CODEC - Detecting Linear Correlations in Dense Clusters using coMAD-based PCA

Maximilian Archimedes Xaver Hünemörder, Daniyal Kazempour, Anna Beer, and Thomas Seidl

Ludwig-Maximilians-Universität München, Munich, Germany
{huenemoerder,kazempour,beer,seidl}@dbs.ifi.lmu.de

Abstract. The coMAD (co-median absolute deviation) is a measure for the joint median of two random variables. Previous experiments have shown that a coMAD-based PCA is more robust towards noise and outliers, yielding eigenvectors which represent linear correlation better than its covariance-based competitors. In this preliminary work we introduce CODEC - COrrelations in DEnse Clusters - a method for detecting linear correlations in dense clusters utilizing a coMAD-based PCA. The idea of CODEC is intriguingly simple: first a density-based clustering is performed using the well established clustering method DBSCAN. Then on each of the clusters PCA is performed. Instead of using the covariance matrix we use the coMAD matrix as a basis for performing PCA.

Keywords: Comedian · Correlation Clustering · Principal Component Analysis · CoMAD.

1 Introduction

As Data Analysis is becoming more and more important in recent years, a multitude of possible interesting properties of datasets have emerged. Clustering, which constitutes a huge field of data analysis with diverse sub-categories, leverages these properties to find sets of points which are similar to each other, but dissimilar to points of other sets or clusters. This similarity can be based on density, on the distance to centroids, or how well they fit to a certain distribution. On the other hand, points which are correlated only in certain dimensions can also be interpreted as similiar, which is covered by subspace and correlation clustering algorithms. Surprisingly only few algorithms combine both concepts, and, to the best of our knowledge, none of them investigate found clusters further regarding possible correlations. Especially density-based clusters, which can be of any shape, can contain correlated data (rather than centroid-based clusters for example, which tend to have similar extensions in every dimension). Thus, we introduce CODEC, a new prototype algorithm which finds correlations in density-based clusters found by DBSCAN using an improved version of PCA to

account for dispersions. We have found that using the coMAD matrix instead of the covariance matrix, we find less distorted main components of correlation clusters.

We provide an overview over related work in Section 2, including an introduction of the improved PCA using the coMAD (co-median absolute deviation) matrix. In Section 3 the algorithm is explained in detail and tested in Section 4. Section 5 concludes this short paper and gives ideas for future work.

2 Related Work

There already exist several sophisticated methods which detect linear correlated clusters, such as ORCLUS[2], 4C[3] or CASH[1]. ORCLUS and 4C rely on Principal Component Analysis (PCA). ORCLUS combines the PCA with k-Means and 4C uses PCA and a DBSCAN[4]-like approach. In this work-in-progress, we aim to harness the robustness of the coMad, orginally introduced as the "comedian" in [5] to detect linear correlations in dense clusters.

3 Method

Since the coMAD is one of the core elements of our method, we will first elaborate on the definition of the coMAD. Suppose we are given a data matrix D of dimensionality d, where the rows represent a data record and their respective columns represent the features $(A_1, ..., A_d)$. If we consider the variance, its analogon in the median context would be the median absolute deviation from the median (MAD):

$$mad(A_i) = med(|A_i - med(A_i)|)$$

Then the analogon to the covariance is the coMAD which is a generalization of the MAD:

$$com(A_i, A_j) := med((A_i - med(A_i))(A_j - med(A_j)))$$

Finally we use the definition of the coMAD to construct the coMAD matrix Λ known as:

$$\Lambda_D = \begin{pmatrix} mad(A_1) & \cdots & com(A_1, A_d) \\ \vdots & \ddots & \vdots \\ com(A_d, A_1) & \cdots & mad(A_d) \end{pmatrix}$$

Based on the coMAD matrix Λ the PCA is performed which yields the corresponding eigenpairs.

Our algorithm then proceeds as follows: First, dense clusters are detected by applying DBSCAN on the data set. Then on each of the dense clusters PCA is applied, using the coMAD instead of the covariance matrix. As a result we obtain a set of eigenvectors for each cluster, which show the direction of linear

correlations within the clusters. The intuition here is that instead of the direction of highest variance (which would be the classical result using the covariance matrix), the eigenvectors now point in the direction of the highest MAD, therefore the direction where most points are situated. This leads to the PCA being less contaminated by noise points that diverge from the main direction of linear correlation.

One may think that a dense cluster should already be almost without any noise. This however depends on the density of the clusters which is implicitly determined by the choice of the hyperparameters of DBSCAN, namely ε-range and *minpts* for the minimum number of objects required to be located within an ε-range. The larger the ε-ranges and the lower *minpts*, the less dense are clusters. Further the so called 'border points' can, depending on the ε-range, have similar effects as outliers and therefore skew the principal components of a PCA. In Section 4 we show a case where PCA based on the covariance matrix yields skewed results compared to PCA using the coMAD matrix.

4 Experiments

As a first experiment of our preliminary work, we constructed a data set with four clusters exhibiting linear correlations, both locally and globally, as it can be seen in Figure 1 (top left). Furthermore, each of these clusters are not perfectly linearly correlated but studded with noise. From a high-level view one could state that the noise should not have any impact on the result of the PCA. However, if we apply a covariance-based PCA on each of the dense clusters, the resulting eigenvectors are significantly skewed, as it can be seen in Figure 1 (top right). Especially in the blue cluster the noise leads to a massive distortion of the expected direction. The effects of using a coMAD-based approach become visible in Figure 1 (bottom), where despite the noise the detected eigenvectors remain robust. The algorithm and data generator used for the experiment were implemented in python and are publically available [1].

5 Conclusion and Future Work

In summary we developed a method to find correlations in density-based clusters accurately and robust to noise and jitter. We showed that using the coMAD matrix for PCA delivers more intuitive results than using the covariance matrix, especially for real-world data which is usually not correlated perfectly. We plan to examine further combinations of correlation clustering and density-based clustering in future work. Investigating the trade-off between efficiency of the computation and improvement of the results using the coMAD matrix is a further subject of future work.

[1] https://github.com/huenemoerder/CODEC

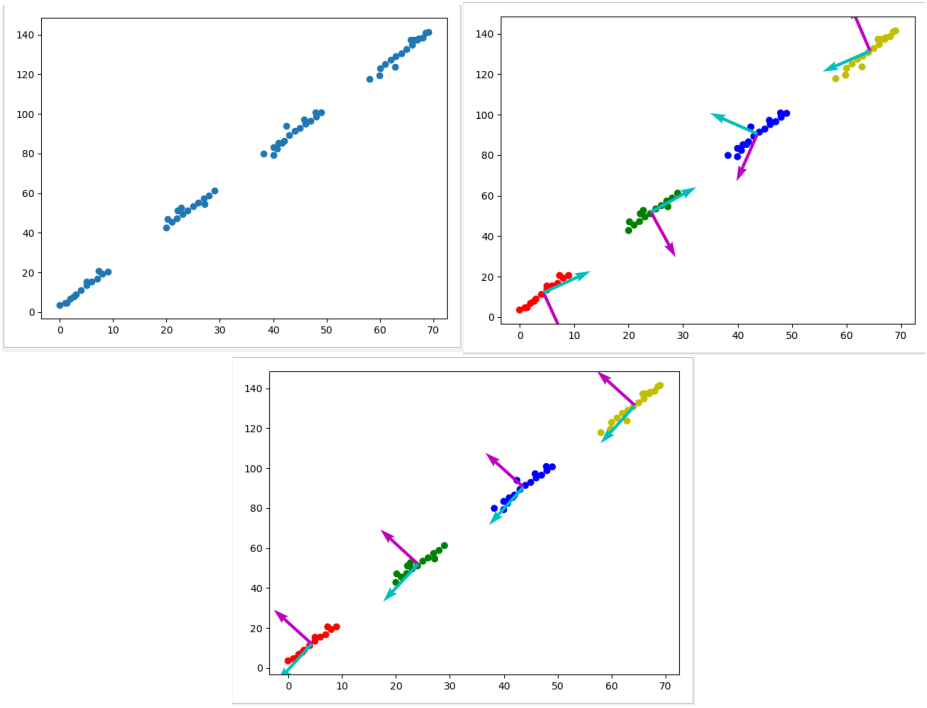

Fig. 1. Top left: test data set; top right: the computed eigenvectors with a covariance-based PCA; bottom: the computed eigenvectors with a coMAD-based PCA

References

1. Achtert, E., Böhm, C., David, J., Kröger, P., Zimek, A.: Global correlation clustering based on the hough transform. Statistical Analysis and Data Mining: The ASA Data Science Journal **1**(3), 111–127 (2008)
2. Aggarwal, C.C., Yu, P.S.: Finding generalized projected clusters in high dimensional spaces, vol. 29. ACM (2000)
3. Böhm, C., Kailing, K., Kröger, P., Zimek, A.: Computing clusters of correlation connected objects. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data. pp. 455–466. ACM (2004)
4. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
5. Falk, M.: On mad and comedians. Annals of the Institute of Statistical Mathematics **49**(4), 615–644 (1997)