Subspace Estimation from Incomplete Observations: A High-Dimensional Analysis

Chuang Wang, Yonina C. Eldar, *Fellow, IEEE* and Yue M. Lu, *Senior Member, IEEE*

***Abstract*—We present a high-dimensional analysis of three popular algorithms, namely, Oja's method, GROUSE and PETRELS, for subspace estimation from streaming and highly incomplete observations. We show that, with proper time scaling, the time-varying principal angles between the true subspace and its estimates given by the algorithms converge weakly to deterministic processes when the ambient dimension n tends to infinity. Moreover, the limiting processes can be exactly characterized as the unique solutions of certain ordinary differential equations (ODEs). A finite sample bound is also given, showing that the rate of convergence towards such limits is $\mathcal{O}(1/\sqrt{n})$. In addition to providing asymptotically exact predictions of the dynamic performance of the algorithms, our high-dimensional analysis yields several insights, including an asymptotic equivalence between Oja's method and GROUSE, and a precise scaling relationship linking the amount of missing data to the signal-to-noise ratio. By analyzing the solutions of the limiting ODEs, we also establish phase transition phenomena associated with the steady-state performance of these techniques.**

***Index Terms*—Subspace tracking, streaming PCA, incomplete data, high-dimensional analysis, scaling limit**

I. Introduction

Subspace estimation is a key task in many signal processing applications. Examples include source localization in array processing, system identification, network monitoring, and image sequence analysis, to name a few. The ubiquity of subspace estimation comes from the fact that a low-rank subspace model can conveniently capture the intrinsic, low-dimensional structures of many large datasets.

In this paper, we consider the problem of estimating and tracking an unknown subspace from *streaming measurements* with many *missing entries*. The streaming setting appears in applications (*e.g.* video surveillances) where high-dimensional data arrive sequentially over time at high rates. It is especially relevant in dynamic scenarios where the underlying subspace to be estimated can be time-varying. Missing data is also a very common issue in practice. Incomplete observations may result from a variety of reasons, such as the limitations of the sensing mechanisms, constraints on power consumption or communication bandwidth, or a deliberate design feature that protects the privacy of individuals by removing partial records.

GROUSE [1] and PETRELS [2] as well as the classical Oja's method [3] are three popular algorithms for solving the above estimation problem. They are all streaming algorithms in the sense that they provide instantaneous, *on-the-fly* updates to their subspace estimates upon the arrival of a data point. The three differ in their update rules: Oja's method and GROUSE perform first-order incremental gradient descent on the Euclidean space and the Grassmannian, respectively, whereas PETRELS can be interpreted as a second-order stochastic gradient descent scheme. These algorithms have been shown to be highly effective in practice, but their performance depends on the careful choice of algorithmic parameters such as the step size (for GROUSE and Oja's method) and the discount parameter (for PETRELS). Various convergence properties of these techniques have been studied in [2], [4]–[7], but a precise analysis of their performance is still an open problem. Moreover, the important question of how the signal-to-noise ratios (SNRs), the amount of missing data, and various other algorithmic parameters affect the estimation performance is not fully understood.

As the main objective of this work, we present a tractable and asymptotically exact analysis of the dynamic performance of Oja's method, GROUSE and PETRELS in the high-dimensional regime. Our contribution is mainly threefold:

1. *Precise analysis via scaling limits.* We show in Theorem 1 and Theorem 2 that the time-varying trajectories of the estimation errors, measured in terms of the principal angles between the true underlying subspace and the estimates given by the algorithms, converge weakly to deterministic processes, as the ambient dimension $n \to \infty$. Moreover, such deterministic limits can be characterized as the unique solutions of certain ordinary differential equations (ODEs). In addition, we provide a finite-size guarantee in Theorem 3, showing that the convergence rate towards such limits is $\mathcal{O}(1/\sqrt{n})$. Numerical simulations verify the accuracy of our asymptotic predictions. The main technical tool behind our analysis is the weak convergence theory of stochastic processes (see [8]–[12] for its mathematical foundations and [13]–[15] for its recent applications in related estimation problems).

2. *Insights regarding the algorithms.* In addition to providing asymptotically exact predictions of the dynamic performance of the three subspace estimation algorithms, our high-

C. Wang is with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA (e-mail: chuangwang@g.harvard.edu).

Y. C. Eldar is with the Department of EE, Technion, Israel Institute of Technology, Haifa, 32000, Israel (e-mail: yonina@ee.technion.ac.il).

Y. M. Lu is with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA (e-mail: yuelu@seas.harvard.edu).

The work of C. Wang and Y. M. Lu was supported in part by the US Army Research Office under contract W911NF-16-1- 0265 and in part by the US National Science Foundation under grants CCF-1319140 and CCF-1718698. The work of Y. Eldar was supported in part by the European Union's Horizon 2020 Research and Innovation Program under Grant 646804-ERCCOG-BNYQ. Preliminary results of this work was presented at the Signal Processing with Adaptive Sparse Structured Representations (SPARS) workshop in 2017.

dimensional analysis leads to several valuable insights. First, the result of Theorem 1 implies that, despite their different update rules, Oja's methods and GROUSE are asymptotically equivalent, with both converging to the *same* deterministic process as the dimension increases. Second, the characterization given in Theorem 2 shows that PETRELS can be examined within a common framework that incorporates all three algorithms, with the difference being that PETRELS uses an adaptive scheme to adjust its effective step sizes. Third, our limiting ODEs also reveal an (asymptotically) exact scaling relationship that links the amount of missing data to the SNR. See the discussions in Section IV-A for details.

3. *Fundamental limits and phase transitions.* Analyzing the limiting ODEs also reveals phase transition phenomena associated with the steady-state performance of these algorithms. Specifically, we provide in Propositions 1 and 2 critical thresholds for setting key algorithm parameters (as a function of the SNR and the subsampling ratio), beyond which the algorithms converge to "noninformative" estimates that are no better than mere random guesses.

The rest of the paper is organized as follows. We start by presenting in Section II-A the exact problem formulation for subspace estimation with missing data. This is followed by a brief review of the three algorithms to be analyzed in this work. The main results are presented in Section III, where we show that the dynamic performance of Oja's method, GROUSE and PETRELS can be asymptotically characterized by the solutions of certain deterministic systems of ODEs. Numerical experiments are also provided to illustrate and verify our theoretical predictions. To place our asymptotic analysis in proper context, we discuss related work in the literature in Section III-D. We consider various implications and insights drawn from our analysis in Section IV. Due to space limitation, we only present informal derivation of the limiting ODEs and proof sketches in Section V. More technical details and the proofs of all the results presented in this paper can be found in the Supplementary Materials [16].

Notation: Throughout the paper, we use $\boldsymbol{I}_d$ to denote the $d \times d$ identity matrix. For any positive semidefinite matrix $\boldsymbol{M}$, its principal squared root is written as $(\boldsymbol{M})^{-\frac{1}{2}}$. Depending on the context, $\|\cdot\|$ denotes either the ℓ_2 norm of a vector or the spectral norm of a matrix. For any $x \in \mathbb{R}$, the floor operation $\lfloor x \rfloor$ gives the largest integer that is smaller than or equal to x. Let $\{X_n\}$ be a sequence of random variables in a general probability space. $X_n \xrightarrow{\mathcal{P}} X$ means that X_n converges in probability to a random variable X, whereas $X_n \xrightarrow{\text{weakly}} X$ means that X_n converges to X weakly (*i.e.* in law). Finally, $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function for an event $\mathcal{A}$.

II. Problem Formulation and Overview of Algorithms

A. Observation Model

We consider the problem of estimating a low-rank subspace using partial observations from a data stream. At any discrete-time k, suppose that a sample vector $\boldsymbol{s}_k \in \mathbb{R}^n$ is generated according to

$$\boldsymbol{s}_k = \boldsymbol{U}\boldsymbol{c}_k + \boldsymbol{a}_k. \tag{1}$$

Here, $\boldsymbol{U} \in \mathbb{R}^{n \times d}$ is an unknown deterministic matrix whose columns form an orthonormal basis of a d-dimensional subspace, and $\boldsymbol{c}_k \in \mathbb{R}^d$ is a random vector representing the expansion coefficients in that subspace. We also assume[1] that the covariance matrix of $\boldsymbol{c}_k$ is

$$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d), \tag{2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are some strictly positive numbers. The noise in the observations is modeled by a random vector $\boldsymbol{a}_k \in \mathbb{R}^n$ with zero mean and a covariance matrix equal to $\boldsymbol{I}_n$. Furthermore, $\boldsymbol{a}_k$ is independent of $\boldsymbol{c}_k$. Since $\{\lambda_\ell\}_{1 \leq \ell \leq d}$ in (2) indicate the "strength" of the subspace component relative to the noise, we shall refer to these parameters as the SNR in our subsequent discussions.

We consider the missing data case, where only a subset of the entries of $\boldsymbol{s}_k$ is available. This observation process can be modeled by a diagonal matrix

$$\boldsymbol{\Omega}_k = \operatorname{diag}(v_{k,1}, v_{k,2}, \ldots, v_{k,n}), \tag{3}$$

where $v_{k,i} = 1$ if the ith component of $\boldsymbol{x}_k$ is observed, and $v_{k,i} = 0$ otherwise. Our actual observation, denoted by $\boldsymbol{y}_k$, may then be written as

$$\boldsymbol{y}_k = \boldsymbol{\Omega}_k \boldsymbol{s}_k. \tag{4}$$

Given a sequence of incomplete observations $\{\boldsymbol{y}_k, \boldsymbol{\Omega}_k\}_{k \geq 0}$ arriving in a stream, we aim to estimate the subspace spanned by the columns of $\boldsymbol{U}$.

B. Oja's Method

Oja's method [3] is a classical algorithm for estimating low-rank subspaces from streaming samples. It was originally designed for the case where the full sample vectors $\boldsymbol{s}_k$ in (1) are available. Given a collection of K such sample vectors, it is natural to use the following optimization formulation to estimate the unknown subspace:

$$\widehat{\boldsymbol{U}} = \operatorname*{arg\,min}_{\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d} \sum_{k=1}^{K} \min_{\boldsymbol{w}_k} \|\boldsymbol{s}_k - \boldsymbol{X}\boldsymbol{w}_k\|^2 \tag{5}$$

$$= \operatorname*{arg\,max}_{\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d} \sum_{k=1}^{K} \boldsymbol{X}^\top \boldsymbol{s}_k \boldsymbol{s}_k^\top \boldsymbol{X}, \tag{6}$$

where the equivalence between (5) and (6) is established by solving the simple quadratic problem $\min_{\boldsymbol{w}_k} \|\boldsymbol{s}_k - \boldsymbol{X}\boldsymbol{w}_k\|^2$ and substituting the solution into (5).

Oja's method is a stochastic projected-gradient algorithm for solving (6). At each step k, let $\boldsymbol{X}_k$ denote the current estimate of the subspace. Then, with the arrival of a new sample vector $\boldsymbol{s}_k$, we first update $\boldsymbol{X}_k$ according to

$$\widetilde{\boldsymbol{X}}_k = \boldsymbol{X}_k + \frac{\tau_k}{n} \boldsymbol{s}_k \boldsymbol{w}_k^\top, \tag{7}$$

[1]The assumption that the covariance matrix is diagonal can be made without loss of generality, after a rotation of the coordinate system. To see that, suppose $\boldsymbol{c}_k$ has a general covariance matrix $\boldsymbol{\Sigma}$, which is diagonalized as $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^\top$. Here, $\boldsymbol{\Phi}$ is an orthonormal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix as in (2). The generating model (1) can then be rewritten as $\boldsymbol{s}_k = (\boldsymbol{U}\boldsymbol{\Phi})(\boldsymbol{\Phi}^\top \boldsymbol{c}_k) + \boldsymbol{a}_k$. Thus, our problem is equivalent to estimating a subspace spanned by $\boldsymbol{U}\boldsymbol{\Phi}$, and $\boldsymbol{\Lambda}$ is the covariance matrix of the new expansion coefficient vector $\boldsymbol{\Phi}^\top \boldsymbol{c}_k$.

where $\boldsymbol{w}_k = \boldsymbol{X}_k^\top \boldsymbol{s}_k$ and $\{\tau_k\}$ is a sequence of positive constants that control the step-size (or learning rate) of the algorithm. We note that, up to a scaling constant, $\boldsymbol{s}_k \boldsymbol{w}_k^T$ in (7) is exactly equal to the gradient of the objective function $\boldsymbol{X}^\top \boldsymbol{s}_k \boldsymbol{s}_k^\top \boldsymbol{X}$ in (6) due to the new sample $\boldsymbol{s}_k$. Next, to enforce the orthogonality constraint, we compute

$$\boldsymbol{X}_{k+1} = \widetilde{\boldsymbol{X}}_k (\widetilde{\boldsymbol{X}}_k^\top \widetilde{\boldsymbol{X}}_k)^{-\frac{1}{2}}, \tag{8}$$

where $(\cdot)^{-\frac{1}{2}}$ stands for the principal square root of a positive semidefinite matrix. In practice, (8) is implemented using the QR-decomposition of $\widetilde{\boldsymbol{X}}_k$.

To handle the case of partially-observed samples, we can modify Oja's method in two ways. First, we estimate the expansion coefficients $\boldsymbol{w}_k$ in (7) by solving a least squares problem that takes into account the missing data model:

$$\widehat{\boldsymbol{w}}_k = \underset{\boldsymbol{w} \in \mathbb{R}^d}{\arg\min} \left\| \boldsymbol{y}_k - \boldsymbol{\Omega}_k \boldsymbol{X}_k \boldsymbol{w} \right\|^2, \tag{9}$$

where $\boldsymbol{y}_k$ is the incomplete sample vector defined in (4), $\boldsymbol{\Omega}_k$ is the corresponding subsampling matrix, and $\boldsymbol{X}_k$ is the current estimate of the subspace. Next, we replace the missing elements in $\boldsymbol{y}_k$ by the corresponding entries in $\boldsymbol{X}_k \widehat{\boldsymbol{w}}_k$. This *imputation* step leads to an estimate of the full vector:

$$\widehat{\boldsymbol{y}}_k = \boldsymbol{y}_k + (\boldsymbol{I}_n - \boldsymbol{\Omega}_k) \boldsymbol{X}_k \widehat{\boldsymbol{w}}_k. \tag{10}$$

Replacing the original vectors $\boldsymbol{y}_k$ and $\boldsymbol{w}_k$ in (7) by their estimated counterparts $\widehat{\boldsymbol{y}}_k$ and $\widehat{\boldsymbol{w}}_k$, we reach the modified Oja's method, the pseudocode of which is summarized in Algorithm 1. Note that, to ensure we have enough observed entries in $\boldsymbol{y}_k$, we first check, with the arrival of a new partially observed vector $\boldsymbol{y}_k$, whether

$$\det(\boldsymbol{X}_k^\top \boldsymbol{\Omega}_k \boldsymbol{X}_k) > \epsilon \det(\boldsymbol{X}_k^\top \boldsymbol{X}_k), \tag{11}$$

where $\epsilon > 0$ is a small positive constant. If this is indeed the case, we do the standard update as described above; otherwise, we simply ignore the new sample vector and do not change the estimate in this step. Note that, under a suitable probabilistic model for the subsampling process (see assumption (A.3) in Section III-C), one can show that (11) is satisfied with high probability as long as $\epsilon < \alpha$, where α denotes the subsampling ratio defined in assumption (A.3).

Algorithm 1 Oja's method [3] with imputation

Require: An initial estimate $\boldsymbol{X}_0$ such that $\boldsymbol{X}_0^\top \boldsymbol{X}_0 = \boldsymbol{I}_d$, a sequence of step-size parameters $\{\tau_k\}$ and a positive constant ϵ.

$k \coloneqq 0$
repeat
  if $\det(\boldsymbol{X}_k^\top \boldsymbol{\Omega}_k \boldsymbol{X}_k) > \epsilon \det(\boldsymbol{X}_k^\top \boldsymbol{X}_k)$ **then**
    $\widehat{\boldsymbol{w}}_k \coloneqq \arg\min_{\boldsymbol{w}} \|\boldsymbol{y}_k - \boldsymbol{\Omega}_k \boldsymbol{X}_k \boldsymbol{w}\|$
    $\widehat{\boldsymbol{y}}_k \coloneqq \boldsymbol{y}_k + (\boldsymbol{I}_n - \boldsymbol{\Omega}_k) \boldsymbol{X}_k \widehat{\boldsymbol{w}}_k$
    $\widetilde{\boldsymbol{X}}_k \coloneqq \boldsymbol{X}_k + \frac{\tau_k}{n} \widehat{\boldsymbol{y}}_k \widehat{\boldsymbol{w}}_k^\top$
    $\boldsymbol{X}_{k+1} \coloneqq \widetilde{\boldsymbol{X}}_k (\widetilde{\boldsymbol{X}}_k^\top \widetilde{\boldsymbol{X}}_k)^{-\frac{1}{2}}$
  else
    $\boldsymbol{X}_{k+1} \coloneqq \boldsymbol{X}_k$
  end if
  $k \coloneqq k + 1$
until termination

Algorithm 2 GROUSE [1]

Require: An initial estimate $\boldsymbol{X}_0$ such that $\boldsymbol{X}_0^\top \boldsymbol{X}_0 = \boldsymbol{I}_d$, a sequence of step-size parameters $\{\tau_k\}$ and a positive constant ϵ.

$k \coloneqq 0$
repeat
  if $\det(\boldsymbol{X}_k^\top \boldsymbol{\Omega}_k \boldsymbol{X}_k) > \epsilon \det(\boldsymbol{X}_k^\top \boldsymbol{X}_k)$ **then**
    $\widehat{\boldsymbol{w}}_k \coloneqq \arg\min_{\boldsymbol{w}} \|\boldsymbol{y}_k - \boldsymbol{\Omega}_k \boldsymbol{X}_k \boldsymbol{w}\|$
    $\boldsymbol{p}_k \coloneqq \boldsymbol{X}_k \widehat{\boldsymbol{w}}_k$
    $\boldsymbol{r}_k \coloneqq \boldsymbol{y}_k - \boldsymbol{\Omega}_k \boldsymbol{p}_k$
    $\theta_k \coloneqq \frac{\tau_k}{n} \|\boldsymbol{r}_k\| \|\boldsymbol{p}_k\|$
    $\boldsymbol{X}_{k+1} \coloneqq \boldsymbol{X}_k + \left[(\cos(\theta_k) - 1) \frac{\boldsymbol{p}_k}{\|\boldsymbol{p}_k\|} + \sin(\theta_k) \frac{\boldsymbol{r}_k}{\|\boldsymbol{r}_k\|} \right] \frac{\widehat{\boldsymbol{w}}_k^\top}{\|\widehat{\boldsymbol{w}}_k\|}$
  else
    $\boldsymbol{X}_{k+1} \coloneqq \boldsymbol{X}_k$
  end if
  $k \coloneqq k + 1$
until termination

C. GROUSE

Similar to Oja's method, Grassmannian Rank-One Update Subspace Estimation (GROUSE) [1] is a first-order stochastic gradient descent algorithm for solving (5). The main difference is that GROUSE solves the optimization problem on the Grassmannian, the manifold of all subspaces with a fixed rank. One advantage of this approach is that it avoids the explicit orthogonalization step in (8), allowing the algorithm to achieve even lower computational complexity.

At each step, GROUSE first finds the coefficient $\widehat{\boldsymbol{w}}_k$ according to (9). It then computes the reconstruction error vector

$$\boldsymbol{r}_k = \boldsymbol{y}_k - \boldsymbol{\Omega}_k \boldsymbol{p}_k, \tag{12}$$

where $\boldsymbol{p}_k \overset{\text{def}}{=} \boldsymbol{X}_k \widehat{\boldsymbol{w}}_k$. Next, it updates the current estimate $\boldsymbol{X}_k$ on the Grassmannian as

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \left[(\cos\theta_k - 1) \frac{\boldsymbol{p}_k}{\|\boldsymbol{p}_k\|} + \sin\theta_k \frac{\boldsymbol{r}_k}{\|\boldsymbol{r}_k\|} \right] \frac{\boldsymbol{w}_k^\top}{\|\boldsymbol{w}_k\|},$$

where

$$\theta_k = \frac{\tau_k}{n} \|\boldsymbol{r}_k\| \cdot \|\boldsymbol{p}_k\|, \tag{13}$$

and $\{\tau_k\}_k$ is a sequence of step-size parameters. The algorithm is summarized in Algorithm 2.

D. PETRELS

When there is no missing data, an alternative to Oja's method is a classical algorithm called Projection Approximation Subspace Tracking (PAST) [17]. This method estimates

Algorithm 3 Simplified PETRELS [2]

Require: An initial estimate of the subspace $\boldsymbol{X}_0$, $\boldsymbol{R}_0 = \frac{\delta}{n}\boldsymbol{I}_d$ for some $\delta > 0$, and positive constants γ and ϵ.

1: $k \coloneqq 0$
2: **repeat**
3: **if** $\det(\boldsymbol{X}_k^\top \boldsymbol{\Omega}_k \boldsymbol{X}_k) > \epsilon \det(\boldsymbol{X}_k^\top \boldsymbol{X}_k)$ **then**
4: $\widehat{\boldsymbol{w}}_k \coloneqq \arg\min_{\boldsymbol{w}} \|\boldsymbol{y}_k - \boldsymbol{\Omega}_k \boldsymbol{X}_k \boldsymbol{w}\|$
5: $\boldsymbol{X}_{k+1} \coloneqq \boldsymbol{X}_k + \boldsymbol{\Omega}_k(\boldsymbol{y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{w}}_k)\widehat{\boldsymbol{w}}_k^\top \boldsymbol{R}_k$
6: $\boldsymbol{v}_k \coloneqq \gamma^{-1}\boldsymbol{R}_k \widehat{\boldsymbol{w}}_k$
7: $\beta_k \coloneqq 1 + \alpha\, \widehat{\boldsymbol{w}}_k^\top \boldsymbol{v}_k$
8: $\boldsymbol{R}_{k+1} \coloneqq \gamma^{-1}\boldsymbol{R}_k - \alpha\, \boldsymbol{v}_k \boldsymbol{v}_k^\top / \beta_k$
9: **else**
10: $\boldsymbol{X}_{k+1} \coloneqq \boldsymbol{X}_k$
11: $\boldsymbol{R}_{k+1} \coloneqq \boldsymbol{R}_k$
12: **end if**
13: $k \coloneqq k + 1$
14: **until** termination

the underlying subspace $\boldsymbol{U}$ by solving an exponentially-weighted least-squares problem

$$\boldsymbol{X}_{k+1} = \underset{\boldsymbol{X} \in \mathbb{R}^{n \times d}}{\arg\min} \sum_{k'=1}^{k} \gamma^{k-k'} \|\boldsymbol{s}_{k'} - \boldsymbol{X}\boldsymbol{w}_{k'}\|^2, \tag{14}$$

where $\boldsymbol{w}_{k'} = \boldsymbol{X}_{k'}^T \boldsymbol{s}_{k'}$ and $\gamma \in (0, 1]$ is a discount parameter. The solution of (14) has a simple recursive update rule

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + (\boldsymbol{s}_k - \boldsymbol{X}_k \boldsymbol{w}_k)\, \boldsymbol{w}_k^\top \boldsymbol{R}_k \tag{15}$$

$$\boldsymbol{R}_{k+1} = (\gamma \boldsymbol{R}_k^{-1} + \boldsymbol{w}_k \boldsymbol{w}_k^\top)^{-1}. \tag{16}$$

Moreover, one can avoid the explicit calculation of the matrix inverse in (16) by using the Woodbury identity and the fact that (16) amounts to a rank-1 update.

Parallel Subspace Estimation and Tracking by Recursive Least Squares (PETRELS) [2] extends PAST to the case of partially-observed data. The main change is that it estimates the coefficient $\boldsymbol{w}_k$ in (14) using (9). In addition, it provides a parallel sub-routine in its calculations so that updates to different coordinates can be computed in a fully parallel fashion. In its most general form, PETRELS needs to maintain and update a different $d \times d$ matrix $\boldsymbol{R}_k^i$ for each of the n coordinates. To reduce computational complexity, a simplified version of PETRELS has been provided in [2], using a common $\boldsymbol{R}_k$ for all the coordinates.

In this paper, we focus on this simplified version of PETRELS, which is summarized in Algorithm 3. Note that we introduce an additional parameter α in lines 7 and 8 of the pseudocode. The simplified algorithm given in [2] corresponds to setting $\alpha = 1$. In our analysis, we set α to be equal to the subsampling ratio defined later in (25). Empirically, we find that, with this modification, the performance of the simplified algorithm matches that of the full PETRELS algorithm when the ambient dimension n is large.

III. Main Results: Scaling Limits

In this section, we present the main results of this work—a tractable and asymptotically exact analysis of the performance of the three algorithms reviewed in Section II.

A. Performance Metric: Principal Angles

We start by defining the performance metric we will be using in our analysis. Recall the generative model defined in (1). The ground truth subspace is represented by the matrix $\boldsymbol{U}$, whose column vectors form an orthonormal basis of that subspace. For Algorithms 1, 2, and 3, the estimated subspace at the kth step is spanned by an orthogonal matrix

$$\widehat{\boldsymbol{U}}_k \overset{\text{def}}{=} \boldsymbol{X}_k (\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1/2}, \tag{17}$$

where $\boldsymbol{X}_k$ is the kth iterand generated by the algorithms. Note that, for Oja's method and GROUSE, $\widehat{\boldsymbol{U}}_k = \boldsymbol{X}_k$ as the matrix $\boldsymbol{X}_k$ is already orthogonal, whereas for PETRELS, generally $\boldsymbol{X}_k^\top \boldsymbol{X}_k \neq \boldsymbol{I}_d$ and thus the step in (17) is necessary.

In the special case of $d = 1$ (*i.e.* rank-one subspaces), both $\boldsymbol{U}$ and $\widehat{\boldsymbol{U}}_k$ are unit-norm vectors. The degree to which these vectors are aligned can be measured by their *cosine similarity*, defined as $\left|\boldsymbol{U}^\top \widehat{\boldsymbol{U}}_k\right|$. This concept can be naturally extended to arbitrary $d \geq 1$. In general, the closeness of two d-dimensional subspaces may be quantified by their d principal angles [18], [19]. In particular, the cosines of the principal angles are uniquely specified as the singular values of a $d \times d$ matrix defined as

$$\boldsymbol{Q}_k^{(n)} \overset{\text{def}}{=} \boldsymbol{U}^\top \widehat{\boldsymbol{U}}_k = \boldsymbol{U}^\top \boldsymbol{X}_k (\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1/2}. \tag{18}$$

In what follows, we shall refer to $\boldsymbol{Q}_k^{(n)}$ as the *cosine similarity matrix*. Since we will be studying the high-dimensional limit of $\boldsymbol{Q}_k^{(n)}$ as the ambient dimension $n \to \infty$, we use the superscript (n) to make the dependence of $\boldsymbol{Q}_k^{(n)}$ on n explicit.

B. The Scaling Limits of Stochastic Processes: Main Ideas

To analyze the performance of Algorithms 1, 2, and 3, our goal boils down to tracking the evolution of the cosine similarity matrix $\boldsymbol{Q}_k^{(n)}$ over time. Thanks to the streaming nature of all three methods, it is easy to see that the dynamics of their estimates $\boldsymbol{X}_k$ can be modeled by homogeneous Markov chains with state space in $\mathbb{R}^{n \times d}$. Thus, being a function of $\boldsymbol{X}_k$ [see (18)], the dynamics of $\boldsymbol{Q}_k^{(n)}$ forms a hidden Markov chain. We then show that, as $n \to \infty$ and with proper time scaling, the family of stochastic processes $\left\{\boldsymbol{Q}_k^{(n)}\right\}_n$ indexed by n converges *weakly* to a deterministic function of time that is characterized as the unique solution of some ODEs. Such convergence is known in the literature as the *scaling limits* [10], [12], [15], [20] of stochastic processes. To present out results, we first consider a simple one-dimensional example that illustrates the underlying ideas behind scaling limits. Our main convergence theorems are presented in Section III-C.

Consider a 1-D stochastic process defined by a recursion

$$q_{k+1} = q_k + \tfrac{\tau}{n} f(q_k) + \tfrac{1}{n^{(1/2)+\delta}} v_k, \tag{19}$$

where $f(\cdot)$ is a Lipschitz function, τ and δ are two positive constants, v_k is a sequence of i.i.d. random variables with zero mean and unit variance, and $n > 0$ is a constant introduced to scale the step-size and the noise variance. (This particular scaling is chosen here because it mimics the actual scaling that appears in the high-dimensional dynamics of $\boldsymbol{Q}_k^{(n)}$ we shall study.)

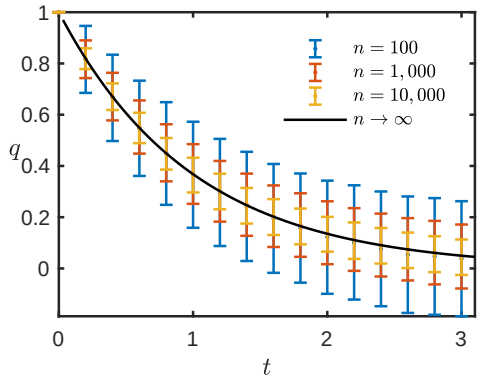


Figure 1. Convergence of the 1-D stochastic process $q_t^{(n)}$ described in Example 1 to its deterministic scaling limit. Here, we use $\delta = 0.25$.

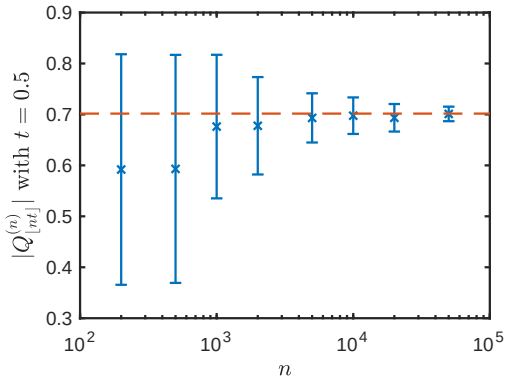


Figure 2. Convergence of the cosine similarity $\boldsymbol{Q}_{\lfloor nt \rfloor}^{(n)}$ associated with GROUSE at a fixed rescaled time $t = 0.5$, as we increase n from 200 to $50,000$. In this experiment, $d = 1$ and thus $\boldsymbol{Q}_k^{(n)}$ reduces to a scalar, denoted by $Q_k^{(n)}$. The error bars show the standard deviation of $Q_{\lfloor nt \rfloor}^{(n)}$ over 100 independent trials. In each trial, we randomly generate a subspace $\boldsymbol{U}$, the expansion coefficients $\{\boldsymbol{c}_k\}$ and the noise vector $\{\boldsymbol{a}_k\}$ as in (1). The red dashed lines is the limiting value predicted by our asymptotic characterization, to be given in Theorem 1.

When n is large, the difference between q_k and q_{k+1} is small. In other words, we will not be able to see *macroscopic* changes unless we observe the process over a large number of steps. To accelerate the time (by a factor of n), we embed $\{q_k\}$ in continuous-time by defining a piecewise-constant process

$$q^{(n)}(t) \stackrel{\text{def}}{=} q_{\lfloor nt \rfloor}, \tag{20}$$

where $\lfloor \cdot \rfloor$ is the floor function. Here, t is the rescaled (accelerated) time: within $t \in [0, 1]$, the original discrete-time process moves n steps. Due to the scaling of the noise term in (19) (with the noise variance equal to $n^{-1-2\delta}$ for some $\delta > 0$), the rescaled stochastic process $q^{(n)}(t)$ converges to a deterministic limit function as $n \to \infty$. We illustrate this convergence behavior using the following example.

Example 1: Let us consider the special case where $f(q) = -q$. We plot in Figure 1 simulations results of $q^{(n)}(t)$ for several different values of n. We see that, as n increases, the rescaled stochastic processes $q^{(n)}(t)$ indeed converge to some limit function (the black line in the figure), which will be denoted by $q(t)$. To prove this convergence, we first expand the recursion (19) (by using the fact that $f(q) = -q$) and get

$$q_k = (1 - \tfrac{\tau}{n})^k q_0 + \Delta_k, \tag{21}$$

where Δ_k is a zero-mean random variable defined as

$$\Delta_k = \frac{1}{n^{(1/2)+\delta}} \sum_{i=0}^{k-1} (1 - \tfrac{\tau}{n})^{k-1-i} v_i.$$

Since $\{v_i\}_i$ are independent random variables with unit variance, $\mathbb{E}\,(\Delta_k)^2 \le \frac{1}{n^{1+2\delta}}(1 - (1 - \frac{\tau}{n})^2)^{-1} = \mathcal{O}(n^{-2\delta})$. It then follows from (21) that, for any $t > 0$,

$$q^{(n)}(t) = q_{\lfloor nt \rfloor} \xrightarrow{\mathcal{P}} \lim_{n\to\infty} (1 - \tfrac{\tau}{n})^{\lfloor nt \rfloor} q_0 = q_0 e^{-\tau t}, \tag{22}$$

where $\xrightarrow{\mathcal{P}}$ stands for convergence in probability.

For general nonlinear function $f(q)$, we can no longer directly simplify the recursion (19) as in (21). However, similar convergence behaviors of $q^{(n)}(t)$ still exist. Moreover, the limit function $q(t)$ can be characterized via an ODE. To see the origin of the ODE, we note that, for any $t > 0$ and $k = \lfloor nt \rfloor$, we may rewrite (19) as

$$\frac{q^{(n)}(t + 1/n) - q^{(n)}(t)}{1/n} = \tau f[q^{(n)}(t)] + n^{(1/2)-\delta} v_k. \tag{23}$$

Taking the $n \to \infty$ limit on both side of (23) and neglecting the noise term $n^{(1/2)-\delta} v_k$, we may then write—at least in a nonrigorous way—the following ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} q(t) = \tau f[q(t)],$$

which always has a unique solution due to the Lipschitz property of $f(\cdot)$. For instance, the ODE associated with the linear setting in Example 1 is $\frac{d}{dt} q(t) = -\tau q(t)$, whose unique solution $q(t) = q_0 e^{-\tau t}$ is indeed the limit established in (22). A rigorous justification of the above steps can be found in the theory of weak convergence of stochastic processes (see, for example, [12], [20]).

Returning from the above detour, we recall that the central objects of our analysis are the cosine similarity matrices $\boldsymbol{Q}_k^{(n)}$ defined in (18). It turns out that, just like the simple 1-D process q_k in (19), the matrix-valued stochastic processes $\boldsymbol{Q}_k^{(n)}$, after a proper time rescaling $k = \lfloor nt \rfloor$, also converge to a deterministic limit as the ambient dimension $n \to \infty$. This phenomenon is demonstrated in Figure 2, where we plot the cosine similarity $\boldsymbol{Q}_{\lfloor nt \rfloor}^{(n)}$ of GROUSE at $t = 0.5$ for different values of n. In this experiment, we set $d = 1$ and thus $\boldsymbol{Q}_{\lfloor nt \rfloor}^{(n)}$ reduces to a scalar. The standard deviations of $\boldsymbol{Q}_{\lfloor nt \rfloor}^{(n)}$ over 100 independent trials, shown as error bars in Figure 2, decrease as n increases. This indicates that the performance of these stochastic algorithms can indeed be characterized by certain deterministic limits when the dimension is high.

C. The Scaling Limits of Oja's, GROUSE and PETRELS

To study the scaling limits of the cosine similarity matrices $\boldsymbol{Q}_{\lfloor nt \rfloor}^{(n)}$, we embed the discrete-time process $\boldsymbol{Q}_k^{(n)}$ into a continuous time process $\boldsymbol{Q}^{(n)}(t)$ via a simple piecewise-constant

interpolation:

$$\boldsymbol{Q}^{(n)}(t) \stackrel{\text{def}}{=} \boldsymbol{Q}^{(n)}_{\lfloor nt \rfloor}. \tag{24}$$

The main objective of this work is to establish the high-dimensional limit of $\boldsymbol{Q}^{(n)}(t)$ as $n \to \infty$. Our asymptotic analysis is carried out under the following technical assumptions on the generative model (1) and the observation model (3).

(A.1) The elements of the noise vector $\boldsymbol{a}_k$ are i.i.d. random variables with zero mean, unit variance, and finite higher-order moments;

(A.2) $\boldsymbol{c}_k$ in (1) is a d-D random vector with zero-mean and a covariance matrix $\boldsymbol{\Lambda}$ as given in (2). Moreover, all the higher-order moments of $\boldsymbol{c}_k$ exist and are finite, and $\{\boldsymbol{c}_k\}$ is independent of $\{\boldsymbol{a}_k\}$;

(A.3) We assume that $\{v_{k,i}\}$ in the observation model (3) is a collection of independent and identically distributed binary random variables such that

$$\mathbb{P}(v_{k,i} = 1) = \alpha, \tag{25}$$

for some constant $\alpha \in (0, 1)$. Throughout the paper, we refer to α as the *subsampling ratio*. We shall also assume that the algorithmic parameter ϵ used in Algorithms 1–3 satisfies the condition that $\epsilon < \alpha$.

(A.4) The subspace matrix $\boldsymbol{U}$ and initial guess $\boldsymbol{X}_0$ are *incoherent* in the sense that

$$\sum_{i=1}^{n}\sum_{j=1}^{d} U_{i,j}^4 \le \frac{C}{n} \text{ and } \sum_{i=1}^{n}\sum_{j=1}^{d} X_{0,i,j}^4 \le \frac{C}{n}, \tag{26}$$

where $U_{i,j}$ and $X_{0,i,j}$ denote the (i,j)th entries of $\boldsymbol{U}$ and $\boldsymbol{X}_0$, respectively, and C is a generic constant that does not depend on n.

(A.5) The initial cosine similarity $\boldsymbol{Q}_0^{(n)}$ converges entrywise and in probability to a deterministic matrix $\boldsymbol{Q}(0)$.

(A.6) For Oja's method and GROUSE, the step-size parameters $\tau_k = \tau(k/n)$, where $\tau(\cdot)$ is a deterministic function such that $\sup_{t\ge 0}|\tau(t)| \le C$ for a generic constant C that does not depend on n. For PETRELS, the discount factor

$$\gamma = 1 - \tfrac{\mu}{n}, \tag{27}$$

for some constant $\mu > 0$.

Assumption (A.4) requires some further explanations. The condition (26) essentially requires the basis matrix $\boldsymbol{U}$ and the initial guess $\boldsymbol{X}_0$ to be generic. To see this, consider a $\boldsymbol{U}$ that is drawn uniformly at random from the Grassmannian for rank-d subspaces. Such a $\boldsymbol{U}$ can be generated as

$$\boldsymbol{U} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}, \tag{28}$$

where $\boldsymbol{X}$ is an $n \times d$ random matrix whose entries are i.i.d. standard normal random variables. For such a generic choice of $\boldsymbol{U}$, one can show that its entries $U_{i,j} \sim \mathcal{O}(1/\sqrt{n})$ and that (26) holds with high probability when n is large.

Theorem 1 (Oja's method and GROUSE): Fix $T > 0$, and let $\{\boldsymbol{Q}^{(n)}(t)\}_{t\in[0,T]}$ be the time-varying cosine similarity matrices associated with either Oja's method or GROUSE over the finite interval $t \in [0, T]$. Under assumptions (A.1)–(A.6), we have

$$\{\boldsymbol{Q}^{(n)}(t)\}_{t\in[0,T]} \stackrel{\text{weakly}}{\longrightarrow} \boldsymbol{Q}(t),$$

where $\stackrel{\text{weakly}}{\longrightarrow}$ stands for weak convergence and $\boldsymbol{Q}(t)$ is a deterministic matrix-valued process. Moreover, $\boldsymbol{Q}(t)$ is the unique solution of the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{Q}(t) = F(\boldsymbol{Q}(t), \tau(t)\boldsymbol{I}_d), \tag{29}$$

where $F : \boldsymbol{R}^{d\times d} \times \boldsymbol{R}^{d\times d} \mapsto \boldsymbol{R}^{d\times d}$ is a matrix-valued function defined as

$$F(\boldsymbol{Q}, \boldsymbol{G}) \stackrel{\text{def}}{=} \Big[\alpha\boldsymbol{\Lambda}^2\boldsymbol{Q} - \tfrac{1}{2}\boldsymbol{Q}\boldsymbol{G} - \boldsymbol{Q}\left(\boldsymbol{I}_d + \tfrac{1}{2}\boldsymbol{G}\right)\boldsymbol{Q}^\top \alpha\boldsymbol{\Lambda}^2\boldsymbol{Q}\Big]\boldsymbol{G}. \tag{30}$$

Here α is the subsampling ratio, and $\boldsymbol{\Lambda}$ is the diagonal covariance matrix defied in (2).

In Section V, we present a (nonrigorous) derivation of the limiting ODE (29). Full technical details and a complete proof can be found in the Supplementary Materials [16]. An interesting conclusion of this theorem is that the cosine similarity matrices $\boldsymbol{Q}^{(n)}(t)$ associated with Oja's method and GROUSE converge to the same asymptotic trajectory. We will elaborate on this point in Section IV-A.

To establish the scaling limits of PETRELS, we need to introduce an auxiliary $d \times d$ matrix

$$\boldsymbol{G}_k^{(n)} \stackrel{\text{def}}{=} (\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-\frac{1}{2}} \boldsymbol{R}_k (\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-\frac{1}{2}}, \tag{31}$$

where the matrices $\boldsymbol{R}_k$ and $\boldsymbol{X}_k$ are those used in Algorithm 3. Similar to (24), we embed the discrete-time process $\boldsymbol{G}_k^{(n)}$ into a continuous-time process:

$$\boldsymbol{G}^{(n)}(t) \stackrel{\text{def}}{=} \boldsymbol{G}^{(n)}_{\lfloor nt \rfloor}. \tag{32}$$

The following theorem, whose proof can be found in the Supplementary Materials [16], characterizes the asymptotic dynamics of PETRELS.

Theorem 2 (PETRELS): For any fixed $T > 0$, let $\{\boldsymbol{Q}^{(n)}(t)\}_{t\in[0,T]}$ be the time-varying cosine similarity matrices associated with PETRELS on the interval $t \in [0, T]$. Let $\{\boldsymbol{G}^{(n)}(t)\}_{t\in[0,T]}$ be the process defined in (32). Under assumptions (A.1) – (A.6) and as $n \to \infty$, we have

$$\{\boldsymbol{Q}^{(n)}(t)\}_{t\in[0,T]} \stackrel{\text{weakly}}{\longrightarrow} \boldsymbol{Q}(t) \text{ and } \{\boldsymbol{G}^{(n)}(t)\}_{t\in[0,T]} \stackrel{\text{weakly}}{\longrightarrow} \boldsymbol{G}(t),$$

where $\{\boldsymbol{Q}(t), \boldsymbol{G}(t)\}$ is the unique solution of the following system of coupled ODEs:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{Q}(t) = F(\boldsymbol{Q}(t), \boldsymbol{G}(t)), \tag{33}$$

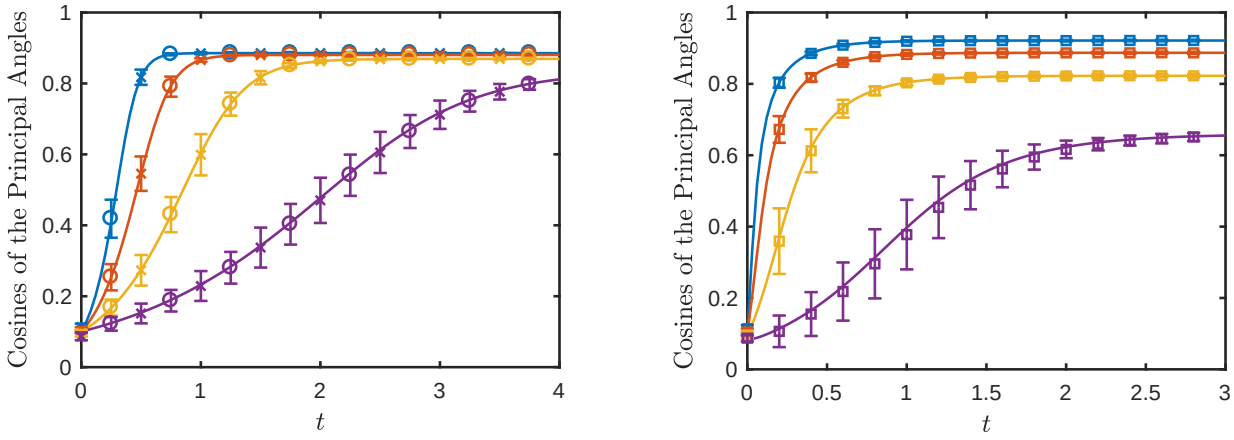
$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{G}(t) = H(\boldsymbol{Q}(t), \boldsymbol{G}(t)). \tag{34}$$

Here, F is the function defined in (30) and H is a function defined by

$$H(\boldsymbol{Q}, \boldsymbol{G}) \stackrel{\text{def}}{=} \boldsymbol{G}\left[\mu - \boldsymbol{G}(\boldsymbol{G} + \boldsymbol{I}_d)(\boldsymbol{Q}^\top \alpha\boldsymbol{\Lambda}^2\boldsymbol{Q} + \boldsymbol{I}_d)\right], \tag{35}$$

where $\mu > 0$ is the constant given in (27).

Theorem 1 and Theorem 2 establish the scaling limits of Oja's method, GROUSE and PETELS, respectively, as $n \to \infty$. In practice, the dimension n is always finite, and thus the actual trajectories of the performance curves will fluctuate around their asymptotic limits. To bound such fluctuations via a finite-sample analysis, we first need to slightly strengthen assumption (A.5) as follows:



(a) GROUSE (crosses) and Oja (circles)

(b) PETRELS

Figure 3. Numerical simulations vs. asymptotic characterizations. (a) Results for Oja's method and GROUSE, where the solid lines are the theoretical predictions of the cosines of 4 principal angles by the solution of the ODE (29). The crosses (for Oja's method) and circles (for GROUSE) show the simulation results averaged over 100 independent trials. In each trial, we randomly generate a subspace $\boldsymbol{U}$ as in (28), the expansion coefficients $\{\boldsymbol{c}_k\}$ and the noise vector $\{\boldsymbol{a}_k\}$. The error bars indicate ± 2 standard deviations. (b) Similar comparisons of numerical simulations and theoretical predictions for PETRELS.

(A.7) Let $\boldsymbol{Q}_0^{(n)}$ be the initial cosine similarity matrices. There exists a fixed matrix $\boldsymbol{Q}(0)$ such that

$$\mathbb{E}\left\|\boldsymbol{Q}_0^{(n)} - \boldsymbol{Q}(0)\right\|_2 \leq Cn^{-1/2},$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix and $C > 0$ is a constant that does not depend on n.

Theorem 3 (Finite Sample Analysis): Let $\boldsymbol{Q}^{(n)}(t)$ be the time-varying cosine similarity matrices associated with Oja's method, GROUSE, or PETELS, respectively. Let $\boldsymbol{Q}(t)$ denote the corresponding scaling limit given in (29), (33) or (34). Fix any $T > 0$. Under assumptions (A.1)–(A.4), (A.6)–(A.7), for any $t \in [0, T]$, we have

$$\sup_{n\geq 1} \mathbb{E}\left\|\boldsymbol{Q}^{(n)}(t) - \boldsymbol{Q}(t)\right\|_2 \leq \frac{C(T)}{\sqrt{n}}, \tag{36}$$

where $C(T)$ is a constant that can depend on the terminal time T but not on n.

The above theorem, whose proof can be found in the Supplementary Materials [16], shows that the rate of convergence towards the scaling limits is $\mathcal{O}(1/\sqrt{n})$.

Example 2: To demonstrate the accuracy of the asymptotic characterizations given in Theorem 1 and Theorem 2, we compare the actual performance of the algorithms against their theoretical predictions in Figure 3. In our experiments, we generate a random orthogonal matrix $\boldsymbol{U}$ according to (28) with $n = 20,000$ and $d = 4$. For Oja's method and GROUSE, we use a constant step size $\tau = 0.5$. For PETRELS, the discount factor is $\gamma = 1 - \mu/n$ with $\mu = 5$, and $\boldsymbol{R}_0 = \frac{\delta}{n}\boldsymbol{I}_d$ with $\delta = 10$. The covariance matrix is set to

$$\boldsymbol{\Lambda} = \mathrm{diag}\,\{5, 4, 3, 2\}$$

and the subsampling ratio is $\alpha = 0.5$. Figure 3(a) shows the evolutions of the cosines of the 4 principal angles between $\boldsymbol{U}$ and the estimates given by Oja's method (shown as crosses) and GROUSE (shown as circles). We compute the theoretical predictions of the principal angles by performing a SVD of the limiting matrices $\boldsymbol{Q}(t)$ as specified by the ODE (29). (In fact, this ODE has a simple analytical solution. See Section IV-B for details.) Figure 3(b) shows similar comparisons between PETRELS and its corresponding theoretical predictions. In this case, we solve the limiting ODEs (33) and (34) numerically.

D. Related Work

The problem of estimating and tracking low-rank subspaces has received a lot of attention recently in the signal processing and learning communities. Under the setting of fully observed data, an earlier work [21] studies a block-version of Oja's method and provides a sample complexity estimate for the case of $d = 1$. Similar analysis is available for general d-dimensional subspaces [22], [23]. The streaming version of Oja's method and its sample complexities have also been extensively studied. See, *e.g.*, [24]–[28].

For the case of incomplete observations, the sample complexity of a block version of Oja's method with missing data is analyzed in [29] under the same generative model as in (1). In [7], the authors provide the sample complexity for learning a low-rank subspace from subsampled data under a nonparametric model much more general than (1): the complete data vectors are assumed to be i.i.d. samples from a general probability distribution on $\mathbb{R}^n$. In the streaming setting, Oja's method, GROUSE, PETRELS are three popular algorithms for tackling the challenge of subspace learning with partial information. Other interesting approaches include online matrix completion methods [30]–[32]. See [33] for a recent review of relevant literature in this area. Local convergence of GROUSE is given in [4], [5]. Global convergence of GROUSE is established in [6] under the noiseless setting. In general, establishing finite sample global performance guarantees for GROUSE and other algorithms such as Oja's and PETRELS in the missing data case is still an open problem.

Unlike most work in the literature that seeks to establish finite-sample performance guarantees for various subspace

estimation algorithms, our results in this paper provide an asymptotically exact characterization of three popular methods in the high-dimensional limit. The main technical tool behind our analysis is the weak convergence of stochastic processes towards their scaling limits that are characterized by ODEs or stochastic differential equations (see, *e.g.*, [8]–[10], [15]).

Using ODEs to analyze stochastic recursive algorithms has a long history [34], [35]. An ODE analysis of an early subspace tracking algorithm was given in [36], and this result was adapted to analyze PETRELS for the nonsubsampled case [2]. Our results in this paper differ from previous analysis not only in that it can handle the more challenging case of incomplete observations. In addition, previous ODE analysis in [2], [36] keeps the ambient dimension n *fixed* and studies the asymptotic limit as the step size tends to 0. The resulting ODEs involve $\mathcal{O}(n)$ variables. In contrast, our analysis studies the limit as the dimension $n \to \infty$, and the resulting ODEs only involve at most $2d^2$ variables, where d is the dimension of the subspace which, in many practical situations, is a small constant. This low-dimensional characterization makes our limiting results more practical to use, especially when the ambient dimension n is large.

It is important to point out a limitation of our asymptotic analysis: we require the initial estimate $\boldsymbol{X}_0$ to be asymptotically correlated with the true subspace $\boldsymbol{U}$. To see why this is an issue, we note that if the initial cosine similarity matrix $\boldsymbol{Q}(0) = \boldsymbol{0}$ (*i.e.*, a fully uncorrelated initial estimate), then the ODEs in Theorems 1 and 2 only provide a trivial solution $\boldsymbol{Q}(t) \equiv 0$, yielding no useful information. In practice, a correlated initial estimate can be obtained by performing a PCA on a small batch of samples; it may also available from additional side information about the true subspace $\boldsymbol{U}$. Therefore, the requirement that $\boldsymbol{Q}(0)$ be invertible is not overly restrictive. Nevertheless, we observe in numerical simulations that, under sufficiently high SNRs, Oja's method, GROUSE and PETRELS can successfully estimate the subspace by starting from random initial guesses that are uncorrelated with $\boldsymbol{U}$. Extending our analysis framework to handle the case of random initial estimates is an important line of future work.

IV. Implications of High-Dimensional Analysis

The scaling limits presented in Section III provide asymptotically exact characterizations of the dynamic performance of Oja's method, GROUSE, and PETRELS. In this section, we discuss implications of these results. Analyzing the limiting ODEs also reveals the fundamental limits and phase transition phenomena associated with the steady-state performance of these algorithms.

A. Algorithmic Insights

By examining Theorem 1 and Theorem 2, we draw the following conclusions regarding the three subspace estimation algorithms.

1. Connections and differences between the algorithms. Theorem 1 implies that, as $n \to \infty$, Oja's method and GROUSE converge to the same deterministic limit process characterized as the solution of the ODE (29). This result is somewhat surprising, as the update rules of the two methods (see Algorithm 1 and Algorithm 2) appear to be sufficiently different.

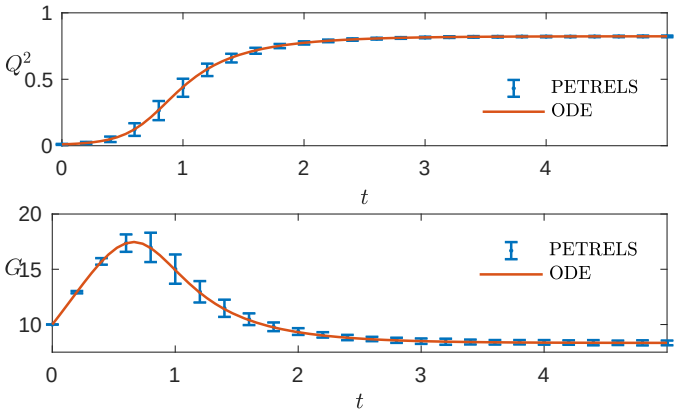


Figure 4. Monte Carlo simulations of the PETRELS algorithm v.s. asymptotic predictions obtained by the limiting ODEs given in Theorem 2 for $d = 1$. In this case, the two matrices $\boldsymbol{Q}(t)$ and $\boldsymbol{G}(t)$ reduce to two scalars $Q(t)$ and $G(t)$. The variable $G(t)$ acts as an effective step-size, which adaptively adjusts its value according to the change in $Q(t)$. The error bars shown in the figures represent one standard deviation over 50 independent trials. The signal dimension is $n = 10^4$.

Theorem 2 shows that PETRELS is also intricately connected to the other two algorithms. Indeed, the ODE (33) of the cosine similarity matrix $\boldsymbol{Q}(t)$ for PETRELS has exactly the same form as the one for GROUSE and Oja's method shown in (29), except for the fact that the nonadaptive step-size $\tau(t)\boldsymbol{I}_d$ in (29) is now replaced by a $d \times d$ matrix $\boldsymbol{G}(t)$, itself governed by an ODE (33). Thus, $\boldsymbol{G}(t)$ in PETRELS can be viewed as an adaptive scheme for adjusting the step-size.

To investigate how $\boldsymbol{G}(t)$ evolves, we run an experiment for $d = 1$. In this case, the quantities $\boldsymbol{Q}(t)$, $\boldsymbol{G}(t)$ and $\boldsymbol{\Lambda}$ reduce to three scalars, denoted by $Q(t)$, $G(t)$, and λ, respectively. Figure 4 shows the dynamics of PETRELS to recover this 1-D subspace. It shows that $G(t)$ increases initially, which helps to boost the convergence speed. As $Q(t)$ increases (meaning the estimates becoming more accurate), however, the effective step-size $G(t)$ gradually decreases, in order to help $Q(t)$ reach a higher steady-state value.

2. Subsampling vs. the SNR. The ODEs in Theorems 1 and 2 also reveal an interesting (asymptotic) equivalence between the subsampling ratio α and the SNR as specified by the matrix $\boldsymbol{\Lambda}$. To see this, we observe from the definition of the two functions F and H in (30) and (35) that α always appears together with $\boldsymbol{\Lambda}$ in the form of a product $\alpha\boldsymbol{\Lambda}^2$. This implies that an observation model with subsampling ratio α and SNR $\boldsymbol{\Lambda}$ will have the same asymptotic performance as a different model with subsampling ratio $\widehat{\alpha}$ and SNR $\sqrt{\alpha/\widehat{\alpha}}\,\boldsymbol{\Lambda}$. In simpler terms, having missing data is asymptotically equivalent to lowering the SNR in the fully-observable setting.

B. Oja's Method and GROUSE: Analytical Solutions and Phase Transitions

Next, we investigate the dynamics of Oja's method and GROUSE by studying the solution of the ODE given in

Theorem 1. To that end, we consider a change of variables by defining

$$\boldsymbol{P}(t) \stackrel{\text{def}}{=} [\boldsymbol{Q}(t)\boldsymbol{Q}^\top(t)]^{-1}. \tag{37}$$

One may deduce from (29) that the evolution of $\boldsymbol{P}(t)$ is also governed by a first-order ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}(t) = \boldsymbol{A}(t) - \boldsymbol{P}(t)\boldsymbol{B}(t) - \boldsymbol{B}(t)\boldsymbol{P}(t), \tag{38}$$

where

$$\boldsymbol{A}(t) = \tau(t)[2+\tau(t)]\alpha\boldsymbol{\Lambda}^2 \tag{39}$$

$$\boldsymbol{B}(t) = \tau(t)\big(\alpha\boldsymbol{\Lambda}^2 - \tfrac{\tau(t)}{2}\boldsymbol{I}_d\big) \tag{40}$$

are two diagonal matrices. Thanks to the linearity of (38), it admits an analytical solution

$$\begin{aligned}\boldsymbol{P}(t) = e^{-\int_0^t \boldsymbol{B}(r)\,\mathrm{d}r}\boldsymbol{P}(0)e^{-\int_0^t \boldsymbol{B}(r)\,\mathrm{d}r} \\ + \int_0^t \boldsymbol{A}(s)e^{-2\int_s^t \boldsymbol{B}(r)\,\mathrm{d}r}\,\mathrm{d}s.\end{aligned} \tag{41}$$

Note that the first two terms on the right-hand side of (41) represent the influence of the initial estimate $\boldsymbol{P}(0) = [\boldsymbol{Q}(0)\boldsymbol{Q}^\top(0)]^{-1}$ on the current state at t. In the special case of the algorithms using a constant step size, *i.e.*, $\tau(t) \equiv \tau > 0$, the solution (41) may be further simplified as

$$\boldsymbol{P}(t) = e^{-t\boldsymbol{B}}\boldsymbol{P}(0)e^{-t\boldsymbol{B}} + \boldsymbol{Z}(t), \tag{42}$$

where $\boldsymbol{Z}(t) = \operatorname{diag}\big\{z_1(t),\ldots,z_d(t)\big\}$ with

$$z_\ell(t) = \frac{(2+\tau)\alpha\lambda_\ell^2}{2\alpha\lambda_\ell^2 - \tau}\left(1 - e^{\tau(2\alpha\lambda_\ell^2-\tau)t}\right) \tag{43}$$

for $1 \le \ell \le d$. Note that if $2\alpha\lambda_\ell^2 - \tau = 0$ for some ℓ, the above expression for z_ℓ is understood via the convention that $(1-e^{-\tau t0})/0 = \tau t$.

The formula (42) reveals a phase transition phenomenon for the steady-state performance of the two algorithms as we change the step-size parameter τ. To see that, we first recall that the eigenvalues of $\boldsymbol{Q}^{(n)}(t)(\boldsymbol{Q}^{(n)}(t))^\top$ are exactly equal to the squared cosines of the d principal angles $\big\{\theta_\ell^n(t)\big\}$ between the true subspace $\boldsymbol{U}$ and the estimate given by the algorithms. We say an algorithm generates an asymptotically *informative solution* if

$$\lim_{t\to\infty}\lim_{n\to\infty}\cos^2(\theta_\ell^n(t)) > 0 \quad \text{for all } 1 \le \ell \le d, \tag{44}$$

i.e., the steady-state estimates of the algorithms achieve nontrivial correlations with all the directions of $\boldsymbol{U}$. In contrast, a *noninformative solution* corresponds to

$$\lim_{t\to\infty}\lim_{n\to\infty}\cos^2(\theta_\ell^n(t)) = 0 \quad \text{for all } 1 \le \ell \le d, \tag{45}$$

in which case the steady-state estimates carry no information about $\boldsymbol{U}$. For $d > 1$, one may also have the third situation where only a subset of the directions of $\boldsymbol{U}$ can be recovered (with nontrivial correlations) by the algorithm.

Proposition 1: Let $\theta_\ell^{(n)}(t)$ denotes the ℓth principal angle between the true subspace and the estimate obtained by Oja's method or GROUSE with a constant step size τ. Under the same assumptions as in Theorem 1, we have

$$\lim_{t\to\infty}\lim_{n\to\infty}\cos^2(\theta_\ell^{(n)}(t)) = \max\left\{0, \tfrac{2\alpha\lambda_\ell^2-\tau}{\alpha\lambda_\ell^2(2+\tau)}\right\}, \tag{46}$$

where $\{\lambda_\ell\}$ are the SNR parameters defined in (2). It follows that the two algorithms provide asymptotically informative solutions *if and only if*

$$\tau < 2\alpha \min_{1\le\ell\le d}\lambda_\ell^2. \tag{47}$$

Proof: Suppose the diagonal matrix $\boldsymbol{B}$ in (40) has d_1 positive diagonal entries (with $0 \le d_1 \le d$), and $d_2 = d - d_1$ negative or zero entries. Without loss of generatively, we may assume that $\boldsymbol{B}$ can be split into a block form $\begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{0}_{d_1\times d_2}\\ \boldsymbol{0}_{d_2\times d_1} & -\boldsymbol{B}_2\end{bmatrix}$ such that $\boldsymbol{B}_1$ only contains the positive diagonal entries, and $-\boldsymbol{B}_2$ only contains the nonpositive entires. Accordingly, we split the other two matrices in (42) as $\boldsymbol{P}(0) = \begin{bmatrix}\boldsymbol{P}_{1,1} & \boldsymbol{P}_{1,2}\\ \boldsymbol{P}_{2,1} & \boldsymbol{P}_{2,2}\end{bmatrix}$ and $\boldsymbol{Z}(t) = \begin{bmatrix}\boldsymbol{Z}_1(t) & \boldsymbol{0}_{d_1\times d_2}\\ \boldsymbol{0}_{d_2\times d_1} & \boldsymbol{Z}_2(t)\end{bmatrix}$. Applying the block matrix inverse formula to (42), we get

$$\boldsymbol{P}^{-1}(t) = \begin{bmatrix}\boldsymbol{W}_{1,1}(t) & \boldsymbol{W}_{1,2}(t)\\ \boldsymbol{W}_{2,1}(t) & \boldsymbol{W}_{2,2}(t)\end{bmatrix}, \tag{48}$$

where

$$\begin{aligned}\boldsymbol{W}_{1,1}^{-1}(t) &= e^{-t\boldsymbol{B}_1}\boldsymbol{P}_{1,1}e^{-t\boldsymbol{B}_1} + \boldsymbol{Z}_1(t)\\ &\quad - e^{-t\boldsymbol{B}_1}\boldsymbol{P}_{1,2}e^{t\boldsymbol{B}_2}\left(e^{t\boldsymbol{B}_2}\boldsymbol{P}_{2,2}e^{t\boldsymbol{B}_2} + \boldsymbol{Z}_2\right)^{-1}e^{t\boldsymbol{B}_2}\boldsymbol{P}_{2,1}e^{-t\boldsymbol{B}_1}.\end{aligned}$$

It is easy to verify from the definitions of $\boldsymbol{B}$ and $\boldsymbol{Z}$ that

$$\lim_{t\to\infty}\boldsymbol{W}_{1,1}(t) = \operatorname{diag}\left\{\tfrac{2\alpha\lambda_1^2-\tau}{\alpha\lambda_1^2(2+\tau)},\ldots,\tfrac{2\alpha\lambda_{d_1}^2-\tau}{\alpha\lambda_{d_1}^2(2+\tau)}\right\}. \tag{49}$$

Similarly, we may verify that

$$\begin{aligned}\lim_{t\to\infty}\boldsymbol{W}_{1,2}(t) &= \boldsymbol{0}_{d_1\times d_2}\\ \lim_{t\to\infty}\boldsymbol{W}_{2,2}(t) &= \boldsymbol{0}_{d2\times d_2}.\end{aligned} \tag{50}$$

Substituting (49) and (50) into (48) and recalling that the eigenvalues of $\boldsymbol{P}^{-1}(t)$ are exactly equal to the squared cosines of the principal angles, we reach (46). Applying the conditions given in (44) and (45) to (46) yields (47). ∎

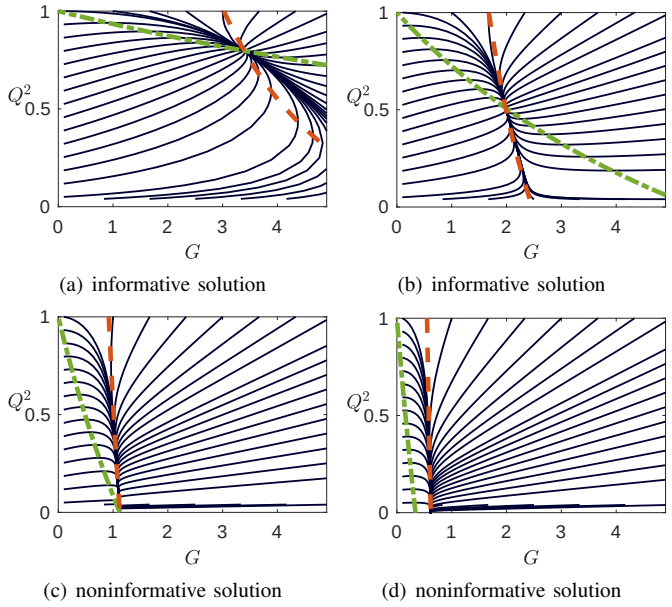
C. Steady-State Analysis of PETRELS

The steady-state property of PETRELS can also be obtained by studying the limiting ODEs as given in Theorem 2. The coupling of $\boldsymbol{Q}(t)$ and $\boldsymbol{G}(t)$ in (33) and (34), however, makes the analysis much more challenging. Unlike the case of Oja's method and GROUSE, we are not able to obtain closed-form analytical solutions of the ODEs for PETRELS. In what follows, we restrict our discussions to the special case of $d = 1$. This simplifies the task, as the matrix-valued ODEs (33) and (34) reduce to scalar-valued ones.

It is not hard to verify that, for any solution $\big\{Q(t), R(t)\big\}$ with an initial condition $\big\{Q(0), R(0)\big\}$, there is a symmetric solution $\big\{-Q(t), G(t)\big\}$ for the initial condition $\big\{-Q(0), G(0)\big\}$. To remove this redundancy, it is convenient to investigate the dynamics of $Q^2(t)$ and $G(t)$, which satisfy the following ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}[Q^2(t)] = GQ^2[2\alpha\lambda^2 - G - 2Q^2(1+\tfrac{1}{2}G)\alpha\lambda^2] \tag{51}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}G(t) = G[\mu - G(G+1)(Q^2\alpha\lambda^2+1)]. \tag{52}$$



(a) informative solution (b) informative solution

(c) noninformative solution (d) noninformative solution

Figure 5. Phase portraits of the nonlinear ODEs in Theorem 2: The black curves are trajectories of the solutions $(Q^2(t), G(t))$ of the ODES starting from different initial values. The green and red curves represent nontrivial solutions of the two stationary equations $\frac{\mathrm{d}Q^2(t)}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}G(t)}{\mathrm{d}t} = 0$. Their intersection point, if it exists, is a stable fixed point of the dynamical system. The fixed-points of the top two figures correspond to $Q^2(\infty) > 0$, and thus the steady-state solutions in these two cases are informative. In contrast, the fixed-points of the bottom two figures are associated with noninformative steady-state solutions with $Q^2(\infty) = 0$.

Figure 5 visualizes several different solution trajectories of these ODEs as the black curves in the Q–G plane. These solutions start from different initial conditions at the borders of the figures, and they converge to certain stationary points. The locations of these stationary points depend on the SNR $\{\lambda_\ell\}$, the subsampling ratio α and the discount parameter μ used by the algorithm. In Figures 5(a) and 5(b), the stationary points correspond to $Q^2 > 0$, and thus the algorithm generates asymptotically informative solutions according to the definition in (44). In contrast, Figure 5(c) and Figure 5(d) show the situations where the steady-state solutions are noninformative.

Proposition 2: Let $d = 1$. Under the same assumptions as in Theorem 2, PETRELS generates an asymptotically informative solution *if and only if*

$$\mu < \left(2\alpha\lambda^2 + \tfrac{1}{2}\right)^2 - \tfrac{1}{4}, \tag{53}$$

where μ is the parameter defined in (27), λ denotes the SNR in (2), and α is the subsampling ratio.

Proof: It follows from Theorem 2 that verifying the conditions (44) and (45) boils down to studying the fixed point of a dynamical system governed by the limiting ODEs (51) and (52). This task is in turn equivalent to setting the left-hand sides of the ODEs to zero and solving the resulting equations.

Let $\{Q^*, G^*\}$ be any solution to the equations $\frac{\mathrm{d}G^2}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}G}{\mathrm{d}t} = 0$. From the forms of the right-hand sides of (51) and (52), we see that $\{Q^*, G^*\}$ must fall into one of the following three cases:

Case I: $G^* = 0$ and Q^* can take arbitrary values;

Case II: $Q^* = 0$ and G^* is the unique positive solution to

$$G^*(G^* + 1) = \mu; \tag{54}$$

Case III: $Q^* \neq 0$ and $G^* \neq 0$.

A local stability analysis, deferred to the end of the proof, shows that the fixed points in Case I are always unstable, in the sense that any small perturbation will make the dynamics move away from these fixed points. Thus, we just need to focus on Case II and Case III, with the former corresponding to an uninformative solution and the latter to an informative one. We will show that, under (53), a fixed point in Case III exists and it is the unique stable fixed point. That solution disappears when (53) ceases to hold, in which case the solution in Case II becomes the unique stable fixed point.

To see why (53) provides the phase transition boundary, we note that a solution in Case III, if it exists, must satisfy $(Q^*)^2 = f(G^*)$ and $(Q^*)^2 = h(G^*)$, where

$$f(G) \overset{\text{def}}{=} \frac{\alpha\lambda^2 + 1}{(1 + \frac{G}{2})\alpha\lambda^2} - \frac{1}{\alpha\lambda^2} \tag{55}$$

$$h(G) \overset{\text{def}}{=} \left(\frac{\mu}{G(G+1)} - 1\right)\frac{1}{\alpha\lambda^2}. \tag{56}$$

The above two equations are derived from $\frac{\mathrm{d}Q^2(t)}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}G(t)}{\mathrm{d}t} = 0$. In Figure 5, the functions $f(G)$ and $h(G)$ are plotted as the green and red dashed lines, respectively.

It is easy to verify from their definitions that $f(G)$ and $h(G)$ are both monotonically decreasing in the feasible region ($0 \leq Q^2 \leq 1$ and $G > 0$). Moreover, $0 = f^{-1}(1) < h^{-1}(1)$, where f^{-1} and h^{-1} denote the inverse function of f and h, respectively. Thus, a solution in Case III exists if $f^{-1}(0) > h^{-1}(0)$, which then leads to (53) after some algebraic manipulations.

Finally, we examine the local stability of the fixed points in Case I and Case II. Note that a fixed point (Q^*, G^*) of the 2-dimensional ODE (51) and (52) is stable if and only if

$$\frac{\partial}{\partial[Q^2]}\left[\frac{\mathrm{d}}{\mathrm{d}t}Q^2(t)\right]|_{Q=Q^*,G=G^*} < 0$$

and

$$\frac{\partial}{\partial G}\left[\frac{\mathrm{d}}{\mathrm{d}t}G(t)\right]|_{Q=Q^*,G=G^*} < 0,$$

where $\frac{\mathrm{d}}{\mathrm{d}t}Q^2(t)$ and $\frac{\mathrm{d}}{\mathrm{d}t}G(t)$ are the functions on the right-hand side of (51) and (52), respectively. It follows that all the Case I fixed points are always unstable, because $\frac{\partial}{\partial G}\left[\frac{\mathrm{d}}{\mathrm{d}t}G(t)\right]|_{G=0} = \mu > 0$. Furthermore, the Case II fixed point is also unstable if (53) holds, because

$$\frac{\partial}{\partial[Q^2]}\left[\frac{\mathrm{d}}{\mathrm{d}t}Q^2(t)\right]|_{Q=0,G=G^*} = 2\alpha\lambda^2 - G^* > 0,$$

where G^* is the value specified in (54). On the other hand, when (53) does not hold, the Case II fixed point becomes stable. ∎

Example 3: Proposition 2 predicts a critical choice of μ (as a function of the SNR λ and the subsampling ratio α) that separates informative solutions from noninformative ones. This prediction is confirmed numerically in Figure 6. In our experiments, we set $d = 1$, $n = 10,000$. We then scan the

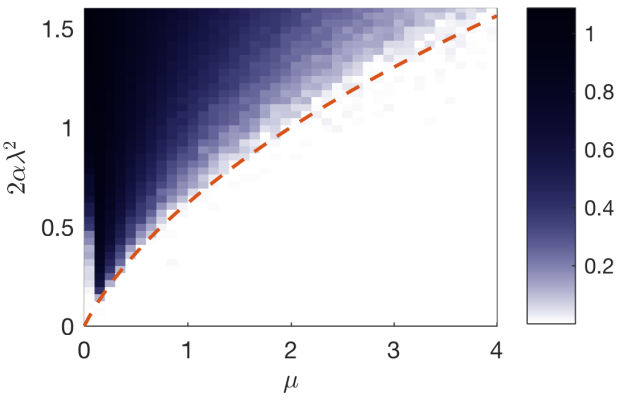


Figure 6. The grayscale in the figure visualizes the steady-state squared cosine similarities of PETRELS corresponding to different values of the SNR λ^2, the subsampling ratio α, and the step-size parameter μ. The red curve is the theoretical prediction given in Proposition 2 of a phase transition boundary, below which no informative solution can be achieved by the algorithm. The theoretical prediction matches numerical results.

parameter space of μ and $\alpha\lambda^2$. For each choice of these two parameters on our search grid, we perform 100 independent trials, with each trial using a different realizations of $\boldsymbol{c}_k$ and $\boldsymbol{a}_k$ in (1) and a different $\boldsymbol{U}$ drawn uniformly at random from the n-D sphere. The grayscale in Figure 6 shows the average value of the squared cosine similarity $Q(t)$ at $t = 10^3$.

V. Derivations of the ODEs and Proof Sketches

In this section, we present a nonrigorous derivation of the limiting ODEs and sketch the main ingredients of our proofs of Theorems 1 and 2. More technical details and the complete proofs can be found in the Supplementary Materials [16].

A. Derivations of the ODE

In what follows, we show how one may derive the limiting ODE in Theorem 1. We focus on GROUSE, but the other two algorithms can be treated similarly. For simplicity, we consider the case in which the subspace dimension is $d = 1$. In this case, the true subspace $\boldsymbol{U}$ in (1) and its estimate $\boldsymbol{X}_k$ given by Algorithm 2 reduce to vectors $\boldsymbol{u}$ and $\boldsymbol{x}_k$, respectively. The covariance matrix $\boldsymbol{\Lambda}$ in (2) also reduces to a scalar λ. Consequently, the weight vector $\widehat{\boldsymbol{w}}_k$ obtained in (9) becomes a scalar $w_k = \boldsymbol{x}_k^\top \boldsymbol{\Omega}_k \boldsymbol{s}_k / \|\boldsymbol{\Omega}_k \boldsymbol{x}_k\|^2$.

Our first observation is that the dynamic of GROUSE can be modeled by a Markov chain $(\boldsymbol{x}_k, \boldsymbol{u}_k)$ on $\mathbb{R}^{2n}$, where $\boldsymbol{u}_k \equiv \boldsymbol{u}$ for all k. The update rule of this Markov chain is

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = \left[\big(\cos(\theta_k) - 1\big) \frac{\boldsymbol{p}_k}{\|\boldsymbol{p}_k\|} + \sin(\theta_k) \frac{\boldsymbol{r}_k}{\|\boldsymbol{r}_k\|} \right] \mathbb{1}_{\mathcal{A}_k}, \tag{57}$$

where $\mathcal{A}_k = \{\|\boldsymbol{\Omega}_k \boldsymbol{x}_k\|^2 > \epsilon\}$. Here, the indicator function $\mathbb{1}_{\mathcal{A}_k}$ encodes the test in line 3 of Algorithm 2. Since we are considering the special case of $d = 1$, the vectors $\boldsymbol{r}_k$ and $\boldsymbol{p}_k$ as originally defined in (12) can be rewritten as

$$\boldsymbol{r}_k = \boldsymbol{\Omega}_k (\boldsymbol{s}_k - \boldsymbol{p}_k) \tag{58}$$

$$\boldsymbol{p}_k = \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top \boldsymbol{\Omega}_k \boldsymbol{s}_k}{\|\boldsymbol{\Omega}_k \boldsymbol{x}_k\|^2}. \tag{59}$$

Multiplying both sides of (57) from the left by $\boldsymbol{u}^\top$, we get

$$Q_{k+1} - Q_k = \tfrac{1}{n} g_k, \tag{60}$$

where

$$g_k \overset{\text{def}}{=} n \left[\big(\cos(\theta_k) - 1\big) \frac{\boldsymbol{u}^\top \boldsymbol{p}_k}{\|\boldsymbol{p}_k\|} + \sin(\theta_k) \frac{\boldsymbol{u}^\top \boldsymbol{r}_k}{\|\boldsymbol{r}_k\|} \right] \mathbb{1}_{\mathcal{A}_k} \tag{61}$$

specifies the increment of the cosine similarity from Q_k to Q_{k+1}.

To derive the limiting ODE, we first rewrite (60) as

$$\frac{Q_{k+1} - Q_k}{1/n} = \mathbb{E}_k\, g_k + (g_k - \mathbb{E}_k\, g_k), \tag{62}$$

where $\mathbb{E}_k$ denotes conditional expectation with respect to all the random elements encountered up to step $k-1$, *i.e.*, $\{\boldsymbol{c}_j, \boldsymbol{a}_j, \boldsymbol{\Omega}_j\}_{0 \le j \le k-1}$ in the generative model (1). One can show that

$$\mathbb{E}\,(g_k - \mathbb{E}_k\, g_k)^2 = \mathcal{O}(1) \tag{63}$$

and

$$\mathbb{E}_k\, g_k = F(Q_k, \tau_k) + \mathcal{O}(1/\sqrt{n}), \tag{64}$$

where $F(\cdot, \cdot)$ is the function defined in (30). Substituting (64) into (62) and omitting the zero-mean difference term $(g_k - \mathbb{E}_k\, g_k)$, we get

$$\frac{Q_{k+1} - Q_k}{1/n} = F(Q_k, \tau_k) + \mathcal{O}(1/\sqrt{n}). \tag{65}$$

Let $Q(t)$ be a continuous-time process defined as in (24), with $t = k/n$ being the rescaled time. In an intuitive but nonrigorous way, we have $\frac{Q_{k+1} - Q_k}{1/n} \to \frac{\mathrm{d}}{\mathrm{d}t} Q(t)$ as $n \to \infty$. This then gives us the ODE in (29).

In what follows, we provide some additional details behind the estimate in (64). To simplify our presentation, we first introduce a few variables. Let

$$\begin{aligned} z_k &\overset{\text{def}}{=} \|\boldsymbol{\Omega}_k \boldsymbol{x}_k\|^2, & \widetilde{z}_k &\overset{\text{def}}{=} \tfrac{1}{n} \|\boldsymbol{\Omega}_k \boldsymbol{s}_k\|^2 \\ p_k &\overset{\text{def}}{=} \boldsymbol{u}^\top \boldsymbol{\Omega}_k \boldsymbol{s}_k, & q_k &\overset{\text{def}}{=} \boldsymbol{x}_k^\top \boldsymbol{\Omega}_k \boldsymbol{s}_k \\ \widetilde{Q}_k &\overset{\text{def}}{=} \boldsymbol{u}^\top \boldsymbol{\Omega}_k \boldsymbol{x}_k. \end{aligned} \tag{66}$$

Since $\|\boldsymbol{u}\| = \|\boldsymbol{x}_k\| = 1$, all these variables are $\mathcal{O}(1)$ quantities when $n \to \infty$. (See Lemma 5 in Supplementary Materials.)

Given its definition in (13), we rewrite θ_k used in (57) as

$$\theta_k^2 = \frac{\tau_k^2}{n} \frac{q_k^2}{z_k^2} \left[\widetilde{z}_k^2 - \frac{q_k^2}{n z_k} \right] \sim \mathcal{O}(1/n).$$

Thus, it is natural to expand the two terms $\cos(\theta_k)$ and $\sin(\theta_k)$ that appear in (57) via a Taylor series expansion, which yields

$$\begin{aligned} \cos(\theta_k) &= 1 - \frac{\tau_k^2 \|\boldsymbol{r}_k\|^2 \cdot \|\boldsymbol{p}_k\|^2}{2n^2} + \mathcal{O}(n^{-2}) \\ \sin(\theta_k) &= \frac{\tau_k}{n} \|\boldsymbol{r}_k\|^2 \cdot \|\boldsymbol{p}_k\|^2 + \mathcal{O}(n^{-3/2}). \end{aligned} \tag{67}$$

Substituting (67) into (61) gives us

$$g_k = \tfrac{\tau_k}{z_k} \big[p_k q_k - \tfrac{1}{z_k} q_k^2 \big(\widetilde{Q}_k + \tfrac{\tau_k}{2} \widetilde{z}_k Q_k \big) \big] \mathbb{1}_{\{z_k > \epsilon\}} + \mathcal{O}(n^{-1/2}). \tag{68}$$

A rigorous justification of this step is presented as Lemma 8 in the Supplementary Materials.

One can show that, as $n \to \infty$, both z_k and $\widetilde{z}_k$ defined in (66) converge to α, the quantity $\widetilde{Q}_k \to \alpha Q_k$, and $\mathbb{1}_{\{z_k > \epsilon\}} \to 1$ (for any $\epsilon < \alpha$). Furthermore, we can also show that

$$\mathbb{E}\left|\mathbb{E}_k\, q_k^2 - \alpha\left(1 + \alpha\lambda^2 Q_k^2\right)\right| \le C/n,$$
$$\mathbb{E}\left|\mathbb{E}_k\, p_k q_k - \alpha Q_k(1 + \alpha\lambda^2)\right| \le C/n,$$

for some constant C. (The convergence of these variables is established in Lemma 6 in the Supplementary Materials.) Finally, by substituting the limiting values of the variables $z_k, \widetilde{z}_k, \widetilde{Q}_k, q_k^2, q_k p_k$ into (68), we reach the estimate in (64).

B. Main Steps of Our Proofs

The proofs of Theorems 1 and 2 follow a standard argument for proving the weak convergence of stochastic processes [8], [9], [11]. For example, to establish the scaling limit of Oja's method and GROUSE as stated in Theorem 1, our proof consists of three main steps.

First we show that the sequence of stochastic processes $\{\{Q^{(n)}(t)\}_{0\le t\le T}\}_{n=1,2,\ldots}$ indexed by n is tight. The tightness property then guarantees that any such sequence must have a converging sub-sequence. Second, we prove that the limit of any converging (sub)-sequence must be a solution of the ODE (29). Third, we show that the ODE (29) admits a *unique* solution. This last property can be easily established from the fact that the function $F(\cdot,\cdot)$ on the right-hand side of (29) is a Lipschitz function (noting that $\big|Q(t)\big| \le 1$ given the initial condition $\big|Q(0)\big| \le 1$). Combining the above three steps, we may then conclude that the entire sequence of stochastic processes $\{\{Q^{(n)}(t)\}_{0\le t\le T}\}_{n=1,2,\ldots}$ must converge weakly to the unique solution of the ODE.

VI. Conclusion

In this paper, we present a high-dimensional analysis of three popular algorithms, namely, Oja's method, GROUSE, and PETRELS, for estimating and tracking a low-rank subspace from streaming and incomplete observations. We show that, with proper time scaling, the time-varying trajectories of estimation errors of these methods converge weakly to deterministic functions of time. Such scaling limits are characterized as the unique solutions of certain ordinary differential equations (ODEs). Numerical simulations verify the accuracy of our asymptotic results. In addition to providing asymptotically exact performance predictions, our high-dimensional analysis yields several insights regarding the connections (and differences) between the three methods. Analyzing the limiting ODEs also reveals and characterizes phase transition phenomena associated with the steady-state performance of these techniques.

References

[1] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. Allerton Conference on Communication, Control and Computing.*, 2010.

[2] Y. Chi, R. Calderbank, and Y. C. Eldar, "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.

[3] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of Mathematical Biology*, vol. 15, no. 3, pp. 267–273, 1982.

[4] 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, Oct. 2015.

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

[6] ——, "Global Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 51, 2016, pp. 1460–1468.

[7] A. Gonen, D. Rosenbaum, Y. C. Eldar, and S. Shalev-Shwartz, "Subspace Learning with Partial Information," *Journal of Machine Learning Research*, vol. 17, pp. 1–21, 2016.

[8] S. Meleard and S. Roelly-Coppoletta, "A propagation of chaos result for a system of particles with moderate interaction," *Stochastic Processes and their Applications*, vol. 26, pp. 317–332, Jan. 1987.

[9] A.-S. Sznitman, "Topics in propagation of chaos," in *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, ser. Lecture Notes in Mathematics, P.-L. Hennequin, Ed. Springer Berlin Heidelberg, 1991, no. 1464, pp. 165–251.

[10] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. Wiley, 1985.

[11] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.

[12] J. Jacod and A. Shiryaev, *Limit theorems for stochastic processes*. Springer, 2003, vol. 288.

[13] C. Wang and Y. M. Lu, "Online learning for sparse pca in high dimensions: Exact dynamics and phase transitions," in *Proc. IEEE Information Theory Workshop (ITW)*, Cambridge, UK, Sep. 2016.

[14] ——, "The Scaling Limit of High-Dimensional Online Independent Component Analysis," in *Advances in Neural Information Processing Systems*, 2017.

[15] C. Wang, J. Mattingly, and Y. M. Lu, "Scaling Limit: Exact and Tractable Analysis of Online Learning Algorithms with Applications to Regularized Regression and PCA," *arXiv:1712.04332*, 2017.

[16] C. Wang, Y. C. Eldar, and Yue M. Lu, "Supplementary materials: Subspace estimation from incomplete observations: A high-dimensional analysis," 2018. [Online]. Available: https://lu.seas.harvard.edu/files/yuelu/files/subspace_supplementary.pdf

[17] B. Yang, "Projection Approximation Subspace Tracking," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 95–107, 1995.

[18] I. C. F. Ipsen and C. D. Meyer, "The Angle Between Complementary Subspaces," *The American Mathematical Monthly*, vol. 102, no. 10, pp. 904–911, 1995.

[19] F. Deutsch, "The angle between subspaces of a Hilbert space," in *Approximation theory, wavelets and applications*. Springer, 1995, pp. 107–130.

[20] P. Billingsley, *Convergence of probability measures*, 2nd ed., ser. Wiley series in probability and statistics. Probability and statistics section. New York: Wiley, 1999.

[21] I. Mitliagkas, C. Caramanis, and P. Jain, "Memory limited, streaming PCA," in *Advances in Neural Information Processing Systems*, 2013, pp. 2886–2894.

[22] M. Hardt and E. Price, "The noisy power method: A meta algorithm with applications," in *Advances in Neural Information Processing Systems*, 2014, pp. 2861–2869.

[23] M.-F. Balcan, S. S. Du, Y. Wang, and A. W. Yu, "An improved gap-dependency analysis of the noisy power method," in *Conference on Learning Theory*, 2016, pp. 284–309.

[24] C. J. Li, M. Wang, H. Liu, and T. Zhang, "Near-optimal stochastic approximation for online principal component estimation," *arXiv preprint arXiv:1603.05305*, 2016.

[25] 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 *Conference on Learning Theory*, 2016, pp. 1147–1164.

[26] A. Balsubramani, S. Dasgupta, and Y. Freund, "The fast convergence of incremental pca," in *Advances in Neural Information Processing Systems*, 2013, pp. 3174–3182.

[27] O. Shamir, "Convergence of stochastic gradient descent for pca," in *International Conference on Machine Learning*, 2016, pp. 257–265.

[28] Z. Allen-Zhu and Y. Li, "First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate," in *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*. IEEE, 2017, pp. 487–492.

[29] I. Mitliagkas, C. Caramanis, and P. Jain, "Streaming PCA with many missing entries," 2014.
[30] 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.
[31] C. Jin, S. M. Kakade, and P. Netrapalli, "Provable efficient online matrix completion via non-convex stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 4520–4528.
[32] B. Lois and N. Vaswani, "Online matrix completion and online robust pca," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 1826–1830.
[33] L. Balzano, Y. Chi, and Y. M. Lu, "A Modern Perspective on Streaming PCA and Subspace Tracking: The Missing Data Case," *Proceedings of the IEEE*, no. to appear, 2018.
[34] T. G. Kurtz, "Solutions of ordinary differential equations as limits of pure jump Markov processes," *Journal of Applied Probability*, vol. 7, no. 1, p. 49, Apr. 1970.
[35] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 22, no. 4, pp. 551–575, 1977.
[36] B. Yang, "Asymptotic convergence analysis of the projection approximation subspace tracking algorithms," *Signal Processing*, vol. 50, no. 1–2, pp. 123–136, Apr. 1996.