Inexact MDL for Linear Manifold Clusters

Robert M. Haralick*, Art Diky†, Xing Su‡
Department of Computer Science
The Graduate Center, CUNY
365 Fifth Avenue, New York, NY 10016
Email: *RHaralick@gradcenter.cuny.edu
†adiky@gradcenter.cuny.edu
‡xsu@gradcenter.cuny.edu

Nancy Y. Kiang
NASA Goddard Institute for Space Studies
2880 Broadway, New York, NY 10025
Email: Nancy.Y.Kiang@nasa.gov

***Abstract*—We present a regularization technique based on the minimum description length (MDL) principle for the linear manifold clustering. We suggest an inexact minimum description length method based on describing the data structure as linear manifold clusters. We examine the behavior of the proposed method and compare it performance against simulated clustering results of various dimensionality and structure. Finally, we empirically evaluate the proposed technique on a climate data.**

I. Introduction

Minimum Description Length (MDL) can be understood to be a technical specification or formalization of the Occam's razor principle to understand a data set. One way to pose the general problem is given a data set, define a language in which to represent the data set so that the data set described in the language is a meaningful description and the number of bits for the representation in the language is minimal. This is different from data compression in the sense that data compression only uses information theoretic methods to minimize the number of bits to represent the data set, but the representation in itself is not meaningful. It does not give any insight into the structure of the data. If there are multiple possible languages for describing the data, minimum description length can be the principle for deciding which is the best language.

Our description language is the language of linear manifolds. Each linear manifold cluster consists of the description of the linear manifold and the coding of the data associated with the linear manifold is given by encoding the orthogonal projection of each data point onto its manifold and the encoding of the difference between its position off the manifold and its orthogonal projection on the manifold. The inexactness of the description arises because the description of the position of each data point off the manifold is described not exactly, but with some controlled error.

Section II is a literature review. Section III is a technical description of the linear manifold clustering stochastic search technique. Section IV describes how the MDL principle is used to determine whether to accept a cluster or not. Section V discusses our results and section VI concludes the paper.

II. Literature Review

Subspace clustering [1] is a special case of linear manifold clustering, where the basis vector set for each cluster is a subset of the natural basis vectors of the space. Data can be well approximated by a mixture of linear manifolds (linear or affine subspaces). Haralick and Harpaz [2] presented a linear manifold clustering algorithm (LMCLUS) which is a strict partitioning clustering algorithm that performs stochastic search on the dataset in order to find best possible location of the linear manifold clusters. Kak [3] used a linear manifold representation of a fixed number of clusters, obtained by sampling the original dataset and minimizing the reconstruction error from point assignments to cluster prototypes. Peng et al. [4] constructed linear manifold cluster prototypes by performing spectral decomposition of small random samples with subsequent assignment of the rest of the dataset points to a nearest subspace cluster prototype. Wang et al. [5] used a mixture of probabilistic PCAs to form a collection of linear manifolds on the dataset.

Moreover, many linear methods fail to provide good performance when applied to nonlinear structures. On the other hand, nonlinear methods, such as nonlinear dimensionality reduction techniques, can be naturally used on linear manifolds [6]–[8].

Rissanen at al. [9] presented an approach where data and noise are separated, and the code length of the model is restricted by a parameter. The hypothesis selected by MDL captures all the structure inherent in the data. Given the hypothesis, the data cannot be distinguished from random noise.

III. Linear Manifold Clustering

The cluster ideal in Linear Manifold Clustering is a linear manifold. A linear manifold of dimension zero is a point. A linear manifold of dimension 1 is a line. A linear manifold of dimension 2 is a plane. In general, a linear manifold is a translated subspace. The dimension of the linear manifold is the dimension of the subspace. K-means is a special case of linear manifold clustering where the linear manifold has dimension zero.

Linear manifold clustering is appropriate in the case that there are linear dependencies among the variables, each cluster having a different number and different kinds of linear dependencies. Let us take an example to make this clear. Suppose that we are in a three dimensional space with three clusters. The first two clusters have one linear dependency.

Their linear manifolds are planes. In our example the planes are parallel. The third cluster has two linear dependencies. Its linear manifold is a line. The observed data points can be thought of as points whose ideals are on their manifolds, but were slightly perturbed off their manifolds, see Fig. 1.

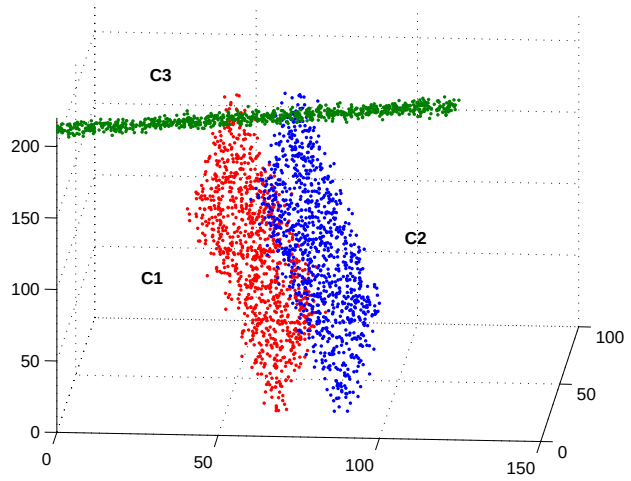

Fig. 1. Example of points forming three linear manifold clusters. The green cluster is one dimensional. The blue and red clusters are two dimensional and parallel [2].

Linear manifold clusters [2] can be found by a stochastic search procedure, beginning with the one dimensional linear manifold clusters and proceeding to higher dimensional clusters. A one dimensional linear manifold is determined by two points. The stochastic search samples two points from the data set, forms the manifold, and then the distances from all points to the discovered manifold are calculated. If the manifold is indeed one that has many data points close to it, the distance histogram will have a peak close to 0 distance followed by a valley and then a rise to a long fat tail or another peak far from the origin, see Fig. 2.

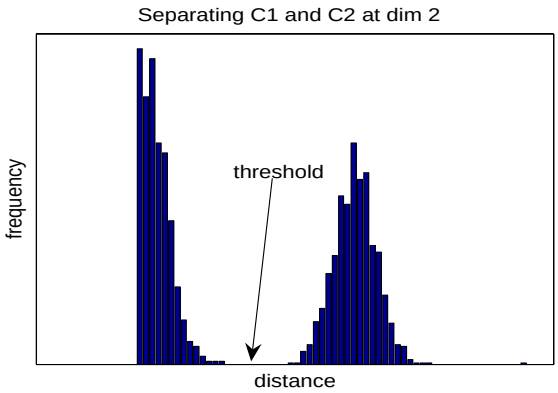

Fig. 2. Example of a distance to manifold histogram that shows that a linear manifold cluster can be formed from the test manifold [2].

If this happens, then a suitable threshold can be found that separates the data points that are near to the manifold from those data points that are far away from the manifold. The data points that are near the manifold are collected together and are used to make a good statistical estimate for the manifold basis and offset. The manifold basis is given by the first M principal components of the cluster data points, where M is the dimension of the manifold. The offset can simply be the mean of the data points in the cluster. Manifolds that do not have the right shaped distance to manifold histograms do not get the chance to form clusters.

Our linear manifold clusters must satisfy two criteria. First, the goodness of a separation between the mode near zero and rest of the point-to-manifold distance histogram modes should be larger then the user specified. This criterion is fully explained in [2]. Second, the cluster compression ratio, defined as the ratio between the linear manifold cluster description length and the uncompressed description length of the cluster, should be larger then the user defined threshold. This criterion, in effect, acts as an internal validation the cluster goodness-of-fit, and it is a new addition to the algorithm described in [2].

IV. MDL Linear Manifold Cluster Description

First we determine the number of bits it takes to encode the translational offset of the linear manifold and then the orthonormal basis vectors spanning the linear manifold. Then we determine the number of bits it takes to encode the points of the candidate linear manifold cluster to within a given squared error.

Let $X = \{x_j \in \mathbb{R}^N | \; j = 1, \ldots, J\}$ be the points associated with the M-dimensional linear manifold cluster $\mathcal{M}$. It is described by a set of orthonormal basis vectors, that span the linear manifold, $B = \{b_m \in \mathbb{R}^N | m = 1, \ldots, M\}$ and a translation vector $\mu \in \mathbb{R}^N$.

a) Model Encoding: The encoding of the translation vector μ requires N numbers.

To represent any vector x, after its translation, we need the basis vectors spanning the manifold and we need the basis vectors orthogonal to the manifold. From the basis vectors spanning the manifold we can determine the relative coordinates of the orthogonal projection of x to the manifold and from the basis vectors spanning the orthogonal complement space, we can determine the orthogonal projection of x to the complement space.

Since the basis vectors of the linear manifold and its orthogonal compliment space are orthonormal, we can represent the basis vectors in less than N^2 numbers. We can use a decoding schema that uses the orthonormal constraints in recovering the N basis vectors. Each basis vector has norm 1. This constitutes N constraints. The orthonormality constraints specify another $N(N-1)/2$ constraints. The total number of orthonormality constraints is then $N(N+1)/2$.

To describe the linear manifold requires N numbers for the offset of the manifold from the origin plus $N^2 - N(N+1)/2$ numbers for basis vectors. Letting P_m be the number of bits used for encoding each component of the offset and each of the numbers required to calculate the basis vectors. Then the total number of bits, $L(H)$, required to specify the structure of a linear manifold and its orthogonal complement space is

$$L(H) = P_m[N + N(N-1)/2)] = P_m N(N+1)/2 \quad (1)$$

b) Data Encoding: Let $B^{N\times M}$ be a matrix whose columns are the orthonormal basis vectors spanning the linear manifold. Then the relative coordinates of the orthogonal projection of a vector $x-\mu$ to the manifold is given by $B^T(x-\mu)$. This is a vector of dimension $M \times 1$. Each of the M components of this vector will be encoded with P_d bits.

Let $\bar{B}^{N\times N-M}$ be a matrix whose columns are the basis vectors spanning the orthogonal complement space. The relative coordinates of the orthogonal projection of a vector $x-\mu$ to the complement space of the manifold is given by $\bar{B}^T(x-\mu)$. This is a vector having $N-M$ components. The reconstruction of that part of x that lies in the orthogonal complement space is given by $\bar{B}(\bar{B}^T(x-\mu))$.

The total number of bits required to encode data D given a model H is

$$L(D|H) = J[P_d M + S(\varepsilon)] \quad (2)$$

where J is number of points in the linear manifold cluster, S is the entropy of the distribution of cluster points, in the orthogonal compliment subspace to the linear manifold of the cluster, calculated to be correct within the fitting error ε.

We assume that each of the $K = N - M$ components of that part of x that lies in the orthogonal complement space is uniformly distributed, but that the interval of the uniform distribution is different for each component. For component k, we let the uniform distribution be defined on the interval $[-A_k/2, A_k/2]$. We will quantize the interval $[-A_k/2, A_k/2]$ into N_k equal length quantizing intervals and encode component k by the index of the quantizing interval into which it lies. Since the intervals are all equal length, knowing the index of the subinterval into which a value falls, permits the value of to be approximated by the mean of the subinterval into which it falls. The squared error is then the variance of a uniform distribution over the subinterval.

Set the log of the total number of quantized choices in the K-dimensional space equal to a given $C = \sum_{k=1}^{K} \log N_k$.

From this it follows that the integer value of N_k can be taken to be the smallest integer N_k satisfying

$$N_k(C) = \lceil A_k e^{(C - \sum_{j=1}^{K} \log A_j)/K} \rceil \quad (3)$$

The interval lengths $A_1, \ldots, A_K$ are given and fixed. The values of $N_1, \ldots, N_K$ are each dependent on the value of C. So we can write the squared error E^2 as the variance of a uniform distribution over all subintervals of the K quantized components,

$$E^2(C) = \frac{1}{12} \sum_{k=1}^{K} \left(\frac{A_k}{N_k(C)} \right)^2 \quad (4)$$

If we operate under the protocol that the quantizing must be done fine enough, such that for the user specified quantization error bound ε, the value of C is small enough to satisfy

$$E^2(C) < \varepsilon^2 \quad (5)$$

then is not hard to show that the value C is defined over the interval

$$[0, K \log N_{max}] + \sum_{j=1}^{K} \log A_j - \min_k \log A_k$$

We can find the optimal number of quantization intervals N_k with a given user defined precision value ε by performing a search for appropriate value of C in the above interval such that it would satisfy condition (5).

Given the value C that satisfies (5), we can calculate the number of bits required to encode the position of the cluster point in the orthogonal complement space of the linear manifold cluster. The value C corresponds to the entropy S of a distribution of cluster points in the orthogonal compliment space, that is required in (2). Since the logarithms are to base e, C does not have the meaning of bits. But

$$\frac{C}{\log 2} = \sum_{k=1}^{K} \log_2 N_k$$

does have the meaning of bits.

Using the above descriptions of model (1) and data message (2) length, the total length of the message for linear manifold cluster (LMC) is calculated as

$$L(\varepsilon) = P_m N(N+1)/2 + J(P_d M + S(\varepsilon)) \quad (6)$$

From (6), we can see that two factors affect description length - the precision constants and the entropy. If simple models of the linear manifold cluster are favored then the entropy and the precision parameters should be proportionate. It would allow stable growth of the description length with respect to the size and the dimensionality of the linear manifold cluster.

If we want to determine a optimal clustering parameters, it is important to use the encoding that does not calculate the data points in the clusters, but the distribution of the data points in each of the clusters. The difference is this: to characterize the data points in the cluster, the number of bits required will increase with the number of data points. However the characterization of the distribution does not depend on the particular number of points: it depends on representing the various parameters of each of the clusters so that from the representation a sample of data points can be generated that would be indistinguishable from the original sample. Or to say this another way, the clustering is to characterize the population from which the observed data has been sampled.

We use model encoding schema as given in (1), as for data encoding is determined based on a the spread of the data on the manifold and as well in the orthogonal complement space.

For an M-dimensional manifold, we can use the first M eigenvalues as the variance of the spread on the manifold.

Since this is from a principal components, the covariance matrix is diagonal. The distribution of the data on the manifold, then can be described as a Normal distribution with the mean being given by the translation vector and the covariance matrix being diagonal with the diagonal entries coming from the first M eigenvalues of the principal components.

For the orthogonal complement subspace, we assume a more general model that allows for a description that is accurate to within a user specified error. We model the distribution based on the quantization of orthogonal complement subspace $N-M$ dimensions. Each of these dimensions has an observed minimum value, a maximum value and number of quantized bins as determined by the entropy calculation and the user specified error. As well, each of the bins has a probability. To generate points in the orthogonal complement space, for each of its dimensions, we can choose a bin in accordance with the bin probabilities and within a bin choose a value uniformly distributed between the quantizing boundaries of the bin.

The M coordinates generated from the manifold and the $N-M$ coordinates chosen from the orthogonal complement space then can be used as coefficients of their respective basis vectors to produce a vector in the N-dimensional space.

$$L(D|H) = P_d \left(N + 2 \sum_{m=M+1}^{N} Q_m \right) \tag{7}$$

where Q_m is the number of quantized levels for orthogonal complement dimension m.

We can assume that because of the MDL in the clustering, regardless of the value of the input parameters that the user set, the clustering gives an appropriate characterization of the distribution of the population from which the observed data set was sampled. The best characterization of the population is the characterization that has fewest bits and calculated as

$$L(\varepsilon) = P_m N(N+1)/2 + P_d \left(N + 2 \sum_{m=M+1}^{N} Q_m(\varepsilon) \right) \tag{8}$$

V. Results

Finally, we would like to understand how well MDL evaluates the goodness of a linear manifold cluster. Suppose, we have a 2D linear manifold cluster in 3D space, how can we guarantee that the particular cluster is actually a 2D cluster? What if this cluster is a 1D linear manifold cluster with wide bounds? What will be the criteria which would provide a distinctive answer on correctness of a structure description of some linear manifold cluster. We claim that MDL value of a linear manifold cluster, calculated with correct assumptions about its structure would yield a minimal value.

In order to test above assumption, we generated a 5D linear manifold cluster in a 10D space, following a similar cluster generation schema as in above experiments. We generated coordinate values of the cluster points from a normal distribution, where for the primary dimensions of the LM cluster, the variance is set 1.0, and for dimensions in the orthogonal complement to the linear manifold, the variance is set to 0.1. The encoding precision constants for a model and data are 24 and 16.

We calculated the MDL value of this cluster as if its dimension is unknown to us, as it happens during a selection of the cluster candidate manifold in LMCLUS algorithm. We specify during the MDL value calculation that our 5D cluster has dimension in range from 1 to 9. Moreover, we use various quantization errors during this experiment to understand how the precision of the cluster description affects the goodness of the selected cluster structure.

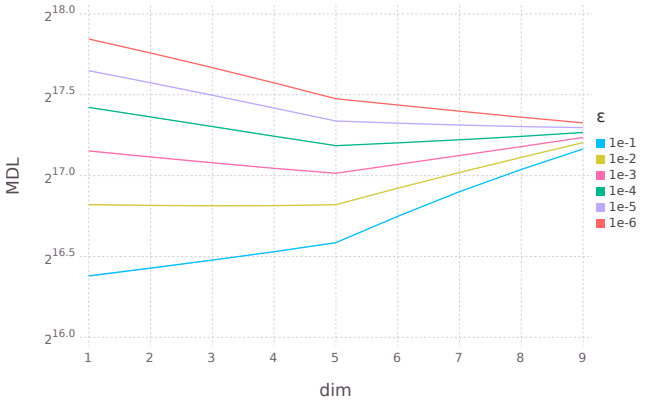

Fig. 3. MDL values, calculated with various cluster dimensionality parameters and quantization error ɛfor a cluster that is actually a 5D LM cluster in a 10D space.

Figure 3 shows that MDL value calculated with the correct structural parameters of the examined linear manifold cluster has a minimum value when the dimension parameter corresponds to the cluster dimensionality. Low values of the quantization error ε will result in a high cluster MDL value. High values of ε will result in a low cluster MDL value. If the quantization error ε is set to a small value, the cluster MDL value will decrease monotonically with the dimension of cluster. If the quantization error ε is set to a large value, the cluster MDL value will increase monotonically with the dimension of cluster. If the quantization error ε is set correctly, the cluster MDL value will decrease until the right number of dimensions is selected after which MDL value increases with increasing number of dimension parameter.

A. MDL of a Zero-Dimensional Manifold Cluster

An interesting case arises when we try to calculate the MDL of a zero-dimensional manifold cluster. Given that a zero-dimensional (ZD) manifold is a point, any cluster characterized only by its center point is considered as a zero-dimensional manifold or spherical cluster . Many clustering algorithms, e.g. k-means, produce zero-dimensional manifold clusters [10].

Any zero-dimensional manifold cluster is a special case of the linear manifold cluster, thus we can use encoding (6) to calculate the MDL value of the cluster given that dimension of the manifold is zero, $M = 0$. Thus, (6) is simplified as follows

$$L(\varepsilon) = P_m N + JS(\varepsilon) \tag{9}$$

Georgieva et al. [11] took a similar approach in describing the MDL of zero-dimensional clusters, produced by the k-means algorithm. However, instead of using the entropy of the quantized distribution of the point positions in particular dimensions, the projection distances to the point were encoded in MDL as follows

$$L = L(H) + L(D|H) = PN + \sum_{i=1}^{J} \sum_{p=1}^{N} \log(d_i^p + 1) \quad (10)$$

where d_i^p corresponds to the projection of the distance d_i of the i-th point to the p-th dimension. Such a description does not provide an informative encoding of coordinates when distances to the center in the cluster are near zero. In such a case, distance is encoded with less than one bit on the average.

We will compare the degenerate case of the inexact encoding of zero-dimensional manifold cluster calculated by (9) on synthetically generated linear manifold and spherical clusters. Such an approach will provide a common ground for comparison between linear manifold and spherical clusters. We also compare the MDL value of a linear manifold clustering with a cumulative MDL of a clustering constructed from zero-dimensional clusters which is a more natural representation of linearly shaped data from the perspective of spherical clustering algorithms.

We used a synthetically generated dataset which has a form of a 1D linear manifold cluster, an elongated dataset along the one axis, in 2D full space. Cluster generation procedure was described above. We performed the MDL value calculation for the 1D manifold following MDL formula (6) and then the 0D manifold case defined by (9) for various quantization errors.

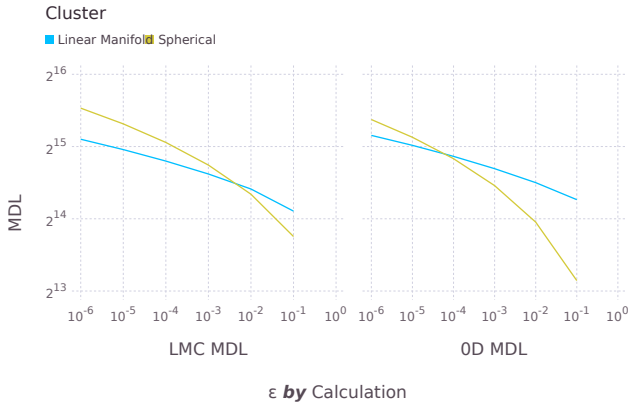

Fig. 4. Linear manifold (1D) and zero-dimensional (0D) MDL calculations for 1D linear manifold and spherical 0D clusters, located in 2D space, with various quantization errors ε.

Figure 4 shows results of linear manifold Eq. (6) and zero-dimensional Eq. (9) MDL value calculations for various types of manifold clusters. For large quantization errors, both approaches to the MDL calculation produce a small MDL value for spherical cluster. However, when the precision of the quantization procedure increases, resulting in a more complete and informative description of the cluster, the MDL value of the linear manifold cluster becomes smaller than the spherical cluster regardless of the selected method of calculation.

Because of the structural difference between linear manifold and spherical clusters, it is hard to come with common criteria for comparison of different types of clusters. We use the MDL value as a measure for heterogeneous cluster comparison. In order to test how the cluster MDL would perform as a comparison score, we calculated MDL values of synthetically generated clusters of different types - linear manifold and spherical.

We generated a 1D linear manifold cluster dataset from a bivariate normal distribution, as in previous experiments, and used the k-means algorithm to synthesize spherical clusters from it. We varied the number of clusters for the k-means algorithm that allowed us to form clusters which gradually obtain a spherical shape, as the linear manifold cluster got partitioned into more clusters.

We perform an evaluation of the MDL value for the linear manifold clusters by Eq. (6) and spherical cluster by Eq. (9). When k-means generated more than one cluster from the original dataset, we summed all the cluster MDL values in the clustering to obtain the MDL score for the original 1D LM cluster represented by the dataset. This is shown in Figure 5.

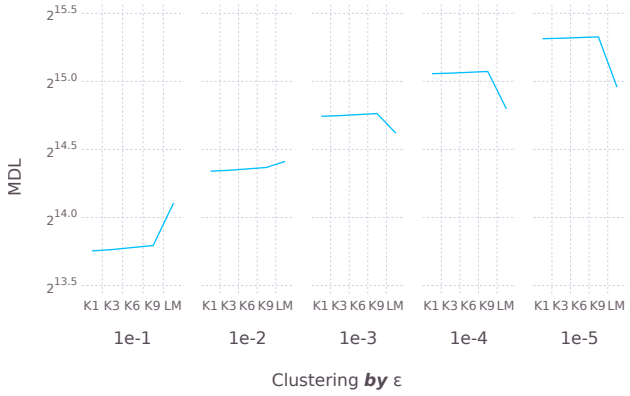

Fig. 5. MDL value of k-means clusterings (K[k]) produced from the 1D linear manifold cluster (LM), located in 2D space, under various quantization errors ε.

We found that the division of the linear manifold on multiple spherical clusters does not provide much difference in the resulted MDL value. As in the previous experiment, the major factor which affects MDL calculations is the quantization error parameter. For a small quantization error, spherical clusters provide a smaller MDL value for the experimental dataset. Moreover, the MDL value of the whole dataset does not increase significantly with the number of clusters in the k-means clustering. However, as the quantization error decreases, the MDL value calculated by Eq. (6) becomes significantly smaller then the spherical cluster MDL value Eq. (9).

This result suggests that for a large quantization error a spherical description of the linear manifold cluster provides more compact MDL value over the linear manifold MDL model. But while the quantization error decreases, giving a better description of the data, the linear manifold MDL model produces more compact encoding of the linear manifold

cluster and outperforms the spherical MDL model regardless of cluster proximity to true spherical representation.

B. Using MDL Heuristic in Climate Data Clustering

We added the linear manifold clustering MDL heuristic into the LMCLUS algorithm, and tested clustering performance on climate datasets.

Our dataset comprised of subset of the CRU 3.22 dataset of monthly global surface temperature averages, and Global Precipitation Climatology Centre (GPCC) dataset of monthly precipitation averages, for a 30 year period form 1951 to 1980. Original datasets have the same $1^\circ \times 1^\circ$ resolution. Both datasets are 12 dimensional, so the combined dataset has 24 dimensions. For each of group of 12 fields, a unit length normalization was performed, by subtracting the minimum value from every point and divided it by the field maximum minus the minimum value. Normalization makes the scale of the disparate temperature and precipitation fields similar.

The Köppen-Geiger (KG) climate classification system is a widely used scheme developed by geographers to classify climate types correlated with observed land ecosystems [12]. It is based on observed limits of these ecosystems relative to seasonal or annual precipitation and temperature. A recent updated version identifies 34 climate classes [13]. The system is not perfect, so variations are often proposed. However, on the hypothesis that ecosystem types are an expression of the climate, the KG system offers a good benchmark for a clustering analysis.

We perform clustering of the above climate dataset using following algorithms: k-Means [10], ORCLUS [14], original LMCLUS [2] and LMCLUS modified with the MDL heuristic.

The k-Means clustering assumes that the data is modeled as a mixture of spherically shaped distributions. In this model, the cluster ideal is a point, the cluster center, which is its mean, and the observations are isotropically perturbed around the mean. Because the number of clusters must be set a priori for k-Means, with the climate data clustering, we set this number to 34 to match the number of Koeppen-Geiger classes. Similarly to k-Means, ORCLUS requires an exact number of clusters as a parameter, but the resulting clusters are linear manifolds of the dimension specified by one of the parameters. We set ORLCUS parameters such that the algorithm would generate 34 1D linear manifold clusters.

Linear Manifold Clustering (LMCLUS) does not have to specify an exact number of clusters in advance, but there are multiple parameters that affect performance of the algorithm. We set only a small group of them, the rest of the parameters were set to their default values. The effect of the parameters on the clustering performance is described in the original paper [2]. In our experiments, the following LMCLUS parameters were set: *best_bound* to 0.4, *sampling_factor* to 0.1, *number_of_clusters* to 34, *min_cluster_size* to 150.

We updated LMCLUS with the MDL heuristic that allowed a goodness evaluation of the prospective manifold cluster before committing to the partitioning of this clusters from the rest of the dataset. We calculated a compression ratio as a ratio between "raw" cluster encoding, as a total number of bits required to encode each point of the cluster with a constant precision, and the cluster MDL encoding. If the compression ratio is larger than a user specified threshold, the cluster is accepted.

In order to compare the goodness of produced clusterings, we calculated the total MDL value of the resulting clustering for each algorithm as a sum of cluster MDL values. We performed the total MDL calculation with various quantization error values to understand how precision affects it.

We calculated value of the approximate MDL Eq. (6) for the climate data clusterings, generated by various algorithms, with the quantization error ε in the interval $[0.001, 0.002, \ldots, 0.01]$. Because the quantization error is a parameter for the MDL heuristic, we calculated separate clusterings for every ε value in the specified interval. Moreover, the quantization error affects the calculation of the compression ratio, thus the compression ratio parameter was selected different for every clustering calculation as well. We used the clustering, generated by the original LMCLUS algorithm, for calculation of an average μ and a standard deviation σ of a cluster compression ratio at every ε value. These statistics were used to bootstrap the compression ratio parameter for the MDL heuristic. The compression ratio set to $\mu - \sigma/2$ value for corresponding ε value.

Figure 6 shows that when MDL heuristics is enabled, it produced clusterings that are slightly different from the clustering generated by original method. We suspected that Eq. (6) does not appropriately reflect the MDL value of the the distribution of the data points in each of the clusters.

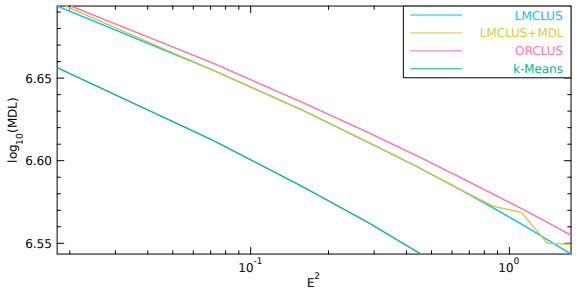

Fig. 6. Clustering optimal quantization MDL value (6) and its squared quantization error for ε in interval $[0.001, 0.002, \ldots, 0.01]$ and various algorithms.

When we switched to the "population" MDL Eq. (8) calculations in our MDL heuristics, with the parameters accordingly recalculated for this algorithm, performance of the clustering algorithm considerably improved.

It became clear that the effect of the MDL heuristics of resulting clustering, see Figure 7, are aligned with results from the synthetic simulation from section V-A. Increasing the precision of the linear manifold MDL calculation results in better goodness-of-fit qualities of clusters and allows the filtering of subpar cluster candidates during the LMCLUS stochastic search, which improves the final clustering.

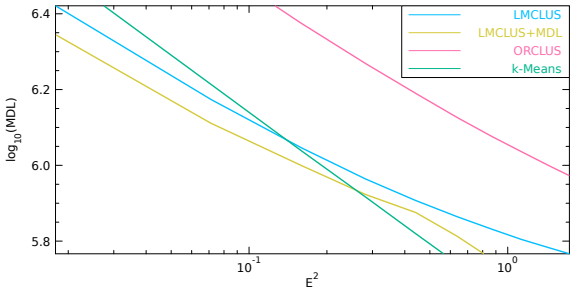

Fig. 7. Clustering population MDL value (8) and its squared quantization error for ε in interval $[0.001, 0.002, \ldots, 0.01]$ and various algorithms.

VI. Conclusion

We described a novel regularization technique for the linear manifold clustering based on the idea that a linear manifold shaped cluster allows efficient compression of the cluster data to the degree allowed by the specified error threshold. This intuitive criterion was formalized as the minimization problem of the description length of a prospective cluster and incorporated into the stochastic search of the clustering algorithm.

In the empirical part of the work we studied the behavior of the proposed MDL encoding, and the effect of the quantization error on it. We confirmed that the described method produced reasonable results for simulated datasets, as well as on the climate data clustering task. We believe that this regularization technique allows creation of clusters that are more informative and comprehensive.

A comprehensive scoring of the clusters with MDL values provides not only a criteria for cluster goodness-of-fit evaluation, but as well can be viewed as a qualitative measure which is used to explore stability of the clustering algorithm [15], or to improve clustering performance by introducing a scoring function for a guided stochastic search, which is left to be explored in our future research.

References

[1] H. Kriegel, P. Kröger, 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*, vol. 3, no. 1, p. 1, 2009.

[2] R. Haralick and R. Harpaz, "Linear manifold clustering in high dimensional spaces by stochastic search," *Pattern recognition*, vol. 40, no. 10, pp. 2672–2684, 2007.

[3] A. Kak, *Clustering Data That Resides on a Low-Dimensional Manifold in a High-Dimensional Measurement Space*, Purdue University, February 2016.

[4] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 430–437.

[5] Y. Wang, Y. Jiang, Y. Wu, and Z. Zhou, "Spectral clustering on multiple manifolds," *Neural Networks, IEEE Transactions on*, vol. 22, no. 7, pp. 1149–1161, 2011.

[6] P. R. Souvenir, R., "Manifold clustering," in *Tenth IEEE International Conference on Computer Vision*, vol. 1, 2005, pp. 648–653.

[7] V. R. Goh, A., "Clustering and dimensionality reduction on riemannian manifolds," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[8] R. Subbarao and P. Meer, "Nonlinear mean shift for clustering over analytic manifolds," in *Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2006, pp. 1168–1175.

[9] J. Rissanen and I. Tabus, "Kolmogorov's structure function in mdl theory and lossy data compression," *Advances in Minimum Description Length: Theory and Applications*, p. 245, 2005.

[10] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[11] O. Georgieva, K. Tschumitschew, and F. Klawonn, "Cluster validity measures based on the minimum description length principle," in *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2011, pp. 82–89.

[12] W. Köppen and R. Geiger, "Das geographische system der klimate," *Handbuch der Klimatologie.*, pp. 1–44, 1936.

[13] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World map of the köppen-geiger climate classification updated," *Meteorologische Zeitschrift*, vol. 15, no. 3, pp. 259–263, 2006.

[14] C. C. Aggarwal and P. S. Yu, "Finding generalized projected clusters in high dimensional spaces," *SIGMOD Rec.*, vol. 29, no. 2, pp. 70–81, May 2000.

[15] U. Von Luxburg and S. Ben-David, "Towards a statistical theory of clustering," in *Pascal workshop on statistics and optimization of clustering*, 2005, pp. 20–26.