STOCHASTIC PRINCIPAL COMPONENT ANALYSIS VIA MEAN ABSOLUTE PROJECTION MAXIMIZATION

*Mayur Dhanaraj and Panos P. Markopoulos**

Dept. of Electrical and Microelectronic Eng., Rochester Institute of Technology, Rochester NY, USA
E-mail: mxd6023@rit.edu, panos@rit.edu

ABSTRACT

Principal-Component Analysis (PCA) is a data processing method with numerous applications in signal processing and machine learning. At the same time, standard PCA has been shown to be very sensitive against faulty/outlying data. On the other hand, L1-norm-based PCA (L1-PCA), seeking to maximize the aggregate absolute projections of the processed data, has demonstrated sturdy corruption resistance. At the same time, in our big data era, there is a need for online (stochastic) algorithms for data analysis with limited storage and computation requirements. To this end, in this paper we extend batch L1-PCA and propose a novel algorithm for stochastic PC calculation based on mean absolute projection maximization, with formal convergence guarantees. Our numerical studies demonstrate the convergence and corroborate the corruption resistance of the proposed method.

1. INTRODUCTION

Principal-Component Analysis (PCA) finds numerous applications in machine learning and signal processing [1] , among other fields. However, PCA is also known to be particularly sensitive against faulty data entries [2], occuring often in modern big datasets, due to various causes such as mislabeling, intermittent sensor malfunction, deliberate jamming, and sensing-environment changes.

To counteract the faulty entries, corruption-resistant PCA approaches have been proposed in the literature. In one standard direction, researchers study methods for Robust PCA (RPCA) that tries to express the processed data matrix as the sum of a low-rank matrix, spanning the sought-after subspace, and a sparse matrix that captures outliers [2]. Typically, RPCA relies on the formulation and solution of a convex optimization problem, with a metric that places ad-hoc weights on dimensionality reduction and outlier sparsity. On a different, arguably more straightforward direction, researchers study reformulations of PCA that are based on absolute projections and the L1-norm (L1-PCA), placing lower emphasis per data point, thus gaining robustness against outliers. For the solution of batch L1-PCA, several exact [3] and efficient solvers [4–6] have been proposed in the literature.

In the advent of the big data era, joint processing of all data measurements as a batch can be computationally prohibitive. In other practical cases, measurements arrive in a streaming fashion, as for example in real-time video processing [7]. In such cases, solving PCA from scratch on an augmented batch each time that a new point arrives is clearly impractical. In order to process a large volume of measurements incrementally, online PCA solvers have been proposed in the literature, typically relying on stochastic optimization [8, 9]. More recently, robust methods for online PCA have also been proposed in order to combine corruption resistance with low-cost online processing [10, 11].

While stochastic PCA has been well studied in the literature, the stochastic formulation of L1-PCA remains to date unexplored, despite its clear robustness in batch processing. In this paper, we formulate a novel stochastic version of L1-PCA, for one principal component, based on mean absolute projection maximization. Then, we propose an incremental algorithm for its solution, based on fundamental stochastic approximation theory. Our method is accompanied by formal convergence guarantees and numerical studies that corroborate its corruption resistance.

2. TECHNICAL BACKGROUND

2.1. Stochastic PCA

Given data distribution $\mathcal{D}$, the stochastic Principal Component (PC) is formulated as

$$\mathbf{q}_{\text{L2}} = \underset{\mathbf{q}\in\mathbb{R}^{D};\ ||\mathbf{q}||_2=1}{\text{argmax}} \ \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\{|\mathbf{x}^\top\mathbf{q}|^2\}. \quad (1)$$

Defining $\mathbf{C} := \mathbb{E}_{\mathbf{x}}\{\mathbf{x}\mathbf{x}^\top\}$, the objective in (1) can be rewritten as $\mathbb{E}_{\mathbf{x}}\{\mathbf{q}^\top\mathbf{x}\mathbf{x}^\top\mathbf{q}\} = \mathbf{q}^\top\mathbf{C}\,\mathbf{q}$ and it is maximized by the dominant eigenvector of $\mathbf{C}$. In practical applications, $\mathbf{C}$ is unknown and sample-average estimated as $\widehat{\mathbf{C}} := \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$, based on a size-$N$ batch of coherent data points $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. Accordingly, substituting $\mathbf{C}$ with $\hat{\mathbf{C}}$, the metric of (1) can be approximated as $\mathbf{q}^\top\hat{\mathbf{C}}\mathbf{q} = \frac{1}{N}\|\mathbf{X}^\top\mathbf{q}\|_2^2$, where the squared L2-norm (or

*Corresponding author.

This research is supported in part by the National Science Foundation, under grant OAC-1808582, and the U.S. Air Force Office of Scientific Research, under the Dynamic Data Driven Application Systems (DDDAS) program.

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

Frobenius-norm) $||.||_2^2$ returns the sum of the squared entries of its vector argument. Similarly, the stochastic PC can be approximated by the batch PC

$$\widehat{\mathbf{q}}_{\text{L2}} = \underset{\mathbf{q}\in\mathbb{R}^D;\ ||\mathbf{q}||_2=1}{\text{argmax}} \quad ||\mathbf{X}^\top \mathbf{q}||_2^2. \tag{2}$$

The solution to (2) can be obtained by Singular-Value Decomposition (SVD) of $\mathbf{X}$ with cost $\mathcal{O}(ND\min(N,D))$. Importantly, as the batch size N tends to infinity, $\hat{\mathbf{C}}$ tends to $\mathbf{C}$ and the batch PC $\widehat{\mathbf{q}}_{\text{L2}}$ tends to the stochastic PC $\mathbf{q}_{\text{L2}}$.

Clearly, for large N, batch PCA in the form of (2) can be computationally prohibitive. Therefore, stochastic PCA methods target at solving (1) incrementally, by processing one sample at a time. One of the earliest solvers [8, 9] initializes at a PC estimate $\mathbf{q}_0$ and, when the t-th sample is processed, for $t = 1, 2, \ldots$, it updates

$$\mathbf{q}_t = \mathcal{P}(\mathbf{q}_{t-1} + \eta \mathbf{x}_t \mathbf{x}_t^\top \mathbf{q}_{t-1}) \tag{3}$$

where, for any $\mathbf{g} \in \mathbb{R}^D \setminus \mathbf{0}_D$, $\mathcal{P}(\mathbf{g}) := \frac{\mathbf{g}}{||\mathbf{g}||_2}$.

In [9] we read that the algorithm in (3) converges almost-surely to the dominant eigenvector of $\mathbf{C}$, when the step-size η decreases across t, e.g., similar to $1/t$. Multiple other works in the literature have also studied the stochastic convergence of (3) [12–14]. Variants of (3), with different step sizes, were proposed in [15, 16]. Recently, [17] proposed a variant of (3) with adaptive step-size (AdaOja), based on the theory of [18]. Other variants of (3) operate on mini-batches, instead of single data points [19].

2.2. Outliers, L1-PCA, and Other Robust Variants

The squared emphasis that batch PCA in (2) places on each data point renders it sensitive against outliers, which are typically peripheral points. A straightforward approach to counteract the impact of outliers is to replace the squared emphasis in (2) by linear emphasis, simply by changing the norm from L2 (sum of squared projections) to L1 (sum of absolute projections). This results in the batch L1-PC

$$\widehat{\mathbf{q}}_{\text{L1}} = \underset{\mathbf{q}\in\mathbb{R}^D;\ ||\mathbf{q}||_2=1}{\text{argmax}} \quad ||\mathbf{X}^\top \mathbf{q}||_1. \tag{4}$$

Batch L1-PCA in (4) has attracted extensive documented interest over the past decade. In [5, 6], authors presented efficient algorithms for its solution based on alternating optimization. In [3], authors solved the problem exactly for the first time. In [4], authors presented a bit-flipping-based iterative solver. Extensions of L1-PCA to tensor (multi-way array) data processing have also been presented in [20–22]. Other algorithms for Robust-PCA were presented, e.g., in [2,23,24].

2.3. Robust Stochastic/Adaptive PCA

Similar to batch PCA, stochastic PCA is sensitive against outliers among the processed data. This has created a need for robust stochastic PCA algorithms. To this end, Online Robust-PCA (OR-PCA) was presented in [25]. Later, [26] presented R-SGD1, an algorithm that conducts robust gradient descent iterations: $\mathbf{q}_t = \mathcal{P}(\mathbf{q}_{t-1} + \eta_t\ \mathbf{x}_t \text{sgn}(\mathbf{x}_t^\top \mathbf{q}_{t-1}))$. A similar streaming algorithm was presented in [32], without stochastic convergence analysis. Recently, an algorithm for adaptive L1-PCA for multiple PCs with online outlier rejection was presented in [27]. Algorithms for incremental L1-PCA were also presented in [28–30]. In this work, for the first time, we present an algorithm for stochastic L1-PCA through mean absolute projection maximization.

3. PROPOSED METHOD

3.1. Problem Formulation

First, we note that the metric of (4) can be equivalently rewritten as $\frac{1}{N}\|\mathbf{X}^\top \mathbf{q}\|_1$. Next, we note that $\frac{1}{N}\|\mathbf{X}^\top \mathbf{q}\|_1$ tends to $\mathbf{E}_{\mathbf{x}}\{|\mathbf{x}^\top \mathbf{q}|\}$, as N tends to infinity. Thus, the stochastic L1-PC takes the form

$$\mathbf{q}_{\text{L1}} = \underset{\mathbf{q}\in\mathbb{R}^D;\ ||\mathbf{q}||_2=1}{\text{argmax}} \quad \mathbb{E}_{\mathbf{x}}\{|\mathbf{x}^\top \mathbf{q}|\}. \tag{5}$$

Incorporating the norm constraint in the objective function, we can rewrite the stochastic L1-PC in (5) as the mean-absolute projection maximization (MaxAP)

$$\mathbf{z}_{\text{MaxAP}} = \underset{\mathbf{z}\in\mathbb{R}^D}{\text{argmax}}\ \mathbb{E}_{\mathbf{x}}\left\{\frac{|\mathbf{x}^\top \mathbf{z}|}{||\mathbf{z}||_2}\right\}, \tag{6}$$

noticing that $\mathcal{P}(\mathbf{z}_{\text{MaxAP}})$ solves (5). In order to ensure the definition and continuous differentiability of the function inside the expectation, we modify it as

$$M(\mathbf{x}; \mathbf{z}, \epsilon) := \frac{\sqrt{|\mathbf{x}^\top \mathbf{z}|^2 + \epsilon}}{\sqrt{||\mathbf{z}||_2^2 + \epsilon}}, \tag{7}$$

for some positive $\epsilon \ll 1$, and rewrite MaxAP as

$$\mathbf{z}_{\text{MaxAP}} = \underset{\mathbf{z}\in\mathbb{R}^D}{\text{argmax}}\ \mathbb{E}_{\mathbf{x}}\left\{M(\mathbf{x}; \mathbf{z}, \epsilon)\right\}. \tag{8}$$

Certainly, for $\epsilon = 0$, (8) coincides with (6). In the sequel, we focus on solving (8).

3.2. MaxAP Solution

At the extrema of the ϵ-modified MaxAP metric in (8), it holds

$$\nabla_{\mathbf{z}^\top}\left[\mathbb{E}_{\mathbf{x}}\left\{M(\mathbf{x}; \mathbf{z}, \epsilon)\right\}\right] = \mathbf{0}_D. \tag{9}$$

In accordance with the regularity conditions on $M(\mathbf{x}; \mathbf{z}, \epsilon)$ [33], we interchange the gradient and expectation operators in (9) as

$$\mathbb{E}_{\mathbf{x}}\{L(\mathbf{x}; \mathbf{z}, \epsilon)\} = \mathbf{0}_D, \tag{10}$$

Proposed Stochastic MaxAP

Input: $\{\mathbf{x}_t\}_{t=1,2,\ldots,N}$, $\{\eta_t\}_{t=1,2,\ldots,N}$, $\mathbf{q}_0 \in \mathbb{R}^D$

for $t = 1, 2, \ldots, N$
  $\mathbf{z}_t \leftarrow \mathbf{z}_{t-1} + \eta_t L(\mathbf{x}_t; \mathbf{z}_{t-1}, \epsilon)$
  $\mathbf{q}_t \leftarrow \mathcal{P}(\mathbf{z}_t)$
end for

Output: $\mathbf{q}_N$ (as estimate of the stochastic L1-PC $\mathbf{q}_{L_1}$)

Fig. 1: Proposed method for PC estimation through MaxAP.

where

$$L(\mathbf{x};\mathbf{z},\epsilon) := \nabla_{\mathbf{z}^\top} M(\mathbf{x};\mathbf{z},\epsilon) \tag{11}$$

$$= \frac{\mathbf{x}^\top \mathbf{z}}{\sqrt{(|\mathbf{x}^\top \mathbf{z}|^2 + \epsilon)(\mathbf{z}^\top \mathbf{z} + \epsilon)}} \mathbf{x} - \frac{\sqrt{|\mathbf{x}^\top \mathbf{z}|^2 + \epsilon}}{(\sqrt{\mathbf{z}^\top \mathbf{z} + \epsilon})^3} \mathbf{z}. \tag{12}$$

In view of (11), we first propose the iteration

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \eta_t L(\mathbf{x}_t; \mathbf{z}_{t-1}, \epsilon),\ t = 1, 2, \ldots, \tag{13}$$

for step sizes $\{\eta_t\}$ that satisfy $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$ (e.g., $\eta_t = \frac{\gamma}{t}$ for some constant $\gamma > 0$). First, we note that, due to the step-size conditions, for any given stream of data, the iteration converges in the argument: $\lim_{t\to\infty} \|\mathbf{z}_t - \mathbf{z}_{t-1}\|_2 = 0$. In addition, based on the fundamental stochastic approximation lemma presented by Robbins-Monro in [31], the iteration attains stochastic convergence, in the mean-square (m.s.) sense, to a root of (10). In fact, similar to gradient ascent for standard Rayleigh quotient maximization [34], (13) stochastically converges to a maximum of $\mathbb{E}_{\mathbf{x}}\{M(\mathbf{x};\mathbf{z},e)\}$ that solves (9). Accordingly, for very low ϵ, the dependent PC sequence

$$\mathbf{q}_t = \mathcal{P}(\mathbf{z}_t),\ t = 1, 2, \ldots, \tag{14}$$

approximates the stochastic L1-PCA solution in (5). For $\epsilon = 0$, the proposed iteration in (13) simplifies to $\mathbf{z}_t = \mathbf{z}_{t-1} + \eta_t \mathbf{P}_t \mathbf{v}_t \text{sgn}(\mathbf{v}_t^\top \mathbf{z}_{t-1})$, $t = 1, 2, \ldots$, where $\mathbf{P}_t := \mathbf{I}_D - \mathbf{z}_{t-1} \frac{1}{\|\mathbf{z}_{t-1}\|_2^2} \mathbf{z}_{t-1}^\top = \mathbf{I}_D - \mathbf{q}_{t-1}\mathbf{q}_{t-1}^\top$ is the projection matrix to the nullspace of the previous PC and $\mathbf{v}_t := \mathbf{x}_t \frac{1}{\|\mathbf{z}_{t-1}\|_2}$ is the normalized new sample. A pseudocode for the proposed method is presented in Fig. (1).

4. NUMERICAL STUDIES

4.1. Synthetic Data

Convergence: We consider $D = 3$ and draw $N = 500$ data points from $\mathcal{N}(\mathbf{0}_3, \mathbf{C})$, where $\mathbf{C} = \begin{bmatrix} 2.05 & 1.05 & 1.08 \\ 1.05 & 0.7 & 0.31 \\ 1.08 & 0.31 & 0.97 \end{bmatrix}$. We set $\epsilon = 0$ and run the proposed iteration with a step size $\eta_t = \frac{1}{t}$. Then, we compute the change magnitude $\|\mathbf{q}_t - \mathbf{q}_{t-1}\|_2^2$ and plot it in Fig. 2(a), versus t. An argument convergence to zero-change is observed after just 5

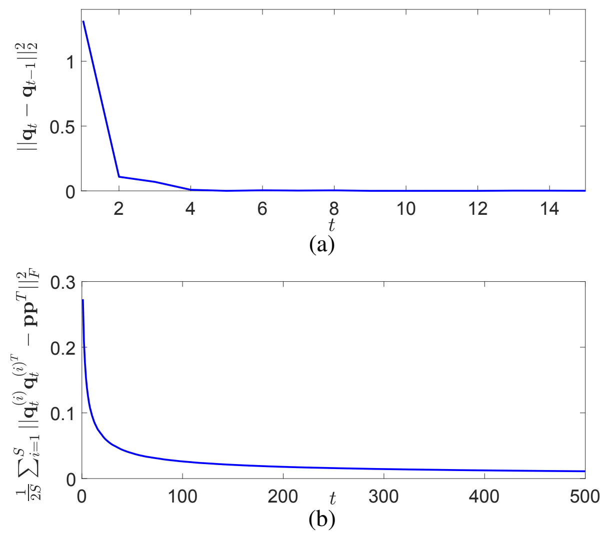

Fig. 2: Convergence of the proposed method: (a) argument convergence in an arbitrary single stream of data; (b) estimated mean-square convergence to the stochastic L1-PC.

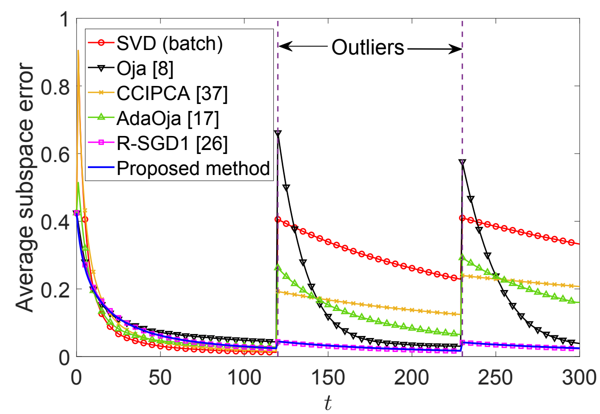

Fig. 3: Average subspace error vs. t.

iterations. Then, we notice that, specifically for Gaussian data $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_3, \mathbf{C})$, $|\mathbf{z}^\top \mathbf{x}|$ follows half-normal distribution with mean $\mathbb{E}_{\mathbf{x}}\{|\mathbf{x}^\top \mathbf{z}|\} = \sqrt{\frac{2}{\pi}\mathbf{z}^\top \mathbf{C}\mathbf{z}}$ [35]. Accordingly, the MaxAP metric becomes $\mathbb{E}_{\mathbf{x}}\{M(\mathbf{x};\mathbf{z},0)\} = \sqrt{\frac{2}{\pi}}\sqrt{\frac{\mathbf{z}^\top \mathbf{C}\mathbf{z}}{\mathbf{z}^\top \mathbf{z}}}$ and the stochastic L1-PC that solves (5) coincides with the dominant eigenvector of $\mathbf{C}$, denoted here by $\mathbf{p}$. To demonstrate the stochastic convergence of the proposed iteration, we draw $S = 1000$ independent realizations of length-N data streams from the above Gaussian distribution and, for the i-th stream, we compute the proposed PC sequence $\{\mathbf{q}_t^{(i)}\}_{t=1,2,\ldots,N}$ according to (14). Then, we average-estimate the m.s. convergence metric as $\frac{1}{2S}\sum_{i=1}^{S} \|\mathbf{q}_t^{(i)}\mathbf{q}_t^{(i)^\top} - \mathbf{p}\mathbf{p}^\top\|_F^2$, for every t, where $\|.\|_F^2$ returns the squared Frobenius-norm of its matrix argument, and plot it in Fig. 2(b). The stochastic convergence of the proposed iteration is clearly documented.

Subspace estimation: We draw $N = 300$ independent points from $\mathcal{N}(\mathbf{0}_3, \mathbf{C})$, where $\mathbf{C} = \begin{bmatrix} 7 & 7 & 6 \\ 7 & 10 & 9 \\ 6 & 9 & 13 \end{bmatrix}$, and form

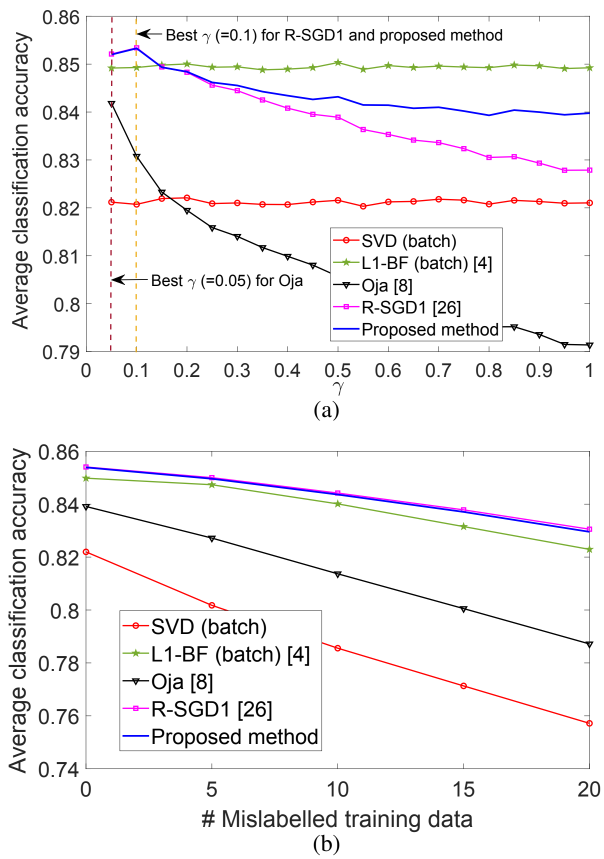

Fig. 4: Average classification accuracy vs. (a) γ and (b) number of mislabeled training points.

data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{3\times 300}$. Then, we add benign zero-mean white Gaussian noise (AWGN) from $2\mathcal{N}(0,1)$ to every entry of $\mathbf{X}$. In addition, we corrupt $\mathbf{x}_{120}$ and $\mathbf{x}_{230}$ with outlying corruption drawn from $\mathcal{N}(\mathbf{0}_3, \mathbf{C}_o)$, where $\mathbf{C}_o = \begin{bmatrix} 1585 & 1039 & -76 \\ 1039 & 1311 & -61 \\ -76 & -61 & 104 \end{bmatrix}$, such that the dominant eigenvector of the outlier covariance matrix $\mathbf{C}_o$, $\mathbf{p}_o$, makes an angle of about 45° with the dominant eigenvector of the nominal covariance matrix $\mathbf{C}$, $\mathbf{p}$. Then, we apply the proposed method to estimate $\mathbf{p}$, from the corrupted data stream. We repeat this task for $S = 10^4$ independent data realizations and plot in Fig. 3 the average-estimated mean subspace error $\frac{1}{S}\sum_{i=1}^{S} ||\mathbf{q}_t^{(i)}\mathbf{q}_t^{(i)\top} - \mathbf{p}\mathbf{p}^\top||_F^2$, versus t. Alongside the proposed algorithm, we also plot the performance of SVD (batch processing –joint calculation on $[\mathbf{X}]_{:,1:t}$ at the t-th iteration), Oja's standard stochastic PCA algorithm of [8], candid covariance-free incremental PCA (CCIPCA) [36], AdaOja [17], and R-SGD1 [26]. For fairness we tune the step sizes of Oja, R-SGD1, and the proposed algorithm to $\eta_t = \frac{0.05}{t^{0.6}}$. We notice that for $t < 120$, all methods perform similarly well and returned subspaces nearly converge to the span of $\mathbf{p}$. For $t \geq 120$, we observe that all L2-based methods (SVD, Oja, CCIPCA, and AdaOja) are significantly affected by the corruptions. On the other hand, the robust solver R-SGD1 [26] and the proposed algorithm show similarly good corruption resistance, maintaining low subspace error.

4.2. Wisconsin Breast Cancer Dataset

In this experiment, we perform nearest subspace (NS) classification on the Wisconsin breast cancer dataset [37], which contains $(D = 9)$-dimensional measurements on healthy and unhealthy cell nuclei. We first split the dataset into 85% training and 15% testing data. In order to identify a preferred learning rate for the stochastic methods, we vary $\gamma \in \{0.05, 0.05, 1\}$ and perform NS classification using the final element of each stochastic PC sequence. Specifically, we first estimate for each method the individual PCs of training data from class 1 (healthy) and class 2 (unhealthy). Next, for each testing point, we measure its squared projection error for each PC and assign it to the class with the PC that attained the lowest error. We repeat this experiment over 10^4 independent data splits and, in Fig. 4(a), we plot the average classification accuracy versus γ (including the batch methods as benchmarks). For the method of [8], $\gamma = 0.1$ is preferred, while for R-SGD1 and the proposed method $\gamma = 0.15$ is preferred. Interestingly, we notice that L1-BF outperforms SVD. In addition, we observe that the proposed method is significantly more robust than its counterparts against inferior choices of γ.

Next, we tune γ to the preferred values found above and mislabel a portion of the training data. We repeat the experiment over 10^3 independent data splits and plot in Fig. 4(b) the average NS classification accuracy versus number of mislabeled training data points from each class. First, we observe that as the number of mislabelings increases, the performance of SVD drops significantly. On the other hand, batch L1-BF, R-SGD1, and the proposed algorithm exhibit similar robustness against mislabeling.

5. CONCLUSIONS

We introduced a method for stochastic L1-PCA based on mean absolute projection maximization, with formal convergence guarantees. Our numerical studies on synthetic and real-world data demonstrate both the convergence and the corruption resistance of the proposed method.

6. REFERENCES

[1] M. B. Christopher, *Pattern recognition and machine learning*. New York, NY: Springer-Verlag, 2016.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, pp. 1–39, May 2011.

[3] P. P. Markopoulos, G. N. Karystinos, and D. A. Pados, "Optimal algorithms for L1-subspace signal processing," *IEEE Trans. Signal Process.*, vol. 62, pp. 5046–5058, Oct. 2014.

[4] P. P. Markopoulos, S. Kundu, S. Chamadia, and D. Pados, "Efficient L1-norm principal-component analysis via bit flipping," *IEEE Trans. Signal Process.*, vol. 65, pp. 4252–4264, Aug. 2017.

[5] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1672–1680, Sep. 2008.

[6] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy L1-norm maximization," in *Proc. Int. Joint Conf. Art. Intell.,* Barcelona, Spain, Jul. 2011, pp. 1433–1438.

[7] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Process. Mag.*, vol. 35, pp. 32–55, Jul. 2018.

[8] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math. Bio.*, vol. 15, pp. 267–273, Nov. 1982.

[9] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. Math. Anal. Appl.*, vol. 106, pp. 69–84, Feb. 1985.

[10] N. Vaswani and P. Narayanamurthy, "Static and dynamic robust PCA via low-rank + sparse matrix decomposition: A review," *arXiv preprint arXiv:1803.00651*, Mar. 2018.

[11] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Provable non-convex robust PCA," in *Adv. Neural Inf. Process. Syst.,* Montreal, Canada, Dec. 2014, pp. 1107–1115.

[12] C. J. Li, M. Wang, H. Liu, and T. Zhang, "Near-optimal stochastic approximation for online principal component estimation," *Math. Prog.*, vol. 167, pp. 75–97, Jan. 2018.

[13] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, "Streaming PCA: matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm," in *Proc. Conf. Learn. Theory,* New York, NY, Jun. 2016, pp. 1147–1164.

[14] O. Shamir, "Convergence of stochastic gradient descent for PCA," in *Proc. Int. Conf. Mach. Learn.,* New York, NY, Jun. 2016, pp. 257–265.

[15] S. Alakkari and J. Dingliana, "An acceleration scheme for memory limited, streaming PCA," *arXiv preprint arXiv:1807.06530*, Jul. 2018.

[16] T. V. Marinov, P. Mianjy, and R. Arora, "Streaming principal component analysis in noisy settings," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 3410–3419.

[17] A. Henriksen and R. Ward, "AdaOja: Adaptive learning rates for streaming PCA," *arXiv preprint arXiv:1905.12115*, May 2019.

[18] R. Ward, X. Wu, and L. Bottou, "AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization," *arXiv preprint arXiv:1806.01811*, June 2018.

[19] I. Mitliagkas, C. Caramanis, and P. Jain, "Memory limited, streaming PCA," in *Adv. Neural Inf. Process. Syst.,* Lake Tahoe, NV, Dec. 2013, pp. 2886–2894.

[20] D. G. Chachlakis, A. Prater-Bennette, and P. P. Markopoulos, "L1-norm Tucker Tensor Decomposition," *arXiv preprint arXiv:1904.06455*, Apr. 2019.

[21] D. G. Chachlakis, M. Dhanaraj, A. Prater-Bennette, and P. P. Markopoulos, "Options for multimodal classification based on L1-Tucker decomposition," in *Proc. SPIE, Big Data: Learning, Analytics, and Applications,* Baltimore, MD, vol. 10989, May. 2019, pp. 0O:1–0O:13.

[22] P. P. Markopoulos, D. G. Chachlakis, and A. Prater-Bennette, "L1-norm higher-order singular-value decomposition," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP),* Anaheim, CA, Nov. 2018, pp. 1353–1357.

[23] H. Guo, C. Qiu and N. Vaswani, "An online algorithm for separating sparse and low-dimensional signal sequences from their sum," *IEEE Trans. Signal Process.*, vol. 62, pp. 4284–4297, Aug., 2014.

[24] Y. Li, L. Q. Xu, J. Morphett, and R. Jacobs "An integrated algorithm of incremental and robust PCA," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, *Barcelona, Spain*, Sep. 2003, pp. 245–248.

[25] J. Feng, H. Xu, S. Mannor, and S. Yan, "Online PCA for contaminated data," in *Adv. Neural Inf. Process. Syst.,* Lake Tahoe, NV, Dec. 2013, pp. 764–772.

[26] J. Goes, T. Zhang, R. Arora, and G. Lerman, "Robust stochastic principal component analysis," in *Proc. Artificial Intell. Statist.,* Reykjavik, Iceland, Apr. 2014, pp. 266–274.

[27] P. P. Markopoulos, M. Dhanaraj, and A. Savakis, "Adaptive L1-norm principal-component analysis with online outlier rejection," *IEEE J. Select. Topics Signal Process.*, vol. 12, pp. 1131–1143, Dec. 2018.

[28] M. Dhanaraj and P. P. Markopoulos, "Novel algorithm for incremental L1-norm principal-component analysis," in *Proc. IEEE Eur. Signal Process. Conf. (EUSIPCO),* Rome, Italy, Sep. 2018, pp. 2020–2024.

[29] D. G. Chachlakis, P. P. Markopoulos, R. J. Muchhala, and A. Savakis, "Visual tracking with L1-Grassmann manifold modeling," in *Proc. SPIE,* Anaheim, CA, vol. 10211, Apr. 2017, pp. 02:1–02:10.

[30] M. Dhanaraj, D. G. Chachlakis, and P. P. Markopoulos, "Incremental complex L1-PCA for direction-of-arrival estimation," in *Proc. IEEE Western NY Image Signal Process. Workshop (WNYISPW),* Rochester, NY, Oct. 2018, pp. 1–5.

[31] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, pp. 400–407, Sep. 1951.

[32] X. Song, "An intuitive and most efficient L1-norm principal component analysis algorithm for big data," in *Proc. IEEE Annu. Conf. Inf. Sci. Syst.,* Baltimore, MD, 2019, pp. 1–4.

[33] S. Kim, "Gradient-based simulation optimization," in *Proc. IEEE Wint. Simul. Conf.,* Monterey, CA, Dec. 2006, pp. 159–167.

[34] R. Mahony, U. Helmke, and J. Moore, "Gradient algorithms for principal component analysis," *The ANZIAM Journal*, vol. 37, no. 4, pp. 430–450, 1996.

[35] F. Leone, L. Nelson, and R. Nottingham, "The folded normal distribution," *J. Technometrics*, vol. 3, pp. 543–550, Nov. 1961.

[36] J. Weng, Y. Zhang, and W.-S. Hwang, "Candid covariance-free incremental principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 1034–1040, Aug. 2003.

[37] Breast Cancer Wisconsin (Original) Data Set [Online]. Available: *http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29.*