Doctoral Dissertation

Study on Learning Invariant Patterns with Second Order Statistics

Satoshi Hara

January 2013

Graduate School of Engineering
Osaka University

JP 16439

Doctoral Dissertation

Study on Learning Invariant Patterns with Second Order Statistics

Satoshi Hara

January 2013

Graduate School of Engineering
Osaka University

Preface

This dissertation presents my research on techniques for learning invariant patterns from multivariate data using second order statistics. The dissertation is the result of the research during the Ph.D. course at the Department of Information and Communication Technology, Graduate School of Engineering, Osaka University. The dissertation is organized as follows.

Chapter 1 describes the background, the motivation, the purpose of this research, and the outline of this dissertation. The key objective of this dissertation is to construct methodologies for finding invariant patterns underlying across multiple datasets sampled from different time points or from several environments. Such techniques allow us to infer the unknown data generating mechanism or to model the target data with an efficient manner. For the purpose, we focus on the second order statistics, one of the most basic parameters representing the properties of multivariate data. In this chapter, we also describe two fundamental models based on the second order statistics, Principal Component Analysis (PCA) model and Graphical Gaussian Model (GGM). These two models form the basis of the dissertation, which we further extend in the upcoming chapters.

Chapter 2 is devoted for the first algorithm that extracts an invariant pattern from the multivariate data. The model we present in this chapter is called Stationary Subspace Analysis (SSA) model and is a specific example of linear source mixing models, that is, a variant of the PCA model. The objective of the SSA problem is to find an invariant pattern across multiple covariance matrices based on a source mixing model. We build a new algorithm Analytic SSA for this problem, which provides a solution by solving a generalized eigenvalue problem. This framework is advantageous compared to an existing algorithm which re-

quires solving a gradient decent based non-convex optimization problem since 1) it requires smaller computational cost, and 2) a global optimal solution can be derived under a certain condition while the prior algorithm guarantees only local optimality of the solution. We also provide theoretical and numerical justifications of this point.

In Chapter 3-5, we describe the second algorithm for discovering an invariant pattern. The major target in these chapters is a GGM, or a conditional dependence structure among random variables. In Chapter 3, we work on convex optimization methods called Dual Augmented Lagrangian (DAL) and Alternating Direction Method of Multipliers (ADMM). We combine the basic idea of these two techniques and formulate the DAL-ADMM algorithm for learning GGM from the data. The advantage of the proposed algorithm is its flexibility. Most existing GGM learning algorithms assume the simplest problem based on an ℓ_1-regularization. On the other hand, our algorithm can treat wider variety of regularization terms including well-known group regularizations. This flexibility is essential for solving more complicated problems arising in Chapter 4 and 5.

In Chapter 4, we consider finding an invariant pattern across multiple GGMs. We formalize the task as a convex optimization problem using sparse regularization techniques, where the proposed formulation can be casted as a generalization of existing GGM learning problems. We also show that the problem can be solved by the DAL-ADMM algorithm. The proposed algorithm is composed of iterative updating steps with each step requiring only simple analytic operations. The validity of the proposed method is verified through numerical simulations and also on an application to an anomaly localization problem.

Chapter 5 describes an anomaly localization problem based on a GGM learning technique. In this chapter, we consider a GGM learning algorithm specialized to this task. One basic finding is that, in an anomaly localization, row/column-wise changes between two precision matrices, or the inverse of covariance matrices, are important. We im-

port this idea and formalize the task as a convex optimization problem. The proposed formulation is a variant of structured sparsity models and requires specific considerations to construct an algorithm. We find that some proper transformations of the problem allow us to treat the problem with DAL-ADMM. Hence, the proposed algorithm requires only simple analytic updating steps. We verify the advantage of our new formulation over existing techniques on an anomaly localization task through a real world data simulation.

Chapter 6 concludes this dissertation.

Acknowledgment

I worked on this dissertation under the supervision of Prof. Takashi Washio. I would first like to thank him for sharing his knowledge, pushing me to work hard and constantly try to improve my work, and giving me the freedom to explore research themes from diverse fields.

I would also like to thank my other co-authors for their help in letting me more productive that I would have been able to on my own: Dr. Yoshinobu Kawahara in Osaka University, Mr. Paul von Bünau in Berlin Institute of Technology, Dr. Terumasa Tokunaga in Research Organization for Information Science and Technology, and Prof. Kiyohumi Yumoto in Kyushu University. My acknowledgment also goes to Dr. Tsuyoshi Idé in IBM Research Tokyo for providing sensor error dataset.

I would like to acknowledge Prof. Tetsuya Takine and Assoc. Prof. Kouji Kozaki in Osaka University for their advice, which were very helpful to revise and improve this dissertation.

During semesters, I had several supports from Prof. Kyo Inoue, Prof. Zenichiro Kawasaki, Prof. Kenichi Kitayama, Prof. Seiichi Sampei, Prof. Tetsuya Takine, Prof. Noboru Babaguchi, and Prof. Riichiro Mizoguchi in Graduate School of Engineering, Osaka University. I would like to acknowledge all of them for their help.

I also had several supports from all of laboratory members. In particular, I would like to thank Ms. Hiroko Okada for her clerical assistance.

Finally and most importantly, I would like to give special thanks to my parents Hiroshi and Kazue, and to all of my friends for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Invariance of the data behavior is a cornerstone assumption in several fields, including statistical learning (Quiñonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2008), signal processing, and control theory. This allows us to model the target problem with simpler formulations which we can manipulate easily. The most well known example of this would be an *independent and identically distributed* (i.i.d.) assumption in statistics. This removes complicated interactions among observations and we can treat each observed sample individually. Another example is a stationarity of a time series. Under the stationarity assumption, we can safely apply the current knowledge to the future prediction. However, these are not always the case in reality and there are several real-world data that changes their behaviors. In such cases, a non-stationary data generating mechanism affects data to have different behaviors in each datasets collected under different conditions, for instance, datasets from several time stamps or the ones sampled under multiple environments. Examples include biomedical measurements (Shenoy, Krauledat, Blankertz, Rao, & Müller, 2006; Blankertz et al., 2008), geophysical data (Mann, 2004; Kaufmann & Stern, 2002), and econometric time series (Engle & Granger, 1987).

Dynamical effects are not just a nuisance for methodology. In fact, understanding temporal changes in data is often the one major point of interest, so that discovering and describing non-stationarities in high-dimensional datasets are a key challenge in explorative data analysis. For instance, there are various approaches to test (Priestley & Rao, 1969; Dickey & Fuller, 1979) and correct for (Quiñonero-Candela et al., 2008; Shimodaira, 2000; Heckman, 1979; Murata, Kawanabe, Ziehe,

Müller, & Amari, 2002) non-stationarities in statistical model. A question arises, however, when the observable data is composed of contributions from both invariant and dynamical effects that are not directly accessible. One naive way is to treat the data as a fully dynamical one since it includes effects from such factors. However, this is not a convenient approach since we discard the fact there exists something invariant in the data, which tends to require complicated modeling. This is unfavorable not only from the methodological aspect, but also from the data analysis perspective. Discerning invariant factors from dynamical ones in the data itself can be one important goal of the analysis. For instance, in electroencephalography (EEG) (Dornhege, Millán, Hinterberger, McFarland, & Müller, 2007), measurements on the scalp capture the activity of a multitude of sources located inside the brain that we cannot measure directly, for technical, medical, or ethical reasons. The observed signal is non-stationary due to the inherent non-stationary dynamics in the brain. However, the EEG signal is not totally contaminated with the non-stationary effects but also reflecting several systematic behaviors in the brain such as kinematic signals. It is natural to assume such systematic behaviors result in some fixed wave forms hidden in the non-stationary observations, which forms an invariant factor across multiple EEG observations sampled under different environments. Finding this hidden wave form in the signal is an important step towards a brain computer interfacing to reflect users intent to a computer control.

In practice, this kinds of underlying partial dynamics of data are captured in several approaches. One way is to explicitly parametrize the invariant and dynamical part in the model (Hamilton, 1994; Durbin & Koopman, 2001). However, this approach tends to require detailed domain knowledge to construct a right model that is scarcely available in most cases. Another way is to impose general and mild assumption on the data. This kind of approach is especially common in multitask learning literatures (Caruana, 1997; Turlach, Venables, & Wright, 2005). In the multitask learning, we exchange the information of each dataset through an invariant factor among them. This allows us to combine multiple tasks into a single problem which efficiently capture the nature of datasets. This dissertation especially focuses on the latter context where the invariant pattern among datasets itself is the objective we want to analyze.

1.2 Main Issues

This dissertation aims to construct methodologies for finding invariant patterns underlying across multiple datasets sampled from different time points or from several environments. A key object for the purpose across the dissertation is the second order statistics obtained from multiple datasets and related Gaussian expressions. We focus on two representative models regarding the second order statistics and introduce the notion of invariance for both of them.

First, we consider a linear mixture model and its relevant invariance. The main problem is to recover latent sources from observations under the linear mixture model. The most prominent example of this would be a Principal Component Analysis (PCA) (Jolliffe, 1986), which uses a sample covariance to derive the solution. We treat an extension of PCA called Stationary Subspace Analysis (SSA) (von Bünau, Meinecke, Király, & Müller, 2009a) where there are two kinds of latent sources which are stationary and non-stationary. The objective of SSA is to recover stationary sources from observations. This is the first invariance we seek for.

Next, we consider the second problem based on a Graphical Gaussian Model (GGM) (Lauritzen, 1996). GGM is defined using an inverse of a covariance matrix, which provides a different perspective to the second order statistics from PCA. The objective of a GGM learning is to infer a graph structure that represents conditional independence relations among random variables. We consider the case when some topological patterns and edge weights are shared between multiple GGMs. We search for this shared pattern in this problem.

We propose algorithms to solve these problems in this dissertation. In Chapter 2, we consider the SSA problem. We propose an Analytic SSA algorithm which is very efficient compared to an existing technique. We also provide detailed discussion regarding the connection of Analytic SSA to well known Independent Component Analysis (ICA) (Hyvärinen, Karhunen, & Oja, 2001) techniques. In Chapter 3-5, we treat the second problem and describe algorithms for finding invariant patterns. Chapter 3 is a preparation for the upcoming Chapter 4 and 5. In this chapter, we work on convex optimization methods called Dual Augmented Lagrangian (DAL) and Alternating Direction Method of Multipliers (ADMM). We combine the basic

idea of these two techniques and formulate the DAL-ADMM algorithm for learning a graphical model. In Chapter 4, we consider the most basic invariant pattern in multiple GGMs. We formulate the task as a convex optimization problem using an ℓ_1 and a group regularization techniques, which we refer as Common Substructure Learning (CSSL). We also show that CSSL can be solved by the DAL-ADMM algorithm. In Chapter 5, we treat an invariant pattern different from the previous chapter, an invariance specific to the anomaly localization problem in sensor signals. Chapter 2-5 are based on (and extend) existing work. In particular, Chapter 2 is based on Hara, Kawahara, Washio, and von Bünau (2010) and Hara, Kawahara, Washio, von Bünau, Tokunaga, and Yumoto (2012), Chapter 3 is based on Hara and Washio (2012b), Chapter 4 is based on Hara and Washio (2011, 2013), Chapter 5 is based on Hara and Washio (2012a).

In the remainder of this chapter, we introduce basic models used across this dissertation. First, we review the second order statistics and the related Gaussian expression of data. We also mention extending this Gaussian expression to multiple datasets, especially for the case of a time series signal. In the sequential two sections, we review two representative models regarding the second order statistics, one is the Principal Component Analysis, the most well known linear mixture model, and the other is the Graphical Gaussian Model, which is one of the most basic graphical model. We also mention that these two expressions are in some sense dual to each other. Finally, we conclude the chapter with a summary of contributions[1].

1.3 Second Order Statistics and Gaussian Distribution

1.3.1 Covariance and Inverse Covariance

In the analysis of multivariate data, the interaction of random variables is the one biggest interest of users. Here, across the dissertation, we consider that a multi-

[1]Each chapter also have an appendix after the concluding section. Some additional remarks and proofs of theorems are described in there.

dimensional random variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ is continuous and is defined on a space $\mathbb{R}^d$. Hence, its distribution is expressed by $p(\boldsymbol{x})$. The most basic property of interactions among continuous random variables is captured by a *covariance* of two variables x_i and x_j defined as

$$\mathrm{Cov}(x_i, x_j) \equiv \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\,(x_j - \mathbb{E}[x_j])\right],$$

where $\mathbb{E}$ denotes an expectation over $p(\boldsymbol{x})$. This measurement is positive if two variables are simultaneously increasing, that is, the greater value of one variable corresponds with the greater value of the other. On the other hand, if one variable gets larger and the other one gets smaller in the same time, the covariance is negative. The zero covariance case is intermediate between these two cases when two variables do not show linear dependencies to each other. For general d-dimensional variable $\boldsymbol{x}$, there exists $\mathcal{O}(d^2)$ combinations of variables and the resulting covariances are represented conveniently in a single matrix $\Sigma \in \mathbb{R}^{d\times d}$ defined as

$$\Sigma \equiv \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])\,(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^\top\right].$$

This matrix Σ is called *covariance matrix* and its (i,j)th entry corresponds to the covariance of x_i and x_j. Importantly, this matrix is symmetric and positive semidefinite from its definition. Note that a variance of x_i, or $\mathrm{Var}(x_i)$, corresponds to the covariance with its own $\mathrm{Var}(x_i) = \mathrm{Cov}(x_i, x_i)$. Hence, the diagonal entries of Σ correspond to the variance of each variable.

Although the sign of covariance is instructive to see how two variables interact, its magnitude heavily depends on the scaling of each variable and is not always meaningful. *Correlation* is an useful alternative to interpret the magnitude of dependencies, which is given by

$$\mathrm{Corr}(x_i, x_j) \equiv \frac{\mathrm{Cov}(x_i, x_j)}{\sqrt{\mathrm{Var}(x_i)\mathrm{Var}(x_j)}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

This is a scaled version of a covariance and its domain is $[-1, 1]$. As the magnitude of the correlation grows, the linear dependency between two variables gets stronger. Hence, the value 1 or -1, two extreme cases of a correlation, implies two variables x_i and x_j are completely linearly dependent to each other: there exists a non-zero

constant c and $x_i = cx_j$ holds where the sign of c corresponds to the sign of a correlation.

Despite its usefulness, the drawback of covariance appears when there are more than two variables. In this situation, covariance captures not only the interaction of two variables but the effects from other variables indirectly. We introduce one simple example describing this drawback through an elementary school children data. Suppose we held examinations for all children in one school and collected data containing three fields: children's age; height; and their test scores. It is obvious that, as children grows, they gets tall. But not only that, they learn more and there test score gets well also. This results in a positive covariance between the height and the score. However, the fact that taller students mark higher scores is against our intuition. This happens because the effect of the age is involved in the covariance between the height and the score. Therefore, we have to remove such indirect effects to observe the essential dependency of two target variables. This leads to the idea of *partial correlation*. Let $\boldsymbol{x}_{\setminus\{i,j\}}$ denote $d-2$ variables in $\boldsymbol{x}$ except x_i and x_j. In partial correlation, the effects of the third variable $\boldsymbol{x}_{\setminus\{i,j\}}$ in x_i and x_j are modeled as a linear function:

$$x_i = r_i + \boldsymbol{w}_i^\top \boldsymbol{x}_{\setminus\{i,j\}},$$
$$x_j = r_j + \boldsymbol{w}_j^\top \boldsymbol{x}_{\setminus\{i,j\}},$$

with some $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$. Here, r_i and r_j are random variables and are statistically independent of $\boldsymbol{x}_{\setminus\{i,j\}}$. This r_i and r_j are essential part of x_i and x_j after removing the effect of the third variable $\boldsymbol{x}_{\setminus\{i,j\}}$. We measure the essential dependency of two variables x_i and x_j as a correlation of r_i and r_j since the effects of the third variable are no longer involved. The following is the definition of a partial correlation between x_i and x_j:

$$\mathrm{PCorr}(x_i, x_j | \boldsymbol{x}_{\setminus\{i,j\}}) \equiv \mathrm{Corr}(r_i, r_j) = \frac{\mathrm{Cov}(r_i, r_j)}{\sqrt{\mathrm{Var}(r_i)\mathrm{Var}(r_j)}}.$$

Hence, the next theorem tells the important connection of a partial correlation to the inverse of a covariance matrix $\Lambda = \Sigma^{-1}$ which is also known as *precision matrix*.

Theorem 1 (Partial Correlation and Precision Matrix (Lauritzen, 1996)). *A partial correlation between x_i and x_j given remaining $d-2$ variables $\boldsymbol{x}_{\setminus\{i,j\}}$ relates to each entry of a precision matrix Λ by*

$$\mathrm{PCorr}(x_i, x_j | \boldsymbol{x}_{\setminus\{i,j\}}) = -\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}}. \tag{1.1}$$

From this result, we can interpret the precision matrix Λ as an unnormalized version of the partial correlation analogous to the relationship between the covariance and the correlation.

In the remainder of the dissertation, we call a covariance matrix Σ and an precision matrix Λ as *second order statistics* since both of them are defined on the second order moment of the probability distribution.

1.3.2 Gaussian Distribution

In data analysis, we often convert observed data into some kind of probability distributions. This allows us to use powerful methods to analyze data more intensively. Here, we assume the mean and the covariance of $\boldsymbol{x}$ is known as $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$ and Σ respectively, which is quite realistic as we see later. The question is what is the most appropriate probability distribution we can use for the data analysis under this situation. One answer to this question is to pick up the distribution with the highest uncertainty, which is also known as the *maximum entropy principle*. This result suggests that a Gaussian distribution is an useful representation of data when the statistics up to second order moments are known.

Theorem 2 (Maximum Entropy Principle (Jaynes, 1957)). *Given mean $\boldsymbol{\mu}$ and covariance matrix Σ, a probability distribution $p(\boldsymbol{x})$ that maximize the following entropy*

$$H(p) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x})\, d\boldsymbol{x},$$

is a Guassian distribution given by

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma) \equiv \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right).$$

The remaining step is to derive the mean $\boldsymbol{\mu}$ and the covariance Σ from the dataset. *Maximum likelihood estimation* is the well-known approach for this problem. We suppose n data points $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ are i.i.d. samples from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with unknown parameters $\boldsymbol{\mu}$ and Σ. The log-likelihood function on the dataset $\mathcal{D}$ is then given by

$$\begin{aligned}\log p(\mathcal{D}; \boldsymbol{\mu}, \Sigma) &= \log \left\{ \prod_{n=1}^N \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right) \right\} \\ &= -\frac{1}{2} \sum_{n=1}^N (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) - \frac{N}{2} \log \det \Sigma - \frac{Nd}{2} \log 2\pi.\end{aligned}$$

We find parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ that maximize this log-likelihood, which are parameters that best fit to the data. First, by setting the derivative over $\boldsymbol{\mu}$ equal to zero, we derive

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^N \boldsymbol{x}_n \,. \tag{1.2}$$

We then optimize the log-likelihood over Σ and derive

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^N (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})^\top \,. \tag{1.3}$$

These two results are the maximum likelihood estimators of the mean and the covariance matrix in a Gaussian distribution, which is also known as an empirical or sample mean and covariance, respectively.

Similarly, we can conduct the maximum likelihood estimation of a precision matrix Λ from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Lambda^{-1})$. Here, the covariance matrix of the distribution is replaced with an inverse of Λ since $\Lambda = \Sigma^{-1}$ from its definition. Again, writing down the log-likelihood function and maximizing it, we derive the maximum likelihood estimators as (1.2) and

$$\hat{\Lambda} = \hat{\Sigma}^{-1}. \tag{1.4}$$

1.3.3 Gaussian Expression of Multiple Datasets

In the previous section, we considered how to approximate a single dataset with a Gaussian distribution. Here, we extend it to a multiple datasets situation. This

extension naturally arises in several studies. For instance, in multitask learning (Caruana, 1997; Turlach et al., 2005), we face a problem of jointly solving multiple tasks where each task has its own distribution. In other example, we need to deal with datasets sampled from different time stamps for a source separation (Matsuoka, Ohoya, & Kawamoto, 1995; Kawamoto, Matsuoka, & Ohnishi, 1998; Pham & Cardoso, 2001; Hyvarinen, 2002; Parra & Sajda, 2003) or for a change point detection (Basseville & Nikiforov, 1993; Siegmund & Venkatraman, 1995; Kohlmorgen, Lemm, Müller, Liehr, & Pawelzik, 1999).

Suppose there are K datasets each consists of i.i.d. data points sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ for $k = 1, 2, \ldots, K$. If we know each dataset have no relations to each other, then we can use an ordinal maximum likelihood estimation for each dataset individually and derive its first and second order statistics as (1.2), (1.3), and (1.4). This is the most simple and naive situation and no special consideration is needed. However, in several real world applications, we know that there exists some kinds of relations among datasets a priori. In such cases, cooperating with user's prior knowledge helps us to improve the data analysis performance. The important point is how to import such knowledge into the naive formulation and improve it. This is the main point across this dissertation and details are described in remaining chapters.

Amongst several applications, one important example of a multiple Gaussian expression is the representation of a time series data. Formally, a time series data is a sequence of data points indexed by a time stamp t and is expressed as $\mathcal{D} = \{\boldsymbol{x}(t)\}_{t=1}^{T}$. Unlike i.i.d. samples, the distribution of a data point $\boldsymbol{x}(t)$ usually have some dependencies on the past data $\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(t-1)$. Another difference is that a time series data can even be non-stationary: the distribution of data itself may change over time. This prohibits us from modeling data as a simple Gaussian distribution. A simple alternative to this problem is to represent a time series as a set of Gaussian distributions. First, we partition a given time series into a set of epochs $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{K}$, $\mathcal{D}_k = \{\boldsymbol{x}(t)\}_{t \in \mathcal{T}_k}$. Here, $\mathcal{T}_k$ is a set of consecutive indices and $\cup_{k=1}^{K} \mathcal{T}_k = \{1, 2, \ldots, T\}$. We then approximate each epoch $\mathcal{D}_k$ with a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$. This corresponds to capturing the dynamics of data as a set of local static distributions. The remaining problem is how to

derive distribution parameters $\boldsymbol{\mu}_n$ and Σ_n from each epoch. In time series data, an ordinary maximum likelihood estimation on i.i.d. samples are no longer useful because of its time dependency structure. However, we can use time average as its alternative under an *ergodicity* condition. An ergodicity is a condition that a population statistic and a time average of a time series meets:

$$\int f(\boldsymbol{x}(t))p(\boldsymbol{x}(t))\,d\boldsymbol{x}(t) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T} f(\boldsymbol{x}(t)),$$

for a given function f[2]. See Hamilton (1994) for further detail. In the current case, a function f is chosen to produce the mean and the covariance. Practically, we only have finite number of samples and therefore we replace the right hand side with their average:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{|\mathcal{T}_k|}\sum_{t\in\mathcal{T}_k}\boldsymbol{x}(t),$$
$$\hat{\Sigma}_k = \frac{1}{|\mathcal{T}_k|}\sum_{t\in\mathcal{T}_k}(\boldsymbol{x}(t)-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}(t)-\hat{\boldsymbol{\mu}}_k)^\top.$$

Using this technique, we can treat a time series data as a set of Gaussian distributions.

1.4 Principal Component Analysis

In this and the next section, we present two most basic models regarding the second order statistics. We first introduce a Principal Component Analysis, or PCA (Jolliffe, 1986) in this section.

Before we present the detail of PCA, we assume that the data is centered, that is, a random variable $\boldsymbol{x} \in \mathbb{R}^d$ has a zero vector $\mathbf{0}_d$ as its mean. Note that we can always transform data to follow this assumption by subtracting sample mean (1.2) from each data point. Therefore, the following discussion can be naturally extended to non-zero mean situations, although we adopt this assumption for simplicity. It also allows us to focus on the role of the second order statistics which is the central target of our analysis in this dissertation.

[2]We assume that the right hand side of this condition exists.

1.4.1 PCA

In the PCA model, or more generally in a linear source mixing model, an i.i.d. observation $\boldsymbol{x}_n$ is modeled as a linear superposition of a m-dimensional ($m < d$) latent variable $\boldsymbol{s}_n$:

$$\boldsymbol{x}_n = A\boldsymbol{s}_n, \tag{1.5}$$

with some matrix $A \in \mathbb{R}^{d\times m}$. It corresponds to assuming that all data points $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ are not fully distributed in $\mathbb{R}^d$ but lie in some low-dimensional subspace in $\mathbb{R}^d$. The objective of PCA is to recover these unknown parameters A and $\{\boldsymbol{s}_n\}_{n=1}^N$ from the data $\mathcal{D}$. We note that there is a linear transformation invariance in this model that the replacement $A \to \hat{A} = AR^{-1}$ and $\boldsymbol{s}_n \to \hat{\boldsymbol{s}}_n = R\boldsymbol{s}_n$ with an arbitrary non-singular matrix $R \in \mathbb{R}^{m\times m}$ produces the same model as (1.5). Therefore, we can restrict ourselves to the orthonormal case $A^\top A = I_m$ without loss of generality where I_m denotes an $m \times m$ identity matrix. This corresponds to limiting R to control only the rotation of a coordinate but not the scaling of $\boldsymbol{s}_n$. Hence, we have an equation

$$\boldsymbol{x}_n = AA^\top \boldsymbol{x}_n,$$

from the model (1.5).

In the above equation, we no longer need to consider the latent variable $\boldsymbol{s}_n$ and can focus only on finding a matrix A that satisfies this condition. Note that however, in practice, data does not follow the model (1.5) exactly, and the above equation holds only approximately. We therefore find a matrix A that minimizes the discrepancy between the left and the right hand side of the above equation. To that end, we adopt the following square metric:

$$\frac{1}{N}\sum_{n=1}^N \left\|\boldsymbol{x}_n - AA^\top \boldsymbol{x}_n\right\|_2^2 = \frac{1}{N}\sum_{n=1}^N \left(\left\|\boldsymbol{x}_n\right\|_2^2 - \left\|A^\top \boldsymbol{x}_n\right\|_2^2\right), \tag{1.6}$$

where we used the orthonormality of A for the equality. Since the first term is independent of A, the resulting optimization problem is summarized as follows:

$$\max_{A^\top A = I_m} \frac{1}{N}\sum_{n=1}^N \left\|A^\top \boldsymbol{x}_n\right\|_2^2 = \max_{A^\top A = I_m} \mathrm{tr}\left[A^\top \hat{\Sigma} A\right], \tag{1.7}$$

where $\hat{\Sigma}$ is the sample covariance matrix given in (1.3).

The solution to the problem (1.7) can be derived using a method of Lagrange multipliers. We rewrite the problem using a Lagrange multiplier $\Gamma \in \mathbb{R}^{m \times m}$ as

$$\max_{A} \min_{\Gamma} \operatorname{tr}\left[A^\top \hat{\Sigma} A\right] - \operatorname{tr}\left[\Gamma\left(A^\top A - I_m\right)\right].$$

By setting the derivative over A equal to zero, we obtain

$$\hat{\Sigma} A - A\Gamma = 0_{d \times m}.$$

Recall that we can always transform A into AR^{-1} with an arbitrary rotation matrix R, we rewrite this equation as

$$\hat{\Sigma} A - AR^{-1}\Gamma R = 0_{d \times m}.$$

Hence, for any Γ, we can always choose a matrix R so that $R^{-1}\Gamma R$ is a diagonal matrix. Here, we define $R^{-1}\Gamma R = \operatorname{diag}(\gamma_1, \gamma_2, \ldots, \gamma_m)$ and $A = \left[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\right]$. The above equation can then be decomposed into m eigenvalue problems:

$$\hat{\Sigma} \boldsymbol{a}_i = \gamma_i \boldsymbol{a}_i \quad (i = 1, 2, \ldots, m).$$

Moreover, from the orthonormality of $\boldsymbol{a}_i$, we have $\gamma_i = \boldsymbol{a}_i^\top \hat{\Sigma} \boldsymbol{a}_i$ and the problem reduces to finding $\gamma_1, \gamma_2, \ldots, \gamma_m$ that maximizes the following objective function:

$$\operatorname{tr}\left[A^\top \hat{\Sigma} A\right] = \sum_{i=1}^{m} \gamma_i,$$

while keeping the orthonormality of A. Obviously, the top m eigenvalues and their corresponding eigenvalues are the solutions of $\gamma_1, \gamma_2, \ldots, \gamma_m$ and $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$, respectively. Once the matrix A is estimated with the above procedure, we can recover latent variables as

$$\boldsymbol{s}_n = A^\top \boldsymbol{x}_n,$$

which follows from the model (1.5) using the orthonormality of A.

This result indicates that leading eigenvalues and eigenvectors of the sample covariance matrix are essential for approximating data points with a low-dimensional expression (1.5). We note that (1.5) is a general model and we can construct

some different models other than PCA. The fundamental differences are the objective function (1.6). For instance, if we introduce a maximum correlation criteria, we derive a Canonical Correlation Analysis (Mardia, Kent, & Bibby, 1979), while introducing an independence criteria leads to an Independent Component Analysis (Hyvärinen et al., 2001). An extension of PCA into a supervised learning literature produces the idea of a Linear Discriminant Analysis (Fisher, 1936; Fukunaga, 1990). In Chapter 2, we introduce a stationarity criteria and derive Stationary Subspace Analysis (SSA) and an algorithm that finds an invariance in the second order statistics.

1.4.2 PCA as Matrix Approximation

In the previous section, we derived a solution of PCA from a minimum projection length criteria (1.6). Here, we derive the PCA solution from a matrix approximation point of view. The objective of a matrix approximation is to derive a matrix $S \in \mathbb{R}^{d \times d}$ that best describes the nature of $\hat{\Sigma}$ from some set of matrices $\mathcal{S}$. A set $\mathcal{S}$ can be usually a set of low-rank matrices (Srebro, Rennie, & Jaakkola, 2005) or a set of sparse matrices (Zou, Hastie, & Tibshirani, 2006) depending on the application. Here, we let $\mathcal{S}$ be a set of rank m matrices defined as $\mathcal{S} = \{S \in \mathbb{R}^{d \times d}; \mathrm{rank}(S) = m\}$. We then find a matrix S that is closest to $\hat{\Sigma}$:

$$\min_{S \in \mathcal{S}} \left\| \hat{\Sigma} - S \right\|_{\mathrm{F}}^2 \tag{1.8}$$

where $\|*\|_{\mathrm{F}}$ denotes a Frobenius norm[3] of a matrix.

To solve the minimization problem above, we derive the lower bound of the objective function (1.8). To begin with, we rewrite the Frobenius norm into the following equivalent form:

$$\begin{aligned} \left\| \hat{\Sigma} - S \right\|_{\mathrm{F}}^2 &= \left\| \hat{\Sigma} \right\|_{\mathrm{F}}^2 - 2\,\mathrm{tr}\left[\hat{\Sigma} S\right] + \left\| S \right\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{d} \sigma_i(\hat{\Sigma})^2 - 2\,\mathrm{tr}\left[\hat{\Sigma} S\right] + \sum_{i=1}^{d} \sigma_i(S)^2, \end{aligned}$$

[3] A Frobenius norm of a matrix A is given by $\|A\|_{\mathrm{F}} \equiv \sqrt{\sum_{i,j} A_{ij}^2}$.

where $\sigma_i(\hat{\Sigma})$ and $\sigma_i(S)$ denote the ith eigenvalues of $\hat{\Sigma}$ and S, respectively. Now, we can apply the von Neumann's trace theorem (Horn & Johnson, 1990) to the second term and derive the lower bound as

$$\begin{aligned}\left\|\hat{\Sigma} - S\right\|_{\mathrm{F}}^2 &\geq \sum_{i=1}^{d} \sigma_i(\hat{\Sigma})^2 - 2\sum_{i=1}^{d} \sigma_i(\hat{\Sigma})\sigma_i(S) + \sum_{i=1}^{d} \sigma_i(S)^2 \\ &= \sum_{i=1}^{d} \Big(\sigma_i(\hat{\Sigma}) - \sigma_i(S)\Big)^2 .\end{aligned}$$

Note that the equality holds when the eigenvectors of $\hat{\Sigma}$ and S are the same. Since the rank of S is limited to m, we have to find the m non-zero eigenvalues of S that minimizes this lower bound. It is obvious that choosing the top m eigenvalues and setting remaining $d - m$ values to be zeros minimizes the bound, which is the same result as PCA derived in the previous section.

From this rank restricted matrix approximation problem, we find that leading eigenvalues and eigenvectors are essential to approximate the matrix. Note that the above discussion holds not only for a sample covariance but for general square matrices[4].

1.5 Graphical Gaussian Model

In the previous section, we observed that the larger eigenvalues of the sample covariance matrix play an important role in PCA. Here, we introduce another important model based on the second order statistics, a Graphical Gaussian Model (GGM).

A graphical model (Lauritzen, 1996) represents a dependency structure among multiple random variables. There are two types of it, a directed model and an undirected one. GGM belongs to one specific case of the latter one, a pairwise undirected graphical model. Therefore, we briefly introduce a pairwise undirected graphical model first and GGM in the next.

[4]It can also be generalized to rectangular matrices using a singular value decomposition.

1.5.1 Pairwise Undirected Graphical Model

Undirected graphical model, which is also known as a Markov random field, represents conditional dependency structure among random variables using a graph. Here, we focus on one specific case, a pairwise undirected graphical model. In the pairwise undirected graphical model, a graph is composed of nodes corresponding to each random variable x_i and a set of edges E spanning between random variables. Using this edge set, a distribution of a random variable $\boldsymbol{x} \in \mathbb{R}^d$ is modeled as a product of non-negative potential functions $\phi_i(x_i)$ and $\phi_{ij}(x_i, x_j)$:

$$p(\boldsymbol{x}) \equiv \frac{1}{Z} \prod_{i=1}^{d} \phi_i(x_i) \prod_{(i,j)\in E} \phi_{ij}(x_i, x_j), \tag{1.9}$$

where Z is a normalization constant defined as

$$Z \equiv \int \prod_{i=1}^{d} \phi_i(x_i) \prod_{(i,j)\in E} \phi_{ij}(x_i, x_j)\, d\boldsymbol{x}.$$

This model is called pairwise since it is defined on each pair of variables and no higher order effects exist. Here, we assume that a pairwise potential $\phi_{ij}(x_i, x_j)$ cannot be expressed as the product of two unary functions. If this is not the case, we can include such unary functions into $\phi_i(x_i)$ and $\phi_j(x_j)$ and remove the corresponding index pairs from E without loss of generality. We also note that a constant function $\phi_{ij}(x_i, x_j) = c$ with some constant c is the special case of the above discussion, so that the valid pairwise potential function is a non-constant, non-decomposable one.

The model (1.9) implies that we can express the conditional distribution over x_i and x_j given remaining $d-2$ variables $\boldsymbol{x}_{\setminus\{i,j\}}$ fixed as

$$\begin{aligned} p(x_i, x_j|\boldsymbol{x}_{\setminus\{i,j\}}) &= \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}_{\setminus\{i,j\}})} \\ &= \frac{\prod_{i'=1}^{d} \phi_{i'}(x_{i'}) \prod_{(i',j')\in E} \phi_{i'j'}(x_{i'}, x_{j'})}{\int \prod_{i'=1}^{d} \phi_{i'}(x_{i'}) \prod_{(i',j')\in E} \phi_{i'j'}(x_{i'}, x_{j'})\, dx_i dx_j} \\ &= \frac{f_i(x_i|\boldsymbol{x}_{\setminus\{i,j\}}) f_j(x_j|\boldsymbol{x}_{\setminus\{i,j\}}) g_{ij}(x_i, x_j)}{\int f_i(x_i|\boldsymbol{x}_{\setminus\{i,j\}}) f_j(x_j|\boldsymbol{x}_{\setminus\{i,j\}}) g_{ij}(x_i, x_j)\, dx_i dx_j}, \end{aligned}$$

where

$$f_i(x_i|\boldsymbol{x}_{\setminus\{i,j\}}) = \phi_i(x_i)\prod_{j'\neq j}^{d}\phi_{ij'}(x_i,x_{j'}),$$

$$f_j(x_j|\boldsymbol{x}_{\setminus\{i,j\}}) = \phi_j(x_j)\prod_{i'\neq i}^{d}\phi_{i'j}(x_{i'},x_j),$$

$$g_{ij}(x_i,x_j) = \begin{cases}\phi_{ij}(x_i,x_j) & \text{if } (i,j)\in E,\\ 1 & \text{if } (i,j)\notin E.\end{cases}$$

If $(i,j)\notin E$, we can further transform the expression into

$$\begin{aligned}p(x_i,x_j|\boldsymbol{x}_{\setminus\{i,j\}}) &= \frac{f_i(x_i|\boldsymbol{x}_{\setminus\{i,j\}})}{\int f_i(x_i|\boldsymbol{x}_{\setminus\{i,j\}})\,dx_i}\frac{f_j(x_j|\boldsymbol{x}_{\setminus\{i,j\}})}{\int f_j(x_j|\boldsymbol{x}_{\setminus\{i,j\}})\,dx_j}\\ &= p(x_i|\boldsymbol{x}_{\setminus\{i,j\}})p(x_j|\boldsymbol{x}_{\setminus\{i,j\}}),\end{aligned}$$

where the last equality follows from the model (1.9) and the definition of a conditional distribution. This result is exactly the definition of a conditional independence, and is one specific example of Hammersley-Clifford theorem (Clifford, 1990). It suggests that as long as potential functions $\phi_{ij}(x_i,x_j)$ are not constants nor decomposable, pairs of random variables indicated by an edge set E coincides with a set of conditionally dependent variables pairs. Or alternatively, we can say that the absence of an edge between two random variables implies these variables are conditionally independent. Because of this property, a pairwise undirected model is a useful tool to model the pairwise conditional independence between random variables. As we mentioned before, GGM is one of the most well known example of this model. Note that Ising model (Lauritzen, 1996), a well known distribution on binary variables, also belongs to this class.

Two extreme cases of the model (1.9) are the fully connected graph and the fully disjoint graph. The former one is the case when all variables are dependent to each other while the latter one is the case when all variables are mutually independent. Most models belong to the intermediate of these two cases that have some conditionally dependent variable pairs connected by edges while some edges are absent expressing corresponding variable pairs are conditionally independent.

1.5.2 Graphical Gaussian Model and Precision Matrix

GGM is one specific example of a pairwise undirected graphical model where the marginal distribution of a variable $\boldsymbol{x}$ is given by a Gaussian distribution. Again, as the previous section, we assume that data points are centered and the distribution is expressed as $\mathcal{N}(\mathbf{0}_d, \Lambda^{-1})$. Here, we used a precision matrix Λ instead of a covariance matrix Σ since Λ plays the central role in GGM. From the definition of a Gaussian distribution, we can write the probability distribution as

$$\begin{aligned} p(\boldsymbol{x}) &= \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Lambda \boldsymbol{x}\right) \\ &= \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \prod_{i=1}^{d} \exp\left(-\frac{1}{2}\Lambda_{ii} x_i^2\right) \prod_{i<j} \exp\left(-\Lambda_{ij} x_i x_j\right). \end{aligned}$$

From this formulation, we find the potential functions in GGM are given by

$$\begin{aligned} \phi_i(x_i) &= \exp\left(-\frac{1}{2}\Lambda_{ii} x_i^2\right), \\ \phi_{ij}(x_i, x_j) &= \exp\left(-\Lambda_{ij} x_i x_j\right). \end{aligned}$$

Recall the discussion on a general pairwise undirected graphical model, we know that if $\phi_{ij}(x_i, x_j)$ is a constant function, then x_i and x_j are conditionally independent. From the above function, it happens only when $\Lambda_{ij} = 0$. Therefore, we can conclude that the conditional independence in GGM and the precision matrix entry have the following correspondence:

$$x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{\setminus\{i,j\}} \;\Leftrightarrow\; \Lambda_{ij} = 0,$$

where $\perp\!\!\!\perp$ denotes statistical independence. Reflecting back this result into the basic nature of a pairwise undirected graphical model, we find that the edge pattern of the GGM corresponds to the zero pattern in the precision matrix since the absence of edges implies conditional independence. See Figure 1.1 for an example.

In precision matrix, as like the case of PCA, leading eigenvalues and eigenvectors are essential to approximate the matrix. Since a precision matrix Λ relates to a covariance matrix Σ through an inverse $\Lambda = \Sigma^{-1}$, leading eigenvalues of Λ corresponds to minor eigenvalues of Σ. In this sense, discarded components in PCA

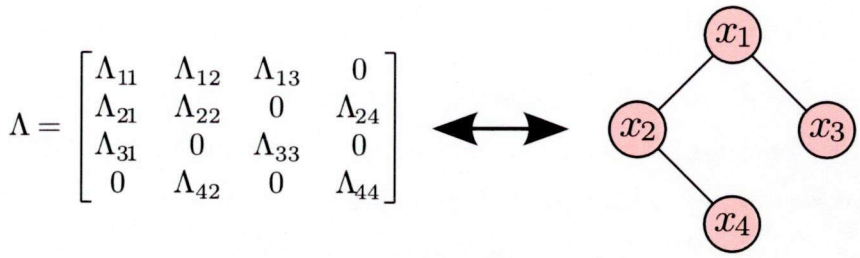


Figure 1.1: An example of GGM. Zero/Non-zero patterns in a precision matrix Λ corresponds to the presence/absence of each edge in GGM.

play the central role in a GGM context, which indicates that they have an opposite nature even though both of them are defined on the second order statistics.

1.5.3 GGM Learning via ℓ_1-Regularization

An important problem when dealing a GGM is how to derive the model from the data where the edge set E is not known a priori. To construct the model, we have to find the proper edge set E by examining the conditional independence between random variables. This problem originates with Dempster (1972) which is referred as *covariance selection*. According to the discussion above, we know the conditional independence structure of GGM is tightly connected to the entries of the precision matrix Λ. Therefore, we can cast the task as estimating Λ from the dataset. The most naive way would be to use the maximum likelihood estimator (1.4). From the statistical perspective, this is the most appropriate estimator explaining the data. However, from the law of large numbers, this estimator is a dense matrix under a finite number of samples, that is, no matrix entries are exactly equal to zero with probability one. It implies even if the (i, j)th entry of Λ is zero in truth, the maximum likelihood estimation provides a non-zero estimator $\hat{\Lambda}_{ij} \neq 0$. This property is unfavorable for covariance selection since the objective is to find conditionally independent pairs of variables, or equivalently, zero entries in Λ.

To overcome the problem, in classical studies, some entries of a precision matrix are fixed as zeros and the remaining non-zero entries are estimated, where the zero pattern is optimized in a combinatorial manner. However, this combinatorial problem is not feasible for high-dimensional data. In recent studies, the use of

an ℓ_1-regularization has been shown to be practical for covariance selection. The first such study was conducted by Meinshausen and Bühlmann (2006). In their approach, the solution is obtained by solving the Lasso (Tibshirani, 1996). Here, let an $N \times d$ matrix $X = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_N \end{bmatrix}^\top$ denote d-dimensional data with N data points. We also define X_i as the ith column and $X_{\setminus i}$ as the remaining $d-1$ columns of X. For each column, we solve the following Lasso:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \left\| X_i - X_{\setminus i}\boldsymbol{\theta} \right\|_2^2 + \rho \left\| \boldsymbol{\theta} \right\|_1, \tag{1.10}$$

where $\rho \geq 0$ is a regularization parameter and $\|\boldsymbol{\theta}\|_1 \equiv \sum_j |\theta_j|$. We then set zero patterns of $\boldsymbol{\theta}$ to be the ith column of Λ. Meinshausen and Bühlmann (2006) have also showed the asymptotic convergence of their estimator to the true graph structure under a proper condition. This approach was later reformulated as an ℓ_1-regularized maximum likelihood problem (M. Yuan & Lin, 2007; Banerjee, El Ghaoui, & d'Aspremont, 2008):

$$\begin{aligned} &\max_{\Lambda \in \mathcal{S}^+} \ell(\Lambda; \hat{\Sigma}) - \rho \left\| \Lambda \right\|_1, \\ &\ell(\Lambda; \hat{\Sigma}) \equiv \log\det\Lambda - \mathrm{tr}\left[\hat{\Sigma}\Lambda\right]. \end{aligned} \tag{1.11}$$

Here, $\ell(\Lambda; \hat{\Sigma})$ is a log-likelihood of a Gaussian distribution (up to a constant), $\mathcal{S}^+$ is a set of symmetric positive definite matrices $\mathcal{S}^+ = \{A \in \mathbb{R}^{d \times d}; A \succ 0\}$, and $\|\Lambda\|_1$ is an element-wise ℓ_1-norm $\|\Lambda\|_1 \equiv \sum_{i,j=1}^d |\Lambda_{ij}|$. We refer to this problem as Sparse Inverse Covariance Selection (SICS) following Scheinberg, Ma, and Goldfarb (2010). The resulting precision matrix of (1.11) has some zero entries owing to the effect of an additional ℓ_1-regularization term. Several efficient optimization techniques are available for solving this problem. Examples include GLasso (Friedman, Hastie, & Tibshirani, 2008), PSM (Duchi, Gould, & Koller, 2008), IPM (Li & Toh, 2010), SINCO (Scheinberg & Rish, 2010), ADMM (X. Yuan, 2009; Scheinberg et al., 2010), and QUIC (Hsieh, Sustik, Dhillon, & Ravikumar, 2011).

1.6 Summary of Contributions

Below, we briefly summarize the contributions of each chapter:

- **Chapter 2**: We consider a model called Stationary Subspace Analysis (SSA) which is a variant model of PCA. The objective of SSA is to find an invariant pattern across multiple covariance matrices based on a source mixing model. We build a new algorithm Analytic SSA for this problem, which provides a solution by solving one generalized eigenvalue problem. This simplicity is advantageous compared to an existing algorithm which requires solving a gradient decent based non-convex optimization problem since 1) it requires smaller computational cost, and 2) a global optimal solution can be derived under a certain condition while the prior algorithm guarantees only local optimality of the solution. We also provide theoretical and numerical justifications of this point.

- **Chapter 3**: In this chapter, we work on convex optimization methods called Dual Augmented Lagrangian (DAL) and Alternating Direction Method of Multipliers (ADMM). We combine the basic idea of these two techniques and formulate the DAL-ADMM algorithm for learning GGM from the data. The advantage of the proposed algorithm is its flexibility. Most existing GGM learning algorithms assume the simplest problem based on an ℓ_1-regularization. On the other hand, our algorithm can treat wider variety of regularization terms including well-known group regularizations. This flexibility is essential for solving more complicated problems arising in Chapter 4 and 5.

- **Chapter 4**: We consider finding an invariant pattern across multiple GGMs. We formalize the task as a convex optimization problem using spare regularization techniques, where the proposed formulation can be casted as a generalization of SICS (1.11) and other existing GGM learning techniques. We also show the problem can be solved by DAL-ADMM algorithm presented in Chapter 3 with each updating step requiring only simple analytic operations. The validity of the proposed method is verified through numerical simulations and also on an application to an anomaly localization problem.

- **Chapter 5**: This chapter is devoted for extending the model in Chapter 4. In this chapter, we focus on an anomaly localization problem and considers a

GGM learning algorithm specialized to this task. One basic finding is that, in an anomaly localization, row/column-wise changes between two precision matrices are important. We import this idea and formalize the task as a convex optimization problem. The proposed formulation is a variant of structured sparsity models and requires specific considerations to construct an algorithm. We find that some proper transformations of the problem allow us to treat the problem with DAL-ADMM. Hence, the proposed algorithm requires only simple analytic updating steps. We verify the advantage of our new formulation over existing techniques on an anomaly localization task through a real world data simulation.

1.7 Proofs of Theorems

1.7.1 Proof of Theorem 1

Let us recall the model for the partial correlation:

$$x_i = r_i + \boldsymbol{w}_i^\top \boldsymbol{x}_{\backslash\{i,j\}},$$
$$x_j = r_j + \boldsymbol{w}_j^\top \boldsymbol{x}_{\backslash\{i,j\}},$$

where r_i, r_j and $\boldsymbol{x}_{\backslash\{i,j\}}$ are statistically independent. Here, we define the expectations of x_i and $\boldsymbol{x}_{\backslash\{i,j\}}$ as

$$\overline{x}_i = \int x_i p(x_i)\, dr_i\,,$$
$$\overline{\boldsymbol{x}}_{\backslash\{i,j\}} = \int \boldsymbol{x}_{\backslash\{i,j\}} p(\boldsymbol{x}_{\backslash\{i,j\}})\, d\boldsymbol{x}_{\backslash\{i,j\}}\,.$$

We also denote the expectation of r_i by $\overline{r}_i$. From the independence, we then have

$$\begin{aligned}
\mathbf{0}_{d-2} &= \int \left(r_i - \overline{r}_i\right)\left(\boldsymbol{x}_{\backslash\{i,j\}} - \overline{\boldsymbol{x}}_{\backslash\{i,j\}}\right) p(r_i, \boldsymbol{x}_{\backslash\{i,j\}})\, dr_i d\boldsymbol{x}_{\backslash\{i,j\}} \\
&= \int \left\{\left(x_i - \overline{x}_i\right) - \boldsymbol{w}_i^\top \left(\boldsymbol{x}_{\backslash\{i,j\}} - \overline{\boldsymbol{x}}_{\backslash\{i,j\}}\right)\right\}\left(\boldsymbol{x}_{\backslash\{i,j\}} - \overline{\boldsymbol{x}}_{\backslash\{i,j\}}\right) p(x_i, \boldsymbol{x}_{\backslash\{i,j\}})\, dx_i d\boldsymbol{x}_{\backslash\{i,j\}} \\
&= \boldsymbol{a}_i - B_{ij}\boldsymbol{w}_i
\end{aligned}$$

where

$$\boldsymbol{a}_i = \int \left(x_i - \overline{x}_i\right)\left(\boldsymbol{x}_{\setminus\{i,j\}} - \overline{\boldsymbol{x}}_{\setminus\{i,j\}}\right) p(x_i, \boldsymbol{x}_{\setminus\{i,j\}})\, dx_i d\boldsymbol{x}_{\setminus\{i,j\}},$$
$$B_{ij} = \int \left(\boldsymbol{x}_{\setminus\{i,j\}} - \overline{\boldsymbol{x}}_{\setminus\{i,j\}}\right)\left(\boldsymbol{x}_{\setminus\{i,j\}} - \overline{\boldsymbol{x}}_{\setminus\{i,j\}}\right)^\top p(\boldsymbol{x}_{\setminus\{i,j\}})\, d\boldsymbol{x}_{\setminus\{i,j\}}\,.$$

From the equation, we derive parameters $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ as

$$\boldsymbol{w}_i = B_{ij}^{-1}\boldsymbol{a}_i,$$
$$\boldsymbol{w}_j = B_{ij}^{-1}\boldsymbol{a}_j,$$

where $\boldsymbol{a}_j$ is defined accordingly to $\boldsymbol{a}_i$. Hence, we have equations

$$r_i = x_i - \boldsymbol{a}_i^\top B_{ij}^{-1}\boldsymbol{x}_{\setminus\{i,j\}},$$
$$r_j = x_j - \boldsymbol{a}_j^\top B_{ij}^{-1}\boldsymbol{x}_{\setminus\{i,j\}}.$$

We now turn to explicitly writing down the formula of a partial correlation. From the definition of a partial correlation, we have

$$\mathrm{PCorr}(x_i, x_j|\boldsymbol{x}_{\setminus\{i,j\}}) = \frac{\mathrm{Cov}(r_i, r_j)}{\sqrt{\mathrm{Var}(r_i)\mathrm{Var}(r_j)}}.$$

The numerator can be computed as

$$\begin{aligned}\mathrm{Cov}(r_i, r_j) &= \mathrm{Cov}(x_i - \boldsymbol{a}_i^\top B_{ij}^{-1}\boldsymbol{x}_{\setminus\{i,j\}}, x_j - \boldsymbol{a}_j^\top B_{ij}^{-1}\boldsymbol{x}_{\setminus\{i,j\}}) \\ &= \sigma_{ij} - \boldsymbol{a}_i B_{ij}^{-1}\boldsymbol{a}_j,\end{aligned}$$

where

$$\sigma_{ij} = \int (x_i - \overline{x}_i)(x_j - \overline{x}_j) p(x_i, x_j)\, dx_i dx_j.$$

Each component of the denominator can also be given by

$$\begin{aligned}\mathrm{Var}(r_i) &= \mathrm{Var}(x_i - \boldsymbol{a}_i^\top B_{ij}^{-1}\boldsymbol{x}_{\setminus\{i,j\}}) \\ &= \sigma_{ii} - \boldsymbol{a}_i^\top B_{ij}^{-1}\boldsymbol{a}_i, \\ \mathrm{Var}(r_j) &= \sigma_{jj} - \boldsymbol{a}_j^\top B_{ij}^{-1}\boldsymbol{a}_j.\end{aligned}$$

Using these results, we derive the partial correlation as

$$\mathrm{PCorr}(x_i, x_j|\boldsymbol{x}_{\setminus\{i,j\}}) = \frac{\sigma_{ij} - \boldsymbol{a}_i B_{ij}^{-1}\boldsymbol{a}_j}{\sqrt{\left(\sigma_{ii} - \boldsymbol{a}_i B_{ij}^{-1}\boldsymbol{a}_i\right)\left(\sigma_{jj} - \boldsymbol{a}_j B_{ij}^{-1}\boldsymbol{a}_j\right)}}.$$

Next, we compute the right hand side of (1.1). From the definition of a covariance matrix Σ, it can be represented as

$$\Sigma = \begin{bmatrix} B_{ij} & \boldsymbol{a}_i & \boldsymbol{a}_j \\ \boldsymbol{a}_i^\top & \sigma_{ii} & \sigma_{ij} \\ \boldsymbol{a}_j^\top & \sigma_{ji} & \sigma_{jj} \end{bmatrix},$$

where we rotated rows and columns simultaneously so that the original ith and jth rows/columns to be the last two rows/columns. From this expression, we first have

$$\begin{aligned} \Lambda_{ij} = \left(\Sigma^{-1}\right)_{ij} &= \left(\sigma_{ij} - \begin{bmatrix} \boldsymbol{a}_i^\top & \sigma_{ii} \end{bmatrix} \begin{bmatrix} B_{ij} & \boldsymbol{a}_i \\ \boldsymbol{a}_j^\top & \sigma_{ji} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{a}_j \\ \sigma_{jj} \end{bmatrix} \right)^{-1} \\ &= -\frac{\sigma_{ij} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_j}{\left(\sigma_{ii} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_i\right)\left(\sigma_{jj} - \boldsymbol{a}_j^\top B_{ij}^{-1} \boldsymbol{a}_j\right) - \left(\sigma_{ij} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_j\right)^2}. \end{aligned}$$

We also have

$$\begin{aligned} \Lambda_{ii} = \left(\Sigma^{-1}\right)_{ii} &= \left(\sigma_{ii} - \begin{bmatrix} \boldsymbol{a}_i^\top & \sigma_{ij} \end{bmatrix} \begin{bmatrix} B_{ij} & \boldsymbol{a}_j \\ \boldsymbol{a}_j^\top & \sigma_{jj} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{a}_i \\ \sigma_{ji} \end{bmatrix} \right)^{-1} \\ &= \frac{\sigma_{jj} - \boldsymbol{a}_j^\top B_{ij}^{-1} \boldsymbol{a}_j}{\left(\sigma_{ii} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_i\right)\left(\sigma_{jj} - \boldsymbol{a}_j^\top B_{ij}^{-1} \boldsymbol{a}_j\right) - \left(\sigma_{ij} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_j\right)^2}, \\ \Lambda_{jj} = \left(\Sigma^{-1}\right)_{jj} &= \frac{\sigma_{ii} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_i}{\left(\sigma_{ii} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_i\right)\left(\sigma_{jj} - \boldsymbol{a}_j^\top B_{ij}^{-1} \boldsymbol{a}_j\right) - \left(\sigma_{ij} - \boldsymbol{a}_i^\top B_{ij}^{-1} \boldsymbol{a}_j\right)^2}. \end{aligned}$$

Since the above three values have a common denominator, they cancels out and

$$-\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}} = \frac{\sigma_{ij} - \boldsymbol{a}_i B_{ij}^{-1} \boldsymbol{a}_j}{\sqrt{\left(\sigma_{ii} - \boldsymbol{a}_i B_{ij}^{-1} \boldsymbol{a}_i\right)\left(\sigma_{jj} - \boldsymbol{a}_j B_{ij}^{-1} \boldsymbol{a}_j\right)}},$$

holds, which is equal to the partial correlation. □

1.7.2 Proof of Theorem 2

The distribution $p(\boldsymbol{x})$ has to satisfy the following three conditions:

$$\begin{aligned} 1 &= \int p(\boldsymbol{x})\, d\boldsymbol{x}, \\ \boldsymbol{\mu} &= \int \boldsymbol{x} p(\boldsymbol{x})\, d\boldsymbol{x}, \\ \Sigma &= \int (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top p(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned}$$

Using a method of Lagrange multipliers with u, $\boldsymbol{v}$ and W, we have the problem as

$$\max_{p} \min_{u,\boldsymbol{v},W} -\int p(\boldsymbol{x}) \log p(\boldsymbol{x})\, d\boldsymbol{x} + u\left(\int p(\boldsymbol{x})\, d\boldsymbol{x} - 1\right) + \boldsymbol{v}^\top \left(\int \boldsymbol{x} p(\boldsymbol{x})\, d\boldsymbol{x} - \boldsymbol{\mu}\right) + \operatorname{tr}\left[W^\top \left(\int (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top p(\boldsymbol{x})\, d\boldsymbol{x} - \Sigma\right)\right].$$

From the variational method, the optimal $p(\boldsymbol{x})$ is given by

$$p(\boldsymbol{x}) = \exp\{-1 + u + \boldsymbol{v}^\top \boldsymbol{x} + (\boldsymbol{x}-\boldsymbol{\mu})^\top W (\boldsymbol{x}-\boldsymbol{\mu})\}.$$

Substituting this result into the above three conditions, we derive the result. □

Chapter 2

Finding Stationary Sources with a Generalized Eigenvalue Problem

2.1 Introduction

In Section 1.4, we derived PCA from the linear mixing model (1.5). In the PCA model, all data points in the dataset are assumed to be independent and identically distributed. The main point of this chapter is to extend the idea of the linear mixing model into a multiple datasets situation where the distributions in each dataset may no longer be identical to each other. The objective is to find an invariance in the second order statistics across datasets based on the source mixing model. In particular, we focus on a time series data where the multiple datasets expression in Section 1.3.3 captures the non-stationarity nature of the observation. However, note that the i.i.d. assumption is involved as the specific case of a time dependency, and thus the discussions in this chapter are naturally applicable to the ordinal multiple datasets setting where each dataset is composed of i.i.d. observations.

The basic model we consider in this chapter is a mixture of stationary and non-stationary sources that are not directly accessible. It is a plausible model when the structure of the data generating system is less understood: there may exist several latent factors affecting the observation, as in stock market analysis (Engle & Granger, 1987) for instance. Some of these latent factors may be stationary while others are non-stationary. The existence of stationary sources is not discernible from the mixed signals since a single non-stationary source can render all variables of a multivariate time series non-stationary and thus mask the presence of time-invariant behavior. Conversely, non-stationary components with low signal power

can remain hidden among strong stationary sources. It is therefore important to discern the stationary and the non-stationary group of components in the mixed signals. However, standard Blind Source Separation (BSS) methods (Hyvärinen et al., 2001; Lee & Seung, 2001; Ziehe, Laskov, Nolte, & Müller, 2004) are not helpful in this respect since BSS algorithms such as Independent Component Analysis (ICA) (Hyvärinen et al., 2001) separate sources by independence but not by stationarity or non-stationarity. In particular, the stationary and non-stationary sources need not be independent.

To that end, the Stationary Subspace Analysis (SSA) paradigm (von Bünau et al., 2009a) has been proposed. In the SSA model, the observed time series $\boldsymbol{x}(t)$ is generated as a linear mixture of stationary sources $\boldsymbol{s}^{\mathrm{s}}(t)$ and non-stationary sources $\boldsymbol{s}^{\mathrm{n}}(t)$ with a time-constant mixing matrix A,

$$\boldsymbol{x}(t) = A \begin{bmatrix} \boldsymbol{s}^{\mathrm{s}}(t) \\ \boldsymbol{s}^{\mathrm{n}}(t) \end{bmatrix},$$

and the aim is to recover these two groups of underlying sources given only samples from $\boldsymbol{x}(t)$. The separation of stationary and non-stationary sources is useful in many circumstances. First of all, SSA can uncover stationary components in seemingly non-stationary time series. Moreover, SSA allows to study the stationary and the non-stationary part independently. For instance, in change-point detection (Basseville & Nikiforov, 1993; Siegmund & Venkatraman, 1995; Kohlmorgen et al., 1999), contributions from the stationary sources are not informative and can be removed to reduce the number of dimensions (Blythe, von Bünau, Meinecke, & Müller, 2012). Conversely, one may be interested in the estimated stationary signals that reflect constant relationships between variables (Engle & Granger, 1987; Meinecke, von Bünau, Kawanabe, & Müller, 2009). Moreover, if the channels of the time series $\boldsymbol{x}(t)$ are spatially distributed, the estimated mixing matrix $\hat{A}$ can be visualized to reveal the characteristic patterns of stationary and non-stationary contributions, as in EEG analysis (Dornhege et al., 2007; von Bünau, Meinecke, Scholler, & Müller, 2010) for instance.

In this paper, we propose a novel SSA algorithm, Analytic SSA (ASSA), where the solution is obtained by solving a generalized eigenvalue problem. The solution

to ASSA is guaranteed to be optimal under the assumption that the covariance between stationary and non-stationary sources is time-constant. Thanks to the analytic form, the algorithm requires a much lower computational cost than the state-of-the-art method KL-SSA (von Bünau et al., 2009a), does not require the selection of algorithmic parameters such as step size and convergence criterion, and is numerically stable. Moreover, ASSA finds a sequence of projections, ordered by their degree of stationarity, and therefore does not need to repeat the procedure for different numbers of stationary sources. In our simulations on synthetic data, we demonstrate that ASSA outperforms KL-SSA and ICA over a wide range of settings, even when the covariance between stationary and non-stationary sources changes over time. Moreover, we apply ASSA to geomagnetic data, namely Pi2 pulsation time series (Yumoto & the CPMN Group, 2001; Tokunaga, Kohta, Yoshikawa, Uozumi, & Yumoto, 2007), which are highly non-stationary and known to involve several sources corresponding to the geophysical mechanisms. In this case, the independence assumption of ICA is not suitable to recover the sources of interest. ASSA successfully decomposes the signals into meaningful global and local modes, which is in agreement with geophysical theory, and more plausible than the decomposition obtained by ICA (Tokunaga et al., 2007).

The remainder of this chapter is organized as follows. First of all, we introduce the SSA model and the state-of-the-art algorithm KL-SSA in Section 2.2. In Section 2.3, we derive our novel method ASSA and study its theoretical properties. The relationship to similar methods is discussed in Section 2.4. Section 2.5 contains extensive numerical simulations to show its validity and a comparison to KL-SSA and ICA. The application to geophysical data analysis is presented in Section 2.6. Our conclusion and outlook are summarized in the last Section 2.7.

2.2 Stationary Subspace Analysis

Stationary Subspace Analysis (SSA) models the observed signal $\boldsymbol{x}(t) \in \mathbb{R}^d$ as a linear superposition of stationary sources $\boldsymbol{s}^{\mathrm{s}}(t) \in \mathbb{R}^m$ and non-stationary sources

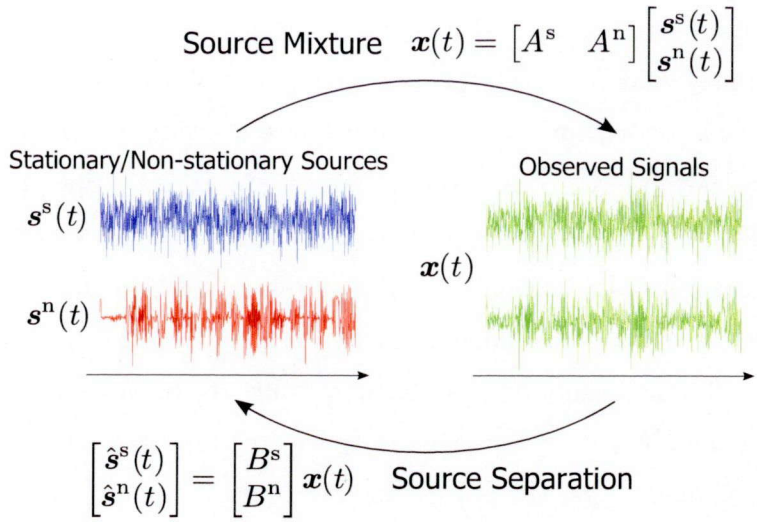


Figure 2.1: An illustrative example of SSA with one-dimensional stationary and non-stationary sources.

$\boldsymbol{s}^{\mathrm{n}}(t) \in \mathbb{R}^{d-m}$ (von Bünau et al., 2009a):

$$\boldsymbol{x}(t) = A\boldsymbol{s} = \begin{bmatrix} A^{\mathrm{s}} & A^{\mathrm{n}} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}^{\mathrm{s}}(t) \\ \boldsymbol{s}^{\mathrm{n}}(t) \end{bmatrix}, \tag{2.1}$$

where A is a time-constant invertible mixing matrix. We refer to the span of $A^{\mathrm{s}} \in \mathbb{R}^{d \times m}$ and $A^{\mathrm{n}} \in \mathbb{R}^{d \times (d-m)}$ as the stationary and the non-stationary subspace, respectively. The aim of SSA is to factorize the observed time series $\boldsymbol{x}(t)$ into stationary and non-stationary sources. That is, SSA estimates the inverse mixing matrix A^{-1} as $B = \begin{bmatrix} B^{\mathrm{s}\top} & B^{\mathrm{n}\top} \end{bmatrix}^{\top}$ such that $\hat{\boldsymbol{s}}^{\mathrm{s}}(t) = B^{\mathrm{s}}\boldsymbol{x}(t)$ and $\hat{\boldsymbol{s}}^{\mathrm{n}}(t) = B^{\mathrm{n}}\boldsymbol{x}(t)$ are the estimated stationary and non-stationary sources, respectively. We refer to the matrix $B^{\mathrm{s}} \in \mathbb{R}^{m \times d}$ and $B^{\mathrm{n}} \in \mathbb{R}^{(d-m) \times d}$ as the stationary and the non-stationary projection, respectively. See Figure 2.1 for an example.

The demixing matrix B is not unique, because the factorization into a group of stationary and a group of non-stationary sources is not unique (von Bünau et al., 2009a; von Bünau, Meinecke, Király, & Müller, 2009b). First of all, any linear transformation within the two groups of sources yields another valid demixing. Secondly, adding stationary components to the estimated non-stationary sources leaves their non-stationary nature intact, whereas the converse is not true. This means that we cannot identify the true non-stationary sources $\boldsymbol{s}^{\mathrm{n}}(t)$ from the mixing. Formally, if

we apply the demixing to the mixed sources,

$$\begin{bmatrix} \hat{\boldsymbol{s}}^{\mathrm{s}}(t) \\ \hat{\boldsymbol{s}}^{\mathrm{n}}(t) \end{bmatrix} = BA\boldsymbol{s}(t) = \begin{bmatrix} B^{\mathrm{s}}A^{\mathrm{s}} & B^{\mathrm{s}}A^{\mathrm{n}} \\ B^{\mathrm{n}}A^{\mathrm{s}} & B^{\mathrm{n}}A^{\mathrm{n}} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}^{\mathrm{s}}(t) \\ \boldsymbol{s}^{\mathrm{n}}(t) \end{bmatrix}, \tag{2.2}$$

we see that by the preceding argument, a solution to the SSA problem is fully characterized by the condition $B^{\mathrm{s}}A^{\mathrm{n}} = 0_{m\times(d-m)}$, that is, a stationary projection B^{s} must eliminate all non-stationary contributions in the estimated stationary sources. This is equivalent to the condition that the rows of the stationary projection are orthogonal on the non-stationary subspace,

$$\mathrm{span}(B^{\mathrm{s}\top}) \perp \mathrm{span}(A^{\mathrm{n}}),$$

where $\mathrm{span}(*)$ denotes the column span of a matrix. In terms of subspaces, this means that the orthogonal complement of the estimated stationary projection is equal to the true non-stationary subspace. Thus we conclude that we can identify the true stationary sources $\boldsymbol{s}^{\mathrm{s}}(t)$ (up to the linear transformation $B^{\mathrm{s}}A^{\mathrm{s}}$) and, equivalently, the true non-stationary subspace. On the other hand, the recovered sources $\hat{\boldsymbol{s}}^{\mathrm{n}}(t)$ are kept non-stationary for several different values of $B^{\mathrm{n}}A^{\mathrm{s}}$ and therefore the true non-stationary sources and the true stationary subspace are not identifiable (von Bünau et al., 2009a, 2009b).

Note that the SSA model (2.1) itself does not specify a notion of stationarity. Both the KL-SSA algorithm (von Bünau et al., 2009a) and our novel ASSA algorithm are based on the so-called weak stationarity (Hamilton, 1994). A possible extension would be to take time structure into account, for instance, the delayed covariance or the autocorrelation (Hamilton, 1994). The notion of stationarity is usually determined by the application domain and numerical considerations.

2.2.1 The KL-SSA Algorithm

The first SSA algorithm (von Bünau et al., 2009a), that we will refer to as KL-SSA[1], is based on the notion of weak stationarity (Hamilton, 1994) without time structure.

[1]KL stands for the Kullback-Leibler divergence (Kullback & Leibler, 1951), which is used to measure the stationarity of the estimated sources by comparing epoch distributions.

That is, a time series $\boldsymbol{u}(t)$ is considered stationary if its mean and covariance remain constant over time, or equivalently

$$\mathbb{E}\big[\boldsymbol{u}(t)\big] = \mathbb{E}\big[\boldsymbol{u}(t+\tau)\big]\,,$$
$$\mathbb{E}\big[\boldsymbol{u}(t)\boldsymbol{u}(t)^\top\big] = \mathbb{E}\big[\boldsymbol{u}(t+\tau)\boldsymbol{u}(t+\tau)^\top\big]\,,$$

for all $t, \tau \in \mathbb{R}$.

To apply this criterion in practice, we first divide a time series into K epochs as we discussed in Section 1.3.3. Here, we let $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ denote consecutive index sets. We then consider the time series $\boldsymbol{u}(t)$ to be stationary if the corresponding epoch means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ and covariance matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_K$ are identical, that is,

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_{k'} \text{ and } \Sigma_k = \Sigma_{k'},$$

for all pairs of epochs $k, k' = 1, 2, \ldots, K$. Now this formulation involves $\mathcal{O}(K^2)$ equality conditions between epochs, which we can reduce to $\mathcal{O}(K)$ by using the equivalent condition that each epoch's mean and covariance matrix is equal to the average,

$$\boldsymbol{\mu}_k = \overline{\boldsymbol{\mu}} \text{ and } \Sigma_k = \overline{\Sigma} \;\; (k = 1, 2, \ldots, K), \tag{2.3}$$

where $\overline{\boldsymbol{\mu}}$ and $\overline{\Sigma}$ are the average epoch mean and covariance matrix, respectively:

$$\overline{\boldsymbol{\mu}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\mu}_k, \quad \overline{\Sigma} = \frac{1}{K}\sum_{k=1}^{K} \Sigma_k.$$

Let us now turn to the algorithm which finds the stationary projection according to this definition. We have observed samples from the time series $\boldsymbol{x}(t) = A\boldsymbol{s}(t)$ which we have divided into K epochs along the time index. The choice of epochs $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ (for instance, non-overlapping consecutive blocks or sliding window) is a model parameter that is selected by the user according to the specific application. For example, the epoch length determines the time-scale on which non-stationarities can be detected or it may be desirable to align the epochs to an experimental paradigm. The number of epochs K needs to be large enough in order to avoid spurious solutions. See Section 2.2.2 for lower bound that guarantees this.

The aim of KL-SSA is to find the stationary projection B^{s} such that the estimated stationary sources $\hat{\boldsymbol{s}}^{\mathrm{s}}(t) = B^{\mathrm{s}}\boldsymbol{x}(t)$ are weakly stationary according to the condition (2.3). Since the first two moments of the estimated stationary sources $\hat{\boldsymbol{s}}^{\mathrm{s}}(t)$ can be written as the projected moments of the input $\boldsymbol{x}(t)$, this means that we aim to find B^{s} such that

$$B^{\mathrm{s}}\boldsymbol{\mu}_k = B^{\mathrm{s}}\overline{\boldsymbol{\mu}} \text{ and } B^{\mathrm{s}}\Sigma_k B^{\mathrm{s}\top} = B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top}, \tag{2.4}$$

for all epochs $k = 1, 2, \ldots, K$. In order to find this projection B^{s}, KL-SSA aims to minimize the distance between each epoch mean and covariance matrix and their respective averages. This distance is measured using the Kullback-Leibler divergence D_{KL} (Kullback & Leibler, 1951) between Gaussian distributions[2]. Since stationary sources can only be determined up to a linear transformation, we can require that $B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top} = I_m$ without loss of generality. This constraint determines the scaling, avoids degenerate solutions, and reduces the number of parameters in the optimization problem. The KL-SSA optimization problem (von Bünau et al., 2009a) is

$$\begin{aligned}
&\min_{B^{\mathrm{s}}\in\mathbb{R}^{m\times d}} \frac{1}{K}\sum_{k=1}^{K} D_{\mathrm{KL}}\left[\mathcal{N}(B^{\mathrm{s}}\boldsymbol{\mu}_k, B^{\mathrm{s}}\Sigma_k B^{\mathrm{s}\top}) || \mathcal{N}(B^{\mathrm{s}}\overline{\boldsymbol{\mu}}, B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top})\right] \\
&= \min_{B^{\mathrm{s}}\in\mathbb{R}^{m\times d}} \frac{1}{K}\sum_{k=1}^{K}\left\{\|B^{\mathrm{s}}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})\|_2^2 - \log\det\left(B^{\mathrm{s}}\Sigma_k B^{\mathrm{s}\top}\right)\right\}, \\
&\text{s.t. } B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top} = I_m.
\end{aligned} \tag{2.5}$$

This optimization problem is non-convex and a local solution is found using a gradient-based method (Avriel, 2003; Amari, 1998; Plumbley, 2005). See Müller, von Bünau, Meinecke, Király, and Müller (2011) for an implementation.

Note that the population statistics in (2.5) are replaced with sample estimators such as (1.2) and (1.3) in practice. In this context, advanced techniques such as exponentially weighted moving average (Roberts, 1959; Montgomery, 2007) would be helpful to obtain more accurate estimates, while we use naive estimators for the simulation in Section 2.5 because we are primarily interested in comparing the performance of SSA algorithms.

[2]According to Theorem 2, this is the least restrictive distributional assumption.

2.2.2 Spurious Stationarity in the KL-SSA Algorithm

The feasibility of SSA depends on the number of non-stationary sources $d - m$ and the number of epochs K. If the number of epochs with a distinct distribution of the non-stationary sources is too small, there exist directions in the non-stationary subspace on which the projected moments match – these are called *spurious stationary projections*. See Figure 2.2 for an example. The existence of spurious stationary projections renders the solution to SSA unidentifiable. The following theorem (von Bünau et al., 2009b) provides us how many distinct epochs are necessary, in order to guarantee that there are no spurious stationary projections in the generic case.

Theorem 3 (Spurious Stationarity in KL-SSA). *For the KL-SSA algorithm, given a d-dimensional signal with m stationary sources, the number of distinct epochs K required to avoid the existence of spurious stationary projections is*

$$K > \frac{d-m}{2} + 2. \tag{2.6}$$

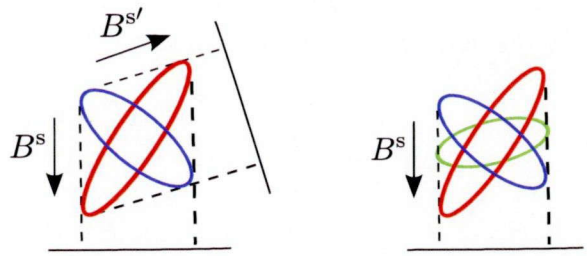
In the special case when the mean is known to be constant for all epochs, this becomes

$$K > d - m + 1. \tag{2.7}$$

Note that in practice, having more epochs of sufficient length is always desirable, as we will see in the results of the simulations in Section 2.5. Unless the number of samples in each epoch becomes too small, additional epochs provide more information about the variation in the non-stationary subspace which makes it easier to identify.

2.3 Analytic SSA

The KL-SSA optimization problem (2.5) is not convex and a local minimum is found by a gradient-based search procedure. Therefore the solution depends on the choice of initial values and algorithmic parameters. Moreover, our stability analysis in Section 2.8.1.3 reveals that the objective function is very flat in the neighborhood of the global solution. This leads to a slow convergence and adds to the computational cost, which is magnified by the need to repeat the optimization to



(a) $K = 2$ (b) $K = 3$

Figure 2.2: Illustrative example of spurious stationarity in $d = 2$. (a) Given two Gaussians with equal means (two ellipsoids), there may exist more than one projection direction on which projected distributions are equal. (b) For three Gaussians (three ellipsoids), this is no longer the case.

avoid local minima. In a fixed-point formulation in Section 2.8.1.2, KL-SSA requires $\mathcal{O}(rKm(m^2+md+d^2))$ operations to solve (2.5) where r is a number of iterations. This computational complexity limits the algorithm's practical utility on large and high-dimensional dataset. In particular, since KL-SSA requires to prespecify the number of stationary sources m, so that it needs to be run repeatedly in order to explore the results for a range of values.

In order to overcome these limitations, we propose a novel SSA algorithm called Analytic SSA (ASSA). Based on an approximate upper bound of the KL-SSA objective function (2.5), it is formulated as a generalized eigenvalue problem, which can be solved efficiently. As such, ASSA does not require any initializations nor algorithmic parameters. In particular, we can show that the solution is optimal when stationary and non-stationary sources have time-constant group-wise covariance. Even when this is not the case, our numerical simulations show that ASSA yields very good results (see Section 2.5).

2.3.1 Analytic SSA Objective Function

The ASSA objective function is based on the following approximate upper bound of the log-term in the KL-SSA objective function (2.5).

Theorem 4 (Approximate Upper Bound of KL-SSA). [3] *Let $f(B^{s})$ denote the unconstrained log-term in the KL-SSA objective function (2.5),*

$$f(B^{s}) = \frac{1}{K}\sum_{k=1}^{K} -\log \frac{\det\left(B^{s}\Sigma_{k}B^{s\top}\right)}{\det\left(B^{s}\overline{\Sigma}B^{s\top}\right)},$$

and B^{s*} is one of the true stationary projections that satisfies (2.4) and the additional constraint $B^{s*}\overline{\Sigma}B^{s*\top} = I_m$. Then the second order Taylor approximation of $f(B^{s})$ in the neighborhood of the solution B^{s*} is upper bounded by the function $g(B^{s})$ defined as

$$g(B^{s}) = \frac{2}{K}\sum_{k=1}^{K} \operatorname{tr}\left[B^{s}\left(\Sigma_{k}-\overline{\Sigma}\right)\overline{\Sigma}^{-1}\left(\Sigma_{k}-\overline{\Sigma}\right)B^{s\top}\right], \tag{2.8}$$

under the constraint that $B^{s}\overline{\Sigma}B^{s\top} = I_m$.

Using this bound, we formulate the following ASSA objective function by replacing the log-term in (2.5),

$$\begin{aligned}
&\min_{B^{s}\in\mathbb{R}^{m\times d}} \frac{1}{K}\sum_{k=1}^{K}\left\{\|B^{s}(\boldsymbol{\mu}_{k}-\overline{\boldsymbol{\mu}})\|_{2}^{2} + 2\operatorname{tr}\left[B^{s}\left(\Sigma_{k}-\overline{\Sigma}\right)\overline{\Sigma}^{-1}\left(\Sigma_{k}-\overline{\Sigma}\right)B^{s\top}\right]\right\} \\
&= \min_{B^{s}\in\mathbb{R}^{m\times d}} \operatorname{tr}\left[B^{s}SB^{s\top}\right], \\
&\text{s.t. } B^{s}\overline{\Sigma}B^{s\top} = I_m,
\end{aligned} \tag{2.9}$$

where the matrix S is given by

$$S = \frac{1}{K}\sum_{k=1}^{K}\left(\boldsymbol{\mu}_{k}\boldsymbol{\mu}_{k}^{\top} + 2\Sigma_{k}\overline{\Sigma}^{-1}\Sigma_{k}\right) - \overline{\boldsymbol{\mu}}\overline{\boldsymbol{\mu}}^{\top} - 2\overline{\Sigma}. \tag{2.10}$$

This objective function can be interpreted as the variance of the mean and covariance across all epochs. The next result ensures the optimality of our approach.

Theorem 5 (Optimality of ASSA). *Let $\hat{B}_{\mathrm{A}}^{s}$ denote the minimizer of (2.9). $\hat{B}_{\mathrm{A}}^{s}$ is then guaranteed to be optimal, that is,* $\operatorname{span}(\hat{B}_{\mathrm{A}}^{s\top}) = \operatorname{span}(B^{s*\top})$, *when the covariance between stationary and non-stationary sources is time-constant.*

[3] Note that Theorem 4 and 5 are valid only when true epoch population means and covariances are available. In practice, we replace those statistics with sample estimators which might lead to a biased result. Nonetheless, as we see in Section 2.5, ASSA shows significant improvement of the resulting errors over KL-SSA.

Moreover, it can be shown that the case of time-constant covariance between stationary and non-stationary sources can be reduced to the equivalent SSA model with group-wise uncorrelated sources. See Lemma 2 in Section 2.8.5.2. This result suggests a canonical choice for the estimated non-stationary projection, which is not identifiable in general (see Section 2.2): the non-stationary projection B^{n} is chosen such that the estimated sources are group-wise uncorrelated, or $B^{\mathrm{n}}\overline{\Sigma}B^{\mathrm{s}\top} = 0_{(d-m)\times m}$. From (2.2), we see that this is equivalent to the condition

$$\operatorname{span}(B^{\mathrm{n}\top}) \perp \operatorname{span}(A^{\mathrm{s}}). \tag{2.11}$$

Thus we conclude that if the stationary and non-stationary sources have time-constant covariance, we can identify the non-stationary sources $\boldsymbol{s}^{\mathrm{n}}(t)$ in the equivalent group-wise uncorrelated model (up to the linear transformation $B^{\mathrm{n}}A^{\mathrm{s}}$) and from (2.11), it follows that under this condition, the stationary subspace $\operatorname{span}(A^{\mathrm{s}})$ can also be identified.

In Section 2.8.4, we provide further discussions about the case when the time-constant covariance assumption is not fulfilled and how the optimality of the ASSA solution is skewed.

2.3.2 ASSA as a Generalized Eigenvalue Problem

The optimization problem (2.9) is known to be equivalent to the minimization of the generalized Rayleigh quotient $\operatorname{tr}\left[\left(B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top}\right)^{-1}\left(B^{\mathrm{s}}SB^{\mathrm{s}\top}\right)\right]$ which appears in several other BSS problems (Jolliffe, 1986; Mardia et al., 1979). The solution is found efficiently by solving the corresponding generalized eigenvalue problem.

The Lagrangian of the ASSA problem (2.9) is given by

$$\mathcal{L}(B^{\mathrm{s}}, \Gamma) = \operatorname{tr}\left[B^{\mathrm{s}}SB^{\mathrm{s}\top}\right] - \operatorname{tr}\left[\Gamma\left(B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top} - I_m\right)\right],$$

where $\Gamma \in \mathbb{R}^{m\times m}$ is the matrix of Lagrange multipliers. By setting its derivative equal to zero, we obtain the following generalized eigenvalue problem:

$$S\boldsymbol{\varphi} = \gamma\overline{\Sigma}\boldsymbol{\varphi}.$$

The solution to this problem is a set of generalized eigenvalues γ_i and generalized eigenvectors $\boldsymbol{\varphi}_i$, $\{\gamma_i, \boldsymbol{\varphi}_i\}_{i=1}^{d}$. The generalized eigenvectors are $\overline{\Sigma}$-orthonormal to

each other, that is, $\boldsymbol{\varphi}_i^\top \overline{\Sigma} \boldsymbol{\varphi}_j = 1$ if $i = j$ and 0 otherwise. This $\overline{\Sigma}$-orthogonality is equivalent to the uncorrelatedness among recovered sources $\hat{s}_i(t) = \boldsymbol{\varphi}_i^\top \boldsymbol{x}(t)$. Hence, each generalized eigenvalue corresponds to the value of the ASSA objective function $\gamma_i = \boldsymbol{\varphi}_i^\top S \boldsymbol{\varphi}_i / \boldsymbol{\varphi}_i^\top \overline{\Sigma} \boldsymbol{\varphi}_i$. Let $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_d$ be the generalized eigenvalues in ascending order. The estimated stationary projection $\hat{B}_{\mathrm{A}}^{\mathrm{s}}$ is then given by the m eigenvectors with m smallest eigenvalues,

$$\hat{B}_{\mathrm{A}}^{\mathrm{s}} = \begin{bmatrix} \boldsymbol{\varphi}_1 & \boldsymbol{\varphi}_2 & \ldots & \boldsymbol{\varphi}_m \end{bmatrix}^\top,$$

and the non-stationary projection $B_{\mathrm{A}}^{\mathrm{n}}$ consists of the remainings,

$$\hat{B}_{\mathrm{A}}^{\mathrm{n}} = \begin{bmatrix} \boldsymbol{\varphi}_d & \boldsymbol{\varphi}_{d-1} & \ldots & \boldsymbol{\varphi}_{m+1} \end{bmatrix}^\top.$$

This solution can be interpreted from a deflation point of view (Hyvärinen et al., 2001), where the ASSA objective function values (eigenvalues) γ_i are interpreted as a non-stationarity score. In the deflation approach, the stationary projections are determined incrementally. In the first step, we select the direction with minimum non-stationarity score γ_1. The $(i+1)$-th stationary projection is then found in the $\overline{\Sigma}$-orthogonal complement of the previously determined stationary projections $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_i$. Thus, in each step, the dimensionality of the input space is deflated by projecting out the newly found stationary projection $\boldsymbol{\varphi}_{i+1}$. In particular, note that the ASSA solution is uniquely determined if all eigenvalues γ_i are different whereas the KL-SSA solution is unique only up to linear transformations.

2.3.3 Spurious Stationarity in ASSA

As in KL-SSA, we need a certain number of distinct epochs in order to avoid the existence of spurious stationary projection, which renders the solution unidentifiable (see Section 2.2.2). We show that this minimum number of epochs is smaller for ASSA, which is useful in practice where data tends to be scarce.

Theorem 6 (Spurious Stationarity in ASSA). *If the covariance between stationary and non-stationary sources is time-constant, the number of epochs K required to guarantee that there exist no spurious stationary projections in d dimensions with*

m stationary sources is

$$K \geq \frac{2(d-m+1)}{\nu+1}, \tag{2.12}$$

where $\nu = \sum_{k=1}^{K} \operatorname{rank}(\Sigma_k - \overline{\Sigma})/K$. *In the special case where the mean is constant over all epochs, this bound becomes*

$$K \geq \frac{2(d-m)+1}{\nu}. \tag{2.13}$$

The requirement of ASSA (2.12) is looser than KL-SSA (2.6) when $\nu \geq 3 - 12/(d-m+4)$. Since ν is the average number of non-stationary sources with different variances among epochs, we can assume $\nu \approx d-m$ in practice and the inequality holds. Again, note that this theorem merely indicates the minimum number of distinct epochs that are necessary to guarantee determinacy. Having more epochs is always desirable to improve the accuracy of the solution.

2.3.4 Computational Complexity

The ASSA algorithm consists of three steps, 1) estimating the K epoch mean vectors and covariance matrices, 2) computing the matrix S, and 3) solving the generalized eigenvalue problem. Let N be the total number of samples $N = \sum_{k=1}^{K} |\mathcal{T}_k|$. Then the first step is in $\mathcal{O}(Nd^2)$, the second step is in $\mathcal{O}(Kd^3)$, and we require $\mathcal{O}(d^3)$ operations to solve the generalized eigenvalue problem, so that the overall complexity is $\mathcal{O}(Nd^2 + Kd^3)$.

The overall computational complexity of KL-SSA, when formulated as a fixed point algorithm, is of the order $\mathcal{O}(Nd^2 + Kd^3 + rKm(m^2 + md + d^2))$ (see Section 2.8.1.2), where r is the number of optimization steps which is expected to be large (for instance, $r > 100$) since KL-SSA converges slowly due to its flatness around the true solution. ASSA is clearly computationally advantageous, which is an important property for an algorithm that is used in the context of explorative data analysis, where results need to be obtained quickly and for different settings.

2.3.5 Choosing the Number of Stationary Sources

In practice, the number of stationary sources m may be unknown and needs to be chosen from the available data. Whereas the KL-SSA algorithm requires to specify

the number m, so that testing every value $m = 1, 2, \ldots, d-1$ would require $d-1$ independent runs of the algorithm, ASSA finds all possible stationary projections in a single step, ordered by their stationarity score. We can then treat the evaluation of each projection in a post-processing stage independently from the source separation.

Since the ASSA objective function (2.9) takes zero for truly stationary sources, one would expect to see a significant jump of the eigenvalue at some level. However, in our empirical studies, we have found that small errors in the estimation of the stationary projections accumulate in the eigenvalues, which make this jump less pronounced.

Apart from the visual inspection of eigenvalues, there exist a wide range of different procedures for testing stationarity (Dickey & Fuller, 1979; Priestley & Rao, 1969) for various types of signals and applications, which is the more suitable approach in practice.

2.4 Relation to Previous Work

2.4.1 Independent Component Analysis

Independent Component Analysis (ICA) (Hyvärinen et al., 2001) finds independent sources from a linear mixture, whereas SSA separates sources by stationarity or non-stationarity. That is, in the ICA mixing model $\boldsymbol{x}(t) = A\boldsymbol{s}(t)$, the sources $\boldsymbol{s}(t)$ are assumed to be independent whereas the general SSA model (2.1) merely presupposes that there exists a group of stationary and a group of non-stationary sources, which may have arbitrary dependence structure among and between themselves.

In order to solve the ICA problem, three major properties of sources are used (Hyvärinen et al., 2001) which are non-Gaussianity (Comon, 1994; Cardoso & Souloumiac, 1993; Hyvarinen, 1999), autocorrelation (Tong, Liu, Soon, & Huang, 1991; Molgedey & Schuster, 1994; Congedo, Gouy-Pailler, & Jutten, 2008), and non-stationarity of the variance (Matsuoka et al., 1995; Kawamoto et al., 1998; Pham & Cardoso, 2001; Hyvarinen, 2002; Parra & Sajda, 2003). The third criterion, the non-stationarity, has some resemblance to the ASSA and the KL-SSA

approach. However, these algorithms impose independence on the sources and do not consider changes of the mean. Moreover, for ASSA, we have shown that it is optimal in the case of time-constant group-wise covariance, which is a less restrictive assumption than the pair-wise independence of the ICA model. We further show the practical distinction of the non-stationarity based ICA to the SSA problem on the simulated experiment in Section 2.5.

Apart from the differences of underlying models, there are some prior works close to ASSA in the ICA context. For example, Parra and Sajda (2003) have formulated the non-stationarity based ICA as a generalized eigenvalue problem. They divide samples into two epochs and diagonalize sample covariance matrices from each epoch simultaneously by solving a generalized eigenvalue problem. The major difference of their approach to ASSA is that the generalized version, joint diagonalization of K covariance matrices from K epochs (Cardoso & Souloumiac, 1993; Belouchrani, Abed-Meraim, Cardoso, & Moulines, 1997; E. Moreau, 2001; Pham & Cardoso, 2001; Choi & Cichocki, 2000), cannot be solved by a generalized eigenvalue problem and requires solving much computationally expensive non-convex optimization problems. Here, we point out that ASSA can be interpreted as a modified version of the above algorithm to the SSA model. In the ASSA context, non-stationary independent sources in ICA are replaced with stationary and non-stationary sources with time-constant group-wise covariance. This difference changes the problem from the joint diagonalization to the joint block-diagonalization (Flury & Neuenschwander, 1994; Belouchrani, Amin, & Abed-Meraim, 1997; Theis & Inouye, 2006; Abed-Meraim & Belouchrani, 2004), which results in the ASSA algorithm when all epoch means are constant. We present the further detail in Section 2.8.3.

2.4.2 Supervised Dimensionality Reduction

The aim of supervised dimensionality reduction, or feature selection, is to find components that are informative for solving a classification or regression task. A common approach is to maximize the difference between the distributions of each class (Blankertz et al., 2008; Blankertz, Tomioka, Lemm, Kawanabe, & Muller, 2007; Fisher, 1936; Fukunaga, 1990) where the most prominent method is Linear

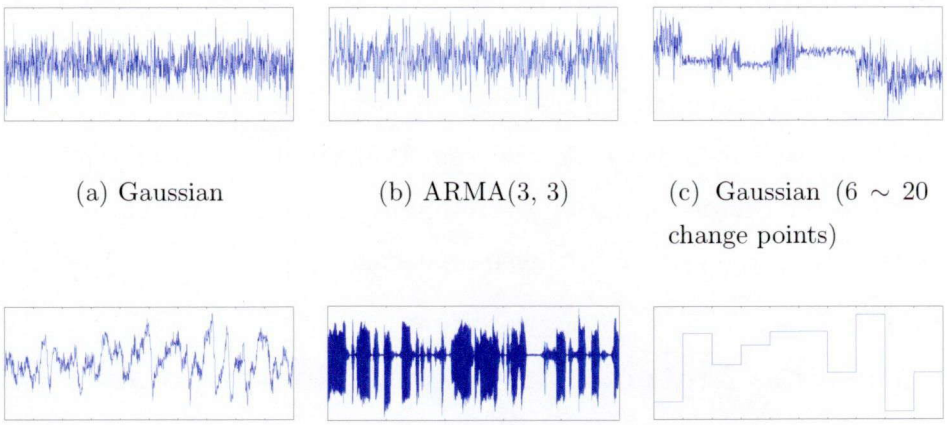
Discriminant Analysis (LDA) (Fisher, 1936; Fukunaga, 1990). LDA finds the direction on which the distance between the class means are maximal under the metric induced by the common covariance matrix. Common Spatial Patterns (CSP) (Koles, 1991; Blankertz et al., 2008, 2007; Grosse-Wentrup & Buss, 2008), a well-known method in EEG analysis (Dornhege et al., 2007), finds the projections such that the difference in variance between two classes is maximized.

If we interpret each epoch $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ of the data $\boldsymbol{x}(t)$ in ASSA as samples from different classes, finding the most non-stationary components is similar to maximizing the difference among class distributions as in LDA and CSP. ASSA can therefore be understood as a generalization of LDA and CSP because it takes both the mean and the variance into account. In particular, LDA is included as a special case of ASSA where all epoch covariances are equal.

2.5 Simulation

2.5.1 Dataset Description

In this section, we investigate the performance of the proposed ASSA algorithm and some existing methods using artificial data generated according to the SSA mixing model (2.1). In order to evaluate the behavior of the algorithms in a realistic setting, we use several types of different sources; see Figure 2.3 for an overview. For the stationary sources, we consider (a) the i.i.d. Gaussian $\mathcal{N}(\boldsymbol{\mu}^{\mathrm{s}}, \Sigma^{\mathrm{s}})$ and (b) the ARMA (Autoregressive Moving Average) (Hamilton, 1994) model of order $(3, 3)$. The parameters of these two models are chosen as follows. Each element of the mean $\boldsymbol{\mu}^{\mathrm{s}}$ and the factors $L^{\mathrm{s}} \in \mathbb{R}^{d \times d}$ of the covariance matrix $\Sigma^{\mathrm{s}} = L^{\mathrm{s}\top} L^{\mathrm{s}}$ are sampled from the standard normal distribution $\mathcal{N}(0, 1)$. The ARMA coefficients are also randomly drawn from Gaussian distributions, where if the resulting set of parameters are producing an unstable process, we discard them and regenerate from the Gaussian till the resulting process gets stable. For the non-stationary sources, we consider (c) an i.i.d. Gaussian model with 6 to 20 change points $\mathcal{N}(\boldsymbol{\mu}_k^{\mathrm{n}}, \Sigma_k^{\mathrm{n}})$ with parameters determined as before, (d) the chaotic Lorenz95 (Lorenz & Emanuel, 1998) process plus white noise, and (e) nine different kinds of real recordings of



(a) Gaussian (b) ARMA(3, 3) (c) Gaussian (6 $\sim$ 20 change points)

(d) Lorenz95 + White Noise (e) Sound, Music and Voice (9 kinds) (f) Constant (6 $\sim$ 20 change points)

Figure 2.3: Examples for candidate processes. (a) and (b) are candidates for stationary sources and (c), (d) and (e) are candidates for non-stationary sources. When (e) is chosen, one of nine recordings is assigned randomly. (b), (d) and (f) are candidates for the time-varying covariance structure (see Section 2.8.2 for further detail).

environment sounds, musics and voices[4]. The initial values of the Lorenz95 process and the nine real recordings are also selected at random.

We also investigate the effect of time-varying covariance between the stationary and non-stationary sources. This is the case when the optimality of ASSA is not guaranteed. For this purpose, we introduce the following model on non-stationary sources:

$$\boldsymbol{s}^{\mathrm{n}}(t) = \boldsymbol{s}^{\mathrm{n}'}(t) + C(t)\boldsymbol{s}^{\mathrm{s}}(t), \tag{2.14}$$

where $\boldsymbol{s}^{\mathrm{n}'}(t)$ are non-stationary sources that are uncorrelated with the stationary sources $\boldsymbol{s}^{\mathrm{s}}(t)$. A time-varying covariance structure between the stationary and the non-stationary sources is induced by the matrix $C(t) \in \mathbb{R}^{(d-m)\times m}$, which is parametrized by a correlation parameter c. It bounds the amplitude of canonical correlations (Mardia et al., 1979) between the two groups of sources and ranges

[4]available here : http://research.ics.tkk.fi/ica/demos.shtml

from zero (correlation is constant) to one (correlation varies from -1 to 1). The details of the data generation can be found in Section 2.8.2.

We set the dimensionality of the observed signal to be $d = 10$, the number of stationary sources to be $m = 5$, and the total number of available samples to be 5000, which are divided into non-overlapping consecutive epochs $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ where we vary their number K in the simulations.

2.5.2 Baseline Methods and Error Measurement

In this simulation, we introduce two baseline methods to contrast with ASSA. The first one is the KL-SSA algorithm, which is implemented as a fixed point algorithm (see Section 2.8.1.1). Since KL-SSA finds only local solutions, we choose the solution with the smallest objective function value among five restarts with random initialization[5]. The second baseline is a non-stationarity based ICA algorithm discussed in Section 2.4.1. Here, we adopt the method proposed by Pham and Cardoso (2001) since it measures the non-stationarity of sources using epoch covariances, which is similar to the approaches by ASSA and KL-SSA. It also suffers from local optima and we therefore choose the best solution among five random restarts as KL-SSA. We then construct the stationary projection $\hat{B}^{\mathrm{s}}_{\mathrm{ICA}}$ in the following manner. Let $W = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \ldots & \boldsymbol{w}_d \end{bmatrix}^\top$ be a $d \times d$ demixing matrix derived by ICA where each row vector corresponds to the source recovering projection $\hat{s}_i(t) = \boldsymbol{w}_i^\top \boldsymbol{x}(t)$. For each vector $\boldsymbol{w}_i$, we heuristically measure the non-stationarity of the recovered source $\hat{s}_i(t)$ using the score:

$$\mathrm{n-score}(\boldsymbol{w}_i) = \sum_{k=1}^{K} (\sigma_{k,i} - \overline{\sigma}_i)^2,$$

where $\sigma_{k,i}$ denotes the standard deviation of $\hat{s}_i(t)$ in the kth epoch and $\overline{\sigma}_i$ is their average across epochs $\overline{\sigma}_i = \sum_{k=1}^{K} \sigma_{k,i}/K$. This criterion achieves the minimum zero for perfectly stationary sources, that is, $\sigma_{k,i} = \sigma_{k',i}$ for all $k \neq k'$, and we choose the resulting projection $\hat{B}^{\mathrm{s}}_{\mathrm{ICA}}$ as a span of m row vectors with top m smallest scores.

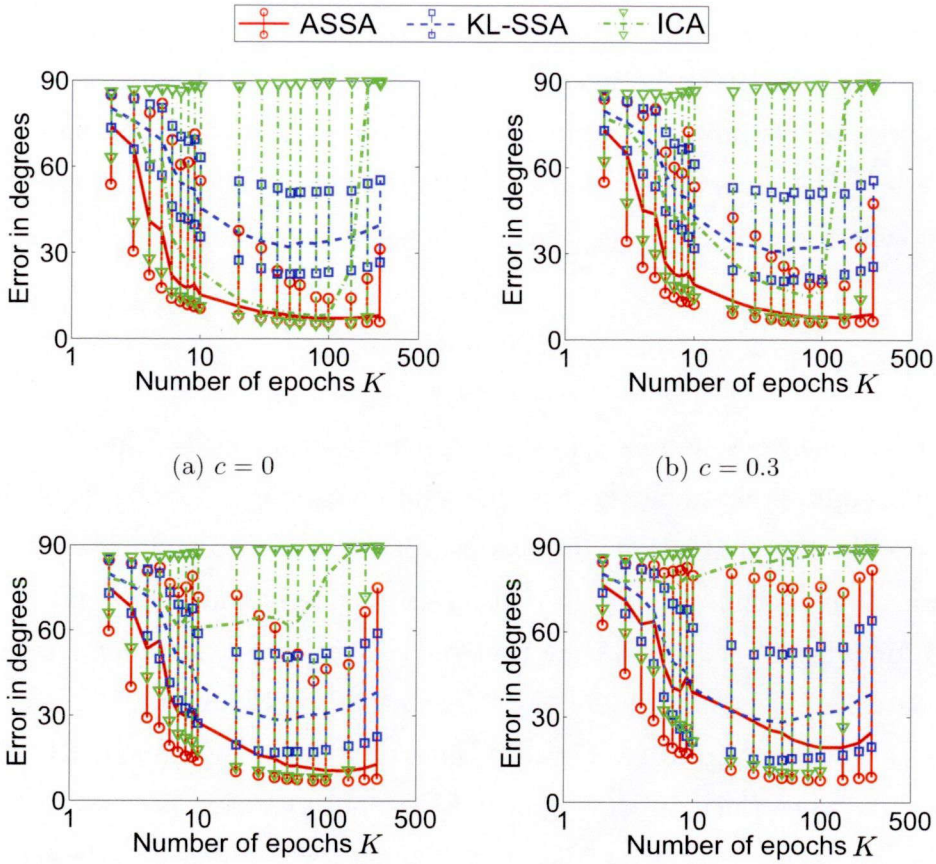
[5]The initial values are generated as $\begin{bmatrix} B^{\mathrm{s}\top} & B^{\mathrm{n}\top} \end{bmatrix}^\top = e^{0.5(M-M^\top)} \overline{\Sigma}^{-\frac{1}{2}}$ where each element of $M \in \mathbb{R}^{d \times d}$ is uniformly random in $[-10, 10]$ and e^A denotes the matrix exponential of A.

In order to evaluate the performance of algorithms, we adopt the smallest canonical angle (Chatelin, 1993) between subspaces θ (in degrees) to measure the difference between the estimated stationary projection $\hat{B}^{\mathrm{s}}$ and the true non-stationary subspace A^{n}. We report the number $90 - \theta(\hat{B}^{\mathrm{s}\top}, A^{\mathrm{n}})$, which is zero for a perfect demixing where stationary projection is orthogonal to the non-stationary subspace.

2.5.3 Result

The results are shown in Figure 2.4. When the number of epochs is small, we observe the effect of spurious stationarity: the true solution cannot be found reliably because it is masked by the presence of spurious stationary projections. In this setting, the minimum required number of epochs K given by the bounds (2.6) and (2.12) are 5 and 3 for KL-SSA and ASSA, respectively. Though we have not analyzed the spurious stationarity condition for the ICA method by Pham and Cardoso (2001), it seems that it is intermediate between the conditions of ASSA and KL-SSA. For any methods, when the number of epochs K is small, there exists spurious solutions which results in the observed median errors above 45°. Moreover, a larger number of epochs is clearly preferable to obtain more accurate solutions. However, note that when the number of samples per epoch gets too small (around $K \geq 250$ in this case), the effect of estimation errors in the epoch mean and covariance matrix leads to deteriorating performance.

Figure 2.4(a) shows the result for the case $c = 0$ (time-constant covariance), in which ASSA is guaranteed to be optimal. We can see that ASSA outperforms both baseline methods. While the median ICA result is achieving the competitive performance with ASSA around $K = 50$ to 100, we can also see its instability from the 75% error quantiles, nearly 90 degree errors meaning totally collapsed solutions. Even for time-varying covariances ($c > 0$), where ASSA is not guaranteed to be optimal, Figure 2.4(b), 2.4(c), and 2.4(d) show that ASSA is consistently outperforming on average (median performance) for all numbers of epochs K. In these cases, the independence assumption of the ICA model is also not fulfilled since the underlying stationary and non-stationary sources are correlated. The figures clearly show that ICA cannot reliably recover the two groups of sources, because



(a) $c = 0$ (b) $c = 0.3$

(c) $c = 0.6$ (d) $c = 0.9$

Figure 2.4: Median errors of ASSA, KL-SSA, and non-stationarity based ICA over 1000 random realizations of the data for different correlation parameters c. The dimensionality of the observed signals, the number of stationary sources, and the signal length are set to be 10, 5, and 5000, respectively. The observations are divided into non-overlapping consecutive epochs. The horizontal axis denotes the number of epochs K and is in a logarithmic scale. The vertical axis denotes a subspace error and the error bars extend from the 25% to the 75% quantile.

its assumption is violated. The ICA results for correlated sources seem to be quite distorted and appropriate stationary projections are found only by chance. We conjecture that the relatively poor performance of KL-SSA is due to its numerical instability (see Section 2.8.1.3).

We further conducted two extensive comparisons of ASSA and KL-SSA. In Ta-

Table 2.1: The median runtime in seconds for ASSA and KL-SSA in the simulation depicted in Figure 2.4(a). We used a Matlab implementation under 64bit Windows7 with a Intel Xeon W3565 CPU. "Pre1" denotes the computation of means and covariances from data. "Pre2" is an individual pre-processing, the computation of the matrix S in ASSA and the whitening in KL-SSA. "Main" is an optimization process, solving the generalized eigenvalue problem in ASSA and the one updating step in KL-SSA. "Step" denotes the median number of updating steps in KL-SSA with five random initializations. "Total" is the overall runtime.

	K	Pre1	Pre2	Main	Step	Total
ASSA	10	.0014	.0002	.0002	-	.0020
KL-SSA	10	.0014	.0002	.0005	340	.1930
ASSA	50	.0049	.0004	.0002	-	.0059
KL-SSA	50	.0049	.0004	.0023	372	.9086
ASSA	100	.0092	.0007	.0002	-	.0107
KL-SSA	100	.0091	.0007	.0046	560	2.6569
ASSA	200	.0178	.0013	.0002	-	.0202
KL-SSA	200	.0178	.0012	.0091	556	5.2513

ble 2.1, we have summarized the computational advantage of ASSA over KL-SSA. Here, we find that ASSA has achieved more than 100 times faster speed than KL-SSA by avoiding an iterative optimization, which is a practical bottleneck of KL-SSA due to its flatness of the objective function (see Section 2.8.1.3). The results of an exhaustive comparison over different degrees of correlation parameter are also shown in Figure 2.5. Here, the median error of ASSA is significantly lower than that of KL-SSA even though the error of KL-SSA slightly improves as a correlation parameter gets larger. However, despite its good median performance of ASSA, we also observe the gradual growth of its 75% error quantile. We conjecture that this is due to the violated assumption. It seems that even though ASSA is performing well on average, its result is distorted for certain cases, and the probability of facing such cases increases as a degree of assumption violation grows[6]. Even so, the result

[6]See Section 2.8.4 for further discussion about this point.

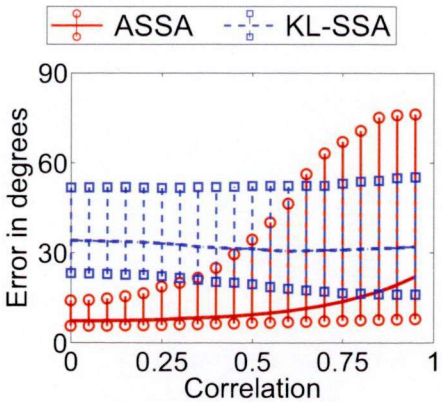


Figure 2.5: Comparison of ASSA and KL-SSA for varying correlation parameter c. In this simulation, the number of epochs K is set to be 100. The vertical axis shows the error measured as the subspace angle to the true solution. The horizontal axis shows the correlation parameter. The median error of ASSA and KL-SSA over 1000 random realizations of the data is plotted along with error bars that extend from the 25% to the 75% quantile.

shows that ASSA is not sensitive to the assumption violation as ICA and is outperforming KL-SSA in terms of the 75% error quantile for correlation parameters smaller than 0.6.

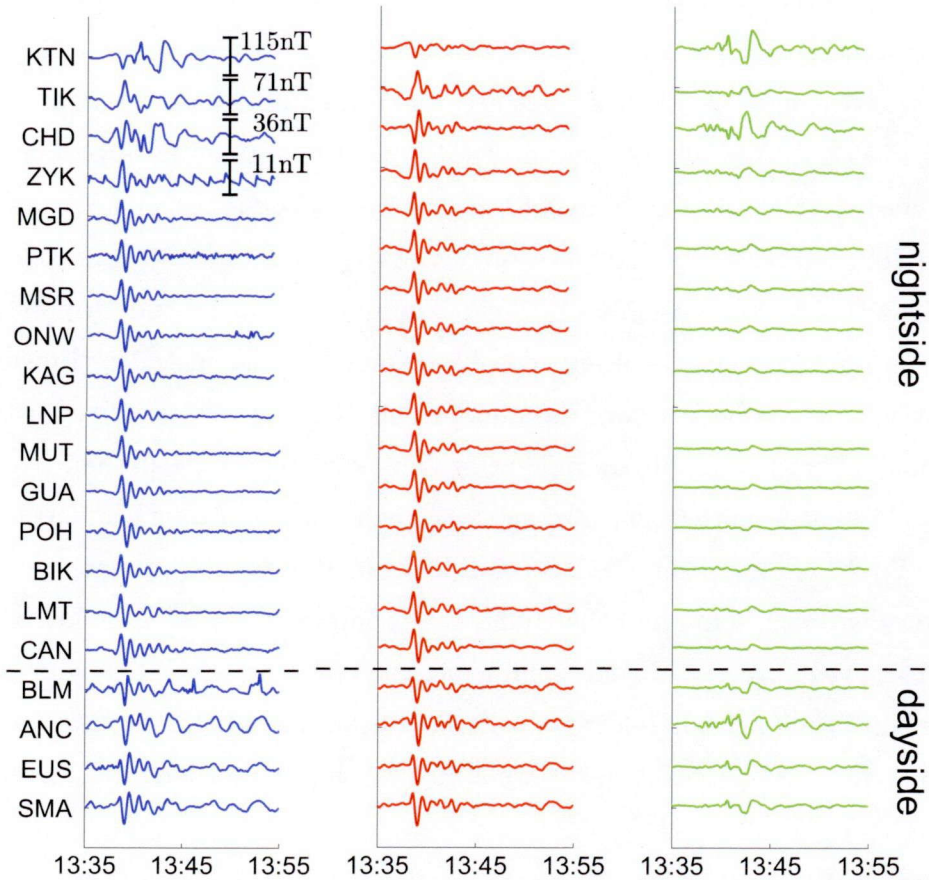
2.6 Application to the Geomagnetic Data Analysis

We now apply ASSA to the investigation of the dynamics of the earth's magnetic field using ground magnetometer data. The geomagnetic phenomenon called Pi2 pulsation (Jacobs, Kato, Matsushita, & Troitskaya, 1964; Saito, 1969) has been studied to reveal the connection to the substorm (Saito, Yumoto, & Koyama, 1976) or the propagation mechanism of magnetohydrodynamic waves in the magnetosphere (Uozumi et al., 2004). However, the observations of Pi2 pulsations on the ground involve several components reflecting 1) propagations of fast and shear Alfvén wave, 2) resonances of plasmaspheric or magnetospheric cavity and magnetic field lines, and 3) transformations to ionospheric current systems (Yumoto

& the CPMN Group, 2001; Sutcliffe & Yumoto, 1991; Yeoman & Orr, 1989; Olson & Rostoker, 1977; Kuwashima & Saito, 1981). It is unclear how they couple with each other and how their signals are distributed at different latitudes. Thus, in order to extract the global system of Pi2 pulsations from the superpositions of several effects, the use of ICA had been proposed (Tokunaga et al., 2007). The result of ICA suggests the existence of two major components in an isolated Pi2 event. One is the global oscillation that is common for all latitudes and the other is the local pulsation that is observed only in some specific latitudes. However, the source-wise independency assumption underlying ICA is too restrictive for this specific problem. From a geophysical point of view, one expects that there are several factors behind each source that interact with each other and thus lead to dependent sources. We have also observed in Section 2.5 that ICA results for such sources can be highly distorted. On the other hand, the components that are closely related to the Pi2 event are those that exhibit strong non-stationary behavior over the selected time window. Therefore, in order to obtain meaningful results, extracting the non-stationary sources seems more plausible than factorizing into independent sources.

The ground magnetometer data was obtained from CPMN stations at the 210° magnetic meridian chain and South America chain (Yumoto & the CPMN Group, 2001). Figure 2.6(a) shows the horizontal direction component of each station[7], which is bandpass-filtered ($25 - 250$s) amplitude-time recording of Pi2 pulsation observed during the time window 13:35-13:55 UT on February 17, 1995 at 400 points in time. Note that the top four signals have larger powers: KTN (115nT), TIK (71nT), CHD (36nT), and ZYK (11nT), the other signals have power around 3nT. The periodic wave in channel ZYK is environmental noise that is not related to the Pi2 event. As shown in Figure 2.6(a), most signals, especially those in low latitude, have similar highly non-stationary waveforms. We therefore expect that a common non-stationary source can be recovered from the signals. Moreover, the most stationary sources would correspond to observation noise and the sources with medium non-stationarity score are probably related to local phenomena. In order to

[7]The names of the stations are abbreviated by three letter codes. See Yumoto and the CPMN Group (2001).



(a) Original Signals (b) Pi2 Component A (c) Pi2 Component B

Figure 2.6: (a) Original signals: horizontal direction component of Pi2 pulsations observed on February 17, 1995 at CPMN stations. The bandpass filter range is $25 - 250$s. The plots are aligned in the descending order of station's latitude from the top. Stations above and below dashed line are the 210° magnetic meridian chain and the South America chain, respectively. The scaling of the vertical axis is around 3nT except for top 4 stations. (b) Separated Pi2 component A as a linear combination of N1, N2, and N3 (see Figure 2.7). (c) Separated Pi2 component B as a linear combination of N4 and N5.

suppress the effect of noise, we first extract the seven Principal Components (Jolliffe, 1986) from the data to which we then apply ASSA.

Figure 2.7 shows the waveforms of the non-stationary sources (Ns) estimated

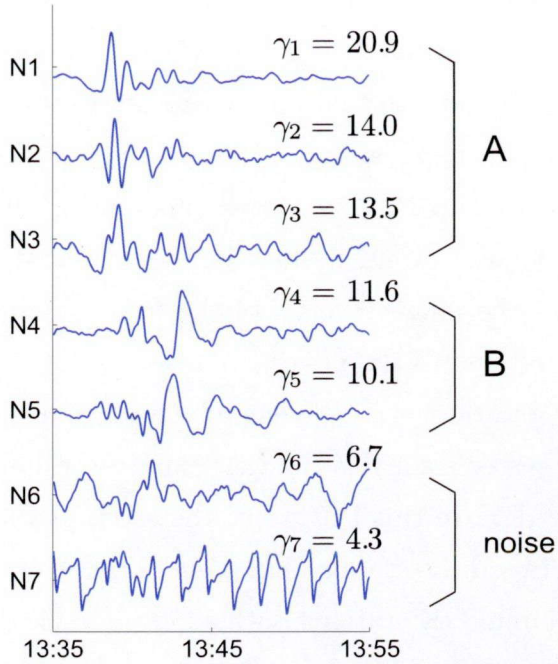


Figure 2.7: Estimated non-stationary sources (Ns) by means of ASSA. The observed signals are divided into $K = 20$ non-overlapping consecutive epochs. The estimated sources are classified into three groups based on their ASSA scores γ. N1, N2, and N3 are classified into Group A. N4 and N5 are classified into Group B. N6 and N7 are noise sources.

by ASSA (using a consecutive partitioning into $K = 20$ epochs) in descending order by their non-stationarity score γ_i. We categorize the seven sources according to their relative non-stationarity into group A (highly non-stationary), group B (medium non-stationarity), and a noise group (virtually stationary). We conjecture that the sources in group A and B are related to the Pi2 event. Figure 2.6(b) and 2.6(c) show the Pi2 components A and B plotted as a linear combination of the sources in group A and group B, respectively. We can see that the Pi2 component A, the global mode, is distributed globally to all latitudes whereas the Pi2 component B, the local mode, occurs only in some specific stations (KTN, CHD), mainly at nightside high latitudes. In past studies, the plasmashperic cavity mode is deemed to be one of the dominant mechanism of Pi2 pulsations at low and middle latitudes (Sutcliffe & Yumoto, 1991; Yeoman & Orr, 1989; Takahashi, Lee,

Nosé, Anderson, & Hughes, 2003). This finding coincides with our Pi2 component A. The KTN station, whose signal does not show contributions from component A, is located in very high latitude, so that one would not expect that it is affected by the plasmapause. The plasmapause is also known to cause a polarization reversal of the substorm associated to Pi2 pulsations (Fukunishi, 1975; Takahashi et al., 2003). In this particular Pi2 event, its location is estimated between the stations CHD and ZYK. Hence the phase reversal of the Pi2 component A between CHD and ZYK is probably related to the polarization reversal.

However, the interpretation of the Pi2 component B is unclear. Potential causes are substorm current systems such as the westward auroral electrojets and oscillations of the current wedge. In this Pi2 event, the estimated location of the aurora break up spot is in between the stations KTN and TIK. This would imply that the signals from the KTN and TIK stations both show large local modes, which is not the case for the component B. The effect of current systems is highly complex and only partly understood. Further analysis will require satellite observations and the investigation of other aspects of the magnetometer data, which is beyond the scope of this study.

The components extracted by ASSA suggest that there are two major sources behind the Pi2 pulsations. Our Pi2 component A corresponds directly to geophysical theory and the findings of other empirical studies. The component B suggests that there are other mechanisms whose understanding requires further investigation of current systems and the auroral breakup. In comparison to the ICA result (Tokunaga et al., 2007), the global mode found by ASSA is more plausible because it has smaller power at the KTN station, which locates north of the auroral breakup and less effects from the plasmapause is expected.

2.7 Conclusion and Future Work

In this chapter, we have proposed the first SSA algorithm, ASSA, whose solution can be obtained in closed form, and we have shown that it is optimal in the case of time-constant group-wise covariance. Thanks to its formulation as a generalized eigenvalue problem, it is more than 100 times faster than the state-of-the-art KL-

SSA and it does not require tuning any algorithmic parameters. We also proved that ASSA has a looser condition for avoiding spurious solutions. Moreover, unlike KL-SSA, we do not need to run ASSA multiple times to derive solutions for different numbers of sources: we can derive a set of solutions in one step. We have demonstrated the performance of ASSA in a realistic set of experiments and applied it to geomagnetic measurements Pi2 pulsations, where it successfully factorizes the observed time series into meaningful components.

A number of open questions remain. First of all, to date there exists no systematic approach for selecting the number of stationary sources m from data. Even though ASSA's eigenvalue spectrum and subsequent hypothesis testing can offer some guidance, a principled model selection technique, such as Information Criterion (Akaike, 1974; Schwarz, 1978), still needs to be developed. Similarly, apart from the lower bound on the number of epochs K, their choice $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ is so far determined heuristically, based, for instance, on the number of samples in each epoch. Most importantly, both KL-SSA and ASSA hinge on the limited notion of weak stationarity, which is a good pragmatic choice for many scenarios. However, an extension towards separating sources by non-stationarities with respect to the time structure would open up a wide field of new applications, where temporal changes in the frequency domain are the main point of interest.

2.8 Appendix

2.8.1 Computational Issues of KL-SSA

2.8.1.1 KL-SSA with Fixed Point Algorithm

To solve the optimization problem (2.5), the combination of natural gradient (Amari, 1998; Plumbley, 2005) and conjugate gradient (Avriel, 2003) had been proposed by von Bünau et al. (2009a). Here, we introduce the use of the fixed point algorithm (Hyvärinen et al., 2001). The fixed point algorithm is simpler since it does not require the tuning of step size as the gradient descent method. It therefore allows us to compare ASSA and KL-SSA more objectively.

In the pre-processing stage, each epoch mean and covariance are centered and

whitened as

$$\boldsymbol{\mu}_k^{\mathrm{w}} = \overline{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}),$$
$$\Sigma_k^{\mathrm{w}} = \overline{\Sigma}^{-\frac{1}{2}}\Sigma_k\overline{\Sigma}^{-\frac{1}{2}},$$

for $k = 1, 2, \ldots, K$. We also factorize the stationary projection B^{s} as $B^{\mathrm{s}} = W\overline{\Sigma}^{-\frac{1}{2}}$, $W \in \mathbb{R}^{m\times d}$ and derive the alternative problem of (2.5):

$$\min_{W\in\mathbb{R}^{m\times d}} \frac{1}{K}\sum_{k=1}^{K}\left\{\|W\boldsymbol{\mu}_k^{\mathrm{w}}\|_2^2 - \log\det\left(W\Sigma_k^{\mathrm{w}}W^\top\right)\right\}, \text{ s.t. } WW^\top = I_m.$$

The fixed point algorithm is based on the fact that solutions to the optimization problem $\min_W f(W)$ with a constraint $WW^\top = I_m$ has the following property (Hyvärinen et al., 2001):

$$\mathrm{span}(W^\top) = \mathrm{span}(dW^\top),$$

where dW denotes the gradient $dW = \partial f(W)/\partial W$. It indicates that the optimal W is proportional to the gradient dW. Therefore, in the fixed point algorithm, we update W by substituting dW and rescaling it so that $WW^\top = I_m$ is kept. The overall procedure is summarized in Algorithm 1. Note that in KL-SSA, the gradient dW is given by

$$dW = \frac{2}{K}\sum_{k=1}^{K}\left\{W\boldsymbol{\mu}_k^{\mathrm{w}}\boldsymbol{\mu}_k^{\mathrm{w}\top} - \left(W\Sigma_k^{\mathrm{w}}W^\top\right)^{-1}W\Sigma_k^{\mathrm{w}}\right\}.$$

In the synthetic experiment in Section 2.5, we stopped the updating iteration when $1 - \mathrm{tr}\left[W_{\mathrm{new}}W_{\mathrm{old}}^\top\right]/m < 10^{-5}$.

2.8.1.2 Computational Complexity

In this section, we derive the computational complexity of KL-SSA in the fixed point formulation (Algorithm 1). In the pre-processing stage, we compute the sample means and covariance matrices which is $\mathcal{O}(Nd^2)$ where N is the total size of epochs $N = \sum_{k=1}^{K}|\mathcal{T}_k|$. The computation of the whitening matrix $\overline{\Sigma}^{-\frac{1}{2}}$ is in $\mathcal{O}(d^3)$ and the whitening of all epochs is of the order $\mathcal{O}(Kd^3)$. In the optimization stage, there appears an inverse of $W\Sigma_k^{\mathrm{w}}W^\top$ which requires $\mathcal{O}(md(m+d))$ for the

Algorithm 1 : KL-SSA with a fixed point algorithm

Input: samples $\{\boldsymbol{x}(t)\}_{t=1}^{T}$, index sets $\{\mathcal{T}_k\}_{k=1}^{K}$, number of stationary sources m

Output: stationary projection $\hat{B}_{\mathrm{KL}}^{\mathrm{s}}$

1: divide samples into epochs by $\{\mathcal{T}_k\}_{k=1}^{K}$;
2: center and whiten the means and the covariances $\{\boldsymbol{\mu}_k^{\mathrm{w}}, \Sigma_k^{\mathrm{w}}\}_{k=1}^{K}$;
3: initialize $W \in \mathbb{R}^{m \times d}$ so that $WW^\top = I_m$;
4: **repeat**
5: compute the gradient dW;
6: update $W \leftarrow dW$;
7: normalize W so that $WW^\top = I_m$;
8: **until** W converges
9: set $\hat{B}_{\mathrm{KL}}^{\mathrm{s}} \leftarrow W\overline{\Sigma}^{-\frac{1}{2}}$;

matrix multiplication and $\mathcal{O}(m^3)$ for the inverse. The cost for all epochs is thus $\mathcal{O}(Km(m^2 + md + d^2))$ and the overall complexity is $\mathcal{O}(Nd^2 + Kd^3 + rKm(m^2 + md + d^2))$ where r is the number of updating steps till convergence.

2.8.1.3 Stability

When solving an optimization problem $\min_{\boldsymbol{u}} f(\boldsymbol{u})$ by an iterative method, its numerical stability is governed by the condition number of the Hessian matrix $\nabla^2 f(\boldsymbol{u})$, where the condition number $\kappa(C)$ for a matrix C is defined as a ratio of its largest singular value to the smallest singular value. If $\kappa(\nabla^2 f(\boldsymbol{u}))$ is large, the contour of f forms a long ellipsoid with large eccentricity. In that case, the optimization procedure tends to require a large number of steps and the solution gets numerically instable (Boyd & Vandenberghe, 2004).

Here, we see the Hessian matrix of KL-SSA (2.5) for the case of $m = 1$, that is, $B^{\mathrm{s}} = \boldsymbol{b}^\top \in \mathbb{R}^{1 \times d}$, for simplicity. The unconstrained KL-SSA objective function $f(\boldsymbol{b})$ is

$$f(\boldsymbol{b}) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\left\|\boldsymbol{b}^\top(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})\right\|_2^2}{\boldsymbol{b}^\top\overline{\Sigma}\boldsymbol{b}} - \log\frac{\boldsymbol{b}^\top\Sigma_k\boldsymbol{b}}{\boldsymbol{b}^\top\overline{\Sigma}\boldsymbol{b}}\right).$$

Let $\boldsymbol{b}^* \in \mathbb{R}^d$ be a true SSA solution satisfying (2.4). We then derive

$$\frac{1}{4}\nabla^2 f(\boldsymbol{b}^*) = \frac{1}{K}\sum_{k=1}^{K} \frac{\Sigma_k \boldsymbol{b}^* \boldsymbol{b}^{*\top} \Sigma_k - \overline{\Sigma}\boldsymbol{b}^* \boldsymbol{b}^{*\top}\overline{\Sigma}}{\boldsymbol{b}^{*\top}\overline{\Sigma}\boldsymbol{b}^*}.$$

Moreover, we can see from (2.18) that this Hessian matrix gets zero when the covariance between stationary and non-stationary sources is time-constant. It implies that f is very flat in the neighborhood of the true solution $\boldsymbol{b}^*$, and the gradient based method may stop far before. When the covariance between the two groups of sources is time-varying, we can express $\Sigma_k \boldsymbol{b}^*$ as

$$\Sigma_k \boldsymbol{b}^* = \begin{bmatrix} A^{\mathrm{s}} & A^{\mathrm{n}} \end{bmatrix} \begin{bmatrix} \Sigma^{\mathrm{s}} \\ \Sigma_k^{\mathrm{ns}} \end{bmatrix} A^{\mathrm{s}\top} \boldsymbol{b}^*,$$

where $\Sigma^{\mathrm{s}} \in \mathbb{R}^{m \times m}$ is a covariance matrix of stationary sources, which is constant across epochs, and $\Sigma_k^{\mathrm{ns}} \in \mathbb{R}^{(d-m)\times m}$ is a covariance matrix between non-stationary and stationary sources in the kth epoch. We then derive

$$\frac{1}{K}\sum_{k=1}^{K} \Sigma^{\mathrm{s}} A^{\mathrm{s}\top} \boldsymbol{b}^* \boldsymbol{b}^{*\top} A^{\mathrm{s}} \left(\Sigma_k^{\mathrm{ns}} - \overline{\Sigma}^{\mathrm{ns}}\right)^\top = 0_{m\times(d-m)},$$

where the equality holds from the definition of $\overline{\Sigma}$. The Hessian matrix is

$$\frac{1}{4}\nabla^2 f(\boldsymbol{b}^*) = A \begin{bmatrix} 0_{m\times m} & 0_{m\times(d-m)} \\ 0_{(d-m)\times m} & Z^{\mathrm{n}} \end{bmatrix} A^\top,$$

where Z^{n} is defined as

$$Z^{\mathrm{n}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\left(\Sigma_k^{\mathrm{ns}} - \overline{\Sigma}^{\mathrm{ns}}\right) A^{\mathrm{s}\top} \boldsymbol{b}^* \boldsymbol{b}^{*\top} A^{\mathrm{s}} \left(\Sigma_k^{\mathrm{ns}} - \overline{\Sigma}^{\mathrm{ns}}\right)^\top}{\boldsymbol{b}^{*\top}\overline{\Sigma}\boldsymbol{b}^*}.$$

It is obvious that the Hessian matrix is rank deficient and thus the condition number is infinite, which again implies that KL-SSA is instable around $\boldsymbol{b}^*$. Note that this result is irrelevant to the parametrization of $\boldsymbol{b}$. Even if we parametrize $\boldsymbol{b}$ as $\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{\theta})$ with some other parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, the Hessian matrix of f over $\boldsymbol{\theta}$, $\nabla_\theta^2 f(\boldsymbol{b}(\boldsymbol{\theta}))$, is expressed as

$$\nabla_\theta^2 f(\boldsymbol{b}(\boldsymbol{\theta})) = J_\theta \nabla^2 f(\boldsymbol{b}) J_\theta^\top,$$

with a Jacobian matrix $J_\theta \in \mathbb{R}^{d\times d}$, and again it is rank deficient.

2.8.2 Data Generation

From the model (2.14), we can see that the correlation between $\boldsymbol{s}^{\mathrm{s}}(t)$ and $\boldsymbol{s}^{\mathrm{n}}(t)$ gets $C(t)$ after a proper scaling. In this simulation, we set $C(t) = R_1 \mathrm{diag}(\boldsymbol{c}(t)) R_2^\top$ where R_1 and R_2 are $m \times m$ and $(d-m) \times (d-m)$ orthogonal matrices and $\mathrm{diag}(\boldsymbol{c}(t)) \in \mathbb{R}^{m \times (d-m)}$ is a matrix with $\boldsymbol{c}(t) \in \mathbb{R}^{\min(m, d-m)}$ on its diagonal. Each component of $\boldsymbol{c}(t)$ is limited to $[-1, 1]$ from the definition of correlation. The process $\boldsymbol{c}(t)$ is also chosen from several different sources in Figure 2.3 which are (b) ARMA(3, 3), (d) Lorenz95, and (f) constant with 6 to 20 change points. The chosen process is scaled so that each component of $\boldsymbol{c}(t)$ belongs to $[-c, c]$ for a given correlation parameter $c \in [0, 1]$. When $c = 0$, the covariance between stationary and non-stationary sources is zero and thus time constant, in which the optimality of ASSA is guaranteed while the ASSA assumption is violated for $c > 0$. The overall data generating procedure is as follows: 1) randomly generate A, R_1 and R_2, 2) randomly assign m processes to stationary sources from two candidates and $d-m$ processes to non-stationary sources from three candidates, 3) generate $\boldsymbol{s}^{\mathrm{s}}(t)$ and $\boldsymbol{s}^{\mathrm{n}'}(t)$ from each assigned processes, 4) randomly assign one from three candidates to $\boldsymbol{c}(t)$ and generate non-stationary sources $\boldsymbol{s}^{\mathrm{n}}(t)$ according to the model (2.14), and 5) generate the observed signal $\boldsymbol{x}(t)$ from the SSA mixture (2.1).

2.8.3 ASSA and Joint Block-Diagonalization

Here, we briefly introduce how the joint block-diagonalization (Flury & Neuenschwander, 1994; Belouchrani, Amin, & Abed-Meraim, 1997; Theis & Inouye, 2006; Abed-Meraim & Belouchrani, 2004) approach can be applied to the SSA problem. As have shown in (2.17), the essential covariance structure of stationary and non-stationary sources is in the block-diagonal form when two sources are group-wise uncorrelated. Therefore, the source recovery can be interpreted as the problem of finding a matrix $B \in \mathbb{R}^{d \times d}$ that makes $B \Sigma_k B^\top$ to be block-diagonal for all K matrices in one time, which is achieved by solving

$$\min_{B \in \mathbb{R}^{d \times d}} \sum_{k=1}^{K} \left\| \text{block-off-diag}(B \Sigma_k B^\top) \right\|_{\mathrm{F}}^2, \quad \text{s.t. } B \overline{\Sigma} B^\top = I_d, \tag{2.15}$$

where block-off-diag$(*)$ denotes block-off-diagonal elements of a matrix. The equivalence of ASSA to (2.15) is summarized in the next theorem.

Theorem 7 (ASSA and Joint Block-Diagonalization). *The problem (2.15) coincides with the ASSA problem (2.9) with a condition* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_K$.

2.8.4 Assumption Violation and ASSA Solution

In Section 2.3, we have constructed the ASSA algorithm based on the assumption that stationary and non-stationary sources have a time-constant covariance. Moreover, even when this assumption is not fulfilled, we have observed that ASSA is quite robust against the violation through simulations in Section 2.5. Here, we provide one theoretical result that gives an insight how the assumption violation affects the ASSA solution.

In the analysis, we consider the simplest case when stationary and non-stationary sources are group-wise uncorrelated. The reason is that we can always construct such a model in a time-constant situation without loss of generality from Lemma 2 (Section 2.8.5.2). Under such a model, we study how the following small perturbation on a covariance between the two groups of sources affects the resulting ASSA solution,

$$\Sigma_k = A \begin{bmatrix} \Sigma^{\mathrm{s}} & \epsilon \Sigma_k^{\mathrm{sn}} \\ \epsilon \Sigma_k^{\mathrm{sn}\top} & \Sigma_k^{\mathrm{n}} \end{bmatrix} A^\top,$$

where $\Sigma^{\mathrm{s}} \in \mathbb{R}^{m \times m}$ is a covariance matrix of stationary sources, which is common across epochs, $\Sigma_k^{\mathrm{n}} \in \mathbb{R}^{(d-m)\times(d-m)}$ is a covariance of non-stationary sources in the kth epoch, and $\Sigma_k^{\mathrm{sn}} \in \mathbb{R}^{m\times(d-m)}$ is a covariance between the two groups. Here, we let $\mathcal{O}(\Sigma_k^{\mathrm{sn}}) = \mathcal{O}(1)$ and explicitly impose a parameter $0 \leq \epsilon \ll 1$ to express that the assumption violation is sufficiently small. Our main result is based on the fact that under this small perturbation, we can also express the matrix S defined in (2.10) as

$$S = A \begin{bmatrix} 0_{m\times m} & \epsilon Q \\ \epsilon Q^\top & P \end{bmatrix} A^\top + \mathcal{O}(\epsilon^2), \tag{2.16}$$

with some $Q \in \mathbb{R}^{m\times d}$ and $P \in \mathbb{R}^{(d-m)\times(d-m)}$. We then have the following theorem.

Theorem 8 (Effect of the Assumption Violation). *For a sufficiently small perturbation* $0 < \epsilon \ll 1$, *the ASSA solution* $\hat{B}_{\mathrm{A}}^{\mathrm{s}}$ *is not orthogonal to the* n*-space* $\mathrm{span}(A^{\mathrm{n}})$ *and its error is in the following order:*

$$\mathcal{O}(\hat{B}_{\mathrm{A}}^{\mathrm{s}} A^{\mathrm{n}}) = \mathcal{O}(\epsilon Q P^{-1}).$$

From the definition of the matrix S, we can interpret matrices Q and P as the non-stationarity degree of the covariance between the two groups and the covariance within non-stationary sources, respectively. The above result shows that even for small ϵ, the assumption violation might cause larger error if QP^{-1} is large. It occurs when Q has larger values in the directions where P has smaller values. The smaller values of P imply that the sources in these directions are non-stationary only slightly, while the larger Q stands that two sources are strongly correlated. We can therefore interpret the above result as that sources with only slight non-stationarity tend to be mixed up with truly stationary sources under the assumption violation. On the other hand, the effects of small correlations with highly non-stationary sources can be negligible in practice. It is in line with our intuition that the significant non-stationarity could be easier to distinguish even when there are some correlations between the two groups.

The above theorem can partly explain the simulation result in Section 2.5. In Figure 2.5, the median error gradually grows along a correlation parameter c while the 75% quantile is rapidly increasing. Note that the parameter c corresponds to the perturbation ϵ in the theorem. The theorem indicates that the assumption violation is not always fatal. There are some cases that has small errors even under large c if QP^{-1} is small. We conjecture this is the reason why the ASSA solution is not entirely collapsed even for large c, but only for some specific cases as observed in the 75% error quantile.

2.8.5 Proofs of Theorems

2.8.5.1 Proof of Theorem 4

Since $f(B^{\mathrm{s}})$ takes its minimum zero at $B^{\mathrm{s}} = B^{\mathrm{s}*}$, its first order derivative vanishes. Therefore the second order Taylor approximation $\tilde{f}(B^{\mathrm{s}}; B^{\mathrm{s}*})$ depends only on the

second order derivative of $f(B^{s})$:

$$\tilde{f}(B^{s};B^{s*}) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i,i'=1}^{m}\sum_{j,j'=1}^{d} \frac{1}{2}\left(\frac{\partial^2\ell(B^{s};\Sigma_k)}{\partial B^{s}_{ij}\partial B^{s}_{i'j'}} - \frac{\partial^2\ell(B^{s};\overline{\Sigma})}{\partial B^{s}_{ij}\partial B^{s}_{i'j'}}\right)_{B^{s}=B^{s*}} \left(B^{s}_{ij} - B^{s*}_{ij}\right)\left(B^{s}_{i'j'} - B^{s*}_{i'j'}\right),$$

where $\ell(B^{s};\Sigma) \equiv -\log\det\left(B^{s}\Sigma B^{s\top}\right)$ and

$$\frac{1}{2}\frac{\partial^2\ell(B^{s};\Sigma)}{\partial B^{s}_{ij}\partial B^{s}_{i'j'}} = -\left(B^{s}\Sigma B^{s\top}\right)^{-1}_{ii'}\Sigma_{jj'} + \left(B^{s}\Sigma B^{s\top}\right)^{-1}_{ii'}\Sigma_{j\cdot}B^{s\top}\left(B^{s}\Sigma B^{s\top}\right)^{-1}B^{s}\Sigma_{\cdot j'} + \left(B^{s}\Sigma B^{s\top}\right)^{-1}_{i\cdot}B^{s}\Sigma_{\cdot j'}\left(B^{s}\Sigma B^{s\top}\right)^{-1}_{i'\cdot}B^{s}\Sigma_{\cdot j}.$$

Here, $C_{i\cdot}$ and $C_{\cdot j}$ denote the ith row and jth column vectors of a matrix C, respectively. From the assumption $B^{s*}\Sigma_k B^{s*\top} = B^{s*}\overline{\Sigma}B^{s*\top} = I_m$, we derive the following simpler expression:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\left(\frac{\partial^2\ell(B^{s};\Sigma_k)}{\partial B^{s}_{ij}\partial B^{s}_{i'j'}} - \frac{\partial^2\ell(B^{s};\overline{\Sigma})}{\partial B^{s}_{ij}\partial B^{s}_{i'j'}}\right)_{B^{s}=B^{s*}}\\
&= \frac{1}{K}\sum_{k=1}^{K}\Big(-I_{m,ii'}\Sigma_{k,jj'} + I_{m,ii'}\overline{\Sigma}_{jj'} + I_{m,ii'}\Sigma_{k,j\cdot}B^{s*\top}B^{s*}\Sigma_{k,\cdot j'} - I_{m,ii'}\overline{\Sigma}_{j\cdot}B^{s*\top}B^{s*}\overline{\Sigma}_{\cdot j'}\\
&\qquad + I_{m,i\cdot}B^{s*}\Sigma_{k,\cdot j'}I_{m,i'\cdot}B^{s*}\Sigma_{k,\cdot j} - I_{m,i\cdot}B^{s*}\overline{\Sigma}_{\cdot j'}I_{m,i'\cdot}B^{s*}\overline{\Sigma}_{\cdot j}\Big)\\
&= \frac{1}{K}\sum_{k=1}^{K}\Big\{I_{m,ii'}\left(\Sigma_k - \overline{\Sigma}\right)_{j\cdot}B^{s*\top}B^{s*}\left(\Sigma_k - \overline{\Sigma}\right)_{\cdot j'}\\
&\qquad + I_{m,i\cdot}B^{s*}\left(\Sigma_k - \overline{\Sigma}\right)_{\cdot j'}I_{m,i'\cdot}B^{s*}\left(\Sigma_k - \overline{\Sigma}\right)_{\cdot j}\Big\},
\end{aligned}$$

where the last equality holds from $\overline{\Sigma} = \sum_{k=1}^{K}\Sigma_k/K$. By using $\sum_{i,i',j,j'}C_{ii'}D_{jj'}X_{ij}X_{i'j'} = \mathrm{tr}\left[C^{\top}XDX^{\top}\right]$ and $B^{s*}\Sigma_k B^{s*\top} = B^{s*}\overline{\Sigma}B^{s*\top}$, we obtain the resulting second order Taylor approximation $\tilde{f}(B^{s};B^{s*})$ as

$$\tilde{f}(B^{s};B^{s*}) = \frac{1}{K}\sum_{k=1}^{K}\Big\{\mathrm{tr}\Big[B^{s}\left(\Sigma_k - \overline{\Sigma}\right)B^{s*\top}B^{s*}\left(\Sigma_k - \overline{\Sigma}\right)B^{s*\top}\Big] + \mathrm{tr}\Big[B^{s}\left(\Sigma_k - \overline{\Sigma}\right)B^{s*\top}B^{s}\left(\Sigma_k - \overline{\Sigma}\right)B^{s*\top}\Big]\Big\}.$$

Then, we first derive the following upper bound:

$$\tilde{f}(B^{s};B^{s*}) \le \frac{2}{K}\sum_{k=1}^{K}\mathrm{tr}\Big[B^{s}\left(\Sigma_k - \overline{\Sigma}\right)B^{s*\top}B^{s*}\left(\Sigma_k - \overline{\Sigma}\right)B^{s\top}\Big],$$

from the Cauchy-Schwarz inequality

$$\left|\langle C_k, C_k^\top \rangle\right| \leq \sqrt{\langle C_k, C_k \rangle \langle C_k^\top, C_k^\top \rangle} = \langle C_k, C_k \rangle,$$

where $\langle C, D \rangle = \mathrm{tr}\left[CD^\top\right]$ is an inner product of matrices C and D, and $C_k = B^{\mathrm{s}}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{s}*\top}$.

Here, let $B^{\mathrm{n}*} \in \mathbb{R}^{(d-m)\times d}$ denote a $\overline{\Sigma}$-orthonormal complement of $B^{\mathrm{s}*}$, that is, $B^{\mathrm{n}*}\overline{\Sigma}B^{\mathrm{s}*\top} = 0_{(d-m)\times m}$, $B^{\mathrm{n}*}\overline{\Sigma}B^{\mathrm{n}*\top} = I_{d-m}$. We then derive the following further upper bound from the fact that $\mathrm{tr}\left[B^{\mathrm{s}}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{n}*\top}B^{\mathrm{n}*}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{s}\top}\right] \geq 0$ holds for any B^{s}:

$$\begin{aligned}
\tilde{f}(B^{\mathrm{s}}; B^{\mathrm{s}*}) &\leq \frac{2}{K}\sum_{i=1}^{K}\Big\{\mathrm{tr}\left[B^{\mathrm{s}}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{s}*\top}B^{\mathrm{s}*}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{s}\top}\right] \\
&\qquad + \mathrm{tr}\left[B^{\mathrm{s}}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{n}*\top}B^{\mathrm{n}*}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{s}\top}\right]\Big\} \\
&= \frac{2}{K}\sum_{k=1}^{K}\mathrm{tr}\left[B^{\mathrm{s}}\left(\Sigma_k - \overline{\Sigma}\right)B^{*\top}B^{*}\left(\Sigma_k - \overline{\Sigma}\right)B^{\mathrm{s}\top}\right],
\end{aligned}$$

where $B^* = \left[B^{\mathrm{s}*\top} \quad B^{\mathrm{n}*\top}\right]^\top$. Moreover, $B^{*\top}B^* = \overline{\Sigma}^{-1}$ follows from $B^*\overline{\Sigma}B^{*\top} = I_d$ and we derive (2.8). □

2.8.5.2 Proof of Theorem 5

The theorem is obvious from following lemmas.

Lemma 1 (ASSA and Uncorrelated SSA). *The ASSA solution $\hat{B}^{\mathrm{s}}_{\mathrm{A}}$ is optimal when stationary and non-stationary sources are group-wise uncorrelated, that is, the covariance between the two groups is zero.*

(proof) If stationary and non-stationary sources are group-wise uncorrelated, Σ_k is in the following block-diagonal form:

$$\Sigma_k = \left[A^{\mathrm{s}} \quad A^{\mathrm{n}}\right]\begin{bmatrix}\Sigma^{\mathrm{s}} & 0_{m\times(d-m)} \\ 0_{(d-m)\times m} & \Sigma^{\mathrm{n}}_k\end{bmatrix}\left[A^{\mathrm{s}} \quad A^{\mathrm{n}}\right]^\top, \tag{2.17}$$

where Σ^{s} is a covariance matrix of stationary sources, which is constant across epochs, and Σ^{n}_k is a covariance of non-stationary sources in the kth epoch. From

this block-diagonal structure and the orthogonality between B^{s*} and A^{n}, $B^{s*}\Sigma_k = B^{s*}A^{s}\Sigma^{s}A^{s\top}$ holds. Since this is independent of the epoch index k, we derive the relations on B^{s*}:

$$B^{s*}\boldsymbol{\mu}_k = B^{s*}\overline{\boldsymbol{\mu}} \text{ and } B^{s*}\Sigma_k = B^{s*}\overline{\Sigma}, \tag{2.18}$$

for $k = 1, 2, \ldots, K$. It is obvious that the ASSA objective function (2.9) gets zero at $B^{s} = B^{s*}$. Since the ASSA objective function is non-negative, B^{s*} is a minimizer. □

Lemma 2 (Equivalent Class of Uncorrelated SSA). *Any SSA model (2.1) with a time-constant covariance between stationary and non-stationary sources can be reduced to the equivalent model with group-wise uncorrelated sources.*

(proof) Let $\Sigma^{sn} \in \mathbb{R}^{m\times(d-m)}$ be a time-constant covariance matrix between stationary and non-stationary sources. The equivalent uncorrelated SSA model is then given by

$$\boldsymbol{x}(t) = \begin{bmatrix} A + B\Sigma^{sn}\Sigma^{s-1} & B \end{bmatrix} \begin{bmatrix} \boldsymbol{s}^{s}(t) \\ \boldsymbol{s}^{n}(t) - \Sigma^{sn\top}\Sigma^{s-1}\boldsymbol{s}^{s}(t) \end{bmatrix},$$

where $\Sigma^{s} \in \mathbb{R}^{m\times m}$ is a covariance matrix of stationary sources and thus time-constant. □

2.8.5.3 Proof of Theorem 6

From the Lemma 2, it is sufficient to prove for the case of group-wise uncorrelated sources. Under the uncorrelated model, the conditions (2.4) is replaced by (2.18). Let $\mathcal{B}$ denote a set of the ASSA solutions that satisfies (2.18) and the constraint $B^{s}\overline{\Sigma}B^{s\top} = I_m$. The uniqueness of the ASSA solution is guaranteed (up to linear transformation) if $\mathrm{span}(B^{s\top}) = \mathrm{span}(B^{s'\top})$ holds for any solutions $B^{s}, B^{s'} \in \mathcal{B}$. It holds when $\boldsymbol{b} \in \bigcup_{B^{s}\in\mathcal{B}} \mathrm{span}(B^{s\top})$ has degrees of freedom equal to or less than $m-1$ where the -1 stems from the constraint. The conditions (2.18) impose the following $K(d+1)$ constraints:

$$(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})^{\top}\boldsymbol{b} = 0, \tag{2.19}$$

$$(\Sigma_k - \overline{\Sigma})\boldsymbol{b} = \mathbf{0}_d, \tag{2.20}$$

where $k = 1, 2, \ldots, K$. However, note that not all of them are independent. Since at most $\mathrm{rank}(\Sigma_k - \overline{\Sigma})$ equations are independent in (2.20), the expected number of independent constraints is $K(\nu+1)$. Conditions (2.19) and (2.20) also include $d - m + 1$ dependent equations since their sums are obviously zeros from the definition of $\overline{\boldsymbol{\mu}}$ and $\overline{\Sigma}$, that is,

$$\left(\sum_{k=1}^{K} \boldsymbol{\mu}_k - K\overline{\boldsymbol{\mu}}\right)^{\top} \boldsymbol{b} = 0 \quad \text{and} \quad \left(\sum_{k=1}^{K} \Sigma_k - K\overline{\Sigma}\right)\boldsymbol{b} = \mathbf{0}_d.$$

Therefore, the total number of independent constraints is $K(\nu+1) - (d-m+1)$ and $\boldsymbol{b}$ has $d - K(\nu+1) + (d+m-1)$ degrees of freedom. Since this has to be equal to or less than $m-1$, (2.12) follows. In the special case when the mean is constant, the condition (2.19) vanishes and we have $K\nu - (d-m)$ independent constraints. The degrees of freedom on $\boldsymbol{b}$ is $d - K\nu + (d-m)$ and (2.13) holds. □

2.8.5.4 Proof of Theorem 7

Let a matrix $B = \left[B^{\mathrm{s}\top} \quad B^{\mathrm{n}\top}\right]^{\top}$. The block-off-diagonal element of $B\Sigma_k B^{\top}$ is then $B^{\mathrm{s}}\Sigma_k B^{\mathrm{n}\top}$. Therefore, the objective function of (2.15) satisfies the following inequality:

$$\begin{aligned}
&\sum_{k=1}^{K} \mathrm{tr}\left[B^{\mathrm{s}}\Sigma_k B^{\mathrm{n}\top} B^{\mathrm{n}}\Sigma_k B^{\mathrm{s}\top}\right] \\
&\leq \sum_{k=1}^{K} \left\{\mathrm{tr}\left[B^{\mathrm{s}}\Sigma_k B^{\mathrm{n}\top} B^{\mathrm{n}}\Sigma_k B^{\mathrm{s}\top}\right] + \mathrm{tr}\left[B^{\mathrm{s}}\Sigma_k B^{\mathrm{s}\top} B^{\mathrm{s}}\Sigma_k B^{\mathrm{s}\top}\right]\right\} \\
&= \sum_{k=1}^{K} \mathrm{tr}\left[B^{\mathrm{s}}\Sigma_k \overline{\Sigma}^{-1}\Sigma_k B^{\mathrm{s}\top}\right], \qquad (2.21)
\end{aligned}$$

where the last equality follows from $B^{\mathrm{s}\top}B^{\mathrm{s}} + B^{\mathrm{n}\top}B^{\mathrm{n}} = B^{\top}B$ and $B\overline{\Sigma}B^{\top} = I_d$. The equivalence of (2.21) to the ASSA objective function (2.9) can be checked with some algebra with a condition $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_K$. Hence, $B^{\mathrm{s}}\Sigma_k B^{\mathrm{s}\top} = B^{\mathrm{s}}\overline{\Sigma}B^{\mathrm{s}\top} = I_m$ is a necessary condition for a minimizer of (2.21) to coincide with the minimizer of (2.15), which is guaranteed by Lemma 1. □

2.8.5.5 Proof of Theorem 8

We first show that the matrix S is given in a form (2.16). From the definition, we have the average covariance $\overline{\Sigma}$ as

$$\overline{\Sigma} = A \begin{bmatrix} \Sigma^{\mathrm{s}} & \epsilon \overline{\Sigma}^{\mathrm{sn}} \\ \epsilon \overline{\Sigma}^{\mathrm{sn}\top} & \overline{\Sigma}^{\mathrm{n}} \end{bmatrix} A^{\top},$$

where $\overline{\Sigma}^{\mathrm{sn}} = \sum_{k=1}^{K} \Sigma_k^{\mathrm{sn}}/K$ and $\overline{\Sigma}^{\mathrm{n}} = \sum_{k=1}^{K} \Sigma_k^{\mathrm{n}}/K$. Hence, its inverse is given by

$$\begin{aligned} \overline{\Sigma}^{-1} &= A^{-\top} \begin{bmatrix} \Sigma^{\mathrm{s}-1} + \epsilon^2 \Sigma^{\mathrm{s}-1} \overline{\Sigma}^{\mathrm{sn}} J \overline{\Sigma}^{\mathrm{sn}\top} \Sigma^{\mathrm{s}-1} & -\epsilon \Sigma^{\mathrm{s}-1} \overline{\Sigma}^{\mathrm{sn}} J \\ -\epsilon J \overline{\Sigma}^{\mathrm{sn}\top} \Sigma^{\mathrm{s}-1} & J \end{bmatrix} A^{-1} \\ &= A^{-\top} \begin{bmatrix} \Sigma^{\mathrm{s}-1} & -\epsilon \Sigma^{\mathrm{s}-1} \overline{\Sigma}^{\mathrm{sn}} \overline{\Sigma}^{n-1} \\ -\epsilon \overline{\Sigma}^{n-1} \overline{\Sigma}^{\mathrm{sn}\top} \Sigma^{\mathrm{s}-1} & \overline{\Sigma}^{n-1} \end{bmatrix} A^{-1} + \mathcal{O}(\epsilon^2), \end{aligned}$$

where $J = \left(\overline{\Sigma}^{\mathrm{n}} - \epsilon^2 \overline{\Sigma}^{\mathrm{sn}\top} \Sigma^{\mathrm{s}-1} \overline{\Sigma}^{\mathrm{sn}} \right)^{-1} = \overline{\Sigma}^{\mathrm{n}-1} + \mathcal{O}(\epsilon^2)$. Using this expression, we can write down the product $\Sigma_k \overline{\Sigma}^{-1} \Sigma_k$ as

$$\Sigma_k \overline{\Sigma}^{-1} \Sigma_k = A \begin{bmatrix} \Sigma^{\mathrm{s}} & \epsilon Q_k \\ \epsilon Q_k^{\top} & \Sigma_k^{\mathrm{n}} \overline{\Sigma}^{\mathrm{n}-1} \Sigma_k^{\mathrm{n}} \end{bmatrix} A^{\top} + \mathcal{O}(\epsilon^2),$$

where $Q_k = \Sigma_k^{\mathrm{sn}} + \Sigma_k^{\mathrm{n}} \overline{\Sigma}^{\mathrm{n}-1} \left(\Sigma_k^{\mathrm{sn}} - \overline{\Sigma}^{\mathrm{sn}} \right)$, and the matrix S takes a form (2.16) with Q and P defined as

$$\begin{aligned} Q &= \frac{2}{K} \sum_{k=1}^{K} \left(\Sigma_k^{\mathrm{sn}} - \overline{\Sigma}^{\mathrm{sn}} \right) \left(I_{d-m} + \Sigma_k^{\mathrm{n}} \overline{\Sigma}^{\mathrm{n}-1} \right), \\ P &= \frac{1}{K} \sum_{k=1}^{K} \left(\boldsymbol{\mu}_k^{\mathrm{n}} \boldsymbol{\mu}_k^{\mathrm{n}\top} + 2 \Sigma_k^{\mathrm{n}} \overline{\Sigma}^{\mathrm{n}-1} \Sigma_k^{\mathrm{n}} \right) - \overline{\boldsymbol{\mu}}^{\mathrm{n}} \overline{\boldsymbol{\mu}}^{\mathrm{n}\top} - 2 \overline{\Sigma}^{\mathrm{n}}. \end{aligned}$$

Here, we used notations $\boldsymbol{\mu}_k = A \left[\boldsymbol{\mu}_k^{\mathrm{s}\top} \quad \boldsymbol{\mu}_k^{\mathrm{n}\top} \right]^{\top}$ and $\overline{\boldsymbol{\mu}} = A \left[\overline{\boldsymbol{\mu}}^{\mathrm{s}\top} \quad \overline{\boldsymbol{\mu}}^{\mathrm{n}\top} \right]^{\top}$.

Now, we turn to proving the main claim. Let $\hat{B}_{\mathrm{A},\epsilon}^{\mathrm{s}}$ denote the ASSA solution under some fixed $\epsilon \geq 0$. Note that $\hat{B}_{\mathrm{A},0}^{\mathrm{s}} = B^{\mathrm{s}*}$ holds from the Theorem 5. Here, we assume that $B^{\mathrm{s}*}$ satisfies $B^{\mathrm{s}*} \overline{\Sigma} B^{\mathrm{s}*\top} = I_m$. Hence, from the continuity of $\hat{B}_{\mathrm{A},\epsilon}^{\mathrm{s}}$ over ϵ, there exists some $0 \leq \eta \ll 1$ and $C \in \mathbb{R}^{m \times d}$ such that $\mathcal{O}(C) = \mathcal{O}(1)$ and $\hat{B}_{\mathrm{A},\epsilon}^{\mathrm{s}} = B^{\mathrm{s}*} + \eta C$ for $0 \leq \epsilon \ll 1$. With this expression, we can write down products

$\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}\overline{\Sigma}\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon}$ and $\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}S\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon}$ as follows:

$$\begin{aligned}
\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}\overline{\Sigma}\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon} &= I_m + \eta\Big(CA^{\mathrm{s}}\Sigma^{\mathrm{s}}A^{\mathrm{s}\top}B^{\mathrm{s}*\top} + B^{\mathrm{s}*}A^{\mathrm{s}}\Sigma^{\mathrm{s}}A^{\mathrm{s}\top}C^{\top}\Big) \\
&\quad + \epsilon\eta\Big(CA^{\mathrm{n}}\overline{\Sigma}^{\mathrm{sn}\top}A^{\mathrm{s}\top}B^{\mathrm{s}*\top} + B^{\mathrm{s}*}A^{\mathrm{s}}\overline{\Sigma}^{\mathrm{sn}}A^{\mathrm{n}\top}C^{\top}\Big) \\
&\quad + \eta^2\Big(CA^{\mathrm{s}}\Sigma^{\mathrm{s}}A^{\mathrm{s}\top}C^{\top} + CA^{\mathrm{n}}\overline{\Sigma}^{\mathrm{n}}A^{\mathrm{n}\top}C^{\top}\Big) + \mathcal{O}(\epsilon\eta^2), \\
\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}S\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon} &= \epsilon\eta\Big(CA^{\mathrm{n}}Q^{\top}A^{\mathrm{s}\top}B^{\mathrm{s}*\top} + B^{\mathrm{s}*}A^{\mathrm{s}}QA^{\mathrm{n}\top}C^{\top}\Big) \\
&\quad + \eta^2CA^{\mathrm{n}}PA^{\mathrm{n}\top}C^{\top} + \mathcal{O}(\epsilon\eta^2).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathrm{tr}\left[\left(\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}\overline{\Sigma}\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon}\right)^{-1}\left(\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}S\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon}\right)\right] &= 2\epsilon\eta\,\mathrm{tr}\left[CA^{\mathrm{n}}Q^{\top}A^{\mathrm{s}\top}B^{\mathrm{s}*\top}\right] \\
&\quad + \eta^2\mathrm{tr}\left[CA^{\mathrm{n}}PA^{\mathrm{n}\top}C^{\top}\right] + \mathcal{O}(\epsilon\eta^2),
\end{aligned}$$

because $\left(\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}\overline{\Sigma}\hat{B}^{\mathrm{s}\top}_{\mathrm{A},\epsilon}\right)^{-1} = I_m + \mathcal{O}(\eta)$. As we have discussed in Section 2.3, the ASSA solution $\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}$ is a minimizer of the above. Note that the problem further reduces to finding an optimal ηC since $B^{\mathrm{s}*}$ is a constant matrix. By setting the derivative over ηCA^{n} equal to zero, we obtain

$$\eta CA^{\mathrm{n}} = -\epsilon B^{\mathrm{s}*}A^{\mathrm{s}}QP^{-1} + \mathcal{O}(\epsilon\eta).$$

This is equivalent to $\hat{B}^{\mathrm{s}}_{\mathrm{A},\epsilon}A^{\mathrm{n}} = (B^{\mathrm{s}*} + \eta C)A^{\mathrm{n}}$ since $B^{\mathrm{s}*}A^{\mathrm{n}} = 0_{m\times(d-m)}$, and we have the claim. □

Chapter 3

Sparse Inverse Covariance Selection with a Dual Augmented Lagrangian Method

In Chapter 3-5, we consider a Graphical Gaussian Model (GGM) learning problem introduced in Section 1.5, and extend it by introducing a notion of invariance into the model. Across chapters, the Sparse Inverse Covariance Selection (SICS) problem (1.11) is the basis of our framework. In this chapter, before we go into the techniques for finding some invariance in GGM, we construct a technical foundation used in upcoming chapters, which is a general convex optimization method for GGM learning problems.

3.1 Introduction

SICS is the maximum likelihood estimation problem of a precision matrix Λ under a sparsity constraint. In (1.11), an element-wise ℓ_1-norm is used as the regularization term. Although this is the most basic formulation considered by a number of authors (Meinshausen & Bühlmann, 2006; M. Yuan & Lin, 2007; Banerjee et al., 2008), it is not the unique regularization term used in several problems. The SICS problem (1.11) can be expressed more generally in the following form:

$$\begin{aligned} &\max_{\Lambda \in \mathcal{S}^+} \ell(\Lambda; \hat{\Sigma}) - \varphi_\rho(\Lambda), \\ &\ell(\Lambda; \hat{\Sigma}) \equiv \log\det\Lambda - \mathrm{tr}\Big[\hat{\Sigma}\Lambda\Big], \end{aligned} \tag{3.1}$$

where $\varphi_\rho(\Lambda)$ is an arbitrary regularization term parametrized by ρ. In most cases, some sparsity inducing norms are used as $\varphi_\rho(\Lambda)$. For instance, Duchi, Gould, and Koller (2008) and Schmidt, Van Den Berg, Friedlander, and Murphy (2009) considered a grouped feature case and introduced a group regularization term instead of an ℓ_1-norm, while Honorio, Ortiz, Samaras, Paragios, and Goldstein (2009) considered a spatial structure in a precision matrix and introduced a fused regularization to promote common constant entries in the estimator. Note that the problem (3.1) is a convex problem as long as $\varphi_\rho(\Lambda)$ is a convex function, and a global solution can be derived with some proper optimization methods[1]. The objective of this chapter is to provide one such algorithm. For the SICS problem (1.11), several optimization procedures have been proposed (Scheinberg et al., 2010; Duchi, Gould, & Koller, 2008; Friedman et al., 2008; X. Yuan, 2009; Scheinberg & Rish, 2010; Hsieh et al., 2011). Amongst these methods, QUIC (Hsieh et al., 2011) would be the most practical state-of-the-art method with some theoretical guarantees. However, the efficiency of QUIC heavily depends on the specific property of the ℓ_1-norm and it is not applicable to the general regularization term. In this chapter, we consider the case when the regularization term $\varphi_\rho(\Lambda)$ is convex and the proximity operator (Rockafellar, 1996) defined on the convex conjugate of $\varphi_\rho(\Lambda)$ can be efficiently computed. This assumption involves $\varphi_\rho(\Lambda) = \rho \|\Lambda\|_1$ as its special case, that is, an algorithm proposed in this chapter, which we call DAL-ADMM, has wider flexibility on the regularization term compared to algorithms specific to the ℓ_1-regularization such as QUIC.

The main scope of this chapter is to propose a new algorithm for the generalized SICS problem (3.1), which can treat general regularization terms other than the ℓ_1-norm. The proposed method relies on the Dual Augmented Lagrangian (DAL) method (Tomioka, Suzuki, & Sugiyama, 2011) which provides an efficient algorithm for convex and sparse regularization problems. We further update the DAL framework by combining the Alternating Direction Method of Multipliers (ADMM) (Scheinberg et al., 2010; X. Yuan, 2009; Boyd, Parikh, Chu, Peleato, &

[1]Some authors also considered non-convex regularization terms, see J. Guo, Levina, Michailidis, and Zhu (2011) for instance. In such cases, the global optimality of the solution is no longer guaranteed in general.

Eckstein, 2011) and propose a DAL-ADMM algorithm. This update makes the entire procedure dramatically simple and helps reducing the practical computational cost.

The remainder of this chapter is organized as follows. In Section 3.2, we review the extended SICS problem with a group structure as a specific example of (3.1). In Section 3.3, we introduce the DAL based optimization method, and then update it by combining ADMM and propose DAL-ADMM algorithm in Section 3.4. The validity of the proposed method is presented through synthetic experiments in Section 3.5. Finally, we conclude the chapter in Section 3.6.

3.2 Sparse Inverse Covariance Selection and Its Group Extension

In this section, we briefly review the extension of SICS into its grouped variant (Duchi, Gould, & Koller, 2008; Schmidt et al., 2009) as one specific example of a generalized formulation (3.1). This extended group SICS model is helpful when we aim to find the dependency between the set of variables.

In group SICS, all d^2 entries in a precision matrix Λ are partitioned into M disjoint groups. Here, let $\mathcal{I}$ be a set of all d^2 indices in Λ, that is, $\mathcal{I} = \{(i,j); i,j = 1,2,\ldots,d\}$. Each of M groups $\mathcal{G}_m\,(m = 1,2,\ldots,M)$ is then represented as a subset of $\mathcal{I}$ where $\mathcal{G}_m \cap \mathcal{G}_{m'} = \phi$ for $m \neq m'$ and $\cup_{m=1}^{M}\mathcal{G}_m = \mathcal{I}$. We also use a notation $\Lambda_{\mathcal{G}_m}$ to represent a vector composed of entries in Λ specified by $\mathcal{G}_m$, that is, $\Lambda_{\mathcal{G}_m} = (\Lambda_{ij})_{(i,j)\in\mathcal{G}_m}$. While the objective of the ordinal SICS is to identify whether each (i,j)th entry of Λ is zero or not, the objective of group SICS is to infer which of $\Lambda_{\mathcal{G}_m}$ gets simultaneously zeros among M groups. For example, this setting is relevant to the identification of dependencies between two sets of genes. In such a case, we partition the entries of Λ into four disjoint groups; two of them corresponds to the block-diagonal entries representing inner group interactions while the other twos specify block-off-diagonal entries related to the interaction between the groups. If latter two entries are simultaneously zeros, it implies that two sets of genes do not involve any interactions between them.

Duchi, Gould, and Koller (2008) and Schmidt et al. (2009) formulated this problem as follows using group regularization techniques (Turlach et al., 2005; M. Yuan & Lin, 2006):

$$\max_{\Lambda \in \mathcal{S}^+} \ell(\Lambda; \hat{\Sigma}) - \sum_{m=1}^{M} \rho_m \left\| \Lambda_{\mathcal{G}_m} \right\|_{p_m}. \tag{3.2}$$

Here, $\left\| \Lambda_{\mathcal{G}_m} \right\|_{p_m}$ is an ℓ_{p_m}-norm[2] of $\Lambda_{\mathcal{G}_m}$ with $p_m \in [1, \infty]$ and parameters ρ_m and p_m are assigned individually to each group. Note that this is one specific variant of the problem (3.1) with $\varphi_\rho(\Lambda) = \sum_{m=1}^{M} \rho_m \left\| \Lambda_{\mathcal{G}_m} \right\|_{p_m}$ and $\rho = (\rho_1, \rho_2, \ldots, \rho_M)^\top$. This can be also seen as a generalization of SICS since setting $\rho_m = \rho$ and $p_m = 1$ results in (1.11). For $p_m > 1$, a set of parameters $\Lambda_{\mathcal{G}_m}$ shrinks to zeros simultaneously owing to the group effect. Hence, the optimal solution Λ^* has a group-wise sparse structure. A parameter p_m is typically set to be 2 or ∞ due to computational considerations.

More generally, in what follows, we assume $\varphi_\rho(\Lambda)$ is a convex possibly non-differentiable function. Therefore we cannot merely apply ordinal gradient ascent based methods to solve the problem (3.1). In addition, we assume that for all $\beta > 0$, $\eta\varphi_\rho(\Lambda) = \varphi_{\rho\beta}(\Lambda)$, that is, a multiplication of β to the regularization term is equivalent to the regularization term parameterized by $\rho\beta$. Note that this is the generalization of a multiplicative form $\varphi_\rho(\Lambda) = \rho f(\Lambda)$. Another important assumption is that the following proximity operator (Rockafellar, 1996) can be computed efficiently:

$$\mathrm{prox}_{\varphi_\rho^*}(B) = \operatorname*{argmin}_Y \varphi_\rho^*(Y) + \frac{1}{2} \left\| Y - B \right\|_{\mathrm{F}}^2,$$

where φ_ρ^* is a convex conjugate[3] of φ_ρ. An ℓ_1-regularization $\varphi_\rho(\Lambda) = \rho \left\| \Lambda \right\|_1$ and a group regularization $\varphi_\rho(\Lambda) = \sum_{m=1}^{M} \rho_m \left\| \Lambda_{\mathcal{G}_m} \right\|_{p_m}$ with $p_m = 2$ are the examples that this proximity operator can be computed analytically. For instance, the convex conjugate of $\varphi_\rho(\Lambda) = \left\| \Lambda \right\|_1$ is an indicator function defined as

$$\varphi_\rho^*(Y) = \begin{cases} 0 & \text{if } \left\| Y \right\|_\infty \leq \rho, \\ \infty & \text{otherwise,} \end{cases}$$

[2] An ℓ_p-norm of a vector $\boldsymbol{x}$ is given by $\left\| \boldsymbol{x} \right\|_p \equiv (\sum_i |x_i|^p)^{\frac{1}{p}}$ for $p \in [1, \infty)$, and $\left\| \boldsymbol{x} \right\|_\infty \equiv \max_i |x_i|$.

[3] A convex conjugate of a function $f(\boldsymbol{x})$ is defined as $f^*(\boldsymbol{y}) \equiv \sup_{\boldsymbol{x}} \boldsymbol{y}^\top \boldsymbol{x} - f(\boldsymbol{x})$.

where $\|Y\|_\infty$ is the dual of an ℓ_1-norm and is given by $\|Y\|_\infty = \max_{i,j} |Y_{ij}|$. The computation of a proximity operator can be casted as an Euclidian projection of B onto the set $\mathcal{A} = \{Y; \|Y\|_\infty \leq \rho\}$. This problem can be factorized into element-wise subproblems:

$$\min_y \frac{1}{2}(y-b)^2, \quad \text{s.t.} \quad -\rho \leq y \leq \rho,$$

for each (i,j)th entry $y = Y_{ij}$ and $b = B_{ij}$. The solution to this problem is analytically given by $y = \min(1, \rho/|b|)b$, and the proximity operator can be expressed as

$$\text{prox}_{\varphi_\rho^*}(B) = \left(\min\left(1, \frac{\rho}{|B_{ij}|}\right) B_{ij} \right)_{i,j=1,2,\ldots,d}.$$

The proximity operator for the group regularization with $p_m = 2$ is derived in the similar manner and is given by

$$\text{prox}_{\varphi_\rho^*}(B) = \left(\min\left(1, \frac{\rho_m}{\|B_{\mathcal{G}_m}\|_2}\right) B_{\mathcal{G}_m} \right)_{m=1,2,\ldots,M}.$$

3.3 Dual Augmented Lagrangian for SICS

Now, we derive the algorithm for generalized SICS (3.1) using Dual Augmented Lagrangian (DAL) (Tomioka et al., 2011). DAL is an algorithm applying an Augmented Lagrangian technique (Boyd et al., 2011) to the dual of the target problem. It is known that DAL is super-linearly convergent, hence it is well suited for sparse regularization problems (Tomioka et al., 2011).

The dual of generalized SICS (3.1) is given by

$$\min_{W \in \mathcal{S}^+, Y} -\log\det W + \varphi_\rho^*(Y), \quad \text{s.t.} \quad W + Y - \hat{\Sigma} = 0_{d\times d}, \tag{3.3}$$

following the Fenchel duality theorem (Rockafellar, 1996). Here, $W \in \mathbb{R}^{d\times d}$ is a dual parameter, which satisfies $W^* = \Lambda^{*-1}$ at its optimal from the duality. We have also introduced the additional parameter Y for the sake of compatibility with the latter discussion. In DAL, we first formulate the following Augmented Lagrangian function:

$$\mathcal{L}_\beta(W, Y, Z) = -\log\det W + \varphi_\rho^*(Y) + \frac{\beta}{2}\left\| W + Y + \frac{1}{\beta}Z - \hat{\Sigma} \right\|_{\text{F}}^2,$$

where $\beta > 0$ is an algorithm parameter and $Z \in \mathbb{R}^{d \times d}$ is a Lagrange multiplier. Note that an Augmented Lagrangian function with $\beta \to 0$ corresponds to the ordinal Lagrangian function. The basic approach of DAL is to relax the equality constraint in (3.3) in the intermediate steps of the algorithm and make it fulfilled at the termination. In DAL, we repeat the following two updating steps till the convergence:

$$\begin{cases} W^{(t+1)}, Y^{(t+1)} \in \underset{W \in \mathcal{S}^+, Y}{\operatorname{argmin}} \mathcal{L}_\beta(W, Y, Z^{(t)}), \\ Z^{(t+1)} = Z^{(t)} + \beta\left(W^{(t+1)} + Y^{(t+1)} - \hat{\Sigma}\right). \end{cases}$$

In every steps, a value of β is also gradually increased so that the super-linear convergence is achieved (Tomioka et al., 2011). For the cases of an ℓ_1 and a group regularization with $p_m = 2$, we can analytically write down $Y^{(t+1)}$ as a function of $W^{(t+1)}$. By plugging-in this analytic expression, we can further reduce the first problem into the following unified form (Tomioka et al., 2011) using Moreau's decomposition (J. J. Moreau, 1965):

$$W^{(t+1)} \in \underset{W \in \mathcal{S}^+}{\operatorname{argmin}} - \log\det W + \frac{1}{2\beta} \left\| \operatorname{prox}_{\varphi_{\rho\beta}}(-\beta W - Z^{(t)} + \beta\hat{\Sigma}) \right\|_{\mathrm{F}}^2.$$

See Sra, Nowozin, and Wright (2011, Section 9.4.1 and 9.8.2) for the detail. This is a smooth convex optimization problem and is solvable with some proper methods such as a quasi-Newton method.

3.4 SICS via DAL-ADMM

The DAL algorithm derived in the preceding section has a super-linear convergence property. This property is based on the simultaneous update of W and Y and a gradual increase of β in every steps. However, SICS involves $\mathcal{O}(d^2)$ free parameters to be optimized and hence the computation of the gradient over W requires $\mathcal{O}(d^3)$ complexity owing to the log-determinant term. This can be too demanding even for middle sized d. Therefore, we need to reduce the number of gradient evaluations so that the entire procedure to become much more efficient. In this section, we tackle this problem by introducing an idea of Alternating Direction Method of Multipliers

(ADMM) (Scheinberg et al., 2010; X. Yuan, 2009; Boyd et al., 2011) and propose a DAL-ADMM algorithm.

In ADMM, we decouple the minimization of W and Y into sequential steps,

$$\begin{cases} W^{(t+1)} \in \underset{W \in \mathcal{S}^+}{\operatorname{argmin}} \mathcal{L}_\beta(W, Y^{(t)}, Z^{(t)}), \\ Y^{(t+1)} \in \underset{Y}{\operatorname{argmin}} \mathcal{L}_\beta(W^{(t+1)}, Y, Z^{(t)}). \end{cases}$$

It means that the optimization of $\mathcal{L}_\beta(W, Y, Z^{(t)})$ over W and Y is solved only in an approximate manner. Under this relaxation, as we see later, we can construct an analytic update procedure for W which requires only one eigenvalue decomposition in every update steps. This modification has another advantage that the second step, an update of Y, is exactly the same as the computation of the proximity operator on φ_ρ^*. Unlike DAL, we do not need to plug-in this result into the larger optimization problem. This allows us to use wider classes of regularizations; for instance, a group regularization with $p_m = \infty$ (Schmidt et al., 2009) which was difficult to treat with DAL. On the other hand, only a linear convergence is guaranteed for DAL-ADMM (He & Yuan, 2012). However, as we see in numerical experiments, a reduction of the number of gradient evaluation overwhelms this drawback and results in the faster computation. In the next subsection, we detail the above two update procedures.

3.4.1 Solutions to Inner Optimization Problems

The inner optimization problem over W is given by

$$\min_{W \in \mathcal{S}^+} -\log\det W + \frac{\beta}{2}\left\| W + Y^{(t)} + \frac{1}{\beta}Z^{(t)} - \hat{\Sigma} \right\|_{\mathrm{F}}^2.$$

By setting the derivative over W equal to zero, we derive the first order optimality condition:

$$W - \left(-Y^{(t)} - \frac{1}{\beta}Z^{(t)} + \hat{\Sigma} \right) - \frac{1}{\beta}W^{-1} = 0_{d \times d}.$$

Here, we consider the eigenvalue decomposition of the second term:

$$-Y^{(t)} - \frac{1}{\beta}Z^{(t)} + \hat{\Sigma} = U \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d) U^\top.$$

The solution $W^{(t+1)}$ is then given by

$$W^{(t+1)} = U\mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_d)U^\top,$$
$$\tilde{\sigma}_i = \frac{1}{2}\left(\sigma_i + \sqrt{\sigma_i^2 + \frac{4}{\beta}}\right).$$

See X. Yuan (2009) for the detail. Note that the positive definiteness of $W^{(t+1)}$ directly follows from this result.

As we already mentioned, an optimization of Y under a fixed W can be casted as the computation of a proximity operator on φ_ρ^*:

$$\begin{aligned}
&\operatorname*{argmin}_Y \varphi_\rho^*(Y) + \frac{\beta}{2}\left\|Y - \left(-W^{(t+1)} - \frac{1}{\beta}Z^{(t)} + \hat{\Sigma}\right)\right\|_{\mathrm{F}}^2 \\
&= \operatorname*{argmin}_Y \varphi_{\rho\beta}^*(\beta Y) + \frac{1}{2}\left\|\beta Y - \beta\left(-W^{(t+1)} - \frac{1}{\beta}Z^{(t)} + \hat{\Sigma}\right)\right\|_{\mathrm{F}}^2 \\
&= \frac{1}{\beta}\mathrm{prox}_{\varphi_{\rho\beta}^*}(-\beta W^{(t+1)} - Z^{(t)} + \beta\hat{\Sigma}).
\end{aligned}$$

Hence, this update step is efficiently computed as long as the proximity operator on φ_ρ^* is computationally cheap.

3.4.2 Convergence

Here, we list two convergence properties of DAL-ADMM under a fixed $\beta > 0$.

1. A sequence $\{Z^{(t)}\}_{t=1}^\infty$ converges to the optimal parameter $Z^* = \Lambda^*$.

2. A function value $\tilde{g}(W, Y) \equiv -\log\det W + \varphi_\rho^*(Y)$ converges linearly to its global minimum $\tilde{g}(W^*, Y^*)$.

These results can be shown as follows. We first get the optimality condition $Z^* = W^{*-1}$ by setting the derivative of $\mathcal{L}_0(W, Y, Z)$ equal to zero. Then, by applying the general theorem for ADMM (Boyd et al., 2011; He & Yuan, 2012) and recalling $W^* = \Lambda^{*-1}$, the claims follow.

3.4.3 Implementation Details

In our implementation of DAL-ADMM, we use following two *gaps* presented by Boyd et al. (2011) for the termination criteria:

$$\text{primal-gap} \equiv \left\| W^{(t+1)} + Y^{(t+1)} - \hat{\Sigma} \right\|_{\mathrm{F}},$$
$$\text{dual-gap} \equiv \beta \left\| Y^{(t+1)} - Y^{(t)} \right\|_{\mathrm{F}}.$$

When both of them are under a given threshold ϵ, we regard that the process has converged and stop the iteration. Here, two gaps measure how much the equality constraint in (3.3) and the optimality of parameters are fulfilled, respectively.

The choice of an algorithm parameter β also needs some consideration in practice. Unlike DAL, we can not merely increase β in every steps since it may lead to a non-optimal solution. In the proposed algorithm, we introduce the following heuristic from Boyd et al. (2011):

$$\beta^{(t+1)} = \begin{cases} 2\beta^{(t)} & \text{if primal-gap} \geq 10\,\text{dual-gap}, \\ 0.5\beta^{(t)} & \text{if dual-gap} \geq 10\,\text{primal-gap}, \\ \beta^{(t)} & \text{otherwise.} \end{cases}$$

This heuristic balances two gaps and makes them small simultaneously.

3.5 Simulation

In this section, we demonstrate the validity of DAL-ADMM through synthetic experiments. All simulations in this section have been conducted on Windows 7 (64bit), Intel Xeon W365 CPU machines with a 6GB RAM.

3.5.1 Data Description

In our simulations, we considered a group regularization problem with $p_m = 2$, that is, $\varphi_\rho(\Lambda) = \sum_{m=1}^{M} \rho_m \|\Lambda_{\mathcal{G}_m}\|_2$. We have generated data in the following manner. First, we give a number of Gaussian variables d and its partition $d_1, d_2, \ldots, d_K$ where $\sum_{k=1}^{K} d_k = d$. For each d_k, we generate elements of a random matrix $U_k \in \mathbb{R}^{d_k \times 5d_k}$

independently from a standard normal distribution $\mathcal{N}(0,1)$. We then generate a positive definite matrix $C_k = L_k L_k^\top$ and set the resulting precision matrix $\Lambda \in \mathbb{R}^{d\times d}$ to be a block-diagonal matrix with $C_1, C_2, \ldots, C_K$ on its block-diagonal. Here, each group $\mathcal{G}_m$ corresponds to a pair of d_k and $d_{k'}$ variables with $k, k' = 1, 2, \ldots, K$ and the total number of groups is $M = K^2$. In the simulation, we consider 3 cases with $d = 20, 60$, and 100. For each case, the number of partition K and a value $d_1 = d_2 = \ldots = d_K = r$ are set to be $(K, r) = (2, 10), (3, 20)$, and $(4, 25)$. After a precision matrix Λ is derived, we generate $n = 5d$ independent samples from a normal distribution $\mathcal{N}(\mathbf{0}_d, \Lambda^{-1})$.

3.5.2 Baseline Methods

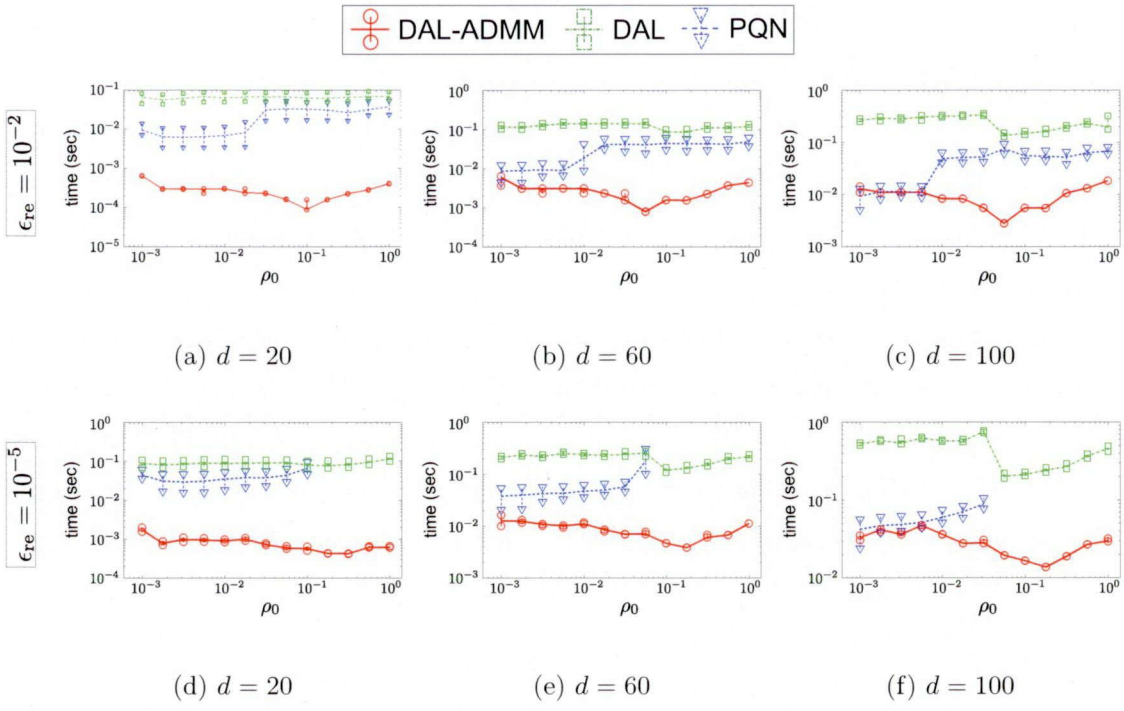
In the simulation, we adopt a PQN algorithm (Schmidt et al., 2009), an algorithm constructed for group SICS, to contrast with DAL-ADMM. We also introduce DAL to compare with DAL-ADMM aiming to observe the advantage of an ADMM relaxation. DAL-ADMM, DAL, and PQN are implemented using MATLAB and C. We used a DAL package[4] and implemented a DAL procedure for group SICS. We have also modified a PQN package[5] and used for our simulation. In the simulation, we set $\rho = d\rho_0$ where ρ_0 varies in 13 different values ranging from 10^{-3} to 10^0 in a logarithmic-scale.

3.5.3 Result

We randomly generated datasets 1000 times for each setting and compared the running time of DAL-ADMM, DAL, and PQN. The results are summarized in Figure 3.1. In the figure, we plot median times that each method achieves a relative error $(g(\Lambda^{(k)}) - g(\Lambda^*))/g(\Lambda^*)$ under tolerance parameters $\epsilon_{\text{gap}} = 10^{-2}$ and 10^{-5} where $g(\Lambda) \equiv -\ell(\Lambda; \hat{\Sigma}) + \sum_{m=1}^{M} \rho_m \|\Lambda_{\mathcal{G}_m}\|_{p_m}$. The vertical bars extend from the 25% to the 75% quantiles of the running time. Note that PQN did not achieve a relative error under $\epsilon_{\text{gap}} = 10^{-5}$ for larger ρ_0 and thus omitted from the graph.

[4]available at `http://www.ibis.t.u-tokyo.ac.jp/ryotat/dal/`

[5]available at `http://www.di.ens.fr/~mschmidt/Software/PQN.html`



(a) $d = 20$ (b) $d = 60$ (c) $d = 100$

(d) $d = 20$ (e) $d = 60$ (f) $d = 100$

Figure 3.1: Median running time until achieving a relative error under $\epsilon_{\mathrm{gap}} = 10^{-2}$ and 10^{-5} with vertical bars extending from the 25% to the 75% quantiles.

In all experimental settings, we observe that DAL-ADMM outperformes other twos. In particular, we can see the gradual decrease of the DAL-ADMM running time for larger ρ_0. We conjecture this property is what original DAL has as an efficient optimization method for sparse regularization problems, and is also inherited to DAL-ADMM. Through simulations, we observe that the inner optimization process in DAL gets a practical bottleneck and it is resolved by an ADMM relaxation resulting in a dramatic improvement. A solution sequence in PQN approaches to the optimal solution in a relatively small running time. However, at some point, this speed drastically decreases and the improvement of the solution seems to be bounded afterward.

3.6 Conclusion

In this chapter, we treated a generalized SICS problem (3.1) where the state-of-the-art method for SICS (1.11) is no longer applicable. Our proposed DAL-ADMM

algorithm is based on DAL and we relaxed it by introducing an ADMM approximation. In synthetic experiments with a group regularization term, we observed that this relaxation dramatically improved the running time against naively applying DAL. A comparison of DAL-ADMM against PQN also showed favorable results that DAL-ADMM is faster and hence works well for larger ρ where PQN tends to require a longer running time.

Several future works have been indicated. The optimal choice of an algorithm parameter β remains as an open problem. In our algorithm, we used a heuristic update which works practically well but does not have any theoretical guarantees. An introduction of a skipping technique proposed by Scheinberg et al. (2010) would be a promising extension of DAL-ADMM to further improve its performance.

Chapter 4

Learning a Common Substructure of Multiple Graphical Gaussian Models

4.1 Introduction

In this chapter, we address the problem of finding an invariant pattern from a set of GGMs obtained from multiple datasets. We provide a technique for finding constant interactions, or dependencies, among variables across several different conditions. An illustrative example of this problem is an engineering system where system errors are observed as dependency anomalies in sensor values (Idé, Lozano, Abe, & Liu, 2009), which are usually caused by a fault in a subsystem. The invariance, which in this example is the remaining healthy subsystems, is captured by a steady dependency over the multiple datasets sampled before and after the error onset. Hence, we can use such information as a clue for finding erroneous subsystems.

Unlike the basic GGM learning problem (1.11) which focuses on recovering the topology of a dependency structure from a single dataset, our objective is to decompose the resulting GGMs from several datasets into common and individual substructures. Hence, this common pattern is the invariance we aim to detect. See Figure 4.1 for an illustration. There are some prior studies on learning a set of GGMs from multiple datasets. Varoquaux, Gramfort, Poline, and Thirion (2010) and Honorio and Samaras (2010) imported the idea of Group-Lasso (M. Yuan & Lin, 2006; Bach, 2008) and Multitask-Lasso (Turlach et al., 2005; Liu, Palatucci, & Zhang, 2009) and extended the framework of a single GGM setting. In both

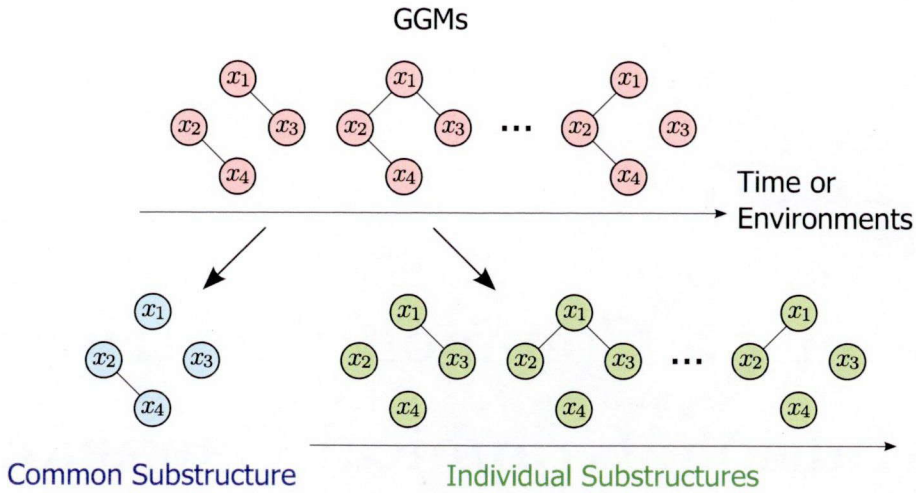


Figure 4.1: A decomposition of multiple GGMs into common and individual substructures. The main objective of this chapter is to propose a methodology that achieves this.

cases, the problem is formulated under the assumption that all precision matrices share the same zero patterns. J. Guo et al. (2011) considered a method to avoid this additional assumption, although the problem then loses convexity. Though these approaches achieved some success in improving the estimation accuracy of graphical models, this does not necessarily mean that they are suitable for finding commonness across datasets as we will see in the simulation. In the context of common substructure detection, Zhang and Wang (2010) proposed using a Fused-Lasso (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005) type of technique to find an invariant pattern between two datasets. As a general framework for K datasets situations, Chiquet, Grandvalet, and Ambroise (2011) considered imposing sign coherence on the resulting structures. In the opposite context where the target is dynamics rather than invariance, Zhou, Lafferty, and Wasserman (2010) proposed using weighted statistics to trace the evolution of a GGM. We note that there are also several related studies in the binary Markov random field literatures (F. Guo, Hanneke, Fu, & Xing, 2007; Ahmed & Xing, 2009). They also use ℓ_1-regularization (Wainwright, Ravikumar, & Lafferty, 2007) and Fused-Lasso type techniques (Ahmed & Xing, 2009) for recovering temporal dependency structures, which are technically quite close to the methodologies developed on GGM.

The contribution of this chapter is twofold. First, we introduce the novel Common Substructure Learning (CSSL) framework that is applicable to a general case of K datasets. Second, we show that the target problem can be solved by the DAL-ADMM algorithm introduced in Chapter 3. In the proposed algorithm, the inner problems for each iterative update are simple and can be solved efficiently which results in fast computation. We confirm the validity of the CSSL approach through simulations on synthetic datasets and on an anomaly localization task in real-world data.

The remainder of the chapter is organized as follows. In Section 4.2, we briefly review some existing GGM learning techniques. In Section 4.3, we present the proposed framework and its theoretical properties. The optimization algorithm based on DAL-ADMM is introduced in Section 4.4. The validity of the proposed method is presented through synthetic experiments in Section 4.5. In Section 4.6, we apply the proposed method to an anomaly localization task on a real world data. Finally, we conclude the chapter in Section 4.7.

4.2 Structure Learning of Graphical Gaussian Model

In this section, we review some prior extension of SICS problem (1.11) into multiple datasets situations (Varoquaux et al., 2010; Honorio & Samaras, 2010; Zhang & Wang, 2010).

4.2.1 Learning a Set of GGMs with Same Topological Patterns

The ordinary SICS problem (1.11) aims to learn one GGM from a single dataset. The extension of this framework to multiple datasets has been studied by Varoquaux et al. (2010) and Honorio and Samaras (2010). The task is to estimate K precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$ from K datasets where the sample covariance matrices for each dataset are $\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_K$. The objective of this multitask extension is

to improve the estimation accuracy of each GGM by incorporating the similarity among datasets. In the framework of the above studies, GGMs from each dataset are assumed to have the same topological patterns, that is, the same edge connection structures while the edge weights might be different among GGMs. They both introduced the following $\ell_{1,p}$-norm of a set of K precision matrices $\{\Lambda_k\}_{k=1}^K$:

$$\|\Lambda\|_{1,p} \equiv \sum_{i,j=1}^{d} \left(\sum_{k=1}^{K} |\Lambda_{k,ij}|^p \right)^{\frac{1}{p}},$$

as a regularization term analogous to the Group-Lasso (M. Yuan & Lin, 2006; Bach, 2008) and Multitask-Lasso (Turlach et al., 2005; Liu et al., 2009) with $p \in [1, \infty]$. Varoquaux et al. (2010) has considered the case $p = 2$ while Honorio and Samaras (2010) used $p = \infty$. These two choices are commonly adopted in many scenarios owing to the computational efficiency. The entire estimation problem is defined as

$$\max_{\{\Lambda_k; \Lambda_k \in \mathcal{S}^+\}_{k=1}^K} \sum_{k=1}^{K} \eta_k \ell(\Lambda_k; \hat{\Sigma}_k) - \rho \|\Lambda\|_{1,p}, \tag{4.1}$$

with non-negative weights $\eta_1, \eta_2, \ldots, \eta_K$. Without loss of generality, we can limit ourselves to the normalized case $\sum_{k=1}^K \eta_k = 1$ since the unnormalized version is just a scaled objective function for some constant. The typical choice of parameters would be $\eta_k = N_k / \sum_{k=1}^K N_k$ where N_k is the number of data points in the kth dataset. We refer to the problem (4.1) as Multitask Sparse Inverse Covariance Selection (MSICS) in the remainder of the chapter.

Note that the MSICS problem (4.1) involves the ordinary SICS (1.11) as a special case when $p = 1$ where the $\ell_{1,1}$-regularization term completely decouples into K individual ℓ_1-norms. In the extended case for $p > 1$, the regularization term enforces the joint structure $\tilde{\Lambda}_{ij} = \left(\sum_{k=1}^K |\Lambda_{k,ij}|^p \right)^{\frac{1}{p}}$ to be sparse, with $\tilde{\Lambda}_{ij} = 0$ indicating that the corresponding (i, j)th entries are zeros across all K precision matrices.

4.2.2 Learning Structural Changes between Two GGMs

Although taking advantage of situations with multiple datasets using the above-mentioned techniques is useful for improving the estimation performances of the resulting GGMs, it only imposes joint zero patterns and does not indicate anything

about the commonness of the non-zero entries. It is therefore not helpful when comparing GGMs representing similar models where we expect that there may exist some common edges whose weights are close to each other. Zhang and Wang (2010) considered the two datasets case and constructed an algorithm using a Fused-Lasso type regularization (Tibshirani et al., 2005) to round these similar values to be exactly the same allowing only significantly different edges between two GGMs to be extracted. Their approach follows the ideas of Meinshausen and Bühlmann (2006) by connecting the update procedure (1.10) for two datasets X_1 and X_2 through a new regularization term for the variation between two parameters $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1$,

$$\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} \sum_{i=1}^{2} \left\{ \frac{1}{2} \left\| X_{i,j} - X_{i,\setminus j}\boldsymbol{\theta}_i \right\|_2^2 + \rho \left\| \boldsymbol{\theta}_i \right\|_1 \right\} + \gamma \left\| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\|_1 , \tag{4.2}$$

where $\gamma \geq 0$ is a regularization parameter for the variation. The new term enforces the variation of some elements in two parameters to shrink to zeros. They also provided a coordinate descent-based optimization procedure for the above problem.

4.3 Learning Common Patterns in Multiple GGMs

The above-mentioned work by Zhang and Wang (2010) adopted the idea of the Fused-Lasso type technique using the specific formulation of the two datasets situation. In this section, we address our new framework, a Common Substructure Learning (CSSL), for finding invariant patterns in multiple dependency structures that is applicable to the general case of K datasets.

4.3.1 Common Substructure Learning Problem

We first formalize what invariance we are aiming to detect in multiple dependency structures. To begin with, we assume that the number of variables in each dataset is the same, so that they are all d-dimensional. Also, the identity of each variable are the same. For instance, a realization of x_1 is always a value from the same sensor while its behavior may change across datasets. We then define a common substructure of multiple GGMs as follows.

Definition 1 (Common Substructure of Multiple GGMs). *Let* Λ_k $(k = 1, 2, \ldots, K)$ *be a precision matrix corresponding to the kth GGM. The common substructure of the GGMs is then expressed by an adjacency matrix* $\Theta \in \mathbb{R}^{d \times d}$ *defined as*

$$\Theta_{ij} = \begin{cases} \Lambda_{1,ij} & \text{if } \Lambda_{1,ij} = \Lambda_{2,ij} = \ldots = \Lambda_{K,ij}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

Note that this is a natural extension of the invariance notion adopted in the prior work by Zhang and Wang (2010) for the case of two datasets. With an ordinary sparsity assumption for GGMs, this definition leads the precision matrices to have sparseness and commonness simultaneously. More specifically,

- Sparseness: $\Lambda_{k,ij} = 0$ for some $k = 1, 2, \ldots, K$ and $i, j = 1, 2, \ldots, d$,
- Commonness: $\Lambda_{1,ij} = \Lambda_{2,ij} = \ldots = \Lambda_{K,ij}$ for some $i, j = 1, 2, \ldots, d$.

Under the above commonness, the basic idea of our framework is to parametrize each precision matrix Λ_k using two components, a common substructure Θ and an individual substructure $\Omega_k \in \mathbb{R}^{d \times d}$:

$$\Lambda_k = \Theta + \Omega_k. \tag{4.4}$$

Here, each individual substructure matrix Ω_k is composed of non-zero entries that are not common across the K precision matrices.

In the formulation (4.2), some entries in the two precision matrices are shrunk to the same value owing to the effect of the term $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1$. In the proposed parameterization, such commonness corresponds to the case when some entries of the individual substructures are simultaneously zero, that is, $\Omega_{1,ij} = \Omega_{2,ij} = \ldots = \Omega_{K,ij} = 0$. Hence, the non-zero common value is expressed by a common substructure matrix Θ. These facts motivate us to regularize the individual substructures through the group regularization $\|\Omega\|_{1,p}$. On the other hand, we expect a common substructure Θ to be sparse so that we can interpret it easily. To that end, we adopt an ordinary ℓ_1-regularization $\|\Theta\|_1$ and the overall problem is summarized as

follows:

$$\begin{aligned} &\max_{\Theta,\{\Omega_k\}_{k=1}^K} \sum_{k=1}^K \eta_k \ell(\Theta+\Omega_k;\hat{\Sigma}_k) - \rho\,\|\Theta\|_1 - \gamma\,\|\Omega\|_{1,p}\,, \\ &\text{s.t. } \Theta+\Omega_k \in \mathcal{S}^+ \quad (k=1,2,\ldots,K), \end{aligned} \tag{4.5}$$

with regularization parameters $\rho \geq 0$ and $\gamma \geq 0$. Since $-\ell(\Theta+\Omega_k;\hat{\Sigma}_k)$, $\|\Theta\|_1$ and $\|\Omega\|_{1,p}$ are all convex, the entire formulation is again a convex optimization problem. We refer to this problem as Common Substructure Learning (CSSL). Note that in the above formulation, we have slightly relaxed the condition of commonness to allow Θ_{ij} and $\Omega_{k,ij}$ to simultaneously become non-zeros which is contrary to Definition (4.3). We correct this point by applying the criterion (4.3) to the resulting precision matrices $\hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_K$ in the post processing stage to extract only truly common entries.

Here, we list two important properties of the CSSL problem (4.5), a dual problem and the bound on eigenvalues. We first present the dual problem, which is essential when we aim to solve the problem through DAL-ADMM.

Proposition 1 (Dual of CSSL). *The dual problem of CSSL (4.5) is*

$$\begin{aligned} &\min_{\{W_k;W_k\in\mathcal{S}^+\}_{k=1}^K} -\sum_{k=1}^K \eta_k \log\det W_k - d, \\ &\text{s.t. } \left|\sum_{k=1}^K \eta_k\left(W_{k,ij} - \hat{\Sigma}_{k,ij}\right)\right| \leq \rho \quad (i,j=1,2,\ldots,d), \\ &\left(\sum_{k=1}^K \eta_k^q |W_{k,ij} - \hat{\Sigma}_{k,ij}|^q\right)^{\frac{1}{q}} \leq \gamma \quad (i,j=1,2,\ldots,d), \end{aligned} \tag{4.6}$$

where q denotes a parameter satisfying $p^{-1}+q^{-1}=1$. The resulting matrices of the dual problem W_k^* are related to the optimal precision matrices Λ_k^* through the inverse, $\Lambda_k^* = W_k^{*-1}$.

In both the primal and the dual formulations (4.5), (4.6), we enforced the positive definiteness constraints, $\Lambda_k = \Theta+\Omega_k \in \mathcal{S}^+$ and $W_k \in \mathcal{S}^+$ so that the matrices are valid precision or covariance matrices. Here, we show that they can be tightened according to the next theorem.

Theorem 9 (Bounds on Eigenvalues). *The optimal precision matrices for the CSSL (4.5)* $\Lambda_1^*, \Lambda_2^*, \ldots, \Lambda_K^*$ *with* $0 < \rho < K^{\frac{1}{p}}\gamma < \infty$ *have bounded eigenvalues* $\lambda_k^{\min} I_d \preceq \Lambda_k^* \preceq \lambda_k^{\max} I_d$, *where the bounding parameters* $\lambda_k^{\min}$ *and* $\lambda_k^{\max}$ *are given by*

$$\lambda_k^{\min} = \frac{\eta_k}{\eta_k \left\| \hat{\Sigma}_k \right\|_{\mathrm{S}} + d\gamma},$$
$$\lambda_k^{\max} = \frac{K^{\frac{1}{p}} d^2}{\rho}.$$

Here, $\|*\|_{\mathrm{S}}$ *denotes a spectral norm of a matrix and is given by* $\|A\|_{\mathrm{S}} \equiv \max_i \sigma_i(A)$ *where* $\sigma_i(A)$ *is an ith singular value of* A.

Using this result, we can replace the constraint $\Lambda_k \in \mathcal{S}^+$ with the tighter one $\Lambda_k \in \tilde{\mathcal{S}}_k^+ = \{A \in \mathbb{R}^{d\times d}; A \succeq \lambda_k^{\min} I_d\}$ and similarly $W_k \in \{A \in \mathbb{R}^{d\times d}; A \succeq \lambda_k^{\max -1} I_d\}$. Note that this update is practically important when constructing an optimization algorithm. Since the new constraint set $\tilde{\mathcal{S}}_k^+$ is closed, we can project points out of the constraint set onto the boundary, which is not possible for the open set $\mathcal{S}^+$.

4.3.2 Interpretations of CSSL

The proposed CSSL problem (4.5) can be interpreted as a generalization of the ordinary SICS problem (1.11) and its multitask extension MSICS (4.1). In the case that $\gamma \to \infty$, the solution to CSSL is $\Omega_1 = \Omega_2 = \ldots = \Omega_K = 0_{d\times d}$, which means that all precision matrices are equal and are represented by a single matrix Θ. We can obtain such a Θ by solving the SICS problem (1.11) with $\hat{\Sigma} = \sum_{k=1}^K \eta_k \hat{\Sigma}_k$. On the other hand, if $\rho \geq K^{\frac{1}{p}}\gamma$, the common substructure Θ becomes zero. This fact follows from the relationship between ℓ_p-norms:

$$\begin{aligned}\gamma \|\Theta + \Omega_k\|_{1,p} &\leq K^{\frac{1}{p}}\gamma \|\Theta\|_1 + \gamma \|\Omega\|_{1,p} \\ &\leq \rho \|\Theta\|_1 + \gamma \|\Omega\|_{1,p}.\end{aligned}$$

Suppose that the common substructure is non-zero, that is, $\Theta \neq 0_{d\times d}$. The above inequality then implies that the update $\Omega_k \leftarrow \Theta + \Omega_k$ and $\Theta \leftarrow 0_{d\times d}$ improves the objective function value (4.5) without changing the resulting precision matrix $\Lambda_k = \Theta + \Omega_k$, and thus the solution must be $\Theta = 0_{d\times d}$. In this situation, the CSSL

problem (4.5) coincides with MSICS (4.1). For proper parameters $\rho < K^{\frac{1}{p}}\gamma < \infty$, the CSSL problem (4.5) is intermediate between those two problems.

The CSSL problem can also be interpreted from a distributional perspective. From the relationship between the Lagrangian expression and the constrained optimization problem, the CSSL problem (4.5) is equivalent to solving a set of K maximum likelihood estimation problems under the additional constraints

$$\|\Theta\|_1 \leq \delta, \quad \|\Omega\|_{1,p} \leq \delta', \tag{4.7}$$

for some properly chosen positive constants δ and δ'. Moreover, we have

$$\begin{aligned}\max_{k,k'=1,2,\ldots,K} \|\Omega_k - \Omega_{k'}\|_1 &\leq \max_{k,k'=1,2,\ldots,K} \sum_{i,j=1}^{d} (|\Omega_{k,ij}| + |\Omega_{k',ij}|) \\ &\leq 2\,\|\Omega\|_{1,\infty} \leq 2\,\|\Omega\|_{1,p}\,,\end{aligned}$$

where the second inequality comes from the fact that exchanging the order of $\max_{k,k'=1,2,\ldots,K}$ and $\sum_{i,j=1}^{d}$ produces the upper bound. The last inequality is an ordinary relationship between ℓ_p-norms. These relations and the fact that $\Lambda_k - \Lambda_{k'} = \Omega_k - \Omega_{k'}$ lead to the bound

$$\max_{k,k'=1,2,\ldots,K} \|\Lambda_k - \Lambda_{k'}\|_1 \leq 2\delta'.$$

Hence, from the result of Honorio (2011, Lemma 23) and general matrix norm rules, the left-hand side of this inequality can be interpreted as the upper bound of the KL divergence between two distributions $p_k(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}_d, \Lambda_k^{-1})$ and $p_{k'}(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}_d, \Lambda_{k'}^{-1})$. With these properties, we can interpret the second constraint in (4.7) as a constraint on the similarity among distributions:

$$\max_{k,k'=1,2,\ldots,K} D_{\mathrm{KL}}[p_k(\boldsymbol{x})||p_{k'}(\boldsymbol{x})] \leq 2\delta' \max_{k=1,2,\ldots,K} \left\|\Lambda_k^{-1}\right\|_{\mathrm{S}},$$

where $D_{\mathrm{KL}}[p_k(\boldsymbol{x})||p_{k'}(\boldsymbol{x})]$ denotes a KL divergence between two distributions $p_k(\boldsymbol{x})$ and $p_{k'}(\boldsymbol{x})$. From Theorem 9, the optimal parameters $\Lambda_1^*, \Lambda_2^*, \ldots, \Lambda_K^*$ have bounded spectral norms for a finite γ, and thus this upper bound on the KL divergence is always valid. Moreover, we can further extend this bound into the extreme case $\gamma \to \infty$ and $\delta' \to 0$. As we have discussed before, this is the case $\Omega_1 = \Omega_2 =$

$\ldots = \Omega_K = 0_{d\times d}$ and the problem is equivalent to solving a single SICS problem for Θ with $\hat{\Sigma} = \sum_{k=1}^{K} \eta_k \hat{\Sigma}_k$. Hence, from Banerjee et al. (2008, Theorem 1), we can see that the resulting precision matrices still have finite eigenvalues for $\rho > 0$, and the right hand side of the above inequality goes to zero. This implies that the resulting distributions represented by precision matrices derived from CSSL (4.5) have to be similar to one another at some level and they can be even identical in the extreme case. Note that MSICS (4.1) is a special case of CSSL when $\Theta = 0_{d\times d}$ and thus the same upper bound holds, although there is the significant distinction that the parameter δ' in MSICS (4.1) also affects the sparsity of the resulting precision matrices while CSSL (4.5) can control the sparsity through the other hyper-parameter ρ.

4.3.3 Connection to Additive Sparsity Models

In this section, we discuss some connections of the CSSL problem (4.5) to *additive sparsity models* (Jalali, Ravikumar, Sanghavi, & Ruan, 2010; Chandrasekaran, Parrilo, & Willsky, 2012; Agarwal, Negahban, & Wainwright, 2011; Candès, Li, Ma, & Wright, 2011; Obozinski, Jacob, & Vert, 2011). In general additive sparsity models, the objective parameter we want to estimate is modeled as the sum of two components, as in (4.4). Hence, these two parameters are estimated using sparsity inducing norms such as an ℓ_1-norm and a trace-norm. In this sense, CSSL can be interpreted as a specific example of additive sparsity models where we use the combination of an ℓ_1-regularization and a group regularization.

Here, we point out two close works from Jalali et al. (2010) and Chandrasekaran et al. (2012). The former considers the multitask least squares regression problem under the combination of ℓ_1 and group regularizations. Their basic idea is quite close to ours in that some regression parameters can be close to each other across datasets. They also proved the advantage of combining two regularizations over using only one both theoretically and numerically. The latter study is on GGMs but with different sparsity assumptions from ours. They show that the additive sparsity model naturally appears in GGM when there are latent variables. In such a situation, the first component in the additive sparsity model corresponds to the precision

matrix between observed variables while the latter component is an interaction between latent variables. This insight is also helpful for interpreting our model (4.4), that is, a common interaction among observed variables is contaminated by the effect of latent variables whose distributions are different across datasets.

4.4 Optimization via DAL-ADMM

In this section, we present the optimization algorithm for solving the CSSL problem (4.6). In a prior study, Tomioka et al. (2011) have shown that DAL is preferable for the case when the primal loss is badly conditioned. See Tomioka et al. (2011, Table 3) and the discussion therein. This is actually the case we are faced with, as summarized in the next theorem.

Theorem 10. *The Hessian matrix of the CSSL primal loss function* $\sum_{k=1}^{K} \eta_k \ell(\Theta + \Omega_k; \hat{\Sigma}_k)$ *is rank-deficient while the Hessian matrix of the CSSL dual loss function* $-\sum_{k=1}^{K} \eta_k \log\det W_k$ *is always full rank for* $0 < \rho < K^{\frac{1}{p}}\gamma < \infty$.

This fact motivates us to solve the problem with DAL, which we modify into DAL-ADMM in Chapter 3 for a computational consideration.

4.4.1 Optimization via DAL-ADMM

To begin with, we rewrite the CSSL dual problem (4.6) in the following equivalent form:

$$\begin{aligned} \min_{\{W_k, Y_k; W_k \in \mathcal{S}^+\}_{k=1}^K} & -\sum_{k=1}^{K} \eta_k \log\det W_k, \\ \text{s.t. } & \eta_k W_k + Y_k - \eta_k \hat{\Sigma}_k = 0 \ \ (k = 1, 2, \ldots, K), \\ & \left| \sum_{k=1}^{K} Y_{k,ij} \right| \leq \rho \ \ (i, j = 1, 2, \ldots, d), \\ & \left(\sum_{k=1}^{K} |Y_{k,ij}|^q \right)^{\frac{1}{q}} \leq \gamma \ \ (i, j = 1, 2, \ldots, d). \end{aligned} \tag{4.8}$$

Although this formulation is slightly different from (3.1), it is still in the scope of DAL-ADMM. Based on the above expression, we define the following Augmented

Lagrangian function:

$$\mathcal{L}_\beta(W,Y,Z) = -\sum_{k=1}^{K} \eta_k \log\det W_k + \varphi^*_{\rho,\gamma}(Y) + \frac{\beta}{2}\left\| HW + Y + \frac{1}{\beta}Z - HS \right\|_{\mathrm{F}}^2, \tag{4.9}$$

where β is a nonnegative parameter, S, W, Y and Z are the concatenated matrices given by $S = \begin{bmatrix}\hat{\Sigma}_1 & \hat{\Sigma}_2 & \dots & \hat{\Sigma}_K\end{bmatrix}^\top$, $W = \begin{bmatrix}W_1 & W_2 & \dots & W_K\end{bmatrix}^\top$, $Y = \begin{bmatrix}Y_1 & Y_2 & \dots & Y_K\end{bmatrix}^\top$, and $Z = \begin{bmatrix}Z_1 & Z_2 & \dots & Z_K\end{bmatrix}^\top$, and H is the matrix constructed as $H = \mathrm{diag}(\eta_1, \eta_2, \dots, \eta_K) \otimes I_d$ where $\otimes$ denotes the Kronecker product. A function $\varphi^*_{\rho,\gamma}(Y)$ is a convex conjugate of a regularization term $\rho\|\Theta\|_1 + \gamma\|\Omega\|_{1,p}$ and is given by

$$\varphi^*_{\rho,\gamma}(Y) = \delta_\rho(Y) + \tilde{\delta}^q_\gamma(Y),$$

$$\delta_\rho(Y) = \begin{cases} 0 & \text{if } \left|\sum_{k=1}^K Y_{k,ij}\right| \le \rho \ \ (i,j = 1,2,\dots,d), \\ \infty & \text{otherwise,} \end{cases}$$

$$\tilde{\delta}^q_\gamma(Y) = \begin{cases} 0 & \text{if } \left(\sum_{k=1}^K |Y_{k,ij}|^q\right)^{\frac{1}{q}} \le \gamma \ \ (i,j = 1,2,\dots,d), \\ \infty & \text{otherwise.} \end{cases}$$

In the Augmented Lagrangian function (4.9), the optimal precision matrix Λ^*_k is represented by the optimal Lagrange multipliers Z^*_k. This can be verified through a simple calculation. We set the derivative of the unaugmented Lagrangian $\mathcal{L}_0(W,Y,Z)$ over W_k equal to zero and find that

$$W^{*-1}_k = Z^*_k,$$

which implies that $\Lambda^*_k = Z^*_k$ from Proposition 1. This follows from the fact that the solution to (4.8) must be the saddle point of the unaugmented Lagrangian function $\mathcal{L}_0(W,Y,Z)$.

We solve problem (4.8) using ADMM by iteratively applying the following three

steps until convergence:

$$\begin{cases} W^{(t+1)} \in \underset{\{W_k; W_k \in \mathcal{S}^+\}_{k=1}^K}{\operatorname{argmin}} \mathcal{L}_\beta(W, Y^{(t)}, Z^{(t)}), \\ Y^{(t+1)} \in \underset{Y}{\operatorname{argmin}} \mathcal{L}_\beta(W^{(t+1)}, Y, Z^{(t)}), \\ Z^{(t+1)} = Z^{(t)} + \beta\big(HW^{(t+1)} + Y^{(t+1)} - HS\big). \end{cases}$$

Hence, using ADMM, convergence of the Lagrange multiplier Z to the optimal parameter Z^* is guaranteed as the number of iterations tends to infinity (Boyd et al., 2011, Section 3.2). Therefore, we can find the optimal precision matrices $\Lambda_1^*, \Lambda_2^*, \ldots, \Lambda_K^*$ using DAL-ADMM. In the following two subsections, we give the update procedures of W and Y.

4.4.2 Inner Optimization Problem: Update of W

The update of W can be factorized into K independent problems where each problem defines an update of W_k:

$$\min_{W_k \in \mathcal{S}^+} -\eta_k \log\det W_k + \frac{\beta}{2}\left\|\eta_k W_k + Y_k^{(t)} + \frac{1}{\beta}Z_k^{(t)} - \eta_k \hat{\Sigma}_k\right\|_{\mathrm{F}}^2.$$

By setting the derivative over W_k equal to zero, we obtain

$$W_k - \left(-\frac{1}{\eta_k}Y_k^{(t)} - \frac{1}{\beta\eta_k}Z_k^{(t)} + \hat{\Sigma}_k\right) - \frac{1}{\beta\eta_k}W_k^{-1} = 0_{d\times d}.$$

Now, we write the eigenvalue decomposition of the second term as

$$-\frac{1}{\eta_k}Y_k^{(t)} - \frac{1}{\beta\eta_k}Z_k^{(t)} + \hat{\Sigma}_k = U\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)U^\top.$$

The above matrix equation then has a solution of the form

$$W_k = U\mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_d)U^\top.$$

The equation on each eigenvalue is

$$\tilde{\sigma}_i - \sigma_i - \frac{1}{\beta\eta_k}\tilde{\sigma}_i^{-1} = 0,$$

for $i = 1, 2, \ldots, d$, which has the analytic solution

$$\tilde{\sigma}_i = \frac{1}{2}\left(\sigma_i + \sqrt{\sigma_i^2 + \frac{4}{\beta\eta_k}}\right).$$

Note that the positive definiteness of W_k is automatically fulfilled since $\tilde{\sigma}_i > 0$ for $\beta > 0$.

4.4.3 Inner Optimization Problem: Update of Y

The update of Y is formulated as

$$\min_{Y} \delta_\rho(Y) + \tilde{\delta}^q_\gamma(Y) + \frac{\beta}{2}\left\| HW^{(t+1)} + Y + \frac{1}{\beta}Z^{(t)} - HS \right\|^2_{\mathrm{F}},$$

or equivalently, the Euclidean projection of a point $Y_0 = -HW^{(t+1)} - Z^{(t)}/\beta + HS$:

$$Y = \mathrm{proj}(Y_0, \mathcal{A}) \equiv \operatorname*{argmin}_{V \in \mathcal{A}} \frac{1}{2}\|V - Y_0\|^2_{\mathrm{F}},$$

where the constraint set is given by $\mathcal{A} = \left\{ Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_K \end{bmatrix}^\top ; \left|\sum_{k=1}^K Y_{k,ij}\right| \leq \rho \ , \left(\sum_{k=1}^K |Y_{k,ij}|^q\right)^{\frac{1}{q}} \leq \gamma \quad (i,j = 1,2,\dots,d) \right\}$. Note that in the current case, as we discussed in Section 3.4.1, the projection function $\mathrm{proj}(Y_0, \mathcal{A})$ corresponds to computing the proximity operator. We can further decompose this problem into $\mathcal{O}(d^2)$ problems over $\boldsymbol{y} = (Y_{1,ij}, Y_{2,ij}, \dots, Y_{K,ij})^\top$ for each (i,j)th entry. Hence, each problem is

$$\boldsymbol{y} = \mathrm{proj}(\boldsymbol{y}_0, \mathcal{C}), \tag{4.10}$$

where $\boldsymbol{y}_0$ is a K-dimensional vector with the kth component equal to $-\eta_k W^{(t+1)}_{k,ij} - Z^{(t)}_{k,ij}/\beta + \eta_k \hat{\Sigma}_{k,ij}$ and the constraint set is $\mathcal{C} = \left\{ \boldsymbol{u} \in \mathbb{R}^K ; |\mathbf{1}_K^\top \boldsymbol{u}| \leq \rho, \ \|\boldsymbol{u}\|_q \leq \gamma \right\}$ with $\mathbf{1}_K$ denoting a vector of all ones.

For any $q \in [1, \infty]$, problem (4.10) has a trivial solution $\boldsymbol{y} = \boldsymbol{y}_0$ if $\boldsymbol{y}_0 \in \mathcal{C}$. In the remaining cases, that is, $|\mathbf{1}_K^\top \boldsymbol{y}_0| > \rho$ or $\|\boldsymbol{y}_0\|_q > \gamma$, the solution is on the boundary of the constraint set $\partial\mathcal{C} = \left\{ \boldsymbol{u}; |\mathbf{1}_K^\top \boldsymbol{u}| = \rho, \ \|\boldsymbol{u}\|_q \leq \gamma \right\} \cap \left\{ \boldsymbol{u}; |\mathbf{1}_K^\top \boldsymbol{u}| \leq \rho, \ \|\boldsymbol{u}\|_q = \gamma \right\}$ owing to the convexity of the objective function. Thus, the problem can be reduced to a search on the boundary. However, even though the constraint set $\mathcal{C}$ is convex, it is an intersection of two sets and the shape of the boundary $\partial\mathcal{C}$ is rather complicated. Therefore, we do not search on the boundary $\partial\mathcal{C}$ directly, but solve a set of simpler problems instead. The basic approach is to classify the boundary into three parts, namely

$$\begin{aligned}
\partial\mathcal{C}_1 &= \left\{ \boldsymbol{u}; |\mathbf{1}_K^\top \boldsymbol{u}| = \rho, \ \|\boldsymbol{u}\|_q \neq \gamma \right\}, \\
\partial\mathcal{C}_2 &= \left\{ \boldsymbol{u}; |\mathbf{1}_K^\top \boldsymbol{u}| \neq \rho, \ \|\boldsymbol{u}\|_q = \gamma \right\}, \\
\partial\mathcal{C}_3 &= \left\{ \boldsymbol{u}; |\mathbf{1}_K^\top \boldsymbol{u}| = \rho, \ \|\boldsymbol{u}\|_q = \gamma \right\}.
\end{aligned}$$

The problems we solve here are modified versions of (4.10), replacing the constraint with $\boldsymbol{y} \in \partial\mathcal{C}_m$ for each $m \in \{1, 2, 3\}$:

$$\boldsymbol{y} = \mathrm{proj}(\boldsymbol{y}_0, \partial\mathcal{C}_m). \tag{4.11}$$

Note that $\partial\mathcal{C}_1$ and $\partial\mathcal{C}_2$ involve infeasible solutions to the problem (4.10). For example, a point $\boldsymbol{y}$ with $\|\boldsymbol{y}\|_q > \gamma$ is infeasible even if $\boldsymbol{y} \in \partial\mathcal{C}_1$, while these three regions covers the entire boundary of the constraint set $\partial\mathcal{C} \subset \cup_{m=1}^{3}\partial\mathcal{C}_m$. This guarantees that we can search on the entire boundary $\partial\mathcal{C}$ indirectly by searching on the sets $\partial\mathcal{C}_m$ $(m = 1, 2, 3)$ instead. Hence, if neither of the solutions to (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_1$ and $\boldsymbol{y} \in \partial\mathcal{C}_2$ are involved in $\mathcal{C}$, the solution to (4.10) is in $\partial\mathcal{C}_3$. We can take advantage of this property to construct an efficient solution procedure. We first solve problems (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_1$ and $\boldsymbol{y} \in \partial\mathcal{C}_2$, respectively, and if neither of solutions are in $\mathcal{C}$, we then solve (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_3$. In this chapter, we focus on the specific cases $q = 1, 2$, and ∞, since efficient solution procedures are available. In Table 4.1, we summarized the solutions to the problem (4.10). For further details, see Section 4.8.1.

4.4.4 Convergence Criteria

In Section 3.4.3, we introduced two gaps as stopping criteria, namely a primal-gap which measures how much the equality constraints in (4.8) is fulfilled,

$$\text{primal-gap} \equiv \left\|HW^{(t)} + Y^{(t)} - HS\right\|_{\mathrm{F}},$$

and a dual-gap which is a degree of the feasibility condition of the solution, defined as

$$\text{dual-gap} \equiv \beta \left\|H(Y^{(t+1)} - Y^{(t)})\right\|_{\mathrm{F}}.$$

Here, we consider another criterion called *duality-gap* which is the difference between the primal and the dual objective function values. Let $f(W)$ be the objective function in (4.6) and let $g(\Theta, \Omega)$ be the one in (4.5). The duality-gap at the tth iteration is then defined as

$$\text{duality-gap} \equiv f(\tilde{W}^{(t)}) - \max_{1 \le t' \le t} g(\tilde{\Theta}^{(t')}, \tilde{\Omega}^{(t')}),$$

Table 4.1: Solutions to problem (4.10) for $q = 1, 2$, and ∞: see corresponding sections for the detail. An operator $T_\gamma(*)$ in $\boldsymbol{y} \in \partial\mathcal{C}_2$ for $q = \infty$ is a thresholding for each $y_{0,i}$, that is, $y_i = \mathrm{sgn}(y_{0,i})\min(|y_{0,i}|, \gamma)$.

	$q = 1$	$q = 2$	$q = \infty$
$\boldsymbol{y}_0 \in \mathcal{C}$	$\boldsymbol{y} = \boldsymbol{y}_0$		
$\boldsymbol{y} \in \partial\mathcal{C}_1$	$\boldsymbol{y} = \boldsymbol{y}_0 - \frac{\mathbf{1}_K^\top \boldsymbol{y}_0 - \rho\,\mathrm{sgn}(\mathbf{1}_K^\top \boldsymbol{y}_0)}{K}\mathbf{1}_K$ (Section 4.8.1.1)		
$\boldsymbol{y} \in \partial\mathcal{C}_2$	Continuous Quadratic Knapsack Problem (Section 4.8.1.2)	$\boldsymbol{y} = \frac{\gamma}{\|\boldsymbol{y}_0\|_2}\boldsymbol{y}_0$ (Section 4.8.1.3)	$\boldsymbol{y} = T_\gamma(\boldsymbol{y}_0)$ (Section 4.8.1.4)
$\boldsymbol{y} \in \partial\mathcal{C}_3$	Continuous Quadratic Knapsack Problem (Section 4.8.1.5)	Analytic Solution (Section 4.8.1.6)	Continuous Quadratic Knapsack Problem (Section 4.8.1.7)

where $\tilde{W}^{(t)}$, $\tilde{\Theta}^{(t)}$, and $\tilde{\Omega}^{(t)}$ denote parameters estimated in the tth step after proper projections and transformations. We need these modifications of variables since the estimators in intermediate steps are not necessarily feasible. For example, $W^{(t)}$ does not need to satisfy the constraints in (4.6) since they are imposed only on a variable Y in the DAL-ADMM setting (4.8). The projected variable $\tilde{W}^{(t)}$ is $\tilde{W}^{(t)} = -H^{-1}\tilde{Y}^{(t)} + S$ where $\tilde{Y}^{(t)} = \mathrm{proj}(Y_0^{(t)}, \mathcal{A})$ and $Y_0^{(t)} = -H(W^{(t)} - S)$. The same goes for $\Lambda^{(t)} = Z^{(t)}$. An estimator $\Lambda_k^{(t)}$ is not necessarily positive definite, and thus we project them as $\tilde{\Lambda}_k^{(t)} = \mathrm{proj}(\Lambda_k^{(t)}, \tilde{\mathcal{S}}_k^+)$. This projection is conducted in the following manner. We first compute an eigenvalue decomposition $\Lambda_k^{(t)} = U\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)U^\top$. The projected matrix is then given by $\tilde{\Lambda}_k^{(t)} = U\mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_d)U^\top$ where each eigenvalue is given by $\tilde{\sigma}_i = \max(\sigma_i, \lambda_k^{\min})$. For computing the value of $g(\tilde{\Theta}^{(t)}, \tilde{\Omega}^{(t)})$, we need to further factorize $\tilde{\Lambda}^{(t)}$ into $\tilde{\Theta}^{(t)}$ and $\tilde{\Omega}^{(t)}$. This can be computed in an element-wise manner. Let $\theta = \tilde{\Theta}_{ij}^{(t)}$, $\Omega_{k,ij}^{(t)} = \tilde{\Lambda}_{k,ij}^{(t)} - \theta$, and $\boldsymbol{\lambda} = (\tilde{\Lambda}_{1,ij}^{(t)}, \tilde{\Lambda}_{2,ij}^{(t)}, \ldots, \tilde{\Lambda}_{K,ij}^{(t)})^\top$. The problem we need to solve is

then given by

$$\min_{\theta} \rho|\theta| + \gamma \left\|\boldsymbol{\lambda} - \theta \mathbf{1}_K\right\|_p .$$

For $p = 1$ and ∞, this function is piecewise linear with breakpoints given by $\{0, \lambda_1, \lambda_2, \ldots, \lambda_K\}$ and $\{0, (\min_k \lambda_k + \max_{k'} \lambda_{k'})/2\}$, respectively. Hence, the optimal θ is one of these breakpoints and can be found by searching over the candidates. For the case $p = 2$, there exists an analytic solution

$$\theta = \frac{1}{K}\left\{ \mathbf{1}_K^\top \boldsymbol{\lambda} - \operatorname{sgn}(\mathbf{1}_K^\top \boldsymbol{\lambda}) \sqrt{(\mathbf{1}_K^\top \boldsymbol{\lambda})^2 - K \frac{\gamma^2 (\mathbf{1}_K^\top \boldsymbol{\lambda})^2 - \rho^2 \left\|\boldsymbol{\lambda}\right\|_2^2}{\gamma^2 K - \rho^2}} \right\}.$$

In our simulations in Sections 4.5 and 4.6, we have evaluated both criteria. We set two threshold parameters ϵ_{pdgap} and ϵ_{gap}, and evaluate the conditions

$$\begin{aligned} \max(\text{primal-gap}, \text{dual-gap}) &\leq \epsilon_{\text{pdgap}}, \\ \text{duality-gap} &\leq \epsilon_{\text{gap}}, \end{aligned}$$

in each iteration. If one of two conditions is fulfilled, we regard the iteration as converged and output the result. In the simulations in Sections 4.5 and 4.6, we set $\epsilon_{\text{pdgap}} = 10^{-5}$ and $\epsilon_{\text{gap}} = 10^{-5} d$.

4.4.5 Computational Complexity

In this section, we summarize the computational complexity of the proposed algorithm. In the W update step, the computational cost is dominated by the eigenvalue decomposition of a $d \times d$ matrix, which requires $\mathcal{O}(d^3)$ operations, so that the overall complexity is $\mathcal{O}(Kd^3)$ for the update of K matrices. In the Y update step, we need a projection $\text{proj}(Y_0, \mathcal{A})$ which is divided into $\mathcal{O}(d^2)$ subproblems. For both $q = 1$ and $q = \infty$, the most computationally expensive procedure is solving the continuous quadratic knapsack problem which requires sorting $\mathcal{O}(K)$ elements and has complexity $\mathcal{O}(K \ln K)$[1]. In the case $q = 2$, the update is analytically available with $\mathcal{O}(K)$ complexity. The overall complexity for the Y update is thus $\mathcal{O}((K \ln K)d^2)$ for $q = 1, \infty$ and $\mathcal{O}(Kd^2)$ for $q = 2$. The complexity for the Z update is $\mathcal{O}(Kd^2)$.

[1] See Section 4.8.1.2, 4.8.1.5, and 4.8.1.7.

In the convergence check, we need to calculate the projection $\mathrm{proj}(\Lambda_k^{(t)}, \tilde{\mathcal{S}}_k^+)$ which has $\mathcal{O}(d^3)$ complexity or $\mathcal{O}(Kd^3)$ for K matrices. We also need the projection $\mathrm{proj}(Y_0^{(t)}, \mathcal{A})$ which is again $\mathcal{O}((K \ln K)d^2)$ for $q = 1, \infty$ and $\mathcal{O}(Kd^2)$ for $q = 2$. Summarizing the above results, we conclude that the computational complexity of one update in DAL-ADMM is $\mathcal{O}(Kd^3 + (K \ln K)d^2)$ for $q = 1, \infty$ and $\mathcal{O}(Kd^3)$ for $q = 2$. In many practical situations, the number of datasets K is in the tens, while the dimensionality of the data d can be a few hundred. In such cases, $\ln K \ll d$ holds, and the entire complexity is approximately $\mathcal{O}(Kd^3)$. We note that this is the least necessary complexity in general. For an unregularized setting, the solution Λ_k^* is a maximum likelihood estimator $\hat{\Sigma}_k^{-1}$, which requires $\mathcal{O}(d^3)$ complexity for a matrix inverse and $\mathcal{O}(Kd^3)$ for K matrices.

4.4.6 Heuristic Choice of Hyper–parameters

In the CSSL problem (4.5), the choice of hyper-parameters ρ and γ affects the resulting precision matrices. There are several approaches for choosing these, such as cross-validation (M. Yuan & Lin, 2007; J. Guo et al., 2011) and the Bayesian information criterion (J. Guo et al., 2011). Apart from selection techniques, the following result gives us some insight into ρ and γ, and is helpful in analyzing the data more intensively.

Proposition 2. *Let the bivariate common substructure Θ and individual substructures Ω_k be in the forms $\Theta = \begin{bmatrix} 0 & \theta \\ \theta & 0 \end{bmatrix}$ and $\Omega_k = \begin{bmatrix} u_k & \omega_k \\ \omega_k & v_k \end{bmatrix}$, and consider the following CSSL problem with regularizations only on off-diagonal entries:*

$$\begin{aligned} &\max_{\Theta, \{\Omega_k\}_{k=1}^K} \sum_{k=1}^K \eta_k \ell(\Theta + \Omega_k; \hat{\Sigma}_k) - 2\rho|\theta| - 2\gamma \left\|\boldsymbol{\omega}\right\|_p, \\ &\text{s.t. } \Theta + \Omega_k \in \mathcal{S}^+ \quad (k = 1, 2, \ldots, K), \end{aligned} \tag{4.12}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_K)^\top$. Then the off-diagonal entries of the resulting precision matrices $\theta, \boldsymbol{\omega}$ have the following property:

$$\max_{k=1,2,\ldots,K} |r_k| \leq \gamma \text{ and } \left|\sum_{k=1}^K \eta_k r_k\right| \leq \rho \Rightarrow \theta = 0, \ \boldsymbol{\omega} = \mathbf{0}_K,$$

where r_k is the off-diagonal entry of $\hat{\Sigma}_k$.

Although the result is specific to the bivariate case, we can use this as a guideline for choosing the hyper-parameters ρ and γ. It also shows that ρ and γ are not independent of each other, but rather they should change simultaneously proportional to $\max_{1\leq k\leq K}|r_k|$ and $\left|\sum_{k=1}^{K}\eta_k r_k\right|$, respectively. In particular, if each matrix $\hat{\Sigma}_k$ is multiplied by some positive constant c, the above condition indicates that ρ and γ also need to be multiplied by c. Such scale invariance is maintained only by a linear model between ρ and γ. Therefore, we construct the following heuristic based on this linear model.

1. Let $u_{ij} = \max_{k=1,2,\ldots,K}|\hat{\Sigma}_{k,ij}|$ and $v_{ij} = \left|\sum_{k=1}^{K}\eta_k\hat{\Sigma}_{k,ij}\right|$. We then assume that the linear relation

$$v_{ij} = u_{ij}s_1 + s_0,$$

   holds for all entries $i, j = 1, 2, \ldots, d$ for some $s_0, s_1 \in \mathbb{R}$.

2. Estimate s_0 and s_1 from the tuples $\{u_{ij}, v_{ij}\}_{i,j=1,2,\ldots,d}$ using a least squares regression.

3. Parameterize ρ and γ as $\rho = \max(\alpha s_1 + s_0, 0)$ and $\gamma = \alpha$ using a parameter α.

This procedure provides an efficient way of tuning ρ and γ simultaneously through a single parameter α.

4.5 Simulation

In this section, we investigate the performance of the proposed CSSL approach on finding common substructures among datasets through numerical simulations.

4.5.1 Generation of Synthetic Data

Here, we briefly summarize the data generation procedure for our simulation. For the synthetic data, we need K precision matrices with sparseness and commonness. We tackle this problem in a two-stage approach. We first generate a single

sparse precision matrix, and then add some non-zero entries to make K matrices where the additional patterns are independent of each other[2]. After K precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$ have been constructed, we generate K datasets from the corresponding Gaussian distributions $\mathcal{N}(\mathbf{0}_d, \Lambda_k^{-1})$ for $k = 1, 2, \ldots, K$.

4.5.2 Baseline Methods and Evaluation Measurements

In the simulation, we adopt SICS (1.11) and MSICS (4.1) as baseline methods to compare with CSSL. Since neither method is designed for finding a common substructure, we apply a heuristic to extract the substructure $\hat{\Theta}$ from the estimated precision matrices $\hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_K$. Note that, in SICS, each $\hat{\Lambda}_k$ is estimated by solving (1.11) individually while the set of matrices is estimated simultaneously in MSICS (4.1). The following is the heuristic criterion used:

$$\hat{\Theta}_{ij} = \begin{cases} \theta_{ij} & \text{if } \max_{k,k'=1,2,\ldots,K} |\hat{\Lambda}_{k,ij} - \hat{\Lambda}_{k',ij}| \leq \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where ϵ is some given threshold. Here, to avoid selecting zero edges as parts of a common substructure, we set $\hat{\theta}_{ij}$ to be zero if $\hat{\Lambda}_{1,ij} = \hat{\Lambda}_{2,ij} = \ldots = \hat{\Lambda}_{K,ij} = 0$ and one otherwise. In our simulation, we select the threshold ϵ from the resulting precision matrices. Specifically, we compute variations of estimators for each entry $\left\{\max_{k,k'=1,2,\ldots,K} |\hat{\Lambda}_{k,ij} - \hat{\Lambda}_{k',ij}|\right\}_{i,j=1,2,\ldots,d}$ and then set ϵ as the $100\epsilon_0\%$ quantile. This corresponds to considering the lower $100\epsilon_0\%$ varied entries as common.

In our simulation, we evaluate the common substructure detection performance through precision, recall, and the F-measure. While these values are defined based on the number of true positive, false positive, and false negative detections, we slightly modify these measurements. The reason is that finding common dependencies with higher amplitudes is much more important than finding very small dependencies which can be approximated as zero in practice. To that end, we adopt following weighted measurements, namely WTP (weighted true positive),

[2]See Section 4.8.2 for further details.

WFP (weighted false positive), and WFN (weighted false negative):

$$\begin{aligned}
\mathrm{WTP} &= \sum_{i<j} \tilde{J}_{\mathrm{c},ij} \tilde{J}_{\mathrm{p},ij} J_{\mathrm{c},ij} \max_{k=1,2,\ldots,K} |\Lambda_{k,ij}|, \\
\mathrm{WFP} &= \sum_{i<j} \tilde{J}_{\mathrm{c},ij} \tilde{J}_{\mathrm{p},ij} \Big(1 - J_{\mathrm{c},ij}\Big) \max_{k=1,2,\ldots,K} |\Lambda_{k,ij}|, \\
\mathrm{WFN} &= \sum_{i<j} \Big\{ \tilde{J}_{\mathrm{c},ij} \Big(1 - \tilde{J}_{\mathrm{p},ij}\Big) + \Big(1 - \tilde{J}_{\mathrm{c},ij}\Big) \Big\} J_{\mathrm{c},ij} \max_{k=1,2,\ldots,K} |\Lambda_{k,ij}|,
\end{aligned}$$

where $\tilde{J}_{\mathrm{c},ij}$, $\tilde{J}_{\mathrm{p},ij}$ and $J_{\mathrm{c},ij}$ are defined as

$$\begin{aligned}
\tilde{J}_{\mathrm{c},ij} &= I(\max_{k,k'=1,2,\ldots,K} |\hat{\Lambda}_{k,ij} - \hat{\Lambda}_{k',ij}| < \epsilon), \\
\tilde{J}_{\mathrm{p},ij} &= I(\max_{k=1,2,\ldots,K} |\hat{\Lambda}_{k,ij}| > 0), \\
J_{\mathrm{c},ij} &= I(\max_{k,k'=1,2,\ldots,K} |\Lambda_{k,ij} - \Lambda_{k',ij}| = 0).
\end{aligned}$$

Here, $I(P)$ is an indicator function that returns 1 for a true statement P and 0 otherwise. The modified measurements in the simulation are defined using these values as

$$\begin{aligned}
\text{Precision} &= \frac{\mathrm{WTP}}{\mathrm{WTP} + \mathrm{WFP}}, \\
\text{Recall} &= \frac{\mathrm{WTP}}{\mathrm{WTP} + \mathrm{WFN}}, \\
\text{F-measure} &= 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.
\end{aligned}$$

In the simulation, we also observe whether the zero pattern in the precision matrices is properly recovered through CSSL, SICS, and MSICS. We use the following F-measure for this evaluation, which we refer to the "F_0-measure" to distinguish it from the one above:

$$\begin{aligned}
\mathrm{F}_0\text{-measure} &= \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \\
\mathrm{TP} &= \sum_{k=1}^{K} \sum_{i<j} I(\Lambda_{k,ij} = 0) I(\hat{\Lambda}_{k,ij} = 0), \\
\mathrm{FP} &= \sum_{k=1}^{K} \sum_{i<j} I(\Lambda_{k,ij} \neq 0) I(\hat{\Lambda}_{k,ij} = 0), \\
\mathrm{FN} &= \sum_{k=1}^{K} \sum_{i<j} I(\Lambda_{k,ij} = 0) I(\hat{\Lambda}_{k,ij} \neq 0).
\end{aligned}$$

4.5.3 Result

We conduct simulations for three cases with data dimensionality $d = 25, 50$, and 100 where the number of datasets is fixed at $K = 5$. For each case, we generate precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$ to have 15% non-zero entries on average. In the simulation, we randomly generate datasets 100 times and apply CSSL, SICS, and MSICS using several different hyper-parameters, where in each run, we set the number of data points in each dataset to be $5d$. For CSSL, we use the heuristic with a parameter α varying from 10^{-2} to 10^{-0} over 41 values. We also evaluate results for $\rho = \alpha$ and $\gamma = \infty$ to see the effect of γ in an extreme case. As discussed in Section 4.3.2, this corresponds to solving a single SICS problem with $\hat{\Sigma} = \sum_{k=1}^{K} \eta_k \hat{\Sigma}_k$ and setting the result to $\hat{\Lambda}_1 = \hat{\Lambda}_2 = \ldots = \hat{\Lambda}_K = \hat{\Lambda}$. For SICS and MSICS, we set the value of ρ to be α. For each method, we adopt the resulting precision matrices with 15% non-zero entries among these 41 values of α. In SICS and MSICS, we also vary the thresholding parameter ϵ_0 among 0.5, 0.7, and 0.9.

We summarize the results in Table 4.2. From the table, we can see the clear advantage of CSSL with $p = 2$ and ∞ over the other methods. These two methods show higher F-measures, which are from their higher precision. This contrasts with SICS and MSICS, which achieve high recall but have relatively poor precision. This implies that structures detected by those methods involve not only true common substructure but also many false detections. This shows the drawback of estimated precision matrices derived through SICS and MSICS, that is, their estimators tend to be highly varied even for true common entries while this is not the case for CSSL. This phenomenon is especially significant in SICS, which can hardly find common substructures owing to its highly varied estimators. The results for MSICS under $p = \infty$ and $\epsilon_0 = 0.9$ are still better than the others, although $\epsilon_0 = 0.9$ means that 90% of estimated non-zero entries are considered common, which is too optimistic. Moreover, we can see that the improvement of the F-measure is achieved by the growth of recall by contrasting the results with $\epsilon_0 = 0.5$ and 0.9. This implies that variations on the true common substructure mostly happen in between 50% and 90% of the entire variations of the estimated precision matrices, which are highly varied and can hardly be considered common. Note that despite

Table 4.2: Simulation results for three cases ($d = 25, 50$, and 100) with $K = 5$ datasets evaluated by weighted precision, recall, F-measure, and F_0-measure. The measurements are averaged over 100 random realization of datasets. The numbers in brackets are standard deviations of each measurement. Each of the three rows in SICS and MSICS corresponds to results with $\epsilon_0 = 0.5, 0.7$, and 0.9 from the top. Top three results are highlighted in each measurement.

(a) $d = 25$

	CSSL ($p = 1$)	CSSL ($p = 2$)	CSSL ($p = \infty$)	CSSL ($\gamma = \infty$)	SICS	MSICS ($p = 2$)	MSICS ($p = \infty$)
Prec.	**.84 (.19)**	**.70 (.16)**	**.56 (.19)**	.48 (.20)	.14 (.14)	.38 (.21)	.54 (.23)
					.20 (.16)	.43 (.21)	.49 (.21)
					.33 (.16)	.41 (.19)	.45 (.19)
Rec.	.45 (.32)	.82 (.14)	**.84 (.12)**	**.86 (.11)**	.07 (.07)	.48 (.24)	.60 (.24)
					.23 (.18)	.74 (.19)	.74 (.19)
					.80 (.20)	.83 (.13)	**.86 (.11)**
F	.56 (.22)	**.75 (.14)**	**.66 (.17)**	**.60 (.19)**	.09 (.08)	.41 (.21)	.55 (.23)
					.21 (.16)	.53 (.21)	.58 (.20)
					.45 (.18)	.53 (.19)	.58 (.18)
F_0	.92 (.02)	.92 (.02)	.92 (.02)	.92 (.02)	.92 (.02)	.93 (.02)	.92 (.02)

(b) $d = 50$

	CSSL ($p = 1$)	CSSL ($p = 2$)	CSSL ($p = \infty$)	CSSL ($\gamma = \infty$)	SICS	MSICS ($p = 2$)	MSICS ($p = \infty$)
Prec.	**.87 (.11)**	**.69 (.14)**	.56 (.17)	.47 (.17)	.10 (.13)	.24 (.20)	**.58 (.19)**
					.13 (.14)	.37 (.20)	.52 (.19)
					.27 (.19)	.42 (.18)	.47 (.18)
Rec.	.41 (.20)	.83 (.11)	**.85 (.10)**	**.91 (.05)**	.04 (.04)	.18 (.19)	.60 (.19)
					.10 (.11)	.51 (.21)	.72 (.16)
					.50 (.22)	.81 (.12)	**.86 (.08)**
F	.53 (.20)	**.75 (.12)**	**.66 (.15)**	**.61 (.15)**	.05 (.06)	.20 (.19)	.58 (.19)
					.10 (.11)	.42 (.20)	.59 (.18)
					.34 (.20)	.54 (.17)	.60 (.16)
F_0	.90 (.03)	.90 (.02)	.89 (.02)	.89 (.03)	.89 (.03)	.90 (.02)	.90 (.03)

(c) $d = 100$

	CSSL ($p = 1$)	CSSL ($p = 2$)	CSSL ($p = \infty$)	CSSL ($\gamma = \infty$)	SICS	MSICS ($p = 2$)	MSICS ($p = \infty$)
Prec.	**.91 (.07)**	**.78 (.10)**	.64 (.14)	.53 (.15)	.09 (.11)	.17 (.14)	**.68 (.15)**
					.10 (.12)	.33 (.21)	.62 (.16)
					.22 (.17)	.46 (.18)	.55 (.16)
Rec.	.37 (.18)	.81 (.11)	**.83 (.11)**	**.95 (.02)**	.03 (.10)	.06 (.10)	.59 (.17)
					.06 (.10)	.25 (.21)	.67 (.15)
					.24 (.19)	.67 (.16)	**.82 (.09)**
F	.51 (.19)	**.79 (.10)**	**.72 (.12)**	**.67 (.12)**	.05 (.10)	.08 (.11)	.63 (.16)
					.07 (.10)	.28 (.21)	.64 (.15)
					.22 (.18)	.54 (.17)	.65 (.14)
F_0	.87 (.04)	.87 (.04)	.87 (.03)	.87 (.03)	.87 (.03)	.88 (.04)	.87 (.03)

the significant difference in the common entry detection performances, all methods achieve comparable zero pattern identification performances as shown by the F_0-measure. This shows that finding common entries is a different problem from the ordinary graphical model selection, and that only CSSL works well for both tasks.

We note that CSSL with $p = 1$ and $\gamma = \infty$ give two extreme results. In the former setting, the resulting precision matrices achieve higher precision with lower recall, while it is the opposite in the latter setting. The first result is caused by the difference between the term $\|\Omega\|_{1,p}$ with $p = 1$ and $p > 1$. For $p = 1$, $\|\Omega\|_{1,p}$ completely decouples into ordinary ℓ_1-regularizations and the resulting precision matrices do not necessarily have common zero entries in individual substructures. Intuitively speaking, the results for $p = 1$ have common zero entries only when it is strongly confident, which results in a very conservative performance compared with $p > 1$. On the other hand, if $\gamma = \infty$, the entire structures are considered to be common, which results in fewer false negatives and more false positives.

4.6 Application to Anomaly Localization

In this section, we apply CSSL to an anomaly localization problem. The task is to identify contributions of each variable to the difference between two datasets. *Cor-*

relation anomalies (Idé et al., 2009), or errors on dependencies between variables, are known to be difficult to detect using existing approaches, especially with noisy data. To overcome this problem, the use of sparse precision matrices was proposed by Idé et al. (2009) since the sparse approach reasonably suppresses the pseudo-correlation among variables caused by noise and improves the detection rate. Here, we propose using CSSL. There is a clear indication that the proposed method can further suppress the variation in the estimated matrices. In particular, we expect that dependency structures among healthy variables are estimated to be common, which reduces the risk that such variables are mis-detected and only anomalies are enhanced.

4.6.1 Anomaly Score

We adopt the measurement for correlation anomalies proposed by Idé et al. (2009). This score is based on the KL-divergence between two conditional distributions. Formally, let $\boldsymbol{x}$ be a Gaussian random variable which follows $\mathcal{N}(\mathbf{0}_d, \Lambda_1^{-1})$ before the error onset and $\mathcal{N}(\mathbf{0}_d, \Lambda_2^{-1})$ afterward. We measure the degree of anomaly on the ith variable x_i using the KL-divergence between conditional distributions from before and after the error, which are $p_1(x_i|\boldsymbol{x}_{\backslash i})$ and $p_2(x_i|\boldsymbol{x}_{\backslash i})$, respectively, where $\boldsymbol{x}_{\backslash i}$ is the remaining $d-1$ variables except for x_i. To compute the score, we first divide the precision matrix Λ_1 and its inverse W_1 into a $(d-1)\times(d-1)$ dimensional matrix, a $d-1$ dimensional vector, and a scalar,

$$\Lambda_1 = \begin{bmatrix} L_1 & \boldsymbol{l}_1 \\ \boldsymbol{l}_1^\top & \lambda_1 \end{bmatrix}, \quad W_1 \equiv \Lambda_1^{-1} = \begin{bmatrix} V_1 & \boldsymbol{v}_1 \\ \boldsymbol{v}_1^\top & \sigma_1 \end{bmatrix},$$

where we have rotated the rows and columns of Λ_1 and W_1 simultaneously so that their original ith rows and columns are located at the last rows and columns of the matrices. The matrix Λ_2 and its inverse W_2 are also divided in a same manner. The score is then given by

$$\begin{aligned} d_i^{12} &= \int D_{\mathrm{KL}}\left[p_1(x_i|\boldsymbol{x}_{\backslash i})||p_2(x_i|\boldsymbol{x}_{\backslash i})\right] p_1(\boldsymbol{x}_{\backslash i})\, d\boldsymbol{x}_{\backslash i} \\ &= \boldsymbol{v}_1^\top(\boldsymbol{l}_1 - \boldsymbol{l}_2) + \frac{1}{2}\left(\frac{\boldsymbol{l}_2^\top V_1 \boldsymbol{l}_2}{\lambda_2} - \frac{\boldsymbol{l}_1^\top V_1 \boldsymbol{l}_1}{\lambda_1}\right) + \frac{1}{2}\left\{\ln\frac{\lambda_1}{\lambda_2} + \sigma_1(\lambda_1 - \lambda_2)\right\}. \end{aligned}$$

Here, the KL-divergence is averaged over the remaining $d-1$ variables $\boldsymbol{x}_{\backslash i}$. Since the KL-divergence is not symmetric and $d_i^{12} \neq d_i^{21}$ holds in general, the resulting anomaly score a_j is defined as their maximum:

$$a_i = \max(d_i^{12}, d_i^{21}). \tag{4.13}$$

4.6.2 Simulation Setting

We evaluate the anomaly localization performance using sensor error data (Idé et al., 2009). The dataset comprised 42 sensor values collected from a real car in 79 normal states and 20 faulty states. The fault is caused by mis-wiring of the 24th and 25th sensors, resulting in correlation anomalies. Since sample covariances are rank-deficient in some datasets, we added 10^{-3} on their diagonal to avoid singularities.

For simulation, we randomly sample K_{n} datasets from the normal states and K_{f} datasets from the faulty states, and then estimate sparse precision matrices using six methods, CSSL with $p = 1, 2$, and ∞, SICS (1.11), and MSICS (4.1) with $p = 2$ and ∞. For CSSL, we adopt the heuristic and set $\rho = \max(\alpha s_1 + s_0, 0)$ and $\gamma = \alpha$ for a given α, and for SICS and MSICS, we set $\rho = \alpha$. We test each method for 11 different values of α ranging from $10^{-1.5}$ to $10^{-0.5}$. The weight parameters η_k in CSSL and MSICS are set to be $\eta_k = 1/2K_{\mathrm{n}}$ for normal datasets and $\eta_k = 1/2K_{\mathrm{f}}$ for faulty datasets to balance the effects from the two states. Since the anomaly score is designed only for a pair of datasets, we calculate anomaly scores for each of $K_{\mathrm{n}} \times K_{\mathrm{f}}$ pairs of datasets.

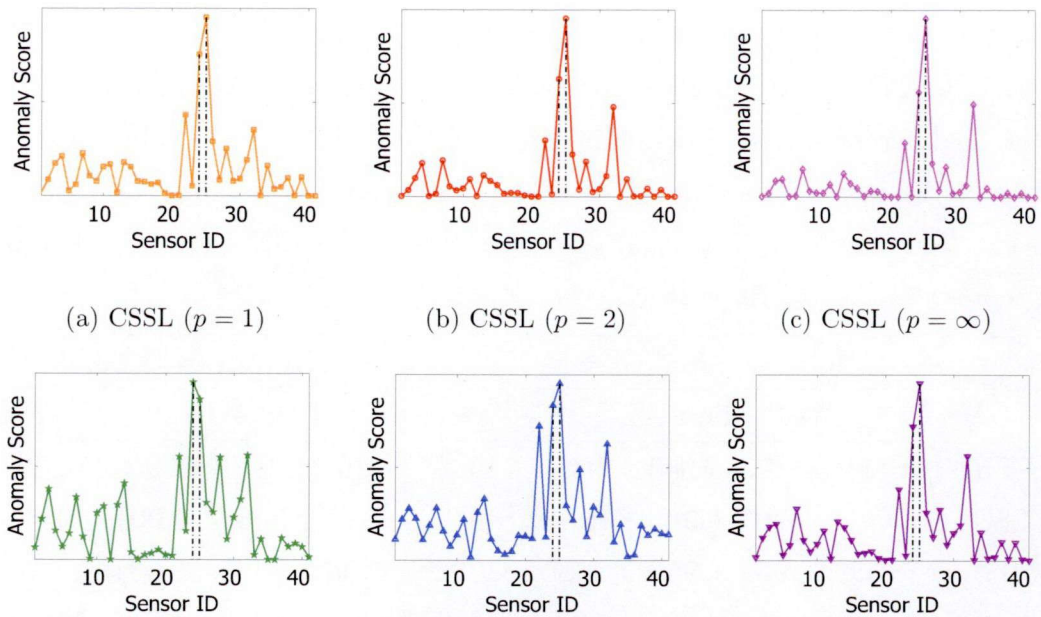
4.6.3 Result

We repeat the above procedure 100 times for 4 different settings, $[K_{\mathrm{n}}, K_{\mathrm{f}}] = [4, 1]$, $[12, 3]$, $[20, 5]$, and $[40, 10]$. For each run, we evaluate the localization performance of each method by drawing an receiver operating characteristic (ROC) curve and measuring the area under the curve (AUC). In Table 4.3, we summarize the best median results for each method and setting. The table shows that CSSL with $p = 2, \infty$ and MSICS with $p = \infty$ achieve better localization performances than the others. In particular, CSSL with $p = 2$ and ∞ achieve AUC $= 1$ as their

Table 4.3: Anomaly localization results under 4 different settings, $[K_n, K_f] = [4, 1], [12, 3], [20, 5]$, and $[40, 10]$. For each method, we compute precision matrices for 11 different values of α ranging from $10^{-1.5}$ to $10^{-0.5}$. The table shows the median of the best AUCs among these 11 results over 100 random realizations of datasets. The numbers in brackets are the 25% and the 75% quantiles. The bold font represents the top three results.

	$[K_n, K_f] = [4, 1]$		$[K_n, K_f] = [12, 3]$	
	best AUC	α	best AUC	α
CSSL ($p = 1$)	.975 (.950 / .987)	$10^{-0.9}$	.975 (.950 / 1.00)	$10^{-0.9}$
CSSL ($p = 2$)	**.987 (.963 / 1.00)**	$10^{-0.9}$	**.987 (.963 / 1.00)**	$10^{-0.9}$
CSSL ($p = \infty$)	**.987 (.963 / 1.00)**	$10^{-0.9}$	**1.00 (.987 / 1.00)**	$10^{-0.9}$
SICS	.975 (.938 / .987)	$10^{-0.5}$	.975 (.938 / .987)	$10^{-0.5}$
MSICS ($p = 2$)	.975 (.950 / .987)	$10^{-0.8}$	.975 (.950 / .987)	$10^{-0.7}$
MSICS ($p = \infty$)	**.987 (.963 / 1.00)**	$10^{-1.1}$	**.987 (.975 / 1.00)**	$10^{-1.2}$
	$[K_n, K_f] = [20, 5]$		$[K_n, K_f] = [40, 10]$	
	best AUC	α	best AUC	α
CSSL ($p = 1$)	.975 (.950 / 1.00)	$10^{-0.9}$	.975 (.963 / 1.00)	$10^{-0.9}$
CSSL ($p = 2$)	**1.00 (.975 / 1.00)**	$10^{-0.8}$	**.987 (.963 / 1.00)**	$10^{-0.8}$
CSSL ($p = \infty$)	**1.00 (.987 / 1.00)**	$10^{-0.9}$	**1.00 (.987 / 1.00)**	$10^{-0.9}$
SICS	.975 (.950 / .987)	$10^{-0.5}$	.975 (.950 / .987)	$10^{-0.5}$
MSICS ($p = 2$)	.975 (.950 / .987)	$10^{-1.0}$	.975 (.950 / .987)	$10^{-1.0}$
MSICS ($p = \infty$)	**.987 (.975 / 1.00)**	$10^{-1.1}$	**.987 (.975 / 1.00)**	$10^{-0.9}$

median performance in some cases. This means that they detect faulty sensors perfectly for more than half of the simulation. To see further differences, we plot the median anomaly scores derived from each method for $[K_n, K_f] = [20, 5]$ in Figure 4.2. From these graphs, we observe a clear distinction between successful methods and others on the significance of healthy sensors. The 22nd and the 28th sensors are relatively highly enhanced in SICS and MSICS with $p = 2$, but are not in CSSL and MSICS with $p = \infty$. We conjecture that this is the major cause of performance differences. Interestingly, not only the 22nd and the 28th sensors but



(a) CSSL ($p = 1$) (b) CSSL ($p = 2$) (c) CSSL ($p = \infty$)

(d) SICS (e) MSICS ($p = 2$) (f) MSICS ($p = \infty$)

Figure 4.2: Median anomaly scores for each method under $[K_{\mathrm{n}}, K_{\mathrm{f}}] = [20, 5]$ with best AUCs. Each plot is normalized so that the maximum is the same. Dotted lines denote true faulty sensors.

most of other healthy sensors also have the same tendencies. That is, CSSL and MSICS with $p = \infty$ reasonably suppress their significance while keeping erroneous sensors enhanced. Moreover, although the differences are subtle, we can see that CSSL with $p = 2$ and ∞ more successfully suppress the significance of sensors 1 to 21 and 33 to 42 than does MSICS with $p = \infty$. Thus, as we expected in the beginning, CSSL reduces the nuisance effects and highlights only variables with correlation anomalies. The remaining peaks at some healthy variables are caused by the effect of the two faulty sensors since their effects may propagate to other healthy yet highly related sensors.

4.7 Conclusion

In this chapter, we formulated the CSSL problem for multiple GGMs. We further showed that the problem can be solved using DAL-ADMM with each updating

step computed efficiently. Numerical results on synthetic datasets indicate the clear advantage of the CSSL approach, in that it can achieve high precision and recall at the same time, which existing GGM structure learning methods can not achieve. We also applied the proposed CSSL technique to the anomaly localization task in sensor error data. Through the simulation, we observed that CSSL could efficiently suppress nuisance effects among variables in noisy sensors and successfully enhanced target faulty sensors.

Several future research topics have been indicated, including analyzing the asymptotic property of the CSSL problem (4.5) and extending the current formulation to the Adaptive-Lasso (Zou, 2006; Fan, Feng, & Wu, 2009) type one to guarantee the *oracle property* (Zou, 2006) of the estimator. Applying the notion of commonness to more general dependency models, such as those with non-linear relations and commonness based on higher-order moment statistics, is also important.

4.8 Appendix

4.8.1 Solutions to (4.10) for $q = 1, 2$, and ∞

Here, we provide detailed derivations of Table 4.1.

4.8.1.1 The solution is in $\partial\mathcal{C}_1$

Problem (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_1$ is formulated as follows:

$$\min_{\boldsymbol{y}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2, \quad \text{s.t.} \quad |\mathbf{1}_K^\top \boldsymbol{y}| = \rho. \tag{4.14}$$

Note that we ignored the constraint $\|\boldsymbol{y}\|_q \neq \gamma$ because it holds for general $\boldsymbol{y}_0$ and γ with probability one. Hence, our interest is whether the solution to (4.14) satisfies $\|\boldsymbol{y}\|_q \leq \gamma$ or not. The additional constraint is not important in this respect.

The problem (4.14) has two possible cases as its solution, $\mathbf{1}_K^\top \boldsymbol{y} = \rho$ and $\mathbf{1}_K^\top \boldsymbol{y} = -\rho$. For each case, we can solve the problem using a method of Lagrange multipliers:

$$\min_{\boldsymbol{y}} \max_{\mu} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2 + \mu(\mathbf{1}_K^\top \boldsymbol{y} - \zeta),$$

where $\zeta \in \{\rho, -\rho\}$. By setting the derivative over $\boldsymbol{y}$ equal to zero, we obtain $\boldsymbol{y} = \boldsymbol{y}_0 - \mu \mathbf{1}_K$. Moreover, by substituting this result into the above, we derive the optimal μ as $\mu = \left(\mathbf{1}_K^\top \boldsymbol{y}_0 - \zeta\right)/K$, and the resulting objective function value is $\left(\mathbf{1}_K^\top \boldsymbol{y}_0 - \zeta\right)^2/2K$. The constraint $\zeta = \rho$ or $\zeta = -\rho$ is chosen so that this objective function value is minimized. Obviously, $\zeta = \rho$ is optimal for the case when $\mathbf{1}_K^\top \boldsymbol{y}_0 \geq 0$, while $\zeta = -\rho$ for $\mathbf{1}_K^\top \boldsymbol{y}_0 < 0$. Thus, the overall solution to problem (4.14) is

$$\boldsymbol{y} = \boldsymbol{y}_0 - \frac{\mathbf{1}_K^\top \boldsymbol{y}_0 - \rho \operatorname{sgn}(\mathbf{1}_K^\top \boldsymbol{y}_0)}{K} \mathbf{1}_K .$$

4.8.1.2 The solution is in $\partial \mathcal{C}_2$ for $q = 1$

When the solution is in $\partial \mathcal{C}_2$, the problem is formulated as

$$\min_{\boldsymbol{y}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2, \quad \text{s.t.} \quad \|\boldsymbol{y}\|_q = \gamma. \tag{4.15}$$

Here, the shape of the constraint boundary changes according to the value of q. For general $q \in [1, \infty]$, there exists several algorithms to solve this problem (Boyd & Vandenberghe, 2004; Sra, 2011). In particular, for $q = 1, 2$, and ∞, we can derive solutions in very efficient manners.

For $q = 1$, Honorio and Samaras (2010) showed that the problem is equivalent to the following *continuous quadratic knapsack problem*:

$$\min_{\boldsymbol{z}} \sum_{k=1}^{K} \frac{1}{2} (z_k - |y_{0,k}|)^2, \quad \text{s.t.} \quad \boldsymbol{z} \geq 0, \ \mathbf{1}_K^\top \boldsymbol{z} = \gamma, \tag{4.16}$$

which relates to $\boldsymbol{y}$ by $y_k = \operatorname{sgn}(y_{0,k}) z_k$. Honorio and Samaras (2010) also provided a solution technique for this problem. From the KKT condition, the solution to (4.16) is $z_k(\nu) = \max(|y_{0,k}| - \nu, 0)$ for some constant ν. Moreover, the optimal ν satisfies $\mathbf{1}_K^\top \boldsymbol{z}(\nu) = \gamma$. Since $\mathbf{1}_K^\top \boldsymbol{z}(\nu)$ is a decreasing piecewise linear function with breakpoints $\{|y_{0,k}|\}_{k=1}^K$, we can find the minimum breakpoint ν_0 that satisfies $\mathbf{1}_K^\top \boldsymbol{z}(\nu_0) \leq \gamma$ by sorting the K breakpoints. The optimal ν is then given by

$$\nu = \frac{\sum_{k \in \mathcal{I}_0} |y_{0,k}| - \gamma}{|\mathcal{I}_0|},$$

where $\mathcal{I}_0 = \{k; |y_{0,k}| - \nu_0 \geq 0\}$. Note that the complexity of this algorithm is $\mathcal{O}(K \log K)$ since we conduct a sorting of K values[3].

[3]We can further reduce this to expected linear time complexity by introducing a randomized algorithm (Duchi, Shalev-Shwartz, Singer, & Chandra, 2008).

4.8.1.3 The solution is in $\partial\mathcal{C}_2$ for $q = 2$

We can derive the solution to the problem (4.15) for $q = 2$ analytically. We solve the problem using a method of Lagrange multipliers:

$$\min_{\boldsymbol{y}} \max_{\lambda} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2 + \frac{\lambda}{2}\left(\|\boldsymbol{y}\|_2^2 - \gamma^2\right).$$

By setting the derivative over $\boldsymbol{y}$ equal to zero, we derive $\boldsymbol{y} = \boldsymbol{y}_0/(1+\lambda)$. Moreover, from the constraint $\|\boldsymbol{y}\|_2 = \gamma$, the solution is

$$\boldsymbol{y} = \frac{\gamma}{\|\boldsymbol{y}_0\|_2}\boldsymbol{y}_0.$$

4.8.1.4 The solution is in $\partial\mathcal{C}_2$ for $q = \infty$

The solution of (4.15) for the case $q = \infty$ is much simpler. The problem is just a box-constrained least squares with a solution given by

$$y_k = \begin{cases} \gamma & \text{if } y_{0,k} > \gamma, \\ y_{0,k} & \text{if } -\gamma < y_{0,k} < \gamma, \\ -\gamma & \text{if } y_{0.k} < -\gamma, \end{cases}$$

which is equivalent to $y_k = \text{sgn}(y_{0,k})\min(|y_{0,k}|, \gamma)$.

4.8.1.5 The solution is in $\partial\mathcal{C}_3$ for $q = 1$

We provide the solution procedure for (4.11) when $\boldsymbol{y} \in \partial\mathcal{C}_3$ and $q = 1$ based on the next theorem.

Theorem 11. *Let $\tilde{\boldsymbol{y}}$ be the solution to problem (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_1$ and suppose it is infeasible in the original problem (4.10). Then the solution to (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_3$ has same signs with $\tilde{\boldsymbol{y}}$, that is, $\tilde{y}_k y_k \geq 0$ for $k = 1, 2, \ldots, K$.*

From this result, we can factorize the variable indices into two parts, $\mathcal{I}_+ = \{k; \tilde{y}_k \geq 0\}$ and $\mathcal{I}_- = \{k; \tilde{y}_k < 0\}$. Using this factorization, we can rewrite the

problem as

$$\min_{\boldsymbol{y}} \frac{1}{2}\sum_{k\in\mathcal{I}_+}(y_k - y_{0,k})^2 + \frac{1}{2}\sum_{k\in\mathcal{I}_-}(y_k - y_{0,k})^2,$$
$$\text{s.t.} \sum_{k\in\mathcal{I}_+} y_k + \sum_{k\in\mathcal{I}_-} y_k = \zeta,$$
$$\sum_{k\in\mathcal{I}_+} y_k - \sum_{k\in\mathcal{I}_-} y_k = \gamma,$$

where $\zeta \in \{\rho, -\rho\}$. This can be divided into two independent problems defined on two sets of variables $\{y_k^+; k \in \mathcal{I}_+\}$ and $\{y_k^-; k \in \mathcal{I}_-\}$, respectively, given by

$$\min_{\boldsymbol{y}^+} \frac{1}{2}\sum_{k\in\mathcal{I}_+}\left(y_k^+ - y_{0,k}\right)^2, \text{ s.t. } \boldsymbol{y}^+ \geq 0, \sum_{k\in\mathcal{I}_+} y_k^+ = \frac{\gamma+\zeta}{2},$$
$$\min_{\boldsymbol{y}^-} \frac{1}{2}\sum_{k\in\mathcal{I}_-}\left(y_k^- + y_{0,k}\right)^2, \text{ s.t. } \boldsymbol{y}^- \geq 0, \sum_{k\in\mathcal{I}_-} y_k^- = \frac{\gamma-\zeta}{2}.$$

The solutions to these problems relate to $\boldsymbol{y}$ in that $y_k = y_k^+$ for $k \in \mathcal{I}_+$ and $y_k = -y_k^-$ for $k \in \mathcal{I}_-$. These problems are continuous quadratic knapsack problems and the solution can be found by using the same algorithm as in problem (4.16). We derive the final solution by solving these problems for the two cases $\zeta = \rho$ and $\zeta = -\rho$, and choosing the one with the smaller objective function value in (4.11).

4.8.1.6 The solution is in $\partial\mathcal{C}_3$ for $q = 2$

We can derive the solution to the case $\boldsymbol{y} \in \partial\mathcal{C}_3$ and $q = 2$ analytically. We use a method of Lagrange multipliers:

$$\min_{\boldsymbol{y}} \max_{\mu,\lambda} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2 + \mu(\mathbf{1}_K^\top \boldsymbol{y} - \zeta) + \frac{\lambda}{2}(\|\boldsymbol{y}\|_2^2 - \gamma^2),$$

where $\zeta \in \{\rho, -\rho\}$. By setting the derivative over $\boldsymbol{y}$ equal to zero, we derive $\boldsymbol{y} = (\boldsymbol{y}_0 - \mu\mathbf{1}_K)/(1+\lambda)$. If $\rho = 0$, we have $\mu = \mathbf{1}_K^\top \boldsymbol{y}_0/K$ from the constraint $\mathbf{1}_K^\top \boldsymbol{y} = 0$. Hence, from $\|\boldsymbol{y}\|_2 = \gamma$, we obtain the optimal $\boldsymbol{y}$ as

$$\boldsymbol{y} = \frac{\gamma}{\|\boldsymbol{y}_0 - \mu\mathbf{1}_K\|_2}(\boldsymbol{y} - \mu\mathbf{1}_K).$$

For the case of $\rho > 0$, we have $1/(1+\lambda) = \zeta/(\mathbf{1}_K^\top \boldsymbol{y}_0 - K\mu)$ from the constraint $\mathbf{1}_K^\top \boldsymbol{y} = \zeta$. Hence, we have a quadratic equation in μ from the constraint $\|\boldsymbol{y}\|_2^2 = \gamma^2$:

$$\rho^2\|\boldsymbol{y}_0 - \mu\mathbf{1}_K\|_2^2 = \gamma^2(\mathbf{1}_K^\top \boldsymbol{y}_0 - K\mu)^2.$$

Solving this equation gives the optimal $\boldsymbol{y}$ as

$$\boldsymbol{y} = \frac{\zeta}{\mathbf{1}_K^\top \boldsymbol{y}_0 - K\mu}(\boldsymbol{y}_0 - \mu \mathbf{1}_K),$$

where

$$\mu = \frac{1}{K}(\mathbf{1}_K^\top \boldsymbol{y}_0 \pm \sqrt{\tau}),$$
$$\tau = (\mathbf{1}_K^\top \boldsymbol{y}_0)^2 - K\frac{\gamma^2(\mathbf{1}_K^\top \boldsymbol{y}_0)^2 - \rho^2 \|\boldsymbol{y}_0\|_2^2}{\gamma^2 K - \rho^2}.$$

By substituting this result into $\|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2$, we obtain

$$\|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2 = \frac{1}{K}(\zeta - \mathbf{1}_K^\top \boldsymbol{y}_0)^2 + \frac{K\|\boldsymbol{y}_0\|_2^2 - (\mathbf{1}_K^\top \boldsymbol{y}_0)^2}{K\tau}(\zeta \pm \sqrt{\tau})^2.$$

Since $K\|\boldsymbol{y}_0\|_2^2 - (\mathbf{1}_K^\top \boldsymbol{y}_0)^2 \geq 0$, the minimum of this value is achieved by choosing ζ and a sign in μ as $\zeta = \mathrm{sgn}(\mathbf{1}_K^\top \boldsymbol{y}_0)\rho$ and $-\mathrm{sgn}(\mathbf{1}_K^\top \boldsymbol{y}_0)$, respectively. Thus, the overall result is given by

$$\boldsymbol{y} = \mathrm{sgn}(\mathbf{1}_K^\top \boldsymbol{y}_0)\frac{\rho}{\mathbf{1}_K^\top \boldsymbol{y}_0 - K\mu}(\boldsymbol{y}_0 - \mu \mathbf{1}_K),$$
$$\mu = \frac{1}{K}(\mathbf{1}_K^\top \boldsymbol{y}_0 - \mathrm{sgn}(\mathbf{1}_K^\top \boldsymbol{y}_0)\sqrt{\tau}).$$

4.8.1.7 The solution is in $\partial\mathcal{C}_3$ for $q = \infty$

The solution for (4.11) with $\boldsymbol{y} \in \partial\mathcal{C}_3$ and $q = \infty$ has two possible cases, $\mathbf{1}_K^\top \boldsymbol{y} = \rho$ and $\mathbf{1}_K^\top \boldsymbol{y} = -\rho$, where for each case the problem is given by

$$\min_{\boldsymbol{y}} \sum_{k=1}^{K} \frac{1}{2}(y_k - y_{0,k})^2, \quad \text{s.t. } \mathbf{1}_K^\top \boldsymbol{y} = \zeta, \ -\gamma \mathbf{1}_K \leq \boldsymbol{y} \leq \gamma \mathbf{1}_K, \tag{4.17}$$

with $\zeta \in \{\rho, -\rho\}$. Here, the constraint $\|\boldsymbol{y}\|_\infty = \gamma$ is relaxed to $\|\boldsymbol{y}\|_\infty \leq \gamma$. However, if the solution to (4.17) satisfies $\|\boldsymbol{y}\|_\infty < \rho$, it has to be already found as a solution to (4.11) for $\boldsymbol{y} \in \partial\mathcal{C}_1$ and therefore this relaxation does not affect the overall procedure.

Since problem (4.17) is a variant of the continuous quadratic knapsack problem, we can take a strategy similar to (4.16). From the KKT condition, the solution to (4.17) is of the form $y_k(\nu) = \mathrm{sgn}(y_{0,k} - \nu)\min(|y_{0,k} - \nu|, \gamma)$ for some constant

ν. Moreover, the optimal ν satisfies $\mathbf{1}_K^\top \boldsymbol{y}(\nu) = \zeta$. Since $\mathbf{1}_K^\top \boldsymbol{y}(\nu)$ is a decreasing piecewise linear function with breakpoints $\{y_{0,k} - \gamma, y_{0,k} + \gamma\}_{k=1}^K$, we can find the minimum breakpoint ν_0 that satisfies $\mathbf{1}_K^\top \boldsymbol{y}(\nu_0) \leq \zeta$ by sorting the $2K$ breakpoints. The optimal ν is then given by

$$\nu = \begin{cases} \dfrac{\sum_{k \in \mathcal{I}_2} y_{0,k} + \gamma(|\mathcal{I}_1| - |\mathcal{I}_3|) - \zeta}{|\mathcal{I}_2|} & \text{if } \mathcal{I}_2 \neq \phi, \\ \nu_0 & \text{if } \mathcal{I}_2 = \phi, \end{cases}$$

where $\mathcal{I}_1 = \{k; y_{0,k} - \nu_0 \geq \gamma\}$, $\mathcal{I}_2 = \{k; -\gamma \leq y_{0,k} - \nu_0 < \gamma\}$, and $\mathcal{I}_3 = \{k; y_{0,k} - \nu_0 < -\gamma\}$.

4.8.2 Generation of Synthetic Precision Matrices

Here, we present the detailed procedure used to generate sparse precision matrices with a common substructure in Section 4.5. The procedure is composed of two sequential steps. We first generate a single precision matrix, which is the common substructure in the resulting K matrices. We then add some non-zero entries to get a matrix Λ_k. This additional pattern is chosen to be unique for each matrix so that the resultant matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$ satisfy the additive model assumption (4.4). In the following two subsections, we explain the above steps.

4.8.2.1 Generation of a Sparse Precision Matrix

In several previous studies, synthetic sparse precision matrices are generated in a naive manner, that is, just adding a properly scaled identity matrix to a sparse symmetric matrix so that the resulting matrix is sparse and positive definite (Banerjee et al., 2008; Wang, Sun, & Toh, 2009; Li & Toh, 2010). In our simulations, we take a different approach to generating a sparse precision matrix for compatibility with the next step.

Our approach is based on an eigenvalue decomposition $\Lambda = UDU^\top$, where D is a matrix with eigenvalues on its diagonal and U is an orthonormal matrix such that $U^\top U = UU^\top = I_d$. Here, we use the fact that Λ is sparse if U is sufficiently sparse and the problem can be reduced to generating a sparse orthonormal matrix U. This can be done easily by applying a Givens rotation (Golub & Van Loan,

1996) to an identity matrix I_d. Formally, we let $U^{(0)} = I_d$ and apply the following procedure repeatedly until the desired sparsity is achieved.

1. Randomly pick two indices i, j from $\{1, 2, \ldots, d\}$.

2. Randomly generate θ from a uniform distribution from 0 to 2π.

3. Update the $(i,i), (i,j), (j,i)$ and (j,j)th entries of $U^{(t)}$ as

$$\begin{bmatrix} U_{ii}^{(t+1)} & U_{ij}^{(t+1)} \\ U_{ji}^{(t+1)} & U_{jj}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} U_{ii}^{(t)} & U_{ij}^{(t)} \\ U_{ji}^{(t)} & U_{jj}^{(t)} \end{bmatrix}.$$

4. Keep the remaining entries $U_{i'j'}^{(t+1)} = U_{i'j'}^{(t)}$ for $(i', j') \notin \{(i,i), (i,j), (j,i), (j,j)\}$.

In our simulations, we generated each eigenvalue from a uniform distribution from 0 to 1.

4.8.2.2 Generation of Sparse Precision Matrices with a Common Substructure

Here, we turn to imposing commonness on the resulting precision matrices. To begin with, we generate small sparse precision matrices $\Psi_1, \Psi_2, \ldots, \Psi_a$ in the above-mentioned manner and construct a sparse block-diagonal precision matrix as $\Lambda_0 = \text{block-diag}(\Psi_1, \Psi_2, \ldots, \Psi_a)$. We then add some non-zero entries to Λ_0 and generate K precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_K$. At this stage, we keep the original non-zero entries Λ_0 unchanged so that they form a common substructure at the end. Note that the addition of non-zero entries can not be done randomly since this might destroy the positive definiteness of matrices.

We describe the procedure for the case $a = 2$. Let the eigenvalue decompositions of Ψ_1 and Ψ_2 be $\Psi_1 = U_1 D_1 U_1^\top$ and $\Psi_2 = U_2 D_2 U_2^\top$. Note that U_1 and U_2 are sparse since they are generated to be so. Now, let matrix Λ_k be of the form

$$\Lambda_k = \begin{bmatrix} \Psi_1 & \Phi_k \\ \Phi_k^\top & \Psi_2 \end{bmatrix}.$$

The objective is to generate a sparse non-zero matrix Φ_k while keeping the positive definiteness of Λ_k. This corresponds to keeping a determinant of Λ_k positive. Here, we choose Φ_k of the form

$$\Phi_k = \tilde{U}_1^b \Xi_k \tilde{U}_2^{b\top},$$

where Ξ_i is a $b \times b$ diagonal matrix and $\tilde{U}_1^b$ and $\tilde{U}_2^b$ are matrices composed of b columns in U_1 and U_2, respectively. Specifically, we let $U_1 = \begin{bmatrix} \boldsymbol{u}_{1,1} & \boldsymbol{u}_{1,2} & \dots & \boldsymbol{u}_{1,d_1} \end{bmatrix}$ and $U_2 = \begin{bmatrix} \boldsymbol{u}_{2,1} & \boldsymbol{u}_{2,2} & \dots & \boldsymbol{u}_{2,d_2} \end{bmatrix}$, where d_1 and d_2 denote the dimensionality of each matrix. Matrices $\tilde{U}_1^b$ and $\tilde{U}_2^b$ are then given by

$$\tilde{U}_1^b = \begin{bmatrix} \boldsymbol{u}_{1,\pi_{1,1}} & \boldsymbol{u}_{1,\pi_{1,2}} & \dots & \boldsymbol{u}_{1,\pi_{1,b}} \end{bmatrix},$$
$$\tilde{U}_2^b = \begin{bmatrix} \boldsymbol{u}_{2,\pi_{2,1}} & \boldsymbol{u}_{2,\pi_{2,2}} & \dots & \boldsymbol{u}_{2,\pi_{2,b}} \end{bmatrix},$$

for some index sets $\{\pi_{1,1}, \pi_{1,2}, \dots, \pi_{1,b}\} \subseteq \{1, 2, \dots, d_1\}$, $\{\pi_{2,1}, \pi_{2,2}, \dots, \pi_{2,b}\} \subseteq \{1, 2, \dots, d_2\}$. Then, from a general matrix property, we can express the determinant of Λ_k as

$$\begin{aligned} \det \Lambda_k &= \det\left(\Psi_1 - \Phi_k \Psi_2^{-1} \Phi_k^\top\right) \\ &= \det\left(D_1 - U_1^\top \Phi_k U_2 D_2^{-1} U_2^\top \Phi_k^\top U_1\right) \\ &= \prod_{i=1}^{b} \left(\sigma_{1,\pi_{1,i}} - \frac{\xi_{k,i}^2}{\sigma_{2,\pi_{2,i}}}\right), \end{aligned}$$

where $D_1 = \mathrm{diag}(\sigma_{1,1}, \sigma_{1,2}, \dots, \sigma_{1,d_1})$, $D_2 = \mathrm{diag}(\sigma_{2,1}, \sigma_{2,2}, \dots, \sigma_{2,d_2})$, and $\Xi_k = \mathrm{diag}(\xi_{k,1}, \xi_{k,2}, \dots, \xi_{k,b})$. Hence, the positive definiteness of Λ_k is guaranteed if $\xi_{k,i}^2 < \sigma_{1,\pi_{1,i}} \sigma_{2,\pi_{2,i}}$ is satisfied for $i = 1, 2, \dots, b$. Moreover, this inequality provides us a guideline on choosing index sets. Since we want non-zero entries of Φ_k to be larger, which can be achieved by larger $|\xi_{k,i}|$, we choose index sets so that $\sigma_{1,\pi_{1,i}} \sigma_{2,\pi_{2,i}}$ gets large. This corresponds to choosing leading eigenvalues and eigenvectors of Ψ_1 and Ψ_2. In our simulations, we pick $b = 2$ indices at random from those with eigenvalues in the top 1/3. We also generate $\xi_{k,i}$ as $\xi_{k,i} = \xi_{0,k,i} \sqrt{\sigma_{1,\pi_{1,i}} \sigma_{2,\pi_{2,i}}}$, where $\xi_{0,k,i} = \kappa \upsilon$ and κ takes a value 1 or -1 with equally likely, and υ follows a uniform distribution from 0.5 to 0.8.

For general $a > 2$ cases, we first construct a matrix $\Lambda_k^{(1)}$ from Ψ_1 and Ψ_2. We then iteratively apply the above procedure to generate $\Lambda_k^{(2)}$ from $\Lambda_k^{(1)}$ and Ψ_3, $\Lambda_k^{(3)}$

from $\Lambda_k^{(2)}$ and Ψ_4, until $\Lambda_k = \Lambda_k^{(a-1)}$ is derived. In the simulation in Section 4.5, we set the number of modules to be $a = 2$ for $d = 25$, $a = 3$ for $d = 50$, and $a = 4$ for $d = 100$.

4.8.3 Proofs of Theorems

4.8.3.1 Proof of Proposition 1

Let E and F_k be non-negative $d \times d$ matrices satisfying $-E_{ij} \leq \Theta_{ij} \leq E_{ij}$ and $-F_{k,ij} \leq \Omega_{k,ij} \leq F_{k,ij}$, respectively, for all $k = 1, 2, \ldots, K$ and $i, j = 1, 2, \ldots, d$. Then, using Lagrange multipliers Γ, Γ_0, and $\{\Delta_k, \Delta_{0,k}\}_{k=1}^K$, the CSSL problem (4.5) is expressed as

$$\begin{aligned}
\max_{\Theta, E, \{\Omega_k, F_k\}_{k=1}^K} \min_{\Gamma, \Gamma_0, \{\Delta_k, \Delta_{0,k}\}_{k=1}^K} & \sum_{k=1}^K \eta_k \Big\{ \log\det(\Theta + \Omega_k) - \mathrm{tr}\Big[\hat{\Sigma}_k(\Theta + \Omega_k)\Big] \Big\} \\
& - \sum_{i,j=1}^d \left\{ \rho E_{ij} + \gamma \left(\sum_{k=1}^K F_{k,ij}^p \right)^{\frac{1}{p}} \right\} \\
& - \mathrm{tr}\big[\Gamma\Theta\big] + \mathrm{tr}\big[\mathrm{abs}(\Gamma)E\big] + \mathrm{tr}\big[\Gamma_0 E\big] \\
& - \sum_{k=1}^K \big\{ \mathrm{tr}\big[\Delta_k \Omega_k\big] - \mathrm{tr}\big[\mathrm{abs}(\Delta_k) F_k\big] - \mathrm{tr}\big[\Delta_{0,k} F_k\big] \big\},
\end{aligned}$$

$$\text{s.t. } \Gamma_{0,ij} \geq 0, \ \Delta_{0,k,ij} \geq 0 \ \ (k = 1, 2, \ldots, K, \ i, j = 1, 2, \ldots, d).$$

By changing the order of maximization and minimization above, we derive the dual problem. Now, we optimize each variable Θ, E, Ω_k, and F_k by setting each derivative equal to zero:

$$\begin{aligned}
\sum_{k=1}^K \eta_k \Big\{ (\Theta + \Omega_k)^{-1} - \hat{\Sigma}_k \Big\} - \Gamma &= 0_{d \times d}, \\
-\rho \mathbf{1}_d \mathbf{1}_d^\top + \mathrm{abs}(\Gamma) + \Gamma_0 &= 0_{d \times d}, \\
\eta_k \Big\{ (\Theta + \Omega_k)^{-1} - \hat{\Sigma}_k \Big\} - \Delta_k &= 0_{d \times d} \ \ (k = 1, 2, \ldots, K), \\
-\gamma \left(\sum_{k=1}^K F_{k,ij}^p \right)^{\frac{1-p}{p}} F_{k,ij} + |\Delta_{k,ij}| + \Delta_{0,k,ij} &= 0 \ \ (k = 1, 2, \ldots, K, \ i, j = 1, 2, \ldots, d).
\end{aligned}$$

As the result of these equations, we obtain

$$\Delta_k = \eta_k \Big\{ (\Theta + \Omega_k)^{-1} - \hat{\Sigma}_k \Big\},$$
$$\left| \sum_{k=1}^{K} \Delta_{k,ij} \right| \leq \rho \ \ (i,j = 1,2,\ldots,d),$$
$$\left(\sum_{k=1}^{K} |\Delta_{k,ij}|^q \right)^{\frac{1}{q}} \leq \gamma \ \ (i,j = 1,2,\ldots,d),$$

and the dual problem is given by (4.6) where we set $W_k = (\Theta + \Omega_k)^{-1} = \Delta_k/\eta_k + \hat{\Sigma}_k$. □

4.8.3.2 Proof of Theorem 9

We first prove the lower-bound. Let $W_k = \Delta_k/\eta_k + \hat{\Sigma}_k$ in the dual problem (4.6). We then have $\left|\sum_{k=1}^{K} \Delta_{k,ij}\right| \leq \rho$ and $\left(\sum_{k=1}^{K} |\Delta_{k,ij}|^q\right)^{\frac{1}{q}} \leq \gamma$, and hence

$$\begin{aligned}
\left\| \frac{1}{\eta_k} \Delta_k + \hat{\Sigma}_k \right\|_{\mathrm{S}} &\leq \frac{1}{\eta_k} \left\| \Delta_k \right\|_{\mathrm{S}} + \left\| \hat{\Sigma}_k \right\|_{\mathrm{S}} \\
&\leq \frac{d}{\eta_k} \max_{i,j=1,2,\ldots,d} |\Delta_{k,ij}| + \left\| \hat{\Sigma}_k \right\|_{\mathrm{S}} \\
&\leq \frac{d}{\eta_k} \max_{k=1,2,\ldots,K} \max_{i,j=1,2,\ldots,d} |\Delta_{k,ij}| + \left\| \hat{\Sigma}_k \right\|_{\mathrm{S}} \\
&\leq \frac{d\gamma}{\eta_k} + \left\| \hat{\Sigma}_k \right\|_{\mathrm{S}},
\end{aligned}$$

where the last inequality comes from the general relationship between ℓ_p-norms $\max_{k=1,2,\ldots,K} |\Delta_{k,ij}| \leq \left(\sum_{k=1}^{K} |\Delta_{k,ij}|^q\right)^{\frac{1}{q}}$. Since $W_k^* = \Delta_k^*/\eta_k + \hat{\Sigma}_k = \Lambda_k^{*-1}$ holds at the optimum, we have the lower-bound.

We now turn to proving the upper-bound. From strong duality, the duality-gap is zero at the optimal solution to the primal and the dual problems (4.5) and (4.6), and therefore we have

$$\rho \left\| \Theta^* \right\|_1 + \gamma \left\| \Omega^* \right\|_{1,p} = d - \sum_{k=1}^{K} \eta_k \mathrm{tr}\Big[\hat{\Sigma}_k (\Theta^* + \Omega_k^*) \Big].$$

Moreover, from $0 < \rho < K^{\frac{1}{p}} \gamma < \infty$, $\mathrm{tr}\Big[\hat{\Sigma}_k (\Theta^* + \Omega_k^*) \Big] \geq 0$, and the general ℓ_p-norm

rule $\left(\sum_{k=1}^{K} |\Omega^*_{k,ij}|^p\right)^{\frac{1}{p}} \geq \max_{k=1,2,\ldots,K} |\Omega^*_{k,ij}|$,

$$\|\Theta^*\|_1 + K^{-\frac{1}{p}} \|\Omega^*\|_{1,\infty} \leq \frac{d}{\rho},$$

holds. Since $K^{\frac{1}{p}} \geq 1$ for $p \geq 1$, we obtain

$$\|\Theta^*\|_1 + \|\Omega^*\|_{1,\infty} \leq \frac{K^{\frac{1}{p}} d}{\rho}.$$

We use this inequality to derive the upper-bound. From the definition, the precision matrix is given by $\Lambda^*_k = \Theta^* + \Omega^*_k$, and hence we have

$$\begin{aligned}
\|\Lambda^*_k\|_{\mathrm{S}} &\leq \|\Theta^*\|_{\mathrm{S}} + \|\Omega^*_k\|_{\mathrm{S}} \\
&\leq \|\Theta^*\|_{\mathrm{S}} + d \max_{i,j=1,2,\ldots,d} |\Omega^*_{k,ij}| \\
&\leq \|\Theta^*\|_{\mathrm{S}} + d \max_{k=1,2,\ldots,K} \max_{i,j=1,2,\ldots,d} |\Omega^*_{k,ij}| \\
&\leq \|\Theta^*\|_{\mathrm{S}} + d \|\Omega^*\|_{1,\infty} \\
&\leq d \left(\|\Theta^*\|_{\mathrm{S}} + \|\Omega^*\|_{1,\infty} \right) \\
&\leq d \left(\|\Theta^*\|_1 + \|\Omega^*\|_{1,\infty} \right) \\
&\leq \frac{K^{\frac{1}{p}} d^2}{\rho}.
\end{aligned}$$

Here, we have used the relationship $\|\Theta^*\|_{\mathrm{S}} \leq \|\Theta^*\|_2 \leq \|\Theta^*\|_1$. □

4.8.3.3 Proof of Theorem 10

The Hessian matrix of the CSSL primal loss $\sum_{k=1}^{K} \eta_k \ell(\Theta + \Omega_k; \hat{\Sigma}_k)$ is given by

$$\mathcal{H}_{\text{primal}} = - \begin{bmatrix}
\sum_{k=1}^{K} \eta_k Q_k & \eta_k Q_1 & \eta_2 Q_2 & \cdots & \eta_K Q_K \\
\eta_1 Q_1 & \eta_1 Q_1 & 0_{d^2 \times d^2} & \cdots & 0_{d^2 \times d^2} \\
\eta_2 Q_2 & 0_{d^2 \times d^2} & \eta_2 Q_2 & & \vdots \\
\vdots & \vdots & & \ddots & 0_{d^2 \times d^2} \\
\eta_K Q_K & 0_{d^2 \times d^2} & \cdots & 0_{d^2 \times d^2} & \eta_K Q_K
\end{bmatrix},$$

where $Q_k = (\Theta + \Omega_k)^{-1} \otimes (\Theta + \Omega_k)^{-1}$. It is easy to verify that $[-I_d, \mathbf{1}_N^\top \otimes I_d]^\top$ spans a null space of $\mathcal{H}_{\text{primal}}$ and thus $\mathcal{H}_{\text{primal}}$ is always rank-deficient.

On the other hand, the matrix of the CSSL dual loss $-\sum_{k=1}^{K}\eta_k \log\det W_k$ is the block-diagonal matrix

$$\mathcal{H}_{\text{dual}} = \text{block-diag}(\eta_1\tilde{Q}_1, \eta_2\tilde{Q}_2, \ldots, \eta_K\tilde{Q}_K),$$

where $\tilde{Q}_k = W_k^{-1} \otimes W_k^{-1}$. From Theorem 9, we know that the CSSL solution has bounded eigenvalues and thus the above Hessian matrix is always strictly positive definite for any feasible W_k. □

4.8.3.4 Proof of the Proposition 2

Let $\hat{\Sigma}_k$ be the covariance matrix $\hat{\Sigma}_k = \begin{bmatrix} a_k & r_k \\ r_k & b_k \end{bmatrix}$. We then have an upper-bound of (4.12) given by

$$\begin{aligned}
&\sum_{k=1}^{K}\eta_k\left\{\log(u_kv_k - (\theta+\omega_k)^2) - (a_ku_k + b_kv_k + 2r_k\theta + 2r_k\omega_i)\right\} - 2\rho|\theta| - 2\gamma\left\|\boldsymbol{\omega}\right\|_p \\
&\leq \sum_{k=1}^{K}\eta_k\left\{\log(u_kv_k - (\theta+\omega_k)^2) - (a_ku_k + b_kv_k) - 2\left(r_k\omega_k + \gamma|\omega_k|\right)\right\} \\
&\quad - 2\left(\sum_{k=1}^{K}\eta_k r_k\theta + \rho|\theta|\right),
\end{aligned}$$

from the relationship $\sum_{k=1}^{K}\eta_k|\omega_k| \leq \left\|\boldsymbol{\omega}\right\|_\infty \leq \left\|\boldsymbol{\omega}\right\|_p$. Moreover, this upper-bound coincides with the original problem when $\boldsymbol{\omega} = \mathbf{0}_K$. Therefore, if $\boldsymbol{\omega} = \mathbf{0}_K$ is a maximizer of this upper-bound, it is also a maximizer of (4.12). From the derivative of the upper-bound over ω_k, we get that $\omega_k = 0$ is a maximizer if the following condition holds:

$$-(\gamma + r_k) \leq \frac{\theta}{u_kv_k - \theta^2} \leq (\gamma - r_k).$$

This is a sufficient condition for the original problem (4.12) to have $\omega_k = 0$ as its solution. Under this condition, problem (4.12) can be expressed as

$$\begin{aligned}
&\max_{\theta,\tilde{u},\tilde{v},u_k,v_k} \log(\tilde{u}\tilde{v} - \theta^2) - (\tilde{a}\tilde{u} + \tilde{b}\tilde{v}) - 2(\tilde{r}\theta + \rho|\theta|), \\
&\text{s.t. } \tilde{u}\tilde{v} - \theta^2 > 0, \\
&\quad -(\gamma + r_k) \leq \frac{\theta}{u_kv_k - \theta^2} \leq (\gamma - r_k) \quad (k = 1, 2, \ldots, K),
\end{aligned}$$

for some properly chosen $\tilde{a}, \tilde{b}$, and $\tilde{r} = \sum_{k=1}^{K} \eta_k r_k$. Hence, since the additional condition involves $\theta = 0$ irrelevant to the value of u_k and v_k if $\max_{k=1,2,\ldots,K} |r_k| \leq \gamma$ holds, we have $\theta = 0$ when $|\tilde{r}| \leq \rho$ from Idé et al. (2009, Proposition 1). □

4.8.3.5 Proof of Theorem 11

Let $h(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2/2$ and $\boldsymbol{y}'$ be one of the feasible solutions to the original problem (4.10). Moreover, since $\tilde{\boldsymbol{y}}$ is infeasible for the original problem (4.10), $\|\tilde{\boldsymbol{y}}\|_q > \gamma$ holds. Then, for $\boldsymbol{y}'' = \boldsymbol{y}' + \epsilon(\tilde{\boldsymbol{y}} - \boldsymbol{y}')$ with $0 < \epsilon \leq 1$, $h(\boldsymbol{y}'') \leq h(\boldsymbol{y}')$ holds from the convexity of h. Therefore, $\boldsymbol{y}''$ is a better solution to the problem (4.10) as long as the following two constraints are fulfilled:

$$|\mathbf{1}_K^\top \boldsymbol{y}''| \leq \rho,$$
$$\|\boldsymbol{y}''\|_q \leq \gamma.$$

The first condition always holds because

$$|\mathbf{1}_K^\top \boldsymbol{y}''| \leq (1-\epsilon)|\mathbf{1}_K^\top \boldsymbol{y}'| + \epsilon|\mathbf{1}_K^\top \tilde{\boldsymbol{y}}| \leq \rho.$$

On the other hand, the latter condition $\|\boldsymbol{y}''\|_q = \left(\sum_{k=1}^{K} |y_k''|^q\right)^{\frac{1}{q}} \leq \gamma$ is no longer valid if $\|\boldsymbol{y}'\|_q = \gamma$ and $\mathrm{sgn}(y_k') = \mathrm{sgn}(\tilde{y}_k - y_k')$, which results in $\tilde{y}_k y_k' \geq 0$. This is a necessary condition for the solution to (4.10). Otherwise, we can always improve the solution by the above procedure, which contradicts its optimality. □

Chapter 5

Structure Learning for Anomaly Localization

5.1 Introduction

The main scope of this chapter is an anomaly localization problem which we considered in Section 4.6 as a benchmark application. This is the one important technical field of data mining; related topics involve a change-point detection of time series (Basseville & Nikiforov, 1993; Siegmund & Venkatraman, 1995; Kohlmorgen et al., 1999) and an outlier detection (Hodge & Austin, 2004). A localization of anomalous variables is a key task toward characterizing the cause of the change (Idé, Papadimitriou, & Vlachos, 2007; Idé et al., 2009; Hirose, Yamanishi, Nakata, & Fujimaki, 2009; Jiang, Fei, & Huan, 2011). This is an essential technology for finding erroneous sensors automatically, for instance. The importance of anomaly localization techniques is especially high in engineering systems (Idé et al., 2009) and on sensor networks (Hirose et al., 2009), where the number of possible faulty sensors can be large. In such situations, the localization of errors requires professional investigations and tends to be costly. There are two fundamental directions on the study of anomaly localization; one is a graph based approach (Idé et al., 2007, 2009; Sun, Qu, Chakrabarti, & Faloutsos, 2005; Papadimitriou, Sun, & Yu, 2006; Sun, Xie, Zhang, & Faloutsos, 2007) as we considered in Section 4.6 and the other is a PCA based approach (Hirose et al., 2009; Jiang et al., 2011) which seeks a subspace where anomalies occur.

As in Section 4.6, the graph based approach consists of two stages, 1) estimating GGMs from datasets sampled before and after the error onset, and 2) finding

anomalous variables by contrasting these GGMs using a KL divergence based metric (4.13). The most important part in this method is an estimation of GGMs where the estimation error in this stage may mask faulty variables and lead to higher false detection rates. The aim of this chapter is to improve the anomaly localization performance by providing good estimates of GGMs. To that end, we consider an invariance specific to this task. The proposed method is based on this newly defined pattern; we introduce a new regularization term that penalizes a difference of precision matrices in a row/column-wise manner, which we show in the simulation that it is more suitable to the anomaly localization than the one we considered in Chapter 4.

The major challenge of this study is how to deal with the new regularization term and solve the estimation problem. The difficulty lies in two fundamental parts, that is, 1) the new term is the sum of group regularization terms with overlapping supports between the groups, and 2) the penalty is symmetric up to a matrix transpose. In particular, the first difficulty makes the computation of the proximity operator on our new regularization term inefficient and thus DAL-ADMM is not directly applicable to the problem. However, we show that these two difficulties can be avoided by formulating the problem properly. Hence, we can apply DAL-ADMM after the transformation. The resulting algorithm requires only analytic operations in each updating step.

The remainder of this chapter is organized as follows. In Section 5.2, we formalize the GGM based anomaly localization problem. In Section 5.3, we present the proposed invariant pattern and formulate the GGM learning problem. The algorithm with DAL-ADMM is also described in this section. The validity of the proposed method is presented through an experiment using sensor error data in Section 5.4. Finally, we conclude the chapter in Section 5.5.

5.2 Anomaly Localization with GGMs

In this section, we revisit the GGM based anomaly localization problem and provide its detailed formalization. In an anomaly localization task, we have two datasets, where one is sampled before the error onset and the other after that. The goal is

to identify contributions of each of d random variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$ to the difference between these two datasets. In the following, we assume the next two points on the dataset.

- The number of variables in each dataset is the same and they are all d-dimensional.
- The identity of each variable are the same, for instance, a realization of x_1 is always a value from the same sensor.

Under this condition, Idé et al. (2009) proposed to represent data with GGMs and score the degree of anomaly for each variable using a KL divergence[1]. The underlying assumption on GGM in this approach is as follows (Idé et al., 2009).

Assumption 1 (Neighborhood Preservation). *If the system is working normally, the neighborhood graph of each node (variables) is almost invariant against the fluctuations of experimental conditions.*

Formally, let $\Lambda_1, \Lambda_2 \in \mathbb{R}^{d\times d}$ be precision matrices from two datasets and their partitions be $\Lambda_k = \begin{bmatrix} L_k & \boldsymbol{l}_k \\ \boldsymbol{l}_k^\top & \lambda_k \end{bmatrix}$ with $k = 1, 2$ where $\boldsymbol{l}_k$ and λ_k correspond to the original ith row/column of matrices after permuting rows and columns of matrices simultaneously. The above assumption indicates that if there are no errors occurring on the ith variable x_i between two datasets, the pairs $\{\boldsymbol{l}_1, \lambda_1\}$ and $\{\boldsymbol{l}_2, \lambda_2\}$ are almost identical. The comparison of these two pairs corresponds to contrasting two conditional distributions and an anomaly score (4.13) arises as its metric. The score (4.13) marks higher values when the neighborhood structure on x_i changes along the error.

From the definition of the anomaly score (4.13), it is obvious that providing good estimates of Λ_1 and Λ_2 from data is an essential step to estimate the anomaly score accurately. The use of SICS (1.11) for this purpose was firstly introduced by Idé et al. (2009). In Section 4.6, we show that the CSSL estimator provides better localization performance than the one of SICS. Note that when there are only two

[1]See Section 4.6.1 for the detail.

datasets, the CSSL problem can be simplified as follows under slight modifications:

$$\max_{\Lambda_1,\Lambda_2\in\mathcal{S}^+} \sum_{k=1}^{2} \ell(\Lambda_k;\hat{\Sigma}_k) - \rho\,\|\Lambda_1-\Lambda_2\|_1\,, \tag{5.1}$$

where $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are sample covariance matrices obtained from each dataset.

5.3 Anomalous Neighborhood Selection

Now, we turn to providing the proposed formulation using a row/column-wise regularization. We also show that, with a proper transformation, the problem can be solved through DAL-ADMM.

5.3.1 Row/Column-wise Regularization

In the CSSL formulation (5.1), we regularized the difference of two matrices in an element-wise manner. However, the neighborhood preservation assumption indicates that if no error is occurring on a variable x_i, its neighborhood graphs on two GGMs may be kept almost constant across two datasets. Or alternatively, an error on x_i causes some changes on its neighborhood graphs. In a precision matrix literature, this corresponds that two matrices have row/column-wise changes before and after the error onset. Therefore, it is much more appropriate to find row/column-wise differences between matrices rather than element-wise changes.

We formalize this problem by introducing a row/column-wise regularization term. Specifically, we model the difference of Λ_1 and Λ_2 as the sum of d components $\Omega_1, \Omega_2, \ldots, \Omega_d \in \mathbb{R}^{d\times d}$ given by

$$\Lambda_1 - \Lambda_2 = \sum_{i=1}^{d} \Omega_i,$$

where each Ω_i has a support $\mathrm{supp}(\Omega_i) = \{(j,j'); j = i \vee j' = i\}$, that is, the (j,j')th entry of Ω_i is zero for any $(j,j') \notin \mathrm{supp}(\Omega_i)$. See Figure 5.1 for an illustrative image. In this parametrization, each Ω_i corresponds to the row/column-wise change caused by an error on the variable x_i. The condition $\Omega_i \neq 0_{d\times d}$ implies that the difference

$$\Lambda_1 - \Lambda_2 = \sum_{i=1}^{d} \Omega_i$$

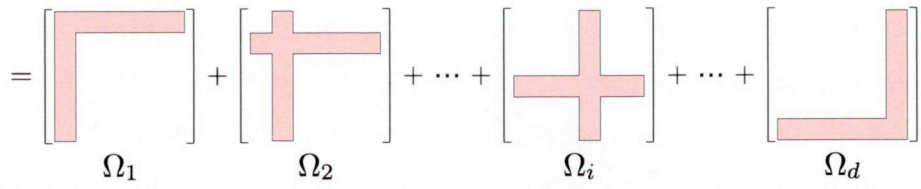


Figure 5.1: Row/column-wise parametrization of a difference between two precision matrices. Each matrix Ω_i has a support on the ith row/column denoted by colored regions.

$\Lambda_1 - \Lambda_2$ has a non-zero ith row/column and therefore the ith variable is anomalous, while $\Omega_i = 0_{d\times d}$ indicates that the ith variable is healthy.

To make the estimators to have this group-wise zero/non-zero structures, we penalize each Ω_i in a group-wise manner using a group regularization term (M. Yuan & Lin, 2006):

$$\phi(\Omega) \equiv \sum_{i=1}^{d} \sqrt{\frac{1}{2}\Omega_{i,ii}^2 + \sum_{(j,j')\in \mathrm{off}(\Omega_i)} \Omega_{i,jj'}^2},$$

where $\Omega_{i,jj'}$ denotes the (j,j')th entry of Ω_i, $\mathrm{off}(\Omega_i) = \mathrm{supp}(\Omega_i)\backslash\{(i,i)\}$ is an off-diagonal support, and we halved the effect of a diagonal term to make the optimization process simple. With this term, we define the following convex optimization problem:

$$\max_{\Lambda_1,\Lambda_2\in\mathcal{S}^+,\{\Omega_i\}_{i=1}^d} \sum_{k=1}^{2} \ell(\Lambda_k;\hat{\Sigma}_k) - \rho\phi(\Omega), \quad \text{s.t. } \Lambda_1 - \Lambda_2 = \sum_{i=1}^{d} \Omega_i, \tag{5.2}$$

which we call Anomalous Neighborhood Selection (ANS).

5.3.2 Optimization via DAL-ADMM

In solving ANS (5.2), there are two fundamental difficulties both owing to the term $\phi(\Omega)$. The first difficulty is that each support of Ω_i overlaps to one another which promotes some redundancies in the model. Suppose Λ_1 and Λ_2 are both composed of non-zero diagonal entires with one non-zero off-diagonal value on the (i,i')th

entry. Identifying whether this off-diagonal non-zero entry is induced by a term Ω_i or $\Omega_{i'}$ is not always possible (Obozinski et al., 2011). This kind of model is known as Latent-Group-Lasso and its properties are analyzed by Obozinski et al. (2011). The second difficulty is that Λ_1, Λ_2 as well as Ω_1, Ω_2, ..., Ω_d are all symmetric matrices. Explicitly imposing symmetricity constraints in the problem will make the entire optimization process complicated and might even harm the computational efficiency. If there is only the first constraint, the problem (5.2) is one specific example of Latent-Group-Lasso, and a covariate duplication technique (Obozinski et al., 2011) will be a possible approach to solve the problem while the additional second constraint makes the problem more difficult.

We tackle this problem by using DAL-ADMM. To begin with, we derive the dual problem of ANS (5.2).

Proposition 3 (Dual Problem of ANS). *The dual problem of ANS (5.2) is given by*[2]

$$\begin{aligned} &\min_{Y=Y^\top} -\log\det(\hat{\Sigma}_1 - Y) - \log\det(\hat{\Sigma}_2 + Y), \\ &\text{s.t. } 2Y_{ii}^2 + \sum_{(j,j')\in \mathrm{off}(\Omega_i)} Y_{jj'}^2 \leq \rho^2 \quad (i = 1, 2, \ldots, d). \end{aligned} \tag{5.3}$$

Here, $Y \in \mathbb{R}^{d\times d}$ is a dual variable and its optimal value Y^* relates to the optimal primal solutions Λ_1^* and Λ_2^* through $\Lambda_1^* = (\hat{\Sigma}_1 - Y^*)^{-1}$ and $\Lambda_2^* = (\hat{\Sigma}_2 + Y^*)^{-1}$.

To deal this problem with DAL-ADMM, we need to compute a proximity operator. Let $\mathcal{C} = \{Y \in \mathbb{R}^{d\times d}; 2Y_{ii}^2 + \sum_{(j,j')\in \mathrm{off}(\Omega_i)} Y_{jj'}^2 \leq \rho^2 \quad (i = 1, 2, \ldots, d)\}$. The proximity operator defined on the convex conjugate of $\varphi_\rho(\Lambda) = \rho\phi(\Omega)$ is then given by

$$\mathrm{prox}_{\varphi_\rho^*}(B) = \mathrm{proj}(B, \mathcal{C}).$$

This is a convex optimization problem and the solution can be found using some proper algorithms. The question is whether that computation can be conducted efficiently or not. Unfortunately, the shape of a set $\mathcal{C}$ is quite complicated because of some variable overlaps between inequalities, which makes the computation of

[2]We explicitly included the constraint $Y = Y^\top$ to show the symmetricity.

this projection less obvious. We tackle this problem not directly, but with some modifications, where the resulting problem requires much simpler operations only.

In DAL-ADMM, or more generally in ADMM, the problem is composed of two convex functions and some linear constraints defined between two groups of variables (Boyd et al., 2011). Our basic idea is to design these functions and constraints so that the resulting algorithm becomes simple. We first introduce two additional parameters $W_1, W_2 \in \mathbb{R}^{d\times d}$ that satisfy $W_1 = \hat{\Sigma}_1 - Y$ and $W_2 = \hat{\Sigma}_2 + Y$, respectively. We then combine the symmetricity constraint into these two equations and derive two additional equations $W_1 = \hat{\Sigma}_1 - Y^\top$ and $W_2 = \hat{\Sigma}_2 + Y^\top$. Note that one of the above four equations is redundant but we deal them equally to make the entire expression symmetric. Since these four new equations now involve the symmetricity constraint, we no longer need to impose the symmetricity on Y explicitly. This allows us to rewrite the inequality constraint into the following form:

$$\sum_{j=1}^{d} Y_{ji}^2 \leq \frac{\rho^2}{2} \quad (i = 1, 2, \ldots, d).$$

Note that this new constraint no longer have any parameter overlaps and hence it is not symmetric over Y. Together with the above four equalities, we can derive the equivalent dual problem with (5.3) but with simpler constraints as

$$\begin{aligned} &\min_{W_1, W_2 \in \mathcal{S}^+, Y} -\sum_{k=1}^{2} \log\det W_k, \\ &\text{s.t. } R_1 = R_2 = R_3 = R_4 = 0_{d\times d}, \\ &\quad \sum_{j=1}^{d} Y_{ji}^2 \leq \frac{\rho^2}{2} \quad (i = 1, 2, \ldots, d), \end{aligned} \tag{5.4}$$

where

$$\begin{aligned} R_1 &= W_1 + Y - \hat{\Sigma}_1, \\ R_2 &= W_1 + Y^\top - \hat{\Sigma}_1, \\ R_3 &= W_2 - Y - \hat{\Sigma}_2, \\ R_4 &= W_2 - Y^\top - \hat{\Sigma}_2. \end{aligned}$$

Now, we turn to solving the problem (5.4) with ADMM. First, we introduce the following Augmented Lagrangian function:

$$\mathcal{L}_\beta(W, Y, Z) = -\sum_{k=1}^{2} \log\det W_k + \delta(Y) + \frac{\beta}{2}\sum_{m=1}^{4}\left\| R_m + \frac{1}{\beta} Z_m \right\|_{\mathrm{F}}^2,$$

where Z_m is a Lagrange multiplier and β is a non-negative parameter. The function $\delta(Y)$ is an indicator function on Y, which is defined as

$$\delta(Y) = \begin{cases} 0 & \text{if } \sum_{j=1}^{d} Y_{ji}^2 \leq \rho^2/2 \ \ (i = 1, 2, \ldots, d), \\ \infty & \text{otherwise.} \end{cases}$$

Using this Augmented Lagrangian, we repeat the following three steps:

$$\begin{cases} W_1^{(t+1)}, W_2^{(t+1)} \in \underset{W_1, W_2 \in \mathcal{S}^+}{\operatorname{argmin}} \mathcal{L}_\beta(W, Y^{(t)}, Z^{(t)}), \\ Y^{(t+1)} \in \underset{Y}{\operatorname{argmin}} \mathcal{L}_\beta(W^{(t+1)}, Y, Z^{(t)}), \\ Z_m^{(t+1)} = Z_m^{(t)} + \beta R_m^{(t+1)} \ \ (m = 1, 2, 3, 4), \end{cases}$$

where $R_m^{(t+1)}$ is an R_m with $W^{(t+1)}$ and $Y^{(t+1)}$. In the next three subsections, we show that the above update processes on W and Y can be solved analytically, and the optimal solutions of ANS (5.2) can be derived as the result of ADMM.

5.3.2.1 Update of W

Here, we detail the update process for W_1 as an example. The update of W_2 can be done in the same manner.

First, the optimization problem about W_1 is given by[3]

$$\begin{aligned} &\min_{W_1 \in \mathcal{S}^+} -\log\det W_1 + \beta \left\| W_1 - A_1 \right\|_{\mathrm{F}}^2, \\ &A_1 = \hat{\Sigma}_1 - \frac{1}{2}\left(Y^{(t)} + Y^{(t)\top}\right) - \frac{1}{2\beta}\left(Z_1^{(t)} + Z_2^{(t)}\right). \end{aligned} \tag{5.5}$$

Here, we note that once we initialize $Z_1^{(0)} + Z_2^{(0)}$ to be a symmetric matrix, a matrix $Z_1^{(t)} + Z_2^{(t)}$ is also symmetric for any $t \geq 0$. It can be verified easily by the induction.

[3]We use $A_2 = \hat{\Sigma}_2 + (Y^{(t)} + Y^{(t)\top})/2 - (Z_3^{(t)} + Z_4^{(t)})/2\beta$ instead of A_1 for the update of W_2.

From the ADMM updating rules on Z_1 and Z_2, we have

$$Z_1^{(t+1)} + Z_2^{(t+1)} = Z_1^{(t)} + Z_2^{(t)} + \beta\left(Y^{(t)} + Y^{(t)\top}\right) - 2\beta\hat{\Sigma}_1,$$

which is symmetric if $Z_1^{(t)} + Z_2^{(t)}$ is symmetric.

Under the symmetric initialization on $Z_1^{(0)} + Z_2^{(0)}$, the matrix A_1 is also guaranteed to be symmetric from its definition. Hence, the first order optimality condition of (5.5) is given by the following matrix equation:

$$W_1 - \frac{1}{2\beta}W_1^{-1} - A_1 = 0_{d\times d}.$$

Here, let an eigenvalue decomposition of A_1 be $A_1 = U\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)U^\top$. The solution to the matrix equation is then $W_1^{(t+1)} = U\mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_d)U^\top$ where $\tilde{\sigma}_i$ is a solution to the quadratic equation $\tilde{\sigma}_i - \tilde{\sigma}_i^{-1}/2\beta - \sigma_i = 0$ and is given by

$$\tilde{\sigma}_i = \frac{1}{2}\left(\sigma_i + \sqrt{\sigma_i + \frac{2}{\beta}}\right).$$

5.3.2.2 Update of Y

We first define a matrix B as

$$B = \frac{1}{2}\left\{\left(\hat{\Sigma}_1 - W_1^{(t+1)}\right) - \left(\hat{\Sigma}_2 - W_2^{(t+1)}\right)\right\} - \frac{1}{4\beta}\left(Z_1^{(t)} + Z_2^{(t)\top} - Z_3^{(t)} - Z_4^{(t)\top}\right).$$

The optimization problem over Y is then defined as follows:

$$\min_Y \frac{1}{2}\|Y - B\|_{\mathrm{F}}^2, \quad \text{s.t.} \quad \sum_{j=1}^{d} Y_{ji}^2 \leq \frac{\rho}{2} \quad (i = 1, 2, \ldots, d).$$

This problem can be further decomposed into individual problems defined on each column of matrices. Here, let $Y = \begin{bmatrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \ldots & \boldsymbol{y}_d\end{bmatrix}$ and $B = \begin{bmatrix}\boldsymbol{b}_1 & \boldsymbol{b}_2 & \ldots & \boldsymbol{b}_d\end{bmatrix}$. Each subproblem is then defined as

$$\min_{\boldsymbol{y}_i} \frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{b}_i\|_2^2, \quad \text{s.t.} \quad \|\boldsymbol{y}_i\|^2 \leq \frac{\rho^2}{2}. \tag{5.6}$$

This problem has two possible cases as its solution. First, when $\|\boldsymbol{b}_i\|_2^2 \leq \rho^2/2$ holds, the solution is $\boldsymbol{y}_i = \boldsymbol{b}_i$. On the other hand, if $\|\boldsymbol{b}_i\|_2^2 > \rho^2/2$, the solution is on the

boundary and $\|\boldsymbol{y}_i\|^2 = \rho^2/2$ holds from the convexity of the quadratic objective function. This time, we use a method of Lagrange multipliers and solve

$$\min_{\boldsymbol{y}_i} \max_{\mu} \frac{1}{2} \|\boldsymbol{y}_i - \boldsymbol{b}_i\|_2^2 + \frac{\mu}{2}\left(\|\boldsymbol{y}_i\|_2^2 - \frac{\rho^2}{2} \right).$$

From this problem, we have that the optimal $\boldsymbol{y}_i$ is in the form of $\boldsymbol{y}_i = \boldsymbol{b}_i/(1+\mu)$. Hence, together with the constraint $\|\boldsymbol{y}_i\|_2^2 = \rho^2/2$, we derive the solution as

$$\boldsymbol{y}_i = \frac{\rho}{\sqrt{2}\,\|\boldsymbol{b}_i\|_2} \boldsymbol{b}_i.$$

Thus, the overall solution to the problem (5.6) is given by

$$\boldsymbol{y}_i = \min\left(1, \frac{\rho}{\sqrt{2}\,\|\boldsymbol{b}_i\|_2}\right) \boldsymbol{b}_i.$$

5.3.2.3 Convergence

Here, we note a convergence property of the ADMM iterative update. First, it is guaranteed that a sequence $\{Z_m^{(k)}\}_{k=1}^{\infty}$ converges to the optimal parameter Z_m^* (Boyd et al., 2011). Second, we have two conditions on optimal parameters W_1^*, W_2^* and $Z_1^*, Z_2^*, Z_3^*, Z_4^*$ as

$$W_1^{*-1} = Z_1^* + Z_2^*,$$
$$W_1^{*-2} = Z_3^* + Z_4^*,$$

which follows from the first order optimality conditions of W_1 and W_2 on an unaugmented Lagrangian function $\mathcal{L}_0(W, Y, Z)$. Together with the primal-dual optimality $\Lambda_1^* = (\hat{\Sigma}_1 - Y^*)^{-1}$, $\Lambda_2^* = (\hat{\Sigma}_2 + Y^*)^{-1}$, and linear constraints $W_1^* = \hat{\Sigma}_1 - Y^*$, $W_2^* = \hat{\Sigma}_2 + Y^*$, we derive the optimal primal parameters as

$$\Lambda_1^* = Z_1^* + Z_2^*,$$
$$\Lambda_2^* = Z_3^* + Z_4^*.$$

It indicates that we can derive the optimal precision matrices to the problem (5.2) by using the resulting Lagrange multipliers Z_m^* from the ADMM iterative update. Note that only precision matrices Λ_1 and Λ_2 are derived from ADMM while the difference parameters Ω_i are not.

5.4 Simulation

In this section, we verify the validity of ANS (5.2) through an anomaly localization simulation using a real world data.

5.4.1 Simulation Setting

In the simulation, we use the same sensor error data (Idé et al., 2009) as in Section 4.6. The dataset comprised 42 sensor values collected from a real car in 79 normal states and 20 faulty states. The fault is caused by mis-wiring of the 24th and 25th sensors, resulting in erroneous behaviors. In the simulation, we transform each dataset into a sample covariance matrix, so that we have 79 and 20 matrices from normal and faulty states, respectively. Since sample covariances are rank-deficient in some datasets, we added 10^{-3} on their diagonal to avoid singularities.

We adopt SICS (1.11) and CSSL (5.1) as baseline methods to contrast with ANS. In this simulation, we consider the two datasets case different from Section 4.6 since ANS is designed under such a situation. Therefore, the result here cannot be directly compared with those in Section 4.6 where we considered a general multiple datasets situation.

5.4.2 Result

We conducted the simulation for all 79×20 normal-faulty pairs of datasets. In each run, we have datasets from two different states and estimated two precision matrices Λ_1 and Λ_2 with three different methods, which are SICS (1.11), CSSL (5.1), and ANS (5.2). After precision matrices are estimated, we calculated the anomaly score (4.13) for each variable using estimated matrices from each of three methods. The anomaly localization performance is evaluated by drawing an ROC curve and measuring the AUC, which achieves the best result 1 if two erroneous sensors mark top two anomaly scores. The overall performance for each of three matrix estimation methods is measured as the median AUC of all 79×20 runs of the simulation.

We summarize the best median AUC results for each method among 41 different

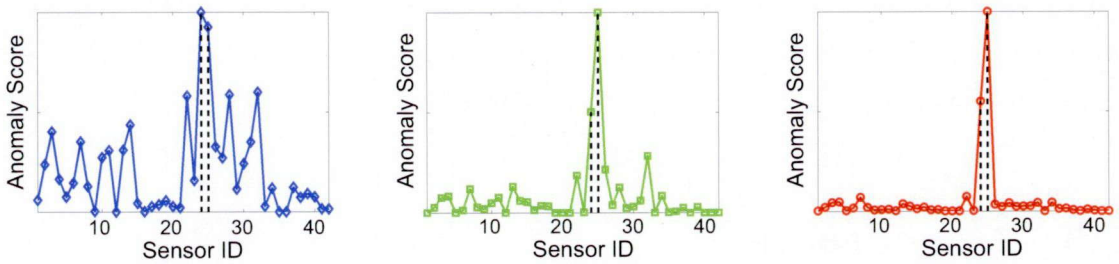
Table 5.1: Anomaly localization results for SICS, CSSL, and ANS. For each method, we compute precision matrices for 41 different values of ρ ranging from 10^{-2} to 10^{0}. The table shows the median of the best AUCs among these 41 results over all 79×20 pairs of normal-faulty datasets. The numbers in brackets are the 25% and the 75% quantiles.

	best median AUC (25% / 75% quantiles)	optimal ρ
SICS	0.9875 (0.9500 / 1.0000)	$10^{-0.50}$
CSSL	0.9875 (0.9500 / 1.0000)	$10^{-1.05}$
ANS	1.0000 (0.9750 / 1.0000)	$10^{-0.05}$

values of ρ ranging from 10^{-2} to 10^{0} in Table 5.1. The result shows the significant success of ANS that achieves AUC $= 1$ as its median performance. It means that ANS could detect faulty sensors perfectly for more than half of the 79×20 cases. To see further differences, we plot the median anomaly scores derived from each method in Figure 5.2. It is obvious that ANS successfully extract only faulty variables while the anomaly scores on other healthy variables kept almost zero. This makes sharp contrast to other methods whose scores have some peaks on some healthy sensors. The results in Figure 5.3 also support this tendency that only ANS could successfully highlight matrix entries related to anomalous sensors. In other two methods, it is hard to observe such clear evidence of errors in estimated matrices. From these results, we can conclude that ANS is the superior method to others both on anomaly localization performances and also on an interpretability of the result.

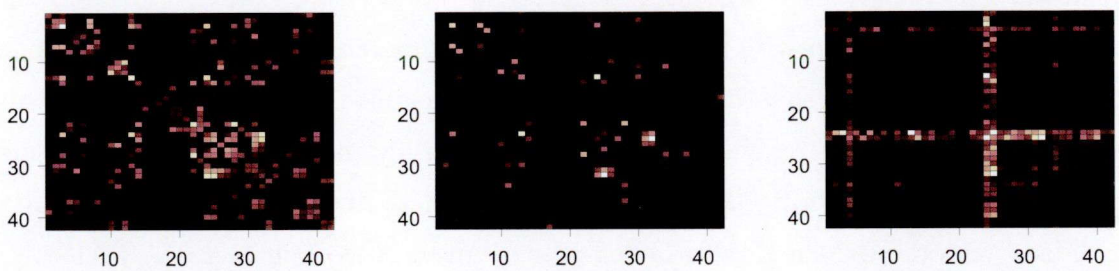
5.4.3 Discussion

Through the simulation, we observed the advantage of ANS over other two existing precision matrix learning methods. This advantage is caused by our new regularization term on the row/column-wise difference between two matrices. From the definition of an anomaly score (4.13), we can see that the score gets small when $\{\boldsymbol{l}_1, \lambda_1\}$ and $\{\boldsymbol{l}_2, \lambda_2\}$ are similar to each other. Hence, the equivalence of these two pairs results in the zero anomaly score. In the SICS problem (1.11), this kind of



(a) SICS (b) CSSL (c) ANS

Figure 5.2: Median anomaly scores for each method with best AUCs. Dotted lines denote true faulty sensors.



(a) SICS (b) CSSL (c) ANS

Figure 5.3: An example of the difference between two estimated precision matrix entries. Darker/Lighter means lower/higher discrepancies.

properties on precision matrices are not considered, which results in high variations between two matrix entries even on healthy sensors (Figure 5.3(a)). Such variations produce higher anomaly scores not only on truly faulty sensors as we can observe in Figure 5.2(a). This would be the reason why the SICS estimators have inferior anomaly localization performances. On the other hand, the CSSL problem (5.1) considers the variation between two precision matrices. Compared to the result of SICS (Figure 5.3(a)), we can observe that the resulting matrices derived through CSSL have less variations (Figure 5.3(b)). However, the regularization is applied in an element-wise manner on the variation and seems not be sufficient to extract only error related changes as we can see on some peaks in Figure 5.2(b). In the result of ANS (Figure 5.3(c)), some healthy sensor related rows/columns are also

extracted as candidates of anomalies, though their magnitude are sufficiently small and their effects are almost negligible in the anomaly score (Figure 5.2(c)). This would be the cause of the significant success of ANS.

5.5 Conclusion

In this chapter, we proposed a precision matrix estimation technique ANS (5.2) for an anomaly localization task. We focused on the neighborhood preservation assumption and considered that a row/column-wise similarity would be an appropriate invariant pattern representing healthy variables. Based on this idea, we introduced a row/column-wise regularization on the difference of two matrices, which is much more effective than existing element-wise regularization techniques for this specific task. The new regularization term has overlapping support structures and hence it is symmetric up to a matrix transpose. These difficulties can be efficiently avoided by modifying the dual problem which can be solved through DAL-ADMM. We showed that each updating step of ADMM can be computed analytically and the iterative update steps converge to the optimal parameter. We also verified the effectiveness of ANS through a real world data simulation, which shows higher anomaly localization performances and a higher interpretability.

5.6 Proofs of Theorems

5.6.1 Proof of Proposition 3

We first introduce matrices $\Gamma_i \in \mathbb{R}^{d\times d}$ $(i = 1, 2, \ldots, d)$ satisfying

$$\Gamma_{i,jj'} \geq 0 \text{ and } -\Gamma_{i,jj'} \leq \Omega_{i,jj'} \leq \Gamma_{i,jj'} \ (i, j, j' = 1, 2, \ldots, d). \tag{5.7}$$

Using these matrices, we can rewrite the problem (5.2) as

$$\max_{\Lambda_1,\Lambda_2\in\mathcal{S}^+,\{\Omega_i,\Gamma_i\}_{i=1}^d} \sum_{k=1}^{2} \ell(\Lambda_k; \hat{\Sigma}_k) - \rho\phi(\Gamma), \text{ s.t. (5.7) and } \Lambda_1 - \Lambda_2 = \sum_{i=1}^{d} \Omega_i.$$

We further introduce Lagrange multipliers P and Q_i $(i = 1, 2, \ldots, d)$ and rewrite the problem as

$$\max_{\Lambda_1, \Lambda_2 \in \mathcal{S}^+, \{\Omega_i, \Gamma_i\}_{i=1}^d} \min_{P,Q,Y} \sum_{k=1}^{2} \ell(\Lambda_k; \hat{\Sigma}_k) - \rho\phi(\Gamma)$$
$$+ \sum_{i=1}^{d} \sum_{(j,j') \in \mathrm{off}(\Omega_i)} (P_{i,jj'}\Omega_{i,jj'} - |P_{i,jj'}|\Gamma_{i,jj'} - Q_{i,jj'}\Gamma_{i,jj'})$$
$$+ \mathrm{tr}\left[Y^\top \left(\Lambda_1 - \Lambda_2 - \sum_{i=1}^{d} \Omega_i \right) \right],$$
$$\text{s.t. } Q_{i,jj'} \geq 0 \ \ (i, j, j' = 1, 2, \ldots, d).$$

By exchanging the order of the maximization and the minimization, we derive the dual problem. First, we optimize the above over Λ_1 and Λ_2 by setting the derivatives equal to zero and derive

$$\Lambda_1^{-1} - \hat{\Sigma}_1 + Y = 0_{d \times d},$$
$$\Lambda_2^{-1} - \hat{\Sigma}_2 - Y = 0_{d \times d}.$$

Secondly, from the optimization over Ω and Γ, we have the condition

$$2Y_{ii}^2 + \sum_{(j,j') \in \mathrm{off}(\Omega_i)} Y_{jj'}^2 \leq \rho^2.$$

Finally, by substituting these results into the above dual problem, we derive the result (5.3). □

Chapter 6

Conclusion

This dissertation investigated methodologies for learning invariant patterns hidden across multiple datasets. For the purpose, we focused on the second order statistics, one of the most primitive parameters representing natures of multivariate random variables. In particular, we considered two models, a linear mixing model and a graphical model, as the basis of our framework.

First, we worked on a model called Stationary Subspace Analysis (SSA), which is a variant of linear mixing models. This model assumes that the observation is a linear mixture of two kinds of latent sources, which are stationary and non-stationary. We built up the proposed algorithm, Analytic SSA (ASSA), which recovers these latent sources from the data based on the fact that the problem can be formulated as a generalized eigenvalue problem under proper conditions. The advantages of ASSA over other existing algorithms have been verified both theoretically and numerically.

Next, we considered finding invariant patterns across multiple Graphical Gaussian Models (GGMs). We first derived a general convex optimization algorithm DAL-ADMM to solve GGM learning problems. This algorithm allows us to work on a wider class of problems where existing methods could not treat. We then considered two invariant patterns on multiple GGMs, or corresponding precision matrices; the first one is an element-wise commonness across multiple precision matrices, while the latter one is a row/column-wise heterogenisity, a specific pattern for an anomaly localization task. Each of these two patterns are incorporated with a GGM learning problem by introducing new regularization terms. Hence, these problems can be solved by DAL-ADMM procedure.

Apart from the remaining problems raised in each chapter, we point out two

general issues as the further improvement directions of this study. The first one is an establishment of the general framework for an invariant pattern learning. Currently, two fundamental models, SSA and GGMs, are introduced based on the second order statistics. However, these two problems are defined on quite different principles. Introduction of a general model that unifies these problems would be needed to further improve the invariant pattern learning problem. The second issue is on the practical aspect, an introduction of task specific invariant patterns. One of this problem is already considered in Chapter 5 where we observed that the algorithm specific to the target task is superior to general methodologies. Investigations of practical invariant patterns and their learning algorithms would be a possible future direction.

References

Abed-Meraim, K. & Belouchrani, A. (2004). Algorithms for joint block diagonalization. *Proceedings of the 12th European Signal Processing Conference*, 209–212.

Agarwal, A., Negahban, S., & Wainwright, M. (2011). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Proceedings of the 28th International Conference on Machine Learning*, 1129–1136.

Ahmed, A. & Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, *106*(29), 11878–11883.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*(2), 251–276.

Avriel, M. (2003). *Nonlinear Programming: Analysis and Methods.* Dover Publications.

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*, 1179–1225.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, *9*, 485–516.

Basseville, M. & Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application.* Prentice Hall.

Belouchrani, A., Abed-Meraim, K., Cardoso, J. F., & Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, *45*(2), 434–444.

Belouchrani, A., Amin, M. G., & Abed-Meraim, K. (1997). Direction finding in correlated noise fields based on joint block-diagonalization of spatio-temporal correlation matrices. *IEEE Signal Processing Letters*, *4*(9), 266–268.

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Nikulin, V., & Müller, K.-R. (2008). Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. *Advances in Neural Information Processing Systems*, *20*, 113–120.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Muller, K.-R. (2007). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, *25*(1), 41–56.

Blythe, D. A. J., von Bünau, P., Meinecke, F. C., & Müller, K.-R. (2012). Feature extraction for change-point detection using stationary subspace analysis. *IEEE Transactions on Neural Networks*, *23*(4), 631–643.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(3), 11:1–11:37.

Cardoso, J. F. & Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings F, Radar & Signal Processing*, *140*(6), 362–370.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Chandrasekaran, V., Parrilo, P. A., & Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Annals of Statistics*, *40*(4), 1935–1967.

Chatelin, F. (1993). *Eigenvalues of Matrices.* John Wiley and Sons.

Chiquet, J., Grandvalet, Y., & Ambroise, C. (2011). Inferring multiple graphical structures. *Statistics and Computing*, *21*(4), 537–553.

Choi, S. & Cichocki, A. (2000). Blind separation of nonstationary and temporally correlated sources from noisy mixtures. *Proceedings of the 2000 IEEE Signal Processing Society Workshop*, *1*, 405–414.

Clifford, P. (1990). Markov random fields in statistics. *Disorder in physical systems*, 19–32.

Comon, P. (1994). Independent component analysis, a new concept? *Signal*

Processing, *36*(3), 287–314.

Congedo, M., Gouy-Pailler, C., & Jutten, C. (2008). On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. *Clinical Neurophysiology*, *119*(12), 2677–2686.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*(1), 157–175.

Dickey, D. A. & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431.

Dornhege, G., Millán, J. d. R., Hinterberger, T., McFarland, D. J., & Müller, K.-R. (2007). *Toward Brain-Computer Interfacing.* The MIT Press.

Duchi, J., Gould, S., & Koller, D. (2008). Projected subgradient methods for learning sparse Gaussians. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 145–152.

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, 272–279.

Durbin, J. & Koopman, S. J. (2001). *Time Series Analysis by State Space Methods.* Oxford University Press.

Engle, R. F. & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, *55*(2), 251–276.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics*, *3*(2), 521–541.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Flury, B. D. & Neuenschwander, B. E. (1994). Simultaneous diagonalization algorithms with applications in multivariate statistics. *Proceedings of the Conference on Approximation and Computation*, 179–205.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Academic Press.

Fukunishi, H. (1975). Polarization changes of geomagnetic Pi2 pulsations associated

with the plasmapause. *Journal of Geophysical Research*, *80*(1), 98–110.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations.* The Johns Hopkins University Press.

Grosse-Wentrup, M. & Buss, M. (2008). Multiclass common spatial patterns and information theoretic feature extraction. *IEEE Transactions on Biomedical Engineering*, *55*(8), 1991–2000.

Guo, F., Hanneke, S., Fu, W., & Xing, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. *Proceedings of the 24th International Conference on Machine Learning*, 321–328.

Guo, J., Levina, E., Michailidis, G., & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, *98*(1), 1–15.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton Univ. Press.

Hara, S., Kawahara, Y., Washio, T., & von Bünau, P. (2010). Stationary subspace analysis as a generalized eigenvalue problem. *Neural Information Processing. Theory and Algorithms, Lecture Notes in Computer Science*, *6443*, 422–429.

Hara, S., Kawahara, Y., Washio, T., von Bünau, P., Tokunaga, T., & Yumoto, K. (2012). Separation of stationary and non-stationary sources with a generalized eigenvalue problem. *Neural Networks*, *33*, 7–20.

Hara, S. & Washio, T. (2011). Common substructure learning of multiple graphical Gaussian models. *Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Artificial Intelligence*, *6912*, 1–16.

Hara, S. & Washio, T. (2012a). Anomalous neighborhood selection. *Proceedings of the 12th IEEE International Conference on Data Mining Workshops*, 474–480.

Hara, S. & Washio, T. (2012b). Group sparse inverse covariance selection with a dual augmented Lagrangian method. *Neural Information Processing, Lecture Notes in Computer Science*, *7665*, 108–115.

Hara, S. & Washio, T. (2013). Learning a common substructure of multiple graphical Gaussian models. *Neural Networks*, *38*, 23–38.

He, B. & Yuan, X. (2012). On the $\mathcal{O}(1/n)$ convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, *50*, 700–709.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–162.

Hirose, S., Yamanishi, K., Nakata, T., & Fujimaki, R. (2009). Network anomaly detection based on eigen equation compression. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1185–1194.

Hodge, V. & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Honorio, J. (2011). Lipschitz parametrization of probabilistic graphical models. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 347–354.

Honorio, J., Ortiz, L., Samaras, D., Paragios, N., & Goldstein, R. (2009). Sparse and locally constant Gaussian graphical models. *Advances in Neural Information Processing Systems*, *22*, 745–753.

Honorio, J. & Samaras, D. (2010). Multi-task learning of Gaussian graphical models. *Proceedings of the 27th International Conference on Machine Learning*, 447–454.

Horn, R. A. & Johnson, C. R. (1990). *Matrix Analysis.* Cambridge University Press.

Hsieh, C. J., Sustik, M. A., Dhillon, I., & Ravikumar, P. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems*, *24*, 2330–2338.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634.

Hyvarinen, A. (2002). Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, *12*(6), 1471–1474.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis.* John Wiley and Sons.

Idé, T., Lozano, A. C., Abe, N., & Liu, Y. (2009). Proximity-based anomaly detection using sparse structure learning. *Proceedings of the 2009 SIAM In-*

ternational Conference on Data Mining, 97–108.

Idé, T., Papadimitriou, S., & Vlachos, M. (2007). Computing correlation anomaly scores using stochastic nearest neighbors. *Proceedings of th 7th IEEE International Conference on Data Mining*, 523–528.

Jacobs, J. A., Kato, Y., Matsushita, S., & Troitskaya, V. A. (1964). Classification of geomagnetic micropulsations. *Journal of Geophysical Research*, *69*(1), 341–342.

Jalali, A., Ravikumar, P., Sanghavi, S., & Ruan, C. (2010). A dirty model for multi-task learning. *Advances in Neural Information Processing Systems*, *23*, 964–972.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Jiang, R., Fei, H., & Huan, J. (2011). Anomaly localization for network data streams with graph joint sparse PCA. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 886–894.

Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer.

Kaufmann, R. K. & Stern, D. I. (2002). Cointegration analysis of hemispheric temperature relations. *Journal of Geophysical Research: Atmospheres*, *107*(D2), 4012.

Kawamoto, M., Matsuoka, K., & Ohnishi, N. (1998). A method of blind separation for convolved non-stationary signals. *Neurocomputing*, *22*(1-3), 157–171.

Kohlmorgen, J., Lemm, S., Müller, K.-R., Liehr, S., & Pawelzik, K. (1999). Fast change point detection in switching dynamics using a hidden Markov model of prediction experts. *Proceedings of the 9th International Conference on Artificial Neural Networks*, 204–209.

Koles, Z. J. (1991). The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalography and Clinical Neurophysiology*, *79*, 440–447.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86.

Kuwashima, M. & Saito, T. (1981). Spectral characteristics of magnetic Pi2 pul-

sations in the auroral region and lower latitudes. *Journal of Geophysical Research: Space Physics*, *86*(A6), 4686–4696.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press.

Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, *13*, 556–562.

Li, L. & Toh, K. C. (2010). An inexact interior point method for L_1-regularized sparse covariance selection. *Mathematical Programming Computation*, *2*(3–4), 291–315.

Liu, H., Palatucci, M., & Zhang, J. (2009). Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. *Proceedings of the 26th International Conference on Machine Learning*, 649–656.

Lorenz, E. N. & Emanuel, K. A. (1998). Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, *55*(3), 399–414.

Mann, M. E. (2004). On smoothing potentially non-stationary climate time series. *Geophysical Research Letters*, *31*(7), L07214.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press.

Matsuoka, K., Ohoya, M., & Kawamoto, M. (1995). A neural net for blind separation of nonstationary signals. *Neural Networks*, *8*(3), 411–419.

Meinecke, F. C., von Bünau, P., Kawanabe, M., & Müller, K.-R. (2009). Learning invariances with stationary subspace analysis. *Proceedings of the 12th IEEE International Conference on Computer Vision Workshops*, 87-92.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Molgedey, L. & Schuster, H. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, *72*(23), 3634–3637.

Montgomery, D. (2007). *Introduction to Statistical Quality Control.* John Wiley and Sons.

Moreau, E. (2001). A generalization of joint-diagonalization criteria for source separation. *IEEE Transactions on Signal Processing*, *49*(3), 530–541.

Moreau, J. J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, *93*, 273–299.

Müller, J. S., von Bünau, P., Meinecke, F. C., Király, F. J., & Müller, K.-R. (2011). The stationary subspace analysis toolbox. *Journal of Machine Learning Research*, *12*, 3065–3069.

Murata, N., Kawanabe, M., Ziehe, A., Müller, K.-R., & Amari, S. (2002). On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks*, *15*(4-6), 743–760.

Obozinski, G., Jacob, L., & Vert, J. P. (2011). Group lasso with overlaps: The latent group lasso approach. *Arxiv preprint arXiv:1110.0413*.

Olson, J. V. & Rostoker, G. (1977). Latitude variation of the spectral components of auroral zone Pi2. *Planetary and Space Science*, *25*(7), 663–671.

Papadimitriou, S., Sun, J., & Yu, P. (2006). Local correlation tracking in time series. *Proceedings of the 6th IEEE International Conference on Data Mining*, 456–465.

Parra, L. & Sajda, P. (2003). Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, *4*, 1261–1269.

Pham, D. T. & Cardoso, J. F. (2001). Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing*, *49*(9), 1837–1848.

Plumbley, M. D. (2005). Geometrical methods for non-negative ICA: Manifolds, Lie groups and toral subalgebras. *Neurocomputing*, *67*, 161–197.

Priestley, M. B. & Rao, T. S. (1969). A test for non-stationarity of time-series. *Journal of the Royal Statistical Society: Series B*(31), 140–149.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. (2008). *Dataset Shift in Machine Learning*. The MIT Press.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *42*(1), 239–250.

Rockafellar, R. (1996). *Convex Analysis*. Princeton University Press.

Saito, T. (1969). Geomagnetic pulsations. *Space Science Reviews*, *10*, 319–412.

Saito, T., Yumoto, K., & Koyama, Y. (1976). Magnetic pulsation Pi2 as a sensitive indicator of magnetospheric substorm. *Planetary and Space Science*, *24*(11),

1025–1029.

Scheinberg, K., Ma, S., & Goldfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. *Advances in Neural Information Processing Systems*, *23*, 2101–2109.

Scheinberg, K. & Rish, I. (2010). Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach. *Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science*, *6323*, 196–212.

Schmidt, M., Van Den Berg, E., Friedlander, M., & Murphy, K. (2009). Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 456–463.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N., & Müller, K.-R. (2006). Towards adaptive classification for BCI. *Journal of Neural Engineering*, *3*, R13–R23.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244.

Siegmund, D. & Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Annals of Statistics*, *23*(1), 255–271.

Sra, S. (2011). Fast projections onto $\ell_{1,q}$-norm balls for grouped feature selection. *Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Artificial Intelligence*, *6913*, 305–317.

Sra, S., Nowozin, S., & Wright, S. (2011). *Optimization for Machine Learning.* The MIT Press.

Srebro, N., Rennie, J., & Jaakkola, T. (2005). Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, *17*(5), 1329–1336.

Sun, J., Qu, H., Chakrabarti, D., & Faloutsos, C. (2005). Neighborhood formation and anomaly detection in bipartite graphs. *Proceedings of the 5th IEEE International Conference on Data Mining*, 418–425.

Sun, J., Xie, Y., Zhang, H., & Faloutsos, C. (2007). Less is more: Compact matrix decomposition for large sparse graphs. *Proceedings of the 2007 SIAM International Conference on Data Mining*, 366–377.

Sutcliffe, P. R. & Yumoto, K. (1991). On the cavity mode nature of low-latitude Pi 2 pulsations. *Journal of Geophysical Research: Space Physics*, *96*(A2), 1543–1551.

Takahashi, K., Lee, D. H., Nosé, M., Anderson, R. R., & Hughes, W. J. (2003). CRRES electric field study of the radial mode structure of Pi2 pulsations. *Journal of Geophysical Research: Space Physics*, *108*(A5), 1210.

Theis, F. J. & Inouye, Y. (2006). On the use of joint diagonalization in blind signal processing. *Proceedings of the 2006 IEEE International Symposium on Circuits and Systems*, 3589–3592.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, *67*(1), 91–108.

Tokunaga, T., Kohta, H., Yoshikawa, A., Uozumi, T., & Yumoto, K. (2007). Global features of Pi 2 pulsations obtained by independent component analysis. *Geophysical Research Letters*, *34*, L14106.

Tomioka, R., Suzuki, T., & Sugiyama, M. (2011). Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation. *Journal of Machine Learning Research*, *12*, 1537–1586.

Tong, L., Liu, R. W., Soon, V. C., & Huang, Y. F. (1991). Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, *38*(5), 499–509.

Turlach, B. A., Venables, W. N., & Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, *47*(3), 349–363.

Uozumi, T., Yumoto, K., Kawano, H., Yoshikawa, A., Ohtani, S., Olson, J. V., Akasofu, S. I., Solovyev, S. I., Vershinin, E. F., Liou, K., Meng, C. I. (2004). Propagation characteristics of Pi 2 magnetic pulsations observed at ground high latitudes. *Journal of Geophysical Research: Space Physics*, *109*.

Varoquaux, G., Gramfort, A., Poline, J. B., & Thirion, B. (2010). Brain covariance selection: Better individual functional connectivity models using population prior. *Advances in Neural Information Processing Systems*, *23*, 2334–2342.

von Bünau, P., Meinecke, F. C., Király, F. C., & Müller, K.-R. (2009a). Finding stationary subspaces in multivariate time series. *Physical Review Letters*, *103*(21), 214101.

von Bünau, P., Meinecke, F. C., Király, F. C., & Müller, K.-R. (2009b). Finding stationary subspaces in multivariate time series, supplemental materials. *EPAPS Document* (E-PRLTA-103-014948).

von Bünau, P., Meinecke, F. C., Scholler, S., & Müller, K.-R. (2010). Finding stationary brain sources in EEG data. *Proceedings of the 32nd Annual Conference of the IEEE Engineering in Medicine and Biology Society*, 2810-2813.

Wainwright, M. J., Ravikumar, P., & Lafferty, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. *Advances in Neural Information Processing Systems*, *19*, 1465–1472.

Wang, C., Sun, D., & Toh, K. (2009). Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM Journal on Optimization*, *20*, 2994–3013.

Yeoman, T. K. & Orr, D. (1989). Phase and spectral power of mid-latitude Pi2 pulsations: Evidence for a plasmaspheric cavity resonance. *Planetary and Space Science*, *37*(11), 1367–1383.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, *68*(1), 49–67.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*, 19–35.

Yuan, X. (2009). Alternating direction methods for sparse covariance selection. *Preprint available at* `http://www.optimization-online.org/DB\_FILE/2009/09/2390.pdf`.

Yumoto, K. & the CPMN Group. (2001). Characteristics of Pi2 magnetic pulsations observed at the CPMN stations: A review of the step results. *Earth, Planets Space*, *53*, 981–992.

Zhang, B. & Wang, Y. (2010). Learning structural changes of Gaussian graphical models in controlled experiments. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 701–708.

Zhou, S., Lafferty, J., & Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, *80*(2), 295–319.

Ziehe, A., Laskov, P., Nolte, G., & Müller, K.-R. (2004). A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, *5*, 777–800.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.

List of Publications

Journal Papers

- Hara, S. & Washio, T. (2013). Learning a common substructure of multiple graphical Gaussian models. *Neural Networks*, *38*, 23–38.

- Hara, S., Kawahara, Y., Washio, T., von Bünau, P., Tokunaga, T., & Yumoto, K. (2012). Separation of stationary and non-stationary sources with a generalized eigenvalue problem. *Neural Networks*, *33*, 7–20.

Conference Proceedings

- Hara, S. & Washio, T. (2012a). Anomalous neighborhood selection. *Proceedings of the 12th IEEE International Conference on Data Mining Workshops*, 474–480.

- Hara, S. & Washio, T. (2012b). Group sparse inverse covariance selection with a dual augmented Lagrangian method. *Neural Information Processing, Lecture Notes in Computer Science*, *7665*, 108–115.

- Hara, S. & Washio, T. (2011). Common substructure learning of multiple graphical Gaussian models. *Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Artificial Intelligence*, *6912*, 1–16.

- Hara, S., Kawahara, Y., Washio, T., & von Bünau, P. (2010). Stationary subspace analysis as a generalized eigenvalue problem. *Neural Information Processing. Theory and Algorithms, Lecture Notes in Computer Science*, *6443*, 422–429.