การลดจำนวนตัวแปรแบบแฝงในแบบจำลองมิมิคสำหรับการระบุปัจจัยกำหนดโครงสร้างเงินทุน

นางสาวชลลดา เลาหพันธ์ศักดา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมการเงิน ภาควิชาการธนาคารและการเงิน
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

A REDUCTION OF LATENT VARIABLES IN MIMIC MODELS TO IDENTIFY DETERMINANTS OF CAPITAL STRUCTURE

Miss Chollada Laohaphansakda

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Financial Engineering
Department of Banking and Finance
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2017

Thesis Title	A REDUCTION OF LATENT VARIABLES IN MIMIC MODELS TO IDENTIFY DETERMINANTS OF CAPITAL STRUCTURE
By	Miss Chollada Laohaphansakda
Field of Study	Banking and Finance
Thesis Advisor	Assistant Professor Jitkomut Songsiri, Ph.D.
Thesis Co-Advisor	Assistant Professor Thaisiri Watewai, Ph.D.

Accepted by the Faculty of Commerce and Accountancy, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

....Dean of the Faculty of Commerce and Accountancy
(Associate Professor Pasu Decharin, Ph.D.)

THESIS COMMITTEE

....Chairman
(Anant Chiarawongse, Ph.D.)

....Thesis Advisor
(Assistant Professor Jitkomut Songsiri, Ph.D.)

....Thesis Co-Advisor
(Assistant Professor Thaisiri Watewai, Ph.D.)

....Committee
(Pisit Jarumaneeroj, Ph.D.)

ชลลดา เลาหพันธ์ศักดา: การลดจำนวนตัวแปรแฝงในแบบจำลองมิมิคสำหรับการระบุปัจจัยกำหนดโครงสร้างเงินทุน (A REDUCTION OF LATENT VARIABLES IN MIMIC MODELS TO IDENTIFY DETERMINANTS OF CAPITAL STRUCTURE)
อ. ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. จิตโกมุท ส่งศิริ, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร. ไทยศิริ เวทไว 92 หน้า

โครงสร้างเงินทุนอธิบายถึงอัตราส่วนหนี้สินและส่วนของผู้ถือหุ้นของบริษัทในลักษณะของแหล่งเงินทุนต่างๆ อัตราส่วนวิเคราะห์นโยบายทางการเงินเป็นตัวชี้วัดของโครงสร้างเงินทุน ปัจจัยกำหนดโครงสร้างเงินทุนที่ใช้อธิบายตัวชี้วัดของโครงสร้างเงินทุนมีตัวอย่างดังเช่น อัตราการเจริญเติบโต, ลักษณะเฉพาะของสินค้า, ความผันผวน, ความสามารถในการทำกำไร, ผลประโยชน์ทางภาษีที่ไม่ใช่หนี้สิน, มูลค่าของหลักทรัพย์ที่ใช้ค้ำประกัน, และประเภทของอุตสาหกรรม ปัจจัยกำหนดโครงสร้างเงินทุนเหล่านี้ถูกเรียกว่าตัวแปรแฝงซึ่งไม่สามารถสังเกตได้โดยตรงแต่ถูกอธิบายได้จากตัวแปรสาเหตุของปัจจัยกำหนดโครงสร้างเงินทุนที่สังเกตได้ เนื่องจากโครงสร้างเงินทุนเกี่ยวข้องกับการประเมินประสิทธิภาพของบริษัท การศึกษาความสัมพันธ์ระหว่างตัวแปรดังกล่าวข้างต้นจึงมีความสำคัญต่อการจัดการโครงสร้างเงินทุน ความสัมพันธ์ระหว่างตัวแปรเหล่านี้สามารถอธิบายได้ผ่านแบบจำลองมิมิคที่เป็นแบบจำลองถดถอยหลายตัวแปร และใช้ตัวแปรชี้วัดกับตัวแปรสาเหตุที่สังเกตได้ในการอธิบายตัวแปรแฝง โดยทั่วไปนั้น ผู้ใช้อาจจะกำหนดโครงสร้างเงินทุนโดยการรวมตัวแปรแฝงที่เป็นไปได้ซึ่งมักจะประกอบไปด้วยตัวแปรแฝงทั้งที่สำคัญและไม่สำคัญ การลดตัวแปรแฝงที่ไม่สำคัญสามารถลดความแปรปรวนของคำตอบจากการหาความสัมพันธ์ที่แท้จริงระหว่างตัวชี้วัดและปัจจัยกำหนดโครงสร้างเงินทุน วิทยานิพนธ์นี้จึงนำเสนอ 2 รูปแบบปัญหา ได้แก่ 1) การคัดเลือกตัวแปรแฝงที่สำคัญ และ 2) การประมาณแบบจำลองมิมิคที่ถูกลดรูปหลังจากลดจำนวนตัวแปรแฝงที่ไม่สำคัญ รูปแบบปัญหาการคัดเลือกตัวแปรแฝงที่สำคัญนั้นป็นการประมาณแบบกำลังสองต่ำสุดที่มีฟังก์ชันลงโทษแบบนอร์ม-1 เพื่อบังคับให้เกิดโครงสร้างศูนย์ในแบบจำลองซึ่งจะอธิบายถึงตัวแปรแฝงที่ไม่สำคัญ รูปแบบปัญหาที่สองเป็นการประมาณแบบกำลังสองต่ำสุดที่มีเงื่อนไขเชิงเส้น เนื่องจากทั้งสองรูปแบบปัญหามีลักษณะเป็นไบคอนเวกซ์ งานวิจัยนี้ประยุกต์ใช้การหาค่าต่ำสุดแบบสลับในการแก้ปัญหา ในแต่ละปัญหาย่อยของการแก้ปัญหาซึ่งมีลักษณะเป็นคอนเวกซ์จะประกอบด้วยปัญหาแลซโซแบบกลุ่ม และวิธีการประมาณแบบกำลังสองต่ำสุด ซึ่งสามารถประยุกต์ใช้ขั้นตอนเชิงเลขได้หลายวิธี โครงสร้างความสัมพันธ์ของระหว่างตัวแปรแฝงและตัวแปรชี้วัดจะแปรเปลี่ยนไปตามพารามิเตอร์ลงโทษในปัญหาการคัดเลือกตัวแปรแฝงที่สำคัญ เราจึงเลือกพารามิเตอร์นั้นจากเกณฑ์การเลือกแบบจำลองต่างๆ ตามความต้องการของผู้ใช้ ประสิทธิภาพของปัญหาที่นำเสนอนั้นได้ถูกทดสอบด้วยข้อมูลที่สังเคราะห์ขึ้น ผลลัพธ์ที่ได้แสดงให้เห็นว่ารูปแบบการประมาณที่เราได้นำเสนอมีประสิทธิภาพในการคัดเลือกตัวแปรแฝงมากกว่ารูปแบบการประมาณแบบกำลังสองต่ำสุดเมื่อแบบจำลองจริงมีลักษณะเบาบาง นอกจากนี้ เราได้ประยุกต์ใช้งานกับข้อมูลจริงจากอุตสาหกรรม 7 ประเภทในประเทศอเมริกาตอนเหนือ เกณฑ์การเลือกแบบจำลองทุกเกณฑ์ให้ผลว่า อัตราการเจริญเติบโตเป็นปัจจัยกำหนดโครงสร้างเงินทุนที่มีความสำคัญมากที่สุด ในทางกลับกันความผันผวนเป็นปัจจัยที่ไม่มีความสำคัญต่อโครงสร้างเงินทุน ผลการวิจัยนี้ถูกเปรียบเทียบกับสองทฤษฎีหลักของโครงสร้างเงินทุน ได้แก่ ทฤษฎีโครงสร้างเงินทุนที่เหมาะสม และทฤษฎีการจัดหาเงินทุนตามลำดับขั้น ซึ่งพบว่าทิศทางของความสัมพันธ์ระหว่างอัตราส่วนหนี้สิน กับ i) อัตราการเจริญเติบโต ii) ลักษณะเฉพาะของสินค้า และ iii) ผลประโยชน์ทางภาษีที่ไม่ใช่หนี้สิน มีความสอดคล้องกับทฤษฎีโครงสร้างเงินทุนที่เหมาะสม ในขณะที่ทิศทางของอัตราส่วนหนี้สินกับความสามารถในการทำกำไรมีความสอดคล้องกับทฤษฎีการจัดหาเงินทุนตามลำดับขั้น

ภาควิชา การธนาคารและการเงิน .... **ลายมือชื่อนิสิต** ....
สาขาวิชา วิศวกรรมการเงิน .... **ลายมือชื่อ อ.ที่ปรึกษา** ....
ปีการศึกษา 2560 ....

5882913626 : MAJOR FINANCIAL ENGINEERING

KEYWORDS : CAPITAL STRUCTURE / MIMIC MODEL / GROUP LASSO / BICONVEX / ALTERNATING MINIMIZATION

CHOLLADA LAOHAPHANSKDA : A REDUCTION OF LATENT VARIABLES IN MIMIC MODELS TO IDENTIFY DETERMINANTS OF CAPITAL STRUCTURE ADVISOR : ASST. PROF. JITKOMUT SONGSIRI, Ph.D., CO-ADVISOR : ASST. PROF. THAISIRI WATEWAI, Ph.D., 92 pp.

A capital structure describes the proportions of debt and equity used by a firm as different sources of funds. Financial leverage ratios are the measures of capital structure. The determinants of capital structure are used to explain the measures of capital structure, for example, growth, volatility, profitability, non-debt tax shields, collateral value of assets, and industry classification. These determinants of capital structure are typically regarded as latent variables since they cannot be directly observed, but can be explained by some observed variables which are the causes of determinants of capital structure. Since the capital structure relates to a firm's performance, studying a relationship among these variables is significant for the capital structure management. The relation among variables in the capital structure can be explained by a Multiple Indicators Multiple Causes (MIMIC) model which is simply a multivariate regression equation where latent variables are explained by both effects and causes from observed variables, *i.e.*, the measures and causes of determinants of capital structure. In general, one can preliminarily include all possible latent variables in the model but they may contain both relevant and irrelevant latent variables. Consequently, a reduction of those ineffective latent variables can reduce the variance of estimated model parameters for an exploration of a true relationship between the measures of capital structure and their determinants. This thesis proposes two estimation formulations: a latent variable selection and an estimation of the reduced MIMIC model after eliminating ineffective latent variables. The problem of latent variable selection is a least-squares problem with a 1-norm penalty to induce a zero structure in the model which further describes ineffective latent variables. The second formulation is a least-squares problem with linear constraints. The proposed problems are biconvex and we apply the alternating minimization to solve for numerical solutions. In each step of alternating minimization, the optimization problems are convex having a form of group lasso and linear least-squares problems which can be solved by many existing efficient numerical algorithms. Since our formulations provide a set of models whose relationship structures vary upon the regularization parameter, we apply model selection criterions to select an appropriate model based on preferences and objectives of users. The performance of the proposed formulations is demonstrated via simulation experiments. The results show that our proposed method can remove ineffective latent variables more correctly than methods based on least-squares when true model is sparse. We apply our estimation formulations to applications of the capital structure from seven industries in the North America. All model selection criterions totally agree that growth is the most important determinant of capital structure; in contrast, volatility is an insignificant factor. The results from our work are compared to two main capital structure theories: the trade-off theory and pecking-order theory. The directions of relationships between debt ratios and i) growth ii) uniqueness, and iii) non-debt tax shields are consistent with the trade-off theory while the direction between the debt ratios and profitability is consistent with the pecking-order theory.

Department : Banking and Finance
Field of Study : Financial Engineering
Academic Year : 2017

Student's Signature ....
Advisor's Signature ....

Acknowledgements

Firstly, I would like to express my deepest and sincerest thanks to my thesis advisor Assistant Professor Jitkomut Songsiri for everything she has done for me. She has helped and taught me a lot of things about the research development and analysis. I truly appreciate her patient guidance, dedication, and intention in doing the very best in our works. She does not neglect even a slight mistake so that such an efficient research is achieved. This is the most difficult work I have ever done but I have never been afraid of any obstacle since I always have continuous support from her. She is my inspiration to improve myself all the time.

My grateful thanks are also extended to Mr. Anupon Pruttiakaravanich, my friend from Control Systems Research Laboratory. He is another person who always assists me about mathematical problems, techniques, and algorithms. Thank you for being the best buddy, especially, giving dedicated help, advices, and encouragement. Thank you for staying by my side through the good and bad times.

My sincere thanks also goes to my thesis co-advisor Assistant Professor Thaisiri Watewai for good counsel and intensive knowledge in a financial field. Thank you for beneficial suggestion and critical discussion; moreover, providing new perspectives that open my mind to learn new things. Moreover, I would like to thank my thesis committee Dr. Anant Chiarawongse and Dr. Pisit Jarumaneeroj for their insightful comments and sharing helpful opinions. In addition, I am deeply appreciative of my family's support and encouragement throughout all the time of my study.

In particular, I am thankful Faculty of Commerce and Accountancy, Master of Science in Financial Engineering program for providing facilities and data resources to fulfill the completion of my thesis. Lastly, I am grateful to Bangkok Bank Public Company Limited for providing the scholarship throughout my master's study.

Contents

Page

CHAPTER

List of Figures

CHAPTER I

INTRODUCTION

1.1 Introduction

A capital structure tells us how a firm operates its capital for running the business based on debt and equity by using different sources of funds, *e.g.*, loans from financial institutions, bonds, shares, etc. Management of the capital structure of a company is very significant since it relates to the performance of corporation; moreover, it also affects to the main objective of the company that is to maximize the profit. The fund of the company comes from debt and equity. The corporation tries to find a right balance between the two to minimize bankruptcy with the optimal capital structure [Modigliani and Miller, 1958]. The company can be in trouble and may go bankrupt if the firm chooses capital structure inappropriately. On the other hand, the company can enlarge its firm market value if the cost of capital is reduced. Consequently, studying the determinants of capital structure is extremely significant in the sense that it has an influence to a proportion of debt and equity on capital structure and promotes a firm's administrator to make an optimal decision about capital structure based on characteristics which are related to debt and equity financing. However, the theoretical determinants of capital structure are sometimes determined by unobserved factors that illustrate abstract concepts in each attribute. Unobserved factors cannot be directly observed but they are explained by some observed factors which have characteristics as same as in the unobserved factors.

In order to manage the capital structure, a relationship between measures and determinants of capital structure is needed. The measures of capital structure are related to financial leverage ratios, *e.g.*, long-term debt, short-term debt, and convertible debt divided by market or by book values of equity. Growth, volatility, profitability, non-debt tax shields, collateral value of assets, industry classification which are abstract concepts and cannot directly be observed are considered to be examples of the determinants of capital structure. These abstract concepts are explained by some observed factors; for example, profitability (unobserved variable) is explained by operating income-to-total assets ratio and operating income-to-total sales ratio (observed variables). Since unobserved variables are induced to the model and under an assumption that variables in the model are explained in the multiple linear regression, Structural Equation Modeling (SEM) is a popular method to estimate the relationship among variables.

Structural Equation Modeling (SEM) is a statistical modeling procedure that is widely used to find a causal relationship among variables. This method is divided into two types: exploratory and confirmatory modeling. To search for a structure of a statistical causal model, the exploratory modeling is applied and the confirmatory modeling is used to verify whether a model is promoted by a sample of data or not. When there is no significant theoretical information to support the model, one can start applying *exploratory factor analysis* (EFA) to a model for seeking the number and the feature of unobserved variables, *i.e.*, exogenous and endogenous variables by using given observations. After the model is proposed, *confirmatory factor analysis* (CFA) is applied for testifying the model constructed by EFA with another set of observations. There are two types of random variables in this model that are observed variables and latent variables. The variables which are directly measured and can extrapolate the unobserved variables are defined as observed variables. On the other hand, latent variables are indirectly measured but they are extrapolated from observed variables [Lomax and Schumacker, 2012, §5]. In measurement models, the latent variables that demonstrate highly abstract concepts, *e.g.*, anxiety, intelligence, happiness, merge physical realities which can be observed, *e.g.*, age, weight, pressure, to a single term [Bollen, 2014, §6]. One constructs the latent variables based on the idea of the similarity of physical variables since each latent variable combines common characteristics of their observed variables. For example, operating income-to-total assets ratio and operating income-to-sales ratio are the ratios (observed variables) that represent the concept of profitability of a firm (latent variable).

Since latent variables may be constructed in the model, their signification should be explained by the relationship among observed variables that may present as both effects and causes of latent variables or either one of the two [Jöreskog and Goldberge, 1975]. Multiple Indicators Multiple Causes model or MIMIC model, a special type of SEM model, is suggested to explain this type of causal model. In MIMIC model, the multiple indicators are the observed outcome variables determining the latent variables in the multiple linear regression. Besides, the multiple causes are mentioned to the multiple predictors of latent variables that are also explained in the multiple linear regression (see the path diagram of MIMIC model in Figure 1.1. MIMIC model was first proposed in 1975 by [Jöreskog and Goldberge, 1975] and they apply the maximum-likelihood in order to estimate the parameters in MIMIC model with a single latent variable and this model is improved with multiple latent variables by [Stapleton, 1978].

The mathematical representation of a MIMIC model is given by

$$\eta = B^T x + \epsilon_\eta \tag{1.1}$$

$$y = A\eta + \epsilon_y \tag{1.2}$$

where observed variables $x \in \mathbf{R}^p$ and $y \in \mathbf{R}^q$ are causes and indicators of latent variable $\eta \in \mathbf{R}^r$, respectively. $A = \begin{bmatrix} A_1 & \cdots & A_r \end{bmatrix} \in \mathbf{R}^{q \times r}$ and $B = \begin{bmatrix} B_1 & \cdots & B_r \end{bmatrix} \in \mathbf{R}^{p \times r}$ are the coefficient matrices that show the relation of y to η and η to x, respectively. $\epsilon_\eta \in \mathbf{R}^r$ is the disturbance of η and $\epsilon_y \in \mathbf{R}^q$ is the measurement error of y. Note that x, ϵ_η, and ϵ_y are assumed to have a normal distribution and x is independent of ϵ_η and ϵ_y.

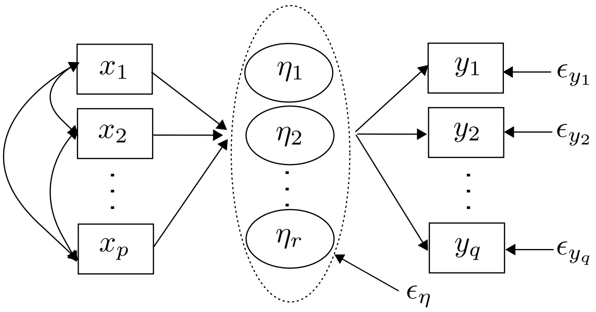


Figure 1.1: Path diagram of a MIMIC model.

An example of MIMIC model applied to a capital structure research can be found in [Titman and Wessels, 1988]. In their research, y, the firm's debt-equity choices, is explained by η, the determinants of capital structure which is measured by x. They denote latent attribute η to be non-debt tax shields, growth, uniqueness, industry classification, size, earning volatility, and profitability; furthermore, the measure of capital structure, y, is given by short-term debt, long-term debt, and convertible debt divided by market and by book value of equity.

[Titman and Wessels, 1988] use SEM to explore the relationship among observed and latent variables, A and B. Since there are too many latent variables compared to observed variables, they have to add more constraints to get an identifiable model. An effect from adding more constraints to the model leads to an estimated solution that misses the relevant relations to its true value, *i.e.*, the estimator from constrained problem is more biased than the estimator from an unconstrained problem. They suggest that the results can be developed by finding variables which have a stronger linkage between observed and latent variables.

To solve the problem in [Titman and Wessels, 1988], [Chang et al., 2009] apply the reduced form of MIMIC model that comes from substituting (1.1) into (1.2). The reduced form of MIMIC model is given by

$$y = A(B^T x + \epsilon_\eta) + \epsilon_y = AB^T x + A\epsilon_\eta + \epsilon_y = Fx + \epsilon \tag{1.3}$$

where $F = AB^T$ and $\epsilon = A\epsilon_\eta + \epsilon_y$. Instead of using model (1.1) and (1.2) to find A and B, [Chang et al., 2009] use the reduced form of MIMIC model (1.3) to find F which is defined as the indirect effect between x and y while neglecting the effect from latent variables. To interpret a relation between η and y, they first read off the relationship between x and y represented by F. Then, they calculate the relative impact, *i.e.*, a standardized total effect of y for each x (the calculation is provided in the section 5.2.2). After that, they sort such relative impact in descending order. Under the assumption of knowing that each η is explained by some x's, they consider that the most effective relative impact of x (in each η) to y can be interpreted as the relative impact of such η to y. For example, η_1 is explained by x_1 and x_2, and η_2 is explained by x_3, x_4 and x_5. Suppose the relative impact of x in descending order is sorted by x_1, x_3, x_4, x_5 and x_2, respectively. According to [Chang et al., 2009] 's interpretation, η_1 (effect form x_1) is more effective than η_2 (effect form x_3). However, in fact, η_2 may have more influence than η_1 if effects from x_4 and x_5 are considered. Consequently, this scheme may not work in the sense of neglecting effect from other x's in each η. Considering the most effective relative impact of x (in each η) to y, then assuming that it is the relative impact of such η to y may lead to wrong interpretation since it lacks information of other x's in each η.

In general, estimation of MIMIC model is to search for A and B of model (1.1) and (1.2). However, a limitation of estimation A and B is that typical methods lead to non-uniqueness solutions; therefore, more constraints are added in the model in order to get an identifiable model. As a result, many estimated coefficients are biased from their true value [Stapleton, 1978,Dell'Anno et al., 2007,Gallo et al., 1994]. Consequently, estimation of F in the reduced form of MIMIC model (1.3) is considered because the number of parameters in the model is less and it is possible to get an identifiable model. Nevertheless, F which represents the indirect effect between x and y is unclear to interpret the relation between latent and observed variables. Accordingly, interpretation of results from A and B are preferred to F because they provide the direct relationship among latent and observed variables.

Since many possible latent variables can be added to the model from a viewpoint of investigator, in other words, the capital structure can be explained by many determinants or latent factors which may contain some irrelevant latent variables, a latent variable reduction is very useful for estimating the true relationship between capital structure and its determinants. There are many techniques for latent variable reduction, *e.g.*, regularization techniques, sequential selection; moreover, each method provides different benefits and drawbacks that are discussed in section 3.4. Therefore, we propose a scheme to select significant latent variables. We apply a regularized least-squares for removing insignificant latent variables (η) which affect all components of y based on the model (1.2). In other words, we ignore eliminating latent variables which affect each y_j individually but we are focusing on removing latent variables that simultaneously hardly influence overall y , *i.e.*, we do not consider η and y as scalar but we consider them as vectors. Then we estimate A and B based on remaining highly efficient latent variables. Knowing coefficient matrices A and B provides us some ideas about the relative importance among the determinants. After A and B are estimated we rank the relative impact of the determinants of capital structure, then we can notice how much each determinant affects debt-to-equity ratios which are measures of the capital structure. Note that we do not directly use coefficient matrices A and B for quantitative analysis. In other words, we do not try to estimate debt-to-equity ratios estimation nor do we try to estimate how much we finance according to each observed predictor ratio since there are other factors that affect capital structure management, but knowing a relative impact of the determinants of capital structure is very beneficial for a future capital structure management. Two main theories that attempt to explain capital structure decision are the trade-off theory and the pecking order theory that are discussed in section 2.1.

1.1.1 Assumptions

Under our conditions that the zero structure of A and B is given by Figure 1.2, each of latent variable, η_k, $k = 1, \ldots, r$ affects all y_j, $j = 1, \ldots, q$ and each η_k is explained by some x_i, $i = 1, \ldots, q$. A relationship between each η and X is expressed by $\eta_k = b_k^T x + \epsilon_{\eta_k}$, *i.e.*, row vector in B^T provides the relationship from X to η. According to zero structure of B, each η is not influenced from all x's so row vectors in B^T are not dense vectors. For example, according to the Figure 1.2, for zero structure of B, x_1 to x_6 have an influence to η_1 so only coefficient elements b_{11} to b_{61} in B are nonzero. On the contrary, there is no zero structure in A, *i.e.*, A is dense since all η's explain all y's. In this work, there are three assumptions as follows:

- MIMIC model (1.1) and (1.2) are static models, *i.e.*, the model states at specific time instance.
- A relation among observed and latent variables is constructed as a linear model.
- The structure of A and B is given by Figure 1.2 based on researchers' prior knowledge about capital structure.

1.1.2 Objectives

- To estimate A and B that are the direct effect between observed and latent variables in MIMIC model (1.1) and (1.2).
- To present a scheme to select significant latent variables.

To reach our objectives, we need to achieve two main processes:

1. To select significant latent variables: a number of columns of A and B are a number of latent variables involved in the model suggested by researchers. Such latent variables may contain irrelevant latent variables leading to more variance of solutions. Consequently, we aim to decrease the number of latent variables (r) to m remaining highly significant latent variables, in other words, we reduce columns of A and B from r to m. Obviously, when some latent variables are removed, x related to such removed latent variables are also deleted, saying that variables in x and rows of B are reduced from p to $\tilde{p}$.
2. To estimate A and B based on remaining latent variables: we will find the coefficient matrices $A \in \mathbf{R}^{q \times m}$ and $B \in \mathbf{R}^{\tilde{p} \times m}$ representing directly the relationship between observed and latent variables instead of using $F \in \mathbf{R}^{q \times p}$ in order to obtain the stronger interpretation of solutions.

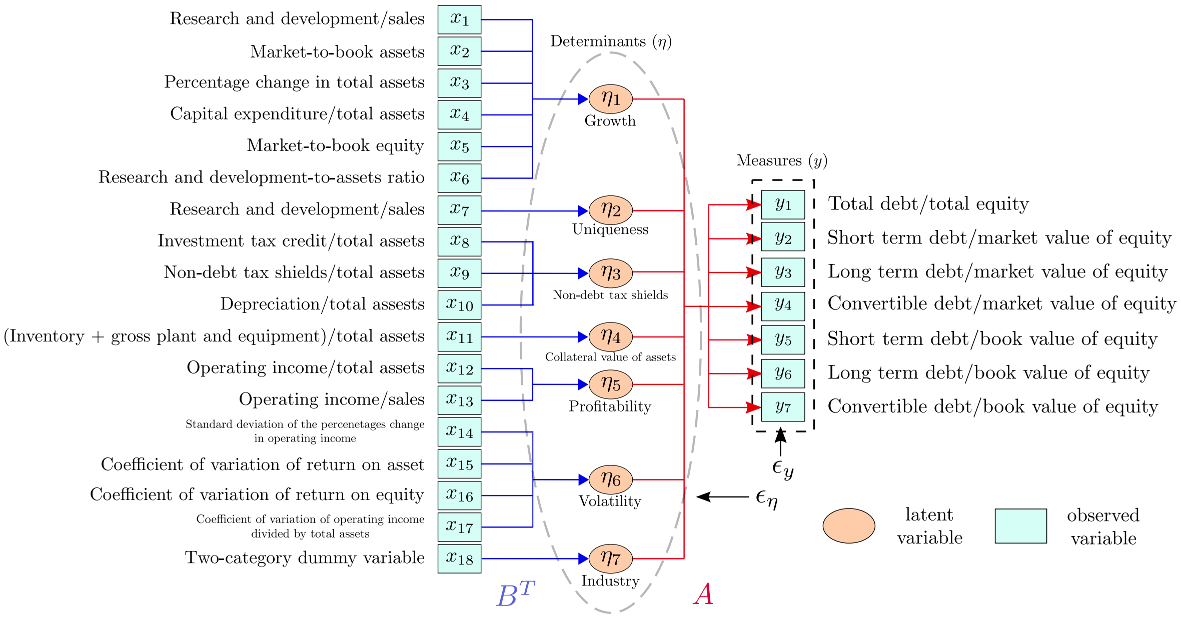

Research and development/sales
Market-to-book assets
Percentage change in total assets
Capital expenditure/total assets
Market-to-book equity
Research and development-to-assets ratio
Research and development/sales
Investment tax credit/total assets
Non-debt tax shields/total assets
Depreciation/total assests
(Inventory + gross plant and equipment)/total assets
Operating income/total assets
Operating income/sales
Standard deviation of the percenetages change in operating income
Coefficient of variation of return on asset
Coefficient of variation of return on equity
Coefficient of variation of operating income divided by total assets
Two-category dummy variable
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
x_{10}
x_{11}
x_{12}
x_{13}
x_{14}
x_{15}
x_{16}
x_{17}
x_{18}
Determinants (η)
η_1
Growth
η_2
Uniqueness
η_3
Non-debt tax shields
η_4
Collateral value of assets
η_5
Profitability
η_6
Volatility
η_7
Industry
B^T
A
ϵ_η
Measures (y)
y_1
y_2
y_3
y_4
y_5
y_6
y_7
Total debt/total equity
Short term debt/market value of equity
Long term debt/market value of equity
Convertible debt/market value of equity
Short term debt/book value of equity
Long term debt/book value of equity
Convertible debt/book value of equity
ϵ_y
latent variable
observed variable

Figure 1.2: Path diagram of MIMIC model illustrates the relationship among variables in capital structure and shows structure of A and B in our work.

1.1.3 Scope of thesis

- Provide a scheme to identify effective determinants of capital structure.
- Provide a formulation that selects highly effective latent variables and a formulation for estimating the reduced MIMIC model.
- Provide numerical methods for solving the two proposed formulations.
- Apply our formulations to real application data and interpret the results.

1.1.4 Expected results

- Estimation formulations of MIMIC model to identify effective determinants of capital structure.
- Numerical methods for solving the two proposed formulations.

1.1.5 Thesis Outline

Our thesis is organized as follows. Chapter 2 explains the background on capital structure, including, capital structure decision, explanation about measures and determinants of capital structure and also a relationship of variables in capital structure. A background on MIMIC model is described in chapter 3. Section 3.2 provides the background of MIMIC model estimation techniques, especially for regularization least-squares estimation which is a fundamental of reducing variables in a model and applied in our work. Section 3.3 explains other alternative methods for MIMIC model estimation via the matrix factorization which could be omitted if readers are familiar with this approach. The methodologies consisting of latent variable selection and least-squares estimation for reduced MIMIC model are stated in chapter 4. Moreover, model selection and numerical methods are provided. Chapter 5 presents experimental results from simulation process including illustration of latent variable selection and performance evaluation. Moreover, we apply our formulation to real data from seven industries in the North America and provide results interpretation. Lastly, the conclusion of this work is shown in chapter 6.

CHAPTER II

BACKGROUND ON CAPITAL STRUCTURE

This chapter demonstrates background on capital structure including capital structure decision, measures and determinants of capital structure, and relationships of variables in capital structure. In this work, we use seven debt-to-equity ratios to be the measures of capital structure consisting of debt-to-equity ratios which are total debt divided by total equity, long-term, short-term, and convertible debt divided by market and by book value of equity as provided in [Titman and Wessels, 1988]. Besides, the determinants of capital structure provided in [Chang et al., 2009] consist of seven latent variables, including, growth, uniqueness, non-debt tax shields, collateral value of assets, profitability, volatility, and industry classification and each of the latent variable can be approximated by some observed variables. First of all, in order to understand materials easier, we have to know the basic accounting equation:

$$\text{Assets} = \text{Liability} + \text{Owner's Equity} \tag{2.1}$$

where assets are resources of a firm or things that the firm owns, *e.g.*, cash, land, buildings, inventory, etc. Liability is an arrangement between a firm and debt holders to borrow money under the condition that the firm will repay the principal with interest at the certain maturity. Equity that expresses ownership of the firm is the value of assets minus the cost of the liabilities.

2.1 Capital structure decision

Capital structure describes how a firm finances its assets through equities and long term debts. Sources of capital structure funds consist of internal funds, *e.g.*, profitability, depreciation, and external funds, *e.g.*, loans from financial institutions, bonds. Since capital structure affects overall operations of a firm, managers select capital structure which generates the highest firm value since firm's shareholders will gain the most profit from such capital structure [Hillier et al., 2010].

[Modigliani and Miller, 1958] convince that changing a firm's capital structure does not have an influence on the firm value in a perfect capital market. In other words, in the world without taxes, the value of the levered and unlevered firm are not different. However, when taxes are induced, [Modigliani and Miller, 1963] dispute that capital structure becomes relevant to firm value due to tax savings from debt. When debt financing is used, bankruptcy costs, transactions costs, and taxes are created and always interpreted by the two main theories of capital structure, *i.e.*, the trade-off theory and the pecking order theory. These two theories attempt to describe a financial decision of a firm. The trade-off theory chooses an optimal capital structure based on a balance between debts and equities by trading off the benefits of debt (taxes reduction) against the costs of debt (bankruptcy costs) [Frank and Goyal, 2011]. As debt increases, risk of a firm raises; however, the expected return grows up as well. On the other hand, the pecking order theory claims that because of the costs of information asymmetry, insiders have more information about investment opportunities, return, risk, and value of the firm than outsiders. Accordingly, financing of a firm will come from internal funds as the first priority, then debt and followed by an issue of new equity [Myers, 1984].

The two well-known optimal capital structure theories, the trade-off theory and pecking order theory, hypothesize that capital structure which is expressed by leverage level is affected by some unobserved determinants. The quantity of funds that a firm aims to borrow from financial institutions or investors depends on changes of determinants of capital structure. In order to determine the firm's capital structure, determinants of capital structure should be recognized since they have an influence on an appropriate proportion of debt and equity on capital structure. Table 2.1 shows the theoretical prediction of direction-relation between debt ratio and determinants of capital structure from the trade-off theory and the pecking-order theory [Mazur, 2007].

Table 2.1: Predicted relationship between debt ratio and latent variable. Note that + shows a positive relationship, - shows a negative relationship, and blank space shows that relation is not provided in the theory.

Determinants of capital structure	The trade-off theory	The pecking order theory
Assets structure	+	-
Profitability	+	-
Growth	-	+
Liquidity	+	-
Size	+	+/-
Uniqueness	-	
Business risk	-	
Non-debt tax shields	-	
Business risk		-
Dividend policy		+

The comparison of the direction-relation between empirical results and the theoretical results provides an explanation of firms' financing behavior. It is beneficial to policy makers in the sense that they obtain a policy guide to determine a suitable policy to decrease bankruptcy and information asymmetry problems [Bany-Ariffin and Jr, 2012]. In this work, we provide an empirical relationship between debt ratios and determinants of capital structure from our formulation applied to real data in section 5.4.

2.2 Measures of capital structure

Since the capital structure expresses debt and equity trade-off of the firm that is described by debt-to-equity ratios [Swanson et al., 2003], in this work, we apply debt-to-equity ratios to be the measures of capital structure. The debt-to-equity ratios we consider here are total debt divided by total equity, long-term, short-term, and convertible debt divided by market and by book value of equity as summarized by [Titman and Wessels, 1988].

Short term debts, *e.g.*, short-term bank loans, accounts payable, and long term debts, *e.g.*, notes payable, bonds payable, are debts that their maturity for repaying principal with interest dues within one year and more than one year or beyond the current business year, respectively. Convertible debt is a bond that can be converted to stock. Firms issue this bond to avoid the situation where the market comprehends an overvalued perspective of the firm's stock price when the firm decides to issue stock [Whitehurst, 2003].

The market value of equity can be measured by market capitalization which is a product of the number of outstanding shares of the firm and a current share price. The book value of equity is the value of business based on equity of stockholders in a financial statement. It is equal to a difference between assets and liabilities of a firm. In general, since the market value of equity does not express capital resources of the firm, financial institutions do not lend money based on market value of equity; however, it is beneficial for investors to notice the development and size of the firm.

The performance of the firm is indicated by debt-to-equity ratios in the sense of evaluating a firm's potentiality for repaying a commitment. Low debt-to-equity ratios (less than one) are preferred for investors since this means equity is greater than debt, *i.e.*, greater protection to their business. Conversely, investors do not prefer high debt-to-equity ratios (greater than one) because equity is less than debt. Knowing what causes affect debt-to-equity ratios is beneficial for a firm to manage the capital.

2.3 Determinants of capital structure

In order to determine a decision for capital structure, firms need to recognize the determinants of capital structure. The theoretical determinants of capital structure are sometimes determined by unobserved factors that illustrate abstract concepts in each attribute. Latent variables cannot be directly observed but they are explained by some observed variables which have characteristics as same as in the latent variables. The capital structure choice is influenced by determinants of capital structure that are derived from various theories such as the trade-off theory and the pecking order theory. In this work, we apply seven determinants of capital structure that are growth, uniqueness, non-debt tax shields, collateral value of assets, profitability, volatility, and industry classification following the summary in [Chang et al., 2009]. The following section will concisely explain the definition of each determinant and its effect on the capital structure choice.

Growth

Growth is really a meaningful factor to determine the capital structure since to reach the proposes of firms, firms have to grow up which is related to capital decision-making. Because of more investment opportunities, growth firms tend to demand higher fund and external financing, leverage, is always used. Because high growth firms have high cash flow volatility, leverage ratios from their business should be reduce over a period of time leading have a negative relation between growth and debt-ratios. Since MBA, MBE are commonly used to illustrate growth of firms; moreover, cash flow is represented by capital expenditures and research and development, RD/S, CE/TA, GTA, MBA, MBE, and RD/TA are applied to measure the growth of a firm.

- Research and development/sales (RD/S)

  Research and development is a fund provided by a firm in order to improve existing products and systems of the firm for obtaining new and better products. RD/S that is the proportion of research and development to sales illustrates the percentage of the fund which the firm uses in developed activities to total sales.

- Capital expenditure/total assets (CE/TA)

  Capital expenditure is a budget which is used to do new projects for improving the performance of a firm, for example, developing the physical assets, *e.g.*, equipment, buildings, properties. CE/TA indicates how much capital expenditure the firm uses compared with the firm's the total assets.

- Percentage change in total assets (GTA)

  GTA demonstrates a change of a firm's assets over period of time we considered since the assets of a firm are sold and bought overtime. The higher GTA shows the more profits a firm can generate.

- Market-to-book assets (MBA)

  MBA is the ratio of market value of assets and book value of assets. The interpretation of this ratio is that i) MBA is less than one: the current value is less than the started value of a firm, *i.e.*, the worth of a firm is considered to be undervalued, ii) MBA is greater than one: the current value is greater than the started value of a firm, *i.e.*, the worth of a firm is considered to be overvalued, and iii) MBA is equal than one: the worth of a firm cannot be interpreted that it is better or worse than the started value.

- Market-to-book equity (MBE)

  MBE is the ratio of market value of equity and book value of equity. The interpretation of this ratio is the same as MBA.

- Research and development-to-assets ratio (RD/TA)

  RD/TA that is the proportion of research and development to total assets illustrates the percentage of funds which a firm uses in developed activities to total assets.

Uniqueness

[Titman and Wessels, 1988] state that firms' liquidation decision and its bankruptcy status are correlated, therefore, liquidation costs is meaningful to capital structure of the firm. When firms with high level of uniqueness liquidate, customers, workers, and suppliers, respectively, will be difficult to find other alternative products, jobs, and buyer, respectively. This is the reason why debt ratios and uniqueness are expected to have negatively related. To measure the uniqueness of products, RD/S is considered.

- Research and development/sales (RD/S)

  Research and development is a fund provided by a firm in order to improve existing products and systems of the firm for obtaining new and better products. It can measure the uniqueness in the sense that firm RD/S that is the proportion of research and development to sales illustrates the percentage of the fund which the firm uses in developed activities to total sales. It can measure the uniqueness in the sense that the firm with high research and development can always provides new and unique products.

Non-debt tax shields

Non-debt tax shields are a reduction in income taxes due to non-debt quantity, *e.g.*, depreciation expenses, investment tax credits. If non-debt tax shields are large, a firm will have the less debt in capital structure due to the tax benefits of debt financing [DeAngelo and Masulis, 1980]. Besides, [Fama and French, 2002] and [Berger et al., 1997] apply depreciation and investment tax credits, respectively, to represent non-debt tax shields. Consequently, NDT/TA, ITC/TA, and DEP/TA are applied to be indicators of non-debt tax shields.

- Non-debt tax shields/total assets (NDT/TA)

  NDT/TA illustrates the percentage of non-debt tax shields to the total assets of a firm.

- Investment tax credit/total assets (ITC/TA)

  Investment tax credit is an amount of money that is approved by the government to reduce taxes from investment of a firm so that the firm can use such amount of money to reinvest. ITC/TA expresses the percentage of investment tax credit to the total assets of the firm.

- Depreciation/total assets (DEP/TA)

  Depreciation is the decrease in value of tangible assets that has been run out over time, it does not indicate a cash transaction. DEP/TA shows the percentage of depreciation to the total assets of a firm.

Collateral value of assets

Collateral value of assets expresses the estimation of loan collateral that affects capital structure. The more collateral value of assets, the more debt that the firm wants to issue in order to take advantage of the low cost [Myers and Majluf, 1984]. Since inventory, gross plant, and equipment are consider to be collateral value of assets of the firm, IGP/TA is applied to be the indicator of collateral value of assets.

- (Inventory + gross plant and equipment)/total assets (IGP/TA)

  IGP/TA illustrates the proportion of uneasily liquidated properties, *e.g.*, inventory, gross plant, and equipment, to total assets.

Profitability

Profitability is a potentiality of a firm to generate profit which is an important factor to determine capital structure. [Booth et al., 2001] argue that growth of a firm is financed by its profitability which comes from maintaining income as fixing debt ratios to be constant. Consequently, a firm with less profitability will be forced to use debt financing. According to their work, ten developing counties with low debt can generate high profitability. OI/TA and OI/S are considered to be the indicators of profitability.

- Operating income/total assets (OI/TA)

  OI/TA demonstrates how operating income from business operation that a firm can create based on total assets investment.

- Operating income/sales (OI/S)

  OI/S shows return on sales, *i.e.*, the percentage of operating income that a firm can generate based on the total value of sales.

Volatility

Volatility expresses earning variability of firms. An optimal debt level of a firm and its volatility of earning are negatively related since firms with high volatility are always revealed agency and bankruptcy costs leading to not completely use benefits of debt in the capital structure [Bradley et al., 1984]. Four observed factors illustrating volatility are STDGOI, CV(ROA), CV(ROE), and CV(OITA).

- Standard deviation of the percentage change in operating income (STDGOI)

  STDGOI indicates variability from operating income that is realized profit from an operation that is eliminated by operating expenses.

- Coefficient of variation of return on asset (CV(ROA))

  CV(ROA) illustrates the dispersion of ROA which is the proportion of net income to total asset. It shows how much profit firms can generate from their total asset.

- Coefficient of variation of return on equity (CV(ROE))

  CV(ROE) shows the dispersion of ROE which is the proportion of net income to shareholder's equity. It represents how much profit firms can generate from shareholder's equity of such firms.

- Coefficient of variation of operating income divided by total assets (CV(OITA))

  CV(OITA) represents the dispersion of operating income divided by total assets. It illustrates how much operating income firms can generate from their total asset.

Industry

Industry of firms influences how the firms manage their capital structure. [Titman, 1984] state that high cost of liquidation is found in firms that produce equipment and machine, therefore, these firms are less to finance with debt. two-category dummy variable (IND) is used to separate industry classification. IND is equal to one for firms that produce equipment and machine, and equal to zero otherwise.

- Two-category dummy variable (IND)

  In this research, the dummy variable IND is equal to one for firms that produce equipment and machine, and equal to zero otherwise.

2.4 Relationships of variables in capital structure

As we mentioned, determinants of capital structure affect a proportion of debt and equity on capital structure; thereby, studying the relationship between measures and determinants of capital structure is significant in the sense that it helps managers to perform an appropriate capital structure decision. This section provides literature reviews about techniques for finding the relationship between measures and determinants of capital structure.

Ordinary least squares (OLS) method is generally applied to search for the relationship between measures and determinants of capital structure under the assumption that each determinant of capital structure (latent variable) is explained by a single observed variable. For example, [Deesomsak et al., 2004] investigate the determinants of capital structure of firms operating in the Asia Pacific region, in four countries, including, Thailand, Malaysia, Singapore and Australia. They apply firm leverage to be the dependent variable and predictors consist of the tangibility, profitability, firm size, growth opportunity, non-debt tax shield, the liquidity, earnings volatility/risk, share price performance. [Matemilola et al., 2013] examine the significance of latent variables in firm-specific effects. They apply total debt and long term debt to be measures and fixed assets, profit, size, growth opportunity, and non-debt tax shield to be determinants of capital structure. [Hardiyanto et al., 2015] investigate determinants of capital structure and ownership in public listed companies in Indonesia by using debt/asset ratio as a measure of capital structure and its determinants are total assets, fixed assets, tax shield from interest expenses, net cash flow volatility, interest expenses, and intangible assets. [Serghiescu and Văidean, 2014] also examine the determinants of the capital structure decisions for Romanian firms listed by applying debt/asset ratio as measures of capital structure and profitability, size, tangibility of the assets, liquidity of the assets, and asset turnover as determinants.

However, using OLS method faces many problems such as multicollinearity in predictor variables and violation of error assumptions, *e.g.*, error of variables may be correlated. [Titman and Wessels, 1988] specify issues for applying OLS as follows: i) sets of latent variables induced by researchers are not unique so they select variables based on statistical goodness-of-fit criterions leading to bias explanation, ii) the representation of a single observed variable to each latent variable is imperfect since, in fact, there are many observed variables which fulfill a latent variable representation, and iii) the disturbance of dependent and independent variables may be correlated.

Consequently, they apply SEM approach to find a relationship between measures and determinants of capital structure since SEM allows theoretical determinants to have several observable variables as indicators without multicollinearity problems; moreover, this technique can control measurement error. However, there are too many latent variables compared to observed variables; as a result, they have to add more constraints to get an identifiable model leading to an estimated solution that misses the relevant relations to its true value and many coefficients that indicates the relationship among observed and latent variables are statistically insignificant. They suggest that the results can be improved by finding variables which have stronger linkage between observed and latent variables.

To solve the problem in [Titman and Wessels, 1988], [Chang et al., 2009] apply the reduced form of MIMIC model (1.3) to find the relationship between indicators and causes of determinants of capital structure but this scheme may not work in because it neglects the effect from latent variables. They rank a relative impact between x and y and assume that the most effective x in each η shows the relationship between such η and y. Consequently, we propose two formulations to estimate A and B from reduced MIMIC model and remove some ineffective latent variables. The detail is provided in section 4.

CHAPTER III

BACKGOUND ON MIMIC MODEL

This chapter provides background on MIMIC model including MIMIC model identifiability that explains uniqueness of model parameters, MIMIC model estimation that shows various ways of parameters estimation in the MIMIC model, and latent variable reduction which provides some methods to remove ineffective latent variables in the model.

3.1 MIMIC model identifiability

When a MIMIC model is proposed, the uniqueness of model parameters in implied covariance matrix of x and y of the model based on sample covariance matrix should be identified. This section explains the identifiability of MIMIC model about how to conclude the uniqueness of model parameters and some restrictions for model identification which can be added when the model is not unique. Firstly, we describe implied covariance matrix to investigate the structure of model parameters.

Implied covariance matrix

The model implied covariance matrix, Σ, for x and y written as a function of free model parameters in θ is derived form (1.3) as

$$\begin{aligned}\Sigma &= \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix} = \begin{bmatrix}\mathbf{E}[xx^T] & \mathbf{E}[x(Fx+\epsilon)^T]\\ \mathbf{E}[(Fx+\epsilon)x^T] & \mathbf{E}[(Fx+\epsilon)(Fx+\epsilon)^T]\end{bmatrix}\\ &= \begin{bmatrix}\Phi & \Phi F^T\\ F\Phi & F\Phi F^T+\Psi\end{bmatrix} = \begin{bmatrix}\Phi & \Phi BA^T\\ AB^T\Phi & AB^T\Phi BA^T + A\Psi_\eta A^T + \Psi_y\end{bmatrix}\end{aligned} \tag{3.1}$$

where Φ is the covariance matrix of x and Ψ is the covariance matrix of ϵ. Additionally, Ψ_η and Ψ_y are covariance matrices of ϵ_η and ϵ_Y, respectively [Bollen, 2014, §4].

Given the sample covariance matrix of (x,y) which can be calculated from observations, sample covariance (S) is partitioned as

$$S = \begin{bmatrix}S_{xx} & S_{xy}\\ S_{yx} & S_{yy}\end{bmatrix}$$

When we match the sample covariance, S, to the implied covariance matrix, Σ, that we get from MIMIC model, the sample covariance matrix which can be observed is the function of the parameters and perfectly estimated since the best estimator of Σ is S. In this case, we obtain

$$\begin{aligned} S &= \Sigma \\ \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} &= \begin{bmatrix} \Phi & \Phi F^T \\ F\Phi & F\Phi F^T + \Psi \end{bmatrix} = \begin{bmatrix} \Phi & \Phi B A^T \\ A B^T \Phi & A B^T \Phi B A^T + A\Psi_\eta A^T + \Psi_y \end{bmatrix} \end{aligned} \tag{3.2}$$

According to (3.2), if Φ, F, Ψ are the model parameters with condition $S = \Sigma(F, \Phi, \Psi)$, the unique explicit solutions are obtained as below:

$$\begin{aligned} \Phi &= S_{xx} \\ F &= S_{yx} {S_{xx}}^{-1} \\ \Psi &= S_{yy} - {S_{xx}}^{-1} S_{xy} S_{yx}. \end{aligned} \tag{3.3}$$

To examine the model identification when $\Sigma = S$ is to consider two vectors of unknown parameters, θ_1 and θ_2, then construct the implied covariance matrices, Σ_1 and Σ_2. If the model is identified, then for $\Sigma_1 = \Sigma_2$, $\theta_1 = \theta_2$. If $\Sigma_1 = \Sigma_2$ and $\theta_1 \neq \theta_2$, then the model is unidentified [Bollen, 2014, §4].

In order to be easy to understand, we utilize the value of degree of freedom (*df*) that is denoted as below [Raykov and Marcoulides, 2012, §1]:

$$\begin{aligned} df = \; & \text{the number of equations relating the elements of the sample covariance matrix, } S \\ & \text{- the number of parameters in implied covariance matrix, } \Sigma \end{aligned} \tag{3.4}$$

The necessary condition for model identification is nonnegative *df*, *i.e.*, when the model is identified, *df* is nonnegative because the sample covariance matrix, S, provides enough information to solve for the parameters. In contrast, if the model is unidentified, then *df* is negative since Σ contains more parameters than equations that leads to a lack of information in the sample covariance matrix, S [Raykov and Marcoulides, 2012, §1].

According to (3.2), for $\theta = (F, \Phi, \Psi)$, the positive *df* is provided, *i.e.*, the number of equations are not less than the number of parameters. In sum, F is obtained as (3.3). On the other hand, for $\theta = (A, B, \Phi, \Psi_\eta, \Psi_y)$, there is extremely high chance which the number of equations that is $\frac{(p+q)(p+q+1)}{2}$ are less than the number of unknown parameters, $\Psi, A, B, \epsilon_\eta$ and ϵ_y, that leads to negative *df*, so this situation provides the non-uniqe solution of the unknown parameters.

When the unidentified model happens, we cannot interpret the value of parameters; consequently, two common ways to restrict necessary for providing nononegative *df* are i) setting some parameters to zero or some constant in order to reduce the unknown parameters, in other words, to change the model to have positive *df* and ii) setting coefficient matrix of ϵ or error in MIMIC model to identity matrix that means each ϵ shows in only one equation with a coefficient of one. Moreover, we have to scale the latent variables for interpretability by determining the variance of latent variables to constant or scaling it to one of the observed variables [Bollen, 2014, §4]. As a result, two restrictions will turn the unidentification model to identification model [Stapleton, 1978, Dell'Anno et al., 2007, Gallo et al., 1994].

3.2 MIMIC model estimation

There are various ways of parameters estimation in MIMIC model. One way is to apply SEM formulation since the goal of this optimization problem is to minimize Kullback-Liebler (KL) divergence function which represents the distance between the sample covariance matrix, S, and the model implied covariance matrix Σ with a special structure defined in (3.2). Another way to estimate the parameters of the model is called maximum likelihood estimation that is considered to be the popular way to search for parameters of a statistical model such that the statistical likelihood function is maximized. The least squares method is also considered in the sense that minimizes the difference between the predictors and the dependent observed variables, *i.e.*, residues of the model [Jamesh et al., 2013]. Last but not least, regularized least-squares estimation is the method that applies the regularization and the least squares method, *i.e.*, the solution from this procedure is estimated like the least square method but some elements are zero due to the regularization term. In this section, we will explain the typical interesting estimation methods which we mentioned above to search for the parameter F in (1.3).

3.2.1 SEM formulation

Let $\theta = (F, \Phi, \Psi)$ be parameters of model. The Structural Equation Modeling fitting function is presented by [Jöreskog, 1970] under multivariate normal distribution variables with covariance-based. The principle of this method is that the implied covariance matrix, $\Sigma(\theta)$ should be closed to the actual sample covariance matrix, S with regard to the Kullback-Leiber divergence which finds the divergence measures between the probability density of the samples and the probability density of the model [Penny et al., 2004] that is equivalent to minimize the objective function of the optimization problem (3.5). Moreover, such function is improved by [Jöreskog and Goldberge, 1975] for a single latent variable and by [Stapleton, 1978] for more latent variables. The optimization problem for the model estimation approach is defined as

$$\begin{array}{ll} \text{minimize} & \log \det\Sigma + \mathbf{tr}(S\Sigma^{-1}) - \log \det(S) - (p+q) \\ \text{subject to} & \Sigma = \begin{bmatrix} \Phi & \Phi F^T \\ F\Phi & F\Phi F^T + \Psi \end{bmatrix} \end{array} \tag{3.5}$$

with variables $\Sigma \in \mathbf{R}^{(p+q)\times(p+q)}$, $F \in \mathbf{R}^{q\times p}$, $\Phi \in \mathbf{R}^{p\times p}$, $\Psi \in \mathbf{R}^{q\times q}$.

3.2.2 Maximum likelihood estimation

Maximum likelihood estimation (MLE) will search for the unknown parameters in the sense that joint density (probability), namely, likelihood function of the data is maximized. When the sample size increase to infinity, maximum likelihood estimators have three main good properties that are i) consistency: the estimators converge in probability to its true unknown parameter (θ_0), ii) asymptotic normality: the maximum likelihood estimators (θ) tend to have normal distribution with mean $= \theta_0$ and covariance matrix is equal to the inverse of the Fisher information called asymptotic variance, and the last property is iii) efficiency: the maximum likelihood estimators have the smallest asymptotic variance among $\sqrt{N}$ consistent estimators or the lower bounded of all possible variance [Cameron and Trivedi, 2005].

Denote the vectors of independent random variables, $x \in \mathbf{R}^p$ and $y \in \mathbf{R}^q$, have N observations as ordered pairs $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ where x and y are jointly $(p+q)$-dimensional Gaussian vector as $\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\Bigg(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \Bigg)$.

The sample covariance matrix of (x, y) which can be calculated from sample measurement is defined as

$$S = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} x^{(i)} - \mu_x \\ y^{(i)} - \mu_y \end{bmatrix}^T \begin{bmatrix} x^{(i)} - \mu_x \\ y^{(i)} - \mu_y \end{bmatrix}.$$

The joint probability density function of N independent observations of (x, y) is

$$\begin{aligned} f(y, x|\theta) &= f(y^{(1)}, x^{(1)}|\theta) f(y^{(2)}, x^{(2)}|\theta) \ldots f(y^{(N)}, x^{(N)}|\theta) \\ &= \frac{1}{(2\pi)^{N(p+q)/2} (\det\Sigma)^{N/2}} \exp\Bigg(-\frac{1}{2} \sum_{i=1}^{N} \begin{bmatrix} x^{(i)} - \mu_x \\ y^{(i)} - \mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x^{(i)} - \mu_x \\ y^{(i)} - \mu_y \end{bmatrix} \Bigg). \quad (3.6) \end{aligned}$$

Since a log function is a monotonic function, to maximize likelihood function is equable to maximize log-likelihood function that is

$$\begin{aligned} \log f(y, x|\theta) &= -\frac{1}{2} \sum_{i=1}^{N} \begin{bmatrix} x^{(i)} - \mu_x \\ y^{(i)} - \mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_i - \mu_x \\ y_i - \mu_y \end{bmatrix} - \frac{N}{2} \log \det\Sigma - \frac{N(p+q)}{2} \log 2\pi \\ &= -\frac{N}{2} \Bigg(\mathbf{tr}\, S\Sigma^{-1} + \log \det\Sigma \Bigg) - \frac{N(p+q)}{2} \log 2\pi. \quad (3.7) \end{aligned}$$

Knowing that constant term does not affect to optimization problem and the $(N/2)$ term has no influence on the choice of θ, we can consider maximizing log-likelihood function (3.7) as minimizing the fitted function (3.8) [Bollen, 2014, §4]:

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, S\Sigma^{-1} + \log \det\Sigma \\ \text{subject to} & \Sigma = \begin{bmatrix} \Phi & \Phi F^T \\ F\Phi & F\Phi F^T + \Psi \end{bmatrix} \end{array} \quad (3.8)$$

with variables $\Sigma \in \mathbf{R}^{(p+q)\times(p+q)}$, $F \in \mathbf{R}^{q\times p}$, $\Phi \in \mathbf{R}^{p\times p}$, $\Psi \in \mathbf{R}^{q\times q}$.

We can notice that the optimization problem (3.8) is equivalent to (3.5). The solutions of the problem (3.5) and (3.8) are provided when the objective function is zero, *i.e.*, $\Sigma = S$. As a result, if $\Sigma = S$ has solutions, such solutions given by (3.3) are the solutions of the problems (3.5) and (3.8).

3.2.3 Least-squares estimation

Let $x \in \mathbf{R}^p$, and $y \in \mathbf{R}^q$ be independent random variables and $(x^{(i)}, y^{(i)})$ be i^{th} sample data where $i = 1, \ldots, N$. Denote $X = \begin{bmatrix} x^{(1)} & \cdots & x^{(N)} \end{bmatrix} \in \mathbf{R}^{p \times N}$ and $Y = \begin{bmatrix} y^{(1)} & \cdots & y^{(N)} \end{bmatrix} \in \mathbf{R}^{q \times N}$ be matrices with $x^{(i)}$ and $y^{(i)}$, respectively, as their i^{th} column. The reduced MIMIC model (1.3) is written as

$$Y = FX + E \tag{3.9}$$

For least square error method, in order to estimate the coefficient F representing the relationship between observed variables (X, Y) from MIMIC model (3.9) with N observations, we can think that FX based on the value of X be the prediction of dependent variable Y, then the residue (E) is the difference between Y and FX that is $E = Y - FX$. Since we want FX to be closed to Y, in other words, to minimize the squares of the Frobenius norm $(\|.\|_F^2)$ of errors. Sum square errors which are the sum of the squares of the deviations of the actual values from the predicted values for regression problem is needed to minimize so we have the objective function as follows:

$$\text{minimize} \quad \|Y - FX\|_F^2 \tag{3.10}$$

with variable $F \in \mathbf{R}^{q \times p}$.

$$F = \left(\frac{YX^T}{N}\right)\left(\frac{XX^T}{N}\right)^{-1} = S_{YX}{S_{XX}}^{-1} \tag{3.11}$$

Consequently, F which is predicted by maximum likelihood function (3.8) and by least square error (3.11) is the same when we assume that the observed variables (x, y) are normally distributed.

3.2.4 Regularized least-squares estimation

In case of high correlation of the two variables x and y, for the value of coefficient $F_{ij} \in F$ for $i \in 1, 2, ..., q$ and $j \in 1, 2, ..., p$ representing the relationship between two variables (x, y), the high variance model is constructed because the coefficient of one variable that has wildly positive value can compensate with the highly negative coefficient on the other variable [Friedman et al., 2001]. Regularization or shrinkage method that shrinks the coefficient towards zero using least-squares to fit are the alternative method to reduce the mentioned problem. ℓ_2-regularization and ℓ_1-regularization that are the well-known techniques of regularization method [Jamesh et al., 2013, §10]. They resemble to least-squares problem (3.9), excepting adding the ℓ_2-norm and ℓ_1-norm penalty terms, respectively, for promoting the entries in F toward zero denoting by the problems (3.12) and (3.14) where $\gamma \geq 0$. When γ = 0, the penalty term has no impact to F but when γ goes to infinity, the coefficient estimated F will approach to zero. In this paper, we define the notation of (a, b)-norm of matrix as $\|F\|_{a,b} = \left(\sum_{j=1}^{p} \left(\sum_{i=1}^{q} \|F_{ij}\|^b \right)^{a/b} \right)^{1/a}$.

ℓ_2-Regularization

The ℓ_2-regularization or *ridge regression* estimates F by the following objective function:

$$\text{minimize} \quad \|Y - FX\|_F^2 + \gamma\|F\|_{1,2}^2 \tag{3.12}$$

with variable $F \in \mathbf{R}^{q \times p}$.

The advantage of ℓ_2-regularization over least-squares is based on the bias-variance tradeoff [Jamesh et al., 2013, §6]. When γ is high, the relaxibility of model is low, leading to decreased variance but increased bias; accordingly, we will search for γ that provides the minimum Mean Square Error (MSE) of the model. However, the interpretation may have problems if we want the zero structure in F since the main drawback of ℓ_2-regularization is that its penalty will shrink F towards zero but not exactly zero; therefore, ℓ_1-regularization is chosen to solve such problem. For unconstrained convex optimization (3.12), the closed form solution of this problem is calculated by differentiation with respect to F and set those to zero [Boyd and Vandenberghe, 2004, §6]that is

$$F = YX^T(XX^T + \gamma I)^{-1} = S_{YX}\left(S_{XX}{}^{-1} + \frac{NI}{\gamma}\right) \tag{3.13}$$

ℓ_1-Regularization

The ℓ_1-regularization often called *lasso* estimates F by the following objective function:

$$\text{minimize} \quad \|Y - FX\|_F^2 + \gamma\|F\|_{1,1} \tag{3.14}$$

with variable $F \in R^{q \times p}$.

If we choose sufficiently large γ, the influence of ℓ_1-penalty will force some elements of F to be exactly zero that simpler and easier to interpret than ℓ_2-regularization. However, because the convex optimization problem (3.14) is non-differentiable when F_{ij} equal zero, the closed form solution cannot be attained like the way we do with ℓ_2-regularization. Consequently, the constrained formulation (3.15), most conscientious problem to demonstrate the problem (3.14), is considered.

$$\begin{array}{ll} \text{minimize} & \|Y - FX\|_F^2 \\ \text{subject to} & \|F\|_{1,1} \le t \end{array} \tag{3.15}$$

with variable $F \in \mathbf{R}^{q \times p}$.

Various techniques are used to solve the convex optimization problem (3.15) since 1996 [Schmidt, 2005]; for example, converting the constrain in the problem (3.15) into a set of linear constrains [Tibshirani, 1996], Interior point method and non-negative variables with log barrier [Chen et al., 2001] [Sardy et al., 2000], Active set method and Local linearization [Osborne et al., 2000b] [Osborne et al., 2000a], Iterated Ridge Regression [Fan and Li, 2001], Grafting [Perkins et al., 2003], Gauss-Seidel [Shevade and Keerthi, 2003], and also Shooting method [Fu, 1998].

3.3 Matrix factorization in MIMIC model

According to the section 3.2, the solution of MIMIC model, F, which is the coefficient matrix representing the indirect effect relationship between x and y can be found by several techniques, including, SEM formulation, maximum likelihood estimation, least-squares estimation, and regularized least-squares estimation. However, in many applications, finding A and B that expresses the coefficient matrices of direct effect of x to η and η to y, respectively, is more significant than finding F since beneficial interpretation from A and B will be presented. Therefore, in this section, as we know that $F = AB^T$, we will search for the decomposed matrices A and B from F which is known from the section 3.2 by applying interesting methods that provide the different advantages for the explanation of the solutions, including, i) adding linear constraints in coefficient matrices representing the relationship between observed and latent variables which is common approach in MIMIC model, ii) sparse factorization, and iii) nuclear norm regularization in rank minimization problem that are available methods for matrix factorization.

3.3.1 Adding linear constraints on matrix factors

According to the model identification, in case that perfect model fit to the data, if $S = \Sigma(F, \Phi, \Psi)$ is regarded, the explicit uniquely solution is provided as shown in (3.3). On the contrary, when we consider $S = \Sigma(A, B, \Phi, \Psi_\eta, \Psi_y)$, non-unique solutions may be provided because of negative *df*. Thereby, some parameters should be set to be constant, saying zero, *e.g.*, adding linear constraints in A and B, in order to remove indeterminacy leading to having nonnegative *df*.

Focusing on finding A and B from $S = \Sigma(A, B, \Phi, \Psi_\eta, \Psi_y)$, to get nonnegative *df*, the number of equations relating the elements of S must be greater than or equal to the number of unknown parameters in Σ, in other words, $\frac{(p+q)(p+q+1)}{2} \geq qm + pm + p^2 + m^2 + q^2$ so some elements have to be set to constant based on suitability of data, then the unknown parameters can be found by solving (3.2). For example, [Stapleton, 1978] assume that B is dense and set some parameters of A to zero, *e.g.*, for the first column of A that relates to the first latent variable (η_1), there are $A_{1,1}$ to $A_{5,1}$ that have the relationship with η_1 so they set other entries in this column to zero and do the same method with the other columns.

3.3.2 Sparse factorization

According to the problem (3.15), since we want $F \in \mathbf{R}^{q \times p}$ to sparse and can be factorized into AB^T where $A \in \mathbf{R}^{q \times r} = \begin{bmatrix} A_1 & A_2 & \cdots A_r & \end{bmatrix}$ and $B \in \mathbf{R}^{p \times r} = \begin{bmatrix} B_1 & B_2 & \cdots & B_r \end{bmatrix}$, our procedures are to find $Z = AB^T \in \mathbf{R}^{q \times p}$ which has zero structure by applying ℓ_1-regularization and is close to observed variable F from the first process. [Bach et al., 2008] suggest the sparse composition norms by using sparse factorization. Given an observed matrix $F \in \mathbf{R}^{q \times p}$ and we search for factorization form $Z = AB^T$ that is close to F where A or B is sparse, *i.e.*, penalizing each column of A and B by the following objective function:

$$\text{minimize} \quad \sum_{i=1}^{q}\sum_{j=1}^{p}(F_{ij}-(AB^T)_{ij})^2+\gamma\sum_{k=1}^{r}(\|A_k\|_C^2+\|B_k\|_R^2) \tag{3.16}$$

with variables $A \in \mathbf{R}^{q\times r}, B \in \mathbf{R}^{p\times r}$ where $\|\cdot\|_C$ and $\|\cdot\|_R$ are any norms on $\mathbf{R}^q$ and $\mathbf{R}^p$ (on the column space and row space of original matrix Z).

Let m or number of latent variables grow to infinity and $\|Z\|_D = \|A_k\|_C^2 + \|B_k\|_R^2$ we can consider the below problem (3.17) as a convex optimization problem which is equivalent to the problem (3.16)

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{p}(F_{ij}-Z_{ij})^2+\gamma\|Z\|_D \tag{3.17}$$

with variables $Z \in \mathbf{R}^{q\times p}$.

[Bach et al., 2008] provide the closed form solution of Z in (3.17) as

$$Z(i,:)=\max\{\|F(i,:)^T\|_2-\gamma,0\}\frac{F(i,:)}{\|F(i,:)^T\|_2}$$

when $\|\cdot\|_C = \|\cdot\|_1$ and $\|\cdot\|_R = \|\cdot\|_2$ for $i \in \{1,...,q\}$ and

$$Z(i,j)=\max\{|F(i,j)|-\gamma,0\}\frac{F(i,j)}{|F(i,:)|}$$

when $\|\cdot\|_C = \|\cdot\|_1$ and $\|\cdot\|_R = \|\cdot\|_1$ for $i \in \{1,...,q\}$ and $j \in \{1,...,p\}$.

Sparse factorization has been proposed in a variety of the approaches that have been discussed for factorizing a matrix into a product of two matrices. For example, [Mairal et al., 2010] provide an online stochastic optimization algorithm based on a stochastic approximation for dictionary learning that applies to sparse coding. Besides, by inviting this process to non negative matrix factorization and sparse principal component analysis formulation, the results are practically answered. The main advantages of this work are that the process of the algorithm is fast and large data set can be considered. [Zhang et al., 2012] propose sparse principal component analysis that seeks for a principal component which is fixed the number of nonzero coefficients by applying orthogonal transformation and maximizing variance in the data. However, this problem is difficult to solve; therefore, they generate the better estimated version by using the convex relaxation method, including, relaxation with ℓ_0 and ℓ_1 penalization. [Gillis, 2012] discuss about nonnegative matrix factorization (NMF) based on nonnegative data matrix procedure since it has properties to extract significant characteristics that are very useful in machine learning field. In this work, they provide the sparser solution of the better aspect of NMF problems. Moreover, the algorithm of sparse matrix factorization for linear version and investigate its relationship under randomness and sparsity assumptions is examined by [Neyshabur and Panigrahy, 2013]. In deep learning network's perspective, searching for values of hidden units and edges in different layers coincide with seeking for matrix factorization. [Richard et al., 2014] illustrate tight convex relaxations of sparse factorization for low-rank matrices estimation by assuming we know the non-zero entries in the matrix we want to factorize. They suppose that the matrix which they want to factorize is a product of k non-zero entries column vector and q non-zero entries row vector. Then, they define the (k, q)-rank of a matrix and relax such matrix producing convex relaxation of (k, q)-trace norm in order to be easier to manage.

3.4 Latent variable reduction

Researchers often introduce plenty of latent variables to the model. However, such latent variables may contain hardly important latent variables leading to non-significant results. Consequently, we aim to decrease the number of latent variables (r) until remaining highly significant latent variables (m), in other words, we reduce columns of A and B from r to $m.$ This section provides the common techniques to select highly effective latent variables based on rank minimization, *i.e.*, nuclear norm regularization, sparse latent semantic analysis, and based on a sequential selection.

Rank minimization

As previously mentioned, we require to find A and B from $F = AB^T$ that explains the direct effect between observed and latent variables, sometimes, we need the highly significant latent variables from all latent variables we investigate.

Suppose we want the most m important latent variables, *i.e.*, A and B have m columns, there are six cases that matrices A and B are possible to be (see Figure 3.1). There are only cases 1.) and 4.) that $q > m$ and $p > m$, *i.e.*, the number of latent variables are less than the number of observed variables for both A and B; however, for the others, the number of latent variables is greater than the number of observed variables for A or B.

In the case that $m < \mathbf{rank}(F)$, this cannot be occurred since $\mathbf{rank}(F) = \mathbf{rank}(AB^T) \leq \min\{\mathbf{rank}(A), \mathbf{rank}(B)\}, \mathbf{rank}(A) \leq \min\{q, m\}$ and $\mathbf{rank}(B) \leq \min\{p, m\}$; therefore, $\mathbf{rank}(F)$ $\leq \min\{q, p, m\}$. Because we want only a few essentially important latent variables, we need the structure of A and B like items 1.) and 4.), *i.e.*, A and B are skinny matrices. However, at the beginning, the rank of F may be equal $\min\{q, p\}$ (full rank) that leads to the cases of 2.), 3.), 5.), and 6.) so we have to reduce the rank of F from finding the low rank matrix Z that is close to F by applying the rank minimization problem.

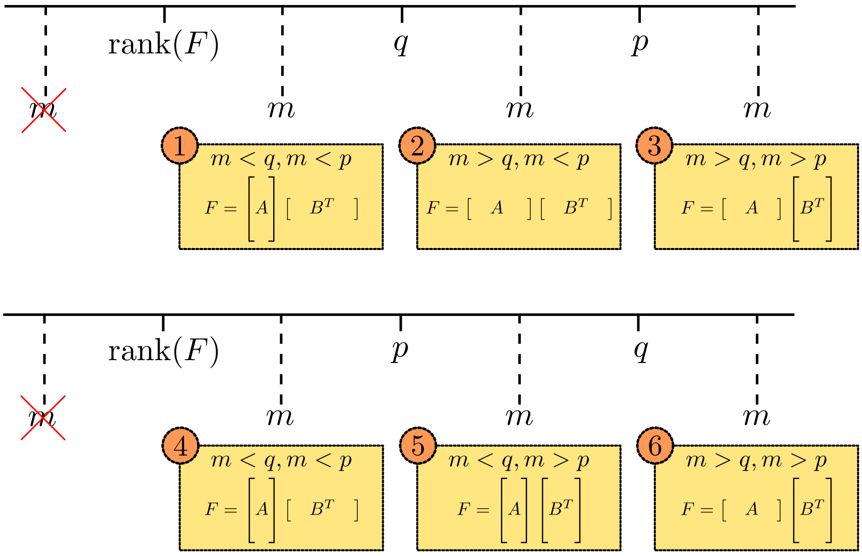


Figure 3.1: Possible A and B from matrix factorization.

Singular Value Decomposition (SVD)

When we want to approximate the low rank matrix Z that is close to the known matrix F, the singular value decomposition is a good application for estimation since this choice can be applied for square matrix and rectangular matrix; moreover, we can determine for every rank of Z. The singular value decomposition (SVD) is a factorization of matrix $F \in \mathbf{R}^{q\times p}$ that has rank r as

$$F = UDV^T = \sum_{i=1}^{r} d_i u_i v_i^T$$

where $D \in \mathbf{R}^{r\times r}$ is a diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_r > 0$ that are the singular values of F, corresponded to the rank of F, *i.e.*, the number of positive singular values that are calculated by $d_i = \sqrt{\lambda_i(FF^T)}$, *e.g.*, the eigenvalues of FF^T. $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_r \end{bmatrix} \in \mathbf{R}^{q\times r}$ and $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix} \in \mathbf{R}^{p\times r}$ are unitary matrices, called a left and a right singular vector matrices, respectively. Besides, the columns of U and V are the eigenvectors of FF^T and F^TF, respectively [Horn and Johnson, 2012, §7].

The singular values of F are unique but the singular vector matrices, U and V, are not unique. However, U and V are unique in the case that the singular values are all different. If there are some identical singular values, the singular vector matrices are not unique since the subspace from spanning of coincident singular vectors can be applied as the singular vectors.

In this part, we are talking about rank minimization problem that the objective is to find the low rank matrix $Z \in \mathbf{R}^{q\times p}$ that represents the observed data matrix $F \in \mathbf{R}^{q\times p}$ and has generalization training error bound, $\delta \geq 0$ as follows:

$$\begin{array}{ll} \text{minimize} & \textbf{rank}(Z) \\ \text{subject to} & \sum_{i,j} (Z_{ij} - F_{ij})^2 \leq \delta \end{array} \tag{3.18}$$

with variable $Z \in \mathbf{R}^{q \times p}$. However, it is difficult to solve the problem (3.18) due to the non-convexity of the cost objective function. A convex relaxation of the low rank optimization (3.18) is represented by [Fazel, 2002] who provide the nuclear norm optimization problem, *i.e.*, the nuclear norm, $\|Z\|_*$, is the sum of the singular values of Z as follow:

$$\begin{array}{ll} \text{minimize} & \|Z\|_* \\ \text{subject to} & \sum_{i,j}(Z_{ij} - F_{ij})^2 \leq \delta \end{array} \tag{3.19}$$

with variable $Z \in \mathbf{R}^{q \times p}$. We notice that the left side of a constraint of the problem (3.19) is bounded by some positive constant, in other words, we want to regularize its value to be less than or equal to δ.

3.4.1 Nuclear norm regularization

A nuclear norm regularization which is convex relaxation for the rank minimization problem is widely used for low rank matrix approximation, *i.e.*, we want to find low rank matrix Z that closely estimates the data matrix F. Moreover, nuclear norm regularization is the Lagrange version of the problem (3.19) as

$$\text{minimize} \quad \frac{1}{2}\|Z_{ij} - F_{ij}\|_F^2 + \gamma\|Z\|_* \tag{3.20}$$

with variable $Z \in \mathbf{R}^{q \times p}$ and $\gamma \geq 0$ is a regularization parameter.

A more general form for the problem (3.20) is suggested by [Mazumder et al., 2010] who find Z from the problem (3.20) for $(i, j) \subset \Omega$ where $\Omega \subset \{1, ..., q\} \times \{1, ..., p\}$ expresses the indices of observed entries in F by the following problem:

$$\text{minimize} \quad \frac{1}{2}\|P_\Omega(Z) - P_\Omega(F)\|_F^2 + \gamma\|Z\|_* \tag{3.21}$$

with variable $Z \in \mathbf{R}^{q \times p}$ and $\gamma \geq 0$ is a regularization parameter and the orthogonal projector P onto the span of matrices missing outside of Ω has the $(i, j)^{\text{th}}$-component of $P_\Omega(W)$ equals W_{ij} when $(i, j) \in \Omega$ and zero otherwise. The solution of nuclear norm regularization is given by lemma 1 [Mazumder et al., 2010].

Lemma 1. *[Nuclear norm regularization] Suppose the matrix $F \in \mathbf{R}^{q \times p}$ has rank r, SVD of F is $F = UDV^T$ and $D = \mathbf{diag}\left[d_1, ..., d_r\right]$. The solution to the optimization problem*

$$\underset{Z}{\text{minimize}} \quad \frac{1}{2}\|Z - F\|_F^2 + \gamma\|Z\|_* \tag{3.22}$$

is given by $Z = UD_\gamma V^T$ with $D_\gamma = \mathbf{diag}\left[\max\{d_1 - \gamma, 0\}, ..., \max\{d_r - \gamma, 0\}\right]$.

Saying that the solution of nuclear norm minimization problem is the closed form which is similar to the SVD of F excepting the D, *i.e.*, applying D_γ instead where D_γ is a diagonal matrix with the soft-threshold function diagonal elements that are the maximum of $d_i - \gamma$ and zero for $i = 1, \ldots, r$.

Now, the low rank matrix Z is provided by lemma 1, we sometimes want to factorize Z into the product of two matrices, *e.g.*, $Z = AB^T$ in order to interpret each composition, A and B where the number of columns in A and B or the number of the factors equals to the rank of F. This idea is suggested by lemma 2 [Srebro et al., 2005] [Mazumder et al., 2010].

Lemma 2. *[Low rank factorization] For any matrix Z the following holds:*

$$\begin{aligned} \|Z\|_* = &\quad \mathrm{minimize}(1/2)(\|A\|_F^2 + \|B\|_F^2) \\ \mathrm{subjectto} &\quad Z = AB^T \end{aligned} \tag{3.23}$$

with variable $Z \in \mathbf{R}^{q\times p}$, $A \in \mathbf{R}^{q\times m}, B \in \mathbf{R}^{p\times m}$. If $\mathbf{rank}(Z) = m \leq \min\{q, p\}$, then the minimum above is attained at a factor decomposition $Z = AB^T$.

When the factorization from lemma 2 is occurred, the rank of Z corresponds to the number of latent variables, m; consequently, we can reduce the number of all considered latent variables until it reaches our desired number of latent variables, m, by specifying the rank of Z to be m. According to lemma 2, we can consider (3.22) as below optimization problem called "Maximum Margin Matrix Factorization (MMMF)" :

$$\mathrm{minimize} \quad \|F - AB^T\|_F^2 + \gamma(\|A\|_F^2 + \|B\|_F^2) \tag{3.24}$$

with variables $A \in \mathbf{R}^{q\times m}$ and $B \in \mathbf{R}^{p\times m}$.

Generally, the restriction of rank minimization problem is the number of allowed factor or the size of A and B. Instead of dimensional limitation, constraining on the norms of A and B covers the overall strength of the factors rather than their dimension. This restriction is introduced by the method of Maximum Margin Matrix Factorization (MMMF) [Srebro et al., 2005]. In many circumstances, there are a large number of latent variables effecting the observed variables but there are only a few very important factors, saying latent variables. MMMF is the good choice to factorize Z that is close to F to get A and B with low rank constraint of Z since this method can reduce the number of latent variables based on the rank minimization [Rennie and Srebro, 2005].

However, the solutions A and B from this technique need not be unique and we can not identify remaining latent variables since removed latent variables are not specified. Consequently, nuclear norm minimization is not suitable to apply to reduce the number of latent variables for the MIMIC model.

3.4.2 Sparse principal component analysis

Principal component analysis or PCA which was first proposed by [Pearson, 1901] is a common technique for variable reduction. The principle of this method is to reduce the number of variables to lower dimension of new variables by vector space transform ($v \in \mathbf{R}^r$) called principal components correspond to maximum sample variance directions where new variables are a linear combination of initial variables [Friedman et al., 2001, §14]. When the number of variables is high, sparse principal component analysis or sparse PCA is applied to set some elements in principal components to be zero in order to remove some variables [Hastie et al., 2015, §8].

In other words, new lower dimension $\tilde{\eta}$ is a linear combination of latent variables η as

$$\begin{aligned}
\tilde{\eta}_1 &= a_{11}\eta_1 + a_{12}\eta_2 + \cdots + a_{1r}\eta_r \\
\tilde{\eta}_2 &= a_{21}\eta_1 + a_{22}\eta_2 + \cdots + a_{2r}\eta_r \\
\vdots & \quad \vdots \\
\tilde{\eta}_m &= a_{m1}\eta_1 + a_{m2}\eta_2 + \cdots + a_{mr}\eta_r.
\end{aligned}$$

If sparse PCA is applied to latent variable reduction, *i.e.*, we reduce the number of latent variables from r to m by applying sparse PCA that is also explained by maximizing variance of latent variables along a direction of v. For the first principal component the optimization problem is defined as:

$$\begin{array}{ll}
\text{maximize} & v^T \mathbf{cov}(\eta) v \\
\text{subject to} & \|v\|_2 = 1 \\
& \|v\|_0 \leq t
\end{array} \tag{3.25}$$

with variables $v \in \mathbf{R}^r$ where ℓ_0-norm constraint indicates nonzero components limitation in v that is dependent of a nonnegative constant t, saying that there are nonzero entries in v which less than or equal to t.

For finding further i^{th} principal component, optimization problem (3.25) is continuously solved for $i = 1, \ldots, m$ but $\mathbf{cov}(\eta)$ is replaced by

$$\mathbf{cov}(\eta)_{i+1} = \mathbf{cov}(\eta)_i - (v_i^T \mathbf{cov}(\eta)_i v_i) v_i v_i^T.$$

The concept of sparse PCA can be applied to reduce latent variables as we propose; however, we cannot explain the physical meaning of new latent variables which come from a linear combination of initial latent variables. Moreover, according to the problem (3.25), sparse PCA for latent variable reduction requires to calculate $\mathbf{cov}(\eta)$ which is dependent of B and covariance of ϵ_η (from model (1.1)) that are unknown variables. Consequently, performing sparse PCA for latent variable reduction on the MIMIC model may need further analysis. Therefore, we do not explore nor apply this technique to our work.

3.4.3 Sparse latent semantic analysis

Latent semantic analysis (LSA) proposed by [Deerwester et al., 1990] is one of the regular techniques for a matrix dimension reduction. The concept of LSA is to project a high dimensional vector space represented in $Y \in \mathbf{R}^{q\times N}$ to a lower dimensional latent space represented in latent variables $U \in \mathbf{R}^{r\times N}$ where N is a number of observations, q is a number of variables of Y, and r is a number of variables of U. Singular value decomposition (SVD) is applied in LSA to construct a rank r estimation of Y. In other words, $Y \approx V^T DU$ where $U \in \mathbf{R}^{r\times N}$ and $V \in \mathbf{R}^{r\times q}$ are orthogonal matrices and $D \in \mathbf{R}^{r\times r}$ is diagonal matrix. Given that $A = V^T D$ is a projection matrix that maps input feature space to latent space and $\epsilon_Y \in \mathbf{R}^{q\times N}$ be the noise of Y, AU is rank r estimation of Y. The linear relationship between observed variables Y and latent variables U is represented as:

$$Y = AU + \epsilon_Y$$

To obtain a projection matrix (A) and latent variables (U), minimizing error from rank r estimation of Y is considered with the orthogonality constraint on U for independent latent variables. LSA is proposed as following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|Y - AU\|_F^2 \\ \text{subject to} & UU^T = I \end{array} \tag{3.26}$$

with variables $A \in \mathbf{R}^{q\times r}$ and $U \in \mathbf{R}^{r\times N}$.

However, when a small number of latent variables is required for a better interpretation among variables, *i.e.*, highly effective latent variables are needed, [Chen et al., 2011] present sparse LSA that extends the method of LSA by adding ℓ_1- regularization term on A in order to restrict a number of latent variables as following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|Y - AU\|_F^2 + \gamma\|A\|_{1,1} \\ \text{subject to} & UU^T = I \end{array} \tag{3.27}$$

with variables $A \in \mathbf{R}^{q\times r}$ and $U \in \mathbf{R}^{r\times N}$ where $\gamma \geq 0$ is a regularization parameter to control a sparsity of A.

The results from [Chen et al., 2011] show that when an initial number of latent variables is high, sparse LSA works effectively since it reconstructs model by lowering the number of latent variables. On the contrary, when an initial number of latent variables is low, LSA is better than sparse LSA because sparse model may miss important information.

Note that ℓ_1- regularization term on A which shrinks some entries A_{ij} to zero which means that latent variable η_j does not affect to y_i and will be removed. The regularization term which limits a number of latent variables may remove different latent variables for each y_i individually leading that latent variables which affect to all of y do not decrease. In other words, a pattern of influence from

η_j to is dissimilar for all i; consequently, it is difficult to interpret an implication of such remaining latent variables. However, our objectives are to remove some insignificant latent variables which simultaneously hardly influence *overall* y and also to identify the meaning of each remaining latent variable. Consequently, we do not focus on reducing latent variables by this technique since it does not achieve our two purposes.

3.4.4 Sequential selection

The process of variable selection is to select an optimal subset from a set of relevant variables for more efficient model construction. In this section, examples of fundamental sequential search algorithms, *i.e.*, sequential backward selection (SBS) and sequential forward selection (SFS) are discussed.

SBS algorithm instructed by [Marill and Green, 1963] is a top-down search starting with the full set of variables, then removing a variable that least reduces model fit criterion, *e.g.*, maximum likelihood and minimum error of a model, and repeating this step until no improvement of a goodness-of-fit of a model. SBS has a good performance when the optimal subset is large; however, a disadvantage of this method is that SBS cannot reevaluate a benefit from removed variables.

[Whitney, 1971] suggest SFS algorithm that is a bottom-up search starting with no variables and gradually adds the most important variable by model fit criterion until no improvement of model fit. SFS has a good performance when the optimal subset is small; however, this algorithm has a drawback in the sense that SFS only adds variables into the model but does note remove variables from feature set and a redundancy between variables is not analyzed.

Some algorithms in this approach are time-consuming since they add or delete only one variable in each step. In other words, if a number of optimal variables is high for SFS or is low for SBS, the process for variable reduction spend a long time to find the optimal variables.

Overall, according to the method of nuclear norm regularization, sparse PCA and sparse LSA, those procedures can reduce latent variables based on parameter regularization. However, they cannot specify each latent variable leading that such latent variables have no meaning in the sense of capital structure and interpretation of results from those latent variables is not reliable. For sequential selection method, nesting problem is occurred and calculation of goodness of fit of the model for several numerously sets of latent variables has to be found which consumes very long time and memories.

If the mentioned weakness can be remedied, searching for determinants of capital structure will be more effective; moreover, the better explication and interpretation from results are provided. In order to specify remaining latent variables after reduction, in the next chapter, we provide the formulation for latent variable selection which applies the least-squares problem with sparsity term that forces some columns of matrix A to be zero in order to remove latent variables.

CHAPTER IV

METHODOLOGIES

Remind that MIMIC model is represented by

$$\eta = B^T x + \epsilon_\eta \tag{4.1}$$

$$y = A\eta + \epsilon_y \tag{4.2}$$

where observed variables $x \in \mathbf{R}^p$ and $y \in \mathbf{R}^q$ are causes and indicators of latent variable $\eta \in \mathbf{R}^r$, respectively. $\epsilon_\eta \in \mathbf{R}^r$ and $\epsilon_y \in \mathbf{R}^q$ are the disturbance of η and y, respectively.
Besides, the reduced MIMIC model becomes

$$y = Fx + \epsilon \tag{4.3}$$

where $F = AB^T \in \mathbf{R}^{q \times p}$ and $\epsilon = A\epsilon_\eta + \epsilon_y \in \mathbf{R}^q$.

Since MIMIC model may consist of a mix of relevant and irrelevant latent variables, we would like to reduce the number of latent variables until highly effective remaining latent variables are provided. Beginning with selecting m significant latent variables in the section 4.1, we provide an optimization problem which shrinks some columns of A to zero and the latent variables associated with the zero columns of A will be removed. Moreover, the columns of B which have relationships with such removed latent variables will also be zero automatically. Since we decrease the number of latent variables (r) to m remaining effective latent variables, columns of A and B are reduced from r to m. Obviously, when some latent variables are removed, x that related to such removed latent variables are also deleted, saying that variables in x and rows of B are reduced from p to $\tilde{p}$. The next step shown in the section 4.2 is to estimate $A \in \mathbf{R}^{q \times m}$ and $B \in \mathbf{R}^{\tilde{p} \times m}$ which are the coefficient matrices illustrated the relation of y to η and η to x, respectively, when the unimportant latent variables are removed. A and B from this step are estimated by the least-squares problem with linear constraints. Since our formulation provides a set of models that vary upon a number of selected latent variables, a model selection is performed in the section 4.3. Moreover, numerical methods to solve the proposed formulation are provided.

4.1 Latent variable selection

We need to know which latent variables are eliminated in order to properly interpret the solution. [Chen and Huang, 2012] offer predictors selection of linear model (3.9) where the coefficient matrix F is factorized into AB^T motivated from [Yuan and Lin, 2006]. A similar approach can be found in [Kharratzadeh and Coates, 2016]. They propose the least-squares problem with sparsity term that forces some rows of a matrix B to be zero for removing some predictor variables. They proposed the formulation as follows:

$$\begin{array}{ll} \text{minimize} & \|Y - AB^TX\|_F^2 + \gamma \sum_{i=1}^{r} \|B_i\|_2 \\ \text{subject to} & A^TA = I \end{array} \tag{4.4}$$

with variables $A \in \mathbf{R}^{q \times r}, B \in \mathbf{R}^{p \times r}$ (B_i expresses i^{th} row of matrix B) where $Y \in \mathbf{R}^{q \times N}$ and $X \in \mathbf{R}^{p \times N}$ are response and predictor matrices, respectively.

However, the solutions from the problem (4.4) might not be unique but the zero rows that are provided by the penalty term are uniquely determined. In other words, the optimization problem (4.4) chooses the set of meaningful variables uniquely because the solutions of the problem (4.4) have the following properties.

Properties of solutions of the problem (4.4)

1. The solution to the optimization problem (4.4) is unique up to an $r \times r$ orthogonal matrix. More precisely, suppose (A, B) and $(\tilde{A}, \tilde{B})$ are also a solution of the problem (4.4) if and only if there is an orthogonal matrix Q such that $\tilde{A} = AQ$ and $\tilde{B} = BQ$.

2. Zero rows of B_i of the optimization problem (4.4) is uniquely determined.

Proposed formulation

[Chen and Huang, 2012] forces some rows of B to be zero for predictors selection but our objective is latent variable selection. Consequently, we use a similar idea to remove hardly important latent variables that affect y by providing the optimization problem (4.5) that is the least-squares problem with a regularized term that forces some *columns of a matrix* A to be zero. The zero columns illustrate that the latent variables related to such columns are not important. Consequently, we remove those latent variables. The formulation that we propose is the follows:

$$\begin{array}{ll} \text{minimize} & \|Y - AB^TX\|_F^2 + \gamma \sum_{i=1}^{r} \|A_i\|_2 \\ \text{subject to} & \mathrm{P}(B) = 0 \end{array} \tag{4.5}$$

with variables $A \in \mathbf{R}^{q \times r}, B \in \mathbf{R}^{p \times r}$ (A_i expresses the i^{th} column of A) where $Y \in \mathbf{R}^{q \times N}$ and $X \in \mathbf{R}^{p \times N}$ are response and predictor matrices, respectively. Given *projection operator* is a function $\mathrm{P} : \mathbf{R}^{p \times r} \to \mathbf{R}^{p \times r}$, defined as $\mathrm{P}(B) = B_{ij}$ when $(i, j) \in$ set of zero entries in B and $\mathrm{P}(B) = 0$, otherwise. In other words, $\mathrm{P}(B)$ is a linear projection transformation mapping the entries of B that are supposed to be zero based on the structure of B showed in Figure 1.2. Note that the constraints $\mathrm{P}(B) = 0$ comes from the structure of B which is assumed from prior knowledge for grouping same characteristic of x in each η. Applying sum of 2-norm penalty is characterized as an ℓ_1 - norm minimization that is regularized least-squares estimation explained in the section 3.2.4. Group lasso is an extension of the lasso in the sense that promotes some *group* of parameters simultaneously to zero [Hastie et al., 2015, §3]. The problem (4.5) connects to the group lasso problem on columns of A since the penalty term depending on γ forces some columns of A to zero. Regularization parameter, γ, controls a sparsity pattern in the sense that it increases the weight of penalized term. As a result, when γ is large, A contains many zero columns. Besides, the proposed formulation (4.5) is biconvex problem in (A, B) (see the details in the section of numerical method 4.4).

4.2 Least-squares estimation for reduced MIMIC model

After insignificant latent variables are removed, A and B have smaller number columns. The penalty term in the formulation (4.5) introduces more bias to the model. Consequently, to reduce bias of solutions from the formulation (4.5), this part provides the formulation (4.6) to estimate of $A \in \mathbf{R}^{q \times m}$ and $B \in \mathbf{R}^{\tilde{p} \times m}$ which are the coefficient matrices that show the relation of y to η and η to x, respectively.

Proposed formulation

We provide the formulation that estimates A and B as:

$$\begin{array}{ll} \text{minimize} & \|Y - AB^T\tilde{X}\|_F^2 \\ \text{subject to} & \mathrm{P}(B) = 0 \end{array} \tag{4.6}$$

with variables $A \in \mathbf{R}^{q \times m}$ and $B \in \mathbf{R}^{\tilde{p} \times m}$ where $\tilde{X} \in \mathbf{R}^{\tilde{p} \times N}$ is X that is deleted some observed variables related to removed latent variables, *i.e.*, deleting some rows of X until it has $\tilde{p}$ rows. The proposed formulation (4.6) is biconvex problem in (A, B) (see the details in the section 4.4).

4.3 Model selection

According to the section 4.1 and 4.2, a different γ can provide various remaining latent variables leading to a model with various structures. Since candidate models have different structures, to select an appropriate model is depended on choosing a suitable γ. There are several criterions to select γ from a set of candidate γ's which provide different model structure. Information criterion which considers the trade-off between the goodness of fit and the complexity of the model is widely used for model selection. In this research, Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC), and Kullback Information Criterion (KIC), and corrected Kullback Information Criterion (KICc) are applied to select γ with different roles and different asymptotic assumption. The expression of AIC, AICc, BIC, KIC, KICc are

$$\begin{aligned} \text{AIC} &= -2\mathcal{L} + 2d, \\ \text{AICc} &= -2\mathcal{L} + \frac{2dN}{N-d-1}, \\ \text{BIC} &= -2\mathcal{L} + d\,\log N, \\ \text{KIC} &= -2\mathcal{L} + 3d, \\ \text{KICc} &= -2\mathcal{L} + \frac{(d+1)(3N-d-2)}{N-d-2} + \frac{d}{N-d}. \end{aligned} \tag{4.7}$$

where $\mathcal{L}$ is a log-likelihood function value of the model,

d is the number of effective parameters in the model. In this work, d is calculated by $qm + \tilde{p}$ (a number of all entries in $\hat{A}$ and nonzero entries in $\hat{B}$),

N is the number of observations.

According to (4.7), the first component (negative log-likelihood function) expresses the goodness of fit of the model and the second one depending on a number of parameters explains the complexity of the model. Since all information criterions select an appropriate model based on the bias-variance trade-off of the model, for each criterion, the best model is the one having the lowest information criterion score due to minimal combination of bias and variance. AIC [Akaike, 2011, Akaike, 1987, Akaike et al., 1998] tends to select a complex model since the complexity term depends on only the double number of estimated parameters, regardless of sample size. However, when a size of sample is small compared to number of parameters, AIC may perform poorly based on Kullback-Leibler divergence deviation. Consequently, the AICc, an unbiased estimator of AIC, is introduced to improve bias adjustment on a small-sample setting where the complexity term is a slightly heavier penalty depending on both sample size and a number of parameters [Anderson and Burnham, 2002]. BIC [Schwarz, 1978] tends to choose a simpler model than the other criterions since the model complexity is penalized by the sample size (log N). Furthermore, when sample size goes to infinity, it will select the correct model with probability approaching one [Friedman et al., 2001, §7]. KIC [Cavanaugh, 1999] was suggested under the similar penalty as AIC in the sense of depending on an only number of parameters. This criterion serves as an unbiased estimator of Kullback's asymmetric divergence. Like AIC, KIC is biased and underestimated when the sample size is large with respect to a number of parameters. KICc [Seghouane, 2006] is proposed to improve the performance of model selection in KIC for a small-sample setting and also provide a bias reduction. Note that the penalty term of KIC and KICc is heavier than AIC and AICc but is lighter than BIC; therefore, the model selected by KIC or KICc is simpler than AIC and AICc but is more complex than BIC.

For real world data, to select an appropriate model depends on objectives and asymptotic assumptions of users. Although all criterions have the same goal to select the best model based on the bias-variance trade-off of the model, each criterion provides different weights in model complexity component interpreted by the second term. Because of different roles of complexity selection in each criterion, we cannot compare their goodness. For example, if the users need such a complex model, AIC and AICc are more suitable choice; in contrast, BIC is applied when the users need a simple model. If the users prefer not too complex or too simple model, KIC and KICc are considered. Consequently, we provide these five information criterions to the users in order to select the model based on their preference and objectives.

According to (4.7), in order to reduce and eliminate redundant inputs, we provide normalized information criterions scores as follows (see the derivation in Appendix 7.2):

$$\begin{aligned}
\text{Normalized AIC} &= \log\,\det\hat{\Sigma} + \frac{2d}{N}, \\
\text{Normalized AICc} &= \log\,\det\hat{\Sigma} + \frac{2d}{N-d-1}, \\
\text{Normalized BIC} &= \log\,\det\hat{\Sigma} + \frac{d\,\log N}{N}, \\
\text{Normalized KIC} &= \log\,\det\hat{\Sigma} + \frac{3d}{N}, \\
\text{Normalized KICc} &= \log\,\det\hat{\Sigma} + \frac{(d+1)(3N-d-2)}{N(N-d-2)} + \frac{d}{N(N-d)}
\end{aligned} \tag{4.8}$$

Figure 4.1 illustrates the summary of the methodologies including the parts of 4.1, 4.2, and 4.3. Beginning with *latent variable selection* in the section 4.1, the formulation (4.5) is proposed to select m effective latent variables by applying the least-squares problem with penalty term that shrinks some columns of A to zero. Latent variables related to such zero columns are removed. Since the penalty term is controlled by regularization parameter γ, we solve the problem (4.5) by varying $\gamma \in [\gamma_0, \gamma_{\max}]$ in order to vary the sparsity patterns of A where $\gamma_{\max}$ is the γ that penalizes all entries in A to become zero. Besides, the columns of B related with such removed latent variables will also be zero automatically. After that, we remove zero columns that illustrate ineffective latent variables in A and B, saying, columns of A and B are reduced from r to m. Obviously, x related to such removed latent variables expressing by rows of B are also deleted, saying, rows of B are reduced from p to $\tilde{p}$. As a result, from this step, A and B with different structure are provided depending on a various value of γ. Following by the section 4.2, *least-squares estimation for reduced MIMIC model* is proposed to find A and B based on provided structure of A and B from latent variable selection step. Because of biconvexity of latent variables reduction and least-squares estimation for reduced MIMIC model steps problems, alternating minimization is commonly applied to solve these problems. As we know, various structures of A and B are given depending on γ. Lastly, in the section 4.3, we perform a *model selection* via information criterions to select γ from a set of candidate γ's which provide different model structure.

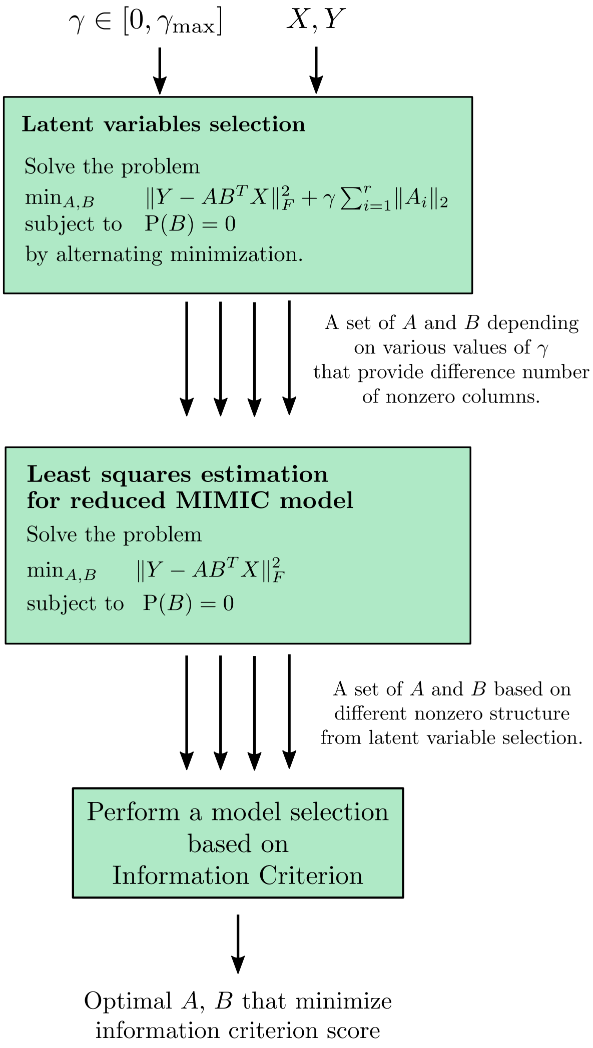

$\gamma \in [0, \gamma_{\max}]$
X, Y
Latent variables selection
Solve the problem
$\min_{A,B} \quad \|Y - AB^T X\|_F^2 + \gamma \sum_{i=1}^{r} \|A_i\|_2$
subject to $\quad \mathrm{P}(B) = 0$
by alternating minimization.
A set of A and B depending on various values of γ that provide difference number of nonzero columns.
Least squares estimation for reduced MIMIC model
Solve the problem
$\min_{A,B} \quad \|Y - AB^T X\|_F^2$
subject to $\quad \mathrm{P}(B) = 0$
A set of A and B based on different nonzero structure from latent variable selection.
Perform a model selection based on Information Criterion
Optimal A, B that minimize information criterion score

Figure 4.1: The diagram of the proposed method.

4.4 Numerical Methods

This section provides numerical methods to solve the proposed formulations, including, latent variable selection (4.5) and Least-squares estimation for reduced MIMIC model (4.6). These two problems are the biconvex optimization, *i.e.*, when B is fixed, the problem is convex in A; conversely, when A is fixed, the problem is convex in B. To solve a biconvex problem, a commonly known alternating minimization is applied.

Biconvex problem and Alternating minimization

Alternating minimization is widely used to solve a biconvex problem. Technically, it fixes one variable to be constant and optimizes over the other variable. A common stopping criterion for alternating minimization is to alternate the problem until the solution and the objective function value converges. The following provides the definition of biconvex problem and the process of alternating minimization.

For nonempty and convex sets $\mathcal{A} \subseteq \mathbf{R}^m$ and $\mathcal{B} \subseteq \mathbf{R}^n$, biconvex set $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{B}$ and given $\mathcal{C}_{a^\star} := \{b \in \mathcal{B} \mid (a^\star, b) \in \mathcal{C}\}$ and $\mathcal{C}_{b^\star} := \{a \in \mathcal{A} \mid (a, b^\star) \in \mathcal{C}\}$. According to the form of *biconvex* problem,

$$\underset{a,b}{\text{minimize}} \quad \{f(a,b) : (a,b) \in \mathcal{C}\}. \tag{4.9}$$

To solve the biconvex problem, *alternating minimization* is applied. Denote (a_i, b_i) be the solution of the problem at iteration i. Start with an initial point $z_0 = (a_0, b_0) \in \mathcal{C}$ and set iteration index $i = 0$.

- When b_i is fixed, solve the convex problem:

$$\underset{a}{\text{minimize}} \quad \{f(a, b_i) : a \in \mathcal{C}_{b_i}\}. \tag{4.10}$$

If there exists an optimal solution $a^\star \in \mathcal{C}_{b_i}$ to this problem, set $a_{i+1} = a^\star$, otherwise, STOP.

- When a_{i+1} is fixed, solve the convex problem:

$$\underset{b}{\text{minimize}} \quad \{f(a_{i+1}, b) : b \in \mathcal{C}_{a_{i+1}}\}. \tag{4.11}$$

If there exists an optimal solution $b^\star \in \mathcal{C}_{a_{i+1}}$ to this problem, set $b_{i+1} = b^\star$, otherwise, STOP.
Next, set $z_{i+1} = (a_{i+1}, b_{i+1})$. If a stopping criterion is satisfied, then stop, otherwise set $i = i + 1$.

The solutions from this method are in general not assured to converge to the global minimum but those solutions are guaranteed that when they converge, they converge to a partial optimum [Hastie et al., 2015, §5]. Let $f : \mathcal{C} \to \mathbf{R}$ be a given function and let $(a^\star, b^\star) \in \mathcal{C}$. Then, $(a^\star, b^\star)$ is called a **partial optimum** of f, if

$$f(a^\star, b^\star) \leq (a^\star, b) \text{ for all } \; b \in \mathcal{C}_{a^\star} \quad \text{and} \quad f(a^\star, b^\star) \leq (a, b^\star) \text{ for all } \; a \in \mathcal{C}_{b^\star}.$$

Although the solution from alternating minimization is generally not guaranteed a convergence, the convergence is guaranteed under some conditions, *i.e.*, a unique solution from each step of solving, stated in lemma (3).

Lemma 3. *[Gorski et al., 2007] Let the optimization problems* (4.10) *and* (4.11) *be solvable. If for each accumulation point* $z^\star = (a^\star, b^\star)$ *of the sequence* $\{z_i\}_{i\in\mathbb{Z}}$*, the optimal solutions of both problems* (4.10) *with* $a = a^\star$ *and* (4.11) *with* $b = b^\star$ *are unique, then*

$$\lim_{i\to\infty} \|z_{i+1} - z_i\| = 0.$$

In other words, the convergence of the sequence $\{z\}_{i\in\mathbb{Z}}$ *is provided.*

In conclusion, the solutions of two subproblems solved by alternating minimization converge under the assumptions that the problem is solvable and whose solution is uniquely obtained in each step of alternating minimization. Our two proposed formulations (4.5) and (4.6) are biconvex that is solved by alternating minimization. Since in each step of alternating minimization is convex, a solution is unique if the problem is strictly convex. Consequently, the solution is guaranteed to converge. Next, we provide a numerical method to solve the problem of latent variable selection and least-square estimation for reduced MIMIC model.

4.4.1 Latent variable selection

The formulation (4.5) is the problem:

$$\begin{array}{ll} \text{minimize} & \|Y - AB^TX\|_F^2 + \gamma \sum_{i=1}^r \|A_i\|_2 \\ \text{subject to} & \mathrm{P}(B) = 0 \end{array}$$

with variables $A \in \mathbf{R}^{q\times r}, B \in \mathbf{R}^{p\times r}$. The two alternating steps can be described as follows.

- When B is fixed, the optimization problem (4.5) is reduced to the following group lasso problem:

$$\text{minimize} \quad \|Y - AB^TX\|_F^2 + \gamma \sum_{i=1}^r \|A_i\|_2 \tag{4.12}$$

with variables $A \in \mathbf{R}^{q\times r}$.

This problem can be solved by many existing efficient convex program solvers such as MATLAB package CVX [Grantl et al., 2014]. Moreover, [Songsiri, 2015] apply a fast alternating directions method of multipliers (ADMM) algorithm to solve a problem which is class of group fused lasso formulation. This method provides advantage in the sense of efficiency of the ADMM algorithm in a high-dimensional setting. Consequently, we apply ADMM following the methodology and codes in [Songsiri, 2015] (the detail is provided in the section 7.3).

The uniqueness of the solution of (4.12) can be obtained if its cost function is strictly convex. Since the penalty term is a norm of A and hence, it is strictly convex (a norm is a strictly convex function). Moreover, a sum of strictly convex and convex functions is strictly convex. Therefore, (4.12) has a strictly convex cost objective and therefore has a unique solution.

- When A is fixed, the optimization problem (4.5) is reduced to

$$\begin{array}{ll} \text{minimize} & \|Y - AB^T X\|_F^2 \\ \text{subject to} & \mathrm{P}(B) = 0 \end{array} \tag{4.13}$$

with variables $B \in \mathbf{R}^{p \times r}$. The problem (4.13) can be considered as linear least-squares problem with linear constraints and it can be reduced to another unconstrained least squares (4.14). To eliminate the constraints, we opt to form the problem in a vector form as an unconstrained least-squares problem:

$$\underset{\beta}{\text{minimize}} \quad \|w - Z\beta\|_2^2. \tag{4.14}$$

The solution of this problem has a closed form of $\beta = (Z^T Z)^{-1} Z^T w$ where

the entries in	$Z \in \mathbf{R}^{qN \times p}$	are functions of A and X,
	$w \in \mathbf{R}^{qN}$	is derived from vectorization of Y,
	$\beta \in \mathbf{R}^{p}$	is derived from vectorization of B (see the details in Appendix 7.1).

These unconstrained least-squares problems are solved by least-squares method. Note that a solution β is unique under the condition that Z is full rank and skinny. In problem (4.14), the matrix Z is typically skinny $(qN > p)$ because N, the sample size, is generally large. It is thus left to check that Z that has structure shown in Appendix 7.1 is full rank or not and the condition depends on A and X.

We find A and B from the problem (4.12) and (4.13), respectively, as varies $\gamma \in \left[0, \gamma_{\max}\right]$ where $\gamma_{\max}$ corresponds to the value of γ that results in zero solution of A in the problem (4.12). Suppose we obtain convergence from solving the problem (4.5), the zero columns of A indicate that the latent variables. Consequently, the columns of B and x's that are related to such latent variables must be further eliminated Accordingly, A and B after we removed unimportant latent variables will have m columns and variables in x and rows of B are reduced from p to $\tilde{p}$. In other words, $A \in \mathbf{R}^{q \times m}$ and $B \in \mathbf{R}^{\tilde{p} \times m}$ are provided, *i.e.*, there are m highly effective latent variables.

4.4.2 Least-squares estimation for reduced MIMIC model

The formulation (4.6) is the problem:

$$\begin{array}{ll} \text{minimize} & \|Y - AB^T\tilde{X}\|_F^2 \\ \text{subject to} & \mathrm{P}(B) = 0 \end{array}$$

with variables $A \in \mathbf{R}^{q\times m}$ and $B \in \mathbf{R}^{\tilde{p}\times m}$. The two alternating steps are as follows.

- When B is fixed, the optimization problem (4.6) is reduced to the following problem:

$$\text{minimize} \quad \|Y - AB^T\tilde{X}\|_F^2 \tag{4.15}$$

with variables $A \in \mathbf{R}^{q\times m}$.

- When A is fixed, the optimization problem (4.6) is reduced to

$$\begin{array}{ll} \text{minimize} & \|Y - AB^T\tilde{X}\|_F^2 \\ \text{subject to} & \mathrm{P}(B) = 0 \end{array} \tag{4.16}$$

with variables $B \in \mathbf{R}^{\tilde{p}\times m}$.

To obtain the closed from solution of the problems (4.15) and (4.16), we can reduce the problem to unconstrained least-squares problems with a different number of variables. Consequently, we opt to convert a matrix form into a vector form of (4.14) (see the details in Appendix 7.1).

For the problem (4.15), the entries in $Z \in \mathbf{R}^{qN\times qm}$ are functions of B and $\tilde{X}$,
$w \in \mathbf{R}^{qN}$ is a vectorization of Y,
$\beta \in \mathbf{R}^{qm}$ is a vectorization of A.

In the same way,
for the problem (4.16), the entries in $Z \in \mathbf{R}^{qN\times\tilde{p}}$ are functions of A and $\tilde{X}$,
$w \in \mathbf{R}^{qN}$ is a vectorization of Y,
$\beta \in \mathbf{R}^{\tilde{p}}$ is a vectorization of B.

Note that a solution β is unique under the condition that Z is full rank and skinny. The structure of Z provided in the Appendix 7.1. Z is skinny when $N > m$ for the problem (4.15) and $qN > \tilde{p}$ for the problem (4.16) and those two conditions are easily satisfied when N is large. Equivalently, (4.15) and (4.16) have unique solutions and these two unconstrained least-squares problems are solved by least-squares method.

From our experimental result for solving the problem (4.5) and (4.6), we found that sometimes, the solution changes following initial condition that we determine since the solution is local optima of nonlinear optimization problem. Therefore, a meaningful initial point should be chosen. Next, we suggest an initialization method used in the alternating minimization.

4.4.3 Initialization of alternating minimization

When solving a nonconvex problem by any iterative methods, different choices of initial value may lead to different local optima. This is then quite a general issue when alternating minimization is applied. Typically, a meaningful initial guess is chosen though there are several ways to specify an initial point denoted by (A_0, B_0). Typically we solve the latent variable selection problem (4.5) for several values of γ, denoted by $\gamma^{(1)}, \dots, \gamma^{(M)}$ to obtain M models with various complexities. We first suggest to solve (4.5) with $\gamma^{(1)} = 0$ by a special choice of (A_0, B_0) and use the obtained optimal solution (A, B) as the initial (A_0, B_0) when solving (4.5) with the next value of γ, saying $\gamma^{(2)}$. This procedure repeats until solving (4.5) with the last γ and will be referred to as a *warm start* method in our experiment.

Speaking of choosing a good (A_0, B_0) when solving (4.5) with $\gamma = 0$, note that this is essentially solving a least-squares problem in (A, B). The cost function can be regarded as $\|Y - FX\|_F$ where $F = AB^T$. We can then propose to firstly solve the least-square for the solution F_0, factorize F_0 as $A_0B_0^T$, and use this (A_0, B_0) as an initial guess when solving (4.5) by alternating minimization. To factorize F_0 as $A_0B_0^T$, we examine the rank of F as follows. Since $\mathbf{rank}(F) = \mathbf{rank}(AB^T) \leq \min\{\mathbf{rank}(A), \mathbf{rank}(B)\} = \min\{q, p, r\}$, we know that $\mathbf{rank}(F) \leq r$ in general. If $\mathbf{rank}(F) = r$, F_0 can be factorized using the singular value decomposition (SVD) that $F_0 = UDV^T$ (the detail is provided in the section 3.4), and let $A_0 = U$ and $B_0 = VD^T$. In the other case, when $\mathbf{rank}(F) < r$, we cannot factorize F using SVD since the factor U would have the number of columns less than r and defining $A_0 = U$ is not what we desire as A_0 must have r columns. Therefore, we opt to choose a dense A_0 randomly and obtain B_0 from $F_0 = A_0B_0^T$.

According to the initial point for other γ's, we apply warm start, *i.e.*, the solution of the formulation with previous γ will be initial condition of the formulation with next γ. Moreover, it guarantees the consistency of zero columns in A and B, in other words, zero columns will not return to be nonzero when γ increases when solving the group lasso problem by ADMM algorithm (see the proof in Appendix 7.3).

According to the problem (4.6), we initialize (A_0, B_0) from the optimal (A, B) from latent variable selection that removes zero columns in A and B and removes some rows of B related to ineffective latent variables. Because this initial point is optimal solution from latent variable selection, it will be a good choice to be the initial point of least-squares estimation for reduced MIMIC model.

CHAPTER V

EXPERIMENTAL RESULTS

This chapter provides all numerical experiments and results of the two proposed formulations: latent variable selection and least-squares estimation for reduced MIMIC model. We provide simulation results, real data application results and their interpretations. In particular, the performance of the model estimation indicated by ROC curves and accuracy of model prediction are obtained.

According to simulation experiments, we assume a ground-truth model parametrized by A_{true} and B_{true}, with some zero columns illustrating ineffective latent variables. The objective of simulation process is to show the performance of our formulations by showing that our formulations predict effective and ineffective latent variables correctly or not. The simulation process is that we generate standard normal random variable $X \in \mathbf{R}^{p\times N}$, $A_{\text{true}} \in \mathbf{R}^{q\times r}$ with $A_2, A_4, A_6 = 0$, $B_{\text{true}} \in \mathbf{R}^{p\times r}$ with zero structure showed in Figure 1.2 where $p = 7, q = 18, r = 7$ and $N = 200$. Next, we generate $Y_{\text{true}} \in \mathbf{R}^{q\times N}$ following the equation $Y_{\text{true}} = A_{\text{true}}B_{\text{true}}^T X + \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$ with $\text{var}(y) = 10.8967$ averaged from 50 samples. Moreover, we vary 100 values of $\gamma \in [0, \gamma_{\max}]$ where $\gamma_{\max}$ corresponds to the value of γ that results in the zero solution of A in the problem (4.5).

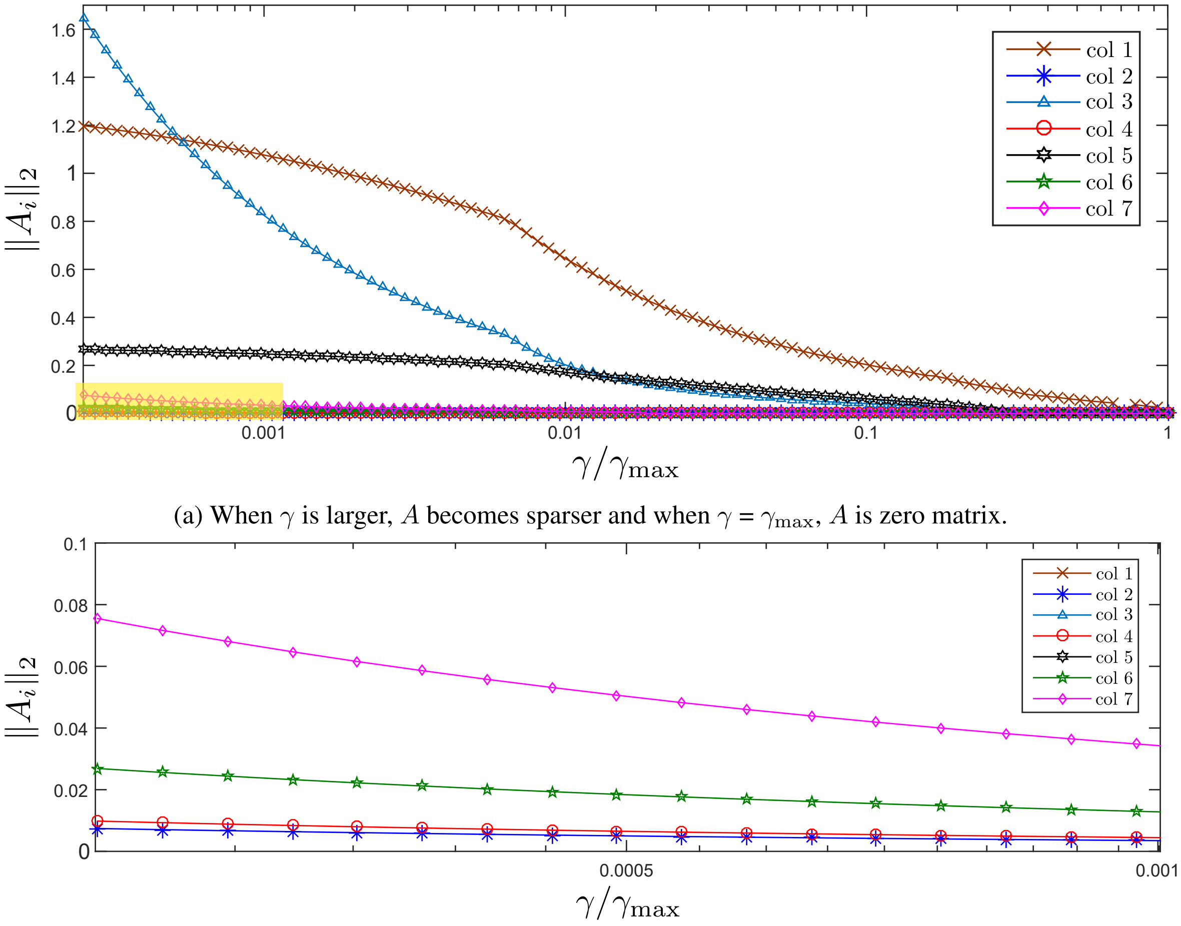
5.1 Illustration of Latent variable selection

This section demonstrates the numerical results of illustrative sample of the proposed formulations (4.5) and (4.6) that is applied to find optimal A and B for each γ. We investigate the estimated value of a sparsity pattern of A and BIC scores as γ varies.

Sparsity patterns of A as γ varies

One way to investigate zero columns of A is to calculate the Euclidean norm of each column as γ varies shown in Figure 5.1a. According to Figure 5.1a, we notice that the norms of each column of A are shrunk to zero when γ increases and all columns of A are zero at $\gamma_{\max}$. However, the norms of some columns of A is close to zero rapidly; therefore, it is difficult to investigate the result. Figure 5.1b shows the larger version of the highlight area in Figure 5.1a when $\gamma/\gamma_{\max} \in [0, 0.001]$ in order to notice the convergence of norm for some A_i. The result shows that A_2, A_4, and A_6 which are zero columns in A_{true} are the first three columns approaching zero.

The same conclusion can be illustrated as follows. We use a binary matrix including colored squares and colorless squares illustrating nonzero entries and zero entries in A, respectively. According to Figure 5.2, beginning with $\gamma = 0$, A is dense and while $\gamma = \gamma_{\max}$, A is a zero matrix. We can notice that when γ is larger, sparsity of A is increased. As previously mentioned, the 2^{nd}, 4^{th}, and 6^{th} columns of A_{true} that we generate are zero. The result shows that using an appropriate value of γ gives A that has the zero columns as same as A_{true}.



(a) When γ is larger, A becomes sparser and when γ = $\gamma_{\max}$, A is zero matrix.

(b) The figure is zoomed to show the highlight area of the figure 5.1a.

Figure 5.1: Norm of the i^{th} column of A as γ varies.

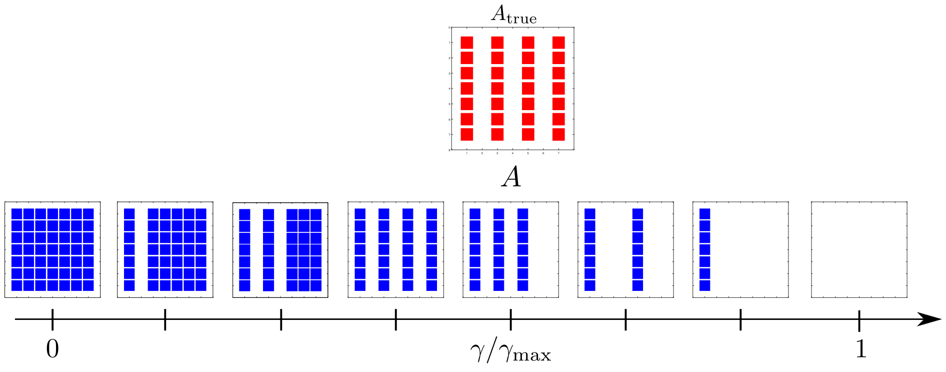


Figure 5.2: Trend of zero pattern in A as γ varies: when γ increases, A becomes sparser.

BIC scores as γ varies

When γ is large, A contains many zero columns leading to low value of penalty term in the problem (4.5) and has low goodness of fit. More zero columns structure induces more bias to the model although variance is lower. Figure 5.3 (top) illustrates the relationship between bias and variance indicated by column sum norm of A (red star line) and norm of prediction error (blue square line), respectively. We observe that when the value of γ is varied from 0 to the middle range of $\gamma_{\max}$ (log-scale), the norm of error is quite low as the number of zero column in A is similar to A_{true}. In that case, the value of prediction error is low implying low bias but the variance of parameter is high as shown by the high value of the column sum norm of A. However, when the value of γ is closer to $\gamma_{\max}$, A becomes sparser and the norm of error becomes larger. The penalty term, *i.e.*, the sum of norm of A_i, monotonically decreases when the value of γ increases because of a larger number of zero columns in A. This case demonstrates low variance and high bias of the model. Information criterion which considers a trade-off between goodness of fit and complexity of the model is suitable for model selection that is to choose γ.

We follow the methodology shown in Figure 4.1, and note that we choose to apply BIC for model selection. γ that corresponds to the minimum BIC score is selected since it minimizes the combination of bias and variance. Figure 5.3 (bottom) shows BIC scores as γ varies. At γ which provides efficient latent variables as same as in the true model, BIC score is lowest showing the optimal trade-off between bias and variance of the model. When γ is larger, the estimated model is more discrepant from the true model. Therefore, A and B from such γ is selected to be an appropriate solution of the proposed formulation (4.6).

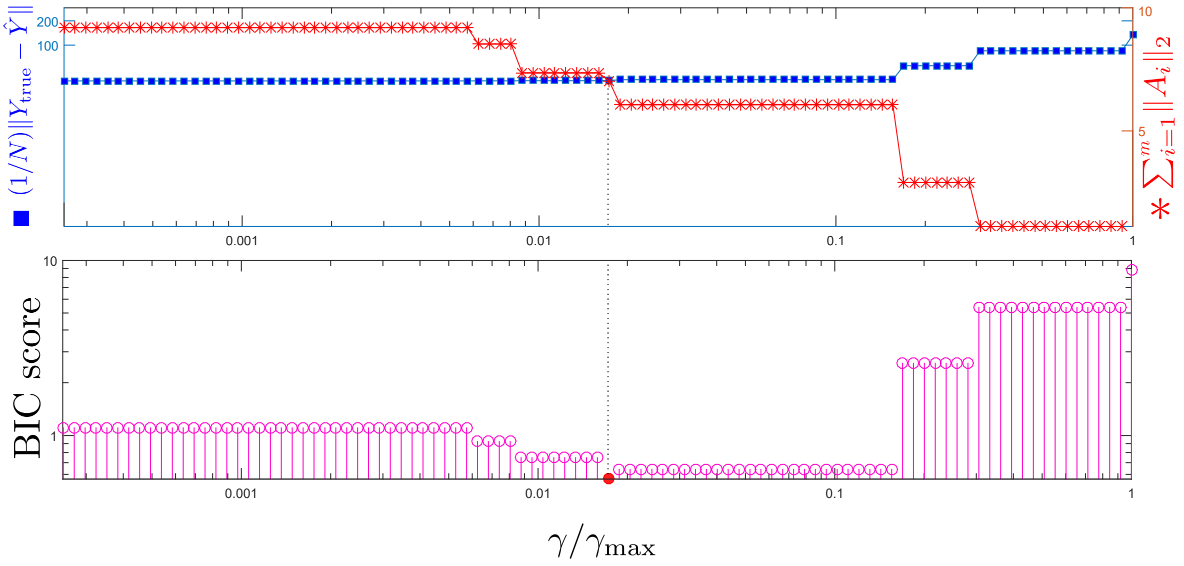


Figure 5.3: As γ varies, norm of error and value of penalty term that are inverse illustrates the trade-off between goodness of fit and model complexity. At an appropriate γ, BIC is minimum (the red circle) and such γ is selected to be the suitable penalty parameter of the model.

5.2 Performance of latent variable selection

This section illustrates a performance of latent variable selection by showing that our formulations can predict zero columns of A correctly or not. We provide Receiver Operating Characteristic (ROC) curve and accuracy of predicted latent variable selection to evaluate the performance of the proposed formulations.

5.2.1 Receiver Operating Characteristic (ROC) curve

Given *positive* is nonzero column on A and *negative* is zero column in A. The four outcomes that are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are defined as:

True positive (TP): the number of correctly identified columns as nonzero
(nonzero columns in $\hat{A}$ that are in A_{true}),

False positive (FP): the number of incorrectly identified columns as nonzero
(nonzero columns in $\hat{A}$ that are not in A_{true}),

True negative (TN): the number of correctly identified columns as zero
(zero columns in $\hat{A}$ that are in A_{true}),

False negative (FN): the number of incorrectly identified columns as zero
(zero columns in $\hat{A}$ that are not in A_{true}).

To illustrate the performance of identifying the zeros on A, common measures are

True positive rate (TPR) = TP/(TP+FN), and False positive rate (FPR) = FP/(FP+TN).

TPR shows correct positive columns occur among all positive samples and FPR shows incorrect positive columns occur among all negative samples.

Receiver Operating Characteristic (ROC) curve is a plot of TPR versus FPR by varying γ from 0 to $\gamma_{\max}$. It illustrates the effectiveness of the formulation by showing an ability to predict positions of zero and nonzero columns. At $\gamma = 0$, $\hat{A}$ is dense so nonzero columns in A_{true} are all correctly identified and zero columns in A_{true} are all incorrectly identified leading to having high TPR and FPR (top-right corner). When γ increases, a number of zero columns in $\hat{A}$ tends to also increase, we anticipate that FPR will be decreased. Accordingly, a good performance of the formulation provides ROC curve lying close to the top-left corner and at least lies above the diagonal line, saying that there are some values of regularized parameter giving high TPR and low FPR simultaneously, *i.e.*, the accuracy of prediction is high. We evaluate the performance proposed formulation via ROC curves with various settings including ROC curves with different initialized method, different sample size, and different number of zero columns in A_{true} as γ varies. Each point on ROC curve is generated by plotting TPR against FPR averaged over 50 runs at each γ and ROC curve is obtained by varying γ from 0 to $\gamma_{\max}$.

ROC curves with different initialization methods

In general, in the case of applying the same initial point for all γ, zero columns in A and B can return to be nonzero when γ increases. However, using warm start guaranteed the consistency of zero columns in A and B, in other words, zero columns will not return to be nonzero when γ increases explained in section (4.4.3). Figure 5.4 illustrates ROC curves of results from our formulations when we use the same initial point for all γ versus initialize with warm start. The results show that the performance of both methods yield similarly good results. According to the method of the same initial point for all γ, the trend of ROC curve is not smooth since nonzero columns of A do not monotonically decrease when γ increases, meaning that zero column of A can be returned to nonzero column leading to a fluctuation of FPR. However, initialization of warm start guarantees the consistency of estimated zero columns; consequently, we apply warm start method to our experiment.

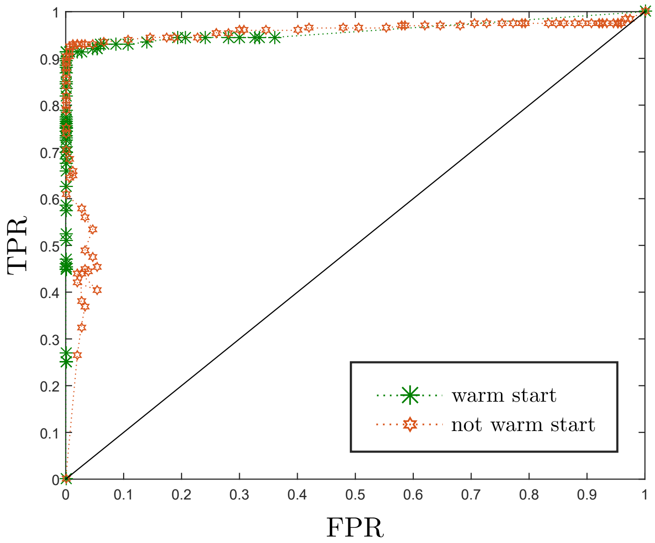


Figure 5.4: A comparison of ROC curves between two methods: initialization with the same initial point for all γ and warm start.

ROC curves with different sample sizes

We investigate the ability of latent variable selection when sample of measurement (N) is varied via ROC curves in the Figure 5.5. In this experiment, we vary three sample sizes as 50, 200, and 2000, respectively. The result shows that three ROC curves lie close to the top-left corner illustrating a good performance of latent variable selection. Moreover, the performance of latent variable selection increases when the sample size increases because N can reduce variance of $\hat{A}$ leading to higher efficiency for correctly prediction.

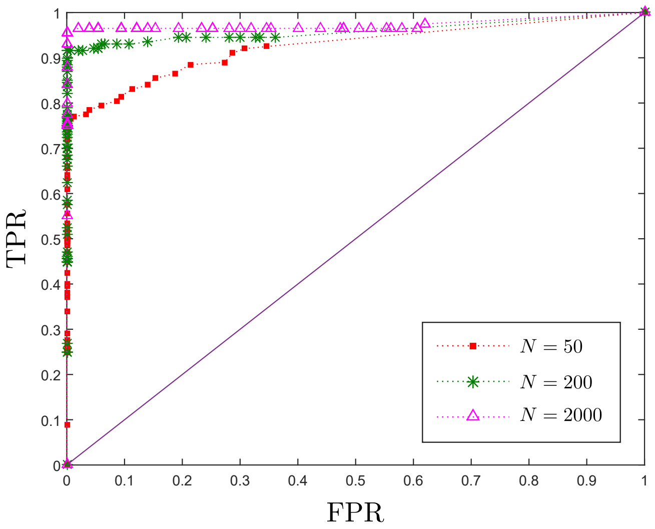


Figure 5.5: ROC curve illustrates True Positive Rate (TPR) versus False Positive Rate (FPR) as the sample size (N) of data varies.

ROC curves with different zero columns in A_{true}

Figure 5.6 shows ROC curves when the number of zero columns in A_{true} is varied. We vary three different zero columns number in A_{true} as 1, 3, and 5 columns which are calculated to a sparsity of 14.29%, 42.86%, and 71.43%, respectively, with $N = 200$. Similar to the previous results, three ROC curves lie close to the top-left corner illustrating a good performance of latent variable selection. Moreover, we can notice that the performance of the proposed formulations with sparser A_{true} outperforms dense A_{true}. In other words, our formulations can select significant latent variables more correctly when the number of true insignificant latent variables is large.

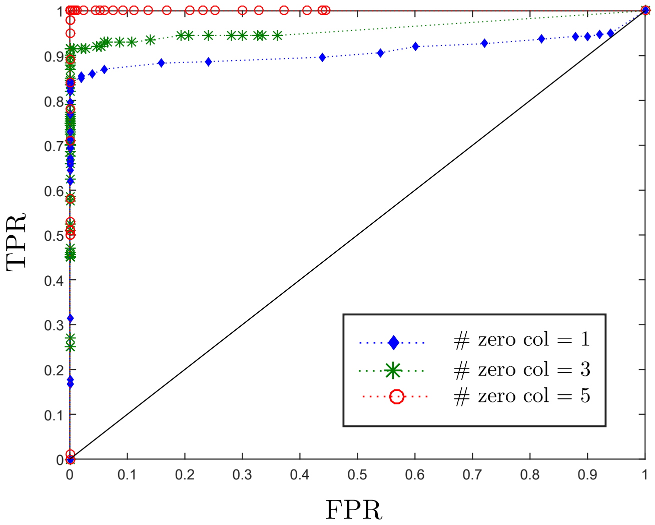


Figure 5.6: ROC curve illustrates True Positive Rate (TPR) versus False Positive Rate (FPR) as the number of zero columns in A_{true} varies.

5.2.2 Error of predicted latent variables

In addition, we provide the error of predicted latent variables, *i.e.*, the portion of number of incorrectly predicted latent variable compared between our method and [Chang et al., 2009] 's method. Since [Chang et al., 2009] apply F in reduced MIMIC model (4.3) representing the relationship between x and y to identify relative impacts of η to y that is calculated by the most effective relative impact of x (in each η) to y. To be easier to understand, firstly, we provide the mathematical representation of relative impacts as following. Denote

$$F = \begin{bmatrix} F_1^{(1)} & \cdots & F_{n_1}^{(1)} & F_1^{(2)} & \cdots & F_{n_2}^{(2)} & \cdots & F_1^{(r)} & \cdots & F_{n_r}^{(r)} \end{bmatrix}$$

where n_i is a number of x's in each η_i, superscript shows the index of η and subscript shows the index of x related to η. The relative impact from [Chang et al., 2009] 's solution is measured as:

$$\text{Relative impacts of } \eta_i \text{ to } y \ (RI_i) = \max \ \{\|F_1^{(i)}\|_2, \|F_2^{(i)}\|_2, \ldots, \|F_{n_i}^{(i)}\|_2\}. \tag{5.1}$$

In our method, A illustrates a direct relationship between η and y but [Chang et al., 2009] do not provide a direct measure in the model. Consequently, to compare the result between our method and [Chang et al., 2009] 's method, we need to use the same measure. In [Chang et al., 2009] 's method, we let A_i be $F_j^{(i)}$ that maximizes RI_i and it is served as a proxy to explain the relationship from η_i to y. Then we compare the performance via A from our method and [Chang et al., 2009] 's method. Let $j = 1, \ldots, n_i$, in [Chang et al., 2009] 's method, A_i is defined as:

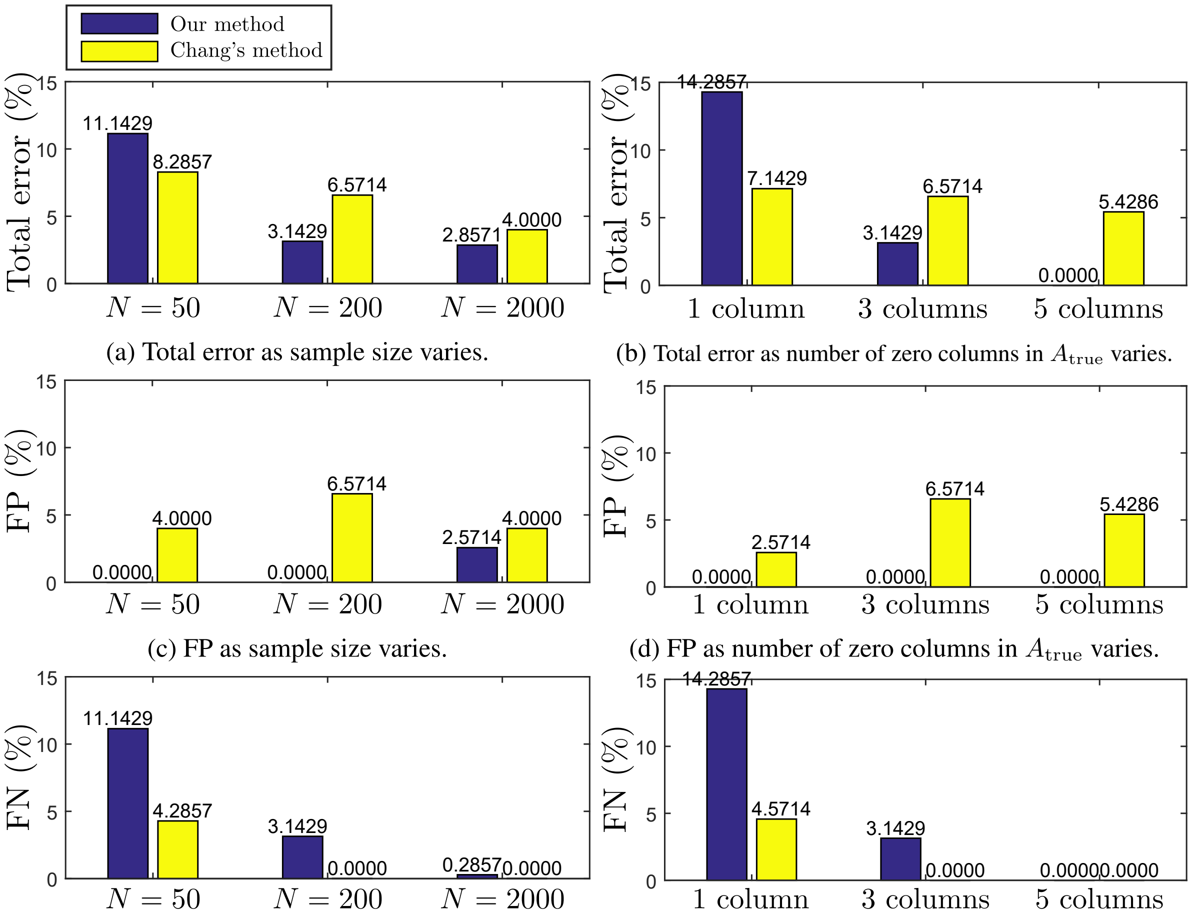
$$A_i = \underset{F_j^{(i)}}{\operatorname{argmax}} \ RI_i. \tag{5.2}$$

Note that for [Chang et al., 2009] 's method, the Wald statistical test is applied to test the significance of column with a significance level of 0.05. We use A_i as in (5.2) and estimated A from our method to calculate the portion of incorrectly number of predicted latent variables (total error) that is calculated by number of both False positive (FP) (number of incorrectly predicted nonzero columns of A) and False negative (FN) (number of incorrectly predicted zero columns of A), then divided by total number of all columns of A. Figure 5.7 illustrates the total error, FP, and FN, compared between our method and [Chang et al., 2009] 's method when N varies and number of zero columns in A_{true} varies, averaged over 50 runs.

Figure 5.7a, 5.7c, and 5.7e illustrate the total error, FP, and FN, respectively, compared between our method and [Chang et al., 2009] 's method when sample size N varies as 50, 200, and 2000, respectively. When N is larger, total error from both our method and [Chang et al., 2009] 's method tend to decrease, in other words, insignificant latent variables can be removed more correctly since using large N can reduce variance of $\hat{A}$ leading to higher efficiency for correctly prediction. When N is middle to large, total error from our method is less than [Chang et al., 2009] 's. However, [Chang et al., 2009] 's method slightly outperforms ours when N is small because N affects to model selection criterion, BIC. When N is small, BIC may provides A which has different zero structure from A_{true} leading to have large FN shown by Figure 5.7e.

Figure 5.7b, 5.7d, and 5.7f show the total error, FP, and FN, respectively, compared between our method and [Chang et al., 2009] 's method when number of zero columns in A_{true} varies as 1, 3, and 5 columns, respectively, from 7 columns with $N = 200$. The results show that whatever number of zero columns in A_{true}, our method still predicts insignificant latent variables correctly shown by zero FP for all three cases in Figure 5.7d. We can investigate that the higher number of zero columns in A_{true} lowers the portion of number of incorrectly predicted latent variables from our method. For interpretation, since we choose BIC score to select the model, A with a sparser pattern is selected with high possibility. Consequently, if A_{true} is sparser, our method will predict more correctly zero columns as we expected. If A_{true} has a few sparse pattern, to choose AIC or KIC score may provide more accurate prediction of zero columns. This is a reason why the total error from our method is higher than [Chang et al., 2009] 's in the case that A is dense. Referring to the result from [Chang et al., 2009] 's experiments, the total error of latent variable prediction dose not significantly change because the Wald statistical test is applied to test the significance of parameter with a significance level of 0.05. Consequently, the performance of [Chang et al., 2009] 's is still the same whatever number of zero columns in A_{true}.

In summary, our method provides the very low percentages of FP (zero FP in almost cases) illustrating a good performance of the proposed formulations for ineffective latent variables prediction. While [Chang et al., 2009] 's method provides low FN, both methods perform relatively well because of low error in almost cases. The performance of our method depends on sample size and true zero columns in the sense that more sample size or more true zero columns introduces more correct prediction. Although [Chang et al., 2009] 's method outperform ours in the case of dense true model, it depends on the model selection criterion used to select γ in the latent variable selection problem in the sense that if we choose suitable criterion, the prediction is more correct. Moreover, since our formulation is principally proposed for removing ineffective latent variables, it performs well if a true model is sparse.



(a) Total error as sample size varies.

(b) Total error as number of zero columns in A_{true} varies.

(c) FP as sample size varies.

(d) FP as number of zero columns in A_{true} varies.

(e) FN as sample size varies.

(f) FN as number of zero columns in A_{true} varies.

Figure 5.7: The portion of number of incorrectly predicted latent variables, including, total error, FP, and FN, compared between our method and [Chang et al., 2009] 's method from 50 runs.

5.3 Relative impact of selected latent variables

In this section, we show normalized relative impacts of η to y in order to investigate the effect of each latent variable to y. We compare normalized relative impacts between the true solution, our proposed formulation solution and [Chang et al., 2009] 's reduced MIMIC model solution from various settings. The relative impacts from true solution and our solution are calculated by column norm of A as follows:

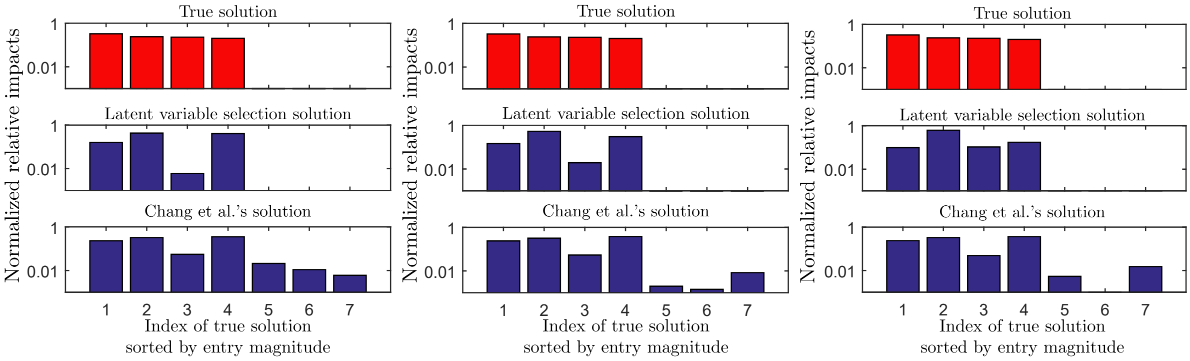
$$\text{Relative impacts of } \eta_i \text{ to } y\ (RI_i) = \|A_i\|_2. \tag{5.3}$$

Note that, according to [Chang et al., 2009] 's solution, the relative impact is calculated by (5.1) since they apply F which is the indirect relationship between x and y to refer the relationship between η and y. While true solution and our solution have A that is the direct relationship between η and y so the relative impacts of both solution can be calculated by column norm of A. To be easier to compare the effect of latent variable to y, we provide the normalized relative impact of η_i to y which is calculated from RI_i divided by sum of all RI as follows:

$$\text{Normalized relative impact of } \eta_i \text{ to } y = \frac{RI_i}{\sum\limits_i RI_i}. \tag{5.4}$$

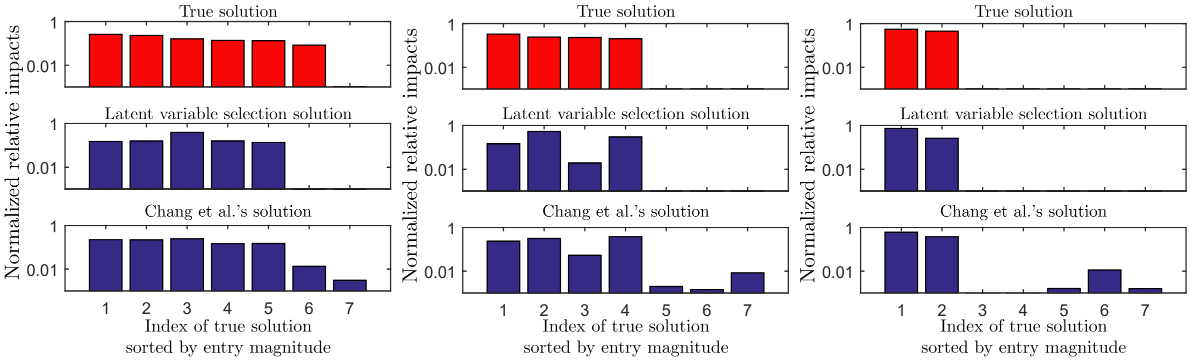
Next, we rank the true relative impacts of each η in descending order, then rank the relative impacts averaged from 50 trials from our solution and [Chang et al., 2009] 's solution following the index of the true one as shown in the Figure 5.8 and 5.9. Figure 5.8a, 5.8b, and 5.8c illustrate the normalized relative impacts with sample size of 50, 200, and 2000, respectively. Figure 5.9a, 5.8b, and 5.9c illustrate the normalized relative impacts with one, three, and five zero columns of A_{true}, respectively, by using a moderate sample size of 200.

In overview, the trend of nonzero relative impacts from η, our solution is almost similar to [Chang et al., 2009] 's solution but both solutions are different from the true solution. The zero relative impacts of η from our solution and the true solution in almost all cases come from the same latent variables. Likewise, for [Chang et al., 2009] 's solution, the least influential relative impacts come from such ineffective η. Although the least influential relative impacts from [Chang et al., 2009] 's solution come from η which has true zero relative impacts, its existence of those small relative impact leads to an ambiguous interpretation that we should consider such η or not if we apply [Chang et al., 2009] 's method. Nevertheless, our results overcome their weakness because we apply group lasso that can remove ineffective latent variable. In other words, our method has a good performance for removing ineffective latent variables but cannot provide a good performance to estimate the relative impact for effective latent variables. According to Figure 5.8, when N is larger, some nonzero relative impacts from our solution is closer to the true solution than the case of small N (obviously notice from the 3rd latent variable in Figure 5.8). Moreover, some zero relative impacts from [Chang et al., 2009] 's solution is closer to the true solution (notice from the 5th and 6th indices in Figure 5.8). Moreover, our formulation provides less FP than [Chang et al., 2009] 's in the sense of ineffective latent variable prediction since it can predict ineffective latent variable correctly for all runs, saying that FP is zero. Figure 5.9a shows the relative impacts when A_{true} is dense, our formulation provides FN in 6^{th} index, saying that our formulation selects denser model than the true model. Since we applied BIC which generally selects sparse model for model selection, [Chang et al., 2009] 's method outperforms ours in this case. However, according to Figure 5.9b and 5.9c , our method can predict ineffective latent variables correctly for all runs when number of zero columns of A_{true} is larger, in other words, there is no FP. Besides, when number of zero columns of A_{true} is higher, the nonzero relative impacts from our solution and [Chang et al., 2009] 's solution are closer to A_{true} shown by Figure 5.9c.



(a) Sample size of 50. (b) Sample size of 200. (c) Sample size of 2000.

Figure 5.8: With different sample size (N), normalized relative impact of η's to y (log-scale) sorted by entry magnitude of true solution: the comparison between the true solution with latent variable selection solution and [Chang et al., 2009] 's solution.



(a) *One* zero column of A_{true}. (b) *Three* zero columns of A_{true}. (c) *Five* zero columns of A_{true}.

Figure 5.9: With different zero structure of A_{true}, normalized relative impact of η's to y (log-scale) sorted by entry magnitude of true solution: the comparison between the true solution with latent variable selection solution and [Chang et al., 2009] 's solution.

Simulation experiments, including, illustrative of latent variable selection and performance evauation had already been obtained in this section. In the next section, we apply our proposed formulations to real application data to identify effective determinants of capital structure.

5.4 Identification of determinants of capital structure

In this section, we apply our proposed formulations to real application data to identify effective determinants of capital structure. Besides, the relations among determinants and measures of capital structure and relative impacts of the determinants of capital structure are provided. Moreover, the direction of relationship between determinants and measures of capital structure from our framework are compared with the trade-off theory and the pecking order theory. This section is divided into two parts including data description and results as follow.

5.4.1 Data description

We collect 11,382 observations from North America Fundamental Annual Updates in *Compustat - Capital IQ* based on 28-year pooled sample for the period 1988-2015. According to the unconstrained least-squares problems of (4.13) and (4.16), one assumption to obtain a unique solution is that x_i for $i = 1, \ldots, p$ must be independent based on the structure of Z; see the discussion in the section 4.4.1. However, there is a duplicate variable RD/S in both η_1 (growth) and η_2 (uniqueness) so we remove RD/S in η_1 (growth) in order to get rid of the problem of dependent variables. Moreover, MBA and IND are not provided in the database. Consequently, we do not consider these three variables and we provide a new path diagram of MIMIC model for real sample data set illustrating the structure of A and B in the Figure 5.10. The sample data sets consist of seven industries based on four-digit Standard Industrial Classification (SIC) code, including i) Agriculture, Forestry and Fishing, ii) Mining, iii) Construction, iv) Manufacturing, v) Transportation, Communications, Electric, Gas and Sanitary service, vi) Wholesale Trade, and vii) Retail Trade.

Since we apply the method of least squares in our formulations and assume that ϵ has a normal distribution, before estimating A and B, we standardize all variables by transforming the data into the normal scores, *i.e.*, subtracting the data by the sample mean and dividing the data by the sample standard deviation.

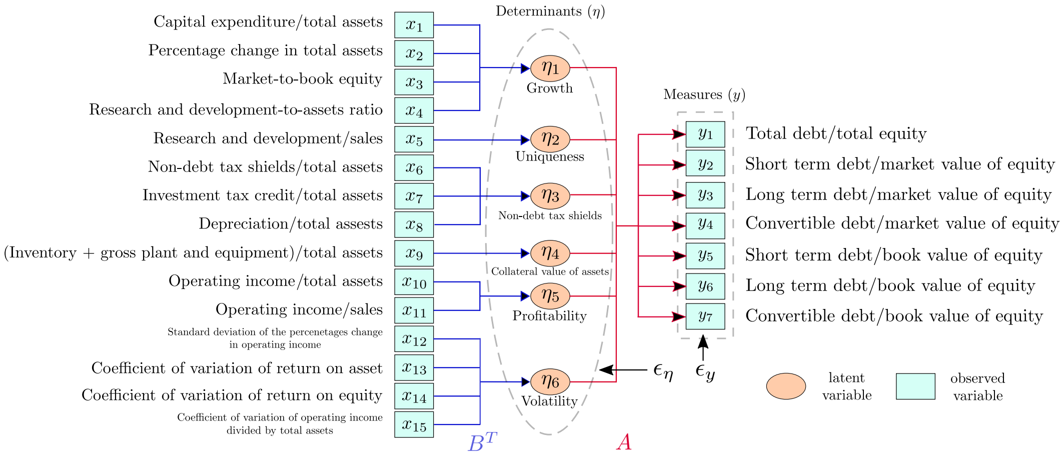

Determinants (η)
Capital expenditure/total assets
Percentage change in total assets
Market-to-book equity
Research and development-to-assets ratio
Research and development/sales
Non-debt tax shields/total assets
Investment tax credit/total assets
Depreciation/total assests
(Inventory + gross plant and equipment)/total assets
Operating income/total assets
Operating income/sales
Standard deviation of the percenetages change in operating income
Coefficient of variation of return on asset
Coefficient of variation of return on equity
Coefficient of variation of operating income divided by total assets
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
x_{10}
x_{11}
x_{12}
x_{13}
x_{14}
x_{15}
η_1
Growth
η_2
Uniqueness
η_3
Non-debt tax shields
η_4
Collateral value of assets
η_5
Profitability
η_6
Volatility
Measures (y)
y_1
y_2
y_3
y_4
y_5
y_6
y_7
Total debt/total equity
Short term debt/market value of equity
Long term debt/market value of equity
Convertible debt/market value of equity
Short term debt/book value of equity
Long term debt/book value of equity
Convertible debt/book value of equity
ϵ_η
ϵ_y
latent variable
observed variable
B^T
A

Figure 5.10: Path diagram of MIMIC model for real data sets illustrating the structure of A and B

5.4.2 Results

As we mentioned that we provide five model selection criterions to select latent variables following a zero structure of $\hat{A}$, different criterions may give different results based on its assumptions and objective. Figure 5.11 illustrates AIC, AICc, BIC, KIC, and KICc scores as γ varies. For each criterion, γ's that minimize the information criterion score represented by the blue-circled point will be selected to use in the formulation of latent variable selection.

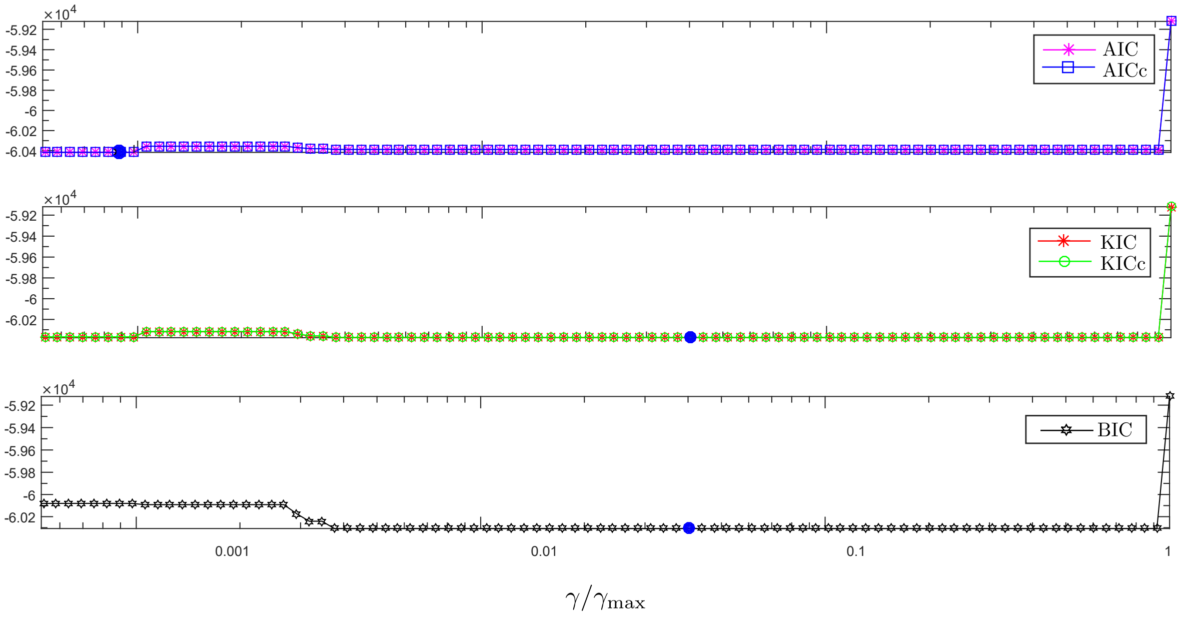


Figure 5.11: AIC, AICc, BIC, KIC, and KICc scores as γ varies: γ's that minimize the information criterion score represented by the blue-circled point will be selected to use in the formulation of latent variable selection.

Figure 5.12 represents the best structure of $\hat{A}$ for each criterion selected by γ that minimize each information criterion score. We can investigate that AIC and AICc choose a dense $\hat{A}$ consisted of six latent variables which makes the model very complex. While BIC, KIC and KICc select a simpler model than AIC and AICc, *i.e.*, only one latent variable providing the simplest model. Table 5.1 represents the determinants of capital structure which are selected by each criterion. *growth* is only one determinant which is selected by all information criterions, in other words, all criterions totally agree that *growth* is the effective determinant of capital structure. On the contrary, *volatility* is not selected by any criterion implying it is not an effective determinant of capital structure in the seven industries we are interested in the North America.

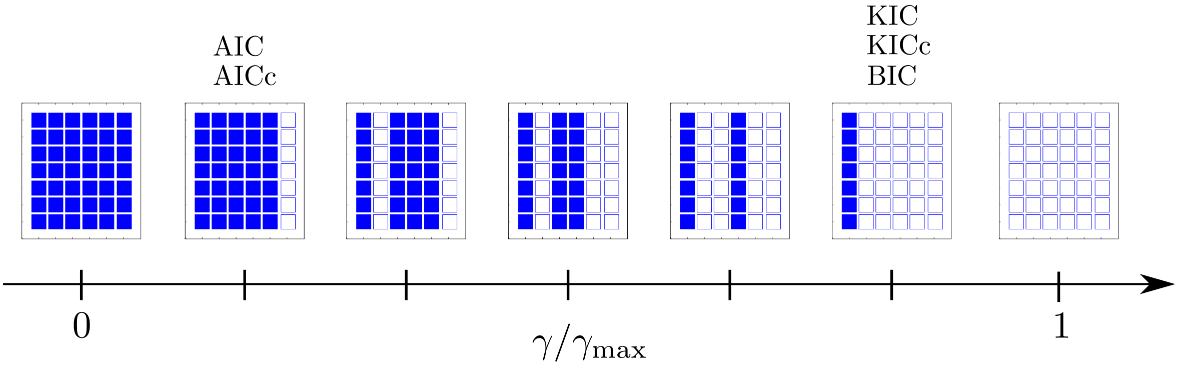


Figure 5.12: Structure of $\hat{A}$ as γ varies: AIC and AICc choose very complex $\hat{A}$, while KIC, KICc and BIC select the simpler one.

Table 5.1: Latent variables which all criterions choose: AIC and AICc choose five determinants that are growth, uniqueness, non-debt tax shields, collateral value of assets, and profitability. While BIC, KIC and KICc chooses only growth.

Determinants of capital structure	AIC, AICc	BIC	KIC, KICc
Growth	●	●	●
Uniqueness	●		
Non-debt tax shields	●		
Collateral value of assets	●		
Profitability	●		
Volatility			

Table 5.2 illustrates the estimated coefficient A selected by various criterions that provide different ineffective latent variables. Figure 5.13 shows the normalized relative impacts of the determinants of capital structure selected by varying information criterions, including, AIC, AICc, BIC, KIC, and KICc. According to AIC and AICc, the relative impacts are described in descending order as growth, non-debt tax shields, collateral value of assets, uniqueness, profitability, and volatility. For BIC, KIC and KICc only one determinant, growth, is selected. Apparently, all information criterions select growth to be the most proficient determinant of capital structure and volatility is not chosen by any criterion so it is unimportant determinant in this case. Moreover, the normalized relative impacts calculated by the approach from [Chang et al., 2009] 's is provided. Their result corresponds to the five information criterions in the sense that growth and volatility are the most significant and the ineffective determinants of capital structure, respectively.

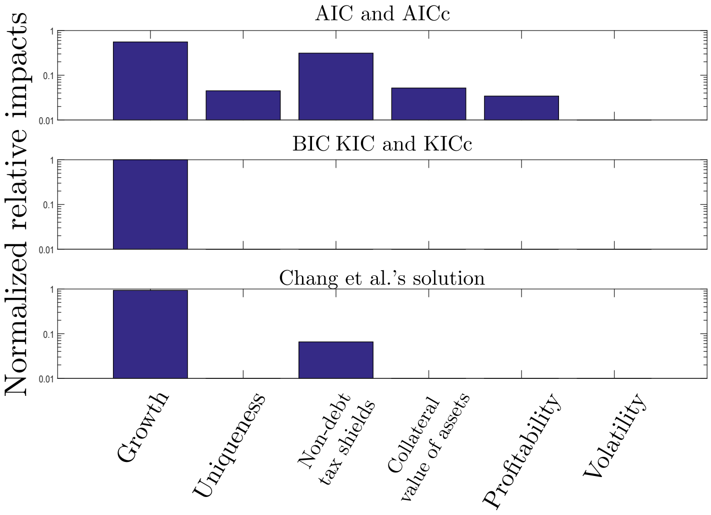


Figure 5.13: Normalized relative impact of the determinants of capital structure selected by varying information criterions and [Chang et al., 2009] 's solutions: growth is the most effective determinant while volatility does not influence to the capital structure based on the result from information criterions and is the least effective based on [Chang et al., 2009] 's solutions.

One important interpretation is to investigate the predicted sign of the estimated coefficient A in order to get a direction of relationship between measures and determinants of capital structure. The direction of relationship between η_i to y is the sign of sum of all entries in A_i. Compared with Table 2.1, Table 5.3 provides the direction of relationship between debt ratios and i) growth (-), ii) uniqueness (-), and iii) non-debt tax shields (-) that are consistent with the trade-off theory while profitability (+) is consistent with the pecking order theory. The negative relationship between **growth** and debt to equity ratios is supported by [Jensen and Meckling, 1976, Myers, 1977]. Since growing firms with more investment opportunities have lower debt ratio, they will have less leverage and use more equity financing to avoid underinvestment and asset substitution problems [Jensen and Meckling, 1976, Myers, 1977]. Moreover, [Myers, 1984] argues that since growth opportunity is intangible assets, the growing firm has to use less debt to prevent bankruptcy situations. **Volatility** which is not selected by any criterion does not affect capital structure according to these kind of industries we apply. This result is consistent with [Titman and Wessels, 1988] 's result whose volatility and non-debt tax shields are statistically insignificant. According to a positive relation of **Collateral value of asset** to debt ratio, a firm with high collateral value of asset issues more debt since it has a benefit from low cost of debt [Myers and Majluf, 1984]. Moreover, because of high collateral value of assets from fixed assets, a firm is not necessary to reveal all information for obtain long-term debt capital from financial institutions. **Uniqueness** is negatively related to debt to equity ratio is supported by [Titman and Wessels, 1988, Chang et al., 2009]. This can imply that a high uniqueness firm will have high research and development expenditure used to improve the existing products and provides more liquidity to a firm in the sense that customers will difficult to find other products. **Non-debt tax shields** is consistent with the trade-off theory in the sense that it is negatively related to debt to equity ratio.

It is the reduction in income taxes due to non-debt quantity that are allowed to subtract from taxable income, *e.g.*, depreciation expenses, investment tax credits. If non-debt tax shields are large, a firm will have the less debt due to the tax benefits of debt financing [DeAngelo and Masulis, 1980]. Besides, the negative relation between **Profitability** and leverage is supported by pecking order theory. A firm with high profitability uses internal fund first because of more retained earnings, then issues securities, and is followed by issuing new equity. In contrast, a low profit firm uses more debt because of insufficient internal fund.

As we mentioned in the section 2.1 that the direction of relation between debt ratios and determinants of capital structure provides the benefit about financial policy formulation, when we know that the empirical results match to either the trade-off theory or the pecking order theory, the capital structure behavior of a firm is also known. If the direction are consistent with the trade-off theory, it means that the financing of a firm is considered by the trade-off between the benefit and cost of debt. If the direction is consistent with the pecking order theory, the capital structure of a firm is formulated based on asymmetry information. When a firm's financing behavior is recognized, policy maker will obtain a policy guide to create a reasonable financing policy. For example, if the government would like to boost the economic of the capital market, a policy maker may increase taxes to have more investments based on the trade-off theory. Future research may extend this work to serve as a policy guide for policy formulation by investigating the behavior of firm's capital structure in order to reduce bankruptcy and information asymmetry problems.

Table 5.2: Estimated coefficient A showing the relationship between the measures and the determinants of capital structure selected by various information criterions.

Criterion	η y	Growth	Uniqueness	Non-debt tax shields	Collateral value of asset	Profitablity	Volatility
AIC, AICc	TD/TE	-0.0254	0.0011	0.0022	-0.0002	0.0003	
	ST/MVE	0.0001	0.0003	-0.0283	0.0049	0.0001	
	LT/MVE	0.0001	-0.0000	0.0128	0.0022	-0.0001	
	C/MVE	0.0001	-0.0002	-0.0069	0.0011	0.0000	
	ST/BVE	-0.0605	-0.0051	0.0100	-0.0011	-0.0035	
	LT/BVE	-0.0323	0.0008	0.0056	-0.0005	-0.0000	
	C/BVE	-0.0140	-0.0028	-0.0245	0.0042	-0.0029	
KIC, KICc	TD/TE	-0.0220					
	ST/MVE	0.0001					
	LT/MVE	0.0000					
	C/MVE	0.0001					
	ST/BVE	-0.0525					
	LT/BVE	-0.0280					
	C/BVE	-0.0122					
BIC	TD/TE	-0.0220					
	ST/MVE	0.0001					
	LT/MVE	0.0000					
	C/MVE	0.0001					
	ST/BVE	-0.0525					
	LT/BVE	-0.0280					
	C/BVE	-0.0122					

Table 5.3: Empirical relationship between debt ratio and determinants of capital structure from our formulation applied to real data. Note that $\surd$: consistent with theory, X: inconsistent with theory, blank space: not provide in the theory.

Determinants of capital structure	Empirical Relationship	Trade-off theory	Pecking order theory
Growth	-	$\surd$	X
Uniqueness	-	$\surd$	
Non-debt tax shields	-	$\surd$	
Collateral value of assets	+		
Profitability	-	X	$\surd$
Volatility	0	$\surd$	$\surd$

CHAPTER VI

CONCLUSION

This thesis provide a scheme to identify the determinants of capital structure whose involved factors are related via a MIMIC model. Two formulations are proposed to select effective determinants of capital structure and to estimate the relationship between the determinants and the measures of capital structure based on the MIMIC model. The first formulation is applied to select highly effective determinants (latent variables). It is a least-squares problem with a 1-norm penalty, says group lasso problem, to induce a zero structure in the model. Consequently, the procedure of this formulation is that unimportant latent variables will be removed based on the sparsity pattern of parameters of the model. When ineffective latent variables are removed, the second formulation is least-squares estimation of reduced MIMIC model to find the remaining parameters of the model illustrating the relationship between the remaining latent and observed variables. Since the two proposed formulations are biconvex, they are solved by commonly well-known alternating minimization. According to the simulation experiment, as we assume true coefficient matrices with some zero columns, our proposed method can remove latent variables correctly with the use of some suitable regularization parameter. The results show that the performance of latent variable selection increases when the sample size or zero columns in A_{true} increases based on applying BIC to select a model. Moreover, we provide the comparison of performance between our solution and [Chang et al., 2009] 's solution. [Chang et al., 2009] 's method and ours perform relatively well with low total error of number of predicted latent variables. In particularly, our solution apparently outperforms [Chang et al., 2009] 's in the case of sparse A_{true} since our formulation is principally proposed for removing ineffective latent variables, it performs well if a true model is sparse. Moreover, our method provides very low False Positive indicating a good performance of ineffective latent variables prediction. According to empirical results, to select an appropriate model in real application data from seven industries in the North America, we apply AIC, AICc, BIC, KIC, and KICc depending on preferences of users. The normalized relative impacts of determinants of capital structure show that growth is the most influential determinants agreed by all information criterions and [Chang et al., 2009] 's result. On the contrary, volatility is ineffective determinant since it is not selected by any criterion. Moreover, we compare our results to the trade-off theory and pecking order theory to perform an appropriate decision to manage capital in a firm. The direction of relationship between debt ratios and i) growth (-), ii) uniqueness (-), and iii) non-debt tax shields (-) are consistent with the trade-off theory while profitability (+) is consistent with the pecking-order theory.

Bibliography

[Akaike, 1987] Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3): 317–332.

[Akaike, 2011] Akaike, H. (2011). Akaike's Information Criterion. In *International Encyclopedia of Statistical Science*, 25–25. Springer.

[Akaike et al., 1998] Akaike, H., Petrov, B., and Csaki, F. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, 199–213. Springer.

[Anderson and Burnham, 2002] Anderson, D. and Burnham, K. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, 66(3): 912–918.

[Bach et al., 2008] Bach, F., Mairal, J., and Ponce, J. (2008). Convex sparse matrix factorizations. *arXiv:0812.1869*.

[Bany-Ariffin and Jr, 2012] Bany-Ariffin, A. and Jr, C. M. (2012). Trade off theory against pecking order theory of captial structure in a nested model: Panel GMM evedence from South Africa. *The Global Journal of Finance and Economics*, 9(2): 133–147.

[Berger et al., 1997] Berger, P., Ofek, E., and Yermack, D. (1997). Managerial entrenchment and capital structure decisions. *The Journal of Finance*, 52(4): 1411–1438.

[Bollen, 2014] Bollen, K. (2014). *Structural Equations with Latent Variables*. John Wiley & Sons.

[Booth et al., 2001] Booth, L., Aivazian, V., Demirguc-Kunt, A., and Maksimovic, V. (2001). Capital structures in developing countries. *The Journal of Finance*, 56(1): 87–130.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press.

[Bradley et al., 1984] Bradley, M., Jarrell, G., and Kim, E. (1984). On the existence of an optimal capital structure: Theory and evidence. *The Journal of Finance*, 39(3): 857–878.

[Cameron and Trivedi, 2005] Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. Cambridge university press.

[Cavanaugh, 1999] Cavanaugh, J. (1999). A Large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, 42(4): 333–343.

[Chang et al., 2009] Chang, C., Lee, A., and Lee, C. (2009). Determinants of capital structure choice: A Structural equation modeling approach. *The Quarterly Review of Economics and Finance*, 49(2): 197–213.

[Chen and Huang, 2012] Chen, L. and Huang, J. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500): 1533–1545.

[Chen et al., 2001] Chen, S., Donoho, D., and Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, 43(1): 129–159.

[Chen et al., 2011] Chen, X., Qi, Y., Bai, B., Lin, Q., and Carbonell, J. (2011). Sparse latent semantic analysis. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, 474–485. SIAM.

[DeAngelo and Masulis, 1980] DeAngelo, H. and Masulis, R. (1980). Optimal capital structure under corporate and personal taxation. *Journal of Financial Economics*, 8(1): 3–29.

[Deerwester et al., 1990] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391.

[Deesomsak et al., 2004] Deesomsak, R., Paudyal, K., and Pescetto, G. (2004). The Determinants of capital structure: Evidence from the Asia Pacific region. *Journal of Multinational Financial Management*, 14(4): 387–405.

[Dell'Anno et al., 2007] Dell'Anno, R., Gómez-Antonio, M., and Pardo, A. (2007). The Shadow economy in three Mediterranean countries: France, Spain and Greece. A MIMIC approach. *Empirical Economics*, 33(1): 51–84.

[Fama and French, 2002] Fama, E. and French, K. (2002). Testing trade-off and pecking order predictions about dividends and debt. *The Review of Financial Studies*, 15(1): 1–33.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360.

[Fazel, 2002] Fazel, M. (2002). *Matrix rank minimization with applications*. PhD thesis, Stanford University.

[Frank and Goyal, 2011] Frank, M. and Goyal, V. (2011). Trade-off and pecking order theories of debt. *Handbook of Empirical Corporate Finance*, 2: 135.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer series in statistics Springer, Berlin.

[Fu, 1998] Fu, W. (1998). Penalized regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3): 397–416.

[Gallo et al., 1994] Gallo, J., Anthony, J., and Muthén, B. (1994). Age differences in the symptoms of depression: A Latent trait analysis. *Journal of Gerontology*, 49(6): 251–264.

[Gillis, 2012] Gillis, N. (2012). Sparse and unique nonnegative matrix factorization through data preprocessing. *Journal of Machine Learning Research*, 13: 3349–3386.

[Gorski et al., 2007] Gorski, J., Pfeuffer, F., and Klamroth, K. (2007). Biconvex sets and optimization with biconvex functions: A Survey and extensions. *Mathematical Methods of Operations Research*, 66(3): 373–407.

[Grantl et al., 2014] Grantl, M., Boyd, S., and Ye, Y. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

[Hardiyanto et al., 2015] Hardiyanto, A., Achsani, N., Sembel, R., and Maulana, N. (2015). Ownership and determinants capital structure of public listed companies in Indonesia: A Panel data analysis. *International Research Journal of Business Studies*, 6(1): 29–43.

[Hastie et al., 2015] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

[Hillier et al., 2010] Hillier, D., Ross, S., Westerfield, R., Jaffe, J., and Jordan, B. (2010). *Corporate Finance*. McGraw Hill.

[Horn and Johnson, 2012] Horn, R. and Johnson, C. (2012). *Matrix Analysis*. Cambridge university press.

[Jamesh et al., 2013] Jamesh, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

[Jensen and Meckling, 1976] Jensen, M. and Meckling, W. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4): 305–360.

[Jöreskog, 1970] Jöreskog, K. (1970). A General method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2): i–41.

[Jöreskog and Goldberge, 1975] Jöreskog, K. and Goldberge, A. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association*, 70(351a): 631–639.

[Kharratzadeh and Coates, 2016] Kharratzadeh, M. and Coates, M. (2016). Sparse multivariate factor regression. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, 1–5. IEEE.

[Lomax and Schumacker, 2012] Lomax, R. and Schumacker, R. (2012). *A Beginner's Guide to Structural Equation Modeling*. Routledge Academic New York, NY.

[Mairal et al., 2010] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11: 19–60.

[Marill and Green, 1963] Marill, T. and Green, D. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9(1): 11–17.

[Matemilola et al., 2013] Matemilola, B., Bany-Ariffin, A., and Jr, C. M. (2013). Unobservable effects and firm's capital structure determinants. *Managerial Finance*, 39(12): 1124–1137.

[Mazumder et al., 2010] Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11: 2287–2322.

[Mazur, 2007] Mazur, K. (2007). The Determinants of capital structure choice: Evidence from Polish companies. *International Advances in Economic Research*, 13(4): 495–514.

[Modigliani and Miller, 1958] Modigliani, F. and Miller, M. (1958). The Cost of capital, corporation finance and the theory of investment. *The American Economic Review*, 48(3): 261–297.

[Modigliani and Miller, 1963] Modigliani, F. and Miller, M. (1963). Corporate income taxes and the cost of capital: A Correction. *The American Economic Review*, 53(3): 433–443.

[Myers, 1977] Myers, S. (1977). Determinants of corporate borrowing. *Journal of Financial Economics*, 5(2): 147–175.

[Myers, 1984] Myers, S. (1984). The Capital structure puzzle. *The Journal of Finance*, 39(3): 574–592.

[Myers and Majluf, 1984] Myers, S. and Majluf, N. (1984). Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics*, 13(2): 187–221.

[Neyshabur and Panigrahy, 2013] Neyshabur, B. and Panigrahy, R. (2013). Sparse matrix factorization. *arXiv:1311.3315*.

[Osborne et al., 2000a] Osborne, M., Presnell, B., and Turlach, B. (2000a). A New approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3): 389–403.

[Osborne et al., 2000b] Osborne, M., Presnell, B., and Turlach, B. (2000b). On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2): 319–337.

[Parikh et al., 2014] Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3): 127–239.

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572.

[Penny et al., 2004] Penny, W., Stephan, K., Mechelli, A., and Friston, K. (2004). Modelling functional integration: A Comparison of structural equation and dynamic causal models. *Neuroimage*, 23: S264–S274.

[Perkins et al., 2003] Perkins, S., Lacker, K., and Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 3: 1333–1356.

[Raykov and Marcoulides, 2012] Raykov, T. and Marcoulides, G. (2012). *A First Course in Structural Equation Modeling*. Routledge.

[Rennie and Srebro, 2005] Rennie, J. and Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, 713–719. ACM.

[Richard et al., 2014] Richard, E., Obozinski, G., and Vert, J. (2014). Tight convex relaxations for sparse matrix factorization. In *Advances in Neural Information Processing Systems*, 3284–3292.

[Sardy et al., 2000] Sardy, S., Bruce, A., and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics*, 9(2): 361–379.

[Schmidt, 2005] Schmidt, M. (2005). Least squares optimization with ℓ_1-norm regularization.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464.

[Seghouane, 2006] Seghouane, A. (2006). Vector autoregressive model-order selection from finite samples using Kullback's symmetric divergence. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(10): 2327–2335.

[Serghiescu and Văidean, 2014] Serghiescu, L. and Văidean, V. (2014). Determinant factors of the capital structure of a firm- an empirical analysis. *Procedia Economics and Finance*, 15: 1447–1457.

[Shevade and Keerthi, 2003] Shevade, S. and Keerthi, S. (2003). A Simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17): 2246–2253.

[Songsiri, 2015] Songsiri, J. (2015). Learning multiple granger graphical models via group fused Lasso. In *Proceedings of the 10th Asian Control Conference (ASCC), 2015*, 1–6. IEEE.

[Srebro et al., 2005] Srebro, N., Rennie, J., and Jaakkola, T. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, 1329–1336.

[Stapleton, 1978] Stapleton, D. (1978). Analyzing political participation data with a MIMIC model. *Sociological Methodology*, 15(1): 52–74.

[Swanson et al., 2003] Swanson, Z., Srinidhi, B., and Seetharaman, A. (2003). *The Capital Structure Paradigm: Evolution of Debt/Equity Choices*. Greenwood Publishing Group.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58: 267–288.

[Titman, 1984] Titman, S. (1984). The Effect of capital structure on a firm's liquidation decision. *Journal of Financial Economics*, 13(1): 137–151.

[Titman and Wessels, 1988] Titman, S. and Wessels, R. (1988). The Determinants of capital structure choice. *The Journal of Finance*, 43(1): 1–19.

[Whitehurst, 2003] Whitehurst, D. (2003). *Corporate Finance*. McGraw-Hill/Irwin.

[Whitney, 1971] Whitney, A. (1971). A Direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, 100(9): 1100–1103.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67.

[Zhang et al., 2012] Zhang, Y., d′Aspremont, A., and Ghaoui, L. E. (2012). Sparse PCA: Convex relaxations, Algorithms and Applications. *Handbook on Semidefinite, Conic and Polynomial Optimization*, 166: 915–940.

APPENDICES

APPENDICES

7.1 The vectorized form of the proposed formulation

This section will explain the transformation of matrix form of (4.15), and (4.16) to the vector form as (4.14) by vectorization. Let

$$\begin{aligned} A &= \begin{bmatrix} A_1 & A_2 & \cdots & A_m \end{bmatrix} = [a_{ij}], \text{for } i = 1, \dots, q \text{ and } j = 1, \dots, m, \\ B &= \begin{bmatrix} B_1 & B_2 & \cdots & B_m \end{bmatrix} = [b_{ij}], \text{for } i = 1, \dots, \tilde{p} \text{ and } j = 1, \dots, m, \\ Y_{\text{vec}} &= \begin{bmatrix} y_{11} & \cdots & y_{q1} & y_{12} & \cdots & y_{q2} & \cdots & y_{1N} & \cdots & y_{qN} \end{bmatrix}^T. \end{aligned}$$

The problem (4.13),

$$\begin{aligned} \underset{B}{\text{minimize}} \quad & \|Y - AB^T\tilde{X}\|_F^2, \\ \text{subject to} \quad & \mathrm{P}(B) = 0, \end{aligned}$$

can be written as a vector form as

$$\begin{aligned} \underset{\tilde{\beta}}{\text{minimize}} \quad & \|w - \tilde{Z}\tilde{\beta}\|_2^2 \\ \text{subject to} \quad & \tilde{\beta}_i = 0 \quad \forall i \in \Omega, \end{aligned} \tag{7.1}$$

where Ω is a set of zero entries in B explained from the Figure 1.2.

Consider $AB^TX =$

$$\begin{bmatrix} x_{11}A_1b_{11} & \cdots & x_{p1}A_1b_{p1} & x_{11}A_2b_{12} & \cdots & x_{p1}A_2b_{p2} & \cdots & x_{11}A_rb_{1r} & \cdots & x_{p1}A_rb_{pr} \\ x_{12}A_1b_{11} & \cdots & x_{p2}A_1b_{p1} & x_{12}A_2b_{12} & \cdots & x_{p2}A_2b_{p2} & \cdots & x_{12}A_rb_{1r} & \cdots & x_{p2}A_rb_{pr} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1N}A_1b_{11} & \cdots & x_{pN}A_1b_{p1} & x_{1N}A_2b_{12} & \cdots & x_{pN}A_2b_{p2} & \cdots & x_{1N}A_rb_{1r} & \cdots & x_{pN}A_rb_{pr} \end{bmatrix}.$$

Consequently, $\mathbf{vec}(Y - AB^TX) =$

$$\underbrace{\begin{bmatrix} y_{11} \\ \vdots \\ y_{q1} \\ \vdots \\ y_{1N} \\ \vdots \\ y_{qN} \end{bmatrix}}_{w} - \underbrace{\begin{bmatrix} x_{11}A_1 & \cdots & x_{p1}A_1 & x_{11}A_2 & \cdots & x_{p1}A_2 & \cdots & x_{11}A_r & \cdots & x_{p1}A_r \\ x_{12}A_1 & \cdots & x_{p2}A_1 & x_{12}A_2 & \cdots & x_{p2}A_2 & \cdots & x_{12}A_r & \cdots & x_{p2}A_r \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1N}A_1 & \cdots & x_{pN}A_1 & x_{1N}A_2 & \cdots & x_{pN}A_2 & \cdots & x_{1N}A_r & \cdots & x_{pN}A_r \end{bmatrix}}_{\tilde{Z}} \underbrace{\begin{bmatrix} b_{11} \\ \vdots \\ b_{p1} \\ \vdots \\ b_{1r} \\ \vdots \\ b_{pr} \end{bmatrix}}_{\tilde{\beta}}.$$

Then, we plug $\tilde{\beta}_i = 0$ into the objective function and the problem (7.1) is transformed to unconstrained problem as the problem (4.14) where β is $\tilde{\beta}$ with its zero removed and Z is $\tilde{Z}$ that is removed columns related to zero of $\tilde{\beta}$ and zero columns of A from the problem (4.12). Consequently, the problem (4.13) can be vectorized to the problem (4.14).

For the problem (4.15):

$$\underset{A}{\text{minimize}} \quad \|Y - AB^T\tilde{X}\|_F^2,$$

it can be written as a vector form as:

$$\underset{\beta}{\text{minimize}} \quad \|w - Z\beta\|_2^2.$$

Suppose (i, j) is an index of matrix that we consider, the problem (4.15) can be rewritten as

$$\begin{aligned} \|Y - AB^T\tilde{X}\|_F^2 &= \textstyle\sum_{i,j}(y_{ij} - (AB^T\tilde{X})_{ij})^2 \\ &= \textstyle\sum_{i,j}(y_{ij} - \sum_{k=1}^{m} a_{ik}g_{kj})^2 \\ &= \textstyle\sum_{i,j}(y_{ij} - \sum_{k=1}^{m} g_{kj}a_{ik})^2 \end{aligned} \qquad (7.2)$$

where $g_{kj} = \sum_{l=1}^{\tilde{p}} b_{kl}x_{lj}$.
If we sort (7.2) sorted by $i = 1, \ldots, q$ and fixed j can be written in a vector form as

$$\mathbf{vec}(Y - AB^TX) = \begin{bmatrix} y_{1j} \\ y_{1j} \\ \vdots \\ y_{qj} \end{bmatrix} - \begin{bmatrix} g_{1t} & & & & \cdots & g_{mt} & & & \\ & g_{1t} & & & \cdots & & g_{mt} & & \\ & & \ddots & & \cdots & & & \ddots & \\ & & & g_{1t} & \cdots & & & & g_{mt} \end{bmatrix} \begin{bmatrix} a_{11} \\ \vdots \\ a_{q1} \\ \vdots \\ a_{1m} \\ \vdots \\ a_{qm} \end{bmatrix} . \qquad (7.3)$$

When (7.3) is sorted by $j = 1, \ldots, N$, we get

$$\mathbf{vec}(Y - AB^TX) = \underbrace{\begin{bmatrix} y_{11} \\ \vdots \\ y_{q1} \\ \vdots \\ y_{1N} \\ \vdots \\ y_{qN} \end{bmatrix}}_{w} - \underbrace{\begin{bmatrix} g_{11} & & & & \cdots & g_{m1} & & & \\ & g_{11} & & & \cdots & & g_{m1} & & \\ & & \ddots & & \cdots & & & \ddots & \\ & & & g_{11} & \cdots & & & & g_{m1} \\ g_{12} & & & & \cdots & g_{m2} & & & \\ & g_{12} & & & \cdots & & g_{m2} & & \\ & & \ddots & & \cdots & & & \ddots & \\ & & & g_{12} & \cdots & & & & g_{m2} \\ & & & \vdots & & & & \vdots & \\ g_{1N} & & & & \cdots & g_{mN} & & & \\ & g_{1N} & & & \cdots & & g_{mN} & & \\ & & \ddots & & \cdots & & & \ddots & \\ & & & g_{1N} & \cdots & & & & g_{mN} \end{bmatrix}}_{Z} \underbrace{\begin{bmatrix} a_{11} \\ \vdots \\ a_{q1} \\ \vdots \\ a_{1m} \\ \vdots \\ a_{qm} \end{bmatrix}}_{\beta}.$$

Consequently, the problem (4.15) can be vectorized to the problem (4.14).

The vector form of (4.16) is derived in the same way as the problem (4.13) except that the dimensions of A, B, and X are reduced based on zero columns in A and B.

7.2 Log-likelihood of the model

This section describes the derivation of log-likelihood of the model which is required in model selection criterion. According to the reduced MIMIC model in term of A and B,

$$y = AB^Tx + \epsilon,$$

given N independent observations of data set $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ where $x^{(i)} \in \mathbf{R}^p$ and $y^{(i)} \in \mathbf{R}^q$ and $\epsilon \sim \mathcal{N}(0, \Sigma)$. The likelihood of y for parameter θ is defined as:

$$\begin{aligned} f(y \mid x; \theta) &= f(y \mid x; A, B, \Sigma) \\ &= f(y^{(1)}, \ldots, y^{(N)} | x^{(1)}, \ldots, x^{(N)}; A, B, \Sigma) \\ &= \frac{1}{(2\pi)^{N/2}(\det\Sigma)^{N/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^N (y^{(i)} - AB^Tx^{(i)})^T \Sigma^{-1} (y^{(i)} - AB^Tx^{(i)}) \right) \\ &= \frac{1}{(2\pi)^{N/2}(\det\Sigma)^{N/2}} \exp\left(-\frac{1}{2} \mathbf{tr}(Y - AB^TX)^T \Sigma^{-1} (Y - AB^TX) \right). \end{aligned}$$

Log-likelihood function of the model for parameters A, B, and Σ is written as:

$$\begin{aligned}
\mathcal{L}(A,B,\Sigma) \quad &= \log f(y|x;A,B,\Sigma) \\
&= -\frac{1}{2}\Bigg(\mathbf{tr}((Y-AB^TX)^T\Sigma^{-1}(Y-AB^TX)) + N\log\det\Sigma + N\log 2\pi \Bigg) \\
-2\mathcal{L}(A,B,\Sigma) \quad &= N\Bigg(\mathbf{tr}\,\big(\frac{(Y-AB^TX)(Y-AB^TX)^T}{N}\Sigma^{-1}\big) + \log\det\Sigma + \log 2\pi \Bigg) \\
&= N\Bigg(\mathbf{tr}(\Sigma\Sigma^{-1}) + \log\det\Sigma + \log 2\pi \Bigg) \\
&= Nq + N\log\det\Sigma + N\log 2\pi. \qquad (7.4)
\end{aligned}$$

If we choose $\hat{\Sigma}$ to be the maximum likelihood estimators;

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\hat{\epsilon}^{(i)}\hat{\epsilon}^{(i)T} = \frac{1}{N}(Y-\hat{A}\hat{B}^TX)(Y-\hat{A}\hat{B}^TX)^T.$$

To calculate AIC, AICc, BIC, KIC, and KICc, we plug (7.4) into (4.7). However, constant terms do not affect to the minimization of information criterion score so we ignore the constant terms, then divided by N to scale down the information criterion score as:

$$\begin{aligned}
\text{Normalized AIC} \quad &= \quad \log\det\hat{\Sigma} + \tfrac{2d}{N} \\
\text{Normalized AICc} \quad &= \quad \log\det\hat{\Sigma} + \tfrac{2d}{N-d-1} \\
\text{Normalized BIC} \quad &= \quad \log\det\hat{\Sigma} + \tfrac{d\,\log N}{N} \\
\text{Normalized KIC} \quad &= \quad \log\det\hat{\Sigma} + \tfrac{3d}{N} \\
\text{Normalized KICc} \quad &= \quad \log\det\hat{\Sigma} + \tfrac{d}{N(N-d)}.
\end{aligned}$$

7.3 Consistency of zero column in warm start method

This section explain a consistency of estimating zero columns of the solution in the problem (4.12) solving by Alternating Direction Method of Multipliers (ADMM) with warm start that we mention in 4.4.3.

ADMM is a well-known numerical algorithm to solve convex optimization problem by separating the objective function into two parts of convex function and inviting some auxiliary variables. [Songsiri, 2015] apply ADMM which is a general efficient numerical algorithm to solve the group lasso problem characterized as an ℓ_1 regularized problem that is convex optimization problem and we follow methodology from this paper.

According to the group lasso problem (4.12),

$$\underset{A}{\text{minimize}} \quad \|Y-AB^TX\|_F^2 + \gamma\textstyle\sum_{i=1}^{r}\|A_i\|_2,$$

it can be written in a vector form as following:

$$\underset{a}{\text{minimize}} \quad \|Ga - y_{\text{vec}}\|_2^2 + \gamma\|a\|_{2,1} \tag{7.5}$$

where $\|\cdot\|_{2,1}$ is a sum of 2-norm of block vector, for example, denote $a = (a_1, \dots, a_L)$,

$$\|a\|_{2,1} = \sum_{i=1}^{L} \|a_i\|_2.$$

Besides, the entries in $G \in \mathbf{R}^{qN \times pr}$ are functions of B and X (it is derived in the same way as Z in the problem (4.15)), $y_{\text{vec}} \in \mathbf{R}^{qN}$ is derived from vectorization of Y, and $a \in \mathbf{R}^{pr}$ is derived from vectorization of A.

To derive consistency of zero column in warm start method, we prove that if A_j and B_j equal zero at γ_k, then A_j and B_j equal zero at γ_{k+1} under the assumption of structure of G. Since we apply alternating minimization to firstly solve A and then solve B, if $A_j = 0$, then $B_j = 0$ automatically. Thereby, we can only prove that if $B_j = 0$ at γ_k, then $A_j = 0$, saying $a_j = 0$ at γ_{k+1}.

Before solving (7.5), we have to rearrange this problem into ADMM format as

$$\begin{array}{ll} \underset{x_1, x_2}{\text{minimize}} & \frac{1}{2}\|Gx_1 - y_{\text{vec}}\|_2^2 + \gamma\|x_2\|_{2,1} \\ \text{subject to} & x_1 = x_2. \end{array} \tag{7.6}$$

In each update step, ADMM algorithm minimizes the augmented Lagrangian defined as following:

$$L(x_1, x_2, z) = \frac{1}{2}\|Gx_1 - y_{\text{vec}}\|_2^2 + \gamma\|x_2\|_{2,1} + z^T(x_1 - x_2) + \frac{\rho}{2}\|x_1 - x_2\|_2^2$$

with variables $x_1 \in \mathbf{R}^{pr}, x_2 \in \mathbf{R}^{pr}$, and $z \in \mathbf{R}^{pr}$ (dual variable), respectively, where the penalty parameter $\rho > 0$ controls a speed of convergence. This problem has been proposed in [Songsiri, 2015], therefore we follow the update steps from this paper and it has been described as follows:

$$\begin{array}{rcl} x_1^+ & = & \underset{x_1}{\text{argmin}} \quad \frac{1}{2}\|Gx_1 - y_{\text{vec}}\|_2^2 + z^T(x_1 - x_2) + \frac{\rho}{2}\|x_1 - x_2\|_2^2 \\ & = & (G^T G + \rho I)^{-1}(G^T y_{\text{vec}} - z + \rho x_2), \\ x_2^+ & = & \underset{x_2}{\text{argmin}} \quad \gamma\|x_2\|_{2,1} + z^T(x_1 - x_2) + \frac{\rho}{2}\|x_1 - x_2\|_2^2 \\ & = & \mathbf{prox}_{\gamma/2\rho, L}\big(x_1^+ + \frac{z}{\rho}\big) \quad \text{where } \mathbf{prox} \text{ is proximal operator of } \|x_2\|_{2,1} \text{ [Parikh et al., 2014]}, \\ z^+ & = & z + \rho(x_1^+ - x_2^+) \end{array} \tag{7.7}$$

until stopping criterion is satisfied.

According to the update step, we will show that if $B_j = 0$, then $(x_1^+)_j$ and $(x_2^+)_j = 0$ at γ_k since this can imply the statement that if $B_j = 0$ at γ_k, then $A_j = 0$, saying $a_j = 0$ at γ_{k+1}.

Let us start with x_1-update step. Suppose $B_j = 0$, we will show that $(x_1^+)_j = 0$. Denote j^{th} block row of G, saying $(G)_j$, is $(jq - q + 1)^{\text{th}}$ to $(jq)^{\text{th}}$ rows of G. Suppose j^{th} column of B is zero, we are interested in j^{th} block row of $(G^TG + \rho I)^{-1}$. Consider a permutation matrix P that permute the 1st and j^{th} block row, *e.g.*, $P = \begin{bmatrix} e_j & e_2 & \cdots & e_1 & e_{j+1} & \cdots \end{bmatrix}^T$ that permutes $G^TG + \rho I$ so that the zero block rows and columns become the first block rows and columns. Let 0 be a zero sub-matrix and $\square$ be a sub-matrix which does not involved in derivation. If $B_j = 0$, then

$$\begin{aligned} P(G^TG + \rho I)P^T &= \begin{bmatrix} \rho I & 0 \\ 0 & \square \end{bmatrix}, \\ \big(P(G^TG + \rho I)P^T\big)^{-1} &= \begin{bmatrix} \rho^{-1} I & 0 \\ 0 & \square \end{bmatrix}, \\ \big(P(G^TG + \rho I)P^T\big)_1^{-1} &= \begin{bmatrix} \rho^{-1} I & 0 \end{bmatrix}. \end{aligned} \tag{7.8}$$

From $P = P^T$ and $P = P^{-1}$, P_1 is the first row of P,

$$\begin{aligned} \big(P(G^TG + \rho I)P^T\big)_1^{-1} &= P_1\big(P(G^TG + \rho I)P\big)^{-1}P \\ &= e_j^T\big(P(G^TG + \rho I)P\big)^{-1}P \\ &= \big(G^TG + \rho I\big)_j^{-1}P \\ \text{so we obtain} \quad \big(G^TG + \rho I\big)_j^{-1}P &= \begin{bmatrix} \rho^{-1} I & 0 \end{bmatrix} \qquad \text{(from (7.8))}. \\ \text{Consequently,} \quad \big(G^TG + \rho I\big)_j^{-1} &= \begin{bmatrix} \rho^{-1} I & 0 \end{bmatrix} P^{-1} \\ &= \begin{bmatrix} \rho^{-1} I & 0 \end{bmatrix} P \\ &= \begin{bmatrix} \rho^{-1} I & 0 \end{bmatrix} \begin{bmatrix} e_j & e_2 & \cdots & e_1 & e_{j+1} & \cdots \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & \cdots & \rho^{-1} I & \cdots & 0 & 0 \end{bmatrix}. \end{aligned}$$

where $\rho^{-1}I$ is in j^{th} block columns. Therefore, we can always permutes $G^TG + \rho I$ so that the zero block rows and columns become the first block rows and columns and if $B_j = 0$, we can conclude that

$$\begin{aligned} \big(G^TG + \rho I\big)_j^{-1} &= \begin{bmatrix} 0 & 0 & \cdots & \rho^{-1} I & \cdots & 0 & 0 \end{bmatrix}, \qquad \text{and} \\ (G^Tb)_j &= G_j^Tb \;=\; 0. \end{aligned} \tag{7.9}$$

Note that, when $B_j = 0$, we have $(x_2)_j = 0$. Next, we need $(z)_j = 0$ to get $(x_1^+)_j = 0$. We can show that $(z)_j = 0$ at k^{th} iteration when $k \geq 1$ for any z_0. From (7.7) and denote superscript show iteration of update step, $(z^1)_j = (z^0)_j + \rho(x_1^1)_j = 0$ since $(x_1^1)_j = \frac{-(z^0)_j}{\rho}$ from the structure of

$\big(G^TG+\rho I\big)_j^{-1}$ in (7.9). Consequently, when $B_j = 0$, then

$$\begin{aligned}
(x_1^+)_j &= \big(G^TG+\rho I\big)_j^{-1}(G^Tb+\rho x_2 - z)\\
&= \begin{bmatrix}0 & 0 & \cdots & \rho^{-1}I & \cdots & 0 & 0\end{bmatrix}\left(\underbrace{\begin{bmatrix}\square\\ \square\\ 0\\ \square\\ \square\end{bmatrix}}_{G^Tb} + \rho\underbrace{\begin{bmatrix}\square\\ \square\\ 0\\ \square\\ \square\end{bmatrix}}_{x_2} - \underbrace{\begin{bmatrix}\square\\ \square\\ 0\\ \square\\ \square\end{bmatrix}}_{z}\right)\\
&= 0,\\
\text{and}\qquad (x_2^+)_j &= \big(\mathbf{prox}_{\gamma/2\rho,L}\big(x_1^+ + \tfrac{z}{\rho}\big)\big)_j\\
&= \max\left\{1 - \frac{\gamma/2\rho}{\|(x_1^+)_j + \frac{(z)_j}{\rho}\|_2^2}\,,\ 0\right\}\left((x_1^+)_j + \frac{(z)_j}{\rho}\right)\\
&= 0 \qquad \text{since } (x_1^+)_j \text{ and } (z)_j \text{ are zero.}
\end{aligned}$$

Since $(x_1^+)_j = 0$ and $(x_2^+)_j = 0$, we can conclude that $a_j^+ = 0$, in other words, if $B_j = 0$ at k^{th} iteration then $a_j = 0$ at $(k+1)^{\text{th}}$ iteration, *i.e.*, zero columns will not return to be nonzero when γ increases.

7.4 MATLAB codes of the proposed formulations

This section provides MATLAB codes of functions: *latent_selection()* to solve the first proposed formulation (4.5) "latent variable selection", *reduced_mimic()* to solve the second proposed formulation (4.6) "least-squares estimation for reduced MIMIC model", and example of MATLAB codes for model selection. Note that the function *latent_selection()* contains subfunctions *norm21()*, *prox_sumof2norm*, and *group_lasso()* following the method and codes from [Songsiri, 2015]; moreover, the function *shape()* is runned before applying *reduced_mimic()* since it removes zero columns of A and B from *latent_selection()*. These codes are used in the simulation process and experiment for real data described in the following.

7.4.1 MATLAB codes of latent variable selection

In this part, MATLAB codes for solving the formulation of latent variable selection (4.5) is provided corresponding to the numerical method in the section 4.4.1 as follows:

```
function [model_latent.A ,model_latent.B,history_latent] = latent_reduction(X,Y,gamma,Ainit,Binit,IND,rho);

% solves the following problem via Alternating minization:
% minimize || Y-AB'X ||_F^2 + \gamma sum(norm(Ai))
% where Ai is i-th column of A
% with variables A and B

% it requires input as follows.
% 1) X is p*N matrix where N is number of observations and p is number of variables in x
% 2) Y is q*N matrix where q is number of variables in y
% 3) gamma is regularization parameter
% 4) Ainit is a q*r initial condition matrix for A
% 5) Binit is a p*r initial condition matrix for B
% 6) IND is  linear indices of nonzero elements in B
% 7) rho is the augmented Lagrangian parameter used in group lasso

% and it returns output
% 1) model_latent.A and model_latent.B: A and B that are optimal solutions.
% 2) history_latent is a structure that contains
% history_latent.obj_value: the objective value,
% history_latent.uniqueB: flag for checking uniqueness of B (return 1 when the solution is unique and 0 when the solution is not unique),
% history_latent.conv_A and history_latent.conv_B: relative error of A and B for each iteration,


N=length(X(1,:));
[q r]=size(Ainit);
[p r]=size(Binit);

epsilon_obj=1e-4;   %relative tolerance for objective value
epsilon_A=1e-2;     %relative tolerance for A
epsilon_B=1e-2;     %relative tolerance for B
itermax=10000;

```

```
y_vec = Y(:);       %vector form of Y
X_rep=repmat(kron(X',ones(q,1)),1,r);
%use for solving B via alternating minimization

i=1;                %iteration count
while i<itermax
  if i==1;
    A = Ainit;
    B = Binit;
    history_latent.obj_value(i)  = ,...
    (norm(Y-A*B'*X,'fro'))^2+gamma*sum(norms(A));
    conv_obj_value(i) = history_latent.obj_value(i);

% check uniqueness of B by vectorizing the
% formulation given that A is fixed as follows:
% min || y_vec - H_tilde b ||_2^2 with variable 'b'
% where b is a vectorized form of B and checking the condition
% that H_tilde is full rank and skinny

% calculate H_tilde
    A_rep=kron(kron(ones(N,1),A),ones(1,p));
    H=A_rep.*X_rep;
    ind_z_B=find(B==zeros);     %zero elements in B
    H_tilde=H;
    H_tilde(:,ind_z_B)=[]; %H_tilde is H that remove zero col
    [rowH_tilde  colH_tilde]=size(H_tilde);
    rank_H_tilde=nnz(svd(H_tilde)~=0);
    %calculate rank(H_tilde) by SVD to avoid numerical error
     if rank_H_tilde == min(size(H_tilde)) && ,...
        rowH_tilde > colH_tilde %H_tilde is full rank and skinny
        history_latent.uniqueB(i)=1; %B is unique
     else
        history_latent.uniqueB(i)=0; %B is not unique
     end
 else

% solve A by fixing B and vetorizes the formulation as follows: min || y_vec - Ga ||_2^2 +\gamma sum(norm21(a,q)) with variable 'a' where a is a vectorized form of A

% calculate G for applying in lasso algorithm
     G=kron(X'*B,eye(q));
     ind_zero_G=find(norms(G)==0);
     G(:,ind_zero_G)=[];  %remove zero col in G

     ind_zero_col_B=find(norms(B)==0);    % index of zero col in B
     A_new=ones(q,r);
     A_new(:,ind_zero_col_B)=zeros;       % zero col of B
     if gamma==0
        a2_b=G\y_vec;  %when gamma=0; the problem is least squares.
     else
        [a1_b, a2_b,history]=group_lasso(G, y_vec,gamma,q,rho);
        %a2_b contains nz entries in A but may contains zero
     end
     ind_nz_A=find(A_new);   %nonzero entries in A
     A_new(ind_nz_A)=a2_b;

% solve B by fixing A and vetorize the formulation as
% follows: min || y_vec - H_tilde b ||_2^2 with variable 'b'

% calculate H_tilde
     A_rep=kron(kron(ones(N,1),A_new),ones(1,p));
     H=A_rep.*X_rep;
     ind_zero_col_A=find(norms(A_new)==0); % index of zero col in A
     B_new=ones(p,r);
     B_new(:,ind_zero_col_A)=zeros;
     %force zero col in B following zero col in A
     B_new(setdiff([1:p*r],IND))=zeros;
     %plug the constraint of zero path in B
     ind_z_B=find(B_new==zeros); %zero elements in B
     ind_nz_B=setdiff(1:p*r,ind_z_B); %nonzero elements in B

     H_tilde=H;
     H_tilde(:,ind_z_B)=[];  %H_tilde is H that remove zero col

% check uniqueness of B by checking the condition that
% H_tilde is skinny and full rank

     [rowH_tilde,colH_tilde]=size(H_tilde);
     rank_H_tilde=nnz(svd(H_tilde)~=0);
     if rank_H_tilde == min(size(H_tilde))&& ,...
     rowH_tilde > colH_tilde %H_tilde is full rank and skinny
        history_latent.uniqueB(i)=1; %B is unique
     else
        history_latent.uniqueB(i)=0; %B is not unique
     end
```

```

% calculate 'b' by least squares
     b=H_tilde\y_vec;
     B_new=zeros(p,r);
     B_new(ind_nz_B)=b;

% For convergence condition of objective function value
     history_latent.obj_value(i) = ,...
     (norm(Y-A_new*B_new'*X,'fro'))^2+gamma*sum(norms(A_new));
     conv_obj_value(i) = abs(history_latent.obj_value(i) - history_latent.obj_value(i-1))/abs(history_latent.obj_value(i-1));

     if (A == zeros)
        history_latent.conv_A(i)=0;
        history_latent.conv_B(i)=0;
     else
      history_latent.conv_A(i) = norm(A_new-A,'fro')/norm(A,'fro');
      history_latent.conv_B(i) = norm(B_new-B,'fro')/norm(B,'fro');
     end
         A=A_new;
         B=B_new;

     if (conv_obj_value(i) <= epsilon_obj) && (history_latent.conv_A(i) <=epsilon_A) && (history_latent.conv_B(i) <= epsilon_B)
          history_latent.A=A;
          history_latent.B=B;
          break;
     end
  end
     i=i+1;
end
```

- *group_lasso()* is applied to solve the group lasso problem (4.12) based on the vector form.

```
function [x1, x2,history] = group_lasso(G, b,lambda,p,rho, varargin)
% group_lassooff  Solve group lasso problem via ADMM
% [x, history] = group_lassooff(G, b,p,lambda, rho);
% solves the following problem via ADMM:
%
%   minimize || Gx - b ||_2^2 + \lambda sum(norm(x,2))
%
% nn is the total length of x
% p is the length of subblocks in x
%
% The solution is returned in the vector x, and the sparse version is in x2
% history is a structure that contains the objective value, the primal and
% dual residual norms, and the tolerances for the primal and dual residual
% norms at each iteration.
%
% rho is the augmented Lagrangian parameter.
%
% BLOCK_SIZE_SUM2NORM is an integer indicating the block size when computing the sum of norm
%
% varargin is 'initial condition' for x (optional)

PRINT_RESULT = 1;
FREQ_PRINT = 10;
MAXITERS = 20000;
ABSTOL = 1e-6;
RELTOL = 1e-6;

% store variables
nn = size(G,2);

% nn = n^2*p; np = (n^2-n)*p;
Gtb = G'*b;

L = chol(sparse(G'*G+rho) ,'lower');
L = sparse(L); U = L';


```

```
%% ADMM solver

optargin = size(varargin,2);

if optargin == 0,
    x1 = zeros(nn,1);
else
    x1 = varargin{1};
end
x2 = x1;
z = zeros(nn,1);

if ~PRINT_RESULT
    fprintf('%3s\t%10s\t%10s\t%10s\t%10s\t%10s\n', 'iter', ...
      'r norm', 'eps pri', 's norm', 'eps dual', 'objective');
end

for k = 1:MAXITERS
    % x1-update
    q = Gtb + (rho*x2-z);     % temporary value
    x1 = U \ (L \ q); % x1 is not generally sparse

    % x2-update
    x2old = x2;
    x2 = prox_sumof2norm(x1+z/rho,p,lambda/2/rho);

    % z-udpate
    z = z + rho*(x1 - x2);

    % stopping criterion
    obj = 0.5*norm(G*x1-b)^2+lambda/2*norm21(x2,p);

    history.objval(k) = obj;
    history.r_norm(k) = norm(x1-x2);
    history.s_norm(k)  = norm(rho*(x2-x2old));
    history.eps_pri(k) = sqrt(nn)*ABSTOL + RELTOL*max(norm(x1), norm(x2));
    history.eps_dual(k)= sqrt(nn)*ABSTOL + RELTOL*norm(z);

    if (PRINT_RESULT && mod(k,FREQ_PRINT) == 0)
        fprintf('%3d\t%10.4f\t%10.4f\t%10.4f\t%10.4f\t%10.2f\n', k, ...
            history.r_norm(k), history.eps_pri(k), ...
            history.s_norm(k), history.eps_dual(k), history.objval(k));
    end

    if (history.r_norm(k) < history.eps_pri(k) && ...
       history.s_norm(k) < history.eps_dual(k))
         break;
    end
end
end
```

- *norm21()* is applied to calculate a group norm of vector in *group_lasso()* .

```
function[y,w] = norm21(x,m)
% NORM21 return the group norm of an n-dimensional vector x
% where x = (x1,x2,...,xM) and x1,x2,...,xM are vectors of size m
%     In other words, we chop x in to k subvectors and each subvector has size m
%
% Therefore, mod(n,m) must be zero
%
% y = NORM21(x,m) = sum_{j=1}^M || xj ||_2
% w = [ ||x1|| ||x2|| ... ||xM|| ]

n = length(x);
if mod(n,m)~= 0
    error('mod(n,m) must be zero. Enter a new m');
else
    M = floor(n/m);
end

if m==1
    y = norm(x,1); % typical l1-norm
    w = abs(x);
else
    z = reshape(x,m,M); % z = [x1 x2 ... xM]
    w = norms(z,2)'; % use norms in CVX
    y = sum(w);
end
```

- *prox_sumof2norm* is applied to calculate the proximal operator in *group_lasso()*.

```
function [x] = prox_sumof2norm(u,p,a)

% PROX_SUMOF2NORM computes the proximal operator of a*f where
% f(x) = sum_{k=1}^K ||xk ||_2
% where xk is p x 1 and 'a' is a scalar
% The proximal operator of f is the block soft thresholding
% prox_{af}(u) _{kth block} = max(1- a/||uk||_2, 0 )* uk
% USAGE: [x] = prox_sumof2norm(u,p,a)
% u = (u1,u2,...,uN) where uk has size p x 1

n = length(u);
if mod(n,p)~= 0
    error('mod(n,p) must be zero. Enter a new p');
else
    M = floor(n/p);
end
z = reshape(u,p,M); % z = [u1 u2 ... uM]
w = norms(z,2); % use norms in CVX

z = z.*repmat(pos(1 - a./w),p,1);
x = z(:);
```

- *shape()* is applied to remove zero columns of A and B from latent variable selection to eliminate ineffective latent variables and also remove some rows of B related to such insignificant latent variables.

```
function [A_new_res,B_new_res,X_res] = shape(A,B,X,index_B)
%function shape is applied to delete zero columns of A and B
%and also some rows of B related to ineffective latent variables

%remove zero column of A
ind_zero_col_A = find(sum(abs(A)) == 0);
ind_zero_col_A =sort(ind_zero_col_A,'descend');
A_new_res = A;
A_new_res( :, ~any(A_new_res,1) ) = [];

%insert index_B
%remove zero columns of B (and also remove the rows corresponding to to those columns)
B_new_res = B;
X_res = X;
h=length(ind_zero_col_A);

  %remove row
for i=1:h
    for j=1:length(A)
        if ind_zero_col_A(:,i)==j
            ind_B_res=find(index_B(:,2)==j); % index
            B_new_res(ind_B_res ,:) = [];
            X_res(ind_B_res,:) = [];
        end
    end
end
  %remove col later
    B_new_res( :, ~any(B_new_res,1) ) = [];
end
```

7.4.2 MATLAB codes of least-squares estimation for reduced MIMIC model

In this part, MATLAB codes for solving the formulation of least-squares estimation for reduced MIMIC model (4.6) is provided corresponding to the numerical method in the section 4.4.2 as follows:

```
[model.A,model.B,model.Sigma,model.L,model.d,history_reducedMIMIC] = reduced_mimic(Y,X_res,Ainit,Binit,index_B)

% solves the following problem via Alternating minization:
%   minimize || Y-AB'X ||_F^2    with variables A and B

% it requires input as follows.
% 1) Y is q*N matrix where p is number of variables in y
% 2) X_res is p_tilde*N matrix where p_tilde is number of remaining variables in x
% 3) Ainit is a q*m initial condition matrix for A
% 4) Binit is a p*m initial condition matrix for B
% 5) index_B is indeices of B_ij that are fixed to be nonzero according to MIMIC path diagram by inseting first column for i and second column for j eg. nonzero elements are B_11, B_23, B_56 index_B=[1 1;2 3; 5 6].

% and it returns output
% 1) model.A and model.B: A and B that are optimal solutions,
% 2) model.Sigma: covariance matrix of prediction error,
% 3) model.L: likelihood value,
% 4) model.d: no.of free parameters.
% 5) history_reducedMIMIC is a structure that contains
% history_reducedMIMIC.conv_A and  history_reducedMIMIC.conv_B: relative error of A and B for each iteration,
% history_reducedMIMIC.obj_value: objective value,
% history_reducedMIMIC.uniqueA and B: flag for checking uniqueness of A and B (return 1 when the solution is unique  and 0 when the solution is not unique),


epsilon_obj=1e-4;   %relative tolerance for objective value
epsilon_A=1e-2;     %relative tolerance for A
epsilon_B=1e-2;     %relative tolerance for B
itermax=10000;

y_vec = Y(:);       %vector form of Y
N=length(Y(1,:));
```

```
i=1;                %iteration count
while i<itermax
  if i==1;
    A_res=Ainit;
    B_res=Binit;
    [q m]=size(A_res);
    [p m]=size(B_res);
    history_reducedMIMIC.obj_value(i)  = ,...
    (norm(Y-A_res*B_res'*X_res,'fro'))^2;
    conv_obj_value(i) = history_reducedMIMIC.obj_value(i);

% check uniqueness of A by vectorizing the
% formulation given that B is fixed as follows:
% min || y_vec - Ga ||_2^2 with variable 'a'
% where a is a vectorized form of A and checking the condition
% that G is full rank and skinny

% calculate G
     G=kron(X_res'*B_res,eye(q));
     [rowG colG]=size(G);
     rank_G=nnz(svd(G)~=0);
      if rank_G == min(size(G))&& rowG > colG %G is full rank and skinny
        history_reducedMIMIC.uniqueA(i)=1; %A is unique
     else
        history_reducedMIMIC.uniqueA(i)=0; %A is not unique
     end

% check uniqueness of B by vectorizing the
% formulation given that A is fixed as follows:
% min || y_vec - H_tilde b ||_2^2 with variable 'b'
% where b is a vectorized form of B and checking the condition
% that H_tilde is full rank and skinny

% calculate H_tilde
     X_res_rep=repmat(kron(X_res',ones(q,1)),1,m);
     A_res_rep=kron(kron(ones(N,1),A_res),ones(1,p));
     H=A_res_rep.*X_res_rep;
     ind_z_B_res=find(B_res==zeros);    %zero elements in B

     H_tilde=H;
     H_tilde(:,ind_z_B_res)=[]; %H_tilde is H that remove zero col
     [rowH_tilde  colH_tilde]=size(H_tilde);
     rank_H_tilde=nnz(svd(H_tilde)~=0);
     %calculate rank(H_tilde) by SVD to avoid numerical error
     if rank_H_tilde == min(size(H_tilde))&& rowH_tilde > colH_tilde %H_tilde is full rank and skinny
        history_reducedMIMIC.uniqueB(i)=1; %B is unique
     else
        history_reducedMIMIC.uniqueB(i)=0; %B is not unique
     end
  else
% solve A by fixing B and vetorize the formulation as
% follows: min || y_vec - Ga ||_2^2  with variable 'a'
% where 'a' is a vectorized form of A
% check uniqueness of A by checking the condition that
% G is full rank and skinny

% calculate G
     G=kron(X_res'*B_res,eye(q));
     [rowG colG]=size(G);
     rank_G=nnz(svd(G)~=0);
     if rank_G == min(size(G))&& rowG > colG %G is full rank and skinny
       history_reducedMIMIC.uniqueA(i)=1; %A is unique
     else
       history_reducedMIMIC.uniqueA(i)=0; %A is not unique
     end
%calculate 'a' by least squares
     a=G\y_vec;
     A_new_res=reshape(a,q,m);

% solve B by fixing A and vetorize the formulation as
% follows: min || y_vec - H_tilde b ||_2^2 with variable 'b'

% calculate H_tilde
     X_res_rep=repmat(kron(X_res',ones(q,1)),1,m);
     A_res_rep=kron(kron(ones(N,1),A_new_res),ones(1,p));
     H=A_res_rep.*X_res_rep;

     ind_z_B_res=find(B_res==0);              %zero elements in B
     ind_nz_B_res=setdiff(1:p*m,ind_z_B_res);%nonzero elements in B

     H_tilde=H;
     H_tilde(:,ind_z_B_res)=[];  %H_tilde is H that remove zero col

```

```
%calculate 'b' by least squares
     b=H_tilde\y_vec;
     B_new_res=zeros(p,m);
     B_new_res(ind_nz_B_res)=b;

% check uniqueness of B by checking the condition that
 % H_tilde is skinny and full rank
     [rowH_tilde  colH_tilde]=size(H_tilde);
     rank_H_tilde=nnz(svd(H_tilde)~=0);
     if rank_H_tilde == min(size(H_tilde))&& rowH_tilde > colH_tilde %H_tilde is full rank and skinny
        history_reducedMIMIC.uniqueB(i)=1; %B is unique
     else
        history_reducedMIMIC.uniqueB(i)=0; %B is not unique
     end

% For convergence condition of objective function value
     history_reducedMIMIC.obj_value(i) = (norm(Y-A_new_res*B_new_res'*X_res,'fro'))^2;
     conv_obj_value(i) = abs(history_reducedMIMIC.obj_value(i) -
     history_reducedMIMIC.obj_value(i-1))/abs(history_reducedMIMIC.obj_value(i-1));

     if (Ainit == zeros)
        history_reducedMIMIC.conv_A(i)=0;
        history_reducedMIMIC.conv_B(i)=0;
     else
        history_reducedMIMIC.conv_A(i) = norm(A_new_res-A_res,'fro')/norm(A_res,'fro');
        history_reducedMIMIC.conv_B(i) = norm(B_new_res-B_res,'fro')/norm(B_res,'fro');
     end
        A_res=A_new_res;
        B_res=B_new_res;

     if (conv_obj_value(i) <= epsilon_obj) && (history_reducedMIMIC.conv_A(i) <=epsilon_A)
        && (history_reducedMIMIC.conv_B(i) <=epsilon_B)
         Sigma=Y-A_res*B_res'*X_res;
         model.Sigma=Sigma*Sigma'/N; %covariance matrix of prediction error
         model.L=(-N/2)*(q+log(det(model.Sigma))+log(2*pi)); % likelihood value
         model.A=A_res;
         model.B=B_res;
         break;
     end
  end
  i=i+1;
end
```

7.4.3 MATLAB codes of model selection

In this part, example of MATLAB codes for model selection is provided used the section 4.3 in order to select the model that minimizes information criterion score as follows:

```
% examples of MATLAB codes for model selection

% it requires input as follows.
% 1) X is p*N matrix where N is number of observations
% and q is number of variables in x
% 2) Y is q*N matrix where p is number of variables in y
% 3) gamma \in [0,gamma_max] is regularization parameter containg n value of gamma
% 4) IND is linear indices of nonzero elements in B
% 5) rho is the augmented Lagrangian parameter used in group lasso containg n value of rho
% 6) r is a number of latent variables

% and it returns output
%  A_opt and B_opt that are optimal solutions.

[p N]=size(X);
q=length(Y(:,1));

%gamma
gamma_max =5000;              %penalty parameter that makes all columns of A be zero
gamma=gamma_max*[0 0.5*logspace(-3.6,0,94) linspace(0.7,1.7,5)]; %varies 100 gammas
n=length(gamma);

%IND
%structure of B
%nonzero elements in B where the element B_ij is shown by B_i ( index of
%i-th row) and B_j (index of j-th column)
%eg. nonzero elements are B_11, B_13, B_56, B_i=[1;1;5] and B_j =[1;3;6]
B_i =[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18];
B_j =[1,1,1,1,1,1,2,3,3,3,4,5,5,6,6,6,6,7];
%linear indices of nonzero elements in B
IND=sub2ind([length(B_i),length(B_j)],B_i,B_j);

%rho
rho=logspace(-3,log(gamma_max/70)/log(10),100);

% apply SVD method for initialization in the case of rank(F)=r
F=X'\Y'; F=F';
[U D V]=svd(F);

Ainit=U;
Binit=V*D';

r=nnz(svd(F)~=0);      %calculate rank(F) by SVD to avoid numerical error

% latent variable selection
for j=1:n

    [model_latent.A ,model_latent.B,history_latent] = latent_reduction(X,Y,gamma(j),Ainit,Binit,IND,rho(j));

    A_save_opt(:,:,j)=model_latent.A;  %optimal A for gamma j
        B_save_opt(:,:,j)=model_latent.B;  %optimal B for gamma j

    %apply warm start
    Ainit=A;
    Binit=B;

end

% least-squares estimation for reduced MIMIC model
for j=1:n
  %remove zero col in A and B and zero row related to zero col in B and X
  %then use such A and B to be initial point
  [Ainit,Binit,X_res] = shape(A_save_opt(:,:,j),B_save_opt(:,:,j),X,index_B);
  X_save_opt_res{j,1}=X_res;
  if A_save_opt(:,:,j)==zeros(q,m);
      %if A from latent selection is zero, A,B from this formulation is zero
         A_save_opt_res{j,1}=0;
         B_save_opt_res{j,1}=0;
         X_save_opt_res{j,1}=0;

  else

     [model.A,model.B,model.Sigma,model.L,model.d, history_reducedMIMIC] = reduced_mimic(Y,X_res,Ainit,Binit, index_B);
     A_save_opt_res{j,1}=model.A;
         B_save_opt_res{j,1}=model.B;

     % model selection
     % We solve the problem of least-squares for reduced MIMIC model  by varying 100 gammas \in [0,gamma_max]
     % then perform model selection by selecting A_opt and B_opt that minimize BIC scores
     % BIC=-2*L + d log(N)
     BIC(j,1)= -2*model.L+model.d*log(N);
  end
end

ind=find(BIC==min(BIC)); %index of gamma that provide minimun BIC score
A_opt=A_save_opt_res{ind,1};
B_opt=B_save_opt_res{ind,1};
```

Biography

Chollada Laohaphansakda was born in Chiangmai, Thailand, in 1992. She was a Mathematically Gifted high school students at the Prince Royal's College, Chiangmai and graduated in 2010. She received a full-ride scholarship at Mahidol University and finished her Bachelor of Science in Mathematics with the first class honor in 2015. She completed her Master's degree in Financial Engineering Program, Faculty of Commerce and Accountancy at Chulalongkorn University, Thailand in 2017. Her study in Master's degree was supported by a full-ride scholarship from Bangkok Bank Public Company Limited. She enjoys music, cooking, eating, traveling, and exercising. She is hard working and rises up to challenges.

List of Publications

1. C. Laohaphansakda, J. Songsiri, and T. Watewai, “A Reduction of ineffective determinants of capital structure," *Proc. of Universal Academic Cluster International Autumn Conference in Tokyo*, Japan, 2017.