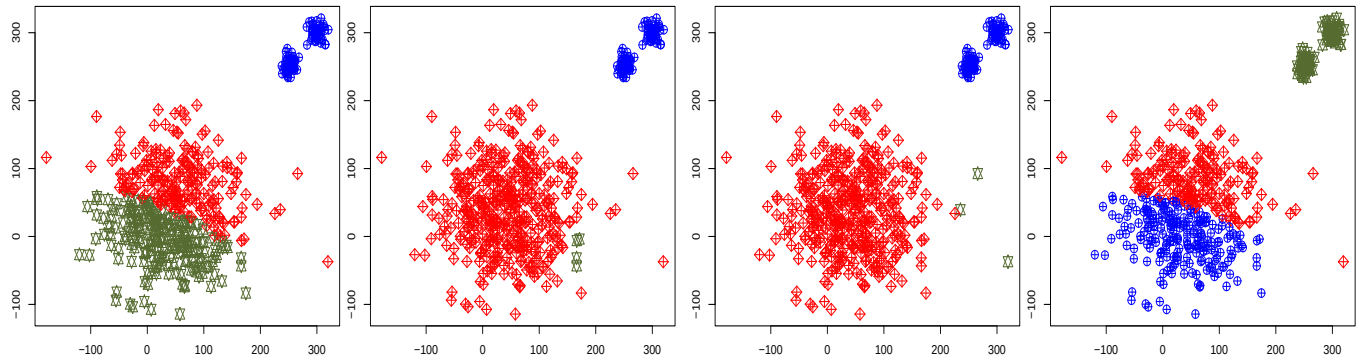
A number of somewhat ad hoc approaches are currently used to assign starting values to partition-optimization algorithms. The statistical package Splus obtains initializing centers for k-means from a hierarchically clustered grouping of the data, while R [28] uses k randomly chosen observation points as the default initializer. Lozano *et al* [29] show that this is, on the average, a viable strategy for initialization: however, obtained clusterings can sometimes be very erroneous, as seen in the two-dimensional example of Figure 1 which provides the results on four successive calls to 7-means, using R's default settings. If the number of clusters is known to be seven, the correct grouping for most observations is very apparent. Further, the groups appear to be quite homogeneous so that k-means is a good choice for partitioning the dataset. Indeed, with initial seeds from the correct clusters, the algorithm provides the most apparent solution. But the random-starts strategy has a fair chance of not getting initial representatives from different clusters in the above example, making the task of ultimately obtaining the correct grouping more difficult.

A common strategy suggested to alleviate the above problem is to run the algorithm with several random starts and to choose the best solution. Such a strategy is not always practical, especially for high-dimensional datasets. As a refinement, Bradley and Fayyad [30] and Fayyad, Reina and Bradley [31] proposed clustering several sub-samples of the data, using k-means or EM, depending on the clustering methodology being used and random-starts. The cluster centers obtained from each sub-sample are again partitioned using k-means, with initial values serially provided by the final cluster centers from each sub-sample. Each exercise provides a further set of cluster means, and the initializing values for the k-means algorithm are chosen from amongst this last set to be closest in a least-squares sense to all of them. Both methods give



(a) (b) (c) (d)

Fig. 1. Results of running the k-means algorithm with seven clusters and using four different starting points, each randomly chosen from the dataset. Each identified group is identified by a different plotting character and color. The role of the plotting character and color here and in Figures 2 and 4 is nominal. Note the widely divergent results.



(a) (b) (c) (d)

Fig. 2. Results of running the EM-algorithm with three clusters, initialized from parameter estimates obtained using hierarchical clustering with (a) Ward's minimum variance criterion, (b) single, (c) average and (d) likelihood gain (with Mclust) linkages.

rise to some concern about bringing in a random component to an essentially deterministic algorithm, with no clear way of assessing and controlling for variability in the estimation process. Further, the logic behind the choice of the final step in Fayyad *et al* [31] appears unclear and has the strong potential for providing infeasible starting values and for eliminating smaller clusters that are wide apart but which should be natural candidates to be picked up by any reasonable algorithm.

The alternative approach of using hierarchical clustering into seven groups (using any merge criterion) to obtain initial seeds for k-means and EM approaches performs very well in Figure 1. However, consider the example in Figure 2, modified from the one in Kettenring [8], with three clear clusters, one dominant and accounting for 90% of the dataset. The other two clusters are smaller, equal-sized, considerably farther away from the larger cluster and closer to each other. In keeping with practical scenarios, the two smaller clusters have substantially lower dispersion than the larger one. Performance of k-means is understandably poor in this setting, even when initialized with the true centers, so performance using EM on multi-Gaussian mixtures is explored. Figure 2 shows the results upon initializing the algorithm with hierarchical clusterings obtained with four different linkages. In all four cases, performance is poor. Note that the likelihood gain merge criterion is used in the popular Mclust algorithm [32].

The choice of hierarchical clustering to initialize partition-optimization algorithms is in itself debatable, given the inherent assumption of a regimentation structure in the dataset with the number of groups determined by the resolution at which the dataset is summarized (or in practical terms, at the height at which the tree is cut). Partition-optimization algorithms do not subscribe to this limited world-view, but using hierarchical clustering to initialize them results in a *de facto* prescription. Further, most hierarchical clustering algorithms combine in essentially a binary fashion, with dissimilarities measured in terms of combinations of *paired* inter-point distances rather

than some comprehensive metric. More practically, a dissimilarity matrix requires computations an order of magnitude higher than the number of observations, so the size of the datasets that can be handled is severely restricted. In many modern applications, such as the complete microarray dataset introduced in Section I-A.1, this is too severe a restriction.

Other methods have also been proposed and investigated in the context of different applications. Tseng and Wong [33] proposed an initialization approach in the context of finding centers of tight clusters, which cuts the tree into $k \times p$ groups, with p representing dimensionality of the observations, and then choosing the k most populated clusters. Al-Daoud [34] suggests choosing as initial set of means the data-points closest to the modes of the projection of the data on the first principal component. An inherent assumption behind clustering is that an unknown grouping is the dominant source of variability in the dataset: the suggestion is unsatisfactory when the first principal component is inadequate in capturing most of the variability in the data. In the context of EM algorithms for Gaussian clustering, Biernacki, Celeux and Govaert [35] do a detailed analysis of several initialization methods and find that using several short runs of the EM initialized with *valid* random starts as parameter estimates – with validity defined by existence of likelihood – provides the best initializer in terms of ultimately maximizing the likelihood. They call this the em-EM approach. Specifically, each short run consists of stopping the EM algorithm, initialized with a valid random start, according to a lax convergence criterion. The procedure is repeated until an overall number of total iterations is exhausted, at which point the solution with the highest log-likelihood value is declared to be the initializer for the long EM. This approach is computationally intensive and suffers from the same comments on random starts mentioned above. Further, note that just using one random start of the EM, as well as choosing from several random initializing points and deciding on the set with highest likelihood are special cases of the above method of initialization. Intuitively, it is debatable whether computer time is not better utilized in increasing the number of random starts (and no short runs of the EM). Indeed, our experiments indicate this as an acceptable alternative, even with ignoring the computational advantage and keeping the same number of random starts as the total number of iterations of the short run iterations. In fact, a larger number of random starts can actually be considered because the computational overhead from the task of actually performing short runs of the EM-step is completely removed. We call this approach Rnd-EM and illustrate that its performance is comparable even when the number of random starts is kept the same as the total number of short em iterations in em-EM.

This paper provides a staged deterministic approach to initializing partition-optimization algorithms. The basic idea, detailed in Section 2, is to obtain a large number of modes of the dataset, and then to choose representatives from the most widely-separated ones. In a specific implementation of the above, we first propose finding modes in the one-dimensional projections along the direction of the singular vectors. For datasets with all singular values positive, we also find modes in the each standardized dimension of the original dataset. The product set of all these modes, after thinning for points with no representation in the data, is fed into a k-means algorithm to obtain a large number of local multivariate modes, which are then explored for representatives from the ones that are farthest apart. For the k-means algorithms, the result provides the initial values, while the initializers for EM are provided by parameters estimated from the classification of the data to the closest mode. Performance evaluations on test experiments are detailed in Section 3. The gene expression and the mercury release datasets introduced in this section are analyzed in Section 4. The paper concludes with some discussion.

II. Methodology

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$ be a random sample, each from one of an unknown number K of p-variate populations. In a model-based setup, given K, the density of each $\boldsymbol{X}_i$ is assumed to be given by $f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k g_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$, where $g_k(\cdot)$ denotes the kth sub-population multivariate density with parameter $\boldsymbol{\theta}_k$ and π_k is the mixing proportion of the kth subpopulation. A commonly-used density, and the one considered throughout this article, is the multivariate Gaussian, with $\boldsymbol{\theta}_k$ representing the corresponding mean and dispersion parameters. One goal is to estimate the number of sub-populations K, the parameters $\boldsymbol{\theta}_k$s and π_ks; however our primary interest is in classifying the observations into their correct group. The EM approach to solving the problem, given K, assumes class membership as the missing information, and proceeds to set up an iterative scheme to estimate the parameters, starting with a set of initial parameter estimates. Once these parameters are estimated, they are used to classify observations into groups based on their posterior probabilities of inclusion. (Note that, as very kindly pointed out by a referee, the Gaussian means are all assumed to be different: this assumption, while reasonable from a practical standpoint in clustering, also obviates identifiability issues in parameter estimation.)

The k-means algorithm, on the other hand makes no overt distributional assumptions even though it can be formulated as a solution to the likelihood equation for the means and class identities of a fixed-partition model of multinormal distributions with equal and spherical dispersion matrices. Formally, here we have a random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$ with joint likelihood given by $\prod_{i=1}^{n} \sum_{i=1}^{K} \eta_{i,k} \phi(\boldsymbol{X}_i; \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})$ where $\phi(\cdot)$ is the multivariate normal density and $\eta_{i,k}$ is one if $\boldsymbol{X}_i$ is in the kth subpopulation, zero otherwise. In this setup, the $\eta_{i,k}$'s and $\boldsymbol{\mu}_k$'s are parameters to be estimated, σ is a nuisance parameter. A local solution to this likelihood problem is provided by k-means, which starts with a set of initial centers and partitions the dataset by assigning each observation to the closest (according to some metric) cluster center, updates the cluster center in each partition and iterates till convergence. Like EM, convergence is to a local optimum in a neighborhood of the initializer and can be far from the global solution. Initialization methods that perform well are important: we propose a class of such algorithms next.

A. A Multi-Stage Initializer

Let $\boldsymbol{X}$ be the $n \times p$ data matrix with rows given by the observations $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$. Our objective is

to find initial seeds for partitioning algorithms to group the dataset into K clusters, assuming that K is known. Consider the following multi-stage algorithm:

1) Obtain the singular value decomposition (SVD) of the *centered* data $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$, where $\boldsymbol{D}$ is the diagonal matrix of the m positive singular values $d_1 \geq d_2 \geq \ldots \geq d_m$, and $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices of order $n \times m$ and $p \times m$, both with orthonormal columns (in n- and p-dimensional space, respectively). For a given m^*, consider the reduced $n \times m^*$ projection given by $\boldsymbol{U}_*$ consisting of the first m^* columns of $\boldsymbol{U}$ given by $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{m^*}$. We propose working in the reduced space.
2) For each coordinate in the reduced space, we obtain an appropriate number of local modes. We choose more modes in those coordinates with higher singular values (or standard deviations of the principal components), under the assumption that information in the dataset is more concentrated along those projections corresponding to higher values, and therefore these would contain more information about the clusters. Specifically, we propose choosing the number of modes, k_j in the jth reduced-space coordinate to be equal to $\lceil (c_{m-m^*}K)^{\frac{1}{m^*}} \rceil d_j/d_{m^*}$ rounded to the nearest integer, with $\lceil x \rceil$ denoting the smallest integer greater than or equal to x, and c_k is non-decreasing and concave in k. While one could use Fisher's [16] computationally demanding prescription for one-dimensional partitions, we propose one-dimensional k-means to determine the modes in the jth reduced coordinate data space initialized using the quantiles corresponding to the k_j equal increments in probabilities in $(0, 1)$. The choice of k-means is appropriate because the goal here is to find a large number of univariate local modes for input into the next step.
3) Form the set of candidate multivariate local modes in the reduced space by taking the product set of all the one-dimensional modes. Eliminate all those candidates from the product set which are not closest to any observation in $\boldsymbol{U}_*$. The remaining k^* modes are used as initial points for a k-means algorithm that provides us with k^* local modes. Note that typically, $k^* >> k$.
4) Obtain the k^* local modes of the dataset using the k-means algorithm with the starting points provided from above. Also, classify the observations, and obtain the corresponding group means in the original domain.
5) At this point, we have k^* local modes of the dataset in the reduced space and the corresponding group centers in the original space. The goal is to obtain k representative points from the above which are as far as possible from each other. We use hierarchical clustering with single-linkage on these k^* modes and cut the tree into k groups. Since a single-linkage merge criterion combines groups based on the least minimum pairwise distance between its members, its choice in the hierarchical clustering algorithm here means that we obtain k groups of local modes (from out of k^*) that are as far apart in the transformed space as possible. Means, and if needed, relative frequencies and dispersions, of the observations in the dataset assigned to each of the k grouped modes are calculated: these provide the necessary initialization points for the partition-optimization algorithms.

The above algorithm is practical to implement: perhaps the most computer-intensive step in this exercise in the singular value decomposition, requiring about $O(p^3)$ computations (see page 237–40 of Demmel [36]. The series of univariate mode-finding steps can be executed quite easily using existing quicksort algorithms for finding the appropriate quantiles, and then using k-means in the one-dimension. The product set of the univariate modes can be obtained by direct enumeration, essentially an $O(k)$ operation. Using k-means to provide for the k^* local modes keeps the algorithm efficient, while hierarchical clustering is done on the k^* modes and therefore, a distance (dissimilarity) matrix of around $O(k^2)$-elements needs to be stored and computed. Thus, unlike the case when hierarchical clustering is used to obtain initializing values, the size of the dataset n is not a major limiting factor in the computation.

The above strategy can throw up initial groups, some with less than p data-points. In the EM context with general variance-covariance matrix structures, this is problematic because it results in singular initial dispersion estimates for those clusters. When such a case arises, we estimate a common $\boldsymbol{\Sigma}$ for those groups for which the initial dispersion matrices are singular. This is done by taking the average of the initial variance-covariance matrices for those clusters with nonsingular initial estimates of dispersions. This is done only for the initialization. This approach was used in the experimental evaluations reported in this paper. Finally in this regard, we note that when there is more specific information about the covariance structures in the mixture model, that can be used in obtaining the initializers.

The choice of m^* is crucial in the suggested setup. Our recommendation, which we adopt in our experiments and software, is to run the algorithm values for different values of m^* which d_ls are positive, and then to choose, for k-means, those starting values for which $\mid SSP_W \mid$ is minimum. For the EM algorithm, we choose the set of initial values maximizing the likelihood. When $\boldsymbol{X}$ is of full column rank, we also compare with the above method modified for use in the original coordinate-space. Specifically, we standardize the variables to have zero mean and unit variance in each coordinate, obtain the product set of the univariate modes using one-dimensional k-means started from the product set of the quantiles and then use hierarchical clustering to get representatives from the K most widely separated groups of local modes. The observations are then classified and initializing parameters obtained. The goal behind this strategy is to insure against possible masking effects, but also to use projections to drive us to good initializing values when projections are more helpful.

Our next comment pertains to invariance. The specific implementation above is invariant to both rotation and translation. but not to arbitrary linear transformations on the data. The method is also not scale-invariant in the reduced domains,

though the standardizing operations on the original domain makes it so for those computations. In general, we believe that our approach will provide initializing values when the affinely transformed data preserves the grouping – it is unclear whether clustering itself makes a lot of sense in other contexts. In general, if invariance to affine transformations is an important consideration, one may consider other approaches to finding the candidate set of local modes, such as the bump-hunting of Friedman and Fisher [37].

Another comment pertains to the choice of c_k, which should be such that more local univariate candidates are chosen for lower-dimensional projections of the data, while at the same time ensuring that the number does not grow very rapidly. Our experiments use $c_k = \lceil k + 1 \rceil$, though other choices for c_k (such as $\lceil \log k + 1 \rceil$), and thus k_j, satisfying the general philosophy of (2) could also be considered. An additional possibility, which we do not implement in our experiments in this paper, is to try out different candidate sets of c_k's and choose the initializer with lowest $\mid SSP_W \mid$ (for k-means) or highest log-likelihood for EM-clustering. Finally, although also not implemented in the software for this paper, this approach can be readily parallelized for computational speed and efficiency.

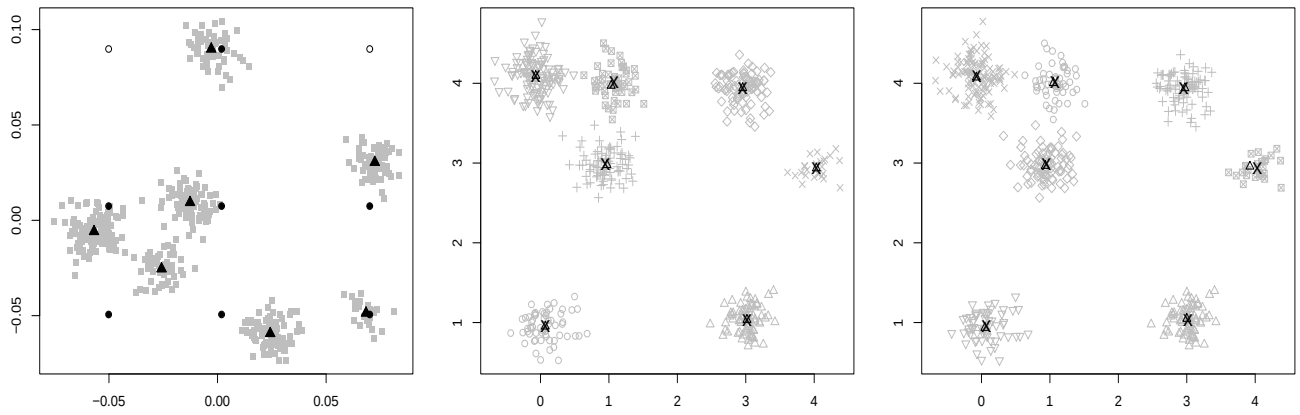
III. Experimental Evaluations

The performance of the initialization scheme is evaluated in a series of simulation experiments, starting with the two bivariate examples introduced in Section 1. We then follow through with a range of simulations on a variety of test case scenarios in many dimensions and degrees of separation between clusters. We evaluate the performance of our methods numerically via the adjusted Rand [38] measure of similarity between two partitions. (One criticism of the Rand [39] measure is that it is more likely to be close to 1 and does not weigh disagreements in grouping adequately. The adjusted Rand measure is designed to spread the weight around more equitably so we report only this measure.) Our comparisons are with groupings obtained from the unlikely scenario of running the partition-optimization algorithm initialized with the true parameter values.

Clustering using our suggested initialization strategy is compared with some of the common or better-performing methods used in the literature. In the k-means context, we compare with (1) initialization from the best (in terms of lowest $\mid SSP_W \mid$) of p^2 randomly chosen data-points (*Rnd-KM*), (2) hierarchical clustering (*Hclust-KM*) into k groups using Ward's criterion (experimentally shown to be among the most competitive for a range of k-means initializers [40]), (3) the extension of hierarchical clustering (*Hclust-TW-KM*) proposed by Tseng and Wong [33] and (4) Bradley and Fayyad's [30] approach of choosing the centroid of centers obtained from p^2 random starts (*BF-KM*). For EM, our comparisons are with clusterings obtained using (1) initialization from the best (in terms of highest log-likelihood values) of p^2 valid sets of randomly chosen data-points (*Rnd-EM*), (2) Mclust which uses hierarchical clustering and likelihood gain and (3) initialization using p^2 short EM's initialized from a set of randomly chosen valid data-points (em-EM). We also tried using Fayyad *et al*'s (1998) suggestions to initialize EM, but were very often unable to come up with a valid set of means, and dispersions even after several attempts. We use Ward's and likelihood gain criterion respectively in the hierarchical clustering initialization using k-means and EM, because these criteria are used in the corresponding objective function for which the respective optima are desired. The latter is used by the R package Mclust, and we use this to denote the strategy. The choice of p^2 replications in the cases above corresponds on the average to about 2-3 times the time taken by our suggested approach to find an initializer. Note that k-means is done all through using Hartigan and Wong's [41] implementation. Further, although not a focus of this paper, we also study the proposed initialization scheme in the context of estimating the number of clusters using Marriott's [42] criterion for k-means or the Bayes Information Criterion (BIC) of Schwarz [43], assuming a Gaussian mixture model (Fraley and Raftery [32]. Finally, simulation variability in the performance evaluations is accounted for by replicating each experiment 25 times.

A. Bivariate Examples

Our first bivariate example, also used to demonstrate our methodology, uses 500 observations (Figure 1) from seven bivariate normal clusters with same dispersion (correlation $\rho = 0$, standard deviations $\sigma = 2$ in both dimensions). The clusters were centered at (1.05, 4), (3,1.05), (2.97,3.94), (-0.08,4.07), (0.93,2.96), (0.05, 0.93) and (3.92,2.96) with mixing probabilities of 0.1, 0.16, 0.2, 0.21, 0.17, 0.11 and 0.06 respectively. Figure 3a illustrates the application of our proposed methodology in the projection space for $m^* = 2$. The final centers and classification (Figure 3b) obtained using our suggested initialization strategy are essentially indistinguishable from those obtained upon running the k-means algorithm using the true centers as starting points (Figure 3c). The adjusted Rand ($\mathcal{R}_a$) measure of similarity was 1.0, indicating a perfect match between the two clusterings. While not displayed, the results match mixtures-of-Gaussians model-based clustering done using the EM algorithm and our suggested starting points. Finally, both Marriott's criterion (when using k-means) and BIC (upon using EM) identified seven as the optimal number of clusters. This is also true upon using k-means initialized using *Hclust-KM*. However, *Hclust-TW* ($\mathcal{R}_a = 0.952$), *Rnd-KM* ($\mathcal{R}_a = 0.782$) and *BF-KM* ($\mathcal{R}_a = 0.797$) initialization perform much worse in the context of 7-means. Moreover, in these cases, Marriot's criterion finds 9, 13 and 10 clusters ($\mathcal{R}_a = 0.958, 0.922$ and $R_a = 0.943$, respectively), with the latter often yielding initializers proximate to no data-point and hence not consistent with k-means starting values. In the EM context, Mclust is also perfect whether or not k is known, while *Rnd-EM* and *emEM* both identify 9 optimal clusters with $\mathcal{R}_a = 0.978$ in each case. When the number of clusters is provided, $\mathcal{R}_a$ was 0.818 and 0.794 respectively. For the 25 replicated datasets, when using k-means, our suggested algorithm and *Hclust-KM* were always perfect in both partitioning and estimating



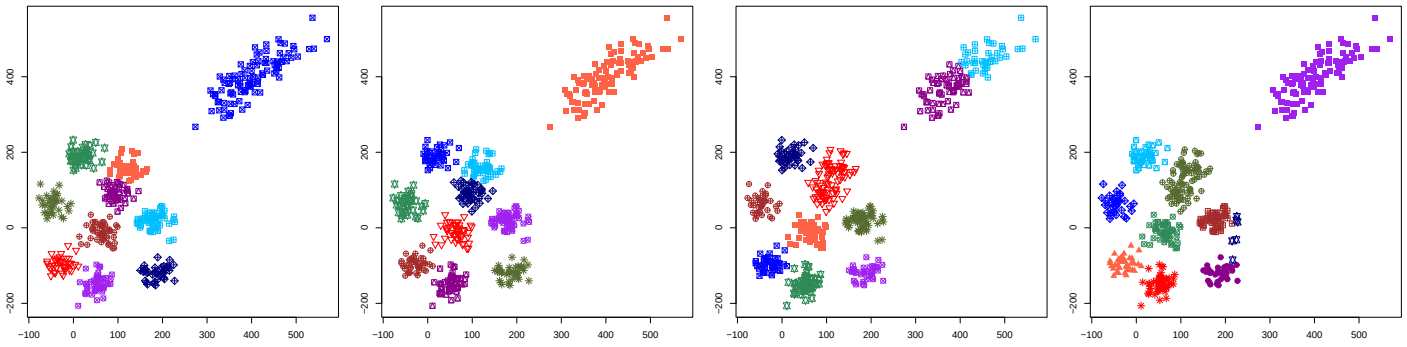
(a) (b) (c)

Fig. 3. Illustration of proposed algorithm. (a) Projection of centered dataset along left singular vectors, with unfilled circles denoting product set of centers of partitions along each projection. Filled circles represent the thinned subset to initialize the k-means algorithm in Step 4, the output of which is represented by unfilled triangles. Filled triangles represent the centers in projection space obtained upon using hierarchical clustering with single-linkage on the centers with unfilled triangles. (b) The starting values (unfilled triangles) and final centers (x) for k-means in original space, together with the derived classification. (c) Final centers (x) and grouping arrived at using k-means initialized by the true centers (unfilled triangles).

k, while *Rnd-KM*, *BF-KM* and *Hclust-TW* were less so, with median adjusted Rand ($\mathcal{R}_{a;0.5}$) values of 0.944, 0.949 and 0.962 for 7-means and with inter-quartile ranges (IQR_a) of 0.011, 0.051 and 0.018 respectively. With k unknown, the corresponding median adjusted Rand values ($\mathcal{R}_{a;0.5,k_{opt}}$), inter-quartile ranges ($IQR_{a;k_{opt}}$) and median number of estimated clusters ($k_{opt;0.5}$) were (0.978, 0.02, 8), (0.982, 0.032, 8) and (0.982, 0.046, 8). With EM, our suggested algorithm and Mclust were essentially perfect ($\mathcal{R}_{a;0.5} = \mathcal{R}_{a;0.5,k_{opt}}$ = 1.0, $IQR_a = IQR_{a;k_{opt}} = 0$, $k_{opt} = 7$) but *Rnd-EM* and *emEM* were less so whether k was given ($\mathcal{R}_{a;0.5}$ = 0.942 and 0.948, IQR_a = 0.066 and 0.040) or estimated ($\mathcal{R}_{a;0.5,k_{opt}}$ = 0.984 and 0.982, $IQR_{a;k_{opt}}$ = 0.016 and 0.0.021, $k_{opt;0.5} = 8$ in both cases).

Our second experiment revisits the case study of Figure 2. Here, there are 400 bivariate normal observations from the larger cluster, centered at (40,36), with equal variances of 4,000 and a correlation coefficient of 0.015. The remaining observations were simulated in equal parts from bivariate normal populations with equal variances in both dimensions of 80, but with different centers, at (250,250) and (300,300) and correlations of 0.075 and -0.05, respectively. Initializing the EM algorithm with our methodology for three clusters correctly identified the grouping and matched exactly that obtained by initializing the algorithm with the true parameters. Expectedly, $\mathcal{R}_a = 1.0$. Further, BIC identified correctly the number of clusters in the data. On the other hand, performance using *Rnd-EM* initialization is substantially worse, whether the number of clusters is known ($\mathcal{R}_a = 0.392$) or unknown (6 clusters identified; $\mathcal{R}_a = 0.417$). The same pattern is repeated with *emEM* regardless of whether the number of clusters is known to be three ($\mathcal{R}_a = 0.356$) or optimally estimated using BIC (5 clusters; $\mathcal{R}_a = 0.422$). Mclust is no better whether the number of clusters is known ($\mathcal{R}_a = 0.356$) or optimally estimated via BIC (5 clusters; $\mathcal{R}_a = 0.4058$). Finally, performance evaluations on the 25 replicated datasets were uneven and reflected Kettenring's (2006) comments on the difficulty of the clustering problem: when the number of clusters were given, our suggested algorithm had $\mathcal{R}_{a;0.5} = 1$, $\mathcal{I}QR_a = 0.588$, Mclust had $\mathcal{R}_{a;0.5} = 0.367$, $\mathcal{I}QR_a = 0.468$, *Rnd-EM* had $\mathcal{R}_{a;0.5} = 0.374$, $\mathcal{I}QR_a = 0.015$ and *emEM* had $\mathcal{R}_{a;0.5} = 0.386$, $\mathcal{I}QR_a = 0.071$. When the number of clusters is required to be estimated from the data, the measures for ($\mathcal{R}_{a;0.5,k_{opt}}, IQR_{a;k_{opt}}, k_{opt}$) are (1.0, 0.491, 7) for our suggested strategy, (0.421, 0.06, 5) with Mclust, (0.454, 0.563, 4) with *Rnd-EM* and (0.576, 0.582, 4) when using *emEM*.

Our final bivariate experimental setup (Figure 4), suggested very kindly by a referee, is in some ways the antithesis of the previous one. Here, there are nine small (each with mixing proportion 0.09) but overlapping clusters centered at (62, -15), (18, 189), (186, -115), (95, 86), (-25, -103), (-39, 67), (125, 161), (176, 18) and (50, -150), and with common dispersions ($\rho = 0.05; \sigma = 20$ in both dimensions). A larger elongated cluster is centered at (400, 400) and has dispersion given by $\sigma = \sqrt{3000}$ in both dimensions and $\rho = 0.8$. It is readily obvious that this is not an easy problem to cluster, especially when the number of clusters is not known. Performance of k-means on this dataset of 500 observations is unsurprisingly poor, even when initialized using the true centers, with the algorithm thrown off-keel by the point (273.824, 267.431). However, clustering via EM initialized using our strategy (Figure 4a) was indistinguishable ($\mathcal{R}_a = 1.0$) with that started from the true parameter values (Figure 4b) and performed very well. BIC chose the number of clusters correctly. On the other hand, Mclust had difficulty whether the number of clusters was given ($\mathcal{R}_a = 0.795$) or unknown ($\mathcal{R}_a = 0.780; k_{opt} = 9$). In both cases, the elongated cluster was split. Performance was better using $emEM$ or $Rnd - EM$ when the number



(a) (b) (c) (d)

Fig. 4. Clustering obtained using the EM-algorithm with ten clusters, initialized using the (a) suggested methodology, (b) true parameter values, (c) Mclust and (d) $emEM$.

of clusters was estimated (8 clusters, $\mathcal{R}_a = 0.944$ and 6 clusters, $\mathcal{R}_a = 0.926$ respectively) or given ($\mathcal{R}_a = 0.919$ – see Figure 4d – and 0.914). The stray point (273.824, 267.431) for this dataset is what hurts the performance of the alternative strategies: performance evaluations over 25 replicated datasets are somewhat different. These yielded $(\mathcal{R}_{a;0.5}\mathcal{I}QR_a) = (0.966, 0.013)$ for our suggested algorithm when the $K = 10$ was given. On the same replicates, Mclust reported $\mathcal{R}_{a;0.5} = 0.969$, $\mathcal{I}QR_a = 0.034$ and *Rnd-EM* had $\mathcal{R}_{a;0.5} = 0.955$ and $\mathcal{I}QR_a = 0.02$ and *emEM* had $\mathcal{R}_{a;0.5} = 0.0.949$ and $\mathcal{I}QR_a = 0.048$. Corresponding measures for $(\mathcal{R}_{a;0.5,k_{opt}}, IQR_{a;k_{opt}}, k_{opt})$ with unknown number of clusters were (0.967, 0.024, 10) using our suggested strategy, (0.978, 0.009, 12) using Mclust and (0.967, 0.019, 9) and (0.968, 0.021, 9) with *Rnd-EM* and *emEM*.

The performance of the different initialization strategies above points to their pitfalls. If the true cluster parameters are not close to being represented at the initialization stage itself, there is often very little chance of their recovery. The suggested approach on the other hand seems to do a better job more consistently. We now report comparative performance evaluations in higher dimensions.

B. Experiments in Higher Dimensions

The next suite of experiments detail evaluations on datasets of size 500, 1,000 and 1,500 generated from 7-, 9- and 11-component multivariate normal mixtures in 5-, 7- and 10-dimensional space, respectively. In each case cluster means, mixing proportions, and dispersion matrices were generated randomly but satisfying certain criteria of representation of and separation between clusters. Following Dasgupta [44], two p-variate Gaussian densities $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ are defined to be c-separated if $\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \| \geq c\sqrt{p \max(\lambda_{\max}(\boldsymbol{\Sigma}_i), \lambda_{\max}(\boldsymbol{\Sigma}_j))}$, where $\lambda_{\max}(\boldsymbol{\Sigma})$ denotes the largest eigenvalue of the variance-covariance matrix $\boldsymbol{\Sigma}$. We modify this definition to add equality in the above for at least one pair (i, j). This means that we are insisting on exact-c-separation for at least two of the clusters. According to Dasgupta [44], there is significant overlap between at least two clusters for $c = 0.5$ and $c = 1.0$ and good separation between all clusters for $c > 2.0$. For instance, in the experiment of Figure 1, $c = 0.37$ and indeed, there is considerable overlap between some of the clusters. There is not much overlap between the clusters in Figure 2 for which $c = 3.327$. It is a bit surprising however to note that $c = 2.787$ for the dataset of Figure 4. This last measure points to a major shortcoming in the definition above in that the degree of separation between clusters as defined above depends only on the means and the largest eigenvalues of the cluster dispersions, regardless of their orientation. Further, other factors (such as mixing proportions) beyond separation also play a role in determining the degree of difficulty of clustering: hence clustering difficulty of a dataset is only partially captured by the measure of c-separation between two clusters.

Using R code, available on request, we implement exact-c-separation between groups in our experiments via the additional restriction that there is at least one pair of cluster densities for which equality is very close (at a value of between c and $c+0.005$) to being attained. For each of our experiments, we choose our parameters to satisfy c-values of 0.8, 1.2, 1.5 or 2.0, corresponding to a sliding scale of overlap from the substantial to the very modest. Further, our mixing proportions were stipulated to each be at least 0.056, 0.032 and 0.025 for the 5-, 7- and 10-dimensional experiments, respectively. These minimum values were chosen so that there were almost surely at least p observations in each of the clusters.

Table I summarizes the performance of k-means clustering over the replicated datasets initialized using the different strategies. When k is known, *Hclust-KM* is the top performer for all experiments: however, our suggested algorithm is very close. The performance of the other competing algorithms is mixed. While *BF-KM* does well in smaller dimensions and higher separation, that advantage evaporates in higher dimensions. Among the weakest performed for all sets of experiments are *Hclust-TW-KM* and *Rnd-KM*. Interestingly with unknown k, the suggested algorithm outperforms *Hclust-KM* and all others in terms of the grouping obtained. This is encouraging because most often (and as in the expression microarray application), the true k is unknown but interest still is on the grouping obtained. The results of these simulation experiments provide confidence that k-means initialized using our suggested algorithm yields groupings closest to the true.

The performance of Gaussian clustering using EM initialized with the competing strategies is summarized in Table II.

TABLE I

SUMMARIZED ADJUSTED RAND SIMILARITY MEASURES ($\mathcal{R}_a$) OF GROUPINGS OBTAINED OVER 25 REPLICATIONS FOR EACH SETTING WITH k-MEANS USING DIFFERENT INITIALIZATION STRATEGIES (**starts**). SIMILARITIES ARE CALCULATED FROM GROUPINGS OBTAINED BY RUNNING k-MEANS FROM THE TRUE PARAMETER VALUES. THE SUMMARY STATISTICS REPRESENTED ARE THE MEDIAN $\mathcal{R}_a$ WHEN k IS KNOWN ($\mathcal{R}_{a;0.5}$) OR ESTIMATED ($\mathcal{R}_{a,k_{opt};0.5}$) AND THE CORRESPONDING INTERQUARTILE RANGES ($\mathcal{IQR}_a$ AND $\mathcal{IQR}_{a,k_{opt}}$). FOR THE LATTER CASE, THE MEDIAN OPTIMAL NUMBER OF CLUSTERS ($k_{opt;0.5}$) ESTIMATED IS ALSO PROVIDED. FINALLY, $\#tops$ ALSO REPRESENTS THE NUMBER OF REPLICATIONS (OUT OF 25) FOR WHICH THE GIVEN INITIALIZATION STRATEGY DID AS WELL AS THE BEST STRATEGY WITH k KNOWN AND UNKNOWN (IN ITALICS).

Starts	**Statistic**	$p=5, k=7, n=500$				$p=7, k=9, n=1,000$				$p=10, k=11. n=1,500$			
		separation c				**separation** c				**separation** c			
		0.8	1.2	1.5	2.0	0.8	1.2	1.5	2.0	0.8	1.2	1.5	2.0
Rnd-KM	$\mathcal{R}_{a;0.5}$	0.928	0.952	0.930	0.958	0.940	0.948	0.934	0.938	0.953	0.960	0.965	0.947
	$\mathcal{IQR}_{a;0.5}$	0.059	0.052	0.064	0.092	0.053	0.033	0.052	0.039	0.042	0.023	0.038	0.043
	$\mathcal{R}_{a,k_{opt};0.5}$	0.931	0.948	0.957	0.952	0.939	0.953	0.942	0.962	0.964	0.976	0.966	0.964
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.033	0.046	0.036	0.052	0.040	0.025	0.048	0.035	0.033	0.028	0.032	0.022
	$k_{opt;0.5}$	10	9	10	9	12	11	12	11	13	12	12	12
	#tops	4,*8*	3,*5*	3,*7*	8,*3*	1,*3*	3,*4*	0,*1*	1,*2*	2,*7*	2,*5*	0,*4*	0,*3*
BF-KM	$\mathcal{R}_{a;0.5}$	0.950	1.0	1.0	1.0	0.961	0.970	0.983	0.978	0.976	0.988	0.985	0.978
	$\mathcal{IQR}_{a;0.5}$	0.059	0.026	0	0.058	0.036	0.039	0.046	0.040	0.038	0.018	0.024	0.048
	$\mathcal{R}_{a,k_{opt};0.5}$	0.931	0.936	0.946	0.925	0.937	0.945	0.944	0.963	0.971	0.980	0.982	0.977
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.015	0.044	0.032	0.042	0.028	0.031	0.037	0.034	0.024	0.023	0.017	0.026
	$k_{opt;0.5}$	10	10	9	10	12	12	12	11	13	13	13	12
	#tops	9,*1*	16,*3*	19,*3*	17,*1*	6,*2*	5,*1*	8,*1*	7,*5*	8,*6*	6,*3*	7,*4*	3,*7*
Hclust-KM	$\mathcal{R}_{a;0.5}$	0.966	1.0	1.0	1.0	0.998	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	$\mathcal{IQR}_{a;0.5}$	0.033	0	0	0	0.040	0	0	0	0	0	0	0
	$\mathcal{R}_{a,k_{opt};0.5}$	0.927	0.924	0.932	0.924	0.936	0.938	0.930	0.947	0.964	0.967	0.965	0.963
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.021	0.028	0.025	0.021	0.029	0.019	0.029	0.018	0.013	0.016	0.015	0.013
	$k_{opt;0.5}$	10	10	10	10	12	12	12	12	13	13	13	13
	#tops	17,*1*	24,*3*	25,*1*	25,*0*	13,*0*	23,*0*	25,*0*	25,*0*	19,*0*	25,*0*	25,*0*	25,*0*
Hclust-TW-KM	$\mathcal{R}_{a;0.5}$	0.931	0.943	0.949	0.933	0.934	0.933	0.923	0.924	0.946	0.950	0.942	0.939
	$\mathcal{IQR}_{a;0.5}$	0.028	0.047	0.041	0.067	0.051	0.044	0.037	0.046	0.020	0.043	0.047	0.032
	$\mathcal{R}_{a,k_{opt};0.5}$	0.925	0.932	0.945	0.927	0.940	0.940	0.924	0.949	0.959	0.960	0.963	0.951
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.020	0.041	0.039	0.047	0.024	0.029	0.041	0.050	0.033	0.032	0.043	0.036
	$k_{opt;0.5}$	10	10	9	10	12	11	11	11	13	13	12	12
	#tops	2,*3*	3,*1*	3,*3*	1,*4*	0,*4*	1,*1*	0,*0*	0,*2*	0,*0*	1,*0*	1,*1*	0,*1*
suggested	$\mathcal{R}_{a;0.5}$	0.960	1.0	1.0	1.0	0.968	1.0	1.0	1.0	0.978	1.0	1.0	1.0
	$\mathcal{IQR}_{a;0.5}$	0.084	0.020	0	0	0.050	0.006	0	0	0.027	0	0	0
	$\mathcal{R}_{a,k_{opt};0.5}$	0.941	0.966	0.989	0.991	0.961	0.998	1.0	0.986	0.973	0.994	0.992	0.992
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.044	0.041	0.053	0.033	0.042	0.019	0.012	0.019	0.032	0.013	0.016	0.014
	$k_{opt;0.5}$	10	9	8	8	12	10	9	10	13	13	12	12
	#tops	8,*13*	18,*16*	23,*14*	25,*18*	10,*16*	18,*20*	24,*24*	25,*16*	4,*12*	20,*18*	25,*17*	25,*14*

Interestingly, Mclust is the best performer for cases with substantial overlap, but its performance surprisingly degrades with increased separation. The suggested algorithm is however a consistently strong performer and shows excellent performance with higher degrees of separation. The slightly worse performance of our approach in the case of substantial overlap can be explained by the fact that it is based on locating multiple local modes in the dataset: when there is very little separation, it is harder for it to correctly locate their widely separated representatives. Note that, as mentioned in Section I-B, *Rnd-EM* and *emEM* have similar performance in almost all cases. Further, when the number of clusters is not known, EM initialized by our approach and in conjunction with BIC does an excellent job in identifying the number of clusters and, more importantly, the correct grouping. This is of course, a consequence of the benefits of a good partitioning, which flows from the good performance of our initialization strategy in these experiments.

C. Protein Localization Data

Our final experiment explores applicability of clustering on a multivariate protein localization dataset for which class information is available. This dataset, publicly available from the University of California Irvine's Machine Learning Repository [45] concerns identification of protein localization sites for the *E. coli* bacteria, an important early step for finding remedies [46]. Although information is available on 336 protein sequences from eight sites, we only use data on 324 sequences from five sites as in Maitra [47], because the others have too little information to perform (supervised)

TABLE II

SUMMARIZED ADJUSTED RAND SIMILARITY MEASURES ($\mathcal{R}_a$) OF GROUPINGS OBTAINED OVER 25 REPLICATIONS FOR EACH SETTING WITH GAUSSIAN CLUSTERING USING EM WITH DIFFERENT INITIALIZATION STRATEGIES (**starts**). SIMILARITIES ARE CALCULATED FROM GROUPINGS OBTAINED BY RUNNING EM FROM THE TRUE PARAMETER VALUES. THE SUMMARY STATISTICS REPRESENTED ARE AS IN TABLE I.

Starts	**Statistic**	$p=5, k=7, n=500$				$p=7, k=9, n=1,000$				$p=10, k=11. n=1,500$			
		separation c				**separation** c				**separation** c			
		0.8	1.2	1.5	2.0	0.8	1.2	1.5	2.0	0.8	1.2	1.5	2.0
Rnd-EM	$\mathcal{R}_{a;0.5}$	0.949	0.941	0.933	0.928	0.965	0.946	0.946	0.943	0.962	0.956	0.959	0.942
	$\mathcal{IQR}_{a;0.5}$	0.060	0.059	0.087	0.064	0.052	0.060	0.036	0.064	0.062	0.042	0.067	0.059
	$\mathcal{R}_{a,k_{opt};0.5}$	0.984	0.992	0.998	0.994	0.983	0.991	0.992	0.994	0.990	0.994	0.994	0.988
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.024	0.014	0.023	0.016	0.026	0.015	0.017	0.027	0.013	0.012	0.012	0.013
	$k_{opt;0.5}$	7	8	8	8	9	10	11	10	11	12	12	12
	#tops	5,*6*	1,*3*	3,*10*	3,*10*	3,*3*	1,*4*	3,*4*	0,*6*	1,*3*	1,*3*	0,*5*	0,*2*
emEM	$\mathcal{R}_{a;0.5}$	0.971	0.941	0.943	0.940	0.953	0.944	0.945	0.941	0.959	0.953	0.953	0.953
	$\mathcal{IQR}_{a;0.5}$	0.053	0.076	0.046	0.067	0.042	0.047	0.082	0.056	0.070	0.045	0.047	0.047
	$\mathcal{R}_{a,k_{opt};0.5}$	0.993	0.978	0.992	0.990	0.984	0.994	0.992	0.992	0.994	0.994	0.990	0.992
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.016	0.033	0.021	0.029	0.025	0.021	0.015	0.017	0.007	0.010	0.011	0.012
	$k_{opt;0.5}$	7	8	8	8	10	10	10	10	12	12	12	12
	#tops	8,*7*	1,*6*	3,*12*	2,*7*	2,*6*	1,*7*	1,*3*	1,*6*	3,*7*	1,*5*	0,*2*	0,*2*
Mclust	$\mathcal{R}_{a;0.5}$	1.0	1.0	0.980	0.982	1.0	0.992	0.959	0.963	1.0	0.995	0.984	0.973
	$\mathcal{IQR}_{a;0.5}$	0.037	0.017	0.043	0.055	0.041	0.032	0.073	1.183	0.02	0.048	1.135	1.150
	$\mathcal{R}_{a,k_{opt};0.5}$	1.0	1.0	0.992	0.982	1.0	0.994	0.994	0.986	1.0	0.997	0.993	0.989
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.012	0	0.028	0.030	0.015	0.020	0.040	0.015	0.002	0.010	0.015	0.017
	$k_{opt;0.5}$	7	7	8	7	3	3	9	8	11	11	11	10
	#tops	18,*15*	17,*24*	11,*11*	12,*12*	16,*15*	13,*12*	5,*5*	4,*4*	20,*19*	11,*10*	6,*6*	4,*4*
suggested	$\mathcal{R}_{a;0.5}$	0.940	1.0	1.0	1.0	0.960	1.0	1.0	1.0	0.948	1.0	1.0	1.0
	$\mathcal{IQR}_{a;0.5}$	0.073	0.046	0	0	0.066	0.002	0	0	0.224	0.008	0	0
	$\mathcal{R}_{a,k_{opt};0.5}$	0.987	1.0	1.0	1.0	0.978	1.0	1.0	1.0	0.975	1.0	1.0	1.0
	$\mathcal{IQR}_{a,k_{opt};0.5}$	0.025	0.002	0	0	0.023	0.004	0	0	0.091	0.004	0	0
	$k_{opt;0.5}$	8	7	7	7	9	9	9	9	11	11	11	11
	#tops	7,*8*	16,*17*	24,*24*	25,*25*	6,*6*	17,*18*	24,*24*	25,*25*	4,*3*	18,*18*	25,*25*	25,*25*

classification via quadratic discriminant analysis (QDA), much less (unsupervised) clustering using Gaussian-mixture assumptions. (This dataset was shown in Maitra [47] to perform considerably better using QDA rather than linear discriminant analysis (LDA) so clustering using a k-means algorithm was not considered.) See Horton and Nakai [48] and the references therein for details on the attributes and the protein localization sites.

QDA on this dataset provides us with a correct classification rate of 90.74%, with $\mathcal{R}_a = 0.942$ vis-a-vis the site-grouping. The measure is obtained using a supervised learning method under multi-Gaussian-mixtures assumptions and so may be regarded, in some sense, as an upper benchmark for the performance of any clustering algorithm applied on this dataset. Using BIC with EM initialized using *Rnd-EM* and *emEM* identified 4 and 8 optimal locations, but the obtained grouping was poor with $\mathcal{R}_a$ equal to 0.511 and 0.489, respectively. Initializing EM via our suggested methodology, determined five groups as optimal for the dataset. The corresponding clustering had $\mathcal{R}_a = 0.914$. However, Mclust did somewhat better, choosing 5 clusters ($\mathcal{R}_a = 0.927$). Despite being slightly out-performed by Mclust, it is clear that using mixture-of-Gaussians EM-clustering with a good initialization is captures the classification and is an appropriate tool to use in such experiments.

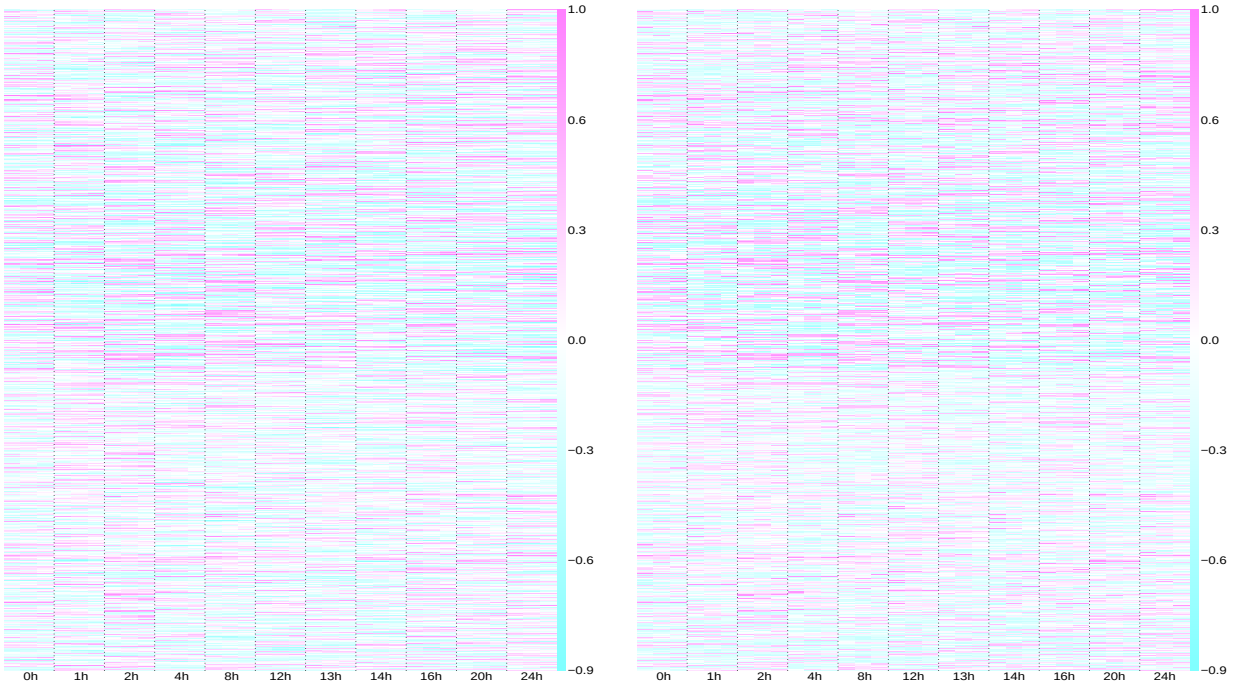
The above experiments indicate excellent performance of partition-optimization clustering approaches when initialized with the algorithm outlined in this paper. Further, clustering in the context of unknown number of groups perform very well using our approach. We next apply the above methodology to the microarray and mercury release datasets.

IV. APPLICATION TO STARCH AND MERCURY RELEASE DATASETS

A. Grouping Similar-Acting Genes in the Diurnal Starch Cycle

As described in Section I-A.1, the dataset is on expression levels of 22,810 genes collected over eleven time-points. Interest centers only on those genes that show activity during the stages of the day-night cycle, so our first step was to use RCBD analysis to identify such genes. Controlling for the expected false discovery rate (FDR) of Benjamini and Hochberg [49] at $q = 0.05$ identified 4,513 most significantly active genes. We analyzed this reduced set of genes in order to understand their role in the diurnal starch cycle.

Genes were clustered using the expression data on the mean of the two replicates. Each replicate was also clustered separately to get an assessment of the variability in our clustering procedure and then compared with the result



(a) (b)

Fig. 5. Heatmap of standardized expression data and class centers of active genes that correspond to starch synthesis and degradation in the *Arabidopsis* diurnal cycle. To facilitate comparison and variability assessment, each time-period is displayed using three measurements representing (a) the two replicates with their mean sandwiched between them and (b) the centers of the separately clustered replicates, with those of the mean sandwiched between them.

obtained previously. As explained earlier, data on the means and the replicates were separately centered and sphered in the temporal dimension to provide observations that are on the eleven-dimensional zero-centered unit sphere, restricted to be orthogonal to the eleven-dimensional unit vector. Figure 5a provides a heatmap display of the standardized microarray gene expression data used in our analysis. In order to portray variability in measurement, the observed replicated data at each time-point are displayed along with the mean. The display indicates moderate variability – this is not entirely surprising, given that these are the most significant genes in the entire dataset.

The initialization algorithm suggested in this paper was used to set up initial seeds for the k-means algorithm, for given numbers of clusters, with Marriott's criterion used to choose the optimal number. Note that resampling approaches such as the Gap statistic [50] are not directly applicable here because of the constraints brought in by the standardization of the dataset.) Further, SSP_W is singular because of the centering, so we use the (11, 11) minor as a surrogate for $\mid SSP_W \mid$. (The (11, 11) minor is the determinant of a square matrix obtained after eliminating its 11th row and the 11th column.) For the two replication datasets, Marriott's criterion identified 21 and 22 as the optimal number of clusters, while partitioning with 21 clusters was optimal for the mean expression. The slight difference in the number of groups is indicative of the variability in the clusterings. Figure 5b is a corresponding heatmap display of Figure 5a of the mean values of the cluster to which each gene is assigned by the k-means algorithm. The k cluster means have been standardized to lie on the eleven-dimensional zero-centered unit sphere. The figure indicates some variability in the assignments, even though many of the assigned group centers are virtually indistinguishable from the others at the same time-point. However, it is encouraging to note that where there is variability, the values of the centers assigned on the basis of clustering the means are intermediate to the ones assigned on the basis of the individual replications. Finally, we also performed pairwise comparisons on the three clusterings: the two groupings on the replicated samples reported $\mathcal{R}_a = 0.925$. Thus there is some discrepancy in the two groupings. Comparing the grouping provided by clustering the mean gene expression data to the two individual replicates provided similar values of 0.932 for $\mathcal{R}_a$. These reported measures help quantify, in some sense, the variability in the derived groupings.

In the sequel, we discuss briefly the groupings obtained upon partitioning the data on the replicated expression levels $\mathcal{D}_1$, $\mathcal{D}_2$ and their means $\mathcal{M}$. Among the genes mentioned in Section I-A.1, we note that PGM1, STS4, SBE3, ISA3 and DPE1 are identified as having similar activity patterns in the diurnal cycle, whether in clustering $\mathcal{D}_1$, $\mathcal{D}_2$ or $\mathcal{M}$. These genes are among those that encode phosphoglucomutase, starch synthase IV, starch branching enzyme III, starch debranching enzyme - isoamylase II and glucanotransferase. Further genes encoding both plastidial (PHS1) and cytostolic glucan phosphorylase (PHS2) are also identified as acting together in all three groupings. The clustering of both replicates identifies them as acting together with the previous group of

genes: however, the grouping based on $\mathcal{M}$ identifies these genes as acting together in a separate second group that is a bit more distinctive than the first. The grouping for glucan water dikinase 1 (GWD1 and SEX1) is a bit less certain: it agrees with the first group of genes in the clustering of $\mathcal{D}_2$ and $\mathcal{M}$, but with the second grouping in the partitioning of the replicates. Genes encoding starch synthase I (STS1) are identified as acting together with the second group in the clustering of the two replicates but not of $\mathcal{M}$. Similar is the case for ADP-Glc phosphorylase (large subunit 2) (APL2) and transglucosidase (DPE2) which are identified as acting in concert in all partitionings, and also in concert with the second set of genes above in partitionings of both $\mathcal{D}_1$ and $\mathcal{M}$ but not of $\mathcal{D}_2$. Further, genes encoding starch debranching enzyme: isoamylase I (ISA1) share the same partition as APL2 and DPE2 for $\mathcal{D}_2$ and $\mathcal{M}$ but not for $\mathcal{D}_1$. Further, among those genes in Table 1 of Smith *et al* [21] whose assignment is not supported by experimental evidence, glucan water dikinase-like 3 (GWD3) acts in conjunction with the first group mentioned above in all three partitionings, and β-amylase 9 (BAM9 and BMY3) is also identified in all clusterings to act together with β-amylase 9 (BAM3, BMY8 and ctBMY). There is also some evidence that α-glucosidase-like 2 (AGL2) acts together and in conjunction with the first group above, being in the same partitions of $\mathcal{D}_1$ and $\mathcal{M}$, but not of $\mathcal{D}_2$. Similar is the case for β-amylase 6 (BAM6) which is identified as acting in concert with the gene encoding starch debranching enzyme: isoamylase II (ISA2 and DBE1) from clusterings of $\mathcal{D}_1$ and $\mathcal{D}_2$ but not $\mathcal{M}$. Also, α-glucosidase-like 5 (AGL5) is identifying as sharing activity patterns with ISA1, APL2 and DPE2 in case of both $\mathcal{D}_2$ and $\mathcal{M}$ but not $\mathcal{D}_1$. These derived partitions can be further analyzed for potentially deepening understanding into the complex process of starch synthesis and breakdown over a twenty-four-hour period. The fact that there is variability in some of these groupings also indicates that there is need, in some cases, for more experiments and further analysis. In any case, it may be noted that the above results are dependent on the choice of the most significant genes (by setting the choice of the expected FDR at 5% above). Indeed, PGI1 and several other genes mentioned in Smith *et al* [21] are not in the reduced set. An alternative may be to use the tight clustering approach for clustering in the presence of scatter, using perhaps a refinement (or a possible refinement thereabout) of Tseng and Wong [33] on the entire dataset. However, we note that even then, initialization remains an issue, since their suggestion of using hierarchical clustering to obtain starting values is not practical on a dataset with 22,810 genes.

B. Profiling Industrial Facilities Based on Mercury Releases

Given the severely skewed nature of the release data, we modeled a transformation $\boldsymbol{Z} = \psi(\boldsymbol{X})$ of the reported releases from each facility $\boldsymbol{X}$ in terms of $f(\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\phi(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian density with mean $\boldsymbol{\mu}_k$ and common dispersion $\boldsymbol{\Sigma}$ evaluated at $\boldsymbol{z}$, $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)$ and $Z_i = \log(1 + \log(1 + X_i))$. Marginally, the density reduces to a mixture of distributions similar to the loglognormal introduced by Lin [51]. We adopt this distribution based on descriptive analyses and its inherent simplicity. Even then, there are some very outlying observations, so we also postulate a homogeneous dispersion assumption on the *transformed data*. As a result, we can identify both groupings and extreme observations in one stroke.

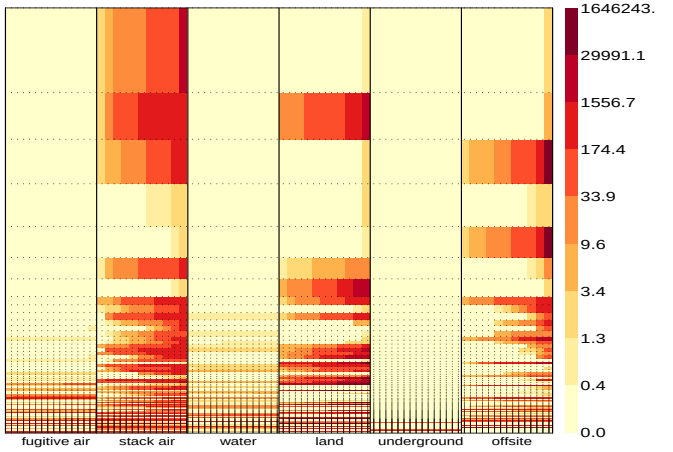


Fig. 6. Distribution of members in each identified group from largest (top) to smallest (bottom), with area of the displayed rectangles proportional to group size. For each coordinate, the imaged intensities represent, on a common loglog scale, the deciles of the marginal release data for each group and release type.

For different numbers of clusters, we initialized the EM algorithm using our suggested methodology and iterated till convergence. BIC identified 113 optimal clusters, and the corresponding grouping had classes ranging in size from a singleton to 281 (20.01%) reporting facilities. A substantial number of extreme observations are identified. All 113 classes, and the marginal distributions of their memberships, in terms of their deciles, are displayed in Figure 6. In order to be able to portray the quantiles, the intensities are mapped on a shifted loglog scale, corresponding to the ψ-transformation above. The display indicates widely divergent distribution of release characteristics for the different groups and visually provides confidence in the clustering obtained. About 48 of these groups are singletons, representing records that are far unlike any other. The 17 largest groups are enough to account for at least 80% of the data. While a study of outlying and very dissimilar release profiles may be of interest, our focus was on characterizing the many different aspects of these releases. We now describe the nature and composition of the five largest groups accounting for just under 59% of all facilities.

The largest group had air emissions – like most others almost entirely from the stack – reporting a median ($\tilde{\mu} = 24.55$ lbs.) and a mean ($\mu = 88$ lbs.). Only a few facilities in this group disposed very small amounts of mercury via land or offsite. Seventy-six, or 51% of all facilities manufacturing stone, clay, glass and concrete products form the largest bloc in this group, followed by electrical services facilities (66 reports) and chemicals and petroleum-refining facilities (41 and 21 reports, respectively). Given the composition of the facilities in this group, it is not very surprising that Texas (31) accounts

for the largest number, followed by Pennsylvania (22) and a number of midwestern and southern states.

The second largest group was characterized largely by facilities that report substantially large amounts of (stack) air emissions ($\mu = 249.17$ lbs., $\tilde{\mu} = 177$ lbs.) and large releases to land ($\mu = 186.11$ lbs., $\tilde{\mu} = 81.88$ lbs.). Over 76.62% (118) of the reports were from facilities in the electrical services industry: in particular, all but one of these were from fossil-combusting facilities. Further, most of these facilities were from the midwestern and southern states. The third largest group consists of facilities that report moderate stack air releases ($\mu = 79.26$ lbs., $\tilde{\mu} = 31$ lbs.) and offsite disposal ($\mu = 4488.18$ lbs., $\tilde{\mu} = 19.46$ lbs.). About (56%) of the 148 facilities in this group are from the electric generation services, while 14 are from petroleum refining and allied industries.

The fourth largest group can be categorized as the "cleanest", characterized by low stack air emissions ($\mu = 0.54$ lbs., $\tilde{\mu} = 0.15$ lbs.) and low land and offsite disposals. This group had very similar numbers of facilities from the major industries, with no clear pattern. Out of the 19 facilities from the fossils-combusting electricity generating facilities, 12 were from Puerto Rico, California and Massachusetts, with the remaining seven scattered over several states.

The fifth largest group of reports are characterized by large quantities of offsite mercury disposals ($\mu = 1,253.55$ lbs., $\tilde{\mu} = 30$ lbs.) coupled with low releases to land or to air through the stack. Perhaps unsurprisingly, refuse systems formed the largest category here. These facilities are all in an industry category called "Business Services". However, this group also had a few facilities manufacturing chemicals and allied products as well as coal- and ol-fired electrical utilities.

Analysis of the major groups in the dataset suggests that coal- and gas-powered electric services facilities dominated groups characterized by high volumes of mercury releases to air. While some preponderance is expected, given that they filed almost 35% of all mercury release reports in 2000, their presence in these groups was disproportionately high. Further down, the seventh largest group had releases, almost all to land ($\mu = 843.3$ lbs., $\tilde{\mu} = 74.25$ lbs), and with the majority composed by mining facilities. In both cases, some facilities of the same types were also in groups with low releases, indicating that local factors may influence their release behaviors, and should be studied further in formulating appropriate policies for public health.

The use of our methodology to initialize partition-optimization algorithms in the two areas above is promising. While k-means and EM-clustering algorithms could be used here using any initialization strategy, the performance evaluations of Section III provide some confidence in the groupings picked up in these two applications. At the same time, many of the classes identified in each of these applications are intuitive and lend further strength to conclusions on inferred group properties and behaviors.

V. Discussion

The main contribution of this paper is the development of a computationally feasible deterministic algorithm for initializing greedy partition-optimization algorithms. Preliminary results on an extensive suite of test experiments and a classification dataset are very promising. We also provide detailed simulation studies to evaluate performance of several commonly-used initializers. Our evaluations are over a vast range of cases, from scenarios ranging from substantially-overlapping to well-separated groups with differing inclusion probabilities. It is very unlikely that any algorithm can uniformly be the best among all initializers in all cases: indeed, the performance evaluations in this paper strongly suggest otherwise. However, they also indicate that the suggested approach can be added to the toolbox of good candidates for use as an initializer. Implements from this toolbox can be used on clustering problems and the best chosen from amongst them. We have also developed ISO/ANSI-compliant C software for our algorithm. Our algorithm is computer-intensive, but practical, and took no more than a few seconds to suggest an initializer, even on a moderately-sized dataset. Further, as a consequence of the excellent performance in partitioning the experimental datasets, algorithms initialized via our suggested methodology performed very well in the difficult task of clustering when the number of groups was unknown. We also applied our algorithm with k-means to analyze microarray gene expression data on the diurnal starch cycle in Arabidopsis leaves. We also used it, together with a model-based EM clustering approach to group different industrial facilities that reported releases of mercury or mercury compounds in the year 2000. While the goal behind the first application is to obtain a better understanding of the pathways and processes involved in starch synthesis and breakdown, the second application was geared towards characterizing the different kinds of mercury releases in order to effectively frame policies to improve public health. In each case, the results arrived at were both interpretable and provided additional insight into the underlying factors governing the two datasets. Clustering is widely used in several applications in the health and biological sciences, and partition-optimization algorithms are often employed to achieve the task. There is therefore need for effective initialization strategies for these algorithms and this paper suggests methodology for this purpose. At the same time, there is need for adaptation of methodology for cases where the distance metric for clustering can not be reduced to a transformation of Euclidean distance, or when the underlying mixing distributions are not Gaussian. Another area not addressed is the issue of when we have a large number of coordinates, which renders the specific implementation of our algorithm using SVD computationally prohibitive. Thus, while our suggested methodology can be regarded as an important statistical contribution towards strategies for effective initialization, there are issues that require further attention.

Acknowledgment

The author thanks Dan Nettleton for his patient explanations of the many different aspects of microarray data collection and analysis. Thanks are also due to the editor and two anonymous referees for helpful suggestions and for comments that considerably improved this article.

REFERENCES

[1] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, pp. 803–821, 1993.

[2] G. Brossier, "Piece-wise hierarchical clustering," *Journal of Classification*, vol. 7, pp. 197–216, 1990.

[3] G. Celeux and Govaert, "Gaussian parsimonious clustering models," *Computational Statistics and Data Analysis*, vol. 28, pp. 781–93, 1995.

[4] W. F. Eddy, A. Mockus, and S. Oue, "Approximate single linkage cluster analysis of large datasets in high-dimnesional spaces," *Computational Statistics and Data Analysis*, vol. 23, pp. 29–43, 1996.

[5] B. S. Everitt, "A finite mixture model for the clustering of mixed-mode data," *Statistics and Probability Letters*, vol. 6, pp. 305–309, 1988.

[6] I. J. Good, "The clustering of random variables," *Journal of Statistical Computing and Simulation*, vol. 9, pp. 241–248, 1979.

[7] J. Hartigan, "Statistical theory in clustering," *Journal of Classification*, vol. 2, pp. 63–76, 1985.

[8] J. R. Kettenring, "The practice of cluster analysis," *Journal of classification*, vol. 23, pp. 3–30, 2006.

[9] F. Murtagh, *Multi-dimensional clustering algorithms*. Berlin; New York: Springer-Verlag, 1985.

[10] D. B. Ramey, "Nonparametric clustering techniques," in *Encyclopedia of Statistical Science*. New York: Wiley, 1985, vol. 6, pp. 318–319.

[11] B. D. Ripley, "Classification and clustering in spatial and image data," in *Analyzing and Modeling Data and Knowledge*. Berlin; New York: Springer-Verlag, 1991, pp. 85–91.

[12] M. J. Symons, "Clustering criteria and multivariate normal mixtures," *Biometrics*, vol. 37, pp. 35–43, 1981.

[13] R. J. Tibshirani and G. Walther, "Cluster validation by prediction strength," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, pp. 511–528, 2005.

[14] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, vol. 62, pp. 1159–1178, 1967.

[15] A. J. Scott and M. J. Symons, "Clustering methods based on likelihood ratio criteria," *Biometrics*, vol. 27, pp. 387–397, 1971.

[16] W. D. Fisher, "On grouping for maximum homogeneity," *Journal of the American Statistical Association*, vol. 53, pp. 789–798, 1958.

[17] L. Kaufman and P. J. Rousseuw, *Finding Groups in Data*. New York: John Wiley and Sons, Inc., 1990.

[18] D. Steinley, "Local optima in k-means clustering: what you don't know may hurt you," *Psychological Methods*, vol. 8, pp. 294–304, 2003.

[19] S. C. Zeeman, A. Tiessen, E. Pilling, K. L. Kato, A. M. Donald, and A. M. Smith, "Starch synthesis in arabidopsis. granule synthesis, composition, and structure," *Plant Physiology*, vol. 129, pp. 516–529, 2002.

[20] A. M. Smith, S. C. Zeeman, D. Thorneycroft, and S. M. Smith, "Starch mobilization in leaves," *Journal of Experimental Botany*, vol. 54, pp. 577–83, 2003.

[21] S. M. Smith, D. C. Fulton, T. Chia, D. Thorneycroft, A. Chapple, H. Dunstan, C. Hylton, S. C. Zeeman, and A. M. Smith, "Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in arabidopsis leaves," *Plant Physiology*, vol. 136, pp. 2687–2699, 2004.

[22] P. Grandjean, P. Weihe, R. F. White, F. Debes, S. Araki, K. Yokoyama, K. Murata, N. Sörensen, R. Dahl, and P. J. Jorgensen, "Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury," *Neurotoxicology and Teratology*, vol. 19, no. 6, pp. 417–428, 1997.

[23] T. Kjellstrom, P. Kennedy, S. Wallis, and C. Mantell, *Physical and mental development of children with prenatal exposure to mercury from fish. Stage 1: Preliminary tests at age 4*. Sweden: Swedish National Environmental Protection Board, 1986.

[24] N. Sörensen, K. Murata, E. Budtz-Jrgensen, P. Weihe, and P. Grandjean, "Prenatal methylmercury exposure as a cardiovascular risk factor at seven years of age," *Epidemiology*, vol. 10, no. 4, pp. 370–375, 1999.

[25] N. A. of Sciences, *Toxicological effects of methylmercury*. Washington DC: National Academy Press, 2000.

[26] U. S. E. P. Agency, *America's Children and the Environment: Measures of Contaminants, Body Burdens, and Illnesses*. Washington DC: United States Environmental Protection Agency, 2003.

[27] W. F. Fitzgerald, D. R. Engstrom, R. P. Mason, and E. A. Nater, "The case for atmospheric mercury contamination in remote areas," *Environmental Science and Technology*, vol. 32, pp. 1–7, 1998.

[28] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2006, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[29] J. A. Lozano, J. M. P. na, and P. L. naga, "An empirical comparison of four initialization methods for the k-means algorithm," *Pattern Recognition Letters*, vol. 20, pp. 1027–1040, 1999.

[30] P. S. Bradley and U. M. Fayyad, "Refining initial points for K-Means clustering," in *Proc. 15th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 1998, pp. 91–99.

[31] U. Fayyad, C. Reina, and P. Bradley, "Initialization of iterative refinement clustering algorithms," in *Proc. of the 4'th International Conference on Knowledge Discovery and Data Mining*, New York, 1998, pp. 194–198.

[32] C. Fraley and A. E. Raftery, "How many clusters? which cluster method? answers via model-based cluster analysis," *Computer Journal*, vol. 41, pp. 578–588, 1998.

[33] G. C. Tseng and W. H. Wong, "Tight clustering: A resampling-based approach for identifying stable and tight patterns in data," *Biometrics*, vol. 61, pp. 10–16, 2005.

[34] M. B. Al-Daoud, "A new algorithm for cluster initialization," *Transactions on Engineering, Computing and Technology*, vol. V4, pp. 74–76, 2005.

[35] C. Biernacki and G. C. amd G. Govaert, "Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models," *Computational Statistics and Data Analysis*, vol. 413, pp. 561–575, 2003.

[36] J. W. Demmel, *Applied Numerical Linear Algebra*. Philadelphia: SIAM, 1977.

[37] J. Friedman and N. Fisher, "Bump-hunting in high-dimensional data," *Statistics and Computing*, vol. 9, no. 2, pp. 1–20, 1999.

[38] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[39] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[40] D. Steinley and M. J. Brusco, "Initializing k-means batch clustering: A critical evaluation of several techniques," *Journal of Classification*, vol. 24, pp. 99–121, 2007.

[41] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[42] F. H. Marriott, "Practical problems in a method of cluster analysis," *Biometrics*, vol. 27, pp. 501–514, 1971.

[43] G. Schwarz, "Estimating the dimensions of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[44] S. Dasgupta, "Learning mixtures of gaussians," in *Proc. IEEE Symposium on Foundations of Computer Science*, New York, 1999, pp. 633–644.

[45] C. B. D.J. Newman, S. Hettich and C. Merz. (1998) UCI repository of machine learning databases. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[46] K. Nakai and M. Kinehasa, "Expert sytem for predicting protein localization sites in gram-negative bacteria," *PROTEINS: Structure, Function, and Genetics*, vol. 11, pp. 95–110, 1991.

[47] R. Maitra, "A statistical perspective to data mining," *Journal of the Indian Society of Probability and Statistics*, vol. 6, pp. 28–77, 2002.

[48] P. Horton and K. Nakai, "A probablistic classification system for predicting the cellular localization sites of proteins," *Intelligent Systems in Molecular Biology*, pp. 109–115, 1985.

[49] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, pp. 289–300, 1995.

[50] R. J. Tibshirani, G. Walther, and T. J. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *Journal of the Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2003.

[51] S. H. Lin, "Statistical behavior of rain attenuation," *Bell System Technical Journal*, vol. 52, no. 4, pp. 557–581, 1973.