2015 IEEE Symposium Series on Computational Intelligence

Maximum Clusterability Divisive Clustering

David Hofmeyr
Department of Mathematics and Statistics
Lancaster University
Lancaster, LA1 4YF, UK
Email: d.hofmeyr@lancaster.ac.uk

Nicos Pavlidis
Department of Management Science
Lancaster University
Lancaster, LA1 4YX, UK
Email: n.pavlidis@lancaster.ac.uk

***Abstract*—The notion of clusterability is often used to determine how strong the cluster structure within a set of data is, as well as to assess the quality of a clustering model. In multivariate applications, however, the clusterability of a data set can be obscured by irrelevant or noisy features. We study the problem of finding low dimensional projections which maximise the clusterability of a data set. In particular, we seek low dimensional representations of the data which maximise the quality of a binary partition. We use this bi-partitioning recursively to generate high quality clustering models. We illustrate the improvement over standard dimension reduction and clustering techniques, and evaluate our method in experiments on real and simulated data sets.**

I. Introduction

Clustering is one of the fundamental problems in data mining, machine learning and statistics. Clustering deals with finding structure in data by identifying groups of similar points, without any explicit information regarding group associations of any of the data. This has applications in diverse areas from bioinformatics to computer vision to marketing.

The notion of *clusterability* refers to the strength, or conclusiveness of the cluster structure within a given set of data [1]. A variety of measures of clusterability have been proposed in the literature (the readers are directed to [1] for a thorough study). Generally in order to determine clusterability, a clustering model over the given data set is required, and as such determining the clusterability of a data set has the same complexity as finding the optimal clustering model. Furthermore, it has been shown [1] that even determining whether or not clusterability exceeds a given threshold is in some cases NP hard.

In the context of multivariate data analysis, the potential irrelevance of certain features in the data makes inference and knowledge discovery especially challenging, and clustering is no exception. In the clustering context these irrelevant features can significantly obscure the cluster structure present in the data, and so even if an optimal clustering model is available, the clusterability of the data under that model may be misleading. Dimension reduction techniques seek to mitigate the effect of irrelevant features by attempting to identify subspaces which contain the most information from the data. However traditionally, dimension reduction and clustering have been performed in isolation from one another. Popular techniques such as Principal and Independent Component Analysis (PCA, and ICA) have shown good performance in a number of cases [2], however the information retained by these methods might not be relevant to the cluster structure in the data and it is trivial to construct examples in which these methods will fail. More recent approaches have been designed with the notion of cluster separation in mind, and so are able to find low dimensional subspaces which retain information relevant to the clustering objective. In statistics, clusters are often associated with the modes of a probability distribution. Dimension reduction in this context has included maximising the departure from unimodality [4] and minimising the integrated density on a separating hyperplane admitted by a univariate projection [5]. Dimension reduction for spectral clustering [6] seeks to find subspaces which minimise the connectivity of a data set as measured by spectral graph theory. Maximum margin clustering [7], [8] can also be viewed in the context of dimension reduction as the univariate subspace admitting the largest margin hyperplane through the data.

In this paper we propose a combined dimension reduction and clustering algorithm, motivated by recursively finding the univariate subspace which maximises the 2 way clusterability of (subsets of) the data within that subspace. This recursive bi-partitioning results in a hierarchical divisive clustering model. We focus on a common measure of clusterability known as the Variance Ratio, first introduced in [3]. Variance Ratio clusterability is given by the ratio of the between cluster variability to the within cluster variability. If the between cluster variability is large relative to the within cluster variability, then the data set is well clusterable. The variance ratio is closely connected to the K-means objective, in that the optimal K-means solution is that which results in the highest variance ratio clusterability. Let $C_1, ..., C_k$ be a k clustering of a data set X, then the variance ratio for this clustering is given by,

$$VR(C_1, ..., C_k|X) = \frac{\sum_{i=1}^{k} \frac{|C_i|}{|X|}\|\mu_{C_i} - \mu_X\|^2}{\sigma_X^2 - \sum_{i=1}^{k} \frac{|C_i|}{|X|}\|\mu_{C_i} - \mu_X\|^2}, \quad (1)$$

where $\mu_X = \frac{1}{|X|}\sum_{x \in X} x$ and $\sigma_X^2 = \frac{1}{|X|}\sum_{x \in X} \|x - \mu_X\|^2$ are the mean and variance of the points in X respectively. Henceforth we write only $VR(C_1, ..., C_k)$, noting that the data set giving rise to the clusters $C_1, ..., C_k$ will be apparent from the context.

For a data set $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$, the bi-partitioning subproblem associated with our clustering algorithm is given by,

$$\max_{v \in \mathbb{R}^d \setminus \{\mathbf{0}\}, C \subset X} VR\left(\{v^\top x | x \in C\}, \{v^\top x | x \in X \setminus C\}\right). \quad (2)$$

The vector v parameterises the univariate subspace, while the set C determines the 2 way clustering of X into $C, X \setminus C$. One

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.116

IEEE computer society

of the benefits of considering univariate subspaces is that the corresponding optimal clustering can be efficiently computed, as we discuss in Section II. In what follows we will use the following notation. For a univariate data set $R = \{r_1, ..., r_n\}$ we write $R_{(i)}$ for the i-th order statistic of R and $R_{(i):(j)} = \{R_{(i)}, \ldots R_{(j)}\}$ for $i \leq j$. We break ties for order statistics by the order of the original indices in R. For multivariate data set $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ and projection vector $v \in \mathbb{R}^d$ we write $v^\top X$ for the projected data set $\{v^\top x_1, \ldots, v^\top x_n\}$. We use $x^{v,i}$ to denote the element in X which corresponds to the i-th order statistic of $v^\top X$.

The remainder of the paper is organised as follows. Section II details our methodology for projection pursuit for maximum clusterability, and how iterating the resulting bi-partitioning process is used to build clustering models. In Section III we present the results of experiments on simulated and real data sets utilising our proposed method. Finally we give concluding remarks in Section IV.

II. Methodology

In this section we provide details of our methodology for divisive clustering using maximum clusterability projections. We refer to this method as Maximum Clusterability Divisive Clustering (MCDC). Divisive clustering algorithms recursively bi-partition (subsets) of a set of data until a desired number of clusters results. There are two main components to such an algorithm, (i) how to implement a binary division of a given set of data and (ii) given a collection of disjoint data sets (the clusters so far discovered) how to determine which should be partitioned next. The first is addressed in Section II-A, where we formulate the optimisation subproblem, Eq. (2), in the context of projection pursuit. The second presents a challenging problem, and we provide a heuristic argument inspired by analysis of variance in Section II-B. In Section II-C we formally describe the MCDC algorithm and discuss its computational complexity.

A. Maximum Clusterability Binary Partitions

Here we describe how we find optimal projections based on the variance ratio clusterability criterion. We formulate the problem in the context of projection pursuit, for which the projection index, $\Phi(v|X)$, is given by the maximum clusterability over all binary partitions of the projected data set. The variance ratio (2) is unbounded if $\exists v \in \mathbb{R}^d, C \subset X$ s.t. $v^\top C$ and $v^\top (X \setminus C)$ each take on only a single value. For subspace optimisation we therefore consider a slight modification as follows,

$$\max_{v \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \Phi(v|X)$$
$$\Phi(v|X) := \max_{C \subset X} VR'(v^\top C, v^\top (X \setminus C)), \quad (3)$$

where $VR'(\cdot, \cdot)$ is exactly $VR(\cdot, \cdot)$ as before, except that we replace σ_X in the denominator with $\frac{|X|}{|X|-1}\sigma_X$. We note that these are equivalent in that for a fixed data set, X, the same clustering maximises both VR and VR'. In addition, for two data sets X_1, X_2 with $|X_1| = |X_2|$ the more clusterable of the two is the same under both VR and VR'. As a result (3) has the same solution as the original optimisation problem (2).

We show that this projection index, when parameterised using polar coordinates, is Lipschitz continuous and continuously differentiable almost everywhere. Thus the negative of the projection index satisfies the conditions for almost sure convergence to a local minimum using the Gradient Sampling algorithm (GS, [9]). GS is a generalised gradient descent algorithm that works by sampling points in a shrinking radius around the current iterate. The smallest element in the convex hull of the gradients evaluated at the sampled points is an approximate steepest descent direction for the current iterate, which is used as a search direction. When $\mathbf{0}$ lies within this convex hull, then the current iterate is close to an ϵ Clarke-stationary point of the objective and the algorithm terminates. For full details of the gradient sampling algorithm, the reader is directed to [9].

Throughout this section we will use the fact that for a univariate data set $R = \{r_1, ..., r_n\}$, if $C_1, C_2 \subset R$ satisfy $\mu_{C_1}, \mu_{C_2} \geq \mu_R$ then,

$$\begin{aligned} & VR'(C_1, R \setminus C_1) > VR'(C_2, R \setminus C_2) \\ \iff & \frac{1}{\sqrt{|C_1|(|R \setminus C_1|)}} \sum_{r_i \in C_1} r_i > \frac{1}{\sqrt{|C_2|(|R \setminus C_2|)}} \sum_{r_i \in C_2} r_i. \end{aligned} \quad (4)$$

The following lemma establishes that the optimal 2 way clustering of a univariate data set, R, partitions the data above/below an order statistic.

Lemma 1: Let $R = \{r_1, \ldots, r_n\} \subset \mathbb{R}$. Then $\exists i \in \{1, \ldots, n-1\}$ s.t.

$$VR'(R_{(1):(i)}, R_{(i+1):(n)}) = \max_{C \subset R} VR'(C, R \setminus C).$$

Proof: Let $C \subset R$ be such that $\exists r_i, r_j \in C, r_k \in R \setminus C$ with $r_i \leq r_k \leq r_j$. If $\mu_C \geq \mu_R$ then define $C' = (C \setminus \{r_i\}) \cup \{r_k\}$. Then,

$$\frac{1}{\sqrt{|C'|(|R \setminus C'|)}} \sum_{r_l \in C'} r_l \geq \frac{1}{\sqrt{|C|(|R \setminus C|)}} \sum_{r_l \in C} r_l$$
$$\Rightarrow VR'(C', R \setminus C') \geq VR'(C, R \setminus C).$$

If instead we have $\mu_C < \mu_R$ then set $C' = (C \setminus \{r_j\}) \cup \{r_k\}$ and the same inequality holds. This process can be iterated until no triples $r_i, r_j \in C', r_k \notin C'$ with $r_i < r_k < r_j$ exist. At this point $\exists i \in \{1, \ldots, n-1\}$ s.t. $VR'(C', R \setminus C') = VR'(R_{(1):(i)}, R_{(i+1):(n)})$ and $VR'(C', R \setminus C') \geq VR'(C, R \setminus C)$. Since C was arbitrary, this proves the result. ■

The importance of this lemma is that it tells us that in order to determine the optimal partition of a projected data set, $v^\top X$, we need only consider $n-1$ possible partitions, corresponding to the order statistics of $v^\top X$. In the context of projection pursuit, we can therefore define the objective function, $\Phi(v|X)$, as

$$\Phi(v|X) = \max_{i \in \{1,\ldots,n-1\}} VR'_i(v|X) \quad (5)$$
$$VR'_i(v|X) := VR'\left((v^\top X)_{(1):(i)}, (v^\top X)_{(i+1):(n)}\right). \quad (6)$$

Now, for $i \in \{1, \ldots, n-1\}$ let v be such that $(v^\top X)_{(i)} \neq (v^\top X)_{(j)}$ for $j \neq i$. Then $VR'_i(v|X)$ is differentiable at v, and

$$\nabla_v VR_i'(v|X) = \sum_{k=1}^{n} \frac{\alpha_k - \beta_k}{\left(\frac{|X|}{|X|-1}\sigma^2_{v^\top X} - BC_{v^\top X,i}\right)^2} x^{v,k} \tag{7}$$

$$\alpha_k := \begin{cases} \frac{2}{n}\left(\mu_{(v^\top X)_{(1):(i)}} - \mu_{v^\top X}\right)\frac{|X|}{|X|-1}\sigma^2_{v^\top X}, & k \leq i \\ \frac{2}{n}\left(\mu_{(v^\top X)_{(i+1):(n)}} - \mu_{v^\top X}\right)\frac{|X|}{|X|-1}\sigma^2_{v^\top X}, & k > i \end{cases} \tag{8}$$

$$\beta_k := \frac{2}{n-1}\left(v^\top x^{v,k} - \mu_{v^\top X}\right) BC_{v^\top X,i} \tag{9}$$

$$\begin{aligned} BC_{v^\top X,i} := {} & \frac{i}{n}\left(\mu_{(v^\top X)_{(1):(i)}} - \mu_{v^\top X}\right)^2 \\ & + \frac{n-i}{n}\left(\mu_{(v^\top X)_{(i+1):(n)}} - \mu_{v^\top X}\right)^2 . \end{aligned} \tag{10}$$

Furthermore we can see that $\nabla_v VR_i'(v|X)$ is continuous for such v. Therefore, if v is such that $i' = \mathrm{argmax}_{i\in\{1,...,n-1\}} VR_i'(v|X)$ is a singleton and $v^\top X$ contains only unique points, then $\nabla_v \Phi(v|X)$ exists and is continuous at v. The following lemma shows that under a reasonable assumption, such v occur almost everywhere for any continuous sampling distribution.

Lemma 2: Suppose $X = \{x_1, ..., x_n\}$ satisfies the property that for $S, S' \subset X$ with $S \neq S'$ we have

$$\frac{1}{\sqrt{|S|(|X|-|S|)}}\sum_{x_i\in S} x_i \neq \frac{1}{\sqrt{|S'|(|X|-|S'|)}}\sum_{x_i\in S'} x_i. \tag{11}$$

Then the function $\Phi(v|X)$ is continuously differentiable a.e. v for any continuous sampling distribution.

Proof: No generality is lost in assuming $\frac{1}{n}\sum_{i=1}^n x_i = \mathbf{0}$, since clusterability is invariant to translation. Notice also that a consequence of the above assumption is that X contains no repeated points. We prove the result by showing that the set

$$\begin{aligned} D := \{v \in \mathbb{R}^d \big| & v^\top X \text{ contains only unique points and} \\ & \mathrm{argmax}_{i\in I} VR'\left((v^\top X)_{(1):(i)}, (v^\top X)_{(i+1):(n)}\right) \\ & \text{is a singleton.}\} \end{aligned}$$

is open and dense in $\mathbb{R}^d$.

D is clearly open since $VR'(\cdot,\cdot)$ is continuous in its arguments and the elements of $v^\top X$ are continuous in v. To see that it is dense, take $w \notin D$ and $\epsilon > 0$. Since X contains no repetitions $\exists v \in \mathcal{B}_\epsilon(w)$ s.t. $v^\top X$ contains only unique points. If in addition we have that $\mathrm{argmax}_{i\in I} VR'\left((v^\top X)_{(1):(i)}, (v^\top X)_{(i+1):(n)}\right)$ is a singleton, then $\mathcal{B}_\epsilon(w) \cap D \neq \emptyset$ and we are done. Suppose then that it is not and let $I' = \mathrm{argmax}_{i\in I} VR\left((v^\top X)_{(1):(i)}, (v^\top X)_{(i+1):(n)}\right)$. Define

$$i' \in \mathrm{argmax}_{i\in I'} \frac{1}{\sqrt{i(n-i)}}\left\| \sum_{j=i+1}^{n} x^{v,j} \right\|,$$

and set $v' = \sum_{j=i'+1}^n x^{v,j}$. $\exists \delta > 0$ s.t. the following conditions hold

1) $v + \delta v' \in \mathcal{B}_\epsilon(w)$
2) $\mathrm{order}(v^\top X) = \mathrm{order}((v+\delta v')^\top X)$
3) $VR'\left(((v+\delta v')^\top X)_{(1):(i)}, ((v+\delta v')^\top X)_{(i+1):(n)}\right) > VR'\left(((v+\delta v')^\top X)_{(1):(j)}, ((v+\delta v')^\top X)_{(j+1):(n)}\right)$ for all $i \in I', j \notin I'$.

Take $j \in I'$, $j \neq i'$. We know $\frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^n x^{v,k} \neq \frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^n x^{v,k}$. However, $\frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^n v^\top x^{v,k} = \frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^n v^\top x^{v,k}$, and therefore $\sum_{k=i'+1}^n x^{v,k} \not\propto \sum_{k=j+1}^n x^{v,k} \Rightarrow v'^\top \sum_{k=j+1}^n x^{v,k} < \|v'\| \|\sum_{k=j+1}^n x^{v,k}\|$. So,

$$\begin{aligned} & \frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^{n}((v+v')^\top X)_{(k)} \\ & = \frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^{n}(v^\top X)_{(k)} + \frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^{n} v'^\top x^{v,k} \\ & = \frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^{n}(v^\top X)_{(k)} + \frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^{n} v'^\top x^{v,k} \\ & < \frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^{n}(v^\top X)_{(k)} + \frac{1}{\sqrt{j(n-j)}}\|v'\|\left\|\sum_{k=j+1}^{n} x^{v,k}\right\| \\ & \leq \frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^{n}(v^\top X)_{(k)} + \frac{1}{\sqrt{i'(n-i')}}\|v'\|\left\|\sum_{k=i'+1}^{n} x^{v,k}\right\| \\ & = \frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^{n}((v+v')^\top X)_{(k)}. \end{aligned}$$

The equality $\frac{1}{\sqrt{j(n-j)}}\sum_{k=j+1}^n (v^\top X)_{(k)} = \frac{1}{\sqrt{i'(n-i')}}\sum_{k=i'+1}^n (v^\top X)_{(k)}$ holds since both $i', j \in I'$. Therefore $VR'(((v+v')^\top X)_{(1):(j)}, ((v+v')^\top X)_{(j+1):(n)}) < VR'(((v+v')^\top X)_{(1):(i')}, ((v+v')^\top X)_{(i'+1):(n)})$. Since $j \in I'$ was arbitrary we have $\mathrm{argmax}_{i\in I} VR'\left(((v+v')^\top X)_{(1):(i)}, ((v+v')^\top X)_{(i+1):(n)}\right)$ is a singleton, and thus $D \cap \mathcal{B}_\epsilon(w) \neq \emptyset$. Since ϵ was arbitrary, we have that D is dense in $\mathbb{R}^d$. ■

Notice that the assumption given in the previous lemma, (11), holds with probability 1 if X is a sample of realisations of a continuous random variable.

Now, for $v \in D$ as in the above proof, we have $\nabla_v \Phi(v|X) = \nabla_v VR_{i'}'(v|X)$, where $i' = \mathrm{argmax}_{i\in\{1,...,n-1\}} VR_i(v|X)$. If we consider Eq. (7) we see that $\|\nabla_v \Phi(v|X)\|$ scales inversely with $\|v\|$, and so $\Phi(v|X)$ is not Lipschitz continuous. It is quite straightforward to show, however, that for $v, w \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ we have

$$|\Phi(v|X) - \Phi(w|X)| \leq K\frac{(\|v\| + \|w\|)\|v-w\|}{\|v\|\|w\|^2}, \tag{12}$$

where $K = \frac{4n^2 \mathrm{Diam}(X)^2 \lambda_M}{\lambda_m^2}$ and λ_M, λ_m are the largest and smallest eigenvalues of the covariance matrix of X respectively. Since the variance ratio is scale invariant we lose no generality by restricting v to lie on the surface of the unit sphere. This leads us to consider a reparameterisation of v in terms of its polar co-ordinates as follows. Let $\Theta = [0, \pi)^{d-1}$

and for $\theta \in \Theta$ define $v(\theta)$ by,

$$v(\theta)_i = \begin{cases} \cos(\theta_i) \prod_{j=1}^{i-1} \sin(\theta_j), & i = 1, \ldots, d-1 \\ \prod_{j=1}^{d-1} \sin(\theta_j), & i = d-1. \end{cases} \tag{13}$$

Notice that lemma 2 still applies if we sample elements of Θ via a continuous sampling distribution and consider their mapped values in $\mathbb{R}^d$ via Eq. (13). In addition, it is clear that $v(\theta)$ is Lipschitz continuous in θ, as a collection of products of Lipschitz functions, and that since $\|v(\theta)\| = 1$, Eq. (12) tells us that $\Phi(v(\theta)|X)$ is Lipschitz continuous as a composition of Lipschitz functions.

The variance ratio is closely connected with the K-means objective, and so we initialise the projection pursuit by setting $v_0 = \mu_1 - \mu_2$, where μ_1 and μ_2 are the centers of a 2-means clustering of X. We then use the gradient sampling algorithm to find a local maximum of the projection index $\Phi(v|X)$, which also automatically determines a locally optimal bi-partition of X.

B. Which Cluster to Split

The question of which of a collection of data sets (the clusters discovered so far by the divisive algorithm) should be split next boils down to how we should compare variance ratios. A direct comparison can be misleading especially when the number of dimensions is high relative to the number of data. In the extreme case, for data set, X, of size n in $d \geq n-2$ independent dimensions and for any partition of X into two clusters C_1, C_2 $\exists v \in \mathbb{R}^d$ s.t. $v^\top C_1$ and $v^\top C_2$ each take on only a single value. The variance ratio can therefore be made infinite for any partition of the data. Even when d is slightly larger than the number of data the variance ratio can be misleading in determining cluster structure within a univariate subspace. We require a measure which factors in the number of data as well as their dimensionality.

Here we provide a heuristic for the comparison of variance ratios. Despite the motivation coming from statistical theory we do not propose this as a statistically robust selection technique, and offer it rather as a rule of thumb. While the reasoning underpinning our argument is based on an overly simplified case of a mixture of two homoscedastic multivariate Gaussian components, we have found it to perform adequately in varied cases. It is well established that high dimensional data projected into low dimensional spaces have a tendency to appear as a mixture of Gaussians [10], [11], due to the aggregation of the features resulting in a sort of central limit effect. This observation has been addressed theoretically in the context of random projections [10]. The more questionable of our simplifications might then be that of homoscedasticity.

The variance ratio objective, Eq. (2), minimises the scaled sum of square deviations from two parallel hyperplanes. This can be viewed roughly as a random effects linear model in which the class identities are latent variables. The number of free parameters in this model depends on the number of data, n. If $n \leq d$ then the data can be embedded in an $n-1$ dimensional subspace without affecting their relative structure. The maximum number of free parameters in v is therefore $\min\{n-2, d-1\}$, since v is forced to lie on the surface of the unit ball, which has 1 dimension fewer. The total number of free parameters, including the two intercepts, is therefore $\min\{n, d+1\}$. If the class labels are known, rather than being latent variables, then the variance ratio, if appropriately scaled follows a non-centralised F distribution. When the classes are well separated along the optimal projection, v, the means of the projected data assigned to each class will closely approximate the means of the projected data arising from the true classes and so even in the latent variable model the variance ratio is approximately distributed non-centralised F.

The non-centrality parameter, γ, is unknown, and we use a single reference rule for all cases. We note that this parameter has little effect on performance provided it is not set too large. We choose the infimum value from the collection of bimodal two component Gaussian mixtures with equal variance and mixing proportion, under the assumption that class labels are known. This leads to $\gamma = n$.

From a collection of clusters $\{C_1, ..., C_k\}$, the next to be split is that which minimises

$$\mathbb{P}(F_{\alpha_{C_i}, \beta_{C_i}, \gamma_{C_i}} > f_{C_i}) \tag{14}$$

$$f_{C_i} = \frac{\beta_{C_i}}{\alpha_{C_i}} \max_{v \in \mathbb{R}^d, C \subset C_i} VR(v^\top C, v^\top (C_i \setminus C)) \tag{15}$$

$$\alpha_{C_i} = \min\{|C_i|, d+1\} \tag{16}$$

$$\beta_{C_i} = \max\{0, |C_i| - d - 1\} \tag{17}$$

$$\gamma_{C_i} = |C_i|, \tag{18}$$

where $F_{\alpha,\beta,\gamma}$ is a random variable with non-centralised F distribution with α and β degrees of freedom and non-centrality parameter γ. We note that these should not be interpreted as probabilities, but rather we use the probability function as a measure of how much more separable a data set is than the supremally overlapping case which still has discernable clusters.

C. The MCDC Algorithm

Algorithm 1 contains the algorithmic structure of the MCDC algorithm. The objects $F_{\alpha_{C'}, \beta_{C'}, \gamma_{C'}}$ and $f_{C'}$ can be found in Eq.'s (14)-(18). The optimisation determining the optimal partition, C, is described in Section II-A

Algorithm 1 MCDC

```
Input: Data set $X$, number of clusters $K$

$\mathcal{C} \leftarrow \{X\}$
while $|\mathcal{C}| < K$ do
    $C' \leftarrow \text{argmax}_{C \in \mathcal{C}} \mathbb{P}(F_{\alpha_C, \beta_C, \gamma_C} > f_C)$
    $[v, C] \leftarrow \text{argmax}_{w \in \mathbb{R}^d, B \subset C'} VR(w^\top B, w^\top (C' \setminus B))$
    $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C'\}) \cup \{C, C' \setminus C\}$
end while
return $\mathcal{C}$
```

The maximum clusterability binary partition of a univariate data set is determined by a single order statistic. The variance ratio of such a partition can easily be computed from the cumulative sum vector of the sorted data set, $CS(R) := (R_{(1)}, \sum_{i=1}^{2} R_{(i)}, ..., \sum_{i=1}^{n-1} R_{(i)}, \sum_{i=1}^{n} R_{(i)})$, with a computational cost $\mathcal{O}(1)$. The total cost of determining the optimal bi-partition of a univariate data set is thus $\mathcal{O}(n \log(n))$, with the log factor resulting only from the sorting algorithm. We note that the order of the data projected onto consecutive iterates in the gradient descent algorithm tends to be similar, and so the sorting cost can be reduced by warm starting using the previous order. The gradient sampling algorithm requires at least $d+1$ sampled gradients for each iteration. The total cost for performing a binary partition is therefore $\mathcal{O}(d^2 n \log(n) i)$, where i is the number of iterations in the gradient descent algorithm. Total complexity for the MCDC algorithm is thus $\mathcal{O}(d^2 n \log(n) \sum_{j=1}^{K-1} i_j)$ where i_j is the number of iterations in the j-th gradient descent scheme.

The bottleneck in terms of computation lies in the gradient sampling. For high dimensional examples we alleviate this burden by checking with each iteration whether the current solution is close to a point of non-differentiability. If the second largest variance ratio is at least 99% of the largest, or the order statistic defining the largest variance ratio is within 0.001 of another projected datum then an iteration of the GS algorithm is performed. If not, then a standard gradient step is done instead. This approach still has a worst case complexity as above, however we have found empirically that it tends to improve computational speed significantly. We also note that in no cases did this approach fail to converge to a local maximum.

III. Experimental Results

In this section we provide the results from experiments using the proposed method. We asses the performance as a dimension reduction method and as a clustering algorithm.

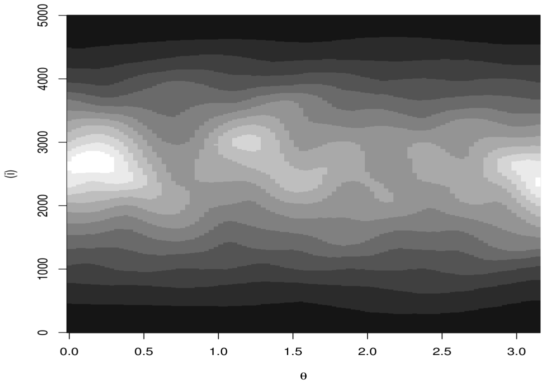
A. MCDC Dimension Reduction

In this subsection we consider the dimension reduction aspect of the MCDC method. We compare our method with PCA, projection pursuit for maximisation of the Dip statistic [4] (Dip), dimensionality reduction for spectral clustering [6] (DRSC) and iterative Support Vector Regression [8] (iterSVR). The latter is an algorithm for finding maximum margin hyperplanes for clustering.

Table I shows the univariate density plots for the various dimension reduction methods applied to three benchmark data sets from the UCI machine learning repository [15]. The vertical lines indicate the binary partition of the data. We include the class components (unknown to the algorithms) to illustrate the class separability within the subspaces. All objective driven dimension reduction methods find substantially stronger cluster structure than standard PCA in general. MCDC and iterSVR visually appear to find the best subspaces to separate classes.

The poor performance of DRSC on the Yeast data set highlights an important feature of dimension reduction for clustering. In many cases the optimal partition based on various cluster definitions will only separate singletons, or small groups of data which do not constitute entire clusters. As such dimension reduction techniques such as the minimum density [5] and large margin methods [8] have been designed to take as input a balancing parameter, good values of which will not always be known in practice. The MCDC method does not appear to suffer this limitation. Figure 1 shows the variance ratio clusterability of the 2 dimensional S1 data set [12]. This 2 dimensional example allows us to visualise the objective as a function of a single projection angle, θ, and the order statistic (i), where lighter colour indicates higher variance ratio clusterability. We can see for all projections the optimal partition leads to a roughly balanced split of the data

Fig. 1. Variance ratio clusterability of the S1 data set [12]. Lighter colour indicates higher variance ratio value.



B. Clustering Performance

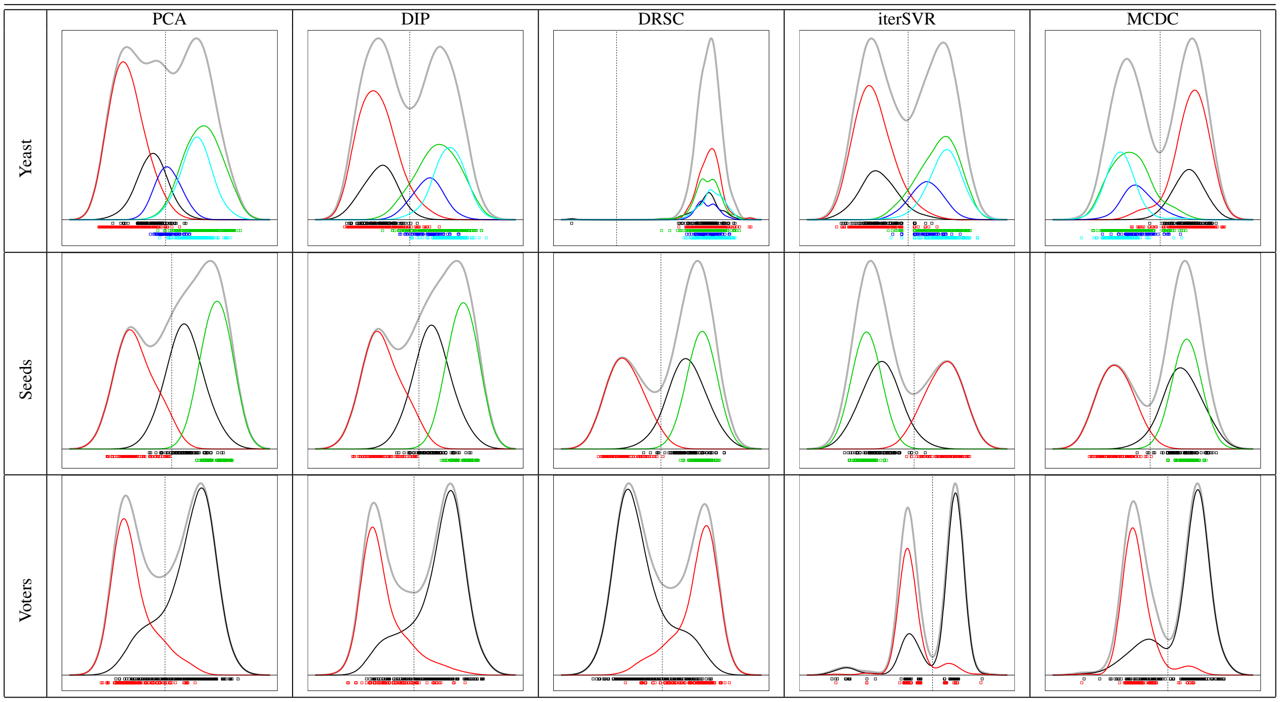
In this subsection we present the results from experiments on publicly available data sets. We compare the clustering results of the following algorithms

1) K-means (KM): We use the `R` implementation of the standard K-means algorithm
2) PCA K-means (PCA KM): Standard K-means applied to the data projected onto the first $K-1$ principal components
3) Spectral Clustering (SC): We use the implementation in the `R` package `kknn`, which is based on the symmetric normalised Laplacian with affinity matrix based on 7 mutual nearest neighbours.
4) PCA Spectral Clustering (PCA SC): Spectral clustering as above applied to the data projecetd onto the first $K-1$ principal components
5) Iterative Support Vector Regression [8] (iterSVR): We set the balancing parameter $\ell = 0.3$ as in [8], where we argue the balance of classes will not be known in general. We split the cluster with the most data at each iteration. We use the linear kernel as this gives the most meaningful comparison with our method.
6) Maximum Clusterability Divisive Clustering (MCDC): Our proposed method

All algorithms are given the correct number of clusters as input. All cases of the K-means algorithm are based on the best solution from 10 initialisations.

Table II reports the clustering performance from a collection of benchmark data sets from the UCI machine learning repository [15]. The data sets range in dimensionality from 4 to 72 and in size from 150 to $\approx$ 700. The number of clusters ranges from 2 to 6. Algorithms are compared based on Purity [13] and V-Measure [14]. Both measures compare the clustering result with the true class labels, with high values indicating a better agreement between the two. The table

TABLE I. UNIVARIATE DENSITY PLOTS



reports the average clustering performance (± one standard deviation) from 10 replications. The results are encouraging and show that the MCDC algorithm is capable of building high quality clustering models in various environments. In almost all cases MCDC outperforms standards K-means and PCA K-means, which is an important comparison due to their similar objectives. MCDC also achieves the best performance of all algorithms in most cases considered.

C. Image Segmentation

One of the key tasks in image segmentation lies in identifying the focal objects in the image while including as little noise from the background as possible. In this section we present the results of image segmentation tasks using the MCDC algorithm, iterSVR, 2-means and PCA 2-means. Two way clustering was performed on the RGB pixel values. Spectral clustering was not performed in these tasks due to the high computational cost.

Table III shows the results of the image segmentation task on three images from the Berkeley image database [16]. The first row shows the original image. The second shows the results from 2-means and the third row shows PCA 2-means. The fourth row shows the result from the iterSVR algorithm, and finally MCDC resutls are in row 5. MCDC clearly does the best job separating the objects from background in these examples.

IV. Conclusion

We proposed a novel combined dimension reduction and clustering algorithm which is based on maximising the variance ratio clusterability of the data projected into a univariate subspace. We showed that the sufficient conditions for convergence to a local optimum using the gradient sampling algorithm are satisfied. We proposed a heuristic method for comparing variance ratios within projected subspaces which allowed us to order the clusters discovered using a binary divisive procedure and thereby iterate such binary divisions to generate high quality clustering models.

We evaluated the proposed methodology on tasks in dimension reduction, clustering and image segmentation. We found the method to be versatile in its varied applications. The results indicated strong performance of the proposed method in comparison with existing methods.

A limitation of the proposed approach lies in the fact that the number of clusters is assumed to be known. However this knowledge will not always be available in practice. Determining the number of clusters in a data set is a challenging problem, and remains a very active area of research. One of the benefits of divisive hierarchical clustering models is that the problem of determining the number of clusters is equivalent to determining whether a data set contains one cluster or more than one cluster, since this defines a stopping criterion for the divisive procedure. In many cases this is a less challenging problem than determining overall the number of clusters present. Determining such a stopping criterion represents a promising direction to improve the current method in the future.

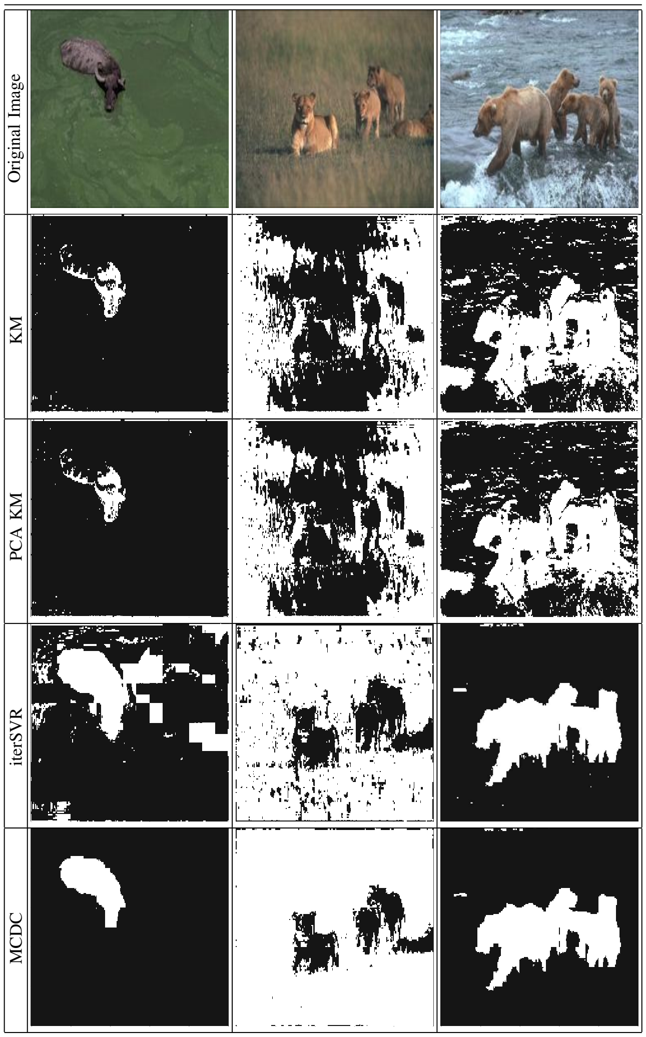
Acknowledgment

David Hofmeyr gratefully acknowledges funding from both the Engineering and Physical Sciences Research Council (EPSRC) and the Oppenheimer Memorial Trust.

TABLE II. Comparative Performance on Publicly Available Data Sets

		KM	PCA KM	SC	PCA SC	iterSVR	MCDC
Iris	Purity	0.833±0.000	0.833±0.000	0.820±0.000	0.773±0.000	0.827±0.000	**0.847±0.000**
	V-Measure	0.659±0.000	0.659±0.000	0.627±0.000	0.586±0.000	0.653±0.000	**0.673±0.000**
Ionosphere	Purity	0.709±0.000	0.700±0.000	0.640±0.000	0.683±0.000	0.700±0.000	**0.711±0.000**
	V-Measure	0.129±0.000	0.115±0.000	0.030±0.000	0.085±0.000	0.121±0.000	**0.134±0.000**
Yeast	Purity	0.707±0.002	0.715±0.002	0.739±0.000	0.639±0.000	**0.746±0.004**	0.744±0.000
	V-Measure	0.527±0.001	0.529±0.009	**0.560±0.000**	0.403±0.000	0.547±0.006	0.555±0.011
Seeds	Purity	0.919±0.000	0.924±0.000	0.938±0.000	0.900±0.000	0.947±0.002	**0.948±0.000**
	V-Measure	0.728±0.000	0.738±0.000	0.774±0.000	0.689±0.000	0.799±0.005	**0.824±0.000**
Dermatology	Purity	0.870±0.027	0.910±0.019	0.962±0.000	0.943±0.000	0.803±0.000	**0.966±0.004**
	V-Measure	0.868±0.012	0.876±0.009	0.924±0.000	0.901±0.000	0.769±0.001	**0.944±0.007**
Chart	Purity	0.715±0.051	0.707±0.052	0.667±0.000	**0.823±0.000**	0.720±0.000	0.820±0.005
	V-Measure	0.759±0.016	0.753±0.013	0.795±0.000	**0.871±0.000**	0.664±0.000	0.758±0.004

TABLE III. Image Segmentation Results



References

[1] M. Ackerman and S. Ben David. Clusterability: A Theoretical Study. In *Proceedings of AISTATS-09*, JMLR: W&CP, 5:1-8, 2009.

[2] H.-P. Kriegel, P. Kroger, and A Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data, 3(1):158, 2009.

[3] B. Zhang. Dependence of Clustering Algorithm Performance on Clustered-ness of Data. HP Labs Technical Report HPL-2001-91, Hewlett-Packard Laboratories, 2001.

[4] A. Krause and V. Liebscher. Multimodal Projection Pursuit Using the Dip Statistic. Technical Report 13, Universit at Greifswald, 2005.

[5] N. Pavlidis, D. Hofmeyr, and S. Tasoulis. Minimum Density Hyperplane: An Unsupervised and Semi-Supervised Classifier. *arXiv preprint arXiv:1507.04201*, 2015.

[6] D. Niu, J. G. Dy, and M. I. Jordan. Dimensionality Reduction for Spectral Clustering. In *Proceedings of AISTATS-2011*, 552-560, 2011.

[7] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum Margin Clustering. In *Proceedings of NIPS-2005*, 17, 2005.

[8] K. Zhang, I. W. Tsang and J. Kwok. Maximum Margin Clustering Made Practical. *Proceedings of ICML-2007*, 2007

[9] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751-779, 2006.

[10] P. Diaconis and D. Freedman. Asymptotics of Graphical Projection Pursuit. *Annals of Statistics*, 12:793-815, 1984.

[11] E. Meckes. Quantitative Asymptotics of Graphical Projection Pursuit. *Electronic Communications in Probability*. 14(17):176-185, 2009.

[12] P. Franti and O. Virmajoki. Iterative Shrinking Method for Clustering Problems. In *Pattern Recognition*, 39(5):761775, 2006. doi: `http://dx.doi.org/10.1016/j.patcog.2005.09.012`

[13] Y. Zhao and G. Karypis. Criterion Functions for Document Clustering: Experiments and Analysis. In *Machine Learning*, 2001.

[14] A. Rosenberg and J. Hirschberg. V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning-2007*, 410- 420, 2007.

[15] K. Bache and M. Lichman. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. `http://archive.ics.uci.edu/m`

[16] P. Arbelaez, C. Fowlkes, and D. Martin. Berkeley Image Database. University of Berkeley, California. `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds`