High Dimensional Covariance Estimation for Spatio-Temporal Processes

by

Kristjan Greenewald

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering-Systems)
in The University of Michigan
2017

Doctoral Committee:

Professor Alfred O. Hero III, Co-chair
Assistant Professor Shuheng Zhou, Co-chair
Assistant Professor Honglak Lee
Assistant Professor Raj Rao Nadakuditi

Kristjan H. Greenewald

greenewk@umich.edu

ORCID iD: 0000-0001-9038-1975

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

High Dimensional Covariance Estimation for Spatio-Temporal Processes

by

Kristjan Greenewald

Chairs: Alfred O. Hero III and Shuheng Zhou

High dimensional time series and array-valued data are ubiquitous in signal processing, machine learning, and science. Due to the additional (temporal) direction, the total dimensionality of the data is often extremely high, requiring large numbers of training examples to learn the distribution using unstructured techniques. However, due to difficulties in sampling, small population sizes, and/or rapid system changes in time, it is often the case that very few relevant training samples are available, necessitating the imposition of structure on the data if learning is to be done. The mean and covariance are useful tools to describe high dimensional distributions because (via the Gaussian likelihood function) they are a data-efficient way to describe a general multivariate distribution, and allow for simple inference, prediction, and regression via classical techniques.

In this work, we develop various forms of multidimensional covariance structure that explicitly exploit the *array structure* of the data, in a way analogous to the widely used low rank modeling of the mean. This allows dramatic reductions in the number of training samples required, in some cases to a single training sample.

Covariance models of this form have been increasing in interest recently, and statistical performance bounds for high dimensional estimation in sample-starved scenarios are of great relevance.

This thesis focuses on the high-dimensional covariance estimation problem, exploiting spatio-temporal structure to reduce sample complexity. Contributions are made in the following areas: (1) development of a variety of rich Kronecker product-based covariance models allowing the exploitation of spatio-temporal and other structure with applications to sample-starved real data problems, (2) strong performance bounds for high-dimensional estimation of covariances under each model, and (3) a strongly adaptive online method for estimating changing optimal low-dimensional metrics (inverse covariances) for high-dimensional data from a series of similarity labels.

CHAPTER I

Introduction

1.1 Multidimensional Modeling

Multidimensional processes are ubiquitous, and modeling them is a critical task in machine learning, signal processing, and statistics. A multidimensional process is a process that can be expressed as a function of multiple variables, i.e.

$$x(t_1, t_2, \ldots, t_K).$$

As a common example, if a set of variables x_i evolve in time t, we have the spatiotemporal process

$$x(i, t) = x_i(t).$$

If time is sampled discretely at times τ_i, the process becomes *array valued*:

$$\mathbf{X} = \begin{bmatrix} x_1(\tau_1) & \ldots & x_1(\tau_q) \\ \vdots & & \vdots \\ x_p(\tau_1) & \ldots & x_p(\tau_q) \end{bmatrix} \in \mathbb{R}^{p \times q}$$

Spatio-temporal (or bi-dimensional) processes are universal, and include video (pixels vs. time), EEG/MEG (electrodes vs. time), FMRI, sensor networks, and a large vari-

ety of meteorological, physical, and biological data. In high dimensional applications, however, unstructured modeling of the data becomes difficult since the number of variables grows as pq. This severely limits the applicability of covariance-based methods in high dimensional problems. As a result, many types of structure have been proposed, from ignoring the temporal dimension, including separability constraints (*Allen and Tibshirani*, 2010; *Werner et al.*, 2008), to application-specific generative models. Many of these forms of structure are restrictive (e.g. separability), lack explicit spatio-temporal structure (e.g. sparsity), or are somewhat application specific (generative models) (*Tsiligkaridis and Hero*, 2013; *Kalaitzis et al.*, 2013).

In recent years, a great deal of success has been achieved by modeling bi-dimensional processes as *low rank*, i.e. as a sum of r separable functions. Example applications include recommender systems (*Bell et al.*, 2007; *Allen and Tibshirani*, 2010), video modeling (*Moore et al.*, 2014; *He et al.*, 2012), radar (*Newstadt et al.*, 2014), and many others. This does not require application-specific parameterizations, and can drastically reduce the number of parameters to $O(r(p+q))$. This can make estimation simpler, and often reduces the required number of training examples to one.

Low rank methods, however, only model the first moment of the data process. While a variety of relatively ad-hoc methods exist for using the low rank model to perform prediction, missing data inference, etc. (e.g. (*Bell et al.*, 2007)), the statistically principled approach to such methods requires a model of the joint distribution of the spatiotemporal data (*Allen and Tibshirani*, 2010).

In this thesis, we thus consider methods that extend the "low-rank modeling" concept to the second moment, i.e. the bi-dimensional data covariance. In particular, we consider modeling the covariance as a *sum of separable covariances* (*Tsiligkaridis and Hero*, 2013). Even greater gains are possible than with the first moment, since the number of elements in the unstructured covariance grows as p^2q^2. We consider the covariance because (via the Gaussian likelihood function) it is a data-efficient way to

describe a general multivariate distribution, and allows simple inference, prediction, regression, anomaly detection, and classification. Significant reductions in the number of parameters describing the covariance allows for effective modeling of the rich higher-dimensional data sources ubiquitous in the age of big data. The added structure allows significant reductions in the required number of iid training samples, without compromising the bias of the estimate.

We introduce several models, incorporating aspects such as robustness (*Chandrasekaran et al.*, 2009), block Toeplitz structure (*Kamm and Nagy*, 2000), estimation of sparse graphical models (*Banerjee et al.*, 2008), REPLACE time varying spatial covariances (*Zhou et al.*, 2010a), and the modeling of higher dimensional processes. In addition to classical unlabeled covariance estimation, we also consider labeled "covariance learning", where similarity labels are used to learn a time-varying regularized metric ("inverse covariance") efficiently describing high-dimensional data and its low-dimensional structure. For each, we apply the model to real world datasets, and derive strong performance guarantees. Both the real data experiments and the theory demonstrate that significant reductions in the required number of training samples are achieved.

We note that in this work we are focused on the low-sample regime, where the celebrated convolutional neural network (CNN) and recurrent neural network (RNN) approaches do not perform well (*Srivastava et al.*, 2014). This focus allows us to consider highly nonstationary and/or highly individualized data sources, where large numbers of samples are either impossible (e.g. due to nonstationarity) or costly (e.g. biological data) to obtain. Further research into hybrid approaches is of interest, for example using our nonstationary methods on data-modality-specific features learned by a neural net to provide individualized learning, or by adding neural net structure to nonlinearize our models. Our last chapter in particular focuses on task-specific nonstationary online learning of an optimal data embedding, the application of which

to pre-learned nonlinear features could be especially fruitful.

1.2 Overview of Contributions

This thesis aims to address the problem of estimating covariances of high-dimensional array-valued data, by developing strong, generally applicable forms of structure that naturally model array valued processes. This structure should allow major reductions in the required number of training samples to levels that allow practical application of covariance methods to high-dimensional problems, particularly in the case where distributions are changing in time, limiting the number of available training samples.

In (*Tsiligkaridis and Hero*, 2013), the covariance was modeled as a sum of r (separable) Kronecker products. This dramatically reduced the number of parameters required to model a spatiotemporal covariance. In practice, however, there is almost always sparsely correlated noise, and/or outliers or missing entries in spatiotemporal data. These phenomena are not well modeled by r Kronecker products, and generally result in the estimate of r being increased dramatically and/or damaging the condition number of the covariance estimate. To address this problem, Chapter III presents the Robust KronPCA model of the covariance as a sum of r (separable) Kronecker products and a sparse term to model noise and anomalies. A nuclear norm plus L1 norm penalized Frobenius norm objective function is introduced, and an algorithm derived to find the global minimum. Additional block Toeplitz constraints are also optionally incorporated, giving even greater reduction in the number of parameters. Nonasymptotic bounds on the Frobenius norm of the error of the resulting estimator are then derived using random matrix theory. These bounds and simulations confirm that the required number of training samples is significantly reduced in high dimensions for small r. Finally, an application to the discovery of periodicity in cell cycle data is presented, demonstrating the ability of the estimator to estimate the covariance in high noise, corrupted settings with a minimum of training samples.

An application of Kronecker structure to the synthetic aperture radar space-time adaptive processing (STAP) problem is discussed in Chapter IV. STAP estimates the subspace associated with stationary clutter and removes it, thus enhancing the signatures of moving targets buried in the clutter and noise. Based on the physics of the problem, a Kronecker product covariance model with low rank factors is derived, and an iterative estimation algorithm presented. The resulting covariance estimate is used to design optimal clutter cancelation filters, which we prove achieve high moving target SINR with as few as $O(1)$ training samples. Results on the Gotcha multichannel SAR dataset are shown, demonstrating dramatically improved moving target detection performance.

Up to this point, we have considered data arranged in a two-dimensional array, i.e. a matrix. Many data sources, however, can be arranged in arrays of higher dimensions, i.e. a K dimensional tensor. Examples include video (x vs. y vs. time), EEG/FMRI data (spatial location vs. time vs. subject), and synthetic aperture radar (range vs. time vs. channel vs. pass), and many others. Chapter V presents our TEnsor gRAphical Lasso (TeraLasso) model, where the inverse covariance of a K dimensional tensor process is modeled by K-way Kronecker sum of sparse factors, allowing for non-separable estimation exploiting both Kronecker structure and sparsity in the inverse, i.e. estimation of a graphical model. This is opposed to KronPCA, which gives a more general model, but is limited to $K = 2$ and cannot efficiently estimate a graphical model in the inverse covariance. The TeraLasso solves a convex L1 penalized objective function to find the maximum-entropy sparse Kronecker sum approximation to the precision matrix given projections of the data along each tensor mode. We derive non-asymptotic bounds for the TeraLasso estimator that give one-sample convergence in high dimensions, and successfully apply the model to meteorological and brain EEG datasets.

Complex, high dimensional datasets are often very rich, admitting a wide variety

of classification tasks and/or interpretations. In addition, real data sources often arise from distributions that change with time. In both cases, it is desirable to find a set of features that allow embedding the data in a low-dimensional space that preserves only the structure needed for the task at hand. This optimal embedding may change, due to changes in distribution and/or changes in task. In chapter VI, we consider dynamic metric learning, where we use regularization and adaptive online learning methods to track a changing optimal low rank metric (low rank embedding) on a data space. Learning is based on a time series of similarity labels on pairs of points. This allows us to perform feature selection on the fly, finding an optimal embedding of the data that best emphasizes the desired structure at any given time or task. We call our approach Online Convex Ensemble StrongLy Adaptive Dynamic Learning (OCELAD), and it is broadly applicable beyond the realm of metric learning. Regret bounds are derived that show OCELAD is *strongly adaptive*, implying low regret on every subinterval in the face of a changing optimal metric. We apply our methods to both simulated data, and to tracking changing political discussion on Twitter. Both confirm the theoretical results and demonstrate the ability of OCELAD to rapidly adapt and track changes in the optimal metric.

1.3 List of publications

Journal Publications

1. K. Greenewald, S. Zhou, and A. Hero, "The Tensor Graphical Lasso (TeraLasso)," nearing submission to JMLR.

2. K. Greenewald, S. Kelley, B. Oselio, and A. Hero, "Similarity Function Tracking using Pairwise Comparisons," submitted to IEEE Transactions on Signal Processing.

3. K. Greenewald, E. Zelnio, and A. Hero, "Robust SAR STAP via Kronecker

Decomposition," *IEEE Transactions on Aerospace and Electronic Systems*, Dec 2016.

4. K. Greenewald and A. Hero, "Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation," *IEEE Transactions on Signal Processing*, 2015.

5. K. Moon, K. Sricharan, K. Greenewald, and A. Hero, "Nonparametric Ensemble Estimation of Distributional Functionals," Submitted to *IEEE Transactions on Information Theory*, arXiv preprint arXiv:1601.06884.

Conference Publications

1. K. Greenewald, S. Kelley, and A. Hero, "Dynamic Metric Tracking from Pairwise Comparisons," *Allerton Conference on Control, Communication, and Computing*, 2016.

2. K. Moon, K. Sricharan, K. Greenewald, and A.O. Hero III, "Improving Convergence of Divergence Functional Ensemble Estimators," *IEEE ISIT* 2016.

3. K. Greenewald, E. Zelnio, and A. Hero, "Robust Kronecker STAP," *Proceedings of SPIE DSS*, 2016.

4. K. Greenewald and A. Hero, "Regularized Block Toeplitz Covariance Matrix Estimation via Kronecker Product Expansions," *IEEE Workshop on Statistical Signal Processing (SSP)*, 2014 (invited).

5. K. Greenewald and A. Hero, "Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation," *International Conference on Partial Least Squares and Related Methods (PLS)*, 2014 (invited).

6. K. Greenewald and A. Hero, "Kronecker PCA based spatio-temporal modeling of video for dismount classification," *Proceedings of SPIE DSS*, 2014.

7. K. Greenewald, T. Tsiligkaridis, and A. Hero, "Kronecker Sum Decompositions of Space-Time Data," *IEEE Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, December 2013 (invited).

CHAPTER II

Overview

In this chapter, we establish the mathematical setting, introduce the notation, and cover the main results that will propel the thesis by chapter.

2.1 Kronecker Structured Covariance Estimation

In the first part of this work, we develop methods for structured estimation of spatio-temporal covariances and apply them to multivariate time series modeling and parameter estimation. The covariance for spatio-temporal processes is manifested as a multiframe covariance, i.e. as the covariance not only between variables or features in a single frame (time point), but also between variables in a set of p_t nearby frames. If each frame contains p_s spatial variables, then the spatio-temporal covariance at time t is described by a $p_t p_s$ by $p_t p_s$ matrix

$$\mathbf{\Sigma}_t = \mathrm{Cov}\left[\{\mathbf{I_n}\}_{n=t-p_t}^{t-1}\right], \tag{2.1}$$

where $\mathbf{I}_n$ denotes the p_s variables or features of interest in the nth frame.

As $p_s p_t$ can be very large, even for moderately large p_s and p_t the number of degrees of freedom $(p_s p_t(p_s p_t + 1)/2)$ in the covariance matrix can greatly exceed the number n of training samples available to estimate the covariance matrix. One way

to handle this problem is to introduce structure and/or sparsity into the covariance matrix, thus reducing the number of parameters to be estimated.

A natural non-sparse option is to introduce structure by modeling the covariance matrix $\mathbf{\Sigma}$ as the Kronecker product of two smaller symmetric positive definite matrices, i.e.

$$\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}. \tag{2.2}$$

When the measurements are Gaussian with covariance of this form they are said to follow a matrix-normal distribution (*Dutilleul*, 1999; *Dawid*, 1981; *Tsiligkaridis et al.*, 2013). This model lends itself to coordinate decompositions (*Tsiligkaridis and Hero*, 2013), such as the decomposition between space (variables) vs. time (frames) natural to spatio-temporal data (*Tsiligkaridis and Hero*, 2013; *Greenewald et al.*, 2013). In the spatio-temporal setting, the $p_s \times p_s$ $\mathbf{B}$ matrix is the "spatial covariance" and the $p_t \times p_t$ matrix $\mathbf{A}$ is the "time covariance," both identifiable up to a multiplicative constant.

As it stands, the model (2.2) is somewhat insufficient for most practical applications. Separability of the covariance implies that predicting a variable in time depends only on previous samples form that same variable, completely ignoring its neighbors. While for some datasets separable models are useful, for many problems non-separable models are critical to be able to perform effective prediction. The focus of our covariance estimation work has thus been the exploitation of this natural Kronecker space-time structure and using it to develop a rich variety of models, most of which are nonseparable, yet are able to exploit different forms of Kronecker structure to provably achieve significant gains in sample complexity. Each method has its own advantages and is applicable to its own set of problems.

2.1.1 Kronecker PCA

The first model we consider is an extension to the representation (2.2). In (*Tsiligkaridis and Hero*, 2013), the covariance matrix is approximated using a sum of Kronecker product factors

$$\mathbf{\Sigma} = \sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i, \tag{2.3}$$

where r is the separation rank, $\mathbf{A}_i \in \mathbb{R}^{p_t \times p_t}$, and $\mathbf{B}_i \in \mathbb{R}^{p_s \times p_s}$. We call this the Kronecker PCA (KronPCA) covariance representation.

This type of model (with $r > 1$) has been used successfully in various applications by us and others, including

- video modeling (*Greenewald et al.*, 2013)
- gender classification in video (*Greenewald and Hero*, 2014a)
- sensor network anomaly detection (*Greenewald and Hero*, 2014b)
- gene expression modeling (Chapter III)
- wind speed modeling (*Tsiligkaridis and Hero*, 2013)
- synthetic aperture radar (*Rucci et al.*, 2010)
- MEG/EEG covariance modeling (*De Munck et al.*, 2002, 2004; *Bijma et al.*, 2005; *Jun et al.*, 2006).

In (*Loan and Pitsianis*, 1993) it was shown that any covariance matrix can be represented in this form with sufficiently large r. This allows for more accurate approximation of the covariance when it is not in Kronecker product form but most of its energy can be accounted for by a few Kronecker components. An algorithm (Permuted Rank-penalized Least Squares (PRLS)) for fitting the model (2.3) to a measured sample covariance matrix was introduced in (*Tsiligkaridis and Hero*, 2013) and was shown to have strong high dimensional MSE performance guarantees.

The basic Kronecker PCA model does not naturally accommodate additive noise since the diagonal elements (variances) must conform to the Kronecker structure of the matrix. To address this issue, in (*Greenewald et al.*, 2013; *Greenewald and Hero*, 2014b) we extended this KronPCA model, and the PRLS algorithm of (*Tsiligkaridis and Hero*, 2013), by adding a structured diagonal matrix to (2.3). This model is called Diagonally Loaded Kronecker PCA (DL-KronPCA) and, although it has an additional $p_s p_t$ parameters, it was shown that for fixed r it performs significantly better for inverse covariance estimation in cases where there is additive measurement noise (*Greenewald et al.*, 2013).

$$\mathbf{\Sigma} = \left(\sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i\right) + \mathbf{U} = \mathbf{\Theta} + \mathbf{U}, \tag{2.4}$$

where the diagonal matrix $\mathbf{U}$ is called the "diagonal loading matrix."

Chapter III extends DL-KronPCA to the case where $\mathbf{U}$ in (2.4) is a sparse loading matrix that is not necessarily diagonal. In other words, we model the covariance as the sum of a low separation rank matrix $\mathbf{\Theta}$ and a sparse matrix $\mathbf{\Gamma}$:

$$\mathbf{\Sigma} = \left(\sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i\right) + \mathbf{\Gamma} = \mathbf{\Theta} + \mathbf{\Gamma}. \tag{2.5}$$

DL-KronPCA is obviously a special case of this model. The motivation behind the extension (2.5) is that while the KronPCA model (2.3) may provide a good fit to most entries in $\mathbf{\Sigma}$, there are sometimes a few variables (or correlations) that cannot be well modeled using KronPCA, due to complex non-Kronecker structured covariance patterns, e.g. sparsely or uncorrelated additive noise, sensor failure, or corruption. Thus, inclusion of a sparse term in (2.5) allows for a better fit with lower separation rank r, thus reducing the overall number of model parameters. This notion of adding a sparse correction term to a regularized covariance estimate is found in the Robust PCA literature, where it is used to allow for more robust and parsimonious approx-

imation to data matrices (*Chandrasekaran et al.*, 2009, 2010; *Candès et al.*, 2011; *Yang and Ravikumar*, 2013). Robust KronPCA differs from Robust PCA in that it replaces the outer product with the Kronecker product. Sparse correction strategies have also been applied in the regression setting where the sparsity is applied to the first moments instead of the second moments (*Peng et al.*, 2010; *Otazo et al.*, 2014).

The model (2.5) is called the Robust Kronecker PCA (Robust KronPCA) model, and we propose regularized least squares based estimation algorithms for fitting the model. In particular, we propose a singular value thresholding (SVT) approach using a rearranged nuclear norm.

We derive high dimensional consistency results for the SVT-based algorithm that specify the MSE tradeoff between covariance dimension and the number of samples. Specifically, for n training samples and $\mathbf{\Theta}$ a sum of r Kronecker products, we derive non-asymptotic results that show that

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F = O\left(\max\left\{r\frac{p_t^2 + p_s^2 + \log M}{n}, \sqrt{r\frac{p_t^2 + p_s^2 + \log M}{n}}, \sqrt{\frac{s\log p_t p_s}{n}}\right\}\right)$$

as opposed to the unstructured SCM rate of $\|\widehat{\mathbf{\Sigma}}_{SCM} - \mathbf{\Sigma}\|_F = O\left(\sqrt{\frac{p_s^2 p_t^2}{n}}\right)$.

We also allow for the enforcement of a temporal block Toeplitz constraint, which corresponds to a temporally stationary covariance and results in a further reduction in the number of parameters when the process under consideration is temporally stationary and the time samples are uniformly spaced. This gives the improved rate

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F = O\left(\max\left\{r\frac{2p_t + p_s^2 + \log M}{n}, \sqrt{r\frac{2p_t + p_s^2 + \log M}{n}}, \sqrt{\frac{s\log p_t p_s}{n}}\right\}\right)$$

2.1.2 Kronecker STAP

In chapter IV, we consider a practical application for which Kronecker structure arises naturally from the physical model, and enables significant performance improvements.

In computer vision, the detection (and tracking) of moving objects is an important task for scene understanding, as motion often indicates human related activity (*Newstadt et al.*, 2014). Radar sensors are uniquely suited for this task, as object motion can be discriminated via the Doppler effect. We propose a spatio-temporal decomposition method of detecting ground based moving objects in airborne Synthetic Aperture Radar (SAR) imagery, also known as SAR GMTI (SAR Ground Moving Target Indication).

Radar moving target detection modalities include MTI radars (*Newstadt et al.*, 2014; *Ender*, 1999), which use a low carrier frequency and high pulse repetition frequency to directly detect Doppler shifts. This approach has significant disadvantages, however, including low spatial resolution, small imaging field of view, and the inability to detect stationary or slowly moving targets. The latter deficiency means that objects that move, stop, and then move are often lost by a tracker.

SAR, on the other hand, typically has extremely high spatial resolution and can be used to image very large areas, e.g. multiple square miles in the Gotcha data collection (*Scarborough et al.*, 2009). As a result, stationary and slowly moving objects are easily detected and located (*Ender*, 1999; *Newstadt et al.*, 2014). Doppler, however, causes smearing and azimuth displacement of moving objects (*Jao*, 2001), making them difficult to detect when surrounded by stationary clutter. Increasing the number of pulses (integration time) simply increases the amount of smearing instead of improving detectability (*Jao*, 2001). Several methods have thus been developed for detecting and potentially refocusing (*Cristallini et al.*, 2013; *Cerutti-Maori et al.*, 2012) moving targets in clutter. Our goal is to remove the disadvantages of MTI and

SAR by combining their strengths (the ability to detect Doppler shifts and high spatial resolution) using space time adaptive processing (STAP) with a novel Kronecker product spatio-temporal covariance model, as explained below.

SAR systems can either be single channel (standard single antenna system) or multichannel. Standard approaches for the single channel scenario include autofocusing (*Fienup*, 2001) and velocity filters. Autofocusing works only in low clutter, however, since it may focus the clutter instead of the moving target (*Fienup*, 2001; *Newstadt et al.*, 2014). Velocity filterbank approaches used in track-before-detect processing (*Jao*, 2001) involve searching over a large velocity/acceleration space, which often makes computational complexity excessively high. Attempts to reduce the computational complexity have been proposed, e.g. via compressive sensing based dictionary approaches (*Khwaja and Ma*, 2011) and Bayesian inference (*Newstadt et al.*, 2014), but remain computationally intensive.

Multichannel SAR has the potential for greatly improved moving target detection performance (*Ender*, 1999; *Newstadt et al.*, 2014). Standard multiple channel configurations include spatially separated arrays of antennas, flying multiple passes (change detection), using multiple polarizations, or combinations thereof (*Newstadt et al.*, 2014).

Several techniques exist for using multiple radar channels (antennas) to separate the moving targets from the stationary background. SAR GMTI systems have an antenna configuration such that each antenna transmits and receives from approximately the same location but at slightly different times (*Scarborough et al.*, 2009; *Ender*, 1999; *Newstadt et al.*, 2014; *Cerutti-Maori et al.*, 2012). Along track interferometry (ATI) and displaced phase center array (DPCA) are two classical approaches (*Newstadt et al.*, 2014) for detecting moving targets in SAR GMTI data, both of which are applicable only to the two channel scenario. Both ATI and DPCA first form two SAR images, each image formed using the signal from one of the antennas.

To detect the moving targets, ATI thresholds the phase difference between the images and DPCA thresholds the magnitude of the difference. A Bayesian approach using a parametric cross channel covariance generalizing ATI/DPCA to p channels was developed in (*Newstadt et al.*, 2014), and a unstructured method fusing STAP and a test statistic in (*Cerutti-Maori et al.*, 2012). Space-time Adaptive Processing (STAP) learns a spatio-temporal covariance from clutter training data, and uses these correlations to filter out the stationary clutter while preserving the moving target returns (*Ender*, 1999; *Ginolhac et al.*, 2014; *Klemm*, 2002).

In this chapter, we focus on the SAR GMTI configuration and propose a covariance-based STAP algorithm with a customized Kronecker product covariance structure. The SAR GMTI receiver consists of an array of p phase centers (antennas) processing q pulses in a coherent processing interval. Define the array $\mathbf{X}^{(m)} \in \mathbb{C}^{p\times q}$ such that $X_{ij}^{(m)}$ is the radar return from the jth pulse of the ith channel in the mth range bin. Let $\mathbf{x}_m = \text{vec}(\mathbf{X}^{(m)})$. The target-free radar data $\mathbf{x}_m$ is complex valued and is assumed to have zero mean. Define

$$\mathbf{\Sigma} = \text{Cov}[\mathbf{x}] = E[\mathbf{x}\mathbf{x}^H]. \tag{2.6}$$

The training samples, denoted as the set $\mathcal{S}$, used to estimate the SAR covariance $\mathbf{\Sigma}$ are collected from n representative range bins. The sample covariance matrix (SCM) is given by

$$\mathbf{S} = \frac{1}{n}\sum_{m\in\mathcal{S}} \mathbf{x}_m\mathbf{x}_m^H. \tag{2.7}$$

If n is small, $\mathbf{S}$ may be rank deficient or ill-conditioned (*Newstadt et al.*, 2014; *Ginolhac et al.*, 2014; *Greenewald et al.*, 2013; *Greenewald and Hero*, 2014b), and it can be shown that using the SCM directly for STAP requires a number n of training samples that is at least twice the dimension pq of $\mathbf{S}$ (*Reed et al.*, 1974). In this data rich

case, STAP performs well (*Newstadt et al.*, 2014; *Ender*, 1999; *Ginolhac et al.*, 2014). However, with p antennas and q time samples (pulses), the dimension pq of the covariance is often very large, making it difficult to obtain a sufficient number of target-free training samples. This so-called "small n large pq" problem leads to severe instability and overfitting errors, compromising STAP cancelation performance.

By introducing structure and/or sparsity into the covariance matrix, the number of parameters and the number of samples required to estimate them can be reduced. As the spatiotemporal clutter covariance $\mathbf{\Sigma}$ is low rank (*Brennan and Staudaher*, 1992; *Ginolhac et al.*, 2014; *Rangaswamy et al.*, 2004; *Ender*, 1999), Low Rank STAP (LR-STAP) clutter cancelation estimates a low rank clutter subspace from $\mathbf{S}$ and uses it to estimate and remove the rank r clutter component in the data (*Bazi et al.*, 2005; *Ginolhac et al.*, 2014), reducing the number of parameters from $O(p^2q^2)$ to $O(rpq)$. Efficient algorithms, including some involving subspace tracking, have been proposed (*Belkacemi and Marcos*, 2006; *Shen et al.*, 2009). Other methods adding structural constraints such as persymmetry (*Ginolhac et al.*, 2014; *Conte and De Maio*, 2003), and robustification to outliers either via exploitation of the SIRV model (*Ginolhac et al.*, 2009) or adaptive weighting of the training data (*Gerlach and Picciolo*, 2011) have been proposed. Fast approaches based on techniques such as Krylov subspace methods (*Goldstein et al.*, 1998; *Honig and Goldstein*, 2002; *Pados et al.*, 2007; *Scharf et al.*, 2008) and adaptive filtering (*Fa and De Lamare*, 2011; *Fa et al.*, 2010) exist. All of these techniques remain sensitive to outlier or moving target corruption of the training data, and generally still require large training sample sizes (*Newstadt et al.*, 2014).

Instead, for SAR GMTI we propose to exploit the explicit space-time arrangement of the covariance by modeling the clutter covariance matrix $\mathbf{\Sigma}_c$ as the Kronecker

product of two smaller matrices

$$\mathbf{\Sigma}_c = \mathbf{A} \otimes \mathbf{B}, \tag{2.8}$$

where $\mathbf{A} \in \mathbb{C}^{p \times p}$ is rank 1 and $\mathbf{B} \in \mathbb{C}^{q \times q}$ is low rank. In this setting, the $\mathbf{B}$ matrix is the "temporal (pulse) covariance" and $\mathbf{A}$ is the "spatial (antenna) covariance."

In this work, an iterative L2 based algorithm is proposed to directly estimate the low rank Kronecker factors from the observed sample covariance. We then introduce the Kron STAP filter, which projects away both the spatial and temporal clutter subspaces. This projects away a higher dimensional subspace than does LR-STAP, thereby achieving improved noise and clutter cancelation. Theoretical results indicate significantly fewer training samples are required to achieve high signal-to-interference-plus-noise ratio (SINR), specifically that the SINR loss ρ is $O(1-\frac{1}{n})$, as opposed to the standard LR-STAP rate of $O(1-\frac{r}{n})$ where $r \sim q$ is the rank of the clutter subspace.

It is also shown that the proposed approach improves robustness to corrupted training data. Critically, robustness allows significant numbers of moving targets to remain in the training set. Finally, we apply KronSTAP to both synthetic data and a real SAR dataset, demonstrating significant performance improvement as expected.

2.1.3 Tensor Graphical Lasso (TeraLasso)

So far, we have focused on covariance modeling for matrix-valued data. In Chapter V, we focus on developing structured, sparse inverse covariance models for high-dimensional tensor data. Consider the K-order data tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ (*Kolda and Bader*, 2009). For convenience, define a dimension vector $\mathbf{p} = [d_1, \ldots, d_K]$, and set

$$p = \prod_{k=1}^{K} d_k, \quad m_k = \prod_{i \neq k} d_i = \frac{p}{d_k}.$$

We propose the TEnsor gRAphical Lasso (TeraLasso) model

$$(\text{Cov}[\text{vec}(X^T)])^{-1} = \Sigma^{-1} = \Omega = \underbrace{\Psi_1 \oplus \cdots \oplus \Psi_K}_{K \text{ terms}}, \tag{2.9}$$

where the Ψ_k are sparse, each corresponding to a graph across the kth dimension of the data tensor. We have used $\text{vec}(X)$ as in (*Kolda and Bader*, 2009), and defined $X^T \in \mathbb{R}^{d_K \times \cdots \times d_1}$ by analogy to the matrix transpose, i.e. $[X^T]_{i_1,\ldots,i_K} = X_{i_K,\ldots,i_1}$.

Many methods for first-moment modeling of tensor-valued data have been proposed (*Kolda and Bader*, 2009). Many of these involve low-rank factor decompositions, including PARAFAC and CANDECOMP (*Harshman and Lundy*, 1994; *Faber et al.*, 2003) and Tucker decomposition-based methods such as (*Tucker*, 1966) and (*Hoff et al.*, 2016). Recently, several works have found that such modeling can be improved by taking into account the second moment of the data (i.e. covariance), which so far is typically modeled using Kronecker products (*Xu et al.*, 2011; *Zhe et al.*, 2015; *Pouryazdian et al.*, 2016).

As the second moment, or covariance, encodes relationships and interactions between variables, it is a powerful tool for modeling multivariate distributions, allowing inference, likelihood calculation, and prediction. For tensor-valued data, however, the very large number of variables prohibits the estimation and use of the $O(\prod_{k=1}^{K} d_k^2)$ element unstructured covariance in many situations. As a result, there has been increasing interest in developing structured covariance models appropriate for matrix- and tensor-valued data (*Tsiligkaridis and Hero*, 2013; *Zhou*, 2014; *Werner et al.*, 2008; *Sun et al.*, 2015; *Xu et al.*, 2011; *Greenewald and Hero*, 2015; *Allen and Tibshirani*, 2010). As we have discussed in earlier chapters, the Kronecker product covariance

$$\Sigma = A_1 \otimes A_2 \otimes \cdots \otimes A_K \tag{2.10}$$

exploits the natural tensor arrangement of the variables and forms a joint model

from K models each along one tensor axis. When the covariance (2.10) describes a Gaussian distribution, this model is known as the *matrix normal* distribution (*Dawid*, 1981) ($K = 2$) and *tensor normal* for $K > 2$ (*Sun et al.*, 2015; *Xu et al.*, 2011), with a penalized version in (*Allen and Tibshirani*, 2010) called the transposable covariance model.

In a Gaussian graphical model, edges correspond to nonzero entries in the precision matrix $\Omega = \Sigma^{-1}$. The Kronecker product graphical model (*Zhou*, 2014; *Tsiligkaridis et al.*, 2013; *Sun et al.*, 2015) estimates K sparse factor precision matrices $\Psi_k \in \mathbb{R}^{d_k \times d_k}$ (Figure 5.1(a-c)), setting $\Omega = \Psi_1 \otimes \cdots \otimes \Psi_K$. This model has excellent statistical convergence results, but creates an overall graph where each edge in the final model is the product of K separate edges from the factor graphs Ψ_k. A proliferation of inter-related edges is thus created, illustrated in Figure 5.1 (right), with each edge in the factor models affecting up to m_k^2 total edges in the final graph.

Instead of the Kronecker product, one may imagine that it could be desirable to have each edge in the factor models map directly to edges in the final model. One such model, shown in Figure 2.1 (left) would map the i, jth edge in the kth factor Ψ_k to edges between nodes in ith and jth position along the kth tensor mode. This type of structure implies that conditional dependence moves along axes or modes, in a manner analogous to several popular forms of Markov random fields used in computer vision and other applications (*Wang et al.*, 2013; *Diebel and Thrun*, 2005).

In order to estimate sparse Kronecker sum precision matrices, we derive a joint, convex objective function that is an L1 penalization of the joint maximum likelihood function:

$$Q(\Psi_1, \ldots, \Psi_K) = -\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \sum_{k=1}^{K} m_k \left(\langle S_k, \Psi_k \rangle + \rho_k |\Psi_k^-|_1 \right), \qquad (2.11)$$

where $\langle A, B \rangle = \mathrm{tr}(A^T B)$. Due to the Kronecker sum structure, we have been able to

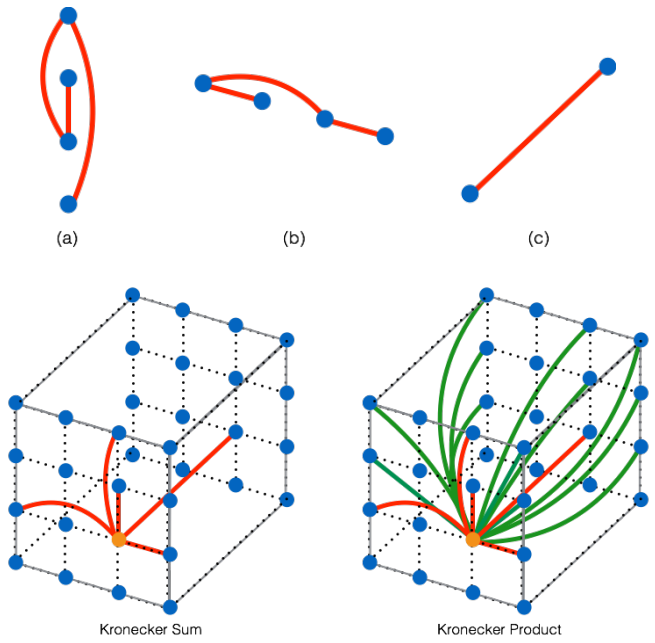


Figure 2.1: Tensor graphical models on a $4 \times 4 \times 2$ Cartesian node grid. Consider three graphical models, one along each axis (a-c). The Kronecker sum and Kronecker product of these graphs are computed, with only the edges emanating from the orange node shown. The Kronecker sum (64 total edges) preserves the sparsity of the axis graphs (a-c), forming a joint model where each edge is associated with a single edge in an axis graph. The Kronecker product (184 total edges), on the other hand, creates an explosion of edges (marked green) each with a weight a multiple of three separate weights from the axis graphs. Hence, we argue that in many situations the Kronecker sum is a more natural and interpretable tensor expansion of sparse graphical models.

eliminate all inner products involving the full sample covariance, replacing them with $d_k \times d_k$ inner products.

Using this form of the objective, we derive a fast, scalable first-order algorithm that is guaranteed to geometrically converge to the global minimum of the objective. This algorithm enjoys a computational cost of $O(p + d_k^3)$ per iteration, as compared to the $O(p^3)$ computation per iteration required for the GLasso.

We also derive nonasymptotic statistical convergence rates, that imply the follow-

ing bounds.

$$\|\widehat{\Omega} - \Omega\|_F^2 = O_p\left((s+p)\frac{\log p}{n \min_k m_k}\right),$$
$$\sum_{k=1}^{K} \frac{\|\Delta_k^-\|_F^2}{d_k} = O_p\left(\left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right)\frac{\log p}{n \min_k m_k}\right),$$
$$\|\widehat{\Omega} - \Omega\|_2 = O_p\left(\sqrt{\left(\max_k \frac{d_k}{m_k}\right)\left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right)\frac{\log p}{n}}\right),$$

where $\Delta_k^- = \widehat{\Psi}_k^- - \Psi_{0,k}^-$.

Observe that factorwise and spectral norm convergence in the fixed sample regime holds whenever $s_k \leq O(d_k)$ and $m_k > d_k$. Finally, we confirm these rates by applying this estimator to synthetic data, as well as to meteorological and EEG seizure prediction datasets.

2.2 Strongly Adaptive Online Metric Learning

The effectiveness of many machine learning and data mining applications rely on an appropriate measure of pairwise distance between data points that accurately reflects the objective, e.g., prediction, clustering or classification. In settings with clean, appropriately-scaled spherical Gaussian data, standard Euclidean distance can be utilized. However, when the data is heavy tailed, multimodal, contaminated by outliers, irrelevant or replicated features, or observation noise, Euclidean inter-point distance can be problematic, leading to bias or loss of discriminative power.

Many unsupervised, data-driven approaches for identifying appropriate distances between points have been proposed. These methodologies, broadly taking the form of dimensionality reduction or data "whitening", aim to utilize the data itself to learn a transformation of the data that embeds it into a space where Euclidean distance is appropriate. Examples of such unsupervised techniques include Principal Component

Analysis (*Bishop*, 2006), Multidimensional Scaling (*Hastie et al.*, 2005), covariance estimation (*Hastie et al.*, 2005; *Bishop*, 2006), and manifold learning (*Lee and Verleysen*, 2007). Such unsupervised methods do not have the benefit of human input on the distance metric, and rely on prior assumptions, e.g., local linearity or smoothness.

For unimodal Gaussian data, the ideal metric is the Mahalonobis distance based on the inverse covariance, i.e. $d^2(\mathbf{x}, \mathbf{z}) = (\mathbf{x} - \mathbf{z})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{z})$. The generalized Mahalonobis distance is parameterized by $\mathbf{M}$ as

$$d_M^2(\mathbf{x}, \mathbf{z}) = (\mathbf{x} - \mathbf{z})^T \mathbf{M} (\mathbf{x} - \mathbf{z}) \tag{2.12}$$

where $\mathbf{M} \in \mathbb{R}^{n \times n} \succeq 0$. Our goal in this chapter is a generalized "inverse covariance" estimation, finding a regularized matrix $\mathbf{M}$ that *best separates clusters* rather than simply *best explaining variation*.

In other words, one seeks to learn linear transformations of the data that are well matched to a particular task specified by the user. In this case, point labels or constraints indicating point similarity or dissimilarity are used to learn a transformation of the data such that similar points are "close" to one another and dissimilar points are distant in the transformed space. Learning distance metrics in this manner allows a more precise notion of distance or similarity to be defined that is related to the task at hand.

Many supervised and semi-supervised distance metric learning approaches have been developed (*Kulis*, 2012). This includes online algorithms (*Kunapuli and Shavlik*, 2012) with regret guarantees for situations where similarity constraints are received in a stream. In this chapter, we propose a new way of formulating the distance metric learning task. We assume the underlying ground-truth distance metric from which constraints are generated is evolving over time. This problem formulation suggests an adaptive, online approach to track the underlying metric ("inverse covariance")

as constraints are received. We present an algorithm for estimating time-varying distance metrics inspired by recent advances in composite objective mirror descent for metric learning (*Duchi et al.*, 2010b) (COMID) and a framework proposed for handling discrete nonstationarities (*Daniely et al.*, 2015). We call our framework for strongly adaptive, parameter-free handling of all types of dynamic nonstationarities Online Convex Ensemble StrongLy Adaptive Dynamic Learning, or OCELAD. This framework is widely applicable beyond metric learning.

Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are classic examples of linear transformations for projecting data into more interpretable low dimensional spaces. Unsupervised PCA seeks to identify a set of axes that best explain the variance contained in the data. LDA takes a supervised approach, minimizing the intra-class variance and maximizing the inter-class variance given class labeled data points.

Much of the recent work in Distance Metric Learning has focused on learning Mahalanobis distances on the basis of pairwise similarity/dissimilarity constraints. These methods have the same goals as LDA; pairs of points labeled "similar" should be close to one another while pairs labeled "dissimilar" should be distant. MMC (*Xing et al.*, 2002), a method for identifying a Mahalanobis metric for clustering with side information, uses semidefinite programming to identify a metric that maximizes the sum of distances between points labeled with different classes subject to the constraint that the sum of distances between all points with similar labels be less than some constant. Large Margin Nearest Neighbor (LMNN) (*Weinberger et al.*, 2005) similarly uses semidefinite programming to identify a Mahalanobis distance, however it modifies the constraints to only take into account a small, local neighborhood for each point. Information Theoretic Metric Learning (ITML) (*Davis et al.*, 2007) is another popular Distance Metric Learning technique. ITML minimizes the Kullback-Liebler divergence between an initial guess of the matrix that parameterizes the Mahalanobis

distance and a solution that satisfies a set of constraints.

In a dynamic environment, it is necessary to be able to compute multiple estimates of the changing metric at different times, and to be able to compute those estimates online. Online learning (*Cesa-Bianchi and Lugosi*, 2006) meets these criteria by efficiently updating the estimate every time a new data point is obtained, instead of solving an objective function formed from the entire dataset.

Many online learning methods have regret guarantees, that is, the loss in performance relative to a batch method is provably small (*Cesa-Bianchi and Lugosi*, 2006; *Duchi et al.*, 2010b). In practice, however, the performance of an online learning method is strongly influenced by the learning rate which may need to vary over time in a dynamic environment (*Daniely et al.*, 2015; *McMahan and Streeter*, 2010; *Duchi et al.*, 2010a).

Adaptive online learning methods attempt to address this problem by continuously updating the learning rate as new observations become available. For example, AdaGrad-style methods (*McMahan and Streeter*, 2010; *Duchi et al.*, 2010a) perform gradient descent steps with the step size adapted based on the magnitude of recent gradients. Follow the regularized leader (FTRL) type algorithms adapt the regularization to the observations (*McMahan*, 2014). Recently, a method called Strongly Adaptive Online Learning (SAOL) has been proposed, which maintains several learners with different learning rates and selects the best one based on recent performance (*Daniely et al.*, 2015). Several of these adaptive methods have provable regret bounds (*McMahan*, 2014; *Herbster and Warmuth*, 1998; *Hazan and Seshadhri*, 2007). These typically guarantee low total regret (i.e. regret from time 0 to time t) at every time (*McMahan*, 2014). SAOL, on the other hand, is guaranteed to have low regret on every subinterval, as well as low regret overall (*Daniely et al.*, 2015).

We formulate the metric learning problem as a cooperative dynamic game between the learner and the analyst. Both players' goal is for the learner to learn the internal

metric $\mathbf{M}$ used by the analyst. The metric is changing over time, making the game dynamic. The analyst selects pairs of data points $(\mathbf{x}_t, \mathbf{z}_t)$ in $\mathbb{R}^n$ and labels them as similar ($y_t = 1$) or dissimilar ($y_t = -1$). The labels are assumed to arrive in a temporal sequence, hence the labels at the beginning may have arisen from a different metric than those at the end of the sequence.

Following (*Kunapuli and Shavlik*, 2012), we introduce the following margin based constraints:

$$\begin{aligned} d_M^2(\mathbf{x}_t, \mathbf{z}_t) &\leq \mu - 1, \quad \forall\{t|y_t = 1\} \\ d_M^2(\mathbf{x}_t, \mathbf{z}_t) &\geq \mu + 1, \quad \forall\{t|y_t = -1\} \end{aligned} \tag{2.13}$$

where μ is a threshold that controls the margin between similar and dissimilar points. A diagram illustrating these constraints and their effect is shown in Figure 6.2.

These constraints are softened by penalizing violation of the constraints with a convex loss function ℓ_t. This gives the following objective:

$$\begin{aligned} &\min_{\mathbf{M}\succeq 0, \mu \geq 1} \frac{1}{T} \sum_{t=1}^{T} \ell_t(\mathbf{M}, \mu) + \rho r(\mathbf{M}) \\ &\ell_t(\mathbf{M}, \mu) = \ell(m_t),\ m_t = y_t(\mu - \mathbf{u}_t^T \mathbf{M} \mathbf{u}_t),\ \mathbf{u}_t = \mathbf{x}_t - \mathbf{z}_t \end{aligned} \tag{2.14}$$

where r is the regularizer. Kunapuli and Shavlik propose using nuclear norm regularization ($r(\mathbf{M}) = \|\mathbf{M}\|_*$) to encourage projection of the data onto a low dimensional subspace (feature selection/dimensionality reduction). Other regularization such as sparsity promoting 1-norm ($r(\mathbf{M}) = \|\mathbf{M}\|_1$) is also possible.

This objective can be solved online via Composite Objective Mirror Descent (COMID) (*Kunapuli and Shavlik*, 2012). COMID, however, requires setting a learning rate that decays to zero, thus preventing any further learning after a certain point. In

order to be able to track an arbitrarily changing metric, we use the Strongly Adaptive Online Learning framework, which at every time point maintains a set of learners operating at different learning rates and adaptively selects the most appropriate one based on recent performance. We develop an efficient implementation of SAOL with COMID learners, and apply it to the metric learning problem.

The performance of online algorithms is often quantified using regret bounds, which measure the loss relative to a static batch estimate. A useful generalization of the standard static regret to the dynamic case is as follows. Let $\mathbf{w} = \{\theta_t\}_{t\in[0,T]}$ be an arbitrary sequence of parameters. Then, the *dynamic regret* of an algorithm $\mathcal{B}$ relative to a comparator sequence $\mathbf{w}$ on the interval I is defined as

$$R_{\mathcal{B},\mathbf{w}}(I) = \sum_{t\in I} f_t(\widehat{\theta}_t) - \sum_{t\in I} f_t(\theta_t), \tag{2.15}$$

where $\widehat{\theta}_t$ are generated by $\mathcal{B}$. This allows for a dynamically changing estimate.

In (*Hall and Willett*, 2015) the authors derive dynamic regret bounds that hold over all possible sequences $\mathbf{w}$ such that $\sum_{t\in I} \|\theta_{t+1} - \theta_t\| \leq \gamma$, i.e. bounding the total amount of variation in the estimated parameter. Without this temporal regularization, minimizing the loss would cause θ_t to grossly overfit. In this sense, setting the comparator sequence $\mathbf{w}$ to the "ground truth sequence" or "batch optimal sequence" both provide meaningful intuitive bounds.

Strongly adaptive regret bounds *Daniely et al.* (2015) have claimed that static regret is low on every subinterval, instead of only low in the aggregate. We use the notion of dynamic regret to introduce strongly adaptive dynamic regret bounds, proving that *dynamic regret is low on every subinterval* $I \subseteq [0,T]$ *simultaneously.* Suppose there are a sequence of random loss functions $\ell_t(\theta_t)$. The goal is to estimate a sequence $\widehat{\theta}_t$ that minimizes the dynamic regret.

Theorem II.1. *Let* $\mathbf{w} = \{\theta_1, \ldots, \theta_T\}$ *be an arbitrary sequence of parameters and*

define $\gamma_{\mathbf{w}}(I) = \sum_{q \le t < s} \|\theta_{t+1} - \theta_t\|$ *as a function of* $\mathbf{w}$ *and an interval* $I = [q, s]$. *Choose an ensemble of learners* $\mathcal{B}$ *such that given an interval* I *the learner* $\mathcal{B}_I$ *creates an output sequence* $\theta_t(I)$ *satisfying the dynamic regret bound*

$$R_{\mathcal{B}_I, \mathbf{w}}(I) \le C(1 + \gamma_{\mathbf{w}}(I))\sqrt{|I|} \tag{2.16}$$

for some constant $C > 0$. *Then the strongly adaptive dynamic learner* $OCELAD^{\mathcal{B}}$ *using* $\mathcal{B}$ *as the ensemble creates an estimation sequence* $\widehat{\theta}_t$ *satisfying*

$$R_{OCELAD^{\mathcal{B}}, \mathbf{w}}(I) \le 8C(1 + \gamma_{\mathbf{w}}(I))\sqrt{|I|} + 40\log(s+1)\sqrt{|I|}$$

on every interval $I = [q, s] \subseteq [0, T]$.

In a dynamic setting, bounds of this type are particularly desirable because they allow for changing *drift rate* and guarantee quick recovery from *discrete changes*. For instance, suppose K discrete switches (large parameter changes or changes in drift rate) occur at times t_i satisfying $0 = t_0 < t_1 < \cdots < t_K = T$. Then since $\sum_{i=1}^{K} \sqrt{|t_{i-1} - t_i|} \le \sqrt{KT}$, this implies that the total expected dynamic regret on $[0, T]$ remains low ($O(\sqrt{KT})$), while simultaneously guaranteeing that an appropriate learning rate is achieved on each subinterval $[t_i, t_{i+1}]$.

We are able to confirm the superiority of the adaptive approach via simulation and experiments on highly nonstationary Twitter data.

CHAPTER III

Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation

In this chapter, we develop a method for robust estimation of spatio-temporal covariances and apply it to multivariate time series modeling and parameter estimation. We exploit Kronecker structure by assuming the covariance can be expressed as a sum of r Kronecker products and a sparse term.

3.1 Introduction

Let $\mathbf{X}$ be a $p_s \times p_t$ matrix with entries $\widetilde{x}(m,t)$ denoting samples of a space-time random process defined over a p_s-grid of spatial samples $m \in \{1,\ldots,p_s\}$ and a p_t-grid of time samples $t \in \{1,\ldots,p_t\}$. Let $\mathbf{x} = \mathrm{vec}(\mathbf{X})$ denote the $p_t p_s$ column vector obtained by lexicographical reordering. Define the $p_t p_s \times p_t p_s$ spatiotemporal covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}[\mathbf{x}]$.

As $p_s p_t$ can be very large, even for moderately large p_s and p_t the number of degrees of freedom $(p_s p_t(p_s p_t + 1)/2)$ in the covariance matrix can greatly exceed the number n of training samples available to estimate the covariance matrix. One way to handle this problem is to introduce structure and/or sparsity into the covariance matrix, thus reducing the number of parameters to be estimated. A natural non-

sparse option is to introduce structure by modeling the covariance matrix $\mathbf{\Sigma}$ as the Kronecker product of two smaller symmetric positive definite matrices, i.e.

$$\mathbf{\Sigma} = \mathbf{A} \otimes \mathbf{B}. \tag{3.1}$$

An extension to the representation (5.2), discussed in (*Tsiligkaridis and Hero*, 2013), approximates the covariance matrix using a sum of Kronecker product factors

$$\mathbf{\Sigma} = \sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i, \tag{3.2}$$

where r is the separation rank, $\mathbf{A}_i \in \mathbb{R}^{p_t \times p_t}$, and $\mathbf{B}_i \in \mathbb{R}^{p_s \times p_s}$. We call this the Kronecker PCA (KronPCA) covariance representation.

In (*Loan and Pitsianis*, 1993) it was shown that any covariance matrix can be represented in this form with sufficiently large r. This allows for more accurate approximation of the covariance when it is not in Kronecker product form but most of its energy can be accounted for by a few Kronecker components. An algorithm (Permuted Rank-penalized Least Squares (PRLS)) for fitting the model (5.5) to a measured sample covariance matrix was introduced in (*Tsiligkaridis and Hero*, 2013) and was shown to have strong high dimensional MSE performance guarantees. It should also be noted that, as contrasted to standard PCA, KronPCA accounts specifically for spatio-temporal structure, often provides a full rank covariance, and requires significantly fewer components (Kronecker factors) for equivalent covariance approximation accuracy. Naturally, since it compresses covariance onto a more complex (Kronecker) basis than PCA's singular vector basis, the analysis of Kron-PCA estimation performance is more complicated.

The standard Kronecker PCA model does not naturally accommodate additive noise since the diagonal elements (variances) must conform to the Kronecker structure of the matrix. To address this issue, in (*Greenewald et al.*, 2013) we extended this

KronPCA model, and the PRLS algorithm of (*Tsiligkaridis and Hero*, 2013), by adding a structured diagonal matrix to (5.5). This model is called Diagonally Loaded Kronecker PCA (DL-KronPCA) and, although it has an additional $p_s p_t$ parameters, it was shown that for fixed r it performs significantly better for inverse covariance estimation in cases where there is additive measurement noise (*Greenewald et al.*, 2013).

The DL-KronPCA model (*Greenewald et al.*, 2013) is the $r+1$-Kronecker model

$$\mathbf{\Sigma} = \left(\sum\nolimits_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i\right) + \mathbf{U} = \mathbf{\Theta} + \mathbf{U}, \tag{3.3}$$

where the diagonal matrix $\mathbf{U}$ is called the "diagonal loading matrix." Following Pitsianis-VanLoan rearrangement of the square $p_t p_s \times p_t p_s$ matrix $\mathbf{\Sigma}$ to an equivalent rectangular $p_s^2 \times p_t^2$ matrix (*Tsiligkaridis and Hero*, 2013; *Werner et al.*, 2008), this becomes an equivalent matrix approximation problem of finding a low rank plus diagonal approximation (*Greenewald et al.*, 2013; *Tsiligkaridis and Hero*, 2013). The DL-KronPCA estimation problem was posed in (*Greenewald and Hero*, 2014b; *Greenewald et al.*, 2013) as the rearranged nuclear norm penalized Frobenius norm optimization

$$\min_{\mathbf{\Sigma}} \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{SCM}\|_F^2 + \lambda \|\mathcal{R}(\mathbf{\Theta})\|_* \tag{3.4}$$

where the minimization is over $\mathbf{\Sigma}$ of the form (3.3), $\mathcal{R}(\cdot)$ is the Pitsianis-VanLoan rearrangement operator defined in the next section, and $\|\cdot\|_*$ is the nuclear norm. A weighted least squares solution to this problem is given in (*Greenewald et al.*, 2013; *Greenewald and Hero*, 2014b).

This work extends DL-KronPCA to the case where $\mathbf{U}$ in (3.3) is a sparse loading matrix that is not necessarily diagonal. In other words, we model the covariance as

the sum of a low separation rank matrix $\mathbf{\Theta}$ and a sparse matrix $\mathbf{\Gamma}$:

$$\mathbf{\Sigma} = \left(\sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i\right) + \mathbf{\Gamma} = \mathbf{\Theta} + \mathbf{\Gamma}. \tag{3.5}$$

DL-KronPCA is obviously a special case of this model. The motivation behind the extension (3.5) is that while the KronPCA models (5.5) and (3.3) may provide a good fit to most entries in $\mathbf{\Sigma}$, there are sometimes a few variables (or correlations) that cannot be well modeled using KronPCA, due to complex non-Kronecker structured covariance patterns, e.g. sparsely correlated additive noise, sensor failure, or corruption. Thus, inclusion of a sparse term in (3.5) allows for a better fit with lower separation rank r, thus reducing the overall number of model parameters. In addition, if the underlying distribution is heavy tailed, sparse outliers in the sample covariance will occur, which will corrupt Kronecker product estimates (5.5) and (3.3) that don't have the flexibility of absorbing them into a sparse term. This notion of adding a sparse correction term to a regularized covariance estimate is found in the Robust PCA literature, where it is used to allow for more robust and parsimonious approximation to data matrices (*Chandrasekaran et al.*, 2009, 2010; *Candès et al.*, 2011; *Yang and Ravikumar*, 2013). Robust KronPCA differs from Robust PCA in that it replaces the outer product with the Kronecker product. KronPCA and PCA are useful for significantly different applications because the Kronecker product allows the decomposition of spatio-temporal processes into (full rank) spatio-temporally *separable* components, whereas PCA decomposes them into deterministic basis functions with no explicit spatio-temporal structure (*Tsiligkaridis and Hero*, 2013; *Greenewald et al.*, 2013; *Werner et al.*, 2008). Sparse correction strategies have also been applied in the regression setting where the sparsity is applied to the first moments instead of the second moments (*Peng et al.*, 2010; *Otazo et al.*, 2014).

The model (3.5) is called the Robust Kronecker PCA (Robust KronPCA) model,

and we propose regularized least squares based estimation algorithms for fitting the model. In particular, we propose a singular value thresholding (SVT) approach using the rearranged nuclear norm. However, unlike in robust PCA, the sparsity is applied to the Kronecker decomposition instead of the singular value decomposition. We derive high dimensional consistency results for the SVT-based algorithm that specify the MSE tradeoff between covariance dimension and the number of samples. Following (*Greenewald and Hero*, 2014b), we also allow for the enforcement of a temporal block Toeplitz constraint, which corresponds to a temporally stationary covariance and results in a further reduction in the number of parameters when the process under consideration is temporally stationary and the time samples are uniformly spaced. We illustrate our proposed robust Kronecker PCA method using simulated data and a yeast cell cycle dataset.

The rest of the chapter is organized as follows: in Section 3.2, we introduce our Robust KronPCA model and introduce an algorithm for estimating covariances described by it. Section 3.3 provides high dimensional convergence theorems. Simulations and an application to cell cycle data are presented in Section 4.5, and our conclusions are given in Section 4.6.

3.2 Robust KronPCA

Consider the model (3.5) for the covariance as the sum of a low separation rank matrix $\mathbf{\Theta}$ and a sparse matrix $\mathbf{\Gamma}$:

$$\mathbf{\Sigma} = \mathbf{\Theta} + \mathbf{\Gamma}. \tag{3.6}$$

Define $\mathbf{M}(i,j)$ as the i,jth $p_s \times p_s$ subblock of $\mathbf{M}$, i.e., $\mathbf{M}(i,j) = [M]_{(i-1)p_s+1:ip_s,(j-1)p_s+1:jp_s}$. The invertible Pitsianis-VanLoan rearrangement operator $\mathcal{R}(\cdot)$ maps $p_t p_s \times p_t p_s$ matrices to $p_t^2 \times p_s^2$ matrices and, as defined in (*Tsiligkaridis and Hero*, 2013; *Werner*

et al., 2008) sets the $(i-1)p_t + j$th row of $\mathcal{R}(\mathbf{M})$ equal to $\text{vec}(\mathbf{M}(i,j))^T$, i.e.

$$\mathcal{R}(\mathbf{M}) = [\ \mathbf{m}_1 \quad \dots \quad \mathbf{m}_{p_t^2}\ ]^T, \tag{3.7}$$
$$\mathbf{m}_{(i-1)p_t+j} = \text{vec}(\mathbf{M}(i,j)), \quad i,j = 1,\dots,p_t.$$

After Pitsianis-VanLoan rearrangement the expression (3.6) takes the form

$$\mathcal{R}(\mathbf{\Sigma}) = \sum\nolimits_{i=1}^{r} \mathbf{a}_i \mathbf{b}_i^T + \mathbf{S} = \mathbf{L} + \mathbf{S}, \tag{3.8}$$

where $\mathbf{a}_i = \text{vec}(\mathbf{A}_i)$ and $\mathbf{b}_i = \text{vec}(\mathbf{B}_i)$. In the next section we solve this Robust Kronecker PCA problem (low rank + sparse + noise) using sparse approximation, involving a nuclear and 1-norm penalized Frobenius norm loss on the rearranged fitting errors.

3.2.1 Estimation

Similarly to the approach of (*Greenewald et al.*, 2013; *Tsiligkaridis and Hero*, 2013), we fit the model (3.6) to the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{SCM} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$, where $\overline{\mathbf{x}}$ is the sample mean and n is the number of samples of the space time process $\mathbf{X}$. The best fit matrices $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{\Gamma}$ in (3.6) are determined by minimizing the objective function

$$\min_{\widehat{\mathbf{\Theta}},\widehat{\mathbf{\Gamma}}} \|\widehat{\mathbf{\Sigma}}_{SCM} - \widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Gamma}}\|_F^2 + \lambda_\Theta \|\mathcal{R}(\widehat{\mathbf{\Theta}})\|_* + \lambda_\Gamma \|\widehat{\mathbf{\Gamma}}\|_1. \tag{3.9}$$

We call the norm $\|\mathcal{R}(\mathbf{\Theta})\|_*$ the rearranged nuclear norm of $\mathbf{\Theta}$. The regularization parameters λ_Θ and λ_Γ control the importance of separation rank deficiency and sparsity, respectively, where increasing either increases the amount of regularization. The

objective function (3.9) is equivalent to the rearranged objective function

$$\min_{\widehat{\mathbf{L}},\widehat{\mathbf{S}}} \|\mathbf{R} - \widehat{\mathbf{L}} - \widehat{\mathbf{S}}\|_F^2 + \lambda_\Theta \|\widehat{\mathbf{L}}\|_* + \lambda_\Gamma \|\widehat{\mathbf{S}}\|_1, \tag{3.10}$$

with $\mathbf{R} = \mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM})$. The objective function is minimized over all $p_t^2 \times p_s^2$ matrices $\widehat{\mathbf{R}} = \widehat{\mathbf{L}} + \widehat{\mathbf{S}}$. The solutions $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{S}}$ correspond to estimates of $\mathcal{R}(\mathbf{\Theta})$ and $\mathcal{R}(\mathbf{\Gamma})$ respectively. As shown in (*Greenewald et al.*, 2013), the left and right singular vectors of $\widehat{\mathbf{L}}$ correspond to the (normalized) vectorized $\mathbf{A}_i$ and $\mathbf{B}_i$ respectively, as in (3.8).

This nuclear norm penalized low rank matrix approximation is a well-studied optimization problem (*Mazumder et al.*, 2010), where it is shown to be strictly convex. Several fast solution methods are available, including the iterative SVD-based proximal gradient method on which Algorithm 1 is based (*Moore et al.*, 2014). If the sparse correction $\mathbf{\Gamma}$ is omitted, equivalent to setting $\lambda_\Gamma = \infty$, the resulting optimization problem can be solved directly using the SVD (*Tsiligkaridis and Hero*, 2013). The minimizers $\widehat{\mathbf{L}}$, $\widehat{\mathbf{S}}$ of (4.15) are transformed to the covariance estimate by the simple operation

$$\widehat{\mathbf{\Sigma}} = \mathcal{R}^{-1}\left(\widehat{\mathbf{L}} + \widehat{\mathbf{S}}\right), \tag{3.11}$$

where $\mathcal{R}^{-1}(\cdot)$ is the inverse of the permutation operator $\mathcal{R}(\cdot)$. As the objective function in Equation (4.15) is strictly convex and is equivalent to the Robust PCA objective function of (*Moore et al.*, 2014), Algorithm 1 converges to a unique global minimizer.

Algorithm 1 is an appropriate modification of the iterative algorithm described in (*Moore et al.*, 2014). It consists of alternating between two simple steps: 1) soft thresholding of the singular values of the difference between the overall estimate and the estimate of the sparse part ($\mathbf{SVT}_\lambda(\cdot)$), and 2) soft thresholding of the entries of the difference between the overall estimate and the estimate of the low rank part

($\mathbf{soft}_\lambda(\cdot)$). The soft singular value thresholding operator is defined as

$$\mathbf{SVT}_\lambda(\mathbf{M}) = \mathbf{U}\left(\mathrm{diag}(\sigma_1, \ldots, \sigma_{\min(m_1,m_2)}) - \lambda\mathbf{I}\right)_+ \mathbf{V}^T, \tag{3.12}$$

where $\mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_{\min(m_1,m_2)})\mathbf{V}^T$ is the singular value decomposition of $\mathbf{M} \in \mathbb{R}^{m_1 \times m_2}$ and $(\cdot)_+ = \max(\cdot, 0)$. The entrywise soft thresholding operator is given by

$$[\mathbf{soft}_\lambda(\mathbf{M})]_{ij} = \mathrm{sign}(M_{ij})(|M_{ij}| - \lambda)_+. \tag{3.13}$$

Algorithm 1 Proximal Gradient Robust KronPCA

$\mathbf{R} = \mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM})$
Initialize $\mathbf{M}, \mathbf{S}, \mathbf{L}$, choose step sizes τ_k.
while $\mathcal{R}^{-1}(\mathbf{L} + \mathbf{S})$ not converged **do**
  $\mathbf{L}^k = \mathbf{SVT}_{\tau_k \lambda'_\Theta}(\mathbf{M}^{k-1} - \mathbf{S}^{k-1})$
  $\mathbf{S}^k = \mathbf{soft}_{\tau_k \lambda'_\Gamma}(\mathbf{M}^{k-1} - \mathbf{L}^{k-1})$
  $\mathbf{M}^k = \mathbf{L}^k + \mathbf{S}^k - \tau_k(\mathbf{L}^k + \mathbf{S}^k - \mathbf{R})$
end while
$\widehat{\mathbf{\Sigma}} = \mathcal{R}^{-1}(\mathbf{L} + \mathbf{S})$
return $\widehat{\mathbf{\Sigma}}$

Algorithm 2 Proximal Gradient Robust Toeplitz KronPCA

$\widetilde{\mathbf{R}} = \mathbf{P}\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM})$
Initialize $\mathbf{M}, \mathbf{S}, \mathbf{L}$, choose step sizes τ_k.
while $\mathcal{R}^{-1}\left(\mathbf{P}^T(\mathbf{L} + \mathbf{S})\right)$ not converged **do**
  $\mathbf{L}^k = \mathbf{SVT}_{\tau_k \lambda'_\Theta}(\mathbf{M}^{k-1} - \mathbf{S}^{k-1})$
  for $j \in \mathcal{I}$ **do**
    $\mathbf{S}^k_{j+p_t} = \mathbf{soft}_{\tau_k \lambda'_\Gamma c_j}(\mathbf{M}^{k-1}_{j+p_t} - \mathbf{L}^{k-1}_{j+p_t})$
  end for
  $\mathbf{M}^k = \mathbf{L}^k + \mathbf{S}^k - \tau_k(\mathbf{L}^k + \mathbf{S}^k - \widetilde{\mathbf{R}})$
end while
$\widehat{\mathbf{\Sigma}} = \mathcal{R}^{-1}\left(\mathbf{P}^T(\mathbf{L} + \mathbf{S})\right)$
return $\widehat{\mathbf{\Sigma}}$

3.2.2 Block Toeplitz Structured Covariance

Here we extend Algorithm 1 to incorporate a block Toeplitz constraint. Block Toeplitz constraints are relevant to stationary processes arising in signal and image processing. For simplicity we consider the case that the covariance is block Toeplitz with respect to time, however, extensions to the cases of Toeplitz spatial structure and having Toeplitz structure simultaneously in both time and space are straightforward. The objective function (4.15), is to be solved with a constraint that both $\widehat{\mathbf{\Theta}}$ and $\widehat{\mathbf{\Gamma}}$ are temporally block Toeplitz.

For low separation rank component $\mathbf{\Theta} = \sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i$, the block Toeplitz constraint is equivalent to a Toeplitz constraint on the temporal factors $\mathbf{A}_i$. The Toeplitz constraint on $\mathbf{A}_i$ is equivalent to (*Kamm and Nagy*, 2000; *Pitsianis*, 1997)

$$[\mathbf{a}_i]_k = v^{(i)}_{j+p_t},\ \forall k \in \mathcal{K}(j), j = -p_t + 1, \ldots, p_t - 1, \tag{3.14}$$

for some vector $\mathbf{v}^{(i)}$ where $\mathbf{a}_i = \mathrm{vec}(\mathbf{A}_i)$ and

$$\mathcal{K}(j) = \{k :\ (k-1)p_t + k + j \in [-p_t + 1,\ p_t - 1]\}. \tag{3.15}$$

It can be shown, after some algebra, that the optimization problem (4.15) constrained to block Toeplitz covariances is equivalent to an unconstrained penalized least squares problem involving $\mathbf{v}^{(i)}$ instead of $\mathbf{a}_i$. Specifically, following the techniques of (*Kamm and Nagy*, 2000; *Pitsianis*, 1997; *Greenewald and Hero*, 2014b) with the addition of the 1-norm penalty, the constrained optimization problem (4.15) can be shown to be equivalent to the following unconstrained optimization problem:

$$\min_{\widetilde{\mathbf{L}},\widetilde{\mathbf{S}}} \|\widetilde{\mathbf{R}} - \widetilde{\mathbf{L}} - \widetilde{\mathbf{S}}\|_F^2 + \lambda'_\Theta \|\widetilde{\mathbf{L}}\|_* + \lambda'_\Gamma \sum_{j \in \mathcal{I}} c_j \|\widetilde{\mathbf{S}}_{j+p_t}\|_1, \tag{3.16}$$

where $\widetilde{\mathbf{S}}_j$ denotes the jth row of $\widetilde{\mathbf{S}}$, $\widetilde{\mathbf{L}} = \mathbf{P}\widehat{\mathbf{L}}$, $\widetilde{\mathbf{S}} = \mathbf{P}\mathbf{S}$, and $\widetilde{\mathbf{R}} = \mathbf{P}\mathbf{R}$. The summation

indices $\mathcal{I}$, the 1-norm weighting constants c_j, and the $(2p_t - 1) \times p_t^2$ matrix $\mathbf{P}$ are defined as

$$\begin{aligned} \mathcal{I} =& \{-p_t + 1, \ldots, p_t - 1\} \\ c_j =& 1/\sqrt{p_t - |j|} \\ P_{p_t+j,i} =& \begin{cases} \frac{1}{\sqrt{p_t-|j|}} & i \in \mathcal{K}(j) \\ 0 & o.w. \end{cases} \end{aligned} \tag{3.17}$$

where the last line holds for all $j = -p_t + 1, \ldots, p_t - 1, i = 1, \ldots, p_s^2$. Note that this imposition of Toeplitz structure also results in a significant reduction in computational cost primarily due to a reduction in the size of the matrix in the singular value thresholding step (*Greenewald and Hero*, 2014b). The block Toeplitz estimate is given by

$$\widehat{\mathbf{\Sigma}} = \mathcal{R}^{-1}\left(\mathbf{P}^T\left(\widetilde{\mathbf{L}} + \widetilde{\mathbf{S}}\right)\right), \tag{3.18}$$

where $\widetilde{\mathbf{L}}$, $\widetilde{\mathbf{S}}$ are the minimizers of (3.16). Similarly to the non-Toeplitz case, the block Toeplitz estimate can be computed using Algorithm 2, which is the appropriate modification of Algorithm 1. As the objective function in Equation (4.15) is strictly convex and is equivalent to the Robust PCA objective function of (*Moore et al.*, 2014), Algorithm 2 converges to a unique global minimizer.

The non-Toeplitz and Toeplitz objective functions (4.15) and (3.16), respectively, are both invariant with respect to replacing $\mathbf{\Theta}$ with $\mathbf{\Theta}^T$ and $\mathbf{\Gamma}$ with $\mathbf{\Gamma}^T$ because $\mathbf{\Sigma}_{SCM}$ is symmetric. Furthermore, $\|(\mathbf{M} + \mathbf{M}^T)/2\| \leq \frac{1}{2}(\|\mathbf{M}\| + \|\mathbf{M}^T\|) = \|\mathbf{M}\|$ (for both the weighted 1-norm and nuclear norm) by the triangle inequality. Hence the symmetric $(\mathbf{\Sigma} + \mathbf{\Sigma}^T)/2 = (\mathbf{\Theta} + \mathbf{\Theta}^T)/2 + (\mathbf{\Gamma} + \mathbf{\Gamma}^T)/2$ will always result in at least as low an objective function value as would $\mathbf{\Sigma}$. By the uniqueness of the global optimum, the Robust KronPCA covariance estimates $\widehat{\mathbf{\Theta}}$ and $\widehat{\mathbf{\Gamma}}$ are therefore symmetric for both the

Toeplitz and non-Toeplitz cases.

3.3 High Dimensional Consistency

In this section, we impose the additional assumption that the training data $\mathbf{x}_i$ is Gaussian with true covariance given by

$$\mathbf{\Sigma}_0 = \mathbf{\Theta}_0 + \mathbf{\Gamma}_0, \tag{3.19}$$

where $\mathbf{\Theta}_0$ is the low separation rank covariance of interest and $\mathbf{\Gamma}_0$ is sparse.

A norm $\mathcal{R}_k(\cdot)$ is said to be *decomposable* with respect to subspaces $(\mathcal{M}_k, \bar{\mathcal{M}}_k)$ if (*Yang and Ravikumar*, 2013)

$$\mathcal{R}_k(u+v) = \mathcal{R}_k(u) + \mathcal{R}_k(v), \quad \forall u \in \mathcal{M}_k, v \in \bar{\mathcal{M}}_k^{\perp}. \tag{3.20}$$

We define subspace pairs $\mathcal{M}_{\mathbf{Q}}, \bar{\mathcal{M}}_{\mathbf{Q}}$ (*Yang and Ravikumar*, 2013) with respect to which the rearranged nuclear ($\mathbf{Q} = \mathbf{\Theta}$) and 1-norms ($\mathbf{Q} = \mathbf{\Gamma}$) are respectively decomposable (*Yang and Ravikumar*, 2013). These are associated with the set of either low separation rank ($\mathbf{Q} = \mathbf{\Theta}$) or sparse ($\mathbf{Q} = \mathbf{\Gamma}$) matrices. For the sparse case, let S be the set of indices on which $\text{vec}\{\mathbf{\Gamma}\}$ is nonzero. Then $\mathcal{M}_\Gamma = \bar{\mathcal{M}}_\Gamma$ is the subspace of vectors in $\mathbb{R}^{p_t^2 p_s^2}$ that have support contained in S, and $\bar{\mathcal{M}}_\Gamma^{\perp}$ is the subspace of vectors orthogonal to $\bar{\mathcal{M}}_\Gamma^{\perp}$, i.e. the subspace of vectors with support contained in S^c.

For the rearranged nuclear norm, note that by (*Loan and Pitsianis*, 1993) any $pq \times pq$ matrix $\mathbf{\Theta}$ can be decomposed as

$$\mathbf{\Theta} = \sum_{i=1}^{\min(p_t^2, p_s^2)} \sigma_i \mathbf{A}_\Theta^{(i)} \otimes \mathbf{B}_\Theta^{(i)} \tag{3.21}$$

where for all i, $\|\mathbf{A}_\Theta^{(i)}\|_F = \|\mathbf{B}_\Theta^{(i)}\|_F = 1$, $\sigma_i \geq 0$ and nonincreasing, the $p_t \times p_t$ $\{\mathbf{A}_\Theta^{(i)}\}_i$ are all linearly independent, and the $p_s \times p_s$ $\{\mathbf{B}_\Theta^{(i)}\}_i$ are all linearly independent. It is easy to show that this decomposition can be computed by extracting and rearranging the singular value decomposition of $\mathcal{R}(\mathbf{\Theta})$ (*Loan and Pitsianis*, 1993; *Tsiligkaridis and Hero*, 2013; *Werner et al.*, 2008) and thus the σ_i are uniquely determined. Let r be such that $\sigma_i = 0$ for all $i > r$. Define the matrices

$$\mathbf{U}_A = [\text{vec}(\mathbf{A}_\Theta^{(1)}), \ldots, \text{vec}(\mathbf{A}_\Theta^{(r)})],$$
$$\mathbf{U}_B = [\text{vec}(\mathbf{B}_\Theta^{(1)}), \ldots, \text{vec}(\mathbf{B}_\Theta^{(r)})].$$

Then we define a pair of subspaces with respect to which the nuclear norm is decomposable as

$$\begin{aligned} \mathcal{M}_\Theta &= \text{range}(\mathbf{U}_A \otimes \mathbf{U}_B), \\ \bar{\mathcal{M}}_\Theta^\perp &= \text{range}(\mathbf{U}_A^\perp \otimes \mathbf{U}_B^\perp). \end{aligned} \tag{3.22}$$

It can be shown that these subspaces are uniquely determined by $\mathbf{\Theta}$.

Consider the covariance estimator that results from solving the optimization problem in Equation (4.15). As is typical in Robust PCA, an incoherence assumption is required to ensure that $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ are distinguisable. Our incoherence assumption is as follows:

$$\begin{aligned} &\max\left\{\sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Theta}}}\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Gamma}}}\right), \sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Theta}}^\perp}\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Gamma}}}\right), \right. \\ &\left. \sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Theta}}}\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Gamma}}^\perp}\right), \sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Theta}}^\perp}\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{\Gamma}}^\perp}\right)\right\} \leq \frac{16}{\Lambda^2} \end{aligned} \tag{3.23}$$

where

$$\Lambda = 2 + \max\left\{\frac{3\beta\sqrt{2r}}{\lambda\sqrt{s}}, \frac{3\lambda\sqrt{s}}{\beta\sqrt{2r}}\right\}, \tag{3.24}$$

$\mathcal{P}_{\bar{\mathcal{M}}_{\mathbf{Q}}}$ is the matrix corresponding to the projection operator that projects onto the subspace $\bar{\mathcal{M}}_{\mathbf{Q}}$ and σ_{max} denotes the maximum singular value.

By way of interpretation, note that the maximum singular value of the product of projection matrices measures the "angle" between the subspaces. Hence, the incoherence condition is imposing that the subspaces in which $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ live be sufficiently "orthogonal" to each other i.e., "incoherent." This ensures identifiability in the sense that a portion of $\mathbf{\Gamma}$ (a portion of $\mathbf{\Theta}$) cannot be well approximated by adding a small number of additional terms to the Kronecker factors of $\mathbf{\Theta}$ ($\mathbf{\Gamma}$). Thus $\mathbf{\Theta}$ cannot be sparse and $\mathbf{\Gamma}$ cannot have low separation rank. In (*Yang and Ravikumar*, 2013) it was noted that this incoherency condition is significantly weaker than other typically imposed approximate orthogonality conditions.

Suppose that in the robust KronPCA model the n training samples are multivariate Gaussian distributed and IID, that $\mathbf{L}_0 = \mathcal{R}(\mathbf{\Theta}_0)$ is at most rank r, and that $\mathbf{S}_0 = \mathcal{R}(\mathbf{\Theta}_0)$ has s nonzero entries (3.8). Choose the regularization parameters to be

$$\lambda_\Theta = k\|\mathbf{\Sigma}_0\| \max(\alpha^2, \alpha), \quad \lambda_\Gamma = 32\rho(\mathbf{\Sigma}_0)\sqrt{\frac{\log p_t p_s}{n}}, \tag{3.25}$$

where $\rho(\mathbf{\Sigma}) = \max_j \Sigma_{jj}$ and k is smaller than an increasing function of t_0 given in the proof ((3.41)). We define α below.

Given these assumptions, we have the following bound on the estimation error (defining $M = \max(p_t, p_s, n)$).

Theorem III.1 (Robust KronPCA)**.** *Let* $\mathbf{L}_0 = \mathcal{R}(\mathbf{\Theta}_0)$. *Assume that the incoherence assumption* (3.23) *and the assumptions in the previous paragraph hold, and the regularization parameters* λ_Θ *and* λ_Γ *are given by Equation* (3.25) *with* $\alpha = \sqrt{t_0(p_t^2 + p_s^2 + \log M)/n}$ *for any chosen* $t_0 > 1$. *Then the Frobenius norm error of*

the solution to the optimization problem (4.15) *is bounded as:*

$$\|\widehat{\mathbf{L}}-\mathbf{L}_0\|_F \leq \tag{3.26}$$
$$6\max\left\{k\|\mathbf{\Sigma_0}\|\sqrt{r}\max(\alpha^2,\alpha), 32\rho(\mathbf{\Sigma_0})\sqrt{\frac{s\log p_t p_s}{n}}\right\}$$

with probability at least $1 - c\exp(-c_0 \log p_t p_s)$*, where* c, c_0 *are constants, and* c_0 *is dependent on* t_0 *but is bounded from above by an absolute constant.*

The proof of this theorem can be found in Appendix 3.6.1. Note that in practice the regularization parameters will be chosen via a method such as cross validation, so given that the parameters in (3.25) depend on $\mathbf{\Sigma}_0$, the specific values in (3.25) are less important than how they scale with p_t, p_s, n.

Next, we derive a similar bound for the Toeplitz Robust KronPCA estimator.

It is easy to show that a decomposition of $\mathbf{\Theta}$ of the form (3.21) exists where all the $\mathbf{A}_i$ are Toeplitz. Hence the definitions of the relevant subspaces ($\bar{\mathcal{M}}_{\mathbf{\Gamma}}$, $\bar{\mathcal{M}}_{\mathbf{\Theta}}$, $\mathcal{M}_{\mathbf{\Gamma}}$, $\mathcal{M}_{\mathbf{\Theta}}$) are of the same form as for the non Toeplitz case. In the Gaussian robust Toeplitz KronPCA model (3.16), further suppose $\widetilde{\mathbf{L}}_0 = \mathbf{P}\mathbf{L}_0$ is at most rank r and $\widetilde{\mathbf{S}}_0 = \mathbf{P}\mathbf{S}_0$ has at most s nonzero entries.

Theorem III.2 (Toeplitz Robust KronPCA)**.** *Assume that the assumptions of Theorem III.1 hold and that* $\mathbf{P}\mathbf{L}_0$ *is at most rank* r *and that* $\mathbf{P}\mathbf{S}_0$ *has at most* s *nonzero entries. Let the regularization parameters* λ_Θ *and* λ_Γ *be as in* (3.25) *with* $\alpha = \sqrt{t_0(2p_t + p_s^2 + \log M)/n}$ *for any* $t_0 > 1$*. Then the Frobenius norm error of the solution to the Toeplitz Robust KronPCA optimization problem* (4.15) *with coefficients given in* (3.17) *is bounded as:*

$$\|\widehat{\mathbf{L}}-\mathbf{L}_0\|_F \leq \tag{3.27}$$
$$6\max\left\{k\|\mathbf{\Sigma}_0\|\sqrt{r}\max(\alpha^2,\alpha), 32\rho(\mathbf{\Sigma}_0)\sqrt{\frac{s\log p_t p_s}{n}}\right\}$$

with probability at least $1 - c\exp(-c_0 \log p_t p_s)$*, where* c, c_0 *are constants, and* c_0 *is dependent on* t_0 *but is bounded from above by an absolute constant.*

The proof of this theorem is given in Appendix 3.6.1.

Comparing the right hand sides of (3.26) and (3.27) in Theorems III.1 and III.2 we see that, as expected, prior knowledge of Toeplitz structure reduces the Frobenius norm error of the estimator $\widehat{\mathbf{L}}$ from $O(p_t^2)$ to $O(p_t)$.

3.4 Results

3.4.1 Simulations

In this section we evaluate the performance of the proposed robust Kronecker PCA algorithms by conducting mean squared covariance estimation error simulations. For the first simulation, we consider a covariance that is a sum of 3 Kronecker products ($p_t = 10$, $p_s = 50$), with each term being a Kronecker product of two autoregressive (AR) covariances. AR processes with AR parameter a have covariances $\mathbf{\Psi}$ given by

$$\psi_{ij} = ca^{|i-j|}. \tag{3.28}$$

For the $p \times p$ temporal factors $\mathbf{A}_i$, we use AR parameters $[0.5, 0.8, 0.05]$ and for the $q \times q$ spatial factors $\mathbf{B}_i$ we use $[0.95, 0.35, 0.999]$. The Kronecker terms are scaled by the constants $[1, 0.5, 0.3]$. These values were chosen to create a complex covariance with 3 strong Kronecker terms with widely differing structure. The result is shown in Figure 3.1. We ran the experiments below for 100 cases with randomized AR parameters and in every instance Robust KronPCA dominated standard KronPCA and the sample covariance estimators to an extent qualitatively identical to the case shown below.

To create a covariance matrix following the non-Toeplitz "KronPCA plus sparse" model, we create a new "corrupted" covariance by taking the 3 term AR covariance

and deleting a random set of row/column pairs, adding a diagonal term, and sparsely adding high correlations (whose magnitude depends on the distance to the diagonal) at random locations. Figure 3.2 shows the resulting corrupted covariance. To create a "corrupted" block Toeplitz "KronPCA plus sparse" covariance, a diagonal term and block Toeplitz sparse correlations were added to the AR covariance in the same manner as in the non-Toeplitz case.

Figures 3.3 and 3.4 show results for estimating the Toeplitz corrupted covariance, and non-Toeplitz corrupted covariance respectively, using Algorithms 1 and 2. For both simulations, the average MSE of the covariance estimate is computed for a range of Gaussian training sample sizes. Average 3-steps ahead prediction MSE loss using the learned covariance to form the predictor coefficients ($\widehat{\mathbf{\Sigma}}_{yx}\widehat{\mathbf{\Sigma}}_{xx}^{\dagger}$) is shown in Figure 3.5. MSE loss is the prediction MSE using the estimated covariance minus $E[(y - E[y|x])^2]$, which is the prediction MSE achieved using an infinite number of training samples. The regularization parameter values shown are those used at $n = 10^5$ samples, the values for lower sample sizes are set proportionally using the n-dependent formulas given by (3.25) and Theorem III.1. The chosen values of the regularization parameters are those that achieved best average performance in the appropriate region. Note the significant gains achieved using the proposed regularization, and the effectiveness of using the regularization parameter formulas derived in the theorems. In addition, note that separation rank penalization alone does not maintain the same degree of performance improvement over the unregularized (SCM) estimate in the high sample regime, whereas the full Robust KronPCA method maintains a consistent advantage (as predicted by the Theorems III.1 and III.2).

3.4.2 Cell Cycle Modeling

As a real data application, we consider the yeast (S. cerevisiae) metabolic cell cycle dataset used in (*Deckard et al.*, 2013). The dataset consists of 9335 gene probes

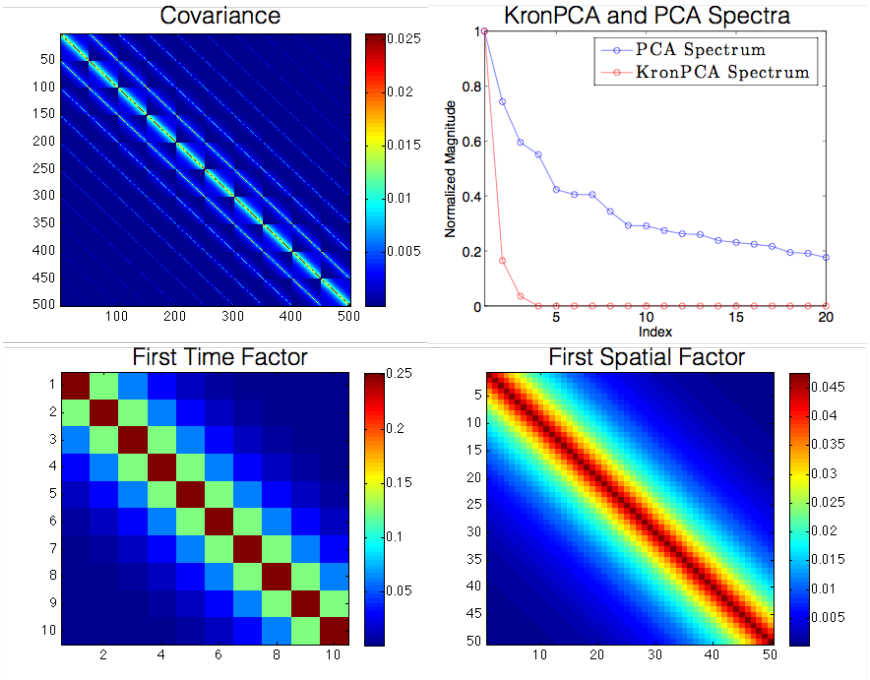


Figure 3.1: Uncorrupted covariance used for the MSE simulations. Subfigures (clockwise starting from upper left): $r = 3$ KronPCA covariance with AR factors $\mathbf{A}_i$, $\mathbf{B}_i$ (5.5); its KronPCA and PCA spectra; and the first spatial and temporal factors of the original covariance.

sampled approximately every 24 minutes for a total of 36 time points, and about 3 complete cell cycles (*Deckard et al.*, 2013).

In (*Deckard et al.*, 2013), it was found that the expression levels of many genes exhibit periodic behavior in the dataset due to the periodic cell cycle. Our goal is to establish that periodicity can also be detected in the temporal component of the Kronecker spatio-temporal correlation model for the dataset. Here the spatial index is the label of the gene probe. We use $p_t = 36$ so only one spatio-temporal training sample is available. Due to their high dimensionality, the spatial factor estimates have very low accuracy, but the first few temporal $\mathbf{A}_i$ factors (36×36) can be effectively estimated (bootstrapping using random sets of 20% genes achieved less than 3% RMS variation) due to the large number of spatial variables. We learn the spatiotemporal covariance (both space and time factors) using Robust KronPCA and then analyze the estimated time factors ($\mathbf{A}_i$) to discover periodicity. This allows us to consider

the overall periodicity of the gene set, taking into account relationships between the genes, as contrasted to the univariate analysis as in (*Deckard et al.*, 2013). The sparse correction to the covariance allows for the partial or complete removal of genes and correlations that are outliers in the sense that their temporal behavior differs from the temporal behavior of the majority of the genes.

Figure 3.6 shows the quantiles of the empirical distribution of the entries of the sample covariance versus those of the normal distribution. The extremely heavy tails motivate the use of a sparse correction term as opposed to the purely quadratic approach of standard KronPCA. Plots of the first row of each temporal factor estimate are shown in Figure 3.7. The first three factors are shown when the entire 9335 gene dataset is used to create the sample covariance. Note that 3 cycles of strong temporal periodicity are discovered, which matches our knowledge that approximately 3 complete cell cycles are contained in the sequence. Figure 3.8 displays the estimates of the first temporal factor when only a random 500 gene subset is used to compute the sample covariance. Note that the standard KronPCA estimate has much higher variability than the proposed robust KronPCA estimate, masking the presence of periodicity in the temporal factor. This is likely due to the heavy tailed nature of the distribution, and to the fact that robust KronPCA is better able to handle outliers via the sparse correction of the low Kronecker-rank component.

3.5 Conclusion

This chapter proposed a new robust method for performing Kronecker PCA of sparsely corrupted spatiotemporal covariances in the low sample regime. The method consists of a combination of KronPCA, a sparse correction term, and a temporally block Toeplitz constraint. To estimate the covariance under these models, a robust PCA based algorithm was proposed. The algorithm is based on nuclear norm penalization of a Frobenius norm objective to encourage low separation rank, and high

dimensional performance guarantees were derived for the proposed algorithm. Finally, simulations and experiments with yeast cell cycle data were performed that demonstrate advantages of our methods, both relative to the sample covariance and relative to the standard (non-robust) KronPCA.

3.6 Appendix

3.6.1 Robust KronPCA: Derivation of High Dimensional Consistency

In this section we first prove Theorem III.1 and then prove Theorem III.2, which are the bounds for the non-Toeplitz and Toeplitz cases respectively.

A general theorem for decomposable regularization of this type was proven in (*Yang and Ravikumar*, 2013). In (*Yang and Ravikumar*, 2013) the theorem was applied to Robust PCA directly on the sample covariance, hence when appropriate we follow a similar strategy for our proof for the Robust KronPCA case.

Consider the more general M-estimation problem

$$\min_{(\theta_k)_{k\in I}} \mathcal{L}\left(\sum_{k\in I}\theta_k\right) + \sum_{k\in I}\lambda_k \mathcal{R}_k(\theta_k), \tag{3.29}$$

where $\mathcal{L}$ is a convex differentiable loss function, the regularizers $\mathcal{R}_k$ are norms, with regularization parameters $\lambda_k \geq 2\mathcal{R}_k^*(\nabla_{\theta_k}\mathcal{L}(\theta^*))$. $\mathcal{R}_k^*$ is the dual norm of the norm $\mathcal{R}_k$, ∇_θ denotes the gradient with respect to θ, and θ^* is the true parameter value.

To emphasize that $\mathcal{L}$ depends on the observed training data $\mathbf{X}$ (in our case through the sample covariance), we also write $\mathcal{L}(\theta; \mathbf{X})$. Let $\mathcal{M}_k$ be the model subspace associated with the constraints enforced by $\mathcal{R}_k$ (*Yang and Ravikumar*, 2013). Assume the following conditions are satisfied:

1. The loss function $\mathcal{L}$ is convex and differentiable.

2. Each norm $\mathcal{R}_k$ ($k \in \mathcal{I}$) is decomposable with respect to the subspace pairs

$(\mathcal{M}_k, \bar{\mathcal{M}}_k^{\perp})$, where $\mathcal{M}_k \subseteq \bar{\mathcal{M}}_k$.

3. (Restricted Strong Convexity) For all $\mathbf{\Delta} \in \Omega_k$, where Ω_k is the parameter space for parameter component k,

$$\begin{aligned}\delta\mathcal{L}(\mathbf{\Delta}_k; \theta^*) &:= \mathcal{L}(\theta^* + \mathbf{\Delta}_k) - \mathcal{L}(\theta^*) - \langle \mathbf{\Delta}_\theta \mathcal{L}(\theta^*), \mathbf{\Delta}_k \rangle \\ &\geq \kappa_{\mathcal{L}} \|\mathbf{\Delta}_k\|^2 - g_k \mathcal{R}_k^2(\mathbf{\Delta}_k), \end{aligned} \tag{3.30}$$

where $\kappa_{\mathcal{L}}$ is a "curvature" parameter, and $g_k \mathcal{R}_k^2(\mathbf{\Delta}_k)$ is a "tolerance" parameter.

4. (Structural Incoherence) For all $\mathbf{\Delta}_k \in \Omega_k$,

$$\begin{aligned}&|\mathcal{L}(\theta^* + \sum_{k \in I} \mathbf{\Delta}_k) + (|I| - 1)\mathcal{L}(\theta^*) - \sum_{k \in I} \mathcal{L}(\theta^* + \mathbf{\Delta}_k)| \\ &\leq \frac{\kappa_{\mathcal{L}}}{2} \sum_{k \in I} \|\mathbf{\Delta}_k\|^2 + \sum_{k \in I} h_k \mathcal{R}_k^2(\mathbf{\Delta}_k). \end{aligned} \tag{3.31}$$

Define the *subspace compatibility constant* as $\Psi_k(\mathcal{M}, \|\cdot\|) := \sup_{u \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}_k(u)}{\|u\|}$. Given these assumptions, the following theorem holds (Corollary 1 in (*Yang and Ravikumar*, 2013)):

Theorem III.3. *Suppose that the subspace-pairs are chosen such that the true parameter values $\theta_k^* \in \mathcal{M}_k$. Then the parameter error bounds are given as:*

$$\|\widehat{\theta} - \theta^*\| \leq \left(\frac{3|I|}{2\bar{\kappa}} \right) \max_{k \in I} \lambda_k \Psi_k(\bar{\mathcal{M}}_k). \tag{3.32}$$

where

$$\begin{aligned}\bar{\kappa} &:= \frac{\kappa_{\mathcal{L}}}{2} - 32\bar{g}|I| \left(\max_{k \in I} \lambda_k \Psi_k(\bar{\mathcal{M}}_k) \right)^2, \\ \bar{g} &:= \max_k \frac{1}{\lambda_k} \sqrt{g_k + h_k}. \end{aligned}$$

We can now prove Theorem III.1.

Proof of Theorem III.1. To apply Theorem III.3 to the KronPCA estimation problem, we first check the conditions. In our objective function (4.15), we have a loss $\mathcal{L}(\mathbf{\Sigma}; \mathbf{X}) = \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{SCM}\|_F^2$, which of course satisfies condition 1. It was shown in (*Yang and Ravikumar*, 2013) that the nuclear norm and the 1-norm both satisfy Condition 2 with respect to $\mathcal{M}_\Theta, \bar{\mathcal{M}}_\Theta$ and $\mathcal{M}_\Gamma, \bar{\mathcal{M}}_\Gamma$ respectively. Hence we let the two $\mathcal{R}_k$ be the nuclear norm ($\mathcal{R}_\Theta(\cdot) = \|\cdot\|_*$) and the 1-norm ($\mathcal{R}_\Gamma(\cdot) = \|\cdot\|_1$) terms in (4.15). The restricted strong convexity condition (Condition 3) holds trivially with $\kappa_\mathcal{L} = 1$ and $g_k = 0$ (*Yang and Ravikumar*, 2013).

It was shown in (*Yang and Ravikumar*, 2013) that for the linear Frobenius norm mismatch term ($\mathcal{L}(\mathbf{\Sigma}) = \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{SCM}\|_F^2$) that we use in (4.15), the following simpler structural incoherence condition implies Condition 4 with $h_k = 0$:

$$\begin{aligned} \max \Big\{ &\sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_\Theta}\mathcal{P}_{\bar{\mathcal{M}}_\Gamma}\right), \sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_\Theta}\mathcal{P}_{\bar{\mathcal{M}}_\Gamma^\perp}\right), \\ &\sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_\Theta^\perp}\mathcal{P}_{\bar{\mathcal{M}}_\Gamma}\right), \sigma_{max}\left(\mathcal{P}_{\bar{\mathcal{M}}_\Theta^\perp}\mathcal{P}_{\bar{\mathcal{M}}_\Gamma^\perp}\right)\Big\} \leq \frac{1}{16\Lambda^2} \end{aligned} \tag{3.33}$$

where $\Lambda = \max_{k_1,k_2}\left\{2 + \frac{3\lambda_{k_1}\Psi_{k_1}(\bar{\mathcal{M}}_{k_1})}{\lambda_{k_2}\Psi_{k_2}(\bar{\mathcal{M}}_{k_2})}\right\}$.

The subspace compatibility constants are as follows (*Yang and Ravikumar*, 2013):

$$\begin{aligned} \Psi_{\mathbf{\Theta}}(\bar{\mathcal{M}}_{\mathbf{\Theta}}) &= \sup_{\mathbf{\Delta}\in\bar{\mathcal{M}}_{\mathbf{\Theta}}\setminus\{0\}} \frac{\|\mathbf{\Delta}\|_*}{\|\mathbf{\Delta}\|_F} \leq \sqrt{2r}, \\ \Psi_{\mathbf{\Gamma}}(\bar{\mathcal{M}}_{\mathbf{\Gamma}}) &= \sup_{\mathbf{\Delta}\in\bar{\mathcal{M}}_{\mathbf{\Gamma}}\setminus\{0\}} \frac{\|\mathbf{\Delta}\|_1}{\|\mathbf{\Delta}\|_F} \leq \sqrt{s}, \end{aligned} \tag{3.34}$$

where r is the rank of $\mathbf{\Theta}$ and s is the number of nonzero entries in $\mathbf{\Gamma}$. The first follows from the fact that for all $\mathbf{\Theta} \in \bar{\mathcal{M}}_\Theta$, $\text{rank}(\mathbf{\Theta}) \leq 2r$ since both the row and column spaces of $\mathbf{\Theta}$ must be of rank r (*Yang and Ravikumar*, 2013). Hence, we have that

$$\Lambda = 2 + \max\left\{\frac{3\beta\sqrt{2r}}{\lambda\sqrt{s}}, \frac{3\lambda\sqrt{s}}{\beta\sqrt{2r}}\right\}. \tag{3.35}$$

Finally, we need to show that both of the regularization parameters satisfy $\lambda_k \geq 2\mathcal{R}_k^*(\nabla_{\theta_k}\mathcal{L}(\theta^*;\mathbf{X}))$, i.e.

$$\begin{aligned}\lambda_\Theta &\geq 2\mathcal{R}_\Theta^*(\nabla_\Theta\mathcal{L}(\mathbf{\Theta}_0+\mathbf{\Gamma}_0;\mathbf{X})) \\ \lambda_\Gamma &\geq 2\mathcal{R}_\Gamma(\nabla_\Gamma\mathcal{L}(\mathbf{\Theta}_0+\mathbf{\Gamma}_0;\mathbf{X}))\end{aligned} \tag{3.36}$$

with high probability. Since the 1-norm is invariant under rearrangement, the argument from (*Yang and Ravikumar*, 2013) still holds and we have that

$$\lambda_\Gamma = 32\rho(\mathbf{\Sigma}_0)\sqrt{\frac{\log p_t p_s}{n}} \tag{3.37}$$

satisfies (3.36) with probability at least $1-2\exp(-c_2\log p_s p_t)$.

For the low rank portion, (3.36) will hold if (*Yang and Ravikumar*, 2013)

$$\lambda_\Theta \geq 4\|\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM}-\mathbf{\Sigma}_0)\|. \tag{3.38}$$

From (*Tsiligkaridis and Hero*, 2013) we have that for $t_0 \geq f(\epsilon) = 4C\log(1+\frac{2}{\epsilon})$ (C absolute constant given in (*Tsiligkaridis and Hero*, 2013)), C an absolute constant, and $\alpha \geq 1$

$$\|\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM}-\mathbf{\Sigma}_0)\| \leq \frac{\|\mathbf{\Sigma}_0\|t_0}{1-2\epsilon}\frac{p_t^2+p_s^2+\log M}{n} \tag{3.39}$$

with probability at least $1-2M^{-t_0/4C}$ and otherwise

$$\|\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM}-\mathbf{\Sigma}_0)\| \leq \frac{\|\mathbf{\Sigma}_0\|\sqrt{t_0}}{1-2\epsilon}\sqrt{\frac{p_t^2+p_s^2+\log M}{n}} \tag{3.40}$$

with probability at least $1-2M^{-t_0/4C}$. Thus our choice of λ_θ satisfies (3.38) with high probability. To satisfy the constraints on t, we need $t_0 > f^2(\epsilon)$. Clearly, ϵ can

be adjusted to satisfy the constraint and

$$k = 4/(1-2\epsilon). \tag{3.41}$$

Recalling the sparsity probability $1 - 2\exp(-c_2 \log p_t p_s)$, the union bound implies (3.36) is satisfied for both regularization parameters with probability at least $1 - 2\exp(-c_2 \log p_t p_s) - 2\exp(-(t_0/4C)\log M) \geq 1 - c\exp(-c_0 \log p_t p_s)$ and the proof of Theorem III.1 is complete. □

Next, we present the proof for Theorem III.2, emphasizing only the parts that differ from the non Toeplitz proof of Theorem III.1, since much of the proof for the previous theorem carries over to the Toeplitz case. Let

$$\begin{aligned} \boldsymbol{\Delta}_n &= \mathcal{R}(\mathbf{W}) \\ \mathbf{W} &= \widehat{\boldsymbol{\Sigma}}_{SCM} - \boldsymbol{\Sigma}_0. \end{aligned} \tag{3.42}$$

We require the following corollary based on an extension of a theorem in (*Tsiligkaridis and Hero*, 2013) to the Toeplitz case:

Corollary III.4. *Suppose $\boldsymbol{\Sigma}_0$ is a $p_t p_s \times p_t p_s$ covariance matrix, $\|\boldsymbol{\Sigma}_0\|_2$ is finite for all p_t, p_s, and let $M = \max(p_t, p_s, n)$. Let $\epsilon' < 0.5$ be fixed and assume that $t_0 \geq f(\epsilon)$ and $C = \max(C_1, C_2) > 0$. We have that*

$$\|\boldsymbol{\Delta}_n\|_2 \leq \frac{\|\boldsymbol{\Sigma}_0\|}{1-2\epsilon'} \max\left\{t_0\alpha^2, \sqrt{t_0}\alpha\right\} \tag{3.43}$$

with probability at least $1 - 2M^{-\frac{t_0}{4C}}$, where

$$\alpha = \frac{2p_t + p_s^2 + \log M}{n}. \tag{3.44}$$

The proof of this result is in Appendix 3.6.2.

Proof of Theorem III.2. Adjusting for the objective in (3.16), let the regularizers $\mathcal{R}_k$ be $\mathcal{R}_\Theta(\cdot) = \|\cdot\|_*$ and $\mathcal{R}_\Gamma(\mathbf{M}) = \sum_{j\in\mathcal{I}} c_j \|\mathbf{M}_{j+p_t}\|_1$. Condition 1 still holds as in the general non-Toeplitz case, and Condition 2 holds because $\mathcal{R}_\Gamma$ is a positively weighted sum of norms, forming a norm on the product space (which is clearly the entire space). $\mathcal{R}_\Gamma$ is decomposable because the 1-norm is decomposable and the overall model subspace is the product of the model subspaces for each row. The remaining two conditions trivially remain the same from the non Toeplitz case.

The subspace compatibility constant remains the same for the nuclear norm, and for the sparse case we have for all $\mathbf{\Delta}$

$$\mathcal{R}_\Gamma(\mathbf{\Delta}) \leq \|\mathbf{\Delta}\|_1 \tag{3.45}$$

hence, the supremum under the 1-norm is greater than the supremum under the row weighted norm. Thus, the subspace compatibility constant is still less than or equal to $\sqrt{s}$, where s is now the number of nonzero entries in $\mathbf{P}\mathcal{R}(\mathbf{\Gamma})$. A tighter bound is achievable if the degree of sparsity in each row is known.

We now show that the regularization parameters chosen satisfy (3.36) with high probability. For the sparse portion, we need to find the dual of $\mathcal{R}_\Gamma$, defined as

$$\mathcal{R}^*_\Gamma(\mathbf{Z}) = \sup\left\{\langle\langle \mathbf{Z}, \mathbf{X}\rangle\rangle | \mathcal{R}_\Gamma(\mathbf{X}) \leq 1\right\}, \tag{3.46}$$

where $\langle\langle \mathbf{Z}, \mathbf{X}\rangle\rangle = \text{trace}\{\mathbf{Z}^T\mathbf{X}\}$. Let the matrix $\mathbf{P}_1 = \text{diag}\{\{\sqrt{p_t - |j|}\}_{j=-p_t+1}^{p_t-1}\}$. Define the matrices $\mathbf{X}'$ such that $\mathbf{X}' = \mathbf{P}_1^{-1}\mathbf{X}$. Then $\mathcal{R}_\Gamma(\mathbf{X}) = \|\mathbf{X}'\|_1$ and

$$\begin{aligned}\mathcal{R}^*_\Gamma(\mathbf{Z}) &= \sup\left\{\langle\langle \mathbf{P}_1\mathbf{Z}, \mathbf{X}'\rangle\rangle \mid \|\mathbf{X}'\|_1 \leq 1\right\} \\ &= \|\mathbf{P}_1\mathbf{Z}\|_\infty\end{aligned} \tag{3.47}$$

since the dual of the 1-norm is the ∞-norm. From (*Agarwal et al.*, 2012), (3.36) now

takes the form

$$\lambda_\Gamma \geq 4\|\mathbf{P}_1\mathbf{P}\mathbf{W}\|_\infty = \|\widetilde{\mathbf{W}}\|_\infty \tag{3.48}$$

where

$$\widetilde{W}_{j+p_t,i} = \sum_{\ell \in \mathcal{K}(j)} W_{\ell,i}. \tag{3.49}$$

Hence

$$\begin{aligned} |\widetilde{W}_{j+p_t,i}| &\leq (p_t - |j|)\|\mathbf{W}\|_\infty \\ \|\widetilde{\mathbf{W}}\|_\infty &\leq p_t\|\mathbf{W}\|_\infty. \end{aligned} \tag{3.50}$$

From (*Agarwal et al.*, 2012) (via the union bound), we have

$$\Pr\left(\|\mathbf{W}\|_\infty > 8\rho(\mathbf{\Sigma})\sqrt{\frac{\log p_t p_s}{n}}\right) \leq 2\exp(-c_2\log(p_t p_s)), \tag{3.51}$$

giving

$$\Pr\left(\|\widetilde{\mathbf{W}}\|_\infty > 8\rho(\mathbf{\Sigma})p_t\sqrt{\frac{\log p_t p_s}{n}}\right) \leq 2\exp(-c_2\log(p_t p_s)), \tag{3.52}$$

which demonstrates that our choice for λ_Γ is satisfactory with high probability.

As before, for the low rank portion (3.36) will hold if (*Yang and Ravikumar*, 2013)

$$\lambda_\Theta \geq 4\|\mathbf{P}\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM} - \mathbf{\Sigma}_0)\|. \tag{3.53}$$

From Corollary III.4 we have that for $t \geq f(\epsilon)$, C an absolute constant, and $\alpha \geq 1$

$$\|\mathbf{P}\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM} - \mathbf{\Sigma}_0)\| \leq \frac{\|\mathbf{\Sigma}_0\|t_0}{1-2\epsilon}\frac{2p_t + p_s^2 + \log M}{n} \tag{3.54}$$

with probability at least $1 - 2M^{-t_0/4C}$ and otherwise

$$\|\mathbf{P}\mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM} - \mathbf{\Sigma}_0)\| \leq \frac{\|\mathbf{\Sigma}_0\|\sqrt{t_0}}{1-2\epsilon}\sqrt{\frac{2p_t + p_t^2 + \log M}{n}} \tag{3.55}$$

with probability at least $1-2M^{-t_0/4C}$. Hence, in the same way as in the non-Toeplitz case we have with high probability

$$\|\widehat{\widetilde{\mathbf{L}}} - \widetilde{\mathbf{L}}_0\|_F \leq 6\max\left\{k\|\mathbf{\Sigma}_0\|\sqrt{r}\max(\alpha^2, \alpha), 32\rho(\mathbf{\Sigma}_0)\sqrt{\frac{s\log p_t p_s}{n}}\right\} \tag{3.56}$$

and since $\mathbf{L} = \mathbf{P}^T\widetilde{\mathbf{L}}$ the theorem follows.

□

3.6.2 Gaussian Chaos Operator Norm Bound

We first note the following corollary from (*Tsiligkaridis and Hero*, 2013):

Corollary III.5. *Let* $\mathbf{x} \in \mathcal{S}_{p_t^2-1}$ *and* $\mathbf{y} \in \mathcal{S}_{p_s^2-1}$. *Let* $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{\Sigma}_0)$, $i = 1, \ldots, n$ *be* $p_t p_s$ *dimensional iid training samples. Let* $\mathbf{\Delta}_n = \mathcal{R}(\widehat{\mathbf{\Sigma}}_{SCM} - \mathbf{\Sigma}_0) = \mathcal{R}(\frac{1}{n}\sum_i \mathbf{z}_i\mathbf{z}_i^T - \mathbf{\Sigma}_0)$. *Then for all* $\tau > 0$,

$$\Pr(|\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \tau) \leq \exp\left(\frac{-n\tau^2/2}{C_1\|\mathbf{\Sigma}_0\|_2^2 + C_2\|\mathbf{\Sigma}_0\|_2\tau}\right) \tag{3.57}$$

where C_1, C_2 *are absolute constants.*

The proof (appropriately modified from that of a similar theorem in (*Tsiligkaridis and Hero*, 2013)) of Corollary III.4 then proceeds as follows:

Proof. Define $\mathcal{N}(\mathcal{S}_{d'-1}, \epsilon')$ as an ϵ' net on $\mathcal{S}_{d'-1}$. Choose $\mathbf{x}_1 \in \mathcal{S}_{2p_t-2}, \mathbf{y}_1 \in \mathcal{S}_{p_s^2-1}$ such that $|\mathbf{x}_1^T\mathbf{P}\mathbf{\Delta}_n\mathbf{y}_1| = \|\mathbf{P}\mathbf{\Delta}_n\|_2$. By definition, there exists $\mathbf{x}_2 \in \mathcal{N}(\mathcal{S}_{2p_t-2}, \epsilon'), \mathbf{y}_2 \in$

$\mathcal{N}(\mathcal{S}_{p_s^2-1}, \epsilon')$ such that $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon', \|\mathbf{y}_1 - \mathbf{y}_2\|_2 \leq \epsilon'$. Then

$$\begin{aligned} |\mathbf{x}_1^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_1| - |\mathbf{x}_2^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_2| &\leq |\mathbf{x}_1^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_1 - \mathbf{x}_2^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_2| \\ &\leq 2\epsilon' \|\mathbf{P}\boldsymbol{\Delta}_n\|_2. \end{aligned} \tag{3.58}$$

We then have

$$\begin{aligned} &\|\mathbf{P}\boldsymbol{\Delta}_n\|_2 (1 - 2\epsilon') \\ &\leq \max \left\{ |\mathbf{x}_2^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_2| : \mathbf{x}_2 \in \mathcal{N}(\mathcal{S}_{2p_t-2}, \epsilon'), \right. \\ &\qquad \left. \mathbf{y}_2 \in \mathcal{N}(\mathcal{S}_{p_s^2-1}, \epsilon'), \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon', \|\mathbf{y}_1 - \mathbf{y}_2\|_2 \leq \epsilon' \right\} \\ &\leq \max \left\{ |\mathbf{x}_2^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_2| : \mathbf{x}_2 \in \mathcal{N}(\mathcal{S}_{2p_t-2}, \epsilon'), \right. \\ &\qquad \left. \mathbf{y}_2 \in \mathcal{N}(\mathcal{S}_{p_s^2-1}, \epsilon') \right\} \end{aligned} \tag{3.59}$$

since $|\mathbf{x}_1^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}_1| = \|\mathbf{P}\boldsymbol{\Delta}_n\|_2$. Hence

$$\|\mathbf{P}\boldsymbol{\Delta}_n\|_2 \leq \frac{1}{1 - 2\epsilon'} \max_{\mathbf{x} \in \mathcal{N}(\mathcal{S}_{2p_t-2}, \epsilon'), \mathbf{y} \in \mathcal{N}(\mathcal{S}_{p_s^2-1}, \epsilon')} |\mathbf{x}^T \mathbf{P}\boldsymbol{\Delta}_n \mathbf{y}|. \tag{3.60}$$

From (*Tsiligkaridis and Hero*, 2013)

$$\text{card}(\mathcal{N}(\mathcal{S}_{d'-1}, \epsilon')) \leq \left(1 + \frac{2}{\epsilon'}\right)^{d'} \tag{3.61}$$

which allows us to use the union bound.

$$
\begin{aligned}
&\Pr(\|\mathbf{P}\boldsymbol{\Delta}_n\|_2 > \epsilon') \qquad (3.62)\\
&\leq \Pr\left(\max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{2p_t-2},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{p_s^2-1},\epsilon')} |\mathbf{x}^T\mathbf{P}\boldsymbol{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right)\\
&\leq \Pr\left(\bigcup_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{2p_t-2},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{p_s^2-1},\epsilon')} |\mathbf{x}^T\mathbf{P}\boldsymbol{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right)\\
&\leq \mathrm{card}(\mathcal{N}(\mathcal{S}_{2p_t-2},\epsilon'))\mathrm{card}(\mathcal{N}(\mathcal{S}_{p_s^2-1},\epsilon'))\\
&\quad\times \max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{2p_t-2},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{p_s^2-1},\epsilon')} \Pr(|\mathbf{x}^T\mathbf{P}\boldsymbol{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon'))\\
&\leq \left(1+\frac{2}{\epsilon'}\right)^{2p_t+p_s^2}\\
&\quad\times \max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}_{2p_t-2},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}_{p_s^2-1},\epsilon')} \Pr(|\mathbf{x}^T\mathbf{P}\boldsymbol{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')).
\end{aligned}
$$

Note that

$$
\begin{aligned}
\|\mathbf{x}^T\mathbf{P}\|_2^2 &= \sum_j \frac{x_{j+p_t}^2}{p_t-|j|}(p_t-|j|) \qquad (3.63)\\
&= \sum_j x_j^2 = \|\mathbf{x}\|_2^2 = 1,
\end{aligned}
$$

so $\mathbf{x}^T\mathbf{P} \in \mathcal{S}_{p_t^2-1}$. We can thus use Corollary III.5, giving

$$
\begin{aligned}
&\Pr(\|\mathbf{P}\boldsymbol{\Delta}_n\|_2 > \epsilon') \qquad (3.64)\\
&\leq 2\left(1+\frac{2}{\epsilon'}\right)^{2p_t+p_s^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{C_1\|\boldsymbol{\Sigma}_0\|_2^2 + C_2\|\boldsymbol{\Sigma}_0\|_2\epsilon(1-2\epsilon')}\right).
\end{aligned}
$$

Two regimes emerge from this expression. The first is where $\epsilon \leq \frac{C_1\|\boldsymbol{\Sigma}_0\|_2}{C_2(1-2\epsilon')}$, which allows

$$
\Pr(\|\mathbf{P}\boldsymbol{\Delta}_n\|_2 > \epsilon) \leq 2\left(1+\frac{2}{\epsilon'}\right)^{2p_t+p_s^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{2C_1\|\boldsymbol{\Sigma}_0\|_2^2}\right). \qquad (3.65)
$$

Choose

$$\epsilon = \frac{\sqrt{t_0}\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{2p_t + p_s^2 + \log M}{n}}. \tag{3.66}$$

This gives:

$$\begin{aligned}
&\Pr\left(\|\mathbf{P}\mathbf{\Delta}_n\|_2 > \frac{\sqrt{t_0}\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{2p_t + p_s^2 + \log M}{n}}\right) \\
&\leq 2\left(1+\frac{2}{\epsilon'}\right)^{2p_t+p_s^2}\exp\left(\frac{-t^2(2p_t+p_s^2+\log M)}{4C_1}\right) \\
&\leq 2\left(\left(1+\frac{2}{\epsilon'}\right)e^{-\frac{t_0}{4C_1}}\right)^{2p_t+p_s^2} M^{-t_0/(4C_1)} \\
&\leq 2M^{-t_0/(4C_1)}.
\end{aligned} \tag{3.67}$$

The second regime ($\epsilon > \frac{C_1\|\mathbf{\Sigma}_0\|_2}{C_2(1-2\epsilon')}$) allows us to set ϵ to

$$\epsilon = \frac{t_0\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\frac{2p_t + p_s^2 + \log M}{n} \tag{3.68}$$

which gives

$$\begin{aligned}
&\Pr\left(\|\mathbf{P}\mathbf{\Delta}_n\|_2 > \frac{t_0\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\frac{2p_t + p_s^2 + \log M}{n}\right) \\
&\leq 2\left(1+\frac{2}{\epsilon'}\right)^{2p_t+p_s^2}\exp\left(\frac{-t(2p_t+p_s^2+\log M)}{4C_2}\right) \\
&\leq 2M^{-t_0/(4C_2)}.
\end{aligned} \tag{3.69}$$

Combining both regimes (noting that $t_0 > 1$ and $\sqrt{t_0}C_1/C_2 > 1$) completes the proof.

□

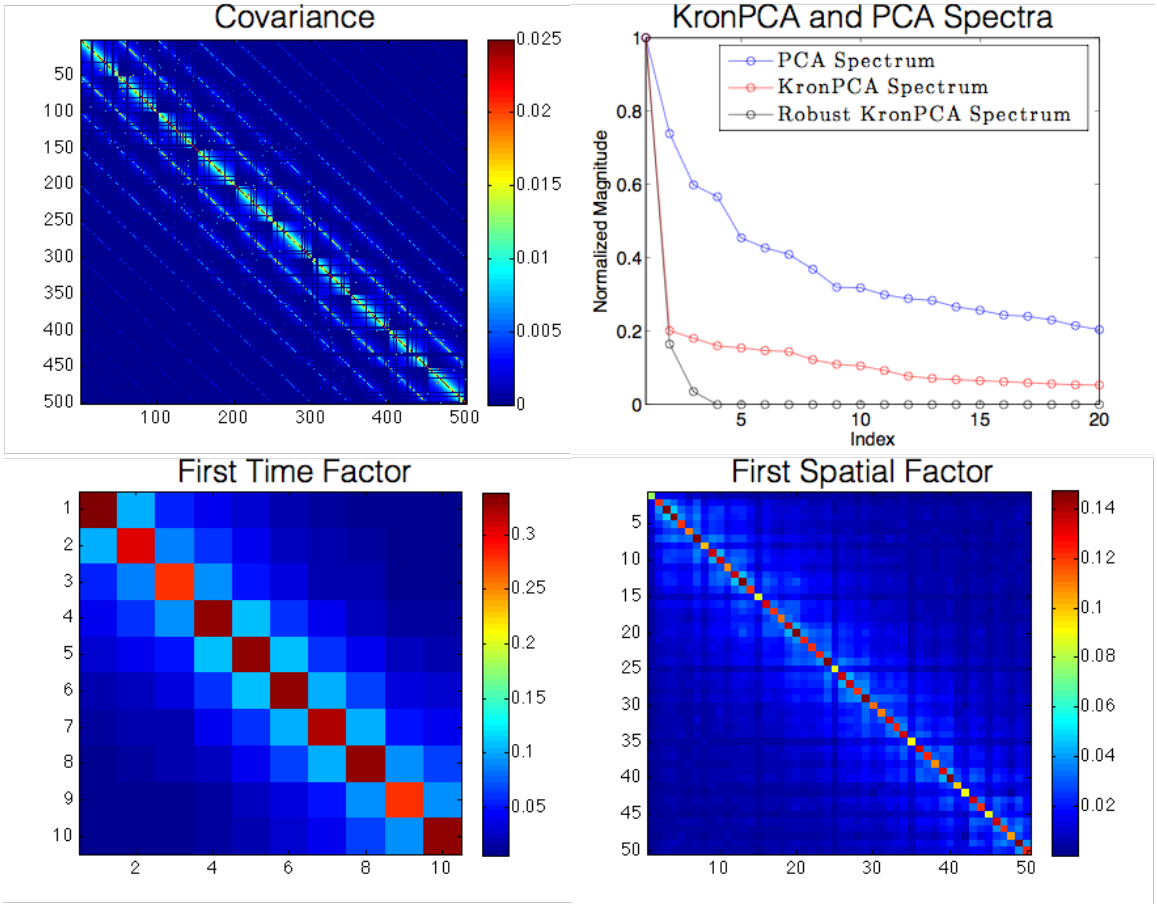


Figure 3.2: Corrupted version of the $r = 3$ KronPCA covariance (Figure 3.1), used to test the robustness of the proposed Robust KronPCA algorithm. Subfigures (Clockwise from upper left): covariance with sparse corruptions (3.5); its KronPCA, Robust KronPCA, and PCA spectra; and the (non-robust) estimates of the first spatial and temporal factors of the corrupted covariance. Note that the corruption spreads the KronPCA spectrum and the significant corruption of the first Kronecker factor in the non-robust KronPCA estimate.

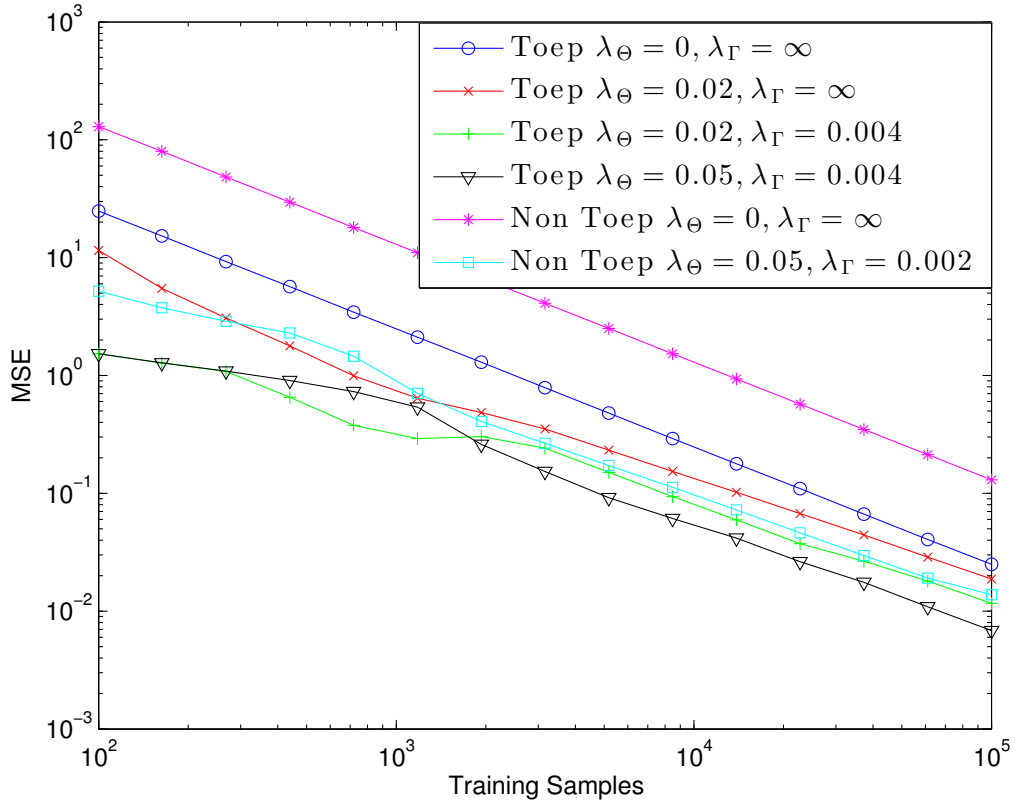


Figure 3.3: MSE plots for both Toeplitz Robust KronPCA (Toep) and non Toeplitz Robust KronPCA (Non Toep) estimation of the Toeplitz corrupted covariance, as a function of the number of training samples. Note the advantages of using each of Toeplitz structure, separation rank penalization, and sparsity regularization, as proposed in this chapter. The regularization parameter values shown are those used at $n = 10^5$ samples, the values for lower sample sizes are set proportionally using the n-dependent formulas given by (3.25) and Theorems III.1 and III.2.

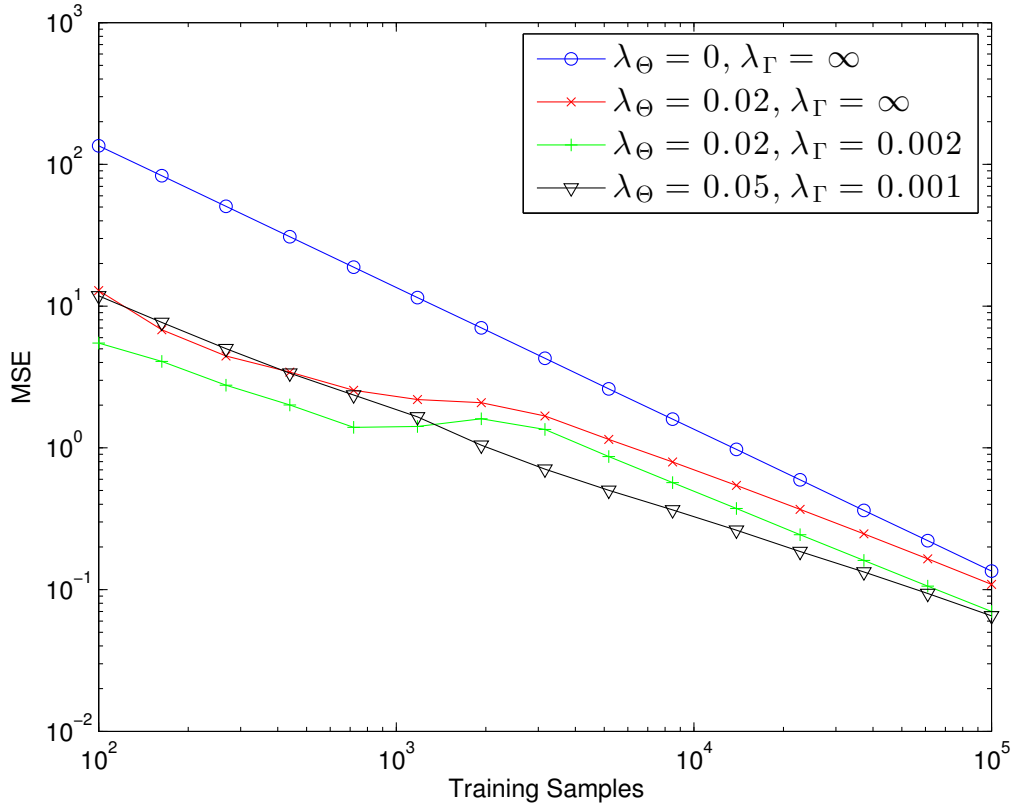


Figure 3.4: MSE plots for non-Toeplitz Robust KronPCA estimation of the corrupted covariance, as a function of the number of training samples. Note the advantages of using both separation rank and sparsity regularization. The regularization parameter values shown are those used at $n = 10^5$ samples, the values for lower sample sizes are set proportionally using the n-dependent formulas given by (3.25) and Theorem III.1.

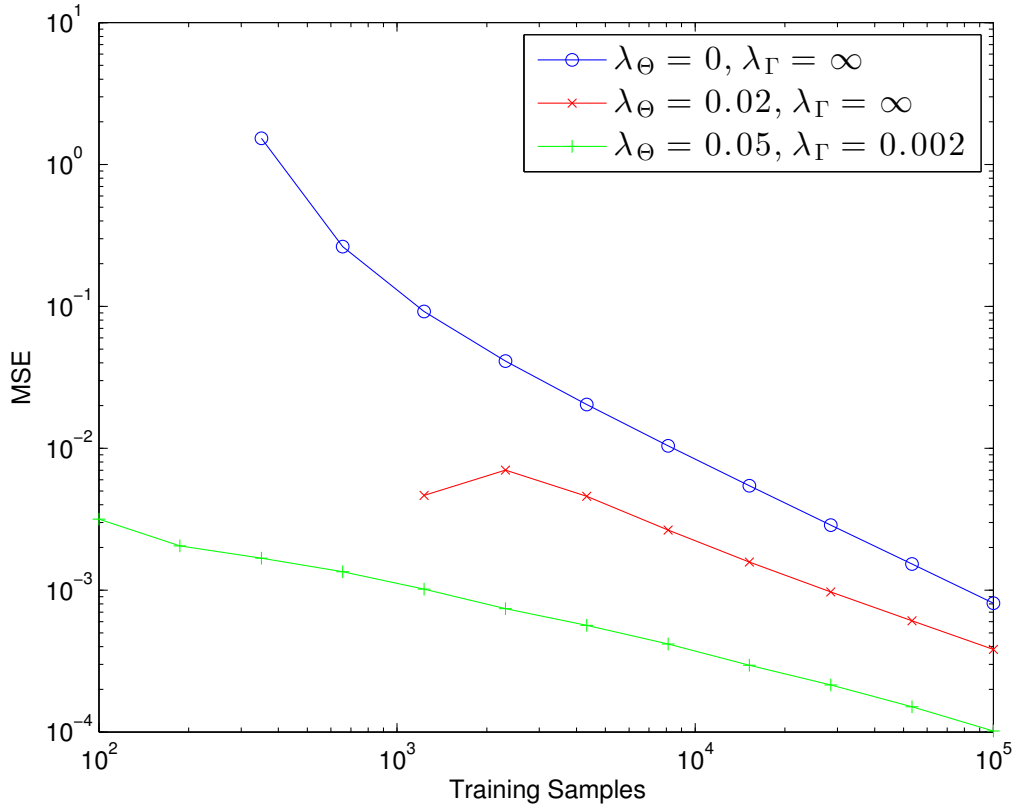


Figure 3.5: 3-ahead prediction MSE loss plots using the OLS predictor with corrupted covariance estimated by non-Toeplitz Robust KronPCA. Note the advantages of using both separation rank and sparsity regularization. The regularization parameter values shown are those used at $n = 10^5$ samples, the values for lower sample sizes are set proportionally using the n-dependent formulas given by (3.25) and Theorem III.1. The sample covariance ($\lambda_\Theta = 0$, $\lambda_\Gamma = \infty$) and standard KronPCA ($\lambda_\Theta = 0.02$, $\lambda_\Gamma = \infty$) curves are cut short due to aberrant behavior in the low sample regime.

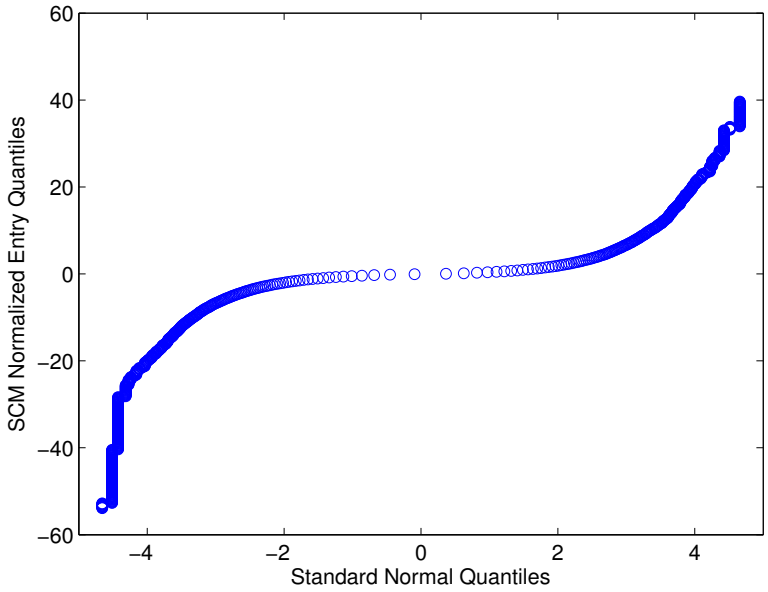


Figure 3.6: Plot of quantiles of the empirical distribution of the sample covariance entries versus those of the normal distribution (QQ). Note the very heavy tails, suggesting that an 2-norm based approach will break down relative to the Robust KronPCA approach allowing for sparse corrections.

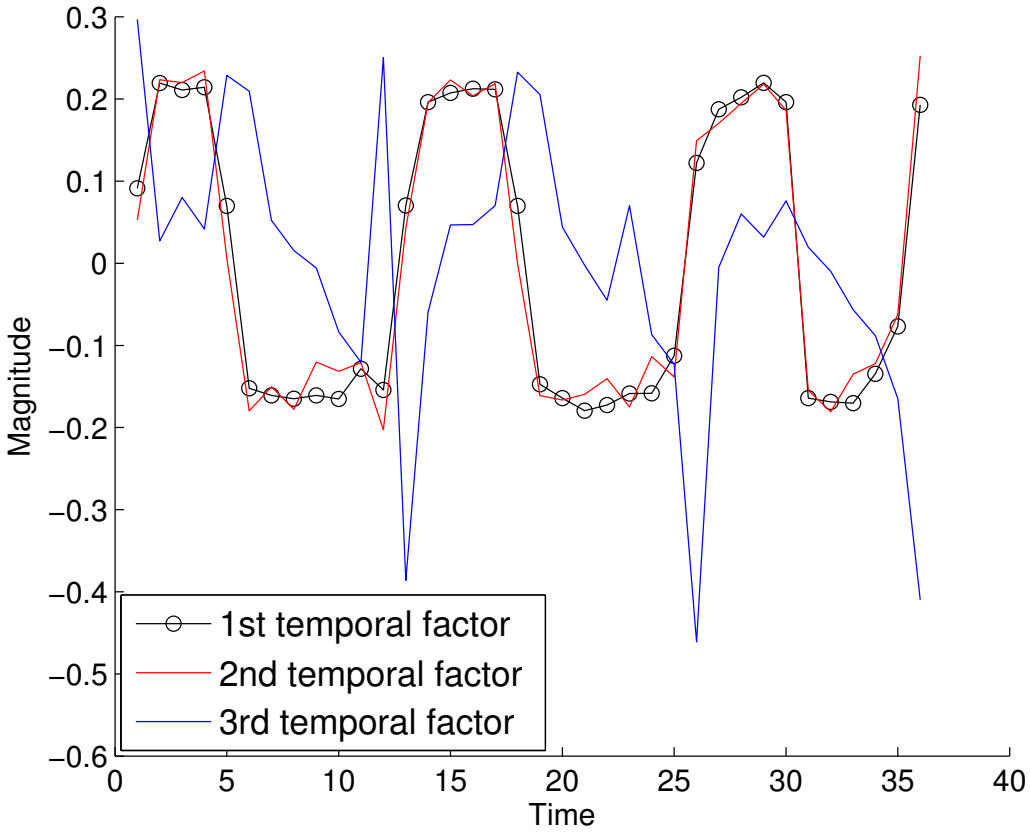


Figure 3.7: Plots of the temporal covariance factors estimated from the entire cell cycle dataset. Shown are the first rows of the first three temporal factors (excluding the first entry). Note the strong periodicity of the first two.

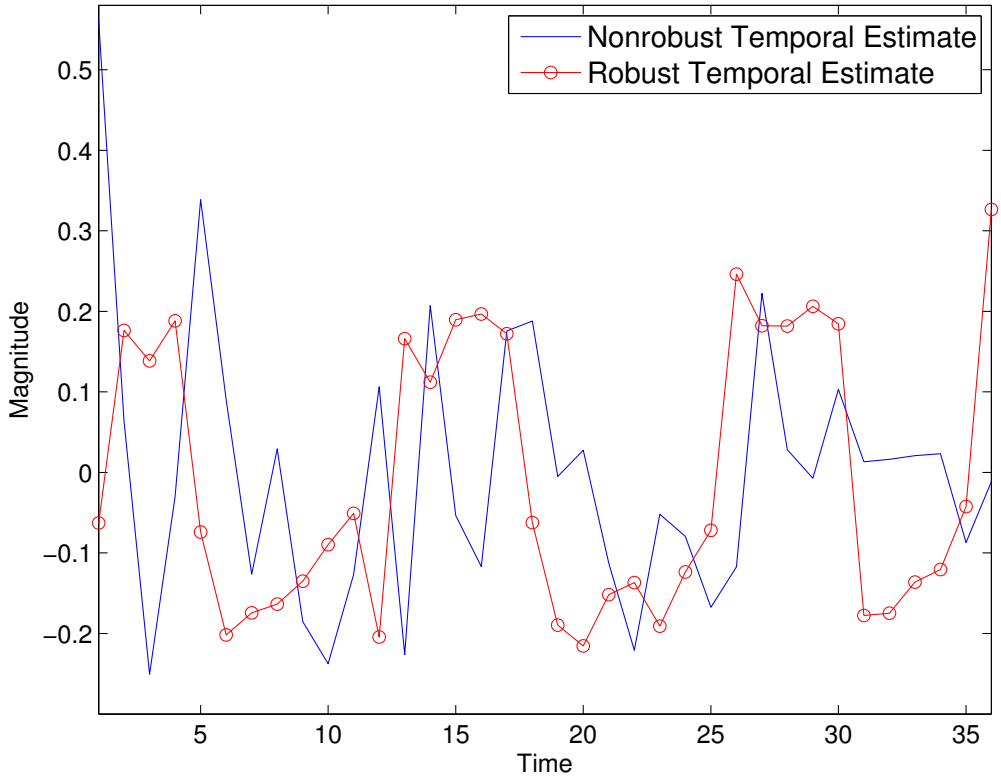


Figure 3.8: Plots of the first temporal covariance factors (excluding the first entry) estimated from the highly subsampled (spatially) cell cycle dataset using robust and standard KronPCA. Note the ability of Robust KronPCA to discover the correct periodicity.

CHAPTER IV

Robust SAR STAP via Kronecker Product Decomposition

In the previous chapter, we focused on using Kronecker structure as a method of approximating and estimating certain classes of spatiotemporal covariances. In this chapter, we consider a highly practical vision application for which Kronecker structure arises naturally from the physical model. By exploiting this structure, we are able to achieve significant performance gains both in synthetic and challenging real data.

4.1 Introduction

The detection (and tracking) of moving objects is an important task for scene understanding, as motion often indicates human related activity (*Newstadt et al.*, 2014). Radar sensors are uniquely suited for this task, as object motion can be discriminated via the Doppler effect. In this work, we propose a spatio-temporal decomposition method of detecting ground based moving objects in airborne Synthetic Aperture Radar (SAR) imagery, also known as SAR GMTI (SAR Ground Moving Target Indication).

Radar moving target detection modalities include MTI radars (*Newstadt et al.*,

2014; *Ender*, 1999), which use a low carrier frequency and high pulse repetition frequency to directly detect Doppler shifts. This approach has significant disadvantages, however, including low spatial resolution, small imaging field of view, and the inability to detect stationary or slowly moving targets. The latter deficiency means that objects that move, stop, and then move are often lost by a tracker.

SAR, on the other hand, typically has extremely high spatial resolution and can be used to image very large areas, e.g. multiple square miles in the Gotcha data collection (*Scarborough et al.*, 2009). As a result, stationary and slowly moving objects are easily detected and located (*Ender*, 1999; *Newstadt et al.*, 2014). Doppler, however, causes smearing and azimuth displacement of moving objects (*Jao*, 2001), making them difficult to detect when surrounded by stationary clutter. Increasing the number of pulses (integration time) simply increases the amount of smearing instead of improving detectability (*Jao*, 2001). Several methods have thus been developed for detecting and potentially refocusing (*Cristallini et al.*, 2013; *Cerutti-Maori et al.*, 2012) moving targets in clutter. Our goal is to remove the disadvantages of MTI and SAR by combining their strengths (the ability to detect Doppler shifts and high spatial resolution) using space time adaptive processing (STAP) with a novel Kronecker product spatio-temporal covariance model, as explained below.

SAR systems can either be single channel (standard single antenna system) or multichannel. Standard approaches for the single channel scenario include autofocusing (*Fienup*, 2001) and velocity filters. Autofocusing works only in low clutter, however, since it may focus the clutter instead of the moving target (*Fienup*, 2001; *Newstadt et al.*, 2014). Velocity filterbank approaches used in track-before-detect processing (*Jao*, 2001) involve searching over a large velocity/acceleration space, which often makes computational complexity excessively high. Attempts to reduce the computational complexity have been proposed, e.g. via compressive sensing based dictionary approaches (*Khwaja and Ma*, 2011) and Bayesian inference (*Newstadt et al.*, 2014),

but remain computationally intensive.

Multichannel SAR has the potential for greatly improved moving target detection performance (*Ender*, 1999; *Newstadt et al.*, 2014). Standard multiple channel configurations include spatially separated arrays of antennas, flying multiple passes (change detection), using multiple polarizations, or combinations thereof (*Newstadt et al.*, 2014).

4.1.1 Previous Multichannel Approaches

Several techniques exist for using multiple radar channels (antennas) to separate the moving targets from the stationary background. SAR GMTI systems have an antenna configuration such that each antenna transmits and receives from approximately the same location but at slightly different times (*Scarborough et al.*, 2009; *Ender*, 1999; *Newstadt et al.*, 2014; *Cerutti-Maori et al.*, 2012). Along track interferometry (ATI) and displaced phase center array (DPCA) are two classical approaches (*Newstadt et al.*, 2014) for detecting moving targets in SAR GMTI data, both of which are applicable only to the two channel scenario. Both ATI and DPCA first form two SAR images, each image formed using the signal from one of the antennas. To detect the moving targets, ATI thresholds the phase difference between the images and DPCA thresholds the magnitude of the difference. A Bayesian approach using a parametric cross channel covariance generalizing ATI/DPCA to p channels was developed in (*Newstadt et al.*, 2014), and a unstructured method fusing STAP and a test statistic in (*Cerutti-Maori et al.*, 2012). Space-time Adaptive Processing (STAP) learns a spatio-temporal covariance from clutter training data, and uses these correlations to filter out the stationary clutter while preserving the moving target returns (*Ender*, 1999; *Ginolhac et al.*, 2014; *Klemm*, 2002).

A second configuration, typically used in classical GMTI, uses phase coherent processing of the signals output by an antenna array for which each antenna receives

spatial reflections of the same transmission at the same time. This contrasts with the above configuration where each antenna receives signals from different transmissions at different times. In this second approach the array is designed such that returns from different angles create different phase differences across the antennas (*Klemm*, 2002; *Ginolhac et al.*, 2014; *Rangaswamy et al.*, 2004; *Kirsteins and Tufts*, 1994; *Haimovich*, 1996; *Conte and De Maio*, 2003). In this case, the covariance-based STAP approach, described above, can be applied to cancel the clutter (*Rangaswamy et al.*, 2004; *Ginolhac et al.*, 2014; *Haimovich*, 1996).

In this chapter, we focus on the first (SAR GMTI) configuration and propose a covariance-based STAP algorithm with a customized Kronecker product covariance structure. The SAR GMTI receiver consists of an array of p phase centers (antennas) processing q pulses in a coherent processing interval. Define the array $\mathbf{X}^{(m)} \in \mathbb{C}^{p\times q}$ such that $X_{ij}^{(m)}$ is the radar return from the jth pulse of the ith channel in the mth range bin. Let $\mathbf{x}_m = \text{vec}(\mathbf{X}^{(m)})$. The target-free radar data $\mathbf{x}_m$ is complex valued and is assumed to have zero mean. Define

$$\mathbf{\Sigma} = \text{Cov}[\mathbf{x}] = E[\mathbf{x}\mathbf{x}^H]. \tag{4.1}$$

The training samples, denoted as the set $\mathcal{S}$, used to estimate the SAR covariance $\mathbf{\Sigma}$ are collected from n representative range bins. The standard sample covariance matrix (SCM) is given by

$$\mathbf{S} = \frac{1}{n}\sum_{m\in\mathcal{S}} \mathbf{x}_m\mathbf{x}_m^H. \tag{4.2}$$

If n is small, $\mathbf{S}$ may be rank deficient or ill-conditioned (*Newstadt et al.*, 2014; *Ginolhac et al.*, 2014; *Greenewald et al.*, 2013; *Greenewald and Hero*, 2014b), and it can be shown that using the SCM directly for STAP requires a number n of training samples that is at least twice the dimension pq of $\mathbf{S}$ (*Reed et al.*, 1974). In this data rich

case, STAP performs well (*Newstadt et al.*, 2014; *Ender*, 1999; *Ginolhac et al.*, 2014). However, with p antennas and q time samples (pulses), the dimension pq of the covariance is often very large, making it difficult to obtain a sufficient number of target-free training samples. This so-called "small n large pq" problem leads to severe instability and overfitting errors, compromising STAP tracking performance.

By introducing structure and/or sparsity into the covariance matrix, the number of parameters and the number of samples required to estimate them can be reduced. As the spatiotemporal clutter covariance $\mathbf{\Sigma}$ is low rank (*Brennan and Staudaher*, 1992; *Ginolhac et al.*, 2014; *Rangaswamy et al.*, 2004; *Ender*, 1999), Low Rank STAP (LR-STAP) clutter cancelation estimates a low rank clutter subspace from $\mathbf{S}$ and uses it to estimate and remove the rank r clutter component in the data (*Bazi et al.*, 2005; *Ginolhac et al.*, 2014), reducing the number of parameters from $O(p^2q^2)$ to $O(rpq)$. Efficient algorithms, including some involving subspace tracking, have been proposed (*Belkacemi and Marcos*, 2006; *Shen et al.*, 2009). Other methods adding structural constraints such as persymmetry (*Ginolhac et al.*, 2014; *Conte and De Maio*, 2003), and robustification to outliers either via exploitation of the SIRV model (*Ginolhac et al.*, 2009) or adaptive weighting of the training data (*Gerlach and Picciolo*, 2011) have been proposed. Fast approaches based on techniques such as Krylov subspace methods (*Goldstein et al.*, 1998; *Honig and Goldstein*, 2002; *Pados et al.*, 2007; *Scharf et al.*, 2008) and adaptive filtering (*Fa and De Lamare*, 2011; *Fa et al.*, 2010) exist. All of these techniques remain sensitive to outlier or moving target corruption of the training data, and generally still require large training sample sizes (*Newstadt et al.*, 2014).

Instead, for SAR GMTI we propose to exploit the explicit space-time arrangement of the covariance by modeling the clutter covariance matrix $\mathbf{\Sigma}_c$ as the Kronecker

product of two smaller matrices

$$\mathbf{\Sigma}_c = \mathbf{A} \otimes \mathbf{B}, \tag{4.3}$$

where $\mathbf{A} \in \mathbb{C}^{p \times p}$ is rank 1 and $\mathbf{B} \in \mathbb{C}^{q \times q}$ is low rank. In this setting, the $\mathbf{B}$ matrix is the "temporal (pulse) covariance" and $\mathbf{A}$ is the "spatial (antenna) covariance," both determined up to a multiplicative constant. We note that this model is not appropriate for classical GMTI STAP, as in that configuration the covariance has a different spatio-temporal structure that is not separable.

Both ATI and DPCA in effect attempt to filter deterministic estimates of $\mathbf{A}$ to remove the clutter, and the Bayesian method (*Newstadt et al.*, 2014) uses a form of this model and incorporates the matrix $\mathbf{A}$ in a hierarchical clutter model. Standard SAR GMTI STAP approaches and the method of (*Cerutti-Maori et al.*, 2012) do not exploit this structure when estimating the spatiotemporal covariance. To our knowledge, this work is the first to exploit spatio-temporal structure to estimate a full low-rank spatio-temporal clutter covariance.

In this chapter, an iterative L2 based algorithm is proposed to directly estimate the low rank Kronecker factors from the observed sample covariance. Theoretical results indicate significantly fewer training samples are required, and it is shown that the proposed approach improves robustness to corrupted training data. Critically, robustness allows significant numbers of moving targets to remain in the training set. We then introduce the Kron STAP filter, which projects away both the spatial and temporal clutter subspaces. This projects away a higher dimensional subspace than does LR-STAP, thereby achieving improved noise and clutter cancelation.

To summarize, the main contributions of this chapter are: 1) the exploitation of the inherent Kronecker product spatio-temporal structure of the clutter covariance; 2) the introduction of the low rank Kronecker product based Kron STAP filter; 3) an

algorithm for estimating the spatial and temporal clutter subspaces that is highly robust to outliers due to the additional Kronecker product structure; and 4) theoretical results demonstrating improved signal-to-interference-plus-noise-ratio.

The remainder of the chapter is organized as follows. Section 4.2, presents the multichannel SIRV radar model. Our low rank Kronecker product covariance estimation algorithm and our proposed STAP filter are presented in Section 4.3. Section 4.4 gives theoretical performance guarantees and Section 4.5 gives simulation results and applies our algorithms to the Gotcha dataset.

In this work, we denote vectors as lower case bold letters, matrices as upper case bold letters, the complex conjugate as a^*, the matrix Hermitian as $\mathbf{A}^H$, and the Hadamard (elementwise) product as $\mathbf{A} \odot \mathbf{B}$.

4.2 SIRV Data Model

Let $\mathbf{X} \in \mathbb{C}^{p \times q}$ be an array of radar returns from an observed range bin across p channels and q pulses. We model $\mathbf{x} = \text{vec}(\mathbf{X})$ as a spherically invariant random vector (SIRV) with the following decomposition (*Yao*, 1973; *Rangaswamy et al.*, 2004; *Ginolhac et al.*, 2014, 2013):

$$\mathbf{x} = \mathbf{x}_{target} + \mathbf{x}_{clutter} + \mathbf{x}_{noise} = \mathbf{x}_{target} + \mathbf{n}, \tag{4.4}$$

where $\mathbf{x}_{noise}$ is Gaussian sensor noise with $\text{Cov}[\mathbf{x}_{noise}] = \sigma^2\mathbf{I} \in \mathbb{C}^{pq \times pq}$ and we define $\mathbf{n} = \mathbf{x}_{clutter} + \mathbf{x}_{noise}$. The signal of interest $\mathbf{x}_{target}$ is the sum of the spatio-temporal returns from all moving objects, modeled as non-random, in the range bin. The return from the stationary clutter is given by $\mathbf{x}_{clutter} = \tau\mathbf{c}$ where τ is a random positive scalar having arbitrary distribution, known as the *texture*, and $\mathbf{c} \in \mathbb{C}^{pq}$ is a multivariate complex Gaussian distributed random vector, known as the *speckle*. We define $\text{Cov}[\mathbf{c}] = \mathbf{\Sigma}_c$. The means of the clutter and noise components of $\mathbf{x}$ are zero.

The resulting clutter plus noise ($\mathbf{x}_{target} = 0$) covariance is given by

$$\mathbf{\Sigma} = E[\mathbf{n}\mathbf{n}^H] = E[\tau^2]\mathbf{\Sigma}_c + \sigma^2\mathbf{I}. \tag{4.5}$$

The ideal (no calibration errors) random speckle $\mathbf{c}$ is of the form (*Newstadt et al.*, 2014; *Ender*, 1999; *Cerutti-Maori et al.*, 2012)

$$\mathbf{c} = \mathbf{1}_p \otimes \widetilde{\mathbf{c}}, \tag{4.6}$$

where $\widetilde{\mathbf{c}} \in \mathbb{C}^q$. The representation (4.6) follows because the antenna configuration in SAR GMTI is such that the kth antenna receives signals emitted at different times at approximately (but not necessarily exactly) the same point in space (*Newstadt et al.*, 2014; *Scarborough et al.*, 2009). This is achieved by arranging the p antennas in a line parallel to the flight path, and delaying the kth antenna's transmission until it reaches the point x_i in space associated with the ith pulse. The representation (4.6) gives a clutter covariance of

$$\mathbf{\Sigma}_c = \mathbf{1}\mathbf{1}^T \otimes \mathbf{B}, \qquad \mathbf{B} = E[\widetilde{\mathbf{c}}\widetilde{\mathbf{c}}^H], \tag{4.7}$$

where $\mathbf{B}$ depends on the spatial characteristics of the clutter in the region of interest and the SAR collection geometry (*Ender*, 1999). While in SAR GMTI $\mathbf{B}$ is not exactly low rank, it is approximately low rank in the sense that significant energy concentration in a few principal components is observed over small regions (*Borcea et al.*, 2013).

Due to the long integration time and high cross range resolution associated with SAR, the returns from the general class of moving targets are more complicated, making simple Doppler filtering difficult. During short intervals for which targets have constant Doppler shift f (proportional to the target radial velocity) within a

range bin, the return has the form

$$\mathbf{x} = \alpha\mathbf{d} = \alpha\mathbf{a}(f) \otimes \mathbf{b}(f), \tag{4.8}$$

where α is the target's amplitude, $\mathbf{a}(f) = [\ 1 \quad e^{j2\pi\theta_1(f)} \quad \ldots \quad e^{j2\pi\theta_p(f)}\ ]^T$, the θ_i depend on Doppler shift f and the platform speed and antenna separation (*Newstadt et al.*, 2014), and $\mathbf{b} \in \mathbb{C}^q$ depends on the target, f, and its cross range path. The unit norm vector $\mathbf{d} = \mathbf{a}(f) \otimes \mathbf{b}(f)$ is known as the *steering vector*. For sufficiently large $\theta_i(f)$, $\mathbf{a}(f)^H\mathbf{1}$ will be small and the target will lie outside of the SAR clutter spatial subspace. The overall target return can be approximated as a series of constant-Doppler returns, hence the overall return should lie outside of the clutter spatial subspace. Furthermore, as observed in (*Fienup*, 2001), for long integration times the return of a moving target is significantly different from that of uniform stationary clutter, implying that moving targets generally lie outside the temporal clutter subspace (*Fienup*, 2001) as well.

In practice, the signals from each antenna have gain and phase calibration errors that vary slowly across angle and range (*Newstadt et al.*, 2014). It was shown in (*Newstadt et al.*, 2014) that in SAR GMTI these calibration errors can be accurately modeled as constant over small regions. Let the calibration error on antenna i be $h_i e^{j\phi_i}$ and $\mathbf{h} = [\ h_1 e^{j\phi_1}, \quad \ldots, \quad h_p e^{j\phi_p}\ ]$, giving an observed return $\mathbf{x}' = (\mathbf{h} \otimes \mathbf{I}) \odot \mathbf{x}$ and a clutter covariance of

$$\widetilde{\mathbf{\Sigma}}_c = (\mathbf{h}\mathbf{h}^H) \otimes \mathbf{B} = \mathbf{A} \otimes \mathbf{B} \tag{4.9}$$

implying that the $\mathbf{A}$ in (5.2) has rank one.

4.2.1 Space Time Adaptive Processing

Let the vector $\mathbf{d}$ be a spatio-temporal "steering vector" (*Ginolhac et al.*, 2014), that is, a matched filter for a specific target location/motion profile. For a measured array output vector $\mathbf{x}$ define the STAP filter output $y = \mathbf{w}^T\mathbf{x}$, where $\mathbf{w}$ is a vector of spatio-temporal filter coefficients. By (4.4) and (4.8) we have

$$y = \mathbf{w}^H\mathbf{x} = \alpha\mathbf{w}^H\mathbf{d} + \mathbf{w}^H\mathbf{n}. \tag{4.10}$$

The goal of STAP is to design the filter $\mathbf{w}$ such that the clutter is canceled ($\mathbf{w}^H\mathbf{n}$ is small) and the target signal is preserved ($\mathbf{w}^H\mathbf{d}$ is large). For a given target with spatio-temporal steering vector $\mathbf{d}$, an optimal clutter cancellation filter is defined as the filter $\mathbf{w}$ that maximizes the SINR (signal to interference plus noise ratio), defined as the ratio of the power of the filtered signal $\alpha\mathbf{w}^H\mathbf{d}$ to the power of the filtered clutter and noise (*Ginolhac et al.*, 2014)

$$\text{SINR}_{out} = \frac{|\alpha|^2|\mathbf{w}^H\mathbf{d}|^2}{E[\mathbf{w}^H\mathbf{n}\mathbf{n}^H\mathbf{w}]} = \frac{|\alpha|^2|\mathbf{w}^H\mathbf{d}|^2}{\mathbf{w}^H\mathbf{\Sigma}\mathbf{w}}, \tag{4.11}$$

where $\mathbf{\Sigma}$ is the clutter plus noise covariance in (4.5).

It can be shown (*Ender*, 1999; *Ginolhac et al.*, 2014) that, if the clutter covariance is known, under the SIRV model the optimal filter for targets at locations and velocities corresponding to the steering vector $\mathbf{d}$ is given by the filter

$$\mathbf{w} = \mathbf{F}_{opt}\mathbf{d}, \quad \mathbf{F}_{opt} = \mathbf{\Sigma}^{-1}. \tag{4.12}$$

Since the true covariance is unknown, we consider filters of the form

$$\mathbf{w} = \mathbf{F}\mathbf{d}, \tag{4.13}$$

and use the measurements to learn an estimate of the best $\mathbf{F}$.

For both classical GMTI radars and SAR GMTI, the clutter covariance has low rank r (*Brennan and Staudaher*, 1992; *Newstadt et al.*, 2014; *Ender*, 1999). Clutter subspace processing finds a *clutter subspace* $\{\mathbf{u}_i\}_{i=1}^r$ using the span of the top r principal components of the clutter sample covariance (*Ender*, 1999; *Ginolhac et al.*, 2014). This gives a clutter cancelation filter $\mathbf{F}$ that projects onto the space orthogonal to the estimated clutter subspace:

$$\mathbf{F} = \mathbf{I} - \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^H. \tag{4.14}$$

Since the sample covariance requires a relatively large number of training samples, obtaining sufficient numbers of target free training samples is difficult in practice (*Newstadt et al.*, 2014; *Ginolhac et al.*, 2014). In addition, if low amplitude moving targets are accidentally included in training, the sample covariance will be corrupted. In this case the resulting filter will partially cancel moving targets as well as clutter, which is especially problematic in online STAP implementations (*Newstadt et al.*, 2014; *Belkacemi and Marcos*, 2006). The proposed Kronecker STAP approach discussed below mitigates these problems as it directly takes advantage of the inherent space vs. time product structure of the clutter covariance $\mathbf{\Sigma}_c$.

4.3 Kronecker STAP

4.3.1 Kronecker Subspace Estimation

In this section we develop a subspace estimation algorithm that accounts for spatio-temporal covariance structure and has low computational complexity. In a high-dimensional setting, performing maximum likelihood on low-rank Kronecker product covariance estimation is computationally intensive under the Gaussian model or its SIRV extensions, and existing approximations combining Kronecker products

with Tyler's estimator (*Greenewald and Hero*, 2014b) do not give low rank estimates.

Similarly to the constrained least squares approaches of (*Werner et al.*, 2008; *Greenewald et al.*, 2013; *Tsiligkaridis and Hero*, 2013; *Greenewald and Hero*, 2014b), we fit a low rank Kronecker product model to the sample covariance matrix $\mathbf{S}$. Specifically, we minimize the Frobenius norm of the residual errors in the approximation of $\mathbf{S}$ by the low rank Kronecker model (4.9), subject to $\text{rank}(\mathbf{A}) \leq r_a, \text{rank}(\mathbf{B}) \leq r_b$, where the goal is to estimate $E[\tau^2]\mathbf{\Sigma}_c$. The optimal estimates of the Kronecker matrix factors $\mathbf{A}$ and $\mathbf{B}$ in (4.9) are given by

$$\widehat{\mathbf{A}}, \widehat{\mathbf{B}} = \arg \min_{\text{rank}(\mathbf{A}) \leq r_a, \text{rank}(\mathbf{B}) \leq r_b} \|\mathbf{S} - \mathbf{A} \otimes \mathbf{B}\|_F^2. \tag{4.15}$$

The minimization (4.15) will be simplified by using the patterned block structure of $\mathbf{A} \otimes \mathbf{B}$. In particular, for a $pq \times pq$ matrix $\mathbf{M}$, define $\{\mathbf{M}(i,j)\}_{i,j=1}^{p}$ to be its $q \times q$ block submatrices, i.e. $\mathbf{M}(i,j) = [\mathbf{M}]_{(i-1)q+1:iq,(j-1)q+1:jq}$. Also, let $\overline{\mathbf{M}} = \mathbf{K}_{p,q}^T \mathbf{M} \mathbf{K}_{p,q}$ where $\mathbf{K}_{p,q}$ is the $pq \times pq$ permutation operator such that $\mathbf{K}_{p,q}\text{vec}(\mathbf{N}) = \text{vec}(\mathbf{N}^T)$ for any $p \times q$ matrix $\mathbf{N}$.

The invertible Pitsianis-VanLoan rearrangement operator $\mathcal{R}(\cdot)$ maps $pq \times pq$ matrices to $p^2 \times q^2$ matrices and, as defined in (*Tsiligkaridis and Hero*, 2013; *Werner et al.*, 2008) sets the $(i-1)p+j$th row of $\mathcal{R}(\mathbf{M})$ equal to $\text{vec}(\mathbf{M}(i,j))^T$, i.e.

$$\mathcal{R}(\mathbf{M}) = [\ \mathbf{m}_1 \quad \dots \quad \mathbf{m}_{p^2}\ ]^T, \tag{4.16}$$

$$\mathbf{m}_{(i-1)p+j} = \text{vec}(\mathbf{M}(i,j)), \quad i,j = 1, \dots, p.$$

The unconstrained (i.e. $r_a = p, r_b = q$) objective in (4.15) is shown in (*Werner et al.*, 2008; *Tsiligkaridis and Hero*, 2013; *Greenewald et al.*, 2013) to be equivalent to a rearranged rank-one approximation problem, with a global minimizer given by

$$\widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}} = \mathcal{R}^{-1}(\sigma_1 \mathbf{u}_1 \mathbf{v}_1^H), \tag{4.17}$$

where $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^H$ is the first singular component of $\mathcal{R}(\mathbf{S})$. The operator $\mathcal{R}^{-1}$ is the inverse of $\mathcal{R}$, given by

$$\begin{aligned} \mathcal{R}^{-1}(\mathbf{M}) &= \mathbf{N} \in \mathbb{C}^{pq \times pq}, \\ \mathbf{N}(i,j) &= \mathrm{vec}_{q,q}^{-1}((\mathbf{M}_{(i-1)p+j,1:q^2})^T), \quad i,j = 1,\ldots,p, \end{aligned} \tag{4.18}$$

where $\mathrm{vec}_{q,q}^{-1}(\cdot)$ is the inverse of the vectorization operator on $q \times q$ matrices, i.e. if $\mathbf{m} = \mathrm{vec}(\mathbf{M}) \in \mathbb{C}^{q \times q}$, $\mathbf{M} = \mathrm{vec}_{q,q}^{-1}(\mathbf{m})$.

When the low rank constraints are introduced, there is no closed-form solution of (4.15). An iterative alternating minimization algorithm is derived in Appendix 4.7.1 and is summarized by Algorithm 3. In Algorithm 3, $\mathrm{EIG}_r(\mathbf{M})$ denotes the matrix obtained by truncating the Hermitian matrix $\mathbf{M}$ to its first r principal components, i.e.

$$\mathrm{EIG}_r(\mathbf{M}) := \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{u}_i^H, \tag{4.19}$$

where $\sum_i \sigma_i \mathbf{u}_i \mathbf{u}_i^H$ is the eigendecomposition of $\mathbf{M}$, and the (real and positive) eigenvalues σ_i are indexed in order of decreasing magnitude.

The objective (4.15) is not convex, but since it is an alternating minimization algorithm, it can be shown (Appendix 4.7.1) that Algorithm 3 will monotonically decrease the objective at each step, and that convergence of the estimates $\mathbf{A}_k, \mathbf{B}_k$ to a stationary point of the objective is guaranteed. We initialize LR-Kron with either $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}$ from the unconstrained estimate (4.17). Monotonic convergence then guarantees that LR-Kron improves on this simple closed form estimator.

We call Algorithm 3 low rank Kronecker product covariance estimation, or LR-Kron. In Appendix 4.7.1 it is shown that when the initialization is positive semidefinite Hermitian the LR-Kron estimator $\widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}}$ is positive semidefinite Hermitian and is thus a valid covariance matrix of rank $r_a r_b$.

Algorithm 3 LR-Kron Covariance Estimation

1: $\mathbf{S} = \mathbf{\Sigma}_{SCM}$, form $\mathbf{S}(i,j)$, $\overline{\mathbf{S}}(i,j)$.
2: Initialize $\mathbf{A}_0$ (or $\mathbf{B}_0$) using (4.17), with $\mathbf{A}_0$ s.t. $\|\mathbf{A}_0\|_F = 1$ (correspondingly $\mathbf{B}_0$).
3: **while** Objective $\|\mathbf{S} - \mathbf{A}_k \otimes \mathbf{B}_k\|_F^2$ not converged **do**
4: $\quad \mathbf{R}_B = \frac{\sum_{i,j}^{p} a_{k,ij}^* \overline{\mathbf{S}}(i,j)}{\|\mathbf{A}_k\|_F^2}$
5: $\quad \mathbf{B}_{k+1} = \mathrm{EIG}_{r_b}(\mathbf{R}_B)$
6: $\quad \mathbf{R}_A = \frac{\sum_{i,j}^{q} b_{k+1,ij}^* \mathbf{S}(i,j)}{\|\mathbf{B}_{k+1}\|_F^2}$
7: $\quad \mathbf{A}_{k+1} = \mathrm{EIG}_{r_a}(\mathbf{R}_A)$
8: **end while**
9: **return** $\widehat{\mathbf{A}} = \mathbf{A}_k, \widehat{\mathbf{B}} = \mathbf{B}_k$.

4.3.2 Robustness Benefits

Besides reducing the number of parameters, Kronecker STAP enjoys several other benefits arising from associated properties of the estimation objective (4.15).

The clutter covariance model (4.9) is low rank, motivating the PCA singular value thresholding approach of classical STAP. This approach, however, is problematic in the Kronecker case because of the way low rank Kronecker factors combine. Specifically, the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ has the SVD (*Loan and Pitsianis*, 1993)

$$\mathbf{A} \otimes \mathbf{B} = (\mathbf{U}_B \otimes \mathbf{U}_B)(\mathbf{S}_A \otimes \mathbf{S}_B)(\mathbf{U}_A^H \otimes \mathbf{U}_B^H) \tag{4.20}$$

where $\mathbf{A} = \mathbf{U}_A \mathbf{S}_A \mathbf{U}_A^H$ and $\mathbf{B} = \mathbf{U}_B \mathbf{S}_B \mathbf{U}_B^H$ are the SVDs of $\mathbf{A}$ and $\mathbf{B}$ respectively. The singular values are $s_A^{(i)} s_B^{(j)}$, $\forall i, j$. As a result, a simple thresholding of singular values is not equivalent to separate thresholding of the singular values of $\mathbf{A}$ and $\mathbf{B}$ and hence won't necessarily adhere to the space vs. time structure.

For example, suppose that the set of training data is corrupted by inclusion of a sparse set of w moving targets. By the model (4.8), the ith moving target gives a return (in the appropriate range bin) of the form

$$\mathbf{z}_i = \alpha_i \mathbf{a}_i \otimes \mathbf{b}_i, \tag{4.21}$$

where $\mathbf{a}_i, \mathbf{b}_i$ are unit norm vectors.

This results in a sample data covariance created from a set of observations $\mathbf{n}_m$ with $\text{Cov}[\mathbf{n}_m] = \boldsymbol{\Sigma}$, corrupted by the addition of a set of w rank one terms

$$\mathbf{S} = \left(\frac{1}{n} \sum_{m=1}^{n} \mathbf{n}_m \mathbf{n}_m^H \right) + \frac{1}{n} \sum_{i=1}^{w} \mathbf{z}_i \mathbf{z}_i^H. \tag{4.22}$$

Let $\widetilde{\mathbf{S}} = \frac{1}{n} \sum_{m=1}^{n} \mathbf{n}_m \mathbf{n}_m^H$ and $\widetilde{\mathbf{T}} = \frac{1}{n} \sum_{i=1}^{w} \mathbf{z}_i \mathbf{z}_i^H$. Let $\lambda_{S,k}$ be the eigenvalues of $\boldsymbol{\Sigma}_c$, $\lambda_{S,min} = \min_k \lambda_{S,k}$, and let $\lambda_{T,max}$ be the maximum eigenvalue of $\widetilde{\mathbf{T}}$. Assume that moving targets are indeed in a subspace orthogonal to the clutter subspace. If $\lambda_{T,max} > O(\lambda_{S,min})$, performing rank r PCA on $\mathbf{S}$ will result in principal components of the moving target term being included in the "clutter" covariance estimate.

If the targets are approximately orthogonal to each other (i.e. not coordinated), then $\lambda_{T,max} = O(\frac{1}{n}|\alpha_i|^2)$. Since the smallest eigenvalue of $\boldsymbol{\Sigma}_c$ is often small, this is the primary reason that classical LR-STAP is easily corrupted by moving targets in the training data (*Newstadt et al.*, 2014; *Ginolhac et al.*, 2014).

On the other hand, Kron-STAP is significantly more robust to such corruption. Specifically, consider the *rearranged* corrupted sample covariance:

$$\mathcal{R}(\mathbf{S}) = \frac{1}{n} \sum_{m=1}^{w} \text{vec}(\mathbf{a}_i \mathbf{a}_i^H) \text{vec}(\mathbf{b}_i \mathbf{b}_i^H)^H + \mathcal{R}(\widetilde{\mathbf{S}}). \tag{4.23}$$

This also takes the form of a desired sample covariance plus a set of rank one terms. For simplicity, we ignore the rank constraints in the LR-Kron estimator, in which case we have (4.17)

$$\widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}} = \mathcal{R}^{-1}(\widehat{\sigma}_1 \mathbf{u}_1 \mathbf{v}_1^H), \tag{4.24}$$

where $\widehat{\sigma}_1 \mathbf{u}_1 \mathbf{v}_1^H$ is the first singular component of $\mathcal{R}(\mathbf{S})$. Let σ_1 be the largest singular value of $\mathcal{R}(\widetilde{\mathbf{S}})$. The largest singular value $\widehat{\sigma}_1$ will correspond to the moving target term only if the largest singular value of $\frac{1}{n} \sum_{m=1}^{w} \text{vec}(\mathbf{a}_i \mathbf{a}_i^H) \text{vec}(\mathbf{b}_i \mathbf{b}_i^H)^H$ is greater than

$O(\sigma_1)$. If the moving targets are uncoordinated, this holds if for some i, $\frac{1}{n}|\alpha_i|^2 > O(\sigma_1)$. Since σ_1 models the entire clutter covariance, it is on the order of the total clutter energy, i.e. $\sigma_1^2 = O(\sum_{k=1}^{r} \lambda_{S,k}^2) \gg \lambda_{S,min}^2$. In this sense Kron-STAP is much more robust to moving targets in training than is LR-STAP.

4.3.3 Kronecker STAP Filters

Once the low rank Kronecker clutter covariance has been estimated using Algorithm 3, it remains to identify a filter $\mathbf{F}$, analogous to (4.14), that uses the estimated Kronecker covariance model. If we restrict ourselves to subspace projection filters and make the common assumption that the target component in (4.4) is orthogonal to the true clutter subspace, then the optimal approach in terms of SINR is to project away the clutter subspace, along with any other subspaces in which targets are not present. If only target orthogonality to the joint spatio-temporal clutter subspace is assumed, then the classical low-rank STAP filter is the projection matrix:

$$\mathbf{F}_{classical} = \mathbf{I} - \mathbf{U}_A\mathbf{U}_A^H \otimes \mathbf{U}_B\mathbf{U}_B^H, \tag{4.25}$$

where $\mathbf{U}_A, \mathbf{U}_B$ are orthogonal bases for the rank r_a and r_b subspaces of the low rank estimates $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, respectively, obtained by applying Algorithm 3. This is the Kronecker product equivalent of the standard STAP projector (4.14), though it should be noted that (4.25) will require less training data for equivalent performance due to the assumed structure.

The classical low-rank filter $\mathbf{F} = \mathbf{I} - \mathbf{U}\mathbf{U}^H$ is, as noted in section 4.2.1, merely an approximation to the SINR optimal filter $\mathbf{F} = \mathbf{\Sigma}^{-1}$. We note, however, that this may not be the only possible approximation. In particular, the inverse of a Kronecker product is the Kronecker product of the inverses, i.e. $\mathbf{A} \otimes \mathbf{B} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$. Hence, we consider using the low rank filter approximation on $\widehat{\mathbf{A}}^{-1}$ and $\widehat{\mathbf{B}}^{-1}$ directly. The

resulting approximation to $\mathbf{F}_{opt}$ is

$$\mathbf{F}_{KSTAP} = (\mathbf{I} - \mathbf{U}_A\mathbf{U}_A^H) \otimes (\mathbf{I} - \mathbf{U}_B\mathbf{U}_B^H) = \mathbf{F}_A \otimes \mathbf{F}_B. \tag{4.26}$$

We denote by Kron-STAP the method using LR-Kron to estimate the covariance and (4.26) to filter the data. This alternative approximation has significant appeal. Note that it projects away both the spatial and temporal clutter subspaces, instead of only the joint spatio-temporal subspace. This is appealing because by (4.8), no moving target should lie in the same spatial subspace as the clutter, and, as noted in Section 4.2, if the dimension of the clutter temporal subspace is sufficiently small relative to the dimension q of the entire temporal space, moving targets will have temporal factors ($\mathbf{b}$) whose projection onto the clutter temporal subspace are small. Note that in the event r_b is very close to q, either truncating r_b to a smaller value (e.g., determined by cross validation) or setting $\mathbf{U}_B = 0$ is recommended to avoid canceling both clutter and moving targets.

Our clutter model has spatial factor rank $r_a = 1$ (4.9), implying that the $\mathbf{F}_{KSTAP}$ defined in (4.26) projects the array signal $\mathbf{x}$ onto a $(p-1)(q-r_b)$ dimensional subspace. This is significantly smaller than the $pq - r_b$ dimensional subspace onto which (4.25) and unstructured STAP project the data. As a result, much more of the clutter that "leaks" outside the primary subspace can be canceled, thus increasing the SINR and allowing lower amplitude moving targets to be detected.

4.3.4 Computational Complexity

Once the filters are learned, the computational complexity depends on the implementation of the filter and does not depend on the covariance estimation method that determined the filter.

The computational complexity of learning the LR-STAP filter is dominated by

the computation of the clutter subspace, which is $O(p^3q^3)$. Our LR-Kron estimator (Algorithm 1) is iterative, with each iteration having $O(p^2q^2) + O(q^3) + O(q^2p^2) + O(p^3) = O(p^2q^2 + p^3 + q^3)$ computations. If the number of iterations needed is small and p, q are large, there will be significant computational gains over LR-STAP.

4.4 SINR Performance

For a STAP filter matrix $\mathbf{F}$ and steering vector $\mathbf{d}$, the data filter vector is given by (4.13): $\mathbf{w} = \mathbf{Fd}$ (*Ginolhac et al.*, 2014). With a target return of the form $\mathbf{x}_{target} = \alpha\mathbf{d}$, the filter output is given by (4.10), and the SINR by (4.11).

Define SINR_{max} to be the optimal SINR, achieved at $\mathbf{w}_{opt} = \mathbf{F}_{opt}\mathbf{d}$ (4.12).

Suppose that the clutter has covariance of the form (4.9). Assume that the target steering vector $\mathbf{d}$ lies outside both the temporal and spatial clutter subspaces as justified in (*Ginolhac et al.*, 2014). Suppose that LR-STAP is set to use r principal components. Suppose further that Kron STAP uses 1 spatial principal component and r temporal components, so that the total number of principal components of LR-STAP and Kron STAP are equivalent. Under these assumptions, if the noise variance σ^2 approaches zero the SINR achieved using LR-STAP, Kron STAP or spatial Kron STAP with infinite training samples achieves SINR_{max} (*Ginolhac et al.*, 2014).

We analyze the asymptotic convergence rates under the finite sample regime. Define the SINR Loss ρ as the loss of performance induced by using the estimated STAP filter $\widehat{\mathbf{w}} = \widehat{\mathbf{F}}\mathbf{d}$ instead of $\mathbf{w}_{opt}$:

$$\rho = \frac{\text{SINR}_{out}}{\text{SINR}_{max}}, \tag{4.27}$$

where SINR_{out} is the output signal to interference ratio when using $\widehat{\mathbf{w}}$.

It is shown in (*Ginolhac et al.*, 2014) that for large n and small σ, the expected

SINR Loss of LR-STAP is

$$E[\rho] = 1 - \frac{r}{n}. \tag{4.28}$$

This approximation is obtained specializing the result in (*Ginolhac et al.*, 2014, Prop. 3.1) to the case of small σ.

We now turn to Kron STAP. Note that the Kron STAP filter can be decomposed into a spatial stage (filtering by $\mathbf{F}_{spatial}$) and a temporal stage (filtering by $\mathbf{F}_{temp}$):

$$\mathbf{F}_{KSTAP} = \mathbf{F}_A \otimes \mathbf{F}_B = \mathbf{F}_{spatial}\mathbf{F}_{temp} \tag{4.29}$$

where $\mathbf{F}_{spatial} = \mathbf{F}_A \otimes \mathbf{I}$ and $\mathbf{F}_{temp} = \mathbf{I} \otimes \mathbf{F}_B$ (4.26). When the clutter covariance fits our model, either the spatial or the temporal stage is sufficient to project away the clutter subspace. Assume one adopts the naive estimator

$$\widehat{\mathbf{A}} = \mathrm{EIG}_1\left(\frac{1}{q}\sum_i \mathbf{S}(i,i)\right) = \widehat{\psi}\widehat{\mathbf{h}}\widehat{\mathbf{h}}^H \tag{4.30}$$

for the spatial subspace $\mathbf{h}$ ($\|\mathbf{h}\|_2 = 1$). For large n and small σ, the expected SINR Loss of Kron STAP using the estimator (4.30) for the spatial subspace is given by

$$E[\rho] = 1 - \frac{1}{n}. \tag{4.31}$$

This result is established in (*Greenewald and Hero III*, 2015, Theorem IV.2). The proof is based on applying the LR-STAP result (4.28) to an equivalent rank-one estimator. Since by (4.7) the full clutter covariance has rank $r \sim q$, the gains of using Kron STAP over LR-STAP (which decays linearly with r) can be quite significant.

Next we establish the robustness of the proposed Kron STAP algorithm to estimation errors, for which the SINR loss (4.27) can only be empirically estimated

from training data. Specifically, consider the case where the spatial covariance has estimation errors, either due to subspace estimation error or to $\mathbf{A}$ having a rank greater than one, e.g., due to spatially varying calibration errors. Specifically, suppose the estimated (rank one) spatial subspace is $\widetilde{\mathbf{h}}$, giving a Kron STAP spatial filter $\mathbf{F}_{spatial} = (\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H) \otimes \mathbf{I}$. Suppose further that spatial filtering of the data is followed by the temporal filter $\mathbf{F}_{temp}$ based on the temporal subspace $\mathbf{U}_B$ estimated from the training data. Define the resulting SINR loss $\rho_t|\widetilde{\mathbf{h}}$ as

$$\rho_t|\widetilde{\mathbf{h}} = \frac{\text{SINR}_{out}}{\text{SINR}_{max}(\widetilde{\mathbf{h}})} \tag{4.32}$$

where $\text{SINR}_{max}(\widetilde{\mathbf{h}})$ is the maximum achievable SINR given that the spatial filter is fixed at $\mathbf{F}_{spatial} = (\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H) \otimes \mathbf{I}$.

We then can obtain the following.Suppose that a value for the spatial subspace estimate $\widetilde{\mathbf{h}}$ (with $\|\widetilde{\mathbf{h}}\|_2 = 1$) and hence $\mathbf{F}_{spatial}$ is fixed. Let the steering vector for a constant Doppler target be $\mathbf{d} = \mathbf{d}_A \otimes \mathbf{d}_B$ per (4.8), and suppose that $\mathbf{d}_A$ is fixed and $\mathbf{d}_B$ is arbitrary. Then for large n and small σ, the SINR loss from using an estimate of $\mathbf{U}_B$ follows

$$E[\rho_t|\widetilde{\mathbf{h}}] \approx 1 - \frac{\kappa r_b}{n}, \qquad \kappa = \frac{\widetilde{\mathbf{d}}_A^H \mathbf{A} \widetilde{\mathbf{d}}_A}{\widetilde{\mathbf{h}}^H \mathbf{A} \widetilde{\mathbf{h}}}. \tag{4.33}$$

where $\widetilde{\mathbf{d}}_A = \frac{(\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H)\mathbf{d}_A}{\|(\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H)\mathbf{d}_A\|_2}$. A proof sketch of this result is in Appendix 4.7.2, and more details are given in (*Greenewald and Hero III*, 2015, Theorem IV.3).

Note that in the $n \gg p$ regime (relevant when $q \gg p$), $\widetilde{\mathbf{h}} \approx \mathbf{h}$, where $\mathbf{h}$ is the first singular vector of $\mathbf{A}$. This gives $\widetilde{\mathbf{h}}^H \mathbf{A} \widetilde{\mathbf{h}} \approx s_A^{(1)}$ and $\kappa \to 0$ if $\mathbf{A}$ is indeed rank one. Hence, κ can be interpreted as quantifying the adverse effect of mismatch between $\mathbf{A}$ and its estimate. From (4.33) it is seen that cancelation of the moving targets is avoided when $r_b \ll q$. Furthermore, since in the ideal large sample regime all the clutter is removed by the spatial stage, r_b can be smaller than rank($\mathbf{B}$), resulting in

higher SINR.

In the next section, we provide empirical finite sample validation of these asymptotic results on robustness of the proposed Kron STAP algorithm.

4.5 Numerical Results

4.5.1 Dataset

For evaluation of the proposed Kron STAP methods, we use measured data from the 2006 Gotcha SAR GMTI sensor collection (*Scarborough et al.*, 2009). This dataset consists of SAR passes through a circular path around a small scene containing various moving and stationary civilian vehicles. The example images shown in the figures are formed using the backprojection algorithm with Blackman-Harris windowing as in (*Newstadt et al.*, 2014). For our experiments, we use 31 seconds of data, divided into 1 second (2171 pulse) coherent integration intervals.

As there is no ground truth for all targets in the Gotcha imagery, target detection performance cannot be objectively quantified by ROC curves. We rely on non ROC measures of performance for the measured data, and use synthetically generated data to show ROC performance gains. In several experiments we do make reference to several higher amplitude example targets in the Gotcha dataset. These were selected by comparing and analyzing the results of the best detection methods available.

4.5.2 Simulations

We generated synthetic clutter plus additive noise samples having a low rank Kronecker product covariance. The covariance we use to generate the synthetic clutter via the SIRV model was learned from a set of example range bins extracted from the Gotcha dataset, letting the SIRV scale parameter τ^2 in (4.5) follow a chi-square distribution. We use $p = 3$, $q = 150$, $r_b = 20$, and $r_a = 1$, and generate both n training

samples and a set of testing samples. The rank of the left Kronecker factor $\mathbf{A}$, r_a, is 1 as dictated by the spatially invariant antenna calibration assumption and we chose $r_b = 20$ based on a scree plot, i.e., 20 was the location of the knee of the spectrum of $\mathbf{B}$. Spatio-temporal Kron-STAP, Spatial-only Kron-STAP, and LR-STAP were then used to learn clutter cancelation filters from the training clutter data.

The learned filters were then applied to testing clutter data, the mean squared value (MS Residual) of the resulting residual (i.e. $(1/M)\sum_{m=1}^{M} \|\mathbf{F}\mathbf{x}_m\|_2^2$) was computed, and the result is shown in Figure 4.1 as a function of n. The results illustrate the much slower convergence rate of unstructured LR-STAP. as compared to the proposed Kron STAP, which converges after $n = 1$ sample. The mean squared residual does not go to zero with increasing training sample size because of the additive noise floor.

As an example of the convergence of the Algorithm 1, Figure 5.6 shows logarithmic plots of $F_i - \lim_{i\to\infty} F_i$ as a function of iteration i, where $F_i = \|\mathbf{S} - \widehat{\mathbf{A}}_i \otimes \widehat{\mathbf{B}}_i\|_F$. Shown are the results for a sample covariance used in the generation of Figure 4.1 ($n = 50$, noise standard deviation σ_0), and the results for the case of significantly higher noise (noise standard deviation $10\sigma_0$). The zeroth iteration corresponds to the SVD-based initialization in step 2 of Algorithm 1. In both cases, note the rapid convergence of the algorithm, particularly in the first iteration.

To explore the effect of model mismatch due to spatially variant antenna calibration errors ($r_a > 1$), we simulated data with a clutter spatial covariance $\mathbf{A}$ having rank 2 with non-zero eigenvalues equal to 1 and $1/30^2$. The STAP algorithms remain the same with $r_a = 1$, and synthetic range bins containing both clutter and a moving target are used in testing the effect of this model mismatch on the STAP algorithms. The STAP filter response, maximized over all possible steering vectors, is used as the detection statistic. The AUC of the associated ROC curves is plotted in Figure 4.3 as a function of the number of training samples. Note again the poor performance

and slow convergence of LR-STAP, and that spatio-temporal Kron-STAP converges very quickly to the optimal spatial Kron-STAP performance.

Finally, we repeat the AUC vs. sample complexity experiment described in the previous paragraph where 5% of the training data now have synthetic moving targets with random Doppler shifts. The results are shown in Figure 4.4. As predicted by the theory in Subsection 4.3.2, the Kronecker methods remain largely unaffected by the presence of corrupting targets in the training data until the very low sample regime, whereas significant losses are sustained by LR-STAP. This confirms the superior robustness of the proposed Kronecker structured covariance used in our Kron STAP method.

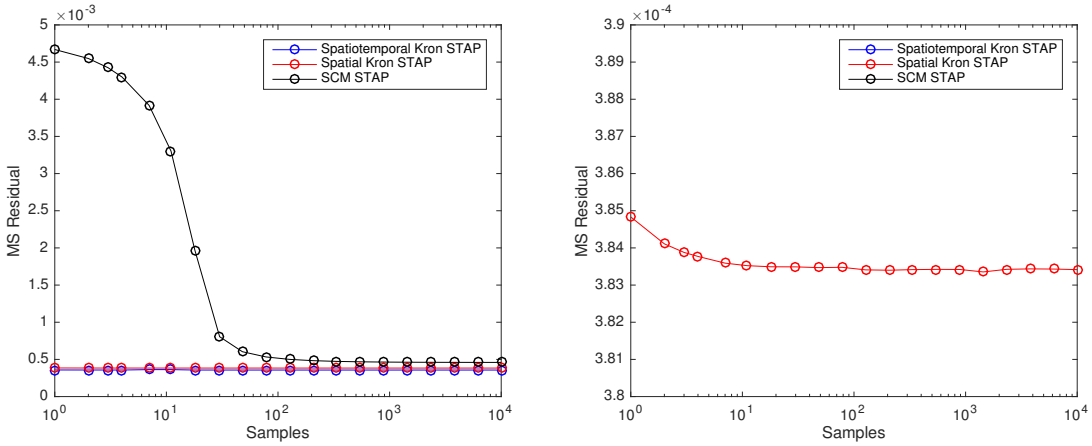


Figure 4.1: Left: Average mean squared residual (MSR), determined by simulations, as a function of the number of training samples, of noisy synthetic clutter filtered by spatio-temporal Kron STAP, spatial only Kron STAP, and unstructured LR-STAP (SCM STAP) filters. On the right a zoomed in view of a Kron STAP curve is shown. Note the rapid convergence and low MSE of the Kronecker methods.

4.5.3 Gotcha Experimental Data

In this subsection, STAP is applied to the Gotcha dataset. For each range bin we construct steering vectors $\mathbf{d}_i$ corresponding to 150 cross range pixels. In single antenna

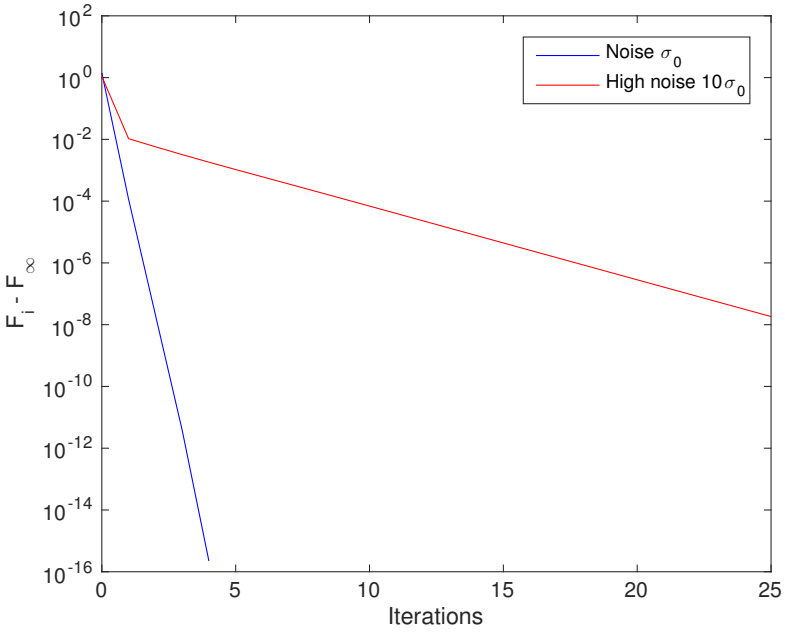


Figure 4.2: Convergence of the LR-Kron algorithm for estimation of the covariance of Figure 1 with $n = 50$. The baseline noise (standard deviation σ_0) case is shown, along with a high noise example with noise standard deviation $10\sigma_0$. Shown are logarithmic plots of $F_i - \lim_{i\to\infty} F_i$ where $F_i = \|\mathbf{S} - \mathbf{A}_i \otimes \mathbf{B}_i\|_F$ as a function of iteration i. Note the rapid convergence of the algorithm.

SAR imagery, each cross range pixel is a Doppler frequency bin that corresponds to the cross range location for a stationary target visible at that SAR Doppler frequency, possibly complemented by a moving target that appears in the same bin. Let $\mathbf{D}$ be the matrix of steering vectors for all 150 Doppler (cross range) bins in each range bin. Then the SAR images at each antenna are given by $\widetilde{\mathbf{x}} = \mathbf{I} \otimes \mathbf{D}^H \mathbf{x}$ and the STAP output for a spatial steering vector $\mathbf{h}$ and temporal steering $\mathbf{d}_i$ (separable as noted in (4.8)) is the scalar

$$y_i(\mathbf{h}) = (\mathbf{h} \otimes \mathbf{d}_i)^H \mathbf{F}\mathbf{x} \tag{4.34}$$

Due to their high dimensionality, plots for all values of $\mathbf{h}$ and i cannot be shown. Hence, for interpretability we produce images where for each range bin the ith pixel

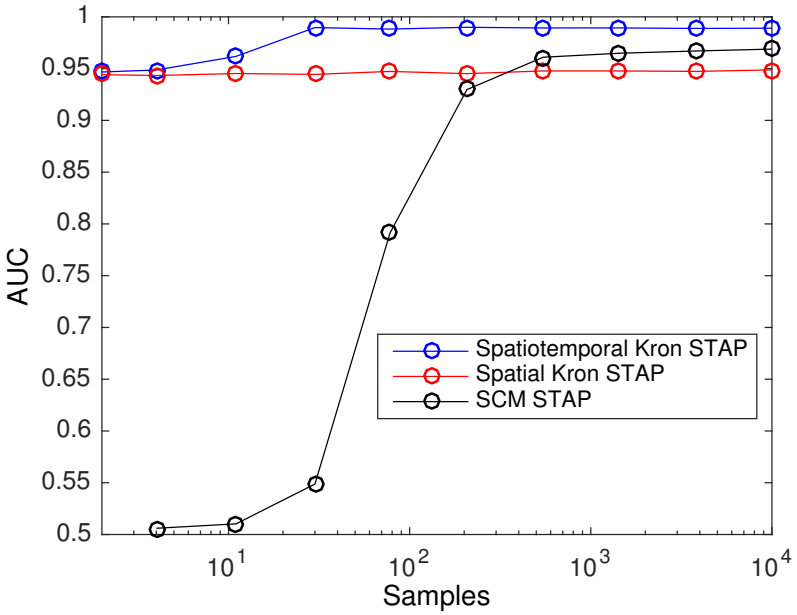


Figure 4.3: Area-under-the-curve (AUC) for the ROC associated with detecting a synthetic target using the steering vector with the largest return, when slight spatial nonidealities exist in the true clutter covariance. Note the rapid convergence of the Kronecker methods as a function of the number of training samples, and the superior performance of spatio-temporal Kron STAP to spatial-only Kron STAP when the target's steering vector $\mathbf{d}$ is unknown.

is set as $\max_{\mathbf{h}} |y_i(\mathbf{h})|$. More sophisticated detection techniques could invoke priors on $\mathbf{h}$, but we leave this for future work.

Shown in Figure 4.6 are results for several examplar SAR frames, showing for each example the original SAR (single antenna) image, the results of spatio-temporal Kronecker STAP, the results of Kronecker STAP with spatial filter only, the amount of enhancement (smoothed dB difference between STAP image and original) at each pixel of the spatial only Kronecker STAP, standard unstructured STAP with $r = 25$ (similar rank to Kronecker covariance estimate), and standard unstructured STAP with $r = 40$. Note the significantly improved contrast of Kronecker STAP relative to the unstructured methods between moving targets (high amplitude moving targets marked in red in the figure) and the background. Additionally, note that both spatial and temporal filtering achieve significant gains. Due to the lower dimensionality, LR-

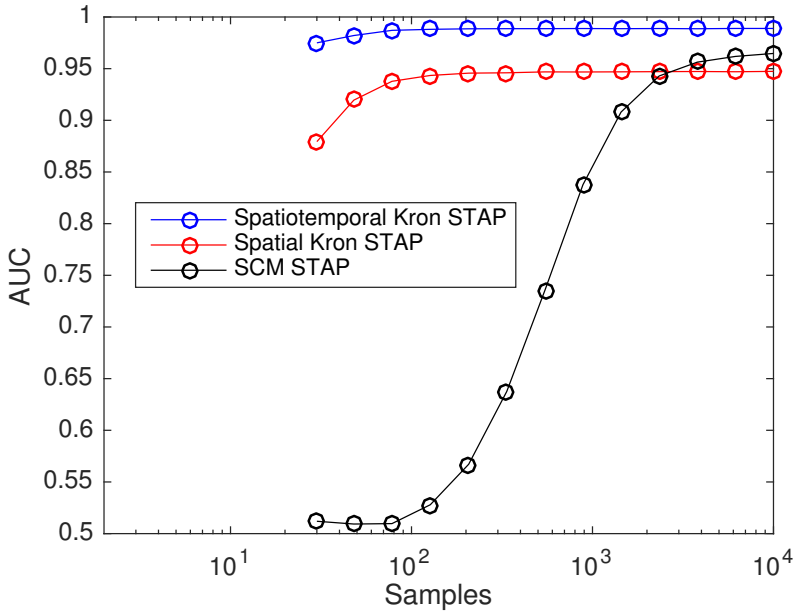


Figure 4.4: Robustness to corrupted training data: AUCs for detecting a synthetic target using the maximum steering vector when (in addition to the spatial nonidealities) 5% of the training range bins contain targets with random location and velocity in addition to clutter. Note that relative to Figure 4.3 LR-STAP has degraded significantly, whereas the Kronecker methods have not.

STAP achieves its best performance for the image with fewer pulses, but still remains inferior to the Kronecker methods.

To analyze convergence behavior, a Monte Carlo simulation was conducted where random subsets of the (bright object free) available training set were used to learn the covariance and the corresponding STAP filters. The filters were then used on each of the 31 1-second SAR imaging intervals and the MSE between the results and the STAP results learned using the entire training set were computed (Figure 4.5). Note the rapid convergence of the Kronecker methods relative to the SCM based method, as expected.

Figure 4.5 (bottom) shows the normalized ratio of the RMS magnitude of the 10 brightest filter outputs $y_i(\mathbf{h})$ for each ground truthed target to the RMS value of the background, computed for each of the STAP methods as a function of the number

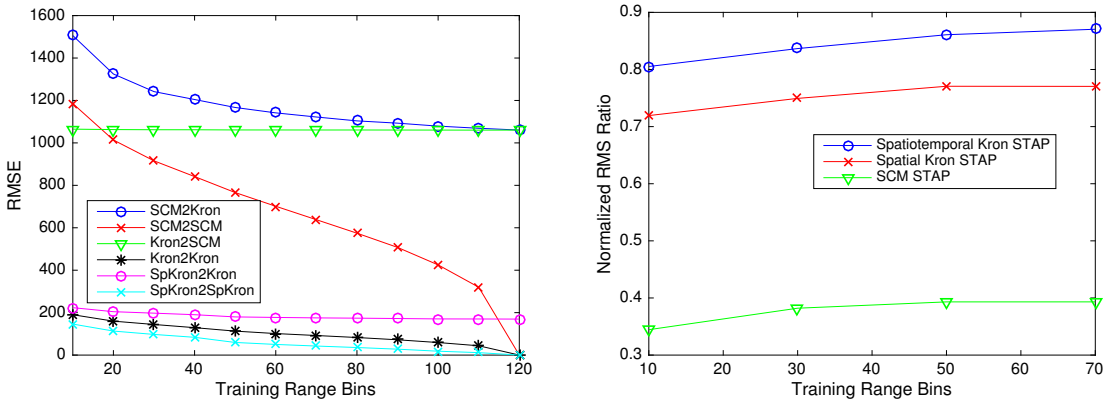


Figure 4.5: Gotcha dataset. Left: Average RMSE of the output of the Kronecker, spatial only Kronecker, and unstructured STAP filters relative to each method's maximum training sample output. Note the rapid convergence and low RMSE of the Kronecker methods. Right: Normalized ratio of the RMS magnitude of the brightest pixels in each target relative to the RMS value of the background, for the output of each of Kronecker STAP, spatial Kronecker STAP, and unstructured STAP.

of training samples. This measure is large when the contrast of the target to the background is high. The Kronecker methods clearly outperform LR-STAP.

4.6 Conclusion

In this chapter, we proposed a new method for clutter rejection in high resolution multiple antenna synthetic aperture radar systems with the objective of detecting moving targets. Stationary clutter signals in multichannel single-pass radar were shown to have Kronecker product structure where the spatial factor is rank one and the temporal factor is low rank. Exploitation of this structure was achieved using the Low Rank KronPCA covariance estimation algorithm, and a new clutter cancelation filter exploiting the space-time separability of the covariance was proposed. The resulting clutter covariance estimates were applied to STAP clutter cancelation, exhibiting significant detection performance gains relative to existing low rank covariance estimation techniques. As compared to standard unstructured low rank STAP meth-

ods, the proposed Kronecker STAP method reduces the number of required training samples and enhances the robustness to corrupted training data. These performance gains were analytically characterized using a SIRV based analysis and experimentally confirmed using simulations and the Gotcha SAR GMTI dataset.

4.7 Appendix

4.7.1 Derivation of LR-Kron Algorithm 3

We have the following objective function:

$$\min_{\text{rank}(\mathbf{A})=r_a,\text{rank}(\mathbf{B})=r_b} \|\mathbf{S}-\mathbf{A}\otimes\mathbf{B}\|_F^2. \tag{4.35}$$

To derive the alternating minimization algorithm, fix $\mathbf{B}$ (symmetric) and minimize (4.35) over low rank $\mathbf{A}$:

$$\begin{aligned}
&\arg\min_{\text{rank}(\mathbf{A})=r_a} \|\mathbf{S}-\mathbf{A}\otimes\mathbf{B}\|_F^2 \\
&= \arg\min_{\text{rank}(\mathbf{A})=r_a} \sum_{i,j}^{q} \|\mathbf{S}(i,j)-b_{ij}\mathbf{A}\|_F^2 \\
&= \arg\min_{\text{rank}(\mathbf{A})=r_a} \sum_{i,j}^{q} |b_{ij}|^2\|\mathbf{A}\|_F^2 - 2\text{Re}[b_{ij}\langle\mathbf{A},\mathbf{S}^*(i,j)\rangle] \\
&= \arg\min_{\text{rank}(\mathbf{A})=r_a} \|\mathbf{A}\|_F^2 - 2\text{Re}\left[\left\langle\mathbf{A},\frac{\sum_{i,j}^{q} b_{ij}\mathbf{S}^*(i,j)}{\|\mathbf{B}\|_F^2}\right\rangle\right] \\
&= \arg\min_{\text{rank}(\mathbf{A})=r_a} \left\|\mathbf{A}-\frac{\sum_{i,j}^{q} b_{ij}^*\mathbf{S}(i,j)}{\|\mathbf{B}\|_F^2}\right\|_F^2
\end{aligned} \tag{4.36}$$

where b_{ij} is the i,jth element of $\widehat{\mathbf{B}}$ and b^* denotes the complex conjugate of b. This last minimization problem (4.36) can be solved by the SVD via the Eckart-Young

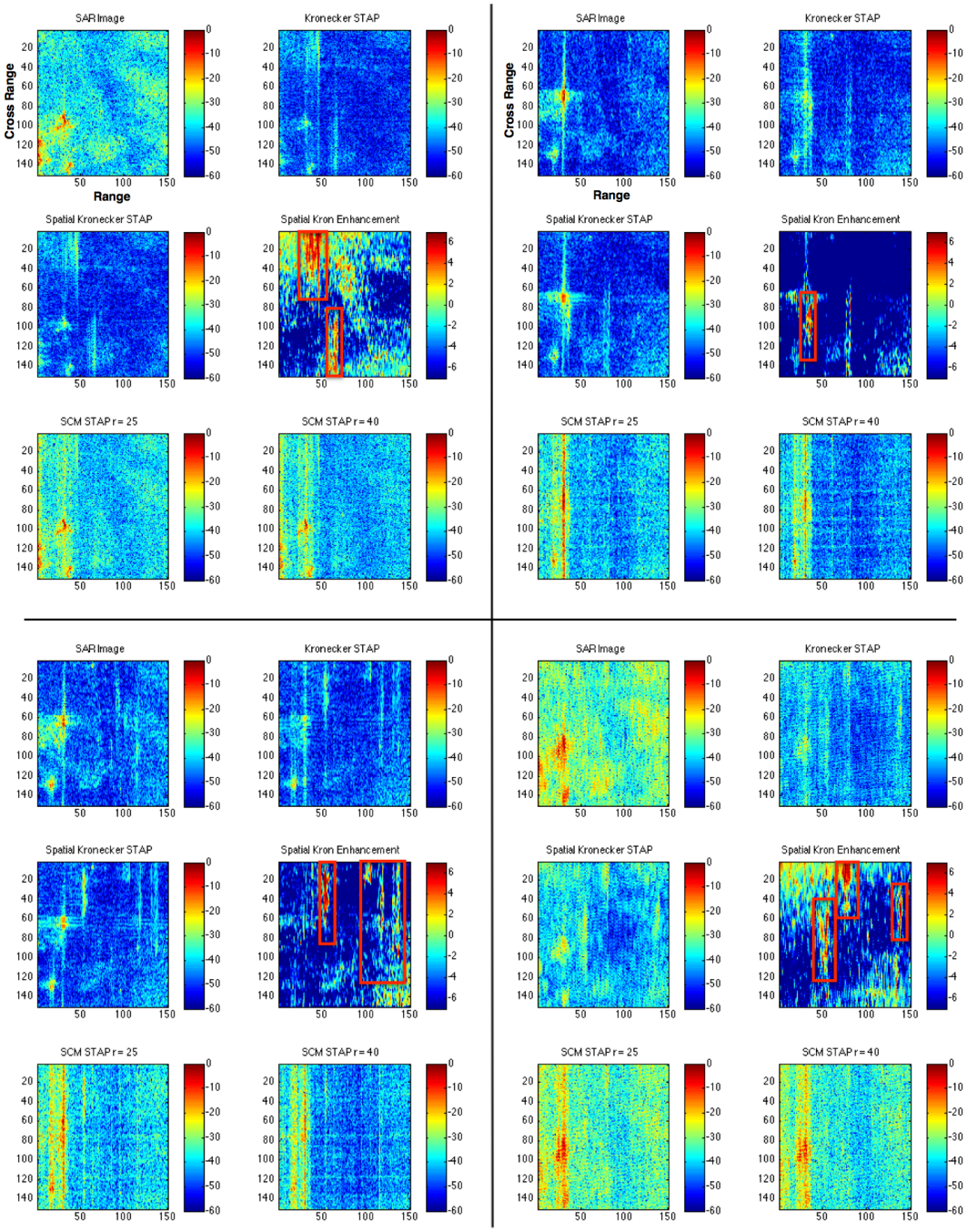


Figure 4.6: Four example radar images from the Gotcha dataset along with associated STAP results. The lower right example uses 526 pulses, the remaining three use 2171 pulses. Several moving targets are highlighted in red in the spatial Kronecker enhancement plots. Note the superiority of the Kronecker methods. Used Gotcha dataset "mission" pass, starting times: upper left, 53 sec.; upper right, 69 sec.; lower left, 72 sec.; lower right 57.25 sec.

theorem (*Eckart and Young*, 1936). First define

$$\mathbf{R}_A = \frac{\sum_{i,j}^{q} b_{ij}^* \mathbf{S}(i,j)}{\|\mathbf{B}\|_F^2}, \tag{4.37}$$

and let $\mathbf{u}_i^A, \sigma_i^A$ be the eigendecomposition of $\mathbf{R}_A$. The eigenvalues are real and positive because $\mathbf{R}_A$ is positive semidefinite (psd) Hermitian if $\mathbf{B}$ is psd Hermitian (*Werner et al.*, 2008). Hence by Eckardt-Young the unique minimizer of the objective (4.36) is

$$\widehat{\mathbf{A}}(\mathbf{B}) = \mathrm{EIG}_{r_a}(\mathbf{R}_A) = \sum_{i=1}^{r_a} \sigma_i \mathbf{u}_i^A (\mathbf{u}_i^A)^H. \tag{4.38}$$

Note that unless either $\mathbf{S}$ or $\mathbf{B}$ is identically zero, since $\mathbf{B}$ is psd $\mathbf{R}_A$ and hence $\widehat{A}(\mathbf{B})$ will be nonzero.

Similarly, minimizing (4.35) over $\mathbf{B}$ with fixed positive semidefinite Hermitian $\mathbf{A}$ gives the unique minimizer

$$\widehat{\mathbf{B}}(\mathbf{A}) = \mathrm{EIG}_{r_b}(\mathbf{R}_B) = \sum_{i=1}^{r_b} \sigma_i^B \mathbf{u}_i^B (\mathbf{u}_i^B)^H, \tag{4.39}$$

where now $\mathbf{u}_i^B, \sigma_i^B$ describes the eigendecomposition of

$$\mathbf{R}_B = \frac{\sum_{i,j}^{p} a_{ij}^* \bar{\mathbf{S}}(i,j)}{\|\mathbf{A}\|_F^2}. \tag{4.40}$$

Iterating between computing $\widehat{\mathbf{A}}(\mathbf{B})$ and $\widehat{\mathbf{B}}(\mathbf{A})$ completes the alternating minimization algorithm.

By induction, initializing with either a psd Hermitian $\mathbf{A}$ or $\mathbf{B}$ and iterating until convergence will result in an estimate $\widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}}$ of the covariance that is psd Hermitian since the set of positive semidefinite Hermitian matrices is closed.

Since for nonzero $\mathbf{S}$ a nonzero $\mathbf{B}_k$ implies a nonzero $\mathbf{A}_{k+1}$ and vice versa, $\mathbf{A}_k$ and

$\mathbf{B}_k$ will never go to zero. Hence, the closed-form factorwise minimizers (4.38) and (4.39) are always uniquely defined, and cannot increase the value of the objective. Thus monotonic convergence of the objective to a value b is ensured (*Byrne*, 2013). Since the coordinatewise minimizers are always unique, if (4.38) or (4.39) result in either $\mathbf{A}_{k+1} \neq \mathbf{A}_k$ or $\mathbf{B}_{k+1} \neq \mathbf{B}_k$ respectively, then the objective function must strictly decrease. Thus, cycles are impossible and $\mathbf{A}_k, \mathbf{B}_k$ must converge to values $\mathbf{A}_*, \mathbf{B}_*$. The value of the objective at that point must be a stationary point by definition, else $\mathbf{A}_*, \mathbf{B}_*$ would not be coordinatewise minima.

4.7.2 KronSTAP SINR: Proof Sketch of Theorem (4.33)

This is a proof sketch, the full proof can be found in our technical report (*Greenewald and Hero III*, 2015, Theorem IV.3).

After the spatial stage of Kron STAP projects away the estimated spatial subspace $\widetilde{\mathbf{h}}$ ($\|\widetilde{\mathbf{h}}\|_2 = 1$) the remaining clutter has a covariance given by

$$((\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H)\mathbf{A}(\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H)) \otimes \mathbf{B}. \tag{4.41}$$

By (4.8), the steering vector for a (constant Doppler) moving target is of the form $\mathbf{d} = \mathbf{d}_A \otimes \mathbf{d}_B$. Hence, the filtered output is

$$\begin{aligned} y = \mathbf{w}^H\mathbf{x} &= \mathbf{d}^H\mathbf{F}\mathbf{x} \\ &= (\mathbf{d}_A^H \otimes \mathbf{d}_B^H)(\mathbf{F}_A \otimes \mathbf{F}_B)\mathbf{x} \\ &= ((\mathbf{d}_A^H\mathbf{F}_A) \otimes (\mathbf{d}_B^H\mathbf{F}_B))\mathbf{x} \\ &= \mathbf{d}_B^H\mathbf{F}_B \left(\left(\mathbf{d}_A^H \left(\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H\right)\right) \otimes \mathbf{I}\right)\mathbf{x} \end{aligned} \tag{4.42}$$

Let $\widetilde{\mathbf{d}}_A = (\mathbf{I} - \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H)\mathbf{d}_A$ and define $\widetilde{\mathbf{c}} = \left(\widetilde{\mathbf{d}}_A^H \otimes \mathbf{I}\right)\mathbf{c}$. Then

$$y = \mathbf{d}_B^H \mathbf{F}_B(\tau\widetilde{\mathbf{c}} + \widetilde{\mathbf{n}}), \tag{4.43}$$

where $\widetilde{\mathbf{n}} = (\widetilde{\mathbf{d}}_A \otimes \mathbf{I})\mathbf{n}$ and

$$\begin{aligned} \mathrm{Cov}[\widetilde{\mathbf{c}}] =& (\widetilde{\mathbf{d}}_A^H \mathbf{A} \widetilde{\mathbf{d}}_A)\mathbf{B} \\ \mathrm{Cov}[\widetilde{\mathbf{n}}] =& \sigma^2 \mathbf{I}, \end{aligned} \tag{4.44}$$

which are proportional to $\mathbf{B}$ and $\mathbf{I}$ respectively. The scalar $(\widetilde{\mathbf{d}}_A^H \mathbf{A} \widetilde{\mathbf{d}}_A)$ is small if $\mathbf{A}$ is accurately estimated, hence improving the SINR but not affecting the SINR loss. Thus, the temporal stage of Kron STAP is equivalent to single channel LR-STAP with clutter covariance $(\widetilde{\mathbf{d}}_A^H \mathbf{A} \widetilde{\mathbf{d}}_A)\mathbf{B}$ and noise variance σ^2.

Given a fixed $\widehat{\mathbf{A}} = \widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H$, Algorithm 3 dictates (4.40), (4.39) that

$$\begin{aligned} \mathbf{R}_B &= \sum_{i,j}^{p} \widetilde{h}_i^* \widetilde{h}_j^* \bar{\mathbf{S}}(i,j) \\ \widehat{\mathbf{B}} &= \mathrm{EIG}_{r_b}(\mathbf{R}_B), \end{aligned} \tag{4.45}$$

which is thus the low rank approximation of the sample covariance of

$$\mathbf{x}_h = \mathbf{x}_{c,h} + \mathbf{n}_h = (\widetilde{\mathbf{h}} \otimes \mathbf{I})^H (\mathbf{x}_c + \mathbf{n}). \tag{4.46}$$

Since $\mathbf{x}_c = \tau\mathbf{c}$, $\mathbf{x}_{c,h} = \tau(\widetilde{\mathbf{h}} \otimes \mathbf{I})^H \mathbf{c}$ is an SIRV (Gaussian random vector $(\widetilde{\mathbf{h}} \otimes \mathbf{I})^H \mathbf{c}$ scaled by τ) with

$$\mathrm{Cov}[\mathbf{x}_{c,h}] = \tau^2 (\widetilde{\mathbf{h}}^H \mathbf{A} \widetilde{\mathbf{h}})\mathbf{B} \tag{4.47}$$

Furthermore, $\mathbf{n}_h = (\widetilde{\mathbf{h}} \otimes \mathbf{I})^H \mathbf{n}$ which is Gaussian with covariance $\sigma^2\mathbf{I}$. Thus, in both training and filtering the temporal stage of Kron STAP is exactly equivalent to single

channel LR STAP. Hence we can directly apply the methods used to prove the bound for LR STAP, which after some work results in (4.33) as desired.

CHAPTER V

The Tensor Graphical Lasso (TeraLasso)

The sum of Kronecker products model in Chapter III is an expressive model and incorporates significant structure. It is not trivial, however, to extend it to tensor valued data (low separation rank approximation of tensor covariances is in fact NP hard), or to incorporate sparsity in the low-separation rank part of the model. The K-way Kronecker product allows for both modeling of K-way tensors and incorporation of sparsity, but the assumption that the process is separable in every dimension is very strong, and hence not ideal.

Instead, in this chapter we consider a sum of K-way Kronecker products model that gives a convex penalized maximum likelihood objective function, allows for sparse factors, is nonseparable, and significantly less restrictive than the Kronecker product. We demonstrate single-sample convergence of our estimator and apply it to meteorological and brain EEG datasets.

5.1 Introduction

Learning useful structured models of high-dimensional datasets from relatively few training samples is an important task in signal processing and machine learning. Many high-dimensional problems involve data that is matrix- or tensor-valued, and exploiting that structure is critical in the learning process. Some common examples of tensor-valued data include spatial and spatio-temporal data such as videos, meteorological data, geolocated time series, fMRI, EEG and other medical imaging modalities, synthetic aperture radar, gene expression data, wireless communication applications (*Werner et al.*, 2008) and sensor networks. Applications of tensor modeling are far-ranging, including signal processing, numerical linear algebra, computer vision, numerical analysis, data mining, graph analysis, and neuroscience. Many such applications are discussed in (*Kolda and Bader*, 2009).

Covariance modeling is a fundamental problem in multivariate statistical analysis. In this work, we focus on developing structured, sparse inverse covariance models for high-dimensional tensor data. Our notation and terminology follows that of (*Kolda and Bader*, 2009). Consider the K-order data tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_K}$. For convenience, define $\mathbf{p} = [d_1, \ldots, d_K]$, and set

$$p = \prod_{k=1}^{K} d_k, \quad m_k = \prod_{i \neq k} d_i = \frac{p}{d_k}.$$

We propose the TEnsor gRAphical Lasso (TeraLasso) model

$$(\mathrm{Cov}[\mathrm{vec}(X^T)])^{-1} = \Sigma^{-1} = \Omega = \underbrace{\Psi_1 \oplus \cdots \oplus \Psi_K}_{K \text{ terms}}, \tag{5.1}$$

where the Ψ_k are sparse, each corresponding to a graph across the kth dimension of the data tensor, and $\oplus$ denotes the kronecker sum, $A \oplus B = A \otimes I_n + I_m B \otimes$, for $m \times m$ and $n \times n$ matrices A and B, respectively, I_m is the $m \times m$ identity matrix,

and $\otimes$ denotes the Kronecker product operator. We have used $\mathrm{vec}(X)$ as in (*Kolda and Bader*, 2009), and defined $X^T \in \mathbb{R}^{d_K \times \cdots \times d_1}$ by analogy to the matrix transpose, i.e. $[X^T]_{i_1,\ldots,i_K} = X_{i_K,\ldots,i_1}$.

Many methods for first-moment modeling of tensor-valued data have been proposed (*Kolda and Bader*, 2009). Many of these involve low-rank factor decompositions, including PARAFAC and CANDECOMP (*Harshman and Lundy*, 1994; *Faber et al.*, 2003) and Tucker decomposition-based methods such as (*Tucker*, 1966) and (*Hoff et al.*, 2016). Recently, several works have found that such modeling can be improved by taking into account the second moment of the data (i.e. covariance), which has typically been modeled using Kronecker products (*Xu et al.*, 2011; *Zhe et al.*, 2015; *Pouryazdian et al.*, 2016).

As the covariance encodes relationships and interactions between variables, it is a powerful tool for modeling multivariate distributions, allowing inference, likelihood calculation, and prediction. For tensor-valued data, however, the very large number of free parameters, of order $O(\prod_{k=1}^{K} d_k^2)$, makes the unstructured covariance model impractical. As a result, there has been increasing interest in developing structured covariance models appropriate for matrix- and tensor-valued data (*Tsiligkaridis and Hero*, 2013; *Zhou*, 2014; *Werner et al.*, 2008; *Sun et al.*, 2015; *Xu et al.*, 2011; *Greenewald and Hero*, 2015; *Allen and Tibshirani*, 2010). As the most common example, the Kronecker product covariance

$$\Sigma = A_1 \otimes A_2 \otimes \cdots \otimes A_K \tag{5.2}$$

exploits the natural tensor arrangement of the variables and forms a joint model from K lower-dimensional models each defined along one tensor axis. When the covariance (5.2) describes a Gaussian distribution, this model is known as the *matrix normal* distribution (*Dawid*, 1981) ($K = 2$) and *tensor normal* for $K > 2$ (*Sun et al.*,

2015; *Xu et al.*, 2011), with a penalized version in (*Allen and Tibshirani*, 2010) called the transposable covariance model. Note that the inverse covariance $\Omega = \Sigma^{-1}$ of a Kronecker product covariance (5.2) also has a Kronecker product representation (*Laub*, 2005).

In a Gaussian graphical model edges correspond to nonzero entries in the precision matrix $\Omega = \Sigma^{-1}$. The Kronecker product graphical model (*Zhou*, 2014; *Tsiligkaridis et al.*, 2013; *Sun et al.*, 2015) estimates K sparse factor precision matrices $\Psi_k \in \mathbb{R}^{d_k \times d_k}$ (Figure 5.1(a-c)), setting $\Omega = \Psi_1 \otimes \cdots \otimes \Psi_K$. This model has excellent statistical convergence results, but creates an overall graph where each edge in the final model is the product of K separate edges from the factor graphs Ψ_k. A proliferation of inter-related edges is thus created, illustrated in Figure 5.1 (right), with each edge in the factor models affecting up to m_k^2 total edges in the final graph.

The Kronecker product model became popular because of the separability interpretation of the factors, hence it is perhaps not surprising that it is not that natural of a parameterization of a sparse graphical model. Indeed, while the separable structure of (5.2) is intuitive for a wide variety of real matrix-valued and spatio-temporal processes, the imposed structure is often restrictive and inaccurate in a wider set of applications (*Tsiligkaridis and Hero*, 2013; *Kalaitzis et al.*, 2013; *Greenewald and Hero*, 2015). As a result, there has been significant recent effort on modeling covariances using other Kronecker representations, with the goal of achieving comparable reductions in the number of parameters while expanding their applicability (*Tsiligkaridis and Hero*, 2013; *Rudelson and Zhou*, 2015; *Greenewald and Hero*, 2015).

Instead of the Kronecker product, it is more desirable to have each edge in the factor model map directly to edges in the final model. One such model is the Kronecker sum model, shown in Figure 5.1 (left), that maps the i, jth edge in the kth factor Ψ_k to edges between nodes in ith and jth position along the kth tensor mode. This type of structure implies that conditional dependence moves along axes or modes, in a

manner analogous to several popular forms of Markov random fields used in computer vision and other applications (*Wang et al.*, 2013; *Diebel and Thrun*, 2005).

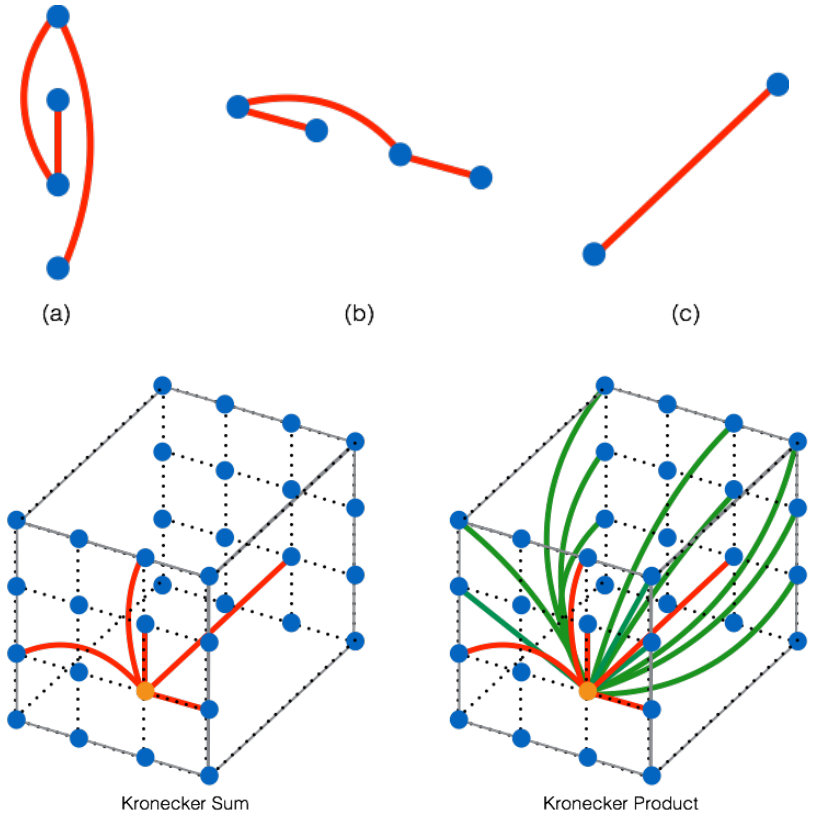


Figure 5.1: Tensor graphical models on a $4 \times 4 \times 2$ Cartesian node grid. Consider three graphical models, one along each axis (a-c). The Kronecker sum and Kronecker product of these graphs are shown at the bottom left and right of the figure, with only the edges emanating from the orange node indicated. The Kronecker sum (64 total edges) preserves the sparsity of the axis graphs (a-c), forming a joint model where each edge is associated with a single edge in an axis graph. The Kronecker product (184 total edges), on the other hand, creates an explosion of edges (marked green) each with a weight a multiple of three separate weights from the axis graphs. Hence, in many situations the Kronecker sum is a more natural and interpretable tensor expansion of sparse graphical models.

It turns out that this hypothesized structure (*Kalaitzis et al.*, 2013) is exactly equivalent to modeling the offdiagonal of Ω as a Kronecker sum $\Psi_1^- \oplus \cdots \oplus \Psi_K^-$, where $\Psi^- = \Psi - \text{diag}(\Psi)$ is the result of setting the diagonal entries of a matrix Ψ to zero.

To simplify the multiway Kronecker notation, we let

$$I_{[d_{k:\ell}]} = \underbrace{I_{d_k} \otimes \cdots \otimes I_{d_\ell}}_{\ell-\text{k+1 factors}}$$

where $\ell \geq k$. Using this notation, the K-way Kronecker sum can be written as

$$\Omega^- = \Psi_1^- \oplus \cdots \oplus \Psi_K^- = \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \Psi_k^- \otimes I_{[d_{k+1:K}]}.$$

It remains to select a structured model for the diagonal elements of Ω. The most natural solution is to use the full Kronecker sum model:

$$\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K. \tag{5.3}$$

This Kronecker sum model, as opposed to the Kronecker product model, is the focus of this paper. To illustrate, in Figure 5.2 we show the inverse covariance (left) and covariance (right) corresponding to a $K = 3$ Kronecker sum of autoregressive-1 (AR-1) graphs with $d_k = 4$.

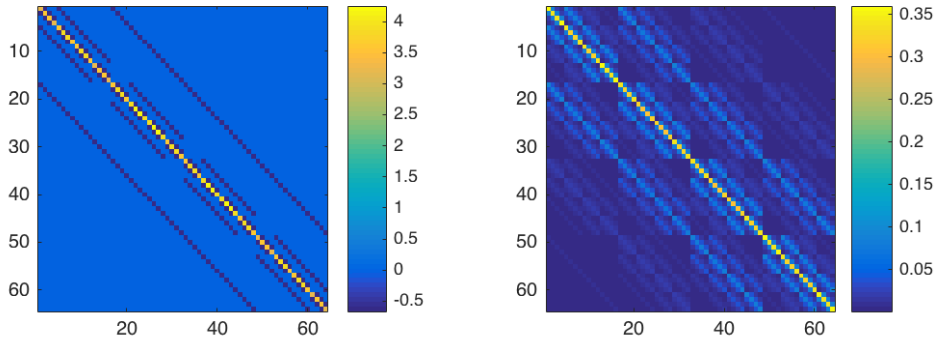


Figure 5.2: Kronecker sum model. Left: Sparse $4 \times 4 \times 4$ Cartesian AR precision matrix $\Omega = \Psi_1 \oplus \Psi_2 \oplus \Psi_3$. Right: Covariance matrix $\Sigma = \Omega^{-1}$. Note the nested block structure, especially of the covariance.

As compared to the Kronecker model, the Kronecker sum model (5.3) has several

attractive properties. As illustrated in Figure 5.1 it provides a sparser representation of the inverse covariance. Furthermore, under the model (5.3) the inverse covariance has Kronecker product eigenvectors with linearly related eigenvalues, as opposed to the multiplicative eigenvalues of the Kronecker product. For simplicity, we consider the $K = 2$ case, but the result generalizes to the full tensor case (see Property 3 in Appendix A). Suppose that $\Psi_1 = U_1\Lambda_1 U_1^T$ and $\Psi_2 = U_2\Lambda_2 U_2^T$ are the eigendecompositions of Ψ_1 and Ψ_2. Then (*Laub*, 2005), if $\Omega = \Psi_1 \oplus \Psi_2$, the eigendecomposition of Ω is

$$\Omega = \Psi_1 \oplus \Psi_2 = (U_1 \otimes U_2)(\Lambda_1 \oplus \Lambda_2)(U_1 \otimes U_2)^T.$$

Thus, the eigenvectors of the Kronecker sum are the Kronecker products of the eigenvectors of each factor. This "block" structure is evident in the inverse Kronecker sum example in Figure 5.2, and is analyzed further in (*Canuto et al.*, 2014). This parallels the eigenvector structure of the Kronecker product - specifically when $\Omega = \Psi_1 \otimes \Psi_2$

$$\Omega = \Psi_1 \otimes \Psi_2 = (U_1 \otimes U_2)(\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2)^T.$$

Hence, use of the Kronecker sum model can be viewed as replacing the nonconvex, relatively unstable multiplicative eigenvalues of the Kronecker product with a stable, convex linear expression. As the tensor dimension K increases, this structural stability compared to the Kronecker product becomes even more increasingly dominant (K term sums instead of K-order products).

Tensor unfolding or matricization of X along the kth-mode is denoted as $X_{(k)} \in \mathbb{R}^{d_k \times \prod_{\ell \neq k} d_\ell}$, and is formed by concatenating the kth mode fibers $X_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_K}$ as columns (*Kolda and Bader*, 2009). In tensor covariance modeling when the dimension p is much larger than the number of samples n, the Gram matrices $S_k = \frac{1}{nm_k}\sum_{i=1}^{n} X_{i,(k)}X_{i,(k)}^T$ are often used to model the rows and columns separately, notably in the covariance estimation approaches of (*Zhou*, 2014; *Kalaitzis et al.*, 2013).

In (*Kalaitzis et al.*, 2013) it was shown that the Kronecker sum model has a rather satisfying connection to these factorwise Gram matrices S_k. Consider (similar to (*Zhou*, 2014)) modeling the likelihood of each S_k independently, i.e. $\ell(X_{(k)}) \propto \exp\{-\langle S_k, \Psi_k\rangle\}$ for some matrix Ψ_k. One may consider forming a joint likelihood (where we undo the normalization of S_k by the unfolding dimension) to jointly maximize the factor likelihoods

$$\ell(X) \propto \exp\left\{-\sum_{k=1}^{K} m_k \langle S_k, \Psi_k\rangle\right\},$$

enabling the joint maximization of the factor likelihoods. Now, in Lemma V.8, we prove

$$L(X) \propto \exp\left\{-\sum_{k=1}^{K} m_k \langle S_k, \Psi_k\rangle\right\} = \exp\left\{-\langle S, \Psi_1 \oplus \cdots \oplus \Psi_K\rangle\right\}.$$

Adding a normalization constant and taking the logarithm, we have

$$\ell(X) = \log L(X) = -\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \langle S, \Psi_1 \oplus \cdots \oplus \Psi_K\rangle \tag{5.4}$$

which is exactly the Gaussian loglikelihood for precision matrices of the form $\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K$. (*Kalaitzis et al.*, 2013) use this form of the model to show that the *maximum entropy precision matrix estimate, given the S_k, is the maximum likelihood Kronecker sum estimate.* In other words, the Kronecker sum estimate makes the least assumptions about the data given the matricized sample covariances S_k. Exploring this connection further, we observe (Lemma V.8) that

$$\langle S, \Psi_1 \oplus \cdots \oplus \Psi_K\rangle = \left\langle (S_1 \oplus \cdots \oplus S_K) - \frac{(K-1)\mathrm{tr}(S)}{p} I_p, \Psi_1 \oplus \cdots \oplus \Psi_K \right\rangle.$$

If the Gram matrices have eigendecompositions $S_k = U_k \Lambda_k U_k^T$, then the eigenvectors

of $S_1 \oplus \cdots \oplus S_K$ are $U_1 \otimes \cdots \otimes U_K$. Since the objective (5.4) is unitarily invariant, we have that the eigenvectors of $\widehat{\Omega} = \widehat{\Psi}_1 \oplus \cdots \oplus \widehat{\Psi}_K$ are also $U_1 \otimes \cdots \otimes U_K$, and thus each factor estimate is of the form

$$\widehat{\Psi}_k = U_k \Gamma_k U_k^T,$$

where Γ_k is diagonal. In other words, the eigenstructure of the Gram matrices are directly incorporated into the associated Kronecker sum factors, making each factor estimate nearly independent.

5.1.1 Related Work

The Kronecker product model has been applied to a wide variety of spatio-temporal problems where structured covariance estimation is required with limited training samples. Previous applications of the model of Equation (5.2) include MIMO wireless channel modeling as a transmit vs. receive decomposition (*Werner and Jansson*, 2007), geostatistics (*Cressie*, 1993), genomics (*Yin and Li*, 2012), multi-task learning (*Bonilla et al.*, 2007), collaborative filtering (*Yu et al.*, 2009), face recognition (*Zhang and Schneider*, 2010), mine detection (*Zhang and Schneider*, 2010), networks (*Leskovec et al.*, 2010), and recommendation systems (*Allen and Tibshirani*, 2010). A variety of estimation algorithms have been proposed, including an SVD based method (*Werner et al.*, 2008) and sparsity-inducing methods (*Tsiligkaridis et al.*, 2013; *Zhou*, 2014).

The Kronecker sum model is related to the sum of Kronecker products model introduced in (*Tsiligkaridis and Hero*, 2013). These authors proposed approximating the covariance matrix using a sum of r Kronecker products

$$\Sigma = \sum\nolimits_{i=1}^{r} A_i \otimes B_i, \tag{5.5}$$

where r is the unknown separation rank and is assumed to be small, $A_i \in \mathbb{R}^{m \times m}$, and $B_i \in \mathbb{R}^{n \times n}$. All of the Kronecker factors A and B are symmetric but not necessarily positive semi-definite matrices. It follows from (*Loan and Pitsianis*, 1993) that any arbitrary covariance matrix can be represented in this form with sufficiently large r. This allows for more accurate approximation of the covariance when most of its energy can be accounted for by the sum of only a few Kronecker factors as in (5.5). An algorithm (Permuted Rank-penalized Least Squares (PRLS)) for fitting the model (5.5) to a measured sample covariance matrix was introduced in (*Tsiligkaridis and Hero*, 2013) and was shown to have strong high dimensional MSE performance guarantees. Additional extensions, such as the inclusion of a sparse correction for robustness (Robust KronPCA) (*Greenewald and Hero*, 2015), and the introduction of improved sketching estimators (*Chi*, 2016), have been developed.

A special case of the sum of Kronecker products model (5.5) is the restriction to $r = 2$ components of the form

$$\Sigma = A \oplus B = A \otimes I_n + I_m \otimes B, \tag{5.6}$$

where $A^{-1} \in \mathbb{R}^{m \times m}$ and $B^{-1} \in \mathbb{R}^{n \times n}$ are sparse (*Rudelson and Zhou*, 2015).

The proposed TeraLasso builds upon the sum of Kronecker products ideas from PRLS, Robust KronPCA, and the Kronecker sum covariance model (*Rudelson and Zhou*, 2015), with the critical difference that TeraLasso creates a sum of Kronecker products in the *inverse covariance matrix*, while the previous methods create a sum of Kronecker products in the *covariance matrix* itself. This difference raises new estimation challenges, but allows direct, identifiable modeling of sparsity in the precision matrix.

The Kronecker sum precision matrix model for $K = 2$ ($\Omega = A \oplus B$) was introduced in (*Kalaitzis et al.*, 2013) as the Bigraphical Lasso. In that work it was assumed that

the diagonals of each of the Kronecker factors were known. Significant reduction of the number of training samples required to learn the covariance was observed, shown experimentally for a variety of real data processes, and comparing favorably with the Kronecker product model (*Kalaitzis et al.*, 2013).

The contributions of this work are as follows. We extend the sparse matrix-variate Bigraphical lasso model to the sparse tensor-variate ($K > 2$) TeraLasso model, allowing modeling of data with arbitrary tensor degree K. We establish nonasymptotic estimator performance bounds for TeraLasso as well as the Bigraphical lasso, implying low sample statistical convergence in the high dimensional regime. We propose a highly scalable, first-order FISTA-based algorithm (TG-ISTA) to solve the TeraLasso objective, prove that it enjoys a geometric convergence rate to the global optimum, and demonstrate its practical advantages on real problems. Finally, we applied the algorithm to synthetic, meteorological, and EEG datasets, demonstrating that TeraLasso significantly improves performance in the low- and single- sample regimes. We argue that the intuitive graphical structure, robust eigenstructure, and the maximum-entropy interpretation of the TeraLasso model makes it superior to the Kronecker product model, and perhaps the ideal candidate for ultra-low sample estimation of graphical models for tensor valued data.

5.1.2 Outline

The remainder of the paper is organized as follows. The TeraLasso objective function is introduced in Section 5.3, and high dimensional consistency results are presented in Section 5.4. Our proposed first order TG-ISTA optimization algorithm in described in Section 5.5, with numerical geometric convergence to the globally optimal estimator proven in Section 5.6. Finally, Sections 5.7 and 5.8 contain simulated and real data results, with Section 5.9 concluding the paper.

5.2 Kronecker Sum Notation

5.2.1 Kronecker Sum Subspace $\widetilde{\mathcal{K}}_{\mathbf{p}}$

We define the set of square matrices expressible as a Kronecker sum $\widetilde{\mathcal{K}}_{\mathbf{p}}$ as

$$\widetilde{\mathcal{K}}_{\mathbf{p}} = \{A \in \mathbb{R}^{p \times p} : \exists\, B_k \in \mathbb{R}^{d_k \times d_k} \text{ s.t. } A = B_1 \oplus \cdots \oplus B_K\}.$$

Observe that $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is a linear sum of K components, and thus $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is linearly spanned by the K components. Thus $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is a linear subspace of $\mathbb{R}^{p \times p}$. Observe that by the definition of the Kronecker sum

$$\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K = \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \Psi_k \otimes I_{[d_{k+1:K}]}.$$

This implies that each entry of Ψ_k appears in $p/d_k = m_k$ entries of Ω. Thus, each parameter appears in a minimum of $\min_k m_k$ entries of Ω. From a geometric perspective, TeraLasso exploits this $\min_k m_k$ repeating structure to correspondingly reduce the variance of the parameter estimates.

It should be noted that $\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K$ does not uniquely determine $\{\Psi_k\}_{k=1}^{K}$, i.e. the Kronecker sum parameterization is not fully identifiable. Specifically, observe that for any c

$$A \oplus B = A \otimes I + I \otimes B = A \otimes I - cI + cI + I \otimes B = (A - cI) \oplus (B + cI),$$

and thus the trace of each factor is nonidentifiable. Observe that Ψ_k^-, Ω^+ are unaffected by the trace ambiguity, where we define the notation $M^- = M - \text{diag}(M)$, and $M^+ = \text{diag}(M)$. Thus, this trace ambiguity does not affect Ω or the off diagonals of

the factors Ψ_k. Hence, the decomposition

$$\Omega = \Omega^+ + (\Psi_1^- \oplus \cdots \oplus \Psi_K^-). \tag{5.7}$$

is identifiable. We will present estimator bounds with respect to this decomposition.

5.2.2 Projection onto $\widetilde{\mathcal{K}}_{\mathbf{p}}$

Since $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is a subspace of $\mathbb{R}^{p\times p}$, we can define a unique projection operator onto $\widetilde{\mathcal{K}}_{\mathbf{p}}$

$$\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = \arg\min_{M\in\widetilde{\mathcal{K}}_{\mathbf{p}}} \|A - M\|_F^2.$$

We first introduce a submatrix notation. Fix a k, and choose $i, j \in \{1, \dots m_k\}$. Let $E_1 \in \mathbb{R}^{\prod_{\ell=1}^{k-1} d_k \times \prod_{\ell=1}^{k-1} d_k}$ and $E_2 \in \mathbb{R}^{\prod_{\ell=k+1}^{K} d_k \times \prod_{\ell=k+1}^{K} d_k}$ be such that $[E_1 \otimes E_2]_{ij} = 1$ with all other elements zero. Observe that $E_1 \otimes E_2 \in \mathbb{R}^{m_k \times m_k}$. For any matrix $A \in \mathbb{R}^{p\times p}$, let $A(i, j|k) \in \mathbb{R}^{d_k \times d_k}$ be the submatrix of A defined via

$$[A(i, j|k)]_{rs} = \mathrm{tr}((E_1 \otimes \mathbf{e}_r\mathbf{e}_s \otimes E_2)A), \qquad r, s = 1, \dots, d_k. \tag{5.8}$$

The submatrix $A(i, j|k)$ is defined for all $i, j \in \{1, \dots m_k\}$ and $k = 1, \dots, K$. When A is a covariance matrix associated with a tensor X, this subblock corresponds to the covariance matrix between the ith and jth slices of X along the kth dimension.

We can now state a closed-form of the projection operator $\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A)$: the projection corresponds to setting each factor k to the coordinatewise $d_k \times d_k$ average of A with a correction that removes trace redundancy:

$$\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = (A_1 \oplus \cdots \oplus A_K) - (K-1)\frac{\mathrm{tr}(A)}{p} I_p,$$

where $A_k = \sum_{i=1}^{m_k} A(i, i|k)$ is the average of A over dimensions $1, \dots, k-1, k+1, \dots, K$ (see Appendix A for details).

5.2.3 Additional Notation

We define the set $\mathcal{K}_{\mathbf{p}}$ of positive semidefinite Kronecker sum matrices $A \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ as

$$\mathcal{K}_{\mathbf{p}} = \{A \succeq 0 | A \in \widetilde{\mathcal{K}}_{\mathbf{p}}\}.$$

Since $\mathcal{K}_{\mathbf{p}}$ is the intersection of the linear subspace $\widetilde{\mathcal{K}}_{\mathbf{p}}$ and the positive semidefinite cone, $\mathcal{K}_{\mathbf{p}}$ is a closed convex set.

In this paper we use the inner product associated with the Frobenius norm, $\langle A, B\rangle = \mathrm{tr}(A^T B) = \mathrm{vec}(A)^T \mathrm{vec}(B)$.

For n independent identically distributed samples $\mathbf{x}_i = \mathrm{vec}(X_i^T)$, $i = 1, \ldots, n$, let $S = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ be its sample covariance. We extract the factor-wise covariances $\Sigma^{(k)} = E[S_k]$ and sample covariances S_k (Gram matrices) of the k-mode matricizations $X_{i,(k)}$ of X_i by taking the average over the other $K-1$ dimensions:

$$S_k = \frac{1}{n m_k}\sum_{i=1}^{n} X_{i,(k)} X_{i,(k)}^T = \sum_{j=1}^{m_k} S(j,j|k), \quad \Sigma^{(k)} = \frac{1}{m_k} E[X_{(k)} X_{(k)}^T] = \sum_{j=1}^{m_k} \Sigma(j,j|k).$$

5.3 Tensor graphical Lasso (TeraLasso)

5.3.1 Subgaussian Model

We first present our generative model for data generated via the TeraLasso model.

For a random variable Y, the sub-gaussian (or ψ_2) norm of Y, denoted by $\|Y\|_{\psi_2}$, is defined as

$$\|Y\|_{\psi_2} = \sup_{q \geq 1} q^{-1/2} (E|Y|^q)^{1/q};$$

$$\text{if } E[Y] = 0, \text{ then } E[\exp(tY)] \leq \exp(Ct^2 \|Y\|_{\psi_2}^2) \quad \forall t \in \mathbb{R}.$$

We define $X \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ to be a K-order subgaussian random tensor when $\mathbf{x} =$

$\text{vec}(X^T)$ is a subgaussian random vector. We say that $\mathbf{x}$ is a subgaussian random vector when

$$\mathbf{x} = \Sigma^{1/2}\mathbf{v},$$

for some covariance matrix $\Sigma \succ 0$ and random vector $\mathbf{v} = [v_1, \ldots, v_p]^T$ with independent, zero mean entries v_j whose variance $E[v_j^2] = 1$ and whose subgaussian norm is bounded $\|v_j\|_{\psi_2} \leq \kappa < \infty$.

In what follows, we assume the data samples X_i are independent identically distributed subgaussian random tensors following the above model. We will present an estimation procedure for $\Sigma^{-1} = \Omega \in \mathcal{K}_{\mathbf{p}}$, where we recall (5.1)

$$\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K.$$

5.3.2 Objective Function

The classic graphical lasso (*Banerjee et al.*, 2008; *Yuan and Lin*, 2007; *Zhou*, 2014; *Zhou et al.*, 2011; *Rolfs et al.*, 2012) estimates Ω by minimizing the objective function (L1 penalized Gaussian loglikelihood)

$$Q(\Omega) = -\log|\Omega| + \langle S, \Omega\rangle + \sum_{ij} \rho_{ij}|\Omega_{ij}|,$$

where S is the sample covariance. Our proposed Tensor graphical Lasso (TeraLasso) estimate of the precision matrix Ω in (5.1) will minimize the GLasso objective function, restricted to Kronecker sum precision matrices $\Omega \in \mathcal{K}_{\mathbf{p}}$:

$$\widehat{\Psi}_1 \oplus \cdots \oplus \widehat{\Psi}_K = \widehat{\Omega} = \arg\min_{\Omega \in \mathcal{K}_{\mathbf{p}}} Q(\Omega),$$

or equivalently substituting in the Kronecker sum parameterization,

$$\arg\min_{\Omega\in\mathcal{K}_{\mathbf{p}}} Q(\Omega) = \arg\min_{\Omega\in\mathcal{K}_{\mathbf{p}}} Q(\Psi_1,\ldots,\Psi_K) = \tag{5.9}$$
$$-\log|\Psi_1\oplus\cdots\oplus\Psi_K| + \langle S, \Psi_1\oplus\cdots\oplus\Psi_K\rangle + \sum_{k=1}^{K}\rho_k m_k|\Psi_k^-|_1.$$

The nondiagonal elements of the factors in a Kronecker sum have disjoint support (Appendix A), hence $\rho_k m_k|\Psi_k^-|_1$ has direct correspondence to $\sum_{ij}\rho_{ij}|\Omega_{ij}|$ for an appropriate choice of ρ_{ij}. This form of the objective function for $K=2$ is related to the bigraphical lasso objective (*Kalaitzis et al.*, 2013), except we do not penalize the nonidentifiable diagonals in the Ψ_k's. Lemma V.9 in Appendix A reveals that the high dimensional sample covariance S only enters into the objective function through its lower dimensional projections S_k onto the Kronecker sum factors:

$$\langle S, \Psi_1\oplus\cdots\oplus\Psi_K\rangle = \sum_{k=1}^{K} m_k\langle S_k, \Psi_k\rangle. \tag{5.10}$$

The simplified objective is

$$Q(\Psi_1,\ldots,\Psi_K) = -\log|\Psi_1\oplus\cdots\oplus\Psi_K| + \sum_{k=1}^{K} m_k\left(\langle S_k,\Psi_k\rangle + \rho_k|\Psi_k^-|_1\right). \tag{5.11}$$

We emphasize that (5.11) and (5.9) are mathematically equivalent, and represent the same objective function. We will usually cite the form (5.11) for simplicity. The objective function Q is jointly convex, and its minimization over $\mathcal{K}_{\mathbf{p}}$ has a unique solution.

Theorem V.1 (Joint Convexity and Uniqueness of the Estimate)**.** *The objective function* (5.11) *is jointly convex in* $\{\Psi_k\}_{k=1}^K$. *Furthermore, define the set* $\mathcal{A} = \{\{\Psi_k\}_{k=1}^K \text{ s.t. } Q(\{\Psi_k\}_{k=1}^K) = Q^*\}$ *where* $Q^* = \min_{\{\Psi_k\}_{k=1}^K} Q(\{\Psi_k\}_{k=1}^K)$. *If this set is nonempty, it maps to a unique inverse covariance* Ω_*, *i.e. there exists a unique*

$\Omega_* \in \mathcal{K}_{\mathbf{p}}$ *that achieves the minimum of* Q *such that*

$$\Psi_1 \oplus \cdots \oplus \Psi_K = \Omega_* \quad \forall \{\Psi_k\}_{k=1}^K \in \mathcal{A} \tag{5.12}$$

Proof. See Appendix 5.12. □

5.4 High Dimensional Consistency

In this section, we derive high dimensional consistency results for the TeraLasso estimator. In addition to the subgaussian generative model detailed above, we make the following assumptions on the true model:

A1 : For all $k = 1, \ldots, K$, let the sets $\mathcal{S}_k = \{(i,j) : i \neq j, [\Psi_k]_{ij} \neq 0\}$. Then $\text{card}(\mathcal{S}_k) \leq s_k \, \forall k$. Observe that the total number of nonzero edges in Ω is thus $s = \sum_{k=1}^K m_k s_k$.

A2 : $\phi_{\min}(\Omega_0) = \sum_{k=1}^K \phi_{\min}(\Psi_k) = \frac{1}{\|\Sigma_0\|} \geq \underline{k}_\Omega$, and $\phi_{max}(\Omega_0) = \sum_{k=1}^K \phi_{max}(\Psi_k) \leq \overline{k}_\Omega$.

Note that these assumptions only involve identifiable parameters of the model. Under these assumptions, we have the following bound on the Frobenius norm error of the estimator, both on the full Ω and on the identifiable parameters.

The unique factor expansions of the off diagonals of Ω_0 and its TeraLasso estimate $\widehat{\Omega}$ are of the form

$$\widehat{\Omega}^- = \Psi_1^- \oplus \cdots \oplus \Psi_K^-$$
$$\Omega_0^- = \Psi_{0,1}^- \oplus \cdots \oplus \Psi_{0,K}^-.$$

Using this notation, we have the following theorems.

Theorem V.2 (TeraLasso estimator: Frobenius error bound)**.** *Suppose the assumptions A1-A2 hold, and that* $\widehat{\Omega}$ *is the minimizer of* (5.11) *with* $\rho_k = \frac{C}{\underline{k}_\Omega}\sqrt{\frac{\log p}{n m_k}}$*. Then*

with probability at least $1 - 2(K+1)\exp(-c\log p)$

$$\|\widehat{\Omega} - \Omega_0\|_F^2 \leq C_1 K^2 (s+p) \frac{\log p}{n \min_k m_k}.$$

Theorem V.3 (TeraLasso estimator: Factorwise and L2 error bounds)**.** *Suppose assumptions A1-A2 hold. If* $\rho_k = \frac{C}{\underline{k}_\Omega}\sqrt{\frac{\log p}{n m_k}}$*, then with probability at least* $1 - 2(K+1)\exp(-c\log p)$ *we have*

$$\frac{\|\widehat{\Omega}^+ - \Omega_0^+\|_2^2}{(K+1)\max_k d_k} + \sum_{k=1}^{K} \frac{\|\Psi_k^- - \Psi_{0,k}^-\|_F^2}{d_k} \leq cK^2 \left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right) \frac{\log p}{n \min_k m_k} \tag{5.13}$$

and

$$\|\widehat{\Omega} - \Omega_0\|_2 \leq cK\sqrt{K+1}\sqrt{\left(\max_k \frac{d_k}{m_k}\right)\left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right)\frac{\log p}{n}}.$$

Theorems V.2 and V.3 are proved in Appendix 5.13 and 5.17 respectively.

Observe that by (5.13) the parameters of Ω can be estimated in the single sample regime when the dimension is large ($m_k > d_k$). Due to the repeating structure and increasing dimension of Ω, the parameter estimates can converge without the overall Frobenius error $\|\widehat{\Omega} - \Omega_0\|_F$ converging. To see this, suppose that $\Delta_\Omega = \Delta \oplus \mathbf{0} \oplus \cdots \oplus \mathbf{0}$ where $\|\Delta\|_F = \epsilon$. Then

$$\|\Delta_\Omega\|_F = \sqrt{m_1}\epsilon.$$

The following corollary immediately follows from Theorems V.2 and 3.

Corollary V.4. *Suppose the assumptions of Theorem V.2 hold, that* K *is fixed, and*

that all the m_k, $k = 1, \ldots, K$ grow without bound.

$$\|\widehat{\Omega} - \Omega\|_F^2 = O_p\left((s+p)\frac{\log p}{n \min_k m_k}\right),$$
$$\frac{\|\widehat{\Omega}^+ - \Omega_0^+\|_2^2}{\max_k d_k} + \sum_{k=1}^{K} \frac{\|\Psi_k^- - \Psi_{0,k}^-\|_F^2}{d_k} = O_p\left(\left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right)\frac{\log p}{n \min_k m_k}\right),$$
$$\|\widehat{\Omega} - \Omega\|_2 = O_p\left(\sqrt{\left(\max_k \frac{d_k}{m_k}\right)\left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right)\frac{\log p}{n}}\right).$$

Hence the TeraLasso estimate converges to the true precision matrix as n increases. Observe further that in the single sample regime ($n = 1$), the TeraLasso estimates of the identifiable parameters and the overall spectral norm error converge as the dimensions d_k increase, whenever s_k/d_k remains bounded and $(d_k \log p)/m_k$ goes to zero. As an example, this latter condition is guaranteed to hold whenever $K \geq 3$ and $\max_k d_k / \min_k d_k$ remains bounded.

For comparison to the GLasso, recall that the subgaussian GLasso rate is (*Zhou et al.*, 2011; *Rothman et al.*, 2008)

$$\|\widehat{\Omega} - \Omega\|_F^2 = O_p\left(\frac{(p+s)\log p}{n}\right),$$

and that the Frobenius norm bound in Corollary V.4 improves on this result by a factor of $\min_k m_k$, clearly exploiting the redundancy across dimensions.

Furthermore, if the dimensions are equal ($d_k = p^{1/K}$ and $s_k = \widetilde{s}$ for all k) and K is fixed, Corollary V.4 implies

$$\|\Delta_k\|_F^2 = O_p\left(\frac{(d_k + s_k)\log p}{m_k n}\right),$$

indicating that TeraLasso with n replicates estimates the identifiable representation of Ψ_k with an error rate equivalent to that of GLasso with $\Omega = \Psi_k$ and nm_k available replicates.

5.5 Algorithm

When the BiGLasso model was proposed (*Kalaitzis et al.*, 2013), a simple alternating-minimization estimation algorithm was given. It required, however, the diagonal elements of the inverse covariance to be known a priori, which is unrealistic in practice. Hence, we derive a joint first-order primal algorithm for BiGLasso and TeraLasso estimation that is not subject to this limitation.

As the TeraLasso objective (5.11) is non-differentiable because of the L1 penalties, we use a iterative soft thresholding (ISTA) method restricted to the convex set $\mathcal{K}_{\mathbf{p}}$ of possible positive semidefinite Kronecker sum precision matrices. We call our approach Tensor Graphical Iterative Soft Thresholding, or TG-ISTA.

5.5.1 Composite gradient descent and proximal first order methods

Our goal is to solve the objective (5.11)

$$Q(\Psi_1, \ldots, \Psi_K) = -\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \sum_{k=1}^{K} m_k \left(\langle S_k, \Psi_k\rangle + \rho_k |\Psi_k^-|_1\right).$$

Note that this objective can be decomposed into the sum of a differentiable function f and a lower semi-continuous but nonsmooth function g:

$$\begin{aligned} Q(\Psi_1, \ldots, \Psi_K) &= f(\Omega) + g(\Omega) \\ f(\Omega) &= -\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \sum_{k=1}^{K} m_k \langle S_k, \Psi_k\rangle \\ &= -\log|\Omega| + \langle S, \Omega\rangle|_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \\ g(\Omega) &= \sum_{k=1}^{K} m_k \rho_k |\Psi_k^-|_1. \end{aligned} \tag{5.14}$$

For objectives of this form (*Nesterov et al.*, 2007) proposed a first order method called composite gradient descent, which has been specialized to the case of $g = |\cdot|_1$ and

is widely known as Iterative Soft Thresholding (ISTA) (*Tseng*, 2010; *Combettes and Wajs*, 2005; *Beck and Teboulle*, 2009; *Nesterov*, 1983, 2004). This method is also applicable to nonconvex regularizers g (*Loh and Wainwright*, 2013).

In the unconstrained setting composite gradient descent is an iterative solution given by the updates

$$\Omega_{t+1} \in \arg\min_{\Omega \in \mathbb{R}^{p \times p}} \left\{ \frac{1}{2} \left\| \Omega - (\Omega_t - \zeta \nabla f(\Omega_t)) \right\|_F^2 + \zeta g(\Omega) \right\},$$

where ζ is a scalar stepsize parameter. This optimization can be solved in closed-form for many common regularizers g, including the one in (5.14).

Paralleling previous applications of ISTA to the positive semidefinite cone (*Rolfs et al.*, 2012), we will derive unrestricted composite gradient descent in the relevant space ($\widetilde{\mathcal{K}}_{\mathbf{p}}$ for TeraLasso) and enforce the positive semidefinite constraint at each step by performing line search to find a suitable stepsize ζ. Convergence to the optimal solution in $\mathcal{K}_{\mathbf{p}}$ is guaranteed (see Section 5.6.2) since the positive definite cone is an open subset of $\mathbb{R}^{p \times p}$ and the objective Q goes to infinity on the boundary of the positive semidefinite cone.

Thus, composite gradient descent within the linear subspace $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is given by

$$\Omega_{t+1} \in \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\nabla f(\Omega_t)) \right) \right\|_F^2 + \zeta g(\Omega) \right\}, \tag{5.15}$$

whenever the initialization $\Omega_0 \in \widetilde{\mathcal{K}}_{\mathbf{p}}$.

5.5.2 TG-ISTA

To apply this form of composite gradient descent to the TeraLasso objective, we first derive the gradient of

$$f(\Omega) = -\log|\Omega| + \langle S, \Omega \rangle \tag{5.16}$$

restricted to the space $\widetilde{\mathcal{K}}_{\mathbf{p}}$.

Since the gradient of $\langle S, \Omega \rangle$ with respect to Ω is S, the gradient of $\langle S, \Omega \rangle$ for Ω in the subspace $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is the projection of S onto the subspace $\widetilde{\mathcal{K}}_{\mathbf{p}}$. This is given by (Lemma V.8)

$$\begin{aligned}
\nabla_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}}(\langle S, \Psi_1 \oplus \cdots \oplus \Psi_k \rangle) &= \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S) \qquad (5.17)\\
&= \left(S_1 - \frac{K-1}{K} \frac{\mathrm{tr}(S_1)}{d_1} I_{d_1} \right) \oplus \cdots \oplus \left(S_K - \frac{K-1}{K} \frac{\mathrm{tr}(S_K)}{d_K} I_{d_K} \right) \\
&= \widetilde{S}_1 \oplus \cdots \oplus \widetilde{S}_K = \widetilde{S} \\
\widetilde{S}_k &= S_k - \frac{K-1}{K} \frac{\mathrm{tr}(S_k)}{d_k} I_{d_k}.
\end{aligned}$$

Recall that in $\mathbb{R}^{p \times p}$ the gradient of $-\log|\Omega|$ is Ω^{-1} (*Boyd and Vandenberghe*, 2009). As the inverse of a Kronecker sum is in general not a Kronecker sum, we project Ω^{-1} onto $\widetilde{\mathcal{K}}_{\mathbf{p}}$ to find the gradient in $\widetilde{\mathcal{K}}_{\mathbf{p}}$. Thus the gradient of the log determinant portion is

$$\nabla_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left(-\log|\Psi_1 \oplus \cdots \oplus \Psi_K| \right) = G_1 \oplus \cdots \oplus G_K = G = \mathrm{proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}} \left((\Psi_1 \oplus \cdots \oplus \Psi_K)^{-1} \right). \qquad (5.18)$$

In Algorithm 5 we show an efficient method of computing this projected inverse by exploiting the eigendecomposition identity in Appendix A Property 3 and Lemma V.21.

Substituting (5.18) and (5.17) into the composite gradient framework (5.15) gives

$$\Omega_{t+1} \in \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta \left(\widetilde{S} - G \right) \right) \right\|_F^2 + \zeta \sum_{k=1}^{K} m_k \rho_k |\Psi_k^-|_1 \right\}. \qquad (5.19)$$

In Appendix 5.18 we show the following:

Lemma V.5 (Decomposition of objective)**.** *For* $\Omega_t, \Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}$*, i.e.*

$$\Omega_t = \Psi_1^t \oplus \cdots \oplus \Psi_K^t, \qquad \Omega = \Psi_1 \oplus \cdots \oplus \Psi_K,$$

the objective (5.19) *can be decomposed into* $K+1$ *fully identifiable, independent optimization problems. Specifically,*

$$[\Psi_k^{t+1}]^- = \arg\min_{\Psi_k^-} \frac{1}{2} \left\| \Psi_k^- - ([\Psi_k^t]^- - \zeta(\widetilde{S}_k^- - G_k^-)) \right\|_F^2 + \zeta\rho_k |\Psi_k^-|_1 \tag{5.20}$$

$$\Omega_{t+1}^+ = \arg\min_{\Omega^+} \frac{1}{2} \left\| \Omega^+ - \left(\Omega_t^+ - \zeta \left(\widetilde{S}^+ - G^+ \right) \right) \right\|_F^2. \tag{5.21}$$

Since the diagonal Ω^+ is unregularized in (5.21), we have

$$\Omega_{t+1}^+ = \Omega_t^+ - \zeta(\widetilde{S}^+ - G^+),$$

i.e.

$$[\Psi_k^{t+1}]^+ = [\Psi_k^t]^+ - \zeta(\widetilde{S}_k^+ - G_k^+). \tag{5.22}$$

For the off diagonal factors, the problem (5.20) is the L1 proximal operator. The solution is given by (*Beck and Teboulle*, 2009)

$$[\Psi_k^t]^- = \text{shrink}_{\zeta\rho_k}^-((\Psi_k^t - \zeta(\widetilde{S}_k - G_k))^-), \tag{5.23}$$

where we define the shrinkage operator $\text{shrink}_\rho^-(\cdot)$ as

$$[\text{shrink}_\rho^-(M)]_{ij} = \begin{cases} \text{sign}(M_{ij})(|M_{ij}| - \rho)_+ & i \neq j \\ M_{ij} & \text{o.w.} \end{cases} \tag{5.24}$$

Combining (5.23) and (5.22) gives a unified expression

$$\Psi_k^{t+1} = \text{shrink}^-_{\zeta_t \rho_k}(\Psi_k^t - \zeta(\widetilde{S}_k - G_k)). \tag{5.25}$$

Our approach to composite gradient descent is summarized in Algorithm 4.

5.5.3 Choice of step size ζ_t

It remains to set the stepsize parameter ζ_t at each step t. We follow the approach of (*Beck and Teboulle*, 2009; *Rolfs et al.*, 2012). We will prove linear convergence when ζ_t is chosen such that

$$f(\Omega_{t+1}) = -\log|\Omega_{t+1}| + \langle S, \Omega\rangle \leq \mathcal{Q}_{\zeta_t}(\Omega_{t+1}, \Omega_t) \tag{5.26}$$

where $\mathcal{Q}_\zeta$ is a quadratic approximation to f given by

$$\begin{aligned}\mathcal{Q}_\zeta(\Omega_{t+1}, \Omega_t) &= f(\Omega_t) + \langle \Omega_{t+1} - \Omega_t, \nabla f(\Omega_t)\rangle + \frac{1}{2\zeta}\|\Omega_{t+1} - \Omega_t\|_F^2 \\ &= -\log|\Omega_t| + \langle \widetilde{S}, \Omega_t\rangle + \langle \Omega_{t+1} - \Omega_t, \widetilde{S} - G\rangle + \frac{1}{2\zeta}\|\Omega_{t+1} - \Omega_t\|_F^2.\end{aligned} \tag{5.27}$$

At each iteration t, we thus select a stepsize ζ_t, and compute the update (5.25). If the resulting Ω_{t+1} is not positive definite or does not decrease the objective sufficiently according to (5.26), we decrease the stepsize to $c\zeta_t$ for $c \in (0, 1)$ and try again. This is guaranteed to find an appropriate step since by construction Ω_t is positive definite, and the positive definite cone is an open set. We continue retesting and decreasing the stepsize by c until the constraints on Ω_t are satisfied. If after a set number of backtracking steps the conditions are still not satisfied, we can always take the safe step

$$\zeta_t = \lambda_{\min}^2(\Omega_t) = \sum_{k=1}^{K} \min_i [\mathbf{s}_k]_i^2.$$

As the safe stepsize is often slow, we use the more aggressive Barzilei-Borwein step to set a starting ζ_t at each time. The Barzilei-Borwein stepsize (*Barzilai and Borwein*, 1988) creates an approximation to the Hessian, in our case given by (see identities in Appendix A)

$$\zeta_{t+1,0} = \frac{\|\Omega_{t+1} - \Omega_t\|_F^2}{\langle \Omega_{t+1} - \Omega_t, G^t - G^{t+1} \rangle} \tag{5.28}$$

The norms and inner products in (5.28) and (5.27) can be efficiently computed factor-wise using the formulas in Appendix A. The complete algorithm is shown in Algorithm 5.

Algorithm 4 TG-ISTA (high level)

1: Input: SCM factors S_k, regularization parameters ρ_i, backtracking constant $c \in (0, 1)$, initial step size $\zeta_{1,0}$, initial iterate Ω_0.
2: **while** not converged **do**
3: Compute the subspace gradient $G_1^t \oplus \cdots \oplus G_K^t$ at Ω_t.
4: Set stepsize ζ_t.
5: **for** $k = 1, \ldots, K$ **do**
6: Perform the composite objective gradient update: $\Psi_k^{t+1} \leftarrow \text{shrink}^-_{\zeta_t \rho_k}(\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t))$.
7: **end for**
8: **end while**
9: Return $\{\Psi_k^{t+1}\}_{k=1}^K$.

5.6 Numerical Convergence

5.6.1 Cost

Due to the representation (5.10), the TG-ISTA algorithm never needs to form the full $p \times p$ covariance. The memory footprint of TG-ISTA is only $O(p + \sum_{k=1}^K d_k^2)$ as opposed to the $O(p^2)$ storage required by unstructured estimators such as the GLasso. Since the training data itself requires $O(np)$ storage, the storage footprint of TG-ISTA is scalable to large values of $p = \prod_{k=1}^K d_k$ when the d_k/p decrease (e.g.

Algorithm 5 TG-ISTA

Input: SCM factors S_k, regularization parameters ρ_i, backtracking constant $c \in (0,1)$, initial step size $\zeta_{1,0}$, initial iterate $\Omega_0 = \Psi_1^0 \oplus \cdots \oplus \Psi_K^0$.
for k = 1,..., K **do**
  $\mathbf{s}_k$, $U_k \leftarrow$ Eigendecomposition of $\Psi_k^0 = U_k \mathrm{diag}(\mathbf{s}_k) U_k^T$.
  $\widetilde{S}_k \leftarrow S_k - I_{d_k} \frac{\mathrm{tr}(S_k)}{d_k} \frac{K-1}{K}$.
end for
while not converged **do**
  $\{\widetilde{\mathbf{s}}\}_{k=1}^K \leftarrow \mathrm{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}} \left(\mathrm{diag} \left(\frac{1}{\mathbf{s}_1 \oplus \cdots \oplus \mathbf{s}_K} \right) \right)$.
  for $k = 1 \ldots K$ **do**
    $G_k^t \leftarrow U_k \mathrm{diag}(\mathbf{s}_k) U_k^T$.
  end for
  for $j = 0, 1, \ldots$ **do**
    $\zeta_t \leftarrow c^j \zeta_{t,0}$.
    for $k = 1, \ldots, K$ **do**
      $\Psi_k^{t+1} \leftarrow \mathrm{shrink}^-_{\zeta_t \rho_k}(\Psi_k^t - \zeta_t(\widetilde{S}_k - G_k^t))$.
      Compute eigendecomposition $U_k \mathrm{diag}(\mathbf{s}_k) U_k^T = \Psi_k^{t+1}$.
    end for
    Compute $\mathcal{Q}_{\zeta_t}(\{\Psi_k^{t+1}\}, \{\Psi_k^t\})$ via (5.27).
    if $f(\{\Psi_k^{t+1}\}) \leq \mathcal{Q}\zeta_t(\{\Psi_k^{t+1}\}, \{\Psi_k^t\})$ as in (5.27) **and** $\min_i([\mathbf{s}_1 \oplus \cdots \oplus \mathbf{s}_K]_i) > 0$ **then**
      Stepsize ζ_t is acceptable; **break**
    end if
  end for
  Compute Barzilei-Borwein stepsize $\zeta_{t+1,0}$ via (5.28)
end while
Return $\{\Psi_k^{t+1}\}_{k=1}^K$.

$d_k = p^{1/K}$).

The computational cost per iteration is dominated by the computation of the gradient, which is performed by doing K eigendecompositions of size $d_1, \ldots, d_K$ respectively and then computing the projection of the inverse of the Kronecker sum of the resulting eigenvalues. The former step costs $O(\sum_{k=1}^K d_k^3)$, and the second costs $O(pK)$. Thus the computational cost per iteration will be

$$O\left(pK + \sum_{k=1}^K d_k^3 \right). \tag{5.29}$$

For $K > 1$ and $d_k/p \ll 1$, this gives a dramatic improvement on the $O(p^3) =$

$O(\prod_{k=1}^{K} d_k^3)$ cost per iteration of unstructured Graphical Lasso algorithms (*Rolfs et al.*, 2012; *Hsieh et al.*, 2014). In addition, for $K \leq 3$ the cost per iteration is comparable to the $O(d_1^3 + d_2^3 + d_3^3)$ cost per iteration of the most efficient ($K = 3$) Kronecker product GLasso methods (*Zhou*, 2014). Given that (*Zhou*, 2014) solves separate factor-wise objective functions, the fact that our joint optimization approach achieves comparable per-iteration computational complexity to the separable Kronecker product model is indicative of the power of the nonseparable Kronecker sum model in high dimension.

5.6.2 TG-ISTA Algorithm Convergence Rate

In Appendix 5.19, we prove the following results demonstrating geometric convergence of the iterates of TG-ISTA.

Theorem V.6. *Assume that the iterates Ω_t of Algorithm 5 satisfy $aI \preceq \Omega_t \preceq bI$, for all t, for some fixed constants $0 < a < b < \infty$. Suppose further that Ω^* is the global optimum. If $\zeta_t \leq a^2$ for all t, then*

$$\|\Omega_{t+1} - \Omega^*\|_F \leq \max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\} \|\Omega_t - \Omega^*\|_F.$$

Furthermore, the step size ζ_t which yields an optimal worst-case contraction bound $s(\zeta_t)$ is $\zeta = \frac{2}{a^{-2}+b^{-2}}$. The corresponding optimal worst-case contraction bound is

$$s(\zeta) = 1 - \frac{2}{1 + \frac{b^2}{a^2}}.$$

Theorem V.7. *Let $\rho_k > 0$ for all k and let Ω_{init} be the initialization of TG-ISTA (Algorithm 5). Let*

$$\alpha = \frac{1}{\sum_{k=1}^{K} \|S_k\|_2 + d_k \rho_k}, \qquad b' = \|\Omega^*\|_2 + \|\Omega_{init} - \Omega^*\|_F,$$

and assume $\zeta_t \leq \alpha^2$ for all t. Then the iterates Ω_t of Algorithm 5 satisfy $\alpha I \preceq \Omega_t \preceq b' I$

for all t, with $b' = \|\Omega^*\|_2 + \|\Omega_{init} - \Omega^*\|_F$.

These two theorems together imply geometric convergence of the precision matrix estimate to the global optimum Ω^* when the stepsize is not excessive.

Note that the contraction rate bound only depends on the dimension through a weighted sum of the dimensions d_k, confirming the scalability of TG-ISTA in the number of iterations as well as in the cost per iteration.

5.7 Synthetic Data

In this section, we verify the performance of TeraLasso and TG-ISTA on synthetic data. See Algorithm 6 in Appendix A for a scalable method of generating the random vector $\mathbf{x} = \text{vec}(X^T)$ under the Kronecker sum model. We created random graphs for each factor Ψ_k using both an Erdos-Renyi (ER) topology and a random grid graph topology. We generated the ER type graphs according to the method of (*Zhou et al.*, 2010b). Initially we set $\Psi = 0.25 I_{n \times n}$, where $n = 100$. Then, we randomly select p edges and update Ψ as follows: for each new edge (i, j), a weight $a > 0$ is chosen uniformly at random from $[0.2, 0.4]$; we subtract a from Ψ_{ij} and Ψ_{ji}, and increase Ψ_{ii}, Ψ_{jj} by a. This keeps Ψ positive definite. We repeat this process until all edges are added. An example 25-node ER graph and precision matrix are shown in Figure 5.3, along with a 225-node precision matrix formed from the Kronecker sum 25 and 9-node ER precision matrices.

The random grid graph is produced in a similar way, with the exception that edges are only allowed between adjacent nodes, where the nodes are arranged on a square grid (Figure 5.4).

5.7.1 Algorithmic Convergence

We first compared our proposed TG-ISTA algorithm to the original BiGLasso algorithm proposed in (*Kalaitzis et al.*, 2013). It should be noted that the BiGLasso

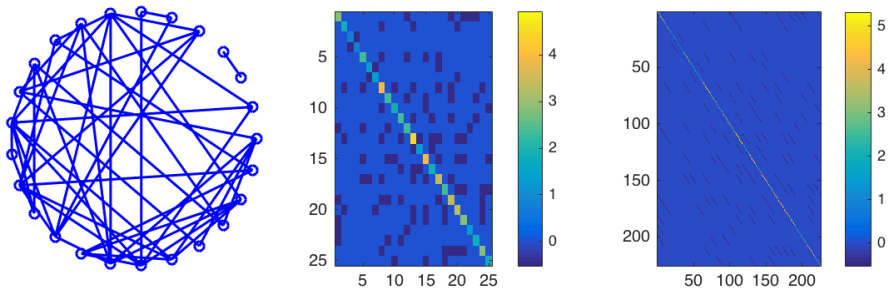


Figure 5.3: Example random Erdos-Renyi (ER) graph with 25 nodes and 50 edges. Left: Graphical representation. Center: Corresponding precision matrix Ψ. Right: Full $K = 2$, 225-node Kronecker sum of Ψ with an ER graph of size 9.

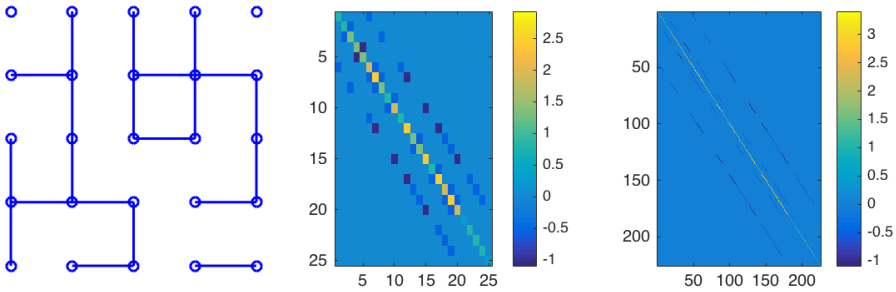
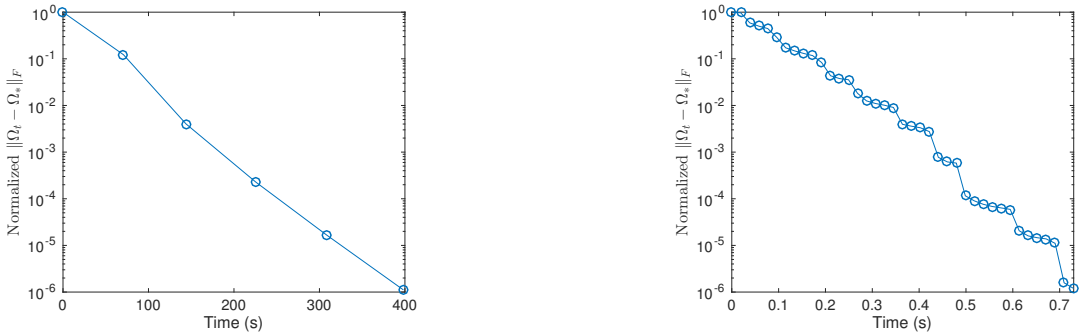


Figure 5.4: Example random grid graph (square) with 25 nodes and 26 edges. Left: Graphical representation. Center: Corresponding precision matrix Ψ. Right: Full $K = 2$, 225-node Kronecker sum of Ψ with a grid graph of size 9.

algorithm does not estimate the diagonal elements of Ω (it assumes they are known), so it cannot strictly be considered to solve the general BiGLasso or TeraLasso objectives. Figure 5.5 shows comparative convergence speeds on a $K = 2$ random ER graph estimation scenario. Observe that TG-ISTA's ability to efficiently exploit the Kronecker sum structure to obtain computational and memory savings allows it to quickly converge to the optimal solution, while the alternating-minimization based BiGLasso algorithm with its heavy per-iteration computation and memory costs is impractically slow.

To confirm the linear convergence theory for the TG-ISTA algorithm, we next

generated Kronecker sum inverse covariance graphs and plotted the Frobenius norm between the inverse covariance iterates Ω_t and the optimal point Ω^*. For simplicity, we used equal d_k and set the Ψ_k to be random ER graphs (see above) with d_k edges. We set the $\rho_k = \rho$ using cross validation. Figure 5.6 shows the results as a function of iteration, for a variety of d_k and K configurations. For comparison, the statistical error of the optimal point is also shown, as optimizing beyond this level provides reduced benefit. As predicted, linear or better convergence to the global optimum is observed, and as expected it does not appear that increasing d_k, p, or K dramatically affect the number of iterations required. Overall, the small number of iterations required combined with the low cost per iteration confirm the efficiency of TG-ISTA.



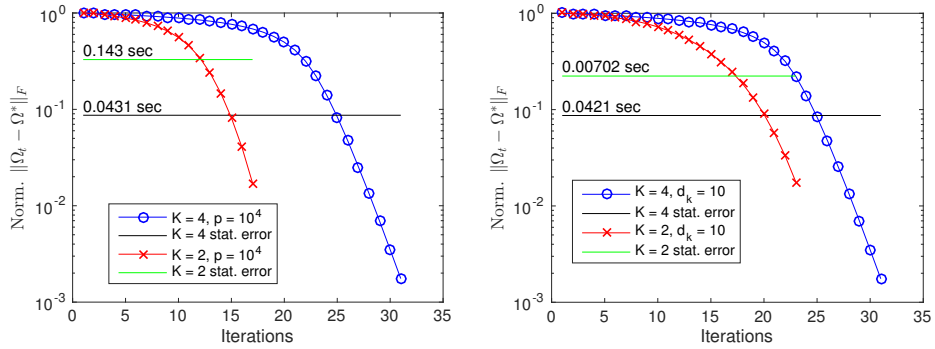
(a) Original BiGLasso

(b) Proposed TG-ISTA

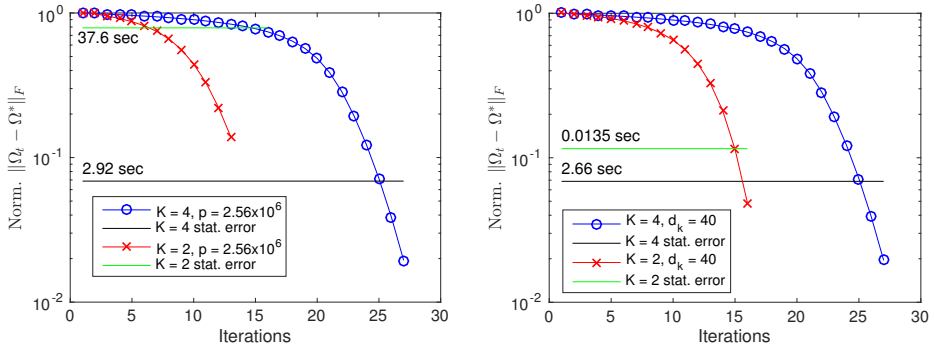
Figure 5.5: BiGLasso algorithm (*Kalaitzis et al.*, 2013) and our TG-ISTA approach, estimating a $K = 2$ Kronecker sum of random ER graphs with $\mathbf{p} = [75, 75]$, i.e. $p = 5625$, and $n = 10$. Normalized Frobenius norm between iterates Ω_t and converged Ω_* are shown for each algorithm. Note the greatly superior speed of TG-ISTA. Recall further that TG-ISTA finds the complete optimum (i.e. $\Omega_* = \Omega_{opt}$) and uses $O(p + d_k^2)$ storage, while BiGLasso does not estimate the diagonal elements of Ω, is limited to $K = 2$, and requires $O(p^2)$ storage.

5.7.2 Tuning Parameters

In the formulation of the TeraLasso objective (5.11) and the TG-ISTA algorithm, the sparsity of the estimate is controlled by K tuning parameters ρ_k for $k = 1, \ldots, K$. However, Theorem V.2 seems to indicate that the ρ_k can be set formulaically, using



(a) $n = 100$ samples



(b) $n = 1$ sample

Figure 5.6: Linear convergence of TG-ISTA. Shown is the normalized Frobenius norm $\|\Omega_t - \Omega^*\|_F$ of the difference between the estimate at the tth iteration and the optimal Ω^*. On the left are results comparing $K = 2$ and $K = 4$ on the same data with the same value of p (different d_k), on the right they are compared for the same value of d_k (different p). Also included are the statistical error levels, and the computation times required to reach them. Observe the consistent and rapid linear convergence rate, not strongly depending on K or dimension d_k.

a single tuning parameter. Specifically, we can expect good performance when we set

$$\rho_k = \bar{\rho}\sqrt{\frac{\log p}{nm_k}} \tag{5.30}$$

with $\bar{\rho}$ being the single tuning parameter. Below, we experimentally validate the reliability of this tuning parameter expression (5.30) for a case of $K = 3$ factors.

As performance metrics, we consider the Frobenius norm ($\|\widehat{\Omega} - \Omega_0\|_F$) and spectral norm ($\|\widehat{\Omega} - \Omega_0\|_2$ errors of the precision matrix estimate $\widehat{\Omega}$, and use the Matthews

correlation coefficient to quantify the edge support estimation performance. Let the number of true positive edge detections be TP, true negatives TN, false positives FP, and false negatives FN. The Matthews correlation coefficient is defined as (*Matthews*, 1975)

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where we consider each nonzero off diagonal element of Ψ_k as a single edge. Increasing values of MCC imply better edge estimation performance, with MCC $= 0$ implying complete failure and MCC $= 1$ implying perfect edge set estimation.

We conduct experiments to verify the single tuning parameter approach, shown in Figure 5.7. Shown are the MCC, normalized Frobenius error, and spectral norm error as functions of $\bar{\rho}_1$ and $\bar{\rho}_2$ where

$$\bar{\rho}_k = \frac{\rho_k}{\sqrt{\frac{\log p}{n m_k}}}. \tag{5.31}$$

If $\bar{\rho}_1 = \bar{\rho}_2 = \bar{\rho}_3$ is near optimal, then indeed the formula (5.30) is successful and using the single tuning parameter will give reliable results.

5.7.3 Empirical Validation of Statistical Convergence

Having verified the single tuning parameter approach, hereafter we will cross-validate only $\bar{\rho}$ in our plots, using the formulaic variation with respect to $\mathbf{p}$ and n. We next verify that our bounds on the rate of convergence are tight. In this experiment, we will hold $\|\Sigma_0\|_2$ and s/p constant. Following the full form bound on the Frobenius error $\|\widehat{\Omega} - \Omega_0\|_F$ in Lemma V.15, we use

$$\rho_k = C\sqrt{\frac{\log p}{n m_k}} \tag{5.32}$$

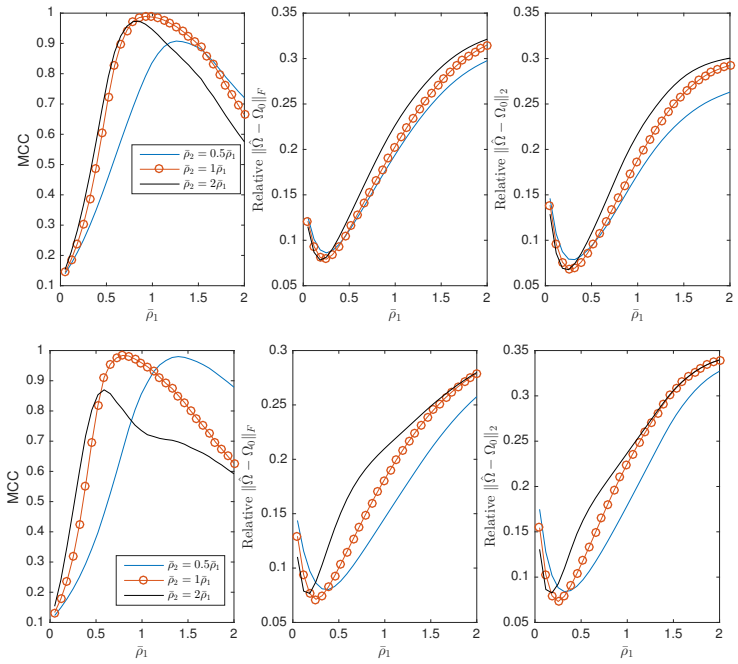


Figure 5.7: Setting tuning parameters with $K = 3$, $n = 1$, and $d_1 = d_3 = 64$. Shown are the MCC, relative Frobenius error, and relative L2 error of the estimate as the scaled tuning parameters are varied (5.31). Shown are deviations of $\bar{\rho}_2$ from the theoretically dictated $\bar{\rho}_2 = \bar{\rho}_1 = \bar{\rho}_3$. Top: Equal dimensions, $d_1 = d_2 = d_3$. First and third factors are random ER graphs with d_k edges, and the second factor is random grid graph with $d_k/2$ edges. Bottom: Dimensions $d_2 = 2d_1$, each factor is a random ER graph with d_k edges. Notice that in all these scenarios, using $\bar{\rho}_1 = \bar{\rho}_2$ is near optimal, confirming the viability of using a theory-motivated single-tuning-parameter approach guided by (5.30). This fact significantly simplifies the problem of choosing the tuning parameters.

where C is an absolute constant. By Lemma V.15, this implies an "effective number of samples" proportional to the inverse of the bound on $\|\widehat{\Omega} - \Omega_0\|_F^2/p$:

$$N_{eff} \propto \frac{n}{\left(\sum_{k=1}^{K} \sqrt{\frac{\log p}{m_k}}\right)^2}. \tag{5.33}$$

For each experiment, we varied K and d_2 over a total of 6 scenarios. To ensure that the constants in the bound were minimally affected, we held Ψ_1 constant over all (K, d_2) scenarios, and let $\Psi_3 = 0$ and $d_3 = d_1$ when $K = 3$. We let d_2 vary by powers of 2, allowing us to create a fixed matrix B and set $\Psi_2 = I_{d_2/d_{2,base}} \otimes B$ to ensure the

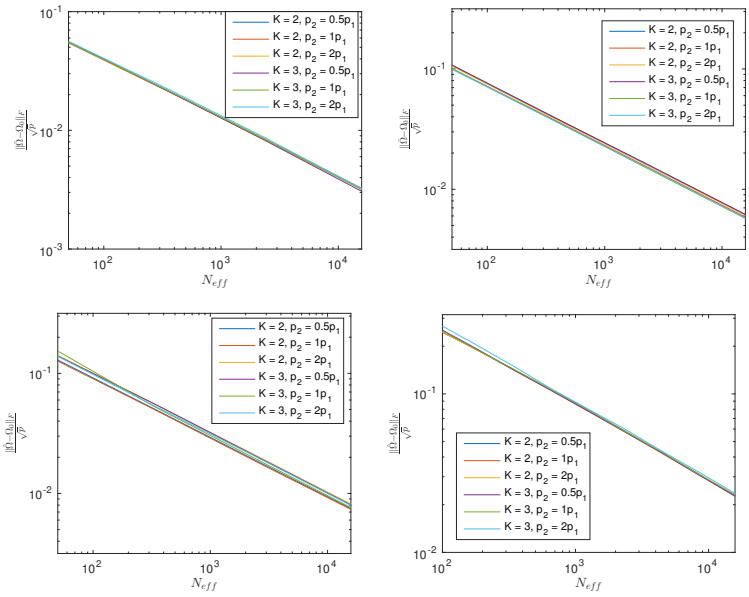


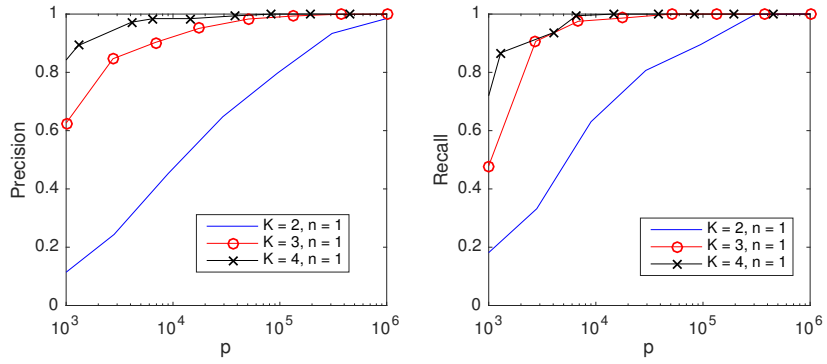
Figure 5.8: Frobenius norm convergence rate verification for proposed TeraLasso. Shown (ordered by increasing difficulty) are results for AR graphs with $d_1 = 40$ (top left), random ER graphs with $d_1 = 10$ (top right), $d_1 = 40$ (bottom left), and random grid graphs with $d_1 = 36$ (bottom right). For each covariance, 6 different combinations of d_2 and K are considered, and the resulting Frobenius error plotted versus the effective number of samples N_{eff} (5.33). In all cases, ρ_k are set according to their theoretical values in equation (5.32). The closeness of the plots over the six scenarios verifies the tightness of the bounds we derive.

eigenvalues of Ψ_2 and thus $\|\Sigma_0\|_2$ remain unaffected as d_2 changes.

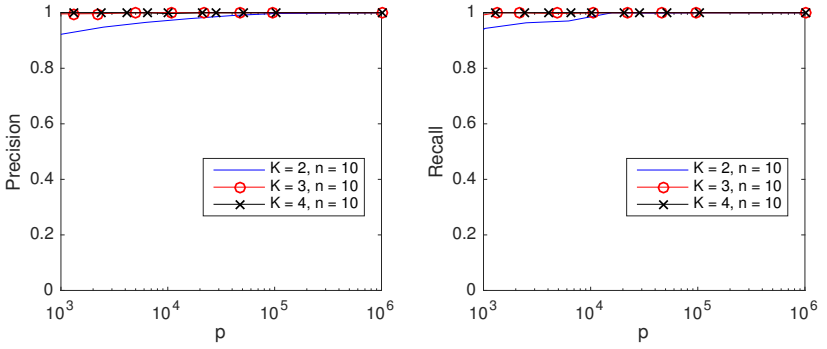
Results averaged over random training data realizations are shown in Figure 5.8 for ER ($d_k/2$ edges per factor), random grid ($d_k/2$ edges per factor), and AR-1 graphs (AR parameter .5 for both factors). Observe that in each case, the curves for the six scenarios are very close in spite of the wide variation in dimension, indicating that our expression for the effective sample size and thus our bound on the rate of convergence is tight. Small variations are not unexpected, since the objective function is joint over all factors, making the several approximations needed in the proof have varying degrees of accuracy.

Figure 6.5 illustrates how increasing dimension p and the increasing structure associated with larger K improves single sample performance. Shown are the average

TeraLasso edge detection precision and recall values for different values of K in the single and 10-sample regimes, all increasing to 1 (perfect structure estimation) as p, K, and n increase.



(a) $n = 1$ sample



(b) $n = 10$ sample

Figure 5.9: Low sample edge support estimation on random ER graphs, with the ρ_k set according to (5.32). Graphical model edge detection precision and recall curves are shown as a function of data dimension $p = \prod_{k=1}^{K} d_k$. For each value of the tensor order K, we set all the $d_k = p^{1/K}$. Observe single sample convergence as the dimension p increases and the structure increases (increasing K).

5.8 Real Data

5.8.1 NCEP Windspeed Data

We illustrate the TeraLasso model on a dataset involving measurements on a spatio-temporal grid, specifically, meteorological data from the US National Center

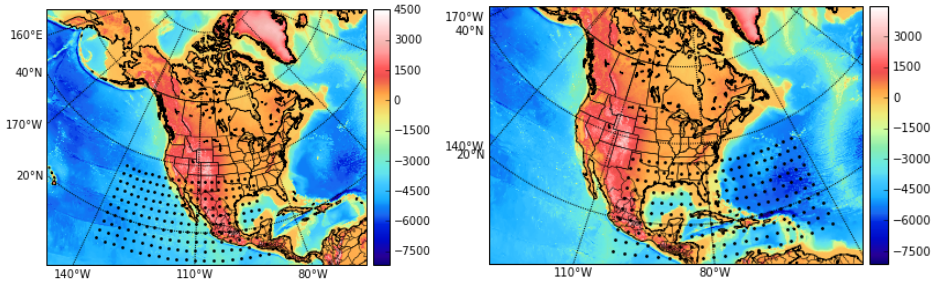


Figure 5.10: Rectangular 10×20 latitude-longitude grids of windspeed locations shown as black dots. Elevation colormap shown in meters. Left: "Western grid", Right: "Eastern grid".

for Environmental Prediction (NCEP). One such data source is meteorological data. The NCEP maintains records of average daily wind velocities in the lower troposphere, with daily readings beginning in 1948. Velocities are recorded globally, in a 144×73 latitude-longitude grid with spacings of 2.5 degrees in each coordinate. Over large but bounded areas, the spacing is approximately a rectangular grid, suggesting a $K = 2$ model (latitude vs. longitude) for the spatial covariance, and a $K = 3$ model (latitude vs. longitude vs. time) for the full spatio-temporal covariance.

We considered the time series of daily-average wind speeds. We first regressed out the mean for each day in the year via a 14-th order polynomial regression on the entire history from 1948-2015 (*Tsiligkaridis and Hero*, 2013). We extracted two 20×10 spatial grids, one from eastern North America, and one from Western North America, with the latter including an expansive high-elevation area and both Atlantic and Pacific oceans (Figure 5.10). Figures 5.11 and 5.12 show precision matrix estimate results trained on the eastern and western grids respectively, using time samples from January in n years following 1948. The TeraLasso estimates are compared to the unstructured shrinkage estimator, the TeraLasso estimator with sparsity parameter $\rho = 0$ (non-sparse Kron. sum estimate), and the maximum likelihood Kronecker product estimator. Note the graphical structure similar to that associated with AR structure in each dimension - a not unreasonable structure given the physical ar-

rangement of the wind stations. The TeraLasso estimate is much more stable than the Kronecker product estimate for much lower sample size n. The estimated latitude and longitude factors are shown in Figure 5.13. Observe the approximately AR structure, and the break in correlation (Figure 5.13, bottom right) in the Western Longitude factor. This break corresponds to the high elevation line of the Rocky Mountains, which is satisfying given the geographical feature's significant impact on weather patterns.

As an example application of the learned precision matrices, we consider likelihood-based season classification. We considered data in the 51-year span from 1948-2009. We considered training spatial precision matrices on n consecutive days in January and June of a training year respectively, and running anomaly detection on 30-day sequences of observations in the remaining 50 testing years. Note that for all results, we cycle through 51 1-vs.-50 train-test partitions of the 51-year data, and average the results. One set of sequences was from summer (June), and the other from winter (January), and we computed the classification error rate for the winter vs. summer classifier obtained by comparing the loglikelihoods $-\sum_{i=1}^{m}(\mathbf{x}_i - \mu_i)^T\widehat{\Omega}(\mathbf{x}_i - \mu_i)$. The $K = 2$ results for TeraLasso are shown in Figure 5.14 (top), with unregularized TeraLasso (ML Kronecker Sum) and maximum likelihood Kronecker product results shown for comparison. Note the superiority and increased single sample robustness of the ML Kronecker Sum estimate as compared to the ML Kronecker product estimate, confirming the better fit of TeraLasso. Spatio-temporal tensor ($K = 3$) results for different sized temporal covariance extents ($T = d_3$) are shown in Figure 5.15 (top) with similar results. We only show the overall classification error rate (instead of the elements of the confusion matrix) because the low number of test samples implies that the apportionment of the total error into the two error types is not statistically meaningful.

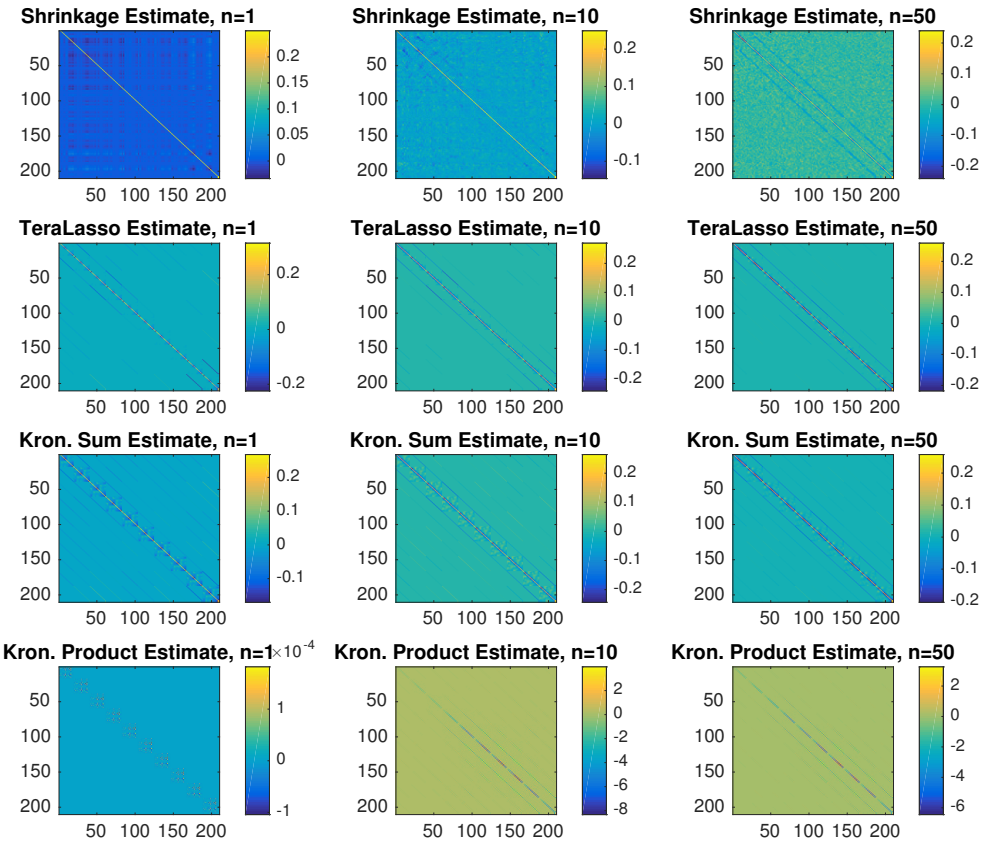


Figure 5.11: Windspeed data, eastern grid. Inverse spatial covariance estimation, comparing TeraLasso to unstructured and Kronecker product techniques. Spatial covariance, $K = 2$. Observe TeraLasso's ability to recover structure with even one sample. For improved contrast, the diagonal elements have been reduced in the plot.

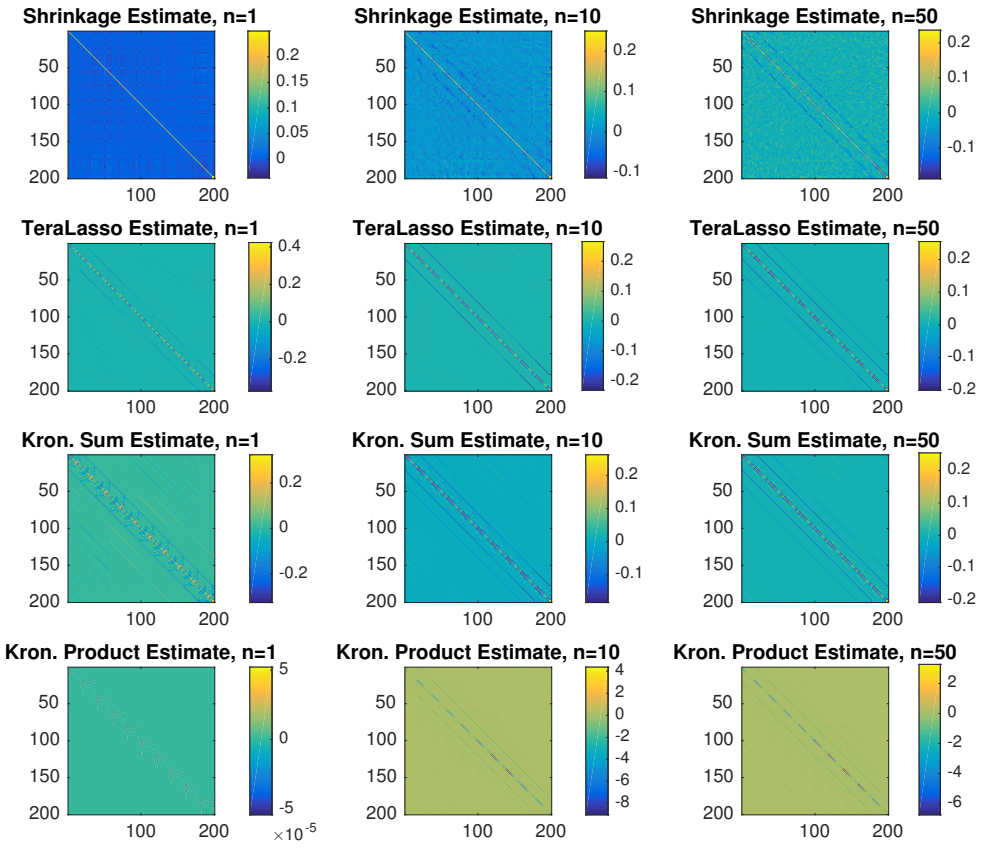


Figure 5.12: Windspeed data, western grid. Inverse spatial covariance estimation, comparing TeraLasso to unstructured and Kronecker product techniques. Spatial covariance, $K = 2$. Observe TeraLasso's ability to recover structure with even one sample.

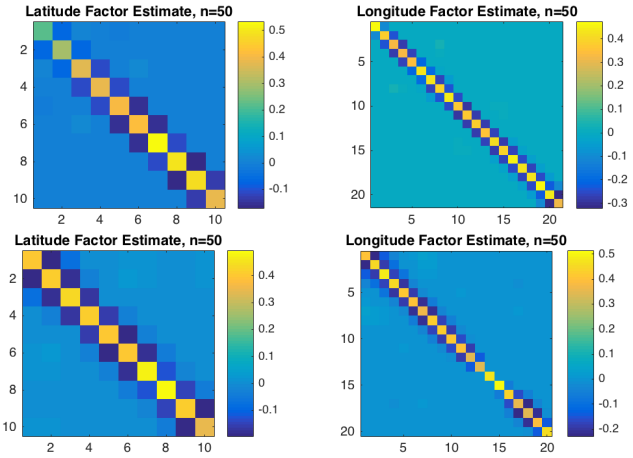
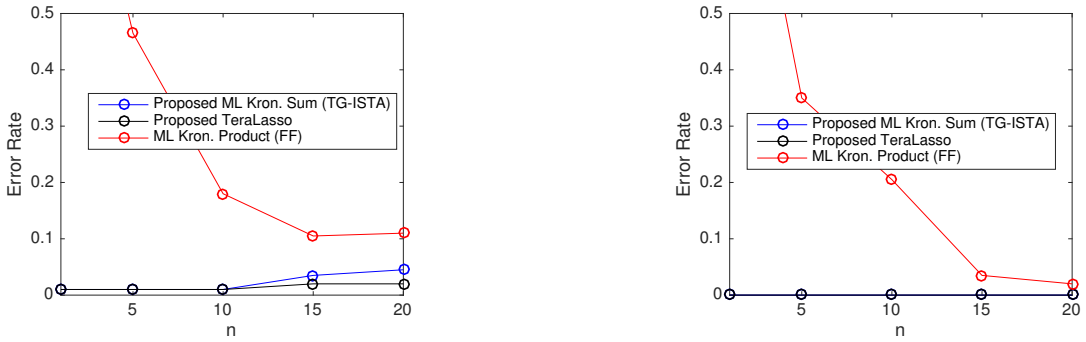


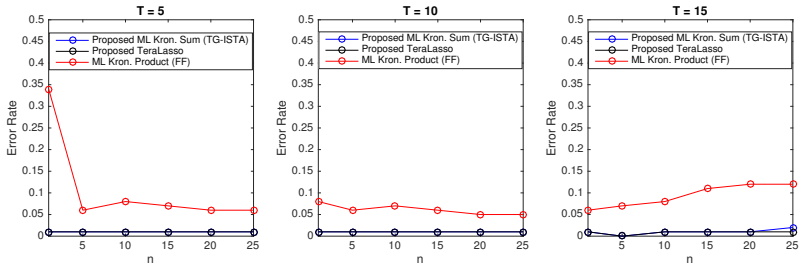
Figure 5.13: TeraLasso estimate factors, $K = 2$. Top: Eastern grid, Bottom: Western grid. Observe the decorrelation (the longitude factor entries connecting nodes 1-13 to nodes 14-20 are essentially zero) in the Western grid longitudinal factor, corresponding to the high-elevation line of the Rocky Mountains.



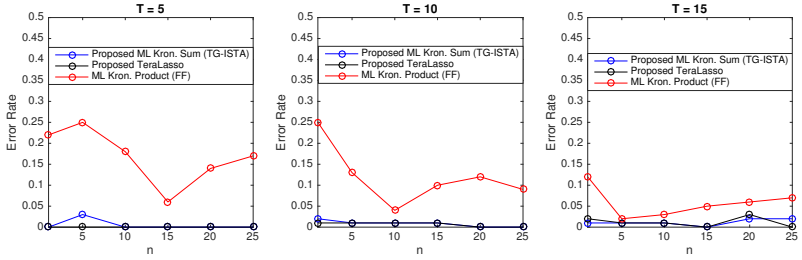
(a) Eastern grid

(b) Western grid

Figure 5.14: Classification using Gaussian loglikelihood and learned spatial ($K = 2$) precision matrices for each season. Shown are windspeed summer vs. winter classification error rates as a function of training samples n. Due to the low number of testing examples (50 per season) and low error rate of the proposed methods, we do not differentiate the two types of error. Note the $n = 1$ stability of the Kronecker sum estimate, and superior performance throughout indicating better model fit.



(a) Eastern grid



(b) Western grid

Figure 5.15: Classification using Gaussian loglikelihood and learned spatio-temporal ($K = 3$) precision matrices for each season, where T is the temporal dimension in days. Shown is windspeed summer vs. winter classification error rate as a function of training samples n and length of temporal window T. Note the $n = 1$ stability of the Kronecker sum estimate, and superior performance throughout indicating better model fit.

5.8.2 EEG Seizure Data

We also consider a canine EEG dataset collected for a Kaggle challenge on seizure prediction. This dataset consists of sets of hour-long 500Hz 16-lead EEG recordings on dogs with epilepsy. The recordings are divided into a set of *interictal* recordings which are temporally isolated from seizure events, and *preictal* recordings recorded from 1 hour, 5 minutes prior to a seizure event up until 5 minutes before the event.

The primary purpose of the dataset is to find local anomalies indicative of an oncoming seizure event. Due to the relative lack of preictal training data, and the a priori impossibility of determining the times at which warning events occur, we focus on detection of anomalous structure. For each one-hour segment, we use an epoch at the beginning of the segment to estimate the spatio-temporal Kronecker sum precision

matrix graph, which we set as a reference. The spatial factor is (16 leads x 16 leads), and we used a (20 samples x 20 samples) temporal factor. In a series of epochs spaced evenly through the remaining segment, we re-estimate the spatio-temporal covariance, and compare the estimated sparse edges with the reference graph. The total number of edge changes between the current estimate and the reference is then calculated. Example plots of the normalized edge changes as a function of time for a preictal segment of Dog #1 and of Dog #2 are shown in Figure 5.17. Note the significant increase in change towards the end of the segment closest to the oncoming seizure event.

Using the same segments from the two dogs, we show estimate graphs for each factor in Figure 5.18 and Figure 5.19 for a variety of number of training samples n. The times (1, 2, and 3) used as examples for each segment are marked on the trajectories in Figure 5.17. The training samples are obtained via a sliding temporal window, and thus are strongly correlated, inflating the value of n required for convergence. Note the increased temporal correlation structure at times 2 and 3 (near the seizure event) relative to the reference estimate at time 1. The changes in spatial structure are more subtle, but present, especially in magnitude.

Given the very small amount of preictal training data, it was impractical to identify a specific structure associated with seizure warning. We instead focused on detecting structural change, forming a reference structure estimate at the start of the episode, and tracking the total edge-change distance for all subsequent steps. The resulting centered interictal and preictal trajectories for both Dog #1 and Dog #2 are shown in Figure 5.20. Declaring a detection when the trajectories cross a threshold gives a preictal vs. interical detection performance with AUCs of .86 and .81 for Dogs #1 and 2 respectively, where all parameters were optimized on Dog #3 only. While our experimental setup is not directly comparable to the classification-based approach used in the Kaggle challenge, the winning entry in that challenge achieved a classifi-

cation AUC of .83, indicating the difficulty of the problem and perhaps the possibility of "warning episodes" not being guaranteed to produce full seizures. The relatively strong performance of our basic detector scheme highlights how TeraLasso's very low sample performance allows it to track rapid, short term changes in spatiotemporal graphical structure.

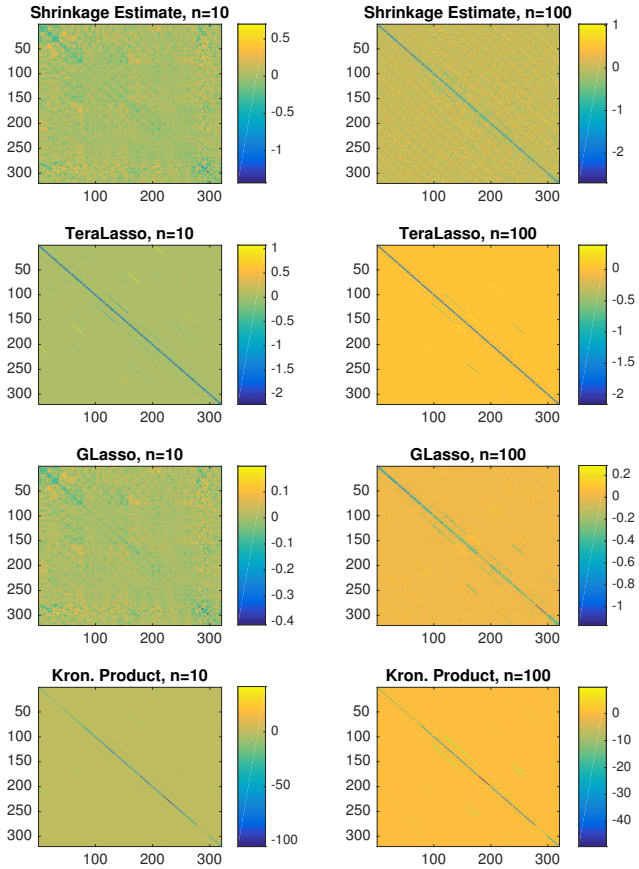


Figure 5.16: EEG data. Precision matrix estimates, with the diagonals set to zero for better contrast. Note the similarity of the GLasso estimate at $n = 100$ and the TeraLasso estimates. The Kronecker product estimate has many spurious positively-weighted edges away from the main diagonal block. The correspondence of the TeraLasso estimate to the high-sample GLasso estimate confirms that the Kronecker sum model fits the data.

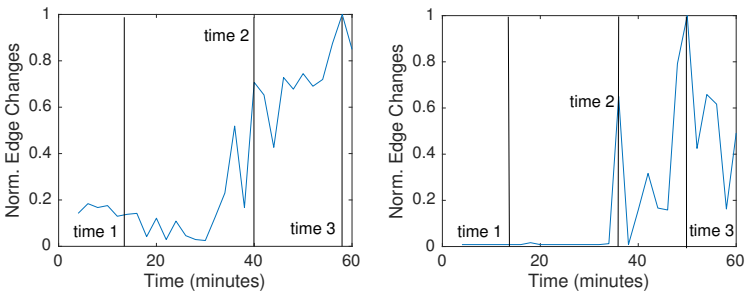


Figure 5.17: Example preictal edge change trajectories of the sparse TeraLasso estimate, computed for dogs #1 (left) and #2 (right). Note the large increase in change towards the end of the epoch, indicating an oncoming seizure. The times used in Figures 5.18 and 5.19 are marked.

5.9 Conclusion

We presented an extension of the Bigraphical Lasso method to tensor valued data, and presented an algorithm for estimation under the TeraLasso model. We derived strong performance guarantees for TeraLasso that showed significant gains over the unstructured approach, and demonstrated single-sample convergence in simulations as well as the tightness of the bounds. Finally, applications to real meteorological data and EEG data were considered, where the TeraLasso was shown to model the data well and enable improved single-sample performance for estimation, prediction, and anomaly detection.

While K in our model should not exceed the order of the data tensor, tensor modes can be combined to form lower-order (smaller K) models with less structure. The question of how and when to combine tensor modes could provide a way to transition between the ultra low sample K-order regime and a less-structured higher sample regime, and is an interesting question for future work. Potential extensions to our approach also include generalizing the first-order approach of TG-ISTA to incorporate Hessian information along the lines of (*Hsieh et al.*, 2014), and extending the inverse Kronecker sum model of TeraLasso to more general sums (along the lines of the sum of Kronecker products model (*Tsiligkaridis and Hero*, 2013; *Greenewald and Hero*,

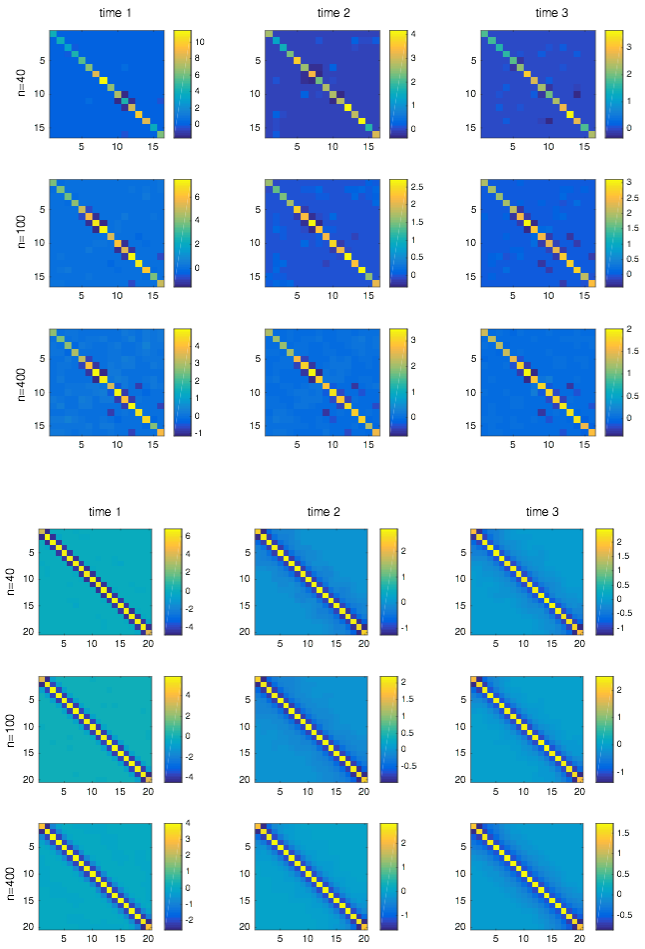


Figure 5.18: Estimated graphs at times 1, 2, and 3 for dog #1 for different values of n. Top: Spatial graph (16 variables) and Bottom: Temporal graph (20 samples). Note the changes in structure between the initial state (time 1) and the states closer to the seizure event (times 2, 3).

2015)) allowing for better approximation while maintaining the gains in sparsity and sample complexity.

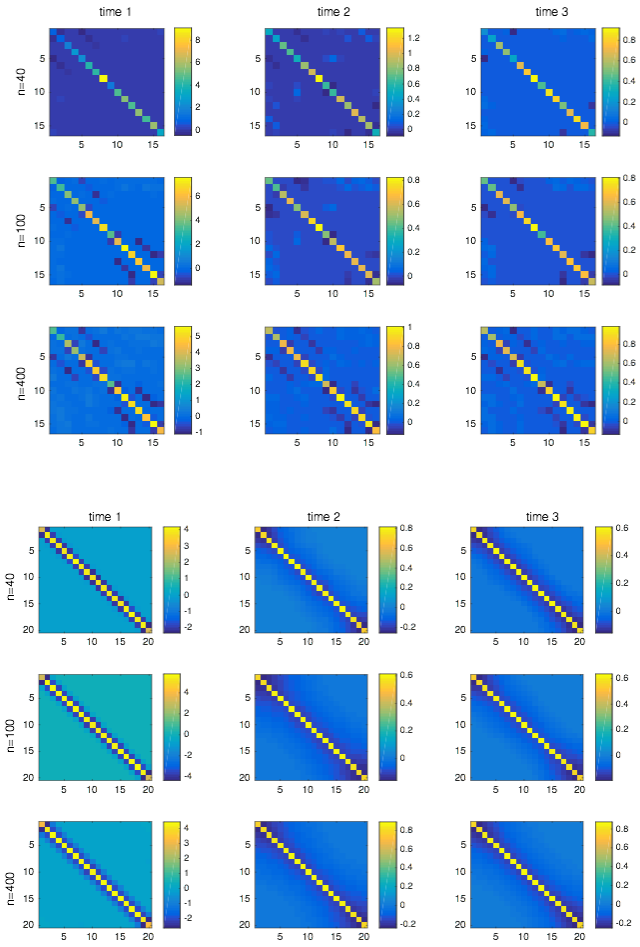


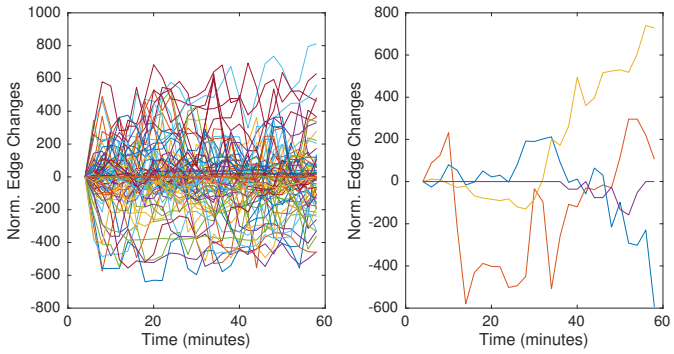
Figure 5.19: Estimated graphs at times 1, 2, and 3 for dog #2 for different values of n. Top: Spatial graph (16 variables) and Bottom: Temporal graph (20 samples). Note the changes in structure between the initial state (time 1) and the states closer to the seizure event (times 2, 3).

5.10 Appendix

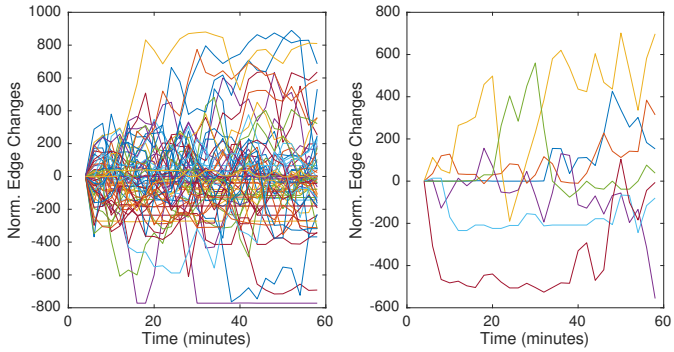
5.11 Useful Properties of the Kronecker Sum

5.11.1 Basic Properties

As the properties of Kronecker sums are not always widely known, we have compiled a list of some fundamental algebraic relations we use.



(a) Dog #1. Pre-seizure activity detection AUC .86.



(b) Dog #2. Pre-seizure activity detection AUC .81.

Figure 5.20: Centered edge-change trajectories vs. time. The total length of each segment is 1 hour, and all parameters were optimized on Dog #3 only. Left: Interictal segments, Right: Preictal. A threshold on absolute change gives a detector of pre-seizure activity with an AUC of .86 for Dog #1 and .81 for Dog #2.

1. Sum or difference of Kronecker sums (*Laub*, 2005):

$$c_A(A_1 \oplus \cdots \oplus A_K) + c_B(B_1 \oplus \cdots \oplus B_K) = (c_A A_1 + c_B B_1) \oplus \cdots \oplus (c_A A_K + c_B B_K).$$

2. Factor-wise disjoint off diagonal support (*Laub*, 2005). By construction, if for any k and $i \neq j$

$$[I_{[d_{1:k-1}]} \otimes A_k \otimes I_{[d_{k+1:K}]}]_{ij} \neq 0,$$

then for all $\ell \neq k$

$$[I_{[d_{1:\ell-1}]} \otimes A_\ell \otimes I_{[d_{\ell+1:K}]}]_{ij} = 0.$$

Thus,

$$|A_1^- \oplus \cdots \oplus A_K^-|_1 = \sum_{k=1}^{K} |I_{[d_{1:k-1}]} \otimes A_k^- \otimes I_{[d_{k+1:K}]}|_1 = \sum_{k=1}^{K} m_k |A_k^-|_1.$$

3. Eigendecomposition: If $A_k = U_k \Lambda_k U_k^T$ are the eigendecompositions of the factors, then (*Laub*, 2005)

$$A_1 \oplus \cdots \oplus A_K = (U_1 \otimes \cdots \otimes U_K)(\Lambda_1 \oplus \cdots \oplus \Lambda_K)(U_1 \otimes \cdots \otimes U_K)^T$$

is the eigendecomposition of $A_1 \oplus \cdots \oplus A_K$. Some resulting identities useful for doing numerical calculations are as follows:

(a) L2 norm:

$$\|A_1 \oplus \cdots \oplus A_K\|_2 = \max\left(\sum_{k=1}^{K} \max_i [\Lambda_k]_{ii}, -\sum_{k=1}^{K} \min_i [\Lambda_k]_{ii}\right) \leq \sum_{k=1}^{K} \|A_k\|_2.$$

(b) Determinant:

$$\log|A_1 \oplus \cdots \oplus A_K| =$$
$$\log|\Lambda_1 \oplus \cdots \oplus \Lambda_K| = \underbrace{\sum_{i_1=1}^{d_1} \cdots \sum_{i_K=1}^{d_K}}_{\text{K sums}} \log\left(\underbrace{[\Lambda_1]_{i_1 i_1} + \cdots + [\Lambda_K]_{i_K i_K}}_{\text{K terms}}\right).$$

(c) Matrix powers (e.g. inverse, inverse square root):

$$(A_1 \oplus \cdots \oplus A_K)^v = (U_1 \otimes \cdots \otimes U_K)(\Lambda_1 \oplus \cdots \oplus \Lambda_K)^v (U_1 \otimes \cdots \otimes U_K)^T.$$

Since the Λ_k are diagonal, this calculation is memory and computation efficient.

5.11.2 Projection onto $\widetilde{\mathcal{K}}_{\mathbf{p}}$

Lemma V.8 (Projection onto $\widetilde{\mathcal{K}}_{\mathbf{p}}$)**.** *For any $A \in \mathbb{R}^{p\times p}$,*

$$\begin{aligned}\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) &= A_1 \oplus \cdots \oplus A_K - (K-1)\frac{\mathrm{tr}(A)}{p} I_p \\ &= \left(A_1 - \frac{K-1}{K}\frac{\mathrm{tr}(A_1)}{d_1} I_{d_1}\right) \oplus \cdots \oplus \left(A_K - \frac{K-1}{K}\frac{\mathrm{tr}(A_K)}{d_K} I_{d_K}\right),\end{aligned}$$

where

$$A_k = \frac{1}{m_k}\sum_{i=1}^{m_k} A(i,i|k).$$

Since the submatrix operator $A(i,i|k)$ is clearly linear, $\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\cdot)$ *is a linear operator.*

Proof. Since $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is a linear subspace, projection can be found via inner products. Specifically, recall that if a subspace $\mathcal{A}$ is spanned by an orthonormal basis U, then

$$\mathrm{Proj}_{\mathcal{A}}(\mathbf{x}) = UU^T\mathbf{x}.$$

Since $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is the space of Kronecker sums, the off diagonal elements are independent and do not overlap across factors (Property 4). The diagonal portion is more difficult as each factor overlaps on the same entries, creating an overdetermined system. We can create an alternate parameterization of $\widetilde{\mathcal{K}}_{\mathbf{p}}$:

$$\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = \bar{A}_1 \oplus \cdots \oplus \bar{A}_K + \tau_A I_p = \tau_A I_p + \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]} \tag{5.34}$$

where we constrain $\mathrm{tr}(\bar{A}_k) = 0$. Each of the $K+1$ terms in this sum is now orthogonal

to all other terms since by construction

$$\begin{aligned}
\langle I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]}, I_{[d_{1:\ell-1}]} \otimes \bar{A}_\ell \otimes I_{[d_{\ell+1:K}]} \rangle \\
&= \frac{p}{d_k d_\ell} \mathrm{tr}((\bar{A}_k \otimes I_{d_\ell})(I_{d_k} \otimes \bar{A}_\ell)) \\
&= \frac{p}{d_k d_\ell} \mathrm{tr}(\bar{A}_k)\mathrm{tr}(\bar{A}_\ell) = 0 \\
\langle \tau_A I_p, I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]} \rangle &= \langle \tau_A I_{[d_{1:k-1}]} \otimes I_{d_k} \otimes I_{[d_{k+1:K}]}, I_{[d_{1:k-1}]} \otimes \bar{A}_k \otimes I_{[d_{k+1:K}]} \rangle \\
&= m_k \langle I_{d_k}, \bar{A}_k \rangle = m_k \mathrm{tr}(\bar{A}_k) = 0
\end{aligned}$$

for $\ell \neq k$ and all possible $\bar{A}_k$, τ_A. Thus, we can form bases for the $\bar{A}_k$ and τ_A independently. To find the $\bar{A}_k$ it suffices to project A onto a basis for $\bar{A}_k$. We can divide this projection into two steps. In the first step, we ignore the constraint on $\mathrm{tr}(\bar{A}_k)$ and create the orthonormal basis

$$\mathbf{u}_k^{(ij)} := \frac{1}{\sqrt{m_k}} I_{[d_{1:k-1}]} \otimes \mathbf{e}_i \mathbf{e}_j^T \otimes I_{[d_{k+1:K}]}$$

for all $i, j = 1, \ldots d_k$. Recall that in a projection of $\mathbf{x}$, the coefficient of a basis component $\mathbf{u}$ is given by $\mathbf{u}^T\mathbf{x} = \langle \mathbf{u}, \mathbf{x} \rangle$. We can thus apply this elementwise to the projection of A. Hence projecting A onto these basis components yields a matrix $B\sqrt{m_k} \in \mathbb{R}^{d_k \times d_k}$ where

$$B_{ij} = \frac{1}{m_k} \langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i \mathbf{e}_j^T \otimes I_{[d_{k+1:K}]} \rangle.$$

To enforce the $\mathrm{tr}(\bar{A}_k) = 0$ constraint, we project away from B the one-dimensional subspace spanned by I_{d_k}. This projection is given by

$$B - \frac{\mathrm{tr}(B)}{d_k} I_{d_k}, \tag{5.35}$$

where by construction

$$\begin{aligned}\frac{\mathrm{tr}(B)}{d_k} &= \frac{1}{d_k m_k}\sum_{i=1}^{d_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_i^T \otimes I_{[d_{k+1:K}]}\rangle \\ &= \frac{1}{p}\langle A, I_p\rangle = \frac{\mathrm{tr}(A)}{p}.\end{aligned}$$

Equation (5.35) completes the projection onto a basis for $\bar{A}_k$, so we can expand the projection $\sqrt{m_k}B$ back into the original space. This yields a $\bar{A}_k$ of the form

$$[\bar{A}_k]_{ij} = \begin{cases} \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_j^T \otimes I_{[d_{k+1:K}]}\rangle & i \neq j \\ \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_i^T \otimes I_{[d_{k+1:K}]}\rangle - \frac{\mathrm{tr}(A)}{p} & i = j \end{cases}$$

Finally, for τ_A we can compute

$$\tau_A = \frac{1}{p}\langle A, I_p\rangle = \frac{\mathrm{tr}(A)}{p}.$$

Combining all these together and substituting into (5.34) allows us to define the projection in terms of matrices $\widetilde{A}_k$, where we split the $\tau_A I_p$ term evenly across the other K factors. Specifically

$$\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = \widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K.$$

where

$$[\widetilde{A}_k]_{ij} = \begin{cases} \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_j^T \otimes I_{[d_{k+1:K}]}\rangle & i \neq j \\ \frac{1}{m_k}\langle A, I_{[d_{1:k-1}]} \otimes \mathbf{e}_i\mathbf{e}_i^T \otimes I_{[d_{k+1:K}]}\rangle - \frac{K-1}{K}\frac{\mathrm{tr}(A)}{p} & i = j \end{cases}.$$

An equivalent representation using factorwise averages is

$$\widetilde{A} = A_k - \frac{K-1}{K}\frac{\mathrm{tr}(A)}{p},$$

where

$$A_k = \sum_{i=1}^{m_k} A(i,i|k).$$

Moving the trace corrections to a last term and putting the result in terms of the A_k yields the lemma.

□

5.11.3 Inner Product in $\widetilde{\mathcal{K}}_{\mathbf{p}}$

Lemma V.9 (Inner Products with S)**.** *Suppose $B \in \mathbb{R}^{p\times p}$. Then for any $A_k \in \mathbb{R}^{d_k\times d_k}$, $k = 1,\ldots,K$,*

$$\langle B, A_1 \oplus \cdots \oplus A_K\rangle = \sum_{k=1}^{K} m_k \langle B_k, A_k\rangle.$$

Proof.

$$\begin{aligned}
\langle B, A_1 \oplus \cdots \oplus A_K\rangle &= \sum_{k=1}^{K} \langle B, I_{[d_{1:k-1}]} \otimes A_k \otimes I_{[d_{k+1:K}]}\rangle \\
&= \sum_{k=1}^{K}\sum_{i=1}^{m_k} \langle B(i,i|k), A_k\rangle \\
&= \sum_{k=1}^{K} \left\langle \sum_{i=1}^{m_k} B(i,i|k), A_k \right\rangle \\
&= \sum_{k=1}^{K} m_k \langle B_k, A_k\rangle.
\end{aligned}$$

where we have used the definition of the submatrix notation $B(i,i|k)$ and the matrices $B_k = \frac{1}{m_k}\sum_{i=1}^{m_k} B(i,i|k)$.

□

5.11.4 Identifiable Parameterization of $\widetilde{\mathcal{K}}_{\mathbf{p}}$

Lemma V.10. *The space $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is linearly, identifiably, and orthogonally parameterized by the quantities $\left(\tau_B \in \mathbb{R}, \left\{\widetilde{A}_k \in \{A \in \mathbb{R}^{d_k \times d_k} | \mathrm{tr}(A) \equiv 0\}\right\}_{k=1}^{K}\right)$. Specifically, any $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ can be identifiably written as*

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K)$$

where $\mathrm{tr}(\widetilde{A}_k) \equiv 0$. By orthogonality, the Frobenius norm can be decomposed as

$$\|B\|_F^2 = p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2 \geq \sum_{k=1}^{K} m_k \left\| \frac{\tau_B}{K} I_{d_k} + \widetilde{A}_k \right\|_F^2,$$

noting that

$$B = \left(\frac{\tau_B}{K} I_{d_1} + \widetilde{A}_1\right) \oplus \cdots \oplus \left(\frac{\tau_B}{K} I_{d_K} + \widetilde{A}_K\right).$$

Corollary V.11 (Construction of Identifiable Parameterization)**.** *Given a $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ with any representation $B = A_1 \oplus \cdots \oplus A_K$, the identifiable parameters $(\tau_B, \{\widetilde{A}_k\}_{k=1}^{K})$ can be computed as*

$$\tau_B = \frac{\mathrm{tr}(B)}{p}, \qquad \widetilde{A}_k = A_k - \frac{\mathrm{tr}(A_k)}{d_k} I_{d_k}.$$

The parameterized result is

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K).$$

5.11.4.1 Identifiable Parameterization

We now prove Lemma V.10.

Proof. Based on the original parameterization

$$B = A_1 \oplus \cdots \oplus A_K,$$

we know that the number of degrees of freedom in B is much smaller than the number of elements p^2. We thus seek a lower-dimensional parameterization of B. The Kronecker sum parameterization is not identifiable on the diagonals, so we seek a representation of B that is identifiable.

Let $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$. By definition, there exists $A_1, \ldots, A_K$ such that

$$\begin{aligned} B = A_1 \oplus \cdots \oplus A_K &= \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes A_k \otimes I_{[d_{k+1:K}]} \\ &= \sum_{k=1}^{K} \left(I_{[d_{1:k-1}]} \otimes (A_k - \tau_k I_{d_k}) \otimes I_{[d_{k+1:K}]} + \tau_k I_p \right) \\ &= \left(\sum_{k=1}^{K} \tau_k \right) I_p + ((A_1 - \tau_1 I_{d_1}) \oplus \cdots \oplus (A_K - \tau_K I_{d_K})). \end{aligned}$$

where $\tau_k = \mathrm{tr}(A_k)/d_k$. Observe that $\mathrm{tr}(A_k - \tau_k I_{d_k}) = 0$ by construction, so we can set $\widetilde{A}_k = A_k - \tau_k I_{d_k}$, creating

$$B = \left(\sum_{k=1}^{K} \tau_k \right) I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K).$$

Note that in this representation, $\mathrm{tr}(\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K) = 0$, so letting $\tau_B = \mathrm{tr}(B)/p$,

$$\tau_B = \sum_{k=1}^{K} \tau_k.$$

Thus, we can write any $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ as

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K),$$

and in the other direction, it is easy to verify any B expressible in this form is in $\widetilde{\mathcal{K}}_{\mathbf{p}}$. Observe that under this parameterization B is a *linear* function of the parameters $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$.

Thus, $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$ parameterizes $\widetilde{\mathcal{K}}_{\mathbf{p}}$. It remains to show that this parameterization is identifiable.

5.11.4.2 Orthogonal Parameterization

We will show that under the linear parameterization of $\widetilde{\mathcal{K}}_{\mathbf{p}}$ by $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$, each of the $K+1$ components are linearly independent of the others.

To see this, we compute the inner products between the components:

$$\begin{aligned}
\langle \tau_B I_p, I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:K}]} \rangle &= \tau_B m_k \mathrm{tr}(\widetilde{A}_k) \equiv 0 \\
\langle I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:K}]}, &I_{[d_{1:\ell-1}]} \otimes \widetilde{A}_\ell \otimes I_{[d_{\ell+1:K}]} \rangle \\
&= \mathrm{tr}\left(I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:\ell-1}]} \otimes \widetilde{A}_\ell \otimes I_{[d_{\ell+1:K}]} \right) \\
&= \frac{p}{d_k d_\ell} \mathrm{tr}(\widetilde{A}_k) \mathrm{tr}(\widetilde{A}_\ell) \equiv 0,
\end{aligned}$$

for all possible values of $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$ and combinations of $k < \ell$. We have recalled that by definition, $\mathrm{tr}(\widetilde{A}_k) \equiv 0$ identically. Since all the inner products are identically zero, the components are orthogonal, thus they are linearly independent. Hence, by the definition of linear independence, this linear parameterization $(\tau_B, \{\widetilde{A}_k\}_{k=1}^K)$ is uniquely determined by $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ (i.e. it is identifiable).

5.11.4.3 Decomposition of Frobenius norm

Using the identifiability and orthogonality of this parameterization, we can find a direct factorwise decomposition of the Frobenius norm on $\widetilde{\mathcal{K}}_{\mathbf{p}}$.

By orthogonality (cross term inner products equal to zero)

$$\begin{aligned}\|B\|_F^2 &= \|\tau_B I_p\|_F^2 + \sum_{k=1}^{K} \|I_{[d_{1:k-1}]} \otimes \widetilde{A}_k \otimes I_{[d_{k+1:K}]}\|_F^2 \\ &= p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2.\end{aligned} \tag{5.36}$$

This completes the first decomposition, representing the squared Frobenius norm as weighted sum of the squared Frobenius norms on each component.

For convenience, we also observe that given any $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ with identifiable parameterization

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K),$$

we can absorb the scaled identity into the Kronecker sum and still bound the Frobenius norm decomposition. Specifically, observe that

$$p\tau_B^2 = pK \sum_{k=1}^{K} \left(\frac{\tau_B}{K}\right)^2 \geq p \sum_{k=1}^{K} \left(\frac{\tau_B}{K}\right)^2.$$

Substituting this into (5.36),

$$\begin{aligned}\|B\|_F^2 = p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2 &\geq p \sum_{k=1}^{K} \left(\frac{\tau_B}{K}\right)^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2 \\ &= \sum_{k=1}^{K} m_k \left(\left\| \frac{\tau_B}{K} I_{d_k} \right\|_F^2 + \|\widetilde{A}_k\|_F^2 \right) \\ &= \sum_{k=1}^{K} m_k \left\| \frac{\tau_B}{K} I_{d_k} + \widetilde{A}_k \right\|_F^2,\end{aligned}$$

where the last term follows because $\operatorname{tr}(\widetilde{A}_k) \equiv 0$ implies that $\langle I_{d_k}, \widetilde{A}_k \rangle \equiv 0$.

Observe that

$$B = \left(\frac{\tau_B}{K} I_{d_1} + \widetilde{A}_1\right) \oplus \cdots \oplus \left(\frac{\tau_B}{K} I_{d_K} + \widetilde{A}_K\right),$$

hence the lemma is proved.

□

5.11.4.4 Spectral Norm Bound

This parameterization reveals that the geometry of $\widetilde{\mathcal{K}}_{\mathbf{p}}$ implies a bound on the spectral norm relative to the Frobenius norm.

Corollary V.12 (Spectral Norm Bound)**.** *For all* $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$,

$$\|B\|_2 \leq \sqrt{\frac{K+1}{\min_k m_k}} \|B\|_F.$$

Proof. Using the identifiable parameterization of B

$$B = \tau_B I_p + (\widetilde{A}_1 \oplus \cdots \oplus \widetilde{A}_K),$$

and the triangle inequality, we have

$$\begin{aligned}
\|B\|_2 \leq |\tau_B| + \sum_{k=1}^{K} \|\widetilde{A}_k\|_2 \leq |\tau_B| + \sum_{k=1}^{K} \|\widetilde{A}_k\|_F &\leq \sqrt{K+1} \sqrt{\tau_B^2 + \sum_{k=1}^{K} \|\widetilde{A}_k\|_F^2} \\
&\leq \sqrt{\frac{K+1}{\min_k m_k}} \sqrt{p\tau_B^2 + \sum_{k=1}^{K} m_k \|\widetilde{A}_k\|_F^2} \\
&\leq \sqrt{\frac{K+1}{\min_k m_k}} \|B\|_F.
\end{aligned}$$

□

5.11.5 Generation of Kronecker Sum Random Tensors

Generating random tensors given a Kronecker sum precision matrix can be made efficient by exploiting the Kronecker sum eigenstructure. Algorithm 6 allows efficient generation of data following the TeraLasso model.

Algorithm 6 Generation of sub-Gaussian tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ under TeraLasso model.

Assume $\Sigma^{-1} = \Psi_1 \oplus \cdots \oplus \Psi_K$.
Input precision matrix factors $\Psi_k \in \mathbb{R}^{d_k \times d_k}$, $k = 1, \ldots, K$.
for $k = 1, \ldots, K$ **do**
    $U_k, \Lambda_k \leftarrow \mathrm{EIG}(\Psi_k)$ eigendecomposition of Ψ_k.
end for
$\mathbf{v} = [v_1, \ldots, v_p] \leftarrow \mathrm{diag}(\Lambda_1) \oplus \cdots \oplus \mathrm{diag}(\Lambda_K) \in \mathbb{R}^p$.
Generate isotropic subgaussian random vector $z \in \mathbb{R}^p$.
$\widetilde{x}_i \leftarrow v_i^{-1/2} z_i$, for $i = 1, \ldots, p$.
for $k = 1, \ldots, K$ **do**
    $\widetilde{\mathbf{x}} \leftarrow (I_{[d_{1:k-1}]} \otimes U_k \otimes I_{[d_{k+1:K}]})\widetilde{\mathbf{x}}$.
end for
Reshape $\widetilde{x}$ into $X \in \mathbb{R}^{d_1 \times \cdots \times d_K}$.

5.12 Proof of Joint Convexity (Theorem V.1)

Proof. Our objective function is (5.11)

$$Q(\{\Psi_k\}) = -\log|\Psi_1 \oplus \cdots \oplus \Psi_K| + \langle S, \Psi_1 \oplus \cdots \oplus \Psi_K \rangle + \sum_k \rho_k d_k \|\Psi_k^-\|_1 \tag{5.37}$$

By definition,

$$\Psi_1 \oplus \cdots \oplus \Psi_K = \Psi_1 \otimes I_{m_1} + \cdots + I_{m_K} \otimes \Psi_K \tag{5.38}$$

is an affine function of $\mathbf{z} = [\mathrm{vec}(\Psi_1); \ldots; \mathrm{vec}(\Psi_K)]$. Thus, since $\log|A|$ is a concave function (*Boyd and Vandenberghe*, 2009), all the terms of Q are convex since convex functions of affine functions are convex and the elementwise L1 norm is convex. Hence Q is jointly convex in $\{\Psi_k\}_{k=1}^K$ on $\mathcal{K}_{\mathbf{p}}$. Hence, every local minima is also global.

We show that a nonempty set of $\{\Psi_k\}_{k=1}^K$ such that $Q(\{\Psi_k\}_{k=1}^K)$ is minimized maps to a unique $\Omega = \Psi_1 \oplus \cdots \oplus \Psi_K$. If only one point $\{\Psi_k\}_{k=1}^K$ exists that achieves the global minimum, then the statement is proved. Otherwise, suppose that two distinct

points $\{\Psi_{k,1}\}_{k=1}^K$ and $\{\Psi_{k,2}\}_{k=1}^K$ achieve the global minimum Q^*. Then, for all k define

$$\Psi_{k,\alpha} = \alpha\Psi_{k,1} + (1-\alpha)\Psi_{k,2} \tag{5.39}$$

By convexity, $Q(\{\Psi_{k,\alpha}\}_{k=1}^K) = Q^*$ for all $\alpha \in [0,1]$, i.e. Q is constant along the specified affine line segment. This can only be true if (up to an additive constant) the first two terms of Q are equal to the negative of the second two terms along the specified segment. Since

$$-\log|A| + \langle S, A\rangle \tag{5.40}$$

is strictly convex and smooth on the positive definite cone (i.e. the second derivative along any line never vanishes) *Boyd and Vandenberghe* (2009) and the sum of the two elementwise $\ell 1$ norms along any affine combination of variables is at most piecewise linear when smooth, this cannot hold when $\Omega_\alpha = \Psi_{1,\alpha} \oplus \cdots \oplus \Psi_{K,\alpha}$ varies with α. Hence, Ω_α must be a constant Ω^* with respect to α. Thus, the minimizing Ω^* is unique and Theorem V.1 is proved.

□

5.13 Statistical Convergence: Proof of Theorem V.2

5.13.1 Notation

Let Ω_0 be the true value of the precision matrix Ω. Since $\Omega, \Omega_0 \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ and $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is convex, we can write

$$\Delta_\Omega = \Omega - \Omega_0 = \Delta_\Omega^+ + (\Delta_{\Psi,1}^- \oplus \cdots \oplus \Delta_{\Psi,K}^-), \tag{5.41}$$

where $\Delta_{\Omega}^{+} = \Omega^{+} - \Omega_{0}^{+}$ and $\Delta_{\Psi,k}^{-} = \Psi_{k}^{-} - \Psi_{0,1}^{-}$ and we recall the identifiable parameterizations

$$\Omega = \Omega^{+} + (\Psi_{1}^{-} \oplus \cdots \oplus \Psi_{K}^{-})$$
$$\Omega_{0} = \Omega_{0}^{+} + (\Psi_{0,1}^{-} \oplus \cdots \oplus \Psi_{0,K}^{-}).$$

due to the identifiability of the off diagonals.

Let $I(\cdot)$ be the indicator function. For an index set A and a matrix $M = [m_{ij}]$, define the operator $\mathcal{P}_{A}(M) \equiv [m_{ij} I((i,j) \in A)]$ that projects M onto the set A. Let $\Delta_{k,S}^{-} = \mathcal{P}_{\mathcal{S}_k}(\Delta_{\Psi,k}^{-})$ be the projection of $\Delta_{\Psi,k}^{-}$ onto the true sparsity pattern of that factor. Let $\mathcal{S}_{k}^{c}$ be the complement of $\mathcal{S}_k$, and $\Delta_{k,S^c} = \mathcal{P}_{\mathcal{S}_{k}^{c}}(\Delta_{\Psi,k}^{-})$.

Furthermore, let

$$\Delta_{\Omega,S} = (\Delta_{1,S}^{-} \oplus \cdots \oplus \Delta_{K,S}^{-})$$

be the projection of Δ_{Ω} onto the full sparsity pattern $\mathcal{S}$. Recall $\mathcal{S}$ does not include the diagonal.

5.13.2 Proof

Proof. Let

$$\begin{aligned}\bar{Q}(\Omega) = Q(\Omega) - Q(\Omega_0) =& \langle \Omega, S \rangle - \log|\Omega| + \sum_{k} \rho_k m_k |\Psi_{k}^{-}|_1 \\ & - \langle \Omega_0, S \rangle + \log|\Omega_0| - \sum_{k} \rho_k m_k |\Psi_{k,0}^{-}|_1 \end{aligned} \tag{5.42}$$

be the difference between the objective function (5.9) at Ω and at Ω_0. This definition is invariant to the decompositions of Ω and Ω_0 since the objective function is (Theorem 1). Let $G(\Delta_{\Omega}) = \bar{Q}(\Omega_0 + \Delta_{\Omega})$. Then clearly $\widehat{\Delta}_{\Omega} = \widehat{\Omega} - \Omega_0$ minimizes $G(\Delta_{\Omega})$, which

is a convex function with a unique minimizer on $\mathcal{K}_{\mathbf{p}}$ as per Theorem V.1.

Define

$$\mathcal{T}_n = \{\Delta_\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}} : \Delta_\Omega = \Omega - \Omega_0, \Omega, \Omega_0 \in \mathcal{K}_{\mathbf{p}}, \|\Delta_\Omega\|_F = M r_{n,\mathbf{p}}\} \tag{5.43}$$

for some $r_{n,\mathbf{p}}$ and

$$M = \frac{1}{2}\phi_{\max}(\Omega_0)\sqrt{\frac{\min_k m_k}{K+1}}. \tag{5.44}$$

We have the following proposition:

Proposition V.13. *In the event that $G(\Delta_\Omega) > 0$ for all $\Delta_\Omega \in \mathcal{T}_n$, we have that*

$$\|\widehat{\Delta}_\Omega\|_F < M r_{n,\mathbf{p}}.$$

Proof. By definition, $G(0) = 0$, so $G(\widehat{\Delta}_\Omega) \leq G(0) = 0$. Thus if $G(\Delta_\Omega) > 0$ on $\mathcal{T}_n$, then by Proposition V.18 (Appendix 5.14.2.1), $\widehat{\Delta}_\Omega \notin \mathcal{T}_n \cup \mathcal{V}_n$ where $\mathcal{V}_n$ is defined therein. The proposition results. □

Thus it remains to show that $G(\Delta_\Omega) > 0$ on $\mathcal{T}_n$ for some $r_{n,\mathbf{p}}$.

In Lemma V.17 in Appendix 5.14.2 we show that if $r_{n,\mathbf{p}} \leq 1$, for all $\Delta_\Omega \in \mathcal{T}_n$ we have that

$$\log|\Omega_0 + \Delta_\Omega| - \log|\Omega_0| \leq \langle \Sigma_0, \Delta_\Omega \rangle - 2\|\Delta_\Omega\|_F^2 / (9\|\Omega_0\|_2^2), \tag{5.45}$$

Substituting into (5.42),

$$
\begin{aligned}
G(\Delta_\Omega) &= \bar{Q}(\Omega_0 + \Delta_\Omega) \qquad (5.46)\\
&= \langle \Omega_0 + \Delta_\Omega, S\rangle - \langle \Omega_0, S\rangle - \log|\Omega_0 + \Delta_\Omega| + \log|\Omega_0| \\
&\quad + \sum_k \rho_k m_k (|\Psi^-_{k,0} + \Delta^-_{\Psi,k}|_1 - |\Psi^-_{k,0}|_1) \\
&\geq \langle \Delta_\Omega, S\rangle - \langle \Delta_\Omega, \Sigma_0\rangle + \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 + \sum_k \rho_k m_k (|\Psi^-_{k,0} + \Delta^-_{\Psi,k}|_1 - |\Psi^-_{k,0}|_1) \\
&= \langle \Delta_\Omega, S - \Sigma_0\rangle + \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 + \sum_k \rho_k m_k (|\Psi^-_{k,0} + \Delta^-_{\Psi,k}|_1 - |\Psi^-_{k,0}|_1).
\end{aligned}
$$

where we have used the bilinearity of the inner product. We next bound the inner product term with high probability.

Lemma V.14. *Let A be the event defined in the proof (Appendix 5.14.1). Then $\Pr(A) \geq 1 - 2(K+1)\exp(-c\log p)$, and when A holds the following holds:*

$$
|\langle \Delta_\Omega, S - \Sigma_0\rangle| \leq C'\|\Sigma_0\|_2 \sum_{k=1}^{K} \left(|\Delta^+_\Omega|_1 + m_k|\Delta^-_{\Psi,k}|_1\right)\sqrt{\frac{\log p}{m_k n}}.
$$

The proof is in Appendix 5.14.1.

Substituting the bound of Lemma V.14 into (5.46), we have that under event A

$$
\begin{aligned}
G(\Delta_\Omega) \geq \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - C'\|\Sigma_0\|_2 \sum_{k=1}^{K} \left(|\Delta^+_\Omega|_1 + m_k|\Delta^-_{\Psi,k}|_1\right)\sqrt{\frac{\log p}{m_k n}} \qquad (5.47)\\
+ \underbrace{\sum_k \rho_k m_k (|\Psi^-_{k,0} + \Delta^-_{\Psi,k}|_1 - |\Psi^-_{k,0}|_1)}_{I}.
\end{aligned}
$$

By the decomposability of the L1 norm and the reverse triangle inequality $|A+B|_1 \geq$

$|A|_1 - |B|_1$, we have

$$\begin{aligned} |\Psi^-_{k,0} + \Delta^-_{\Psi,k}|_1 - |\Psi^-_{k,0}|_1 &= |\Psi^-_{k,0} + \Delta^-_{k,S}|_1 + |\Delta^-_{k,S^c}|_1 - |\Psi^-_{k,0}|_1 \\ &\geq |\Psi^-_{k,0}|_1 - |\Delta^-_{k,S}|_1 + |\Delta^-_{k,S^c}|_1 - |\Psi^-_{k,0}|_1 \\ &\geq |\Delta^-_{k,S^c}|_1 - |\Delta^-_{k,S}|_1 \\ &= |\Delta^-_{\Psi,k}|_1 - 2|\Delta^-_{k,S}|_1, \end{aligned} \tag{5.48}$$

since $\Psi_{k,0}$ is assumed to follow sparsity pattern $\mathcal{S}_k$ by assumption A1. Substituting (5.48) into the expression for term I of (5.47),

$$\begin{aligned} I = \sum_k \rho_k m_k (|\Psi^-_{k,0} + \Delta^-_{\Psi,k}|_1 - |\Psi^-_{k,0}|_1) &\geq \sum_k \rho_k m_k (|\Delta^-_{\Psi,k}|_1 - 2|\Delta^-_{k,S}|_1) \\ &\geq \left(\sum_k \rho_k m_k |\Delta^-_{\Psi,k}|_1 \right) - 2(\max_k \rho_k) \sum_k m_k |\Delta^-_{k,S}|_1 \\ &\geq \left(\sum_k \rho_k m_k |\Delta^-_{\Psi,k}|_1 \right) - 2|\Delta^-_{\Omega,S}|_1 \sum_k \rho_k, \end{aligned} \tag{5.49}$$

where we have recalled the fact that $\sum_k m_k |\Delta^-_{k,S}|_1 = |\Delta^-_{\Omega,S}|_1$ (Appendix A.1 Property 2). Substituting (5.49) into (5.47), under event A

$$\begin{aligned} G(\Delta_\Omega) \geq & \frac{2}{9\|\Omega_0\|_2^2} \|\Delta_\Omega\|_F^2 - C' \|\Sigma_0\|_2 \sum_{k=1}^K \left(|\Delta^+_\Omega|_1 + m_k |\Delta^-_{\Psi,k}|_1 \right) \sqrt{\frac{\log p}{m_k n}} \\ & + \left(\sum_k \rho_k m_k |\Delta^-_{\Psi,k}|_1 \right) - 2|\Delta^-_{\Omega,S}|_1 \sum_k \rho_k. \end{aligned} \tag{5.50}$$

Substituting the expressions $\rho_k = C\|\Sigma_0\|_2\sqrt{\frac{\log p}{nm_k}}$ into (5.51) gives, for $C \geq C'$

$$
\begin{aligned}
G(\Delta_\Omega) \geq & \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - C'\|\Sigma_0\|_2|\Delta_\Omega^+|_1 \sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}} \\
& + \underbrace{\left(C\|\Sigma_0\|_2 \sum_{k=1}^K m_k |\Delta_{\Psi,k}^-|_1 \sqrt{\frac{\log p}{nm_k}}\right) - \left(C'\|\Sigma_0\|_2 \sum_{k=1}^K m_k |\Delta_{\Psi,k}^-|_1 \sqrt{\frac{\log p}{m_k n}}\right)}_{\geq 0} \\
& - 2C\|\Sigma_0\|_2|\Delta_{\Omega,S}^-|_1 \sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}} \\
\geq & \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - C'\|\Sigma_0\|_2|\Delta_\Omega^+|_1 \sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}} - 2C\|\Sigma_0\|_2|\Delta_{\Omega,S}^-|_1 \sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}} \\
= & \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - 2C\|\Sigma_0\|_2|\Delta_{\Omega,S}^- + \Delta_\Omega^+|_1 \sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}}
\end{aligned} \tag{5.51}
$$

where we use the disjoint support of Δ_Ω^+ and $\Delta_{\Omega,S}^-$.

Note that by the properties of the L1 norm $|\Delta_{\Omega,S}^- + \Delta_\Omega^+|_1 \leq \sqrt{s+p}\|\Delta_{\Omega,S}^- + \Delta_\Omega^+\|_F \leq \sqrt{s+p}\|\Delta_\Omega\|_F$, so we have

$$
G(\Delta_\Omega) \geq \frac{2}{9\|\Omega_0\|_2^2}\|\Delta_\Omega\|_F^2 - 2C\|\Sigma_0\|_2\sqrt{s+p}\|\Delta_\Omega\|_F \sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}}.
$$

Since by Assumption A2 $\|\Omega_0\|_2, \|\Sigma_0\|_2$ are upper bounded by constants, observe that $G(\Delta_\Omega)$ is guaranteed to be greater than zero if

$$
\|\Delta_\Omega\|_F^2 \geq c(s+p)\left(\sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}}\right)^2
$$

for some c large enough. Setting $r_{n,\mathbf{p}}^2 = \frac{c'}{M^2}(s+p)\left(\sum_{k=1}^K \sqrt{\frac{\log p}{m_k n}}\right)^2$ in the definition (5.43) of $\mathcal{T}_n$ and recalling Proposition V.13 above implies the following lemma.

Lemma V.15 (Tighter Frobenius norm bound)**.** *Suppose the conditions of Theorem*

V.2. Then with probability at least $1 - 2(K+1)\exp(-c\log p)$,

$$\|\widehat{\Omega} - \Omega_0\|_F^2 \leq C_1(s+p)\left(\sum_{k=1}^{K}\sqrt{\frac{\log p}{m_k n}}\right)^2.$$

This bound is complex, but can be approximated by a more interpretable form. Observe that

$$\left(\sum_{k=1}^{K}\sqrt{\frac{\log p}{m_k n}}\right)^2 \leq K^2 \max_k \frac{\log p}{m_k n}. \tag{5.52}$$

Thus, under event A

$$\|\Delta_\Omega\|_F^2 \leq cK^2(s+p)\frac{\log p}{n\min_k m_k}.$$

This completes the Theorem V.2 Frobenius norm bound on $\widehat{\Omega}$. □

5.14 Proof of Statistical Convergence: Lemmas

5.14.1 Proof of Lemma V.14: Bound on $\langle \Delta_\Omega, S - \Sigma_0\rangle$.

Proof. Using the definition of Δ_Ω (5.41), the projection operator $\text{Proj}_{\widetilde{K}_\mathbf{p}}(\cdot)$, the triangle inequality, and letting $\tau_\Sigma = (K-1)\frac{\text{tr}(S-\Sigma_0)}{p}$, we have that

$$\begin{aligned}
&|\langle \Delta_\Omega, S - \Sigma_0\rangle| = |\langle \Delta_\Omega, \text{Proj}_{\widetilde{K}_\mathbf{p}}(S - \Sigma_0)\rangle| \qquad (5.53)\\
&= \left|\langle \Delta_\Omega^+ + (\Delta_{\Psi,1}^- \oplus \cdots \oplus \Delta_{\Psi,K}^-), (S_1 - \Sigma_0^{(1)}) \oplus \cdots \oplus (S_K - \Sigma_0^{(K)}) - \tau_\Sigma I_p\rangle\right|\\
&\leq \left|\langle \Delta_\Omega^+, (S_1 - \Sigma_0^{(1)}) \oplus \cdots \oplus (S_K - \Sigma_0^{(K)}) - \tau_\Sigma I_p\rangle\right|\\
&+ \left|\langle \Delta_{\Psi,1}^- \oplus \cdots \oplus \Delta_{\Psi,K}^-, (S_1 - \Sigma_0^{(1)}) \oplus \cdots \oplus (S_K - \Sigma_0^{(K)})\rangle\right|,
\end{aligned}$$

where we have used the fact that $\Delta_{\Psi,1}^- \oplus \cdots \oplus \Delta_{\Psi,K}^-$ is zero along the diagonal and thus has zero inner product with I_p.

Expanding the diagonal term out using the definition of the Kronecker sum, and

simplifying the off diagonal term using Lemma V.9, we have

$$
\begin{aligned}
&|\langle \Delta_\Omega, S - \Sigma_0 \rangle| \qquad (5.54)\\
&\leq \left(|\tau_\Sigma \mathrm{tr}(\Delta_\Omega^+)| + \sum_{k=1}^{K} |\langle \Delta_\Omega^+, I_{[d_{1:k-1}]} \otimes (S_k - \Sigma_0^{(k)}) \otimes I_{[d_{k+1:K}]} \rangle| \right) \\
&\quad + \sum_{k=1}^{K} m_k |\langle \Delta_{\Psi,k}^-, S_k - \Sigma_0^{(k)} \rangle| \\
&\leq \left(|\tau_\Sigma| |\Delta_\Omega^+|_1 + \sum_{k=1}^{K} \sum_{i=1}^{p} |[\Delta_\Omega]_{ii}| \cdot |\langle \mathbf{e}_i \mathbf{e}_i^T, I_{[d_{1:k-1}]} \otimes (S_k - \Sigma_0^{(k)}) \otimes I_{[d_{k+1:K}]} \rangle| \right) \\
&\quad + \left(\sum_{k=1}^{K} m_k |\langle \Delta_{\Psi,k}^-, S_k - \Sigma_0^{(k)} \rangle| \right) \\
&\leq \left(|\tau_\Sigma| |\Delta_\Omega^+|_1 + \sum_{k=1}^{K} \sum_{i=1}^{p} |[\Delta_\Omega]_{ii}| \cdot \max_i \left| [S_k - \Sigma_0^{(k)}]_{ii} \right| \right) \\
&\quad + \left(\sum_{k=1}^{K} m_k \sum_{i,j=1}^{d_k} |[\Delta_{\Psi,k}^-]_{ij}| \cdot \max_{ij} \left| [S_k - \Sigma_0^{(k)}]_{ij} \right| \right).
\end{aligned}
$$

Now, by Corollary V.19 we know that for fixed k

$$
\max_{ij} \left| [S_k - \Sigma_0^{(k)}]_{ij} \right| \leq C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}}. \qquad (5.55)
$$

with probability at least $1 - 2\exp(-c' \log p)$. Denoting this event as A_k, we have that $\Pr(A_k) \geq 1 - 2\exp(-c' \log p)$. Note that $E[\mathrm{tr}(S)] = \mathrm{tr}(\Sigma_0)$. Viewing $\frac{1}{p}\mathrm{tr}(\Sigma_0)$ as a 1×1 covariance factor since $\frac{1}{p}\mathrm{tr}(S) = \frac{1}{pn}\sum_{i=1}^{n} \mathrm{vec}(X_i)\mathrm{vec}(X_i)^T$, we can invoke the proof of Corollary V.19 and show that with probability at least $1 - 2\exp(-c' \log p)$

$$
\frac{|\mathrm{tr}(S) - \mathrm{tr}(\Sigma_0)|}{p} = |\tau_\Sigma| \leq C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{pn}}. \qquad (5.56)
$$

We call this event A_0. Let the event that $A_0, A_1, \ldots A_K$ hold be $A = A_0 \cap A_1 \cap \cdots \cap A_K$ By the union bound, we have $\Pr(A) \geq 1 - 2(K+1)\exp(-c \log p)$.

Substituting (5.55) and (5.56) into (5.54), we have that on event A

$$\begin{aligned}
|\langle \Delta_\Omega, S - \Sigma_0 \rangle| &\leq \left(|\Delta_\Omega^+|_1 C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{pn}} + \sum_{k=1}^K |\Delta_\Omega^+|_1 C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} \right) \\
&\quad + \sum_{k=1}^K m_k |\Delta_{\Psi,k}|_1 C \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} \\
&\leq \left(\sum_{k=1}^K |\Delta_\Omega^+|_1 C' \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} \right) + \sum_{k=1}^K m_k |\Delta_{\Psi,k}|_1 C' \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} \\
&= \sum_{k=1}^K \left(|\Delta_\Omega^+|_1 + m_k |\Delta_{\Psi,k}^-|_1 \right) C' \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}}.
\end{aligned}$$

where $C' = 2C$ and we have used $p > m_k$ and recalled the definition of $|\cdot|_1$.

□

5.14.2 Bound on Log Determinant

Recall (5.44)

$$M = \frac{1}{2} \phi_{\max}(\Omega_0) \sqrt{\frac{\min_k m_k}{K+1}}.$$

and let

$$\mathcal{T}_n = \{ \Delta \in \widetilde{\mathcal{K}}_{\mathbf{p}} : \Delta = \Omega - \Omega_0, \Omega, \Omega_0 \in \mathcal{K}_{\mathbf{p}}, \|\Delta\|_F = M r_{n,\mathbf{p}} \} \tag{5.57}$$

We first state the following proposition, modified from a proposition in (*Zhou*, 2014).

Proposition V.16. *Under Assumption A2, for all $\Delta \in \mathcal{T}_n$ for which $r_{n,\mathbf{p}} = o(1)$, we have that*

$$\phi_{\min}(\Omega_0) > 2 M r_{n,\mathbf{p}} = o(1) \tag{5.58}$$

so that $\Omega_0 + v\Delta \succ 0, \forall v \in I \supset [0,1]$, where I is an open interval containing $[0,1]$.

Proof. By Corollary V.12, for all $\Delta \in \mathcal{T}_n$

$$\|\Delta\|_2 \leq \sqrt{\frac{K+1}{\min_k m_k}}\|\Delta\|_F \leq \sqrt{\frac{K+1}{\min_k m_k}} M r_{n,\mathbf{p}} = o(1).$$

The rest of the proposition follows immediately. □

Thus we have that $\log|\Omega_0 + v\Delta|$ is infinitely differentiable on the open interval $I \supset [0,1]$ of v. This allows us to use the Taylor's formula with integral remainder to obtain the following lemma, drawn from (*Zhou*, 2014).

Lemma V.17. *Suppose the conditions of Proposition V.16. Then for all $\Delta \in \mathcal{T}_n$,*

$$\log|\Omega_0 + \Delta| - \log|\Omega_0| \leq \langle \Sigma_0, \Delta \rangle - \frac{2}{9\|\Omega_0\|_2^2}\|\Delta\|_F^2 .$$

Proof. Let us use A as a shorthand for

$$\mathrm{vec}\{\,\Delta\,\}^{\mathrm{T}} \left(\int_0^1 (1-\mathrm{v})(\Theta_0 + \mathrm{v}\Delta)^{-1} \otimes (\Theta_0 + \mathrm{v}\Delta)^{-1} \mathrm{dv} \right) \mathrm{vec}\{\,\Delta\,\},$$

where $\mathrm{vec}\{\,\Delta\,\} \in \mathbb{R}^{\mathrm{p}^2}$ is $\Delta_{p\times p}$ vectorized. Now, the Taylor expansion gives

$$\begin{aligned} \log|\Omega_0 + \Delta| - \log|\Omega_0| &= \frac{d}{dv}\log|\Omega_0 + v\Delta|\Big|_{v=0} \Delta + \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\Omega_0 + v\Delta| dv \\ &= \langle \Sigma_0, \Delta \rangle - A. \end{aligned} \tag{5.59}$$

The last inequality holds because $\nabla_\Omega \log|\Omega| = \Omega^{-1}$ and $\Omega_0^{-1} = \Sigma_0$.

We now bound A, following arguments from (*Zhou et al.*, 2011; *Rothman et al.*,

2008).

$$
\begin{aligned}
A &= \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\Omega_0 + v\Delta|dv \\
&= \mathrm{vec}(\Delta)^T \left(\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1} dv) \right) \mathrm{vec}(\Delta) \\
&\geq \|\Delta\|_F^2 \phi_{\min} \left(\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1} dv \right).
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\phi_{\min} \left(\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1} dv \right) \\
&\geq \int_0^1 (1-v)\phi_{\max}^2((\Omega_0 + v\Delta)^{-1})dv \geq \min_{v\in[0,1]} \phi_{\min}^2((\Omega_0 + v\Delta)^{-1}) \int_0^1 (1-v)dv \\
&= \frac{1}{2} \min_{v\in[0,1]} \frac{1}{\phi_{\max}^2(\Omega_0 + v\Delta)} = \frac{1}{2\max_{v\in[0,1]} \phi_{\max}^2(\Omega_0 + v\Delta)} \\
&\geq \frac{1}{2(\phi_{\max}(\Omega_0) + \|\Delta\|_2)^2}.
\end{aligned}
$$

By Corollary V.12 and using $r_{n,\mathbf{p}} \leq 1$,

$$
\|\Delta\|_2 \leq \sqrt{\frac{K+1}{\min_k m_k}}\|\Delta\|_F \leq \sqrt{\frac{K+1}{\min_k m_k}} M r_{n,\mathbf{p}} \leq \sqrt{\frac{K+1}{\min_k m_k}} M = \frac{1}{2}\phi_{\max}(\Omega_0).
$$

Hence,

$$
\phi_{\min} \left(\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1} dv \right) \geq \frac{2}{9\phi_{\max}^2(\Omega_0)}.
$$

Thus, substituting into (5.59) we have

$$\log|\Omega_0+\Delta|-\log|\Omega_0|\leq\langle\Sigma_0,\Delta\rangle-\frac{2}{9\|\Omega_0\|_2^2}\|\Delta\|_F^2,$$

completing the proof. □

5.14.2.1 Proposition V.18

Proposition V.18. *Let*

$$\mathcal{T}_n=\{\Delta\in\widetilde{\mathcal{K}}_{\mathbf{p}}:\Delta=\Omega-\Omega_0,\Omega,\Omega_0\in\mathcal{K}_{\mathbf{p}},\|\Delta\|_F=Mr_{n,\mathbf{p}}\}.$$

If $G(\Delta)>0$ *for all* $\Delta\in\mathcal{T}_n$*, then* $G(\Delta)>0$ *for all* Δ *in*

$$\mathcal{V}_n=\{\Delta\in\widetilde{\mathcal{K}}_{\mathbf{p}}:\Delta=\Omega-\Omega_0,\Omega,\Omega_0\in\mathcal{K}_{\mathbf{p}},\|\Delta\|_F>Mr_{n,\mathbf{p}}\}.$$

Hence if $G(\Delta)>0$ *for all* $\Delta\in\mathcal{T}_n$*, then* $G(\Delta)>0$ *for all* $\Delta\in\mathcal{T}_n\cup\mathcal{V}_n$*.*

Proof. By contradiction, suppose $G(\Delta')\leq 0$ for some $\Delta'\in\mathcal{V}_n$. Let $\Delta_0=\frac{Mr_{n,\mathbf{p}}}{\|\Delta'\|_F}\Delta'$. Then $\Delta_0=\theta\mathbf{0}+(1-\theta)\Delta'$, where $0<1-\theta=\frac{Mr_{n,\mathbf{p}}}{\|\Delta'\|_F}<1$ by definition of Δ_0. Hence $\Delta_0\in\mathcal{T}_n$ since by the convexity of the positive definite cone $\Omega_0+\Delta_0\succ 0$ because $\Omega_0\succ 0$ and $\Omega_0+\Delta'\succ 0$. By the convexity of $G(\Delta)$, we have that $G(\Delta_0)\leq\theta G(\mathbf{0})+(1-\theta)G(\Delta')\leq 0$, contradicting our assumption that $G(\Delta_0)>0$ for $\Delta_0\in\mathcal{T}_n$. □

5.15 Subgaussian Concentration (Lemma V.19)

Lemma V.19 (Subgaussian Concentration)**.** *Suppose that* $\log p \ll m_k n$ *for all* k*. Then, with probability at least* $1 - 2\exp(-c' \log p)$*,*

$$|\langle \Delta, S_k - \Sigma_0^{(k)} \rangle| \leq C|\Delta|_1 \|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}}$$

for all $\Delta \in \mathbb{R}^{d_k \times d_k}$*, where* c' *is a constant depending on* C *given in the proof.*

Proof. We use a K-way generalization of the invertible Pitsianis-Van Loan type (*Loan and Pitsianis*, 1993) rearrangement operator $\mathcal{R}_k(\cdot)$, which maps $p \times p$ matrices to $d_k^2 \times m_k^2$ matrices. For a matrix $M \in \mathbb{R}^{p \times p}$ we set

$$\mathcal{R}_k(M) = [\ \mathbf{m}_1 \quad \dots \quad \mathbf{m}_{m_k^2}\ ], \tag{5.60}$$
$$\mathbf{m}_{(i-1)m_k + j} = \mathrm{vec}(M(i, j|k)),$$

where we use the $M(i, j|k) \in \mathbb{R}^{d_k \times d_k}$ subblock notation from the introduction.

Consider the inner product $\langle \Delta, S_k - \Sigma_0^{(k)} \rangle$, where Δ is an arbitrary $d_k \times d_k$ matrix. Let

$$\mathbf{h} = \mathrm{vec}(\Delta), \qquad \mathbf{f} = \mathrm{vec}(I_{m_k \times m_k}).$$

By the definition of the factor covariances S_k and the rearrangement operator $\mathcal{R}_k$, it can be seen that

$$\mathrm{vec}(S_k) = \frac{1}{m_k} \mathcal{R}_k(S)\mathbf{f},$$

and that similarly by the definition of the factor covariances $\Sigma_0^{(k)}$

$$\mathrm{vec}(\Sigma_0^{(k)}) = \frac{1}{m_k} \mathcal{R}_k(\Sigma_0)\mathbf{f}.$$

Hence,

$$\begin{aligned}\langle \Delta, S_k - \Sigma_0^{(k)} \rangle &= \frac{1}{m_k} \langle \mathrm{vec}(\Delta), \mathcal{R}_k(S - \Sigma_0)\mathbf{f} \rangle \\ &= \frac{1}{m_k} \mathbf{h}^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f} \\ &= \frac{1}{m_k} \sum_{i=1}^{d_k^2} h_i \mathbf{e}_i^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f}\end{aligned} \tag{5.61}$$

by the linearity of the rearrangement operator and definition of the inner product.

Note that $\mathbf{e}_i$ and $\mathbf{f}$ are deterministic and fixed for fixed i, k. Hence, we can apply Lemma V.20 (Appendix 5.16) and take a union bound over $i = 1, \ldots, d_k^2$. By Lemma V.20,

$$\Pr\left(\left|\mathbf{e}_i^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f}\right| \geq \epsilon \sqrt{m_k} \|\Sigma_0\|_2\right) \leq 2 \exp\left(-c \frac{\epsilon^2 n}{K^4}\right)$$

for $0 \leq \frac{\epsilon}{\sqrt{m_k}} \leq \frac{1}{2}$. Taking the union bound over all i, we obtain

$$\begin{aligned}\Pr\left(\max_i |\mathbf{e}_i^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f}| \geq \epsilon \|\Sigma_0\|_2 \sqrt{m_k}\right) &\leq 2 d_k^2 \exp\left(-c \frac{\epsilon^2 n}{K^4}\right) \\ &\leq 2 \exp\left(2 \log d_k - c \frac{\epsilon^2 n}{K^4}\right).\end{aligned}$$

Setting $\epsilon = C\sqrt{\frac{\log p}{n}}$ for large enough C, with probability at least $1 - 2\exp(-c' \log p)$ we have

$$\max_i |\mathbf{e}_i^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f}| \leq C \|\Sigma_0\|_2 \sqrt{m_k} \sqrt{\frac{\log p}{n}}$$

where we assume $\log p \leq \frac{nm_k}{4C^2}$ and let $c' = \frac{cC^2}{K^4} - 2$. Hence, by (5.61)

$$\begin{aligned}
|\langle \Delta, S_k - \Sigma_0^{(k)} \rangle| &= \frac{1}{m_k} \left| \sum_{i=1}^{d_k^2} h_i \mathbf{e}_i^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f} \right| \\
&\leq \frac{1}{m_k} \sum_{i=1}^{d_k^2} |h_i \mathbf{e}_i^T \mathcal{R}_k(S - \Sigma_0)\mathbf{f}| \\
&\leq C\|\Sigma_0\|_2 \frac{1}{\sqrt{m_k}} \sqrt{\frac{\log p}{n}} \sum_{i=1}^{d_k^2} h_i \\
&= C\|\Sigma_0\|_2 \sqrt{\frac{\log p}{m_k n}} |\Delta|_1
\end{aligned}$$

with probability at least $1 - 2\exp(-c' \log p)$. The first inequality follows from the triangle inequality and the last from the definition of $\mathbf{h} = \text{vec}(\Delta)$ and $|\cdot|_1$.

□

5.16 Concentration Bound

Lemma V.20 (Concentration of Measure)**.** *Let* $\mathbf{u} \in S^{d_k^2 - 1}$ *and* $\mathbf{f} = \text{vec}(I_{m_k})$*. Assume that* $\mathbf{x}_t = \Sigma_0^{1/2} \mathbf{z}_t$ *where* $\mathbf{z}_t$ *has independent entries* $z_{t,f}$ *such that* $Ez_{t,f} = 0$*,* $Ez_{t,f}^2 = 1$*, and* $\|z_{t,f}\|_{\psi_2} \leq K$*. Let* $\Delta_n = S - \Sigma_0$*. Then for all* $0 \leq \frac{\epsilon}{\sqrt{m_1}} < \frac{1}{2}$*:*

$$\Pr(|\mathbf{u}^T \mathcal{R}_k(\Delta_n)\mathbf{f}| \geq \epsilon \sqrt{m_k} \|\Sigma_0\|_2) \leq 2\exp\left(-c \frac{\epsilon^2 n m_k}{K^4}\right)$$

where c *is an absolute constant.*

Proof. We prove the case for $k = 1$. The proofs for the remaining k follow similarly.

By the definition (5.60) of the permutation operator $\mathcal{R}_1$ and letting $\mathbf{x}_t(i) =$

$[x_{t,(i-1)m_1+1}, \ldots, x_{t,im_1}]$,

$$\mathcal{R}_1(S) = \frac{1}{n}\sum_{t=1}^{n} \begin{bmatrix} \text{vec}(\mathbf{x}_t(1)\mathbf{x}_t(1)^T)^T \\ \text{vec}(\mathbf{x}_t(1)\mathbf{x}_t(2)^T)^T \\ \vdots \\ \text{vec}(\mathbf{x}_t(d_1)\mathbf{x}_t(d_1)^T)^T \end{bmatrix} \tag{5.62}$$

Hence,

$$\mathbf{u}^T\mathcal{R}_1(S)\mathbf{f} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t^T(U \otimes I_{m_k})\mathbf{x}_t = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t^T M \mathbf{z}_t \tag{5.63}$$

where $M = \Sigma_0^{1/2}(U \otimes I_{m_k})\Sigma_0^{1/2}$, $U = \text{vec}_{d_1,d_1}^{-1}(\mathbf{u})$.

Thus, by the Hanson-Wright inequality,

$$\begin{aligned} &\Pr(|\mathbf{u}^T\mathcal{R}_1(S)\mathbf{f} - E[\mathbf{u}^T\mathcal{R}_1(S)\mathbf{f}]| \geq \tau) \\ &\leq 2\exp\left[-c\min\left(\frac{\tau^2N^2}{K^4n\|M\|_F^2}, \frac{\tau n}{K^2\|M\|_2}\right)\right] \\ &\leq 2\exp\left[-c\min\left(\frac{\tau^2N}{K^4m_1\|\Sigma_0\|_2^2}, \frac{\tau n}{K^2\|\Sigma_0\|_2}\right)\right] \end{aligned} \tag{5.64}$$

since $\|U \otimes I_{m_1}\|_2 = \|U\|_2 \leq 1$ and $\|U \otimes I_{m_1}\|_F^2 = \|U\|_F^2\|I_{m_1}\|_F^2 = m_1$. Substituting $\epsilon = \frac{\tau}{\sqrt{m_1}\|\Sigma_0\|_2}$

$$\Pr(|\mathbf{u}^T\mathcal{R}_1(\Delta_n)\mathbf{f}| \geq \epsilon\sqrt{m_1}\|\Sigma_0\|_2) \leq 2\exp\left(-c\frac{\epsilon^2 n}{K^4}\right) \tag{5.65}$$

for all $\frac{\epsilon^2 n}{K^4} \leq \frac{\epsilon n\sqrt{m_1}}{K^2}$, i.e. $\epsilon \leq K^2\sqrt{m_1} \leq \frac{\sqrt{m_1}}{2}$, since $K^2 > \frac{1}{2}$ by definition.

□

5.17 Factorwise and Spectral Norm Bounds - Theorem V.3

5.17.1 Factor-wise bound

Proof. From the proof of Theorem V.2, we know that under event A,

$$\|\Delta_\Omega\|_F^2 \leq cK^2(s+p)\frac{\log p}{n \min_k m_k}. \tag{5.66}$$

Furthermore, since the identifiable parameterizations of $\widehat{\Omega}, \Omega_0$ are of the form (Lemma V.10)

$$\widehat{\Omega} = \widehat{\tau} I_p + (\widetilde{\Psi}_1 \oplus \cdots \oplus \widetilde{\Psi}_K)$$
$$\Omega_0 = \tau_0 I_p + (\widetilde{\Psi}_{0,1} \oplus \cdots \oplus \widetilde{\Psi}_{0,K}),$$

we have that the identifiable parameterization of Δ_Ω is

$$\Delta_\Omega = \tau_\Delta I_p + (\widetilde{\Delta}_1 \oplus \cdots \oplus \widetilde{\Delta}_K), \tag{5.67}$$

where $\tau_\Delta = \widehat{\tau} - \tau_0$, $\widetilde{\Delta}_k = \widetilde{\Psi}_k - \widetilde{\Psi}_{0,k}$. Observe that $\mathrm{tr}(\widetilde{\Delta}_k) = \mathrm{tr}(\widetilde{\Psi}_k) - \mathrm{tr}(\widetilde{\Psi}_{0,k}) = 0$.

By Lemma V.10 then,

$$\|\Delta_\Omega\|_F^2 = p\tau_\Delta^2 + \sum_{k=1}^K m_k \|\widetilde{\Delta}_k\|_F^2.$$

Thus, the estimation error on the underlying parameters is bounded by (5.66)

$$p\tau_\Delta^2 + \sum_{k=1}^K m_k \|\widetilde{\Delta}_k\|_F^2 \leq cK^2(s+p)\frac{\log p}{n \min_k m_k},$$

or, dividing both sides by p

$$\tau_\Delta^2 + \sum_{k=1}^{K} \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k} \leq cK^2 \frac{s+p}{p} \frac{\log p}{n \min_k m_k} \tag{5.68}$$
$$= cK^2 \left(\frac{s}{p} + 1\right) \frac{\log p}{n \min_k m_k}.$$

Recall that $s = \sum_{k=1}^{K} m_k s_k$, so $\frac{s}{p} = \sum_{k=1}^{K} \frac{s_k}{d_k}$. Substituting into (5.68)

$$\tau_\Delta^2 + \sum_{k=1}^{K} \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k} \leq cK^2 \left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right) \frac{\log p}{n \min_k m_k}. \tag{5.69}$$

From this, it can be seen that the bound converges as the m_k increase with constant K. To put the bound in the form of the theorem, note that since $\tau_\Delta I_p + (\widetilde{\Delta}_1^+ \oplus \cdots \oplus \widetilde{\Delta}_K^+)$

$$\begin{aligned}
\frac{\|\Delta_\Omega^+\|_2^2}{\max_k d_k} &\leq \frac{\left(\tau_\Delta + \sum_{k=1}^{K} \|\widetilde{\Delta}_k^+\|_2\right)^2}{\max_k d_k} \\
&\leq \frac{K+1}{\max_k d_k} \left(\tau_\Delta^2 + \sum_{k=1}^{K} \|\widetilde{\Delta}_k^+\|_2^2\right) \\
&\leq (K+1) \left(\tau_\Delta^2 + \sum_{k=1}^{K} \frac{\|\widetilde{\Delta}_k^+\|_F^2}{d_k}\right).
\end{aligned}$$

To confirm low sample convergence, we must check that the condition $r_{n,\mathbf{p}} = o(1)$ remains satisfied, i.e. that $r_{n,\mathbf{p}}^2 = \frac{c'}{M^2}(s+p)\left(\sum_{k=1}^{K} \sqrt{\frac{\log p}{m_k n}}\right)^2 \leq 1$. Equivalently, we must show

$$\frac{K+1}{\min_k m_k}(s+p)\left(\sum_{k=1}^{K} \sqrt{\frac{\log p}{m_k n}}\right)^2 = o(1)$$
$$(K+1)K^2 \left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right) \left(\max_k d_k\right) \frac{\log p}{n \min_k m_k} = o(1), \tag{5.70}$$

where we have substituted in $\frac{s}{p} = \sum_{k=1}^{K} \frac{s_k}{d_k}$ and (5.52). The relation (5.70) will hold

whenever the bound (5.68) implies convergence of

$$\left(\max_k d_k\right)\left(\tau_\Delta^2 + \sum_{k=1}^{K} \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k}\right).$$

For $n = 1$ and constant K, convergence is implied whenever $s_k \leq O(d_k)$ and $\min_k m_k > d_k$. Thus the bound in (5.68) implies low sample convergence of the identifiable parameter estimates in this regime, completing the proof of the theorem. □

5.17.2 Spectral norm bound

Proof. The factor-wise bound immediately implies the bound on the spectral norm $\|\Delta_\Omega\|_2$ of the error under event A. We recall the identifiable representation (5.67)

$$\Delta_\Omega = \tau_\Delta I_p + (\widetilde{\Delta}_1 \oplus \cdots \oplus \widetilde{\Delta}_K).$$

By Property 3a in Appendix A and the fact that the spectral norm is upper bounded by the Frobenius norm,

$$\begin{aligned}
\|\Delta_\Omega\|_2 &\leq |\tau_\Delta| + \sum_{k=1}^{K} \|\widetilde{\Delta}_k\|_2 \leq |\tau_\Delta| + \sum_{k=1}^{K} \|\widetilde{\Delta}_k\|_F \\
&\leq \sqrt{K+1}\sqrt{\tau_\Delta^2 + \sum_{k=1}^{K} \|\widetilde{\Delta}_k\|_F^2} \\
&\leq \sqrt{K+1}\sqrt{\max_k d_k}\sqrt{\tau_\Delta^2 + \sum_{k=1}^{K} \frac{\|\widetilde{\Delta}_k\|_F^2}{d_k}} \\
&\leq cK\sqrt{K+1}\sqrt{(\max_k d_k)\left(1 + \sum_{k=1}^{K} \frac{s_k}{d_k}\right)}\sqrt{\frac{\log p}{n \min_k m_k}},
\end{aligned}$$

where in the second line, we have used the fact that for a_k elements of $\mathbf{a} \in \mathbb{R}^K$ the norm relation $\|\mathbf{a}\|_1 \leq \sqrt{K}\|\mathbf{a}\|_2$ implies $(\sum_{k=1}^{K} |a_k|) \leq \sqrt{K}\sqrt{\sum_{k=1}^{K} a_k^2}$. □

5.18 Proof of Lemma V.5

Proof. Expanding out the Kronecker sums, for

$$\Omega_t = \Psi_1^t \oplus \cdots \oplus \Psi_K^t, \qquad \Omega = \Psi_1 \oplus \cdots \oplus \Psi_K,$$

the Frobenius norm term in the objective

$$\Omega_{t+1} \in \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta \left(\widetilde{S} - G \right) \right) \right\|_F^2 + \zeta \sum_{k=1}^{K} m_k \rho_k |\Psi_k^-|_1 \right\}$$

can be decomposed using disjoint support into a sum of a diagonal portion and a factor-wise sum of the off diagonal portions. This holds by Property 2 in Appendix A which states the off diagonal factors Ψ_k^- have disjoint support in Ω. Thus,

$$\begin{aligned}
&\left\| \Omega - \left(\Omega_t - \zeta \left((\widetilde{S}_1 - G_1) \oplus \cdots \oplus (\widetilde{S}_K - G_K) \right) \right) \right\|_F^2 \\
&= \left\| \left(\Psi_1 - (\Psi_1^t - \zeta(\widetilde{S}_1 - G_1)) \right) \oplus \cdots \oplus \left(\Psi_K - (\Psi_K^t - \zeta(\widetilde{S}_K - G_K)) \right) \right\|_F^2 \\
&= \left\| \Omega^+ - \left(\Omega_t^+ - \zeta \left(\widetilde{S}^+ - G^+ \right) \right) \right\|_F^2 + \sum_{k=1}^{K} m_k \left\| \left(\Psi_1 - (\Psi_1^t - \zeta(\widetilde{S}_1 - G_1)) \right)^- \right\|_F^2 .
\end{aligned}$$

Substituting into the objective (5.19), we obtain

$$\begin{aligned}
\Omega_{t+1} \in \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \Bigg\{ & \frac{1}{2} \left\| \Omega^+ - \left(\Omega_t^+ - \zeta \left(\widetilde{S}^+ - G^+ \right) \right) \right\|_F^2 \\
& + \sum_{k=1}^{K} m_k \left(\frac{1}{2} \left\| \left(\Psi_k - (\Psi_k^t - \zeta(\widetilde{S}_k - G_k)) \right)^- \right\|_F^2 + \zeta \rho_k |\Psi_k^-|_1 \right) \Bigg\} .
\end{aligned}$$

This objective is decomposable into a sum of terms each involving either the diagonal Ω^+ or one of the off diagonal factors Ψ_k^-. Thus, we can solve for each portion of Ω independently, giving

$$[\Psi_k^{t+1}]^- = \arg\min_{\Psi_k^-} \frac{1}{2} \left\| \Psi_k^- - ([\Psi_k^t]^- - \zeta(\widetilde{S}_k^- - G_k^-)) \right\|_F^2 + \zeta\rho_k|\Psi_k^-|_1$$

$$\Omega_{t+1}^+ = \arg\min_{\Omega^+} \frac{1}{2} \left\| \Omega^+ - \left(\Omega_t^+ - \zeta \left(\widetilde{S}^+ - G^+ \right) \right) \right\|_F^2 .$$

completing the proof. □

5.19 Convergence Rate

5.19.1 Contraction factor (Theorem V.6)

Proof. Recall that the TG-ISTA update is of the form (5.19)

$$\begin{aligned}
\Omega_{t+1} &= \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta_t \left(\widetilde{S} - G \right) \right) \right\|_F^2 + \zeta_t \sum_{k=1}^K m_k \rho_k |\Psi_k^-|_1 \right\} \\
&= \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left\{ \frac{1}{2} \left\| \Omega - \left(\Omega_t - \zeta_t \left(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S - \Omega_t^{-1} \right) \right) \right\|_F^2 + \zeta_t \sum_{k=1}^K m_k \rho_k |\Psi_k^-|_1 \right\} \\
&= \eta_{\zeta_t}(\Omega_t - \zeta_t \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S - \Omega_t^{-1})),
\end{aligned}$$

where we let $\eta_\zeta(M) = \arg\min_{\Omega \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \left\{ \frac{1}{2} \|\Omega - M\|_F^2 + \zeta \sum_{k=1}^K m_k \rho_k |\Psi_k^-|_1 \right\}$ for $M \in \widetilde{\mathcal{K}}_{\mathbf{p}}$. Since $\sum_{k=1}^K m_k \rho_k |\Psi_k^-|_1$ is a convex function on $\widetilde{\mathcal{K}}_{\mathbf{p}}$, and since $\widetilde{\mathcal{K}}_{\mathbf{p}}$ is a linear subspace, $\eta_\zeta(\cdot)$ is a proximal operator by definition.

By convexity in $\widetilde{\mathcal{K}}_{\mathbf{p}}$ and Theorem V.1, the optimal point Ω_ρ^* is a fixed point of the ISTA iteration (*Combettes and Wajs* (2005), Prop 3.1). Thus,

$$\Omega_\rho^* = \eta_{\zeta_t}(\Omega_\rho^* - \zeta_t \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S - (\Omega_\rho^*)^{-1}).$$

Since proximal operators are not expansive (*Combettes and Wajs*, 2005), we have

$$\begin{aligned}\|\Omega_{t+1} - \Omega_\rho^*\|_F &= \|\eta_{\zeta_t}(\Omega_t - \zeta_t \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(S - \Omega_t^{-1})) - \eta_{\zeta_t}(\Omega_\rho^* - \zeta_t \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(S - (\Omega_\rho^*)^{-1}))\|_F \\ &\leq \|\Omega_t - \zeta_t \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(S - \Omega_t^{-1}) - (\Omega_\rho^* - \zeta_t \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(S - (\Omega_\rho^*)^{-1}))\|_F \\ &= \|\Omega_t + \zeta_t \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(\Omega_t^{-1}) - (\Omega_\rho^* + \zeta_t \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}((\Omega_\rho^*)^{-1}))\|_F.\end{aligned}$$

For $\gamma > 0$ define $h_\gamma : \mathcal{K}_\mathbf{p} \to \mathcal{K}_\mathbf{p}$ by

$$h_\gamma(\Omega) = \text{vec}(\Omega) + \text{vec}(\gamma \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(\Omega^{-1})).$$

Since $\partial\Omega^{-1}/\partial\Omega = -\Omega^{-1} \otimes \Omega^{-1}$,

$$\frac{\partial \text{Proj}_{\widetilde{\mathcal{K}}_\mathbf{p}}(\Omega^{-1})}{\partial \Omega} = -P(\Omega^{-1} \otimes \Omega^{-1})P^T$$

where P is the projection matrix that projects $\text{vec}(\Omega)$ onto the vectorized subspace $\widetilde{\mathcal{K}}_\mathbf{p}$. Thus, we have the Jacobian (valid for all $\Omega \in \mathcal{K}_\mathbf{p}$)

$$J_{h_\gamma}(\Omega) = PP^T - \gamma P(\Omega^{-1} \otimes \Omega^{-1})P^T.$$

Recall that if $h : U \subset \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable mapping, then if $x, y \in U$ and U is convex, then if $J_h(\cdot)$ is the Jacobian of h,

$$\|h(x) - h(y)\| \leq \sup_{c \in [0,1]} \|J_h(cx + (1-c)y)\| \|x - y\|.$$

Thus, letting $Z_{t,c} = \text{vec}(c\Omega_t + (1-c)\Omega_\rho^*)$, for $c \in [0,1]$ we have

$$\|h_{\zeta_t}(x) - h_{\zeta_t}(y)\| \leq \sup_{c \in [0,1]} \|PP^T - \zeta_t P(Z_{t,c}^{-1} \otimes Z_{t,c}^{-1})P^T\| \|\Omega_t - \Omega_\rho^*\|_F.$$

By Weyl's inequality,

$$\lambda_{\max}(Z_{t,c}) \leq \max\{\|\Omega_t\|, \|\Omega_\rho^*\|\}$$

and

$$\lambda_{\min}(Z_{t,c}) \geq \min\{\lambda_{\min}(\Omega_t), \lambda_{\min}(\Omega_\rho^*)\}.$$

Furthermore, note that for any Y and projection matrix P

$$\lambda_{\max}(PYP^T) \leq \lambda_{\max}(Y).$$

We then have

$$\|PP^T - \zeta_t P(Z_{t,c}^{-1} \otimes Z_{t,c}^{-1})P^T\| \leq \|I_{p^2} - \zeta_t Z_{t,c}^{-1} \otimes Z_{t,c}^{-1}\| \leq \max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\},$$

where the latter inequality comes from (*Rolfs et al.*, 2012). Thus,

$$\begin{aligned}\|\Omega_{t+1} - \Omega_\rho^*\|_F &\leq s(\zeta_t)\|\Omega_t - \Omega_\rho^*\|_F \\ s(\zeta) &= \max\left\{\left|1 - \frac{\zeta}{b^2}\right|, \left|1 - \frac{\zeta}{a^2}\right|\right\}\end{aligned}$$

as desired. Algorithm 5 will then converge if $s(\zeta_t) \in (0, 1)$ for all t. The minimum of $s(\zeta)$ occurs at $\zeta = \frac{2}{a^{-2}+b^{-2}}$, completing the proof. □

5.20 Eigenvalue bound on iterates (Theorem V.7)

Proof. We first prove the following properties of the Kronecker sum projection operator.

Lemma V.21. *For any $A \in \mathbb{R}^{p\times p}$ and orthogonal matrices $U_k \in \mathbb{R}^{d_k \times d_k}$, let $U = U_1 \otimes \cdots \otimes U_K \in \widetilde{\mathcal{K}}_{\mathbf{p}}$. Then*

$$\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = U\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(U^T AU)U^T.$$

Furthermore, if the eigendecomposition of A is of the form $A = (U_1 \otimes \cdots \otimes U_K)\Lambda(U_1 \otimes \cdots \otimes U_K)^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, we have

$$\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = U\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Lambda)U^T$$

and

$$\lambda_{\min}(A) \leq \lambda_{\min}(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A)) \leq \lambda_{\max}(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A)) \leq \lambda_{\max}(A).$$

Proof. Recall

$$\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) = \arg\min_{M \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \|A - B\|_F^2 = \arg\min_{B \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \|U^T AU - U^T BU\|_F^2$$

since $U^T AU = \Lambda$ and the Frobenius norm is unitarily invariant. Now, note that for any matrix $B = B_1 \oplus \cdots \oplus B_K \in \widetilde{\mathcal{K}}_{\mathbf{p}}$,

$$\begin{aligned}
&(U_1 \otimes \cdots \otimes U_K)^T B(U_1 \otimes \cdots \otimes U_K) \\
&= \sum_{k=1}^{K} (U_1 \otimes \cdots \otimes U_K)^T (I_{[d_{1:k-1}]} \otimes B_k \otimes I_{[d_{k+1}:K]})(U_1 \otimes \cdots \otimes U_K) \\
&= \sum_{k=1}^{K} I_{[d_{1:k-1}]} \otimes U_k^T B_k U_k \otimes I_{[d_{k+1}:K]} \\
&= (U_1^T B_1 U_1) \oplus \cdots \oplus (U_K^T B_K U_K) \\
&\in \widetilde{\mathcal{K}}_{\mathbf{p}},
\end{aligned}$$

since $U_k^T I_{d_k} U_k = I_{d_k}$. Since $U^T BU \in \widetilde{\mathcal{K}}_{\mathbf{p}}$, the constraint $B \in \widetilde{\mathcal{K}}_{\mathbf{p}}$ can be moved to $C = U^T BU$, giving

$$\begin{aligned}
\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A) &= U(\arg\min_{C \in \widetilde{\mathcal{K}}_{\mathbf{p}}} \|U^T AU - C\|_F^2)U^T \\
&= U(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(U^T AU))U^T.
\end{aligned}$$

If $A = (U_1 \otimes \cdots \otimes U_K)\Lambda(U_1 \otimes \cdots \otimes U_K)^T$, then $U^T A U = \Lambda$, completing the first part of the proof. As shown in the Lemma V.8, $\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Lambda)$ is a diagonal matrix whose entries are weighted averages of the diagonal elements λ_i. Hence

$$\min_i \lambda_i \leq \min_i [\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Lambda)]_{ii} \leq \max_i [\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Lambda)]_{ii} \leq \max_i \lambda_i.$$

Since $\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Lambda)$ gives the eigenvalues of $\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(A)$ by the orthogonality of U, this completes the proof. □

Lemma V.22. *Let $0 < a < b$ be given positive constants and let $\zeta_t > 0$. Assume $aI \preceq \Omega_t \preceq bI$. Then for*

$$\Omega_{t+1/2} := \Omega_t - \zeta_t(\text{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S - \Omega_t^{-1}))$$

we have

$$\lambda_{\min}(\Omega_{t+1/2}) \geq \begin{cases} 2\sqrt{\zeta_t} - \zeta_t \lambda_{\max}(S) & \text{if } a \leq \sqrt{\zeta_t} \leq b \\ \min\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) - \zeta_t \lambda_{\max}(S) & \text{o.w.} \end{cases}$$

and

$$\lambda_{\max}(\Omega_{t+1/2}) \leq \max\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) - \zeta_t \lambda_{\min}(S).$$

Proof. Let $U\Gamma U^T = \Omega_t$ be the eigendecomposition of Ω_t, where $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_p)$. Then all $b \geq \gamma_i \geq a > 0$. Since $\Omega_t \in \widetilde{\mathcal{K}}_{\mathbf{p}}$, by the eigendecomposition property in

Appendix A we have $U = U_1 \otimes \cdots \otimes U_K$ and $\Gamma \in \widetilde{\mathcal{K}}_{\mathbf{p}}$, letting us apply Lemma V.21:

$$\begin{aligned}
\Omega_{t+1/2} &= \Omega_t - \zeta_t(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S) - \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Omega_t^{-1})) \\
&= U\Gamma U^T - \zeta_t(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S) - U\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Gamma^{-1})U^T) \\
&= U\left(\Gamma - \zeta_t(U^T\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(S)U - \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Gamma^{-1}))\right)U^T \\
&= U\left(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Gamma) - \zeta_t\left(\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(U^TSU) - \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\Gamma^{-1})\right)\right)U^T \\
&= \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}\left(U(\Gamma + \zeta\Gamma^{-1} - \zeta_t(U^TSU))U^T\right) \\
&= \mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\widetilde{\Omega}_{t+1/2}),
\end{aligned}$$

where we set $\widetilde{\Omega}_{t+1/2} = U(\Gamma + \zeta\Gamma^{-1} - \zeta_t(U^TSU))U^T$ and recall the linearity of the projection operator $\mathrm{Proj}_{\widetilde{\mathcal{K}}_{\mathbf{p}}}(\cdot)$ (Lemma V.8). By Weyl's inequality, for

$$\gamma_1 + \frac{\zeta_t}{\gamma_1} - \zeta_t\lambda_{\max}(S) \leq \lambda_{\min}(\widetilde{\Omega}_{t+1/2}) \leq \lambda_{\max}(\widetilde{\Omega}_{t+1/2}) \leq \gamma_p + \frac{\zeta_t}{\gamma_p} - \zeta_t\lambda_{\min}(S).$$

By Lemma V.21,

$$\gamma_1 + \frac{\zeta_t}{\gamma_1} - \zeta_t\lambda_{\max}(S) \leq \lambda_{\min}(\Omega_{t+1/2}) \leq \lambda_{\max}(\Omega_{t+1/2}) \leq \gamma_p + \frac{\zeta_t}{\gamma_p} - \zeta_t\lambda_{\min}(S).$$

Note that the only extremum of the function $f(x) = x + \frac{\zeta_t}{x}$ over $a \leq x \leq b$ is a global minimum at $x = \sqrt{\zeta_t}$. Hence

$$\begin{aligned}
\inf_{a\leq x\leq b} x + \frac{\zeta_t}{x} &= \begin{cases} 2\sqrt{\zeta_t} & \text{if } a \leq \sqrt{\zeta_t} \leq b \\ \min\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) & \text{o.w.} \end{cases} \\
\sup_{a\leq x\leq b} x + \frac{\zeta_t}{x} &= \max\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right).
\end{aligned}$$

By our assumption, $a \leq \gamma_1 \leq b$. Thus

$$\lambda_{\min}(\Omega_{t+1/2}) \geq \begin{cases} 2\sqrt{\zeta_t} - \zeta_t \lambda_{\max}(S) & \text{if } a \leq \sqrt{\zeta_t} \leq b \\ \min\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) - \zeta_t \lambda_{\max}(S) & \text{o.w.} \end{cases}$$

$$\lambda_{\max}(\Omega_{t+1/2}) \leq \max\left(a + \frac{\zeta_t}{a}, b + \frac{\zeta_t}{b}\right) - \zeta_t \lambda_{\min}(S).$$

as desired.

□

Lemma V.23. *For $A \in \mathcal{K}_{\mathbf{p}}$ and $\epsilon = [\epsilon_1, \ldots, \epsilon_K]$:*

$$\lambda_{\min}(A) - \sum_{k=1}^{K} d_k \epsilon_k \leq \lambda_{\min}(\eta_\epsilon(A))$$

Proof. Recall that $A \in \mathcal{K}_{\mathbf{p}}$ can be written as

$$A = A^{(1)} \oplus \cdots \oplus A^{(K)},$$

and the Kronecker sum soft thresholding operator can be decomposed as

$$\eta_\epsilon(A) = \eta_{1,\epsilon_1}(A^{(1)}) \oplus \cdots \oplus \eta_{K,\epsilon_K}(A^{(K)}).$$

By the properties of the Kronecker sum, we thus have that

$$\lambda_{\min}(\eta_\epsilon(A)) = \sum_{k=1}^{K} \lambda_{\min}(\eta_{k,\epsilon_k}(A^{(k)})).$$

Via Weyl's inequality and the proof of Lemma 6 in (*Rolfs et al.*, 2012), $\lambda_{\min}(\eta_{k,\epsilon_k}(A^{(k)})) \geq \lambda_{\min}(A^{(k)}) - d_k \epsilon_k$. Hence,

$$\lambda_{\min}(\eta_\epsilon(A)) \geq \sum_{k=1}^{K} \lambda_{\min}(A^{(k)}) - \sum_{k=1}^{K} d_k \epsilon_k = \lambda_{\min}(A) - \sum_{k=1}^{K} d_k \epsilon_k$$

□

Lemma V.24. *Let $\rho = [\rho_1, \ldots, \rho_K]$ with all $\rho_i > 0$. Define*

$$\chi = \sum_{k=1}^{K} d_k \rho_k$$

and let $\alpha = \frac{1}{\|S\|_2 + \chi} < b'$. Assume $\alpha I \preceq \Omega_{t+1}$. Then $\alpha I \preceq \Omega_{t+1}$ for every $0 < \zeta_t < \alpha^2$.

Proof. Since $\zeta_t < \alpha^2$, $\sqrt{\zeta_t} \notin [\alpha, b']$, and $\min\left(\alpha + \frac{\zeta_t}{\alpha}, b' + \frac{\zeta_t}{b'}\right) = \alpha + \frac{\zeta_t}{\alpha}$. Lemma V.22 then implies that

$$\begin{aligned}\lambda_{\min}(\Omega_{t+1/2}) &\geq \min\left(\alpha + \frac{\zeta_t}{\alpha}, b' + \frac{\zeta_t}{b'}\right) - \zeta_t \lambda_{\max}(S) \\ &= \alpha + \frac{\zeta_t}{\alpha} - \zeta_t \lambda_{\max}(S).\end{aligned}$$

By Lemma V.23,

$$\begin{aligned}\lambda_{\min}(\Omega_{t+1}) &= \lambda_{\min}\left(\eta_{\zeta_t \rho}(\Omega_{t+1/2})\right) \\ &\geq \lambda_{\min}(\Omega_{t+1/2}) - \zeta_t \chi \\ &\geq \alpha + \frac{\zeta_t}{\alpha} - \zeta_t \lambda_{\max}(S) - \zeta_t \chi.\end{aligned}$$

Hence, since $\zeta_t > 0$, $\lambda_{\min}(\Omega_{t+1}) \geq \alpha$ whenever

$$\begin{aligned}\zeta_t \left(\frac{1}{\alpha} - \lambda_{\max}(S) - \chi\right) &\geq 0 \\ \frac{1}{\alpha} - \lambda_{\max}(S) - \chi &\geq 0 \\ \alpha &\leq \frac{1}{\|S\|_2 + \chi}.\end{aligned}$$

□

Lemma V.25. *Let χ be as in Lemma V.24 and let $\alpha = \frac{1}{\|S\|_2 + \chi}$. Let $\zeta_t \leq \alpha^2$ for all*

t. We then have $\Omega_t \preceq b'I$ *for all* t *when* $b' = \|\Omega_\rho^*\|_2 + \|\Omega_0 - \Omega_\rho^*\|_F$.

Proof. By Lemma V.24, $\alpha I \preceq \Omega_t$ for every t. Since $\Omega_t \to \Omega_\rho^*$, by strong convexity $\alpha I \preceq \Omega_\rho^*$. Hence $a = \min\{\lambda_{\min}(\Omega_t), \lambda_{\min}(\Omega_\rho^*)\} \geq \alpha$. For $b > a$ and $\zeta_t \leq \alpha^2$,

$$\max\left\{\left|1 - \frac{\zeta_t}{b^2}\right|, \left|1 - \frac{\zeta_t}{a^2}\right|\right\} \leq 1.$$

Hence, by Theorem V.6 $\|\Omega_t - \Omega_\rho^*\|_F \leq \|\Omega_{t-1} - \Omega_\rho^*\|_F \leq \|\Omega_0 - \Omega_\rho^*\|_F$. Thus

$$\|\Omega_t\|_2 - \|\Omega_\rho^*\|_2 \leq \|\Omega_t - \Omega_\rho^*\|_2 \leq \|\Omega_t - \Omega_\rho^*\|_F \leq \|\Omega_0 - \Omega_\rho^*\|_F$$

so

$$\|\Omega_t\|_2 \leq \|\Omega_\rho^*\|_2 + \|\Omega_0 - \Omega_\rho^*\|_F.$$

□

□

CHAPTER VI

Time-varying Metric Learning

So far, we have looked at the estimation of structured spatio-temporal covariances. As the pdf of a Gaussian distribution is determined by the mean and covariance matrix, the covariance (specifically the inverse covariance) is very useful for providing distance between points in a distribution, with applications to anomaly detection, clustering, and classification. However, real data is often not Gaussian distributed, and is often multiclass and multimodal. In this chapter, we abstract the notion of covariance to that of a metric on the data space, and attempt to find the optimal metric ("inverse covariance") that *best emphasizes the structure of the dataset*, e.g. improving separation of clusters, measuring how anomalous a data point is, etc. We allow the optimal metric to change in time, allowing for complex data sources for which the distribution changes rapidly in time. Adaptive online algorithms allow us to efficiently estimate the metric in as few training samples as possible, and regularization allows us to promote low-rank embeddings (low rank metric) and feature selection (sparse metric).

6.1 Introduction

The effectiveness of many machine learning and data mining algorithms depends on an appropriate measure of pairwise distance between data points that accurately

reflects the learning task, e.g., prediction, clustering or classification. The kNN classifier, K-means clustering, and the Laplacian-SVM semi-supervised classifier are examples of such *distance-based* machine learning algorithms. In settings where there is clean, appropriately-scaled spherical Gaussian data, standard Euclidean distance can be utilized. However, when the data is heavy tailed, multimodal, or contaminated by outliers, observation noise, or irrelevant or replicated features, use of Euclidean inter-point distance can be problematic, leading to bias or loss of discriminative power.

To reduce bias and loss of discriminative power of distance-based machine learning algorithms, data-driven approaches for optimizing the distance metric have been proposed. These methodologies, generally taking the form of dimensionality reduction or data "whitening," aim to utilize the data itself to learn a transformation of the data that embeds it into a space where Euclidean distance is appropriate. Examples of such techniques include Principal Component Analysis (*Bishop*, 2006), Multidimensional Scaling (*Hastie et al.*, 2005), covariance estimation (*Hastie et al.*, 2005; *Bishop*, 2006), and manifold learning (*Lee and Verleysen*, 2007). Such unsupervised methods do not exploit human input on the distance metric, and they overly rely on prior assumptions, e.g., local linearity or smoothness.

In distance metric learning one seeks to learn transformations of the data associated with a distance metric that is well matched to a particular task specified by the user. Pairwise labels or "edges" indicating point similarity or dissimilarity are used to learn a transformation of the data such that similar points are "close" to one another and dissimilar points are distant in the transformed space. Learning distance metrics in this manner allows a more precise notion of distance or similarity to be defined that is better related to the task at hand.

Figure 6.1 illustrates this notion. Data points, or nodes, have underlying similarities or distances between them. Absent an exhaustive label set, given an attribute distance function $d(\cdot, \cdot)$ it is possible to infer similarities between nodes as the dis-

tance between their attribute vectors. As an example, the kNN algorithm uses the Euclidean distance to infer similarity. However, the distance function must be specified a priori, and may not match the distance relevant to the task. Distance metric learning proposes a hybrid approach, where one is given a small number of pairwise labels, uses these to learn a distance function on the attribute space, and then uses this learned function to infer relationships between the rest of the nodes.

Many supervised and semi-supervised distance metric learning approaches have been developed for machine learning and data mining (*Kulis*, 2012). This includes online algorithms (*Kunapuli and Shavlik*, 2012) with regret guarantees for situations where similarity constraints are received sequentially.

This paper proposes a new distance metric tracking method that is applicable to the non-stationary time varying case of distance metric drift and has provably *strongly adaptive* tracking performance.

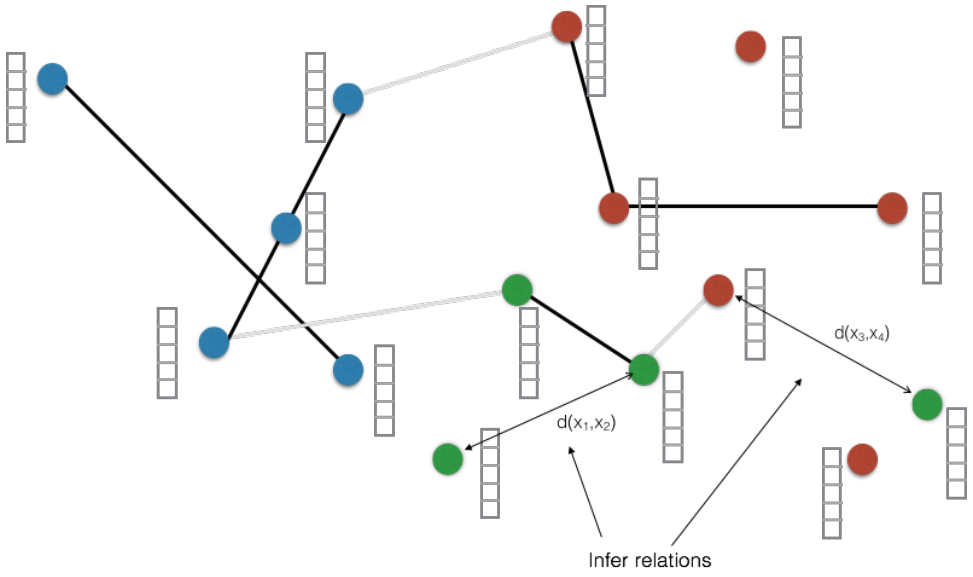


Figure 6.1: Similarity functions on networks, with different clusters indicated by different colored nodes. Attributes of nodes denoted as a 5-element column vector with an unknown similarity function $d(\cdot, \cdot)$ between attributes. Learn and track similarity function implied by observed edges, use result to infer similarities between other nodes.

Specifically, we suppose the underlying ground-truth (or optimal) distance met-

ric from which constraints are generated is evolving over time, in an unknown and potentially nonstationary way. In Figure 6.1, this corresponds to having the relationships between nodes change over time. This can, for example, be caused by changes in the set of features indicative of relations (e.g. polarizing buzzwords in collective discourse), changes in the underlying relationship structure (e.g. evolving communities), and/or changes in the nature of the relationships relevant to the problem or to the user. When any of these changes occur, it is imperative to be able to detect and adapt to them without casting aside previous knowledge.

We propose a strongly adaptive, online approach to track the underlying metric as the constraints are received. We introduce a framework called Online Convex Ensemble StrongLy Adaptive Dynamic Learning (OCELAD), which at every time step evaluates the recent performance of and optimally combines the outputs of an ensemble of online learners, each operating under a different drift-rate assumption. We prove strong bounds on the dynamic regret of every subinterval, guaranteeing strong adaptivity and robustness to nonstationary metric drift such as discrete shifts, slow drift with a widely-varying drift rate, and all combinations thereof. Applying OCELAD to the problem of nonstationary metric learning, we find that it gives excellent robustness and low regret when subjected to all forms of nonstationarity.

Social media provides some of the most dynamic, rapidly changing data sources available. Constant changes in world events, popular culture, memes, and other items of discussion mean that the words and concepts characteristic of subcultures, communities, and political persuasions are rapidly evolving in a highly nonstationary way. As this is exactly the situation our dynamic metric learning approach is designed to address, we will consider modeling political tweets in November 2015, during the early days of the United States presidential primary.

6.1.1 Related Work

Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are classic examples of the use of linear transformations for projecting data into more interpretable low dimensional spaces. Unsupervised PCA seeks to identify a set of axes that best explain the variance contained in the data. LDA takes a supervised approach, minimizing the intra-class variance and maximizing the inter-class variance given class labeled data points.

Much of the recent work in Distance Metric Learning has focused on learning Mahalanobis distances on the basis of pairwise similarity/dissimilarity constraints. These methods have the same goals as LDA; pairs of points labeled "similar" should be close to one another while pairs labeled "dissimilar" should be distant. MMC (*Xing et al.*, 2002), a method for identifying a Mahalanobis metric for clustering with side information, uses semidefinite programming to identify a metric that maximizes the sum of distances between points labeled with different classes subject to the constraint that the sum of distances between all points with similar labels be less than or equal to some constant.

Large Margin Nearest Neighbor (LMNN) (*Weinberger et al.*, 2005) similarly uses semidefinite programming to identify a Mahalanobis distance. In this setting, the algorithm minimizes the sum of distances between a given point and its similarly labeled neighbors while forcing differently labeled neighbors outside of its neighborhood. This method has been shown to be computationally efficient (*Weinberger and Saul*, 2008) and, in contrast to the similarly motivated Neighborhood Component Analysis (*Goldberger et al.*, 2004), is guaranteed to converge to a globally optimal solution. Information Theoretic Metric Learning (ITML) (*Davis et al.*, 2007) is another popular Distance Metric Learning technique. ITML minimizes the Kullback-Liebler divergence between an initial guess of the matrix that parameterizes the Mahalanobis distance and a solution that satisfies a set of constraints. For surveys of the metric

learning literature, see (*Kulis*, 2012; *Bellet et al.*, 2013; *Yang and Jin*, 2006).

In a dynamic environment, it is necessary to track the changing metric at different times, computing a sequence of estimates of the metric, and to be able to compute those estimates online. Online learning (*Cesa-Bianchi and Lugosi*, 2006) meets these criteria by efficiently updating the estimate every time a new data point is obtained instead of minimizing an objective function formed from the entire dataset. Many online learning methods have regret guarantees, that is, the loss in performance relative to a batch method is provably small (*Cesa-Bianchi and Lugosi*, 2006; *Duchi et al.*, 2010b). In practice, however, the performance of an online learning method is strongly influenced by the learning rate, which may need to vary over time in a dynamic environment (*Daniely et al.*, 2015; *McMahan and Streeter*, 2010; *Duchi et al.*, 2010a), especially one with changing drift rates.

Adaptive online learning methods attempt to address the learning rate problem by continuously updating the learning rate as new observations become available. For learning static parameters, AdaGrad-style methods (*McMahan and Streeter*, 2010; *Duchi et al.*, 2010a) perform gradient descent steps with the step size adapted based on the magnitude of recent gradients. Follow the regularized leader (FTRL) type algorithms adapt the regularization to the observations (*McMahan*, 2014). Recently, a method called Strongly Adaptive Online Learning (SAOL) has been proposed for learning parameters undergoing K discrete changes when the loss function is bounded between 0 and 1. SAOL maintains several learners with different learning rates and randomly selects the best one based on recent performance (*Daniely et al.*, 2015). Several of these adaptive methods have provable regret bounds (*McMahan*, 2014; *Herbster and Warmuth*, 1998; *Hazan and Seshadhri*, 2007). These typically guarantee low total regret (i.e. regret from time 0 to time T) at every time (*McMahan*, 2014). SAOL, on the other hand, attempts to have low *static* regret on every subinterval, as well as low regret overall (*Daniely et al.*, 2015). This allows tracking of discrete

changes, but not slow drift. Our work improves upon the capabilities of SAOL by allowing for unbounded loss functions, using a convex combination of the ensemble instead of simple random selection, and providing guaranteed low regret when all forms of nonstationarity occur, not just discrete shifts. All of these additional capabilities are shown in Section 6.6 to be critical for good metric learning performance.

The remainder of this paper is structured as follows. In Section 6.2 we formalize the time varying distance metric tracking problem, and section 6.3 presents the basic COMID online learner and our Retro-Initialized COMID Ensemble (RICE) of learners with dyadically scaled learning rates. Section 6.4 presents our OCELAD algorithm, a method of adaptively combining learners with different learning rates. Strongly adaptive bounds on the dynamic regret of OCELAD and RICE-OCELAD are presented in Section 6.5, and results on both synthetic data and the Twitter dataset are presented in Section 6.6. Section 6.7 concludes the paper.

6.2 Nonstationary Metric Learning

Metric learning seeks to learn a metric that encourages data points marked as similar to be close and data points marked as different to be far apart. The time-varying Mahalanobis distance at time t is parameterized by $\mathbf{M}_t$ as

$$d^2_{M_t}(\mathbf{x}, \mathbf{z}) = (\mathbf{x} - \mathbf{z})^T \mathbf{M}_t (\mathbf{x} - \mathbf{z}) \tag{6.1}$$

where $\mathbf{M}_t \in \mathbb{R}^{n \times n} \succeq 0$.

Suppose a temporal sequence of similarity constraints are given, where each constraint is the triplet $(\mathbf{x}_t, \mathbf{z}_t, y_t)$, $\mathbf{x}_t$ and $\mathbf{z}_t$ are data points in $\mathbb{R}^n$, and the label $y_t = +1$ if the points $\mathbf{x}_t, \mathbf{z}_t$ are similar at time t and $y_t = -1$ if they are dissimilar.

Following (*Kunapuli and Shavlik*, 2012), we introduce the following margin based

constraints for all time points t:

$$\begin{aligned} d^2_{M_t}(\mathbf{x}_t, \mathbf{z}_t) &\leq \mu - 1 \quad \forall y_t = 1 \\ d^2_{M_t}(\mathbf{x}_t, \mathbf{z}_t) &\geq \mu + 1 \quad \forall y_t = -1 \end{aligned} \tag{6.2}$$

where μ is a threshold that controls the margin between similar and dissimilar points. A diagram illustrating these constraints and their effect is shown in Figure 6.2. These constraints are softened by penalizing violation of the constraints with a convex loss function ℓ. This gives a combined loss function

$$\begin{aligned} \mathcal{L}(\{\mathbf{M}_t, \mu\}) &= \frac{1}{T}\sum_{t=1}^{T} \ell(y_t(\mu - \mathbf{u}_t^T \mathbf{M}_t \mathbf{u}_t)) + \lambda r(\mathbf{M}_t) \\ &= \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{M}_t, \mu), \end{aligned} \tag{6.3}$$

where $\mathbf{u}_t = \mathbf{x}_t - \mathbf{z}_t$, r is the regularizer and λ the regularization parameter. Kunapuli and Shavlik (*Kunapuli and Shavlik*, 2012) propose using nuclear norm regularization ($r(\mathbf{M}) = \|\mathbf{M}\|_*$) to encourage projection of the data onto a low dimensional subspace (feature selection/dimensionality reduction), and we have also had success with the elementwise L1 norm ($r(\mathbf{M}) = \|\text{vec}(\mathbf{M})\|_1$). In what follows, we develop an adaptive online method to minimize the loss subject to nonstationary smoothness constraints on the sequence of metric estimates $\mathbf{M}_t$.

6.3 Retro-initialized COMID ensemble (RICE)

Viewing the acquisition of new data points as stochastic realizations of the underlying distribution (*Kunapuli and Shavlik*, 2012) suggests the use of composite objective stochastic mirror descent techniques (COMID). For convenience, we set $\ell_t(\mathbf{M}_t, \mu_t) = \ell(y_t(\mu - \mathbf{u}_t^T \mathbf{M}_t \mathbf{u}_t))$.

For the loss (6.3) and learning rate η_t, application of COMID (*Duchi et al.*, 2010b)

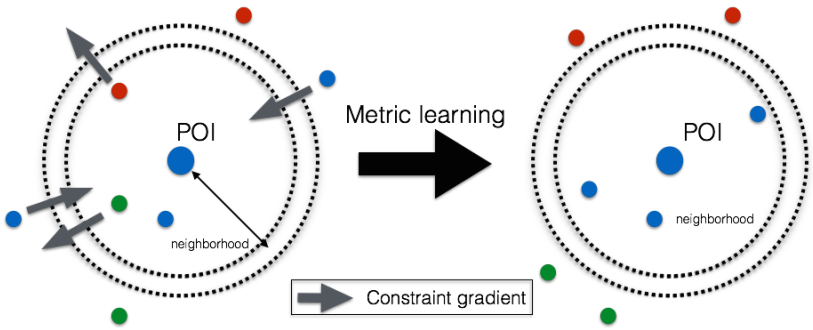


Figure 6.2: Visualization of the margin based constraints (6.2), with colors indicating class. The goal of the metric learning constraints is to move target neighbors towards the point of interest (POI), while moving points from other classes away from the target neighborhood.

gives the online learning update

$$\begin{aligned}\widehat{\mathbf{M}}_{t+1} &= \arg\min_{\mathbf{M}\succeq 0} B_\psi(\mathbf{M}, \widehat{\mathbf{M}}_t) \\ &\quad + \eta_t \langle \nabla_M \ell_t(\widehat{\mathbf{M}}_t, \widehat{\mu}_t), \mathbf{M} - \widehat{\mathbf{M}}_t \rangle + \eta_t \lambda \|\mathbf{M}\|_* \\ \widehat{\mu}_{t+1} &= \arg\min_{\mu \geq 1} B_\psi(\mu, \widehat{\mu}_t) + \eta_t \nabla_\mu \ell_t(\widehat{\mathbf{M}}_t, \widehat{\mu}_t)'(\mu - \widehat{\mu}_t),\end{aligned} \tag{6.4}$$

where B_ψ is any Bregman divergence. As this is an online framework, the t indexing directly corresponds to the received time series of pairwise constraints $(\mathbf{x}_t, \mathbf{z}_t, y_t)$. In (*Kunapuli and Shavlik*, 2012) a closed-form algorithm for solving the minimization in (6.18) with $r(\mathbf{M}) = \|\mathbf{M}\|_*$ is developed for a variety of common losses and Bregman divergences, involving rank one updates and eigenvalue shrinkage.

The output of COMID depends strongly on the choice of η_t. Critically, the optimal learning rate η_t depends on the rate of change of $\mathbf{M}_t$ (*Hall and Willett*, 2015), and thus will need to change with time to adapt to nonstationary drift. Choosing an optimal sequence for η_t is clearly not practical in an online setting with nonstationary drift, since the drift rate is changing. We thus propose to maintain an ensemble of learners with a range of η_t values, whose output we will adaptively combine for optimal nonstationary performance. If the range of η_t is diverse enough, one of the learners in

the ensemble should have good performance on every interval. Critically, the optimal learner in the ensemble may vary widely with time, since the drift rate and hence the optimal learning rate changes in time. For example, if a large discrete change occurs, the fast learners are optimal at first, followed by increasingly slow learners as the estimate of the new value improves. In other words, the optimal approach is fast reaction followed by increasing refinement, in a manner consistent with the attractive $O(1/\sqrt{t})$ decay of the learning rate of optimal nonadaptive algorithms (*Hall and Willett*, 2015).

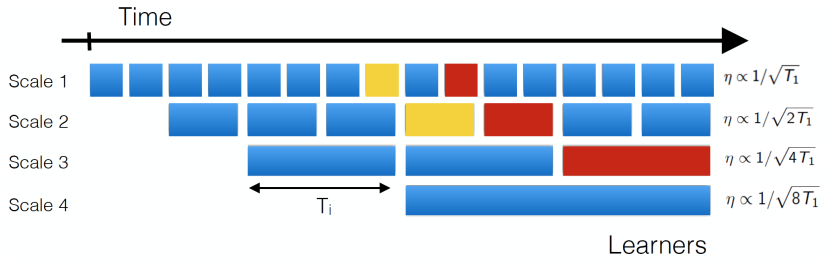


Figure 6.3: Retro-initialized COMID ensemble (RICE). COMID learners at multiple scales run in parallel, with the interval learners learning on the dyadic set of intervals $\mathcal{I}$. Recent observed losses for each learner are used to create weights used to select the appropriate scale at each time. Each yellow and red learner is initialized by the output of the previous learner of the same color, that is, the learner of the next shorter scale.

Define a set $\mathcal{I}$ of intervals $I = [t_{I1}, t_{I2}]$ such that the lengths $|I|$ of the intervals are proportional to powers of two, i.e. $|I| = I_0 2^j$, $j = 0, \ldots,$ with an arrangement that is a dyadic partition of the temporal axis, as in (*Daniely et al.*, 2015). The first interval of length $|I|$ starts at $t = |I|$ (see Figure 6.3), and additional intervals of length $|I|$ exist such that the rest of time is covered.

Every interval I is associated with a base COMID learner that operates on that interval. Each learner (6.18) has a constant learning rate proportional to the inverse square of the length of the interval, i.e. $\eta_t(I) = \eta_0/\sqrt{|I|}$. Each learner (besides the coarsest) at level j ($|I| = I_0 2^j$) is initialized to the last estimate of the next coarsest learner (level $j-1$) (see Figure 6.3). This strategy is equivalent to "backdating"

the interval learners so as to ensure appropriate convergence has occurred before the interval of interest is reached, and is effectively a quantized square root decay of the learning rate. We call our method of forming an ensemble of COMID learners on dyadically nested intervals the Retro-Initialized COMID Ensemble, or RICE, and summarize it in Figure 6.3.

At a given time t, a set $\mathrm{ACT}(t) \subseteq \mathcal{I}$ of $\mathrm{floor}(\log_2 t)$ intervals/COMID learners are active, running in parallel. Because the metric being learned is changing with time, learners designed for low regret at different scales (drift rates) will have different performance (analogous to the classical bias/variance tradeoff). In other words, there is a scale I_{opt} optimal at a given time.

To adaptively select and fuse the outputs of the ensemble, we introduce Online Convex Ensemble StrongLy Adaptive Dynamic Learning (OCELAD), a method that accepts an ensemble of black-box learners and uses recent history to adaptively form an optimal weighted combination at each time.

6.4 OCELAD

To maintain generality, in this section we assume the series of random loss functions is of the form $\ell_t(\theta_t)$ where θ_t is the time-varying unknown parameters. We assume that an ensemble $\mathcal{B}$ of online learners is provided on the dyadic interval set $\mathcal{I}$, each optimized for the appropriate scale. To select the appropriate scale, we compute weights $w_t(I)$ that are updated based on the learner's recent estimated regret. The weight update we use is inspired by the multiplicative weight (MW) literature (*Blum and Mansour*, 2005), modified to allow for unbounded loss functions. At each step, we rescale the observed losses so they lie between -1 and 1, allowing for maximal

weight differentiation while preventing negative weights.

$$r_t(I) = \left(\sum_I \frac{w_t(I)}{\sum_I w_t(I)} \ell_t(\theta_t(I)) \right) - \ell_t(\theta_t(I)) \tag{6.5}$$
$$w_{t+1}(I) = w_t(I) \left(1 + \eta_I \frac{r_t(I)}{\max_{I \in \mathrm{ACT}(t)} |r_t(I)|} \right), \quad \forall t \in I.$$

These hold for all $I \in \mathcal{I}$, where $\eta_I = \min\{1/2, 1/\sqrt{|I|}\}$, $\mathbf{M}_t(I), \mu_t(I)$ are the outputs at time t of the learner on interval I, and $r_t(I)$ is called the estimated regret of the learner on interval I at time t. The initial value of $w(I)$ is η_I. Essentially, (6.5) is highly weighting low loss learners and lowly weighting high loss learners.

For any given time t, the outputs of the learners of interval $I \in \mathrm{ACT}(t)$ are combined to form the weighted ensemble estimate

$$\widehat{\theta}_t = \frac{\sum_{I \in \mathrm{ACT}(t)} w_t(I) \theta_t(I)}{\sum_{I \in \mathrm{ACT}(t)} w_t(I)} \tag{6.6}$$

The weighted average of the ensemble is justified due to our use of a convex loss function (proven in the next section), as opposed to the possibly non-convex losses of (*Blum and Mansour*, 2005), necessitating a randomized selection approach. OCELAD is summarized in Algorithm 1, and the joint RICE-OCELAD approach as applied to metric learning of $\{\mathbf{M}_t, \mu_t\}$ is shown in Algorithm 2.

Algorithm 7 Online Convex Ensemble Strongly Adaptive Dynamic Learning (OCELAD)

1: Provide dyadic ensemble of online learners $\mathcal{B}$.
2: Initialize weight: $w_1(I)$.
3: **for** $t = 1$ to T **do**
4: Observe loss function $\ell_t(\cdot)$ and update $\mathcal{B}$ ensemble.
5: Obtain $|\mathrm{ACT}(t)|$ estimates $\theta_t(I)$ from the $\mathcal{B}$ ensemble.
6: Compute weighted ensemble average $\widehat{\theta}_t$ via (6.6) and set as estimate.
7: Update weights $w_{t+1}(I)$ via (6.5).
8: **end for**
9: Return $\{\widehat{\theta}_t\}$.

Algorithm 8 RICE-OCELAD for Nonstationary Metric Learning

Initialize weight: $w_1(I)$
for $t = 1$ to T **do**
  Obtain constraint $(\mathbf{x}_t, \mathbf{z}_t, y_t)$, compute loss function $\ell_{t,c}(\mathbf{M}_t, \mu_t)$.
  Initialize new learner in RICE if needed. New learner at scale $j > 0$: initialize to the last estimate of learner at scale $j - 1$.
  COMID update $\mathbf{M}_t(I), \mu_t(I)$ using (6.18) for all active learners in RICE ensemble.
  Compute

$$\widehat{\mathbf{M}}_t \leftarrow \frac{\sum_{I \in \mathrm{ACT}(t)} w_t(I) \mathbf{M}_t(I)}{\sum_{I \in \mathrm{ACT}(t)} w_t(I)}$$
$$\widehat{\mu}_t \leftarrow \frac{\sum_{I \in \mathrm{ACT}(t)} w_t(I) \mu_t(I)}{\sum_{I \in \mathrm{ACT}(t)} w_t(I)}.$$

  for $I \in \mathrm{ACT}(t)$ **do**
    Compute estimated regret $r_t(I)$ and update weights according to (6.5) with $\theta_t(I) = \{\mathbf{M}_t(I), \mu_t(I)\}$.
  end for
end for
Return $\{\widehat{\mathbf{M}}_t, \widehat{\mu}_t\}$.

6.5 Strongly Adaptive Dynamic Regret

The standard static regret of an online learning algorithm generating an estimate sequence $\widehat{\theta}_t$ is defined as

$$R_{\mathcal{B},static}(I) = \sum_{t \in I} f_t(\widehat{\theta}_t) - \min_{\theta \in \Theta} \sum_{t \in I} f_t(\theta). \tag{6.7}$$

where $f_t(\theta_t)$ is a loss with parameter θ_t. Since in our case the optimal parameter value θ_t is changing, the static regret of an algorithm $\mathcal{B}$ on an interval I is not useful. Instead, let $\mathbf{w} = \{\theta_t\}_{t \in [0,T]}$ be an arbitrary sequence of parameters. Then, the *dynamic regret* of an algorithm $\mathcal{B}$ relative to any comparator sequence $\mathbf{w} = \{\theta_t\}_{t \in I}$ on the interval I is defined as

$$R_{\mathcal{B},\mathbf{w}}(I) = \sum_{t \in I} f_t(\widehat{\theta}_t) - \sum_{t \in I} f_t(\theta_t), \tag{6.8}$$

where $\widehat{\theta}_t$ are generated by $\mathcal{B}$. This allows for comparison to any possible dynamically changing batch estimate $\mathbf{w} = \{\theta_t\}_{t \in I}$.

In (*Hall and Willett*, 2015) the authors derive dynamic regret bounds that hold over all possible sequences $\mathbf{w}$ such that $\sum_{t \in I} \|\theta_{t+1} - \theta_t\| \leq \gamma$, i.e. bounding the total amount of variation in the estimated parameter. Without this temporal regularization, minimizing the loss would cause θ_t to grossly overfit. In this sense, setting the comparator sequence $\mathbf{w}$ to the "ground truth sequence" or "batch optimal sequence" both provide meaningful intuitive bounds.

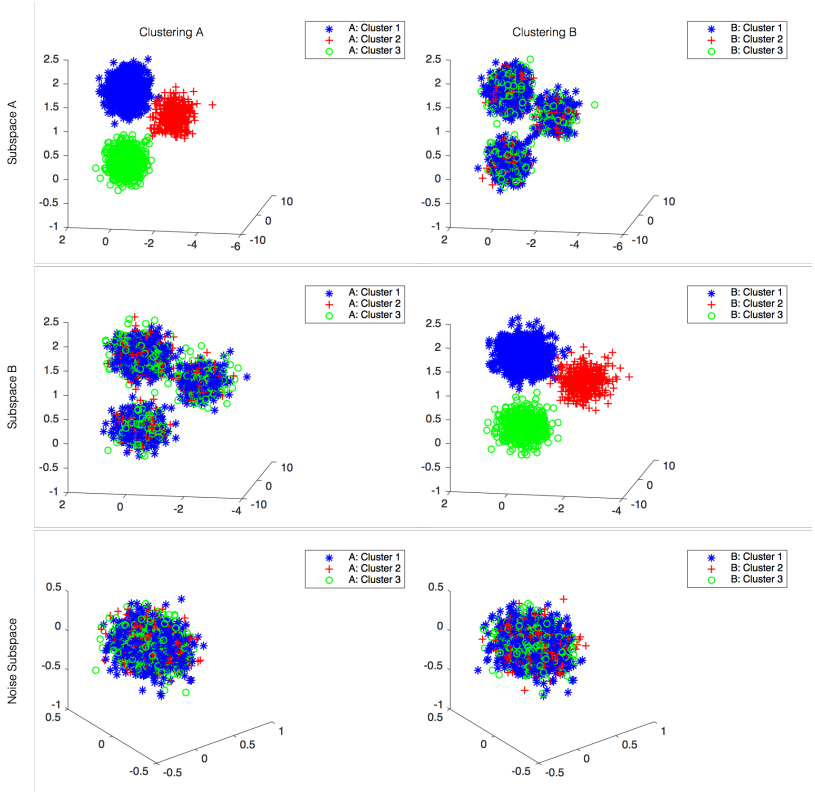


Figure 6.4: 25-dimensional synthetic dataset used for metric learning in Figure 6.5. Datapoints exist in $\mathbb{R}^{25}$, with two natural 3-way clusterings existing simultaneously in orthogonal 3-D subspaces A and B. The remaining 19 dimensions are isotropic Gaussian noise. Shown are the projections of the dataset onto subspaces A and B, as well as a projection onto a portion of the 19 dimensional isotropic noise subspace, with color codings corresponding to the cluster labeling associated with subspaces A and B. Observe that the data points in the left and right columns are identical, the only change is the cluster labels.

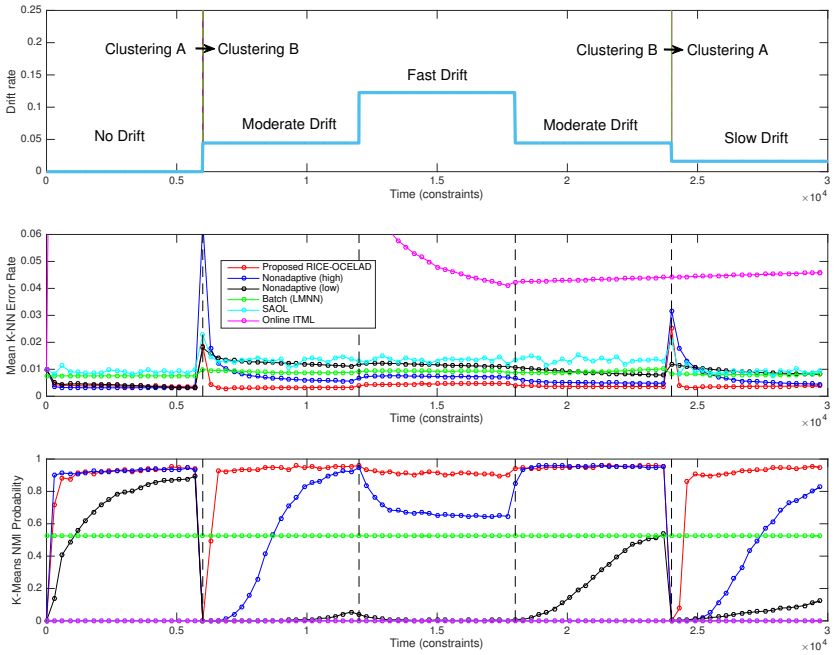


Figure 6.5: Tracking of a changing metric. Top: Rate of change of the data generating random-walk drift matrix $\mathbf{D}_t$ as a function of time. Two discrete changes in clustering labels are marked, causing all methods to have a sudden decrease in performance. Metric tracking performance is computed for RICE-OCELAD, nonadaptive COMID (high learning rate and low learning rate), the batch solution (LMNN), SAOL and online ITML. Shown as a function of time is the mean k-NN error rate (middle) and the probability that the k-means normalized mutual information (NMI) exceeds 0.8 (bottom). Note that RICE-OCELAD alone is able to effectively adapt to the variety of discrete changes and changes in drift rate, and that the NMI of ITML and SAOL fails completely.

Strongly adaptive regret bounds (*Daniely et al.*, 2015) can provide guarantees that static regret is low on every subinterval, instead of only low in the aggregate. We use the notion of dynamic regret to introduce strongly adaptive dynamic regret bounds, proving that *dynamic regret is low on every subinterval* $I \subseteq [0, T]$ *simultaneously*. The following result is proved in the appendix. Suppose there are a sequence of random loss functions $\ell_t(\theta_t)$. The goal is to estimate a sequence $\widehat{\theta}_t$ that minimizes the dynamic regret.

Theorem VI.1 (General OCELAD Regret Framework)**.** *Let* $\mathbf{w} = \{\theta_1, \ldots, \theta_T\}$ *be an arbitrary sequence of parameters and define* $\gamma_{\mathbf{w}}(I) = \sum_{t \in I} \|\theta_{t+1} - \theta_t\|$ *as a function of* $\mathbf{w}$ *and an interval* $I \subseteq [0, T]$*. Choose an ensemble of learners* $\mathcal{B}$ *such that given an interval* I *the learner* $\mathcal{B}_I$ *creates an output sequence* $\theta_t(I)$ *satisfying the dynamic regret bound*

$$R_{\mathcal{B}_I, \mathbf{w}}(I) \leq C(1 + \gamma_{\mathbf{w}}(I))\sqrt{|I|} \tag{6.9}$$

for some constant $C > 0$*. Then the strongly adaptive dynamic learner* $OCELAD^{\mathcal{B}}$ *using* $\mathcal{B}$ *as the ensemble creates an estimation sequence* $\widehat{\theta}_t$ *satisfying*

$$\begin{aligned} R_{OCELAD^{\mathcal{B}}, \mathbf{w}}(I) \leq & 8C(1 + \gamma_{\mathbf{w}}(I))\sqrt{|I|} \\ & + 40 \log\left(1 + \max_{t \in I} t\right)\sqrt{|I|} \end{aligned}$$

on every interval $I \subseteq [0, T]$*.*

In other words, the regret of OCELAD on any finite interval I is sublinear in the length of that interval ($\sqrt{|I|}$), and scales with the amount $\gamma_{\mathbf{w}}(I)$ of variation in true/optimal batch parameter estimates. The logarithmic term in s exists because of the logarithmically increasing number of learners active at time s, required to achieve guaranteed $O(\sqrt{|I|})$ regret on intervals I for which $|I|$ can be up to the order of s.

In a dynamic setting, bounds of this type are particularly desirable because they allow for changing *drift rate* and guarantee quick recovery from *discrete changes*. For instance, suppose a number K of discrete switches (large parameter changes or changes in drift rate) occur at times t_i satisfying $0 = t_0 < t_1 < \cdots < t_K = T$. Then since $\sum_{i=1}^{K} \sqrt{|t_{i-1} - t_i|} \leq \sqrt{KT}$, this implies that the total expected dynamic regret on $[0, T]$ remains low ($O(\sqrt{KT})$), while simultaneously guaranteeing that an appropriate learning rate is achieved on each subinterval $[t_i, t_{i+1}]$.

We emphasize that this type of nonstationarity includes not only changes in model drift rate, but also changes in the time-separation of the incoming labels, as both types

of nonstationarity are equivalent under this model.

Now, reconsider the dynamic metric learning problem of Section II. It is reasonable to assume that the transformed distance between any two points is bounded, implying $\|\mathbf{M}\| \leq c'$ and that $\ell_t(\mathbf{M}_t, \mu_t) \leq k = \ell(c' \max_t \|\mathbf{x}_t - \mathbf{z}_t\|_2^2)$. Thus the loss (and the gradient) are bounded. We can then show the COMID learners in the RICE ensemble have low dynamic regret. The proof of the following result is given in the appendix.

Corollary VI.2 (Dynamic Regret: Metric Learning COMID)**.** *Let the sequence* $\widehat{\mathbf{M}}_t, \widehat{\mu}_t$ *be generated by* (6.18)*, and let* $\mathbf{w} = \{\mathbf{M}_t\}_{t=1}^T$ *be an arbitrary sequence with* $\|\mathbf{M}_t\| \leq c$*. Then using* $\eta_{t+1} \leq \eta_t$ *gives*

$$R_{\mathbf{w}}([0,T]) \leq \frac{D_{max}}{\eta_{T+1}} + \frac{4\phi_{max}}{\eta_T}\gamma + \frac{G_\ell^2}{2\sigma}\sum_{t=1}^{T}\eta_t \tag{6.10}$$

and setting $\eta_t = \eta_0/\sqrt{T}$*,*

$$\begin{aligned} &R_{\mathbf{w}}([0,T]) && (6.11)\\ &\leq \sqrt{T}\left(\frac{D_{max} + 4\phi_{max}(\sum_t \|\mathbf{M}_{t+1} - \mathbf{M}_t\|_F)}{\eta_0} + \frac{\eta_0 G_\ell^2}{2\sigma}\right) \\ &= O\left(\sqrt{T}\left[1 + \sum_{t=1}^{T}\|\mathbf{M}_{t+1} - \mathbf{M}_t\|_F\right]\right). && (6.12)\end{aligned}$$

Since the COMID learners have low dynamic regret on the metric learning problem, we can apply the OCELAD framework to the RICE ensemble.

Theorem VI.3 (Strongly Adaptive Dynamic Regret of RICE-OCELAD applied to metric learning)**.** *Let* $\mathbf{w} = \{\mathbf{M}_t\}_{t\in[0,T]}$ *be any sequence of metrics with* $\|\mathbf{M}_t\| \leq c$ *on the interval* $[0,T]$*, and define* $\gamma_{\mathbf{w}}(I) = \sum_{t\in I}\|\mathbf{M}_{t+1} - \mathbf{M}_t\|$*. Let* $\mathcal{B}$ *be the RICE ensemble with* $\eta_t(I) = \eta_0/\sqrt{|I|}$*. Then the RICE-OCELAD metric learning algorithm*

(Algorithm 2) satisfies

$$R_{OCELAD,\mathbf{w}}(I) \leq \tag{6.13}$$
$$\frac{4}{2^{1/2}-1}C(1+\gamma_{\mathbf{w}}(I))\sqrt{|I|}+40\log(s+1)\sqrt{|I|},$$

for every subinterval $I=[q,s]\subseteq[0,T]$ *simultaneously.* C *is a constant.*

6.6 Results

6.6.1 Synthetic Data

We run our metric learning algorithms on a synthetic dataset undergoing different types of simulated metric drift. We create a synthetic 2000 point dataset with 2 independent three-way clusterings (denoted as clusterings A and B) of the points when projected onto orthogonal 3-dimensional subspaces of $\mathbb{R}^{25}$. The clusterings are formed as 3-D Gaussian blobs with cluster assignment probabilities .5, .3, and .2. The remaining 19 coordinates are filled with isotropic Gaussian noise. Specifically, datapoints $\mathbf{x}_t \in \mathbb{R}^{25}$ are generated as

$$\mathbf{x}_t = \begin{bmatrix} \mathcal{N}(\mathbf{m}_{i_t}, \mathbf{\Sigma}_{i_t}) \\ \mathcal{N}(\mathbf{m}_{j_t}, \mathbf{\Sigma}_{j_t}) \\ \mathcal{N}(0, \sigma_0^2\mathbf{I}_{19\times 19}) \end{bmatrix}$$

$$\Pr(i_t=k)=\Pr(j_t=k)=\begin{cases} .5 & k=1 \\ .3 & k=2 \\ .2 & k=3 \end{cases}$$

where i_t, j_t are independent, σ_0 is the standard deviation of the noise dimensions, and the $\mathbf{m}_k \in \mathbb{R}^3, \mathbf{\Sigma}_k \in \mathbb{R}^{3\times 3}$ are the means and covariances associated with each blob. The label of $\mathbf{x}_t$ under clustering A is i_t, and the label of $\mathbf{x}_t$ under clustering B is j_t.

We create a scenario exhibiting nonstationary drift, combining continuous drifts and shifts between the two clusterings (A and B). To simulate continuous drift, at each time step we perform a random rotation of the dataset, i.e.

$$\widetilde{\mathbf{x}}_t = \mathbf{D}_t \mathbf{x}_t, \quad \widetilde{\mathbf{z}}_t = \mathbf{D}_t \mathbf{z}_t,$$

where $\mathbf{D}_t$ is a random walk (analogous to Brownian motion) on the 25-D sphere of rotation matrices in $\mathbb{R}^{25}$, with $\mathbf{D}_0$ chosen uniformly at random. The time-varying rate of change (random walk stepsize) chosen for $\mathbf{D}_t$ is shown in Figure 6.5, with the small changes in $\mathbf{D}_t$ at each time step accumulating to major changes over longer intervals. For the first interval, partition A is used and the dataset is static, no drift occurs ($\mathbf{D}_t = \mathbf{D}_0$). Then, the partition is changed to B, followed by an interval of first moderate, then fast, and then moderate drift. Finally, the partition reverts back to A, followed by slow drift. The similarity labels y_t are dictated by the partition active at time t. In order to achieve good performance, the online metric learners must be able to track both large discrete changes (change in clustering) as well as the nonstationary gradual drift in $\mathbf{D}_t$.

We generate a series of T constraints from random pairs of points in the dataset $(\widetilde{x}_t, \widetilde{z}_t)$ running each experiment with 3000 random trials. For each experiment conducted in this section, we evaluate performance using two metrics. We plot the K-nearest neighbor error rate, using the learned embedding at each time point, averaging over all trials. We quantify the clustering performance by plotting the empirical probability that the normalized mutual information (NMI) of the K-means clustering of the unlabeled data points in the learned embedding at each time point exceeds 0.8 (out of a possible 1). Clustering NMI, rather than k-NN classification performance, is a more intuitive and realistic indicator of metric learning performance, particularly when finding a relevant embedding in which the clusters are well-separated is the

primary goal.

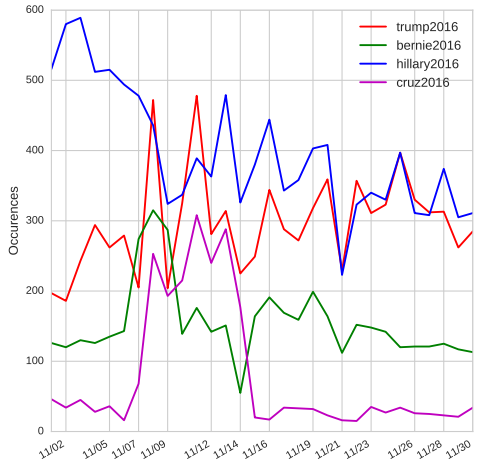
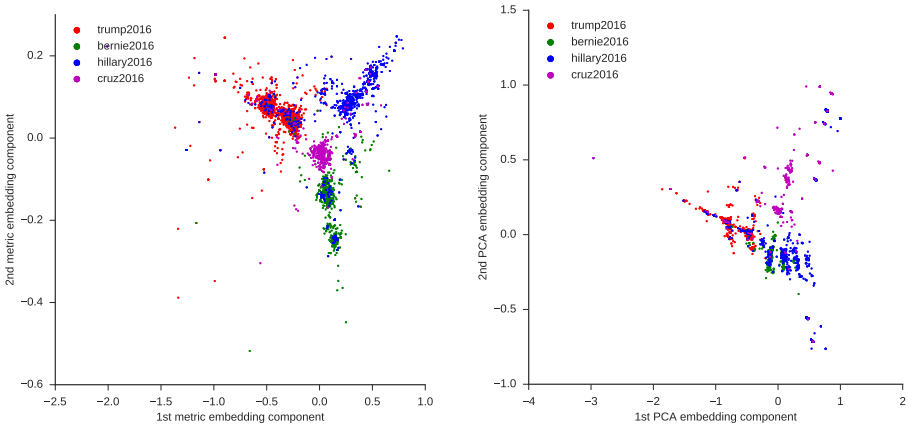


Figure 6.6: Number of tweets per day over the month of November 2015 for four of the US presidential candidates' political hashtags specified in the legend.

In our results, we consider RICE-OCELAD, SAOL with COMID (*Daniely et al.*, 2015), nonadaptive COMID (*Kunapuli and Shavlik*, 2012), LMNN (batch) (*Weinberger et al.*, 2005), and online ITML (*Davis et al.*, 2007).

For RICE-OCELAD, we set the base interval length $I_0 = 1$ time step throughout, and set η_0 via cross-validation in a separate scenario with no drift, emphasizing that the parameters do not need to be tuned for different drift rates. All parameters for the other algorithms were set via cross validation, so as to err on the side of optimism in a truly online scenario. For nonadaptive COMID, we set the high learning rate using cross validation for moderate drift, and we set the low learning rate via cross validation in the case of no drift. The results are shown in Figure 6.5. Online ITML fails due to its bias agains low-rank solutions (*Davis et al.*, 2007), and the batch method and low learning rate COMID fail due to an inability to adapt. The high learning rate COMID does well at first, but as it is optimized for slow drift it cannot adapt to the changes in drift rate as well or recover quickly from the two partition changes. SAOL, as it is designed for mildly-varying bounded loss functions without slow drift and does not use retro-initialized learners, completely fails in this setting



(a) OCELAD Metric Learning (b) Time-windowed PCA

Figure 6.7: Embeddings of political tweets during the last week of November 2015. Shown are the 2-D embeddings using the OCELAD learned metric from the midpoint of the week (a), and using PCA (b). Note the much more distinct groupings by candidate in the OCELAD metric embedding. Using 3-D embeddings, the LOO k-NN error rate is 7.8% in the OCELAD metric embedding and 60.6% in the PCA embedding.

(zero probability of NMI $>$ .8 throughout). RICE-OCELAD, on the other hand, adapts well throughout the entire interval, as predicted by the theory.

6.6.2 Tracking Metrics on Twitter

As noted in the introduction, social media represents a type of highly nonstationary, high dimensional and richly clustered data. We consider political tweets in November 2015, during the early days of the United States presidential primary, and attempt to learn time-varying metrics on the TF-IDF features.

We first extracted all available tweets containing the hashtags #trump2016, #cruz2016, #bernie2016, #hillary2016, representing the two most successful primary candidates from each of the two major parties. We then removed all hashtags from the tweets, and extracted 194 term frequency - inverse document frequency (TF-IDF) stemmed word features. TF-IDF features have been applied to various problems in Twitter data (*Signorini et al.*, 2011; *Antoine et al.*, 2015; *Petrović et al.*, 2010). This provided us with a time series of hashtag-labeled 194-dimensional TF-IDF feature vectors. We

chose to generate pairwise comparisons by considering time-adjacent tweets and labeling them as similar if they shared the same candidate hashtag, and dissimilar if they had different candidate hashtags. This created a time series of 13600 pairwise comparisons, with the time intervals between comparisons highly nonstationary, strongly depending on time of day, day of the week, and various other factors.

We ran RICE-OCELAD metric learning on this time series of pairwise comparisons, with the base interval set at length 1 and base learning rate set at 1. This emphasizes RICE-OCELAD's complete freedom from tuning parameters. To illustrate the learned embedding on the TF-IDF stems, Figure 6.7 shows the projection of tweets from the last week of the month onto the first two principal components of the learned metric $\mathbf{M}_T$ from the midpoint of the last week. Note the clear separation into clusters by political hashtag as desired, with a LOO-kNN error rate of 7.8% in the learned embedding. The standard PCA embedding, on the other hand, is highly disorganized, and suffers a 60.6% LOO-kNN error rate in the same scenario.

Having confirmed that our approach successfully learns the relevant embedding, we illustrate how the learned metric evolves throughout the month in response to changing events. For each metric $\mathbf{M}$, we computed the first two principle component vectors $\mathbf{u}_1$ and $\mathbf{u}_2$. For each feature stem, we found the corresponding entries in $\mathbf{u}_1$, $\mathbf{u}_2$ and used these as (x, y) coordinates in a scatter plot, creating word/stem scatter plots (Figure 6.8). By way of interpretation, the scatter plot location of a word/stem is the point in the 2D embedding to which a tweet containing only that word would be mapped, and quantifies the contribution of each word/stem to the metric.

Figure 6.8 shows word stem scatter plots for the learned metrics at the beginning and end of the month, and the day of and the day after the televised November 10 Republican debate. Only the top 60 terms most relevant to the metric are shown for clarity. Observe the changing structure of the term embeddings, with new terms arising and leaving as the discussion evolves. An alternate view of this experiment

is shown in Figure 6.9, showing the changing relevance of selected individual terms throughout the month. In the captions, we have mentioned explanatory contextual information that can be found in news articles from the period. In both figures, time-varying structure is evident, with Figure 6.8 emphasizing how similar embeddings of words indicate similar meaning/relevance to a candidate, and with Figure 6.9 emphasizing the nonstationary emergence and recession of clustering-relevant terms as the discussion evolves in response to news events.

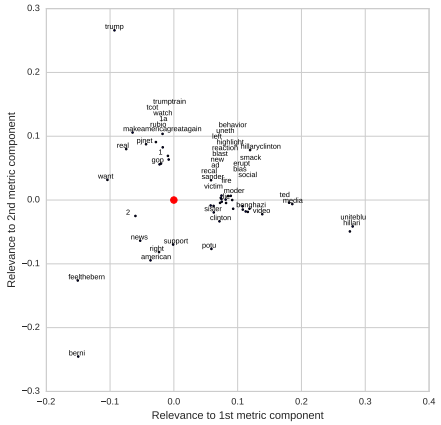
The ability of RICE-OCELAD metric learning, without parameter tuning or specialized feature extraction, to successfully adapt the embedding and identify terms and their relevance to the discussion in this highly nonstationary environment confirms the power of our proposed methodology. RICE-OCELAD allows significant insight into complex, nonstationary data sources to be gleaned by tracking a task-relevant, adaptive, time-varying metric/low dimensional embedding of the data.

6.7 Conclusion

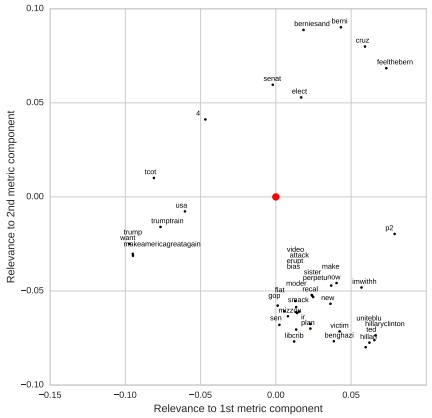
Learning a metric on a complex dataset enables both unsupervised methods and/or a user to home in on the problem of interest while de-emphasizing extraneous information. When the problem of interest or the data distribution is nonstationary, however, the optimal metric can be time-varying. We considered the problem of tracking a nonstationary metric and presented an efficient, strongly adaptive online algorithm (OCELAD), that combines the outputs of any black box learning ensemble (such as RICE), and has strong theoretical regret guarantees. Performance of our algorithm was evaluated both on synthetic and real datasets, demonstrating its ability to learn and adapt quickly in the presence of changes both in the clustering of interest and in the underlying data distribution.

Potential directions for future work include the learning of more expressive metrics beyond the Mahalanobis metric (e.g. nonlinearization via neural network architec-

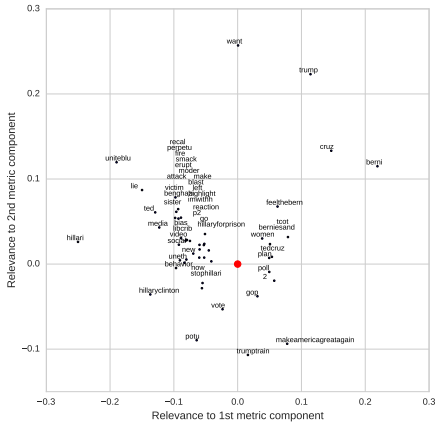
ture), the incorporation of unlabeled data points in a semi-supervised learning framework (*Bilenko et al.*, 2004), and the incorporation of an active learning framework to select which pairs of data points to obtain labels for at any given time (*Settles*, 2012).



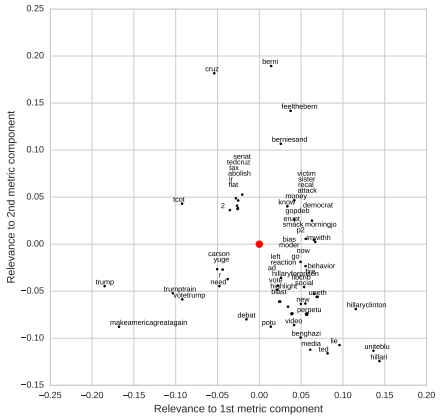
(a) Beginning of the month (Nov 2): Aftermath of Oct 28 Republican debate and revelations from sister of Benghazi victim. Uniteblue campaign to unite Democrats.



(b) End of the month (Nov 30): Continued Benghazi scandal discussion, conservative criticism of University of Missouri protests, Sen. Cruz IRS/tax proposals.

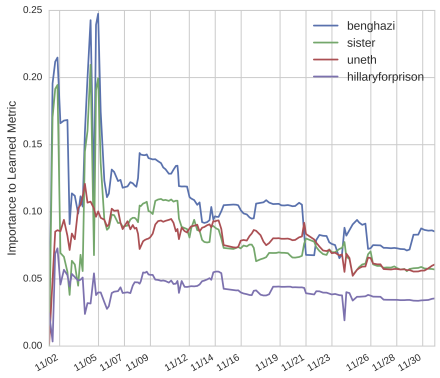


(c) Hours before Nov 10 Republican debate: Discussion of Clinton Benghazi scandal, media bias, Bernie Sanders.

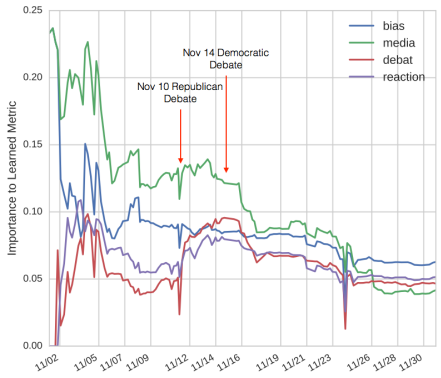


(d) Day after Nov 10 Republican debate: Importance of term "debate", Sen. Cruz's proposals for a flat tax and the abolishing of the IRS, and references to Trump "yuge" and Ben Carson.

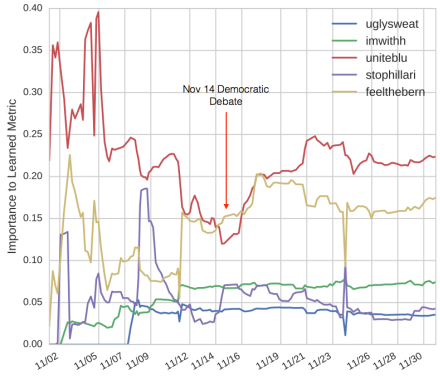
Figure 6.8: Changing metrics on political tweets. Shown are scatter plots of the 60 largest contributions of words to the first two learned metric components. The greater the distance of a word from the origin (marked as a red dot), the larger its contribution to the metric. For readability, we have moved in words with distance from the origin greater than a threshold. Note the changes in relevance and radial groupings of words before and after the Nov 10 Republican debate, and across the entire month.



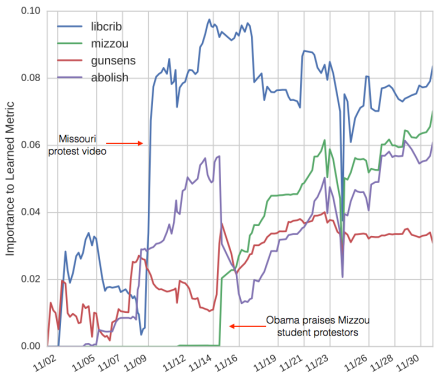
(a) Sister of Benghazi victim spoke out Oct 23, leading to higher relevance early in November.



(b) Accusations of media bias during and after the CNBC Republican debate on Oct 28, but not at the FoxNews Republican debate on Nov 10. Increases in "debate", "reaction", loosely matching the aftermath of those debates, as well as the Nov 14 Democrat debate.



(c) The campaign known as Uniteblue attempted to unify the Democratic party, and ugly sweater promotions for Sanders occurred later in the month. "Uniteblue," "feelthebern," and "stophillary" uptick in relevance during Democratic debate.



(d) On Nov 9 a video of a University of Missouri professor blocking a journalist drew increased attention to liberal protests at that university, related to the rise of the "libcrib" and "mizzou" terms. Cruz policy proposals to limit gun control ("gunsense") and abolish the IRS ("abolish") become informative around and following the Nov 10 Republican debate.

Figure 6.9: Alternate view of the Figure 6.8 experiment, showing as a function of time the relevance (distance from the origin in the embedding) of selected terms appearing in Figure 6.8. The rapid changes in several terms confirms the ability of OCELAD to rapidly adapt the metric to nonstationary changes in the data.

6.8 Appendix

6.8.1 OCELAD - Strongly Adaptive Dynamic Regret

We will prove Theorem 1, giving strongly adaptive dynamic regret bounds. The bound for RICE-OCELAD applied to metric learning follows by combining this general result with Corollary 1.

Define as a function of $I = [q, s] \subseteq [0, T]$

$$\widetilde{w}_t(I) = \begin{cases} 0 & t < q \\ 1 & t = q \\ \widetilde{w}_{t-1}(I)(1 + \eta_I \rho_{t-1} r_{t-1}(I)) & q < t \leq s+1 \\ \widetilde{w}_s(I) & t > s+1 \end{cases} \tag{6.14}$$

and set

$$\rho_t = \frac{1}{\max_I |r_t(I)|}, \ \widetilde{W}_t = \sum_{I \in \mathcal{I}} \widetilde{w}_{t+1}(I). \tag{6.15}$$

Note that $w_t(I) = \eta_I I(t) \widetilde{w}_t(I)$ where $I(t)$ is the indicator function for the interval I, and assume that $\rho_t > c_\rho$, i.e. the estimated regret r_t is bounded, where the bound need not be known.

Recall our definition of the set $\mathcal{I}$ of intervals I such that the lengths $|I|$ of the intervals are proportional to powers of two, i.e. $|I| = I_0 2^j$, $j = 0, \ldots,$ with an arrangement that is a dyadic partition of the temporal axis. The first interval of length $|I|$ starts at $t = |I|$ (see Figure 6.3), and additional intervals of length $|I|$ exist such that the rest of the time axis is covered.

We first prove a pair of lemmas.

Lemma VI.4.

$$\widetilde{W}_t \leq t(\log(t) + 1)$$

for all $t \geq 1$.

Proof. For all $t \geq 1$, by the definition of the set of dyadic intervals $\mathcal{I}$, we have that the number of intervals in $\mathcal{I}$ with endpoint t is given by $|\{[q,s] \in \mathcal{I} : q = t\}| \leq \lfloor \log(t) \rfloor + 1$, where $|\cdot|$ indicates cardinality. Thus summing over all intervals I in the dyadic set of intervals $\mathcal{I}$,

$$\begin{aligned}
\widetilde{W}_{t+1} &= \sum_{I \in \mathcal{I}} \widetilde{w}_{t+1}(I) \\
&= \sum_{I=[q,s] \in \mathcal{I}: q = t+1} \widetilde{w}_{t+1}(I) + \sum_{I=[q,s] \in \mathcal{I}: q \leq t} \widetilde{w}_{t+1}(I) \\
&\leq \log(t+1) + 1 + \sum_{I=[q,s] \in \mathcal{I}: q \leq t} \widetilde{w}_{t+1}(I).
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{I=[q,s] \in \mathcal{I}: q \leq t} \widetilde{w}_{t+1}(I) &= \sum_{I=[q,s] \in \mathcal{I}: q \leq t} \widetilde{w}_t(I)(1 + \eta_I I(t) \rho_t r_t(I)) \\
&= \widetilde{W}_t + \sum_{I \in \mathcal{I}} w_t(I) \rho_t r_t(I).
\end{aligned}$$

Suppose that $\widetilde{W}_t \leq t(\log(t) + 1)$. Furthermore, note that

$$\begin{aligned}
\sum_{I \in \mathcal{I}} w_t(I) \rho_t r_t(I) &= W_t \sum_{I \in \mathcal{I}} p_t(I) \rho_t \left(\ell_t(\widehat{\theta}_t) - \ell_t(\theta_t(I)) \right) \\
&= \rho_t \left(\ell_t \left(\sum_{I \in \mathcal{I}} p_t(I) \theta_t(I) \right) - \sum_{I \in \mathcal{I}} p_t(I) \ell_t(\theta_t(I)) \right) \\
&\leq 0.
\end{aligned}$$

since ℓ_t is convex. Thus

$$\begin{aligned}\widetilde{W}_{t+1} &\leq t(\log(t)+1)+\log(t+1)+1+\rho_t\sum_{I\in\mathcal{I}} w_t(I)r_t(I)\\ &\leq (t+1)(\log(t+1)+1).\end{aligned}$$

Since $\widetilde{W}_1 = \widetilde{w}([1,1]) = 1$, the lemma follows by induction.

□

Lemma VI.5.

$$E\sum_{t=q}^{s} r_t(I) \leq 5\log(s+1)\sqrt{|I|},$$

for every $I=[q,s]\in\mathcal{I}$.

Proof. Fix $I=[q,s]\in\mathcal{I}$. Recall that

$$\widetilde{w}_{s+1}(I) = \prod_{t=q}^{s}(1+\eta_I I(t)\rho_t r_t(I)) = \prod_{t=q}^{s}(1+\eta_I\rho_t r_t(I)).$$

Since $\eta_I\in(0,1/2)$ and $\log(1+x)\geq(x-x^2)$ for all $x\geq -1/2$,

$$\begin{aligned}\log(\widetilde{w}_{s+1}(I)) &= \sum_{t=q}^{s}\log(1+\eta_I\rho_t r_t(I)) \qquad (6.16)\\ &\geq \sum_{t=q}^{s}\eta_I\rho_t r_t(I) - \sum_{t=q}^{s}(\eta_I\rho_t r_t(I))^2\\ &\geq \eta_I\left(\sum_{t=q}^{s}\rho_t r_t(I) - \eta_I|I|\right).\end{aligned}$$

where we have used $|\rho_t r_t(I)| = \frac{|r_t(I)|}{\max_I|r_t(I)|}\leq 1$. By Lemma VI.4 we have

$$\widetilde{w}_{s+1}(I)\leq\widetilde{W}_{s+1}\leq(s+1)(\log(s+1)+1),$$

so

$$\log(\widetilde{w}_{s+1}(I)) \leq \log(\widetilde{w}_{s+1}(I)) \leq \log(s+1) + \log(\log(s+1)+1).$$

Combining with (6.16) and dividing by η_I,

$$\begin{aligned}\sum_{t=q}^{s} \rho_t r_t(I) &\leq \eta_I |I| + \frac{1}{\eta_I}(\log(s+1) + \log(\log(s+1)+1)) \\ &\leq \eta_I |I| + 2\eta_I^{-1} \log(s+1) \\ &= 5 \log(s+1)\sqrt{|I|},\end{aligned}$$

since $x \geq \log(1+x)$ and $\eta_I = \min\{1/2, |I|^{-1/2}\}$. Since $\rho_t > c_\rho > 0$, this implies

$$\sum_{t=q}^{s} r_t(I) \leq \frac{5}{c_\rho} \log(s+1)\sqrt{|I|}.$$

□

Define the restriction of $\mathcal{I}$ to an interval $J \subseteq \mathbb{N}$ as $\mathcal{I}|_J = \{I \in \mathcal{I} : I \subseteq J\}$. Note the following lemma from (*Daniely et al.*, 2015):

Lemma VI.6. *Consider the arbitrary interval $I = [q, s] \subseteq \mathbb{N}$. Then, the interval I can be partitioned into two finite sequences of disjoint and consecutive intervals, given by $(I_{-k}, \ldots, I_0) \subseteq \mathcal{I}|_I$ and $(I_1, I_2, \ldots, I_p) \subseteq \mathcal{I}|_I$, such that*

$$\begin{aligned}|I_{-i}|/|I_{-i+1}| &\leq 1/2, \quad \forall i \geq 1, \\ |I_i|/|I_{i-1}| &\leq 1/2, \qquad \forall i \geq 2.\end{aligned}$$

This enables us to extend the bounds to every arbitrary interval $I = [q, s] \subseteq [0, T]$ and thus complete the proof.

Let $I = \bigcup_{i=-k}^{p} I_i$ be the partition described in Lemma VI.6. Then

$$R_{OCELAD^{\mathcal{B}},\mathbf{w}}(I) \leq \sum_{i\leq 0} R_{OCELAD^{\mathcal{B}},\mathbf{w}}(I_i) + \sum_{i\geq 1} R_{OCELAD^{\mathcal{B}},\mathbf{w}}(I_i). \tag{6.17}$$

By Lemma VI.5 and (6.9),

$$\begin{aligned}
&\sum_{i\leq 0} R_{OCELAD^{\mathcal{B}},\mathbf{w}}(I_i) \\
&\quad \leq C \sum_{i\leq 0} (1+\gamma_{\mathbf{w}}(I_i))\sqrt{|I_i|} + 5\sum_{i\leq 0} \log(s_i+1)\sqrt{I_i} \\
&\quad \leq (C(1+\gamma(I)) + 5\log(s_i+1)) \sum_{i\leq 0} \sqrt{I_i},
\end{aligned}$$

since $\gamma_{\mathbf{w}}(I_i) \leq \gamma_{\mathbf{w}}(I)$ by definition. By Lemma VI.6,

$$\sum_{i\leq 0} \sqrt{|I_i|} \leq \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{|I|} \leq 4\sqrt{|I|}.$$

This bounds the first term of the right hand side of Equation (6.17). The bound for the second term can be found in the same way. Thus,

$$R_{OCELAD^{\mathcal{B}},\mathbf{w}}(I) \leq (8C(1+\gamma(I))\sqrt{|I|} + 40\log(s+1)\sqrt{|I|}.$$

Since this holds for all I, this completes the proof.

6.8.2 Online DML Dynamic Regret

In this section, we derive the dynamic regret of the COMID metric learning algorithm. Recall that the COMID algorithm is given by

$$\begin{aligned}\widehat{\mathbf{M}}_{t+1} &= \arg\min_{\mathbf{M}\succeq 0} B_\psi(\mathbf{M}, \widehat{\mathbf{M}}_t) \\ &\quad + \eta_t \langle \nabla_M \ell_t(\widehat{\mathbf{M}}_t, \mu_t), \mathbf{M} - \widehat{\mathbf{M}}_t \rangle + \eta_t \lambda \|\mathbf{M}\|_* \\ \widehat{\mu}_{t+1} &= \arg\min_{\mu \geq 1} B_\psi(\mu, \widehat{\mu}_t) + \eta_t \nabla_\mu \ell_t(\widehat{\mathbf{M}}_t, \widehat{\mu}_t)'(\mu - \widehat{\mu}_t),\end{aligned} \tag{6.18}$$

where B_ψ is any Bregman divergence and η_t is the learning rate parameter. From (*Hall and Willett*, 2015) we have:

Theorem VI.7.

$$\begin{aligned}G_\ell &= \max_{\theta\in\Theta} \|\nabla f(\theta)\|,\ \phi_{max} = \frac{1}{2}\max_{\theta\in\Theta}\|\nabla\psi(\theta)\| \\ D &= \max_{\theta,\theta'\in\Theta} B_\psi(\theta'\|\theta)\end{aligned}$$

Let the sequence $\widehat{\theta}_t$, $t = 1, \cdots, T$ *be generated via the COMID algorithm, and let* $\mathbf{w}$ *be an arbitrary sequence in* $\mathcal{W} = \{\mathbf{w} | \sum_{t=0}^{T-1} \|\theta_{t+1} - \theta_t\| \leq \gamma\}$. *Then using* $\eta_{t+1} \leq \eta_t$ *gives a dynamic regret*

$$R_{\mathbf{w}}([0,T]) \leq \frac{D}{\eta_{T+1}} + \frac{4\phi_{max}}{\eta_T}\gamma + \frac{G_\ell^2}{2\sigma}\sum_{t=1}^{T}\eta_t \tag{6.19}$$

Using a nonincreasing learning rate η_t, we can then prove a bound on the dynamic regret for a quite general set of stochastic optimization problems.

Applying this to our problem, we have

$$G_\ell = \max_{\|\mathbf{M}\| \le c, t, \mu} \|\nabla(\ell_t(\mathbf{M}, \mu) + \lambda\|\mathbf{M}\|_*)\|_2$$
$$\phi_{max} = \frac{1}{2} \max_{\|\mathbf{M}\| \le c} \|\nabla\psi(\mathbf{M})\|_2,\ D = \max_{\|\mathbf{M}\|, \|\mathbf{M}'\| \le c} B_\psi(\mathbf{M}'\|\mathbf{M}).$$

For $\ell_t(\cdot)$ being the hinge loss and $\psi = \|\cdot\|_F^2$,

$$G_\ell \le \sqrt{(\max_t \|\mathbf{x}_t - \mathbf{z}_t\|_2^2 + \lambda)^2}$$
$$\phi_{max} = c\sqrt{n},\ D = 2c\sqrt{n}.$$

The other two quantities are guaranteed to exist and depend on the choice of Bregman divergence and c. Thus,

Corollary VI.8 (Dynamic Regret: Metric Learning COMID)**.** *Let the sequence* $\widehat{\mathbf{M}}_t, \widehat{\mu}_t$ *be generated by* (6.18)*, and let* $\mathbf{w} = \{\mathbf{M}_t\}_{t=1}^T$ *be an arbitrary sequence with* $\|\mathbf{M}_t\| \le c$*. Then using* $\eta_{t+1} \le \eta_t$ *gives*

$$R_{\mathbf{w}}([0,T]) \le \frac{D}{\eta_{T+1}} + \frac{4\phi_{max}}{\eta_T}\gamma + \frac{G_\ell^2}{2\sigma}\sum_{t=1}^T \eta_t \tag{6.20}$$

and setting $\eta_t = \eta_0/\sqrt{T}$*,*

$$\begin{aligned} &R_{\mathbf{w}}([0,T]) && (6.21)\\ &\le \sqrt{T}\left(\frac{D + 4\phi_{max}(\sum_t \|\mathbf{M}_{t+1} - \mathbf{M}_t\|_F)}{\eta_0} + \frac{\eta_0 G_\ell^2}{2\sigma}\right)\\ &= O\left(\sqrt{T}\left[1 + \sum_{t=1}^T \|\mathbf{M}_{t+1} - \mathbf{M}_t\|_F\right]\right). && (6.22)\end{aligned}$$

Corollary VI.8 is a bound on the regret relative to the batch estimate of $\mathbf{M}_t$ that minimizes the total batch loss subject to a bounded variation $\sum_t \|\mathbf{M}_{t+1} - \mathbf{M}_t\|_F$. Also note that setting $\eta_t = \eta_0/\sqrt{t}$ gives the same bound as (6.22).

In other words, we pay a linear penalty on the total amount of variation in the underlying parameter sequence. From (6.22), it can be seen that the bound-minimizing η_0 increases with increasing $\sum_t \|\mathbf{M}_{t+1} - \mathbf{M}_t\|_F$, indicating the need for an adaptive learning rate.

For comparison, if the metric is in fact static then by standard stochastic mirror descent results (*Hall and Willett*, 2015)

Theorem VI.9 (Static Regret)**.** *If* $\widehat{\mathbf{M}}_1 = 0$ *and* $\eta_t = (2\sigma D_{max})^{1/2}/(G_f\sqrt{T})$*, then*

$$R_{static}([0,T]) \leq G_f(2TD_{max}/\sigma)^{1/2}. \tag{6.23}$$

CHAPTER VII

Conclusion

In conclusion, covariance estimation for array-valued data can be made tractable and efficient by enforcing structure such as sparsity, separation rank deficiency, rank deficiency, bounded nonstationarity, and Kronecker sum structure. In particular, Kronecker structure and strong adaptivity are powerful tools in the modeling of nonstationary spatio-temporal data, and Kronecker models effectively formulate natural spatio-temporal structure and dramatically reduce the required number of training samples. Strong performance bounds were derived, in several cases showing single-sample convergence in high dimensions. In real data applications, orders of magnitude were gained in training sample complexity, resulting in significant performance improvements and enabling greater spatial and temporal resolution of the covariance estimates in highly nonstationary data.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agarwal, A., S. Negahban, M. J. Wainwright, et al. (2012), Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions, *The Annals of Statistics*, *40*(2), 1171–1197.

Allen, G. I., and R. Tibshirani (2010), Transposable regularized covariance models with an application to missing data imputation, *The Annals of Applied Statistics*, *4*(2), 764–790.

Antoine, E., A. Jatowt, S. Wakamiya, Y. Kawai, and T. Akiyama (2015), Portraying collective spatial attention in twitter, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 39–48, ACM, New York, NY, USA, doi:10.1145/2783258.2783418.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008), Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data, *The Journal of Machine Learning Research*, *9*, 485–516.

Barzilai, J., and J. M. Borwein (1988), Two-point step size gradient methods, *IMA Journal of Numerical Analysis*, *8*(1), 141–148.

Bazi, Y., L. Bruzzone, and F. Melgani (2005), An unsupervised approach based on the generalized Gaussian model to automatic change detection in multitemporal SAR images, *Geoscience and Remote Sensing, IEEE Transactions on*, *43*(4), 874–887.

Beck, A., and M. Teboulle (2009), A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM journal on imaging sciences*, *2*(1), 183–202.

Belkacemi, H., and S. Marcos (2006), Fast iterative subspace algorithms for airborne STAP radar, *EURASIP Journal on Advances in Signal Processing*, *2006*.

Bell, R., Y. Koren, and C. Volinsky (2007), Modeling relationships at multiple scales to improve accuracy of large recommender systems, in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 95–104, ACM.

Bellet, A., A. Habrard, and M. Sebban (2013), A survey on metric learning for feature vectors and structured data, *arXiv preprint arXiv:1306.6709*.

Bijma, F., J. C. De Munck, and R. M. Heethaar (2005), The spatiotemporal meg covariance matrix modeled as a sum of kronecker products, *NeuroImage*, *27*(2), 402–415.

Bilenko, M., S. Basu, and R. J. Mooney (2004), Integrating constraints and metric learning in semi-supervised clustering, in *Proceedings of the twenty-first International Conference on Machine learning*, p. 11, ACM.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.

Blum, A., and Y. Mansour (2005), From external to internal regret, in *Learning theory*, pp. 621–636, Springer.

Bonilla, E., K. M. Chai, and C. Williams (2007), Multi-task Gaussian process prediction, in *NIPS*.

Borcea, L., T. Callaghan, and G. Papanicolaou (2013), Synthetic aperture radar imaging and motion estimation via robust principal component analysis, *SIAM Journal on Imaging Sciences*, *6*(3), 1445–1476.

Boyd, S., and L. Vandenberghe (2009), *Convex optimization*, Cambridge university press.

Brennan, L., and F. Staudaher (1992), Subclutter visibility demonstration, *Tech. Rep. RL-TR-92-21*, Adaptive Sensors Incorporated.

Byrne, C. L. (2013), Alternating minimization as sequential unconstrained minimization: a survey, *Journal of Optimization Theory and Applications*, *156*(3), 554–566.

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011), Robust principal component analysis?, *Journal of the ACM (JACM)*, *58*(3), 11.

Canuto, C., V. Simoncini, and M. Verani (2014), On the decay of the inverse of matrices that are sum of kronecker products, *Linear Algebra and its Applications*, *452*, 21–39.

Cerutti-Maori, D., I. Sikaneta, and C. H. Gierull (2012), Optimum sar/gmti processing and its application to the radar satellite radarsat-2 for traffic monitoring, *Geoscience and Remote Sensing, IEEE Transactions on*, *50*(10), 3868–3881.

Cesa-Bianchi, N., and G. Lugosi (2006), *Prediction, learning, and games*, Cambridge University Press.

Chandrasekaran, V., S. Sanghavi, P. Parrilo, and A. Willsky (2009), Sparse and low-rank matrix decompositions, in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pp. 962–967, doi: 10.1109/ALLERTON.2009.5394889.

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2010), Latent variable graphical model selection via convex optimization, in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 1610–1613, IEEE.

Chi, Y. (2016), Kronecker covariance sketching for spatial-temporal data, in *Signal Processing Conference (EUSIPCO), 2016 24th European*, pp. 316–320, IEEE.

Combettes, P. L., and V. R. Wajs (2005), Signal recovery by proximal forward-backward splitting, *Multiscale Modeling & Simulation*, *4*(4), 1168–1200.

Conte, E., and A. De Maio (2003), Exploiting persymmetry for CFAR detection in compound-Gaussian clutter, in *Radar Conference, 2003. Proceedings of the 2003 IEEE*, pp. 110–115, IEEE.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley, New York.

Cristallini, D., D. Pastina, F. Colone, and P. Lombardo (2013), Efficient detection and imaging of moving targets in sar images based on chirp scaling, *Geoscience and Remote Sensing, IEEE Transactions on*, *51*(4), 2403–2416.

Daniely, A., A. Gonen, and S. Shalev-Shwartz (2015), Strongly adaptive online learning, *ICML*.

Davis, J. V., B. Kulis, P. Jain, S. Sra, and I. S. Dhillon (2007), Information-theoretic metric learning, in *Proceedings of the 24th international conference on Machine learning*, pp. 209–216, ACM.

Dawid, A. P. (1981), Some matrix-variate distribution theory: notational considerations and a bayesian application, *Biometrika*, *68*(1), 265–274.

De Munck, J. C., H. M. Huizenga, L. J. Waldorp, and R. M. Heethaar (2002), Estimating stationary dipoles from meg/eeg data contaminated with spatially and temporally correlated background noise, *Signal Processing, IEEE Transactions on*, *50*(7), 1565–1572.

De Munck, J. C., F. Bijma, P. Gaura, C. A. Sieluzycki, M. I. Branco, and R. M. Heethaar (2004), A maximum-likelihood estimator for trial-to-trial variations in noisy meg/eeg data sets, *Biomedical Engineering, IEEE Transactions on*, *51*(12), 2123–2128.

Deckard, A., R. C. Anafi, J. B. Hogenesch, S. B. Haase, and J. Harer (2013), Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data, *Bioinformatics*, *29*(24), 3174–3180.

Diebel, J., and S. Thrun (2005), An application of markov random fields to range sensing, in *NIPS*, vol. 5, pp. 291–298.

Duchi, J. C., E. Hazan, and Y. Singer (2010a), Adaptive subgradient methods for online learning and stochastic optimization, in *COLT*.

Duchi, J. C., S. Shalev-Shwartz, Y. Singer, and A. Tewari (2010b), Composite objective mirror descent, in *COLT*, pp. 14–26, Citeseer.

Dutilleul, P. (1999), The mle algorithm for the matrix normal distribution, *Journal of statistical computation and simulation*, *64*(2), 105–123.

Eckart, C., and G. Young (1936), The approximation of one matrix by another of lower rank, *Psychometrika*, *1*(3), 211–218.

Ender, J. H. (1999), Space-time processing for multichannel synthetic aperture radar, *Electronics & Communication Engineering Journal*, *11*(1), 29–38.

Fa, R., and R. C. De Lamare (2011), Reduced-rank stap algorithms using joint iterative optimization of filters, *Aerospace and Electronic Systems, IEEE Transactions on*, *47*(3), 1668–1684.

Fa, R., R. C. de Lamare, and L. Wang (2010), Reduced-rank STAP schemes for airborne radar based on switched joint interpolation, decimation and filtering algorithm, *Signal Processing, IEEE Transactions on*, *58*(8), 4182–4194.

Faber, N. K. M., R. Bro, and P. K. Hopke (2003), Recent developments in candecomp/parafac algorithms: a critical review, *Chemometrics and Intelligent Laboratory Systems*, *65*(1), 119–137.

Fienup, J. R. (2001), Detecting moving targets in SAR imagery by focusing, *Aerospace and Electronic Systems, IEEE Transactions on*, *37*(3), 794–809.

Gerlach, K., and M. Picciolo (2011), Robust, reduced rank, loaded reiterative median cascaded canceller, *Aerospace and Electronic Systems, IEEE Transactions on*, *47*(1), 15–25, doi:10.1109/TAES.2011.5705656.

Ginolhac, G., P. Forster, J. P. Ovarlez, and F. Pascal (2009), Spatio-temporal adaptive detector in non-homogeneous and low-rank clutter, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2045–2048.

Ginolhac, G., P. Forster, F. Pascal, and J.-P. Ovarlez (2013), Performance of two low-rank STAP filters in a heterogeneous noise, *Signal Processing, IEEE Transactions on*, *61*(1), 57–61, doi:10.1109/TSP.2012.2226162.

Ginolhac, G., P. Forster, F. Pascal, and J. P. Ovarlez (2014), Exploiting persymmetry for low-rank Space Time Adaptive Processing, *Signal Processing*, *97*, 242–251.

Goldberger, J., G. E. Hinton, S. T. Roweis, and R. Salakhutdinov (2004), Neighbourhood components analysis, in *Advances in neural information processing systems*, pp. 513–520.

Goldstein, J., I. S. Reed, and L. Scharf (1998), A multistage representation of the Wiener filter based on orthogonal projections, *Information Theory, IEEE Transactions on*, *44*(7), 2943–2959, doi:10.1109/18.737524.

Greenewald, K., and A. Hero (2014a), Kronecker PCA based spatio-temporal modeling of video for dismount classification, in *Proceedings of SPIE*.

Greenewald, K., and A. Hero (2014b), Regularized block toeplitz covariance matrix estimation via kronecker product expansions, in *Proceedings of IEEE SSP*.

Greenewald, K., and A. Hero (2015), Robust kronecker product pca for spatio-temporal covariance estimation, *Signal Processing, IEEE Transactions on*, *63*(23), 6368–6378, doi:10.1109/TSP.2015.2472364.

Greenewald, K., and A. O. Hero III (2015), Kronecker PCA based robust SAR STAP, *arXiv preprint arXiv:1501.07481*.

Greenewald, K., T. Tsiligkaridis, and A. Hero (2013), Kronecker sum decompositions of space-time data, in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on*, pp. 65–68.

Haimovich, A. (1996), The eigencanceler: adaptive radar by eigenanalysis methods, *Aerospace and Electronic Systems, IEEE Transactions on*, *32*(2), 532–542, doi: 10.1109/7.489498.

Hall, E., and R. Willett (2015), Online convex optimization in dynamic environments, *Selected Topics in Signal Processing, IEEE Journal of*, *9*(4), 647–662.

Harshman, R. A., and M. E. Lundy (1994), Parafac: Parallel factor analysis, *Computational Statistics & Data Analysis*, *18*(1), 39–72.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin (2005), The elements of statistical learning: data mining, inference and prediction, *The Mathematical Intelligencer*, *27*(2), 83–85.

Hazan, E., and C. Seshadhri (2007), Adaptive algorithms for online decision problems, in *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 14.

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

Herbster, M., and M. K. Warmuth (1998), Tracking the best expert, *Machine Learning*, *32*(2), 151–178.

Hoff, P. D., et al. (2016), Equivariant and scale-free tucker decomposition models, *Bayesian Analysis*, *11*(3), 627–648.

Honig, M., and J. Goldstein (2002), Adaptive reduced-rank interference suppression based on the multistage Wiener filter, *Communications, IEEE Transactions on*, *50*(6), 986–994, doi:10.1109/TCOMM.2002.1010618.

Hsieh, C.-J., M. A. Sustik, I. S. Dhillon, and P. D. Ravikumar (2014), Quic: quadratic approximation for sparse inverse covariance estimation., *Journal of Machine Learning Research*, *15*(1), 2911–2947.

Jao, J. K. (2001), Theory of synthetic aperture radar imaging of a moving target, *Geoscience and Remote Sensing, IEEE Transactions on*, *39*(9), 1984–1992.

Jun, S. C., S. M. Plis, D. M. Ranken, and D. M. Schmidt (2006), Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data, *Physics in medicine and biology*, *51*(21), 5549.

Kalaitzis, A., J. Lafferty, N. Lawrence, and S. Zhou (2013), The bigraphical lasso, in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1229–1237.

Kamm, J., and J. Nagy (2000), Opt. kronecker product approx. of block toeplitz matrices, *SIAM Journal on Matrix Analysis and App.*, *22*(1), 155–172.

Khwaja, A. S., and J. Ma (2011), Applications of compressed sensing for sar moving-target velocity estimation and image compression, *Instrumentation and Measurement, IEEE Transactions on*, *60*(8), 2848–2860.

Kirsteins, I., and D. Tufts (1994), Adaptive detection using low rank approximation to a data matrix, *Aerospace and Electronic Systems, IEEE Transactions on*, *30*(1), 55–67, doi:10.1109/7.250406.

Klemm, R. (2002), *Principles of space-time adaptive processing*, 159, IET.

Kolda, T. G., and B. W. Bader (2009), Tensor decompositions and applications, *SIAM review*, *51*(3), 455–500.

Kulis, B. (2012), Metric learning: A survey., *Foundations and Trends in Machine Learning*, *5*(4), 287–364.

Kunapuli, G., and J. Shavlik (2012), Mirror descent for metric learning: a unified approach, in *Machine Learning and Knowledge Discovery in Databases*, pp. 859–874, Springer.

Laub, A. J. (2005), *Matrix analysis for scientists and engineers*, Siam.

Lee, J. A., and M. Verleysen (2007), *Nonlinear dimensionality reduction*, Springer Science & Business Media.

Leskovec, J., D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani (2010), Kronecker graphs: An approach to modeling networks, *Journal of Machine Learning Research*, *11*(Feb), 985–1042.

Loan, C. V., and N. Pitsianis (1993), Approximation with kronecker products, in *Linear Algebra for Large Scale and Real Time Applications*, pp. 293–314, Kluwer Publications.

Loh, P.-L., and M. J. Wainwright (2013), Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima, in *Advances in Neural Information Processing Systems*, pp. 476–484.

Matthews, B. W. (1975), Comparison of the predicted and observed secondary structure of t4 phage lysozyme, *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

Mazumder, R., T. Hastie, and R. Tibshirani (2010), Spectral regularization algorithms for learning large incomplete matrices, *Journal of Machine Learning Research*, *11*, 2287–2322.

McMahan, H. B. (2014), Analysis techniques for adaptive online learning, *arXiv preprint arXiv:1403.3465*.

McMahan, H. B., and M. Streeter (2010), Adaptive bound optimization for online convex optimization, in *COLT*.

Moore, B., R. Nadakutiti, and J. Fessler (2014), Improved robust PCA using low-rank denoising with optimal singular value shrinkage, in *Proceedings of IEEE SSP*.

Nesterov, Y. (1983), A method of solving a convex programming problem with convergence rate o (1/k2), in *Soviet Mathematics Doklady*, vol. 27, pp. 372–376.

Nesterov, Y. (2004), Introductory lectures on convex optimization. applied optimization, vol. 87.

Nesterov, Y., et al. (2007), Gradient methods for minimizing composite objective function, *Tech. rep.*, UCL.

Newstadt, G., E. Zelnio, and A. Hero (2014), Moving target inference with Bayesian models in SAR imagery, *Aerospace and Electronic Systems, IEEE Transactions on*, *50*(3), 2004–2018, doi:10.1109/TAES.2013.130123.

Otazo, R., E. Candès, and D. K. Sodickson (2014), Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components, *Magnetic Resonance in Medicine*, *73*(3), 1125–1136.

Pados, D. A., S. Batalama, G. Karystinos, and J. Matyjas (2007), Short-data-record adaptive detection, in *Radar Conference, 2007 IEEE*, pp. 357–361, IEEE.

Peng, Y., A. Ganesh, J. Wright, W. Xu, and Y. Ma (2010), Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 763–770, doi:10.1109/CVPR.2010.5540138.

Petrović, S., M. Osborne, and V. Lavrenko (2010), Streaming first story detection with application to twitter, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 181–189, Association for Computational Linguistics.

Pitsianis, N. P. (1997), The kronecker product in approximation and fast transform generation, Ph.D. thesis, Cornell University.

Pouryazdian, S., S. Beheshti, and S. Krishnan (2016), Candecomp/parafac model order selection based on reconstruction error in the presence of kronecker structured colored noise, *Digital Signal Processing*, *48*, 12–26.

Rangaswamy, M., F. C. Lin, and K. R. Gerlach (2004), Robust adaptive signal processing methods for heterogeneous radar clutter scenarios, *Signal Processing*, *84*(9), 1653 – 1665, doi:http://dx.doi.org/10.1016/j.sigpro.2004.05.006, special Section on New Trends and Findings in Antenna Array Processing for Radar.

Reed, I., J. Mallett, and L. Brennan (1974), Rapid convergence rate in adaptive arrays, *Aerospace and Electronic Systems, IEEE Transactions on*, *AES-10*(6), 853–863, doi:10.1109/TAES.1974.307893.

Rolfs, B., B. Rajaratnam, D. Guillot, I. Wong, and A. Maleki (2012), Iterative thresholding algorithm for sparse inverse covariance estimation, in *Advances in Neural Information Processing Systems 25*, pp. 1574–1582, Curran Associates, Inc.

Rothman, A. J., P. J. Bickel, E. Levina, J. Zhu, et al. (2008), Sparse permutation invariant covariance estimation, *Electronic Journal of Statistics*, *2*, 494–515.

Rucci, A., S. Tebaldini, and F. Rocca (2010), Skp-shrinkage estimator for sar multibaselines applications, in *Radar Conference, 2010 IEEE*, pp. 701–706, IEEE.

Rudelson, M., and S. Zhou (2015), High dimensional errors-in-variables models with dependent measurements, *arXiv preprint arXiv:1502.02355*.

Scarborough, S. M., et al. (2009), A challenge problem for SAR-based GMTI in urban environments, in *SPIE*, vol. 7337, edited by E. G. Zelnio and F. D. Garber, p. 73370G, [Online]. Available: http://link.aip.org/link/?PSI/7337/73370G/1.

Scharf, L., E. Chong, M. Zoltowski, J. Goldstein, and I. S. Reed (2008), Subspace expansion and the equivalence of conjugate direction and multistage Wiener filters, *Signal Processing, IEEE Transactions on*, *56*(10), 5013–5019, doi:10.1109/TSP.2008.928511.

Settles, B. (2012), Active learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114.

Shen, M., D. Zhu, and Z. Zhu (2009), Reduced-rank space-time adaptive processing using a modified projection approximation subspace tracking deflation approach, *Radar, Sonar Navigation, IET*, *3*(1), 93–100, doi:10.1049/iet-rsn:20080045.

Signorini, A., A. M. Segre, and P. M. Polgreen (2011), The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic, *PLOS ONE*, *6*(5), 1–10, doi:10.1371/journal.pone.0019467.

Srivastava, N., G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014), Dropout: a simple way to prevent neural networks from overfitting., *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Sun, W., Z. Wang, H. Liu, and G. Cheng (2015), Non-convex statistical optimization for sparse tensor graphical model, in *Advances in Neural Information Processing Systems*, pp. 1081–1089.

Tseng, P. (2010), Approximation accuracy, gradient methods, and error bound for structured convex optimization, *Mathematical Programming*, *125*(2), 263–295.

Tsiligkaridis, T., and A. Hero (2013), Covariance estimation in high dimensions via kronecker product expansions, *IEEE Trans. on Sig. Proc.*, *61*(21), 5347–5360.

Tsiligkaridis, T., A. Hero, and S. Zhou (2013), On convergence of kronecker graphical lasso algorithms, *IEEE Trans. Signal Proc.*, *61*(7), 1743–1755, doi: 10.1109/TSP.2013.2240157.

Tucker, L. R. (1966), Some mathematical notes on three-mode factor analysis, *Psychometrika*, *31*(3), 279–311.

Wang, C., N. Komodakis, and N. Paragios (2013), Markov random field modeling, inference & learning in computer vision & image understanding: A survey, *Computer Vision and Image Understanding*, *117*(11), 1610–1627.

Weinberger, K. Q., and L. K. Saul (2008), Fast solvers and efficient implementations for distance metric learning, in *International Conference on Machine Learning*, pp. 1160–1167, ACM.

Weinberger, K. Q., J. Blitzer, and L. K. Saul (2005), Distance metric learning for large margin nearest neighbor classification, in *Advances in Neural Information Processing System*, pp. 1473–1480.

Werner, K., and M. Jansson (2007), Estimation of kronecker structured channel covariances using training data, in *Proceedings of EUSIPCO*.

Werner, K., M. Jansson, and P. Stoica (2008), On estimation of cov. matrices with kronecker product structure, *IEEE Trans. on Sig. Proc.*, *56*(2), 478–491.

Xing, E. P., M. I. Jordan, S. Russell, and A. Y. Ng (2002), Distance metric learning with application to clustering with side-information, in *Advances in Neural Information Processing Systems*, pp. 505–512.

Xu, Z., F. Yan, et al. (2011), Infinite tucker decomposition: Nonparametric bayesian models for multiway data analysis, *arXiv preprint arXiv:1108.6296*.

Yang, E., and P. Ravikumar (2013), Dirty statistical models, in *Advances in Neural Information Processing Systems*, pp. 611–619.

Yang, L., and R. Jin (2006), Distance metric learning: A comprehensive survey, *Michigan State Universiy*, *2*.

Yao, K. (1973), A representation theorem and its applications to spherically-invariant random processes, *Information Theory, IEEE Transactions on*, *19*(5), 600–608.

Yin, J., and H. Li (2012), Model selection and estimation in the matrix normal graphical model, *Journal of Multivariate Analysis*, *107*.

Yu, K., J. Lafferty, S. Zhu, and Y. Gong (2009), Large-scale collaborative prediction using a nonparametric random effects model, in *ICML*, pp. 1185–1192.

Yuan, M., and Y. Lin (2007), Model selection and estimation in the gaussian graphical model, *Biometrika*, *94*(1), 19–35.

Zhang, Y., and J. Schneider (2010), Learning multiple tasks with a sparse matrix-normal penalty, *Advances in Neural Information Processing Systems*, *23*, 2550–2558.

Zhe, S., Z. Xu, X. Chu, Y. A. Qi, and Y. Park (2015), Scalable nonparametric multiway data analysis., in *AISTATS*.

Zhou, S. (2014), Gemini: Graph estimation with matrix variate normal instances, *The Annals of Statistics*, *42*(2), 532–562.

Zhou, S., J. Lafferty, and L. Wasserman (2010a), Time varying undirected graphs, *Machine Learning*, *80*(2-3), 295–319.

Zhou, S., J. Lafferty, and L. Wasserman (2010b), Time varying undirected graphs, *Machine Learning*, *80*(2-3), 295–319.

Zhou, S., P. Rütimann, M. Xu, and P. Bühlmann (2011), High-dimensional covariance estimation based on gaussian graphical models, *The Journal of Machine Learning Research*, *12*, 2975–3026.