MAXIMUM LIKELIHOOD IDENTIFICATION OF AN INFORMATION MATRIX UNDER CONSTRAINTS IN A CORRESPONDING GRAPHICAL MODEL

By

Nan Li

Master Thesis

Submitted to the Faculty

Of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

In

Applied Statistics

December, 2016

APPROVED:

Professor Randy C Paffenroth, Thesis Advisor

Acknowledgements

I would like to express my very great appreciation to my thesis advisor, Dr. Randy C. Paffenroth, who not only provided this meaningful topic but also put so much effort helping me understand the essence of my thesis and making it go the right way. I am also inspired and encouraged to explore and learn new knowledge during this period under his instruction. I also thank Professor Randy C. Paffenroth for taking so much time helping me examine the thesis carefully and patiently. I wish to acknowledge the great programming advice and technical help provided by Chong Zhou. Also, my special thanks are extended to all the professors who imparted me with the valuable statistical knowledge and their assistance for the past five semesters.

Notation

X	Multinormal data set as a $M \times N$ matrix
μ_X	Mean of data set X
Σ	Population covariance
Σ_{XX}	Sample covariance from data set X
Σ_t	Optimal population covariance from convex solver
Σ_{te}	Optimal population covariance from a closed form solution
Σ^{-1}	Population information matrix
Σ_{XX}^{-1}	Sample information matrix from data set X
Σ_t^{-1}	Optimal population information matrix from convex solver
Σ_{te}^{-1}	Optimal population information matrix from a closed form solution
$\mathbb{1}_i$	Column vector that has element "1" at the ith position, "0" at other positions
$\mathbb{1}_{ij}$	A matrix with two columns that has element "1" at the ith position for the first column, element "1" at the jth position for the second column, and "0" at other positions
λ	A Lagrange multiplier
R	$\left[(X-\mu_X)(X-\mu_X)^H\right]$
C	A vertex set
$\mathbf{C}$	A simple undirected graph
$E(\mathbf{C})$	The edge set of graph $\mathbf{C}$
$\tilde{\mathbf{C}}$	Complementary graph of $\mathbf{C}$
$\hat{\Theta}$	Graphical Lasso Estimator of Σ^{-1}
ρ	Graphical Lasso penalization parameter

Abstract

We address the problem of identifying the neighborhood structure of an undirected graph, whose nodes are labeled with the elements of a multivariate normal (MVN) random vector. A semi-definite program is given for estimating the information matrix under *arbitrary* constraints on its elements. More importantly, a *closed-form* expression is given for the maximum likelihood (ML) estimator of the information matrix, under the constraint that the information matrix has pre-specified elements in a given pattern (e.g., in a principal submatrix). The results apply to the identification of dependency labels in a graphical model with neighborhood constraints. This neighborhood structure excludes nodes which are conditionally independent of a given node and the graph is determined by the non-zero elements in the information matrix for the random vector.

A cross-validation principle is given for determining whether the constrained information matrix returned from this procedure is an acceptable model for the information matrix, and as a consequence for the neighborhood structure of the Markov Random Field (MRF) that is identified with the MVN random vector.

Contents

1 Introduction

Markov Random Fields(MRF) and Graphical Models naturally arise in many estimations and machine learning problems [2, 3]. Perhaps even more importantly, *Sparse* graphical models [4] are classically computed using techniques such as the Graphical Lasso [4, 5]. Herein, we extend and generalize such techniques to the case where the underlying information matrix (also called concentration matrix) has a specified pattern of entries. For example, similar to the Graphical Lasso, our techniques provide a closed form solution for the maximum likelihood(ML) estimate of an information matrix, under constraints on the corresponding graphical model. Given empirical measurements from the underlying generative process, they provide a principle for cross-validating the graphical model.

The work in this thesis is closely related to a companion paper *Maximum Likelihood Identification of an Information Matrix Under Constraints in a Corresponding Graphcical Model* [6] which was recently published in the proceedings of the 2016 Asilomar conference on Signals, Systems, and Computers. This paper was produced in collaboration with Professor Randy Paffenroth, from WPI, and Professor Louis Scharf, from Colorado State University. The paper and this thesis share the same structure and results. However, this thesis provides the detailed proofs and applications that are only outlined in the companion paper. Especially, this thesis provides the theoretical basis which are the foundations of our research. Interested readers can see the paper [6] as a summary of this thesis.

Consider the following construction. Given a multivariate normal random vector $\mathbf{x} = [x_1, x_2, \cdots, x_n] \in \mathcal{C}^n$, construct a corresponding undirected graph G [2]whose nodes are identified with the random variables x_i and whose edges are labeled with the elements of the information matrix Σ^{-1}. Assume $\mathbf{x} \sim \mathrm{CN}_n[\mathbf{0}, \Sigma]$ follows a multi-normal distribution. The non-zero elements of the information matrix then code for the neighborhood structure of a Markov Random Field, and according to the Hammersley-Clifford Theorem, for the cliques in the global Gibbs distribution for the multivariate normal distribution [7]. .

Locally, the zero elements of the information matrix determine conditional independence between random variables, and even when the random variables are not multivariate normal, they determine random variables which do not participate in a linear minimum variance unbiased estimator of a given node from all others (sometimes called the BLUE estimator for Best Linear Unbiased Estimator). Thus, the neighborhood of node i in an MRF model for the random vector $\mathbf{x}$ consists of only those random variables that would participate in a BLUE of the random variable x_i. It excludes all random variables x_j that are conditionally linearly independent of x_i, and these excluded random variables are just those for which $\Sigma^{-1}_{ij} = 0$.

So here is the question: given a random sample of the MVN random vector $\mathbf{x}$, what is the maximum likelihood estimator for the information matrix Σ^{-1}, under the constraint that a pattern of entries in Σ^{-1} is specified, and once determined, how is this estimator to be cross-validated? In this thesis we address this question and give the following results:

1. A semidefinite program [8] that returns the constrained ML estimator of Σ^{-1}, where the constraints are an arbitrary pattern of pre-specified entries in Σ^{-1}. A typical pattern is a pattern of zeros defining neighborhood structure in a corresponding graphical model.

2. A *closed form solution* for the ML estimator of the information matrix Σ^{-1} with a constraint on one pair (i, j) of symmetric entries of the information matrix:

$$\arg\max_{\Sigma} -\frac{n}{2}\ln\det(\Sigma) - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}(X-\mu_X)(X-\mu_X)^H\right]$$
$$\text{s.t. } (\Sigma^{-1})_{ij} = (\Sigma^{-1})_{ji} = 0$$

3. A closed form solution for the ML estimator of the information matrix Σ^{-1} with a constraint on any principal submatrix of the information matrix, derived by generalizing the solution method of the previous case.

4. A procedure for cross-validating the constrained ML solutions for Σ^{-1} and Σ as an acceptable model for the multivariate random vector $\mathbf{x}$, and for the corresponding graphical model and MRF encoded into the constraints [2, 3].

It is our intention that these results will extend many practices for identifying Markov Random Fields and Graphical Models. Our closed form solutions for the constrained ML estimators of Σ^{-1} and Σ, under constraints on Σ^{-1} inherited from a graphical model, may be cross-validated with the principle of Expected Likelihood, meaning experimental realizations of $\mathbf{x}$ may be used to cross-validate a hypothetical graphical model. Even further, our closed form solutions of such problems promise to be more rigorous and computationally efficient than standard iterative schemes.

2 Background

For univariate or single valued random variables, we observe the outcome of one random experiment and map the result to one real number. In many problems, however, it is necessary to map the result of one random experiment to multiple random numbers. For example, we have p variables measured on N observations, which are p proteins measured on N cells and, in this case, we would use a $p \times N$ matrix to store these numbers.

There is another way to denote multivariate random variables by using graphical models [2, 3]. Graphical models are ways of representing the relationship between variables, with the nodes represent different variables. There are two main kinds of graphs, namely directed graphs and undirected graphs. Undirected graphs are a set of vertices or nodes that are conset of vertices or nodes that are connected together by edges, where all the edges are bidirectionalnected together by edges, where all the edges are bidirectional and such graphs are widely used to represent multivariate random variables, where they are also known as Graphical Models or Markov Random Fields(MRF). In such a graph, an edge between two

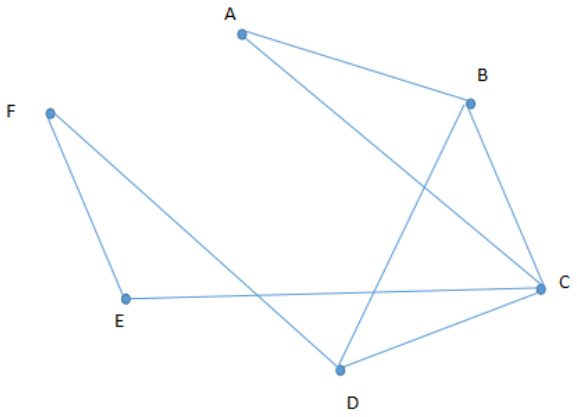

Figure 1: This is an undirected graph, nodes indicate a set of objects and all edges are undirected.

nodes implies these two variables are correlated even when conditions on the other variables. Such a graph is illustrated in Figure 1.

In recent years, many statisticians and researchers have proposed different ways to estimate sparse undirected graphical models [4]. A sparse undirected graph is also known as a conditional independence graph and, in this case, we delete the edge between nodes j and k if the variable $X(j)$ is independent of the variable $X(k)$ given the rest of the variables. Conditional independence graphs provide substantial information about their underlying random variables and the most widely used method to estimate such graphs is the Graphical Lasso [4, 5]. The Graphical Lasso method attempts to learn the structure of a Gaussian graphical model by maximizing the log likelihood of the data, subject to an l_1 penalty on elements of the information matrix (the inverse of the covariance matrix) [4]. As opposed to the Graphical Lasso, which uses an iterative scheme to estimate the maximum likelihood information matrix based upon the measured data balanced against the l_1 penalty, our work provide *closed from solutions* which can be leveraged for future theoretical work.

In this section, we will take a closer look at Undirected Graphical Models, the Hammersley-Clifford Theorem and Graphical Lasso. Most importantly, we will discuss how they related to each other in providing our desired closed form solutions.

2.1 Undirected Graphical Models

In discrete mathematics, especially in graph theory, a graph is a mathematical structure that represents the relationship between objects. Usually a graph $\mathbf{C}$ has two components, vertices and edges, which can be denoted as C and E. In particular, one edge $(x, y) \in E$ means $x, y \in C$ and the presence of such as edge in the graph means that the two corresponding random variables are conditionally dependent [2, 3].

One example of an undirected graph is given in Figure 1. Here we have six objects, denote as A, B, C, D, E, F six nodes in the graph. There are eight edges, which are $(A, B), (A, C), (B, C), (B, D), (C, D), (C, E), (D, F), (E, F)$, the edges here mean that there is some "relation" between these objects.

However, what is the "relation" represented by the edge set E? In other words, when can we put an edge between two vertices? The answer to this question gives the essential part of our research, which is the reason we focused on information matrices of multivariate random variables.

To start, some background knowledge and notations about graphical theory are given. An information matrix, also known as a concentration matrix or a precision matrix, is the inverse of the covariance matrix. Elements in the information matrix can be interpreted in terms of partial correlations. Under an assumption of Gaussianity, non-zero entries in the information matrix imply conditional dependence between corresponding variables given rest variables. This idea implies that, by identifying zero elements in the information matrix we can actually find the conditional independent relationship between variables. This "conditional independent relationship" is actually the "relation" represented by the edge set as we mentioned earlier.

Let's denote the covariance matrix of random variables as Σ, then the information matrix can be written as Σ^{-1}. If the element in the i^{th} row and j^{th} column of the information matrix Σ^{-1} is not zero, we put an edge between nodes i and j in the graph. So the connection between a graph $\mathbf{C}$ and the corresponding information matrix Σ^{-1} can be summarized as [2, 3]

$$\begin{aligned}
\Sigma^{-1}_{ij} = 0 \quad &\Leftrightarrow \text{there is no edge between node } i \text{ and } j. \\
&\Leftrightarrow \text{variable } i \text{ and } j \text{ are conditional independent given other variables.} \\
\Sigma^{-1}_{ij} \neq 0 \quad &\Leftrightarrow \text{there is an edge between node } i \text{ and } j \\
&\Leftrightarrow \text{variable } i \text{ and } j \text{ are conditional dependent.}
\end{aligned}$$

Herein, we follow the notation from [3] and focus on the specific mathematical elements important to our derivations.

As mentioned before, a simple undirected graph is denoted as $\mathbf{C} = (C, E(\mathbf{C}))$, for which C is the vertex set, and $E(\mathbf{C})$ is the edge set. The edge set $E(\mathbf{C})$ contains the unordered pairs of distinct vertices. If $\{\alpha, \beta\} \in E(\mathbf{C})$, then pairs of vertices $\{\alpha, \beta\}$ are said to be adjacent.

We define a *clique* as a maximal set of (≥ 2) vertices in which every pair is adjacent. For any vertex γ we write $\partial\gamma = \{\alpha : \{\alpha, \gamma\} \in E(\mathbf{C})\}$ for the set of neighbors of γ, by which α are the vertices that directly related to γ. We also have $\bar{\gamma} = \gamma \cup \partial\gamma$.

We can then define a *chain* as a sequence $\gamma = \gamma_0, \gamma_1, ..., \gamma_m = \beta$ of vertices such that $\{\gamma_l, \gamma_{l+1} \in E(\mathbf{C})\}$ for $l = 0, 1, ..., m-1$. If $\gamma_0 = \gamma_m$ the chain is called a *cycle*. Another important concept here is the separation of sets of vertices in $\mathbf{C}$. Two sets of vertices a, b are said to be separated by a third set d if every chain connecting an $\alpha \in a$ to a $\beta \in b$ intersects

Figure 2: An undirected graph and its complement have the same node sets but complementary edges set.

d.

The graph $\mathbf{C}$ is said to be triangulated [2] if and only if all cycles $\gamma_0, \gamma_1, ..., \gamma_p = \gamma_0$ of length $p \geq 4$ possess a chord, where a chord is an edge connecting two nonconsecutive vertices of the cycle.

The graph $\tilde{\mathbf{C}}$ is called the compliment of $\mathbf{C}$ when $\tilde{\mathbf{C}}$ has vertex set C and edge set $E(\tilde{\mathbf{C}})$ with the property that $\{\alpha, \beta\} \in E(\tilde{\mathbf{C}})$ if and only if $\alpha \neq \beta$ and $\{\alpha, \beta\} \notin E(\mathbf{C})$.
Example. The graph $\mathbf{C}$ with vertex set $\{1, 2, 3, 4\}$ and edge set $\{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{3, 4\}\}$ could be depicted as the first graph in Figure 2. The second graph in Figure 2 is the complement of graph $\mathbf{C}$. For this graph the set of neighbors of 1 is $\{2, 3, 4\}$; the cliques are $\{1, 2, 3\}, \{1, 3, 4\}$; a chain from $\{2\}$ to $\{4\}$ is 2, 3, 1, 4 and $\{2\}$ is separated from $\{4\}$ by $\{1, 3\}$.

Proposition 1 of [3] relate the conditional independence of a multivariate data set $\mathbf{X}$ to the structure of it its corresponding information matrix Σ^{-1}. In these propositions, following [3], we abbreviate the set intersection $a \cap b$ to ab and write a/b for the complement of b in a. The set C/b will be denoted b'. The characterization of all conditional independence relations consequent upon a given pattern of zeros in Σ^{-1} is:

Proposition 1. For subsets a, b of C with $a \cup b = C$ the following statements are equivalent.
(i) $\Sigma_{a,b} = \Sigma_{a,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b}$.
(i$'$) $\Sigma_{a/b,b/a} = \Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a}$.
(ii) $(\Sigma^{-1})_{a/b,b/a} = 0$
(iii) $\mathbf{X}_a$ and $\mathbf{X}_b$ are conditionally independent given $\mathbf{X}_{ab}$

A detailed proof of Proposition 1 is given in the Appendix.

Other important Corollaries for our work include

Corollary 1. For distinct elements α, β of C, X_α and X_β are conditionally independent given $X_{\{\alpha,\beta\}'}$ if and only if $\Sigma^{-1}(\alpha, \beta) = 0$.

Proof. Put $a = C\ \{\alpha\} = \{\alpha\}'$ and $b = \{\beta\}'$ in Proposition 1.

Proposition 1, provide the theoretical foundation of how the connectivity of an undirected graph are related to the entries in a corresponding information matrix. Figure 3 provide an

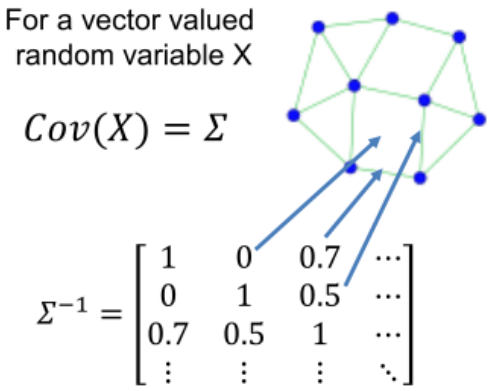

Figure 3: Zero elements of the information matrix means there is no edge between the corresponding nodes in a Markov Random Field, non-zero elements of the information matrix code for the neighborhood structure of a Markov Random Field according to the Hammersley-Clifford theorem.

example of how the zeros in Σ^{-1}, and the corresponding conditional independence relationships, can be encoded in an undirected graph.

2.2 Hammersley-Clifford Theorem and Markov Random Fields

Hammersley-Clifford theorem gives necessary and sufficient conditions under which a positive probability distribution can be represented as a Markov Random Field[7]. It is the fundamental theorem of random fields and the basis for the graphical modeling.

Consider a set of variables corresponding to the nodes of a particular graph. According to Hammersley-Clifford theorem [7], given an undirected graph $\mathbf{C} = (C, E(\mathbf{C}))$, random variables X form a Markov random field with respect to $\mathbf{C}$ if they satisfy the Markov property.

The pairwise Markov property:

$\Sigma^{-1}(\alpha, \beta) = 0$ if $\{\alpha, \beta\} \notin E(\mathbf{C})$ and $\alpha \neq \beta$;

The local Markov property:

For every $\gamma \in C, X_\gamma$ and $\mathbf{X}_{\{\gamma\}'}$ are conditionally independent given $\mathbf{X}_{\partial\gamma}$;

The global Markov property:

For every a, b and d with d separating a from b in $\mathbf{C}, \mathbf{X_a}$ and $\mathbf{X_b}$ are conditionally independent given $\mathbf{X_d}$.

Markov properties reveal connections between sparsity of an undirected graph, depen-

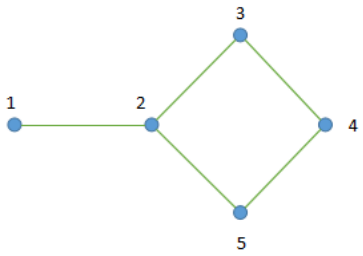

Figure 4: This figure shows an undirected graph corresponding to a given information matrix.

dency between Gaussian random variables and zero entries in information matrix. We will use an example from [3] illustrates Markov Properties and how we can use Hammersley - Clifford theorem to do graph separation.

Example. Suppose Σ^{-1} has the following pattern with $*$ denoting a nonzero element:

$$\begin{bmatrix} * & * & 0 & 0 & 0 \\ * & * & * & 0 & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & * & * & * & * \end{bmatrix}$$

Then the corresponding graph $\mathbf{C}$ would be shown in Figure 4. If we put $\gamma = \{2\}, \partial\gamma = \{1, 3, 5\}$, and use the local Markov property we deduce that X_2 and X_4 are conditionally independent given $\mathbf{X}_{\{1,3,5\}}$. Similarly with $a = \{1\}, b = \{4\}$, and $d = \{2\}$, the global Markov property can be used to assert that X_1 and X_4 are conditionally independent given X_2.

2.3 Graphical Lasso

Recently, there has been a lot of activity on the estimation of undirected graphical models using Graphical Lasso regularization [4, 5]. The Graphical Lasso is an algorithm for estimating a sparse the information matrix from observations of a multivariate Gaussian distribution. It is a widely used method for learning the structure of undirected graphs based on an l_1 regularization technique. By learning the sparsity pattern of the information matrix, Graphical Lasso can estimate the conditional independence between random variables [9].

Consider observations $X_1, X_2, ..., X_n$ from a multivariate Gaussian distribution:

$$X \sim N(\mu, \Sigma)$$

We want to estimate the information matrix $\Theta = \Sigma^{-1}$. The Graphical Lasso estimator is $\hat{\Theta}$ that maximizes the l_1 penalized log-likelihood [4, 5]:

$$\log \det\Theta - \mathrm{tr}(\Sigma_{XX}\Theta) - \rho \left\|\Theta\right\|_1$$

over non-negative definite matrices Θ

Here $\|\Theta\|_1$ is the L_1 norm, which is the sum of the absolute values of the elements of Σ^{-1}

Accordingly, we can write our desired optimization problem as:

$$\hat{\Theta} = \arg\min_{\Theta \geq 0} \left(\text{tr}(\Sigma_{XX}\Theta) - \log \det(\Theta) + \rho \sum |\Theta_{j,k}| \right)$$

Where Σ_{XX} is the sample covariance, and ρ is the penalizing parameter.

Herein, we can consider the problem of estimating sparse graphs by a Lasso penalty applied to the inverse covariance matrix and implementation over all positive, semi-definite, symmetric matrices [4]. However, there is no known closed form solution for Graphical Lasso [4, 5] and the problem is classically solved using an iterative numerical scheme. However, here we focus on producing closed form solution for information matrices under prespecified constraints.

3 Closed form Solution for Optimization

3.1 Lagrange Multipliers

In mathematical optimization, the method of Lagrange Multipliers is a strategy for finding the local maxima and minima of an objectuve function subject to constraints [10]. It is actually easier to explain the geometric basis of Lagrange multiplier for a function of two variables, so address such functions here.

For example, consider the optimization problem with two variables:

$$\text{maximize } f(x, y) \text{ subject to } g(x, y) = c$$

So we want to find the extreme values of $f(x, y)$ when the point (x, y) lie on the curve $g(x, y) = c$.

Figure 5 shows the line $g(x, y) = c$ together with several level curves of $f(x, y)$. These level curves are $f(x, y) = k$, where $k = 5, 7, 8, 9$. To maximize $f(x, y)$ subject to $g(x, y) = c$ is to fine the largest value of k such that the level curves of $f(x, y)$ intersect with $g(x, y) = c$. From Figure 5 we can see that it happens when the curve $g(x, y) = c$ tangent with the curve $f(x, y) = k$. In other words, when these two curves have a common tangent line. This means the gradient vectors are parallel at that point. So for some scalar λ, we have

$$\nabla f(x, y) = \lambda \nabla g(x, y) \tag{3.1}$$

Thus, we define the Lagrange function as

$$L(x, \lambda) = f(x) - \lambda g(x) \tag{3.2}$$

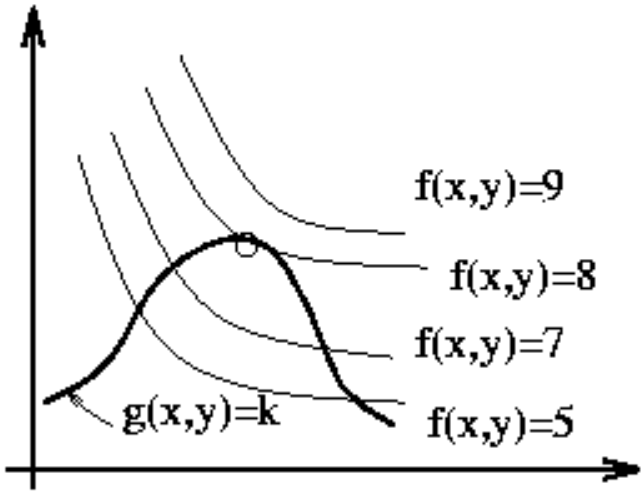

Figure 5: Geometric basis of Lagrange multiplier, $f(x, y) = c$ is the level curve and $g(x, y)$ is the function we want to maximize with constraint $g(x, y) = k$.

Optimal solution can be get by solving

$$\nabla_{xy,\lambda} L(x, \lambda) = 0 \tag{3.3}$$

If we let $f(x)$ is the log-likelihood function and $g(x)$ is the constraints for the information matrix, then our problems is

$$\begin{aligned} &\arg\max_{\Sigma} -\frac{n}{2} \ln \det(\Sigma) - \frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(X-\mu)(X-\mu)^H\right] \\ &\text{s.t. } (\Sigma^{-1})_{ij} = (\Sigma^{-1})_{ji} = 0 \end{aligned} \tag{3.4}$$

We can rewrite our constrained optimization problem as:

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2} \ln \det(\Sigma_{te}^{-1}) + \frac{1}{2} \operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t. } (\Sigma_{te}^{-1})_{ij} = 0 \end{aligned} \tag{3.5}$$

The Gaussian log-likelihood function is a concave function of the information matrix Σ^{-1}. Thus, maximum likelihood estimation in Gaussian models with linear constraints on the information matrix, as for Gaussian graphical models, is actually a convex optimization problem. We will give different closed form solution of the information matrix under different constrains as following sections show.

3.2 One pair of zero constraints

We begin our derivation by considering the simplest case first, namely that of on pair of 0-constraints. In particular, we assume that some oracle has indicate that a pair of specified

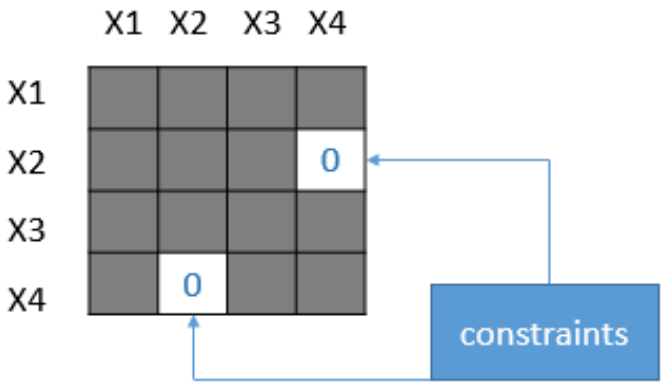

Figure 6: In this figure we show the first case we treat, namely a symmetric pair of 0 constraints in Σ^{-1}.

random variables are conditionally independent and derive closed form solutions to (3.5) that respect this constraint.

Theorem 3.1. If the constraints in Σ^{-1} are one pair of zeros, as in Figure 6, then the constrained optimization problem:

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t. } (\Sigma_{te}^{-1})_{ij} = 0 \end{aligned} \tag{3.6}$$

has an optimal information matrix Σ_{te}^{-1} which can be written as

$$\Sigma_{te}^{-1} = \left(\frac{1}{n}\left[(X-\mu_X)(X-\mu_X)^H\right] + \frac{2}{n}\lambda 1_i 1_j^T + \frac{2}{n}\lambda 1_j 1_i^T\right)^{-1} \tag{3.7}$$

where

$$\lambda = \frac{a}{2(bc-a^2)} \tag{3.8}$$

The elements a, b, c can be define as the elements in matrix

$$R = \left[(X-\mu_X)(X-\mu_X)^H\right] \tag{3.9}$$

where

$$R_{ij}^{-1} = R_{ji}^{-1} = a \tag{3.10}$$

$$R_{ii}^{-1} = c \tag{3.11}$$

$$R_{jj}^{-1} = b \tag{3.12}$$

Proof. Let us first fix some notation, namely let

$$\mathbb{1}_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \tag{3.13}$$

where the 1 is in the i-th position.

So

$$\mathbb{1}_i^T A \mathbb{1}_j \text{ selects the } A_{ij} \text{ entry of matrix } A.$$

Then our problem can be written as

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t. } \mathbb{1}_i^T\Sigma_{te}^{-1}\mathbb{1}_j = 0 \text{ and } \mathbb{1}_j^T\Sigma_{te}^{-1}\mathbb{1}_i = 0 \end{aligned} \tag{3.14}$$

We proceed using the method of Lagrange multipliers and rewrite our constrained optimization as an unconstrained optimization

$$\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = -\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] + \lambda_1\mathbb{1}_i^T\Sigma_{te}^{-1}\mathbb{1}_j + \lambda_2\mathbb{1}_j^T\Sigma_{te}^{-1}\mathbb{1}_i \tag{3.15}$$

By the first order partial derivative of a matrix[10, 11], we have :

$$\frac{\partial \ln|\det(X)|}{\partial X} = (X^{-1})^T \tag{3.16}$$

$$\frac{\partial \operatorname{tr}(XYY^T)}{\partial X} = \frac{\partial \operatorname{tr}(Y^TXY)}{\partial X} = YY^T \tag{3.17}$$

$$\frac{\partial a^TXb}{\partial X} = ab^T \tag{3.18}$$

Let us take some derivatives of our unconstrained optimization formulea as:

$$\begin{aligned} \frac{\partial}{\partial\Sigma_{te}^{-1}}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = \frac{\partial}{\partial\Sigma_{te}^{-1}}\Bigg\{ &-\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &+ \lambda_1\mathbb{1}_i^T\Sigma_{te}^{-1}\mathbb{1}_j + \lambda_2\mathbb{1}_j^T\Sigma_{te}^{-1}\mathbb{1}_i\Bigg\} \end{aligned} \tag{3.19}$$

Since Σ_{te}^{-1} is assumed to be symmetric positive definite (SPD) so

$$\frac{\partial \ln\det(\Sigma_{te}^{-1})}{\partial\Sigma_{te}^{-1}} = \frac{\partial \ln|\det(\Sigma_{te}^{-1})|}{\partial\Sigma_{te}^{-1}} = (\Sigma_{te})^T = \Sigma_{te} \tag{3.20}$$

$$\frac{\partial}{\partial \Sigma_{te}^{-1}} \mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = -\frac{n}{2}\Sigma_{te} + \frac{1}{2}\left[(X-\mu_X)(X-\mu_X)^H\right] + \lambda_1 \mathbb{1}_i \mathbb{1}_j^T + \lambda_2 \mathbb{1}_j \mathbb{1}_i^T. \tag{3.21}$$

Similarly,

$$\frac{\partial}{\partial \lambda_1} \mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = \mathbb{1}_i^T \Sigma_{te}^{-1} \mathbb{1}_j \tag{3.22}$$

$$\frac{\partial}{\partial \lambda_2} \mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = \mathbb{1}_j^T \Sigma_{te}^{-1} \mathbb{1}_i \tag{3.23}$$

Setting the derivatives equal to 0 and moving around terms we get a formula for Σ_{te}^{-1}

$$\Sigma_{te}^{-1} = \left(\frac{1}{n}\left[(X-\mu_X)(X-\mu_X)^H\right] + \frac{2}{n}\lambda_1 \mathbb{1}_i \mathbb{1}_j^T + \frac{2}{n}\lambda_2 \mathbb{1}_j \mathbb{1}_i^T\right)^{-1} \tag{3.24}$$

We can plug this into formulas for λ_1 and λ_2 and multiply out the constants and simplify notation slightly by setting

$$R = \left[(X-\mu_X)(X-\mu_X)^H\right] \tag{3.25}$$

to get

$$0 = \mathbb{1}_i^T \left(R + \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} \begin{bmatrix}2\lambda_1 & 0 \\ 0 & 2\lambda_2\end{bmatrix} \begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix}\right)^{-1} \mathbb{1}_j \tag{3.26}$$

$$0 = \mathbb{1}_j^T \left(R + \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} \begin{bmatrix}2\lambda_1 & 0 \\ 0 & 2\lambda_2\end{bmatrix} \begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix}\right)^{-1} \mathbb{1}_i \tag{3.27}$$

We apply the Woodbury identity [12]

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{3.28}$$

to the middle term givin rise to the two terms

$$0 = \mathbb{1}_i^T R^{-1} \mathbb{1}_j - \mathbb{1}_i^T R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} \left(\begin{bmatrix}\frac{1}{2\lambda_1} & 0 \\ 0 & \frac{1}{2\lambda_2}\end{bmatrix} + \begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix} R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\right)^{-1} \begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix} R^{-1} \mathbb{1}_j \tag{3.29}$$

$$0 = \mathbb{1}_j^T R^{-1} \mathbb{1}_i - \mathbb{1}_j^T R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} \left(\begin{bmatrix}\frac{1}{2\lambda_1} & 0 \\ 0 & \frac{1}{2\lambda_2}\end{bmatrix} + \begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix} R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\right)^{-1} \begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix} R^{-1} \mathbb{1}_i \tag{3.30}$$

Let

$$\begin{bmatrix}\mathbb{1}_j^T \\ \mathbb{1}_i^T\end{bmatrix} R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} = \begin{bmatrix}\mathbb{1}_j^T R^{-1} \mathbb{1}_i & \mathbb{1}_j^T R^{-1} \mathbb{1}_j \\ \mathbb{1}_i^T R^{-1} \mathbb{1}_i & \mathbb{1}_i^T R^{-1} \mathbb{1}_j\end{bmatrix} = \begin{bmatrix}a & b \\ c & a\end{bmatrix} \tag{3.31}$$

Where we leverage the symmetry of R and therefore the symmetry of R^{-1}. We have that

$$\mathbb{1}_j^T R^{-1} \mathbb{1}_i = \mathbb{1}_i^T R^{-1} \mathbb{1}_j = a \tag{3.32}$$

$$\mathbb{1}_i^T R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} = \begin{bmatrix}c & a\end{bmatrix} \tag{3.33}$$

$$\mathbb{1}_j^T R^{-1} \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix} = \begin{bmatrix}a & b\end{bmatrix} \tag{3.34}$$

$$\begin{bmatrix} \mathbb{1}_j^T \\ \mathbb{1}_i^T \end{bmatrix} R^{-1} \mathbb{1}_i = \begin{bmatrix} a \\ c \end{bmatrix} \tag{3.35}$$

$$\begin{bmatrix} \mathbb{1}_j^T \\ \mathbb{1}_i^T \end{bmatrix} R^{-1} \mathbb{1}_j = \begin{bmatrix} b \\ a \end{bmatrix} \tag{3.36}$$

$$\left(\begin{bmatrix} \frac{1}{2\lambda_1} & 0 \\ 0 & \frac{1}{2\lambda_2} \end{bmatrix} + \begin{bmatrix} \mathbb{1}_j^T \\ \mathbb{1}_i^T \end{bmatrix} R^{-1} \begin{bmatrix} \mathbb{1}_i & \mathbb{1}_j \end{bmatrix} \right)^{-1} = \left(\begin{bmatrix} a + \frac{1}{2\lambda_1} & b \\ c & a + \frac{1}{2\lambda_2} \end{bmatrix} \right)^{-1} \tag{3.37}$$

And our two terms are

$$0 = a - \begin{bmatrix} c & a \end{bmatrix} \left(\begin{bmatrix} a + \frac{1}{2\lambda_1} & b \\ c & a + \frac{1}{2\lambda_2} \end{bmatrix} \right)^{-1} \begin{bmatrix} b \\ a \end{bmatrix} \tag{3.38}$$

$$0 = a - \begin{bmatrix} a & b \end{bmatrix} \left(\begin{bmatrix} a + \frac{1}{2\lambda_1} & b \\ c & a + \frac{1}{2\lambda_2} \end{bmatrix} \right)^{-1} \begin{bmatrix} a \\ c \end{bmatrix} \tag{3.39}$$

Moving terms and expanding, we have

$$a^2 + \frac{a}{2\lambda_1} = bc \text{ and } a^2 + \frac{a}{2\lambda_2} = bc \tag{3.40}$$

Which means

$$\lambda_1 = \lambda_2 \tag{3.41}$$

And we can write both of them as λ. So we can write both of the equation as

$$a^2 + \frac{a}{2\lambda} = bc \tag{3.42}$$

by solving the equation, we have

$$\lambda = \frac{a}{2(bc - a^2)} \tag{3.43}$$

Which means the closed form solution of optimization problem with one pair of zeros constraints is

$$\Sigma_{te}^{-1} = \left(\frac{1}{n} \left[(X - \mu_X)(X - \mu_X)^H \right] + \frac{2}{n} \lambda_1 \mathbb{1}_i \mathbb{1}_j^T + \frac{2}{n} \lambda_2 \mathbb{1}_j \mathbb{1}_i^T \right)^{-1} \tag{3.44}$$

□

3.3 One pair of nonzero constraints

Theorem 3.2. If the constraints in Σ^{-1} are one pair of nonzeros, as in Figure 7, then we can state our constrained optimization problem as

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2} \ln\det(\Sigma_{te}^{-1}) + \frac{1}{2} \operatorname{tr}\left[\Sigma_{te}^{-1} (X - \mu_X)(X - \mu_X)^H \right] \\ &\text{s.t. } (\Sigma_{te}^{-1})_{ij} = (\Sigma^{-1})_{ij} \text{ and } (\Sigma_{te}^{-1})_{ji} = (\Sigma^{-1})_{ji} \end{aligned} \tag{3.45}$$

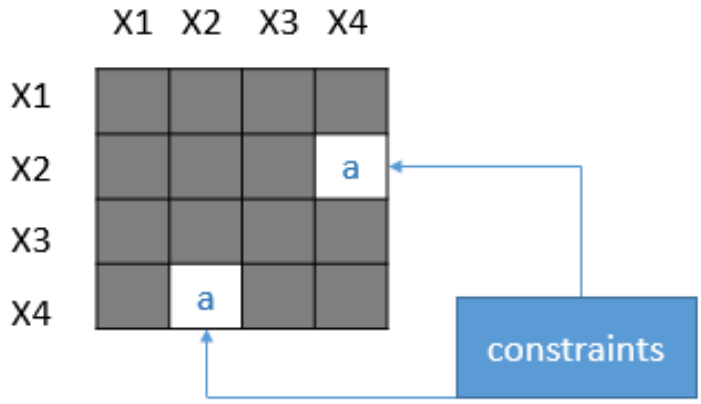

Figure 7: In this figure we show the second case we treat, namely a symmetric pair of nonzero constraints in Σ^{-1}.

Then the closed form solution of optimal information matrix Σ_{te}^{-1} is

$$\Sigma_{te}^{-1} = \left(\frac{1}{n} \left[(X - \mu_X)(X - \mu_X)^H \right] + \frac{2}{n} \lambda \mathbb{1}_i \mathbb{1}_j^T + \frac{2}{n} \lambda \mathbb{1}_j \mathbb{1}_i^T \right)^{-1} \tag{3.46}$$

where

$$\lambda = \frac{-1}{4d} - \frac{a}{2(a^2 - bc)} \pm \frac{1}{4} \sqrt{\frac{1}{d^2} + \frac{4bc}{(a^2 - bc)^2}} \tag{3.47}$$

a, b, c are elements in Equation 3.31, d is defined as

$$(n\Sigma)_{ij}^{-1} = d \tag{3.48}$$

Proof. Our optimization problem can be written as

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2} \ln\det(\Sigma_{te}^{-1}) + \frac{1}{2} \operatorname{tr}\left[\Sigma_{te}^{-1} (X - \mu_X)(X - \mu_X)^H \right] \\ &\text{s.t. } \mathbb{1}_i^T \Sigma_{te}^{-1} \mathbb{1}_j = \mathbb{1}_i^T \Sigma^{-1} \mathbb{1}_j \text{ and } \mathbb{1}_j^T \Sigma_{te}^{-1} \mathbb{1}_i = \mathbb{1}_j^T \Sigma^{-1} \mathbb{1}_i \end{aligned} \tag{3.49}$$

We proceed using the method of Lagrange multipliers and rewrite our constrained optimization as an unconstrained optimization

$$\begin{aligned} \mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = &-\frac{n}{2} \ln\det(\Sigma_{te}^{-1}) + \frac{1}{2} \operatorname{tr}\left[\Sigma_{te}^{-1} (X - \mu_X)(X - \mu_X)^H \right] \\ &+ \lambda_1 \mathbb{1}_i^T (\Sigma_{te}^{-1} - \Sigma^{-1}) \mathbb{1}_j + \lambda_2 \mathbb{1}_j^T (\Sigma_{te}^{-1} - \Sigma^{-1}) \mathbb{1}_i \end{aligned} \tag{3.50}$$

By the first order partial derivative of a matrix, we have

$$\frac{\partial}{\partial \Sigma_{te}^{-1}} \mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = -\frac{n}{2} \Sigma_{te} + \frac{1}{2} \left[(X - \mu_X)(X - \mu_X)^H \right] + \lambda_1 \mathbb{1}_i \mathbb{1}_j^T + \lambda_2 \mathbb{1}_j \mathbb{1}_i^T. \tag{3.51}$$

$$\frac{\partial}{\partial \lambda_1}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = \mathbb{1}_i^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_j \tag{3.52}$$

$$\frac{\partial}{\partial \lambda_2}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2) = \mathbb{1}_j^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_i \tag{3.53}$$

Setting the derivatives equal to 0 and moving around terms we get a formula for Σ_{te}^{-1}

$$\Sigma_{te}^{-1} = \left(\frac{1}{n}\left[(X-\mu_X)(X-\mu_X)^H\right] + \frac{2}{n}\lambda_1 \mathbb{1}_i\mathbb{1}_j^T + \frac{2}{n}\lambda_2 \mathbb{1}_j\mathbb{1}_i^T\right)^{-1} \tag{3.54}$$

We can plug this into the formulas for λ_1 and λ_2 and simplify notation slightly by Equation 3.25, we get

$$0 = \mathbb{1}_i^T\left\{\left(R + \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\begin{bmatrix}2\lambda_1 & 0\\ 0 & 2\lambda_2\end{bmatrix}\begin{bmatrix}\mathbb{1}_j^T\\ \mathbb{1}_i^T\end{bmatrix}\right)^{-1} - (n\Sigma)^{-1}\right\}\mathbb{1}_j \tag{3.55}$$

$$0 = \mathbb{1}_j^T\left\{\left(R + \begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\begin{bmatrix}2\lambda_1 & 0\\ 0 & 2\lambda_2\end{bmatrix}\begin{bmatrix}\mathbb{1}_j^T\\ \mathbb{1}_i^T\end{bmatrix}\right)^{-1} - (n\Sigma)^{-1}\right\}\mathbb{1}_i \tag{3.56}$$

Apply the Woodbury identity [12] in Equation 3.28 to the middle term we have

$$0 = \mathbb{1}_i^T R^{-1}\mathbb{1}_j - \mathbb{1}_i^T R^{-1}\begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\left(\begin{bmatrix}\frac{1}{2\lambda_1} & 0\\ 0 & \frac{1}{2\lambda_2}\end{bmatrix} + \begin{bmatrix}\mathbb{1}_j^T\\ \mathbb{1}_i^T\end{bmatrix}R^{-1}\begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbb{1}_j^T\\ \mathbb{1}_i^T\end{bmatrix}R^{-1}\mathbb{1}_j - \mathbb{1}_i^T(n\Sigma)^{-1}\mathbb{1}_j \tag{3.57}$$

$$0 = \mathbb{1}_j^T R^{-1}\mathbb{1}_i - \mathbb{1}_j^T R^{-1}\begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\left(\begin{bmatrix}\frac{1}{2\lambda_1} & 0\\ 0 & \frac{1}{2\lambda_2}\end{bmatrix} + \begin{bmatrix}\mathbb{1}_j^T\\ \mathbb{1}_i^T\end{bmatrix}R^{-1}\begin{bmatrix}\mathbb{1}_i & \mathbb{1}_j\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbb{1}_j^T\\ \mathbb{1}_i^T\end{bmatrix}R^{-1}\mathbb{1}_i - \mathbb{1}_j^T(n\Sigma)^{-1}\mathbb{1}_i \tag{3.58}$$

Following the notation in Equation 3.31, our two terms are

$$0 = a - \begin{bmatrix}c & a\end{bmatrix}\left(\begin{bmatrix}a+\frac{1}{2\lambda_1} & b\\ c & a+\frac{1}{2\lambda_2}\end{bmatrix}\right)^{-1}\begin{bmatrix}b\\ a\end{bmatrix} - (n\Sigma)_{ij}^{-1} \tag{3.59}$$

$$0 = a - \begin{bmatrix}a & b\end{bmatrix}\left(\begin{bmatrix}a+\frac{1}{2\lambda_1} & b\\ c & a+\frac{1}{2\lambda_2}\end{bmatrix}\right)^{-1}\begin{bmatrix}a\\ c\end{bmatrix} - (n\Sigma)_{ji}^{-1} \tag{3.60}$$

By moving terms and expanding, we get

$$\lambda_1 = \lambda_2 \tag{3.61}$$

Write both of them as λ. Let $(n\Sigma)_{ij}^{-1} = d$, solve λ as

$$\lambda = \frac{-1}{4d} - \frac{a}{2(a^2-bc)} \pm \frac{1}{4}\sqrt{\frac{1}{d^2} + \frac{4bc}{(a^2-bc)^2}} \tag{3.62}$$

Then our optimization problem with one pair of non-zeros constraints can be solved in closed form

$$\Sigma_{te}^{-1} = \left(\frac{1}{n}\left[(X-\mu_X)(X-\mu_X)^H\right] + \frac{2}{n}\lambda_1 \mathbb{1}_i\mathbb{1}_j^T + \frac{2}{n}\lambda_2 \mathbb{1}_j\mathbb{1}_i^T\right)^{-1} \tag{3.63}$$

□

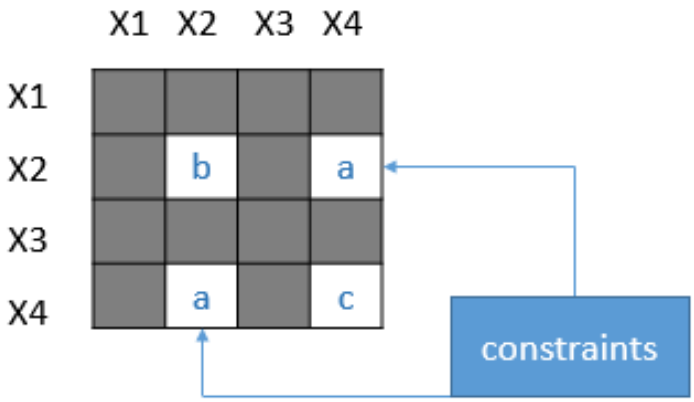

Figure 8: In this figure we show the third case we treat, namely two symmetric pairs of nonzero constraints in Σ^{-1}.

3.4 A 2×2 matrix constraints.

Let's define some notations first.

$$\mathbb{1}_{ij} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 1 & \vdots \\ \vdots & 1 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \tag{3.64}$$

where the 1 is in the i-th and j-th rows.

So

$$\mathbb{1}_{ij}^T A \mathbb{1}_{ij} \text{ selects } A_{ii}, A_{ij}, A_{ji}, A_{jj} \text{ entries of matrix } A.$$
$$\text{Which is a } 2 \times 2 \text{ matrix of } A.$$

Theorem 3.3. If the constraints in Σ^{-1} is a 2×2 matrix, as in Figure 8, then we can start our constrained optimization problem as:

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2} \ln\det(\Sigma_{te}^{-1}) + \frac{1}{2} \operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t.} \quad \mathbb{1}_{ij}^T \Sigma_{te}^{-1} \mathbb{1}_{ij} = \mathbb{1}_{ij}^T (\Sigma^{-1}) \mathbb{1}_{ij} \end{aligned} \tag{3.65}$$

Then our closed form solution is in the form

$$\Sigma_{te} = \Sigma_{XX} + \mathbb{1}_{ij} B \mathbb{1}_{ij}^T \tag{3.66}$$

Where

$$B = \left(\mathbb{1}_{ij}^T \Sigma^{-1} \mathbb{1}_{ij}\right)^{-1} - \left(\mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij}\right)^{-1} \tag{3.67}$$

Proof. We rewrite our optimization problem as

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t.} \quad \mathbb{1}_i^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_i = 0 \text{ and } \mathbb{1}_i^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_j = 0 \\ &\text{and } \mathbb{1}_j^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_i = 0 \text{ and } \mathbb{1}_j^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_j = 0 \end{aligned} \tag{3.68}$$

We proceed using the method of Lagrange multipliers and rewrite our constrained optimization as an unconstrained optimization

$$\begin{aligned} \mathcal{L}(\Sigma_{te}^{-1},\lambda_1,\lambda_2,\lambda_3,\lambda_4) = &-\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &+ \lambda_1\mathbb{1}_i^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_i + \lambda_2\mathbb{1}_i^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_j \\ &+ \lambda_3\mathbb{1}_j^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_i + \lambda_4\mathbb{1}_j^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_j \end{aligned} \tag{3.69}$$

By the first order partial derivative of a matrix, we have

$$\frac{\partial}{\partial\Sigma_{te}^{-1}}\mathcal{L}(\Sigma_{te}^{-1},\lambda_1,\lambda_2,\lambda_3,\lambda_4) = -\frac{n}{2}\Sigma_{te} + \frac{1}{2}\left[(X-\mu_X)(X-\mu_X)^H\right] + \lambda_1\mathbb{1}_i\mathbb{1}_i^T + \lambda_2\mathbb{1}_i\mathbb{1}_j^T + \lambda_3\mathbb{1}_j\mathbb{1}_i^T + \lambda_4\mathbb{1}_j\mathbb{1}_j^T. \tag{3.70}$$

$$\frac{\partial}{\partial\lambda_1}\mathcal{L}(\Sigma_{te}^{-1},\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \mathbb{1}_i^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_i \tag{3.71}$$

$$\frac{\partial}{\partial\lambda_2}\mathcal{L}(\Sigma_{te}^{-1},\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \mathbb{1}_i^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_j \tag{3.72}$$

$$\frac{\partial}{\partial\lambda_3}\mathcal{L}(\Sigma_{te}^{-1},\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \mathbb{1}_j^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_i \tag{3.73}$$

$$\frac{\partial}{\partial\lambda_4}\mathcal{L}(\Sigma_{te}^{-1},\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \mathbb{1}_j^T(\Sigma_{te}^{-1}-\Sigma^{-1})\mathbb{1}_j \tag{3.74}$$

Setting the derivatives equal to 0 and moving around terms we get the closed form solution as

$$\Sigma_{te} = \frac{1}{n}\left[(X-\mu_X)(X-\mu_X)^H\right] + \frac{2}{n}\left[\lambda_1\mathbb{1}_i\mathbb{1}_i^T + \lambda_2\mathbb{1}_i\mathbb{1}_j^T + \lambda_3\mathbb{1}_j\mathbb{1}_i^T + \lambda_4\mathbb{1}_j\mathbb{1}_j^T\right] \tag{3.75}$$

Which can be written as

$$\Sigma_{te} = \Sigma_{XX} + \mathbb{1}_{ij}B\mathbb{1}_{ij}^T \tag{3.76}$$

Where

$$B = \begin{bmatrix} \frac{2}{n}\lambda_1 & \frac{2}{n}\lambda_2 \\ \frac{2}{n}\lambda_3 & \frac{2}{n}\lambda_4 \end{bmatrix} \tag{3.77}$$

We can plug this into the formulas for λ_1, λ_2, λ_3 and λ_4 to get

$$0 = \mathbb{1}_i^T\left\{\left(\Sigma_{XX} + \mathbb{1}_{ij}B\mathbb{1}_{ij}^T\right)^{-1} - \Sigma^{-1}\right\}\mathbb{1}_i \tag{3.78}$$

$$0 = \mathbb{1}_i^T \left\{ \left(\Sigma_{XX} + \mathbb{1}_{ij} B \mathbb{1}_{ij}^T \right)^{-1} - \Sigma^{-1} \right\} \mathbb{1}_j \tag{3.79}$$

$$0 = \mathbb{1}_j^T \left\{ \left(\Sigma_{XX} + \mathbb{1}_{ij} B \mathbb{1}_{ij}^T \right)^{-1} - \Sigma^{-1} \right\} \mathbb{1}_i \tag{3.80}$$

$$0 = \mathbb{1}_j^T \left\{ \left(\Sigma_{XX} + \mathbb{1}_{ij} B \mathbb{1}_{ij}^T \right)^{-1} - \Sigma^{-1} \right\} \mathbb{1}_j \tag{3.81}$$

Which can be combined together and rewritten as:

$$\mathbb{1}_{ij}^T \left\{ \left(\Sigma_{XX} + \mathbb{1}_{ij} B \mathbb{1}_{ij}^T \right)^{-1} \right\} \mathbb{1}_{ij} = \mathbb{1}_{ij}^T \Sigma^{-1} \mathbb{1}_{ij} \tag{3.82}$$

Apply the Woodbury identity [12] in Equation 3.28 to the term

$$\left(\Sigma_{XX} + \mathbb{1}_{ij} B \mathbb{1}_{ij}^T \right)^{-1} \tag{3.83}$$

to get

$$\Sigma_{XX}^{-1} - \Sigma_{XX}^{-1} \mathbb{1}_{ij} \left(B^{-1} + \mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} \right)^{-1} \mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \tag{3.84}$$

Denote

$$\mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} = A \tag{3.85}$$

$$\mathbb{1}_{ij}^T \Sigma^{-1} \mathbb{1}_{ij} = C \tag{3.86}$$

Then the equation

$$\mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} - \mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} \left(B^{-1} + \mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} \right)^{-1} \mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} = \mathbb{1}_{ij}^T \Sigma^{-1} \mathbb{1}_{ij} \tag{3.87}$$

becomes

$$A - A(B^{-1} + A)^{-1} A = C \tag{3.88}$$

and can be written as

$$A^{-1}(A - C)A^{-1} = (B^{-1} + A)^{-1} \tag{3.89}$$

Taking the inverse of both side, we have

$$B^{-1} + A = [A^{-1}(A - C)A^{-1}]^{-1} = A(A - C)^{-1}A \tag{3.90}$$

So B cab be solved as

$$B = [-A + A(A - C)^{-1}A]^{-1} \tag{3.91}$$

Apply the Woodbury identity [12] in Equation 3.28 to the right hand side, we have

$$[-A+A(A-C)^{-1}A]^{-1} = -A^{-1}+A^{-1}A[(A-C)-AA^{-1}A]^{-1}A(-A^{-1}) = -A^{-1}-(-C)^{-1} = C^{-1}-A^{-1} \tag{3.92}$$

So we can write B as

$$B = C^{-1} - A^{-1} = \left(\mathbb{1}_{ij}^T \Sigma^{-1} \mathbb{1}_{ij} \right)^{-1} - \left(\mathbb{1}_{ij}^T \Sigma_{XX}^{-1} \mathbb{1}_{ij} \right)^{-1} \tag{3.93}$$

□

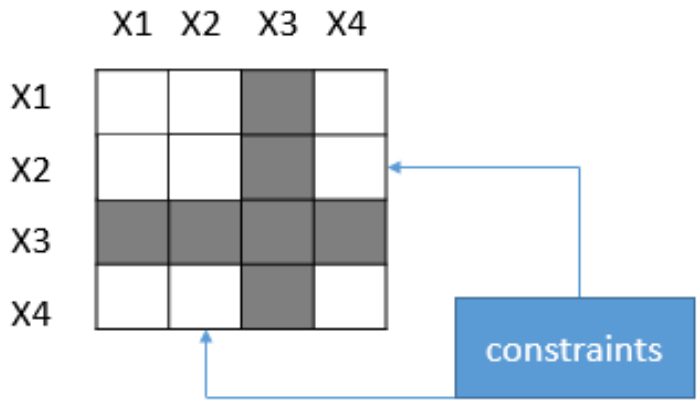

Figure 9: In this figure we show the final case we treat, namely a $k \times k$ principal submatrix of constraints in Σ^{-1}.

3.5 A $k \times k$ matrix of constraints.

We can now proceed to our most general results, a $k \times k$ principle sub-matrix of constraints. We begin with a small measure of notation.

First, denote

$$I_k = \{a_1, a_2, ..., a_k\} \tag{3.94}$$

If A is a $n \times n$ matrix, then

$$A[I_k, I_k] \tag{3.95}$$

is a $k \times k$ submatrix of A, which selected the row $\{a_1, a_2, ..., a_k\}$ and the column $\{a_1, a_2, ..., a_k\}$ of A.

Similarly, $A[\cdot, I_k]$ denote a $n \times k$ submatrix of A, which select every row but only $\{a_1, a_2, ..., a_k\}$ column of A.

Theorem 3.4. If the constraints in Σ^{-1} is a k×k principle sub-matrix, as in Figure 9. Then we can start our constrained optimization problem as:

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t.} \quad \Sigma_{te}^{-1}[I_k, I_k] = \Sigma^{-1}[I_k, I_k] \end{aligned} \tag{3.96}$$

So our closed form solution for Σ_{te} is

$$\Sigma_{te} = \Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T \tag{3.97}$$

Where

$$B = \left(\mathbb{1}_{I_k}^T \Sigma^{-1} \mathbb{1}_{I_k}\right)^{-1} - \left(\mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k}\right)^{-1} \tag{3.98}$$

Proof. Rewrite our optimization problem as

$$\begin{aligned} &\arg\min_{\Sigma_{te}^{-1}} -\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &\text{s.t.} \quad \text{for } \forall a_i, a_j \in I_k, \mathbb{1}_{a_i}^T \Sigma_{te}^{-1} \mathbb{1}_{a_j} = \mathbb{1}_{a_i}^T \Sigma^{-1} \mathbb{1}_{a_j} \end{aligned} \tag{3.99}$$

We proceed using the method of Lagrange multipliers and rewrite our constrained optimization as an unconstrained optimization

$$\begin{aligned} \mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2, ..., \lambda_{k^2}) = &-\frac{n}{2}\ln\det(\Sigma_{te}^{-1}) + \frac{1}{2}\operatorname{tr}\left[\Sigma_{te}^{-1}(X-\mu_X)(X-\mu_X)^H\right] \\ &+ \lambda_1 \mathbb{1}_{a_1}^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_{a_1} + \lambda_2 \mathbb{1}_{a_1}^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_{a_2} + ... + \lambda_{k^2}\mathbb{1}_{a_k}^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_{a_k} \end{aligned} \tag{3.100}$$

Taking derivatives we have

$$\frac{\partial}{\partial \Sigma_{te}^{-1}}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2, ..., \lambda_{k^2}) = -\frac{n}{2}\Sigma_{te} + \frac{1}{2}\left[(X-\mu_X)(X-\mu_X)^H\right] + \lambda_1 \mathbb{1}_{a_1}\mathbb{1}_{a_1}^T + \lambda_2 \mathbb{1}_{a_1}\mathbb{1}_{a_2}^T + ... + \lambda_{k^2}\mathbb{1}_{a_k}\mathbb{1}_{a_k}^T \tag{3.101}$$

$$\frac{\partial}{\partial \lambda_1}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2, ..., \lambda_{k^2}) = \mathbb{1}_{a_1}^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_{a_1} \tag{3.102}$$

$$\frac{\partial}{\partial \lambda_2}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2, ..., \lambda_{k^2}) = \mathbb{1}_{a_1}^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_{a_2} \tag{3.103}$$

$$\vdots$$

$$\frac{\partial}{\partial \lambda_{k^2}}\mathcal{L}(\Sigma_{te}^{-1}, \lambda_1, \lambda_2, ..., \lambda_{k^2}) = \mathbb{1}_{a_k}^T(\Sigma_{te}^{-1} - \Sigma^{-1})\mathbb{1}_{a_k} \tag{3.104}$$

Setting the derivatives equal to 0 and moving around terms we get the closed form solution as

$$\Sigma_{te} = \frac{1}{n}\left[(X-\mu_X)(X-\mu_X)^H\right] + \frac{2}{n}\left[\lambda_1 \mathbb{1}_{a_1}\mathbb{1}_{a_1}^T + \lambda_2 \mathbb{1}_{a_1}\mathbb{1}_{a_2}^T + ... + \lambda_{k^2}\mathbb{1}_{a_k}\mathbb{1}_{a_k}^T\right] \tag{3.105}$$

Which can be rewrite as

$$\Sigma_{te} = \Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T \tag{3.106}$$

Where

$$B = \frac{2}{n}\begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_k \\ \lambda_{k+1} & \lambda_{k+2} & \dots & \lambda_{k+k} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{k(k-1)+1} & \lambda_{k(k-1)+2} & \dots & \lambda_{k^2} \end{bmatrix} \tag{3.107}$$

We can plug this into the formulas for λ_1, λ_2, ..., λ_{k^2} to get

$$0 = \mathbb{1}_{a_1}^T\left\{(\Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T)^{-1} - \Sigma^{-1}\right\}\mathbb{1}_{a_1} \tag{3.108}$$

$$0 = \mathbb{1}_{a_1}^T\left\{(\Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T)^{-1} - \Sigma^{-1}\right\}\mathbb{1}_{a_2} \tag{3.109}$$

$$\vdots$$

$$0 = \mathbb{1}_{a_k}^T \left\{ (\Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T)^{-1} - \Sigma^{-1} \right\} \mathbb{1}_{a_k} \tag{3.110}$$

Which can be combined together and rewritten as:

$$\mathbb{1}_{I_k}^T \left\{ \left(\Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T \right)^{-1} \right\} \mathbb{1}_{I_k} = \mathbb{1}_{I_k}^T \Sigma^{-1} \mathbb{1}_{I_k} \tag{3.111}$$

Apply the Woodbury identity [12] in Equation 3.28 to the term

$$\left(\Sigma_{XX} + \mathbb{1}_{I_k} B \mathbb{1}_{I_k}^T \right)^{-1} \tag{3.112}$$

to get

$$\Sigma_{XX}^{-1} - \Sigma_{XX}^{-1} \mathbb{1}_{I_k} \left(B^{-1} + \mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} \right)^{-1} \mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \tag{3.113}$$

Plug it in to the previous equation:

$$\mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} - \mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} \left(B^{-1} + \mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} \right)^{-1} \mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} = \mathbb{1}_{I_k}^T \Sigma^{-1} \mathbb{1}_{I_k} \tag{3.114}$$

Denote

$$\mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} = A \tag{3.115}$$

$$\mathbb{1}_{I_k}^T \Sigma^{-1} \mathbb{1}_{I_k} = C \tag{3.116}$$

Then the equation is

$$A - A(B^{-1} + A)^{-1} A = C \tag{3.117}$$

which means

$$A^{-1}(A - C)A^{-1} = (B^{-1} + A)^{-1} \tag{3.118}$$

Taking the inverse of both side, we get

$$B^{-1} + A = [A^{-1}(A - C)A^{-1}]^{-1} = A(A - C)^{-1}A \tag{3.119}$$

So B can be solved as

$$B = [-A + A(A - C)^{-1}A]^{-1} \tag{3.120}$$

Apply the Woodbury identity [12] in Equation 3.28 to the right hand side, we have

$$[-A+A(A-C)^{-1}A]^{-1} = -A^{-1}+A^{-1}A[(A-C)-AA^{-1}A]^{-1}A(-A^{-1}) = -A^{-1}-(-C)^{-1} = C^{-1}-A^{-1} \tag{3.121}$$

Thus

$$B = \left(\mathbb{1}_{I_k}^T \Sigma^{-1} \mathbb{1}_{I_k} \right)^{-1} - \left(\mathbb{1}_{I_k}^T \Sigma_{XX}^{-1} \mathbb{1}_{I_k} \right)^{-1} \tag{3.122}$$

□

4 Numerical Experiments

Cross-validation is a primary way of measuring the predictive performance of a statistic or a statistical model. To evaluate the performance of our closed form estimator, a cross-validation procedure was conducted. Also, to compare the accuracy of optimal information matrix we get from our closed form solution and the estimator obtained from the semidefinite procedure we used standard convex optimization solvers [8].

As a computational conveniennce, we randomly generate a 4×4 positive definite symmetric matrix as our covariance matrix Σ. This Σ is used to generate our multivariate data set X. By doing this, we obtain a random sample which follows the multinormal distribution. Thus we have

$$X \sim N(\Sigma, \mathbf{0})$$

With this data set X, we can easily compute the sample covariance Σ_{XX}. Which is

$$\Sigma_{XX} = (X - \mu_{XX})(X - \mu_{XX})^H \tag{4.123}$$

Remember we have the closed form formula for optimal information matrix under different constraints, the Σ_{te}^{-1}. We can plug Σ_{XX}^{-1} into Σ_{te}^{-1}'s expression to get the value.

Another way to obtain optimal solution for information matrix is by running the convex solver in the scripting language Python [13] since our problem can be phrase as a convex optimization problem. So comparison of these two solutions, Σ_{te}^{-1} from closed form solution and Σ_t^{-1} from convex solver, can be easily done by looking at numbers in each entry and by comparing the value of likelihood [8].

We will use the most general case - *$k \times k$ non-zeros constraints* - to illustrate the problem. The constraints here is a 2×2 submatrix ($k = 2$) of the information matrix, which are the $(2,2), (2,3), (3,2), (3,3)$ entries of Σ.

By forcing the corresponding entries in Σ_t and Σ_{te} to equal these constraints, we obtained results of the numerical experiments as following:

$$\Sigma^{-1} = \begin{bmatrix} 2.385 & -2.793 & -0.821 & -1.955 \\ -2.793 & \mathbf{4.444} & \mathbf{0.779} & 3.075 \\ -0.821 & \mathbf{0.779} & \mathbf{0.565} & 0.343 \\ -1.955 & 3.075 & 0.343 & 2.494 \end{bmatrix} \tag{4.124}$$

$$\Sigma_{XX}^{-1} = \begin{bmatrix} 1.657 & -1.582 & -0.75 & -0.521 \\ -1.582 & \mathbf{2.835} & \mathbf{0.465} & 1.776 \\ -0.75 & \mathbf{0.465} & \mathbf{0.954} & -0.614 \\ -0.521 & 1.776 & -0.614 & 2.478 \end{bmatrix} \tag{4.125}$$

$$\Sigma_t^{-1} = \begin{bmatrix} 1.815 & -2.219 & -0.634 & -1.097 \\ -2.219 & \mathbf{4.166} & \mathbf{0.73} & 2.416 \\ -0.634 & \mathbf{0.73} & \mathbf{0.53} & 0.032 \\ -1.097 & 2.416 & 0.032 & 2.206 \end{bmatrix} \tag{4.126}$$

$$\Sigma_{te}^{-1} = \begin{bmatrix} 2.049 & -2.507 & -0.679 & -1.33 \\ -2.507 & \mathbf{4.444} & \mathbf{0.779} & 2.733 \\ -0.679 & \mathbf{0.779} & \mathbf{0.565} & 0.036 \\ -1.33 & 2.733 & 0.036 & 2.569 \end{bmatrix} \tag{4.127}$$

We can see that both Σ_{te}^{-1} and Σ_t^{-1} are quite close to the true information matrix Σ^{-1}, but the submatrix of Σ_{te} is exactly the same as those of Σ. Which means the optimal information matrix we got from the closed form solution is better than what we got from convex solver [13]. Examining the likelihood values produced by the various methods show the same story:

Estimator	Likelihood
Σ^{-1}	-43.1373680097
Σ_{XX}^{-1}	-33.9670608437
Σ_t^{-1}	-36.4135128746
Σ_{te}^{-1}	-36.4043062778

Table 1: The likelihoods of each of the estimate covariance matrices.

From the value of likelihoods, we can see that our closed form solution is more rigorous than the experiment solution from convex solver since it obtains a larger likelihood. As we expect the sample information matrix Σ_{XX} shows the largest likelihood since it is calculated from the real data. By repeating the experiment, our closed form solution keeps showing a larger likelihood than the solution from the convex solver and our solution is better than the solution from the convex solver.

4.1 Test Statistics

Our work has been inspired by a particular class of test statistics for hypothesis tests between sample and postulated exact covariance matrices[1]. These results state that if S is either the true covariance Σ (perfect side information) or a sample covariance Σ_{XX} (no side information)

then the statistic γ is independent of the true covariance Σ where

$$\gamma(\Sigma^{-1}S) = \frac{det(\Sigma^{-1/2}S\Sigma^{-1/2})}{[\frac{1}{M}tr(\Sigma^{-1/2}S\Sigma^{-1/2})]^M} \tag{4.128}$$

which can be proved by using properties in [14]. In the future, we want to use our closed form solutions to see whether we can construct appropriate null-distributions to cross-validate the given side information since the following figure shows that the distribution of γ is invariant of the optimal information matrix as it does for the true information matrix.

The reason we brought up statistic γ here is that we want to use the distribution of γ to validate our closed form solution again. What we did is to use the convex solver in Python [13] to get optimal information matrices Σ_t^{-1}, then plug them into the γ function to get the γ value for each data set, which is

$$\gamma(\Sigma_t^{-1}\Sigma_{XX}) = \frac{det(\Sigma_t^{-1/2}\Sigma_{XX}\Sigma_t^{-1/2})}{[\frac{1}{M}tr(\Sigma_t^{-1/2}\Sigma_{XX}\Sigma_t^{-1/2})]^M} \tag{4.129}$$

For the same data set we get the Σ_{te}^{-1} using the closed form solution this time, and also plug it in to the γ function to get the γ values as

$$\gamma(\Sigma_{te}^{-1}\Sigma_{XX}) = \frac{det(\Sigma_{te}^{-1/2}\Sigma_{XX}\Sigma_{te}^{-1/2})}{[\frac{1}{M}tr(\Sigma_{te}^{-1/2}\Sigma_{XX}\Sigma_{te}^{-1/2})]^M} \tag{4.130}$$

Figure 10 shows the histograms for both γs. We can see that these two histogram overlap with each other, and the Kolmogorov-Smirnov test gave us a very high p-value, which is actually close to 0.999999999, which means that our closed form solution is valid, and is almost the same as what Python optimal solver gave us.

5 Future Work

There are many future directions where one could proceed based upon the foundation of our research. Some detailed proof work is still needed since some of the numerical results support conjectures about the invariance for the test statistic γ. In this section we provide some preliminary theoretical work we were already done and conjectures for possible future work.

5.1 Proofs of results in [1]

Before proving results in [1], we will give some lemmas which are used in the following proofs.

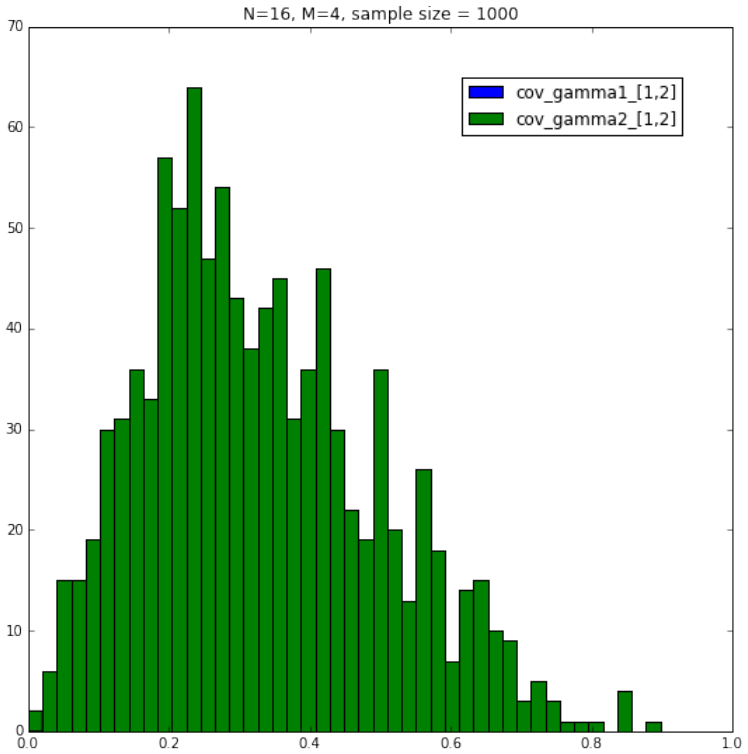

Figure 10: Histograms of statistics γ. The green one is the histogram of γ we calculated using our closed form solution Σ_{te}^{-1}, the blue one is the histogram of γ we calculated using optimal information matrix Σ_t^{-1} given by Python optimal solver. We can see that the green one and the blue one are overlap with each other.

Lemma 1. For a matrix B, if B^{-1} exists then

$$\det\,(B^{-1}AB) = \det\,(A) \tag{5.131}$$

Proof.

$$\begin{aligned}\det\,(B^{-1}AB) &= \det\,(B^{-1})\det\,(A)\det\,(B)\\ &= \big(\det\,(B^{-1})\det\,(B)\big)\det\,(A)\\ &= \det\,(A)\end{aligned} \tag{5.132}$$

□

Lemma 2. If Λ is a diagonal matrix and Λ^{-1} exists, then

$$\text{trace}\,\left(\Lambda^{-1}A\Lambda\right) = \text{trace}\,(A) \tag{5.133}$$

Proof. Since Λ is diagonal matrix, then

$$\left(\Lambda^{-1}A\Lambda\right)_{ii} = \Lambda_{ii}^{-1}A_{ii}\Lambda_{ii} = \Lambda_{ii}^{-1}\Lambda_{ii}A_{ii} = A_{ii} \tag{5.134}$$

so the diagonal of $(\Lambda^{-1}A\Lambda)$ is the same as the diagonal of A. Thus,

$$\text{trace}\,\left(\Lambda^{-1}A\Lambda\right) = \text{trace}\,(A) \tag{5.135}$$

□

Lemma 3. If U is unitary then λ is an eigenvalue of A if and only if λ is an eigenvalue of UAU^T.

Proof. Let λ be the eigenvalue of A with corresponding eigenvector x. Then

$$Ax = \lambda x \tag{5.136}$$

$$AIx = \lambda x \tag{5.137}$$

$$AU^TUx = \lambda x \tag{5.138}$$

$$UAU^TUx = \lambda Ux \tag{5.139}$$

$$UAU^T(Ux) = \lambda(Ux) \tag{5.140}$$

So λ is the eigenvalue of UAU^T with corresponding eigenvector Ux. □

Let $\Sigma \in \mathbb{R}^{N\times N}$ be semi-symmetric positive definite (SSPD) matrix. We can write $\Sigma = U\Lambda\Lambda U^T$ where U is unitary and Λ is diagonal matrix by singular value decomposition (SVD).

Since $X \sim N(0,\Sigma)$ and $X \in \mathbb{R}^{N\ \times M}$ we know that $\Sigma_{XX} = XX^T$ has the property that

$$\Sigma_{XX} = U\Lambda\Theta_{XX}\Lambda U^T \tag{5.141}$$

where $\Theta_{XX} \sim W(N,M,I)$

Then we can show results in [1] as following:

Theorem 5.1. Let $X \sim N(0, \Sigma)$ and Σ_{XX} be the sample covariance then the distribution of the statistic γ where

$$\gamma = \frac{\det\ (\Sigma^{-1/2}\Sigma_{XX}\Sigma^{-1/2})}{\left[\frac{1}{M}\ \text{trace}\ (\Sigma^{-1/2}\Sigma_{XX}\Sigma^{-1/2})\right]^M} \tag{5.142}$$

is independent of Σ.

Proof. Since

$$\det\ (\Sigma^{-1/2}\Sigma_{XX}\Sigma^{-1/2}) = \det\ (\Sigma^{-1/2}\Sigma^{-1/2}\Sigma_{XX}) = \det\ \Sigma^{-1}\Sigma_{XX} \tag{5.143}$$

$$\text{trace}\ (\Sigma^{-1/2}\Sigma_{XX}\Sigma^{-1/2}) = \text{trace}\ (\Sigma^{-1/2}\Sigma^{-1/2}\Sigma_{XX}) = \text{trace}\ \Sigma^{-1}\Sigma_{XX} \tag{5.144}$$

So we can change the theorem to prove that the distribution of

$$\gamma(\Sigma^{-1}\Sigma_{XX}) = \frac{\det\ (\Sigma^{-1}\Sigma_{XX})}{\left[\frac{1}{M}\ \text{trace}\ (\Sigma^{-1}\Sigma_{XX})\right]^M} \tag{5.145}$$

is independet of Σ.

We can write

$$\begin{aligned}\Sigma^{-1}\Sigma_{XX} &= U\Lambda^{-1}\Lambda^{-1}U^T U\Lambda\Theta_{XX}\Lambda U^T \\ &= U\Lambda^{-1}\Lambda^{-1}\Lambda\Theta_{XX}\Lambda U^T \\ &= U\Lambda^{-1}\Theta_{XX}\Lambda U^T\end{aligned} \tag{5.146}$$

Since U is unitary and Λ is diagonal, so

$$\det\ (U\Lambda^{-1}\Theta_{XX}\Lambda U^T) = \det\ \Theta_{XX} \tag{5.147}$$

$$\text{trace}\ (U\Lambda^{-1}\Theta_{XX}\Lambda U^T) = \text{trace}\ (\Theta_{XX}) \tag{5.148}$$

Which are both independent of Σ. □

Let's do some numerical experiments to shown what we just proved resultin [1]. The following results are got from two data sets, we can see clearly that even though the Σ and Σ_{XX} are not the same, but we end up with same determinate and trace of $\Sigma^{-1/2}\Sigma_{XX}\Sigma^{-1/2}$.

$$\Sigma_1 = \begin{bmatrix} 3.791 & 0.034 & 4.016 & 2.377 \\ 0.034 & 2.45 & -1.644 & -2.769 \\ 4.016 & -1.644 & 7.336 & 4.166 \\ 2.377 & -2.769 & 4.166 & 5.106 \end{bmatrix} \tag{5.149}$$

$$\Sigma_{XX1} = \begin{bmatrix} 33.736 & 0.566 & 30.602 & 23.039 \\ 0.566 & 13.787 & -3.781 & -14.429 \\ 30.602 & -3.781 & 61.261 & 16.291 \\ 23.039 & -14.429 & 16.291 & 35.244 \end{bmatrix} \tag{5.150}$$

$$\det_1 = 1936.86418972 \tag{5.151}$$

$$\text{trace}_1 = 37.2879281384 \tag{5.152}$$

$$\Sigma_2 = \begin{bmatrix} 1.162 & -0.047 & 1.071 & -0.283 \\ -0.047 & 5.849 & -0.495 & 1.28 \\ 1.071 & -0.495 & 1.968 & -1.075 \\ -0.283 & 1.28 & -1.075 & 2.195 \end{bmatrix} \tag{5.153}$$

$$\Sigma_{XX2} = \begin{bmatrix} 17.113 & -1.004 & 15.666 & 3.516 \\ -1.004 & 33.297 & -9.446 & 11.889 \\ 15.666 & -9.446 & 19.079 & -4.307 \\ 3.516 & 11.889 & -4.307 & 23.999 \end{bmatrix} \tag{5.154}$$

$$\det_2 = 1936.86418972 \tag{5.155}$$

$$\text{trace}_2 = 37.2879281384 \tag{5.156}$$

Also, we can show the following result:

Theorem 5.2. Let X, Y be two data set which following the same distribution as $X, Y \sim N(0, \Sigma)$ with sample covariance Σ_{XX} and Σ_{YY} respectively. Then the distribution of γ where

$$\gamma = \frac{\det\left(\Sigma_{YY}^{-1/2}\Sigma_{XX}\Sigma_{YY}^{-1/2}\right)}{\left[\frac{1}{M}\ \text{trace}\left(\Sigma_{YY}^{-1/2}\Sigma_{XX}\Sigma_{YY}^{-1/2}\right)\right]^M} \tag{5.157}$$

is independent of Σ.

Proof. Because of the same reason in the previous theorem, we can change our theorem to prove

$$\gamma(\Sigma_{YY}^{-1}\Sigma_{XX}) = \frac{\det\left(\Sigma_{YY}^{-1}\Sigma_{XX}\right)}{\left[\frac{1}{M}\ \text{trace}\left(\Sigma_{YY}^{-1}\Sigma_{XX}\right)\right]^M} \tag{5.158}$$

is independent of Σ. We can write

$$\begin{aligned} \Sigma_{YY}^{-1}\Sigma_{XX} &= U\Lambda^{-1}\Theta_{YY}^{-1}\Lambda^{-1}U^T U\Lambda\Theta_{XX}\Lambda U^T \\ &= U\Lambda^{-1}\Theta_{YY}^{-1}\Lambda^{-1}\Lambda\Theta_{XX}\Lambda U^T \\ &= U\Lambda^{-1}\Theta_{YY}^{-1}\Theta_{XX}\Lambda U^T \end{aligned} \tag{5.159}$$

Since U is unitary and Λ is diagonal so

$$\det\left(U\Lambda^{-1}\Theta_{YY}^{-1}\Theta_{XX}\Lambda U^T\right) = \det\left(\Theta_{YY}^{-1}\Theta_{XX}\right) \tag{5.160}$$

$$\text{trace}\left(U\Lambda^{-1}\Theta_{YY}^{-1}\Theta_{XX}\Lambda U^T\right) = \text{trace}\left(\Theta_{YY}^{-1}\Theta_{XX}\right) \tag{5.161}$$

which are both independent of Σ. □

Here are some numerical results to shown what we just proved. We can see clearly that even though the Σ, Σ_{XX} and $\Sigma_{YY}$are not the same, but we end up with same determinate and trace of $\Sigma_{YY}^{-1/2}\Sigma_{XX}\Sigma_{YY}^{-1/2}$.

$$\Sigma_1 = \begin{bmatrix} 3.791 & 0.034 & 4.016 & 2.377 \\ 0.034 & 2.45 & -1.644 & -2.769 \\ 4.016 & -1.644 & 7.336 & 4.166 \\ 2.377 & -2.769 & 4.166 & 5.106 \end{bmatrix} \tag{5.162}$$

$$\Sigma_{XX1}^{-1} = \begin{bmatrix} 33.736 & 0.566 & 30.602 & 23.039 \\ 0.566 & 13.787 & -3.781 & -14.429 \\ 30.602 & -3.781 & 61.261 & 16.291 \\ 23.039 & -14.429 & 16.291 & 35.244 \end{bmatrix} \tag{5.163}$$

$$\Sigma_{YY1}^{-1} = \begin{bmatrix} 11.742 & -0.368 & 9.68 & 9.368 \\ -0.368 & 9.82 & -4.513 & -11.874 \\ 9.68 & -4.513 & 23.837 & 9.493 \\ 9.368 & -11.874 & 9.493 & 21.944 \end{bmatrix} \tag{5.164}$$

$$\det{}_1 = 0.00815864972649 \tag{5.165}$$

$$\text{trace}_1 = 1.85334124231 \tag{5.166}$$

$$\Sigma_2 = \begin{bmatrix} 1.162 & -0.047 & 1.071 & -0.283 \\ -0.047 & 5.849 & -0.495 & 1.28 \\ 1.071 & -0.495 & 1.968 & -1.075 \\ -0.283 & 1.28 & -1.075 & 2.195 \end{bmatrix} \tag{5.167}$$

$$\Sigma_{XX2} = \begin{bmatrix} 17.113 & -1.004 & 15.666 & 3.516 \\ -1.004 & 33.297 & -9.446 & 11.889 \\ 15.666 & -9.446 & 19.079 & -4.307 \\ 3.516 & 11.889 & -4.307 & 23.999 \end{bmatrix} \tag{5.168}$$

$$\Sigma_{YY2} = \begin{bmatrix} 7.858 & 1.841 & 6.915 & 2.208 \\ 1.841 & 19.214 & 0.113 & 3.815 \\ 6.915 & 0.113 & 7.217 & -0.997 \\ 2.208 & 3.815 & -0.997 & 9.178 \end{bmatrix} \tag{5.169}$$

$$\det{}_2 = 0.00815864972649 \tag{5.170}$$

$$\text{trace}_2 = 1.85334124231 \tag{5.171}$$

5.2 Generalization the Invariance of Σ^{-1} to the Invariance of Σ_{te}^{-1}

The question here is that since we already proved $\gamma(\Sigma^{-1}\Sigma_{XX})$ and $\gamma(\Sigma_{YY}^{-1}\Sigma_{XX})$ are independent of Σ, is the statistic $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$ also independent of Σ?

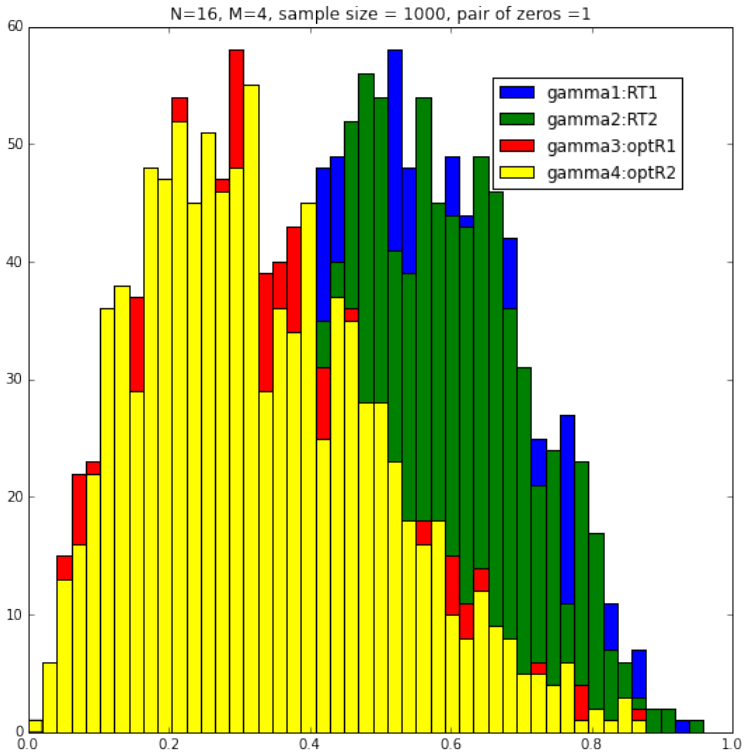

Figure 11: In this figure, the blue and green histograms are $\gamma(\Sigma_1^{-1}\Sigma_{XX1})$ and $\gamma(\Sigma_2^{-1}\Sigma_{XX2})$ respectively. The red and yellow histograms are $\gamma(\Sigma_{te1}^{-1}\Sigma_{XX1})$ and $\gamma(\Sigma_{te2}^{-1}\Sigma_{XX2})$ respectively. The constraints are one pair of zeros.

Before providing details conjectures, we would give the numerical experiment results first. What we did is using two different true population covariance to generate two sets of data, then using these data to get two sets of optimal information matrix Σ_{te}^{-1} using our closed form solution. We then plug these values into the $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$ statistic function to get two sets of $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$s. We can then plot the histogram of these $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$s. Here the constraints we used to calculate $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$ is one pair of zeros, which is

$$\Sigma_{ij} = \Sigma_{ji} = 0$$

Below is is the figure of distribution of γs.

In the figure above, the blue and green histograms are $\gamma(\Sigma_1^{-1}\Sigma_{XX1})$ and $\gamma(\Sigma_2^{-1}\Sigma_{XX2})$ respectively. The Kolmogorov-Smirnov test gave us a p-value:

blue-green p-value 0.822841332825

Which verified our earlier statement that the statistic $\gamma(\Sigma^{-1}\Sigma_{XX})$ is invariant of the true population covariance Σ.

In the figure above, the red and yellow histograms are $\gamma(\Sigma_{te1}^{-1}\Sigma_{XX1})$ and $\gamma(\Sigma_{te2}^{-1}\Sigma_{XX2})$ respectively, they are also overlap with each other very well. The Kolmogorov-Smirnov test

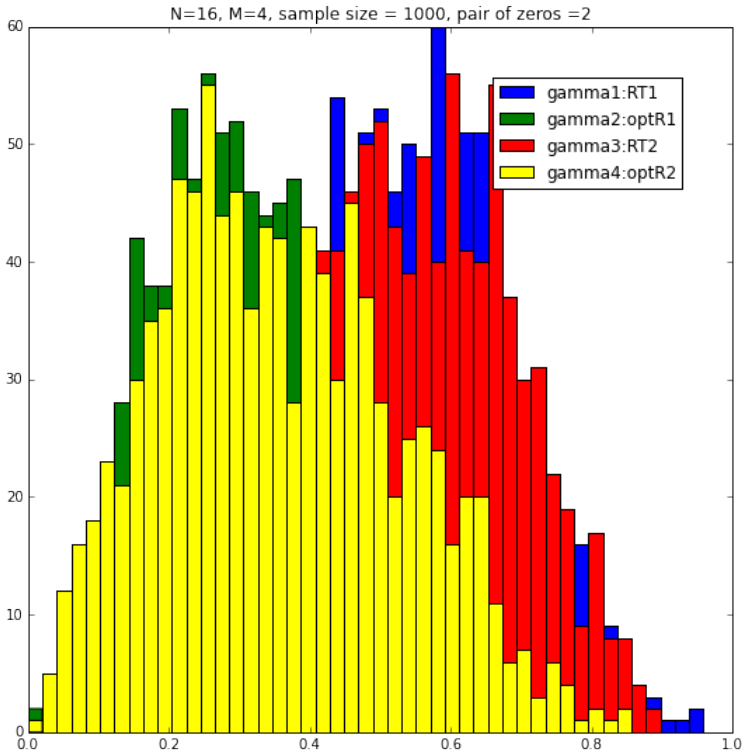

Figure 12: In this figure, the blue and red histograms are $\gamma(\Sigma_1^{-1}\Sigma_{XX1})$ and $\gamma(\Sigma_2^{-1}\Sigma_{XX2})$ respectively. The green and yellow histograms are $\gamma(\Sigma_{te1}^{-1}\Sigma_{XX1})$ and $\gamma(\Sigma_{te2}^{-1}\Sigma_{XX2})$ respectively. The constraints are two pairs of zeros.

gave us a p-value:

red-yellow p-value 0.951890168048

The numerical result above seem to support our conjecture that the statistic $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$ is also invariant of Σ.

Let's give another numerical results by changing the constraints to two pair of zeros, which are

$$\Sigma_{ij} = \Sigma_{ji} = 0 \text{ and } \Sigma_{kl} = \Sigma_{lk} = 0$$

Below is is the figure of distribution of $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$s combined with $\gamma(\Sigma^{-1}\Sigma_{XX})$s.

And the Kolmogorov-Smirnov test p-values are:

blue-red p-value 0.911041018381

green-yellow p-value 0.861365206773

The conclusions we draw from the p-value here are the same as in the previous example.

Thus, we want to prove that the statistic $\gamma(\Sigma_{te}^{-1}\Sigma_{XX})$ is also invariant of Σ, however a detailed proof is beyond the scope of the current work.

For further study, let us make a few definitions and more conjectures here.

Definition 1. Two symmetric matrices A and B are *relabeling* if there exists a permutation matrix P such that $A = PBP^T$.

The following two matrices are relabeling.

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 3 & 4 \\ 0 & 4 & 5 \end{bmatrix} \; B = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 1 & 0 \\ 4 & 0 & 5 \end{bmatrix} \tag{5.172}$$

with the permutation

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.173}$$

Definition 2. Two symmetric matrices A and B are 0-*permutations* if there exists a permutation matrix P such that $C = PBP^T$ and $A_{ij} = 0$ if, and only if, $C_{ij} = 0$.

For example, the matrices below are 0-permutations

$$A = \begin{bmatrix} 1 & 2 & 0 & 6 \\ 2 & 3 & 4 & 0 \\ 0 & 4 & 5 & 7 \\ 6 & 0 & 7 & 8 \end{bmatrix} \; B = \begin{bmatrix} 5 & 8 & 1 & 0 \\ 8 & 3 & 0 & 2 \\ 1 & 0 & 5 & 3 \\ 0 & 2 & 3 & 4 \end{bmatrix} \tag{5.174}$$

with the permutation

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{5.175}$$

Definition 3. Two symmetric matrices A and B are 0-*count equivalent* if the number of 0 entries in A is the same as the number of 0 entries in B.

Note, the following two matrices are *not* 0-permutations, but the are 0-count equivalent.

$$A = \begin{bmatrix} 1 & 2 & 0 & 6 \\ 2 & 3 & 4 & 0 \\ 0 & 4 & 5 & 7 \\ 6 & 0 & 7 & 8 \end{bmatrix} \; B = \begin{bmatrix} 5 & 8 & 1 & 0 \\ 8 & 3 & 2 & 0 \\ 1 & 2 & 5 & 3 \\ 0 & 0 & 3 & 4 \end{bmatrix} \tag{5.176}$$

We can now state three conjectures of increasing strength.

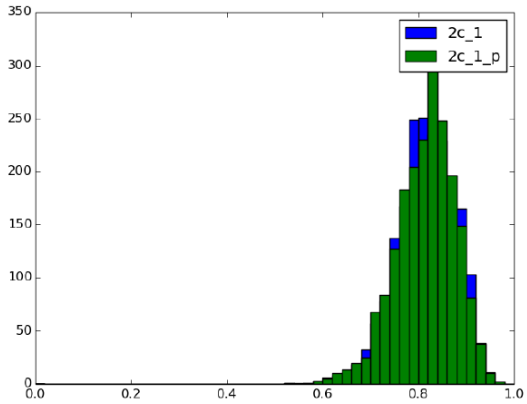

Figure 13: A test of the "very weak" conjecture. $M = 5$, $N = 128$, and we have 2 pairs of zeros.

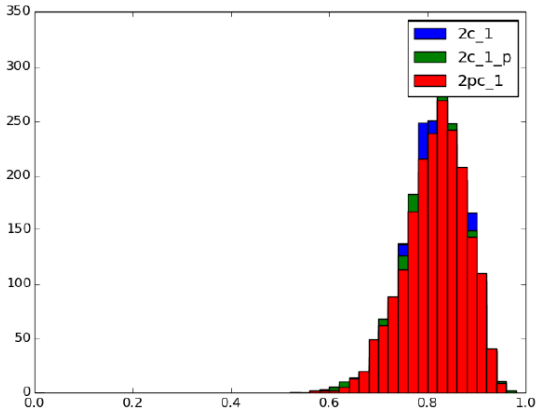

Figure 14: A test of the "weak" conjecture. $M = 5$, $N = 128$, and we have 2 pairs of zeros.

Conjecture 1. The "very weak version": the distribution of Γ is invariant between two sequences of experiments, one with Σ_1 and one with Σ_2, if Σ_1^{-1} and Σ_2^{-1} are *relabeling*.

Conjecture 2. The "weak version": the distribution of Γ is invariant between two sequences of experiments, one with Σ_1 and one with Σ_2, if Σ_1^{-1} and Σ_2^{-1} are 0-*permutations*.

Conjecture 3. The "strong version": the distribution of Γ is invariant between two sequences of experiments, one with Σ_1 and one with Σ_2, if Σ_1^{-1} and Σ_2^{-1} are 0-*count equivalent*.

We feel that the study of these conjectures would a fruitful path for future research.

6 Conclusion

The most general finding of this thesis is an exact formula for the estimator of the information matrix that maximizes multivariate normal likelihood, under a constraint on values

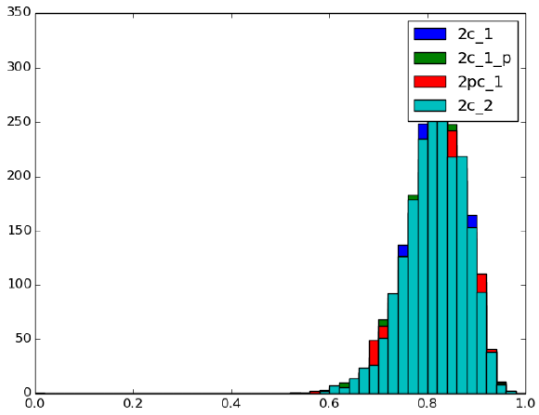

Figure 15: A test of the "strong" conjecture. $M = 5$, $N = 128$, and we have 2 pairs of zeros.

of the information matrix in a principal submatrix. Numerical simulations demonstrate the performance of a convex solver against this exact result, and show that even though the likelihood returned by the convex solver is near the exact likelihood, its returned solution for the information matrix does not well-approximate the constraints, at least for a reasonable number of iterations.

A Proofs of various propositions

Herein be provide detailed proofs of some statements from the main text.

Prove [2] Proposition 5.2

Proposition. Assume that $Y \sim N_{|\Gamma|}(\xi, \Sigma)$, where Σ is regular. Then it holds for $\gamma, \mu \in \Gamma$ with $\gamma \neq \mu$ that

$$Y_\gamma \perp\!\!\!\perp Y_\mu | Y_{\Gamma/\{\gamma,\mu\}} \;\Leftrightarrow\; k_{\gamma\mu} = 0 \tag{A.177}$$

where $K = \{k_{\alpha\beta}\}_{\alpha,\beta\in\Gamma} = \Sigma^{-1}$ is the concentration matrix of the distribution.

Proof. Suppose $V = \mathbb{R}^n$ and assume the random vector $\mathbf{X}$ partitioned into components $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1 \in \mathbb{R}^p$ and $\mathbf{X}_2 \in \mathbb{R}^q$ with $p + q = n$.
The mean vector and covariance matrix can then be partitioned accordingly into blocks as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{A.178}$$

such that Σ_{11} has dimensions $p \times p$ and so on.

Let $\mathbf{X}$ be distributed as $N_n(\xi, \Sigma)$. Then we can show that the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = x_2$ is $N_p(\xi_{1|2,\Sigma_{1|2}})$ where

$$\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}x_2 - \xi_2 \text{ and } \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{A.179}$$

Since

$$f_{\xi,\Sigma}(x) = (2\pi)^{-p/2}(\det\Sigma)^{1/2}e^{-<x-\xi,k(x-\xi)>/2} \tag{A.180}$$

So

$$f(x_1|x_2) \propto f_{\xi,\Sigma}(x) \propto e^{-<x-\xi,k(x-\xi)>/2} \tag{A.181}$$

Since

$$\begin{aligned}
< x-\xi, k(x-\xi) > &= \left\langle \begin{pmatrix} x_1-\xi_1 \\ x_2-\xi_2 \end{pmatrix}, \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{pmatrix} x_1-\xi_1 \\ x_2-\xi_2 \end{pmatrix} \right\rangle \\
&= \left\langle \begin{pmatrix} x_1-\xi_1 \\ x_2-\xi_2 \end{pmatrix}, \begin{pmatrix} k_{11}(x_1-\xi_1)+k_{12}(x_2-\xi_2) \\ k_{21}(x_1-\xi_1)+k_{22}(x_2-\xi_2) \end{pmatrix} \right\rangle \\
&= \begin{pmatrix} x_1-\xi_1 \\ x_2-\xi_2 \end{pmatrix}^T \begin{pmatrix} k_{11}(x_1-\xi_1)+k_{12}(x_2-\xi_2) \\ k_{21}(x_1-\xi_1)+k_{22}(x_2-\xi_2) \end{pmatrix} \\
&= (x_1-\xi_1)^T[k_{11}(x_1-\xi_1)+k_{12}(x_2-\xi_2)] + (x_2-\xi_2)^T[k_{21}(x_1-\xi_1)+k_{22}(x_2-\xi_2)] \\
&= (x_1-\xi_1)^T k_{11}(x_1-\xi_1) + (x_1-\xi_1)^T k_{12}(x_2-\xi_2) \\
&+ (x_2-\xi_2)^T k_{21}(x_1-\xi_1) + (x_2-\xi_2)^T k_{22}(x_2-\xi_2)
\end{aligned} \tag{A.182}$$

Since $(x_2-\xi_2)^T k_{21}(x_1-\xi_1)$ is singular, so $[(x_2-\xi_2)^T k_{21}(x_1-\xi_1)]^T = (x_2-\xi_2)^T k_{21}(x_1-\xi_1)$.
So

$$\begin{aligned}
< x-\xi, k(x-\xi) > &= (x_1-\xi_1)^T k_{11}(x_1-\xi_1) + (x_1-\xi_1)^T k_{12}(x_2-\xi_2) \\
&+ [(x_2-\xi_2)^T k_{21}(x_1-\xi_1)]^T + (x_2-\xi_2)^T k_{22}(x_2-\xi_2) \\
&= (x_1-\xi_1)^T k_{11}(x_1-\xi_1) + (x_1-\xi_1)^T k_{12}(x_2-\xi_2) \\
&+ (x_1-\xi_1)^T k_{12}(x_2-\xi_2) + (x_2-\xi_2)^T k_{22}(x_2-\xi_2) \\
&= (x_1-\xi_1)^T k_{11}(x_1-\xi_1) + 2(x_1-\xi_1)^T k_{12}(x_2-\xi_2) + (x_2-\xi_2)^T k_{22}(x_2-\xi_2)
\end{aligned} \tag{A.183}$$

So

$$f(x_1|x_2) \propto \exp\left\{-(x_1-\xi_1)^T k_{11}(x_1-\xi_1)/2 - (x_1-\xi_1)^T k_{12}(x_2-\xi_2)\right\} \tag{A.184}$$

Since

$$\begin{aligned}
&-(x_1-\xi_1)^T k_{11}(x_1-\xi_1)/2 - (x_1-\xi_1)^T k_{12}(x_2-\xi_2) \\
=&(-x_1^T+\xi_1^T)k_{11}(x_1-\xi_1)/2 + (-x_1^T+\xi_1^T)k_{12}(x_2-\xi_2) \\
=&(-x_1^T k_{11}+\xi_1^T k_{11})(x_1-\xi_1)/2 + (-x_1^T k_{12}+\xi_1^T k_{12})(x_2-\xi_2) \\
=&\frac{-x_1^T k_{11}x_1}{2} + \frac{x_1^T k_{11}\xi_1}{2} + \frac{\xi_1^T k_{11}x_1}{2} - \frac{\xi_1^T k_{11}\xi_1}{2} - x_1^T k_{12}x_2 + x_1^T k_{12}\xi_2 + \xi_1^T k_{12}x_2 - \xi_1^T k_{12}\xi_2
\end{aligned} \tag{A.185}$$

Since $\left[\frac{\xi_1^T k_{11}x_1}{2}\right]^T = \frac{\xi_1^T k_{11}x_1}{2} = \frac{x_1^T k_{11}\xi_1}{2}$, so the linear term involving x_1 has coefficient

$$\frac{k_{11}\xi_1}{2} + \frac{k_{11}\xi_1}{2} - k_{12}x_2 + k_{12}\xi_2 = k_{11}\xi_1 - k_{12}(x_2-\xi_2) = k_{11}[\xi_1 - k_{11}^{-1}k_{12}(x_2-\xi_2)] \tag{A.186}$$

Since

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1} & -E{-}1G \\ -FE^{-1} & D^{-1}+FE^{-1}G \end{pmatrix} \tag{A.187}$$

where $E = A - BD^{-1}C, F = D^{-1}C, G = BD^{-1}$.
Since

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} \tag{A.188}$$

So

$$k_{11}^{-1} = E = A - BD^{-1}C = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{A.189}$$

$$K_{12} = -E^{-1}G \tag{A.190}$$

So

$$k_{11}^{-1}k_{12} = E(-E^{-1}G) = -G = -BD^{-1} = -\Sigma_{12}\Sigma_{22}^{-1} \tag{A.191}$$

So

$$\begin{aligned} f(x_1|x_2) &\propto \exp\left\{\frac{-x_1^T k_{11}x_1}{2} + x_1^T k_{11}[\xi_1 - k_{11}^{-1}k_{12}(x_2-\xi_2)]\right\} \\ &\propto -\frac{1}{2k_{11}^{-1}}\left\{x_1^T x_1 - 2x_1^T[\xi_1 - k_{11}^{-1}k_{12}(x_2-\xi_2)]\right\} \\ &\sim N\left(\xi_1 - k_{11}^{-1}k_{12}(x_2-\xi_2), k_{11}^{-1}\right) \\ &\sim N\left(\xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\xi_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) \\ &\sim N_p\left(\xi_{1|2}, \Sigma_{1|2}\right) \end{aligned} \tag{A.192}$$

Now we know $\Sigma_{1|2} = k_{11}^{-1}$, since $\mathbf{K}_{\{\gamma,\mu\}} = \begin{pmatrix} k_{\gamma\gamma} & k_{\gamma\mu} \\ k_{\mu\gamma} & k_{\mu\mu} \end{pmatrix}$
So we have

$$\Sigma_{\gamma,\mu|\Gamma/\{\gamma,\mu\}} = k_{\{\gamma,\mu\}}^{-1} = \frac{1}{\det\mathbf{K}_{\{\gamma,\mu\}}} = \begin{pmatrix} k_{\mu\mu} & -k_{\gamma\mu} \\ -k_{\mu\gamma} & k_{\gamma\gamma} \end{pmatrix} \tag{A.193}$$

Now back to the proposition:
($\Rightarrow$)
If $Y_\gamma \perp\!\!\!\perp Y_\mu | Y_{\Gamma/\{\gamma,\mu\}}$, that means the nondiangonal entries of the matrix $\Sigma_{\gamma,\mu|\Gamma/\{\gamma,\mu\}}$are zero. Which means $k_{\gamma\mu}$ and $k_{\mu\gamma}$ are zero.
($\Leftarrow$)
It is easy to proof by going through the opposite direction.

□

Proof for Proposition 1

Proposition. For subsets a, b of C with $a \cup b = C$ the following statements are equivalent.
(i) $\Sigma_{a,b} = \Sigma_{a,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b}$.
(i′) $\Sigma_{a/b,b/a} = \Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a}$.
(ii) $(\Sigma^{-1})_{a/b,b/a} = 0$
(iii) $\mathbf{X}_a$ and $\mathbf{X}_b$ are conditionally independent given $\mathbf{X}_{ab}$

Prove Statements (i) and (i$^{'}$) are equivalent.

Proof. Define $\overline{\Sigma_{m,n}}$ as extending $\Sigma_{m,n}$ into $a \times b$ dimensions, keeping entries in rows m and columns n do not change, filling other entries zeros. So, we know

$$\begin{aligned}\Sigma_{a,b} &= \Sigma_{a/b\cup ab\,,\,b/a\cup ab} \\ &= \overline{\Sigma_{a/b,b/a}} + \overline{\Sigma_{a/b,ab}} + \overline{\Sigma_{ab,b/a}} + \overline{\Sigma_{ab,ab}}\end{aligned}$$

From (i') we have

$$\begin{aligned}&= \overline{\Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a}} + \overline{\Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,ab}} + \overline{\Sigma_{ab,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a}} + \overline{\Sigma_{ab,ab}\Sigma_{ab}^{-1}\Sigma_{ab,ab}} \\ &= \overline{\Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a}} + \overline{\Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab}} + \overline{\Sigma_{ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a}} + \overline{\Sigma_{ab}\Sigma_{ab}^{-1}\Sigma_{ab}} \\ &= \overline{\Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a\cup ab}} + \overline{\Sigma_{ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a\cup ab}} \\ &= \overline{\Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b}} + \overline{\Sigma_{ab}\Sigma_{ab}^{-1}\Sigma_{ab,b}} \\ &= \Sigma_{a/b\cup ab,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b} \\ &= \Sigma_{a,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b}\end{aligned} \tag{A.194}$$

So (i) is established from (i$^{'}$).
It is easy to get (i$^{'}$) from (i) by going through the opposite direction.
Hence, statements (i) and (i$^{'}$) are equivalent. □

Prove Statements (i$^{'}$) and (ii)are equivalent.

Proof. Since

$$\Sigma_{a/b,b/a} = \Sigma_{a/b,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a} \tag{A.195}$$

So

$$\mathrm{cov}(\mathbf{X}_{a/b}, \mathbf{X}_{b/a}|\mathbf{X}_{ab}) = \Sigma_{a/b,b/a} - \Sigma_{a/,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b/a} = 0 \tag{A.196}$$

Which means

$$\mathbf{X}_{a/b} \perp\!\!\!\perp \mathbf{X}_{b/a}|\mathbf{X}_{ab} \tag{A.197}$$

Also

$$ab \text{ is the rest of } C \text{ exclude } a/b \text{ and } b/a \tag{A.198}$$

Hence, from [2]

$$(\Sigma^{-1})_{a/b,b/a} = 0 \tag{A.199}$$

So (ii) is established from (i$^{'}$).
It is easy to get (i$^{'}$) from (ii) by going through the opposite direction.
Hence, statements (ii) and (i$^{'}$) are equivalent. □

Prove Statements (i) and (iii) are equivalent.

Proof. Since

$$\Sigma_{a,b} = \Sigma_{a,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b} \tag{A.200}$$

So

$$\mathrm{cov}(\mathbf{X}_a, \mathbf{X}_b|\mathbf{X}_{ab}) = \Sigma_{a,b} - \Sigma_{a,ab}\Sigma_{ab}^{-1}\Sigma_{ab,b} = 0 \tag{A.201}$$

So

$$\mathbf{X}_a \perp\!\!\!\perp \mathbf{X}_b|\mathbf{X}_{ab} \tag{A.202}$$

It is easy to get (i) from (iii) by going through the opposite direction. □

References

[1] Y. I. Abramovich, N. K. Spencer, and A. Y. Gorokhov, "GLRT-based threshold detection-estimation performance improvement and application to uniform circular antenna arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 20–31, 2007. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4034093

[2] S. Lauritzen, *Graphical models*, 1996.

[3] T. Speed and H. Kiiveri, "Gaussian Markov distributions over finite graphs," *The Annals of Statistics*, 1986. [Online]. Available: http://www.jstor.org/stable/2241271

[4] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the lasso," *Biostatistics*, pp. 1–14, 2007. [Online]. Available: http://biostatistics.oxfordjournals.org/content/9/3/432.short http://arxiv.org/abs/0708.3517

[5] D. Witten, J. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *Journal of Computational and*, 2011. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1198/jcgs.2011.11051a

[6] R. Paffenroth, N. Li, and L. Scharf, "MAXIMUM LIKELIHOOD IDENTIFICATION OF AN INFORMATION MATRIX UNDER CONSTRAINTS IN A CORRESPONDING GRAPHICAL MODEL," in *50th Asilomar Conference on Signals, Systems and Computers*, 2016.

[7] J. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," 1971. [Online]. Available: http://www.citeulike.org/group/14833/article/8970271

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004, vol. 25, no. 3.

[9] J. Hastie, Trevor, Tibshirani, Robert, Friedman, *The Elements of Statistical Learning-Data Mining, Inference, and Prediction, Second Edition*, 2009.

[10] J. Stewart, *Calculus*, 2015.

[11] M. S. Pedersen, B. Baxter, B. Templeton, C. Rishøj, D. L. Theobald, E. Hoeghrasmussen, G. Casteel, J. B. Gao, K. Dedecius, K. Strim, L. Christiansen, L. K. Hansen, L. Wilkinson, L. He, M. Bar, O. Winther, P. Sakov, S. Hattinger, K. B. Petersen, and C. Rishø j, "The Matrix Cookbook," *Matrix*, vol. M, pp. 1–71, 2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.139.3165&rep=rep1&type=pdf

[12] M. Woodbury, "Inverting modified matrices," Tech. Rep. 42, 1950.

[13] S. Diamond and S. Boyd, "CVXPY: a python-embedded modeling language for convex optimization," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2909–2913, 2016.

[14] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, 2002. [Online]. Available: http://qpsy.snu.ac.kr/teaching/multivariate/Matrix.pdf