An Initial Study on High-Dimensional Data Visualization Through Subspace Clustering

A. Barbosa*, F. Sadlo** and L. G. Nonato*
*ICMC — Universidade de São Paulo, São Carlos, Brazil
**IWR — Heidelberg University, Heidelberg, Germany
barbosa@icmc.usp.br, filip.sadlo@iwr.uni-heidelberg.de, gnonato@icmc.usp.br

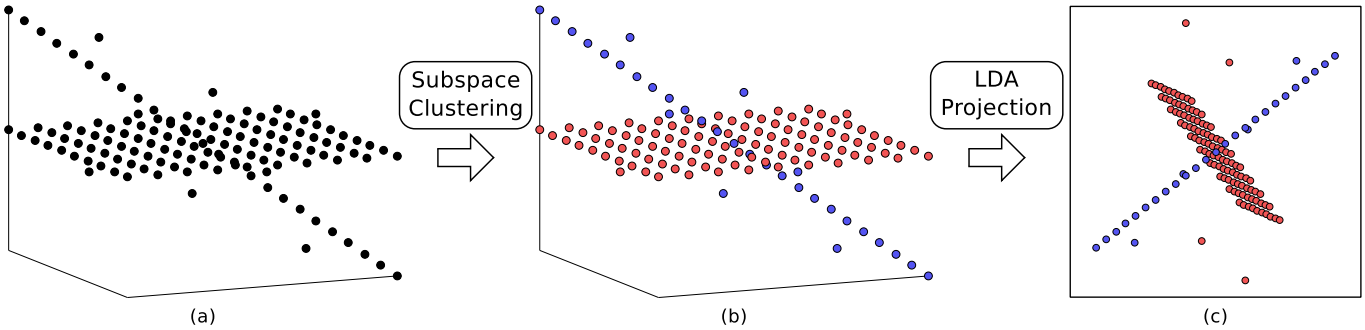


Fig. 1. Our pipeline: high-dimensional data without labels (a). Subspace clustering discovered the subspace structure within the data and segmented it (b). Data visualization using Linear Discriminant Analysis (LDA) (c).

***Abstract*—High-dimensional data are typically handled as laying in a single subspace of the original high-dimensional space. However, data involved in real applications are usually spread around in distinct subspaces which may even have different dimensions. An interesting aspect, that has not been properly exploited so far, is to understand whether information about the subspaces can be used to improve visualization tasks. In this work, we present a study where subspace clustering techniques are employed to detect data subspaces, and the information about such subspaces is used to assist visualizations based on multi-dimensional projection. Our results show that the information about subspaces can improve projection layouts, giving rise to more reliable visualizations.**

***Keywords*-Multidimensional Projection; Subspace Clustering;**

I. Introduction

High-dimensional data can be collected from a variety of sources, including field research, physical experiments, and image collections. This kind of data are usually described in terms of coordinates in a high-dimensional space. As we lack the ability of directly visualizing this data to understand their structure and arrangement, we resort to mathematical and computational methods such as multidimensional projection which are capable of processing and presenting high-dimensional data in a meaningful manner.

Conventional multidimensional projection techniques assume that the data are drawn from a single low-dimensional subspace of a high-dimensional space [1]. However, since the data are potentially drawn from multiple subspaces, we wish to take advantage of this fact. Subspace clustering techniques are capable of finding this subspace structure even when the number of subspaces and their dimensions are unknown.

In this work, we provide an initial study on how the label information given by subspace clustering techniques, such as the Low-Rank Representation (LRR) [2], can be used to aid multidimensional projection. We use the label information as input to Linear Discriminant Analysis (LDA) [3], [4], and compare the quality of these projections with previous techniques such as Local Affine Multidimensional Projection (LAMP) [5] and t-SNE [6]. We also propose a straightforward modification to LAMP to make use of the label information.

Contributions:

- We use subspace clustering techniques combined with visualization techniques for dimensionality reduction;
- We study the effectiveness of the labeling on the visualization by comparing previous multidimensional projection techniques with techniques that may take advantage of the label information;
- A rather straightforward modification of LAMP to make use of the label information.

A. Related work

In many problems data are represented as instances in a high-dimensional space, although they often lie in the neighborhood of subspaces with much lower dimension. Motivated by this fact, a variety of techniques has been developed to detect subspace structure. For example, Principal Component Analysis (PCA) [1] assumes that the data are drawn from a single linear space with dimension smaller than the original space. However, this assumption can be very restrictive. Subspace clustering techniques assume that the data are drawn from multiple independent linear subspaces, but the number

and the dimension of the subspaces, as well as the data membership, are unknown. When the number of subspaces is equal to one, the problem reduces to PCA. The goal of subspace clustering techniques is to find the number and the dimension of each subspace and then to segment the data according to the subspaces. These techniques can be divided into algebraic [7], [8], [9], iterative [10], [11], [12], statistical [13], [14], and spectral methods [2], [15]. A good tutorial on subspace clustering can be found in the work by Vidal [16]. LRR [2] aims to find a low-rank representation of the data by solving for a matrix with minimum rank, subject to some restrictions. We use this technique in our implementation.

The work by Liu et al. [17] uses LRR to compute the subspace clustering. The basis and dimension of each subspace are estimated involving the Grassmannian distance, allowing for an interactive visual exploration of the data trough dynamic projections. We, on the other hand, use the clustering given by LRR as input to LDA, wich produces a static projection. Other work related to subspace clustering in visualization use different approaches [18], [19], [20]. The work by Müller et al. [21] provides an overview on the employed techniques.

B. Technique Overview

Given a high-dimensional data set, we wish to visualize patterns and intrinsic structures of the data. Suppose the data are drawn from multiple independent linear subspaces. We use LRR to compute the subspaces and find the data membership. The next step is to use the subspace membership information found by LRR as labels for LDA, which projects data to a visual space based on the labels. LDA performs the projection using a linear mapping that separates data with distinct labels.

II. Technical Background

Here, we detail our visualization pipeline, comprising two main steps: subspace clustering [2] and LDA projection [4].

A. Subspace Clustering

Given a data set $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$, with n instances in a d-dimensional space, suppose the data can be decomposed as $X = X_0 + E_0$, where X_0 are the data drawn from the independent linear subspaces and E_0 is the "error" of the data, due to corruptions such as noise or outliers. The method aims to find a low-rank matrix X_0 from the given data set X corrupted by errors E_0. The solution is given by the following minimization problem:

$$\min_{Z,E} \text{rank}(Z) + \lambda \|E\|_l, \text{ s.t. } X = AZ + E, \tag{1}$$

where A is a "basis" for the space where the data lie, and $\|\cdot\|_l$ is a matrix norm that may vary depending on what kind of error we wish to filter.

As the problem (1) may not have a unique solution, the following problem is solved instead:

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_l, \text{ s.t. } X = AZ + E,$$

where $\|\cdot\|_*$ is the nuclear norm (sum of singular values).

When the data are affected by sample-specific corruptions or outliers, we can use the norm $\|E\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} \|E_{ij}\|^2}$, which is sensible to this kind of error.

The minimizer Z is a matrix with information about the data membership, and the non-zero entries of the matrix E represent data corruption. Let $Z = UDV^\top$ be the singular value decomposition of Z. To perform the segmentation, we compute the affinity matrix W defined by

$$W_{ij} = (\tilde{U}\tilde{U}^\top)_{ij}, \tag{2}$$

where $\tilde{U}$ is formed by $UD^{\frac{1}{2}}$ with normalized rows, and then apply the Normalized Cuts [22] clustering algorithm.

B. LDA Projection

Given data $\{x_1, x_2, \ldots, x_n\}$ with labels $\{y_1, y_2, \ldots, y_n\}$, let C_i denote the set of instances with label i, n_i the cardinality of C_i, and $\bar{x}_i = \frac{1}{n_i} \sum_{x_j \in C_j} x_j$ be the centroid of label i. From this, we can compute the "between classes scatter matrix" S_B and the "within classes scatter matrix" S_W, defined by

$$S_B = \frac{1}{n} \sum_{j=1}^{n} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^\top, \tag{3}$$

$$S_W = \frac{1}{n} \sum_{l=1}^{k} \sum_{x_j \in C_l} (x_j - \bar{x}_l)(x_j - \bar{x}_l)^\top, \tag{4}$$

where k is the number of classes, and $\bar{x}$ is the centroid (mean) of the entire data.

Maximization of the objective function

$$J(P) = \frac{P^\top S_B P}{P^\top S_W P} \tag{5}$$

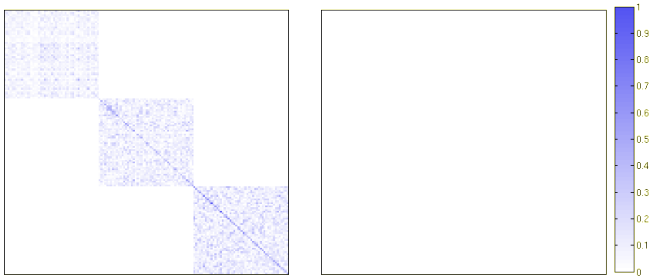
gives us a projection matrix such that the data projection has small variance in each class, and large variance between class centroids. This problem is solved by the generalized eigenvalue problem $S_B P = S_W P \Gamma$. Where the columns of P are the eigenvectors associated to the eigenvalues that form the diagonal matrix Γ.

We then use the first p columns of P (which are associated to the p absolutely largest eigenvalues) as the projection matrix to reduce the dimension of the data, mapping them to a p-dimensional space. For visualization, we usually use $p = 2$.

III. Experiments

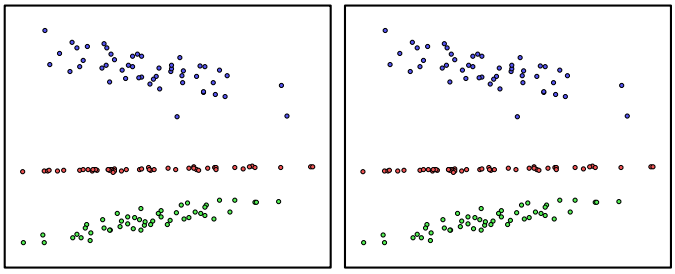
Artificial Data Set: We generated an artificial data set consisting of 50 instances of data drawn from $\mathbb{R}^3$, 50 instances drawn from $\mathbb{R}^7$, and 50 instances drawn from $\mathbb{R}^{10}$, and embedded all 150 instances in $\mathbb{R}^{30}$. The result of LRR, with $\lambda = 0.5$, is shown in Fig. 2. Notice that matrix E is empty, once the data set does have any instance with error.

Fig. 3 shows the projection of the data using LDA with labels given by LRR segmentation. Notice that as LRR performed the segmentation very well, the projections using the colors from the real class are identical to the segmentation.



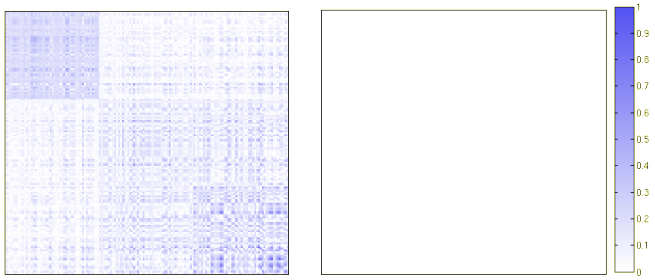
(a) Z matrix (b) E matrix

Fig. 2. LRR minimizer on Artificial data set.



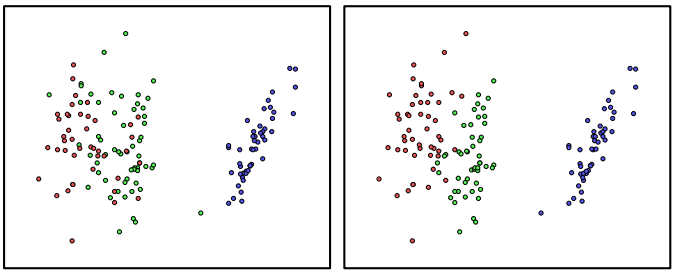
(a) Color given by segmentation (b) Real classes color

Fig. 3. LDA projection of Artificial data set.



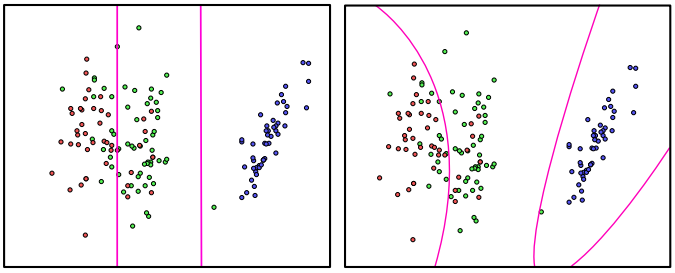
(a) Z matrix (b) E matrix

Fig. 4. LRR minimizer on Iris data set.



(a) Color given by segmentation (b) Real classes color

Fig. 5. LDA projection of the Iris data set.



(a) Fisher linear classification (b) Fisher quadratic classifier

Fig. 6. Fisher classification of the Iris data set.

Iris Data Set: For our second test, we use the well-known Iris data set [23]. It consists of 150 instances of dimension 4 divided in 3 classes, each one of the species of the flower iris. The results of LRR, with $\lambda = 0.5$, are shown in Fig. 4. Notice that one of the classes is well identified in the block diagonal matrix Z, while the other two classes are more difficult to distinguish. The matrix E is empty due to the nature of the data set and because we are using the norm $\|\cdot\|_{2,1}$, which is sensible to sample-specific corruptions.

As in the previous experiment, we projected the data using LDA with labels given by the LRR segmentation (Fig. 5a) and the real data set labels (Fig. 5b).

Using the Fisher classifier [4], we are also able to determine a linear (Fig. 6a) and quadratic (Fig. 6b) classifier that divide the space, allowing us to easily classify any new instance based on its position relative to the pink lines.

IV. Results, Discussion, and Limitations

We evaluate the method by comparing the projections generated by LDA with three other techniques: LAMP [5], t-SNE [6], and a modified version of LAMP. The modification we applied to LAMP is to add the label information for the computation of weights. Originally, we had $\alpha_i = \frac{1}{\|x - x_i\|^2}$, where x is the instance to be projected, and x_i is a control point. In the modified version, we have:

$$\alpha_i = \begin{cases} \frac{1}{\|x - x_i\|^2}, & \text{if } x \text{ and } x_i \text{ have the same label} \\ 0, & \text{otherwise.} \end{cases}$$

The quality of the projections generated by our approach is evaluated using four metrics: stress, neighborhood preservation, and two silhouettes. The stress function we use is given by $\frac{1}{\sum_{ij} d_{ij}} \sum_{ij} \left(d_{ij} - \bar{d}_{ij}\right)^2 / d_{ij}^2$, where d_{ij} is the distance in the original space and $\bar{d}_{ij}$ is the distance in the visual space. For each instance of data, the neighborhood preservation measures how many k-nearest neighbors in the original space are among the k-nearest neighbors in the visual space. The silhouette measures cohesion and separation between clusters. It is given by $\frac{1}{n} \sum_i \frac{b_i - a_i}{\max\{a_i, b_i\}}$, where a_i (the cohesion) is calculated as the average of the distances between a projected instance y_i (projection of x_i) and all other projected instances belonging to the same cluster as y_i, and b_i (the separation) is the minimum distance between y_i and all other projected instances belonging to other clusters. To compute the silhouette, we need to know the labels of the data. We use both, the real labels (silh1) and the labels given by LRR (silh2) to compute the silhouettes.

We use the data sets described in the Table I. Table II summarizes the results. Compared to LAMP, the modified

TABLE I

DATA SETS USED IN THE RESULTS, FROM LEFT TO RIGHT THE COLUMNS CORRESPOND TO THE DATA SET NAME, SIZE, DIMENSION (NUMBER OF ATTRIBUTES), AND SOURCE.

Name	Size	Dim	Source
Iris	150	4	[23]
Synthetic	150	4	[24]
Artificial	150	30	*
Wine	178	13	[23]
Mammals	1000	72	[23]

TABLE II

RESULTS, FROM LEFT TO RIGHT THE COLUMNS CORRESPOND TO THE DATA SET NAME, TECHNIQUE AND METRICS: STRESS, NEIGHBORHOOD PRESERVATION, AND SILHOUETTES. BOLD VALUES ARE THE BEST FOR EACH DATA SET AND METRIC.

Data set	Technique	Stress	NP (%)	Silh1	Silh2
Iris	LAMP	**0.0418**	81.8	0.6371	0.3437
	LAMP (M)	0.0791	77.8	0.6032	0.4221
	LDA	0.3095	63.6	0.6889	**0.6758**
	t-SNE	1.71e+6	**86.9**	**0.7633**	0.3392
Synthetic	LAMP	0.0597	80.9	0.8584	0.8584
	LAMP (M)	**0.0521**	82.0	0.9045	0.9045
	LDA	0.0862	85.7	0.9299	0.9299
	t-SNE	6.2266	**89.4**	**0.9956**	**0.9956**
Artificial	LAMP	**0.0539**	85.4	0.6770	0.6770
	LAMP (M)	0.0749	86.0	0.7787	0.7787
	LDA	0.3749	81.2	**0.9492**	**0.9492**
	t-SNE	0.2962	**90.9**	0.8961	0.8961
Wine	LAMP	**0.0383**	90.7	0.2174	0.3629
	LAMP (M)	0.1371	86.9	0.2269	0.4491
	LDA	0.9802	53.3	0.2694	**0.5314**
	t-SNE	0.9312	**94.3**	**0.3139**	0.4262
Mammals	LAMP	**0.0112**	**87.9**	0.9825	0.9825
	LAMP (M)	0.0172	85.5	**0.9924**	**0.9924**
	LDA	1.0000	81.4	0.9311	0.9311
	t-SNE	0.3829	87.7	0.9653	0.9653

version of LAMP performs better in terms of silh2 (with labels given by LRR), with a small difference in terms of stress. While the stress of LDA is bigger than LAMP and modified LAMP (which is expected, because the objective of LDA is to find the subspace with better separability between classes), it gives a good result in terms of silh2. The results of LDA indicate that the combination of LRR and LDA can be a good choice for dimensionality reduction and unsupervised classification problems where the true label is unknown.

A. Limitations

Subspace clustering techniques assume that the data are drawn from independent subspaces, but this may not be always true in real world data sets. We have extensively tested some examples for many parameters, with no success in finding any reasonable subspace structure. In these cases, we assume that such a subspace structure does not exist and thus that the method cannot be applied properly.

V. CONCLUSION

In this paper, we started the study of visualization aided by subspace clustering. Subspace clustering techniques have been shown to be a promising way to account for the possible intrinsic subspace structure of data. On the other hand, for data with no subspace structure, they cannot be applied properly. The use of subspace clustering allows us to use LDA to perform dimension reduction and classification tasks with good quality in terms of the metrics we have tested.

As future work, we plan to study the estimation of the dimension and basis for each subspace, and how it can be applied to aid multidimensional projection.

ACKNOWLEDGMENT

The authors acknowledge financial support from CAPES.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, ser. Computer science and scientific computing. Elsevier Science, 2013.

[2] H. Zhang, Z. Lin, C. Zhang, and J. Gao, "Robust latent low rank representation for subspace clustering," *Neurocomputing*, vol. 145, pp. 369–373, 2014.

[3] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[4] G. Seber, *Multivariate Observations*, ser. Wiley Series in Probability and Statistics. Wiley, 2009.

[5] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato, "Local affine multidimensional projection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2563–2571, 2011.

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[7] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Proceedings of the IEEE Workshop on Visual Motion*, 1991, pp. 179–186.

[8] J. Costeira, T. Kanade, and M. A. Invariants, "A multi-body factorization method for independently moving objects," *International Journal of Computer Vision*, vol. 29, pp. 159–179, 1997.

[9] C. W. Gear, "Multibody grouping from motion images," *International Journal of Computer Vision*, vol. 29, no. 2, pp. 133–150, Aug. 1998.

[10] P. K. Agarwal and N. H. Mustafa, "K-means projective clustering," in *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '04. New York, NY, USA: ACM, 2004, pp. 155–165.

[11] P. S. Bradley, O. L. Mangasarian, and P. Pardalos, "k-plane clustering," *Journal of Global Optimization*, vol. 16, no. 1, pp. 249–252, 2000.

[12] P. Tseng, "Nearest q-flat to m points," *Journal of Optimization Theory and Applications*, vol. 105, no. 1, pp. 249–252, 2000.

[13] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[14] H. Derksen, Y. Ma, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy coding and compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, pp. 1546–1562, 2007.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.

[16] R. Vidal, "A tutorial on subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2010.

[17] S. Liu, B. Wang, J. J. Thiagarajan, P.-T. Bremer, and V. Pascucci, "Visual Exploration of High-Dimensional Data through Subspace Analysis and Dynamic Projections," *Computer Graphics Forum*, pp. 271–280, 2015.

[18] A. Tatu, F. Maas, I. Farber, E. Bertini, T. Schreck, T. Seidl, and D. Keim, "Subspace search and visualization to make sense of alternative

clusterings in high-dimensional data," in *IEEE Conference on Visual Analytics Science and Technology*, 2012, pp. 63–72.

[19] J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf, "Bicluster viewer: a visualization tool for analyzing gene expression data," in *Advances in Visual Computing*. Springer, 2011, pp. 641–652.

[20] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," *Computer Graphics Forum*, vol. 31, no. 3, pp. 1335–1344, 2012.

[21] E. Müller, S. Günnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data," vol. 2, no. 1. VLDB Endowment, 2009, pp. 1270–1281.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.

[23] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[24] I. Guyon, "Design of experiments of the NIPS 2003 variable selection benchmark," in *NIPS 2003 workshop on feature extraction and feature selection*, 2003. [Online]. Available: http://www.nipsfsc.ecs.soton.ac.uk/datasets/