Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

To Interpret or Not to Interpret PCA? This Is Our Question

Dan Vilenchik,[1] **Barak Yichye,**[1] **Maor Abutbul**[1]
[1] Ben-Gurion University of the Negev, Israel
vilenchi@bgu.ac.il, maor.abut@gmail.com, barak.yichye@gmail.com

Abstract

Principal Component Analysis (PCA) is a central tool for analyzing data and social media data in particular. Typically, the data is projected on the first two PCs to obtain a two-dimensional view, and trends and patterns are being examined. A key to making sense of the projected data is the semantic interpretation of the new axes (the PCs). To label the PCs, one usually looks at the top k vector entries in absolute value and assigns meaning according to them. The choice of k is done by "eyeballing" the vector. In this work we provide a computational framework to support this process and suggest an *interpretability score*, which measures how sensitive the interpretation step could be to the choice of k. Furthermore we give a visual method to choose the optimal k. We study our methodology in four social media platforms and discover that in two of them, Twitter and Instagram, interpretation can be done in a carefree manner, but in Steam and LinkedIn there is no natural labeling of the axes. This separation is clearly reflected in the interpretability score that each dataset received.

1 Introduction

Computing Principal Component Analysis (PCA) is one of the first steps that a researcher performs when studying data, and especially when the data is high-dimensional. The data is typically projected on the leading two PCs and a search for patterns or trends takes place. The crux of PCA is the labeling of the new axes in order to make sense of the observed patterns. In social media data, the axes may point in directions that correspond to characteristics of users, e.g. a measure of popularity, or to modes of behaviour and interaction, e.g. a measure of botness/spam, or a measure of content activity (Canali, Casolari, and Lancellotti 2012; Viswanath et al. 2014; Vilenchik 2019).

Unfortunately, there is no golden rule for labeling the PCs, but rather this step is performed using intuitive and ad-hoc choices. The most widely used method is probably the "interpret-by-top-k" rule. The rule says first sort the PC vector entries in descending order of absolute values, then assign the PC its label according to the top k features, ignoring entries with smaller values. While this practice is useful in many cases, the choice of k is subjective and may affect the interpretation. In addition, choosing smaller k values makes interpretation easier, as fewer features are involved, but possibly at the cost of semantic validity.

To exemplify the trickiness in choosing the right k let us consider an example using data that we collected from four social media platforms: Twitter, Instagram, LinkedIn and Steam. Applying the top-k rule to the leading PC in Steam gives the following result. Up to $k = 3$ the label is "measure of gaming proficiency" (top three features are: number of games owned by a user, Steam experience, and number of badges owned by the user). At $k \geq 4$ this semantic meaning blurs as features that have nothing to do with proficiency are added (e.g. the number of friends, the number of groups the user belongs to). In fact we could not find a natural simple label for the leading PC according to the top $k \geq 4$ features. We checked whether the heavy hitters in the direction of the leading PC can give a clue as to its label. This set turned out to be a mix of very popular yet not very active users and vice versa, and the spectrum in between. Therefore one could confirm different labels (or none) depending on the users that were sampled. A similar phenomenon occurred in our LinkedIn dataset. The interpret-by-top-$k = 4$ rule would label the leading PC as a certain measure of "professional activity" (number of jobs, degrees, certifications, volunteering projects) , however at $k = 5, 6$ two features that measure feedback from others join in (the number of skills - as endorsed by others, and the number of groups that the user was authorized to join).

1.1 Our contribution.

We develop a toolbox to assist the interpret-by-top-k rule by defining an *interpretability score* for every PC. The interpretability score is a number between 0 and 1 which measures how sensitive the interpret-by-top-k rule is to the choice of k. A low interpretability score will indicate that it is not advisable to use this rule to interpret that PC. To compute this score we draw on the following intuition: suppose that the PC has only one non-zero entry, then interpret-by-top-k would work perfectly for any k. More generally, the sparser the PC the less probable it is that subsets of non-zero features will have a different meaning than the full set of non-zero features. In reality, if only for numeric issues, all PC entries will typically be non-zero. The interpretability score may be viewed as a certain normalized sparsity level

of the vector.

The way we suggest to compute the interpretability score, explained in Section 3, produces an *interpretability curve* which is similar in spirit to a scree plot, see Figure 1. One can use Cattell's popular cut-at-the-knee rule to chose the optimal value of k for the interpret-by-top-k rule. By optimal we mean the smallest semantically-consistent k.

We applied our toolbox to data that we collected from Twitter, Instagram, LinkedIn and Steam. Our full findings are given in Section 4, here we review the highlights. We found that the interpretability score of the leading two PCs in Twitter (0.89,0.81) and Instagram (0.85,0.85) was considerably higher than Steam (0.77, 0.7) and LinkedIn (0.76,0.76). As a baseline, the interpretability scores of the leading two PCs of Gaussian white noise were (0.7, 0.65). This matched nicely with the fact that it was straightforward to interpret the leading two PCs in Twitter (a measure of popularity and a measure of botness) and Instagram (a measure of popularity,and a certain measure of posting activity). On the other hand, in Steam and LinkedIn, we could not assign a natural simple label to the leading PCs. Furthermore, the interpretability curve, Figure 1, pointed clearly to $k = 3$ as the right choice for Twitter and $k = 4$ for Instagram. For Steam and LinkedIn there was no knee in the curve to cut.

We compared our results against the top-k rule by using it to compute an interpretability score. For Twitter we obtained very similar results (0.89,0.83) but for Steam we obtained (0.77,0.78), which is similar for the first PC but much higher for the second. Results obtained in (Vilenchik 2019) confirm that the lower score, 0.7, is a better estimation of interpretability for the second PC in Steam, than 0.78. Details in Section 4.3.

2 Related Work

The problem of user characterization in online social media platforms is typically approached as a supervised learning classification problem. For example (Pennacchiotti and Popescu 2011) try to predict the user's ethnicity and political affiliation, (Rao et al. 2010) deal with gender, age, regional origin and (Preotiuc-Pietro, Lampos, and Aletras 2015) with occupational class.

Far less was done using unsupervised learning approaches. (Eirinaki, Monga, and Sundaram 2012) suggested a PageRank-like measure which they call ProfileRank. ProfileRank is computed as some learned linear combination of various user statistics. They tested their score on Facebook and MySpace and showed how the rank can be used to identify influential users. A PCA-based approach was suggested by (Canali, Casolari, and Lancellotti 2012) to characterize users in YouTube and Flickr. Their main result is that the top PCs encode labels that correspond to measures of popularity and activity in the network. In fact the PCA-based approach may be viewed as a ProfileRank measure in which the weights of the linear combination of features are set according to the PCs.

PCA was also used successfully in a closely related task – *anomaly detection*. Viswanath et al. (Viswanath et al. 2014) used PCA in order to classify Facebook users as either "normal" or "anomalous" (user is considered anomalous if its behaviour was tagged as such by Facebook).

The validity of the PC labeling in (Canali, Casolari, and Lancellotti 2012) was done against the results of other algorithms and in (Viswanath et al. 2014) against a given ground truth (labeled data). We are not aware of any work that included a self-contained mechanism to quantify the validity of the interpretation step.

3 Methodology

Our working hypothesis is that the sparser the PC the easier it is to interpret it, and the less sensitive the interpretation will be to the number of features that take part in the process. The interpretability score, which is a number in $[0, 1]$, may be viewed as a normalized sparsity level of the vector, where values closer to 1 represent the fact that the vector may be safely regarded as a sparse vector for the sake of interpenetration, hence using the top-k rule is less sensitive to the exact choice of k.

In reality all the entries of the PC will be non-zero, if only for numeric reasons, but perhaps some entries will be much smaller than others, and it is safe to treat them as zeros. It is desirable that zeroing out entries will be done in a rigorous manner rather than a threshold decided ad-hoc and subjectively. It is also informative to measure how much information was lost when zeroing out entries.

At the basis of our approach lies a variant of PCA called *sparse PCA*. In the sparse PCA problem, rather than finding the leading eigenvector of the covariance matrix, one looks for the unit vector $\mathbf{v}$ with at most k non-zero entries, k is fixed in advance, such that the variance in the direction of $\mathbf{v}$ is maximal; $\mathbf{v}$ is called the leading k-sparse eigenvector (although it is not necessarily an eigenvector of the covariance matrix). The remaining k-sparse eigenvectors are computed in a similar way to the standard PCA.We used R's `nsprcomp` library to compute sparse PCA.

Given a PCA solution $\mathcal{P}$ of a p-dimensional dataset, the first task is to decide which PCs will be considered for interpretation. If a 2D plot is desired then the answer is usually the leading two PCs. More generally, the Guttman-Kaiser (GK) criterion is typically used to discard PCs that explain less than $1/p$-fraction of the variance. Let r be the number of candidate PCs.

3.1 The interpretability curve

The interpretability score is derived from the interpretability curve. So we first describe how to compute this curve for a fixed PC, call it $\mathbf{v}$.

- Compute a series of k-sparse PCA solutions for $k = 1, \ldots, p$, denote them by $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_p = \mathcal{P}$.
- For every solution $\mathcal{P}_k$ find the vector in $\mathcal{P}_k$ that is closest to $\mathbf{v}$, in cosine measure, among the top r k-sparse PCs in $\mathcal{P}_k$. Let s_k be their cosine similarity. (We look for the closest vector rather than use the same index since the original order of the PCs may change when adding sparsity constraints). Doing so for $k = 1, \ldots, p$ we obtain a vector of similarities $\mathbf{s} = (s_1, \ldots, s_p = 1)$.

- Define the curve ℓ by the p points $\ell = \{(\frac{k}{p}, s_k) : k = 1, \ldots, p\}$.

The x-axis of ℓ is normalized by p to enable comparison between data sets with different dimensions. Also note that the curve is expected to be monotonically increasing. See Figure 1 for example.

3.2 The interpretability score

The interpretability score is the area under ℓ given by the trapezoid area estimation:

$$\sum_{k=1}^{p-1} \frac{s_{k+1} + s_k}{2p}. \quad (1)$$

For intuition consider two extreme cases: very concentrated signal and white noise. If the PCA solution $\mathcal{P}$ happens to be 1-sparse then the curve ℓ is $y = 1$. Its corresponding score will be $1 - \frac{1}{p}$, which is also its maximum value. On the other hand, simulations that we did with white noise, a p-dimensional standard Gaussian with an identity covariance matrix, and $n \gg p$ (as in our case), show that the curve never plateaus but rather is a straight line. More generally, the sooner the curve plateaus at $y \approx 1$, the larger the score will be. Figure 1 demonstrates this nicely: Twitter and Instagram plateau fast, while LinkedIn and Steam never plateau.

3.3 Choosing k

The interpretability curve gives a visual way to choose the right k for the top-k rule: cut right above the "knee", similarly to Cattell's method for choosing the number of relevant factors in a scree plot. In our example, this yields $k = 3$ for Twitter and $k = 4$ for Instagram (Figure 1).

Finally, let us note that the interpertability score could have been computed with the top-k rule rather than sparse PCA, by zeroing out the smallest $p-k$ entries and normalizing. In Section 4.3 we show that using sparse PCA produces more reliable scores. One intuitive reason for this is that if the top k entries are not significant, even if their total weight is relatively large, then sparse PCA may find a k-sparse vector farther from the PC, resulting in a lower interpertability score, as should be the case.

4 Evaluation

We now demonstrate how the framework described in Section 3 supports PCA interpretation in various social media platforms.

4.1 Data collection

We crawled the network in a snowball approach, which is commonly used in the literature (Mislove et al. 2007). Crawling starts from a list of randomly selected users and proceeds in a BFS manner. At each step the crawler pops a user v from the queue, explores its outgoing links and adds them to the queue. In Twitter there is a link from v to w if v follows w. In Instagram the set of friends is private in most cases. We say that w is an outgoing link from v if w commented on v's pictures. In Steam the list of friends is public. In LinkedIn the list of friends (called connections) is private. As a proxy for v's friends we used the "People Also Viewed" box which tells what recent profiles w were viewed by people who viewed v.

We collected between 11 to 15 features per network that describe the user's activity in the network and feedback that a user receives from other users. Feedback features included for example the number of users following me, the number of retweets of my tweets by others, the number of likes I received or comments left on my pictures. The activity features included the volume of activity (e.g. posts per day, total number of posts), activity types (e.g. percentage of video vs pictures, urls vs. pure text), social activity (number of friends, number of likes I gave, number of tweets I retweeted). The complete set of features can be found in (Vilenchik 2019). Similar features were used to find influential users in MySpace and Facebook (Eirinaki, Monga, and Sundaram 2012) or in YouTube and Flickr (Canali, Casolari, and Lancellotti 2012) for user classification. We collected a total of 284,758 Twitter accounts, 52,574 in Instagram, 127,830 in Steam and 12,000 in LinkedIn. Different numbers stem from varying levels of technical difficulty in crawling each network and from time constraints.

To check whether bot profiles tilted our distribution, which is relevant for Twitter and Instagram, we removed heavy hitters (users whose projection on either of the top two PCs was in the fourth quartile) and recomputed the leading PC for the new dataset. The cosine similarity between the new PC and the original PC was around 0.95 both in Twitter and Instagram. Therefore we may safely contain the existence of bots in our data, at least for the sake of computing the interpretability scores.

4.2 Computing the interpretability score

For each of the four datasets we computed the covariance matrix and performed PCA. For each PC we recorded its vector entry values, and the percentage of variance it explains. In all networks the top three to five PCs passed the GK criterion. In what follows we focus on the leading two PCs which were the most significant in terms of explained variance.

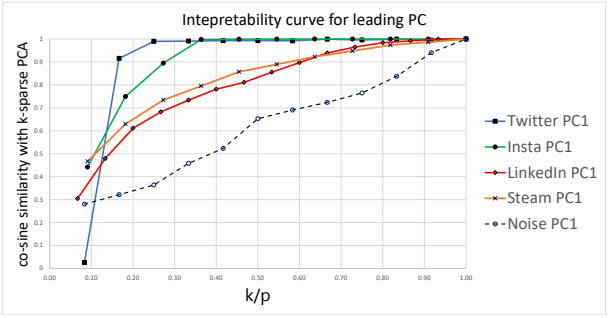

Figure 1: The interpretability curve for various datasets. Twitter's and Instagram's curve (top two lines) plateaus fast at $y \approx 1$, Steam and LinkedIn (next two lines) follow a rather straight line, closer to white noise (lowest line).

Interp. score	PC1	PC1 avg	PC2	PC2 avg
Twitter	0.89	0.86 ± 0.02	0.8	0.78 ± 0.02
Instagram	0.85	0.85 ± 0.009	0.85	0.85 ± 0.01
LinkedIn	0.76		0.76	
Steam	0.77	0.76 ± 0.02	0.7	0.7 ± 0.03
Noise	0.7	0.7 ± 0.04	0.65	0.65 ± 0.05

Table 1: Interpretability score for different datasets. The average score and std are taken over 20 random subsamples of size 3000. White noise is a 12-variate Gaussian with identity population covariance matrix and $n = 3,000$ samples. No average for LinkedIn due to small size of dataset

Figure 1 shows the interpretability curves of the leading PC in different networks. Clearly Twitter and Instagram have a very different curve type than LinkedIn and Steam. This is reflected in the interpretability score given in Table 1, which also presents average and standard deviation over 20 random subsamples of size 3,000 users. The series of 20 scores were verified to follow a normal distribution using Shapiro-Wilk with confidence level $\leq 5\%$. A t-test verified that Steam's and LinkedIn's score is significantly lower than Twitter's and Instagram's (p-value for the null hypothesis was practically zero).

4.3 Comparing against the top-k-rule

We repeated the computation of the interpretablity score, this time using the top-k rule rather than sparse PCA. Namely, given a PC $\mathbf{v}$, we zeroed out its $p - k$ smallest entries, normalized to obtain a unit vector, and used $\mathbf{v}$-truncated to compute the interpretability curve and score. For Twitter, very similar results were obtained (0.89,0.83). This is expected as the score is high, which we take as evidence that the leading two PCs of Twitter are indeed sparse, or in other words, the largest entries are significant. In Steam, the top-k rule yielded (0.77,0.78), compared with (0.77,0.7) using sparse PCA. The results in (Vilenchik 2019) suggest that the top PCs in Steam behave like random vectors with respect to the properties of activity and popularity (which are measured by the 11 Steam features). Therefore, the closer the score to the Noise benchmark (0.65) the more reliable the result, which is given by sparse PCA (0.7 vs 0.78 for the second PC).

In addition, we checked the cosine similarity between the solution of k-sparse PCA and truncation by the top-k rule. Averaged over all k's, the similarity was 0.85 for the leading two PCs in LinkedIn, and 0.99,0.67 respectively for the leading two PCs in Steam. In Twitter and Instagram the similarity was 0.99 for the leading PC.

5 Conclusion

We developed a framework to assist in PCA interpretation. Examining how the framework applies to Twitter, where the highest scores were obtained, we see that the "knee" in Figure 1 occurs between $k = 2$ and $k = 3$. This suggests that sparsity level $k = 3$ is enough to derive the PC label of the leading PC. The three support features of the leading PC all measure popularity: the likes given to the user, the number of retweets of his tweets, and the number of followers. Looking at our sample, indeed the top users in the direction of PC1 are teen pop-idols like Justin Bieber, Zayn Malik, the Kardashians and other celebrities.

LinkedIn and Steam received lower scores, which suggests cautious (if not avoiding) interpretation. This picture is consistent with the results obtained in (Vilenchik 2019), where the semantics of the PCs in these social media platforms was studied. The conclusion in (Vilenchik 2019) was that while in Twitter and Instagram, the top PCs have a clear semantic direction (either popularity or activity), the leading PCs in Steam have a random semantic direction with respect to these properties, and in LinkedIn the picture is mixed (the leading PC having random direction as well).

Finally, turning to the question posed in the title of the paper, the emerging suggestion is: Twitter and Instagram – interpret, Steam - do not interpret, LinkedIn – tread cautiously and perhaps be assisted with other exploratory means.

Acknowledgements

This work was supported by the ISF grant number 1388/16. We thank Shalom Toledo and Yaniv Costica for helping with the data collection.

References

Canali, C.; Casolari, S.; and Lancellotti, R. 2012. A quantitative methodology based on component analysis to identify key users in social networks. *Int. J. Social Network Mining* 1(1):27–50.

Eirinaki, M.; Monga, S. P. S.; and Sundaram, S. 2012. Identification of influential social networkers. *Int. J. Web Based Communities* 8(2):136–158.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM Conference on Internet Measurement*, 29–42.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. In *Proc. of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, volume 11, 281–288,.

Preotiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through twitter content. In *ACL*, 1754–1764.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd Int. Workshop on Search and Mining User-generated Contents*, 37–44.

Vilenchik, D. 2019. Simple statistics are sometime too simple: A case study in social media data. *IEEE Transactions on Knowledge and Data Engineering*.

Viswanath, B.; Bashir, M. A.; Crovella, M.; Guha, S.; Gummadi, K. P.; Krishnamurthy, B.; and Mislove, A. 2014. Towards detecting anomalous user behavior in online social networks. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, 223–238.