MODELLING LARGE SCALE DATASETS USING PARTITIONING-BASED PCA

Salaheddin Alakkari and John Dingliana

Graphics, Vision and Visualisaltion Group (GV2),
School of Computer Science and Statistics,
Trinity College Dublin

Fig. 1. Reconstructed sample from Labeled Faces in the Wild (LFW) dataset using 200 holistic eigenfaces compared to only 20 cell eigenfaces.

ABSTRACT

In this study, we propose an efficient approach for modelling and compressing large-scale datasets. The main idea is to subdivide each sample into smaller partitions where each partition constitutes a particular subset of attributes and then apply PCA to each partition separately. This simple approach enjoys several key advantages over the traditional holistic scheme in terms of reduced computational cost and enhanced reconstruction quality. We study two variants of this approach, namely, cell-based PCA for image datasets where samples are spatially divided into smaller blocks and the more general band-based PCA where attributes are partitioned based on their values distribution.

Index Terms— Large-scale Data, PCA, Pattern Extraction, Image Data Representation, Compression.

1. INTRODUCTION AND MODEL DESCRIPTION

Principal Component Analysis (PCA) is one of the most well-known unsupervised learning techniques used for dimensionality reduction and pattern extraction. The main task of PCA is to compute low-dimensional basis vectors that capture most variability of the input dataset $X \in \mathbb{R}^{d\times n}$. These basis vectors correspond to the k most significant eigenvectors $V \in \mathbb{R}^{d\times k}, k \ll \min(n,d)$ of the covariance matrix $C = \frac{1}{n-1}X_c X_c^T$ where X_c is the dataset after subtracting the sample mean. Such holistic linear representation is optimal in terms of the mean-squared-error [1]. However, finding such basis vectors requires $\mathcal{O}(nd\min(n,d))$ FLOPs of computation and $\mathcal{O}(\min(n,d)^2)$ memory space. Hence, analyzing large-scale datasets (of very large $\min(n,d)$) becomes computationally infeasible. Almost all machine learning approaches nowadays (such as Autoencoders, CNNs, etc.) suffer such scalability problems. Our solution is to subdivide each sample into smaller partitions where each partition constitutes a particular subset of attributes and then apply PCA to each partition separately. While this idea is simple, it enjoys many major advantages over the traditional holistic approach. By subdividing samples into p partitions of equal sizes, one can note that the resulting p covariance matrices are smaller than the holistic one since $p.(d/p)^2 = d^2/p < d^2$ and the computational complexity becomes $\mathcal{O}(nd^2/p)$ instead of $\mathcal{O}(nd^2)$ (assuming $d < n$) reducing both space and computational cost. Moreover, since each partition is processed independently, the approach is embarrassingly parallel and the computation can be further reduced to $\mathcal{O}(n(d/p)^2)$. Having these advantages on hand, such mode of computation raises two main questions. First of all, on what basis are attributes mapped to different partitions? The second question concerns the relation between the partitioned eigenvectors and the holistic ones. We propose two different strategies for assigning attributes to different partitions which we refer to as cell-based PCA and band-based PCA. We asses each strategy in terms of the average mean-squared-error and SSIM between reconstructed samples and their original counterparts in addition to the average run-time per cell using CPU and GPU implementations. We find that using the proposed partitioning strategies significantly enhances reconstruction and dramatically reduces the run-time. In addition, experimental results indicate that when assigning attributes to each partition randomly, the combination of the resulting partial eigenvectors becomes analogous to the holistic solution.

2. RELATED WORK

Partitioning-based PCA has become an active research area in the last two decades specifically for applications in distributed data analysis and computer graphics. Partitioning-based PCA algorithms can be categorized into two types: Sample-based partitioning and attribute-based partitioning. The sample-based partitioning is the most commonly used approach with

extensive use in domain applications including subspace clustering and distributed systems. In such approaches, the data is divided into smaller subsets of samples for two different goals depending on the type of application. In subspace clustering, the aim is to compute a set of subspaces where each subspace optimally fits a subset of samples based on their distribution [2]. On the other hand, distributed PCA addresses the problem of analyzing data partitioned across multiple distributed servers. It can be either used as sample based partitioning where each server possesses $n_i < n$ samples of high dimension or attribute-based partitioning where each server streams $d_i < d$ channels of partial attributes to a global coordinator. The first scenario is typical when considering very high dimensional data distributed in star-topology networks where each machine performs a local SVD on their samples followed by global processing using the center coordinator. Such a problem has been accentuated by the machine learning community in [3, 4, 5]. A more challenging scenario is that when servers are connected in mesh-topology networks which was discussed by Wu et al. [6] and Fellus et al. [7]. In this setting, the power iteration is applied where after each local update a globalization procedure is performed using a Gossip communication protocol [8, 9, 10]. One can note that in the sample-based partitioning in distributed PCA, the general problem remains a batch type of learning. On the other hand, attribute-based distributed PCA is more commonly used in streaming applications particularly in the field of multi-sensor signal processing. Attribute-based partitioning was also investigated for learning appearance of 3D mesh models using eigen-textures [11] where PCA is applied to each texture segment of a rotating object.

3. CELL-BASED PCA

In this section, we address the first strategy for assigning attributes to partitions, which we refer to as cell-based PCA. The technique is inspired by the JPEG compression standard [12] where images are subdivided into small blocks (cells) of size 8×8. Each block is then projected onto the 2D Discrete Cosine Transform (DCT) basis functions. These basis functions form a global representation for any 8×8 gray-scale images by defining spatial frequencies in the 2D space. Each block is represented in terms of its projection values on these 64 basis functions. In order to achieve compression, only a few projection values are chosen for representation, those corresponding to lower frequency basis functions. One can note that the main drawback in such representation is the difficulty of prioritizing the low frequency basis functions in the X and Y directions. This becomes even more problematic when dealing with larger block sizes.

Unlike JPEG compression, cell-based PCA is a data-driven representation and compression approach. In this case, basis functions of each individual block are computed by applying PCA to pixels within the region occupied by a corresponding cell (block). This brings with it the advantage of finding optimal basis functions of each block appropriately ordered based on their significance leading to minimal reconstruction errors (in terms of MSE) [1] for fixed number of basis functions (eigencells). The downside of this method is that each block will have its own basis functions (eigenvectors) based on the distribution of the input dataset. In our implementation, we include all color channels for computing eigencells. This is a somehow trivial choice since one would expect the distribution of colors in a single block to be consistent. Computing cell eigenvectors for larger cell sizes may not be possible using batch PCA. For such scenarios, streaming PCA approaches can be applied instead [13].

4. MODELLING NON-SPATIALLY LOCALIZED DATA USING BAND-BASED PCA

One main limitation of cell-based PCA is that it works only for spatially localized samples of image format. In this section, we present an alternative partitioning technique that can be applied to more general types of datasets. The main premise is to assign attributes to non-spatialized partitions, called *bands*, based on their values distribution instead of their spatial locations. We will discuss two strategies for assigning attributes in the following subsections.

4.1. Mapping Attributes Based on Sample Mean

A key assumption in JPEG and cell-based PCA is that neighboring pixels usually have similar values. However, if the data on hand is not spatially organized, this assumption is no longer valid. In this section, we make no spatial assumptions when assigning attributes to their corresponding bands. Rather, we group attributes that are close in terms of their mean values which can be formally described as follows

$$B = \{a \in x \mid l_B \leq \mathbb{E}(a) \leq u_B\},$$

where a is an attribute value of x and l_B and u_B are lower and upper bounds defining the interval of the corresponding band. Typically, the difference between these bounds should be small. This can be done by subdividing the range of values in the sample mean into s sub-intervals. This will lead to a non-uniform number of attributes per interval. In order to limit the number of attributes per band, attributes belonging to the same interval are further sorted based on their mean values and then grouped into bands with maximum allowed length L. The main problem is that computing the sample mean requires a pre-processing data pass which is prohibitive in the case of large-scale and streaming data scenarios. One can solve this problem by computing such statistics from a single mini-batch of the input data assuming samples are independent and identically distributed.

4.2. Mapping Attributes Based on Mean and Variance

Using only sample mean as a basis for mapping attributes may neglect important criteria that attributes of the same band must share. Furthermore, using the mean by itself does not constitute the distribution of values that a particular attribute may have. It is well-known that PCA is most efficient when samples obey a multivariate normal distribution [14]. Hence, we may assume that attributes of the same band are normally distributed and share similar parameter values in terms of mean and standard deviation. More formally, this can be expressed as follows

$$B = \left\{ a \sim \mathcal{N}\left(\mu, \sigma^2\right) \in x \mid l_B^m \leq \mu \leq u_B^m \wedge l_B^s \leq \sigma \leq u_B^s \right\}$$

where $\mu = \mathbb{E}(a)$ and $\sigma^2 = \text{Var}(a)$ are mean and variance respectively of the attribute a. In order to achieve such conditions, we subdivide the range of values in the sample mean into s sub-intervals. Then, we sort attributes within the same sub-interval based on the variance (instead of the mean) and then group each L ordered attributes into one band.

5. RESULTS

We will evaluate the performance of the aforementioned techniques on two large-scale face datasets, namely, Labeled Faces in the Wild (LFW) [15] and CelebA [16]. While generalizable to more general types of datasets, our choice of such datasets is to build upon an application where the impact of the holistic PCA has a well-earned reputation. In addition, this allows for visual inspection of reconstruction quality. It is worth mentioning that computing the holistic eigenspace for such datasets is infeasible. However, we approximated the first eigenfaces (eigenvectors of face images) using state-of-the-art reduced-complexity streaming PCA according to [17].

5.1. Reconstruction Quality for Different Cell Sizes

We first study the effect of varying the cell-size on the quality of reconstruction when maintaining the same compression ratio of 15:1. Table 1 contains the average MSE and SSIM scores for different cell sizes compared to the reconstruction of 1,000 holistic eigenfaces. We can clearly note that the cell-based PCA results are much better than the holistic scheme. This is consistent with Fig. 1 which shows reconstructed sample from LFW using 200 holistic eigenfaces compared to the reconstruction of only 20 cell eigenfaces (eigenvectors per cell) of cell-size $10 \times 10 \times 3$. It is also apparent that for a fixed compression ratio, expanding the cell size enhances the reconstruction quality at the expense of increasing the number of cell eigenvectors. This is well-reflected in Fig. 2 where cell boundary artifacts are reduced when increasing cell-size from 5×5 to 25×25.

Table 1. Cell-based PCA vs. holistic PCA in terms of reconstruction quality.

	LFW			CelebA		
Partition-size	**25× 25 × 3**	**1,875**	**Holistic**	**25× 25 × 3**	**1,875**	**Holistic**
# of eigenvectors per part	125	125	1,000	125	125	1,000
MSE	**0.00027**	0.00062	0.0029	**0.0007**	0.0011	0.003
SSIM	**0.9425**	0.89	0.71	**0.89**	0.85	0.72

5.2. Computation Run-time

We compare performance of CPU and GPU implementations in terms of average run-time per cell in order to reflect the speedup that can be gained when considering a perfect parallel setting (with no overhead communications). The tests were run on a workstation equipped with a 2.6 GHz Intel Core i7-6700HQ CPU and GeForce GTX 960M GPU. We find that for small cell-sizes the CPU takes shorter run-times than the GPU. For larger cell-sizes, the GPU implementation becomes faster. This is depicted in Table 2 where average run-times per cell are reported. While computing the holistic eigenvectors for such datasets is infeasible using standard PCA, approximating the top 1,000 eigenvectors using accelerated version of PCA costs 21.2 hours for LFW and 4.6 days for CelebA.

Table 2. CPU vs. GPU average run-times per cell in seconds.

	LFW			CelebA		
Cell-size	**5× 5 × 3**	**10× 10 × 3**	**25× 25 × 3**	**5× 5 × 3**	**10× 10 × 3**	**25× 25 × 3**
CPU run-time/cell	**0.0427**	**0.16**	7.87	**0.44**	1.92	63.3
GPU run-time/cell	0.3273	0.295	**3.19**	0.9	**1.2845**	**4.39**

5.3. Comparing Band-based PCA Performance for Different Mapping Strategies

We now compare the two mapping strategies for band-based PCA. For image datasets, due to the redundancy in the color channels, we assign pixels to different bands based on one color channel (in this paper, we used the R channel but a more general approach would be to use a grayscale transformation). Since we are addressing streaming and large-scale data in this study, we estimate the sample mean and variance using a subset of the dataset which we refer to as an estimating subset. We study the reconstruction quality for different estimating subset sizes. As we found earlier in this section that increasing cell size enhances reconstruction results, we set the maximum pixels allowed per band, L, to 625 and number of intervals, s, to 50. We report reconstruction quality in terms of MSE and SSIM when using 125 band eigenfaces. Due to the way our mapping strategy assigns attributes, the number of pixels per band may be smaller than L for many bands resulting in different compression ratios than the desired target (15:1). We also tested band-based PCA when applying random mapping where attributes are assigned to different bands in random manner. Table 3 compares reconstruction results between baseline random mapping and the two proposed mapping strategies when applied to the LFW dataset. Clearly, the proposed mapping techniques are much

Fig. 2. Two reconstructed samples from LFW and CelebA datasets when maintaining 15:1 compression ratio using cell eigenfaces of different cell-sizes.

better at reconstruction than the baseline model. It is also evident that mapping attributes using mean and variance gives better quality results than the mean-based technique. In addition, increasing the estimating subset size enhances the results. However, the reconstruction results for cell-based PCA are still better despite using lower compression ratio. Fig. 3 shows many images reconstructed using different mapping strategies of band based PCA. Clearly, random mapping results in poor reconstruction quality whereas mean mapping produces some dithering artifacts. These dithering artifacts are reduced when applying mean-variance mapping.

5.4. Analogy with the Holistic Eigenspace

Fig. 4 presents a comparison between holistic eigenfaces and combined band eigenvectors (band eigenfaces) resulting from the random mapping. Interestingly, the random mapping eigenfaces have a high resemblance to the holistic ones. This suggests that the worst case instance of band-based PCA, achieved when applying the baseline random mapping model, produces an eigenspace that is analogous to the holistic solution. Both models were shown to produce poor reconstruction results in comparison to the other techniques.

6. CONCLUSION

We proposed two methods for modelling large-scale datasets by dividing data attributes into smaller subsets and then applying PCA to each partition separately. We show that these have several advantages over the standard holistic approach

Table 3. Reconstruction performance using different band-based PCA strategies.

	LFW	CelebA
band size	1875	1875
# of eigenfaces	125	125
MSE	0.00062	0.0011
SSIM	0.89	0.85

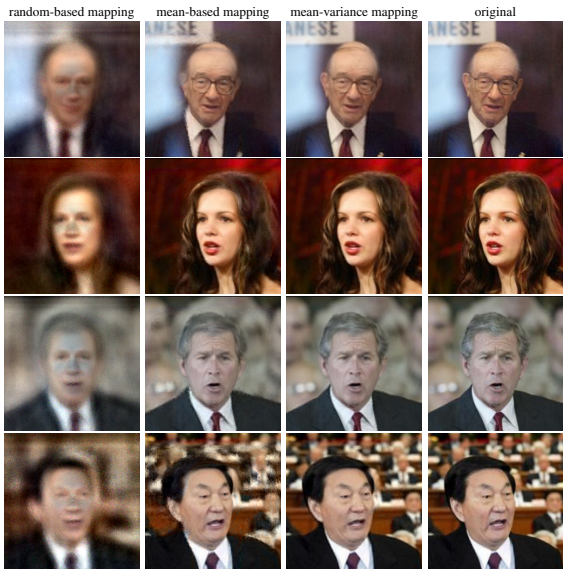

Fig. 3. Reconstructed samples from the LFW dataset using different mapping strategies for band-based PCA.

including enhanced reconstruction quality and increased computation speed, parallelism and scalability. The first model, cell-based PCA, is inspired by the JPEG standard but enjoys better guarantees in terms of reconstruction errors. The second approach, band-based PCA, maps attributes based on their values distribution rather than their spatial locations. We show that mapping pixels using the mean and variance is better in terms of reconstruction quality than mapping based only on sample mean. Both mappings were shown to be superior to the random mapping model. We also found that the baseline performance produced using random mapping is analogous to the holistic PCA solution, but entails lower memory costs and computation run time. Not only are our proposed methods beneficial for data compression, but they may also provide light-weight data representation for further learning tasks. Investigating other partitioning strategies for non-normally distributed data is an interesting avenue for future research.

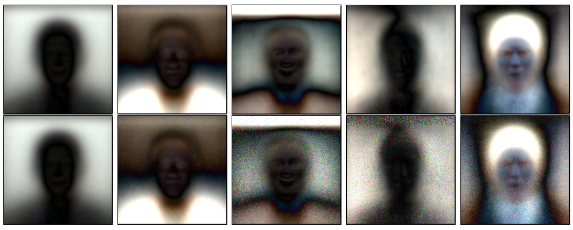

Fig. 4. Comparison of the first five holistic (top row) and random-mapping band eigenfaces of the LWF dataset.

7. ACKNOWLEDGEMENTS

This research has been conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers 13/IA/1895 and 13/RC/2106.

8. REFERENCES

[1] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[2] René Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[3] Yongming Qu, George Ostrouchov, Nagiza Samatova, and Al Geist, "Principal component analysis for dimension reduction in massive distributed data sets," in *Proceedings of IEEE International Conference on Data Mining (ICDM)*, 2002.

[4] Ravi Kannan, Santosh Vempala, and David Woodruff, "Principal component analysis and higher correlations for distributed data," in *Conference on Learning Theory*, 2014, pp. 1040–1057.

[5] Dan Garber, Ohad Shamir, and Nathan Srebro, "Communication-efficient algorithms for distributed stochastic principal component analysis," *arXiv preprint arXiv:1702.08169*, 2017.

[6] Sissi Xiaoxiao Wu, Hoi-To Wai, Lin Li, and Anna Scaglione, "A review of distributed algorithms for principal component analysis," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1321–1340, 2018.

[7] Jérôme Fellus, David Picard, and Philippe-Henri Gosselin, "Dimensionality reduction in decentralized networks by gossip aggregation of principal components analyzers," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2014, pp. 171–176.

[8] John Nikolas Tsitsiklis, "Problems in decentralized decision making and computation.," Tech. Rep., Massachusetts Inst. of Technology Cambride Lab for Informaiton and Decision Systems, 1984.

[9] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[10] Alexandros G Dimakis, Soummya Kar, José MF Moura, Michael G Rabbat, and Anna Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[11] Ko Nishino, Yoichi Sato, and Katsushi Ikeuchi, "Eigentexture method: Appearance compression based on 3D model," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* IEEE, 1999, vol. 1.

[12] Gregory K Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[13] Hervé Cardot and David Degras, "Online principal component analysis in high dimension: Which algorithm to choose?," *International Statistical Review*, vol. 86, no. 1, pp. 29–50, 2018.

[14] Fang Han and Han Liu, "Principal component analysis on non-gaussian dependent data," in *International Conference on Machine Learning*, 2013, pp. 240–248.

[15] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller, "Learning to align from scratch," in *NIPS*, 2012.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[17] Raman Arora, Andrew Cotter, Karen Livescu, and Nathan Srebro, "Stochastic optimization for PCA and PLS," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 861–868.