From Covariance to Comode in context of Principal Component Analysis

Daniyal Kazempour, Long Mathias Yan, and Thomas Seidl

Ludwig-Maximilians-Universität München, Munich, Germany
{kazempour,seidl}@dbs.ifi.lmu.de, l.yan@campus.lmu.de

Abstract. When it comes to the task of dimensionality reduction, the Principal Component Analysis (PCA) is among the most well known methods. Despite its popularity, PCA is prone to outliers which can be traced back to the fact that this method relies on a covariance matrix. Even with the variety of sophisticated methods to enhance the robustness of the PCA, we provide here in this work-in-progress an approach which is intriguingly simple: the covariance matrix is replaced by a so-called comode matrix. Through this minor modification the experiments show that the reconstruction loss is significantly reduced. In this work we introduce the comode and its relation to the MeanShift algorithm, including its bandwidth parameter, compare it in an experiment against the classic covariance matrix and evaluate the impact of the bandwidth hyperparameter on the reconstruction error.

Keywords: Covariance · Comode · Principal Component Analysis.

1 Introduction

In cases where we have to deal with high-dimensional data, a common strategy is to perform a Principal Component Analysis (PCA)[5]. A PCA yields eigenpairs consisting of eigenvectors and their corresponding eigenvalues. In the use-case of dimensionality reduction, projecting given data down to the eigenvectors with the top-k eigenvalues, comes with a reconstruction error when projecting it back to the full dimensional data. This reconstruction error decreases if more principal components are incorporated. Nevertheless, despite using more principal components, the reconstruction error can still be significantly high. Outliers can heavily skew the results which is originated in the mere observation that an outlier can skew the mean for each of the features of a data set. More robust measures of central tendency are the median and the mode. In this work we propose the so-called comode matrix as an alternative to the covariance matrix on which the eigenvalues and eigenvectors are computed in a PCA. Our contribution in this work-in-progress shows that it is more robust towards outliers, but at the same time dependant on the choice of a hyperparameter known as bandwidth.

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2454/paper_65.pdf

2 Related Work

Many efforts have been made to make the PCA more robust towards noise. As such in the work of [6] the authors describe a robust M-estimation algorithm for capturing multivariate representations of high-dimensional data exemplary on images, known as Robust Principal Component Analysis. The authors elaborate that methods such as RANSAC[3] and Least Median Squares would be more robust compared to M-estimation yet remain unclear in their application on high-dimensional data. In a classical approach by Candes et al. [1] the outliers are modeled by a sparse matrix based on the assumption that the data matrix can be expressed in terms of a sum of a low rank and a sparse matrix. The so far mentioned methods rely on complex and sophisticated methods. We challenge the task of robust PCA by asking: What if we simply replace the covariance matrix by a comode matrix? Since the mode is insensitive against noise, so should be the comode according to our hypothesis.

3 Comode

Given a data matrix D where each of its rows represents a data record and its columns represent the features $(A_1, ..., A_d)$. When performing a PCA, first a covariance matrix Σ is computed. The covariance is a generalization of the variance. For computing the covariance from every feature, the expected value E (mean) is subtracted. Since the mean is sensitive towards outliers, the computation of the covariance matrix is also sensitive against outliers. A more robust measure is the mode. Generalizing the mode like variance is generalized to the covariance leads to the comode:

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

Building on the definition of the comode, we define a comode matrix Ω:

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

But how do we actually compute $com(A_i, A_j)$ which represents the mode of a variable A_i in dependence of another variable A_j? One solution we propose here relies on the so called MeanShift[2] algorithm. MeanShift is a method which was primarily used for locating the maxima, which are basically the modes of a given density function. It is well known as being a mode-seeking algorithm and its applications have been extended to tasks like cluster analysis. For any given data object x the shifted data object at one iteration is computed as:

$$m(x) = \frac{\sum_{i=1}^{N(x)} K(x - x_i)x_i}{\sum_{i=1}^{N(x)} K(x - x_i)} \tag{1}$$

where $K(X) = k\left(\left\|\frac{x}{\sigma}\right\|^2\right)$ defines a kernel and N(x) are all objects in the neighborhood of x within a bandwidth σ. The bandwidth follows the intuition of an ϵ-range. It defines the local neighborhood of an data object x. The *mean shift* of x is the difference $m(x) - x$. At this point it may still be unclear how the MeanShift may lead to the detection of the comodes? A comode $com(A_i, A_j)$ is determined by computing the MeanShift $\mu_{shift}(D_{A_i,A_j})$ on a given dataset D considering only its features (= random variables) A_i and A_j.

The comode matrix can thus be expressed in terms of MeanShift computations:

$$com(A_i, A_j) := mode((A_i - mode(A_i))(A_j - mode(A_j)))$$
$$= \mu_{shift}((A_i - \mu_{shift}(D_{A_i,A_j}))(A_j - \mu_{shift}(D_{A_i,A_j})))$$

What remains unclear so far is: how do we approach the fact that there can be not only one (unimodal) but several (multimodal) modes? In this work-in-progress we have focused on taking the mode with the highest frequency, meaning we take the top mode which has the highest number of objects belonging to it. In the case where we have several equally high top-modes, we randomly select one of them. Due to the brevity of this paper, we pursue it in future work to elaborate and evaluate the impact of choosing different top-modes.

4 Experiments

In our experiments we use the MeanShift with a flat kernel from sklearn[1]. For reproducibility purposes, the code is made available on github[2]. We conduct our experiments on the iris dataset[3] which consists of 150 instances and 4 features. We compute the PCA using the comode instead of the covariance matrix. We discard all principal components from the eigenvector matrix U except the first one, and project the data D down to its lower (one)dimensional representation Y through: $Y = D \cdot U_{k=i}$. We reconstruct the data by projecting it back to its original full-dimensional representation through: $Z = Y \cdot U_{k=i}$. By projecting the data down and back again, enables us to compute the Mean Squared Error (MSE) which is defined as $MSE(D, Z) = \frac{\sum_{i=1}^{n}(d_i - z_i)^2}{n}$ where $d_i \in D$, $z_i \in Z$ and n denoted the number of objects for which holds $n = |D| = |Z|$. We apply this procedure for taking the second, third,...,d-th principal component. Since we want to investigate the effect of the choice of the bandwidth we further perform the comode computation choosing the bandwidths $\sigma = (0.1, 4.0, 5.5)$. The results can be seen in Figure 4 where the horizontal axis represents the principal components in descending order of their corresponding eigenvalues. The vertical axis represents the MSE. It can be seen from Figure 4 that PCA with comode and a bandwidth of $\sigma = 0.1$ yields an MSE (~ 1) which is significantly below that of the classical PCA with a covariance matrix (~ 7) taking only the first

[1] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html
[2] https://github.com/hamilton-function/comode
[3] https://archive.ics.uci.edu/ml/datasets/iris

principal component. By taking additionally the second principal component the MSE of the covariance-based PCA drops even slightly below the MSE of the comode variant with $\sigma = 0.1$ which holds for taking three principal components as well. A larger bandwidth of $\sigma = 4.0$ yields to MSE values which exhibit a course similar to that of the covariance-based PCA but with overall higher MSE values. At this point it is interesting for further research to investigate if there is an 'even'-point regarding the bandwidth where the MSE for different number of principal components equals the MSE of the covariance-based PCA. A bandwidth of $\sigma = 5.5$ yields MSE results which are by far worse compared to the other settings.

However, we have to take the MSE results with a grain of salt for the following observation which has been made in [4]. Here the insights were that a low MSE (or in that work: MAE) does not imply a high robustness towards outlier. In fact, the opposite is the case: a low MSE indicates that even outliers can be reconstructed well, which in turn means that the computed principal component is distorted by the outlier. In contrast, a high MSE can indicate that a principal component has been computed which does not take outliers into account, giving therefore more weight to the majority of the objects following a direction with largest variance. As we stated a high MSE *can* indicate a better principal component. For this purpose in future work it is vital do develop methods by which the reconstruction without the outliers is computed. With the bandwidth we have a control unit to tune the comode computation such that either the MSE is minimized, or the overall deviation from the principal component is minimized.

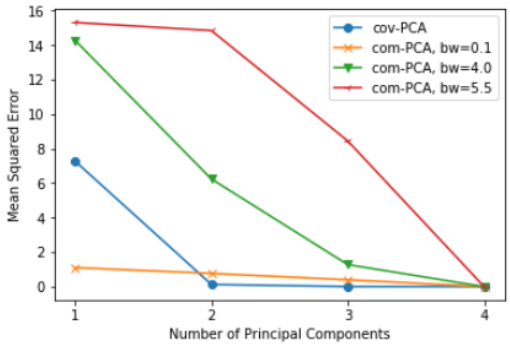

Fig. 1. MSE with increasing number of principal components on the iris data set. orange: comode with $\sigma = 0.1$; green: comode with $\sigma = 4.0$; red: comode with $\sigma = 5.5$; blue: covariance variant

5 Conclusion and Future Work

In this work-in-progress we have presented the Comode in context of PCA. A first experiment showed promising results, outperforming a covariance-based PCA while being intriguingly simple. There are however interesting aspects which demand further research: first and foremost, it remains an open question on how to deal with multimodal cases. What are the effects on the PCA if we choose the second or third strongest mode(s)? How are good bandwidths determined? For these aspects we may seek previous works which aimed at estimating good bandwidths for MeanShift. Further it is of interest to investigate if feature-specific bandwidths have any impact on the robustness of the Comode-based PCA. As for now, we have one bandwidth which is valid for all features, neglecting feature-specific bandwidths. We hope to stimulate further research on the Comode, revealing its limitations as well as its potentials.

Acknowledgement

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

References

1. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM (JACM) **58**(3), 11 (2011)
2. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence **17**(8), 790–799 (1995)
3. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
4. Kazempour, D.; Hünemörder, M.A.X.; Seidl, T.: On coMADs and Principal Component Analysis. SISAP 2019 - Springer Lecture Notes in Computer Science (2019)
5. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901)
6. De la Torre, F., Black, M.J.: Robust principal component analysis for computer vision. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 1, pp. 362–369. IEEE (2001)