A NON-CONVEX FRAMEWORK FOR STRUCTURED NON-STATIONARY COVARIANCE RECOVERY THEORY AND APPLICATION

BY

KATHERINE TSAI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Adviser:

Assistant Professor Oluwasanmi Koyejo

ABSTRACT

Flexible, yet interpretable, models for the second-order temporal structure are needed in scientific analyses of high-dimensional data. The thesis develops a structured time-indexed covariance model for non-stationary time-series data by decomposing them into sparse spatial and temporally smooth components. Traditionally, time-indexed covariance models without structure require a large sample size to be estimable. While the covariances factorization results in both domain interpretability and ease of estimation from the statistical perspective, the resulting optimization problem used to estimate the model components is non-convex. We design an optimization scheme with a carefully tailored spectral initialization, combined with iteratively refined alternating projected gradient descent. We prove a linear convergence rate for the proposed descent scheme and establish sample complexity guarantees for the estimator. As a motivating example, we consider the neuroscience application of estimation of dynamic brain connectivity. Empirical results using simulated and real brain imaging data illustrate that our approach improves time-varying covariance estimation as compared to baselines.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First, I would like to thank my thesis advisor, Professor Sanmi Koyejo, for giving me the opportunity to do research. His advisement and dedication to research have been constant inspirations for both my professional and personal growth. Although he is always busy, he still manages to make time to meet with every student, bring up novel research ideas, and provide insightful suggestions. I am deeply grateful for having valuable discussions with him.

I also wish to thank Professor Mladen Kolar for his mentoring and patience. In particular, the analysis part of this thesis is mainly completed with his guidance. His wealth of knowledge and sharp insights on problems have fostered this thesis.

Furthermore, I would like to thank Professor Grace Gao. Her enthusiasm and persistence have been positive inspirations for me to pursue my research interests. I occasionally recall meaningful discussions with her, and it always motivates me to work harder.

During my Master's degree at the University of Illinois Urbana-Champaign, I had the pleasure to work with fellow research group members, past and present, Juho Kim, Gaurush Hiranandani, Olawale Salaudeen, Chase Duncan, Raj Kiriti, Sameer Manchanda, Haroun Habeeb, Bliss Chapman, Forest Yang, Brando Miranda, Cong Xie, Yibo Zhang, Peiye Zhuang, Patrick Cole, and Andrew Chen. Our lab diverse interests had inspired me to explore different topics, and I enjoyed having random discussions after meetings. I am looking forward to continuing my Ph.D. here and the collaborations ahead!

I was lucky to work with research fellows, Sriramya Bhamidipati, Shubhendra Chauhan, Shubh Gupta, Ashwin Kanhere, Tara Mina, Pulkit Rustagi, Akshay Shetty, Siddharth Tanwar, and Matthew Peretic for a semester. I had a great time working with the intelligent, thoughtful, and entertaining people here, and our friendship extended from research projects.

I would like to extend my thanks to Sameer Manchanda, Michael Andersen,

and James Macquarie Shine for providing the preprocessed Human Connectome Project (HCP) dataset and code of prior work. The use of these data and codes help me to compare and analyze the objectives of this thesis.

Finally, I would like to express my gratitude to my family for their enormous love and support. My academic accomplishments, along with this thesis, are dedicated to my family. I would like to thank my father, who has been a role model for me, for his words have always encouraged me when I faced obstacles. I would like to thank my mother, my lifetime friend, for putting faith in me.

The material of this thesis is funded by NSF IIS 1934986 and an SRI grant from the Grainger College of Engineering, the University of Illinois at Urbana-Champaign.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Dynamic covariances appear prominently in the analysis of non-stationary time series data and provide fundamental insights into complex systems. Dynamic covariance models are used in applications ranging from computational finance and economics [1, 2] to epidemiology [3] and neuroscience [4, 5]. This thesis is motivated by the need to study dynamic Functional brain Network Connectivity (dFNC), which has been applied to the scientific study of cognition [6], and human behavior [7, 8], among other scientific and clinical questions. Estimation of dynamic covariance matrices is challenging since the number of parameters to estimate is $O(TP^2)$, where T is the temporal length and P is the data dimension, while only data from N subjects are available, with $N \ll P$, even possibly $N = 1$, as is the case in many scientific experiments.

Parsimonious models that leverage structure underlying complex data are frequently used to deal with data scarcity. For example, [9] used structural penalty functions and [10, 11] used constraint sets when estimating model parameters, while [12] incorporated structural priors for probabilistic models. In this thesis, we estimate dynamic covariance matrices that are temporally smooth, spatially sparse, and low-rank, which is motivated by the study of dynamic functional brain network connectivity where collected time-series data exhibit such structures in the second moment [5, 13]. We use a factor model to encode the low-rank structure and further, restrict the factors to convex or non-convex constraint sets. The sparse spatial structure is imposed via iterative hard-thresholding and the smooth temporal structure via a temporal kernel projection.

Taken together, the optimization program used to estimate the dynamic covariance model is non-convex. With non-convex optimization, there is a risk of only finding a local optimum; however, there is growing evidence that for certain structured problems, convergence to the global optimum is guaranteed

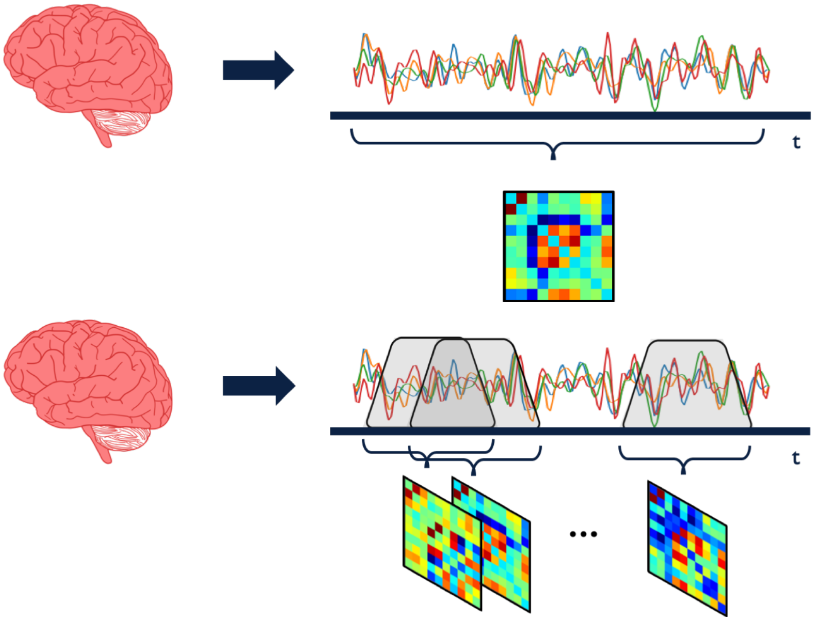

Figure 1.1: **Top**: An illustration of sample covariance of brain time-series signals. The static covariance assumes that the samples are i.i.d. across time. **Bottom**: An illustration of sliding window sample covariance. Each covariance matrix is computed by taking the average of the sample covariances within a window.

under suitable regularity conditions and initialization [14, 15, 16]. That is, if the distance between the initialization point and the optimal point is bounded within a ball of finite radius, then we can use the local regularity conditions to ensure the convergence. Moreover, estimates obtained through non-convex optimization are faster to compute and have better statistical performance [17, 18, 19]. Encouraged by the above line of work, we propose a two-stage procedure for estimating the dynamic covariance model. In the first stage, we use a spectral method to initialize our estimate, which is subsequently refined in the second stage using projected gradient descent.

1.1 A Motivating Example and Limitations

One application of dynamic covariances estimations is the study of dynamic functional brain network connectivity. We will begin by introducing the background of functional connectivity and then discuss the "dynamic" version in the following paragraph. The functional connectivity of brain networks represents temporal connections between different brain regions. Apart from the structural connectivity that studies anatomical brain structures, the functional connectivity is based upon fluctuations of signals generated by brain networks. Since we could not directly observe the signals in our brains, we rely on measurement protocols such as ElectroEncephaloGraphy (EEG) and functional Magnetic Resonance Imaging (fMRI). However, these protocols are notoriously known to be noisy, impeding us from recovering the right brain networks. To resolve this dilemma, we resort to statistical analysis: conduct a series of brain network experiments and summarize the frequency of particular events. A classical approach to measuring the connectivity between two brain regions is to extract a sequence of time-series data and compute the sample covariance is shown in the top of Figure 1.1. This approach is based on the implicit assumption that the data points are independent and identically distributed (i.i.d.) across time. If the assumption is correct, the sample covariance has a desirable property that it would converge to the population covariance as the number of data points goes to infinity. The i.i.d assumption, which implies that the measured signals are stationary, does not hold in the real-world setting. In fact, brain networks exhibit highly complex dynamics and are considered non-stationary, leading to the study of dynamic functional connectivity. An adaptive approach to estimate the dynamic functional connectivity is to compute the sliding window covariance matrix. The idea is as follows; we look at $\mathbf{\Delta}$ samples ahead time t and $\mathbf{\Delta}$ samples after time t to compute the covariance at time t and then compute the average of the sample covariance within this window. As we will see shortly, the choice of the window length $2\mathbf{\Delta} + 1$ is tricky, and improper selection results in spurious fluctuations [20]. The following example is reproduced from the work [20].

Let us first assume that the functional connectivity pattern of brain regions x and y change over time. The sequence temporal random signals we detect from both regions are $x(t) = \sqrt{2}\cos(2\pi ft + \theta_x)$ and $y(t) = \sqrt{2}\cos(2\pi ft + \theta_y)\cos(2\pi f_0 t)$, where the phases are two independent random variables that follow identical uniform distribution $U([-\pi, \pi])$. Here $x(t)$ is a stationary signal and $y(t)$ is a non-stationary signal. The equation of the sliding-window covariance at time t is $C_{xy}(t) = \frac{1}{w}\frac{\sin(\pi f_0 w)}{\sin(\pi f_0)}\cos(2\pi f_0 t)$, where w is the window length. As we see the result from Figure 1.2 that when we change window length, the covariance value changes. This would lead to spurious fluctuations

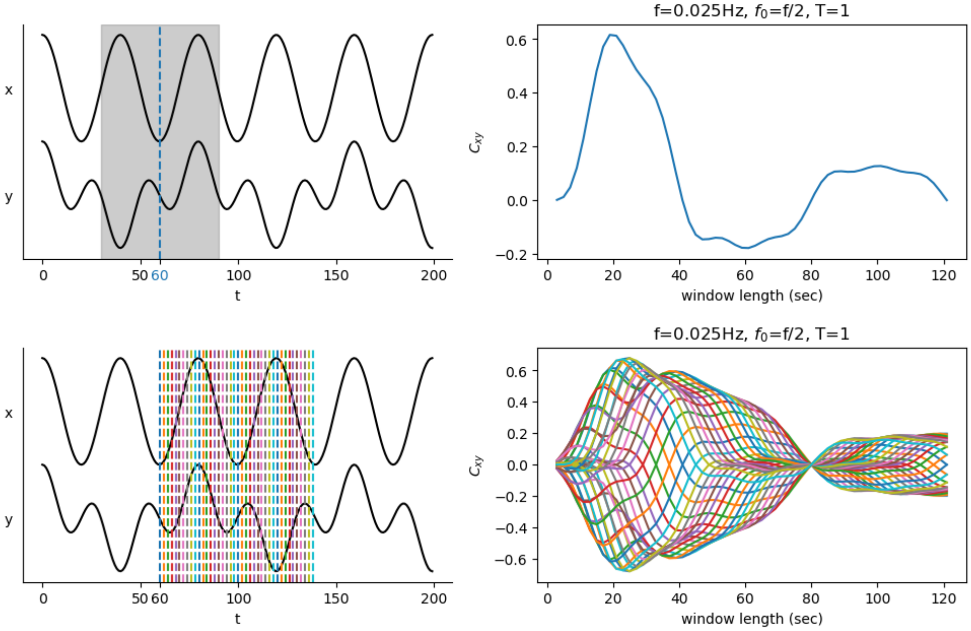

Figure 1.2: **Top left**: The blue dashed line on the left figure denotes the time at $t = 60$ and the shaded area denotes the sliding-window covering area. **Top right**: The covariance of x and y at $t = 60$ where x-axis denotes the sliding window length. **Bottom left**: Repeat the same process from $t = 60$ to $t = 120$ with sampling interval equals 2. **Bottom right**: This shows (1) different window length leads to different C_{xy} and (2) same window length at different positions t have different C_{xy}.

of estimations and bias in interpretations. An naive solution to address this problem is to compute the instantaneous sample covariance, and yet this would lead to poor estimates because of the small sample size at each time point.

Given the examples addressed above, we see that it is challenging to estimate the covariance without any model structures and purely relying on random measurements. Alternatively, we could improve estimates by imposing structures on covariances. Our goal is to recover the ground truth, presented at the end of the thesis, if the covariance matrices are low-rank.

1.2 Related Work

We focus on two main categories of prior work that are closely related to this work, namely factor models and the autoregressive model class. The factor model class constructs latent factors to capture the spatial and temporal structures, which implicitly imposes the low-rank structure. One common approach to model the temporal structure is by introducing latent kernel-regularized factors (or Gaussian process priors) [21, 22, 23, 24]. Other structures, such as sparsity and group sparsity, can also be encouraged by selecting proper priors [25]. The problem of estimating the probabilistic hierarchical models is that the underlying posterior distributions are often intractable, and existing inference methods are computationally demanding. Another closely related method is dictionary learning [26], which encodes data as the product of the temporal component and spatial component. On the other hand, the autoregressive model [2, 27, 28] encodes the temporal structure by modeling the current data point as linear combinations of previous data points and additional noise. The linear coefficients are encapsulated in the matrix, often referred to as the transition matrix. The kernel-smoothing estimator (KSE) [29] is an autoregressive-based model whose covariances are approximated by kernel-weighted functions. Moreover, building structured transition matrices has shown to improve the computation efficiency as well as the prediction accuracy [30, 31]. Other standard approaches which are popular in neuroscience uses are the sliding window (SW) [20, 32, 33] and the hidden Markov model (HMM) [5]. Although structured dynamic covariance estimation is a long-standing problem in neuroscience, to the best of our knowledge, existing methods have focused exclusively on the probabilistic models, and optimization and statistical properties have rarely been studied.

1.3 Contributions of Thesis

This thesis proposes a non-convex optimization scheme for estimating dynamic structured covariances along with the theoretical analysis.

- We prove the linear convergence of the proposed algorithm to the global optimum up to a finite statistical error.
- We propose and prove the sample complexity of the spectral initialization using the Davis-Kahan sinΘ theorem [34], Bernstein matrix concentration inequality, and the Corant-Fischer min-max theorem.
- We show in experiments that our model can successfully recover the temporal smoothness and detect temporal change induced by task activation.

CHAPTER 2

PROPOSED MODEL

2.1 Problem Statement and Notation

Consider an experiment with N subjects, where for each subject we collect T observations recorded at times $t = 1, 2, \ldots, T$. For a subject n at time t we observe $\mathbf{x}_t^{(n)} \in \mathbb{R}^P$, which is independent and mean zero. The sample matrix for a subject is denoted as $\mathbf{X}^{(n)} = [\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_T^{(n)}] \in \mathbb{R}^{P \times T}$. Let $\mathbf{S}_{N,t} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_t^{(n)} \mathbf{x}_t^{(n)\top}$ be the sample covariance of N subjects at time t. We assume that the population covariance has the following structure

$$\mathbb{E}[\mathbf{S}_{N,t}] = \mathbf{\Sigma}_t^\star + \mathbf{E}_t = \mathbf{V}^\star \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{V}^{\star\top} + \mathbf{E}_t, \quad \forall t \in [T] \tag{2.1}$$

where $\mathbf{\Sigma}_t^\star$ is at most rank K and $\mathbf{E}_t$ is a noise matrix. In the factorization of $\mathbf{\Sigma}_t^\star$, we assume that the spatial components $\mathbf{V}^\star = [\mathbf{v}_1^\star, \ldots, \mathbf{v}_K^\star] \in \mathbb{R}^{P \times K}$ are time-invariant and orthogonal to each other, while the temporal components $\mathbf{A}^\star = [\mathbf{a}_1^\star, \ldots, \mathbf{a}_T^\star] = [\tilde{\mathbf{a}}_1^\star, \ldots, \tilde{\mathbf{a}}_K^\star]^\top \in \mathbb{R}^{K \times T}$ are time-dependent. We assume that the columns of $\mathbf{V}^\star$ are sparse and belong to the set $\mathcal{C}_{\mathbf{V}^\star} = \{\mathbf{v} \in \mathbb{R}^P : \|\mathbf{v}\|_0 \leq s^\star, \|\mathbf{v}\|_2 = 1\}$, while the rows of $\mathbf{A}^\star$ are smooth and bounded, and belong to the set $\mathcal{C}_{\mathbf{A}^\star} = \{\tilde{\mathbf{a}} \in \mathbb{R}^T : b \leq \tilde{a}_i \leq c, \tilde{\mathbf{a}}^\top \mathbf{G}^{-1} \tilde{\mathbf{a}} \leq \gamma^\star\}$, where $0 \leq b < c$ and $\mathbf{G}$ is a $T \times T$ positive definite kernel matrix. Moreover, $G_{i,j} = \kappa(i,j)$ and $\kappa(\cdot,\cdot)$ is the kernel metric. These structures are based on the assumption that data are spatially sparse and temporally smooth as shown in Figure 2.1.

Under the model in (2.1), we estimate $\mathbf{V}$ and $\mathbf{A}$ from samples $\{\mathbf{X}^{(n)}\}_{n \in [N]}$ by minimizing the following objective

$$\min f_N(\mathbf{Z}) = \min_{\substack{\mathbf{V} \in \mathcal{C}_V \\ \mathbf{A} \in \mathcal{C}_A}} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{2} \|\mathbf{S}_{N,t} - \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \mathbf{V}^\top\|_F^2 \tag{2.2}$$

where $\mathbf{Z} = [\mathbf{V}^\top, \mathbf{A}]^\top$. Although f_N is non-convex with respect to $\mathbf{Z}$, the loss

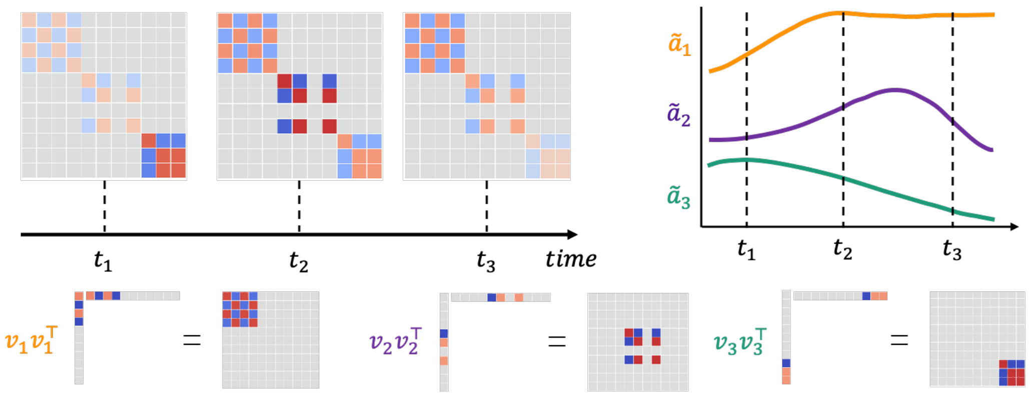

Figure 2.1: **Upper left**: Example of dynamic covariances. **Upper right**: The smooth temporal coefficients $\tilde{\mathbf{a}}_k$ associated with corresponding $\mathbf{v}_k$. **Bottom**: The sparse spatial components $\mathbf{v}_k$.

$\ell_{N,t}(\mathbf{\Sigma}_t) = \frac{1}{2}\|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t\|_F^2$ is μ-strongly convex and L-smooth with respect to the product $\mathbf{\Sigma}_t = \mathbf{V}\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^\top$, where in this case $\mu = L = 1$. Notice that in the optimization problem (2.2), we do not enforce the columns of $\mathbf{V}$ to be orthogonal and yet, with proper initialization, our algorithm will output $\mathbf{V}$ with nearly orthogonal columns. This is because $\mathbf{V}$ will be close to $\mathbf{V}^\star$, whose columns are orthogonal, under some columnwise permutation. To impose the structures onto $\mathbf{V}$ and $\mathbf{A}$, we project columns of $\mathbf{V}$ to $\mathcal{C}_{\mathbf{V}} = \{\mathbf{v} \in \mathbb{R}^P : \|\mathbf{v}\|_0 \leq s, \|\mathbf{v}\|_2 = 1\}$, where $s > s^*$ and rows of $\mathbf{A}$ to $\mathcal{C}_{\mathbf{A}} = \mathcal{C}_{\mathbf{A}^\star}$.

Since there is more than one factorization for each covariance, that is, $\mathbf{\Sigma} = \mathbf{V}\operatorname{diag}(\mathbf{a})\mathbf{V}^\top$ can also be factorized as $\mathbf{V}\mathbf{R}\mathbf{R}^\top\operatorname{diag}(\mathbf{a}_1)\mathbf{R}\mathbf{R}^\top\mathbf{V}^\top$, where $\mathbf{R}$ is permutation matrix, we use the follow metric to evaluate the distance between $\mathbf{Z}^\top = [\mathbf{V}^\top\ \mathbf{A}]^\top$ and $\mathbf{Z}^{\star\top} = [\mathbf{V}^{\star\top}\ \mathbf{A}^\star]^\top$.

Definition 1. *The distance between* $\mathbf{Z}$ *and* $\mathbf{Z}^\star$ *is defined as*

$$\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) = \sum_{t=1}^{T} \min_{\mathbf{R}_t \in \mathcal{P}(K)} \left\{ \|\mathbf{V} - \mathbf{V}^\star\mathbf{R}_t\|_F^2 + \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2 \right\} \quad (2.3)$$

where $\mathcal{P}(K) \ni \mathbf{R}_t$ is the set of $K \times K$ permutation matrices and $\operatorname{diag}(\cdot) : \mathbb{R}^T \to \mathbb{R}^{T\times T}$ converts a T-dimensional vector to a $T \times T$ diagonal matrix. This metric is commonly used in matrix factorization problems, where $\mathcal{P}(K)$ is replaced by the set of rotation matrices. Here, we limit the set of matrices to be permutation matrix because $\mathbf{R}^\top\operatorname{diag}(\mathbf{a}_t)\mathbf{R}$ should also be a diagonal matrix. Throughout the thesis, we sometimes write $d^2(\mathbf{Z}_t, \mathbf{Z}_t^\star)$ as $\min_{\mathbf{R}_t\in\mathcal{P}(K)} \|\mathbf{V} - \mathbf{V}^\star\mathbf{R}_t\|_F^2 + \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2$, where $\mathbf{Z}_t = [\mathbf{V}^\top\operatorname{diag}(\mathbf{a}_t)]^\top$ and $\mathbf{Z}_t^\star = [\mathbf{V}^{\star\top}\operatorname{diag}(\mathbf{a}_t^\star)]^\top$.

In many high-dimensional statistical problems, we allow a small statistical error [35]. Thus, if the algorithm converges to the stationary point, it may deviate away from the global optimum up to finite multiple of the statistical error. In this thesis, we define the statistical error as follows.

Definition 2.

$$\varepsilon_{stat} = \sup_{\|\mathbf{\Delta}_t\|=1 \mathbf{\Delta}_t \in \Xi(2K,2s+s^\star)} \langle \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star), \mathbf{\Delta}_t \rangle \tag{2.4}$$

where $\Xi(2K, 2s+s^\star) = \{\mathbf{\Sigma} \in \mathbb{R}^{P\times P} : \text{rank}(\mathbf{\Sigma}) \leq 2K, \|\mathbf{v}_k\|_0 \leq 2s+s^\star, \|\mathbf{v}_k\|_1 = 1\ \forall k \in [2K]\}$. Here we restrict the differences of the estimation updates and the optimal solutions to lie within the constraint set Ξ. The statistical error describes the geometric landscape around the optimum: it quantifies the gradient magnitude of the empirical loss function $\ell_{N,t}$ $\forall t \in [T]$ in the vicinity of the population optimal $\mathbf{\Sigma}_t^\star$ $\forall t \in [T]$ and implicitly depends on the sample size as discussed later in Chapter 3.

2.2 Proposed Algorithm

The objective (2.2) is non-convex, which implies that there may exist multiple local optima; however, convergence close to global optimum can be guaranteed through careful initialization, proper learning rate selection, subject to local regularity conditions. To this end, our algorithm consists of two parts: spectral initialization and local iterative refinement.

2.2.1 Spectral Initialization

By large sample theory, the sample covariance will converge to the population covariance. Thus, $\mathbf{V}$ and $\mathbf{A}$ estimated via eigendecompositions of $\{\mathbf{S}_{N,t}\}_{t\in[T]}$ will be close to $\mathbf{V}^\star$ and $\mathbf{A}^\star$, respectively, as the number of sample N increases. Although large sample sizes are not considered practical in many scientific experiments, we exploit the structure of the ground truth matrices $\{\mathbf{\Sigma}_t^\star\}_{t\in[T]}$, which share the same underlying eigensubspace, to aggregate the effective sample size. As Algorithm 1 shows, we sum the sample covariance across time and employ a joint eigendecomposition to estimate the initial value of $\mathbf{V}$, denoted as $\mathbf{V}^0$. In the following step, the initial estimate of the temporal coefficient $\mathbf{A}$, denoted as $\mathbf{A}^0$, is estimated by projecting $\{\mathbf{S}_{N,t}\}_{t\in[T]}$ to the estimated subspace $\mathbf{V}^0$. We will analyze the sample complexity of this initialization method in Chapter 4.

```
Algorithm 1: Spectral Initialization
  Input: Data sequences:$\{\mathbf{X}^{(n)}\}_{n\in[N]}$, number of components $K$
  Output: Initial spatial components $\mathbf{V}^0$, initial temporal component
          $\mathbf{A}^0$
  $\mathbf{M}_N = \sum_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_t^{(n)} \mathbf{x}_t^{(n)\top}$
  $\mathbf{\Sigma}, \mathbf{U} \leftarrow$ Eigendecomposition($\mathbf{M}_N$)
  $\mathbf{V}^0 = [\mathbf{v}_1^0, \mathbf{v}_2^0, \ldots, \mathbf{v}_k^0] \leftarrow$ top $K$ eigenvectors of $\mathbf{U}$
  for $t = 1, 2, \ldots, T$ do
      $\mathbf{S}_{N,t} = \frac{1}{N} \sum_{n=1}^{n} \mathbf{x}_t^{(n)} \mathbf{x}_t^{(n)\top}$
      for $k = 1, 2, \ldots, K$ do
          $a_{k,t}^0 \leftarrow \mathbf{v}_k^{0\top} \mathbf{S}_{N,t} \mathbf{v}_k^0$
      end
  end
  $\mathbf{A}^0 = [a_{k,t}^0]$
  return $\mathbf{V}^0, \mathbf{A}^0$
```

2.2.2 Local Refinement

After initialization, we iteratively refine estimates of $\mathbf{V}$ and $\mathbf{A}$ via alternating minimization, as shown in Algorithm 2, where η denotes the step size. We scale down the step size for $\mathbf{V}$ by T to balance the scale of the gradient as T increases. We enforce the sparse structure of $\mathbf{V}$, and smooth structure of $\mathbf{A}$ via projection to constraint sets $\mathcal{C}_{\mathbf{V}}$ and $\mathcal{C}_{\mathbf{A}}$, respectively. Although $\mathcal{C}_{\mathbf{V}}$ is a non-convex constraint set, projection to this set can be computed efficiently by picking the top-s largest entries in magnitude and then projecting the constructed vector to the unit sphere. The proof and the expansion coefficient ρ induced by projecting to non-convex set $\mathcal{C}_{\mathbf{V}}$ are presented in Appendix A.1. In general, the value of ρ depends on the difference of sparsity levels $s-s^\star$ and the initialization distance. On the other hand, projection to $\mathcal{C}_{\mathbf{A}}$ can be solved efficiently via convex programming. In practice, we use an alternating projection method projecting to two convex sets $\{\tilde{\mathbf{a}} \in \mathbb{R}^T : b \leq \tilde{a}_i \leq c \ \forall i \in [T]\}$ and $\{\tilde{\mathbf{a}} \in \mathbb{R}^T : \tilde{\mathbf{a}}^\top \mathbf{G}^{-1} \tilde{\mathbf{a}} \leq \gamma\}$. By von-Neumann's theorem [11], alternating projection to convex sets will converge to their intersection; hence it does not affect the main analysis. Additionally, choosing a proper step size is important when solving non-convex problems as the estimates can escape the local region if the step is too large. We will discuss the selection of step size in Chapter 4.

Algorithm 2: Dynamic Covariance Learning

```
Input: Data sequences $\{\mathbf{X}^{(n)}\}_{n\in[N]}$, number of components $K$,
       tolerance $\varepsilon$
Output: Spatial Dictionary $\mathbf{V}$, Temporal Dictionary $\mathbf{A}$
$\mathbf{V}^0, \mathbf{A}^0 \leftarrow$Spectral Initialization$\big(\{\mathbf{X}^{(n)}\}_{n=1}^{N}\big)$
for $t = 1, 2, \dots, T$ do
    $\mathbf{S}_{N,t} \leftarrow \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_t\mathbf{x}_t^\top$
end
while $|f_N(\mathbf{Z}^{i-1})| - f_N(\mathbf{Z}^{i-2})| > \varepsilon$ do
    $\widehat{\mathbf{A}}^i \leftarrow \mathbf{A}^{i-1} - \eta\left(\frac{1}{T}\mathbf{W}^{i-1}\right)$
    $\mathbf{A}^i \leftarrow$ Project rows of $\widehat{\mathbf{A}}^i$ to $\mathcal{C}_\mathbf{A}$
    $\widehat{\mathbf{V}}^i \leftarrow$
      $\mathbf{V}^{i-1} - \frac{\eta}{T}\left(\frac{2}{T}\sum_{t=1}^{T}(\mathbf{V}^{i-1}\operatorname{diag}(\mathbf{a}_t^{i-1})\mathbf{V}^{i-1\top} - \mathbf{S}_{N,t})\mathbf{V}^{i-1}\operatorname{diag}(\mathbf{a}_t^{i-1})\right)$
    $\mathbf{V}^i \leftarrow$ Project columns of $\widehat{\mathbf{V}}^i$ to $\mathcal{C}_\mathbf{V}$
end
```

$\mathbf{W}^{i-1} = [w_{k,t}], \quad w_{k,t} = \mathbf{v}_k^{i-1\top}(\mathbf{V}^{i-1}\operatorname{diag}(\mathbf{a}_t^{i-1})\mathbf{V}^{i-1\top} - \mathbf{S}_{N,t})\mathbf{v}_k^{i-1}$

CHAPTER 3

CONVERGENCE ANALYSIS

This chapter discusses the convergence of the proposed algorithm. The proof is an extension of the Factored Gradient Descent (FGD) [14, 36, 37] analysis. We begin by stating the underlying assumptions, followed by the main results and the sketch of proofs.

3.1 Assumptions

Assumption A:

The initial estimate of $\{\mathbf{Z}_t^0\}_{t=1}^T$ *is required to to satisfy*

$$d^2(\mathbf{Z}_t^0, \mathbf{Z}_t^\star) \leq 2r^2, \forall t \in [T] \tag{3.1}$$

and

$$r^2 \leq \min\left(\frac{L\mu}{4\xi^2(L+\mu)^2}\frac{1}{(4+c^2)}, 1\right) \tag{3.2}$$

where $\xi^2 = \min_{t\in[T]}\left(\frac{1}{\sigma_K(\mathbf{\Sigma}_t^\star)}\right)^2 + \left(1 + 3\frac{c}{\sigma_K(\mathbf{\Sigma}_t^\star)}\right)^2$

Assumption B:

The step size satisfies

$$\eta \leq \min_{t\in[T]} \frac{T}{64(L+\mu)\|\mathbf{Z}_t^0\|_2^2} \tag{3.3}$$

Note that the step size depends on the initial estimate but remains constant throughout the training.

Assumption C:

The choices of η and ρ satisfy the inequality

$$\beta = \rho\left(1 - \frac{\eta L\mu}{2T\xi^2(L+\mu)}\right) < 1 \tag{3.4}$$

Moreover, Lemma A.3 shows that $\rho \leq \frac{1}{1-r}\left(1 + \frac{2\sqrt{s^\star}}{\sqrt{s-s^\star}}\right)$. β is the contraction coefficient of the distance metric $\text{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$ carried throughout iterations.
Assumption D:

The square of the statistical error satisfies the following inequality

$$\varepsilon_{stat}^2 \leq 2Tr^2 \frac{L\mu}{3\rho\eta(L+\mu)} \tag{3.5}$$

Assumption A ensures that the distance of initiate estimates and the global solutions are bounded within the ball of radius $\sqrt{2}r$. In general, to achieve the inequality, we need enough samples. The minimum sample size required could be estimated by computing concentration inequality and will be discussed in detail in Chapter 4. For most non-convex optimization problems, a careful choice of step size is required to constrain the estimates within the ball [14]. To obtain $\beta < 1$, we need to leverage between the choice of η and ρ. With r fixed, ρ increases as the sparsity level s increases. A small η will lead to the choice of a small ρ, which could be achieved by increasing the sparsity level s. Lastly, Assumption D ensures that the estimates do not escape the local region of the initial ball. Note that $\varepsilon_{stat} \leq \max_t \|\nabla \ell_t(\mathbf{\Sigma}_t^\star)\|_2 = \max_t \|\mathbf{\Sigma}_t^\star - \mathbf{S}_{N,t}\|$, which depends on the given samples. One sufficient condition to hold the inequality is $\frac{2TL\mu}{3\rho\eta(L+\mu)} \geq 1$, and Assumption A is satisfied. In general, $\frac{2TL\mu}{3\rho\eta(L+\mu)} \geq 1$ is immediately satisfied given Assumptions B and C hold. In the case where the inequality is invalid, we can increase the sample size N so that the statistical error ε_{stat} is small.

3.2 Main Results and Proof Sketches

Given that Assumptions A-D hold, we are ready to state the following result.

Theorem 3.1 (Linear Convergence)**.** *Assume that Assumptions A-D are satisfied. Furthermore, the step size η_V for $\mathbf{V}$ is $\frac{\eta}{T}$ and the step size η_A for $\mathbf{a}_t$ $\forall t \in [T]$ is η. With the fact that $\ell_{N,t}$ $\forall t \in [T]$ is μ-strongly convex and L smooth with respect to $\mathbf{\Sigma}_t$ $\forall t \in [T]$, then*

$$\beta \operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) + \rho\eta \frac{3(L+\mu)}{L\mu} \varepsilon_{stat}^2 \geq \operatorname{dist}^2(\mathbf{Z}^+, \mathbf{Z}^\star) \tag{3.6}$$

where $\mathbf{Z}$ denotes the estimation of the current step and $\mathbf{Z}^+$ denotes the estimation of the following step.

The Theorem 3.1 shows that with proper initialization the distance metric converges up to the statistical error scaled by the constant $\rho\eta\frac{3(L+\mu)}{L\mu}$. This also implies that $f_N(\mathbf{Z})$ converges close to $f_N(\mathbf{Z}^\star)$ as the following corollary states.

Corollary 3.1.1. *Assume the setting is the same as Theorem 3.1, then after i iterations, we have*

$$\sum_{t=1}^{T} \|\mathbf{\Sigma}_t^i - \mathbf{\Sigma}_t^\star\|_F^2 \leq 2Q^2 \left(\beta^i \operatorname{dist}(\mathbf{Z}^0, \mathbf{Z}^\star) + \rho\eta \frac{3(L+\mu)}{L\mu} \varepsilon_{stat}^2 \right) \tag{3.7}$$

where $Q = \max_t \frac{1}{2}\|\mathbf{Z}_t^i\|_2^2 + \|\mathbf{Z}_t^\star\|_2$, is bounded.

Corollary 3.1.1 shows the linear convergence up to the statistical error. The main proof consists of two steps. We begin with showing the following Lemma.

Lemma 3.1. *Suppose Assumption A holds. The step size η_V for $\mathbf{V}$ is $\frac{\eta}{T}$ and the step size η_A for $\mathbf{a}_t$ $\forall t \in [T]$ is η. Let $\eta \leq \min_t \frac{T}{32(L+\mu)\|\mathbf{Z}_t\|_2^2}$, then $\beta \operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) + \rho\eta\frac{3(L+\mu)}{L\mu}\varepsilon_{stat}^2 \geq \operatorname{dist}^2(\mathbf{Z}^+, \mathbf{Z}^\star)$.*

The proof is given in Appendix B.1. Note that the step size is dependent on the current estimate of $\mathbf{Z}$. The next step is to find a constant upper bound for η.

Lemma 3.2. *Given that the Assumption A holds, the step size in Assumption B satisfies $\eta \leq \min_t \frac{T}{32(L+\mu)\|\mathbf{Z}_t\|_2^2}$.*

The proof will be presented in Appendix B.2. Combining two lemmas, along with the Assumptions A-D, we can conclude the Theorem 3.1.

CHAPTER 4

SPECTRAL INITIALIZATION

In this chapter, we will discuss the underlying sample complexity to satisfy Assumption A. The main tools used in the proof are the Davis-Kahan sinΘ theorem [34, 38], Bernstein matrix concentration inequality, and the Corant-Fischer min-max theorem. We will first give a brief overview of these theorems and then proceed to state the main result.

4.1 Background Knowledge of Main Theorems

In this section, we will introduce two main theorems used in the proof of sample complexity: Davis-Kahan $\sin\Theta$ theorem [34, 38] and Corant-Fischer min-max theorem. The former theorem is used to bound the distance of the initiate $\mathbf{V}^0$ and the ground truth $\mathbf{V}^\star$. The later is used to bound the distance of initial $\mathbf{A}^0$ and the ground truth $\mathbf{A}^\star$.

4.1.1 Davis-Kahan $\sin\Theta$ Theorem

Consider two P-dimensional subspace spanned by columns of the orthogonal matrices $\widehat{\mathbf{V}}$ and $\mathbf{V}^\star$, respectively. Note that given a particular subspace, it could be represented by more than one set of orthonormal vectors. For example, consider the xy-plane of a three dimensional space, both the orthonormal set $\left\{[1\ 0\ 0]^\top, [0\ 1\ 0]^\top\right\}$ and $\left\{[\frac{1}{\sqrt{2}}\ \frac{1}{\sqrt{2}}\ 0]^\top, [\frac{-1}{\sqrt{2}}\ \frac{1}{\sqrt{2}}\ 0]^\top\right\}$ spanned the same subspace. To measure the distance between two subspace, on the other hand, we can measure the difference between the two projection operators to the subspace. Note that the projection operator to a subspace is unique and is the outer product of the orthogonal matrix. The distance of two projection operators is written as, $\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F$. To upper bound the distance $\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F$ we use the Davis-Kahan $\sin\Theta$ theorem stated as follows.

Lemma 4.1 (Davis-Kahan $\sin\Theta$ theorem, adapted from [38])**.**
Let $\mathbf{M}^\star = \sum_{t=1}^T \mathbf{\Sigma}_t^\star$ *and* $\mathbf{M} = \sum_{t=1}^T \mathbf{S}_{N,t}$*.* $\mathbf{V}^\star$ *is the matrix whose columns are top-K eigenvectors of* $\mathbf{M}^\star$*, and* $\widehat{\mathbf{V}}$ *is the matrix whose columns are the top-K eigenvectors of* $\widehat{\mathbf{M}}$*. The corresponding eigenvalues of* $\mathbf{V}^\star$ *are* $\|\tilde{\mathbf{a}}_1^\star\|_1 \geq \ldots \geq \|\tilde{\mathbf{a}}_K^\star\|_1 > \|\tilde{\mathbf{a}}_{K+1}^\star\|_1 \geq \ldots \geq \|\tilde{\mathbf{a}}_P^\star\|_1$*. Assume that the eigengap* $\|\tilde{\mathbf{a}}_K^\star\|_1 - \|\tilde{\mathbf{a}}_{K+1}^\star\|_1 = g > 0$ *is bounded away from zero. Then*

$$\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F \leq \frac{2\sqrt{K}}{g}\|\widehat{\mathbf{M}} - \mathbf{M}^\star\|_2 \tag{4.1}$$

Note that the theorem required an eigengap between the K-th and $K+1$-th component.

4.1.2 Courant-Fischer min-max Theorem

Courant-Fischer min-max theorem is a useful theorem when considering the perturbation of eigenvalues. It states that Rayleigh–Ritz quotient, $R_M(\mathbf{v}) = \frac{\langle \mathbf{Mv},\mathbf{v}\rangle}{\langle \mathbf{v},\mathbf{v}\rangle}$, $\forall \mathbf{v} \in \mathbb{R}^P/\{\mathbf{0}\}$ lies between the smallest eigenvalue and the largest eigenvalue of $\mathbf{M}$. Moreover, we could use the same scheme to show the following property.

Lemma 4.2 (Corollary of Courant-Fischer min-max Theorem)**.** *Let* $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K] \in \mathbb{R}^{P\times K}$*, where* $P \geq K$*, have orthonormal columns. Assume* $\mu_k = \mathbf{v}_k^\top \mathbf{\Sigma}\mathbf{v}_k \quad \forall k \in [K]$ *and* $\mu_1 \geq \ldots \geq \mu_K$*. Let eigenvalues of* $\mathbf{\Sigma}$ *be* $\lambda_1 \geq \ldots \geq \lambda_K$ *, then* $\lambda_i \geq \mu_i \quad i = 1, 2, \ldots, K$*.*

4.2 Main Results and Proof Sketches

From Assumption A, we know that the initialization distance is bounded by $2Tr^2$. It suffices to show that $\|\mathbf{V}^0 - \mathbf{V}^\star\mathbf{R}_t\|_F^2 \leq r^2 \ \forall t \in [T]$ and $\|\operatorname{diag}(\mathbf{a}_t^0) - \mathbf{R}^\top \operatorname{dist}(\mathbf{a}_t^\star)\mathbf{R}\|_F^2 \leq r^2 \ \forall t \in [T]$. Then, we use this idea to show the Theorem 4.1.

Theorem 4.1 (Sample Bound of Spectral Initialization)**.** *Consider* N *independent zero mean samples* $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)} \in \mathbb{R}^{P\times T}$*. Suppose that* $\|\mathbf{x}_t^{(n)}\|_2 \leq c \ \forall n \in [N], t \in [T]$*. Let* $\mathbf{M} = \sum_{t=1}^T \mathbf{S}_{N,t}$ *and* $\mathbf{M}^\star = \sum_{t=1}^T \mathbf{\Sigma}_t^\star$*. Assume columns of* $\mathbf{V}^\star \in \mathbb{R}^{P\times K}$ *are top-K eigenvectors of* $\mathbf{M}^\star$ *with corresponding* *eigenvalues* $\|\tilde{\mathbf{a}}_1^\star\|_1 \geq \ldots \geq \|\tilde{\mathbf{a}}_K^\star\|_1$ *and the eigengap of* K *and* $K+1$ *components is* $\|\tilde{\mathbf{a}}_K^\star\|_1 - \|\tilde{\mathbf{a}}_{K+1}^\star\|_1 = g > 0$. *Let* $\mathbf{Z}^{0\top} = [\mathbf{V}^{0\top}\mathbf{A}^0]^\top$ *be the matrix obtained via initialization method and* $\zeta = \frac{8KT}{g^2}\sum_{t=1}^T \|\mathbf{\Sigma}_t^\star\|_2^2 + 2K$. *Then,* $\forall \delta > 0$, $\operatorname{dist}^2(\mathbf{Z}^0, \mathbf{Z}^\star) \leq 2Tr^2$ *with probability at least* $1 - 2T\delta$, *if*

$$N \geq \max\left(-\frac{10Kc^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{gr}{\sqrt{5K}})}{(gr)^2}\log\frac{T\delta}{4P}, -\frac{2\zeta c^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{r}{\sqrt{\zeta}})}{r^2}\log\frac{\delta}{4P}\right) \quad (4.2)$$

The complete proof is presented C.1 and is consisted of three steps. The first step is to bound $\|\mathbf{V}^0 - \mathbf{V}^\star\mathbf{R}\|_F$ in terms of $\|\mathbf{M}^0 - \mathbf{M}^\star\|_2$ by Davis-Kahan sinΘ theorem [34]. The second step is to use Courant-Fischer min-max theorem to bound $\sum_{t=1}^T \|\operatorname{diag}(\mathbf{a}_t^0) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F$ in terms of $\sum_{t=1}^T \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2$. Finally, we can estimate the sample complexity by applying Bernstein matrix concentration inequality to $\|\mathbf{M}^0 - \mathbf{M}^\star\|_2$ and $\sum_{t=1}^T \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2$.

Step 1: Recall that our estimates of $\{\mathbf{\Sigma}_t\}_{t=1}^T$ share same $\mathbf{V}$ and columns of the initiate estimate $\mathbf{V}^0$ are orthogonal by construction.[1] Consequently, we can look at the distance $\|\mathbf{V}^0 - \mathbf{V}^\star\mathbf{R}\|_F$ as the distance between the two eigensubspaces: the eigensubspace spanned by $\mathbf{V}^0$ and the eigensubspace spanned by $\mathbf{V}^\star$. To upper bound $\|\mathbf{V}^0 - \mathbf{V}^\star\mathbf{R}\|_F$, we can first apply Davis-Kahan sinΘ theorem [34] to upper bound $\|\mathbf{V}^0\mathbf{V}^{0\top} - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F^2$. Then we could apply Lemma 5.4 in [39] to obtain the following bound $\min_{\mathbf{R}} \|\mathbf{V}^0 - \mathbf{V}\mathbf{R}\|_F^2 \leq \frac{1}{2(\sqrt{2}-1)}\|\mathbf{V}^0\mathbf{V}^{0\top} - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F^2$, where $\mathbf{V}^0$ is the initial estimate of $\mathbf{V}$ by spectral initialization. Of particular note is that the theorem required an eigengap $g > 0$ between the K-th and $K+1$-th component, which is a general assumption in retrieval of maximum eigenvectors problems [40].

Step 2: Now, we proceed to bound $\sum_{t=1}^T \|\operatorname{diag}(\mathbf{a}_t^0) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|$ with the Courant-Fischer min-max theorem. The idea of the proof is that the minimum $\|\operatorname{diag}(\mathbf{a}_t^0) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F$ $\forall t \in [T]$ over all possible permutation matrices is upper bounded by the case 1 that entries of $\mathbf{a}_t^\star$ $\forall t \in [T]$ are permuted to match the order of $\mathbf{a}_t^0$ $\forall t \in [T]$ in magnitude. Then, we can further apply the Courant-Fischer min-max theorem to upper bound the case 1. We begin with a warm-up example by considering the case of single component and then generalize it to multiple components.

Proposition 4.1 (Bound on perturbed eigenvalue mismatch). *Let* $\mathbf{v} \in \mathbb{R}^P$, $\lambda_i(\cdot)$ *denotes the i-th largest eigenvalue,* $\mathbf{\Sigma}$ *be a* $P \times P$ *real symmetric positive*

[1] Note that columns of $\mathbf{V}$ refined by Algorithm 2 are not necessary orthogonal but remain unit norm.

semi-definite matrix, and $\widetilde{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{E}$*, where* $\mathbf{E}$ *is a symmetric positive semi-definite perturbed matrix. Then*

$$|\mathbf{v}^\top \widetilde{\mathbf{\Sigma}} \mathbf{v} - \lambda_i(\mathbf{\Sigma})|^2 \leq 2 \left(\|\mathbf{\Sigma}\|_F\|^2 \mathbf{v}\mathbf{v}^\top - \mathbf{v}_i^\star \mathbf{v}_i^{\star\top}\|_2^2 + \lambda_1^2(\mathbf{E}) \right) \tag{4.3}$$

Proof. Let $\mathbf{v}_i^\star$ be the i-th eigenvector of $\mathbf{\Sigma}$ corresponding to i-th largest eigenvalue. Then

$$\begin{aligned} |\mathbf{v}^\top \widetilde{\mathbf{\Sigma}} \mathbf{v} - \mathbf{v}_i^{\star\top} \mathbf{\Sigma} \mathbf{v}_i^\star|^2 &= |\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} - \mathbf{v}_i^{\star\top} \mathbf{\Sigma} \mathbf{v}_i^\star + \mathbf{v}^\top \mathbf{E} \mathbf{v}|^2 \\ &= \left| \mathrm{Tr} \left(\mathbf{\Sigma} \mathbf{v}\mathbf{v}^\top - \mathbf{\Sigma} \mathbf{v}_i^\star \mathbf{v}_i^{\star\top} + \mathbf{E}\mathbf{v}\mathbf{v}^\top \right) \right|^2 \\ &\leq 2 \left(\left| \langle \mathbf{\Sigma}, \mathbf{v}\mathbf{v}^\top - \mathbf{v}_i^\star \mathbf{v}_i^{\star\top} \rangle \right|^2 + \lambda_1^2(\mathbf{E}) \right) \\ &\leq 2 \left(\|\mathbf{\Sigma}\|_F^2 \|\mathbf{v}\mathbf{v}^\top - \mathbf{v}_i^\star \mathbf{v}_i^{\star\top}\|_2^2 + \lambda_1^2(\mathbf{E}) \right) \end{aligned} \tag{4.4}$$

□

Lemma 4.3 (Bound on perturbed eigenvalues mismatch)**.** *Let* $\mathbf{\Sigma}$ *be a positive semi-definite symmetric matrix . Let* $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_1, \ldots, \mathbf{v}_K^0] \in \mathbb{R}^{P \times K}$ *has orthonormal columns and* $\mu_k = \mathbf{v}_k^\top \mathbf{\Sigma}^\star \mathbf{v}_k \quad \forall i \in [K]$*. Without loss of generosity, assume* $\mu_1 \geq \ldots \geq \mu_K$*.* $\mathbf{\Sigma}^\star$ *has eigenvalues* $\lambda_1^\star > \lambda_2^\star \geq \ldots \geq \lambda_K^\star$ *with corresponding eigenvectors* $\mathbf{v}_1^\star, \ldots, \mathbf{v}_K^\star$*. Let* $\mathbf{\Sigma} = \mathbf{\Sigma}^\star + \mathbf{E}$*, where* $\mathbf{E}$ *is a symmetric perturbation matrix.* $\lambda_i(\cdot)$ *denotes the i-th largest eigenvalue of the matrix, and* $\sigma_{\mathbf{R}}(\cdot)$ *represents the permutation function belonging to the set* $\mathcal{S}(K)$ *with corresponding permutation matrix* $\mathbf{R} \in \mathcal{P}(K)$*, then*

$$\begin{aligned} \min_{\sigma_{\mathbf{R}}(\cdot) \in \mathcal{S}(K)} & \sum_{k=1}^{k} |\mathbf{v}_k^\top \mathbf{\Sigma} \mathbf{v}_k - \lambda_{\sigma_{\mathbf{R}}(k)}(\mathbf{\Sigma}^\star)|^2 \\ & \leq 2 \left(\|\mathbf{\Sigma}^\star\|_F^2 \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star \mathbf{V}^{\star\top}\|_2^2 + \sum_{k=1}^{K} \lambda_k^2(\mathbf{E}) \right) \end{aligned} \tag{4.5}$$

The complete proof will be presented in Appendix C.3. For each $t \in [T]$, we could write $\|\operatorname{diag}(\mathbf{a}_t^0) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{R}_t\|_F^2$ as $\sum_{k=1}^{k} |\mathbf{v}_k^{0\top} \mathbf{S}_{N,t} \mathbf{v}_k^0 - \lambda_{\sigma_{\mathbf{R}_t}(k)}(\mathbf{\Sigma}^\star)|^2$. Bounding Equation (4.5) and result of Step 1 by Bernstein matrix inequality, we complete the sketch of proof.

CHAPTER 5

EXPERIMENTS

We verify the proposed algorithm with both simulated and real data. In the simulation data, we evaluate the recovery with the distance metric $\mathrm{dist}(\mathbf{Z}, \mathbf{Z}^\star)$ and compare the average log-Euclidean metric with other methods. For real data, since the ground truth is unknown, we focus on the interpretability of the model. To this end, we think the motor task fMRI is a good fit because the onset activation of particular task could be served as the reference to evaluate activities of brain regions. We use the Matérn five-half kernel as the smoothing kernel for all the simulations and tasks on real data. Empirically, we find that tuning the length scale of the kernel is more effective than tuning the hyperparameter γ in terms of producing the smooth weights.

5.1 Simulations

We test the algorithm with a variety of temporal dynamics. We compare our algorithm with sliding window PCA (SWPCA), hidden Markov model (HMM), autoregressive HMM (ARHMM), and sparse dictionary learning [41]. Results are averaged over 20 trials. We generate samples from the Gaussian distribution: $\mathbf{x}_t^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t^\star + \sigma \mathbf{I})\ \forall n \in [N], t \in [T]$, where $\mathbf{\Sigma}_t^\star = \sum_{k=1}^K a_{k,t}^\star \mathbf{v}_k^\star \mathbf{v}_k^{\star\top}$ and $\sigma\mathbf{I}$ is the additive noise.

5.1.1 Simulation 1

The ground truth of $\{\tilde{\mathbf{a}}_k^\star\}_{k=1}^K$ and $\{\mathbf{v}_k^\star \mathbf{v}_k^{\star\top}\}_{k=1}^K$ are shown on the left of Figure 5.1. In Figure 5.1, we test our algorithm under noiseless setting and evaluate the recovery using the distance metric $\mathrm{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$. We use Matérn five-half kernel matrix with amplitude = 2 and length scale = 200. The sparsity level of the spatial components is $s = 7$ and the kernel norm of the

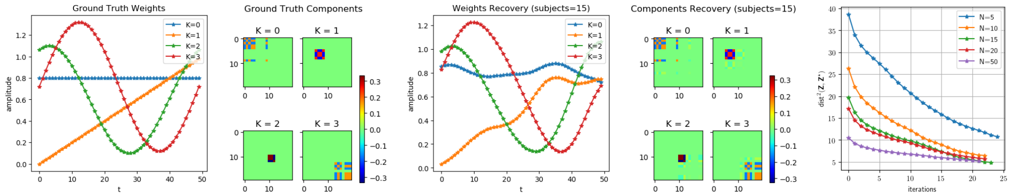

Figure 5.1: Covariance recovery, $K = 4, P = 20, T = 50$. **Left**: The simulation ground truth. **Center**: The recovery. **Right**: The simulation result with different number of subjects. As we increase the sample size, the initialization distance decreases.

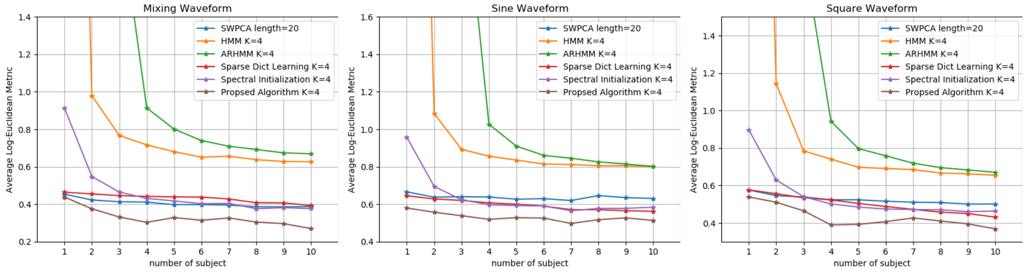

Figure 5.2: Tasks with different temporal structures, $K = 4, P = 16, T = 50$. Ground truth is shown in Figure 5.3. **Left**: The temporal components are the same as Figure 5.1. **Center**: The temporal components are sine waves. **Right**: The temporal components are square waves.

temporal components is no greater than 0.1. Furthermore, the learning rate is $1e - 4$. Here, we select $b = 0$ and $c = 4$.

5.1.2 Simulation 2

In Figure 5.2, we compare the algorithm with other methods with noisy data, where $\sigma = 0.5$. Since some of the baselines do not return factorized estimates, we cannot use the distance $\text{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$ for evaluation. Instead, we use the average log-Euclidean metric [42]: $\frac{1}{T}\sum_{t=1}^{T} \| \log(\mathbf{\Sigma}_t) - \log(\mathbf{\Sigma}_t^\star) \|_F$, where $\log(\mathbf{\Sigma}_t) = \mathbf{U}_t \log(\mathbf{\Lambda}_t) \mathbf{U}_t^\top$ and $\mathbf{U}$ is the eigenvector matrix, and $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix of $\mathbf{\Sigma}_t$. In practice, we truncate the zero eigenvalue and only compute the log of the non-zero eigenvalue so maintain stability of the evaluation.

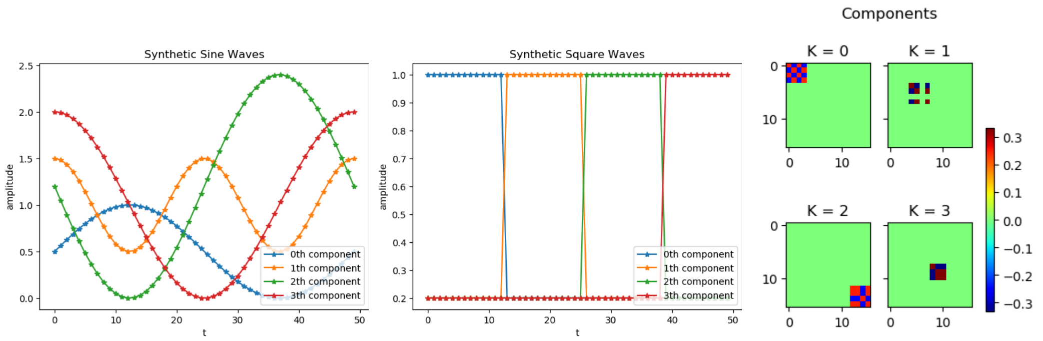

Figure 5.3: **Left**: Synthetic sine waves. **Center**: Synthetic square waves. **Right**: Synthetic components.

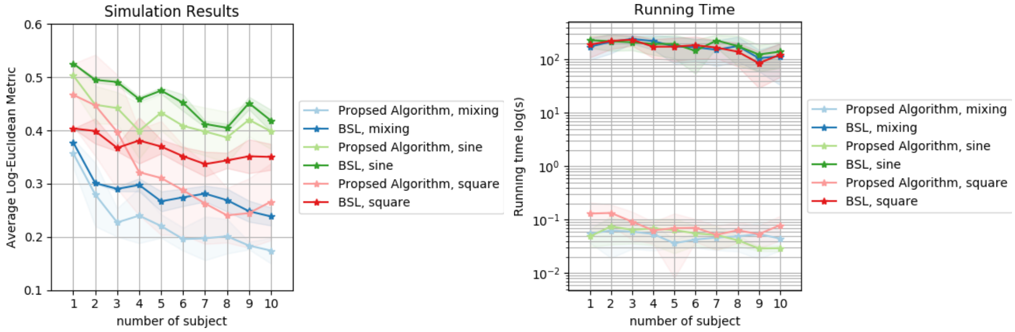

Figure 5.4: **left**: The average LERM of three tasks and the shaded area denotes the variance. $K = 4$, $P = 16$, $T = 50$, and the noise variance $\sigma = 0.05$. **Right**: The running time of the tasks averaged over 10 trials.

5.1.3 Simulation 3

We compare the proposed algorithm with the Bayesian structure learning [22] (BSL). Note that the model structures of both work are similar, but different in optimization schemes. Our work uses alternating projecting gradient descent, where BSL uses variational inference [43]. The simulation results indicate that two work yields comparable results, but our method is more computationally efficient than the counterpart. In the following experiment, we evaluate the average LERM distance of proposed method by taking the average of 20 trails. For the BSL method. We sample 20 samples from the posterior distribution and then compute the mean of the average LERM. Moreover, the evaluation of running time is computed by taking average of 10 trials on both methods. We test two models on three tasks, the mixing waveform, sine waveform, and the square waveform.

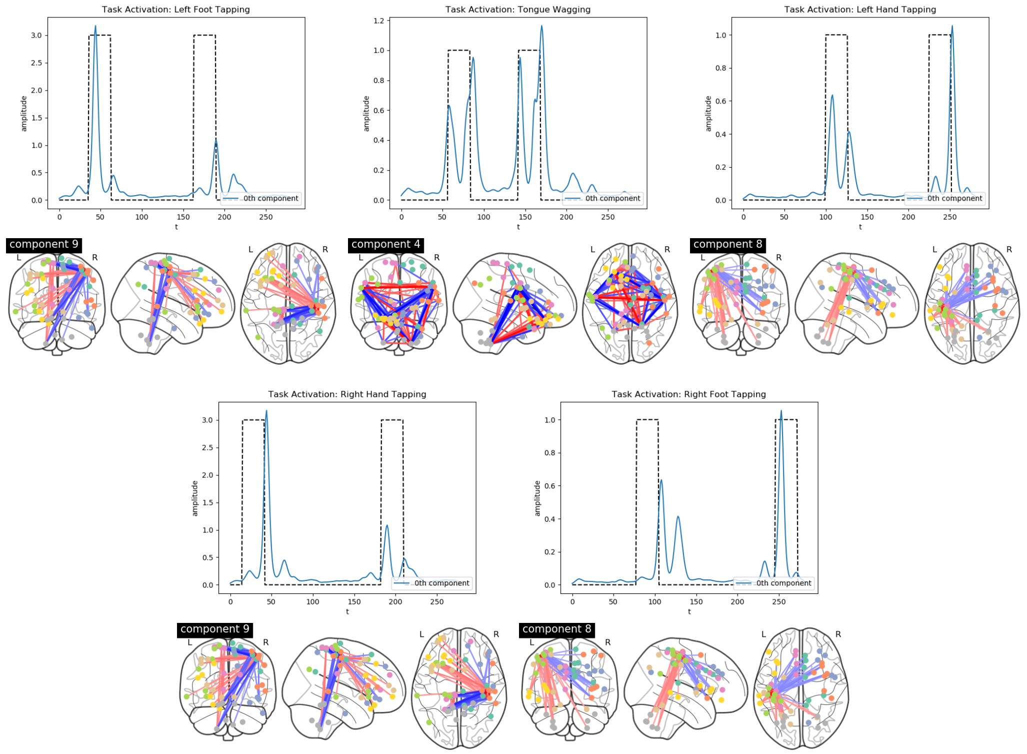

Figure 5.5: **Top (temporal diagram)**: The dashed line denotes the task activation time, and the blue line denotes the estimated temporal weights. **Bottom (connectome)**: The corresponding brain connectivity pattern of the task above. The red line denotes positive connectivity and the blue line denotes negative connectivity.

5.2 Experiment on Task fMRI

We use the Human Connectome Project (HCP) [44] motor task fMRI prepossessed data [45] to validate the algorithm. The data are consisted of five tasks: right hand tapping, left foot tapping, tongue wagging, right foot tapping, and left hand tapping. The training data compose of $N = 20$ subjects, $T = 284$, and $P = 375$. During the session, each task is activated twice, the goal is to analyze the responding brain region and the temporal fluctuation when certain task is activated.

In experiment, we choose $K = 15$ and the result is shown in Figure 5.5, where the sparsity level $s = 54$ and $\gamma = 1.0$. To interpret the model, we compute the correlation of each weight $\tilde{\mathbf{a}}_k, \forall k \in [K]$ with the onset task activation time (see the black dashed line in the top of Figure 5.5), and select the component that has highest correlation value compared with other tasks.

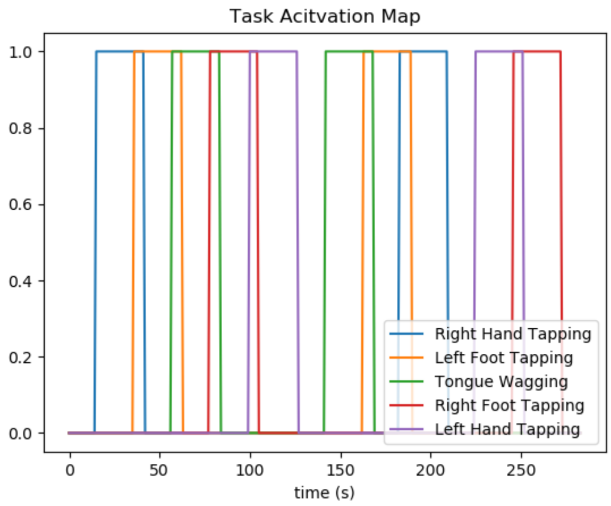

Figure 5.6: The activation map.

Table 5.1: The correlation of the components with task activation.

Task	Rank of the correlation (order from largest to smallest)														
Right Hand Tapping	9	0	6	5	3	2	14	13	12	11	10	7	1	4	8
Left Foot Tapping	9	4	6	2	14	13	12	11	10	3	1	0	5	7	8
Tongue Wagging	4	1	2	7	14	13	12	11	10	3	9	8	0	5	6
Right Foot Tapping	8	7	6	14	13	12	11	10	3	2	4	1	5	0	9
Left Hand Tapping	8	7	5	3	1	2	6	14	13	12	11	10	0	4	9

The result indicates that the proposed algorithm can separate and identify the components of each task. We train the HCP motor task dataset with 200 epochs, learning rate 1.0, and the Matérn five-half kernel matrix with amplitude 2 length scale 5. We then select the best matches component of each task by computing the Pearson correlation of the temporal components with the activation map of the tasks, shown in Figure 5.6. The result is shown in Table 5.1.

In most cases, the neural activities are combinations of multiple components rather than single components. Therefore, for each task, we select the top three components listed in Table 5.1 and plot the connectivity in Figure 5.7.

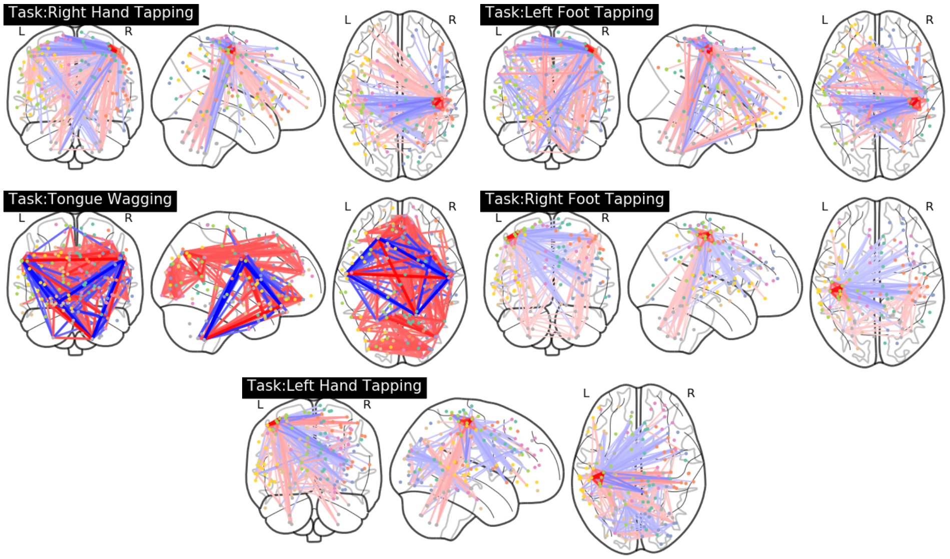

Figure 5.7: The covariance of each task with the highest three components listed in Table 5.1.

CHAPTER 6

DISCUSSION AND CONCLUSION

This thesis proposes a non-convex framework for estimating structured dynamic covariances. The sparse structure is imposed by iterative hard-thresholding, and the smooth structure is imposed by projection to the kernel space. We propose a novel spectral initialization scheme to aggregate the sample size under the assumption of shared spatial structures. While this method improves the initial estimates of spatial components, it does not benefit the initial estimated temporal components. Empirically, we find that good initial spatial estimates will lead to better estimates of temporal components. The underlying sample complexity of the spectral initialization is being shown as well. We consider the worst case of the sample complexity and do not take the estimate's structures into account. The sample complexity may be improved by considering the distance after projecting the initial estimates to the constraint sets $\mathcal{C}_{\mathbf{A}}$ and $\mathcal{C}_{\mathbf{V}}$. Additionally, we have shown, up to the statistical error, the algorithm converges linearly to the global optimum. Our analysis adopts from [14] and extends to estimating structured dynamic matrices. We utilize a new factorization scheme $\mathbf{V}\operatorname{diag}\mathbf{a}_t\mathbf{V}^\top$ to incorporate the temporal structure, which is different than [14] that uses the factorization scheme $\mathbf{U}\mathbf{U}^\top$. We verify the proposed algorithm on both simulations and real data. Our results outperform several methods and recover the temporal dynamics of simulated data. We further compare the Bayesian counterpart and shows that our algorithm has comparable performance with the Bayesian method but converges much faster. Our method converges in $O(10^{-1})$ seconds while the Bayesian method converges in $O(10^2)$ seconds. Finally, we use our model to interpret the temporal and spatial correlation between brain regions of motor task fMRI data. The result shows that our model is able to denoise the data and shows distinct activations corresponding to different tasks.

APPENDIX A

PROJECTION TO $\mathcal{C}_{\mathbf{V}}$ AND $\mathcal{C}_{\mathbf{A}}$

A.1 Projection to $\mathcal{C}_{\mathbf{V}}$

Let $\mathcal{C}_{\mathbf{V}} = \{\mathbf{x} \in \mathbb{R}^P : \|\mathbf{x}\|_0 \leq s, \|\mathbf{x}\|_2 = 1\}$. We want to solve the following problem

$$\arg\min_{\mathbf{x} \in \mathcal{C}_{\mathbf{V}}} \|\mathbf{v}_k - \mathbf{x}\|_2^2 \tag{A.1}$$

Let $\mathcal{S}(\mathbf{x}) = \{i : x_i \neq 0\}$ be the support of $\mathbf{x}$. Given a support $E \subset [P]$, let $\mathbf{x}_E$ be a vector whose i-th entry are $\begin{cases} x_i & i \in E \\ 0 & i \notin E \end{cases}$. The projection of $\mathbf{v}_k$ given a support E is

$$\begin{aligned} d(E) &= \min_{\mathbf{x}} \|\mathbf{v}_k - \mathbf{x}\|_2^2 \\ &\text{subject to } \mathcal{S}(\mathbf{x}) \subseteq E,\ \|\mathbf{x}\|_2 = 1 \end{aligned} \tag{A.2}$$

$$\begin{aligned} d(E) &= \min_{\mathbf{x}} \|\mathbf{v}_k\|_2^2 + \|\mathbf{x}\|_2^2 - 2\langle \mathbf{x}, \mathbf{v}_k \rangle \\ &= \|\mathbf{v}_k\|_2^2 + 1 - 2\max_{\mathbf{x}} \langle \mathbf{x}, \mathbf{v}_k \rangle \\ &= \|\mathbf{v}_k\|_2^2 + 1 - 2\|\mathbf{v}_{k,E}\|_2 \end{aligned} \tag{A.3}$$

Therefore, (A.1) is equivalent as

$$\arg\min_{E:|E| \leq s} d(E) = \arg\max_{E:|E| \leq s} \|\mathbf{v}_{k,E}\|_2 \tag{A.4}$$

This could be solved by finding the top-s entries of $\mathbf{v}_k$ in magnitude, which has computational complexity $O(P \log P)$, and then normalize to unit norm.

A.1.1 Expansion Coefficient of Projection to $\mathcal{C}_{\mathbf{V}}$

Lemma A.1. *Assume that* $\|\mathbf{v}-\mathbf{v}^{\star}\|_2^2 \leq r$,*where* $r<1$, $\|\mathbf{v}^{\star}\|_2 = 1$, $\|\mathbf{v}\|_2 \leq 1$. *Then* $h\|\mathbf{v}-\mathbf{v}^{\star}\|_2^2 \geq \|\frac{\mathbf{v}}{\|\mathbf{v}\|} - \mathbf{v}^{\star}\|_2^2$, *where* $h \leq \frac{1}{1-r}$.

Proof. By expanding $r^2 \geq \|\mathbf{v}-\mathbf{v}^{\star}\|_2^2$, and using the property $\|\mathbf{v}\|_2 \geq 1-r$, we can obtain $\mathbf{v}\cdot\mathbf{v}^{\star} \geq 1-r > 0$. Since $h\|\mathbf{v}-\mathbf{v}^{\star}\|_2^2 - \|\frac{\mathbf{v}}{\|\mathbf{v}\|_2} - \mathbf{v}^{\star}\|_2^2$ can be written as $(h\|\mathbf{v}\|_2^2 + h - 2) + 2\mathbf{v}\cdot\mathbf{v}^{\star}\left(\frac{1}{\|\mathbf{v}\|_2} - h\right)$, it suffices to show that $h\|\mathbf{v}\|_2^2 + h - 2 \geq 0$ and $\frac{1}{\|\mathbf{v}\|_2} - h \geq 0$.

1.

$$h \leq \frac{1}{\|\mathbf{v}\|_2} \leq \frac{1}{1-r} \tag{A.5}$$

2.

$$\begin{aligned} h\|\mathbf{v}\|_2^2 + h - 2 &\geq h(1-r)^2 + h - 2 \\ &\overset{(a)}{\geq} (1-r) + \frac{1}{1-r} - 2 = \frac{1}{1-r} - r - 1 > 0 \end{aligned} \tag{A.6}$$

(a) We can obtain the inequality by plugging Equation (A.5).

Therefore, we can conclude that $h \leq \frac{1}{1-r}$. □

Lemma A.2 (Lemma 4.1 in [46])**.** *Let* $\mathbf{v}^{\star} \in \mathbb{R}^P$ *be a sparse vector such that* $\|\mathbf{v}^{\star}\|_0 \leq s^{\star}$, *and* $\mathcal{H}_s(\cdot) : \mathbb{R}^P \to \mathbb{R}^P$ *be the hard thresholding operator. Given* $s > s^{\star}$, *for any vector* $\mathbf{v} \in \mathbb{R}^P$, *we have*

$$\|\mathcal{H}_s(\mathbf{v}) - \mathbf{v}^{\star}\|_2^2 \leq (1 + \frac{2\sqrt{s^{\star}}}{\sqrt{s-s^{\star}}})\|\mathbf{v}-\mathbf{v}^{\star}\|_2^2 \tag{A.7}$$

Lemma A.3. *Assume that* $\|\mathbf{v}-\mathbf{v}^{\star}\|_2 \leq r$ *and* $\|\mathbf{v}^{\star}\|_0 \leq s^{\star}$. *Let the projection operator to the set* $\mathcal{C}_{\mathbf{V}}$ *defined in Section A.1 be* $\Pi_{\mathcal{C}_{\mathbf{V}}} : \mathbb{R}^P \to \mathbb{R}^P$, *then*

$$\|\Pi_{\mathcal{C}_{\mathbf{V}}}(\mathbf{v}) - \mathbf{v}^{\star}\|_2^2 \leq \rho\|\mathbf{v}-\mathbf{v}^{\star}\|_2^2 \tag{A.8}$$

where $\rho \leq \frac{1}{1-r}\left(1 + \frac{2\sqrt{s^{\star}}}{\sqrt{s-s^{\star}}}\right)$.

Proof. Combining Lemma A.1 and A.2, we arrive at the above conclusion. □

A.2 Projection to $\mathcal{C}_{\mathbf{A}}$

To project to the convex set $\mathcal{C}_{\mathbf{A}}$, we use alternating projection methods.

Step 1: Project $\tilde{\mathbf{a}}_k$ to the hypercube $[b, c]^T$

Step 2: Project to the ellipsoid by solving the following constrained optimization problem.

$$\begin{aligned} &\arg\min_{\mathbf{y}} \quad \|\tilde{\mathbf{a}}_k - \mathbf{y}\|_2^2 \\ &\text{subject to} \\ &\qquad\qquad \mathbf{y}^\top \mathbf{G}^{-1} \mathbf{y} \leq \gamma \end{aligned} \tag{A.9}$$

Let $\mathbf{G}^{-1} = \mathbf{P}\mathbf{\Sigma}\mathbf{P}^T$, where $\mathbf{P}$ is unitary matrix, $\tilde{\mathbf{u}}_k = \mathbf{P}^\top \tilde{\mathbf{a}}_k$, and $\mathbf{z} = \mathbf{P}^\top \mathbf{y}$. Since $\mathbf{P}$ is unitary, $\|\tilde{\mathbf{a}}_k - \mathbf{y}\|_2^2 = \|\mathbf{P}^\top(\tilde{\mathbf{a}}_k - \mathbf{y})\|_2^2 = \|\tilde{\mathbf{u}}_k - \mathbf{z}\|_2^2$. This optimization function is equivalent to the following

$$\begin{aligned} &\arg\min_{\mathbf{z}} \quad \|\tilde{\mathbf{u}}_k - \mathbf{z}\|_2^2 \\ &\text{subject to} \\ &\qquad\qquad \mathbf{z}^\top \mathbf{\Sigma} \mathbf{z} \leq \gamma \end{aligned} \tag{A.10}$$

Now, let $\mathbf{w} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{z}$. Then, we could rewrite the objective function (A.10) as

$$\begin{aligned} &\arg\min_{\mathbf{w}} \quad \|\tilde{\mathbf{u}}_k - \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{w}\|_2^2 \\ &\text{subject to} \\ &\qquad\qquad \mathbf{w}^\top \mathbf{w} \leq \gamma \end{aligned} \tag{A.11}$$

Let the corresponding Lagrangian function be $\mathcal{L}(\mathbf{w}, \lambda) = \|\tilde{\mathbf{u}}_k - \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{w}\|_2^2 + \lambda(\mathbf{w}^\top\mathbf{w} - \gamma)$. $\nabla_{\mathbf{w}}\mathcal{L} = 0$ implies that $\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-\frac{1}{2}}\tilde{\mathbf{u}}_k$. By KKT, if $\tilde{\mathbf{a}}_k^\top \mathbf{G}^{-1}\tilde{\mathbf{a}}_k \leq \gamma$, then $\mathbf{y}^* = \tilde{\mathbf{a}}_k$. Otherwise, $\mathbf{w}^\top\mathbf{w} = \gamma$. This implies that $\sum_t \frac{\tilde{u}_{k,t}^2 \sigma_i}{(1+\lambda\sigma_i)^2} = \gamma$, where σ_i is the i-th diagonal entry of the matrix $\mathbf{\Sigma}$. Using the second-order Taylor expansion, we could write $\sum_t \frac{\tilde{u}_{k,t}^2 \sigma_i}{(1+\lambda\sigma_i)^2} = \gamma$ as

$$3\lambda^2 \sum \sigma_t^3 \tilde{u}_{k,t}^2 - 2\lambda \sum \sigma_t^2 (\tilde{u}_{k,t})^2 + \sum \sigma_t (\tilde{u}_{k,t})^2 - \gamma = 0 \tag{A.12}$$

Then, finding λ is equivalent as finding the roots of the above polynomial function. Then plug λ into $\mathbf{y} = \mathbf{P}\mathbf{\Sigma}^{-1}(\lambda\mathbf{I} + \mathbf{\Sigma}^{-1})^{-1}\mathbf{P}^\top\tilde{\mathbf{a}}_k$.

Step 3: If the projection does not satisfy the box constraints, then repeat Step 1 and Step 2 until a feasible point is found.

In practice, we find that most of the time, one iteration of alternating projections already satisfies both constraint.

APPENDIX B

PROOF OF THEOREM 3.1

B.1 Proof of Lemma 3.1

Preliminaries

1.

$$\begin{aligned}\nabla_V f_N(\mathbf{V}, \mathbf{A}) &= \frac{2}{T} \sum_{t=1}^{T} \nabla \ell_{N,t}(\mathbf{V} \operatorname{diag}(\mathbf{a}_t) \mathbf{V}^\top) \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \\ &= \frac{2}{T} \sum_{t=1}^{T} \nabla \ell_{N,t}(\mathbf{\Sigma}_t) \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \qquad \text{(B.1)}\end{aligned}$$

2.

$$\langle \nabla_V f_N(\mathbf{V}, \mathbf{A}), \mathbf{V} - \mathbf{V}^\star \mathbf{R} \rangle = \frac{2}{T} \sum_{t=1}^{T} \langle \nabla \ell_{N,t}(\mathbf{\Sigma}_t) \mathbf{V} \operatorname{diag}(\mathbf{a}_t), \mathbf{V} - \mathbf{V}^\star \mathbf{R} \rangle$$

$$= \frac{2}{T} \sum_{t=1}^{T} \langle \nabla \ell_{N,t}(\mathbf{\Sigma}_t), \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \mathbf{V}^\top - \mathbf{V}^\star \mathbf{R} \operatorname{diag}(\mathbf{a}_t) \mathbf{V}^\top \rangle \qquad \text{(B.2)}$$

3.

$$\nabla_{a_t} f_N(\mathbf{V}, \mathbf{A}) = \frac{1}{T} \widehat{\operatorname{diag}} \left(\mathbf{V}^\top \nabla \ell_{N,t}(\mathbf{\Sigma}_t) \mathbf{V} \right) \in \mathbb{R}^K \quad \forall t \in [T] \qquad \text{(B.3)}$$

$\widehat{\operatorname{diag}} : \mathbb{R}^{K \times K} \to \mathbb{R}^K$ extracts the diagonal entries and vectorize them

4.

$$\begin{aligned}&\langle \operatorname{diag}(\nabla_{a_t} \ell_{N,t}), \operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{R} \rangle \\ &= \frac{1}{T} \langle \mathbf{V}^\top \nabla \ell_{N,t}(\mathbf{\Sigma}_t) \mathbf{V}, \operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{R} \rangle\end{aligned}$$

$$= \frac{1}{T}\langle \nabla \ell_{N,t}(\mathbf{\Sigma}_t), \mathbf{V}\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^\top - \mathbf{V}\mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\mathbf{V}^\top\rangle \quad \forall t \in [T] \tag{B.4}$$

Proof. Let

$$\mathbf{Z} = \begin{bmatrix} \mathbf{V} \\ \mathbf{A}^\top \end{bmatrix} \qquad \mathbf{Z}^+ = \begin{bmatrix} \mathbf{V}^+ \\ \mathbf{A}^{+\top} \end{bmatrix} \qquad \mathbf{Z}^\star = \begin{bmatrix} \mathbf{V}^\star \\ \mathbf{A}^{\star\top} \end{bmatrix}$$

$$\begin{aligned}
&\operatorname{dist}^2(\mathbf{Z}^+, \mathbf{Z}^\star) \\
&= \sum_{t=1}^{T} \min_{\mathbf{R}_t^+} \|\mathbf{V}^+ - \mathbf{V}^\star \mathbf{R}_t^{+\top}\|_F^2 + \|\operatorname{diag}(\mathbf{a}_t^+) - \mathbf{R}_t^{+\top}\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t^+\|_F^2 \\
&\overset{(a)}{\leq} \sum_{t=1}^{T} \|\mathbf{V}^+ - \mathbf{V}^\star\mathbf{R}_t\|_F^2 + \|\operatorname{diag}(\mathbf{a}_t^+) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2 \\
&= \sum_{t=1}^{T} \|\Pi_{\mathcal{C}_\mathbf{V}}(\mathbf{V} - \eta_V \nabla_V f_N) - \mathbf{V}^\star \mathbf{R}_t\|_F^2 \\
&+ \sum_{t=1}^{T} \|\operatorname{diag}\left(\Pi_{\mathcal{C}_\mathbf{A}}(\mathbf{a}_t - \eta_A \nabla_{a_t}\ell_{N,t})\right) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2 \\
&\overset{(b)}{\leq} \rho \sum_{t=1}^{T} \left\{ \|\mathbf{V} - \eta_V \nabla_V f_N - \mathbf{V}^\star\mathbf{R}_t\|_F^2 \right. \\
&\left. + \|\operatorname{diag}(\mathbf{a}_t) - \operatorname{diag}(\eta_A \nabla_{a_t}\ell_{N,t}) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2 \right\} \\
&\overset{(c)}{=} \rho \operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) + \underbrace{\frac{\eta^2\rho}{T^2}\sum_{t=1}^{T}\|\nabla_v f_N\|_F^2}_{B1} + \underbrace{\eta^2 \rho \sum_{t=1}^{T} \|\operatorname{diag}(\nabla_{a_t}\ell_{N,t})\|_F^2}_{B2} \\
&\underbrace{- \frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla_v f_N, \mathbf{V} - \mathbf{V}\mathbf{R}_t\rangle}_{A1} \\
&\underbrace{- 2\rho\eta \sum_{t=1}^{T} \langle \operatorname{diag}(\nabla_{a_t}\ell_{N,t}), \operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\rangle}_{A2}
\end{aligned} \tag{B.5}$$

(a) $\mathbf{R}_t$ is the optimal solution in previous step. (b) Assume that the $\mathbf{V}^\star$ lies in the constraint set $\mathcal{C}_\mathbf{V}$, by the Lemma A.2, we could obtain the above upper bound. By the non-expansive property of projection to convex sets we could obtain the upper bound of the second term. Since we can select $\rho \geq 1$, we could further multiply the second term by ρ. (c) By the assumption that $\eta_V = \frac{\eta}{T}$ and $\eta_A = \eta$, we get the above equality.

$$\begin{aligned}
B1 &= \frac{\eta^2 \rho}{T^2} \sum_{t=1}^{T} \left\| \frac{2}{T} \sum_{t'=1}^{T} \nabla \ell_{N,t'}(\mathbf{\Sigma}_{t'}) \mathbf{V} \operatorname{diag}(\mathbf{a}_{t'}) \right\|_F^2 \\
&= \frac{\eta^2 \rho}{T} \left\| \frac{2}{T} \sum_{t'=1}^{T} \nabla \ell_{N,t'}(\mathbf{\Sigma}_{t'}) \mathbf{V} \operatorname{diag}(\mathbf{a}_{t'}) \right\|_F^2 \\
&\overset{(a)}{\leq} \frac{4\rho\eta^2}{T^2} \sum_{t=1}^{T} \| \nabla_{N,t}(\mathbf{\Sigma}_t) \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \|_F^2 \\
&= \frac{4\rho\eta^2}{T^2} \sum_{t=1}^{T} \| \left(\nabla \ell_{N,t}(\mathbf{\Sigma}_t) - \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) + \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) \right) \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \|_F^2 \\
&\leq \frac{4\rho\eta^2}{T^2} \sum_{t=1}^{T} \| \left(\nabla \ell_{N,t}(\mathbf{\Sigma}_t) - \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) + \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) \right) \|_F^2 \| \mathbf{V} \operatorname{diag}(\mathbf{a}_t) \|_2^2 \\
&\overset{(b)}{\leq} \frac{8\rho\eta^2}{T^2} \sum_{t=1}^{T} \left(\| \nabla \ell_{N,t}(\mathbf{\Sigma}_t) - \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) \|_F^2 + \varepsilon_{stat}^2 \right) \| \mathbf{V} \|_2^2 \| \operatorname{diag}(\mathbf{a}_t) \|_2^2 \quad \text{(B.6)}
\end{aligned}$$

(a) The ineqaulity is obtained by Cauchy-Schwarz inequality. (b) We arrive at the following inequality by the fact that $\|\nabla \ell_{N,t}(\mathbf{\Sigma})_t\|_F \leq \varepsilon_{stat}$.

$$\begin{aligned}
B2 &= \frac{\rho\eta^2}{T^2} \sum_{t=1}^{T} \left\| \widehat{\operatorname{diag}}(\mathbf{V}^\top \nabla_{N,t}(\mathbf{\Sigma}_t) \mathbf{V}) \right\|_2^2 \\
&\leq \frac{\rho\eta^2}{T^2} \sum_{t=1}^{T} \| \mathbf{V}^\top \nabla \ell_{N,t}(\mathbf{\Sigma}_t) \mathbf{V} \|_F^2 \\
&\leq \frac{\rho\eta^2}{T^2} \sum_{t=1}^{T} \| \nabla \ell_{N,t}(\mathbf{\Sigma}_t) \|_F^2 \| \mathbf{V}\mathbf{V}^\top \|_2^2 \\
&= \frac{\rho\eta^2}{T^2} \sum_{t=1}^{T} \| \nabla \ell_{N,t}(\mathbf{\Sigma}_t) - \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) + \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) \|_F^2 \| \mathbf{V} \|_2^2 \| \mathbf{V} \|_2^2 \\
&\leq \frac{2\rho\eta^2}{T^2} \sum_{t=1}^{T} \left(\| \nabla \ell_{N,t}(\mathbf{\Sigma}_t) - \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star) \|_F^2 + \varepsilon_{stat}^2 \right) \| \mathbf{V} \|_2^2 \| \mathbf{V} \|_2^2 \qquad \text{(B.7)}
\end{aligned}$$

$$B = B1 + B2$$

$$\overset{(a)}{\le} \frac{4\rho\eta^2}{T^2}\sum_{t=1}^{T}\left(\|\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t)-\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t^\star)\|_F^2+\varepsilon_{stat}^2\right)\left(4\|\operatorname{diag}(\mathbf{a}_t)\|_2^2+\|\mathbf{V}\|_2^2\right)$$

$$\overset{(b)}{\le} \frac{16\rho\eta^2}{T^2}\sum_{t=1}^{T}\left(\|\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t)-\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t^\star)\|_F^2+\varepsilon_{stat}^2\right)\left(\|\operatorname{diag}(\mathbf{a}_t)\|_2^2+\|\mathbf{V}\|_2^2\right)$$

$$= \frac{16\rho\eta^2}{T^2}\sum_{t=1}^{T}\left(\|\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t)-\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t^\star)\|_F^2+\varepsilon_{stat}^2\right)\|\mathbf{Z}_t\|_2^2$$

(a) This is because $\|\mathbf{V}\|_2 = \|\mathbf{V}-\mathbf{V}^\star+\mathbf{V}^\star\|_2 \le \|\mathbf{V}-\mathbf{V}^\star\|_F+\|\mathbf{V}^\star\|_2 \le r+1 \le 2$.
(b) We can upper bound the equation by the fact that $\rho \ge 1$.

$$A = A1 + A2$$

$$\overset{(a)}{=} \frac{4\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t)\mathbf{V}\operatorname{diag}(\mathbf{a}_t), \mathbf{V}-\mathbf{V}^\star\mathbf{R}_t\rangle$$

$$+ \frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle \mathbf{V}^\top\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t)\mathbf{V}, \operatorname{diag}(\mathbf{a}_t)-\mathbf{R}_t^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\rangle$$

$$= \frac{4\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t), \mathbf{V}\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^\top-\mathbf{V}^\star\mathbf{R}_t\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^\top\rangle$$

$$+ \frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t), \mathbf{V}\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^\top-\mathbf{V}\mathbf{R}_t^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\mathbf{V}^\top\rangle$$

$$= \underbrace{\frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t)-\nabla\ell_{N,t}(\boldsymbol{\Sigma}_t^\star), \boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_t^\star\rangle}_{(A11)}$$

$$+ \underbrace{\frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t^\star), \boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_t^\star\rangle}_{(A12)}$$

$$+ \underbrace{\frac{4\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t), \overbrace{(\mathbf{V}-\mathbf{V}^\star\mathbf{R}_t)}^{\Delta\mathbf{V}}\overbrace{(\operatorname{diag}(\mathbf{a}_t)-\mathbf{R}_t^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t)}^{\Delta\mathbf{a}_t}\mathbf{V}^\top\rangle}_{(A13)}$$

$$+ \underbrace{\frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla\ell_{N,t}(\boldsymbol{\Sigma}_t), \overbrace{(\mathbf{V}-\mathbf{V}^\star\mathbf{R}_t)}^{\Delta\mathbf{V}}\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\overbrace{(\mathbf{V}-\mathbf{V}^\star\mathbf{R}_t)^\top}^{\Delta\mathbf{V}^\top}\rangle}_{(A14)} \tag{B.8}$$

(a) $A1 = \frac{2\rho\eta}{T}\sum_{t=1}^{T}\langle\frac{2}{T}\sum_{t'=1}^{T}\nabla_{N,t'}(\boldsymbol{\Sigma}_{t'})\mathbf{V}\operatorname{diag}(\mathbf{a}_{t'}), \mathbf{V}-\mathbf{V}^\star\mathbf{R}_t\rangle$. Without loss of generality, we assume that $\{\mathbf{R}_t\}_{t\in T}$ are the same across t, then with little abuse of notation $A1 = \frac{4\rho\eta}{T}\sum_{t=1}^{T}\langle \nabla_{N,t}(\mathbf{\Sigma}_t)\mathbf{V}\operatorname{diag}(\mathbf{a}_t), \mathbf{V}-\mathbf{V}^\star\mathbf{R}_t\rangle$.

By Theorem 2.1.12 in [47], we could obtain the following lower bound of A11

$$\mathrm{A11} \geq \frac{2\rho\eta}{T}\sum_{t=1}^{T}\left(\frac{L\mu}{L+\mu}\|\mathbf{\Sigma}_t-\mathbf{\Sigma}_t^\star\|_F^2+\frac{1}{L+\mu}\|\nabla\ell_{N,t}(\mathbf{\Sigma}_t)-\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|_F^2\right) \tag{B.9}$$

$$\begin{aligned}\mathrm{A12} &\geq -\frac{2\rho\eta}{T}\sum_{t=1}^{T}|\langle\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star), \mathbf{\Sigma}_t-\mathbf{\Sigma}_t^\star\rangle| \\ &\geq -\frac{2\rho\eta}{T}\sum_{t=1}^{T}\varepsilon_{stat}\|\mathbf{\Sigma}_t-\mathbf{\Sigma}_t^\star\|_F \geq -\frac{2\rho\eta}{T}\sum_{t=1}^{T}\frac{\varepsilon_{stat}^2}{2e_1}+\frac{e_1}{2}\|\mathbf{\Sigma}_t-\mathbf{\Sigma}_t^\star\|_F^2\end{aligned} \tag{B.10}$$

The inequlity is obtained by Young's inequality, where $e_1 > 0$.

$$\begin{aligned}A13 \geq & -\frac{4\rho\eta}{T}\sum_{t=1}^{T}|\langle\nabla\ell_{N,t}(\mathbf{\Sigma}_t), \Delta\mathbf{V}\Delta\mathbf{a}_t\mathbf{V}^\top\rangle| \\ \geq & -\frac{4\rho\eta}{T}\sum_{t=1}^{T}\big\{|\langle\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\mathbf{V}\Delta\mathbf{a}_t, \Delta\mathbf{V}\rangle| \\ & +|\langle\left(\nabla\ell_{N,t}(\mathbf{\Sigma}_t)-\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\right)\mathbf{V}\Delta\mathbf{a}_t, \Delta\mathbf{V}\rangle|\big\} \\ \geq & -\frac{4\rho\eta}{T}\sum_{t=1}^{T}\|\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\mathbf{V}\Delta\mathbf{a}_t\|_F\|\Delta\mathbf{V}\|_F \\ & +\|\left(\nabla\ell_{N,t}(\mathbf{\Sigma}_t)-\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\right)\mathbf{V}\Delta\mathbf{a}_t\|_F\|\Delta\mathbf{V}\|_F \\ \geq & -\frac{4\rho\eta}{T}\sum_{t=1}^{T}(\varepsilon_{stat}+\|\nabla\ell_{N,t}(\mathbf{\Sigma}_t)-\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|_F)\|\mathbf{V}\|_2\|\Delta\mathbf{a}_t\|_F\|\Delta\mathbf{V}\|_F \\ \overset{(a)}{\geq} & -\frac{4\rho\eta}{T}\sum_{t=1}^{T}(\varepsilon_{stat}+\|\nabla\ell_{N,t}(\mathbf{\Sigma}_t)-\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|)\mathrm{d}^2(\mathbf{Z}_t,\mathbf{Z}_t^\star) \\ \overset{(b)}{\geq} & -\frac{4\rho\eta}{T}\sum_{t=1}^{T}\frac{1}{e_2}\left(\varepsilon_{stat}^2+\|\nabla\ell_{N,t}(\mathbf{\Sigma}_t)-\nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|_F^2\right) \\ & -\frac{4\rho\eta}{T}e_2\sum_{t=1}^{T}r^2\mathrm{d}^2(\mathbf{Z}_t,\mathbf{Z}_t^\star)\end{aligned} \tag{B.11}$$

(a) $\|\mathbf{V}\|_2\|\Delta\mathbf{a}_t\|_F\|\Delta\mathbf{V}\|_F \leq 2\|\Delta\mathbf{a}_t\|_F\|\Delta\mathbf{V}\|_F$, because $\|\mathbf{V}\|_2 \leq 1+r \leq 2$. Note that $2\|\Delta\mathbf{V}\|_F\|\Delta\mathbf{a}_t\|_F \leq d^2(\mathbf{Z}_t, \mathbf{Z}_t^\star)$, and therefore $\|\mathbf{V}\|_2\|\Delta\mathbf{a}_t\|_F\|\Delta\mathbf{V}\|_F \leq d^2(\mathbf{Z}_t, \mathbf{Z}_t^\star)$. (b) This is by Young's inequality, where $e_2 > 0$. Furthermore, we have $(\varepsilon_{stat} + \|\nabla\ell_{N,t}(\mathbf{\Sigma}_t) - \nabla\ell_{N,t}(\mathbf{\Sigma})_t^\star\|_F)^2 \leq 2(\varepsilon_{stat}^2 + \|\nabla\ell_{N,t}(\mathbf{\Sigma}_t) - \nabla\ell_{N,t}(\mathbf{\Sigma})_t^\star\|_F^2)$, and $d^2(\mathbf{Z}_t, \mathbf{Z}_t^\star) \leq 2r^2$.

$$\begin{aligned} A14 &\geq -\frac{2\rho\eta}{T}\sum_{t=1}^{T}|\langle\nabla\ell_{N,t}(\mathbf{\Sigma}_t), \Delta\mathbf{V}\,\mathrm{diag}(\mathbf{a}_t^\star)\mathbf{R}\Delta\mathbf{V}^\top\rangle| \\ &\geq -\frac{2\rho\eta}{T}\sum_{t=1}^{T}(\varepsilon_{stat} + \|\nabla\ell_{N,t}(\mathbf{\Sigma}_t) - \nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|)\|\,\mathrm{diag}(\mathbf{a}_t^\star)\|_2\|\Delta\mathbf{V}\|_F^2 \\ &\overset{(a)}{\geq} -\frac{2\rho\eta}{T}\sum_{t=1}^{T}\frac{1}{e_3}(\varepsilon_{stat}^2 + \|\nabla\ell_{N,t}(\mathbf{\Sigma}_t) \\ &- \nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|_F^2) - \frac{2\rho\eta}{T}\sum_{t=1}^{T}\frac{1}{2}e_3c^2r^2\mathrm{d}^2(\mathbf{Z}_t, \mathbf{Z}_t^\star) \end{aligned} \tag{B.12}$$

(a) This is by Young's inequality, where $e_3 > 0$. Furthermore, we have $\|\mathbf{a}_t^\star\|_2 \leq c$ and $\|\Delta\mathbf{V}\|_F^2 \leq r\|\Delta\mathbf{V}\|_F$ and $\|\Delta\mathbf{V}\|_F^2 \leq \mathrm{d}^2(\mathbf{Z}_t, \mathbf{Z}_t^\star)$.

$$\begin{aligned} A \geq& \frac{2\rho\eta}{T}\sum_{t=1}^{T}(\frac{L\mu}{L+\mu} - \frac{e_1}{2})\|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F^2 \\ &- \frac{4\rho\eta e_2}{T}\sum_{t=1}^{T}r^2\mathrm{d}^2(\mathbf{Z}_t, \mathbf{Z}_t^\star) - \frac{\rho\eta e_3}{T}c^2r^2\sum_{t=1}^{T}\mathrm{dist}^2(\mathbf{Z}_t, \mathbf{Z}_t^\star) \\ &- \frac{2\rho\eta}{T}\varepsilon_{stat}^2\sum_{t=1}^{T}\left(\frac{1}{2e_1} + \frac{2}{e_2} + \frac{1}{e_3}\right) \\ &+ \frac{2\rho\eta}{T}\sum_{t=1}^{T}\left(\frac{1}{L+\mu} - \frac{2}{e_2} - \frac{1}{e_3}\right)\|\nabla\ell_{N,t}(\mathbf{\Sigma}_t) - \nabla\ell_{N,t}(\mathbf{\Sigma}_t^\star)\|_F^2 \end{aligned} \tag{B.13}$$

Combining A and B we get

$$\begin{aligned} A - B \geq& \frac{\rho\eta}{T}\sum_{t=1}^{T}\underbrace{\left(2\frac{L\mu}{L+\mu} - e_1\right)}_{C1}\|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F^2 \\ &- \frac{\rho}{T}\varepsilon_{stat}^2\sum_{t=1}^{T}\underbrace{\left(\frac{\eta}{e_1} + \frac{4\eta}{e_2} + \frac{2\eta}{e_3} + 16\frac{\eta^2}{T}\|\mathbf{Z}_t\|_2^2\right)}_{C2} \end{aligned}$$

$$+ \frac{\rho}{T}\sum_{t=1}^{T} \underbrace{\left(\frac{2\eta}{L+\mu} - \frac{4\eta}{e_2} - \frac{2\eta}{e_3} - 16\frac{\eta^2}{T}\|\mathbf{Z}_t\|_2^2\right)}_{C3} \|\nabla \ell_{N,t}(\mathbf{\Sigma}_t) - \nabla \ell_{N,t}(\mathbf{\Sigma}_t^\star)\|_F^2$$

$$- \frac{\eta\rho}{T}\left(4e_2 + e_3c^2\right) r^2 \operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) \tag{B.14}$$

where $e_1, e_2 > 0$. Now, we choose $e_1 = \frac{L\mu}{2(L+\mu)}$ $e_2 = 4(L+\mu)$, $e_3 = 4(L+\mu)$. Now, $C1 = \frac{3L\mu}{2(L+\mu)}$ and C3 is $\frac{\eta}{2(L+\mu)} - 16\frac{\eta^2}{T}\|\mathbf{Z}_t\|_2^2$. We want C3 to be nonegative so that we could drop this term, we require $\eta \le \min_t \frac{T}{32(L+\mu)\|\mathbf{Z}_t\|_2^2}$. By Lemma B.1, and therefore we can conclude that $\frac{L\mu}{L+\mu}\sum_{t=1}^{T}\|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F^2 \ge \frac{L\mu}{(L+\mu)\xi^2}\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$. If $r^2 \le \frac{L\mu}{4(L+\mu)^2(4+c^2)\xi^2}$, then

$$A - B \ge \frac{\rho\eta}{T}\frac{L\mu}{2\xi^2(L+\mu)}\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$$

$$- \varepsilon_{stat}^2 \underbrace{\frac{\rho\eta}{T}\sum_{t=1}^{T}\left(\frac{2(L+\mu)}{L\mu} + \frac{3}{2(L+\mu)} + 16\frac{\eta}{T}\|\mathbf{Z}_t\|_2^2\right)}_{D1} \tag{B.15}$$

$$D1 \overset{(a)}{\le} \rho\eta\varepsilon_{stat}^2\left(\frac{2(L+\mu)}{L\mu} + \frac{3}{2(L+\mu)} + \frac{1}{2(L+\mu)}\right) \le \rho\eta\frac{3(L+\mu)}{L\mu}\varepsilon_{stat}^2 \tag{B.16}$$

(a) By $\eta \le \min_t \frac{T}{32(L+\mu)\|\mathbf{Z}_t\|_2^2}$.

Plugging Equation (B.15) into Equation (B.5), we can obtain the following

$$\underbrace{\rho\left(1 - \frac{\eta L\mu}{2T\xi^2(L+\mu)}\right)}_{\beta}\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) + \rho\eta\frac{3(L+\mu)}{L\mu}\varepsilon_{stat}^2 \ge \operatorname{dist}^2(\mathbf{Z}^+, \mathbf{Z}^\star) \tag{B.17}$$

Here completes the proof. □

B.2 Proof of Lemma 3.2

Proof.

$$\|\mathbf{Z}_t\|_2 = \left\|\begin{bmatrix}\mathbf{V}\\ \operatorname{diag}(\mathbf{a}_t)\end{bmatrix} - \begin{bmatrix}\mathbf{V}^\star\mathbf{R}_t\\ \mathbf{R}_t^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\end{bmatrix} + \begin{bmatrix}\mathbf{V}^\star\mathbf{R}_t\\ \mathbf{R}_t^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\end{bmatrix}\right\|_2$$

$$
\begin{aligned}
&\leq \left\| \begin{bmatrix} \mathbf{V} \\ \operatorname{diag}(\mathbf{a}_t) \end{bmatrix} - \begin{bmatrix} \mathbf{V}^\star \mathbf{R}_t \\ \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{R}_t \end{bmatrix} \right\|_2 + \|\mathbf{Z}_t^\star\|_2 \\
&\leq \sqrt{2} r + \|\mathbf{Z}_t^\star\|_2 \\
&\overset{(a)}{\leq} \frac{\sqrt{2}\sigma_K(\mathbf{\Sigma}_t^\star)}{8} + \|\mathbf{Z}_t^\star\|_2 \\
&\overset{(b)}{\leq} \frac{\sqrt{2}+8}{8} \|\mathbf{Z}_t^\star\|_2
\end{aligned} \tag{B.18}
$$

(a) $r \leq \frac{\sigma_K(\mathbf{\Sigma}_t^\star)\sqrt{L\mu}}{2(L+\mu)\sqrt{4+c^2}\sqrt{1+(\sigma_K(\mathbf{\Sigma}_t^\star)+3c)^2}}$, where $L = \mu = 1$ in our case. Using the fact that $\sqrt{L\mu} \leq \frac{1}{2}(L+\mu)$, we can obtain $r \leq \frac{\sigma_K(\mathbf{\Sigma}_t^\star)}{8}$. (b) Since $\|\mathbf{Z}_t^\star\|_2 \geq \sigma_K(\mathbf{\Sigma}_t^\star)$, we can obtain the inequlity.

$$
\begin{aligned}
\|\mathbf{Z}_t^0\|_2 &= \left\| \begin{bmatrix} \mathbf{V}^0 \\ \operatorname{diag}(\mathbf{a}_t^0) \end{bmatrix} - \begin{bmatrix} \mathbf{V}^\star \mathbf{R}_t \\ \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t) \mathbf{R}_t \end{bmatrix} + \begin{bmatrix} \mathbf{V}^\star \mathbf{R}_t \\ \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t) \mathbf{R}_t \end{bmatrix} \right\|_2 \\
&\geq - \left\| \begin{bmatrix} \mathbf{V}^0 \\ \operatorname{diag}(\mathbf{a}_t^0) \end{bmatrix} - \begin{bmatrix} \mathbf{V}^\star \mathbf{R}_t \\ \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{R}_t \end{bmatrix} \right\|_2 + \|\mathbf{Z}_t^\star\|_2 \\
&\geq - \sqrt{2} r + \|\mathbf{Z}_t^\star\|_2 \\
&\geq - \frac{\sqrt{2}\sigma_K(\mathbf{\Sigma}_t^\star)}{8} + \|\mathbf{Z}_t^\star\|_2 \\
&\geq \frac{8-\sqrt{2}}{8} \|\mathbf{Z}_t^\star\|_2
\end{aligned} \tag{B.19}
$$

Combining Equations (B.18) and (B.19), we could obtain

$$
\|\mathbf{Z}_t\|_2 \leq \frac{8+\sqrt{2}}{8-\sqrt{2}} \|\mathbf{Z}_t^0\|_2 \tag{B.20}
$$

This implies that $\|\mathbf{Z}_t\|_2^2 \leq 2\|\mathbf{Z}_t^0\|_2^2$. Therefore

$$
\min_t \frac{T}{32(L+\mu)\|\mathbf{Z}_t\|_2^2} \geq \min_t \frac{T}{64(L+\mu)\|\mathbf{Z}_t^0\|_2^2} \tag{B.21}
$$

η suffices to satisfy $\eta \leq \min_t \frac{T}{64(L+\mu)\|\mathbf{Z}_t^0\|_2^2}$. □

B.3 Proof of Corollary 3.1.1

Proof.

$$\begin{aligned}
\|\mathbf{\Sigma}_t^i - \mathbf{\Sigma}_t^\star\|_F &\leq \|(\mathbf{V}^i \operatorname{diag}(\mathbf{a}_t^i) - \mathbf{V}^\star \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R})\mathbf{R}^\top \mathbf{V}^{\star\top}\|_F \\
&+ \|\mathbf{V}^i \operatorname{diag}(\mathbf{a}_t^i)(\mathbf{V}^i - \mathbf{V}^\star\mathbf{R})^\top\| \\
&\leq \left(\|\mathbf{V}^i\|_2 \|\operatorname{diag}(\mathbf{a}_t^i))\|_2 + \|\operatorname{diag}(\mathbf{a}_t^\star)\|_2\right) \|\mathbf{V} - \mathbf{V}^\star\mathbf{R}\|_F \\
&+ \|\mathbf{V}^i\|_2 \|\operatorname{diag}(\mathbf{a}_t^i) - \mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\|_F \\
&\leq \left(\frac{1}{2}\|\mathbf{Z}_t^i\|_2^2 + \|\operatorname{diag}(\mathbf{a}_t^\star)\|_2\right) \|\mathbf{V} - \mathbf{V}^\star\mathbf{R}\|_F \\
&+ \|\mathbf{V}^i\|_2 \|\operatorname{diag}(\mathbf{a}_t^i) - \mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\|_F \\
&\leq Q\left(\|\mathbf{V} - \mathbf{V}^\star\mathbf{R}\|F + \|\operatorname{diag}(\mathbf{a}_t^i) - \mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\|_F\right) \qquad (B.22)
\end{aligned}$$

where $Q = \max_t \frac{1}{2}\|\mathbf{Z}_t^i\|_2^2 + \|\mathbf{Z}_t^\star\|_2$. Therefore

$$\begin{aligned}
\sum_{t=1}^{T} \|\mathbf{\Sigma}_t^i - \mathbf{\Sigma}_t^\star\|_F^2 &\leq 2Q^2 \operatorname{dist}^2(\mathbf{Z}^i, \mathbf{Z}^\star) \\
&\leq 2Q^2 \left(\beta^i \operatorname{dist}^2(\mathbf{Z}^0, \mathbf{Z}^\star) + \rho\eta \frac{3(L+\mu)}{L\mu} \varepsilon_{stat}^2\right) \qquad (B.23)
\end{aligned}$$

□

B.4 Proof of Lemma B.1

First we show that the following inequality holds.

Lemma B.1.

$$\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) \leq \xi^2 \sum_{t=1}^{T} \|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F^2 \qquad (B.24)$$

where $\xi^2 = \min_{t\in[T]} \left(\left(\frac{1}{\sigma_K(\mathbf{\Sigma}_t^\star)}\right)^2 + \left(1 + 3\frac{c}{\sigma_K(\mathbf{\Sigma}_t^\star)}\right)^2\right)$

With Lemma B.1 in hand, we could rewrite $\sum_{t=1}^{T} \|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F^2$ in terms of the distance metric $\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$ and then show the contraction of $\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star)$ through iterates.

Proof.

$$\begin{aligned}\|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F &\geq \|\mathbf{V}^\star\mathbf{R}\left(\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\right)\mathbf{R}^\top\mathbf{V}^{\star\top}\|_F \\ &\quad - \|\mathbf{V}\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^\top - \mathbf{V}^\star\operatorname{diag}(\mathbf{a}_t)\mathbf{V}^{\star\top}\|_F \\ &\overset{(a)}{\geq} \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\|_F - c\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F \end{aligned} \tag{B.25}$$

(a) $\mathbf{V}^*\mathbf{R}$ is an unitary matrix. By the unitary invariance property of the Frobenious norm, we have $\|\mathbf{V}^\star\mathbf{R}\left(\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\right)\mathbf{R}^\top\mathbf{V}^{\star\top}\|_F = \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\|_F$.

Equation (B.25) implies that $\|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}^\top\operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}\|_F \leq \|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F + c\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F$. Furthermore

$$\begin{aligned}\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F &\leq (\|\mathbf{V}\|_2 + \|\mathbf{V}^\star\|_2)\|\mathbf{V} - \mathbf{V}^\star\mathbf{R}\|_F \\ &\overset{(a)}{\leq} (2+r)\|\mathbf{V} - \mathbf{V}\mathbf{R}\|_F^2 \end{aligned} \tag{B.26}$$

(a) The inequlity is obtained by the fact that $\|\mathbf{V} - \mathbf{V}^\star\|_2 \leq r$ and $\|\mathbf{V}^\star\|_2 = 1$.

By Lemma 3 in [48], we could obtain the following inequality

$$\|\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^\star\|_F^2 \geq \sigma_K(\mathbf{\Sigma}_t^\star)\|\mathbf{V}\sqrt{\operatorname{diag}(\mathbf{a}_t)} - \mathbf{V}^\star\sqrt{\operatorname{diag}(\mathbf{a}_t)^\star}\mathbf{R}_t\|_F^2 \tag{B.27}$$

$$\begin{aligned}&\|\mathbf{V}\sqrt{\operatorname{diag}(\mathbf{a}_t)} - \mathbf{V}^\star\sqrt{\operatorname{diag}(\mathbf{a}_t)^\star}\mathbf{R}_t\|_F^2 = \sum_{k=1}^K \left\|\sqrt{a_{k,t}}\mathbf{v}_k - \sqrt{a^\star_{\psi_{\mathbf{R}_t}(k),t}}\mathbf{v}^\star_{\psi_{\mathbf{R}_t}(k)}\right\|_2^2 \\ &= \sigma_K(\mathbf{\Sigma}_t^\star)\sum_{k=1}^K \left\|\sqrt{\frac{a_{k,t}}{\sigma_K(\mathbf{\Sigma}_t^\star)}}\mathbf{v}_k - \sqrt{\frac{a^\star_{\psi_{\mathbf{R}_t}(k),t}}{\sigma_K(\mathbf{\Sigma}_t^\star)}}\mathbf{v}^\star_{\psi_{\mathbf{R}_t}(k)}\right\|_2^2 \\ &\overset{(a)}{\geq} \sigma_K(\mathbf{\Sigma}_t^\star)\sum_{k=1}^K \left\|\Pi_{\mathcal{C}_B}\left(\sqrt{\frac{a_{k,t}}{\sigma_K(\mathbf{\Sigma}_t^\star)}}\mathbf{v}_k\right) - \Pi_{\mathcal{C}_B}\left(\sqrt{\frac{a^\star_{\psi_{\mathbf{R}_t}(k),t}}{\sigma_K(\mathbf{\Sigma}_t^\star)}}\mathbf{v}^\star_{\psi_{\mathbf{R}_t}(k)}\right)\right\|_2^2 \\ &\overset{(b)}{=} \sigma_K(\mathbf{\Sigma}_t^\star)\sum_{k=1}^K \left\|\mathbf{v}_k - \mathbf{v}^\star_{\psi_{\mathbf{R}_t}(k)}\right\|_2^2 = \sigma_K(\mathbf{\Sigma}_t^\star)\|\mathbf{V} - \mathbf{V}^\star\mathbf{R}_t\|_F^2 \end{aligned} \tag{B.28}$$

where $\Pi_{\mathcal{C}_B}$ denotes the projection to the unit norm ball, $\sigma_K(\mathbf{\Sigma}_t^\star)$ denotes the K-th largest singular value of $\mathbf{\Sigma}_t^\star$, $\psi_{\mathbf{R}_t}(\cdot)$ denotes the permutation function associated with the permutation matrix $\mathbf{R}_t$[1]. (a) The inequality is obtained by the non-expansive property of projection to the convex set. (b) Since $\|\mathbf{v}_k\|_2$, $\|\mathbf{v}_k^\star\|_2$ are 1 for all $k \in [K]$, and $\sqrt{\frac{a_{k,t}}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}}, \sqrt{\frac{a_{k,t}^\star}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}}$ are greater than 1 for all $k \in [K],\ t \in [T]$, then $\Pi_{\mathcal{C}_B}\left(\sqrt{\frac{a_{k,t}}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}}\mathbf{v}_k\right) = \mathbf{v}_k$ and $\Pi_{\mathcal{C}_B}\left(\sqrt{\frac{a^\star_{\psi_{\mathbf{R}_t}(k),t}}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}}\mathbf{v}^\star_{\psi_{\mathbf{R}_t}(k)}\right) = \mathbf{v}_k^\star$ for all $k \in [K],\ t \in [T]$.

Combining Equations (B.27) and (B.28), we can conclude that

$$\|\mathbf{V} - \mathbf{V}^\star\mathbf{R}\|_F \le \min_{t\in[T]} \frac{1}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}\|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t^\star\|_F \tag{B.29}$$

Therefore

$$\begin{aligned}
&\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) = \sum_{t=1}^{T} \min_{\mathbf{R}_t \in \mathcal{P}(K)} \|\mathbf{V} - \mathbf{V}^\star\mathbf{R}_t\|_F^2 + \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2 \\
&\le \min_{t\in[T]} \left(\left(\frac{1}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}\right)^2 + \left(1 + (2+r)\frac{c}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}\right)^2\right) \sum_{t=1}^{T} \|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t^\star\|_F^2 \\
&\le \min_{t\in[T]} \left(\left(\frac{1}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}\right)^2 + \left(1 + 3\frac{c}{\sigma_K(\boldsymbol{\Sigma}_t^\star)}\right)^2\right) \sum_{t=1}^{T} \|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t^\star\|_F^2 \qquad \text{(B.30)}
\end{aligned}$$

□

[1] Note that in Lemma 4.3 we use $\sigma_{\mathbf{R}}(\cdot)$ to denote the permutation function with corresponding permutation matrix $\mathbf{R}$. Here we replace $\sigma_{\mathbf{R}}(\cdot)$ with $\psi_{\mathbf{R}}(\cdot)$ to distinguish between the top-K singular value of the matrix $\sigma_K(\cdot)$.

APPENDIX C

PROOF OF THEOREM 4.1

C.1 Proof of Theorem 4.1

Proof.

$$\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) = \min_{\substack{\mathbf{R}_t \in \mathcal{P}(K) \\ \forall t \in [T]}} \sum_{t=1}^{T} \left[\|\mathbf{V} - \mathbf{V}^\star \mathbf{R}_t\|_F^2 + \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2 \right]$$

$$\overset{\text{(a)}}{\leq} \underbrace{\frac{T}{2(\sqrt{2}-1)} \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F^2}_{S1}$$

$$+ \min_{\substack{\mathbf{R}_t \in \mathcal{P}(K) \\ \forall t \in [T]}} \sum_{t=1}^{T} \|\operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star)\mathbf{R}_t\|_F^2$$

$$\overset{\text{(b)}}{\leq} S1 + 2\sum_{t=1}^{T} \left[\|\mathbf{\Sigma}_t^\star\|_2^2 \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_F^2 + 2K\|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \right]$$

$$\overset{\text{(c)}}{\leq} \frac{4KT}{2(\sqrt{2}-1)g^2} \|\mathbf{M} - \mathbf{M}^\star\|_2^2 + \frac{8K}{g^2} \sum_{t=1}^{T} \|\mathbf{\Sigma}_t^\star\|_2^2 \|\mathbf{M} - \mathbf{M}^\star\|_2^2$$

$$+ 2K \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2$$

$$\overset{\text{(d)}}{\leq} \frac{5KT}{g^2} \|\mathbf{M} - \mathbf{M}^\star\|_2^2 + \left(\frac{8K}{g^2} T \sum_{t=1}^{T} \|\mathbf{\Sigma}_t^\star\|_2^2 + 2K \right) \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \qquad \text{(C.1)}$$

(a) By Lemma 5.4 in [39], (b) is by Lemma 4.3, (c) is by Davis-Kahan $\sin\theta$ theorem (see Lemma 4.1), and (d) is by Cauchy-Schwarz inequality, $\|\mathbf{M} - \mathbf{M}^\star\|_2^2 = \|\sum_{t=1}^{T} \mathbf{S}_{N,t} - \mathbf{\Sigma}_t\|_2^2 \leq T\sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t\|_2^2$.

Let $\zeta = \frac{8K}{g^2} T \sum_{t=1}^{T} \|\mathbf{\Sigma}_t^\star\|_2^2 + 2K$

$$\mathrm{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) \leq \frac{5KT}{g^2} \|\mathbf{M} - \mathbf{M}^\star\|_2^2 + \zeta \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \leq 2Tr^2 \tag{C.2}$$

It suffices to show that $\frac{5KT}{g^2}\|\mathbf{M} - \mathbf{M}^\star\|_2^2 \leq Tr^2$ and $\zeta \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \leq Tr^2$.

$$\begin{aligned}
&\mathbb{P}\left\{\frac{5K}{g^2}\|\mathbf{M} - \mathbf{M}^\star\|_2^2 \geq r^2\right\} = \mathbb{P}\left\{\|\mathbf{M} - \mathbf{M}^\star\|_2 \geq \frac{gr}{\sqrt{5K}}\right\} \\
&\overset{(a)}{\leq} 2P \exp\left(\frac{-n(gr)^2}{10Kc^2(\|\mathbf{M}^\star\|_2 + \frac{gr}{\sqrt{5K}})}\right) = 2P \exp\left(\frac{-n(gr)^2}{10Kc^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{gr}{\sqrt{5K}})}\right)
\end{aligned} \tag{C.3}$$

(a) is by Bernstein's concentration inequality (Corollary 6.20 in [40]).

Similarly

$$\begin{aligned}
&\mathbb{P}\left\{\zeta \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \geq Tr^2\right\} = \mathbb{P}\left\{\sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \geq \frac{Tr^2}{\zeta}\right\} \\
&\leq \mathbb{P}\left\{\bigcup_{t=1}^{T}\left\{\|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \geq \frac{r^2}{\zeta}\right\}\right\} \leq \sum_{t=1}^{T} \mathbb{P}\left\{\|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \geq \frac{r^2}{\zeta}\right\} \\
&\overset{(a)}{\leq} 2P \sum_{t=1}^{T} \exp\left(\frac{-nr^2}{2\zeta c^2(\|\mathbf{\Sigma}_t^\star\|_2 + \frac{r}{\sqrt{\zeta}})}\right) \overset{(b)}{\leq} 2PT \exp\left(\frac{-nr^2}{2\zeta c^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{r}{\sqrt{\zeta}})}\right)
\end{aligned} \tag{C.4}$$

Both (a), (b) is by $\|\mathbf{\Sigma}_t^\star\|_2 \leq \|\mathbf{M}^\star\|_2 \leq \|\tilde{\mathbf{a}}_1^\star\|_1$.

$$\begin{aligned}
&\mathbb{P}\left\{\mathrm{dist}^2(\mathbf{Z}^0, \mathbf{Z}^\star) \leq 2Tr^2\right\} \\
&\geq \mathbb{P}\left\{\left\{\frac{5KT}{g^2}\|\mathbf{M} - \mathbf{M}^\star\|_2^2 \leq Tr^2\right\} \bigcap \left\{\zeta \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \leq Tr^2\right\}\right\} \\
&= 1 - \mathbb{P}\left\{\left\{\frac{5K}{g^2}\|\mathbf{M} - \mathbf{M}^\star\|_2^2 \geq r^2\right\} \bigcup \left\{\zeta \sum_{t=1}^{T} \|\mathbf{S}_{N,t} - \mathbf{\Sigma}_t^\star\|_2^2 \geq Tr^2\right\}\right\} \\
&\geq 1 - \underbrace{2P \exp\left(\frac{-n(gr)^2}{10Kc^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{gr}{\sqrt{5K}})}\right)}_{S2} - \underbrace{2PT \exp\left(\frac{-nr^2}{2\zeta c^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{r}{\sqrt{\zeta}})}\right)}_{S3}
\end{aligned} \tag{C.5}$$

Therefore, to obtain the lower bound of $\mathrm{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) \leq 2Tr^2$ with probability at least $1 - 2T\delta$, it suffices to find the sample bound such that $\mathrm{S2} \leq T\delta$ and $\mathrm{S3} \leq T\delta$. Therefore, the sample size has to be at least larger than

$$n \geq \max\left(-\frac{10Kc^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{gr}{\sqrt{5K}})}{(gr)^2} \log\frac{T\delta}{4P}, -\frac{2\zeta c^2(\|\tilde{\mathbf{a}}_1^\star\|_1 + \frac{r}{\sqrt{\zeta}})}{r^2} \log\frac{\delta}{4P} \right) \tag{C.6}$$

□

C.2 Proof of Lemma 4.2

Proof. Let $\mathcal{V}_{\uparrow i} = \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_i)$ and $\mathcal{V}_{\downarrow i}^\star = \mathrm{span}(\mathbf{v}_i^\star, \ldots, \mathbf{v}_P^\star)$. Then, $\mathcal{V}_{\uparrow i} \cap \mathcal{V}_{\downarrow i}^\star \neq \emptyset$ because $\dim(\mathcal{V}_{\uparrow i} \cup \mathcal{V}_{\downarrow i}^\star) > P$. We can rewrite μ_i as $\mu_i = \min_{\substack{\mathbf{x}\in\mathcal{V}_{\uparrow i} \\ \|\mathbf{x}\|=1}} \mathbf{x}^\top \boldsymbol{\Sigma}^\star \mathbf{x}$

$$\begin{aligned}\mu_i = \min_{\substack{\mathbf{x}\in\mathcal{V}_{\uparrow i} \\ \|\mathbf{x}\|=1}} \mathbf{x}^\top \boldsymbol{\Sigma}^\star \mathbf{x} &\leq \min_{\substack{\mathbf{x}\in\mathcal{V}_{\uparrow i}\cap\mathcal{V}_{\downarrow i}^\star \\ \|\mathbf{x}\|=1}} \mathbf{x}^\top \boldsymbol{\Sigma}^\star \mathbf{x} \\ &\leq \max_{\substack{\mathbf{x}\in\mathcal{V}_{\uparrow i}\cap\mathcal{V}_{\downarrow i}^\star \\ \|\mathbf{x}\|=1}} \mathbf{x}^\top \boldsymbol{\Sigma}^\star \mathbf{x} \leq \max_{\substack{\mathbf{x}\in\mathcal{V}_{\downarrow i}^\star \\ \|\mathbf{x}\|=1}} \mathbf{x}^\top \boldsymbol{\Sigma}^\star \mathbf{x} = \lambda_i^\star \end{aligned} \tag{C.7}$$

□

C.3 Proof of Lemma 4.3

Proof. Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors with norm 1.

$$\begin{aligned}&\min_{\sigma_{\mathbf{R}}(\cdot)\in\mathcal{S}_P} \sum_{k=1}^K |\mathbf{v}_k^\top \boldsymbol{\Sigma} \mathbf{v}_k - \lambda_{\sigma(k)}(\boldsymbol{\Sigma}^\star)|^2 \\ &= \min_{\sigma_{\mathbf{R}}(\cdot)\in\mathcal{S}_P} \sum_{k=1}^K |\mathbf{v}_k^\top \boldsymbol{\Sigma}^\star \mathbf{v}_k + \mathbf{v}_k^\top \mathbf{E} \mathbf{v}_k - \lambda_{\sigma(k)}(\boldsymbol{\Sigma}^\star)|^2 \\ &\leq \min_{\sigma_{\mathbf{R}}(\cdot)\in\mathcal{S}_P} 2\sum_{k=1}^K |\mu_k - \lambda_{\sigma(k)}(\boldsymbol{\Sigma}^\star)|^2 + 2\sum_{k=1}^K |\mathbf{v}_k^\top \mathbf{E} \mathbf{v}_k|^2 \\ &\leq 2\sum_{k=1}^K |\lambda_k^\star - \mu_k|^2 + 2\sum_{k=1}^K |\mathbf{v}_k^\top \mathbf{E} \mathbf{v}_k|^2 \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} 2\left|\sum_{k=1}^{K}(\lambda_k^\star - \mu_k)\right|^2 + 2\sum_{k=1}^{K}\left|\mathbf{v}_k^\top \mathbf{E}\mathbf{v}_k\right|^2 \\
&= 2|\langle \boldsymbol{\Sigma}, \mathbf{V}^\star\mathbf{V}^{\star\top} - \mathbf{V}\mathbf{V}^\top\rangle|^2 + 2\sum_{k=1}^{K}\left|\mathbf{v}_k^\top \mathbf{E}\mathbf{v}_k\right|^2 \\
&\leq 2\|\boldsymbol{\Sigma}\|_2^2\|\mathbf{V}^\star\mathbf{V}^{\star\top} - \mathbf{V}\mathbf{V}^\top\|_F^2 + 2\sum_{k=1}^{K}\lambda_k^2(\mathbf{E}) \qquad \text{(C.8)}
\end{aligned}
$$

(a) The minimum of all permutation matrix is smaller than the permutation of ordering the eigenvalues from largest to smallest. By Lemma 4.2, we know that $\lambda_k^\star - \mu_k \geq 0, \quad \forall k \in [K]$. □

APPENDIX D

FINDING PERMUTATION MATRICES

In the case where K is large, finding the permutation matrix becomes computational expensive and requires searching for K factorial possible solutions. In this case, we could relax the optimization problem (2.3) to the following

$$\operatorname{dist}^2(\mathbf{Z}, \mathbf{Z}^\star) = \sum_{t=1}^{T} \min_{\mathbf{R}_t \in \mathcal{D}(K)} \left\{ \|\mathbf{V} - \mathbf{V}^\star \mathbf{R}_t\|_F^2 + \| \operatorname{diag}(\mathbf{a}_t) - \mathbf{R}_t^\top \operatorname{diag}(\mathbf{a}_t^\star) \mathbf{R}_t \|_F^2 \right\} \tag{D.1}$$

where $\mathcal{D}(K)$ is the set of $K \times K$ doubly stochastic matrices. Note that Equation (D.1) is a constrained convex optimization problem and could be solved in polynomial time. After finding $\{\mathbf{R}_t\}_{t \in [T]}$, we could decompose the doubly stochastic matrices to positive convex combinations of permutation matrices by Birkhoff–von Neumann theorem. Then, we select the permutation matrix with largest coefficient.

REFERENCES

[1] R. F. Engle, O. Ledoit, and M. Wolf, “Large dynamic covariance matrices,” *Journal of Business & Economic Statistics*, vol. 37, no. 2, pp. 363–375, 2019.

[2] Y. Wu, J. M. Hernández-Lobato, and Z. Ghahramani, “Dynamic covariance models for multivariate financial time series,” *arXiv preprint arXiv:1305.4268*, 2013.

[3] E. B. Fox and D. B. Dunson, “Bayesian nonparametric covariance regression,” *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2501–2542, 2015.

[4] D. J. Lurie, D. Kessler, D. S. Bassett, R. F. Betzel, M. Breakspear, S. Kheilholz, A. Kucyi, R. Liégeois, M. A. Lindquist, A. R. McIntosh et al., “Questions and controversies in the study of time-varying functional connectivity in resting fMRI,” *Network Neuroscience*, vol. 4, no. 1, pp. 30–69, 2020.

[5] D. Vidaurre, S. M. Smith, and M. W. Woolrich, “Brain network dynamics are hierarchically organized in time,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12 827–12 832, 2017.

[6] E. Soreq, R. Leech, and A. Hampshire, “Dynamic network coding of working-memory domains and working-memory processes,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[7] R. Liégeois, J. Li, R. Kong, C. Orban, D. Van De Ville, T. Ge, M. R. Sabuncu, and B. T. Yeo, “Resting brain dynamics at different timescales capture distinct aspects of human behavior,” *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[8] D. Lurie, D. Kessler, D. Bassett, R. F. Betzel, M. Breakspear, S. Keilholz, A. Kucyi, R. Liégeois, M. A. Lindquist, A. R. McIntosh et al., “On the nature of time-varying functional connectivity in resting fMRI,” *PsyArXiv*, 2018.

[9] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[10] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proceedings of the 30th International Conference on Machine Learning*, no. CONF, 2013, pp. 427–435.

[11] R. Escalante and M. Raydan, *Alternating Projection Methods.* SIAM, 2011, vol. 8.

[12] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[13] J. M. Shine, P. G. Bissett, P. T. Bell, O. Koyejo, J. H. Balsters, K. J. Gorgolewski, C. A. Moodie, and R. A. Poldrack, "The dynamics of functional brain networks: Integrated network states during cognitive task performance," *Neuron*, vol. 92, no. 2, pp. 544–554, 2016.

[14] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," in *Conference on Learning Theory*, 2016, pp. 530–582.

[15] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[16] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, 2019.

[17] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[18] 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.

[19] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[20] N. Leonardi and D. Van De Ville, "On spurious and real fluctuations of dynamic functional connectivity during rest," *Neuroimage*, vol. 104, pp. 430–436, 2015.

[21] C. J. Paciorek and M. J. Schervish, “Nonstationary covariance functions for Gaussian process regression,” in *Advances in Neural Information Processing Systems*, 2004, pp. 273–280.

[22] M. R. Andersen, O. Winther, L. K. Hansen, R. Poldrack, and O. Koyejo, “Bayesian structure learning for dynamic brain connectivity,” in *21st International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, 2018, pp. 1436–1446.

[23] L. Li, D. Pluta, B. Shahbaba, N. Fortin, H. Ombao, and P. Baldi, “Modeling dynamic functional connectivity with latent factor Gaussian processes,” in *Advances in Neural Information Processing Systems*, 2019, pp. 8261–8271.

[24] R. Li, “Multivariate sparse coding of nonstationary covariances with Gaussian processes,” in *Advances in Neural Information Processing Systems*, 2019, pp. 1610–1619.

[25] G. Kastner, S. Frühwirth-Schnatter, and H. F. Lopes, “Efficient Bayesian inference for multivariate factor stochastic volatility models,” *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 905–917, 2017.

[26] G. Mishne and A. S. Charles, “Learning spatially-correlated temporal dictionaries for calcium imaging,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 1065–1069.

[27] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis.* Springer Science & Business Media, 2005.

[28] D. F. Ahelegbey, M. Billio, and R. Casarin, “Bayesian graphical models for structural vector autoregressive processes,” *Journal of Applied Econometrics*, vol. 31, no. 2, pp. 357–386, 2016.

[29] H. Qiu, F. Han, H. Liu, and B. Caffo, “Joint estimation of multiple graphical models from high dimensional time series,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 2, pp. 487–504, 2016.

[30] R. A. Davis, P. Zang, and T. Zheng, “Sparse vector autoregressive modeling,” *Journal of Computational and Graphical Statistics*, vol. 25, no. 4, pp. 1077–1096, 2016.

[31] A. Skripnikov and G. Michailidis, “Regularized joint estimation of related vector autoregressive models,” *Computational Statistics* & *Data Analysis*, vol. 139, pp. 164 – 177, 2019.

[32] M. G. Preti, T. A. Bolton, and D. Van De Ville, "The dynamic functional connectome: State-of-the-art and perspectives," *Neuroimage*, vol. 160, pp. 41–54, 2017.

[33] A. Zalesky and M. Breakspear, "Towards a statistical test for functional connectivity dynamics," *Neuroimage*, vol. 114, pp. 466–470, 2015.

[34] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.

[35] P.-L. Loh and M. J. Wainwright, "Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 559–616, 2015.

[36] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably," *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2165–2204, 2018.

[37] M. Yu, V. Gupta, M. Kolar et al., "Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 413–457, 2020.

[38] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the Davis–Kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, 2015.

[39] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," *arXiv preprint arXiv:1507.03566*, 2015.

[40] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press, 2019, vol. 48.

[41] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696.

[42] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, 2006.

[43] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[44] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium et al., "The WU-Minn human connectome project: An overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[45] R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis.* Cambridge University Press, 2011.

[46] X. Li, T. Zhao, R. Arora, H. Liu, and J. Haupt, "Stochastic variance reduced optimization for nonconvex sparse learning," in *International Conference on Machine Learning*, 2016, pp. 917–925.

[47] Y. Nesterov, *Introductory lectures on convex optimization: A basic course.* Springer Science & Business Media, 2013, vol. 87.

[48] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 51–96, 2019.