Online Estimation of Coherent Subspaces with Adaptive Sampling

Greg Ongie, David Hong, Dejiao Zhang, Laura Balzano
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48108
e-mail: {gongie,dahong,dejiao,girasole}@umich.edu

***Abstract*—This work investigates adaptive sampling strategies for online subspace estimation from streaming input vectors where the underlying subspace is coherent, i.e., aligned with some subset of the coordinate axes. We adapt the previously proposed Grassmannian rank-one update subspace estimation (GROUSE) algorithm to incorporate an adaptive sampling strategy that substantially improves over uniform random sampling. Our approach is to sample some proportion of the entries based on the leverage scores of the current subspace estimate. Experiments on synthetic data demonstrate that the adaptive measurement scheme greatly improves the convergence rate of GROUSE over uniform random measurements when the underlying subspace is coherent.**

I. Introduction

Subspace estimation and tracking plays a crucial role in several signal processing tasks, including identification of network anomalies [1], beamforming [2], and medical imaging [3], among others. Given a sequence of input vectors, the goal in these problems is to estimate a linear low-dimensional subspace that describes the data well. However, for many applications it is challenging or impossible to achieve full sampling of the input vectors due to their high-dimensionality or due to costs associated with the measurement process. To remedy this issue, several online subspace estimation algorithms have been proposed that can accommodate entrywise undersampling and/or compressive linear measurements of the input stream [4]–[7].

A common sampling strategy for these algorithms is to select entries uniformly at random [4]–[6]. Such a strategy is justified when the underlying subspace is *incoherent*, i.e., the subspace is unaligned with the coordinate axes. For example, under this assumption, local convergence of the Grassmannian Online Rank One Subspace Estimation (GROUSE) algorithm with uniform random sampling has been shown in [8], [9]. However, when the underlying subspace is coherent, existing guarantees in [8], [9] break down. Indeed, we show in the experiments section that GROUSE with uniform random sampling performs poorly for coherent subspaces. Intuitively, this is because vectors drawn from a coherent subspace have most of their energy contained in a few coordinates, and a small uniform random sample will miss these entries with high probability. However, estimating coherent subspaces is essential in a variety of problems involving outliers, for example in recommendation systems with popular items or highly active users, or network monitoring with anomalous hosts.

To overcome this issue, we propose an *adaptive* sampling strategy for coherent subspace estimation using GROUSE. In particular, we propose sampling entries based on the *statistical leverage scores* of the current subspace estimate (see Definition 1). This approach biases the sampling towards entries predicted to contain significant energy of the input vectors. Empirically, we show this greatly improves the convergence rate of GROUSE when the underlying subspace is coherent.

A. Related Work

To the best of our knowledge, the only previous works to consider adaptive sensing (i.e., active learning) for online subspace estimation are [7], [10]. Motivated by multi-armed bandits theory, [7] considers strategies for active selection of entrywise samples of input vectors and proves results on the sampling complexity of their approach. However, their approach is memory and computationally intensive, requiring $O(n^2)$ storage and $O(n^3)$ flops per iteration, where n is the ambient dimension. Our approach, based on the GROUSE algorithm, has $O(nd)$ memory and $O(nd+md^2)$ flops per iteration, where $d \ll n$ is the subspace dimension and $m \leq n$ is the number of sampled entries per time instance. An online adaptive matrix completion algorithm is also proposed in [10]. However, [10] assumes the user has the ability to sample each input vector twice. Their approach is to first sample enough measurements to test whether

a given vector is in the current column space. If it is not, they propose sampling the column fully, and then repeating this process to learn the entire column space, which is not always possible with sensing or memory constraints. In contrast, our approach uses only a small constant fraction of the entries per time instance.

In the matrix completion setting several authors have proposed adaptive sampling strategies for coherent matrix/tensor completion [10]–[13]. This paper can be thought of as an extension of these works to the online setting. In particular, similar to [11], [12] we propose sampling entries based on their statistical leverage scores (see Definition 1), but whereas other works have had access to offline estimates of those scores, we demonstrate that one need only use the scores of the current subspace estimate.

Finally, we recently investigated adaptive sensing for GROUSE assuming arbitrary linear measurements can be taken [14]. The present work extends this line of inquiry to the entrywise sampling model with a focus on coherent subspaces.

II. Problem Formulation and Algorithm

For any matrix $\boldsymbol{V}$ let $\mathcal{R}(\boldsymbol{V})$ denote the range space of $\boldsymbol{V}$, i.e., the linear span of the columns of $\boldsymbol{V}$. We model the ground truth data as a sequence of vectors $\{\boldsymbol{x}_t\}_{t=1}^T$ drawn from a fixed d-dimensional subspace $\mathcal{S} \subset \mathbb{R}^n$ according to the generative model:

$$\boldsymbol{x}_t = \bar{\boldsymbol{U}}\bar{\boldsymbol{w}}_t, \quad t = 1, ..., T, \tag{1}$$

where the columns of $\bar{\boldsymbol{U}} \in \mathbb{R}^{n\times d}$ form an orthonormal basis for $\mathcal{S}$, meaning $\mathcal{S} = \mathcal{R}(\bar{\boldsymbol{U}})$ and $\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}} = \boldsymbol{I}_{d\times d}$, and $\bar{\boldsymbol{w}}_t \in \mathbb{R}^d$ are the subspace weights at time t. We suppose that for each time t we observe $m \geq d$ entries of $\boldsymbol{x}_t$ indexed by $\Omega_t \subset \{1, ..., n\}$:

$$\boldsymbol{y}_t = \boldsymbol{P}_{\Omega_t}\boldsymbol{x}_t \in \mathbb{R}^m; \quad t = 1, ..., T, \tag{2}$$

where $\boldsymbol{P}_{\Omega_t} : \mathbb{R}^n \to \mathbb{R}^m$ denotes the restriction of a vector to its entries indexed by Ω_t. Our goal is to estimate the subspace $\mathcal{S}$ from the sequence $\{\boldsymbol{y}_t\}_{t=1}^T$. The main question we investigate in this work is whether $\mathcal{S}$ can be estimated more efficiently by adaptively choosing the sampling set Ω_t at each time t.

For our online subspace estimation algorithm we investigate a modification of the GROUSE algorithm [4], [9]. GROUSE is designed to approximately minimize the following global cost function in an online fashion:

$$\min_{\boldsymbol{U}, \{\boldsymbol{w}_t\}_{t=1}^T} \sum_{t=1}^{T} \|\boldsymbol{P}_{\Omega_t}\boldsymbol{U}\boldsymbol{w}_t - \boldsymbol{y}_t\|^2 \;\; \text{s.t.} \;\; \mathcal{R}(\boldsymbol{U}) \in \mathcal{G}(n, d) \tag{3}$$

where $\mathcal{G}(n, d)$ denotes the Grassmannian, the set of d-dimensional subspaces in $\mathbb{R}^n$. At each time t the GROUSE algorithm performs one step of block coordinate descent applied to the local cost function

$$\min_{\boldsymbol{U}, \boldsymbol{w}_t} \|\boldsymbol{P}_{\Omega_t}\boldsymbol{U}\boldsymbol{w}_t - \boldsymbol{y}_t\|^2 \;\; \text{s.t.} \;\; \mathcal{R}(\boldsymbol{U}) \in \mathcal{G}(n, d). \tag{4}$$

Let $\boldsymbol{U}_t$ be the current subspace estimate. Fixing $\boldsymbol{U} = \boldsymbol{U}_t$ in (4), the optimal weights $\boldsymbol{w}_t$ are given by

$$\boldsymbol{w}_t = (\boldsymbol{P}_{\Omega_t}\boldsymbol{U}_t)^\dagger \boldsymbol{y}_t \tag{5}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseduoinverse. Then, with the optimal weights $\boldsymbol{w}_t$ fixed, GROUSE updates the subspace representative $\boldsymbol{U}_{t+1}$ by taking a gradient step of the objective (4) along a geodesic on the Grassmannian. These steps are summarized in Algorithm 1. For their derivation see [4]. We make use of the "greedy" gradient step-size for GROUSE proposed in [9].

III. Adaptive Sampling Scheme

To overcome the limitations of uniform random sampling in the case of coherent subspaces, we propose an *adaptive* sampling scheme that biases the sampling towards coordinates where the signal energy is most concentrated. Specifically, we incorporate the statistical leverage scores [15] of our subspace estimates, defined as follows:

Definition 1. Let $\mathcal{S}$ be a d-dimensional subspace in $\mathbb{R}^n$ and let $\boldsymbol{U}$ denote an $n \times d$ basis matrix for $\mathcal{S}$ with orthonormal columns. Let $\boldsymbol{U}(i)$ denote the i-th row of the matrix $\boldsymbol{U}$. Then, the *statistical leverage scores* of $\mathcal{S}$ are given by $\ell_i = \|\boldsymbol{U}(i)\|_2^2$ for $i \in \{1, ..., n\}$.

Note that the statistical leverage scores do not depend on the particular basis $\boldsymbol{U}$, but only on the space spanned by that basis. To see this, let $\boldsymbol{P}_{\mathcal{S}}$ denote the projection matrix onto $\mathcal{S}$, the span of the columns of $\boldsymbol{U}$. Then,

$$\ell_i = \|\boldsymbol{U}(i)\|_2^2 = \|\boldsymbol{U}^T\boldsymbol{e}_i\|_2^2 = \|\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{e}_i\|_2^2 = \|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{e}_i\|_2^2.$$

i.e., the statistical leverage scores are the squared norm of the projection of the canonical basis elements $\{\boldsymbol{e}_i\}_{i=1}^n$ onto $\mathcal{S}$. Since $\boldsymbol{U}$ is orthonormal, we have $\sum_{i=1}^n \ell_i = \|\boldsymbol{U}\|_F^2 = d$. Therefore, we can define a probability distribution over the indices $\{1, ..., n\}$ by $p_i = \ell_i/d$.

We propose the following adaptive sampling procedure for online coherent subspace estimation: compute the normalized statistical leverage scores $p_i = \ell_i/d$ of the current subspace estimate $\boldsymbol{U}_t$, and sample m

indices $\omega \in \{1, ..., n\}$ with replacement according to the probability distribution

$$\mathsf{P}(\omega = i) = \beta\, p_i + (1-\beta)\tfrac{1}{n} \tag{6}$$

for some $\beta \in [0, 1]$. Put in words, we sample indices according to the convex combination of the distribution based on the statistical leverage scores and a uniform distribution. Here β is a tunable parameter that allows us to trade-off between adaptive and non-adaptive sampling strategies: at one extreme ($\beta = 0$) the sampling indices are drawn uniformly at random, at the other extreme ($\beta = 1$) the sampling indices are adaptively chosen according to those previously estimated to have high statistical leverage scores. In our experiments we show that the extremes $\beta = 0$ or 1 perform well only when the subspace is either maximally incoherent or exactly sparse, respectively. For general coherent subspaces our experiments show $0 < \beta < 1$ is a more robust strategy.

The sampling scheme (6) is also motivated by our earlier work [14], which studied an adaptive sensing scheme for GROUSE where one is allowed to take arbitrary linear measurements of the input vectors and for which we provided global convergence guarantees. However, directly applying the adaptive sensing scheme in [14] to the present setting is not possible since our measurement vectors are constrained to be canonical basis vectors. We can approximate the scheme in [14] by finding $k \geq d$ canonical basis vectors indexed by Ω' whose span best approximates the span of $\boldsymbol{U}_t$ in the following sense:

$$\Omega' = \operatorname*{arg\,min}_{|\Omega|=k} \|\boldsymbol{P}_\Omega^T \boldsymbol{P}_\Omega - \boldsymbol{U}_t \boldsymbol{U}_t^T\|_F^2 = \operatorname*{arg\,max}_{|\Omega|=k} \|\boldsymbol{P}_\Omega \boldsymbol{U}_t\|_F^2$$

It is easy to see that Ω' is given by the indices corresponding to the top k statistical leverage scores of $\mathcal{R}(\boldsymbol{U}_t)$. This motivates the sampling scheme (6), which is biased towards indices with high statistical leverage scores.

IV. Experiments

This section illustrates the empirical performance of the proposed adaptive sampling scheme on simulated data. We compare Adaptive GROUSE (Algorithm 1) to GROUSE with non-adaptive uniform random sampling (Algorithm 1 with $\beta = 0$), which we call Non-adaptive GROUSE. We generated data according to the model (1). A random subspace basis $\bar{\boldsymbol{U}} \in \mathbb{R}^{200\times 5}$ was constructed as the left singular vectors of a matrix $\boldsymbol{D}\boldsymbol{U}_0$ where $\boldsymbol{U}_0$ is a random matrix with orthonormal columns and $\boldsymbol{D} = \text{diag}(1^\alpha, 2^\alpha, ..., 200^\alpha)$ for $\alpha \in \{0, 1, 4\}$; larger values of α generate subspaces that are more coherent.

Algorithm 1: Adaptive GROUSE

Choose $0 \leq \beta \leq 1$.

For $t = 1, 2, ...$ do the following.

1. Compute normalized leverage scores:

$p_i = \frac{1}{d}\|\boldsymbol{U}_t(i)\|_2^2,\ i = 1, ..., n.$

2. Draw m indices $\omega \in \{1, ..., n\}$ according to

$P(\omega = i) = \beta\, p_i + (1-\beta)\frac{1}{n}$

3. Update subspace estimate:

update weights: $\boldsymbol{w}_t = (\boldsymbol{P}_{\Omega_t}\boldsymbol{U}_t)^\dagger \boldsymbol{y}_t$

compute projection: $\boldsymbol{p}_t = \boldsymbol{U}_t \boldsymbol{w}_t$

compute residual: $\boldsymbol{r}_t = \boldsymbol{P}_{\Omega_t}^T(\boldsymbol{y}_t - \boldsymbol{P}_{\Omega_t}\boldsymbol{p}_t)$

compute stepsize: $\theta_t = \arctan\left(\frac{\|\boldsymbol{r}_t\|}{\|\boldsymbol{p}_t\|}\right)$

update subspace:

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t + \left(\sin(\theta_t)\frac{\boldsymbol{r}_t}{\|\boldsymbol{r}_t\|} + (\cos(\theta_t) - 1)\frac{\boldsymbol{p}_t}{\|\boldsymbol{p}_t\|}\right)\frac{\boldsymbol{w}_t^T}{\|\boldsymbol{w}_t\|}$$

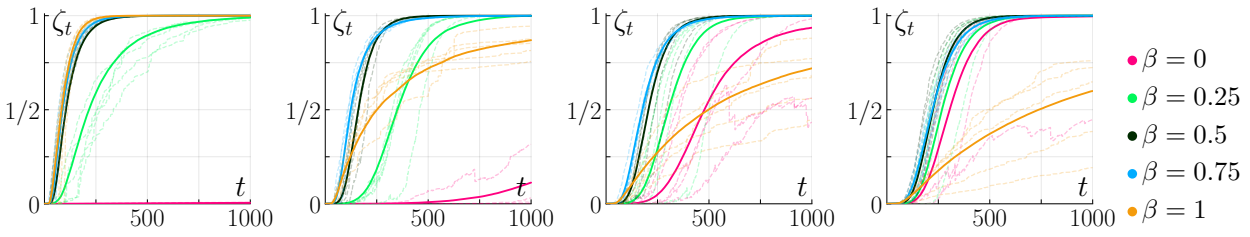
We also considered the "sparse and low-rank" setting where $\bar{\boldsymbol{U}}$ has columns sampled from the identity matrix. We generated subspace weights $\bar{\boldsymbol{w}}_t \sim \mathcal{N}(0, \boldsymbol{I}_{5\times 5})$ and at each time t observed $m = 20$ entries of $\boldsymbol{x}_t$ (10% of the entries). Following [9], we measure recovery performance in terms of the *determinant similarity* $\zeta_t \in [0, 1]$ between the current subspace estimate $\mathcal{R}(\boldsymbol{U}_t)$ and the true subspace $\mathcal{R}(\bar{\boldsymbol{U}})$:

$$\zeta_t := \det(\boldsymbol{U}_t^T \bar{\boldsymbol{U}}\, \bar{\boldsymbol{U}}^T \boldsymbol{U}_t) = \prod_{k=1}^{d} \cos^2(\phi_k) \tag{7}$$

where ϕ_k is the kth principal angle between the two.

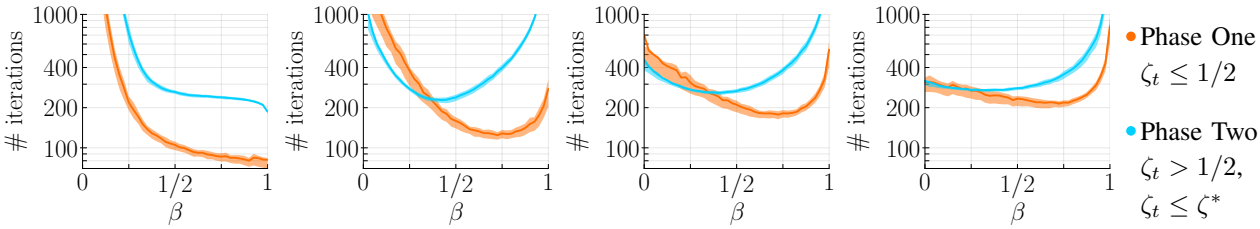
Figure 1 shows the determinant similarity (7) over 1000 iterations of Adaptive GROUSE (Algorithm 1) for various choices of β. In the sparse and low-rank setting, Non-adaptive GROUSE ($\beta = 0$) typically fails but Adaptive GROUSE with $\beta = 1$, i.e., sampling entirely by leverage scores, achieves a target average determinant similarity of $\zeta_t \geq \zeta^* = 0.99$ with $t = 272$ iterations. For increasingly incoherent subspaces, the performance for $\beta = 1$ declines while that for $\beta = 0$ improves. Notably, however, $\beta = 0.5$ and $\beta = 0.75$ perform consistently across the spectrum. For all four levels of coherence, $\beta = 0.5$ achieves the target ζ^* with $t = 365$ iterations.

What is remarkable about these empirical results is that *despite initialization with a random subspace*, the estimate of the leverage scores at each time t seems to provide a useful indicator of where to sample. This is reasonable initially, as a random incoherent subspace estimate will have uniform leverage scores and therefore lead to uniform samples. It seems that any observation we get with nonzero magnitude then biases the leverage



(a) Sparse and Low-rank (b) Coherent: $\alpha = 4$ (c) Coherent: $\alpha = 1$ (d) Incoherent: $\alpha = 0$

Fig. 1: Determinant similarity (7) over the course of one thousand iterations of Adaptive GROUSE (Algorithm 1) for various choices of β and subspaces of varying coherence. The range for β extends from $\beta = 0$, i.e., Non-adaptive GROUSE, to $\beta = 1$, i.e., sampling entirely by leverage scores. For each choice, the traces from five sample runs are shown as dashed lines, and the mean from two hundred runs is shown as a solid line.



(a) Sparse and Low-rank (b) Coherent: $\alpha = 4$ (c) Coherent: $\alpha = 1$ (d) Incoherent: $\alpha = 0$

Fig. 2: Iterations required to pass through two phases of GROUSE convergence versus β for subspaces of varying coherence. Phase one consists of the iterations with determinant similarity $\zeta_t \leq 1/2$; phase two consists of those with determinant similarity $1/2 < \zeta_t \leq \zeta^* = 0.99$. In all cases, the mean number of iterations from two hundred trials is shown as a solid line with the interquartile interval overlaid as a ribbon. Recall that $\beta = 0$ on one end corresponds to Non-adaptive GROUSE and $\beta = 1$ corresponds to sampling entirely according to leverage scores.

scores of the current subspace estimate towards the important coordinates. We note that these experiments are run without added noise and on a relatively small ambient dimension ($d = 200$). Understanding the evolution of the leverage scores in this process is an important topic of future work.

Following the analysis in [9], which identified two distinct phases in the convergence of GROUSE, Figure 2 shows the number of iterations needed to: 1) reach a determinant similarity of $\zeta_t \geq 1/2$ (phase one), and 2) go from there to reach the target determinant similarity of $\zeta_t \geq \zeta^* = 0.99$ (phase two). As in Figure 1, $\beta = 1$ results in the fewest iterations for both phases in the sparse and low-rank setting, and for increasingly incoherent subspaces, the number of iterations for $\beta = 1$ increases while those for $\beta = 0$ decreases. Again β between 0 and 1 provides the best performance in intermediate cases. The difference is particularly significant for coherent subspaces with $\alpha = 4$, where using either $\beta = 0, 1$ requires over a hundred more iterations in both phases. The choice of β that yields the lowest average number of iterations in phase one is also different from that for phase two in all but the sparse and low-rank setting, further highlighting the difference in convergence behavior between the two phases. An interesting area of future work is identifying a data-driven approach to adapting the choice of β over the course of the iterations.

V. Conclusion

This paper shows that an adaptive sampling scheme based on statistical leverage scores can improve the performance of an online subspace estimation algorithm in the case of coherent subspaces. While in this work we focused on the GROUSE subspace estimation algorithm, the adaptive sampling scheme proposed here might also enhance the performance of other recently proposed online subspace estimation and tracking algorithms [5]–[7], [16]. In future work we hope to extend the convergence theory of GROUSE established in [9], [14] to the present setting, as well.

Acknowledgment

This work was supported by DARPA grant 16-43-D3M-FP-037 and NSF Grant ECCS-1508943.

REFERENCES

[1] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4. ACM, 2004, pp. 219–230.

[2] A. M. Haimovich and Y. Bar-Ness, "An eigenanalysis interference canceler," *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 76–84, 1991.

[3] B. A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno, "Activation detection in functional MRI using subspace modeling and maximum likelihood estimation," *IEEE Transactions on Medical Imaging*, vol. 18, no. 2, pp. 101–114, 1999.

[4] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Allerton Conference on Communication, Control, and Computing*. IEEE, 2010, pp. 704–711.

[5] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1568–1575.

[6] Y. Chi, Y. C. Eldar, and R. Calderbank, "PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5947–5959, 2013.

[7] A. Gonen, D. Rosenbaum, Y. C. Eldar, and S. Shalev-Shwartz, "Subspace learning with partial information," *Journal of Machine Learning Research*, vol. 17, no. 52, pp. 1–21, 2016.

[8] L. Balzano and S. J. Wright, "Local convergence of an algorithm for subspace identification from partial data," *Foundations of Computational Mathematics*, vol. 15, no. 5, pp. 1279–1314, 2015.

[9] D. Zhang and L. Balzano, "Convergence of a grassmannian gradient descent algorithm for subspace estimation from undersampled data," *arXiv preprint arXiv:1610.00199*, 2016.

[10] A. Krishnamurthy and A. Singh, "Low-rank matrix and tensor completion via adaptive sampling," in *Advances in Neural Information Processing Systems*, 2013, pp. 836–844.

[11] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Coherent matrix completion," in *International Conference on Machine Learning*, 2014, pp. 674–682.

[12] ——, "Completing any low-rank matrix, provably." *Journal of Machine Learning Research*, vol. 16, pp. 2999–3034, 2015.

[13] A. Eftekhari, M. B. Wakin, and R. A. Ward, "MC^2: A two-phase algorithm for leveraged matrix completion," *arXiv preprint arXiv:1609.01795*, 2016.

[14] G. Ongie, D. Hong, D. Zhang, and L. Balzano, "Enhanced online subspace estimation via adaptive sensing," in *Asilomar Conference on Signals, Systems, and Computers*, 2017.

[15] M. W. Mahoney, "Randomized algorithms for matrices and data," *Foundations and Trends® in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[16] C. Qiu, N. Vaswani, B. Lois, and L. Hogben, "Recursive robust PCA or recursive sparse recovery in large but structured noise," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 5007–5039, 2014.