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 clustering.

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 vectors 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 et 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.

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.

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 is larger the user specified. This criterion is fully explained in [2]. Second, the cluster compression ratio, defined as the ratio between the cluster description length, calculated using the linear manifold cluster model, and the raw description length of the cluster, is larger 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 $X = \{x_j \in \mathbb{R}^N \mid j = 1, \ldots, n\}$ be the points associated with the $M$-dimensional linear manifold cluster $M$. It is described by set of orthonormal basis vectors, that span linear manifold, $B = \{b_m \in \mathbb{R}^N \mid m = 1, \ldots, M\}$ and a translation vector $\mu \in \mathbb{R}^N$.

a) Model Encoding: The encoding of the translation vector $\mu$ 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 $M$ basis vectors are orthonormal, we can represent the basis vectors in less than $MN$ numbers. We can use a decoding schema that uses the orthonormal constraints in recovering the $M$ basis vectors. Each basis vector has norm 1. This constitute $M$ constraints. The orthonormality constraints specify another $M(M - 1)/2$ constraints. The total number of orthonormality constraints is then $M(M + 1)/2$.

To describe the linear manifold requires $N$ numbers for the offset of the manifold from the origin plus $MN - M(M+1)/2$ 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 + M(N - (M + 1)/2)] \tag{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 coded with \( P_d \) bits. Hence the encoding requires \( P_d \times M \) bits. Since the offset \( \mu \) lies in \( \text{col}(B) \), \( BB^T \mu = \mu \) and the reconstruction of that part of \( x \) that lies on the manifold is then given by \( \mu + B(B^T(x - \mu)) \).

Let \( B^{N \times N-M} \) 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 \( B(B^T(x - \mu)) \). This is a vector having \( K = N - M \) components. The reconstruction of that part of \( x \) that lies in the orthogonal complement space is given by \( B(B^T(x - \mu)) \).

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

\[ L(D|H) = n[P_d M + S(\varepsilon)] \tag{2} \]

where \( n \) is number of points in LM cluster \( \mathcal{M} \), \( S \) is an entropy of a distribution of cluster points, in the orthogonal complement subspace to the linear manifold of the cluster, calculated to be correct within an error bound of \( \varepsilon \).

We assume that each of the \( K \) 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 \( C \), s.t. \( \sum_{k=1}^{K} \log N_k = C \).

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

\[ N_k(C) = \left\lfloor A_k e^{(C - \sum_{j=1}^{K} \log A_j)/K} \right\rfloor \tag{3} \]

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,

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

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

\[ E^2(C) < \varepsilon^2 \tag{5} \]

It is not hard to show that the value \( C \) is defined over the interval

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

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

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

\[ C \log 2 = \sum_{k=1}^{K} \log_2 N_k \]

does have the meaning of bits.

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

\[ L(\varepsilon) = P_m[N + M(N - (M + 1)/2)] + n(P_d M + S(\varepsilon)) \tag{6} \]

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.

V. RESULTS

Finally, we would like to understand how well MDL evaluates 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 a primary dimension 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 constants, model and data, are 24 and 16.
From this experiment we observed that the MDL value calculated with 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 the high cluster MDL value. High values of ε will result in the low cluster MDL value. If the quantization error ε set to small value, the cluster MDL value will decreases monotonically with the dimension of cluster. If the quantization error ε set to large value, the cluster MDL value will increases monotonically with the dimension of cluster. If the quantization error ε set correctly, the cluster MDL value will decreases 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 arise 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

\[
L(\varepsilon) = P_nN + nS(\varepsilon) \quad (7)
\]

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 point position in particular dimensions, the projection distances to the point were encoded in MDL as follows

\[
L = L(H) + L(D(H)) = PN + \sum_{i=1}^{J} \sum_{p=1}^{N} \log(d_{i}^{p} + 1) \quad (8)
\]

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

We will compare the degenerate case of the inexact encoding of zero-dimensional manifold cluster calculated by (7) on synthetically generated LM and spherical clusters. Such approach will provide common ground for comparison between LM and spherical clusters. We also compare the MDL value of LM cluster with a cumulative MDL of clustering constructed from zero-dimensional clusters which is more natural representation of linearly shaped data from the perspective of spherical clustering algorithms.

We used synthetically generated dataset which has a form of a 1D linear manifold cluster, an elongated dataset along the one coordinated axis, in 2D full space. Cluster generation procedure was described above. We performed the MDL value calculation for 1D manifold following MDL formula (6) and then 0D manifold case defined by (7) for various quantization errors.

Figure 3 shows results of linear manifold (6) and zero-dimensional (7) MDL value calculations for various types of manifold clusters. For large quantization error, both approaches to the MDL calculation produce a small MDL value for spherical cluster. However, when precision of the quantization procedure increases, resulting in 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 k-means algorithm to synthesize spherical clusters from it. We varied number of clusters for k-means algorithm which allowed us to form clusters which gradually obtained a spherical shape, as the linear manifold cluster got partitioned on more clusters.

We perform evaluation of the MDL value for the linear manifold clusters by (6) and spherical cluster by (7). 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 4.

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 previous experiment, the major factor which affect MDL calculations is a quantization error parameter. For a small quantization error, spherical clusters provide smaller MDL value for the experiment dataset. More-
over, 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 (6) becomes significantly smaller than the spherical cluster MDL value (7).

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 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 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 from 1951 to 1980. Original datasets have the same $1^\circ \times 1^\circ$ resolution. We performed unit length normalization of datasets to archive the similar scale before combining them together in one dataset. Both datasets are 12 dimensional, so the combined dataset has 24 dimensions.

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 two different kinds of classifications on above climate dataset using $k$-Means and Linear Manifold Clustering methods.

$k$-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 Köppen-Geiger classes.

Linear Manifold Clustering (LMCLUS) takes multiple parameters, but we set only a small group of them, the rest of the parameters where 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 0.45, 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 estimate the MDL value of the cluster and calculate a compression ratio given that value. Based on the compression ratio, the heuristic decides whether to form cluster. If the compression ratio is larger than a user specified threshold, the stochastic search is continued. We set the quantization error parameter, mdl_quant_error, to 0.001.

Figure 5 show resulted clusterings plotted on the world map, where patches of the land are associated with particular clusters.

In order to compare 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 the goodness criteria. Figure 6 shows the effect of quantization error on the MDL value of clustering. It is clear that linear manifold clusters, which have intrinsic linear structure, show better goodness-of-fit qualities reflected in the smaller total MDL value than the spherical zero-dimensional clusters produced by $k$-Means algorithm.

### VI. Conclusion

We described a novel regularization technique for the linear manifold clustering based on the idea that a low-dimensional cluster it allows efficient compression of the data in the cluster to the degree allowed by the specified error threshold. This intuitive criterion was formalized as the minimization problem of the minimum description value 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 create 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 evalu-
Fig. 5. Results of clustering of the 24D climate dataset, composed of monthly averages of temperature (CRU) and precipitation (GPCC) during 1951-1980 period, by LMCLUS (a) and \textit{k}-Means (b) algorithms. \textit{Note: Colors represents associations of grid cells to particular clusters. There is no correspondence between colors on displayed plots.}

Fig. 6. Total MDL value for climate dataset clustering, produced by LMCLUS and \textit{k}-Means algorithms calculated for various user-set quantization error values.

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>LMCLUS</th>
<th>\textit{k}-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>23.0</td>
<td>23.5</td>
</tr>
<tr>
<td>0.01</td>
<td>23.5</td>
<td>24.0</td>
</tr>
<tr>
<td>0.001</td>
<td>24.0</td>
<td>24.5</td>
</tr>
<tr>
<td>0.0001</td>
<td>24.5</td>
<td>25.0</td>
</tr>
</tbody>
</table>

\textbf{REFERENCES}


