Enhanced Online Subspace Estimation via Adaptive Sensing

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 the problem of adaptive measurement design for online subspace estimation from compressive linear measurements. We study the previously proposed Grassmannian rank-one online subspace estimation (GROUSE) algorithm with adaptively designed compressive measurements. We propose an adaptive measurement scheme that biases the measurement vectors towards the current subspace estimate and prove a global convergence result for the resulting algorithm. Our experiments on synthetic data demonstrate the effectiveness of the adaptive measurement scheme over non-adaptive compressive random measurements.

I. INTRODUCTION

Subspace estimation and tracking plays an important role in many signal processing tasks such as identification of network anomalies [1] and beamforming [2] among others. Given a sequence of input vectors, the goal in these problems is to estimate a linear lowdimensional 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 [3]–[6].

However, little attention has been given to adaptive or active sampling strategies for online subspace estimation. The typical approach is to undersample the input vectors uniformly at random [3]–[5] or take isotropic random Gaussian measurements [7]. Motivated by the success of adaptive sensing for compressed sensing of sparse vectors [8]–[14] and low-rank matrix/tensor completion [15]–[17], we investigate an adaptive sensing strategy for online subspace estimation from compressive linear measurements using the Grassmannian rank-one update subspace estimation algorithm (GROUSE) [3], [7]. We show that an adaptive sensing strategy that biases the sensing vectors towards the current subspace estimate improves the GROUSE convergence rate relative to nonadaptive random Gaussian measurements. Extending recently developed convergence theory in [7] we also give a global convergence result for GROUSE with the proposed adaptive compressive measurement scheme, providing the first such guarantee for GROUSE with compressive measurements. Finally, we demonstrate the benefits of the adaptive sensing scheme with several numerical simulations.

A. Related Work

To the best of our knowledge, the only previous work to consider adaptive sensing (i.e., active learning) for online subspace estimation is [6]. Motivated by multiarmed bandits theory, the authors of [6] consider strategies for active selection of entrywise samples of input vectors and prove results on the sampling complexity of their approach. The present work differs from [6] in that we consider adaptive sensing specifically for the GROUSE online subspace estimation algorithm and use a different adaptive measurement model. In particular, we assume that we can take arbitrary linear measurements of the input vectors, rather than entrywise samples.

II. PROBLEM FORMULATION

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

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

where the columns of $U \in \mathbb{R}^{n \times d}$ form an orthonormal basis for S, meaning $S = \mathcal{R}(U)$ and $U^T U = I_{d \times d}$, and $w_t \in \mathbb{R}^d$ are the subspace weights at time t. We suppose that for each time t we observe m linear measurements:

$$\boldsymbol{y}_t = \boldsymbol{A}_t \boldsymbol{x}_t \in \mathbb{R}^m; \quad t = 1, ..., T,$$
(2)

where the *measurement matrix* $A_t \in \mathbb{R}^{m \times n}$ can be adaptively designed and $d < m \ll n$. Our goal is to estimate S from the sequence $\{y_t\}_{t=1}^T$. The main question we investigate in this work is whether the subspace S can be estimated more efficiently by choosing an appropriate design of the measurement matrices $\{A_t\}_{t=1}^T$. We will assume the choice of each measurement matrix A_t is unconstrained except that its *sensing energy* $||A_t||_F$ is bounded above by a fixed constant for all times t. In this work we will use the normalization $||A_t||_F \leq m$ for all t, which holds, for example, when the rows of A_t are unit-norm. When A_t is a random matrix, we assume $||A_t||_F \leq m$ holds in expectation.

III. ADAPTIVE GROUSE

A. Compressive GROUSE algorithm

For our online subspace estimation algorithm we investigate a modification of the compressive GROUSE algorithm proposed in [7]. GROUSE is designed to approximately minimize the following global cost function in an online fashion:

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

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

Let U_t be the current estimate of U. Fixing $U = U_t$ in (4), the optimal weights w_t are given by

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

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseduoinverse. Then, with the optimal weights w_t fixed, GROUSE updates the subspace representative U_{t+1} by taking a gradient step of the objective (4) along a geodesic in the Grassmannian. These steps are summarized in Algorithm 1. For their derivation see [3], [7].

B. Proposed adaptive measurement design

Given the subspace estimate U_t at time t, we propose using the adaptive measurement design

$$\boldsymbol{A}_{t} = \begin{bmatrix} \boldsymbol{U}_{t}^{T} \\ \boldsymbol{B}_{t}^{T} \end{bmatrix} \in \mathbb{R}^{m \times n}$$
(6)

where the columns of $B_t \in \mathbb{R}^{n imes (m-d)}$ are drawn randomly from the orthogonal complement of $\mathcal{R}(U_t)$ in a manner specified below. Put in words, we draw d measurement vectors from the current subspace estimate and draw the remaining m - d measurement vectors randomly from its orthogonal complement.

The idea behind this scheme is that including the columns of U_t as measurement vectors gives us the projection of the ground truth data x_t onto the subspace $\mathcal{R}(U_t)$ at each iteration. As the estimate of $\mathcal{R}(U_t)$ improves, the projection of x_t onto $\mathcal{R}(U_t)$ will contain an increasingly larger proportion of the total energy of x_t , yielding a positive feedback effect which enables faster convergence of the algorithm. We quantify this effect in the next section.



IV. MAIN RESULTS

The main theoretical contribution of this work is a global convergence theorem for adaptive GROUSE (Algorithm 1). These results extend the convergence results in [7] given for GROUSE with non-adaptive measurements. Due to space restrictions, we omit the proofs.

We make the following statistical assumptions on the data and measurement models:

Assumption 1. The underlying data is generated according to the model (1), where for each time t the subspace weights $w_t \in \mathbb{R}^d$ are uncorrelated and the entries of w_t have zero mean and unit variance.

Assumption 2. For each time t, the measurements B_t in (6) are generated according to $B_t = (I_{n \times n} - U_t U_t^T) G_t$, where $G_t \in \mathbb{R}^{n \times (m-d)}$ is a random matrix with i.i.d sub-gaussian entries $G_{i,j}$, such that $\mathbb{E}[G_{i,j}] = 0$ and $\operatorname{Var}[G_{i,j}] = \frac{1}{n-d}$.

We state our results in terms of the following similarity measure for subspaces:

Definition 1. The *determinant similarity* $\zeta \in [0,1]$ between subspaces $\mathcal{R}(U)$ and $\mathcal{R}(V)$ is defined as

$$\zeta := \det(\boldsymbol{U}^T \boldsymbol{V} \boldsymbol{V}^T \boldsymbol{U}) = \prod_{k=1}^a \cos^2(\phi_k)$$
(7)

where ϕ_k is the *k*th principal angle between $\mathcal{R}(U)$ and $\mathcal{R}(V)$.

Let ζ_t denote the determinant similarity between the underlying subspace $\mathcal{R}(U)$ and the estimate $\mathcal{R}(U_t)$ at time *t*. Our first result quantifies the expected improvement in ζ_t for each iteration of Algorithm 1.

Lemma 1. For each iteration t of Algorithm 1 the expected improvement in determinant similarity is at least

$$\mathbb{E}\left[\frac{\zeta_{t+1}}{\zeta_t}, \boldsymbol{U}_t\right] \ge 1 + \eta \frac{m-d}{n} \frac{1-\zeta_t}{d}, \qquad (8)$$

where $\eta = 1 - \frac{m+1}{n}$.

Lemma 1 should be compared with Lemma 9 in [7] that shows GROUSE with $m = O(d \log n)$ non-adaptive random Gaussian measurements achieves

$$\mathbb{E}\left[\frac{\zeta_{t+1}}{\zeta_t}, \boldsymbol{U}_t\right] \ge 1 + \eta_1 \frac{m}{n} \frac{1 - \zeta_t}{d},\tag{9}$$

with high probability, where $\eta_1 \approx 1$ is a dimensionallydependent constant. The constant $\eta_1 \frac{m}{n}$ is slightly better than $(1 - \frac{m+1}{n})\frac{m-d}{n}$ in Lemma 1. However, a major difference is that Lemma 1 is non-probabilistic and holds for all m > d, rather than with high probability when $m = O(d \log n)$. This suggests that the adaptive measurement scheme may perform better when the number of measurements per time point m is close to the subspace dimension d.

We now present our main result: the global convergence of the Adaptive GROUSE algorithm. The result is a direct consequence of Lemma 1, and the proof is similar to that of Theorem 7 in [7], which considers the case of fully sampled data.

Theorem 1. Let $0 < \zeta^* \leq 1$ be a desired accuracy and let the initialization $\mathcal{R}(\mathbf{U}_0)$ coincide with the column span of an $n \times d$ random matrix whose entries are i.i.d. standard normal random variables. Then for any p > 0, after

$$K \ge \left(\frac{n}{\eta(m-d)}\frac{2d^2}{\rho} + 1\right)\tau_0\log(n) + \frac{n}{\eta(m-d)}2d\log\left(\frac{1}{2\rho(1-\zeta^*)}\right)$$

iterations of Algorithm 1,

$$\mathbb{P}(\zeta_K \ge \zeta^*) \ge 1 - 2p, \tag{10}$$

where $\tau_0 = 1 + \frac{\log(\frac{(1-\rho/2)}{C}) + d\log(e/d)}{d\log(n)}$, C > 0 is a constant approximately equal to 1, and η is the same as in Lemma 1.

An analogous global convergence result for the nonadaptive version of Algorithm 1 where A_t consists of random Gaussian entries has not been established yet. The main impediment is that the expected improvement in the determinant similarity is not monotonic in that case. The proposed adaptive sensing strategy, on the other hand, ensures monotonic improvement of the determinant similarity in expectation, allowing us to obtain the global convergence guarantee of Theorem 1.

V. EXPERIMENTS

This section illustrates the empirical performance of the proposed adaptive sensing scheme on simulated data. We compare Adaptive GROUSE (Algorithm 1) to GROUSE with non-adaptive random measurements (Algorithm 1 with A_t chosen as a random Gaussian matrix), which we call Non-adaptive GROUSE. We generated data according to the model (1) with random subspace basis U and i.i.d. subspace weights $w_t \sim \mathcal{N}(0, I_{d \times d})$, matching Assumption 1. The measurement matrices for Non-adaptive GROUSE are generated with i.i.d. $\mathcal{N}(0, 1/n)$ entries, and those for Adaptive GROUSE are generated according to (6) with $B_t = (I_{n \times n} - U_t U_t^T)G_t$ where G_t has i.i.d. $\mathcal{N}(0, 1/(n - d))$ entries, matching Assumption 2.

Figures 1 and 2 consider three settings: a) n = 200, d = 5, m = 10, b) n = 200, d = 2, m = 5, and c) n = 200, d = 5, m = 30. Settings (a) and (b) have relatively few measurements whereas setting (c) has around $d \log n$ measurements. Figure 1 shows the determinant similarity (7) over several iterations of the two approaches. In (a) and (b), Adaptive GROUSE converges in fewer iterations and does so more consistently than Non-adaptive GROUSE, which has varied convergence rates in (a) and rarely converges in (b). They perform similarly in (c), which illustrates that the benefit of adaptive measurements diminishes as the number of measurements per time instant increases.

Figure 2 shows the improvement ζ_{t+1}/ζ_t in determinant similarity of a single iteration versus ζ_t for the two approaches. The estimates of the conditional expected improvements are formed by averaging the improvements ζ_{t+1}/ζ_t for all iterates with determinant similarity ζ_t within a 0.01 radius; values of ζ_t with fewer than 400



Fig. 1: Determinant similarity (7) over several iterations of Non-adaptive GROUSE (shown in blue) and Adaptive GROUSE (shown in red). For each approach, the traces from twenty sample runs are shown as dashed lines, and the mean from two hundred runs is shown as a solid line.



Fig. 2: Improvement ζ_{t+1}/ζ_t in determinant similarity of a single iteration versus determinant similarity ζ_t for Nonadaptive GROUSE (shown in blue) and Adaptive GROUSE (shown in red). For each approach, the determinant similarity ζ_t and improvement ζ_{t+1}/ζ_t for three thousand samples are shown as dots, and an estimate of the conditional expected improvement is shown as a solid line. The theoretical bound (8) of Lemma 1 is shown as a solid green line.



(a) First phase: iterations from initialization to $\zeta_t \ge 1/2$.



Fig. 3: Iterations required to pass through two phases of GROUSE convergence versus the number of samples m for Non-adaptive GROUSE (shown in blue) and Adaptive GROUSE (shown in red). For each approach, the mean number of iterations is shown as a solid line with the interquartile interval overlaid as a ribbon.

iterates within a 0.01 radius are excluded. In (a) and (b), Adaptive GROUSE has greater conditional expected improvement than Non-adaptive GROUSE, especially for determinant similarities less than one half. Non-adaptive GROUSE rarely converged in (b) and hence its conditional expected improvements stay below one. The two approaches again perform similarly in (c). In all three settings, the theoretical bound (8) of Lemma 1 closely agrees with the conditional expected improvement when $\zeta_t > 1/2$ and is conservative when $\zeta_t < 1/2$; initial improvement is even better in practice than predicted.

Figure 3 shows, for various numbers of measurements, the number of iterations required to pass through two phases: first, to reach a determinant similarity of $\zeta_t \ge 1/2$, and second, to go from there to a determinant similarity of $\zeta_t \ge \zeta^* = 0.99$. Figure 3a illustrates that Adaptive GROUSE generally requires fewer iterations to pass through the first phase, and the benefit diminishes as the number of measurements per time instant increases. The second phase illustrated in Figure 3b takes a similar number of iterations for the two approaches. Namely, the adaptive approach is primarily helpful early on (in the first phase) and especially so when the number of measurements is small. In the second phase, the adaptive approaches have roughly the same performance.

VI. DISCUSSION AND CONCLUSION

This paper shows that adaptive sensing improves the convergence rate of the GROUSE online subspace estimation algorithm relative to using non-adaptive measurements, especially when very few measurements are available per time instant. However, the adaptive measurement model assumed in this work, which requires the ability to take arbitrary linear measurements, could be unrealistic or infeasible in certain applications. In future work we will extend the present analysis to the case of entrywise sampling of input vectors. More generally, we plan to extend our results to the constrained adaptive sensing setting [14], where the measurement vectors must come from a specified set. Finally, while this work focused on the GROUSE subspace estimation algorithm, the adaptive design proposed in this work might also enhance the performance of other recently proposed online subspace estimation and tracking algorithms [4]-[6], [18].

ACKNOWLEDGMENT

This work was supported by DARPA grant 16-43-D3M-FP-037 and NSF Graduate Research Fellowship DGE #1256260.

REFERENCES

- 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] 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.
- [4] J. He, L. Balzano, and J. Lui, "Online robust subspace tracking from partial information," arXiv preprint arXiv:1109.3827, 2011.
- [5] 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.
- [6] 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.
- [7] D. Zhang and L. Balzano, "Convergence of a grassmannian gradient descent algorithm for subspace estimation from undersampled data," *arXiv preprint arXiv:1610.00199*, 2016.
- [8] J. Haupt, R. M. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6222–6235, 2011.
- [9] M. A. Davenport and E. Arias-Castro, "Compressive binary search," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on.* IEEE, 2012, pp. 1827–1831.
- [10] E. J. Candes and M. A. Davenport, "How well can we estimate a sparse vector?" *Applied and Computational Harmonic Analysis*, vol. 34, no. 2, pp. 317–323, 2013.
- [11] E. Arias-Castro, E. J. Candes, and M. A. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 472–481, 2013.
- [12] M. L. Malloy and R. D. Nowak, "Near-optimal adaptive compressed sensing," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4001–4012, 2014.
- [13] D. Wei and A. O. Hero, "Performance guarantees for adaptive estimation of sparse signals," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 2043–2059, 2015.
- [14] M. A. Davenport, A. K. Massimino, D. Needell, and T. Woolf, "Constrained adaptive sensing," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5437–5449, 2016.
- [15] 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.
- [16] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Completing any low-rank matrix, provably." *Journal of Machine Learning Research*, vol. 16, pp. 2999–3034, 2015.
- [17] A. Eftekhari, M. B. Wakin, and R. A. Ward, "MC²: A twophase algorithm for leveraged matrix completion," *arXiv preprint arXiv:1609.01795*, 2016.
- [18] 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.