Pushing the limits of Random Consensus Robust PCA: an analysis of Noise, Adversarial Outliers, and Big Data on R2PCA performance

3 Team Members

1 Introduction

We focus our research efforts primarily on the R2PCA algorithm [5], which is a random consensus method for robust principal component analysis (PCA). Traditional PCA finds a low-dimensional subspace that approximates a large data matrix by minimizing an L2-norm data consistency term, which makes it especially sensitive to outliers. This motivates the need for robust PCA methods [4, 1, 3]. The random sample consensus (RANSAC) robust estimation algorithm [2] is an established method that enjoys a few advantages over extant robust PCA techniques, namely: (i) working well under noise, (ii) is computationally efficient due to its approach of breaking down subspaces by pieces, and (iii) is easily parallelize. However, incorporating principles of the RANSAC algorithm into PCA directly has its own challenges, whose scope and solution form the basis of the R2PCA and are explained in detail in the following sections.

Robust PCA aims to find a low-dimensional subspace $\mathbf{U}$ that approximates a large data matrix $\mathbf{M}$, in the case where $\mathbf{M}$ contains outliers. In this scenario, we can model $\mathbf{M}$ as the sum of a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$, where $\mathbf{S}$ injects corruption into the entries of the low-rank approximation $\mathbf{L}$. The goal is then to find the subspace $\mathbf{U}$ spanned by the columns of $\mathbf{L}$. Additionally, it is also possible to consider the case where the M is also contaminated with additive noise, in which case

$$\mathbf{M} = \mathbf{L} + \mathbf{S} + \mathbf{W}$$

where $\mathbf{L}$ is rank-r, $\mathbf{S}$ is sparse, and $\mathbf{W}$ represents a noise matrix.

A major motivator behind our focus on this method is its remarkable performance with very large datasets with lots of corruption, which is an extremely common occurrence in naturally-generated data. This is not a trivial problem due to the following: When a large dataset is corrupted, it can be difficult to identify which data points are corrupted and which are not. This makes it difficult to apply traditional data analysis methods, which often assume that the data is clean. In real-world applications, data is often corrupted by noise, measurement errors, or other sources of variation. This can be particularly problematic when the data is generated by complex systems or processes, as these can be difficult to model accurately.

Due to its advances especially in scalability, parallelization, and resistance to noise, we intend to use the R2PCA for foreground-background segmentation tasks such as surveillance, among the many datasets presented in [7], and on especially large datasets such as the MRI time series data from large population studies [8].

The novel contributions of our project will be to investigate (i) how different types of additive noise distributions affect R2PCA performance, (ii) the stability of adversarial R2PCA, in which outliers are selectively inserted to maximize the error of PCA subspace estimation, and (iii) generalizing the noisy-variant R2PCA to cases where the noise level is not known.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

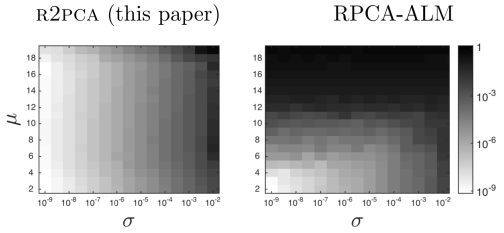
2 Methods

Robust PCA (RPCA) techniques have been developed to address this issue, namely finding a low-dimensional subspace $\mathbf{U}$ that approximates a large data matrix $\mathbf{M}$, when some entries in $\mathbf{M}$ are grossly corrupted. An established approach for robust estimation is with the random sample consensus (RANSAC) algorithm [2], which can be applied to PCA. RANSAC enjoys a few advantages over other robust PCA techniques, as it makes almost no assumptions about the data, does not require unrealistic conditions to succeed, has theoretical guarantees, works well in practice, and has enjoyed many improvements since its inception. The RANSAC version of PCA iteratively searches for $\mathbf{U}$ by randomly selecting a few columns of $\mathbf{M}$ at a time, using these columns to define a candidate subspace, and comparing this subspace to the other columns of $\mathbf{M}$. This approach requires that there are a sufficient number of uncorrupted columns in $\mathbf{M}$ in order to be successful, however, which often may not be true in real-life applications. R2PCA addresses this limitation by extending the principles of RANSAC to RPCA. In contrast to other convex relaxation methods, R2PCA does not rely on assumptions about the coherence or distribution of the sparse outliers in the data. Instead, it assumes sparsity constraints on the number of sparsity of outliers per row and per column, which are reasonable for many real-world applications. As long as the sparsity level of the outliers is much smaller than the dimensions of the data matrix. When the outliers are sufficiently sparse, then R2PCA will succeed in linear complexity in r, where r is the rank of the target low-dimension subspace. Furthermore, the fact that the R2PCA algorithm operates on small blocks of the matrix at a time, makes R2PCA a highly flexible, computationally efficient method for robust PCA analysis of large datasets. The noisy-variant of R2PCA further adds to its versatility, particularly for applications in image analysis.

Although the authors suggest that their method is effective even under adversarial settings, where the outliers are intentionally inserted to complicate success, however this claim has not been thoroughly investigated. To address this gap in the literature, we plan to apply adversarial PCA [6] techniques to the synthetic and real image datasets used in the R2PCA paper, testing the algorithm's resilience to strategically placed outliers. By rigorously evaluating R2PCA technique under these more challenging conditions, we will gain a better understanding of the algorithm's limitations and identify potential areas for further improvement.

Another extension we plan to investigate is the effect of different noise distributions on R2PCA. The authors proposed a noisy-variant of R2PCA, and tested it on synthetic data that was contaminated with noise following a Normal distribution. We are interested to evaluate how R2PCA performs when the noise takes on other distributions. To test this, we will contaminate synthetic (and/or real) image data with noise matrices drawn from Rician (non-central chi-squared), Rayleigh and Erlang distributions with different noise levels/parameters. We plan to replicate the following figure from the paper, in which the estimation error of synthetic data is analyzed in terms of the noise level and data coherence.

Figure 1: Transition diagram of the estimation error of $\mathbf{L}$ as a function of the noise level and the coherence parameter μ, with $p = 5\%$ grossly corrupted entries. The color of each (σ, μ) pixel indicates the average error over 100 trials (the lighter, the better). This shows that R2PCA can consistently estimate $\mathbf{L}$ within the noise level, as long as $\mathbf{S}$ is suciently sparse, regardless of coherence. Other algorithms can also estimate $\mathbf{L}$ within the noise level, but only for a restricted range of matrices with bounded coherence



3 Related Work

The primary motivation behind the random consensus aspect of R2PCA was derived from RANSAC, which was described in a previous section. The main method that R2PCA performance was compared to was the augmented Lagrange multiplier method for robust PCA (RPCA-ALM) [2], which was several studies showed performed better than other methods, such as singular value thresholding and the accelerated proximal gradient method. We also plan to perform comparisons with the RPCA-ALM method in our analyses to serve as a benchmark to a comparable method.

4 Plan of Experiments

The authors of this paper have provided MATLAB and Python code to perform the noise-free R2PCA algorithm. They have not provided code for the noisy variant of R2PCA, which is something we are going to implement based on the description and pseudocode provided in the paper. Given that the noisy R2PCA framework is comparable to the noise-free case and can be developed entirely from readily available MATLAB/Python linear algebra functions, this shouldn't be overly difficult.

The authors tested the R2PCA algorithm on synthetic images and publicly available microscopy and surveillance image data. They have provided the code used to generate the synthetic data, and the links to the microscopy videos, and the surveillance data (Wallflower and IR2 datasets).

We plan on generating and examining several novel synthetic datasets. First, we will use the adversarial PCA method [6], to generate synthetic data with outliers placed in the location(s) that maximize the PCA subspace estimation error. Adversarial PCA [6] was proposed by the same author as the R2PCA paper, and the MATLAB code is available from the same website as the R2PCA's MATLAB code.

For analysis of synthetic data used to investigate different noise distributions and adversarial sparse outliers, we will evaluate performance in terms of the estimation error of the low-rank target subspace $\mathbf{L}$. In these cases, we have a ground truth for the desired result $\mathbf{L}$. Similar to how the authors assessed estimation error of $\mathbf{L}$ with respect to noise level sigma (the standard deviation of the Additive White Gaussian Noise distribution), we will assess the performance under different noise distributions with different noise levels.

We would also like to examine the analysis of big data to test the capabilities of R2PCA, which is something the authors implied but did not test themselves. To this end, we envision analyzing publicly available functional MRI (fMRI) [8] time series data. The HCP has various datasets, each with volumetric fMRI data from thousands of time points and from hundreds of subjects. This offers the potential to generate a matrix with millions or billions of entries, certainly pushing the limits of computational/memory expenses. Beyond demonstrating the capabilities of R2PCA for big data, it would be interesting to analyze fMRI data to see if any trends emerge in terms of functional brain activation across different subject populations.

5 Plan of Project

First, we perform initial setup of the existing R2PCA code and build understanding through building from scratch. This will be done by all 3 authors in parallel, as it ensures everyone is on the same page. We then plan to investigate the adaptation of adversarial PCA to R2PCA, which will be driven by Authors 1 and 3. For the extension phase, Author 1 and Author 2 will be driving the efforts on integrating the choosing, exploring and training on the surveillance data. Author 3 will also be working on fMRI data, as it is closest to their area of expertise. Author 2 will also primarily work on editing and copywriting tasks that arise through the process of refining the paper. Author 3 will primarily participate in the implementation in MATLAB and Python. Through these divisions, we aim to ensure that all 3 authors have a deep understanding of the modifications to be made, both algorithmic and programmatic. We draw upon our built competencies in MATLAB and Python. We anticipate the bulk of the challenges to arise from adapting the novel datasets, especially surveillance segmentation, to be used by the R2PCA algorithm.

We do not anticipate any major problems in replicating the analyses performed in the paper, or in implementing the noisy-variant of R2PCA. We chose several options for extending the method and analyzing novel data, as this is the area we anticipate problems may occur. For the adversarial analysis, we have access to the code for generating adversarial outliers, which should not be challenging to implement. The other extensions are less straightforward, thus in the case we encounter a problem, we will have at least one novel extension working. If necessary, we will find publicly available microscopy data, similar to the videos used in the paper, to include as another novel dataset.

A rough timeline with deliverables is given below:
- March end – Setup and extension phase (adversarial PCA)
- April 1st week – Experimentation phase, extension phase resumes
- April 2nd week – Extension phase (data analysis and dataset integration), documentation starts
- April 3rd week – Documentation and completion of report

References

[1] Fernando De La Torre and Michael J. Black. "A framework for robust subspace learning". In: *International Journal of Computer Vision* (2003).

[2] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.

[3] Qifa Ke and Takeo Kanade. "Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2005.

[4] Ricardo A. Maronna. "Robust M-estimators of multivariate location and scatter". In: *The Annals of Statistics* (1976).

[5] Daniel Pimentel-Alarcón and Robert Nowak. "Random consensus robust PCA". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 344–352.

[6] Daniel L Pimentel-Alarcón, Aritra Biswas, and Claudia R Solıs-Lemus. "Adversarial principal component analysis". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 2363–2367.

[7] G. Sreenu and M. A. Saleem Durai. "Intelligent video surveillance: a review through deep learning techniques for crowd analysis". In: *Journal of Big Data* (2019).

[8] World BankThe World Bank. *Extensively Processed fMRI Data*. https://www.humanconnectome.org/study/hcp-young-adult/document/extensively-processed-fmri-data-documentation. 2017.