SPARSITY IN VARYING-COEFFICIENT REGRESSION AND COVARIANCE MATRIX ESTIMATION

A Dissertation

by

RAKHEON KIM

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Tanya P. Garcia
Co-Chair of Committee,	Mohsen Pourahmadi
Committee Members,	Raymond J. Carroll
	Catherine Yan
Head of Department,	Brani Vidakovic

May 2022

Major Subject: Statistics

ABSTRACT

This dissertation discusses how we can exploit sparsity, a statistical assumption that only a small number of relationships between variables are non-zero, in the model selection for regression and covariance matrix estimation.

In a linear model, the effects from the predictors to the response may vary for each individual. In this case, the purpose of model selection is not only to identify significant predictors but also to understand how their effects on the response differ by individuals. This can be cast as a model selection problem for a varying-coefficient regression. However, this is challenging when there is a pre-specified group structure among variables. We propose a novel variable selection method for a varying-coefficient regression with such structured variables. Our method is empirically shown to select relevant variables consistently. Also, our method screens irrelevant variables better than existing methods. Hence, our method leads to a model with higher sensitivity, lower false discovery rate and higher prediction accuracy than the existing methods. We apply this method to the Huntington disease study and find that the effects from the brain regions to motor impairment differ by disease severity of the patients, indicating the need for customized intervention.

In covariance matrix estimation, current approaches to introduce sparsity do not guarantee positive definiteness or asymptotic efficiency. For multivariate normal distributions, we construct a positive definite and asymptotically efficient estimator when the location of the zero entries is known. If the location of the zero entries is unknown, we further construct a positive definite thresholding estimator by combining iterative conditional fitting with thresholding. We prove our thresholding estimator is asymptotically efficient with probability tending to one. In simulation studies, we show our estimator more closely matches the true covariance and more correctly identifies the non-zero entries than competing estimators. We apply our estimator to Huntington disease and detect non-zero correlations among brain regional volumes. Such correlations are timely for ongoing treatment studies to inform how different brain regions are likely to be affected by these treatments.

DEDICATION

To my wife, Jun Kyung, my daughter, Diana (Dahyun), and my family.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my Ph.D advisors Dr. Tanya Garcia, Dr. Mohsen Pourahmadi and Dr. Samuel Müller. Tanya has provided me with opportunities not only to concentrate on research problems but also to collaborate with other mentors, grow with other Ph.D students, and even mentor some undergraduate/master students since May 10th 2018, the first day I met her. Tanya's opportunities and support have prepared and empowered me to confidently flourish as an independent researcher. I would like to thank Mohsen for being an excellent advisor and for exposing me to the field of covariance estimation. Ever since I received Mohsen's autograph on his book, it has been my go-to textbook for the duration of my Ph.D. program. Moreover, I have been extremely fortunate to have another advisor, Samuel, who has always encouraged, inspired, and assisted me with research and career.

I would like to extend my gratitude to Dr. Raymond Carroll and Dr. Catherine Yan, members of my dissertation committee, for their support and advise on my study. Thanks also to Dr. Irina Gaynanova, Dr. Alan Dabney and Dr. Giovanni Motta for giving me great advice on my academic career. I owe Dr. Irina Gaynanova a debt of gratitude for teaching sparsity and statistical computing classes in which I gained valuable skills for my own research. Dr. Alan Dabney has helped me become a better teacher by coaching me through best practices in pedagogy. I learned a lot while working as a teaching assistant for Dr. Giovanni Motta. Also, as the Graduate Academic Advisor, Andrea Dawson was a great help during my Ph.D. program.

Thanks to my colleagues at the Texas A&M University. Junho Yang and Se Yoon Lee have always been helpful since my first day at Texas A&M for sharing their experience in research, teaching and job search. Kristyn Pantoja was quite generous in sharing her teaching expertise. Also, I learned a lot from a research project with Aramayis Dallakyan who always inspires me. I also enjoyed our seminar series on covariance estimation with Anupam Kundu and some collaborative work with James Dole, Dongbang Yuan and others. I would like to express my gratitude to the Statistics Graduate Student Association for hosting Statcafe and other activities.

This dissertation is dedicated to Jun Kyung Kim, my lovely wife. She was the one who motivated and encouraged me to begin the Ph.D program, which turned out to be one of the best decisions I've ever made, and her unwavering support has continued throughout the program. It is unquestionable that she is the only and the right one for me. Moreover, despite the obstacles posed by the pandemic, my time at Texas A&M University has been a joyful moment of growth and blossom, not only for me as a scholar, but also for my adorable three-year-old daughter, Diana. Finally, my Ph.D. journey would have never been feasible without the love of God and my family in South Korea. Thank you.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Tanya P. Garcia (advisor) of the Department of Biostatistics at the University of North Carolina (Chapel Hill), Dr. Mohsen Pourahmadi (co-advisor) and Dr. Raymond J. Carroll of the Department of Statistics and Dr. Catherine Yan of the Department of Mathematics. The work on varying-coefficient regression was also supported by Dr. Samuel Müller of the Department of Mathematics and Statistics at Macquarie University (Australia).

This study used data from the PREDICT HD Study which received support from the National Institute of Neurological Disorders and Stroke and collected by the PREDICT-HD investigators. We thank the PREDICT-HD investigators and respective coordinators who collected data and/or samples, as well as participants and their families who made this work possible.

All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a teaching assistantship from Texas A&M University and a research assistantship from the National Institute of Neurological Disorders and Stroke (NINDS; K01NS099343) and Australian Research Council (DP210100521).

NOMENCLATURE

MLE	Maximum Likelihood Estimator
OLS	Ordinary Least Squares
GLS	Generalized Least Squares
FGLS	Feasible Generalized Least Squares
EGLS	Estimated Generalized Least Squares
LRT	Likelihood Ratio Test
LASSO	Least Absolute Shrinkage and Selection Operator
SCAD	Smoothly Clipped Absolute Deviation
MCP	Minimax Concave Penalty
ADMM	Alternating Directions Method of Multipliers
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CV	Cross-validation
MSE	Mean Squared Error
ROC	Receiver Operating Characteristic
FDR	False Discovery Rate
HD	Huntington Disease
TMS	Total Motor Score
CAP	CAG-Age-Product

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 A Review of the Sparsity in Regression and Covariance Estimation

With ever increasing data for numerous variables being collected, a key interest is to distinguish scientifically meaningful relationships between variables from noise in the data. For example, in neuroscience, the data for numerous brain regions are analyzed to study the relationships between each region and a disease of the brain but only a handful of regions may be related to the disease. However, such scientific relationships are not clearly distinguished from the noisy relationships between other brain regions and the disease. One statistical assumption to tackle this is *sparsity*, meaning that only a small number of relationships between variables are non-zero and scientifically meaningful. In this section, we review how the sparsity assumption has been used in diverse statistical modeling such as varying-coefficient regression and covariance estimation problems.

1.1.1 Sparsity in Varying-coefficient Regression

In the analysis of linear relationships between a response variable and multiple predictors, we often do not know which predictors are relevant to be included in the linear model to explain the response variable. Since Least Absolute Shrinkage and Selection Operator (LASSO) has been proposed by Tibshirani (1996), several penalization methods such as elastic net (Zou and Hastie, 2005), smoothly clipped absolute deviation (SCAD, Fan and Li (2001)) and minimax concave penalty (MCP, Zhang (2010)) have been proposed as methods for variable selection of the linear modelswhen there is no relationship between the predictors. When there is some structure among the predictors (e.g. a categorical variable expressed with multiple dummy variables), the relationship among the variables necessitates group-wise variable selection. The group LASSO (Yuan and Lin, 2006) and the sparse group LASSO (Simon et al., 2013) addressed pre-defined group structure among predictors. Some literature on structured variable selection (Garcia and Müller, 2014; Garcia et al., 2013; Yuan et al., 2009) considered the structure between main effect terms and other variables such as interaction terms.

However, the effects from the predictors to the response may not be fixed across all individuals but may differ by individuals, for example, differ by gender or age groups. To model such differentiated relationships between a response variable and multiple predictors, we can use the varying-coefficient model from Hastie and Tibshirani (1993). In the varying-coefficient model, the regression coefficient of each predictor, called main predictor, is not fixed, but instead depends on the values of other variables, called modifying variables.

Variable selection of the varying-coefficient models has been focused on the selection of either the main predictors or the modifying variables, but rarely both. Selection of the main predictors has been explored through regularized estimation of the functional coefficients as smooth functions of continuous modifying variables (Wang et al., 2008; Wei et al., 2011) or as constant functions of categorical modifying variables (Gertheiss and Tutz, 2012; Oelker et al., 2014). Selection of the modifying variables has been explored using a tree-based approach such as classification and regression trees (CART) (Wang and Hastie, 2014; Bürgin and Ritschard, 2015; Berger et al., 2017) which captures potentially complex interactions among modifying variables and automatically selects important variables.

Simultaneous selection of main predictors and modifying variables is related to the variable selection for interaction models. These methods use hierarchical constraints to ensure that interaction terms between two predictors are included either when both predictors are in the model (Lim and Hastie, 2015; Haris et al., 2016) or when only one of them is (Bien et al., 2013; She et al., 2018). Tibshirani and Friedman (2019) handles the asymmetric relationship between the main predictors and the modifying variables by proposing pliable LASSO. The pliable LASSO uses hierarchical regularization to identify the significant main predictors first and then select the modifying variables for the significant main predictors.

1.1.2 Sparsity in Covariance Matrix Estimation

Understanding the linear association between variables is important in many applications including genetics (Butte et al., 2000; Rothman et al., 2009), finance (El Karoui et al., 2010; Xue et al., 2012) and climotology (Bickel et al., 2008a). The linear association among variables is

encoded in the covariance matrix and the sample covariance matrix is a popular estimator when the sample size is greater than the number of variables. However, if some covariances are zero, detection of zero covariances from the sample covariance matrix is hard because the sample covariance between any two variables will not be zero due to the noise in the data. Identification of zero entries in the covariance matrix has been studied either by multiple testing (Drton et al., 2007) or by inducing sparsity in the covariance matrix estimator when the variables can be ordered (Wu and Pourahmadi, 2003; Huang et al., 2006; Bickel et al., 2008b) and when there is no ordering to the variables (Bickel et al., 2008a; Rothman et al., 2009; Bien and Tibshirani, 2011).

When the variables can be ordered, such as in time series data, the correlation for the off-diagonal entries far apart from the diagonal are often assumed to be smaller than those closer to the diagonal. This structural information of the covariance matrix can be used to obtain sparse covariance matrix estimators. Typical examples of such sparse covariance matrices include banding estimators (Wu and Pourahmadi, 2003; Bickel et al., 2008b; Bien et al., 2016) and tapering estimators (Furrer and Bengtsson, 2007; Cai et al., 2010).

When there is no ordering to the variables, methods for estimating sparse covariance matrices can be divided into two categories. One class of the methods is based on penalized likelihood. Lam and Fan (2009) studied the theoretical properties including the asymptotic normality of these estimators. Bien and Tibshirani (2011) proposed an algorithm to solve such penalized likelihood when the sample size is greater than the number of variables. The other class of the estimation methods involves thresholding the sample covariance matrix. Bickel et al. (2008a) showed the consistency of the universal thresholding in the operator norm and Cai and Liu (2011) proposed the adaptive thresholding to account for the variability of the individual covariances. Rothman et al. (2009) proposed the generalized thresholding that combines the thresholding with shrinkage and showed operator norm consistency. Rothman et al. (2009) remarked the possibility to develop asymptotic normality for non-zero entries in the thresholding estimators but did not discuss further.

1.2 Research Challenges and Main Contributions

Despite substantial efforts to exploit sparsity, lingering challenges remain including sparsity in the varying-coefficient regression and covariance matrix estimation. Driven by current interdisciplinary problems in neuroscience, I have developed methods that address those challenges.

1.2.1 High-dimensional Varying-coefficient Models

In a linear model, the effects of some predictors to the response may not be the same across all individuals but, for example, differ between men and women or differ by age. Such differentiated effects of predictors to the response can be modeled by a varying-coefficient model, a regression model where each regression coefficient is not fixed but is instead a function of other variables such as gender or age. When we assume sparsity of the varying-coefficient model, we have to select not only the predictors but also the variables within the coefficient of each predictor. This is challenging because the number of parameters in this model can easily exceed the number of samples, leading to a high-dimensional problem. Also, if there is a structure where several variables can be grouped (e.g. dummy variables for a categorical variable), ignorance of the structure may lead to inconsistent model selection by randomly selecting variables from the group of variables.

In this dissertation, we discuss model selection of a varying-coefficient model and proposes a novel penalized regression method called *svReg*.This method uses a hierarchical group penalty which plays two important roles. First, it imposes a hierarchical constraint that allows simultaneous selection of the predictors and the variables within the coefficient of each predictor. Second, it penalizes a group of variables together if those variables can be grouped. This penalty allows us to perform model selection of a varying-coefficient model while considering the grouping structure among variables. Hence our svReg method can be used for customizing the linear model by subgroups or even by individuals. Also, we discover that weighting penalty terms differently according to the size of each group of variables leads to select variables relevant to the response variable more correctly while screening irrelevant variables more effectively. In simulation studies, the weighted hierarchical group penalty in our svReg lower the false discovery rate by 6% and

increase the true positive rate by 2% compared to existing methods. When applied to Huntington disease studies in neuroscience, our method detects differentiated effects from brain regions on disease progression by patients.

1.2.2 Estimation of Covariance Matrices

Sparsity of a covariance matrix is useful to identify linear relationships between variables: zero elements in the matrix mean no linear relationships and non-zero elements indicate the relationship strength. However, constructing a positive definite estimator whose non-zero elements are asymptotically efficient remains a challenge, thus invalidating many statistical analyses such as discriminant analysis or reducing confidence in our estimation. When the locations of the zero elements are known, the Gaussian maximum likelihood estimator is asymptotically efficient and Chaudhuri et al. (2007) devised the iterative conditional fitting algorithm which converges to a positive definite solution to the likelihood equation. However, whether the solution is asymptotically efficient is unclear because the likelihood may have multiple local maxima and asymptotic efficiency of these local solutions has not yet been proven. Moreover, the algorithm is available only in low dimension when the sample size is greater than the number of variables. Lastly, the algorithm does not tell us the locations of the zero elements when they are unknown.

In this dissertation, we first discuss asymptotic efficiency of a sparse covariance matrix estimator when the locations of the zero elements are known. Specifically, we prove that the iterative conditional fitting algorithm produces a positive definite and asymptotically efficient estimator when the algorithm starts from a consistent estimator. We also propose modification to the iterative conditional fitting algorithm for the case when the sample size is smaller than the number of variables. The basic idea of this modification is to combine the algorithm with a shrinkage estimator whose diagonal entries are greater than that of the sample covariance matrix. Since this modification uses multiple uses of the ridge regression, we call this modified algorithm as the iterative conditional ridge algorithm.

Building on this result, we extend to more common situations where the locations of the zero elements are unknown. We propose a new thresholding estimator, COMET (COvariance Maximum-

likelihood Estimation with Thresholding).This involves iterative conditional fitting of non-zero elements determined by thresholding the sample covariance matrix, that is, by forcing sample covariances below a threshold to zero. We prove the COMET is always positive definite and asymptotically efficient with probability tending to one. Also, unlike other thresholding estimators, we can appeal to the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the selection of the threshold, eliminating the needs for computationally expensive cross-validation. In simulation studies, our estimator more closely matches the true covariance and more correctly identifies the non-zero elements than competing estimators. Application to a Huntington disease study detects non-zero correlations among brain regional volumes, which inform which brain regions are likely to be affected by a treatment which targets a specific brain region.

1.3 Outline

The rest of the dissertation is organized as follows. In Chapter 2, we discuss the use of sparsity in varying-coefficient regression problems. In Section 2.1, we pose a real-world problem in Huntington disease study where current statistical methods for varying-coefficient regression have some limitations. In Section 2.2, we propose our main method, structural varying-coefficient regression, to address those limitations of the existing methods. In Section 2.3, we conduct extensive simulation studies to compare our method with other methods. In Section 2.4, we apply our method to PREDICT-HD data from Huntington disease study and discuss its potential for developing customized intervention for patients.

In Chapter 3, we discuss the estimation of a covariance matrix, focusing on the cases where some of the entries in the matrix are exactly zero. In Section 3.1, we review estimation of a covariance matrix with zero entries and discuss current challenges. In Section 3.2 we cast this problem as estimation of the linear covariance model and discuss its interpretation under the linear regression framework. Next, we propose a positive definite and asymptotically efficient estimator of a covariance matrix with zero entries in Section 3.3 and check its efficiency through multiple simulation studies in Section 3.4. However, such an asymptotically efficient estimator is available only when the location of zero entries in a covariance matrix is known. In Section 3.5, we discuss

the implication of unknown zero entries in the matrix.

Chapter 4 extends our discussion further by assuming that the location of the zero entries is unknown. Section 4.1 reviews statistical methods such as thresholding to obtain a sparse covariance matrix estimator and Section 4.2 gives greater details on those methods. Then, in Section 4.3, we propose a new thresholding estimator which is always positive definite and asymptotically efficient with probability tending to one. We compare the performance of this estimator with other thresholding estimators in extensive simulation studies in Section 4.4. In Section 4.5, we apply these methods to PREDICT-HD data from Huntington disease study to identify the relationships among different brain regions.

Chapter 5 concludes this dissertation by summarizing and suggesting a handful of topics for future research. Optimization details of the structural varying-coefficient regression in Chapter 2 can be found in Appendix A. Mathematical proofs of theorems and propositions in Chapter 3 and Chapter 4 are in Appendix B. Additional simulation and data analysis results for Chapter 4 are in Appendix C.

2. HIGH-DIMENSIONAL VARYING-COEFFICIENT MODELS*

2.1 Introduction

For Huntington disease, a genetically inherited neurodegenerative disorder, developing interventions to alleviate the symptoms of the disease is the goal of many clinical trials. One of the main symptoms of the disease is motor impairment (Biglan et al., 2009; Paulsen et al., 2014b; Reilmann et al., 2014) and the motor symptom is known to be related to regional brain atrophy, that is, the loss of cells in some brain regions (Aylward et al., 2013). Hence, one interest in clinical trials is to identify which brain regions are associated with motor impairment and stop or slow atrophy of those regions to prevent motor impairment. For example, the clinical trial SIGNAL determines the effect of an antibody on the regional brain volumes and assesses the motor functions of the participants (Rodrigues and Wild, 2018) by total motor scores (TMS), a score from 0 to 124 with higher indicating more severe impairment (Kieburtz et al., 2001).

Although the relationship between the total motor score and the volume of brain regions is well understood (Aylward et al., 2013), we observed that how the change of brain volumes affects the total motor score may not be the same across all patients but vary for different groups of patients. For example, it is a standard practice in the field to categorize patients into three different groups (high/medium/low) by disease severity, a variable that indicates the risk of being diagnosed with Huntington disease in the next 5 years as in Zhang et al. (2011). In the top panels of Figure 2.1, the effect from the reduction of caudate nucleus to the total motor score is larger for the high disease severity group than for other groups as observed by the steeper regression line. This indicates that patients in the high disease severity group may need different interventions than patients in other groups since their motor function may deteriorate faster than others given a certain amount of change in caudate nucleus volume. Hence, in addition to the identification of brain regions related

*Parts of this section have been modified with permission from [R. Kim, S. Müller and T. Garcia. svReg: Structural Varying-coefficient regression to differentiate how regional brain atrophy affects motor impairment for Huntington disease severity groups. Biometrical Journal. 2021. Volume 63. Pages 1254-1271. (https://doi.org/10.1002/bimj.202000312) Copyright Wiley-VCH GmbH. Reproduced with permission]

to motor impairment, understanding how their effects on motor impairment differ by patient groups will enable us to develop interventions customized for each patient group.

Statistically, identifying brain regions and understanding how their effects on motor impairment differ by patient groups can be cast as a model selection problem of a varying-coefficient model (Hastie and Tibshirani, 1993). A varying-coefficient model is a regression model whose regression coefficients can vary by each individual or group of individuals. To be specific, consider a regression model with the total motor score as a response and the volume of brain regions as main predictors. In a varying-coefficient model, the regression coefficient of each brain region is not fixed but a function of other variables, called modifying variables. For example, if the disease severity is a modifying variable for a brain region, the regression coefficient of that region will take different value for each disease severity group so that we will end up with three different regression models, one for each group. Likewise, other demographic variables such as gender and years of education can also be considered as modifying variables which will divide the patients into smaller subgroups.

A varying-coefficient model is a special form of an interaction model where the interaction terms between main predictors and modifying variables are considered. In Figure 2.1, the interaction effect between the volume of a brain region and the disease severity can be observed through difference in the slope of the regression line for each disease severity group. To the best of our knowledge, in the literature of Huntington disease, the disease severity and other demographic variables such as gender and years of education have been treated as covariates or control variables (Aylward et al., 2013; Biglan et al., 2009; Misiura et al., 2017). However, their interaction effects with brain regions have not been investigated yet. In Tabrizi et al. (2012) and Paulsen et al. (2014a), a different rate of change in brain regional volumes over time was observed for each disease severity group but the effect of the interaction on the total motor score was not considered.

Model selection of a varying-coefficient model includes two tasks: selection of main predictors and selection of modifying variables. In the Huntington disease study, identifying brain regions related to motor impairment corresponds to the selection of main predictors. Understanding how

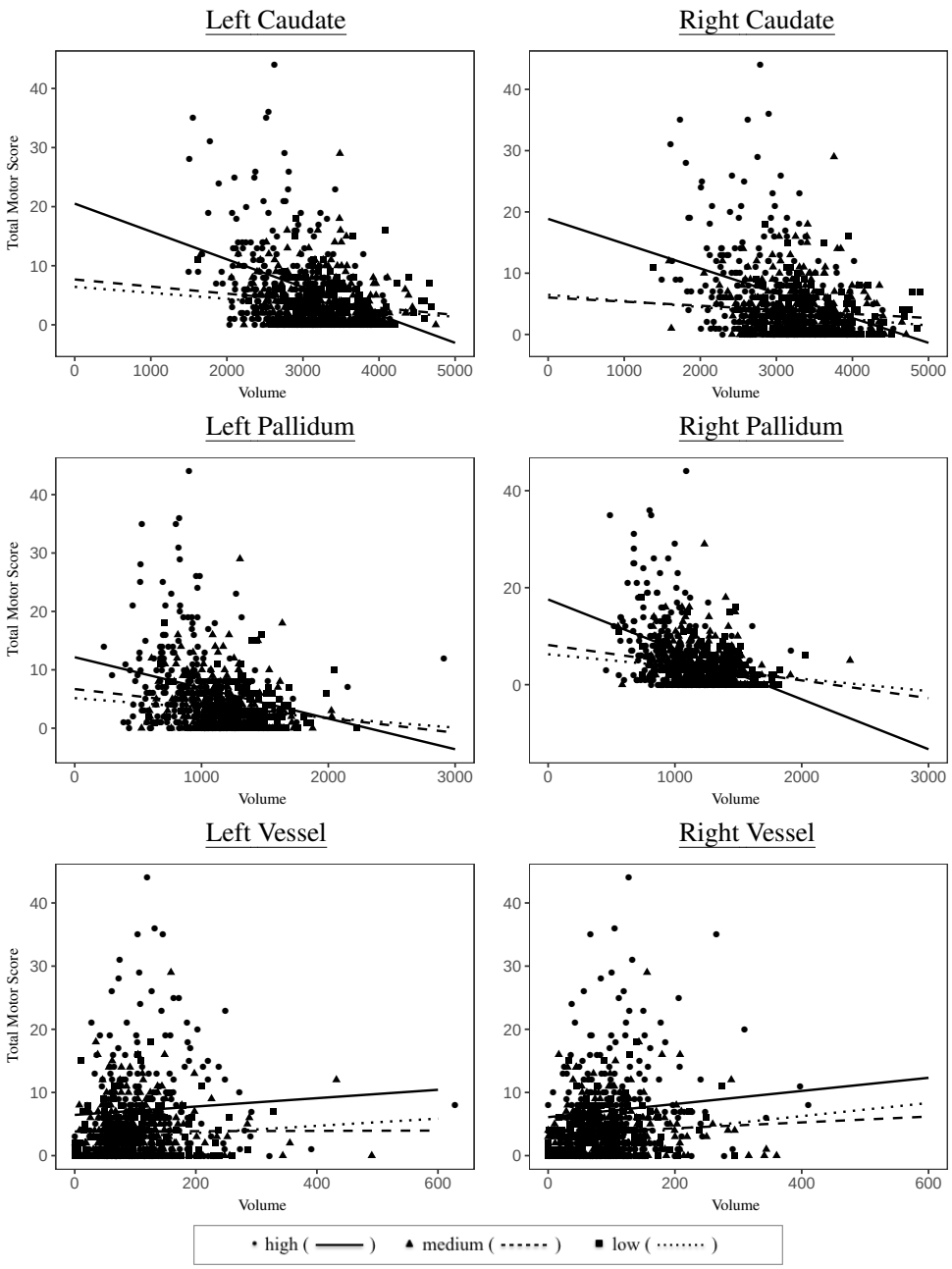

Figure 2.1: Scatter plots between total motor score and volume of brain regions. Least squares fits by the group of scaled CAG-Age-Product (CAP) score, a measure of disease severity, are overlaid. Solid line is the least squares fit of the 'high' disease severity group (circles), dashed line is the least squares fit of the 'medium' disease severity group (triangles) and dotted line is the least squares fit of the 'low' disease severity group (squares). Interaction effects between the volume of some brain regions (left caudate, right caudate, right pallidum) and CAP score are observed through different slopes of the least squares fit for each disease severity group. The difference in slopes is relatively small for the left pallidum and ignorable for the left and the right vessel. The correlation coefficient is 0.94 between the left caudate and the right caudate, 0.77 between the left pallidum and the right pallidum, and 0.48 between the left vessel and the right vessel.

the effects of those brain regions differ by patient groups corresponds to the selection of modifying variables where the possible candidates of modifying variables include disease severity, gender and years of education.

However, the literature on the varying-coefficient model has focused on variable selection of either the main predictors or the modifying variables, but rarely both. Among others, selection of main predictors has been explored when the modifying variable is a continuous variable (Wang et al., 2008; Wei et al., 2011) or a categorical variable (Gertheiss and Tutz, 2012; Oelker et al., 2014). In their work, only one modifying variable is considered so the interest is the selection of main predictors and whether each regression coefficient is fixed or not. Selection among multiple modifying variables has recently been explored through tree-based approaches (Berger et al., 2017; Bürgin and Ritschard, 2015; Wang and Hastie, 2014), which estimate a tree of modifying variables for each main predictor. Alternatively, the varying index coefficient model (Ma and Song, 2015; Na et al., 2019) achieves dimension reduction for multivariate modifying variables by using different loading weight for each modifying variable. However, these approaches focus on the selection of modifying variables and do not consider the selection of main predictors. The selection of main predictors may be considered as an additional procedure to those methods but simultaneous selection of the main predictors and the modifying variables has not been sought. Tibshirani and Friedman (2019) handles the variable selection of a varying-coefficient model by the pliable LASSO (pLASSO), a generalization of the LASSO (least absolute shrinkage and selection operator) that selects both the main predictors and modifying variables, simultaneously.

Additional consideration for Huntington disease application is that there are pre-specified group structures among main predictors and modifying variables. First, measurements of some brain regions can be grouped according to the structural information on a brain and they are often highly correlated. For example, the volume of the left caudate and the right caudate can be considered as a group. Due to their high correlation coefficient $(= 0.94)$, as shown in the top panels of Figure 2.1, the left caudate and the right caudate have similar negative relationship with the total motor score. Second, the disease severity is a categorical variable with three categories (low, medium

and high), expressed in the design matrix for linear regression as a group of two binary dummy variables. Since each of these binary variables contains only partial information for one categorical variable, those two binary variables should be grouped.

The pliable LASSO (Tibshirani and Friedman, 2019) is designed to work well when there is no pre-specified structure among the variables. However, if there is a group structure among the main predictors with high within-group correlation, we claim that the pliable LASSO may lead to inconsistent model selection by randomly selecting variables from those highly correlated variables as the usual LASSO suffers (Zhao and Yu, 2006). This problem of the pliable LASSO will be discussed with a simulation study in Section 2.3. Furthermore, modifying variables may also have a pre-specified group structure as appeared in our Huntington disease problem. Since ignoring such group structure may lead to selecting more variables than necessary (Yuan and Lin, 2006), it is desirable to account for such group structure in model selection.

In this chapter, we propose the novel *structural varying-coefficient regression* (svReg) for a varying-coefficient model with structured variables. This method imposes hierarchical group penalties on each group of main predictors and modifying variables to account for group structures among variables. Such hierarchical group penalties have been studied in other regression settings. To name a few, the group LASSO (Yuan and Lin, 2006) and the sparse group LASSO (Simon et al., 2013) address pre-defined group structure among regressors and the network LASSO (Hallac et al., 2015) extends the group LASSO to a network setting. Zhao et al. (2009) discussed the hierarchical selection of grouped predictors for non-overlapping groups. Some literature on structured variable selection (Garcia and Müller, 2014; Garcia et al., 2013; Yuan et al., 2009) considers the structure between main effect terms and other variables such as interaction terms. Literature on interaction models (Lim and Hastie, 2015; Bien et al., 2013) also considered the hierarchy between main effect and interaction. However, simultaneous selection of main predictors and modifying variables for a varying-coefficient model with group-structured variables has not been explored yet.

The svReg addresses model selection of the varying-coefficient model as the pliable LASSO but differs significantly from that. First, a pre-specified group structure and the within-group cor-

relation of the variables are considered in the svReg, whereas the pliable LASSO ignores such group structure. This feature enables the svReg to be more flexible to the problems where the group structure among the variables exists. Second, as discussed in Garcia et al. (2013, 2016) regarding the use of weighted penalties, we discovered that weighting penalty terms accounting for the different size of each group of variables led to more relevant variables and fewer irrelevant variables being selected into the model. Hence, in addition to being more flexible by accounting for the group structure, penalty terms are differently weighted in the svReg. Third, when some modifying variables are selected in the model, the svReg algorithm identifies the groups of possibly significant modifying variables first and then selects variables from those identified groups to reduce false selection while the pliable LASSO selects variables from the set of all modifying variables. These important differences from the pliable LASSO allow the svReg to select relevant variables consistently and better screen irrelevant variables with higher prediction accuracy. We demonstrate the efficacy of our proposed method on real and simulated data, and provide a publicly available R package `svreg (https://github.com/Tanya-Garcia-Lab/svreg)` for implementation of our method.

2.2 The structural varying-coefficient regression model

2.2.1 Main Model

We consider a varying-coefficient linear model with a response variable, y, and p *main predictors*, $\{x_j\}_{j=1}^{p}$, and K *modifying variables*, $\{z_k\}_{k=1}^{K}$, as below:

$$y = \sum_{j=0}^{p} \left\{ \beta_j + \sum_{k=1}^{K} \theta_{jk} z_k \right\} x_j + \epsilon, \tag{2.1}$$

where $x_0 = 1$, representing a potential intercept term and ϵ is the error term. In this model, $\{z_k\}_{k=1}^{K}$ modify how the j-th predictor x_j affects the response y. When $\theta_{jk} = 0$ for all $j = 0, 1, \ldots, p$ and $k = 1, \ldots, K$, this reduces to a plain linear model with fixed coefficients. The inclusion of $\theta_{jk} z_k$ terms within the coefficient of x_j allows the coefficient to vary depending on the modifying variables $z_1, \ldots, z_K$. For independent subjects $i = 1, \ldots, N$, we denote the response variable, y_i,

and p main predictors, $\{x_{ij}\}_{j=1}^{p}$, and K modifying variables, $\{z_{ik}\}_{k=1}^{K}$.

In the Huntington disease study, our objective is to use the varying-coefficient model to identify main predictors associated with total motor score (y_i) and understand how their effects on total motor score differ by patient groups where patients are grouped by modifying variables. Main predictors in our model will be selected from volume measures of 50 brain regions ($\{x_{ij}\}_{j=1}^{p}$, $p = 50$). However, some of these regions are not independent because they are parts of a larger region. For example, caudate nucleus contains two parts, the left caudate and the right caudate, and correlation coefficient between their volumes is 0.94. That is, these measurements can be considered as a group of size two according to the structure of the brain. Likewise, lots of high correlations among the brain regions can be explained by the structural information of the brain. Hence we consider the structure among the brain regions so that the 50 main predictors are grouped into 34 groups. Potential modifying variables will include gender (z_{i1}), years of education (z_{i2}) and disease severity (z_{i3}, z_{i4}), hence $K = 4$. Here, disease severity is a categorical variable with 3 category levels (low, medium and high) depending on the likeliness of receiving a motor-diagnosis in the next five years. Hence, disease severity is expressed with two binary dummy variables, z_{i3} and z_{i4}, and these two variables should be treated as grouped variables. Our proposed method will properly consider the group structure of the main predictors and the modifying variables by imposing group-wise penalty.

2.2.2 Methodology

We propose a novel modification to the pliable LASSO (Tibshirani and Friedman, 2019) to account for potential structure among the variables (e.g., grouping between variables). The pliable LASSO is a generalization of the LASSO for varying-coefficient models but it ignores potential structure among the variables, such as grouped main predictors (e.g., left and right caudate of the brain could be considered as one group) or grouped modifying variables (e.g., categorical disease severity group). Ignoring such group structure and within-group correlation may lead to inconsistent model selection by randomly selecting variables from those highly correlated variables (Zhao and Yu, 2006) or may lead to selecting more variables than necessary (Yuan and Lin, 2006).

We thus propose a regression method with hierarchical penalties to account for grouped main predictors and grouped modifying variables.

Let $\boldsymbol{y}$ be the N dimensional vector $(y_1, \ldots, y_N)^T$ and let $\mathbf{X}, \mathbf{Z}$ be the $N \times p$ and $N \times K$ matrices containing main predictors and modifying variables respectively. Also, let $\mathbf{x}_j$ be the j-th column of $\mathbf{X}$, $\mathbf{z}_k$ be the k-th column of $\mathbf{Z}$ and let $\mathbf{1}$ be a $N \times 1$ matrix of ones. The varying-coefficient linear model (2.1) can be written in matrix form:

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\theta}_0 + \sum_{j=1}^{p} \{(\beta_j \mathbf{1} + \mathbf{Z}\boldsymbol{\theta}_j) \circ \mathbf{x}_j\} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK})^T$. Here, $\circ$ is component-wise multiplication and captures the impact of the modifying variables by allowing coefficients to vary for each subject. In this model, the coefficient vectors $\{\boldsymbol{\theta}_j\}_{j=1}^p$ exist only within the coefficients of $\{\mathbf{x}_j\}_{j=1}^p$. Hence, for $j = 1, \ldots, p$, if $\mathbf{x}_j$ turns out to be irrelevant (i.e. $\beta_j = 0$), we want $\boldsymbol{\theta}_j$ to be estimated as a zero vector. However, β_j can take a nonzero value even if $\boldsymbol{\theta}_j$ is a zero vector, which results in a fixed coefficient for the j-th predictor. This feature of the varying-coefficient model raises the need to impose an "asymmetric weak hierarchy" constraint: $\boldsymbol{\theta}_j$ can be nonzero only if β_j is nonzero.

Suppose the p main predictors can be grouped into L groups $(L \leq p)$ and the K modifying variables can be grouped into G groups $(G \leq K)$. Each group can contain one or more variables. In our Huntington disease application, there are 50 main predictors of brain regional volumes ($p = 50$) and these predictors can be grouped into 34 groups of brain regions ($L = 34$) according to the pre-specified structure of the brain. For the modifying variables, we have three groups of modifying variables ($G = 3$): gender, years of education and disease severity. The first two groups contain one variable each. The disease severity group contains two dummy variables since disease severity is a categorical variable with three categories. Hence, there are four modifying variables ($K = 4$).

We propose to optimize the following objective function:

$$J^*(\beta_0, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{N} r_i^2 + \lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}), \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\Theta} = (\theta_{jk})_{j=1,k=1}^{p,K}$ is a $p \times K$ matrix, and

$$r_i = y_i - \beta_0 - \mathbf{z}_{i\bullet}\boldsymbol{\theta}_0 - \sum_{\ell=1}^{L} \mathbf{x}_{i[\ell]}(\boldsymbol{\beta}_{[\ell]} + \boldsymbol{\theta}_{[\ell]\bullet}\mathbf{z}_{i\bullet}^T),$$

where $\mathbf{z}_{i\bullet}$ is the i-th row of $\mathbf{Z}$, $\mathbf{x}_{i[\ell]}$ is the ℓ-th group of the main predictors for the i-th row of $\mathbf{X}$, $\boldsymbol{\beta}_{[\ell]}$ is a subset of $\boldsymbol{\beta}$ for the ℓ-th group of the main predictors, $\boldsymbol{\theta}_{[\ell]\bullet}$ is a subset of $\boldsymbol{\Theta}$ for the ℓ-th group of the main predictors and

$$\begin{aligned}\lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}) = (1-\alpha)\lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \left\{ ||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2 + \sum_{g=1}^{G} \frac{\sqrt{p_g}}{\sqrt{1+K}} ||\text{vec}(\boldsymbol{\theta}_{[\ell][g]})||_2 \right\} \\ + \alpha\lambda \sum_{j,k} |\theta_{jk}|_1,\end{aligned}$$

where p_ℓ is the size of the ℓ-th group of the main predictors, p_g is the size of the g-th group of the modifying variables, $\boldsymbol{\theta}_{[\ell][g]}$ is a subset of $\boldsymbol{\Theta}$ for the ℓ-th group of the main predictors and the g-th group of the modifying variables and vec$(\cdot)$ is a vectorization operator. Note that $\boldsymbol{\beta}_{[\ell]}$ is a p_ℓ dimensional column vector, $\boldsymbol{\theta}_{[\ell]\bullet}$ is a $p_\ell \times K$ matrix and $\boldsymbol{\theta}_{[\ell][g]}$ is a $p_\ell \times p_g$ matrix. The first term of the penalty considers the hierarchy constraint between $\boldsymbol{\beta}_{[\ell]}$ and $\boldsymbol{\theta}_{[\ell]\bullet}$ as well as the group structure among the variables through the L_2 penalty. The second term of the penalty gives sparsity to the individual coefficients θ_{jk}'s. The tuning parameter λ determines the magnitude of the overall penalties so that the larger the λ is, the sparser the model is. Another tuning parameter α controls the relative weight on the group penalty terms and penalties on individual components of $\boldsymbol{\Theta}$. For a fixed λ, both the main predictors and the modifying variables will be penalized more as $\alpha \to 0$ whereas $\alpha \to 1$ will result in penalizing more on modifying variables only.

This optimization addresses the model selection of the varying-coefficient model as the pliable

LASSO (Tibshirani and Friedman, 2019) but differs in three ways. First, if there is a pre-specified group structure among the modifying variables, such structure can be considered in the model selection through the penalty terms $||\text{vec}(\boldsymbol{\theta}_{[\ell][g]})||_2$ imposed on each group of modifying variables. The use of such penalty terms differentiates our method from the pliable LASSO because the sparsity at the group level in terms of the modifying variables cannot be attained in the pliable LASSO. For example, dummy variables for the CAP score, a categorical modifying variable, in our Huntington disease study will be treated as stand-alone variables in the pliable LASSO. However, the purpose of the model selection is not to determine the significance of each dummy variable but to identify whether the CAP score is significant or not. Hence, although we have several dummy variables for a categorical modifying variable, those variables should be treated as a group. Second, a group structure among the main predictors can also be considered in our method. For example, our method can be applied to the brain imaging data for Huntington disease study where the left part of the brain regions tend to be highly correlated with the right part. The pliable LASSO does not account for the grouping of such correlated main predictors. Lastly, the penalty terms in our method are weighted differently depending on the size of the group of main predictors and modifying variables. we discovered that such weighted penalty terms led to more relevant variables and fewer irrelevant variables being selected into the model as discussed in Garcia et al. (2013, 2016). We call this method the structural varying-coefficient regression (svReg). Typically, we suggest to standardize the main predictors and modifying variables to have mean zero and variance one as in the LASSO unless all the variables are measured in the same unit (Hastie et al., 2015).

2.2.3 Optimization

We use a blockwise coordinate descent to obtain the global minimum of equation (2.3). Denote $\mathbf{z}_{i[g]}$ as a subset of $\mathbf{z}_{i\bullet}$ for the g-th group of the modifying variables. Also, denote $r_i^{(-\ell)}$ as the partial

residual for the ℓ-th group of the main predictors,

$$r_i^{(-\ell)} = y_i - \sum_{h \neq \ell} \left\{ \mathbf{x}_{i[h]} (\boldsymbol{\beta}_{[h]} + \boldsymbol{\theta}_{[h]\bullet} \mathbf{z}_{i\bullet}^T) \right\}$$

and denote $r_i^{(-\ell)(-g)}$ as the partial residual for the g-th group of modifying variables,

$$r_i^{(-\ell)(-g)} = r_i^{(-\ell)} - \mathbf{x}_{i[\ell]} \sum_{m \neq g} \boldsymbol{\theta}_{[\ell][m]} \mathbf{z}_{i[m]}^T.$$

The procedure for estimating $\{\beta_j\}_{j=0}^p$ and $\{\theta_{jk}\}_{j=0,k=1}^{p,K}$ is given in Algorithm 1. Details of the optimization of the algorithm can be found in Appendix A.

In the step 2-(2)-(b)-(i) of the Algorithm 1, if the variables of the ℓ-th group are uncorrelated with variance one, that is $\sum_{i=1}^N \mathbf{x}_{i[\ell]}^T \mathbf{x}_{i[\ell]} / N = I$, the closed form solution of $\hat{\boldsymbol{\beta}}_{[\ell]}$ is available as below:

$$\hat{\boldsymbol{\beta}}_{[\ell]} = \max \left\{ 1 - \frac{(1-\alpha)\lambda\sqrt{p_\ell}}{||R_\ell||_2}, 0 \right\} \cdot R_\ell$$

where $R_\ell = \sum_{i=1}^N \mathbf{x}_{i[\ell]}^T r_i^{(-\ell)} / N$. Note that this takes the similar form with the solution of the group LASSO proposed by Yuan and Lin (2006). Also, when there is only one predictor, say j-th predictor, in the ℓ-th group, this solution is equivalent to the pliable LASSO (Tibshirani and Friedman, 2019) as below:

$$\hat{\beta}_j = \left(\frac{N}{\sum_{i=1}^N x_{ij}^2} \right) S_{(1-\alpha)\lambda} \left(\frac{1}{N} \sum_{i=1}^N x_{ij} r_i^{(-j)} \right).$$

However, in our Huntington disease study, the main predictors with group structure have high within-group correlation and no closed form solution for $\hat{\boldsymbol{\beta}}_{[\ell]}$ is available. For the group LASSO, Friedman et al. (2010) proposed that the solution for $\hat{\boldsymbol{\beta}}_{[\ell]}$ can be found by sequential optimization of each parameter in $\boldsymbol{\beta}_{[\ell]}$. This one-dimensional search over the parameters in $\boldsymbol{\beta}_{[\ell]}$ uses `optimize` function in the R package, which finds the minimum or maximum of a univariate function using

Algorithm 1 Algorithm for the structural varying-coefficient regression

1. Given initial estimate of $(\boldsymbol{\beta}, \boldsymbol{\Theta})$, compute $\hat{\beta}_0$ and $\hat{\boldsymbol{\theta}}_0$ from the regression of the residual on $\mathbf{Z}$. As the initial estimate, $(\boldsymbol{\beta}, \boldsymbol{\Theta}) = (\mathbf{0}, \mathbf{0})$ can be used. Then, the response variable $\boldsymbol{y}$ is regressed on $\mathbf{Z}$ to compute $\hat{\beta}_0$ and $\hat{\boldsymbol{\theta}}_0$. Otherwise, if several values of λ are tried (e.g. in the cross-validation), the svReg estimate of $(\boldsymbol{\beta}, \boldsymbol{\Theta})$ with the previous λ value can be used as the initial estimate. In this case, the residual $\boldsymbol{y} - \sum_{j=1}^{p} \left\{ (\hat{\beta}_j \mathbf{1} + \mathbf{Z}\hat{\boldsymbol{\theta}}_j) \circ \mathbf{x}_j \right\}$ is regressed on $\mathbf{Z}$ to compute $\hat{\beta}_0$ and $\hat{\boldsymbol{\theta}}_0$.

2. Given λ, α and convergence tolerance ϵ, repeat the following procedure until convergence: $|J^{*(old)}(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Theta}}) - J^{*(new)}(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Theta}})| < \epsilon$ where $J^*(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \boldsymbol{\beta}, \boldsymbol{\Theta})$ is equation (2.3).

   (1) Compute $J^{*(old)}(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Theta}})$ with the current estimate of $(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Theta}})$.

   (2) For a cycle of $\ell = 1, 2, \ldots, L$:

   (a) Check $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell]\bullet}) = 0$ by checking $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell][g]}) = 0$ for all $g = 1, 2, \ldots, G$.

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{i[\ell]}^T r_i^{(-\ell)} \right\|_2 \leq \sqrt{p_\ell}(1-\alpha)\lambda, \text{ and}$$

$$\left\| S_{\alpha\lambda} \left(\frac{1}{N} \sum_{i=1}^{N} \text{vec}(\mathbf{x}_{i[\ell]}^T \mathbf{z}_{i[g]}) r_i^{(-\ell)(-g)} \right) \right\|_2 \leq \sqrt{p_\ell}\left(1 + \frac{\sqrt{p_g}}{\sqrt{1+K}}\right)(1-\alpha)\lambda,$$

   where $S_\lambda(x) = x(1 - \lambda/|x|)_+$ denotes the soft-thresholding operator.

   If all conditions are satisfied, set $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell]\bullet}) = 0$ and skip to (d).

   (b) If $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell]\bullet}) \neq 0$, check $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ by checking $\hat{\boldsymbol{\theta}}_{[\ell][g]} = 0$ for all $g = 1, 2, \ldots, G$.

   (i) First, compute $\hat{\boldsymbol{\beta}}_{[\ell]}$ by one dimensional optimization of each parameter in $\boldsymbol{\beta}_{[\ell]}$ until convergence.

   (ii) Then, check $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ given $\hat{\boldsymbol{\beta}}_{[\ell]}$ by checking $\hat{\boldsymbol{\theta}}_{[\ell][g]} = 0$ for all $g = 1, 2, \ldots, G$.

$$\left\| S_{\alpha\lambda} \left\{ \frac{1}{N} \sum_{i=1}^{N} \text{vec}(\mathbf{x}_{i[\ell]}^T \mathbf{z}_{i[g]}) (r_i^{(-\ell)(-g)} - \mathbf{x}_{i[\ell]} \hat{\boldsymbol{\beta}}_{[\ell]}) \right\} \right\|_2 < (1-\alpha)\lambda \frac{\sqrt{p_g p_\ell}}{\sqrt{1+K}}.$$

   If (ii) is satisfied for all $g = 1, \ldots, G$, set $\boldsymbol{\beta}_{[\ell]} = \hat{\boldsymbol{\beta}}_{[\ell]}$ and $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ and skip to (d).

   (c) If $\hat{\boldsymbol{\beta}}_{[\ell]} \neq 0$ and $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} \neq 0$ (i.e. if there exists g^* such that $\hat{\boldsymbol{\theta}}_{[\ell][g^*]} \neq 0$):

   (i) Use a generalized gradient procedure with approximation through the majorization-minimization algorithm to find $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell][NZ]})$ where $\boldsymbol{\theta}_{[\ell][NZ]}$ denotes the set of nonzero $\boldsymbol{\theta}_{[\ell][g]}$'s

   (ii) With the updated $\hat{\boldsymbol{\beta}}_{[\ell]}$, check the condition in 2-(2)-(b)-(ii) for all $g = 1, 2, \ldots, G$ again to confirm whether $\boldsymbol{\theta}_{[\ell][NZ]}$ contains the same set of $\boldsymbol{\theta}_{[\ell][g]}$'s.

   (iii) If the composition of $\boldsymbol{\theta}_{[\ell][NZ]}$ changed, repeat (i)-(iii) with the updated $\boldsymbol{\theta}_{[\ell][NZ]}$.

   (d) Compute $\hat{\beta}_0$ and $\hat{\boldsymbol{\theta}}_0$ from the regression of the current residual on $\mathbf{Z}$.

   (3) Compute $J^{*(new)}(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Theta}})$ with the current estimate of $(\hat{\beta}_0, \hat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Theta}})$.

golden section search and successive parabolic interpolation. We adopt this approach to compute $\hat{\boldsymbol{\beta}}_{[\ell]}$ in 2-(2)-(b)-(i) of the Algorithm 1. With this procedure, the pliable LASSO can be extended to problems where there are pre-specified group structure among the main predictors.

2.2.4 Comparison with the Pliable LASSO

The pliable LASSO proposed in Tibshirani and Friedman (2019) optimizes the objective function as below:

$$J(\beta_0, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N} \sum_{i=1}^{N} r_i^2 + \lambda P_\alpha(\boldsymbol{\beta}, \boldsymbol{\Theta}), \tag{2.4}$$

where $r_i = y_i - \beta_0 - \mathbf{z}_{i\bullet}\boldsymbol{\theta}_0 - \sum_{j=1}^{p} x_{ij}(\beta_j + \boldsymbol{\theta}_{j\bullet}\mathbf{z}_{i\bullet}^T)$ and

$$\lambda P_\alpha(\boldsymbol{\beta}, \boldsymbol{\Theta}) = (1-\alpha)\lambda \sum_{j=1}^{p} (||(\beta_j, \boldsymbol{\theta}_{j\bullet})||_2 + ||\boldsymbol{\theta}_{j\bullet}||_2) + \alpha\lambda \sum_{j,k} |\theta_{jk}|_1.$$

The penalty term $||\boldsymbol{\theta}_{j\bullet}||_2$ in equation (2.4) is for penalizing the group of all modifying variables as a whole for the "asymmetric weak hierarchy" constraint between the coefficients for the main predictors and the modifying variables and the term $|\theta_{jk}|_1$ is for penalizing each modifying variable. Hence, there is no consideration of the group structure among the modifying variables in equation (2.4). This may lead to spurious selection of irrelevant modifying variables as shown in our simulation study in Section 2.3. Assuming $L = p$ for simplicity, The svReg in equation (2.3) corrects this limitation by replacing the penalty term $||\boldsymbol{\theta}_{j\bullet}||_2$ with the terms $\{||\boldsymbol{\theta}_{j[g]}||_2\}_{g=1}^{G}$, which penalize each group of modifying variables with weight $\sqrt{p_g}/\sqrt{1+K}$. This weight accounts for the size of each group of modifying variables, p_g, and also finds balance between $||(\beta_j, \boldsymbol{\theta}_{j\bullet})||_2$ ($K+1$ parameters) and $||\boldsymbol{\theta}_{j[g]}||_2$ (p_g parameters).

Remark 1. *As noted in Tibshirani and Friedman (2019), the claim that the solutions to the objective function $J(\beta_0, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{\Theta})$ satisfy the asymmetric weak hierarchical property relies on a continuity argument and has not yet been proven rigorously. Likewise, our claim that the solutions to the objective function $J^*(\beta_0, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{\Theta})$ satisfy the asymmetric weak hierarchy has not been proven yet. By construction of the penalties in $J^*(\beta_0, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{\Theta})$, we expect that the probability of the*

continuous random variable $\hat{\beta}_j$ *to be zero while* $\hat{\boldsymbol{\theta}}_{j\bullet} \neq \mathbf{0}$ *will be zero for any* $j = 1, \ldots, p$.

The introduction of the penalties $\{||\boldsymbol{\theta}_{j[g]}||_2\}_{g=1}^{G}$ instead of $||\boldsymbol{\theta}_{j\bullet}||_2$ differentiates the svReg from the pliable LASSO since it not only allows the asymmetric weak hierarchy constraint to be satisfied between β_j and $\boldsymbol{\theta}_j$ as in the pliable LASSO but also allows the group structure of the modifying variables to be considered in the model selection. Also, even when there is no group structure among the modifying variables, the svReg is different from the pliable LASSO since each modifying variable is penalized individually by such penalty. This enables the selection of the modifying variables to be implemented at more granular level in the svReg than the pliable LASSO. We found that such penalties lead to better model selection by increasing the sensitivity and the specificity as shown in the simulation study in Section 2.3.4.

Also, in equation (2.4), the group structure among the main predictors is not considered. This may lead to incorrectly screening true relevant variables when the variables are grouped variables with high within-group correlation as the LASSO which tends to randomly select variables among highly correlated variables (Zhao and Yu, 2006). Our proposed remedy for this inconsistent variable selection is to group the variables using the information on the group structure of the main predictors so that the grouped variables are selected into the model or screened from the model together. Also, all the L_2 penalty terms are weighted differently by $\sqrt{p_\ell}$ in equation (2.3), accounting for different size of each group of the main predictors. This weight is analogous to the weight used in the group LASSO penalty (Yuan and Lin, 2006).

The use of weighted penalties in the svReg accounts for the different size of each group of main predictors and modifying variables. We discovered that imposing such different weights on the penalty terms leads to better variable selection performance as shown in Section 2.3. The improvement in model selection through weighted penalties has also been studied in Garcia et al. (2013, 2016).

In the algorithm 1, the svReg identifies the groups of possibly significant modifying variables first and then selects variables from those identified groups to reduce false selection. This is different from the pliable LASSO algorithm because the pliable LASSO selects variables from the

set of all modifying variables. Such identification of significant group first in the svReg allows the group structure of the modifying variables to be considered in the algorithm. Also, such algorithm is expected to increase the sensitivity in the selection of the grouped modifying variables since the significance of each group of modifying variables is evaluated at the group level, instead of being evaluated at individual variable level.

2.3 Simulation Study

2.3.1 Simulation Design

We compared our structural varying-coefficient regression proposed in Section 2.2 with the LASSO (Tibshirani, 1996) and the pliable LASSO (Tibshirani and Friedman, 2019) in some simulation settings. First, we considered the case when we have both continuous and categorical modifying variables. Second, we additionally considered the correlation between main predictors so that the highly correlated main predictors can be considered as grouped variables.

Setting 1 (Structured modifying variables): We generated 50 standard Gaussian independent predictors with sample size $N = 100$. We also generated twenty modifying variables: ten continuous variables, $z_{i1}, \ldots, z_{i10}$, and ten categorical variables of three categories, $z_{i11}, \ldots, z_{i30}$. Note that each categorical variable is expressed with two dummy variables, hence those two variables can be considered as grouped variables. The continuous modifying variables were generated from the standard Gaussian distribution. The categorical modifying variables were generated from the multinomial distribution with equal probability. The response was generated for $i = 1, \ldots, 100$ from

$$y_i = x_{i1} + x_{i2} + (1 + z_{i1})x_{i4} + (1 - z_{i2} + z_{i11} - z_{i12})x_{i5} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. In Setting 1, the number of parameters is 1,550 (50 in $\boldsymbol{\beta}$ and 1,500 in $\boldsymbol{\Theta}$).

Setting 2 (Structured main predictors & modifying variables): As in Setting 1, we considered 50 main predictors in Setting 2. Let $\{X_i\}_{i=1}^{50}$ denote the i-th main predictor. We generated X_3

and X_6 to be correlated with $\{X_1, X_2\}$ and $\{X_4, X_5\}$, respectively, as follows:

$$x_{i3} = \frac{2}{3}x_{i1} + \frac{2}{3}x_{i2} + \frac{1}{3}\gamma_i \quad \text{and} \quad x_{i6} = \frac{2}{3}x_{i4} + \frac{2}{3}x_{i5} + \frac{1}{3}\delta_i$$

where $\gamma_i \sim N(0,1)$ and $\delta_i \sim N(0,1)$. Other main predictors were standard Gaussian with sample size $N = 100$ and independent to each other. By this construction, x_{i3} and x_{i6} are normally distributed with mean 0 and variance 1 as other main predictors. Given the high correlation, we treated $\{X_1, X_2, X_3\}$ and $\{X_4, X_5, X_6\}$ as grouped variables when we fitted the svReg. This simulation setting is similar to that used in Zhao and Yu (2006) to create dependence between predictors in a model where the model selection result of the LASSO can be inconsistent. The modifying variables and the response were generated as in Setting 1. In Setting 2, the number of parameters is 1,550 (50 in $\boldsymbol{\beta}$ and 1,500 in $\boldsymbol{\Theta}$).

Setting 3 (High dimensional main predictors & structured modifying variables): We generated 200 standard Gaussian independent predictors with sample size $N = 100$. The modifying variables and the response were generated as in Setting 1. In Setting 3, the number of parameters is 6,200 (200 in $\boldsymbol{\beta}$ and 6,000 in $\boldsymbol{\Theta}$).

We applied three methods to the simulated data: the LASSO, the pliable LASSO and the svReg. In the LASSO, all combinations of the interaction between main predictors and modifying variables are considered to avoid model misspecification since the true models contain interaction terms that need to be considered. Since both the LASSO and the pliable LASSO ignore the group structure of the main predictors and the modifying variables, the svReg is expected to perform better than those methods in selecting relevant main predictors and screening irrelevant categorical modifying variables. The LASSO was fitted using the R package `glmnet` and the pliable LASSO and the svReg were fitted using our R package `svReg`.

We ran 100 simulations and used 10-fold cross-validation in each simulation to find the optimal value of the tuning parameter λ. In the cross-validation, we used decreasing λ's from 10 to 0.01 by 0.01 to find the solution path of the parameters. The λ value which minimizes the mean squared

error in the cross-validation was chosen for the model estimation. Such choice of λ using the cross-validation is one choice that allows a 'fair' comparison of the three considered shrinkage procedures. While cross-validation is a common practice for selecting λ, it may not be the best default approach to select the weight parameter α for the pliable LASSO and the svReg. From our experience, cross-validation often chose very large alpha values which forced the solution to converge to the LASSO solution without any modifying variables by shrinking all the interaction terms to zero. Such property of the weight parameter α was also described in Tibshirani and Friedman (2019) for the pliable LASSO. Hence, to avoid such extreme empirical choices of α, we fixed it at 0.5 for our analysis, which gives balanced weight between the penalties on the main predictors and the modifying variables.

2.3.2 Methods for Evaluation

To evaluate the model selection performance of the three methods, we computed the false discovery rate (FDR) (Benjamini and Hochberg, 1995), sensitivity and specificity, the average percentage of time variables are selected, and predictive accuracy as measured by the mean squared errors. We also visualise findings in so-called difference curves as introduced in Garcia et al. (2016).

The FDR is defined as the ratio of the number of irrelevant variables selected over the total number of variables selected. It measures how likely the method makes "false selection" so a high value of FDR is undesirable. Since the pliable LASSO ignores the group structure of the categorical modifying variables and treats the dummy variables separately, it is expected to select more irrelevant modifying variables spuriously than the structural varying-coefficient regression, leading to higher FDR.

Sensitivity is a measure of the "true positive rate" and it is the ratio of the number of relevant variables selected over the number of true relevant variables. Specificity is a measure of the "true negative rate" and it is the ratio of the number of irrelevant variables screened over the number of true irrelevant variables. Both high sensitivity and high specificity are desirable. In addition, we report the geometric mean of sensitivity and specificity ($= \sqrt{\text{Sensitivity} \times \text{Specificity}}$) as used in

Kubat et al. (1998).

We also computed the average percentage of time the variables are selected. The average percentage is computed for the relevant variable group and irrelevant variable group of the main predictors and the modifying variables separately. High percentage of selection is desirable for the relevant variable groups and vice versa for the irrelevant variable groups.

The predictive performance can be evaluated by the mean squared error (MSE) from the V-fold cross-validation. In V-fold cross-validation, the data is split into $(V-1)$ sets for a training set and a test set. The training set is used to fit a model ("training" step) and then, the fitted model is used for calculating the MSE of the response for the test set ("testing" step).

Lastly, we measured the computation time for the simulated data from each setting. We fitted each method to the data for decreasing values of λ from 10 to 0.1 in steps of 0.1. We repeated such task 10 times and measured the average computation time for each method. All timings were carried out using the R package `microbenchmark` on 1.6 GHz Intel Core i5 processor.

For a sample of size N, details of the V-fold cross-validation procedure are written below.

1. Permute the data randomly and label them from 1 to V sequentially. By doing this, each data point is assigned to one of the V groups.

2. Iterate the process below for the v-th group from group 1 to group V.

   (a) Consider the v-th group as the test set and all the other groups as the training set.

   (b) Fit a linear model of the form of equation (2.2) using the training set.

3. Using the fitted models for $v = 1, \ldots, V$, calculate MSE as below:

$$MSE = \frac{1}{N} \sum_{v} \sum_{i=1}^{n_v} (y_i - \hat{y}_i)^2$$

   where n_v is the number of data points for the v-th group and i is the index for the elements of the test set.

2.3.3 Simulation Results

Simulation results are reported in Table 2.1. In this table, we compared the LASSO, the pliable LASSO and the structural varying-coefficient regression with respect to variable selection and prediction accuracy. All models were estimated with the tuning parameter λ which gives the minimum MSE from 10-fold cross-validation.

				Setting 1 (structured modifying variables)			Setting 2 (structured main & modifying variables)			Setting 3 (high-dimensional main predictors)		
metric	covariates			LASSO	pLASSO	svReg	LASSO	pLASSO	svReg	LASSO	pLASSO	svReg
Percentage	Main	Relevant		1.00	1.00	1.00	0.95	0.95	1.00	0.99	1.00	1.00
of selection		Irrelevant		0.48	0.27	0.21	0.47	0.28	0.26	0.19	0.15	0.11
	Modifying	Relevant	continuous	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00
			categorical	0.84	1.00	1.00	0.84	1.00	1.00	0.86	0.98	1.00
		Irrelevant	continuous	0.72	0.78	0.57	0.64	0.80	0.68	0.72	0.72	0.57
			categorical	0.73	0.80	0.56	0.71	0.79	0.70	0.72	0.74	0.54
False discovery rate (FDR)				0.84	0.81	0.75	0.84	0.81	0.79	0.88	0.86	0.82
Sensitivity				0.96	1.00	1.00	0.93	0.98	1.00	0.96	0.99	1.00
Specificity				0.43	0.54	0.66	0.45	0.53	0.59	0.74	0.78	0.84
Geometric mean of sensitivity and specificity				0.63	0.73	0.81	0.64	0.71	0.76	0.84	0.88	0.91
Mean squared error (MSE)				2.57	2.62	2.46	2.55	2.69	2.53	3.37	2.94	2.68
Computation time (sec)				0.03	131	470	0.03	168	426	0.10	319	1,625

Table 2.1: Simulation results for the LASSO, the pliable LASSO (pLASSO) and the structural varying-coefficient regression (svReg). In Setting 1, 50 independent main predictors, 10 continuous modifying variables and 10 categorical modifying variables with 3 categories were generated. In Setting 2, correlation between main predictors were additionally considered. In Setting 3, 200 main predictors were generated for the same setting as Setting 1. All values are the average of the 100 simulations. MSE is computed with the tuning parameter λ which gives minimum MSE from 10-fold cross validation. For the pliable LASSO and the structural varying-coefficient regression, α is set to 0.5.

The pliable LASSO and the svReg select relevant modifying variables better than the LASSO since they correctly specify a varying-coefficient model and treat those modifying variables as the effect modifiers of the main predictors. Both methods also screen irrelevant variables better than the LASSO which leads to lower false discovery rate and higher specificity. Also, both work well even for the case when the number of main predictors is greater than the sample size $(p > n)$.

In Setting 1 and Setting 3, the strength of the svReg over the pliable LASSO is observed in

screening irrelevant variables, which in turn leads to lower FDR by up to 6% points and higher specificity by up to 12% points than the pliable LASSO. Hence, by considering the group structure among the modifying variables, the svReg identifies relevant variables correctly while making fewer inclusion of irrelevant variables than the pliable LASSO, which will eventually lead to a more parsimonious and correct model with easier interpretation.

In Setting 2, additional benefit of the svReg over the LASSO and the pliable LASSO can be found in consistent selection of relevant main predictors when those predictors are structured. As discussed in Zhao and Yu (2006), the LASSO fails to select the relevant main predictors consistently when the predictors are correlated and this is shown in Table 2.1 by the percentage of selection of the relevant main predictors $(= 0.95)$ being less than one. Interestingly, similar pattern is observed in the pliable LASSO. Although model selection consistency of the pliable LASSO is not within the scope of this dissertation, this simulation result indicates that the pliable LASSO also suffers from the problem of inconsistent variable selection when the variables are highly correlated. On the other hand, in the svReg, those correlated variables were grouped to be selected or screened together. Hence, the svReg shows consistent result of variable selection for the relevant main predictors with 2% point higher sensitivity than the pliable LASSO.

In terms of prediction accuracy, the cross-validation MSE of the structural varying-coefficient regression shows an improvement over the pliable LASSO by up to 9% in all simulation settings. This reflects the gain from accounting for the group structure among the modifying variables.

Figure 2.2 compares the receiver operating characteristic (ROC) curves of the LASSO, the pliable LASSO and the structural varying-coefficient regression. The ROC curve compares the true positive rate with the false positive rate over the different values of the penalty parameter, λ. True positive rate measures how well the method selects relevant variables and false positive rate measures the extent of incorrectly including irrelevant variables in the model. The structural varying-coefficient regression (solid red curve) shows higher true positive rate and lower false positive rate than other methods. Thus, we can conclude that structural varying-coefficient regression selects relevant variables more correctly while including fewer irrelevant variables than other

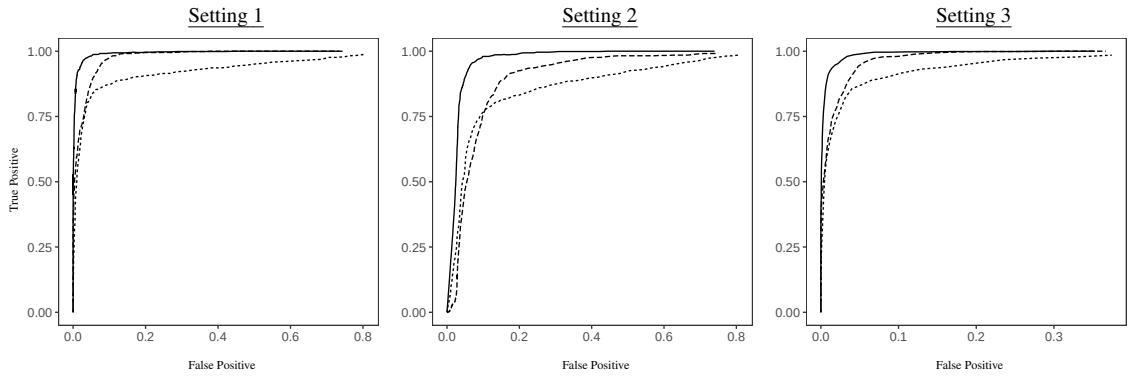

Figure 2.2: Receiver operating characteristic (ROC) curve of the LASSO (dotted curve), the pliable LASSO (dashed curve) and the structural varying-coefficient regression (solid curve) for Setting 1, Setting 2 and Setting 3. The structural varying-coefficient regression shows the lowest false-positive ratio for a fixed true-positive ratio. For the pliable LASSO and the structural varying-coefficient regression, α is set to 0.5.

methods.

Figure 2.3 compares the three methods by plotting the average percentage of selection using a difference curve, a visualisation introduced in Garcia et al. (2016). In a difference curve, the average percentage of time selected for each group of variables is compared to the “ideal" percentage of selection, which is 100% for relevant variables and 0% for irrelevant variables. That is, a better method in terms of variable selection has a lower curve in the plot. In all simulation settings, the curve of the structural varying-coefficient regression is below that of the pliable LASSO, which indicates that the svReg outperforms the pliable LASSO in selecting relevant variables and screening irrelevant variables.

2.3.4 Simulation without Structured Variables

Although the motivation of developing the structural varying-coefficient regression was to deal with the structured main predictors and modifying variables, we can apply our method to the special case when there is no structure among variables. We compared the performance of the svReg with the pliable LASSO. For this purpose, 50 standard Gaussian independent main predictors and 20 continuous modifying variables (Setting 4) or binary modifying variables with equal probability (Setting 5) were generated. The sample size N was 100. The response was generated for

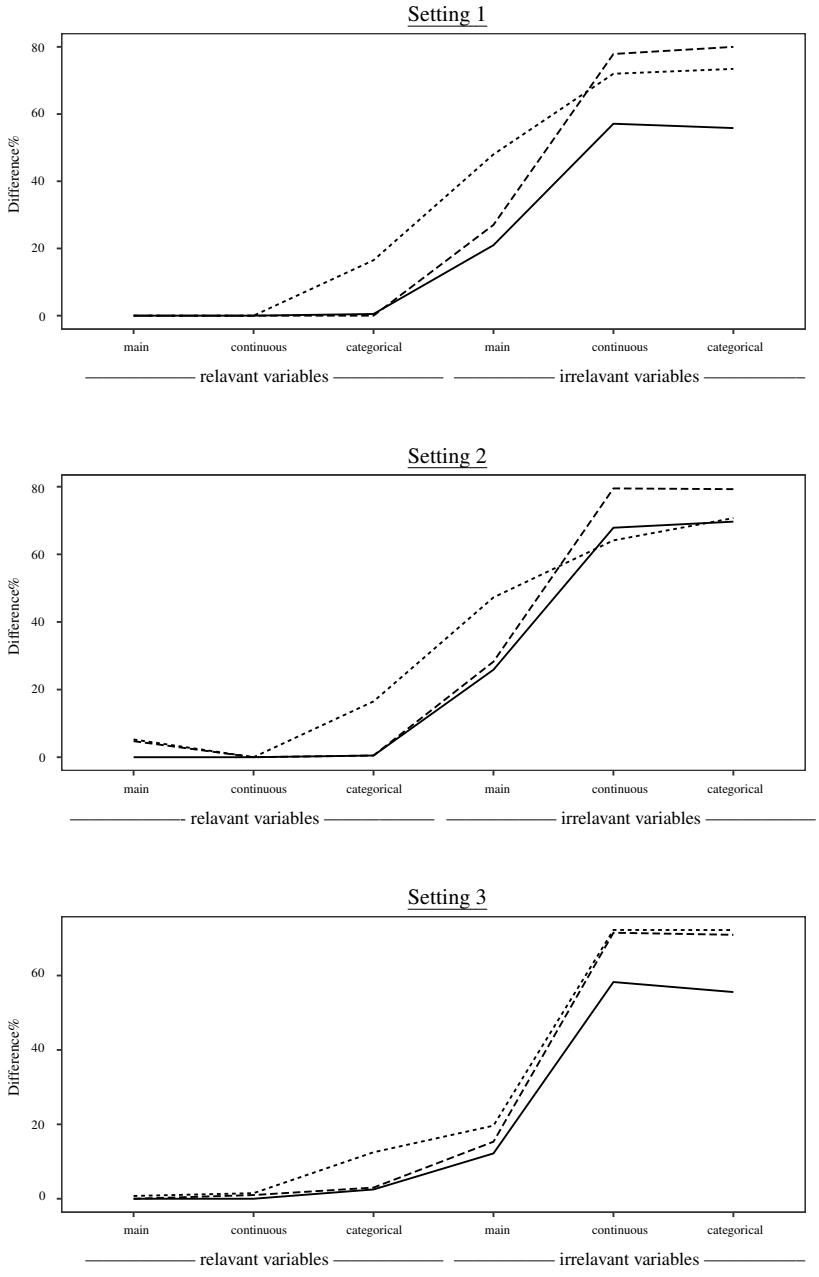

Figure 2.3: Difference curves of the LASSO (dotted curve), the pliable LASSO (dashed curve) and the structural varying-coefficient regression (solid curve) for Setting 1, Setting 2 and Setting 3. In a difference curve, a method with lower curve outperforms a method with upper curve in selecting relevant variables and screening irrelevant variables. In both settings, “main" represents main predictors, “continuous" represents continuous modifying variables and “categorical" represents categorical modifying variables with 3 categories. The structural varying-coefficient regression generally shows lower difference than the LASSO and the pliable LASSO for both settings. For the pliable LASSO and the structural varying-coefficient regression, α is set to 0.5.

$i = 1, \ldots, 100$ from

$$y_i = x_{i1} + x_{i2} + (1 + z_{i1})x_{i3} + (1 - z_{i2})x_{i4} + \epsilon_i$$

where $\epsilon_i \sim N(0, 1)$. In Setting 4 and 5, the number of parameters is 1,050 (50 in $\boldsymbol{\beta}$ and 1,000 in $\boldsymbol{\Theta}$).

The result from this simulation is given in Table 2.2. As in Table 2.1, the svReg selects fewer irrelevant main predictors than the pliable LASSO by 5% to 10% points and fewer irrelevant modifying variables by 17% to 24% points. This leads to lower FDR and higher specificity for the structural varying-coefficient regression. Also, the prediction error of the svReg is lower than that of the pliable LASSO.

			Setting 4 (continuous)		Setting 5 (binary)	
metric	covariates		pLASSO	svReg	pLASSO	svReg
Percentage	Main	Relevant	1.00	1.00	1.00	1.00
of selection		Irrelevant	0.28	0.18	0.21	0.16
	Modifying	Relevant	1.00	1.00	0.91	0.96
		Irrelevant	0.76	0.52	0.63	0.46
False discovery rate (FDR)			0.82	0.75	0.78	0.73
Sensitivity			1.00	1.00	0.97	0.99
Specificity			0.59	0.72	0.67	0.75
Geometric mean of sensitivity and specificity			0.75	0.84	0.80	0.85
Mean squared error (MSE)			2.24	2.11	1.59	1.54
Computation time (sec)			72	320	31	234

Table 2.2: Simulation results for the pliable LASSO (pLASSO) and the structural varying-coefficient regression (svReg) when there is no structure among the main predictors or modifying variables. 50 independent main predictors and 20 continuous (Setting 4) or binary (Setting 5) modifying variables were considered. All values are the average of the 100 simulations. MSE is computed with the tuning parameter λ which gives minimum MSE from 10-fold cross validation. For the pliable LASSO and the svReg, α is set to 0.5.

The reason why the structural varying-coefficient regression outperforms the pliable LASSO

for the variable selection purpose is related to the screening conditions for zero coefficients. In the pliable LASSO, the screening condition for $(\hat{\beta}_j, \hat{\boldsymbol{\theta}}_{j\bullet}) = 0$ involves the calculation of the quantity as below:

$$\left\| S_{\alpha\lambda}\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij}\mathbf{z}_{i\bullet} r_i^{(-j)} \right) \right\|_2, \tag{2.5}$$

and the screening condition is applied to the L_2-norm of the vector of coefficients for all modifying variables as a group (i.e., the size of this group is K). On the other hand, the corresponding condition for the structural varying-coefficient regression involves the calculation of the quantity as below:

$$\left\| S_{\alpha\lambda}\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij}\mathbf{z}_{i[g]} r_i^{(-j)(-g)} \right) \right\|_2. \tag{2.6}$$

Note the $\sum_{i=1}^{N} x_{ij}\mathbf{z}_{i[g]} r_i^{(-j)(-g)}$ takes a scalar value for a continuous modifying variable without any group structure with other modifying variables. In (2.6), each continuous modifying variable is treated as one group of variable (i.e., the size of each group is one) and the screening condition is applied to the coefficient for each modifying variable. Thus, the difference between (2.5) and (2.6) is that (2.6) will penalize each continuous modifying variable individually, while (2.5) will penalize all modifying variables as a group. Even if some elements of the coefficient vector are large and others are small, (2.5) can take large value which leads to possibly non-zero coefficients for all modifying variables whereas (2.6) will take small values for those elements.

Also, once some of the $\{\boldsymbol{\theta}_{j[g]}\}_{g=1}^{G}$ turn out to be zero, the svReg uses gradient descent procedure only for the nonzero $\boldsymbol{\theta}_{j[g]}$'s. This is not the case in the pliable LASSO where the gradient descent is performed for all $\{\theta_{jk}\}_{k=1}^{K}$ if $\boldsymbol{\theta}_{j\bullet}$ is nonzero. This allows the svReg to find the zero coefficients more efficiently than the pliable LASSO.

Figure 2.4 compares those two methods in how the number of nonzero coefficients changes as the algorithm iterates for one simulation case. As the number of iterations increases, both methods keep excluding less significant variables from the model. However, the structural varying-coefficient regression ends up with fewer, but true nonzero β's and θ's than the pliable LASSO. Also, the structural varying-coefficient regression finds those significant variables with fewer itera-

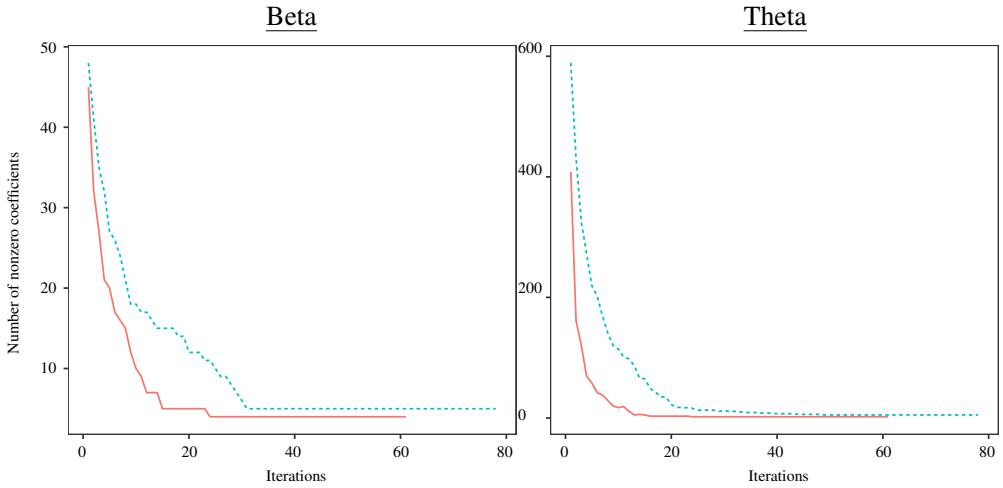

Figure 2.4: Number of nonzero coefficient estimates for main predictors (left panel) and modifying variables (right panel) plotted along the number of iterations of the algorithm for the pliable LASSO (dotted green curve) and the structural varying-coefficient regression (solid red curve). As both algorithms iterate, the coefficients of less significant variables shrink to zero and only significant variables remain in the model with nonzero coefficients. The pliable LASSO algorithm stops iteration at 78 while the structural varying-coefficient regression stops at 61. Also, the speed of excluding insignificant variables is faster in the structural varying-coefficient regression than the pliable LASSO. At the end, the pliable LASSO ends up with five nonzero β's (one false discovery) and five nonzero θ's (three false discoveries). The structural varying-coefficient regression leads to four nonzero β's and two nonzero θ's, only the true significant terms. This is one simulation case from the simulation experiment. The tuning parameter λ is 0.35 and α is 0.5.

tions than the pliable LASSO. Hence, the structural varying-coefficient regression can be preferred over the pliable LASSO regardless of whether there are categorical modifying variables or not.

2.4 Brain Regions Affecting Motor Impairment in Huntington Disease

2.4.1 Clinical Research Problem

We applied our method to the Neurobiological Predictors of Huntington Disease (PREDICT-HD), a large observational study from 2001 to 2013 on potential neurobiological markers of Huntington Disease (HD). We focus on the data of $N = 710$ subjects who are "at risk" of HD with CAG (cytosine, adenine, guanine) repeats greater than or equal to 36. Subjects at risk means that they may or may not exhibit Huntington disease symptoms, whereas those with CAG repeat less than 36 is expected not to develop HD symptoms. The majority of the subjects were female (63.5%).

On average, the subjects were 40.5 years old, had 42.4 CAG repeats (ranges from 37 to 61), and had 14.2 years of education.

In this study, participants enter the study at different phase of the disease. Hence, each participant is subject to different “disease severity" or different proximity to HD diagnosis. As a measure of disease severity, we used the scaled CAG-Age-Product (CAP) score, the product of CAG repeats and age as proposed in Zhang et al. (2011). CAP score is often used as a categorical variable to remove the within-group variability with three categories: low, medium and high. Participants categorized as “high" are regarded as having high probability of being diagnosed with HD based on motor functions in the next 5 years. In our data, about 27% of the subjects are categorized as “low" with CAP score less than 0.67 and about 37% of the subjects as “high" with CAP score greater than 0.85.

In the PREDICT-HD study, the interest is to identify brain regions which are associated with motor impairment. As a measure of motor impairment, we used the total motor score (TMS), a measurement of the overall motor impairment ranging from 0 (no impairment) to 124 (high impairment). As covariates, we used the volume measures of brain regions. Also, as explained above, each subject has different disease severity. If we ignore this feature of the data, the effect from the brain regions on motor impairment will be mixed with the effect of the disease severity and the model will not capture the “pure" effect of the brain regions. For this reason, CAP score has been used as another covariate or control variable (Garcia et al., 2016; Zhang et al., 2011) in addition to the volume measures of brain regions.

However, including the CAP score simply as another covariate assumes that the effects of brain regions on motor impairment are fixed regardless of the CAP score. This assumption is questionable since there may be a different pattern between, for example, the high CAP group and the low CAP group. In Figure 2.1, the least squares regression line between total motor score and volume of brain regions were fitted for the high/medium/low CAP score groups separately. In the top left panel, covariate is the volume of the left caudate and the response variable is the total motor score. It can be clearly observed that the slope of the high CAP group (solid line) is different from

that of low (dotted line) or medium (dashed line) CAP group. This difference in slope indicates that the effect of the left caudate on total motor score depends on whether a participant has high CAP score or not. On the other hand, in the bottom right panel where the covariate is the volume of the right vessel, the difference in slopes is not as clear as in the left caudate. These results indicate that the effects of some brain regions may differ by participant groups but other brain regions may not.

Thus, our interest in this analysis is not only to identify brain regions associated with motor impairment but also to understand how their effects on motor impairment differ by participant groups. This can be achieved by fitting a varying-coefficient regression with the total motor score as a response, volumes of brain regions as main predictors and the CAP score as a modifying variable. In addition, we included gender and years of education as possible modifying variables since the effects of brain regions on motor impairment may also differ by participant groups defined by these variables. Since the CAP score data contains information of both age and CAG repeat by its definition, those two variables were not used as modifying variables.

For estimating the varying-coefficient model, the pliable LASSO (Tibshirani and Friedman, 2019) and the svReg were used. As discussed in Section 2.2, the svReg can consider the pre-specified structure of the variables, whereas the pliable LASSO cannot. Since some main predictors represent the left part and the right part of a brain region (e.g. left caudate vs. right caudate), those main predictors were grouped in the svReg. Also, since the CAP score is expressed as a group of two binary dummy variables, those dummy variables were also regarded as grouped modifying variables in the svReg. Additionally, we considered the LASSO allowing for interaction terms to be selected as in Section 2.3. However, the LASSO is not appropriate for fitting a varying-coefficient model since some main predictors may not be selected even if their interaction terms are selected by the LASSO. Hence, we compared the pliable LASSO and the svReg applied to the PREDICT-HD study. For both methods, the tuning parameter λ was selected based on 10-fold cross-validation and the weight parameter α was set to 0.5.

2.4.2 Analysis Results

Table 2.3 summarizes results for the pliable LASSO (left table) and the svReg (right table). The first column for each method shows the fitted parameters, β, for the main predictors (brain regions) and the other columns show the fitted parameters, θ, for the modifying variables (gender, years of education, CAP score) in the coefficient of each main predictor, as defined in equation (2.2). Here, “CAP(medium)" and “CAP(high)" express the binary variable for the medium CAP score group and the high CAP score group, respectively.

From the nonzero θ estimates for basal ganglia (brain regions related to motor movements including caudate, putamen and pallidum), we can infer that the effects from these brain regions to motor impairment differ by CAP score groups. Particularly, the θ estimates for CAP(high) take negative values, meaning that high CAP score group has steeper slope as observed in Figure 2.1 than low or medium CAP score group. This indicates that the motor function of the high CAP score group may deteriorate faster than other groups given a certain amount of volume change in those brain regions.

Interestingly, the θ for CAP(high) in the coefficient of the left pallidum was determined to be zero by the svReg. This means that the effect of left pallidum on motor impairment may not differ significantly between the high CAP score group and other groups. This is consistent with Figure 2.1 where the differences in slopes are relatively small for the left pallidum. Note that, for the pliable LASSO, this θ estimate is zero simply because the main effect of the left pallidum was not selected. However, the main effect of the left pallidum may have been excluded randomly by the pliable LASSO due to its high correlation with the right pallidum. Hence, the pliable LASSO does not clearly tell us whether the effect of the left pallidum on motor impairment is the same across all participants or differ by disease severity groups whereas the svReg does. The least squares regression of the total motor score on each brain region allowing for interaction with the disease severity also indicates that the interaction between each brain region and CAP(high) is significant for all regions in basal ganglia except for the left pallidum. These least squares regression results can be found in Table 2.4.

The θ's for CAP(medium) in the coefficients of the putamen were determined to be zero by the svReg, whereas the pliable LASSO estimated a positive θ value for CAP(medium) in the coefficient of the right putamen. However, the θ's for CAP(medium) are expected to take negative values as those for CAP(high) because the baseline category is the low CAP group. Thus, the positive θ parameter by the pliable LASSO may have been selected spuriously, meaning that the effect of right putamen on motor impairment may not differ significantly between the low CAP group and the medium CAP group. This can also be inferred from Figure 2.1 where the least squares fit slopes for the low group and the medium group were indistinguishable. Hence, the svReg resulted in selecting fewer irrelevant θ's than the pliable LASSO. This result is consistent with the simulation study in Section 2.3 where the svReg selected fewer irrelevant variables than the pliable LASSO. Correct screening of irrelevant variables will not only result in models with smaller standard errors but also enable clinicians to avoid unnecessary segmentation of the patients in developing customized interventions for patient groups.

To the best of our knowledge, our study is the first to identify the interaction effect between CAP score and the volume of brain regions to motor impairment. This implies the genuine effect from the brain regions to motor impairment can be better understood when the CAP score is taken into account as a modifying variable in a varying-coefficient model. This knowledge can be useful in developing interventions or treatments which target specific group of patients. For example, a newly developed treatment may have some side effect. In this case, we may want to minimize the dosage of the treatment to reduce the risk of the side effect. From our research, we know that the high CAP score group will suffer more severe motor impairment than other groups given some change of the volume of caudate. If the degree of motor impairment is tolerable for low-medium CAP score group but not for high CAP score group, clinicians may need to use the treatment only for the high CAP score group or use different dosage for each group.

2.5 Discussion

In this chapter, we proposed a new variable selection method for a varying-coefficient model with pre-specified group structure among variables. We showed in multiple simulation settings

	Pliable Lasso					svReg				
	β	θ				β	θ			
	Main effect	Gender	Education	CAP (medium)	CAP (high)	Main effect	Gender	Education	CAP (medium)	CAP (high)
Lateral Ventricle	L: 0.063 R: 0	L: -0.087 R: 0		L: -0.053 R: 0	L: 0.034 R: 0	L: 0.063 R: 0.076	L: -0.079 R: -0.094		L: -0.014 R: -0.007	L: 0.009 R: 0.001
Cerebellum Cortex	L: 0 R: -0.039									
Thalamus Proper	L: 0 R: 0.327					L: 0.016 R: 0.134				
Caudate	L: -0.770 R: 0			L: 0.050 R: 0	L: -0.291 R: 0	L: -0.525 R: -0.160			L: 0.014 R: 0.052	L: -0.190 R: -0.125
Putamen	L: 0 R: -0.174			L: 0 R: 0.002	L: 0 R: -0.089	L: -0.102 R: -0.148				L: -0.090 R: -0.125
Pallidum	L: 0 R: -0.669				L: 0 R: -0.077	L: -0.011 R: -0.507				L: 0 R: -0.043
Vessel	L: 0 R: 0.435					L: 0.097 R: 0.315				
Choroid Plexus	L: 0.295 R: 0	L: -0.007 R: 0		L: -0.026 R: 0	L: 0.018 R: 0	L: 0.198 R: -0.045			L: -0.023 R: -0.012	L: 0.015 R: 0
CorticalWhiteMatter	L: -0.230 R: 0					L: -0.122 R: -0.058	L: 0.034 R: 0.001			
3rd Ventricle	0.073			-0.034	0.037					
4th Ventricle	-0.109					-0.111				
CSF	0.163					0.193			-0.021	0.011
WM Hypointensity	0.151			-0.011	0.031	0.140			-0.022	0.070
Optic Chiasm	0.430			-0.059	0.079	0.431			-0.083	0.112
CC Posterior	-0.114	0.037	-0.062		-0.032	-0.144	0.031	-0.087		-0.006

Table 2.3: Parameter estimates of the selected brain regions by the pliable LASSO (pLASSO) and the structural varying-coefficient regression (svReg) for PREDICT-HD data. Parameter values are based on scaled data. Parameters not selected are shown as blank. The first column for each method contains the fixed part of the regression coefficients of main predictors (β's in equation (2.2)). The other columns represent the varying part of the regression coefficients of main predictors (θ's in equation (2.2)). That is, the parameters from the second to fifth columns are the coefficients of the interaction terms between the brain regions (in row) and the modifying variables (in column). For the grouped brain regions (those with two lines), "L" represents the left part of the corresponding brain region and "R" represents the right part of the brain region. Tuning parameter λ is selected from 10-fold cross-validation. α is set to 0.5.

	(model 1)	(model 2)	(model 3)	(model 4)	(model 5)	(model 6)
Left.Caudate	-0.005^{***}					
	(0.001)					
Right.Caudate		-0.004^{***}				
		(0.001)				
Left.Putamen			-0.003^{***}			
			(0.0004)			
Right.Putamen				-0.003^{***}		
				(0.0004)		
Left.Pallidum					-0.005^{***}	
					(0.001)	
Right.Pallidum						-0.010^{***}
						(0.001)
CAPmed	-12.831^{***}	-12.840^{***}	-12.726^{***}	-11.652^{***}	-5.454^{***}	-9.386^{***}
	(2.753)	(2.714)	(2.541)	(2.346)	(1.888)	(2.269)
CAPlow	-14.079^{***}	-12.401^{***}	-14.170^{***}	-13.775^{***}	-7.001^{***}	-11.301^{***}
	(3.267)	(3.182)	(3.191)	(3.080)	(2.311)	(2.662)
Left.Caudate:CAPmed	0.004^{***}					
	(0.001)					
Left.Caudate:CAPlow	0.004^{***}					
	(0.001)					
Right.Caudate:CAPmed		0.003^{***}				
		(0.001)				
Right.Caudate:CAPlow		0.003^{***}				
		(0.001)				
Left.Putamen:CAPmed			0.002^{***}			
			(0.001)			
Left.Putamen:CAPlow			0.003^{***}			
			(0.001)			
Right.Putamen:CAPmed				0.002^{***}		
				(0.001)		
Right.Putamen:CAPlow				0.003^{***}		
				(0.001)		
Left.Pallidum:CAPmed					0.003	
					(0.002)	
Left.Pallidum:CAPlow					0.004^{*}	
					(0.002)	
Right.Pallidum:CAPmed						0.007^{***}
						(0.002)
Right.Pallidum:CAPlow						0.008^{***}
						(0.002)
Constant	20.508^{***}	18.861^{***}	19.500^{***}	18.588^{***}	12.157^{***}	17.570^{***}
	(1.777)	(1.733)	(1.611)	(1.465)	(1.088)	(1.400)
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 2.4: Results for least squares regression of the total motor score on the volumes of basal ganglia regions and disease severity including interaction terms. Standard error for each coefficient is written in parenthesis.

that ignoring this group structure among variables reduced the specificity by up to 12% points and increased the false discovery rate by up to 6% points. It also led to inconsistent selection of relevant main predictors when there is group structure with high within-group correlation and this lowered the sensitivity by 2% points. We applied our method to the Huntington disease study and found that the effect from basal ganglia to motor impairment differs by disease severity of the patients. Such knowledge suggests that different medical interventions might be needed depending on each patient's disease severity.

If other variables in addition to the disease severity are identified as relevant modifying variables in future study, that can be used for extending to the so called personalized interventions which account for the traits of each individual. For example, if gender (male or female) and years of education (integer between 0 and 20) have turned out to be relevant modifying variables, the maximum number of possible models is 126 $(= 3 \times 2 \times 21)$. Each of these models reflects the individual traits determined by the values of the three modifying variables for each patient and this individualized regression model will be useful for developing personalized interventions.

In our application, the group structure among the variables is pre-defined and the number of the groups for the main predictors (L) and the modifying variables (G) are known. Our proposed method has been developed assuming such structural information is available. However, such information may not be available in some other applications. In this case, one additional consideration is the correlation among variables, which needs to be dealt with. For example, the unknown group structure among the variables can be addressed by selecting highly correlated variables together in the model. For a linear model with fixed coefficients, the selection of correlated variables was addressed by the elastic net (Zou and Hastie, 2005) as a modification of the LASSO. For a varying-coefficient model, similar modification to the pliable LASSO might be considered for future research.

In our analysis, we considered only the linear combination of the modifying variables as the varying-coefficient in equation (2.1). This is consistent with the basic setting discussed in Tibshirani and Friedman (2019) but both the pliable LASSO and the svReg can be generalized to

consider a non-parametric form of the varying-coefficient such as splines. The literature on the varying index coefficient model (Ma and Song, 2015; Na et al., 2019) addresses the estimation of non-linear functional coefficient for a family of models in which the model (2.1) is a special case. This branch of literature might serve as a starting point for future research.

Our method is designed for a regression model. However, it can be extended to accommodate survival models or generalized linear models by changing the objective function in equation (2.3). For example, our method can be applied to Cox's proportional hazard model by adding the svReg penalty $\lambda P_{\alpha}^{*}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ in equation (2.3) to the log partial likelihood of the hazard model. A similar attempt has recently been made by Du and Tibshirani (2018) for extending the pliable LASSO to the Cox's proportional hazard model. However, as with the pliable LASSO for a linear model, their method does not account for the pre-specified structure of the variables. The extension of the svReg to the hazard model is expected to select relevant variables consistently and screen irrelevant variables better than the method by Du and Tibshirani (2018) as was the case for the linear model settings and this will be future research.

3. EFFICIENT ESTIMATION OF A COVARIANCE MATRIX WITH ZERO ENTRIES

3.1 Introduction

In multivariate analyses, the strength of relationships between variables is often identified with a covariance matrix and estimation of the covariance matrix is critical in a variety of statistical analyses in genetics, finance and so on. In some applications, covariance matrices are often assumed to contain some zero entries, meaning that there is no linear association between those variables. For example, in genetics, Butte et al. (2000) developed the relevance network among genes assuming that the unconnected genes have zero covariance. Similarly, Rothman et al. (2009) used a covariance matrix with zero entries for gene clustering problems where genes with zero covariance are not clustered together. In finance, covariance matrices with zero entries have been applied to the investment portfolio selection to avoid risk underestimation (El Karoui et al., 2010; Xue et al., 2012). For climate data analysis, Bickel et al. (2008a) exploited the zero covariance between different spatial locations to separate the temperature pattern between continents.

Whether some entries of a covariance matrix are zero may or may not be known a priori. Sometimes, such prior information is available from external sources. For example, in finance and insurance industry, covariance matrices are used for allowing diversification effect between lines of business when quantifying the corporate risk. The regulators of the industry often use “expert judgment" to construct such covariance matrices and may force some entries to zero (BCBS, 2010; Calibration, 2010). On the other hand, when such prior information is unavailable, zero entries of a covariance matrix can be identified from data by some model selection procedures. Such selection procedures have been studied in, for example, Drton and Perlman (2004); Drton et al. (2007); Bickel et al. (2008a); Rothman et al. (2009); Bien and Tibshirani (2011).

Once the zero entries are determined in a covariance matrix, the covariance matrix can be estimated with the zero constraint on those entries. Imposing zero constraint on the entries of a covariance matrix is a special form of the linear covariance model (Anderson, 1970, 1973) where

a covariance matrix is modeled by a linear combination of some known matrices. For a Gaussian random vector, maximum likelihood estimation of the covariance matrix with such linear constraints may lead to an optimization problem with multiple local solutions (Chaudhuri et al., 2007; Zwiernik et al., 2017) due to non-concavity of the log-likelihood function (Bien and Tibshirani, 2011).

Existence of multiple local solutions is problematic because asymptotic properties of the global maximum solution such as consistency and asymptotic efficiency may not be shared by the other local solutions. Since the lack of the knowledge on the asymptotic distribution limits the discussion on the variability of the estimator, asymptotic properties of an estimator should be understood unless the estimator is guaranteed to be a global maximum. Zwiernik et al. (2017) discussed some probabilistic conditions under which the Gaussian log-likelihood function is concave so that a solution from any hill-climbing methods will converge to the global maximum. However, with finite samples, whether the solution from such hill-climbing methods is the maximum likelihood estimator is still not guaranteed.

For maximum likelihood estimation of the Gaussian linear covariance model, Anderson (1973) proposed an iterative scheme which finds a local solution to the normal likelihood equation and proved asymptotic efficiency of the solution from the scheme when the sample size is greater than the number of variables. However, convergence of the iteration is not guaranteed so it may fail to reach a solution. Also, Anderson (1973)'s scheme requires a positive definite estimator as an initial estimator but no easy guidance yet exists to construct this initial estimator. Zou et al. (2017) proposed conditions under which consistency and asymptotic efficiency of the maximum likelihood estimator can be achieved even when the number of variables is greater than the sample size. However, this result is based on the unimodal assumption, that is, it is assumed that only one local maximum exists and it is the unique global maximum.

For covariance matrices with zero constraint, asymptotically efficient estimators have been proposed for some approximations to the maximum likelihood estimation. Kauermann (1996) discussed dual estimation of the covariance matrix with zero constraint which has a unique, positive

definite and asymptotically efficient solution. Wermuth et al. (2006) derived approximations to maximum likelihood estimates that are asymptotically efficient. However, both methods do not exactly give the solution to the maximum likelihood estimation, meaning that there will be other estimators which will more likely generate the observed data than those estimators.

The iterative conditional fitting algorithm proposed by Chaudhuri et al. (2007) finds a local solution to the maximum likelihood estimation of covariance matrices with zero constraint when the sample size is greater than the number of variables. Since it guarantees convergence to a positive definite local maximum or saddle point, it resolves the limitation of Anderson (1973)'s algorithm. However, unlike the solution from the Anderson (1973)'s algorithm, asymptotic bahaviors of the solution from the iterative conditional fitting algorithm have not been understood yet. This lack of understanding on the asymptotic properties has limited the discussion on the variability of the covariance matrix estimator from the iterative conditional fitting algorithm. Also, this algorithm works only when the sample size is greater than the number of variables, which further limits the use of this algorithm in practice.

In this chapter, we combine the advantages of asymptotic efficiency of Anderson (1973) and the convergence property of Chaudhuri et al. (2007) when the location of the zero entries is known. Specifically, we prove that the iterative conditional fitting algorithm will produce a positive definite and asymptotically efficient covariance estimator when the algorithm starts from a consistent estimator as in Anderson (1973). In contrast to Anderson (1973), we suggest an easy and explicit way to construct this initial consistent estimator.

Second, we extend the iterative conditional fitting algorithm to the case when the sample size is smaller than the number of variables. We propose the iterative conditional ridge algorithm which replaces the least squares regression in each iteration of the iterative conditional fitting algorithm with the ridge regression. The solution from the iterative conditional ridge algorithm is positive definite and provides asymptotically efficient estimators of the off-diagonal non-zero entries. This estimator contains some bias in the diagonal entries but extent of the bias can be controlled under any positive constant.

Last, based on the understanding of the asymptotic behavior of the iterative conditional fitting algorithm, we discuss the implication of model underfitting and overfitting when the true covariance model is unknown. Specifically, we claim that model underfitting may induce additional bias to the estimator whereas overfitting increases variability of the estimator.

3.2 Covariance Matrices with Zero Entries

For a random vector $\boldsymbol{y} = (Y_1, Y_2, \ldots, Y_p)^T$ with zero mean vector and unknown covariance matrix $\boldsymbol{\Sigma}$, we consider the estimation of the covariance matrix with a pre-defined zero constraint. This problem can be cast as estimation of the linear covariance model (Anderson, 1973) as below:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\sigma}) = \sigma_1 \boldsymbol{G}_1 + \ldots + \sigma_K \boldsymbol{G}_K \tag{3.1}$$

where each $\boldsymbol{G}_k$ for $k = 1, \ldots, K$ is a $p \times p$ symmetric matrix of 0's and 1's for representing a non-zero element of $\boldsymbol{\Sigma}$ such that 1 in $\boldsymbol{G}_k$ indicates the location of the non-zero element in $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta} = (\sigma_1, \ldots, \sigma_K)^T$ are parameters for estimation. The parameter space for this model is

$$\boldsymbol{\Theta} = \{\boldsymbol{\sigma} : \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \text{ is positive definite}\}.$$

Note that model (3.1) is general enough to include any $p \times p$ unconstrained covariance matrix by $\boldsymbol{\Sigma} = (\sigma_{ij}) = \sum_{i \le j} \sigma_{ij} \mathbf{U}_{ij} = \sum_{k=1}^{K} \sigma_k \boldsymbol{G}_k$ where $\mathbf{U}_{ij}$ contains 1's as the (i, j)-th and (j, i)-th elements and 0's elsewhere, $K = p(p+1)/2$ and each $\sigma_k \boldsymbol{G}_k$ corresponds to one of $\sigma_{ij} \mathbf{U}_{ij}$.

For example, with the constraint that $(1, 3)$-th entry is zero, a 3×3 matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ can be modeled by a linear combination of $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_5$ as below:

$$\boldsymbol{\Sigma} = \sigma_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \sigma_2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \sigma_3 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \sigma_4 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \sigma_5 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

so that $\sigma_1 = \sigma_{11}, \sigma_2 = \sigma_{22}, \sigma_3 = \sigma_{33}, \sigma_4 = \sigma_{12} = \sigma_{21}$ and $\sigma_5 = \sigma_{23} = \sigma_{32}$.

3.2.1 Some Estimators of Covariance Matrices with Zero Entries

3.2.1.1 Ordinary Least Squares (OLS) Estimator

Anderson (1970) discussed the ordinary least squares estimator, $\hat{\boldsymbol{\sigma}}^{OLS}$, as a consistent estimator for $\boldsymbol{\sigma}$ in the model (3.1) as below:

$$\widehat{\boldsymbol{\sigma}}^{OLS} = \text{argmin}_{\boldsymbol{\sigma}} ||\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\sigma})||_F^2.$$

where $\boldsymbol{S}$ is the $p \times p$ sample covariance matrix and $|| \cdot ||_F$ denotes the Frobenius norm. By differentiating the objective function $||\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\sigma})||_F^2$ with respect to $\boldsymbol{\sigma}$, this method finds the solution for the estimating equation for $k = 1, \ldots, K$:

$$\text{trace}\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})\boldsymbol{G}_k\} = \text{trace}(\boldsymbol{S}\boldsymbol{G}_k) \tag{3.2}$$

Using the equation (3.1), the equation (3.2) for all $k = 1, \ldots, K$ can be written as, for $k = 1, \ldots, K$ and $\ell = 1, \ldots, K$,

$$\{\text{trace}(\boldsymbol{G}_k\boldsymbol{G}_\ell)\}_{k\ell}\boldsymbol{\sigma} = \{\text{trace}(\boldsymbol{S}\boldsymbol{G}_k)\}_k \tag{3.3}$$

where $\{\text{trace}(\boldsymbol{G}_k\boldsymbol{G}_\ell)\}_{k\ell}$ is a $K \times K$ matrix and $\{\text{trace}(\boldsymbol{S}\boldsymbol{G}_k)\}_k$ is a K-dimensional column vector. Note that the equation (3.3) is the linear system with a closed form solution as $\widehat{\boldsymbol{\sigma}}^{OLS} = \{\text{trace}(\boldsymbol{G}_k\boldsymbol{G}_\ell)\}_{k\ell}^{-1}\{\text{trace}(\boldsymbol{S}\boldsymbol{G}_k)\}_k$.

For the covariance estimation with zero constraints, one can show that $\widehat{\boldsymbol{\sigma}}^{OLS}$ is the vector of the sample covariances for the unconstrained non-zero entries. That is, $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}) = \sum_{k=1}^{K} \boldsymbol{S} \circ \boldsymbol{G}_k$ where $\circ$ denotes the Hadamard product. Hence, each component of $\widehat{\boldsymbol{\sigma}}^{OLS}$ has the same property as the sample covariance. For example, $\widehat{\boldsymbol{\sigma}}^{OLS}$ is a consistent estimator of $\boldsymbol{\sigma}$ since the (i, j)-th entry of the sample covariance matrix, s_{ij}, converges to σ_{ij} in $\boldsymbol{\Sigma}$ by the law of large numbers.

One problem of the OLS estimator is that it may not lead to a positive definite matrix. For

example, consider a sample covariance matrix as below:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}$$

which is a positive definite matrix. If we impose a zero constraint on the $(2,3)$-th (and $(3,2)$-th) entry, the OLS estimator is

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}) = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0 \\ 0.9 & 0 & 1 \end{bmatrix}$$

and this matrix is not positive definite, hence cannot be a valid covariance matrix. To solve this, Zou et al. (2017) imposed the additional positive definite constraint to the OLS estimator and proposed a numerical algorithm to compute this constrained OLS estimator. The constrained OLS has the same asymptotic distribution as the OLS estimator. For the details, see Zou et al. (2017).

3.2.1.2 Maximum Likelihood Estimator (MLE)

Under the normal assumption, Anderson (1973) proposed the maximum likelihood estimator of $\boldsymbol{\sigma}$ which tries to maximize the log-likelihood as below:

$$\ell(\boldsymbol{\Sigma}) = -\log\det\boldsymbol{\Sigma} - \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) \tag{3.4}$$

where $\boldsymbol{S}$ is the $p \times p$ sample covariance matrix. By differentiating the objective function $\ell(\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\sigma}$, Anderson (1973) tried to find the solution for the score equation for $k = 1, \ldots, K$:

$$\text{trace}\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_k\} = \text{trace}\{\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\}. \tag{3.5}$$

Using the equation (3.1), the score equation (3.5) for all $k = 1, \ldots, K$ can be written as, for $k = 1, \ldots, K$ and $\ell = 1, \ldots, K$,

$$[\text{trace}\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_\ell\}]_{k\ell}\boldsymbol{\beta} = [\text{trace}\{\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\}]_k \tag{3.6}$$

where $[\text{trace}\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_\ell\}]_{k\ell}$ is a $K \times K$ matrix and $[\text{trace}\{\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\}]_k$ is a K-dimensional column vector. Anderson (1973) proposed an iterative scheme which updates $\widehat{\boldsymbol{\sigma}}$ by $[\text{trace}\{(\boldsymbol{\Sigma}^{(i-1)})^{-1}\boldsymbol{G}_k(\boldsymbol{\Sigma}^{(i-1)})^{-1}\boldsymbol{G}_\ell\}]_{k\ell}\widehat{\boldsymbol{\sigma}} = [\text{trace}\{\boldsymbol{S}(\boldsymbol{\Sigma}^{(i-1)})^{-1}\boldsymbol{G}_k(\boldsymbol{\Sigma}^{(i-1)})^{-1}\}]_k$ and sets $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}})$. Anderson (1973) suggested the OLS estimator as the starting point of the iteration. However, neither the OLS estimator nor the following subsequent estimators through the iteration may not be positive definite. Also, the convergence of this algorithm is not guaranteed since the likelihood may decrease through the iteration (Drton and Richardson, 2002).

For the problem of maximum likelihood estimation with zero constraints, Chaudhuri et al. (2007) proposed the iterative conditional fitting algorithm which always converges and gives a positive definite solution. The details of the iterative conditional fitting algorithm will be discussed in Section 3.3.1. Note that the iterative conditional fitting algorithm is applicable only for the covariance matrix estimation with zero constraints and, unlike Anderson (1973)'s algorithm, it cannot be applied to estimate $\boldsymbol{\sigma}$ of model (3.1) in general.

Note that the log-likelihood function $\ell(\boldsymbol{\Sigma})$ takes the form of the sum of a convex function and a concave function (Bien and Tibshirani, 2011). Without the linear constraint (3.1) on the covariance matrix, the log-likelihood function (3.4) is uniquely maximized when $\boldsymbol{\Sigma} = \boldsymbol{S}$ (Watson, 1963; Zwiernik et al., 2017). However, with the linear constraint (3.1), maximization of the log-likelihood (3.4) is not a convex optimization problem and may have multiple solutions of local maxima (Chaudhuri et al., 2007). Hence, a solution of the score equation (3.5) computed from Anderson (1973)'s algorithm or Chaudhuri et al. (2007)'s iterative conditional fitting algorithm may be just one of the multiple local maxima and may not be the MLE.

3.2.1.3 Feasible Generalized Least Squares Estimator (FGLS)

Since the computation of the MLE requires iteration of a numerical algorithm such as Anderson (1973)'s algorithm, Zou et al. (2017) proposed the feasible generalized least squares estimator, $\hat{\boldsymbol{\sigma}}^{FGLS}$, for improving computational efficiency as below:

$$\widehat{\boldsymbol{\sigma}}^{FGLS} = \text{argmin}_{\boldsymbol{\sigma}} \text{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\sigma}))^T (\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}) \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})) \text{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\sigma})).$$

where $\boldsymbol{S}$ is the $p \times p$ sample covariance matrix and vec$(\cdot)$ converts a matrix to a vector by stacking columns of the matrix. The estimating equations can be written as, for $k = 1, \ldots, K$ and $\ell = 1, \ldots, K$,

$$[\text{trace}\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1}\boldsymbol{G}_\ell\}]_{k\ell}\boldsymbol{\sigma} = [\text{trace}\{\boldsymbol{S}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1}\}]_k \qquad (3.7)$$

where $[\text{trace}\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1}\boldsymbol{G}_k\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1}\boldsymbol{G}_\ell\}]_{k\ell}$ is a $K \times K$ matrix. Although the equation (3.6) and (3.7) look similar to each other, the equation (3.7) has a closed form solution and can be solved without numerical iterations.

As in the OLS estimator, one problem of the FGLS estimator is that it may not lead to a positive definite matrix. To solve this, Zou et al. (2017) imposed the additional positive definite constraint to the FGLS estimator and proposed a numerical algorithm to compute this constrained FGLS estimator. The constrained FGLS has the same asymptotic distribution as the FGLS estimator. For the details, see Zou et al. (2017).

3.2.2 Interpretation with the Linear Regression Framework

Given n samples of $(Y_1, \ldots, Y_p)$ which has zero mean vector and unknown covariance matrix $\boldsymbol{\Sigma}$ and given the $n \times p$ matrix $\boldsymbol{Y}$ for samples of size n, let $\boldsymbol{S} = \boldsymbol{Y}^T\boldsymbol{Y}/(n-1)$ be the unbiased estimator of $\boldsymbol{\Sigma}$, that is, the sample covariance matrix. For the model (3.1), define a $p(p+1)/2 \times K$ matrix $\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_K]$ such that $\boldsymbol{g}_k = \text{vech}(\boldsymbol{G}_k)$ for $k = 1, \ldots, K$ where vech$(\cdot)$ converts a symmetric matrix to a vector by stacking columns of the lower diagonal entries of the matrix. For

example, for a 2×2 matrix $\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$, $\text{vech}(\boldsymbol{S}) = (s_{11}, s_{21}, s_{22})^T$. Then we can rewrite the linear covariance model (3.1) as below:

$$\text{vech}(\boldsymbol{\Sigma}) = \boldsymbol{G}\boldsymbol{\sigma}.$$

Since the sample covariance matrix $\boldsymbol{S}$ contains some error in each entry of the matrix, it is explained by the model with error terms as below:

$$\text{vech}(\boldsymbol{S}) = \boldsymbol{G}\boldsymbol{\sigma} + \boldsymbol{e} \tag{3.8}$$

where $\boldsymbol{e}$ is a $p(p+1)/2$-dimensional vector of the errors. Note that the model (3.8) takes the same form as the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{3.9}$$

where $\boldsymbol{y}$ is the n-dimensional vector of response variable, $\mathbf{X}$ is the $n \times p$ design matrix and $\boldsymbol{e}$ is a n-dimensional vector of the errors.

Also, define a $p^2 \times K$ matrix $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K]$ such that $\boldsymbol{q}_k = \text{vec}(\boldsymbol{G}_k)$ for $k = 1, \ldots, K$ where $\text{vec}(\cdot)$ converts a matrix to a vector by stacking columns of the matrix. Then, the estimating equations for the OLS, FGLS and MLE can be expressed in matrix multiplication form as below:

OLS (eq.(3.3)): $\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{\sigma} = \boldsymbol{Q}^T\text{vec}(\boldsymbol{S})$;

FGLS (eq.(3.7)): $\boldsymbol{Q}^T(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1} \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1})\boldsymbol{Q}\boldsymbol{\sigma} = \boldsymbol{Q}^T(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1} \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1})\text{vec}(\boldsymbol{S})$;

MLE (eq.(3.6)): $\boldsymbol{Q}^T(\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \otimes \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1})\boldsymbol{Q}\boldsymbol{\sigma} = \boldsymbol{Q}^T(\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \otimes \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1})\text{vec}(\boldsymbol{S})$.

In this section, we will discuss how these estimating equations can be interpreted as estimating $\boldsymbol{\sigma}$ in model (3.8) with different assumptions on $\boldsymbol{e}$.

3.2.2.1 OLS Estimator for the Gauss-Markov Model

In the linear regression model (3.9), the OLS estimator of $\boldsymbol{\beta}$ is defined as the solution of the normal equation $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\boldsymbol{y}$ and it is the best linear unbiased estimator (BLUE) when the model is correct and the errors are assumed to be independent and identically distributed (i.i.d). That is, the OLS estimator is the BLUE for the Gauss-Markov model which is defined as the model (3.9) with the assumptions that $E(\boldsymbol{e}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$.

Likewise, we can consider the model (3.8) with the assumptions as below:

$$E(\boldsymbol{e}) = \mathbf{0}; \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_{\frac{p(p+1)}{2}\times\frac{p(p+1)}{2}}. \tag{3.10}$$

These assumptions are the exact Gauss-Markov model assumptions for the model (3.8). Hence, the solution to the normal equation for model (3.8)

$$\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{\sigma} = \boldsymbol{G}^T\mathrm{vech}(\boldsymbol{S})$$

is the BLUE for this model. However, this normal equation is different from the equation (3.3) for the OLS estimator. Hence, the solution of this equation is not equal to $\widehat{\boldsymbol{\sigma}}^{OLS}$ in general since it minimizes $||\mathrm{vech}(\boldsymbol{S} - \boldsymbol{\Sigma})||_F^2$ whereas $\widehat{\boldsymbol{\sigma}}^{OLS}$ minimizes $||\mathrm{vec}(\boldsymbol{S} - \boldsymbol{\Sigma})||_F^2$.

Remark 2. *For the linear covariance model with zero constraints, minimization of $||vech(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\sigma}))||_2$ gives the same solution as the minimization of $||vec(\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{\sigma}))||_2$. For the linear covariance model (3.1) in general, they are not equivalent.*

Next, we consider a slight modification to the above Gauss-Markov assumption (3.10) as below:

$$E(\boldsymbol{e}) = \mathbf{0}; \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2(\boldsymbol{D}_p^T\boldsymbol{D}_p)^{-1} = \sigma^2\boldsymbol{D}_p^+\boldsymbol{D}_p^{+T} \tag{3.11}$$

where $\boldsymbol{D}_p$ is a duplication matrix (page 299 of Abadir and Magnus (2005)) which transforms

vech($\boldsymbol{A}$) into vec($\boldsymbol{A}$) for a $p \times p$ symmetric matrix $\boldsymbol{A}$ by $\boldsymbol{D}_p$vech($\boldsymbol{A}$) = vec($\boldsymbol{A}$) and $\boldsymbol{D}_p^+ = (\boldsymbol{D}_p^T\boldsymbol{D}_p)^{-1}\boldsymbol{D}_p^T$ is the Moore-Penrose inverse of $\boldsymbol{D}_p$ (page 317 of Abadir and Magnus (2005)). Denoting $\mathbf{W} = (\boldsymbol{D}_p^T\boldsymbol{D}_p)^{-1}$, the identity matrix $\boldsymbol{I}$ in assumption (3.10) is replaced by $\mathbf{W}$ in assumption (3.11). Note that $\mathbf{W}$ is a $p(p+1)/2 \times p(p+1)/2$ diagonal matrix with diagonal entries 1 (n times) and 0.5 ($n(n-1)/2$ times) (page 314 of Abadir and Magnus (2005)).

With the assumption (3.11), the weighted least sqaures estimator of $\boldsymbol{\beta}$ for model (3.8) can be considered as below:

$$\begin{aligned}
\widehat{\boldsymbol{\sigma}} &= (\boldsymbol{G}^T\mathbf{W}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^T\mathbf{W}^{-1}\text{vech}(\boldsymbol{S}) \\
&= (\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\text{vech}(\boldsymbol{S}) \\
&= (\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\text{vec}(\boldsymbol{S}) \\
&= \widehat{\boldsymbol{\sigma}}^{OLS}.
\end{aligned}$$

Hence, the OLS estimator $\widehat{\boldsymbol{\sigma}}^{OLS}$ is the weigheted least squares estimator and, under the assumption (3.11), it is the BLUE for $\boldsymbol{\sigma}$ (page 83 of Monahan (2008)). The variance of the weighted least squares estimator can be obtained from the diagonal elements of the matrix below (page 165 of Rencher and Schaalje (2008)):

$$\text{Cov}(\widehat{\boldsymbol{\sigma}}^{OLS}) = \sigma^2(\boldsymbol{G}^T\mathbf{W}^{-1}\boldsymbol{G})^{-1} = \sigma^2(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}.$$

3.2.2.2 The Aitken Model with Fixed Error Variance

Now, we add normality assumption for $(Y_1, \ldots, Y_p)$, that is, $(Y_1, Y_2, \ldots, Y_p)^T \sim \mathbb{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. With the normality assumption, it is known that $(n-1)\boldsymbol{S}$ follows the Wishart distribution with $(n-1)$ degress of freedom (Johnson et al., 2002), that is, $(n-1)\boldsymbol{S} \sim W_p(n-1, \boldsymbol{\Sigma})$ or, equivalently, $\boldsymbol{S} \sim W_p(n-1, (n-1)^{-1}\boldsymbol{\Sigma})$, which has mean $\boldsymbol{\Sigma}$. For a Wishart matrix $\mathbf{Z} \sim W_p(n-1, \mathbf{V})$, the covariance of vech($\mathbf{Z}$) is equal to $2(n-1)\boldsymbol{D}_p^+(\mathbf{V} \otimes \mathbf{V})\boldsymbol{D}_p^{+T}$ (page 317 of Abadir and Magnus

(2005)). Hence, in model (3.8),

$$\mathrm{Cov}(\boldsymbol{e}) = \mathrm{Cov}(\mathrm{vech}(\boldsymbol{S})) = \frac{2}{n-1}\boldsymbol{D}_p^{+}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{D}_p^{+T}.$$

Note that, when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ (hence, $\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}$ is an identity matrix), $\mathrm{Cov}(\boldsymbol{e})$ takes the same form as in the assumption (3.11) with σ^2 replaced by $2/(n-1)$.

In estimating $\widehat{\boldsymbol{\sigma}}^{OLS}$, it is assumed that $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{D}_p^{+}\boldsymbol{D}_p^{+T}$. Hence, unless $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\widehat{\boldsymbol{\sigma}}^{OLS}$ is based on the mis-specified $\mathrm{Cov}(\boldsymbol{e})$. When the error variance is mis-specified, the OLS estimator is still unbiased but may no longer be the BLUE estimator (Monahan, 2008; Rencher and Schaalje, 2008). In this case, the variance of the OLS estimator can be obtained from the diagonal elements of the matrix computed as below:

$$\begin{aligned}
\mathrm{Cov}(\widehat{\boldsymbol{\sigma}}^{OLS}) &= (\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\mathrm{Cov}(\mathrm{vech}(\boldsymbol{S}))\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\\
&= \frac{2}{n-1}(\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{D}_p^{+}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{D}_p^{+T}\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\\
&= \frac{2}{n-1}(\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{D}_p^T(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{D}_p\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{D}_p^T\boldsymbol{D}_p\boldsymbol{G})^{-1}\\
&= \frac{2}{n-1}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}.
\end{aligned}$$

The structure of $\mathrm{Cov}(\boldsymbol{e})$ for the normal random vector raises the need for extending the Gauss-Markov model to incorporate various structures of the error variance. One may consider the Aitken model for the linear regression. That is, we can define the Aitken model for the linear covariance model to be the model (3.8) with the assumptions as below:

$$E(\boldsymbol{e}) = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{e}) = \mathbf{V}.$$

where $\mathbf{V}$ is a known $p(p+1)/2 \times p(p+1)/2$ positive definite matrix. For this model, we can construct the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$ as the solution of the equation below:

$$\boldsymbol{G}^T\mathbf{V}^{-1}\boldsymbol{G}\boldsymbol{\sigma} = \boldsymbol{G}^T\mathbf{V}^{-1}\mathrm{vech}(\boldsymbol{S}) \tag{3.12}$$

and the GLS estimator can be shown to be the BLUE (page 83 of Monahan (2008)).

However, the Aitken model cannot be used for the covariance estimation of the normally distributed random vectors since $\mathrm{Cov}(\boldsymbol{e})$ is unknown and contains $\boldsymbol{\Sigma}$ which needs to be estimated. One possible approach to address the unknown error variance is to fix $\mathrm{Cov}(\boldsymbol{e})$ with any estimator of $\boldsymbol{\Sigma}$. For example, we can compute the feasible generalized least squares estimator $\widehat{\boldsymbol{\sigma}}^{FGLS}$ by assuming that the error variance can be obtained by the OLS estimator of $\boldsymbol{\sigma}$ as below:

$$E(\boldsymbol{e}) = \mathbf{0}; \quad \mathrm{Cov}(\boldsymbol{e}) = \frac{2}{n-1}\boldsymbol{D}_p^{+}(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}) \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}))\boldsymbol{D}_p^{+T}. \tag{3.13}$$

With this assumption, the equation (3.12) can be rewritten as

$$\boldsymbol{Q}^T(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1} \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1})\boldsymbol{Q}\boldsymbol{\sigma} = \boldsymbol{Q}^T(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1} \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1})\mathrm{vec}(\boldsymbol{S}).$$

since $\{\boldsymbol{D}_p^{+}(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}) \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS}))\boldsymbol{D}_p^{+T}\}^{-1} = \boldsymbol{D}_p^T(\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1} \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})^{-1})\boldsymbol{D}_p$ and $\boldsymbol{D}_p\boldsymbol{G} = \boldsymbol{Q}$ (Abadir and Magnus, 2005). Note that this equation is equivalent to equation (3.7) by the relation $\mathrm{trace}(\boldsymbol{ABCD}) = \mathrm{vec}(\boldsymbol{D})^T(\boldsymbol{A} \otimes \boldsymbol{C}^T)\mathrm{vec}(\boldsymbol{B}^T)$ (page 283 of Abadir and Magnus (2005)).

The problem with the FGLS estimator is that $\mathrm{Cov}(\boldsymbol{e})$ is fixed based on $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{OLS})$ whereas the final estimator for the $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}^{FGLS})$. Hence, the FGLS estimator still has the problem of misspecification of the error variance structure. However, the FGLS estimator is an asymptotically efficient estimator (Anderson, 1973; Zou et al., 2017).

3.2.2.3 Estimating the Unknown Error Variance using MLE Approach

Given n samples of a random vector $(Y_1, Y_2, \ldots, Y_p)^T \sim \mathbb{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and if we do not make any additional assumption on the unknown error variance in model (3.8), a covariance model for estimating $\boldsymbol{\Sigma}$ can be constructed as the model (3.8) with the assumptions as below:

$$E(\boldsymbol{e}) = \mathbf{0}; \quad \mathrm{Cov}(\boldsymbol{e}) = \frac{2}{n-1}\boldsymbol{D}_p^{+}(\boldsymbol{\Sigma}(\boldsymbol{\sigma}) \otimes \boldsymbol{\Sigma}(\boldsymbol{\sigma}))\boldsymbol{D}_p^{+T}. \tag{3.14}$$

The above assumption on $\mathrm{Cov}(\boldsymbol{e})$ is actually redundant since it is implied by the normality of $(Y_1, Y_2, \ldots, Y_p)$. However, we show the structure of $\mathrm{Cov}(\boldsymbol{e})$ above for comparison with the Gauss-Markov model and the Aitken model. Note that the parameter vector $\boldsymbol{\sigma}$ appears not only as the regression coefficient vector but also determines the covariance structure of the error vector. Hence, although the model (3.1) takes the form of a linear model in terms of $\boldsymbol{\sigma}$, it is strictly not a linear model of $\boldsymbol{\sigma}$ since the errors of the data also depend on $\boldsymbol{\sigma}$.

However, the estimating equation for this model can still be derived similarly as the Aitken model by $\boldsymbol{G}^T\{\mathrm{Cov}(\boldsymbol{e})\}^{-1}\boldsymbol{G}\boldsymbol{\sigma} = \boldsymbol{G}^T\{\mathrm{Cov}(\boldsymbol{e})\}^{-1}\mathrm{vech}(\boldsymbol{S})$ which leads to:

$$\boldsymbol{Q}^T(\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \otimes \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1})\boldsymbol{Q}\boldsymbol{\sigma} = \boldsymbol{Q}^T(\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \otimes \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1})\mathrm{vec}(\boldsymbol{S}) \tag{3.15}$$

which is equivalent to equation (3.6) for estimating the MLE. Hence, the model (3.8) with the assumptions (3.14) provides an alternative approach for obtaining the estimating equation for MLE. Note that, in Section 3.2.1.2, we obtained equation (3.6) by differentiating the log-likelihood. Also, since the MLE is known to be asymptotically efficient (Anderson, 1973; Zou et al., 2017), the solution of the equation (3.15) is an asymptotically efficient estimator of $\boldsymbol{\sigma}$.

Since the estimating equation (3.15) is equivalent to the estimating equation for MLE, it can be solved by the methods for computing the MLE such as Anderson (1973)'s algorithm and, for the special problem of covariance estimation with zero constraints, the iterative conditional fitting algorithm by Chaudhuri et al. (2007). Anderson (1973)'s algorithm is indeed similar to the procedure called the estimated generalized least squares (EGLS, page 84 of Monahan (2008)) which computes the GLS estimator for the Aitken model with unknown error variance by solving the GLS estimating equation (3.12) iteratively.

3.3 An Asymptotically Efficient Estimator of a Covariance Matrix with Zero Entries

3.3.1 Iterative conditional fitting for Gaussian models

Consider n observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ of a random vector $(Y_1, \ldots, Y_p)^T \sim \mathbb{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $n > p$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^p$ is a $p \times p$ positive definite matrix. When some entries of $\boldsymbol{\Sigma}$ are known to be

zero, let $\boldsymbol{\sigma}$ be the vector of parameters for the non-zero entries in $\boldsymbol{\Sigma}$ and denote $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\sigma})$. The log-likelihood function is

$$\ell(\boldsymbol{\sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\sigma})| - \frac{n}{2}\text{tr}[\boldsymbol{S}\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})\}^{-1}] \tag{3.16}$$

where $\boldsymbol{S} = n^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i\boldsymbol{y}_i^T$ is the sample covariance matrix, and the maximum likelihood estimator of $\boldsymbol{\sigma}$ is consistent and asymptotically efficient.

Iterative conditional fitting (Chaudhuri et al., 2007) estimates the non-zero entries in a covariance matrix by solving the normal likelihood equation (3.16). Given the location of the zero entries in $\boldsymbol{\Sigma}$, the algorithm starts from a positive definite matrix with zero values in those entries (e.g. the identity matrix always meets such constraints) and updates the non-zero entries of the jth column of the matrix for $j = 1, \ldots, p$, iteratively until convergence.

The update of the jth column is conducted by the maximization of a conditional likelihood function for Y_j given the probability distribution of $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)^T$. Let σ_{jj} be the variance of Y_j and $\boldsymbol{\Sigma}_{-j,j}$ be the vector of covariances between Y_j and Y_{-j}. If the joint distribution of Y_{-j} is fixed with a known covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{-j,-j}$, the conditional likelihood function is:

$$L(\sigma_{jj}, \boldsymbol{\Sigma}_{-j,j} \mid \widetilde{\boldsymbol{\Sigma}}_{-j,-j}) = \prod_{i=1}^{n}(2\pi\tau_j)^{-\frac{1}{2}}e^{-\frac{(y_{ij} - y_{i,-j}^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\boldsymbol{\Sigma}_{-j,j})^2}{2\tau_j}} \tag{3.17}$$

where $\tau_j = \sigma_{jj} - \boldsymbol{\Sigma}_{-j,j}^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\boldsymbol{\Sigma}_{-j,j}$ and y_{ij} and $y_{i,-j}$ are the ith observation of Y_j and Y_{-j}, respectively. Since the location of the zero entries in $\boldsymbol{\Sigma}_{-j,j}$ is known, only the non-zero entries in $\boldsymbol{\Sigma}_{-j,j}$ and σ_{jj} are estimated via the maximization of (3.17). Such constrained maximization can be circumvented using the so-called 'pseudo-variables' which convert the constrained regression to the standard regression so that the usual least squares technique can be used. The details of the iterative conditional fitting algorithm is described in Algorithm 2.

Chaudhuri et al. (2007) proved convergence of iterative conditional fitting to one of positive definite local solutions. However, due to multiple local solutions (Chaudhuri et al., 2007; Zwiernik

et al., 2017) of the normal likelihood (3.16), different starting values for the iterative conditional fitting algorithm may lead to different local solutions. For example, Chaudhuri et al. (2007) suggested the identity matrix as one choice for the starting value. Since the asymptotic distribution of a local solution may not be equal to that of the maximum likelihood estimator, consistency and asymptotic efficiency of a solution from iterative conditional fitting is unclear.

3.3.2 Asymptotic efficiency of iterative conditional fitting

The theorem below discusses consistency and asymptotic efficiency of the non-zero entries in a covariance matrix computed from iterative conditional fitting.

Theorem 1. *Let $\widehat{\boldsymbol{\sigma}}$ be a solution computed from iterative conditional fitting with a consistent estimator of $\boldsymbol{\Sigma}$ as the starting value. Then, as $n \to \infty$,*

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \to \mathbb{N}(\mathbf{0}, I(\boldsymbol{\sigma})^{-1})$$

where $I(\boldsymbol{\sigma})$ is the Fisher information matrix.

Remark 3. *The Fisher information matrix $I(\boldsymbol{\sigma})$ is derived as negated expectation of the Hessian matrix (Chaudhuri et al., 2007). Using the notation $\boldsymbol{Q}$ in Section 3.2.2, $I(\boldsymbol{\sigma})$ can be written as*

$$I(\boldsymbol{\sigma}) = -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\sigma}^2}\right) = \frac{n}{2}\boldsymbol{Q}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{Q}.$$

Remark 4. *The solution from iterative conditional fitting is a matrix with zero entries whereas $\widehat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}$ are vectors of non-zero estimators and parameters in the matrix, respectively. By $\widehat{\boldsymbol{\sigma}}$ computed from iterative conditional fitting, we mean the vector of non-zero entries in the lower (or upper) triangular part of the solution matrix.*

Remark 5. *By a solution computed from iterative conditional fitting in Theorem 1, we mean the result of applying the iterative conditional fitting algorithm with any finite number of iterations.*

Theorem 1 states that we can find a consistent and asymptotically efficient solution by starting the algorithm from a consistent estimator of $\boldsymbol{\Sigma}$. In addition, the proof of Theorem 1 in Appendix

Algorithm 2 Iterative Conditional Fitting (Chaudhuri et al., 2007)

Consider a random vector $(Y_1, \ldots, Y_p)^T \sim \mathbb{N}_p(\mathbf{0}, \mathbf{\Sigma})$ and construct a $n \times p$ matrix $\boldsymbol{Y} = (y_{ij})_{i=1,j=1}^{n,p}$ for n observations of the random vector. Let $\boldsymbol{Y}^{(j)}$ and $\boldsymbol{Y}^{(-j)}$ denote the columns in Y for Y_j and $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)^T$, respectively. Also, let $sp(j)$ be the set of indices for variables that are marginally dependent with Y_j.

1. Set an initial estimator $\widehat{\mathbf{\Sigma}}^{(0)}$ from the space of positive definite matrices with the zero constraint (e.g. an identity matrix) and set $r = 0$.

2. With any $p \times p$ matrix M, we will use $M_{-j,-j}$ to denote a $(p-1) \times (p-1)$ partitioned matrix of M without the jth column and row. Set $\widehat{\mathbf{\Sigma}}^{(r,0)} = \widehat{\mathbf{\Sigma}}^{(r)}$ and repeat the following updates for $j = 1, \ldots, p$:

   - Set $\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j} = \widehat{\mathbf{\Sigma}}^{(r,j-1)}_{-j,-j}$ and construct a matrix of 'pseudo-variables' as below:

   $$Z^j = \boldsymbol{Y}^{(-j)}(\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j})^{-1}$$

   and let $\mathbf{Z}^j_{sp(j)}$ denotes the matrix of $sp(j)$th columns of Z^j.

   - Update the off-diagonal non-zero elements of the j-th column of $\widehat{\mathbf{\Sigma}}^{(r,j)}$ by

   $$\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j} = \{(\mathbf{Z}^j_{sp(j)})^T(\mathbf{Z}^j_{sp(j)})/n\}^{-1}(\mathbf{Z}^j_{sp(j)})^T\boldsymbol{Y}^{(j)}/n \tag{3.18}$$

   and set the other entries of the j-th column to zero. Then, update the j-th row of $\widehat{\mathbf{\Sigma}}^{(r,j)}$ as the transpose of the updated j-th column of $\widehat{\mathbf{\Sigma}}^{(r,j)}$.

   - Update the j-th diagonal element of $\widehat{\mathbf{\Sigma}}^{(r,j)}$ by

   $$\begin{aligned}\hat{\sigma}_{jj} = &(\boldsymbol{Y}^{(j)} - \mathbf{Z}^j_{sp(j)}\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j})^T(\boldsymbol{Y}^{(j)} - \mathbf{Z}^j_{sp(j)}\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j})/n \\ &+ (\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j})^T(\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j})^{-1}_{sp(j),sp(j)}\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j}\end{aligned} \tag{3.19}$$

3. Set $\widehat{\mathbf{\Sigma}}^{(r+1)} = \widehat{\mathbf{\Sigma}}^{(r,p)}$

4. Iterate step 2 and 3 until a predetermined convergence criterion is met

B.1 does not require the starting value to include any zero entries. Hence, for the choice of the starting value, the sample covariance matrix $\boldsymbol{S}$ can always be used because it is a positive definite and consistent estimator of $\boldsymbol{\Sigma}$. Such starting values without any zero entries were not considered in Chaudhuri et al. (2007) because the convergence of iterative conditional fitting was shown for the parameter space of matrices with zero entries. However, after one cycle of the algorithm starting from any positive definite matrix, the resulting matrix lies within the parameter space by having zero entries as constrained. Hence, the algorithm still converges.

3.3.3 An algorithm for $p > n$ case

The iterative conditional fitting algorithm requires the sample covariance matrix $\boldsymbol{S}$ to be invertible. However, when the sample size n is smaller than the number of variables p, $\boldsymbol{S}$ is not invertible. Hence, the algorithm works only when $p < n$. To remedy this, we consider an objective function as below:

$$\ell^*(\boldsymbol{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\text{tr}\{(\boldsymbol{S} + \epsilon I_p)\boldsymbol{\Sigma}^{-1}\} \tag{3.20}$$

which replaces $\boldsymbol{S}$ in the normal log-likelihood by $\boldsymbol{S} + \epsilon I_p$ for some $\epsilon > 0$. This approach is similar to the optimization of the L_1-penalized likelihood (4.1) by Bien and Tibshirani (2011) when $p > n$.

First, we discuss maximization of the objective function (3.20) without any zero constraint on the entries in $\boldsymbol{\Sigma}$. If the joint distribution of Y_{-j} is fixed with a known covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{-j,-j}$, the next proposition states that the function (3.20) is maximized by the ridge regression. Note that the normal log-likelihood (3.16) is maximized by the least squares regression for β and σ_{jj} given the distribution of $Y_{-j} \sim \mathbb{N}_{p-1}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{-j,-j})$.

Proposition 1. *Suppose the distribution of* $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)^T$ *is* $\mathbb{N}_{p-1}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{-j,-j})$.

Then, $\ell^*(\boldsymbol{\Sigma})$ is maximized by $\widehat{\boldsymbol{\Sigma}}_{-j,j}$ and $\hat{\sigma}_{jj}$ which satisfy

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{-j,j} &= \mathit{argmin}_{\boldsymbol{\Sigma}_{-j,j}} \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - y_{i,-j}^T\beta)^2 + \epsilon||\beta||^2 \\
&= \{\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_{i,-j}\boldsymbol{y}_{i,-j}^T + \epsilon I\big)\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{y}_{i,-j}^T)y_{ij}\} \\
\hat{\sigma}_{jj} &= \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - y_{i,-j}^T\hat{\beta})^2 + \epsilon||\hat{\beta}||^2 + \epsilon + \widehat{\boldsymbol{\Sigma}}_{-j,j}^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\widehat{\boldsymbol{\Sigma}}_{-j,j}
\end{aligned}$$

where $\beta = \widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$ and $\hat{\beta} = \widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\widehat{\boldsymbol{\Sigma}}_{-j,j}$.

Next, we consider the constrained maximization of (3.20) when some entries in $\boldsymbol{\Sigma}$ are known to be zero. In the iterative conditional fitting algorithm, the constrained optimization with zero entries is circumvented by the least squares regression with pseudo variables which are constructed from the data for Y_{-j} and the fixed estimator for $\boldsymbol{\Sigma}_{-j,-j}$. Similarly, we use the data for Y_{-j} and the estimator for $\boldsymbol{\Sigma}_{-j,-j}$ but, based on Proposition 1, we replace the least squares regression with the ridge regression to maximize (3.20) instead of (3.16). Hence, we propose the *iterative conditional ridge* algorithm for the case of $p > n$ in Algorithm 3. The main difference between this algorithm and the iterative conditional fitting algorithm is that the least squares regression is replaced by the ridge regression in each iteration of the algorithm. As the iterative conditional fitting algorithm, this algorithm always converges to a positive definite matrix.

Let $\boldsymbol{\sigma}_\epsilon$ be the vector of parameters for the non-zero entries in $\boldsymbol{\Sigma} + \epsilon I_p$. The next corollary of Theorem 1 shows that the solution computed from iterative conditional ridge is an asymptotically efficient estimator of $\boldsymbol{\Sigma} + \epsilon I_p$. This solution contains bias to the diagonal entries but extent of the bias can be controlled under any positive constant ϵ. Also, it contains no additional bias to off-diagonal entries so can be used to obtain consistent and efficient estimators of the pairwise covariances.

Corollary 1. *Let $\widehat{\boldsymbol{\sigma}}$ be a solution computed from iterative conditional ridge with a consistent*

Algorithm 3 Iterative Conditional Ridge

For a random vector $(Y_1, \ldots, Y_p)^T \sim \mathbb{N}_p(\mathbf{0}, \mathbf{\Sigma})$, let $\boldsymbol{Y} = (y_{ij})_{i=1,j=1}^{n,p}$ denote a $n \times p$ matrix for n independent observations. Let $\boldsymbol{Y}^{(j)}$ and $\boldsymbol{Y}^{(-j)}$ denote the columns in Y for Y_j and $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)^T$, respectively. Also, let $sp(j)$ be the set of indices for variables that are marginally dependent with Y_j.

1. Set an initial estimator $\widehat{\mathbf{\Sigma}}^{(0)}$ from the space of positive definite matrices with the zero constraint (e.g. an identity matrix) and set $r = 0$.

2. With any $p \times p$ matrix M, we will use $M_{-j,-j}$ to denote a $(p-1) \times (p-1)$ partitioned matrix of M without the jth column and row. Set $\widehat{\mathbf{\Sigma}}^{(r,0)} = \widehat{\mathbf{\Sigma}}^{(r)}$ and repeat the following updates for $j = 1, \ldots, p$:

   - Set $\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j} = \widehat{\mathbf{\Sigma}}^{(r,j-1)}_{-j,-j}$ and let $P = (\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j})^{-1}_{sp(j)}$ denotes the matrix of $sp(j)$th columns of $(\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j})^{-1}$.
   - Update the off-diagonal non-zero elements of the j-th column of $\widehat{\mathbf{\Sigma}}^{(r,j)}$ by

     $$\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j} = \{P^T(\boldsymbol{Y}^{(-j)T}\boldsymbol{Y}^{(-j)}/n + \epsilon I_{p-1})P\}^{-1} P^T \boldsymbol{Y}^{(-j)T}\boldsymbol{Y}^{(j)}/n$$

     and set the other entries of the j-th column to zero. Then, update the j-th row of $\widehat{\mathbf{\Sigma}}^{(r,j)}$ as the transpose of the updated j-th column of $\widehat{\mathbf{\Sigma}}^{(r,j)}$.
   - Update the j-th diagonal element of $\widehat{\mathbf{\Sigma}}^{(r,j)}$ by

     $$\begin{aligned}\hat{\sigma}_{jj} &= (\boldsymbol{Y}^{(j)} - \boldsymbol{Y}^{(-j)} P \widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j})^T (\boldsymbol{Y}^{(j)} - \boldsymbol{Y}^{(-j)} P \widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j})/n \\ &\quad + \epsilon(1 + \widehat{\mathbf{\Sigma}}^{(r,j)T}_{sp(j),j} P^T P \widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j}) + (\widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j})^T (\widehat{\mathbf{\Sigma}}^{(r,j)}_{-j,-j})^{-1}_{sp(j),sp(j)} \widehat{\mathbf{\Sigma}}^{(r,j)}_{sp(j),j}\end{aligned}$$

3. Set $\widehat{\mathbf{\Sigma}}^{(r+1)} = \widehat{\mathbf{\Sigma}}^{(r,p)}$

4. Iterate step 2 and 3 until a predetermined convergence criterion is met

estimator of $\boldsymbol{\Sigma} + \epsilon I_p$ as the starting value. Then, as $n \to \infty$,

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}_\epsilon) \to \mathbb{N}(\mathbf{0}, I(\boldsymbol{\sigma}_\epsilon)^{-1})$$

where $I(\boldsymbol{\sigma}_\epsilon)$ is the Fisher information matrix.

Remark 6. *Using the notation $\boldsymbol{Q}$ in Section 3.2.2, $I(\boldsymbol{\sigma})$ can be written as*

$$I(\boldsymbol{\sigma}_\epsilon) = \frac{n}{2}\boldsymbol{Q}^T\{(\boldsymbol{\Sigma} + \epsilon I_p)^{-1} \otimes (\boldsymbol{\Sigma} + \epsilon I_p)^{-1}\}\boldsymbol{Q}.$$

Remark 7. *Corollary 1 assumes that the number of variables p is fixed. Asymptotic efficiency of the iterative conditional ridge algorithm when both n and p increase remains as an open question.*

Remark 8. *For fixed p, $\boldsymbol{S} + \epsilon I_p$ is a consitent estimator of $\boldsymbol{\Sigma} + \epsilon I_p$ and can be considered as the starting value for the iterative conditional ridge algorithm.*

3.4 Simulation Study

When the location of the zero entries in $\boldsymbol{\Sigma}$ is known, we showed that iterative conditional fitting and iterative conditional ridge estimate the non-zero entries with minimum variance asymptotically in Theorem 1 and Corollary 1, respectively. In this section, we check whether the variability of the non-zero entries is reduced by those algorithms with finite samples.

We generate 100 datasets from $\mathbb{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ where the correlation matrix $\boldsymbol{\Sigma}$ is determined by one of the following models.

- Moving average model: the moving average process of order one where $\rho_{ij} = 0.5$ if $|i-j| = 1$ and $\rho_{ij} = 0$ otherwise.
- Banded model: $\rho_{ij} = 0.8$ if $|i-j| = 1$, $\rho_{ij} = 0.6$ if $|i-j| = 2$, $\rho_{ij} = 0.4$ if $|i-j| = 3$, $\rho_{ij} = 0.2$ if $|i-j| = 4$ and $\rho_{ij} = 0$ otherwise.

The moving average model and the banded model are typical covariance models that have been used in many covariance estimation studies such as Rothman et al. (2009) and Qiu and Liyanage

(2019). Both models contain exact zero entries and the proportaion of the zero entries increases as the number of variables p increases. Each simulated dataset contains 25 samples $(n = 25)$ of the normal random vector of dimension 10 or 50 ($p = 10$ or $p = 50$). For $p = 10$, the non-zero entries are estimated by the iterative conditional fitting algorithm. For $p = 50$, the non-zero entries are estimated by the iterative conditional fitting ridge with $\epsilon = 0.01$.

In Figure 3.1, we compare the variability of sample covariances with the solution from iterative conditional fitting or iterative conditional ridge for the moving average model. In upper panels of Figure 3.1, the intervals between the 5th percentile and 95th percentile of the sample covariances are wider than those of the solutions from iterative conditional fitting. Also, for the off-diagonal entries whose true covariances are equal to 0.5, sample covariance take negative values in some simulated datasets whereas iterative conditional fitting estimates positive values in all simulated datasets. Similar pattern is observed for iterative conditional ridge when $p = 50$.

In Figure 3.2, the reduction of the variability is observed more clearly with narrower intervals between the 5th percentile and 95th percentile for the solution from iterative conditional fitting or iterative conditional ridge than the sample covariances. Particularly, the sample covariances for the entries where $\rho_{ij} = 0.2$ often take negative values whereas the solutions from iterative conditional fitting or iterative conditional ridge are mostly positive and concentrated at 0.2. The results in Figure 3.2 and Figure 3.2 suggest that we can estimate the non-zero entries in a covariance matrix with less variability than the sample covariances by using the iterative conditional fitting (or ridge) algorithm.

3.5 Implication of model selection

A fundamental assumption underlying the properties of the maximum likelihood estimator is that the model is correctly specified (White, 1982). Similarly, we assumed that we knew which entries of $\mathbf{\Sigma}$ are zero or non-zero ("correct model"). In practice, the location of the zero entries is often unknown and model selection procedures such as multiple testing (Drton et al., 2007) or thresholding (Bickel et al., 2008a; Rothman et al., 2009) are required to identify the location of the zero entries in the covariance matrix. However, the location of the zero entries selected from such

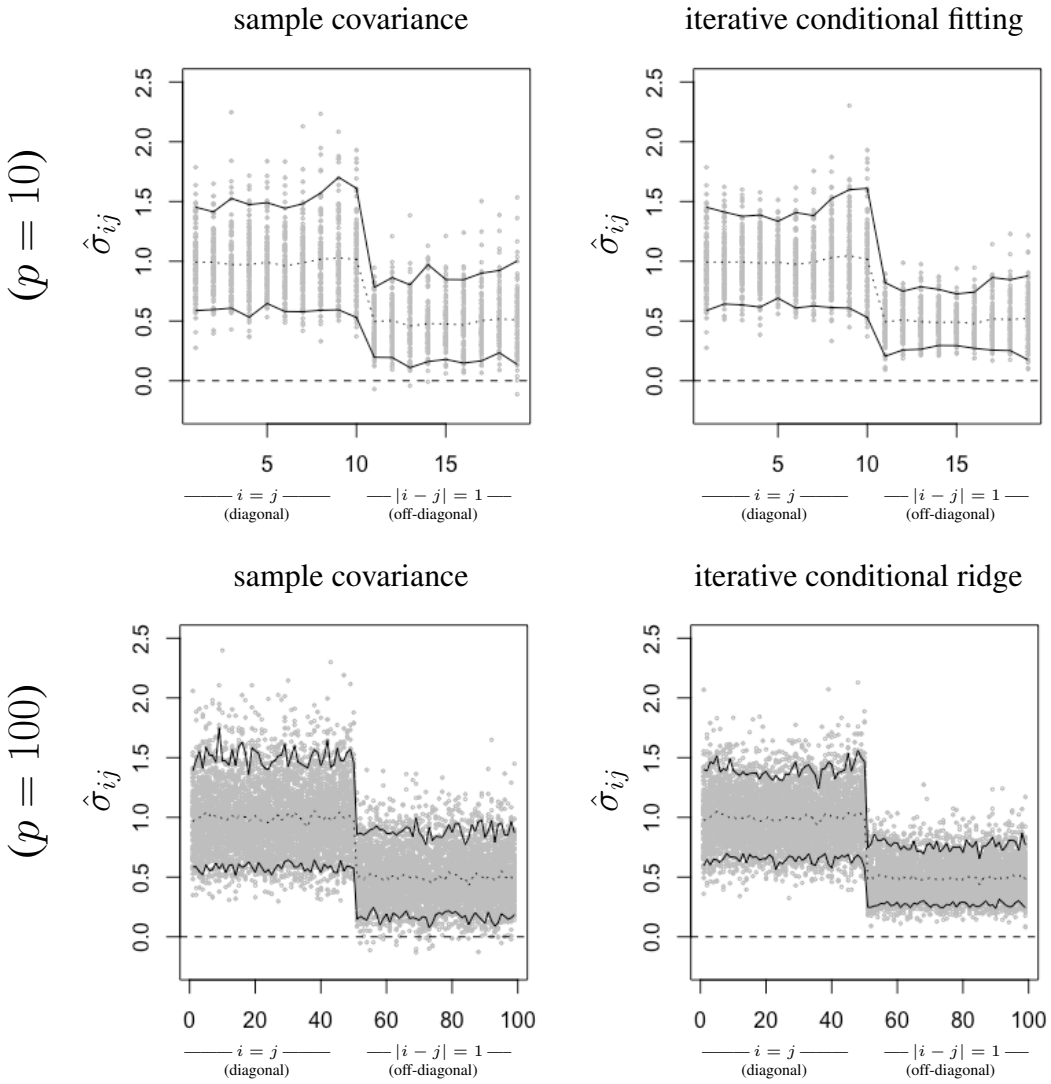

Figure 3.1: Estimates of non-zero parameters of the first-order moving average model with $n = 25$ and $p = 10$ or $p = 50$ are plotted with gray dots for 100 simulated datasets. Diagonal entries are indexed from 1 to p and the first upper off-diagonal entries are indexed from $p + 1$ to $2p - 1$. The x-axis indicates index of each non-zero parameter. Dotted curves represent the mean of 100 estimates for each parameter. The 95th percentile and 5th percentile are drawn by solid curves. For iterative conditional ridge, $\epsilon = 0.01$.

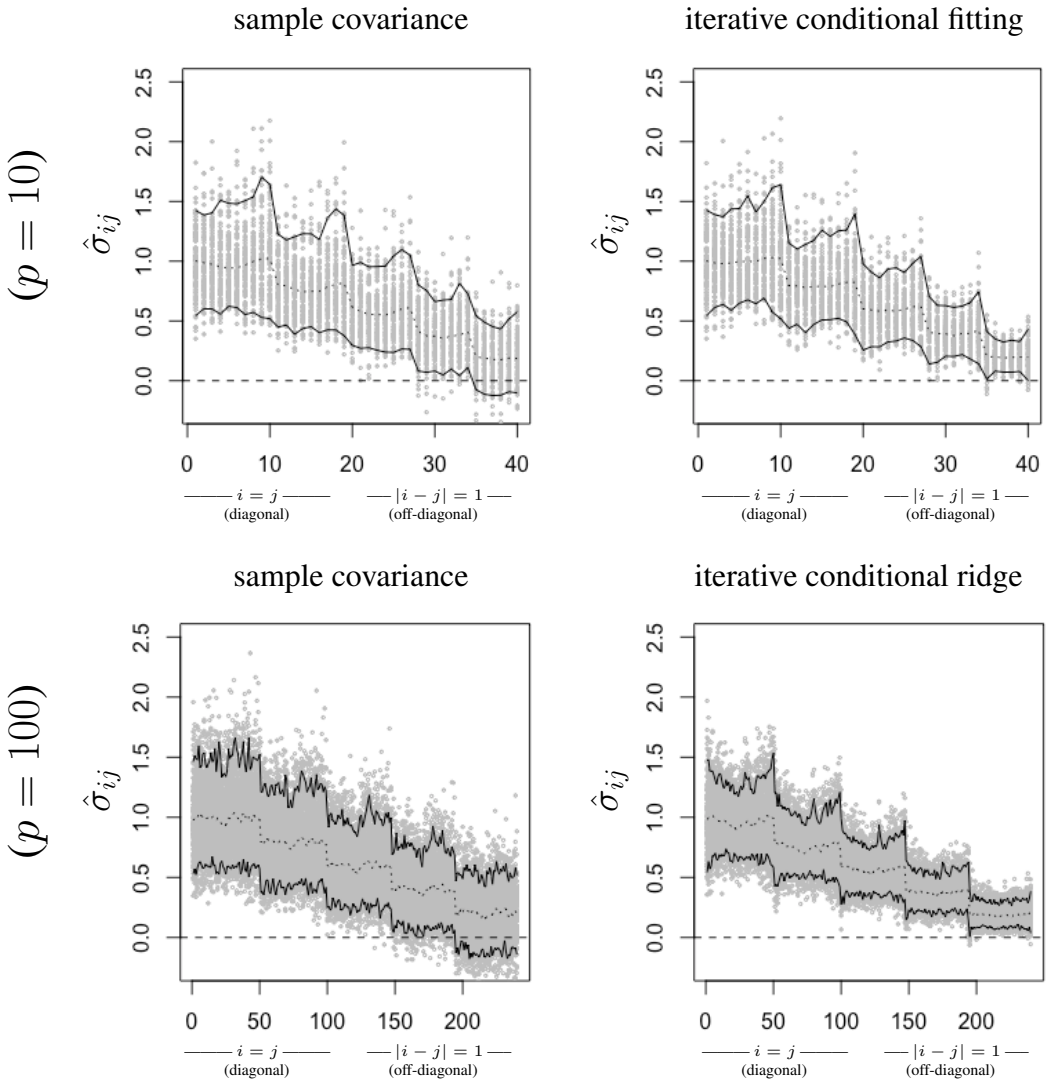

Figure 3.2: Estimates of non-zero parameters of the banded model with $n = 25$ and $p = 10$ or $p = 50$ are plotted with gray dots for 100 simulated datasets. Diagonal entries are indexed from 1 to p, the first upper off-diagonal entries from $p + 1$ to $2p - 1$, the second upper off-diagonal entries from $2p$ to $3p - 2$ and so on. The x-axis indicates index of each non-zero parameter. Dotted curves represent the mean of 100 estimates for each parameter. The 95th percentile and 5th percentile are drawn by solid curves. For iterative conditional ridge, $\epsilon = 0.01$.

procedures may not coincide ("incorrect model") with the true covariance matrix.

The properties of the maximum likelihood estimator under such incorrect models are well known for the linear regression. Specifically, an overfitted model will increase the variability of the estimator while an underfitted model may induce bias (page 76 of Monahan (2008)). White (1982) discussed the properties of the maximum likelihood estimator under a misspecified model, that is, an underfitted model. Such model misspecification has also been considered in the model selection and several information criteria have been proposed for generalized linear models (Lv and Liu, 2014), generalized linear mixed models (Yu et al., 2018) and for time series (Hsu et al., 2019).

To discuss the properties of the maximum likelihood estimator of non-zero entries in a covariance matrix under overfitting or underfitting, we consider three models. The correct model contains the same zero entries as the true covariance matrix and we denote its parameter vector as $\boldsymbol{\sigma}_C$. An overfitted model contains more non-zero parameters in addition to the parameters in the correct model. We donote its parameter vector as $\boldsymbol{\sigma}_O = (\boldsymbol{\sigma}_C, \boldsymbol{\sigma}_{O\setminus C})$ where $\boldsymbol{\sigma}_{O\setminus C}$ represents overfitted parameters. An underfitted model, also referred to as a misspecified model, is a reduced model with fewer parameters than the correct model. We donote its parameter vector as $\boldsymbol{\sigma}_U$ such that $\boldsymbol{\sigma}_C = (\boldsymbol{\sigma}_U, \boldsymbol{\sigma}_{C\setminus U})$ where $\boldsymbol{\sigma}_{C\setminus U}$ represents underfitted parameters. We assume that iterative conditional fitting starts from a consistent estimator of $\boldsymbol{\Sigma}$ as required in Theorem 1.

First, we compare the asymptotic variance of the estimators of the non-zero entries between the correct model and the overfitted model. Similar to Theorem 1, asymptotic normality of the solution computed from iterative conditional fitting for an overfitted model can be shown. That is, denoting $\widetilde{\boldsymbol{\sigma}}_O$ as the solution for the overfitted model, as $n \to \infty$,

$$n^{\frac{1}{2}}(\widetilde{\boldsymbol{\sigma}}_O - \boldsymbol{\sigma}_O) \to \mathbb{N}(\mathbf{0}, I(\boldsymbol{\sigma}_O)^{-1})$$

where $I(\boldsymbol{\sigma}_O)$ is the Fisher information matrix. We can compare the asymptotic variance of the non-zero entries between the correct model and the overfitted model by comparing the diagonal

entries of $I(\boldsymbol{\sigma}_C)^{-1}$ and $I(\boldsymbol{\sigma}_O)^{-1}$. Note that $I(\boldsymbol{\sigma}_C)^{-1}$ is equivalent to $I(\boldsymbol{\sigma})^{-1}$ in Theorem 1. As analogous to linear regression, Proposition 2 below states that overfitting will lead to increased variability in the estimation of the non-zero entries of $\boldsymbol{\Sigma}$.

Proposition 2. *Let $\widehat{\boldsymbol{\sigma}}_C$ and $\widetilde{\boldsymbol{\sigma}}_O = (\widetilde{\boldsymbol{\sigma}}_C, \widetilde{\boldsymbol{\sigma}}_{O\setminus C})$ be solutions computed from iterative conditional fitting for the correct model and an overfitted model, respectively. The standard error of each element in $\widehat{\boldsymbol{\sigma}}_C$ is less than or equal to the standard error of the corresponding element in $\widetilde{\boldsymbol{\sigma}}_C$ asymptotically.*

Proposition 2 implies that, if some zero entries are identified, iterative conditional fitting gives a more efficient estimator of $\boldsymbol{\Sigma}$ than the sample covariance matrix. This result is interesting because the covariance estimation of two normal random variables can be improved in efficiency by considering their covariance with other normal random variables. For a bivariate normal distribution of X and Y, the covariance parameters $\{\sigma_X^2, \sigma_Y^2, \sigma_{XY}\}$ are usually estimated by the sample covariance $\{s_X^2, s_Y^2, s_{XY}\}$. However, if there are other normal random variables, we can find a more efficient estimator of $\{\sigma_X^2, \sigma_Y^2, \sigma_{XY}\}$ by considering the whole covariance matrix of all those variables with zero constraints on some entries of the matrix.

Next, we discuss the properties of iterative conditional fitting for an underfitted model. For this, we define a matrix $\boldsymbol{Q}_C$ with entries of 0 or 1 that satisfies $\text{vec}(\boldsymbol{\Sigma}) = \boldsymbol{Q}_C\boldsymbol{\sigma}_C$ as defined in Chaudhuri et al. (2007) for the correct model. Here, $\text{vec}(\cdot)$ is the vectorization operator which stacks columns of a matrix to a vector. Then, we split columns of $\boldsymbol{Q}_C$ by $[\boldsymbol{Q}_U, \boldsymbol{Q}_{C\setminus U}]$ such that $\boldsymbol{Q}_C\boldsymbol{\sigma}_C = \boldsymbol{Q}_U\boldsymbol{\sigma}_U + \boldsymbol{Q}_{C\setminus U}\boldsymbol{\sigma}_{C\setminus U}$. Proposition 3 below discusses that underfitting may induce bias to the estimator of the non-zero entries of the covariance matrix.

Proposition 3. *Let $\widetilde{\boldsymbol{\sigma}}_U$ be a solution computed from iterative conditional fitting for an underfitted model. Then, as $n \to \infty$,*

$$\widetilde{\boldsymbol{\sigma}}_U \to \boldsymbol{\sigma}_U + (\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U)^{-1}\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_{C\setminus U}\boldsymbol{\sigma}_{C\setminus U}$$

where $\mathbf{W} = \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\sigma}}_U)^{-1} \otimes \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\sigma}}_U)^{-1}$.

In Proposition 3, if $\boldsymbol{\sigma}_{C\setminus U} = \mathbf{0}$, the model is not underfitted and the difference between $\widetilde{\boldsymbol{\sigma}}_U$ and $\boldsymbol{\sigma}_U$ is caused by the randomness of the data, that is, the difference between $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ which converges to zero as the sample size increases. If we multiply $\boldsymbol{Q}_U$ to the left side of the bias term $(\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U)^{-1}\boldsymbol{Q}_U^T\mathbf{W}$, it is the projection matrix of the generalized least squares. Hence, the bias depends on how $\boldsymbol{Q}_{C\setminus U}$ is related to the space of the fitted covariance matrix with the columns of $\boldsymbol{Q}_U$. If $\boldsymbol{Q}_{C\setminus U}$ is orthogonal to the space, $\boldsymbol{Q}_U(\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U)^{-1}\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_{C\setminus U}$ will be a zero matrix and there will be no bias due to underfitting even if $\boldsymbol{\sigma}_{C\setminus U}$ is not a zero vector. Hence, the bias induced by underfitting the model depends on:

- $\boldsymbol{\sigma}_{C\setminus U}$: the magnitude of the missed (or underfitted) components; and
- $\boldsymbol{Q}_U(\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U)^{-1}\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_{C\setminus U}$: how much of the missed components lie in the space of the fitted covariance matrix

Remark 9. *This result is analogue of underfitting the linear regression model that can be found at page 77 of Monahan (2008).*

3.5.1 Example: Estimation bias due to underfitting

We will see several examples of underfitted models below. In some examples, the underfitting does not lead to additional bias. In other examples, the estimated non-zero entries of the covariance matrix contain bias induced by the underfitting. For simplicity, in all examples below, it is assumed that the sample covariance matrix is equal to the true covariance matrix. This eliminates the effects of noise and the difference between $\widetilde{\boldsymbol{\sigma}}_U$ and $\boldsymbol{\sigma}_U$ depends only on the bias due to underfitting.

Example 1: Consider a sample covariance matrix

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

The true covariance matrix $\boldsymbol{\Sigma}$ is equal to $\boldsymbol{S}$ so the correct model for this covariance matrix is as

below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}) \quad = \sigma_1 \boldsymbol{G}_1 + \sigma_2 \boldsymbol{G}_2 + \sigma_3 \boldsymbol{G}_3$$

where

$$\boldsymbol{G}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{G}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{G}_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and the true parameters are $\boldsymbol{\sigma}^T = (\sigma_1, \sigma_2, \sigma_3) = (1, 1, 0.5)$. However, we will fit a model with the constraint that $\sigma_3 = 0$ so our model is an underfitted model as below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}_U) \quad = \sigma_1 \boldsymbol{G}_1 + \sigma_2 \boldsymbol{G}_2$$

where $\boldsymbol{\sigma}_U = (\sigma_1, \sigma_2)^T$. By solving equation (3.5) for this underfitted model with the iterative conditional fitting algorithm or Anderson (1973)'s algorithm, one can obtain the solution $\widehat{\boldsymbol{\sigma}}_U = (1, 1)^T$. Hence, in this example, the underfitting (missing $\boldsymbol{G}_3$ in the model) does not induce bias in the estimation of β_1 and β_2. One can check that $(\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C\setminus U} = (0, 0)^T$ so $\boldsymbol{Q}_U (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C\setminus U}$ gives a zero vector where $\boldsymbol{Q}_U = [\text{vec}(\boldsymbol{G}_1), \text{vec}(\boldsymbol{G}_2)]$, $\boldsymbol{Q}_{C\setminus U} = \text{vec}(\boldsymbol{G}_3)$ and $\mathbf{W} = I_2 \otimes I_2$.

Example 2: Consider a sample covariance matrix

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{bmatrix} (= \boldsymbol{\Sigma}).$$

The correct model for this covariance matrix is as below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}) \quad = \sigma_1 \boldsymbol{G}_1 + \sigma_2 \boldsymbol{G}_2 + \sigma_3 \boldsymbol{G}_3 + \sigma_4 \boldsymbol{G}_4 + \sigma_5 \boldsymbol{G}_5$$

where

$$\boldsymbol{G}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{G}_4 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

and the true parameters are $\boldsymbol{\sigma}^T = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (1, 1, 1, 0.5, 0.5)$. However, we will fit an underfitted model with $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_4$ as below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}_U) \quad = \sigma_1 \boldsymbol{G}_1 + \sigma_2 \boldsymbol{G}_2 + \sigma_3 \boldsymbol{G}_3 + \sigma_4 \boldsymbol{G}_4.$$

Solving equation (3.5) for this underfitted model gives the solution $\widehat{\boldsymbol{\sigma}}_U = (1, 1, 1, 0.5)^T$. Hence, in this example, the underfitting does not induce bias in the estimation of $\boldsymbol{\sigma}_U$. One can check that $(\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C\setminus U} = (0, 0, 0, 0)^T$ so $\boldsymbol{Q}_U (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C\setminus U}$ gives a zero vector where $\boldsymbol{Q}_U = [\text{vec}(\boldsymbol{G}_1), \text{vec}(\boldsymbol{G}_2), \text{vec}(\boldsymbol{G}_3), \text{vec}(\boldsymbol{G}_4)]$, $\boldsymbol{Q}_{C\setminus U} = \text{vec}(\boldsymbol{G}_5)$ and $\mathbf{W} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}_U) \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}_U)$.

Example 3: Consider a sample covariance matrix

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0.25 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0.25 & 0.5 & 1 \end{bmatrix} (= \boldsymbol{\Sigma}).$$

The correct model for this covariance matrix is as below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}) \quad = \sigma_1 \boldsymbol{G}_1 + \sigma_2 \boldsymbol{G}_2 + \sigma_3 \boldsymbol{G}_3 + \sigma_4 \boldsymbol{G}_4 + \sigma_5 \boldsymbol{G}_5 + \sigma_6 \boldsymbol{G}_6 + \sigma_7 \boldsymbol{G}_7 + \sigma_8 \boldsymbol{G}_8$$

where

$$\boldsymbol{G}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \boldsymbol{G}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{G}_5 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_6 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_7 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{G}_8 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

and the true parameters are $\boldsymbol{\sigma}^T = (\sigma_1, \ldots, \sigma_8) = (1, 1, 1, 1, 0.5, 0.5, 0.5, 0.25)$. However, we will fit an underfitted model with $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_7$. Solving equation (3.5) for this underfitted model gives the solution

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}_U) = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.41 & 0 \\ 0 & 0.41 & 0.9 & 0.36 \\ 0 & 0 & 0.36 & 1 \end{bmatrix}$$

Hence, in this example, the underfitting induces bias in the estimation of $\boldsymbol{\beta}^-$. One can check that $(\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C\setminus U} = (0, 0, -0.4, 0, 0, -0.36, -0.55)^T$ so the estimators of σ_3, σ_6 and σ_7 are biased where $\boldsymbol{Q}_U = [\text{vec}(\boldsymbol{G}_1), \ldots, \text{vec}(\boldsymbol{G}_7)]$, $\boldsymbol{Q}_{C\setminus U} = \text{vec}(\boldsymbol{G}_8)$ and $\mathbf{W} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}_U) \otimes \boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}_U)$.

Example 4: Consider a sample covariance matrix

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0.05 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0.05 & 0.5 & 1 \end{bmatrix} (= \boldsymbol{\Sigma}).$$

We fit the same underfitted model as in Example 3. Solving equation (3.5) for this underfitted model gives the solution

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}}_U) = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.48 & 0 \\ 0 & 0.48 & 0.97 & 0.47 \\ 0 & 0 & 0.47 & 1 \end{bmatrix}$$

Note that the same entries in the matrix are biased but the amount of the bias is less than Example 3. This is because the missed component $\boldsymbol{G}_8$ in Example 4 has less covariance ($\sigma_8 = 0.05$) than in Example 3 ($\sigma_8 = 0.25$).

3.5.2 Asymptotic variance of the MLE for an underfitted model

Proposition 2 tells us that eliminating spurious non-zero entries by imposing zero constraints reduces the standard error of the remaining non-zero entries of the MLE. However, if we eliminate some non-zero entries, it does not always reduce the standard error of the remaining non-zero entries. That is, Proposition 2 does not hold for the underfitted model. This is because, in the underfitted model, the standard error of the estimator depends not only on the true covariance matrix but also on the biased estimator whose bias is incurred by the underfitting.

To see this, we will compare the observed Fisher information matrix for the correct model and the underfitted model. The negative Hessian matrix for the likelihood function (3.16) is (Chaudhuri et al., 2007):

$$-\frac{\partial^2 \ell(\boldsymbol{\Sigma})}{\partial \sigma^2} = -\frac{n}{2}\boldsymbol{Q}^T[\{\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\} - \{(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) \otimes \boldsymbol{\Sigma}^{-1}\} - \{\boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1})\}]\boldsymbol{Q}.$$

where $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K]$ is a $p^2 \times K$ matrix such that $\boldsymbol{q}_k = \text{vec}(\boldsymbol{G}_k)$ for $k = 1, \ldots, K$. Denote $\widehat{\boldsymbol{\Sigma}}^C$ as the MLE of the correct model and $\widehat{\boldsymbol{\Sigma}}^U$ as the MLE of the underfitted model. Then, we can prove that the observed Fisher information matrix evaluated at $\widehat{\boldsymbol{\Sigma}}^C$ converges to the Fisher information

matrix because $\widehat{\boldsymbol{\Sigma}}^C$ converges to $\boldsymbol{\Sigma}$ in probability:

$$-E\left(\frac{\partial^2 \ell(\widehat{\boldsymbol{\Sigma}}^C)}{\partial \sigma^2}\right) \to \frac{n}{2}\boldsymbol{Q}^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{Q}.$$

On the other hand, the observed Fisher information matrix evaluated at $\widehat{\boldsymbol{\Sigma}}^U$ does not converge to the Fisher information matrix because $\widehat{\boldsymbol{\Sigma}}^U$ does not converge to $\boldsymbol{\Sigma}$ and has different limit. That is, denoting this limit of $\widehat{\boldsymbol{\Sigma}}^U$ as $\boldsymbol{\Sigma}^*$, we can prove that the observed Fisher information matrix evaluated at $\widehat{\boldsymbol{\Sigma}}^U$ converges to

$$-\frac{n}{2}\boldsymbol{Q}^T[\{\boldsymbol{\Sigma}^{*-1} \otimes \boldsymbol{\Sigma}^{*-1}\} - \{(\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1}) \otimes \boldsymbol{\Sigma}^{*-1}\} - \{\boldsymbol{\Sigma}^{*-1} \otimes (\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1})\}]\boldsymbol{Q}.$$

Example: Consider a sample covariance matrix

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0.5 & -0.4 \\ 0.5 & 1 & 0.5 \\ -0.4 & 0.5 & 1 \end{bmatrix} (= \boldsymbol{\Sigma}).$$

The correct model for this covariance matrix is as below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}) \quad = \sigma_1\boldsymbol{G}_1 + \sigma_2\boldsymbol{G}_2 + \sigma_3\boldsymbol{G}_3 + \sigma_4\boldsymbol{G}_4 + \sigma_5\boldsymbol{G}_5 + \sigma_6\boldsymbol{G}_6$$

where

$$\boldsymbol{G}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{G}_4 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G}_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \boldsymbol{G}_6 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

and the true parameters are $\boldsymbol{\sigma}^T = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6) = (1, 1, 1, 0.5, 0.5, -0.4)$. The MLE is equal to the sample covariance matrix $\boldsymbol{S}$ and standard error of each element of $\boldsymbol{S}$ can be obtained from the observed Fisher information matrix as below:

$$\begin{bmatrix} 0.0141 & 0.0112 & 0.0108 \\ 0.0112 & 0.0141 & 0.0112 \\ 0.0108 & 0.0112 & 0.0141 \end{bmatrix}.$$

However, if we fit an underfitted model with $\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3, \boldsymbol{G}_5, \boldsymbol{G}_6$ as below:

$$\boldsymbol{\Sigma}(\boldsymbol{\sigma}_U) \quad = \sigma_1 \boldsymbol{G}_1 + \sigma_2 \boldsymbol{G}_2 + \sigma_3 \boldsymbol{G}_3 + \sigma_5 \boldsymbol{G}_5 + \sigma_6 \boldsymbol{G}_6,$$

the MLE for this underfitted model is

$$\begin{bmatrix} 1 & 0 & -0.87 \\ 0 & 1 & 0.93 \\ -0.87 & 0.93 & 1.81 \end{bmatrix}$$

and standard error of each element of the MLE can be obtained from the observed Fisher information matrix as below:

$$\begin{bmatrix} 0.0141 & 0 & 0.0132 \\ 0 & 0.0141 & 0.0141 \\ 0.0132 & 0.0141 & 0.0227 \end{bmatrix}.$$

Remark 10. *This result is different with linear regression under the i.i.d error assumption. In the linear regression, the variance of the regression coefficient does not depend on the estimated parameter. In the MLE for the linear covariance model, the variance of the regression coefficient depends on the estimated parameter.*

3.5.3 Likelihood Ratio Test for Model Adequacy

Given a specific covariance model (3.1) determined by the zero constraint, one may need to examine whether the true parameter lies within the parameter space of the specific model. Let $\boldsymbol{\Theta}_0$ be the parameter space of the specific covariance model with the zero constraint and $\boldsymbol{\Theta}_u$ be the parameter space of the unconstrained model so that $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}_u$. Also, let $\boldsymbol{\sigma}$ be the vector of non-zero elements of the true covariance matrix $\boldsymbol{\Sigma}$. The null and alternative hypotheses can be stated as below:

$$\mathrm{H}_0 : \boldsymbol{\sigma} \in \boldsymbol{\Theta}_0 \quad \text{verses} \quad \mathrm{H}_1 : \boldsymbol{\sigma} \notin \boldsymbol{\Theta}_0. \tag{3.21}$$

One way to test the hypotheses (3.21) is to compare the likelihood of the tested model with that of the unconstrained model by the likelihood ratio test (LRT). In LRT, the null hypothesis is rejected if (Johnson et al., 2002)

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_u} L(\boldsymbol{\theta})} < c$$

where $L(\boldsymbol{\theta})$ is the likelihood function and c is a suitably chosen constant. Note that the denominator of Λ is the the likelihood of the sample covariance matrix and Λ is always less than one. The choice of the constant c depends on the sampling distribution of Λ. However, for a large sample size, the sampling distribution of Λ can be approximated by a chi-square distribution as below:

$$-2 \ln \Lambda \sim \chi^2_{\frac{p(p+1)}{2} - K}.$$

where $K = |\boldsymbol{\sigma}|$ is the number of coefficient parameters in the model (3.1).

If the covariance model is correct or overfitted, we expect $-2 \ln \Lambda$ will follow a chi-squared distribution with degrees of freedom $p(p+1)/2 - K$ because the null hypothesis in (3.21) is true under such models. For an underfitted model, however, the null hypothesis is not true. Hence, we expect H_0 in (3.21) to be rejected by the likelihood ratio test if the model is an underfitted model.

Figure 3.3 summarizes a simulation study for the sampling distribution of the likelihood ratio for a correct model, an overfitted model and an underfitted model. For the simulation, 1,000 datasets of sample size $n = 1000$ were generated from a ten-dimensional multivariate normal distribution with the covariance matrix whose (i, j)-th element σ_{ij} is as below:

$$\begin{aligned} \sigma_{ij} &= 0.65 \quad \text{if} \quad |i-j| = 1; \\ \sigma_{ij} &= 0.375 \quad \text{if} \quad |i-j| = 2; \\ \sigma_{ij} &= 0.165 \quad \text{if} \quad |i-j| = 3; \text{and,} \\ \sigma_{ij} &= 0 \quad \text{otherwise.} \end{aligned} \tag{3.22}$$

The correct model imposes zero constraint on the zero entries in the covariance matrix. The overfitted model fits non-zero values for the 4-th off-diagonal entries in addition to the non-zero entries of the correct model. The underfitted model imposes additional zero constraint on the (7,10)-th and (10,7)-th entries in addition to the zero constraint of the correct model.

In the left panels of Figure 3.3, the empirical sampling distribution of $-2\ln\Lambda$ (histogram) is compared to the chi-squared distribution (curve). The right panels of Figure 3.3 show the empirical distribution of p-values from the likelihood ratio test. Under the null-hypothesis, the p-value is known to follow a uniform distribution (Efron, 2012). For both the correct model and the overfitted model, the sampling distribution of $-2\ln\Lambda$ coincides with the chi-squared distribution and the p-value follows the uniform distribution, indicating that the true parameter is within the parameter space of the models considered.

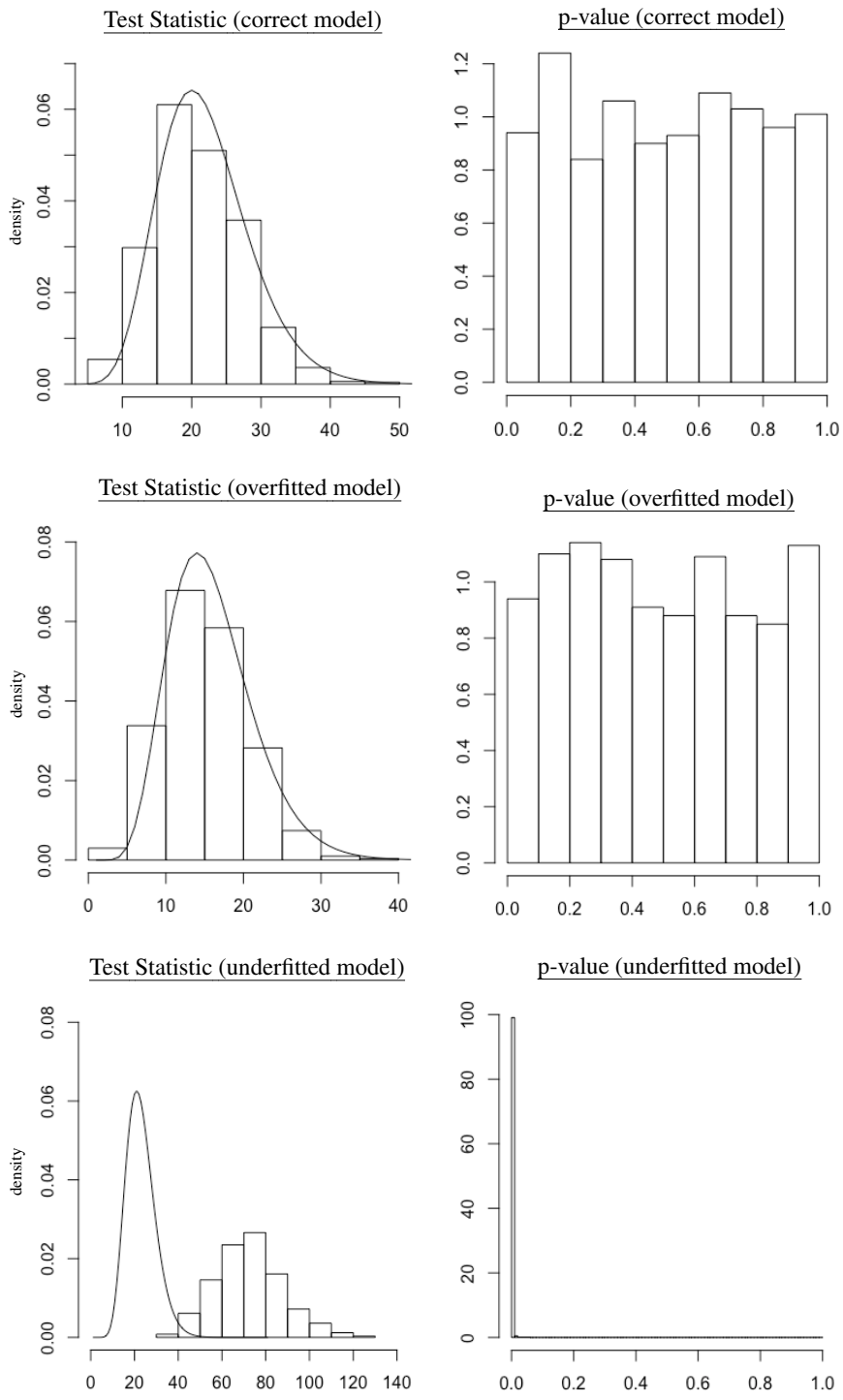

Figure 3.3: The empirical distribution of the test statistic $-2\ln\Lambda$ (left panels) and the p-value (right panels) of the likelihood ratio test for a correct model, an overfitted model and an underfitted model. In the left panels, the curves represent the null hypothesis.

4. A POSITIVE DEFINITE THRESHOLDING ESTIMATOR OF A COVARIANCE MATRIX VIA MAXIMUM LIKELIHOOD

4.1 Introduction

Estimation of a covariance matrix plays an important role in a variety of statistical problems such as classification, clustering and principal component analysis. However, the number of parameters in a covariance matrix grows quadratically as the dimension increases and this leads to high variability in estimation. That is, when there are p variables, the covariance matrix of those p variables contains $p \times (p+1)/2$ parameters to estimate. Hence, unless the sample size n is large enough compared to the number of parameters, the estimate of the covariance matrix from the sample may not be reliable. This suggests that reducing the number of parameters is a critical problem in the estimation of covariance matrix to control the variability.

One way to reduce the number of parameters in a covariance matrix is to assume that the covariance matrix has a certain structure. For example, compound symmetry structure and the first order autoregressive structure requires only one parameter to be estimated in the correlation matrix. Banding (Wu and Pourahmadi, 2003; Bickel et al., 2008b) and tapering (Furrer and Bengtsson, 2007; Cai et al., 2010) of the covariance matrix have also been studied when the correlation for the off-diagonal entries far apart from the diagonal are assumed to be smaller than those closer to the diagonal. These structured covariance matrices are implemented in available software packages (Fitzmaurice et al., 2012), making them readily usable and popular in applications.

However, assuming such structure in a covariance matrix is problematic when the true covariance matrix does not have such structure or is different from the assumed structure. This motivates the estimation of unstructured covariance matrix that allows us to capture any structure in the covariance matrix with flexibility. Using the modified Cholesky decomposition, Pourahmadi (1999) proposed an unconstrained and statistically interpretable reparameterization of a positive definite covariance matrix with the regression coefficients and the variance of innovation. This approach

was combined with the penalized likelihood estimation in Huang et al. (2006) to reduce the number of parameters. Under this approach, the Cholesky factor is assumed to contain many zero off-diagonal entries.

An alternative approach which reduces the number of parameters of an unstructured covariance matrix is to assume sparse covariance matrix. That is, the true covariance matrix is assumed to contain many zero entries and the parameters are estimated for those nonzero entries only. One simple approach to obtain sparse covariance matrix estimator is through thresholding the sample covariance matrix: setting a threshold and simply cutting-off the entries below the threshold to zero and leave the other entries unchanged. This estimator has the advantage of computational simplicity and reduced variability since it avoids estimating small entries so that the noise for those entries are not accumulated to the total noise of the estimator (Fan et al., 2016). This estimator has also been shown to be asymtotically consistent, which leads to a positive definite matrix with probability tending to one (Bickel et al., 2008a; Rothman et al., 2009).

However, with finite sample, the positive definiteness of the thresholding estimator is not guaranteed (Bickel et al., 2008a) since the thresholding the sample covariance matrix can cause the eigenvalues of the matrix to take negative values. This is problematic since positive definiteness is a basic requirement for a valid covariance matrix and losing positive definiteness will invalidate many statistical analyses such as discriminant analysis where covariance matrices are used as an input. The hard-thresholding estimator has been extended to generalized thresholding estimator by Rothman et al. (2009) which encompasses hard-thresholding and soft-thresholding as special cases but they are also not guaranteed to be positive definite.

The positive definiteness problem of the thresholding estimators has been addressed by enforcing a positive constraint on the eigenvalues. In Xue et al. (2012) and Liu et al. (2014), this constraint has been imposed on the soft-thresholding estimator which uses the convex L_1 penalty to shrink the nonzero entries by the amount of the threshold. Wen et al. (2016) proposed a method to find a positive definite solution of other thresholding estimators with non-convex penalty terms such as the hard-thresholding estimator.

In this chapter, we propose a new thresholding estimator that involves iterative conditional fitting (or iterative conditional ridge) of non-zero entries determined by thresholding the sample covariance matrix. The basic idea of this method is to perform constrained maximum likelihood estimation for nonzero entries after thresholding the sample covariance matrix. We prove this thresholding estimator is always positive definite and asymptotically efficient with probability tending to one. By finding the maximum likelihood estimator for the reduced set of parameters, the parameter estimates will feature higher accuracy than other thresholding estimators.

Certainly, a valid concern with any thresholding estimators is that their performance depends on the selection of the threshold (Bickel et al., 2008a). One value of our thresholding estimator is that we can now appeal to the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) since our estimator is based on maximizing the normal likelihood. Our approach not only allows us to select the threshold but also easily answer a question posed by Li and Zou (2016): what is the analogue of AIC or BIC for the covariance matrix estimation? To the best of our knowledge, we are the first to use AIC and BIC to select the threshold for covariance matrix estimation with thresholding. We discuss theoretical properties of AIC and BIC to support our choice of the threshold parameter. In multiple simulation studies, we compare the performance of AIC and BIC with the popular cross-validation approach (Bickel et al., 2008a; Cai and Liu, 2011) and the analytically derived threshold from Qiu and Liyanage (2019).

4.2 Some estimators of a sparse covariance matrix

To address the problem of numerous parameters, sparsity of the true covariance matrix is assumed so that the number of the parameters can be reduced. There are two broad approaches for estimating the sparse covariance matrix. One approach is to consider the natural ordering of the covariates, meaning that the correlation between the covariates far from each other is low. For example, in time-series data, the measurements for each time point can be ordered by the time and the correlation between measurements far from each other in time is often expected to be lower than that between measurements close to each other. Hence, the covariance matrix will have some structure such as Toeplitz matrix. Using this information on the structure can reduce the number

of the parameters to estimate in the covariance matrix.

However, there are many cases when such natural ordering of the covariates does not exist. In this case, we need a “permutation-invariant" method which does not assume any ordering between the covariates. This is the focus of this section and we will discuss three approaches to obtain sparse covariance matrices: thresholding, penalizion and hypothesis testing.

4.2.1 Thresholding estimators

The simplest approach to obtain a sparse covariance matrix estimator is hard-thresholding: simply cutting-off the entries below the threshold to zero and leave the other entries unchanged. This can be obtained by solving the following optimization problem.

$$\widehat{\boldsymbol{\Sigma}} = \text{argmin}(\frac{1}{2}||\boldsymbol{\Sigma} - \boldsymbol{S}||_F^2 + \frac{\lambda^2}{2}||\boldsymbol{\Sigma}||_0)$$

where $\boldsymbol{S}$ is the sample covariance matrix, $||\cdot||_F$ is the Frobenius norm and $||\cdot||_0$ is L_0 norm of the non-diagonal entries of a matrix. This hard-thresholding estimator of the covariance matrix has been shown to be asymtotically positive definite (Bickel et al., 2008a). However, with finite sample, the positive definiteness of the hard-thresholding estimator of the covariance matrix is not guaranteed, restricting the use of the hard-thresholding in practice.

The threshold parameter λ can be either a constant for all elements in $\boldsymbol{\Sigma}$ or different for all elements. The former is called the universal thresholding. Despite its simplicity, the universal thresholding does not take into account the possible heteroscedasticity of the entries of the empirical covariance matrix. To address such different variance for each entry of $\boldsymbol{S}$, Cai and Liu (2011) proposed the adaptive thresholding where different threshold is used for each element in $\boldsymbol{\Sigma}$.

The thresholding estimator was extended to the generalized thresholding estimator (Rothman

et al., 2009) whose (i, j)-th element $\hat{\sigma}_{ij}$ satisfies the following conditions:

$$(i)\ |\hat{\sigma}_{ij}| \leq |s_{ij}|$$
$$(ii)\ \hat{\sigma}_{ij} = 0 \quad \text{if} \quad |s_{ij}| \leq \lambda$$
$$(iii)\ |\hat{\sigma}_{ij} - s_{ij}| \leq \lambda$$

where s_{ij} is the (i, j)-th element of the sample covariance matrix $\boldsymbol{S}$ and λ is the threshold.

The generalized thresholding estimator encompasses hard-thresholding and soft-thresholding as special cases. The most notable estimator is the soft-thresholding estimator which can be obtained by solving the following optimization problem.

$$\widehat{\boldsymbol{\Sigma}} = \text{argmin}(\frac{1}{2}||\boldsymbol{\Sigma} - \boldsymbol{S}||_F^2 + \lambda||\boldsymbol{\Sigma}||_1)$$

where $\boldsymbol{S}$ is the sample covariance matrix, $||\cdot||_F$ is the Frobenius norm and $||\cdot||_1$ is L_1 norm of the non-diagonal entries of a matrix. One can show that the solution to this problem is simply soft-thresholding the non-diagonal entries, that is, cutting-off the entries below the threshold to zero and shrink the other entries by the amount of the threshold. As with the hard-thresholding estimator, the solution to the above optimization problem is not necessarily a positive definite matrix.

4.2.1.1 Spectral projection for positive definiteness

To make a non-positive definite matrix to a positive definite matrix, one can consider a procedure so called spectral projection. Let $\widehat{\boldsymbol{\Sigma}}^{HT}$ be a hard-thresholding estimator of the covariance matrix Σ and let $\widehat{\boldsymbol{\Sigma}}^{HT} = \mathbf{V}\Lambda\mathbf{V}^T$ be the eigen-decomposition of $\widehat{\boldsymbol{\Sigma}}^{HT}$ where Λ is a diagonal matrix of the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{HT}$. If $\widehat{\boldsymbol{\Sigma}}^{HT}$ is not positive definite, some diagonal entries of Λ will take zero or negative values. By replacing those entries with a small number ϵ and multiplying back with $\mathbf{V}$ and $\mathbf{V}^T$, the resulting matrix will be a positive definite matrix. However, after the spectral projection, the result may lose the sparsity pattern of the $\widehat{\boldsymbol{\Sigma}}^{HT}$. That is, some or many of the zero entries of $\widehat{\boldsymbol{\Sigma}}^{HT}$ will take non-zero values after the spectral projection, hence not appropriate for

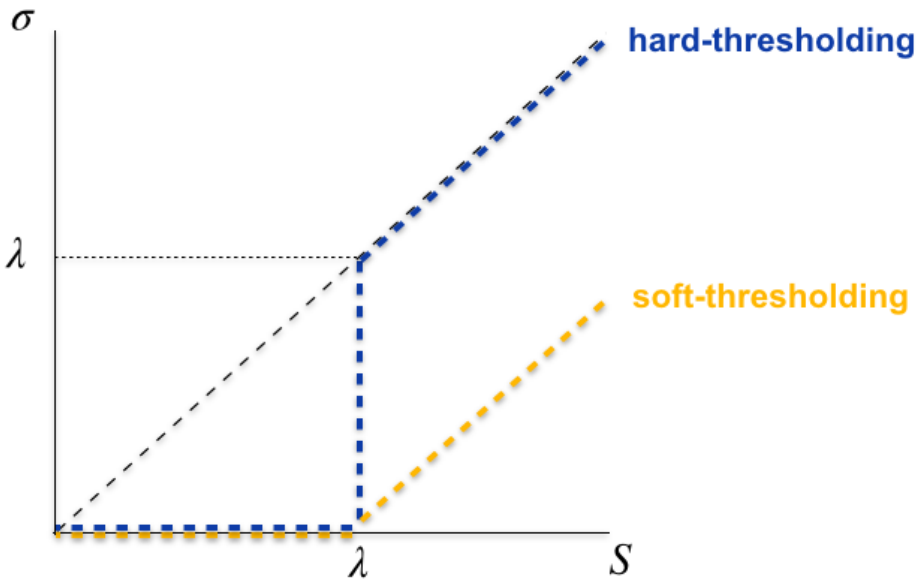

Table 4.1: Illustration of the hard thresholding and soft thresholding the sample covariance at λ

estimating sparse covariance matrix.

4.2.1.2 Positive definite approximation of the thresholding estimators

To remedy the positive definiteness issue, the soft-thresholding estimator with eigenvalue constraint has been proposed by Xue et al. (2012) (for covariance matrix) and Liu et al. (2014) (for correlation matrix). This estimator can be obtained by adding the constraint that the eigenvalues of the soft-thresholding estimator to be larger than a small number ϵ as below:

$$\widehat{\boldsymbol{\Sigma}} = \text{argmin}_{\boldsymbol{\Sigma} \succeq \epsilon I}(\frac{1}{2}||\boldsymbol{\Sigma} - \boldsymbol{S}||_F^2 + \lambda||\boldsymbol{\Sigma}||_1).$$

Note that the constraint $\boldsymbol{\Sigma} \succeq \epsilon I$ has been added to the optimization for the soft thresholding estimator. This constrained optimization problem can be solved by the alternating directions method of multipliers (ADMM) algorithm proposed by Boyd et al. (2011). The eigenvalue constraint is imposed by the iterative spetral projection within each iteration of the ADMM algorithm.

While the soft thresholding estimator is obtained by the convex optimization with L_1 norm

penalty, other generalized thresholding estimators may be obtained by a form of non-convex penalties. For example, the hard thresholding estimator is based on the optimization with L_0 norm penalty. Although such thresholding estimators with non-convex penalties can reduce bias compared to the soft thresholding estimator, optimization of such non-convex problems may be harder than the positive definite optimization of the soft thresholding estimator. For example, Liu et al. (2014) considered the non-convex minimax concave penalty for covariance matrix estimation but the their algorithm often fails to converge. Wen et al. (2016) proposed an algorithm to find a positive definite approximation of thresholding estimators with non-convex penalties and proved convergence of their algorithm.

4.2.2 Penalized likelihood estimators

In thresholding approach, no assumption is made on the distribution of the covariates and Frobenius loss with some penalty terms are minimized. Alternatively, one can consider some distributional assumption such as Gaussian distribution and minimization of the log-likelihood function combined with some penalties on each element of the covariance matrix. Specifically, for the Gaussian model, Lam and Fan (2009) considered minimization of the penalized likelihood

$$\widehat{\boldsymbol{\Sigma}} = \text{argmin}\{\log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + \lambda||\boldsymbol{\Sigma}||_1\} \tag{4.1}$$

and showed the the solution is consistent and asymptotically normal. To solve the problem, Bien and Tibshirani (2011) proposed a majorization-minimization algorithm which finds a positive definite solution to this problem. However, the algorithm finds the exact solution to the penalized likelihood only when the sample size n is greater than the number of variables p. For the case of $n < p$, Bien and Tibshirani (2011) suggested replacing the sample covariance matrix $\boldsymbol{S}$ in the penalized likelihood with $\boldsymbol{S} + \epsilon I_p$ for some $\epsilon > 0$.

4.3 A positive definite thresholding estimator with efficiency

4.3.1 The COMET estimator

We now extend the setup in Section 3.3 by assuming the location of the zero entries in a covariance matrix is unknown and use thresholding to identify the zero entries. Then, the non-zero entries are estimated by iterative conditional fitting (or iterative conditional ridge when $n < p$). We show that, as the sample size increases, thresholding the sample covariance matrix will recover the support for the non-zero entries with probability tending to one so that iterative conditional fitting will compute an asymptotically efficient estimator.

We define COMET (COvariance Maximum-likelihood Estimator with Thresholding) given a $p \times p$ matrix of threshold $\boldsymbol{\Lambda} = (\lambda_{ij})_{i,j=1}^{p}$ as:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Lambda}} = \text{argmax}_{\boldsymbol{\Sigma} \succ 0, \sigma_{ij} \mathbb{1}_{\{|s_{ij}| < \lambda_{ij}\}} = 0} \ell(\boldsymbol{\Sigma}) \tag{4.2}$$

where $\ell(\boldsymbol{\Sigma})$ is the log-likelihood function for $\mathbb{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \succ 0$ constrains $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Lambda}}$ to be a positive definite matrix, $\mathbb{1}$ is the indicator function and σ_{ij} and s_{ij} are the (i,j)th entry in $\boldsymbol{\Sigma}$ and $\boldsymbol{S}$, respectively. Here, λ_{ij} can be either a constant for all (i,j)s or different for each (i,j). The former is called the universal thresholding (Bickel et al., 2008a). Despite its simplicity, the universal thresholding does not take into account the possible heteroscedasticity of the entries of the sample covariance matrix. For example, under the normal assumption with the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^{p}$, $n\boldsymbol{S}$ has a Wishart distribution whose (i,j)th entry has variance $n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})$. To address such different variance for each entry of $\boldsymbol{S}$, Cai and Liu (2011) proposed the adaptive thresholding where different threshold λ_{ij} is used for each s_{ij}.

The definition of COMET means that it is the maximum likelihood estimator for the non-zero entries determined by thresholding the sample covariance matrix. Although the global solution is not guaranteed, we can find a positive definite local solution by iterative conditional fitting. The COMET estimator is implemented in the 'COmet' function in our R package `mgcov` (`https://github.com/Tanya-Garcia-Lab/mgcov`).

The next theorem discusses the asymptotic distribution of COMET when the threshold for the (i,j)th entry takes the form $\lambda_{ij} = C_{ij}n^{-\alpha}$ as proposed in El Karoui et al. (2008). We assume that iterative conditional fitting starts from a consistent estimator of $\mathbf{\Sigma}$.

Theorem 2. *Given the $p \times p$ sample covariance matrix $\boldsymbol{S} = (s_{ij})_{i,j=1}^{p}$ and a matrix of threshold $\mathbf{\Lambda} = (\lambda_{ij})_{i,j=1}^{p}$, define $\boldsymbol{\sigma}_{\mathbf{\Lambda}}$ as the vector of elements in $\{\sigma_{ij} : |s_{ij}| \geq \lambda_{ij}, i \geq j\}$. Let $\widehat{\boldsymbol{\sigma}}_{\mathbf{\Lambda}}$ be the estimator of $\boldsymbol{\sigma}_{\mathbf{\Lambda}}$ computed from iterative conditional fitting. If $\lambda_{ij} = C_{ij}n^{-\alpha}$ for a positive constant C_{ij}, $\alpha = 0.5 - \gamma > 0$ and $\gamma > 0$, then, as $n \to \infty$,*

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\sigma}}_{\mathbf{\Lambda}} - \boldsymbol{\sigma}_{\mathbf{\Lambda}}) \to \mathbb{N}(\mathbf{0}, I(\boldsymbol{\sigma})^{-1})$$

with probability tending to one where $\boldsymbol{\sigma}$ is the non-zero parameter vector in $\mathbf{\Sigma}$ and $I(\boldsymbol{\sigma})$ is the Fisher information matrix.

Remark 11. *By $\widehat{\boldsymbol{\sigma}}_{\mathbf{\Lambda}}$ computed from iterative conditional fitting, we mean the vector of non-zero entries in the lower (or upper) triangular part of $\widehat{\mathbf{\Sigma}}_{\mathbf{\Lambda}}$.*

Remark 12. *When $n < p$, Theorem 2 holds for the solution computed from the iterative conditional ridge. In this case, $\boldsymbol{\sigma}_{\mathbf{\Lambda}}$ represents the vector of elements in $\{\sigma_{ij} : |s_{ij}| \geq \lambda_{ij}, i > j\} \cup \{\sigma_{ij} + \epsilon : i = j\}$.*

4.3.2 Selection of the threshold

When a covariance matrix is estimated by thresholding, the performance of the estimator is crucially dependent on the selection of the threshold parameter (Bickel et al., 2008a). The higher the threshold is, the fewer non-zero entries the estimator has. Although an estimator with fewer non-zero entries is more interpretable in the sense that it is simpler, it may fail to identify some non-zero entries in the true covariance matrix. On the other hand, a model with too low threshold may contain spurious non-zero entries.

4.3.2.1 Cross-validation

Cross-validation has been widely used for the selection of threshold parameter for covariance matrix estimation, for example, Bickel et al. (2008a) for the universal thresholding and Cai and Liu (2011) for the adaptive thresholding. The basic idea of the cross-validation is to minimize the Frobenius risk of the estimator empirically as below.

1. Split the sample of size n randomly into two pieces of size $n_1 = n(1 - \frac{1}{\log n})$ and $n_2 = \frac{n}{\log n}$.
2. Given a threshold λ, construct a thresholding estimator $\widehat{\boldsymbol{\Sigma}}_1(\lambda)$ for the sample of size n_1.
3. Construct the sample covariance matrix $\boldsymbol{S}_2$ for the sample of size n_2.
4. Repeat 1-3 and compute the average of $||\widehat{\boldsymbol{\Sigma}}_1(\lambda) - \boldsymbol{S}_2||_F^2$.
5. Choose $\hat{\lambda}$ which minimizes the average of $||\widehat{\boldsymbol{\Sigma}}_1(\lambda) - \boldsymbol{S}_2||_F^2$.

4.3.2.2 Threshold by Qiu and Liyanage (2019)

Qiu and Liyanage (2019) proposed an analytical form for the adaptive thresholding which is theoretically optimal for minimizing the Frobenius risk.

For n observations $y_1, \ldots, y_n$ of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ with mean zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^p$, the standardized covariance η_{ij} is defined as below:

$$\eta_{ij} = \frac{n^{1/2}}{(\log p)^{1/2}} \sigma_{ij} \theta_{ij}^{-1/2}$$

where $\theta_{ij} = \text{var}(Y_i Y_j)$. Qiu and Liyanage (2019) defined a set of indices whose cardinality is critical to derive the optimal threshold as below:

$$\mathcal{S} = \{(i,j) : i > j, |\eta_{ij}| \in [a_1, 2]\}$$

where $a_1 = 2 - \min\{(2 + \log n / \log p)^{1/2}, 2\}$.

They considered the optimal threshold δ_{opt} for the standardized covariance as the minimizer of the Frobenius risk for the adaptive thresholding estimator as below:

$$\delta_{opt} = \text{argmin}_{\delta} E||\widehat{\mathbf{\Sigma}}(\delta) - \mathbf{\Sigma}||_F^2$$

where $\widehat{\mathbf{\Sigma}}(\delta)$ is an adaptive thresholding estimator (Cai and Liu, 2011) with the threshold δ. They proposed a consistent estimator of δ_{opt} as below:

$$\delta = \sqrt{2\left[2 - \frac{\log\left\{N_2(\log p)^{-1/2}\right\}}{\log p}\right]}$$

where N_2 is the cardinality of $\mathcal{S}$. For details of the estimation of N_2, see Qiu and Liyanage (2019).

4.3.2.3 Information criteria for the selection of COMET threshold

Existing methods for threshold selection such as cross-validation do not make any distributional assumption when selecting the threshold. If a specific distribution such as the Gaussian distribution is assumed, such assumption may be taken into account when the threshold is selected. The COMET estimator allows us to consider such distributional assumptions, providing information criteria as additional tools for selecting the threshold. Given a family of models, one can select a specific model which is optimal under a pre-defined information criterion. Typical choice of such criterion includes Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC can be defined for the covariance matrix with zero entries as below:

$$\text{AIC} = -2\ell(\boldsymbol{\sigma}) + 2|\boldsymbol{\sigma}|$$

$$\text{BIC} = -2\ell(\boldsymbol{\sigma}) + \log(n)|\boldsymbol{\sigma}|$$

where $\ell(\boldsymbol{\sigma})$ is the log-likelihood (3.16) and $|\boldsymbol{\sigma}|$ is the number of non-zero parameters in $\mathbf{\Sigma}$. Although AIC and BIC are commonly used for model selection in general, they have not been used for the thresholding estimators of a covariance matrix because other thresholding estimators (Bickel

et al., 2008a; Rothman et al., 2009) have been estimated regardless of the likelihood. However, the COMET estimator involves maximization of the likelihood. Hence, computation and minimization of AIC and BIC are straightforward for the COMET estimator.

The use of such information criteria can be advocated by their theoretical properties. Particularly, BIC is known to be asymptotically consistent in model selection: given a family of models including the true model, BIC will select the true model with probability tending to one as $n \to \infty$ (Hastie et al., 2009). Thus, in conjunction with the Theorem 1, the COMET estimator selected by BIC will give an asymptotically efficient estimator with high probability when n is large. On the other hand, AIC was shown to be asymptotically equivalent to leave-one-out cross-validation (Stone, 1977). With a small sample, AIC may be preferred to BIC because BIC often chooses too simple models due to higher penalty on the model complexity than AIC.

4.4 Simulation Study

4.4.1 Simulation settings

We compare the performance of COMET with the hard thresholding estimator (Bickel et al., 2008a; Cai and Liu, 2011). The difference between these estimators is in how we estimate the non-zero entries. In the hard thresholding, the non-zero entries are estimated by the sample covariances whereas they are estimated by iterative conditional fitting (when $n > p$) or iterative conditional ridge (when $n \leq p$) in COMET.

We generate 100 datasets, each with sample size $n \in \{25, 50, 100, 200\}$ and the number of variables $p \in \{10, 50\}$ from $\mathbb{N}_p(\mathbf{0}, \mathbf{\Sigma})$. The covariance matrix $\mathbf{\Sigma}$ for each dataset is constructed by $\mathbf{\Sigma} = \boldsymbol{DRD}$ where $\boldsymbol{R}$ is a $p \times p$ correlation matrix and $\boldsymbol{D}$ is a diagonal matrix whose diagonal entries are randomly drawn from the uniform distribution on $(0.1, 10)$. The multiplication of the matrix $\boldsymbol{D}$ to a correlation matrix introduces heteroscedasticity in covariances and such a simulation approach was also used in Qiu and Liyanage (2019). We adopt this approach to mimic the characteristics of our Huntington disease data where the covariance matrix shows strong heteroscedasticity. The correlation matrix $\boldsymbol{R} = (\rho_{ij})_{i,j=1}^{p}$ is determined by one of the following

models.

- Moving average model: the moving average process of order one where $\rho_{ij} = 0.3$ if $|i-j| = 1$ and $\rho_{ij} = 0$ otherwise.
- Autoregressive model: the autoregressive process of order one where $\rho_{ij} = 0.5^{|i-j|}$.
- Block model: the set of indices $\{1, \ldots, p\}$ is partitioned into 5 non-overlapping subsets $S_1, \ldots, S_5$ of equal size and the (i, j)th entry of Σ is $\rho_{ij} = 0.5 I_{(i=j)} + 0.5 \sum_{k=1}^{5} I_{(i \in S_k, j \in S_k)}$.

The moving average model and the autoregressive model are typical time series models that have been used in many covariance estimation studies such as Rothman et al. (2009) and Qiu and Liyanage (2019). In the moving average model, most of the covariances are zero and identifying those zero entries in the covariance matrix can reduce estimation error. On the other hand, the autoregressive model does not contain any zero entry but most of the covariance entries are close to zero for large p such as $p = 50$. Hence, estimating those entries as zero can also reduce estimation error. The block model has also been considered in many studies including Bien and Tibshirani (2011) and Liu et al. (2014). In this model, the groups of variables in different blocks are independent to each other and the maximum likelihood estimator for each block is equal to the sample covariance matrix. Hence, if all zero and non-zero entries are correctly identified by a threshold, the COMET estimator is equal to the hard thresholding estimator.

To account for the heteroscedasticity, the adaptive thresholding (Cai and Liu, 2011) is used for both COMET and the hard thresholding. For the hard thresholding, the threshold is selected by the cross-validation (Bickel et al., 2008a) and by the closed-form threshold (Qiu and Liyanage, 2019). For COMET, AIC, BIC and the closed-form threshold (Qiu and Liyanage, 2019) are used for the threshold selection.

4.4.2 Performance evaluation

Covariance matrix estimators are evaluated in two aspects: covariance estimation and support recovery. Estimation aspect measures how close the estimator is to the true covariance matrix. Sup-

port recovery aspect considers how well the zero entries and nonzero entries in the true covariance matrix are detected in the estimator.

For comparing estimation performance, we use the Frobenius loss, $||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}||_F = \{\sum_{i,j}(\hat{\sigma}_{ij} - \sigma_{ij})^2\}^{1/2}$, which measures the distance between the true covariance matrix $\mathbf{\Sigma}$ and an estimator $\widehat{\mathbf{\Sigma}}$. We also compute the entropy loss, $-\log|\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}| + \text{tr}(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - p$, a measure of the Kullback-Liebler divergence of two multivariate normal densities (Pourahmadi, 2013). Both measures are commonly used for comparing covariance estimation performance (Huang et al., 2006; Bien and Tibshirani, 2011). For both measures, an estimator with lower value is more desirable.

For evaluating support recovery performance, we compare the true positive rate and the false positive rate. The true positive rate is the tendency to correctly estimating non-zero entries in the true covariance matrix as non-zero while the false positive rate is the tendency to falsely estimating zero entries in the true covariance matrix as non-zero. These are standard measures for support recovery (Rothman et al., 2009; Xue et al., 2012) and defined as below:

$$\text{True positive rate} = \frac{\#\{(i,j) : \hat{\sigma}_{ij} \neq 0,\ \sigma_{ij} \neq 0\}}{\#\{(i,j) : \sigma_{ij} \neq 0\}}$$
$$\text{False positive rate} = \frac{\#\{(i,j) : \hat{\sigma}_{ij} \neq 0,\ \sigma_{ij} = 0\}}{\#\{(i,j) : \sigma_{ij} = 0\}}$$

where $\#$ denotes the number of elements in a set. An estimator with higher true positive rate and lower false positive rate is preferred.

We also check whether each estimator is a positive definite matrix, which is a requirement for a valid covariance matrix. Since the solution from the iterative conditional fitting algorithm is always positive definite, we can always find a positive definite COMET estimator. On the other hand, hard thresholding the sample covariance matrix may result in losing the positive definiteness of the matrix. Positive definite approximation of the thresholding estimators was studied in Wen et al. (2016) for the universal thresholding but their method may need modification for the adaptive thresholding which is out of the scope of our analysis.

4.4.3 Simulation results

For estimation performance, we show boxplots of Frobenius loss and entropy loss for the case of $n = 100, p = 50$ in Figure 4.1 and the case of $n = 25, p = 50$ in Figure 4.2. Support recovery performance of the thresholding estimators for different n and p is plotted in Figure 4.3. Simulation results for more cases are presented in Appendix C.1.

First, we compare the COMET estimators with different threshold parameters selected by AIC, BIC and the closed-form threshold (Qiu and Liyanage, 2019). Due to higher penalty on model complexity, BIC selects higher threshold than AIC. Because more zero entries in the true covariance matrix $\mathbf{\Sigma}$ are correctly estimated to be zero by higher threshold, BIC leads to lower false positive rate, Frobenius loss and entropy loss than AIC. The closed-form threshold (Qiu and Liyanage, 2019) tends to select even higher threshold than BIC. When the sample size is greater than the number of variable (e.g. $n = 100, p = 50$), such a high threshold is problematic because many non-zero entries in $\mathbf{\Sigma}$ are estimated to be zero, as seen by low true positive rate in Figure 4.3. This also results in higher Frobenius loss and entropy loss than BIC in Figure 4.1. However, when the sample size is smaller than the number of variable (e.g. $n = 25, p = 50$), many zero entries in $\mathbf{\Sigma}$ are spuriously estimated to be non-zero by BIC, as seen by high false positive rate in the lower panels of Figure 4.3. As a result, the COMET estimators with BIC-threshold tend to show higher Frobenius loss and entropy loss than the COMET estimators with the closed-form threshold in Figure 4.2. Hence, for the selection of COMET threshold, we suggest BIC when $n > p$ and the closed-form threshold when $n \leq p$.

Next, we compare the COMET estimator with the hard thresholding estimator whose threshold parameter is selected by cross-validation or the closed-form threshold (Qiu and Liyanage, 2019). When $n > p$, the COMET estimator with BIC-threshold tends to be closer to the true covariance matrix $\mathbf{\Sigma}$ than hard thresholding estimators as seen by lower Frobenius loss and entropy loss in Figure 4.1. When $n \leq p$, the COMET estimator with the closed-form threshold shows comparable results with hard thresholding estimators in Figure 4.2.

Regarding positive definiteness, the hard thresholding estimators often failed to be positive-

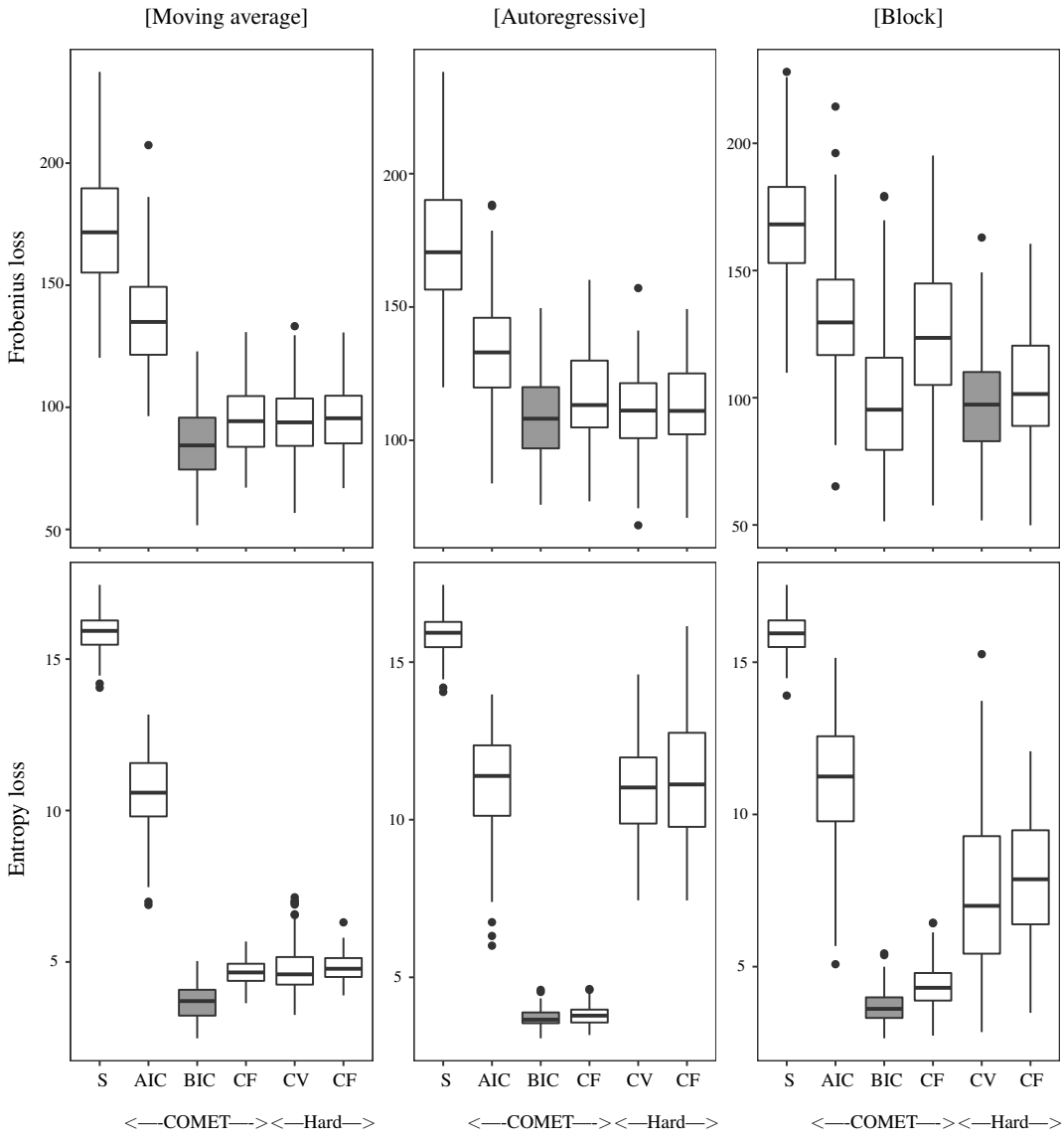

Figure 4.1: Boxplots of Frobenius loss and entropy loss when $n = 100$ and $p = 50$; S, sample covariance matrix; AIC, threshold selected by the AIC; BIC, threshold selected by the BIC; CF, threshold selected by the closed-from threshold; CV, threshold selected by the cross-validation. The estimator with grey box has the lowest mean.

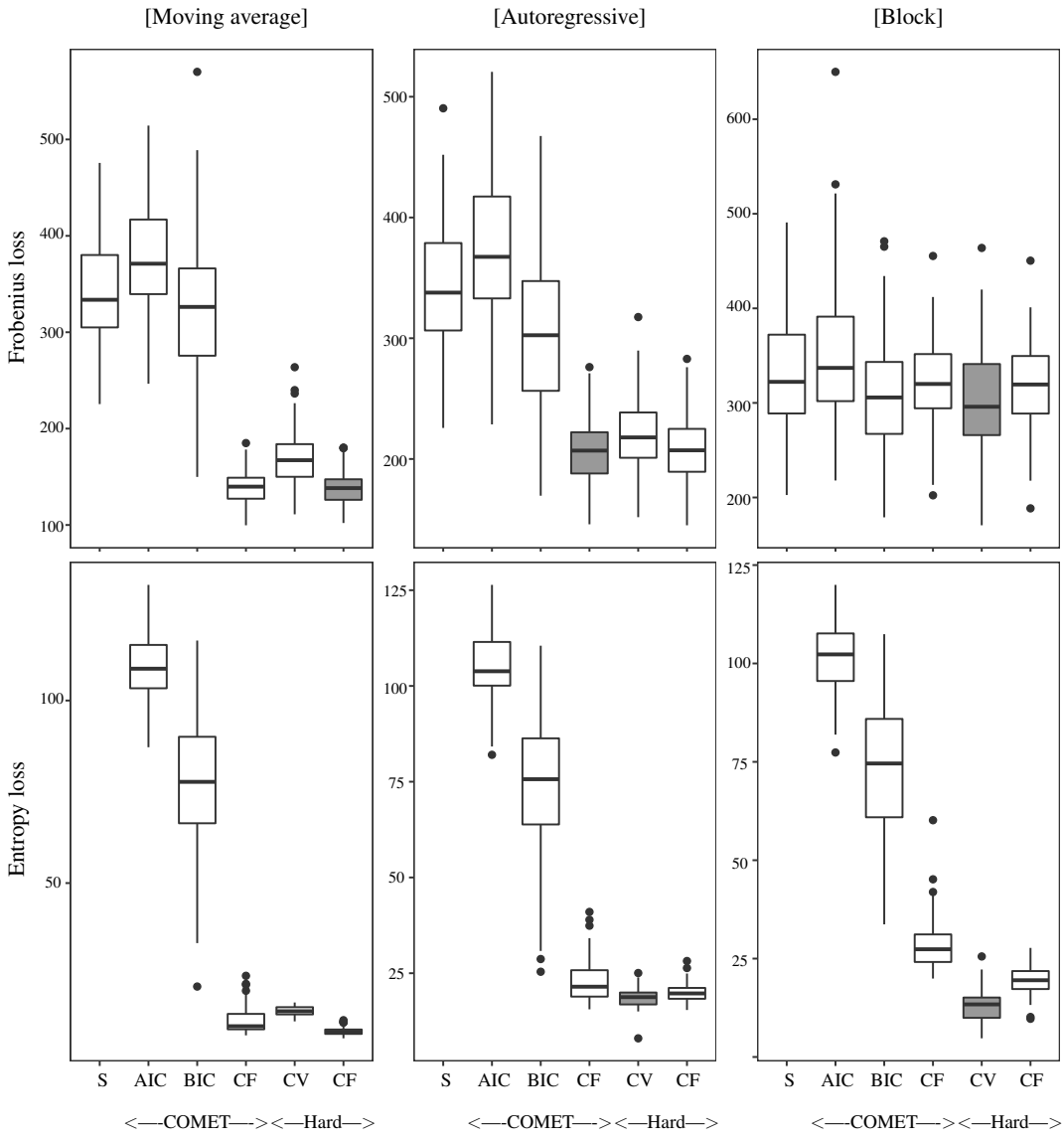

Figure 4.2: Boxplots of Frobenius loss and entropy loss when $n = 25$ and $p = 50$; S, sample covariance matrix; AIC, threshold selected by the AIC; BIC, threshold selected by the BIC; CF, threshold selected by the closed-from threshold; CV, threshold selected by the cross-validation. The estimator with grey box has the lowest mean.

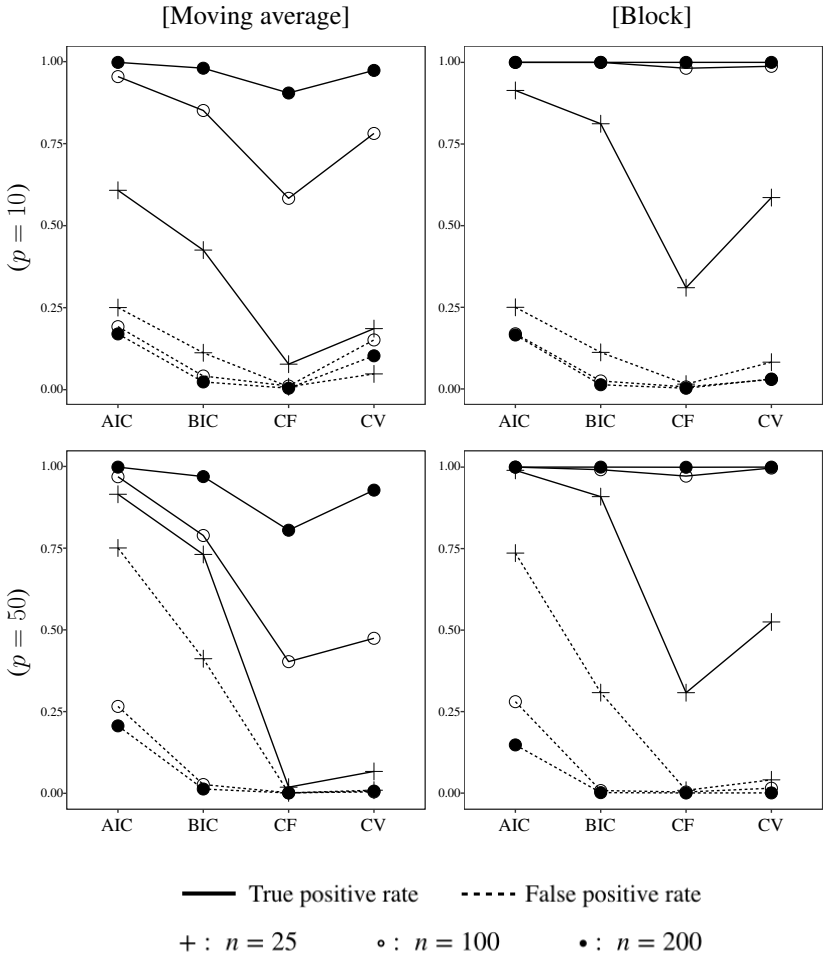

Figure 4.3: True positive rate (solid) and false positive rate (dotted) when $p = 10$ (upper panels) or $p = 50$ (lower panels) and $n = 25$ (+) or $n = 100$ ($\circ$) or $n = 200$ ($\bullet$); AIC, threshold selected by the AIC; BIC, threshold selected by the BIC; CF, threshold selected by the closed-from threshold; CV, threshold selected by the cross-validation. The autoregressive model is not compared since there is no zero entry in the covariance matrix.

definite. For example, when $n = 25$ and $p = 50$, the hard thresholding estimators are not positive definite for 100% of the simulated datasets from the block model. More results on the non-positive definite hard thresholding estimators are presented in Appendix C.1. Since COMET estimators are always positive definite, the COMET estimator is still preferred even when the sample size is smaller than the number of variables.

4.4.4 Simulation for Non-Gaussian Models

One of the main difference between the COMET and other estimators is that the COMET is based on the maximum likelihood estimation for Gaussian distribution. In this section, we compare the COMET with other estimators for non-Gaussian settings.

As non-Gaussian models, we consider a log-normal distribution, Lognormal(μ, σ^2) with $\mu = 0$ and $\sigma = 1$, and an exponential distribution, Exp(λ) with $\lambda = 1$. The samples from these non-Gaussian distributions are drawn by the following procedure.

1. Draw samples $(y_{i1}, \ldots, y_{ip})^T$ from $\mathbb{N}_p(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is equal to one of the correlation matrix described in Section 4.2.1.

2. For each i and j, compute $F^{-1}\{\Phi(y_{ij})\}$ where Φ is the distribution function of the standard normal distribution and F is the distribution function of either the log-normal distribution or the exponential distribution.

From our simulation studies, we do not find strong outperformance of one method over others as seen in Figure 4.4. For the log-normal distribution, the hard thresholding estimators tend to show slightly lower Frobenius loss than the COMET with AIC-threshold or BIC-threshold. For the exponential distribution, however, vice versa. This suggest that the COMET is as competitive as other estimators for non-Gaussian cases.

4.5 Correlations between brain regions for Huntington Disease

We apply our method to the PREDICT-HD study, a large observational study from 2001 to 2013 on potential neurobiological markers of Huntington Disease. Huntington disease is a genetically

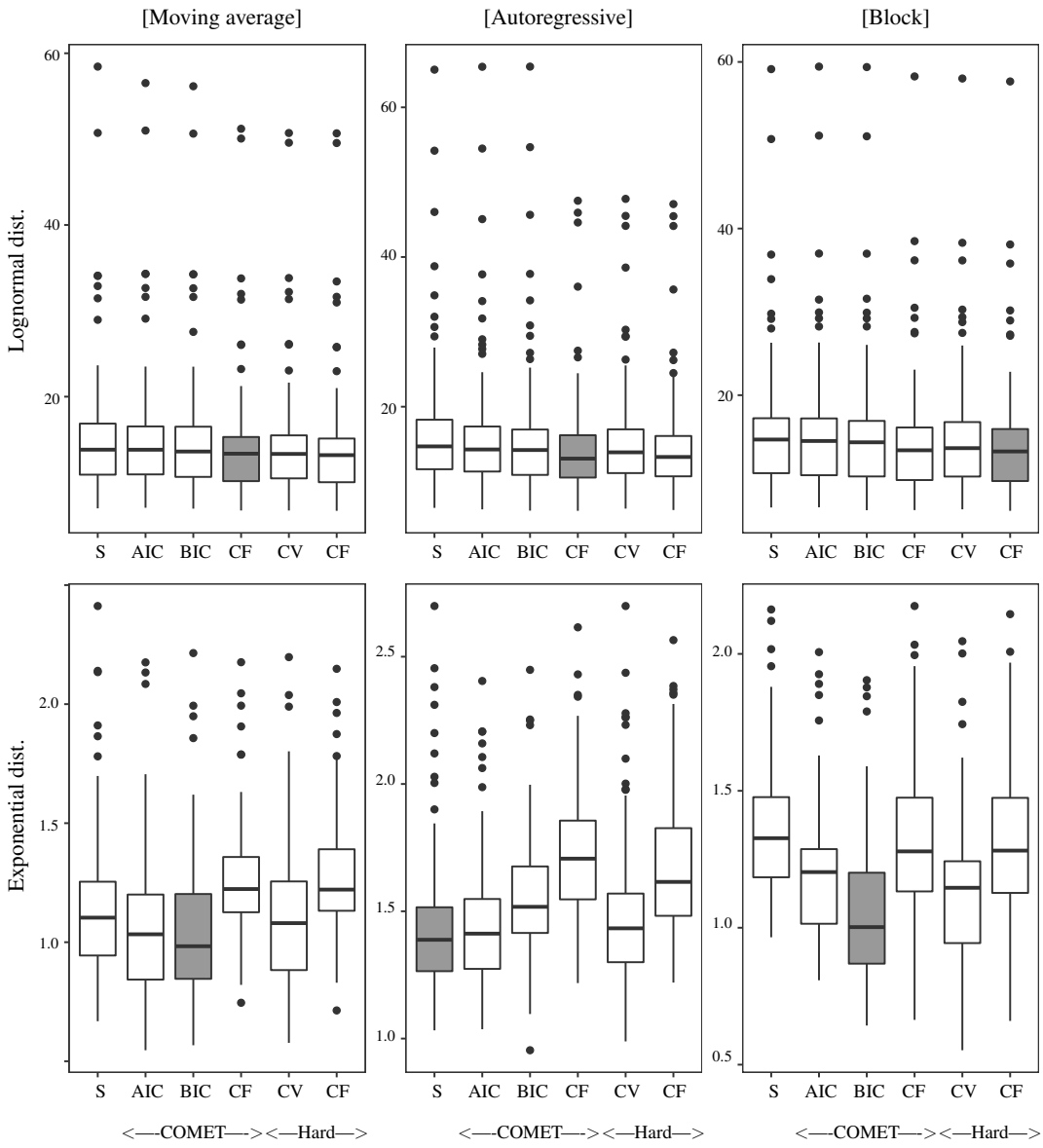

Figure 4.4: Boxplots of Frobenius loss for lognormal distribution and exponential distribution when $n = 100$ and $p = 10$; S, sample covariance matrix; AIC, threshold selected by the AIC; BIC, threshold selected by the BIC; CF, threshold selected by the closed-from threshold; CV, threshold selected by the cross-validation. The estimator with grey box has the lowest mean.

inherited neurodegenerative disorder with cognitive and motor decline. Since those symptoms are caused by volume loss or enlargement of some brain regions, treatments for the disease attempt to interrupt the change of brain regional volumes. However, assessing the effect of a treatment is challenging since brain regions are correlated to each other. For example, a treatment may lower the level of protein called mutant Huntingtin (mHTT) in cerebrospinal fluid, a fluid that surrounds the brain (Tabrizi et al., 2019). The level of mHTT in cerebrospinal fluid was found to be negatively associated with the volume of caudate, one of the brain regions that control motor functions (Rodrigues et al., 2020). In other words, the lower the level of the protein, the slower the volume loss of caudate, thus the slower the motor decline, meaning that the treatment will delay motor impairment. However, the treatment may affect other brain regions as well which are correlated to caudate. Hence, to assess the effect of the treatment, which brain regions are correlated to caudate needs to be identified first by detecting non-zero entries in a correlation matrix.

To detect non-zero correlations, we estimate the covariance matrix of the brain regional volumes by COMET with BIC-threshold. We use the PREDICT-HD data of 710 subjects who are "at risk" of the Huntington disease. Subjects at risk have mutated gene with CAG (cytosine, adenine, guanine) repeats greater than 35 and they will exhibit Huntington disease symptoms in their life time. Volumes of 41 brain regions are measured by MRI scanners for each patient and the 41×41 covariance matrix of the volume measures of those regions are estimated by COMET. Other thresholding estimators discussed in Section 4.4 have also been applied to the data but all hard thresholding estimators were not positive definite. We therefore proceed with our analysis using COMET with BIC-threshold which produced the most reliable positive definite estimator in terms of estimation and support recovery in the simulation studies. Data analysis results for other estimators are summarized in Appendix C.2.

In Figure 4.5, COMET with BIC-threshold identifies brain regions which have non-zero correlations with the basal ganglia, a brain structure that controls motor movement and is known to shrink prominently as the Huntington disease progresses. The regions with positive correlations

coincide with regions which are known to shrink together with the basal ganglia (Reiner et al., 2011). The negative correlations are explained by the enlargement of ventricular system, a set of interconnected cavities, for Huntington disease patients (Reiner et al., 2011; Tabrizi et al., 2019). However, such ventricular enlargement was not captured in earlier covariance studies (Minkova et al., 2016; Coppen et al., 2016) where Bonferroni correction selected non-zero correlations conservatively, detecting only positive correlations. Bonferroni correction result for the PREDICT-HD data is given in Appendix C.2. Our COMET correlation matrix in Figure 4.5 contains both the positive and the negative correlations which inform us how each brain region will be affected by a potential treatment. For example, if a treatment slows the shrinkage of the basal ganglia, the shrinkage of positively correlated regions and the enlargement of negatively correlated regions will also get slower, leading to slower decline of the functions controlled by those regions.

In contrast to the positive correlations, the structure of the negative correlations differs by the regions of the basal ganglia: accumbens, caudate, putamen, and palidium. Figure 4.6 compares the negative correlation structure through network graphs with each node representing each brain region. A negative correlation between two regions is shown by an edge between two nodes, and no edge means that the correlation between two regions is either zero or positive. These network graphs indicate that each region of the basal ganglia affects other brain regions differently. For example, Huntington disease is characterized by the shringkage of the basal ganglia and concomitant enlargement of the lateral ventricles (Reiner et al., 2011; Degnan and Levy, 2014). Our correlation matrix looks deeper into the correlations of lateral ventricles to each region of the basal ganglia and reveals zero correlation to the caudate and non-zero correlations to other regions of the basal ganglia. This implies that a treatment which targets the caudate may not affect the lateral ventricles whereas another treatment which targets the putamen or the pallidum may affect the lateral ventricles. A similar conclusion about the lack of association between the caudate and the lateral ventricles was drawn in Milovanovic et al. (2018) for schizophrenia, a mental disorder which also involves volume loss of the caudate as Huntington disease (Williams, 2016).

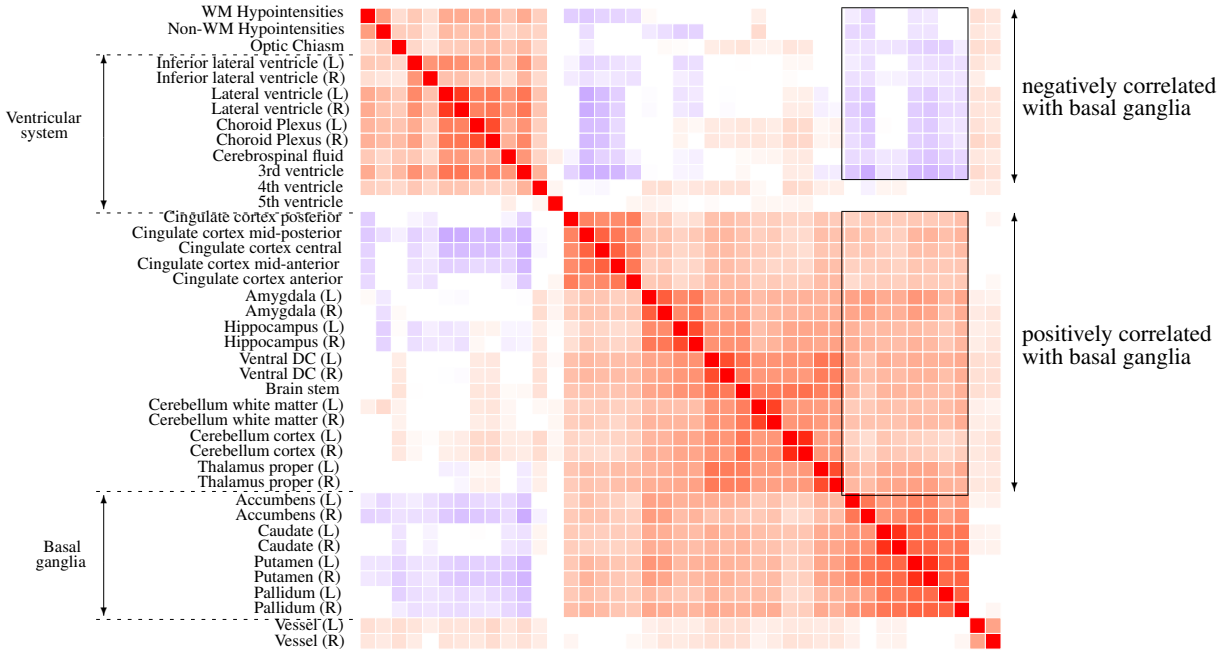

Figure 4.5: Heatmaps of the COMET correlation matrix with BIC-threshold. Positive correlations are shown in red and negative correlations are shown in blue. Zero correlations are shown in white.

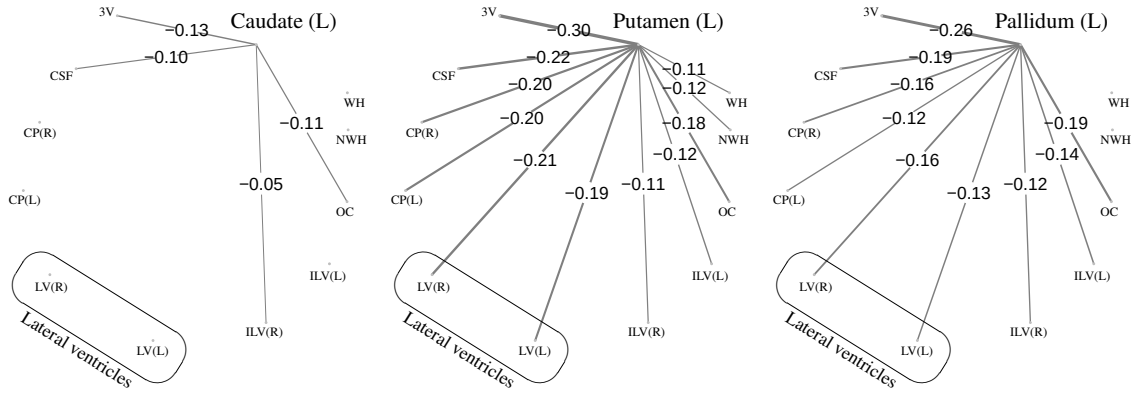

Figure 4.6: Network graphs of the negative correlations by COMET with BIC-threshold for some regions of the basal ganglia: left caudate, left putamen and left pallidum; WH: WM hypointensities; NWH: non-WM hypointensitie; OC: Optic Chiasm; ILV: inferior lateral ventricle; LV: lateral ventricle; CP: Choroid Plexus; CSF: cerebrospinal fluid; 3V: third ventricle. “L" and “R" in parenthesis represent the left and the right part of each brain region, respectively. Magnitude of the negative correlation is shown in the middle of each edge and also expressed by the width of each edge.

4.6 Discussion

The non-zero entries of the COMET estimator can be biased due to model misspecification. An interesting avenue for research is the development of information criteria to mitigate model misspecification. Lv and Liu (2014) studied the model selection in misspecified models and extended AIC and BIC to account for the misspecification bias in generalized linear models. They proposed the generalized BIC with prior probability (GBIC_p) which admits a decomposition of the form

$$\text{goodness of fit} + \text{model complexity} + \text{model misspecification},$$

whose last term has not been considered in AIC and BIC. Such extension to AIC and BIC can be pursued for the selection of the threshold parameter for the COMET estimator.

Although we have focused on thresholding in this chapter, the maximum likelihood estimation for the non-zero entries in a covariance matrix can also be combined with other selection procedures for detecting non-zero entries. For example, we can combine iterative conditional fitting with banding (Bickel et al., 2008b). Such a banding estimator will always be positive definite, a property not owned by other banding estimators in general. We can also appeal to AIC and BIC for the selection of the bandwidth parameter.

5. SUMMARY AND CONCLUSIONS

In this age of big data, data are often collected on numerous variables. For example, in neuroscience, a brain is composed of many regions and the volume of each brain region is measured for monitoring the progress of a neurodegenerative disease. However, as quoted by Rutherford D. Rogers: “We are drowning in information and starving for knowledge", information on each variable is not equally important to explain variables of our interest and there is a growing need to extract essential relationships among the variables in the data.

A statistical assumption to address such problems is sparsity, meaning that only a small number of relationships are non-zero. Driven by current interdisciplinary problems in neuroscience, this dissertation proposes statistical methods that exploit sparsity in varying-coefficient regression and covariance matrix estimation.

We believe that the methods proposed in this dissertation pose many interesting topics for future research. We conclude this dissertation by suggesting some of those topics below.

5.1 Personalized Statistical Modeling and Applications

Although the structural varying-coefficient regression (svReg) was motivated by a neuroscience problem, it can be applied to any areas where personalization is needed in regression modeling. For example, the method can be used as a pricing method for personalized insurance products (presented in Actuarial Research Conference 2021).

Such personalization can also be pursued in other statistical modeling such as covariance matrix estimation. When the data can be split into several subgroups, the covariance matrix among variables is estimated by either a common covariance matrix for all data (when we assume equal covariance matrices between subgroups), or a separate covariance matrix for each subgroup (when we do not assume equal covariance matrices). However, even if we can not assume equal covariances across all subgroups, some covariance entries may be common through all subgroups. Since the penalty terms in the svReg considers differentiated modeling for each subgroup, such

penalization techniques can be extended to the problem of covariance matrix estimation for each subgroup.

5.2 Informed estimation of a covariance matrix

A covariance matrix is essential for the risk management in the finance and insurance industry, for example, for regulatory purposes BCBS (2010); Calibration (2010). The regulators in the industry often specifies some covariance entries with fixed values. However, insurance companies often need to model covariance entries which are not specified by the regulator. When some entries in a covariance matrix is known a priori or provided by external sources (e.g. regulations in finance industry), we need an estimator which is subject to those constraints. Just replacing those entries in a sample covariance matrix with the constrained values is not optimal in any sense (e.g. maximizing the likelihood) and may lose positive definiteness.

This problem can be cast as a matrix completion problem where some entries of the matrix are known and other missing entries need to be estimated. Georgescu et al. (2018) finds a closed-form solution to such covariance matrices based on maximization of the determinant of the matrix. Since this method does not depend on the data, it may be useful if there is no data to fill the missing entries. However, if some data are available, considering those data will improve the estimation. Also, the method in Georgescu et al. (2018) is applicable only when the missing entries are patterned in certain ways.

We can potentially apply the iterative conditional fitting algorithm (Chaudhuri et al., 2007) to such covariance matrix completion problems. For now, the iterative conditional fitting algorithm estimates non-zero entries in a covariance matrix given some constraints that some entries are zero. By some modification to the algorithm, it can be used when we have some "non-zero constraints" (compared to the zero constraints). Because the iterative conditional fitting algorithm finds the MLE for the unconstrained entries while it gaurantees positive definiteness, we can find a valid and optimal solution to the missing entries in terms of likelihood maximization.

5.3 Information Criteria to Address Misspecification for COMET

The non-zero entries of the COMET estimator can be biased due to model misspecification. An interesting avenue for research is the development of information criteria to mitigate model misspecification. Lv and Liu (2014) studied the model selection in misspecified models and extended AIC and BIC to account for the misspecification bias in generalized linear models.

Given n observations, let $\ell_n(\boldsymbol{\theta})$ be the log-likelihood function for the parameter vector $\boldsymbol{\theta}$. Lv and Liu (2014) defined the generalized AIC (GAIC) and the generalized BIC (GBIC) as below:

$$\begin{aligned}
\text{GAIC} &= -2\ell_n(\boldsymbol{\theta}) + 2\text{tr}(\hat{\boldsymbol{H}}_n) \\
\text{GBIC} &= -2\ell_n(\boldsymbol{\theta}) + \log(n) \cdot p - \log|\hat{\boldsymbol{H}}_n| \\
\text{GBIC}_p &= -2\ell_n(\boldsymbol{\theta}) + \log(n) \cdot p + \text{tr}(\hat{\boldsymbol{H}}_n) - \log|\hat{\boldsymbol{H}}_n|
\end{aligned}$$

where p is the number of predictors and $\hat{\boldsymbol{H}}_n = \hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n$ is an estimator of the so-called covariance contrast matrix $\boldsymbol{H}_n$. If the model is correctly specified, $\boldsymbol{H}_n = I_p$ so that $\text{tr}(\boldsymbol{H}_n) = p$ and $\log|\boldsymbol{H}_n| = 0$, thus the GAIC and the GBIC will be close to AIC and BIC, respectively. The estimation of $\hat{\boldsymbol{H}}_n$ will be discussed later.

The GAIC incorporates the effects of model complexity and model misspefication in a single term, $\text{tr}(\hat{\boldsymbol{H}}_n)$. Such a term can be shown to be non-negative, meaning that the term is indeed a penalty term for penalizing the complex models or misspecified models. The form of the GAIC has also been studied by others (Takeuchi, 1976; Stone, 1977). For the GAIC, Lv and Liu (2014) proposed a bootstrap estimator for $\text{tr}(\hat{\boldsymbol{H}}_n)$.

In the GBIC, the model complexity and model misspefication are reflected in separate terms. That is, the term $\log(n) \cdot p$ penalizes the complex models as in the BIC while the term $-\log|\hat{\boldsymbol{H}}_n|$ reflects the model misspecification. However, $-\log|\hat{\boldsymbol{H}}_n|$ is not always non-negative and thus is not necessarily a penalty term. On the other hand, the GBIC_p can be rearranged as below:

$$\text{GBIC}_p = -2\ell_n(\boldsymbol{\theta}) + \{1 + \log(n)\} \cdot p + 2KL\{\mathbb{N}(\mathbf{0}, \hat{\boldsymbol{B}}_n); \mathbb{N}(\mathbf{0}, \hat{\boldsymbol{A}}_n)\}$$

where $KL\{\mathbb{N}(\mathbf{0}, \hat{\boldsymbol{B}}_n); \mathbb{N}(\mathbf{0}, \hat{\boldsymbol{A}}_n)\} = (1/2)\{\text{tr}(\hat{\boldsymbol{H}}_n) - \log|\hat{\boldsymbol{H}}_n| - p\}$ is the Kullback-Leibler divergence of $\mathbb{N}(\mathbf{0}, \hat{\boldsymbol{A}}_n)$ from $\mathbb{N}(\mathbf{0}, \hat{\boldsymbol{B}}_n)$, hence a positive value. That is,

$$\text{GBIC}_p = \text{goodness of fit} + \text{model complexity} + \text{model misspecification}.$$

where both the second and the third penalty terms are all non-negative. Note that the last term for model misspecification has not been considered in our AIC and BIC for the COMET estimator. Such extension to AIC and BIC can be pursued for the selection of the threshold parameter for the COMET estimator.

REFERENCES

Abadir, K. M. and Magnus, J. R. (2005). *Matrix algebra*, volume 1. Cambridge University Press.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in probability and statistics* pages 1–24.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics* **1**, 135–141.

Aylward, E. H., Harrington, D. L., Mills, J. A., Nopoulos, P. C., Ross, C. A., Long, J. D., Liu, D., Westervelt, H. K., and Paulsen, J. S. (2013). Regional atrophy associated with cognitive and motor function in prodromal huntington disease. *Journal of Huntington's disease* **2**, 477–489.

BCBS, I. (2010). Developments in modelling risk aggregation. *Basel Committee on Banking Supervision* .

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 289–300.

Berger, M., Tutz, G., and Schmid, M. (2017). Tree-structured modelling of varying coefficients. *Statistics and Computing* pages 1–13.

Bickel, P. J., Levina, E., et al. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577–2604.

Bickel, P. J., Levina, E., et al. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.

Bien, J., Bunea, F., and Xiao, L. (2016). Convex banding of the covariance matrix. *Journal of the American Statistical Association* **111**, 834–845.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics* **41**, 1111.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**,

807–820.

Biglan, K. M., Ross, C. A., Langbehn, D. R., Aylward, E. H., Stout, J. C., Queller, S., Carlozzi, N. E., Duff, K., Beglinger, L. J., and Paulsen, J. S. (2009). Motor abnormalities in premanifest persons with huntington's disease: the predict-hd study. *Movement Disorders* **24**, 1763–1772.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* **3**, 1–122.

Bürgin, R. and Ritschard, G. (2015). Tree-based varying coefficient regression for longitudinal ordinal responses. *Computational Statistics & Data Analysis* **86**, 65–80.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences* **97**, 12182–12186.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.

Cai, T. T., Zhang, C.-H., Zhou, H. H., et al. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144.

Calibration, C. (2010). Solvency ii calibration paper (15.4. 2010). *URL: https://eiopa. europa. eu/fileadmin/tx_dam/files/publications/submissionstotheec/CEIOPS-Calibration-paper-Solvency-II. pdf* .

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94**, 199–216.

Coppen, E. M., van der Grond, J., Hafkemeijer, A., Rombouts, S. A., and Roos, R. A. (2016). Early grey matter changes in structural covariance networks in huntington's disease. *NeuroImage: Clinical* **12**, 806–814.

Degnan, A. J. and Levy, L. M. (2014). Neuroimaging of rapidly progressive dementias, part 1: neurodegenerative etiologies. *American Journal of Neuroradiology* **35**, 418–423.

Drton, M. and Perlman, M. D. (2004). Model selection for gaussian concentration graphs.

Biometrika **91**, 591–602.

Drton, M., Perlman, M. D., et al. (2007). Multiple testing and error control in gaussian graphical model selection. *Statistical Science* **22**, 430–449.

Drton, M. and Richardson, T. S. (2002). A new algorithm for maximum likelihood estimation in gaussian graphical models for marginal independence. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 184–191. Morgan Kaufmann Publishers Inc.

Du, W. and Tibshirani, R. (2018). A pliable lasso for the Cox model. *arXiv preprint arXiv:1807.06770* .

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

El Karoui, N. et al. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* **36**, 2717–2756.

El Karoui, N. et al. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *The Annals of Statistics* **38**, 3487–3566.

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

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19**, C1–C32.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736* .

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis* **98**, 227–255.

Garcia, T. P. and Müller, S. (2014). Influence of measures of significance based weights in the

weighted lasso. *Journal of the Indian Society of Agricultural Statistics* **68**, 131–144.

Garcia, T. P., Müller, S., Carroll, R. J., Dunn, T. N., Thomas, A. P., Adams, S. H., Pillai, S. D., and Walzem, R. L. (2013). Structured variable selection with q-values. *Biostatistics* **14**, 695–707.

Garcia, T. P., Müller, S., et al. (2016). Cox regression with exclusion frequency-based weights to identify neuroimaging markers relevant to Huntington's disease onset. *The Annals of Applied Statistics* **10**, 2130–2156.

Georgescu, D. I., Higham, N. J., and Peters, G. W. (2018). Explicit solutions to correlation matrix completion problems, with an application to risk management and insurance. *Royal Society open science* **5**, 172348.

Gertheiss, J. and Tutz, G. (2012). Regularization and model selection with categorial effect modifiers. *Statistica Sinica* **22**, 957–982.

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396. ACM.

Haris, A., Witten, D., and Simon, N. (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics* **25**, 981–1004.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 757–796.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hsu, H.-L., Ing, C.-K., Tong, H., et al. (2019). On model selection from a finite family of possibly misspecified time series models. *Annals of Statistics* **47**, 1061–1087.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5.

Prentice hall Upper Saddle River, NJ.

Kauermann, G. (1996). On a dualization of graphical gaussian models. *Scandinavian journal of statistics* pages 105–116.

Kieburtz, K., Penney, J. B., Corno, P., Ranen, N., Shoulson, I., Feigin, A., Abwender, D., Greenarnyre, J. T., Higgins, D., Marshall, F. J., et al. (2001). Unified huntington's disease rating scale: reliability and consistency. *Neurology* **11**, 136–142.

Kubat, M., Holte, R. C., and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* **30**, 195–215.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics* **37**, 4254.

Li, D. and Zou, H. (2016). Sure information criteria for large covariance matrix estimation and their asymptotic properties. *IEEE Transactions on Information Theory* **62**, 2153–2169.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics* **24**, 627–654.

Liu, H., Wang, L., and Zhao, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics* **23**, 439–459.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pages 141–167.

Ma, S. and Song, P. X.-K. (2015). Varying index coefficient models. *Journal of the American Statistical Association* **110**, 341–356.

Milovanovic, N., Damjanovic, A., Milovanovic, S., Duisin, D., Malis, M., Stankovic, G., Rankovic, A., Latas, M., F Filipovic, B., and R Filipovic, B. (2018). Reliability of the bi-caudate parameter in the revealing of the enlarged lateral ventricles in schizophrenia patients. *Psychiatria Danubina* **30**, 150–156.

Minkova, L., Eickhoff, S. B., Abdulkadir, A., Kaller, C. P., Peter, J., Scheller, E., Lahr, J., Roos, R. A., Durr, A., Leavitt, B. R., et al. (2016). Large-scale brain network abnormalities in h untington's disease revealed by structural covariance. *Human brain mapping* **37**, 67–80.

Misiura, M. B., Lourens, S., Calhoun, V. D., Long, J., Bockholt, J., Johnson, H., Zhang, Y., Paulsen, J. S., Turner, J. A., Liu, J., et al. (2017). Cognitive control, learning, and clinical motor ratings are most highly associated with basal ganglia brain volumes in the premanifest huntington's disease phenotype. *Journal of the International Neuropsychological Society* **23**, 159–170.

Monahan, J. F. (2008). *A primer on linear models*. CRC Press.

Na, S., Yang, Z., Wang, Z., and Kolar, M. (2019). High-dimensional varying index coefficient models via stein's identity. *Journal of Machine Learning Research* **20**, 1–44.

Oelker, M.-R., Gertheiss, J., and Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling* **14**, 157–177.

Paulsen, J. S., Long, J. D., Johnson, H. J., Aylward, E. H., Ross, C. A., Williams, J. K., Nance, M. A., Erwin, C. J., Westervelt, H. K., Harrington, D. L., et al. (2014a). Clinical and biomarker changes in premanifest huntington disease show trial feasibility: a decade of the predict-hd study. *Frontiers in aging neuroscience* **6**, 78.

Paulsen, J. S., Long, J. D., Ross, C. A., Harrington, D. L., Erwin, C. J., Williams, J. K., Westervelt, H. J., Johnson, H. J., Aylward, E. H., Zhang, Y., et al. (2014b). Prediction of manifest huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology* **13**, 1193–1201.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons.

Qiu, Y. and Liyanage, J. S. (2019). Threshold selection for covariance estimation. *Biometrics* **75**, 895–905.

Reilmann, R., Leavitt, B. R., and Ross, C. A. (2014). Diagnostic criteria for huntington's disease based on natural history. *Movement Disorders* **29**, 1335–1341.

Reiner, A., Dragatsis, I., and Dietrich, P. (2011). Genetics and neuropathology of huntington's disease. *International review of neurobiology* **98**, 325–372.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Rodrigues, F. B., Byrne, L. M., Tortelli, R., Johnson, E. B., Wijeratne, P. A., Arridge, M., De Vita, E., Ghazaleh, N., Houghton, R., Furby, H., et al. (2020). Longitudinal dynamics of mutant huntingtin and neurofilament light in huntington's disease: the prospective hd-csf study. *medRxiv* .

Rodrigues, F. B. and Wild, E. J. (2018). Huntington's disease clinical trials corner: August 2018. *Journal of Huntington's disease* **7**, 279–286.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177–186.

She, Y., Wang, Z., and Jiang, H. (2018). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association* **113**, 445–454.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 44–47.

Tabrizi, S. J., Leavitt, B. R., Landwehrmeyer, G. B., Wild, E. J., Saft, C., Barker, R. A., Blair, N. F., Craufurd, D., Priller, J., Rickards, H., et al. (2019). Targeting huntingtin expression in patients with huntington's disease. *New England Journal of Medicine* **380**, 2307–2316.

Tabrizi, S. J., Reilmann, R., Roos, R. A., Durr, A., Leavitt, B., Owen, G., Jones, R., Johnson, H., Craufurd, D., Hicks, S. L., et al. (2012). Potential endpoints for clinical trials in premanifest and early huntington's disease in the track-hd study: analysis of 24 month observational data. *The Lancet Neurology* **11**, 42–53.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science* **153**, 12–18.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*

Statistical Society. Series B (Methodological) pages 267–288.

Tibshirani, R. and Friedman, J. (2019). A pliable lasso. *Journal of Computational and Graphical Statistics* pages 1–11.

Wang, J. C. and Hastie, T. (2014). Boosted varying-coefficient regression models for product demand prediction. *Journal of Computational and Graphical Statistics* **23**, 361–382.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* **103**, 1556–1569.

Watson, G. (1963). A note on maximum likelihood. Technical report, JOHNS HOPKINS UNIV BALTIMORE MD.

Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica* **21**, 1515.

Wen, F., Yang, Y., Liu, P., and Qiu, R. C. (2016). Positive definite estimation of large covariance matrix using generalized nonconvex penalties. *IEEE Access* **4**, 4168–4182.

Wermuth, N., Cox, D. R., Marchetti, G. M., et al. (2006). Covariance chains. *Bernoulli* **12**, 841–862.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society* pages 1–25.

Williams, M. (2016). An introduction to the caudate in schizophrenia. *Oruen - The CNS Journal* **2**, 40–42.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–844.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite ℓ1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association* **107**, 1480–1491.

Yu, D., Zhang, X., and Yau, K. K. (2018). Asymptotic properties and information criteria for misspecified generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 817–836.

Yuan, M., Joseph, V. R., Zou, H., et al. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics* **3**, 1738–1757.

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**, 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* **38**, 894–942.

Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W., Paulsen, J. S., Investigators, P.-H., and of the Huntington Study Group, C. (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156**, 751–763.

Zhao, P., Rocha, G., Yu, B., et al. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37**, 3468–3497.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research* **7**, 2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association* **112**, 266–281.

Zwiernik, P., Uhler, C., and Richards, D. (2017). Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1269–1292.

APPENDIX A

OPTIMIZATION OF THE SVREG*

Consider the case where there are L groups of p main predictors and G groups of K modifying variables. The index for each group of main predictors is denoted by $\ell = 1, 2, \ldots, L$ and the index for each group of modifying variables is denoted by $g = 1, 2, \ldots, G$. Assuming no intercept terms for simplicity, the objective function of the structural varying-coefficient regression is

$$J^*(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N} \sum_{i=1}^{N} \left[y_i - \sum_{\ell=1}^{L} \left\{ \mathbf{x}_{i[\ell]} (\boldsymbol{\beta}_{[\ell]} + \boldsymbol{\theta}_{[\ell]\bullet} \mathbf{z}_{i\bullet}^T) \right\} \right]^2 + \lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}),$$

where $\mathbf{z}_{i\bullet}$ is the i-th row of $\mathbf{Z}$, $\mathbf{x}_{i[\ell]}$ is the ℓ-th group of the main predictors for the i-th row of $\mathbf{X}$, $\boldsymbol{\beta}_{[\ell]}$ is a subset of $\boldsymbol{\beta}$ for the ℓ-th group of the main predictors, $\boldsymbol{\theta}_{[\ell]\bullet}$ is a subset of $\boldsymbol{\Theta}$ for the ℓ-th group of the main predictors and

$$\begin{aligned} \lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}) = (1-\alpha)\lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \left\{ ||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2 + \sum_{g=1}^{G} \frac{\sqrt{p_g}}{\sqrt{1+K}} ||\text{vec}(\boldsymbol{\theta}_{[\ell][g]})||_2 \right\} \\ + \alpha\lambda \sum_{j,k} |\theta_{jk}|_1, \end{aligned}$$

where p_ℓ is the size of the ℓ-th group of the main predictors, p_g is the size of the g-th group of the modifying variables, $\boldsymbol{\theta}_{[\ell][g]}$ is a subset of $\boldsymbol{\Theta}$ for the ℓ-th group of the main predictors and the g-th group of the modifying variables and $\text{vec}(\cdot)$ is a vectorization operator.

The first step in the optimization is computing the subgradient equations of the objective function, which we will set to zero. For $g = 1, 2, \ldots, G$, denoting $r_i = y_i - \sum_{\ell=1}^{L} \mathbf{x}_{i[\ell]} (\boldsymbol{\beta}_{[\ell]} + \boldsymbol{\theta}_{[\ell]\bullet} \mathbf{z}_{i\bullet}^T)$,

*Parts of this section have been modified with permission from [R. Kim, S. Müller and T. Garcia. svReg: Structural Varying-coefficient regression to differentiate how regional brain atrophy affects motor impairment for Huntington disease severity groups. Biometrical Journal. 2021. Volume 63. Pages 1254-1271. (https://doi.org/10.1002/bimj.202000312) Copyright Wiley-VCH GmbH. Reproduced with permission]

the subgradient equations are

$$\frac{dJ^*}{d\boldsymbol{\beta}_{[\ell]}} = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T r_i + (1-\alpha)\lambda\sqrt{p_\ell}\mathbf{u}$$
$$\frac{dJ^*}{d\boldsymbol{\theta}_{[\ell][g]}} = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T\mathbf{z}_{i[g]} r_i + (1-\alpha)\lambda\sqrt{p_\ell}(\mathbf{u}_2^{(g)} + \frac{\sqrt{p_g}}{\sqrt{1+K}}\mathbf{u}_3^{(g)}) + \alpha\lambda\mathbf{v}^{(g)}$$

where $\mathbf{z}_{i[g]}$ denotes the g-th group of modifying variables and

$$\begin{aligned}
\mathbf{u} &= \boldsymbol{\beta}_{[\ell]}/||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2 \quad \text{if} \quad (\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet})) \neq 0 \\
&\in \{\mathbf{u} : ||\mathbf{u}||_2 \leq 1\} \quad \text{if} \quad (\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet})) = 0 \\
\mathbf{u}_2^{(g)} &= \boldsymbol{\theta}_{[\ell][g]}/||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2 \quad \text{if} \quad (\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet})) \neq 0 \\
&\in \{\mathbf{u} : ||\text{vec}(\mathbf{u})||_2 \leq 1\} \quad \text{if} \quad (\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet})) = 0 \\
\mathbf{u}_3^{(g)} &= \boldsymbol{\theta}_{[\ell][g]}/||\text{vec}(\boldsymbol{\theta}_{[\ell][g]})||_2 \quad \text{if} \quad \text{vec}(\boldsymbol{\theta}_{[\ell][g]}) \neq 0 \\
&\in \{\mathbf{u} : ||\text{vec}(\mathbf{u})||_2 \leq 1\} \quad \text{if} \quad \text{vec}(\boldsymbol{\theta}_{[\ell][g]}) = 0 \\
\mathbf{v}^{(g)} &\in \text{sign}(\boldsymbol{\theta}_{[\ell][g]})
\end{aligned}$$

Define $r_i^{(-j)}$, the partial residual for j-th coordinate and $r_i^{(-j)(-g)}$, the partial residual for g-th group of modifying variables in the j-th coordinate as below:

$$r_i^{(-\ell)} = y_i - \sum_{h\neq\ell}\left\{\mathbf{x}_{i[h]}(\boldsymbol{\beta}_{[h]} + \boldsymbol{\theta}_{[h]\bullet}\mathbf{z}_{i\bullet}^T)\right\}$$
$$r_i^{(-\ell)(-g)} = r_i^{(-\ell)} - \mathbf{x}_{i[\ell]}\sum_{m\neq g}\boldsymbol{\theta}_{[\ell][m]}\mathbf{z}_{i[m]}^T,$$

where $\mathbf{z}_{i[g]}$ denotes a subset of $\mathbf{z}_{i\bullet}$ for the g-th group of the modifying variables.

Then the objective function can be rewritten as below:

$$\begin{aligned}
J^*(\beta_0, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \boldsymbol{\Theta}) &= \frac{1}{2N}\sum_{i=1}^{N}\left\{r_i^{(-\ell)} - \mathbf{x}_{i[\ell]}(\boldsymbol{\beta}_{[\ell]} + \boldsymbol{\theta}_{[\ell]\bullet}\mathbf{z}_{i\bullet}^T)\right\}^2 + \lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}) \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left[r_i^{(-\ell)} - \mathbf{x}_{i[\ell]}\left\{\beta_{[\ell]} + \sum_{g=1}^{G}\boldsymbol{\theta}_{[\ell][g]}\mathbf{z}_{i[g]}^T\right\}\right]^2 + \lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}) \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left[r_i^{(-\ell)(-g)} - \mathbf{x}_{i[\ell]}\left\{\beta_{[\ell]} + \boldsymbol{\theta}_{[\ell][g]}\mathbf{z}_{i[g]}^T\right\}\right]^2 + \lambda P_\alpha^*(\boldsymbol{\beta}, \boldsymbol{\Theta}).
\end{aligned}$$

Since the minimizer $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}})$ of this objective function should satisfy $\partial J^*/\partial \boldsymbol{\beta}_{[\ell]} = 0$ and $\partial J^*/\partial \boldsymbol{\theta}_{[\ell][g]} = 0$, the following equations hold for all $g = 1, \ldots, G$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T\mathbf{x}_{i[\ell]}(\hat{\boldsymbol{\beta}}_{[\ell]} + \hat{\boldsymbol{\theta}}_{[\ell]\bullet}\mathbf{z}_{i\bullet}^T) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T r_i^{(-\ell)} - (1-\alpha)\lambda\sqrt{p_\ell}\mathbf{u} \\
\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T\mathbf{x}_{i[\ell]}(\hat{\boldsymbol{\beta}}_{[\ell]} + \hat{\boldsymbol{\theta}}_{[\ell][g]}\mathbf{z}_{i[g]}^T)\mathbf{z}_{i[g]} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T\mathbf{z}_{i[g]} r_i^{(-j)(-g)} \\
&\quad - (1-\alpha)\lambda\sqrt{p_\ell}(\mathbf{u}_2^{(g)} + \frac{\sqrt{p_g}}{\sqrt{1+K}}\mathbf{u}_3^{(g)}) - \alpha\lambda\mathbf{v}^{(g)}
\end{aligned}$$

Hence, $(\hat{\beta}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell]\bullet}) = 0$ if and only if $(\hat{\beta}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell][g]}) = 0$ for all $g = 1, \ldots, G$ if

$$\begin{aligned}
\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T r_i^{(-\ell)}\right\|_2 &\le (1-\alpha)\lambda\sqrt{p_\ell}, \text{and} \\
\left\|S_{\alpha\lambda}\left\{\frac{1}{N}\sum_{i=1}^{N}\text{vec}(\mathbf{x}_{i[\ell]}^T\mathbf{z}_{i[g]})r_i^{(-j)(-g)}\right\}\right\|_2 &\le (1 + \frac{\sqrt{p_g}}{\sqrt{1+K}})(1-\alpha)\lambda\sqrt{p_\ell}
\end{aligned}$$

where $S_\lambda(\mathbf{x})$ is a component-wise soft-thresholding operator.

If $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell]\bullet}) \neq 0$, we need to check if $\hat{\boldsymbol{\beta}}_{[\ell]} \neq 0$ and $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$. For this, we first calculate $\hat{\boldsymbol{\beta}}_{[\ell]}$ assuming $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$. For an orthogonal design, that is, if $\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T\mathbf{x}_{i[\ell]}/N = I$, the subgradient

equation $\partial J^*/\partial \boldsymbol{\beta}_{[\ell]} = 0$ given $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ can be reduced as below:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{[\ell]} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{i[\ell]}^T r_i^{(-\ell)} - (1-\alpha)\lambda\sqrt{p_\ell}\mathbf{u} \\ &= \max\left\{1 - \frac{(1-\alpha)\lambda\sqrt{p_\ell}}{||\sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T r_i^{(-\ell)}/N||_2}, 0\right\} \cdot \sum_{i=1}^{N}\mathbf{x}_{i[\ell]}^T r_i^{(-\ell)}/N\end{aligned}$$

In general, however, there is no closed form solution of $\hat{\boldsymbol{\beta}}_{[\ell]}$ given $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ and the solution can be found by sequential optimization of of each parameter in $\boldsymbol{\beta}_{[\ell]}$ using the `optimize` function in the R package as proposed by Friedman et al. (2010). Once the solution of $\hat{\boldsymbol{\beta}}_{[\ell]}$ given $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ is computed, we check if $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} = 0$ using the calculated $\hat{\boldsymbol{\beta}}_{[\ell]}$. From $\partial J^*/\partial \boldsymbol{\theta}_{[\ell][g]} = 0$, we check $\hat{\boldsymbol{\theta}}_{[\ell][g]} = 0$ given $\hat{\boldsymbol{\beta}}_{[\ell]}$ by checking the following conditions for all $g = 1, \ldots, G$.

$$\left\| S_{\alpha\lambda}\left\{\frac{1}{N}\sum_{i=1}^{N}\text{vec}(\mathbf{x}_{i[\ell]}^T\mathbf{z}_{i[g]})(r_i^{(-\ell)(-g)} - \mathbf{x}_{i[\ell]}\hat{\boldsymbol{\beta}}_{[\ell]})\right\}\right\|_2 < (1-\alpha)\lambda\frac{\sqrt{p_g p_\ell}}{\sqrt{1+K}}$$

If $\hat{\boldsymbol{\beta}}_{[\ell]} \neq 0$ and $\hat{\boldsymbol{\theta}}_{[\ell]\bullet} \neq 0$ (i.e. if there exists g^* such that $\hat{\boldsymbol{\theta}}_{[\ell][g^*]} \neq 0$), we use a generalized gradient procedure to find $(\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell][NZ]})$ where $\hat{\boldsymbol{\theta}}_{[\ell][NZ]}$ denotes the collection of nonzero $\hat{\boldsymbol{\theta}}_{[\ell][g]}$'s. This procedure is described below.

Let $\gamma_{[\ell]} = (\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet})) = (\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell][1]}), \ldots, \text{vec}(\boldsymbol{\theta}_{[\ell][G]}))$ and let $\ell(\gamma_{[\ell]})$ be the likelihood part of the objective function $J^*(\boldsymbol{\beta}, \boldsymbol{\Theta})$.

$$\ell(\gamma_{[\ell]}) = \frac{1}{2N}\sum_{i=1}^{N}\left[y_i - \sum_{\ell=1}^{L}\left\{\mathbf{x}_{i[\ell]}(\boldsymbol{\beta}_{[\ell]} + \boldsymbol{\theta}_{[\ell]\bullet}\mathbf{z}_{i\bullet}^T)\right\}\right]^2.$$

The goal of this procedure is to minimize $\ell(\gamma_{[\ell]}) + \lambda P_\alpha^*(\gamma_{[\ell]})$ in terms of $\boldsymbol{\beta}_{[\ell]}$ and nonzero $\boldsymbol{\theta}_{[\ell][g]}$'s where

$$\lambda P_\alpha^*(\gamma_{[\ell]}) = (1-\alpha)\lambda\sqrt{p_\ell}\left\{||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2 + \sum_{g=1}^{G}\frac{\sqrt{p_g}}{\sqrt{1+K}}||\text{vec}(\boldsymbol{\theta}_{[\ell][g]})||_2\right\} + \alpha\lambda\sum_{j\in[\ell],k}|\theta_{jk}|_1$$

for each coordinate ℓ using cyclic coordinate descent algorithm. Here, we use a majorization-

minimization(MM) algorithm for optimization. In MM algorithm, the objective function is majorized with a surrogate function and this surrogate function is minimized instead of the objective function. Let $\tilde{\gamma}_{[\ell]} = (\tilde{\boldsymbol{\beta}}_{[\ell]}, \text{vec}(\tilde{\boldsymbol{\theta}}_{[\ell][1]}), \ldots, \text{vec}(\tilde{\boldsymbol{\theta}}_{[\ell][G]}))$ be the current value of $\gamma_{[\ell]}$ and define $M(\gamma_{[\ell]})$, the surrogate function for $\ell(\gamma_{[\ell]}) + \lambda P_\alpha^*(\gamma_{[\ell]})$ as below.

$$M(\gamma_{[\ell]}) = \ell(\tilde{\gamma}_{[\ell]}) + (\gamma_{[\ell]} - \tilde{\gamma}_{[\ell]})^T \nabla_{\gamma_{[\ell]}} \ell(\tilde{\gamma}_{[\ell]}) + \frac{1}{2t} ||\gamma_{[\ell]} - \tilde{\gamma}_{[\ell]}||_2^2 + \lambda P_\alpha^*(\gamma_{[\ell]})$$

where t is the learning rate and should be sufficiently small to guarantee convergence. $\nabla_{\gamma_{[\ell]}} \ell(\tilde{\gamma}_{[\ell]})$ is the gradient of ℓ with respect to $\gamma_{[\ell]}$.

Minimizing $M(\gamma_{[\ell]})$ is equivalent to minimizing

$$\tilde{M}(\gamma_{[\ell]}) = \frac{1}{2t} ||\gamma_{[\ell]} - \tilde{\gamma}_{[\ell]} + t \nabla_{\gamma_{[\ell]}} \ell(\tilde{\gamma}_{[\ell]})||_2^2 + \lambda P_\alpha^*(\gamma_{[\ell]}).$$

From $\partial \tilde{M} / \partial \boldsymbol{\beta}_{[\ell]} = 0$ and $\partial \tilde{M} / \partial \boldsymbol{\theta}_{[\ell][g]} = 0$ for nonzero $\boldsymbol{\theta}_{[\ell][g]}$'s,

$$\left\{ 1 + \frac{t(1-\alpha)\lambda\sqrt{p_\ell}}{||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2} \right\} \hat{\boldsymbol{\beta}}_{[\ell]} = \tilde{\boldsymbol{\beta}}_{[\ell]} - t \nabla_{\boldsymbol{\beta}_{[\ell]}} \ell(\tilde{\gamma_{[\ell]}})$$

$$\left\{ 1 + \frac{t(1-\alpha)\lambda\sqrt{p_\ell}}{||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2} + \frac{\sqrt{p_g}}{\sqrt{1+K}} \frac{t(1-\alpha)\lambda\sqrt{p_\ell}}{||\text{vec}(\hat{\boldsymbol{\theta}}_{[\ell][g]})||_2} \right\} \hat{\boldsymbol{\theta}}_{[\ell][g]} = S_{t\alpha\lambda} \left\{ \tilde{\boldsymbol{\theta}}_{[\ell][g]} - t \nabla_{\boldsymbol{\theta}_{[\ell][g]}} \ell(\tilde{\gamma_{[\ell]}}) \right\}$$

for nonzero $\boldsymbol{\theta}_{[\ell][g]}$. Note $||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2^2$ is equal to the sum of $||\hat{\boldsymbol{\beta}}_{[\ell]}||_2^2$ and $||\text{vec}(\hat{\boldsymbol{\theta}}_{[\ell][g]})||_2^2$'s for nonzero $\boldsymbol{\theta}_{[\ell][g]}$'s.

Let $a = ||\hat{\boldsymbol{\beta}}_{[\ell]}||_2, b_g = ||\text{vec}(\hat{\boldsymbol{\theta}}_{[\ell][g]})||_2$. Take the norm of both sides in each equation above giving

$$\left\{ 1 + t(1-\alpha)\lambda\sqrt{p_\ell} \frac{1}{\sqrt{a^2 + \sum_{g=1}^G b_g^2}} \right\} a = \|\tilde{\boldsymbol{\beta}}_{[\ell]} - t \nabla_{\boldsymbol{\beta}_{[\ell]}} \ell(\tilde{\gamma_{[\ell]}})\|_2$$

$$\left\{ 1 + t(1-\alpha)\lambda\sqrt{p_\ell} \left(\frac{1}{\sqrt{a^2 + \sum_{g=1}^G b_g^2}} + \frac{\sqrt{p_g}}{\sqrt{1+K}} \frac{1}{b_g} \right) \right\} b_g = \left\| \text{vec} \left[S_{t\alpha\lambda} \left\{ \tilde{\boldsymbol{\theta}}_{[\ell][g]} - t \nabla_{\boldsymbol{\theta}_{[\ell][g]}} \ell(\tilde{\gamma_{[\ell]}}) \right\} \right] \right\|_2$$

for g such that $\boldsymbol{\theta}_{[\ell][g]} \neq 0$. Also, let $r = ||(\boldsymbol{\beta}_{[\ell]}, \text{vec}(\boldsymbol{\theta}_{[\ell]\bullet}))||_2 = \sqrt{a^2 + \sum_{g=1}^{G} b_g^2}$, $c = t(1-\alpha)\lambda\sqrt{p_\ell}$, $h_0 = \|\tilde{\boldsymbol{\beta}}_{[\ell]} - t\nabla_{\boldsymbol{\beta}_{[\ell]}}\ell(\tilde{\gamma_{[\ell]}})\|_2$ and $h_g = \left\| \text{vec}\left[S_{t\alpha\lambda}\left\{ \tilde{\boldsymbol{\theta}}_{[\ell][g]} - t\nabla_{\boldsymbol{\theta}_{[\ell][g]}}\ell(\tilde{\gamma_{[\ell]}}) \right\} \right] \right\|_2$ for g such that $\boldsymbol{\theta}_{[\ell][g]} \neq 0$. Then r satisfies the following quadratic equation.

$$r^2 + 2cr + c^2 - h_0^2 + \sum_{g:\boldsymbol{\theta}_{[\ell][g]}\neq 0} (2ch_g \frac{\sqrt{p_g}}{\sqrt{1+K}} - h_g^2 - c^2\frac{p_g}{1+K}) = 0.$$

Let r^* be the positive root of the above equation. Then

$$\begin{aligned}
\hat{a} &= \frac{h_0 r^*}{r^* + c} \\
\hat{b}_g &= \frac{(h_g - c\frac{\sqrt{p_g}}{\sqrt{1+K}})r^*}{r^* + c} (\text{for } g \text{ such that } \boldsymbol{\theta}_{[\ell][g]} \neq 0).
\end{aligned}$$

If we plug $\hat{a}$ and $\hat{b}_g$'s in the gradient equation of $\tilde{M}$, the solutions $\hat{\boldsymbol{\beta}}_{[\ell]}, \hat{\boldsymbol{\theta}}_{[\ell][g]}$ satisfy

$$\left\{ 1 + t(1-\alpha)\lambda \frac{1}{\sqrt{\hat{a}^2 + \sum_{g=1}^{G}\hat{b}_g^2}} \right\} \hat{\boldsymbol{\beta}}_{[\ell]} = \tilde{\boldsymbol{\beta}}_{[\ell]} - t\nabla_{\boldsymbol{\beta}_{[\ell]}}\ell(\tilde{\gamma_{[\ell]}})$$

$$\left\{ 1 + t(1-\alpha)\lambda \left(\frac{1}{\sqrt{\hat{a}^2 + \sum_{g=1}^{G}\hat{b}_g^2}} + \frac{\sqrt{p_g}}{\sqrt{1+K}}\frac{1}{\hat{b}_g} \right) \right\} \hat{\boldsymbol{\theta}}_{[\ell][g]} = S_{t\alpha\lambda}\left\{ \tilde{\boldsymbol{\theta}}_{[\ell][g]} - t\nabla_{\boldsymbol{\theta}_{[\ell][g]}}\ell(\tilde{\gamma_{[\ell]}}) \right\}$$

for g such that $\boldsymbol{\theta}_{[\ell][g]} \neq 0$. Letting c_1, c_2 be the constants multiplying $\hat{\boldsymbol{\beta}}_{[\ell]}$ and $\hat{\boldsymbol{\theta}}_{[\ell][g]}$ above, we have the update equations

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{[\ell]} &= \frac{\tilde{\boldsymbol{\beta}}_{[\ell]} - t\nabla_{\boldsymbol{\beta}_{[\ell]}}\ell(\tilde{\gamma_{[\ell]}})}{c_1} \\
\hat{\boldsymbol{\theta}}_{[\ell][g]} &= \frac{S_{t\alpha\lambda}\left\{ \tilde{\boldsymbol{\theta}}_{[\ell][g]} - t\nabla_{\boldsymbol{\theta}_{[\ell][g]}}\ell(\tilde{\gamma_{[\ell]}}) \right\}}{c_2}
\end{aligned}$$

APPENDIX B

TECHNICAL PROOFS

B.1 Proof of Theorem 1

We start from the following Lemma which establishes the consistency of the solution computed from iterative conditional fitting.

Lemma 1. *Let $\widehat{\boldsymbol{\sigma}}$ be a solution to the normal likelihood equation computed from iterative conditional fitting with a consistent estimator of $\boldsymbol{\Sigma}$ as the starting value. Then, $\widehat{\boldsymbol{\sigma}}$ is a consistent estimator of $\boldsymbol{\sigma}$.*

Proof of Lemma 1: Suppose the joint distribution of $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)^T$ is fixed with a known covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{-j,-j}$ and $\widetilde{\boldsymbol{\Sigma}}_{-j,-j}$ is a consistent estimator of the covariance matrix of Y_{-j}. When the j-th column of the iterative conditional fitting estimator is updated, the conditional likelihood $L(\sigma_{jj}, \boldsymbol{\Sigma}_{-j,j} | \widetilde{\boldsymbol{\Sigma}}_{-j,-j})$ in equation (2) in the main manuscript is maximized by updating $\boldsymbol{\Sigma}_{-j,j}$ and σ_{jj} with equations (3.18) and (3.19), respectively, given the location of the zero entries in $\boldsymbol{\Sigma}_{-j,j}$. Let $\hat{\sigma}_{jj}$ and $\widehat{\boldsymbol{\Sigma}}_{-j,j}$ be such maximizer of the conditional likelihood computed from the iterative conditional fitting algorithm. To prove Lemma 1, it is sufficient to show that $\hat{\sigma}_{jj}$ and $\widehat{\boldsymbol{\Sigma}}_{-j,j}$ are consistent estimators of σ_{jj} and $\boldsymbol{\Sigma}_{-j,j}$, respectively.

Consider a $n \times p$ design matrix $\boldsymbol{Y} = (y_{ij})_{i=1,j=1}^{n,p}$ for n independent observations of a p-dimensional random vector $(Y_1, \ldots, Y_p)^T \sim \mathbb{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Define sets of indices $V = \{1, \ldots, p\}$, $sp(j) = \{k \in V : \sigma_{jk} \neq 0, k \neq j\}$ and $nsp(j) = \{k \in V : \sigma_{jk} = 0\}$ so that $\{j\} \cup sp(j) \cup nsp(j) = V$. That is, $sp(j)$ and $nsp(j)$ designate the location of the non-zero and zero off-diagonal entries of the j-th column of $\boldsymbol{\Sigma}$, respectively. Let $\boldsymbol{Y}^{(j)}$ and $\boldsymbol{Y}^{(-j)}$ denote the columns in Y for Y_j and $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)^T$, respectively. Then, we construct the matrix $\mathbf{Z}^j_{sp(j)} = \boldsymbol{Y}^{(-j)}(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}_{\bullet,sp(j)}$ where $\bullet$ represents all indices. For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with equal dimension, we will denote $\boldsymbol{A} \xrightarrow{p} \boldsymbol{B}$ if the (i,j)-th element of $\boldsymbol{A}$ converges in probability to the (i,j)-th element of $\boldsymbol{B}$ for all i and j.

The iterative conditional fitting algorithm maximizes $L(\sigma_{jj}, \boldsymbol{\Sigma}_{-j,j}|\widetilde{\boldsymbol{\Sigma}}_{-j,-j})$ under the constraint $\boldsymbol{\Sigma}_{nsp(j),j} = \mathbf{0}$. Hence, the estimation of $\boldsymbol{\Sigma}_{-j,j}$ boils down to estimating $\boldsymbol{\Sigma}_{sp(j),j}$. Based on the least squares regression of $\boldsymbol{Y}_j$ on $\mathbf{Z}^j_{sp(j)}$, the iterative conditional fitting algorithm updates $\boldsymbol{\Sigma}_{sp(j),j}$ by equation (3.18) as below.

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{sp(j),j} &= \{(\mathbf{Z}^j_{sp(j)})^T(\mathbf{Z}^j_{sp(j)})/n\}^{-1}(\mathbf{Z}^j_{sp(j)})^T\boldsymbol{Y}^{(j)}/n \\ &= [\{(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}_{\bullet,sp(j)}\}^T\{(\boldsymbol{Y}^{(-j)})^T\boldsymbol{Y}^{(-j)}/n\}\{(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}_{\bullet,sp(j)}\}]^{-1}\{(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}_{\bullet,sp(j)}\}^T\{(\boldsymbol{Y}^{(-j)})^T\boldsymbol{Y}^{(j)}/n\}.\end{aligned}$$

By the law of large numbers, $(\boldsymbol{Y}^{(-j)})^T\boldsymbol{Y}^{(-j)}/n \xrightarrow{p} \boldsymbol{\Sigma}_{-j,-j}$ and $(\boldsymbol{Y}^{(-j)})^T\boldsymbol{Y}^{(j)}/n \xrightarrow{p} \boldsymbol{\Sigma}_{-j,j}$. Also, by the consistency of $\widetilde{\boldsymbol{\Sigma}}_{-j,-j}$, $\widetilde{\boldsymbol{\Sigma}}_{-j,-j} \xrightarrow{p} \boldsymbol{\Sigma}_{-j,-j}$. For a sequence of square matrices $\boldsymbol{A}_n$, $\lim_{n\to\infty}\boldsymbol{A}_n^{-1} = \boldsymbol{A}^{-1}$ if $\lim_{n\to\infty}\boldsymbol{A}_n = \boldsymbol{A}$ by the continuous mapping theorem. Hence,

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{sp(j),j} &\xrightarrow{p} [\{(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{\bullet,sp(j)}\}^T\boldsymbol{\Sigma}_{-j,-j}(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{\bullet,sp(j)}]^{-1}\{(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{\bullet,sp(j)}\}^T\boldsymbol{\Sigma}_{-j,j} \\ &= \{(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\}^{-1}[\{(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\}^T, \{(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{nsp(j),sp(j)}\}^T]\begin{bmatrix}\boldsymbol{\Sigma}_{sp(j),j}\\ \boldsymbol{\Sigma}_{nsp(j),j}\end{bmatrix} \\ &= \{(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\}^{-1}(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\boldsymbol{\Sigma}_{sp(j),j} \\ &= \boldsymbol{\Sigma}_{sp(j),j}.\end{aligned}$$

The first equality above used the relation that, for a symmetric partitioned matrix $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}^T\\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}$,

$$\begin{bmatrix}\boldsymbol{A}^T & \boldsymbol{B}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}^T\\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix} = \boldsymbol{A},$$

and rearranged the elements of $\boldsymbol{\Sigma}_{-j,j}$ so that $\boldsymbol{\Sigma}^T_{-j,j} = [\boldsymbol{\Sigma}^T_{sp(j),j}, \boldsymbol{\Sigma}^T_{nsp(j),j}]$ without loss of generality. The second equality above holds because all elements in $\boldsymbol{\Sigma}_{nsp(j),j}$ are zero. Since elements of $\widehat{\boldsymbol{\Sigma}}_{nsp(j),j}$ are set to be zero by the iterative conditional fitting algorithm, this proves $\widehat{\boldsymbol{\Sigma}}_{-j,j} \xrightarrow{p} \boldsymbol{\Sigma}_{-j,j}$.

Next, in maximizing $L(\sigma_{jj}, \boldsymbol{\Sigma}_{-j,j}|\widetilde{\boldsymbol{\Sigma}}_{-j,-j})$, the iterative conditional fitting algorithm updates

σ_{jj} by equation (3.19) and its consistency is shown as below.

$$\begin{aligned}
\hat{\sigma}_{jj} &= (\boldsymbol{Y}^{(j)} - \mathbf{Z}^j_{sp(j)}\widehat{\boldsymbol{\Sigma}}_{sp(j),j})^T(\boldsymbol{Y}^{(j)} - \mathbf{Z}^j_{sp(j)}\widehat{\boldsymbol{\Sigma}}_{sp(j),j})/n + (\widehat{\boldsymbol{\Sigma}}_{sp(j),j})^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}_{sp(j),sp(j)}\widehat{\boldsymbol{\Sigma}}_{sp(j),j} \\
&= (\boldsymbol{Y}^{(j)})^T\boldsymbol{Y}^{(j)}/n - 2(\widehat{\boldsymbol{\Sigma}}_{sp(j),j})^T(\mathbf{Z}^j_{sp(j)})^T\boldsymbol{Y}^{(j)}/n + (\widehat{\boldsymbol{\Sigma}}_{sp(j),j})^T\{(\mathbf{Z}^j_{sp(j)})^T\mathbf{Z}^j_{sp(j)}/n\}\widehat{\boldsymbol{\Sigma}}_{sp(j),j} \\
&\quad + (\widehat{\boldsymbol{\Sigma}}_{sp(j),j})^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}_{sp(j),sp(j)}\widehat{\boldsymbol{\Sigma}}_{sp(j),j} \\
&\xrightarrow{p} \sigma_{jj} - 2(\boldsymbol{\Sigma}_{sp(j),j})^T(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\boldsymbol{\Sigma}_{sp(j),j} + 2(\boldsymbol{\Sigma}_{sp(j),j})^T(\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\boldsymbol{\Sigma}_{sp(j),j} \\
&= \sigma_{jj},
\end{aligned}$$

where the convergence in probability is from

$$\begin{aligned}
(\mathbf{Z}^j_{sp(j)})^T\boldsymbol{Y}^{(j)}/n &\xrightarrow{p} (\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}\boldsymbol{\Sigma}_{sp(j),j} \\
(\mathbf{Z}^j_{sp(j)})^T\mathbf{Z}^j_{sp(j)}/n &\xrightarrow{p} (\boldsymbol{\Sigma}_{-j,-j})^{-1}_{sp(j),sp(j)}
\end{aligned}$$

and $\widehat{\boldsymbol{\Sigma}}_{sp(j),j} \xrightarrow{p} \boldsymbol{\Sigma}_{sp(j),j}$. □

Next, to establish the asymptotic normality of the iterative conditional fitting estimator, we will use the following Lemma which results from Theorem 2 of Anderson (1973). For this Lemma, we define a matrix $\boldsymbol{Q}$ with entries of 0 or 1 that satisfies $\text{vec}(\boldsymbol{\Sigma}) = \boldsymbol{Q}\boldsymbol{\sigma}$ as defined in Chaudhuri et al. (2007).

Lemma 2. *Let $\widetilde{\boldsymbol{\sigma}}$ be a consistent estimator of $\boldsymbol{\sigma}$ and $\widetilde{\boldsymbol{\sigma}}^*$ be the solution of the linear equation $\boldsymbol{Q}^T\{\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\sigma}})^{-1} \otimes \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\sigma}})^{-1}\}\boldsymbol{Q}\boldsymbol{\sigma} = \boldsymbol{Q}^T\{\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\sigma}})^{-1} \otimes \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\sigma}})^{-1}\}vec(\boldsymbol{S})$ where $\boldsymbol{S}$ is the sample covariance matrix and $\otimes$ is the kronecker product. Then, as $n \to \infty$,*

$$\sqrt{n}(\widetilde{\boldsymbol{\sigma}}^* - \boldsymbol{\sigma}) \to \mathbb{N}(\mathbf{0}, I(\boldsymbol{\sigma})^{-1})$$

where $I(\boldsymbol{\sigma})$ is the Fisher information matrix.

Proof of Theorem 1: By Lemma 1, $\widehat{\boldsymbol{\sigma}}$ will be a consistent estimator of $\boldsymbol{\sigma}$. Consider a solution

$\widehat{\boldsymbol{\sigma}}^*$ for the following linear equation:

$$\boldsymbol{Q}^T\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}})^{-1}\otimes\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}})^{-1}\}\boldsymbol{Q}\boldsymbol{\sigma}=\boldsymbol{Q}^T\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}})^{-1}\otimes\boldsymbol{\Sigma}(\widehat{\boldsymbol{\sigma}})^{-1}\}\mathrm{vec}(\boldsymbol{S}).$$

Then, since $\widehat{\boldsymbol{\sigma}}$ is a consistent estimator of $\boldsymbol{\sigma}$, $\sqrt{n}(\widehat{\boldsymbol{\sigma}}^*-\boldsymbol{\sigma})\to\mathbb{N}(\mathbf{0},I(\boldsymbol{\sigma})^{-1})$ holds by Lemma 2. Note that $\widehat{\boldsymbol{\sigma}}$ is a solution of the above linear equation because $\widehat{\boldsymbol{\sigma}}$ is a solution of the normal likelihood equation $\partial\ell(\boldsymbol{\sigma})/\partial\boldsymbol{\sigma}=0$, which can be written as below:

$$\boldsymbol{Q}^T\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\otimes\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\}\boldsymbol{Q}\boldsymbol{\sigma}=\boldsymbol{Q}^T\{\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\otimes\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}\}\mathrm{vec}(\boldsymbol{S}).$$

Hence, $\sqrt{n}(\widehat{\boldsymbol{\sigma}}-\boldsymbol{\sigma})\to\mathbb{N}(\mathbf{0},I(\boldsymbol{\sigma})^{-1})$ holds. □

B.2 Proof of Proposition 1

Let $(Y_1,\ldots,Y_{j-1},Y_{j+1},\ldots,Y_p)^T\sim\mathbb{N}_{p-1}(\mathbf{0},\widetilde{\boldsymbol{\Sigma}}_{-j,-j})$ with known $\widetilde{\boldsymbol{\Sigma}}_{-j,-j}$. Also, let $\beta=\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$ and $\lambda_j=\sigma_{jj}-\boldsymbol{\Sigma}_{-j,j}^T\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$. From the blockwise matrix inversion,

$$\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}\boldsymbol{A}&\boldsymbol{B}^T\\\boldsymbol{B}&\boldsymbol{C}\end{bmatrix}=\begin{bmatrix}\lambda_j^{-1}&-\lambda_j^{-1}\beta^T\\-\lambda_j^{-1}\beta&\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}+\lambda_j^{-1}\beta\beta^T\end{bmatrix}\tag{B.1}$$

where

$$\begin{aligned}\boldsymbol{A}&=(\sigma_{jj}-\boldsymbol{\Sigma}_{-j,j}^T\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j})^{-1}\\\boldsymbol{B}&=-(\sigma_{jj}-\boldsymbol{\Sigma}_{-j,j}^T\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j})^{-1}\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}\\\boldsymbol{C}&=\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}+\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}(\sigma_{jj}-\boldsymbol{\Sigma}_{-j,j}^T\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j})^{-1}\boldsymbol{\Sigma}_{j,-j}\widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}.\end{aligned}$$

We derive the solution of $\ell^*(\boldsymbol{\Sigma})$ as below.

$$\begin{aligned}\ell^*(\boldsymbol{\Sigma})&=-\frac{np}{2}\log(2\pi)-\frac{n}{2}\log|\boldsymbol{\Sigma}|-\frac{n}{2}\mathrm{tr}\{(\boldsymbol{S}+\epsilon I)\boldsymbol{\Sigma}^{-1}\}\\&=-\frac{np}{2}\log(2\pi)-\frac{n}{2}\log|\boldsymbol{\Sigma}|-\frac{n}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})-\frac{n\epsilon}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}).\end{aligned}$$

(i) Since $|\boldsymbol{\Sigma}| = |\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}|(\sigma_{jj} - \boldsymbol{\Sigma}^T_{-j,j}\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}\boldsymbol{\Sigma}_{-j,j}) = \lambda_j|\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}|$, $\log|\boldsymbol{\Sigma}| = \log\lambda_j + \log|\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}|$. (ii) Using the blockwise matrix inversion (B.1), we can prove that $y^T_{i\bullet}\boldsymbol{\Sigma}^{-1}y_{i\bullet} = \lambda_j^{-1}(y_{ij} - y^T_{i,-j}\beta)^2 + y^T_{i,-j}\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}y_{i,-j}$ where $\bullet$ represents all indices.

Hence,

$$\begin{aligned}\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) &= \frac{1}{n}\text{tr}(\sum_{i=1}^n y_{i\bullet}y^T_{i\bullet}\boldsymbol{\Sigma}^{-1}) = \frac{1}{n}\sum_{i=1}^n \text{tr}(y_{i\bullet}y^T_{i\bullet}\boldsymbol{\Sigma}^{-1}) = \frac{1}{n}\sum_{i=1}^n \text{tr}(y^T_{i\bullet}\boldsymbol{\Sigma}^{-1}y_{i\bullet}) = \frac{1}{n}\sum_{i=1}^n y^T_{i\bullet}\boldsymbol{\Sigma}^{-1}y_{i\bullet} \\ &= \frac{1}{n}\sum_{i=1}^n \lambda_j^{-1}(y_{ij} - y^T_{i,-j}\beta)^2 + \frac{1}{n}\sum_{i=1}^n y^T_{i,-j}\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}y_{i,-j}\end{aligned}$$

(iii) Using the blockwise matrix inversion (B.1), we can prove that

$$\begin{aligned}\text{tr}(\boldsymbol{\Sigma}^{-1}) &= \lambda_j^{-1} + \text{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}) + \text{tr}(\lambda_j^{-1}\beta\beta^T) \\ &= \lambda_j^{-1} + \text{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}) + \text{tr}(\lambda_j^{-1}\beta^T\beta) \\ &= \lambda_j^{-1} + \lambda_j^{-1}\beta^T\beta + \text{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}).\end{aligned}$$

From (i),(ii) and (iii), considering only the unknown terms β and λ_j,

$$\begin{aligned}\ell^*(\boldsymbol{\Sigma}) &= \ell^*(\beta, \lambda_j) \\ &= -\frac{n}{2}\log\lambda_j - \frac{1}{2}\sum_{i=1}^n \lambda_j^{-1}(y_{ij} - y^T_{i,-j}\beta)^2 - \frac{n\epsilon}{2}\lambda_j^{-1}(1 + ||\beta||^2) + (...).\end{aligned}$$

Note that the above function is the ridge regression in term of β. Using the chain rule, the solution for $\boldsymbol{\Sigma}_{-j,j}$ can be obtained by

$$\widehat{\boldsymbol{\Sigma}}_{-j,j} = \big\{\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}\big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_{i,-j}\boldsymbol{y}^T_{i,-j} + \epsilon I\big)\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}\big\}^{-1}\big\{\frac{1}{n}\sum_{i=1}^n (\widetilde{\boldsymbol{\Sigma}}^{-1}_{-j,-j}\boldsymbol{y}^T_{i,-j})y_{ij}\big\}.$$

The solution for λ_j can be obtained by

$$\hat{\lambda}_j = \frac{1}{n}\sum_{i=1}^n (y_{ij} - y^T_{i,-j}\hat{\beta})^2 + \epsilon||\hat{\beta}||^2 + \epsilon$$

where $\hat{\beta} = \widetilde{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\widehat{\boldsymbol{\Sigma}}_{-j,j}$. Hence the solution for $\hat{\sigma}_{jj}$ is

$$\begin{aligned}\hat{\sigma}_{jj} &= \hat{\lambda}_j + \widehat{\boldsymbol{\Sigma}}_{-j,j}^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\widehat{\boldsymbol{\Sigma}}_{-j,j} \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - y_{i,-j}^T\hat{\beta})^2 + \epsilon||\hat{\beta}||^2 + \epsilon + \widehat{\boldsymbol{\Sigma}}_{-j,j}^T(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\widehat{\boldsymbol{\Sigma}}_{-j,j}.\end{aligned}$$

□

B.3 Proof of Proposition 2

For simplicity, we will prove Proposition 2 for the case when $\boldsymbol{\sigma}_O$ contains one more parameter than $\boldsymbol{\sigma}_C$. For larger $\boldsymbol{\sigma}_O$, Proposition 2 can be shown similarly.

Define a matrix $\boldsymbol{Q}$ with entries of 0 or 1 that satisfies $\text{vec}(\boldsymbol{\Sigma}) = \boldsymbol{Q}\boldsymbol{\sigma}$ as defined in Chaudhuri et al. (2007). Let K be the number of elements in $\boldsymbol{\sigma}_C$. Define a $p^2 \times K$ matrix $\boldsymbol{Q}_C = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K]$ with entries of 0 or 1 that satisfies $\text{vec}(\boldsymbol{\Sigma}) = \boldsymbol{Q}_C\boldsymbol{\sigma}_C$ as defined in Chaudhuri et al. (2007). Also, define a $p^2 \times (K+1)$ matrix $\boldsymbol{Q}_O = [\boldsymbol{Q}_C, \boldsymbol{q}_{K+1}] = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K, \boldsymbol{q}_{K+1}]$. By Theorem 1, $\sqrt{n}(\widehat{\boldsymbol{\sigma}}_C - \boldsymbol{\sigma}_C) \to \mathbb{N}_K(\boldsymbol{0}, I(\boldsymbol{\sigma}_C)^{-1})$ and $\sqrt{n}(\widetilde{\boldsymbol{\sigma}}_O - \boldsymbol{\sigma}_O) \to \mathbb{N}_{K+1}(\boldsymbol{0}, I(\boldsymbol{\sigma}_O)^{-1})$. By denoting $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{A}$ is symmetric and positive-definite. Denote $\boldsymbol{R}_1 = \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{Q}_C$ and $\boldsymbol{r}_2 = \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{q}_{K+1}$. Then,

$$I(\boldsymbol{\sigma}_C) = \frac{1}{2}\boldsymbol{Q}_C^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{Q}_C = \frac{1}{2}\boldsymbol{Q}_C^T\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{Q}_C = \frac{1}{2}\boldsymbol{R}_1^T\boldsymbol{R}_1$$

and

$$\begin{aligned}I(\boldsymbol{\sigma}_O) &= \frac{1}{2}\boldsymbol{Q}_O^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{Q}_O = \frac{1}{2}[\boldsymbol{Q}_C, \boldsymbol{q}_{K+1}]^T\boldsymbol{A}[\boldsymbol{Q}_C, \boldsymbol{q}_{K+1}] \\ &= \frac{1}{2}\begin{bmatrix}\boldsymbol{Q}_C^T\boldsymbol{A}\boldsymbol{Q}_C & \boldsymbol{Q}_C^T\boldsymbol{A}\boldsymbol{q}_{K+1} \\ \boldsymbol{q}_{K+1}^T\boldsymbol{A}\boldsymbol{Q}_C & \boldsymbol{q}_{K+1}^T\boldsymbol{A}\boldsymbol{q}_{K+1}\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\boldsymbol{R}_1^T\boldsymbol{R}_1 & \boldsymbol{R}_1^T\boldsymbol{r}_2 \\ \boldsymbol{r}_2^T\boldsymbol{R}_1 & \boldsymbol{r}_2^T\boldsymbol{r}_2\end{bmatrix}.\end{aligned}$$

Note that the square roots of diagonal elements of $I(\boldsymbol{\sigma}_C)^{-1}/n$ are the standard error of parameters in $\widehat{\boldsymbol{\sigma}}_C$. Also, the square roots of the first K diagonal elements of $I(\boldsymbol{\sigma}_O)^{-1}/n$ are the standard error of parameters in $\widetilde{\boldsymbol{\sigma}}_C$. We claim that the diagonal elements of $I(\boldsymbol{\sigma}_C)^{-1}$ are less than the first K

diagonal elements of $I(\boldsymbol{\sigma}_O)^{-1}$. By Schur complement,

$$I(\boldsymbol{\sigma}_O)^{-1} = 2\begin{bmatrix} \boldsymbol{R}_1^T\boldsymbol{R}_1 & \boldsymbol{R}_1^T\boldsymbol{r}_2 \\ \boldsymbol{r}_2^T\boldsymbol{R}_1 & \boldsymbol{r}_2^T\boldsymbol{r}_2 \end{bmatrix}^{-1} = 2\begin{bmatrix} \{\boldsymbol{R}_1^T\boldsymbol{R}_1 - (\boldsymbol{R}_1^T\boldsymbol{r}_2)(\boldsymbol{r}_2^T\boldsymbol{r}_2)^{-1}(\boldsymbol{r}_2^T\boldsymbol{R}_1)\}^{-1} & (\dots) \\ (\dots) & (\dots) \end{bmatrix}.$$

Hence, our claim is:

$$\text{diag}[(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}] \leq \text{diag}[\{\boldsymbol{R}_1^T\boldsymbol{R}_1 - (\boldsymbol{R}_1^T\boldsymbol{r}_2)(\boldsymbol{r}_2^T\boldsymbol{r}_2)^{-1}(\boldsymbol{r}_2^T\boldsymbol{R}_1)\}^{-1}].$$

Since $\{\boldsymbol{R}_1^T\boldsymbol{R}_1 - (\boldsymbol{R}_1^T\boldsymbol{r}_2)(\boldsymbol{r}_2^T\boldsymbol{r}_2)^{-1}(\boldsymbol{r}_2^T\boldsymbol{R}_1)\}^{-1} =$
$(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1} + (\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)\{(\boldsymbol{r}_2^T\boldsymbol{r}_2) - (\boldsymbol{r}_2^T\boldsymbol{R}_1)(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)\}^{-1}(\boldsymbol{r}_2^T\boldsymbol{R}_1)(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}$,
defining a scalar value $c = (\boldsymbol{r}_2^T\boldsymbol{r}_2) - (\boldsymbol{r}_2^T\boldsymbol{R}_1)(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)$, our claim is equivalent to

$$c^{-1}\text{diag}[(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)(\boldsymbol{r}_2^T\boldsymbol{R}_1)(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}] \geq \mathbf{0}$$

where $\mathbf{0}$ is a K-dimensional zero vector. Because $I(\boldsymbol{\sigma}_O)$ is positive-definite, both $(\boldsymbol{R}_1^T\boldsymbol{R}_1)$ and $c = (\boldsymbol{r}_2^T\boldsymbol{r}_2) - (\boldsymbol{r}_2^T\boldsymbol{R}_1)(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)$ are positive-definite, hence $c > 0$.
Define a K-dimensional column vector $\mathbf{v} = (v_1, \dots, v_K)^T$ to be $\mathbf{v} = (\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)$. Then, $\text{diag}[(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1^T\boldsymbol{r}_2)(\boldsymbol{r}_2^T\boldsymbol{R}_1)(\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}] = \text{diag}(\mathbf{v}\mathbf{v}^T) = (v_1^2, \dots, v_K^2) \geq \mathbf{0}$, hence the claim holds. □

B.4 Proof of Proposition 3

By the definition of $\boldsymbol{Q}_C$ and $\boldsymbol{Q}_U$, $\boldsymbol{Q}_C\boldsymbol{\sigma}_C = \boldsymbol{Q}_U\boldsymbol{\sigma}_U + \boldsymbol{Q}_{C\setminus U}\boldsymbol{\sigma}_{C\setminus U} = \text{vec}(\boldsymbol{\Sigma})$. Hence, if we multiply $\boldsymbol{Q}_U^T\mathbf{W}$ on both sides of this equation,

$$\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U\boldsymbol{\sigma}_U + \boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_{C\setminus U}\boldsymbol{\sigma}_{C\setminus U} = \boldsymbol{Q}_U^T\mathbf{W}\text{vec}(\boldsymbol{\Sigma}). \tag{B.2}$$

Note that $\widetilde{\boldsymbol{\sigma}}_U$ is the solution to the equation $\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U\boldsymbol{\sigma}_U = \boldsymbol{Q}_U^T\mathbf{W}\text{vec}(\boldsymbol{S})$, hence

$$\boldsymbol{Q}_U^T\mathbf{W}\boldsymbol{Q}_U\widetilde{\boldsymbol{\sigma}}_U = \boldsymbol{Q}_U^T\mathbf{W}\text{vec}(\boldsymbol{S}). \tag{B.3}$$

By multiplying $(\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1}$ to equation (B.2) and (B.3),

$$\boldsymbol{\sigma}_U = (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \text{vec}(\boldsymbol{\Sigma}) - (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C \setminus U} \boldsymbol{\sigma}_{C \setminus U}; \text{and}$$
$$\widetilde{\boldsymbol{\sigma}}_U = (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \text{vec}(\boldsymbol{S})$$

By combining these equations,

$$\widetilde{\boldsymbol{\sigma}}_U - \boldsymbol{\sigma}_U = (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \text{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}) + (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C \setminus U} \boldsymbol{\sigma}_{C \setminus U}.$$

and by taking expectation,

$$E(\widetilde{\boldsymbol{\sigma}}_U) - \boldsymbol{\sigma}_U = (\boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_U)^{-1} \boldsymbol{Q}_U^T \mathbf{W} \boldsymbol{Q}_{C \setminus U} \boldsymbol{\sigma}_{C \setminus U}.$$

the bias of $\widetilde{\boldsymbol{\sigma}}_U$ can be quantified as above. □

B.5 Proof of Theorem 2

In the following lemma, we prove that thresholding the (i, j)-th entry of the sample covariance matrix with $\lambda_{ij} = C_{ij} n^{-\alpha}$ will identify both the non-zero entries and zero entries of the true covariance matrix (that is, "recover the support") with probability tending to 1. The proof of this lemma follows the path for the Theorem 2 in Rothman et al. (2009).

Lemma 3. *Let σ_{ij} and s_{ij} be the (i, j)-th entry of $\boldsymbol{\Sigma}$ and the sample covariance matrix $\boldsymbol{S}$, respectively. If $\lambda_{ij} = C_{ij} n^{-\alpha}$ for a positive constant C_{ij}, $\alpha = 0.5 - \gamma > 0$ and $\gamma > 0$, then*

$$|s_{ij}| \leq \lambda_{ij} \text{ for all } (i, j) \text{ such that } \sigma_{ij} = 0; \text{ and}$$
$$|s_{ij}| > \lambda_{ij} \text{ for all } (i, j) \text{ such that } \sigma_{ij} \neq 0$$

with probability tending to 1 as $n \to \infty$.

Proof of Lemma 3: First, we show $|s_{ij}| \leq \lambda_{ij}$ for all (i,j) such that $\sigma_{ij} = 0$ with probability tending to 1. Let s_{ij} be the (i,j)-th element of the sample covariance matrix and let $\lambda^* = \min[\{C_{ij}\}_{i,j=1}^p] \cdot n^{-\alpha}$. Then, $\{(i,j) : |s_{ij}| > \lambda_{ij}, \sigma_{ij} = 0\} \subseteq \{(i,j) : |s_{ij}| > \lambda^*, \sigma_{ij} = 0\}$. Also, since $\{(i,j) : |s_{ij}| > \lambda^*, \sigma_{ij} = 0\} \subseteq \{(i,j) : |s_{ij} - \sigma_{ij}| > \lambda^*\}$, as $n \to \infty$,

$$
\begin{aligned}
P\big(\sum_{i,j} \mathbb{1}_{\{|s_{ij}|>\lambda_{ij}, \sigma_{ij}=0\}} > 0\big) &\leq P\big(\sum_{i,j} \mathbb{1}_{\{|s_{ij}|>\lambda^*, \sigma_{ij}=0\}} > 0\big) \\
&\leq P(\max_{i,j} |s_{ij} - \sigma_{ij}| > \lambda^*) \\
&\leq \sum_{i,j} P(|s_{ij} - \sigma_{ij}| > \lambda^*) \\
&\leq \sum_{i,j} \frac{\text{var}(s_{ij})}{(\lambda^*)^2} \quad \text{(by Chebyshev inequality)} \\
&= \sum_{i,j} \frac{(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})/n}{(\lambda^*)^2} \quad \text{(property of Wishart distribution)} \\
&\leq \frac{C^*}{n^{1-2\alpha}} = C^* n^{-2\delta}.
\end{aligned}
$$

Since the righthand of the above inequality converges to zero as $n \to \infty$, $|s_{ij}| \leq \lambda_{ij}$ with probability tending to 1 if $\sigma_{ij} = 0$.

Next, we show $|s_{ij}| > \lambda_{ij}$ for all (i,j) such that $\sigma_{ij} \neq 0$ with probability tending to 1. Let h be the lower bound for $|\sigma_{ij}|$ for all (i,j) and let $\lambda^{**} = \max[\{C_{ij}\}_{i,j=1}^p] \cdot n^{-\alpha}$. Then, $\{(i,j) : |s_{ij}| \leq \lambda_{ij}, |\sigma_{ij}| > h\} \subseteq \{(i,j) : |s_{ij}| \leq \lambda^{**}, |\sigma_{ij}| > h\}$. Also, by triangle inequality, $\{(i,j) : |s_{ij}| \leq \lambda^{**}, |\sigma_{ij}| > h\} \subseteq \{(i,j) : |s_{ij} - \sigma_{ij}| > |h - \lambda^{**}|\}$. Hence, as $n \to \infty$,

$$
\begin{aligned}
P\big(\sum_{i,j} \mathbb{1}_{\{|s_{ij}|\leq\lambda_{ij}, \sigma_{ij}\neq 0\}} > 0\big) &\leq \sum_{i,j} \frac{(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})/n}{(h - \lambda^{**})^2} \quad \text{(by Chebyshev inequality)} \\
&\leq \frac{C^*}{nh^2 - 2h \cdot \max[\{C_{ij}\}_{i,j=1}^p] \cdot n^{1-\alpha} + \max[\{C_{ij}\}_{i,j=1}^p]^2 \cdot n^{1-2\alpha}}.
\end{aligned}
$$

Since the righthand of the above inequality converges to zero as $n \to \infty$, $|s_{ij}| > \lambda_{ij}$ with probability tending to 1 if $\sigma_{ij} \neq 0$. □

Proof of Theorem 2: By Lemma 3, thresholding the sample covariance matrix with the threshold $\lambda_{ij} = C_{ij} n^{-\alpha}$ will correctly identify the location of the zero entries in $\mathbf{\Sigma}$ with probability tending to 1. Given the correct location of the zero entries, we have shown the asymptotic efficiency of iterative conditional fitting in Theorem 1. □

APPENDIX C

ADDITIONAL NUMERICAL RESULTS

C.1 Additional Simulation Results

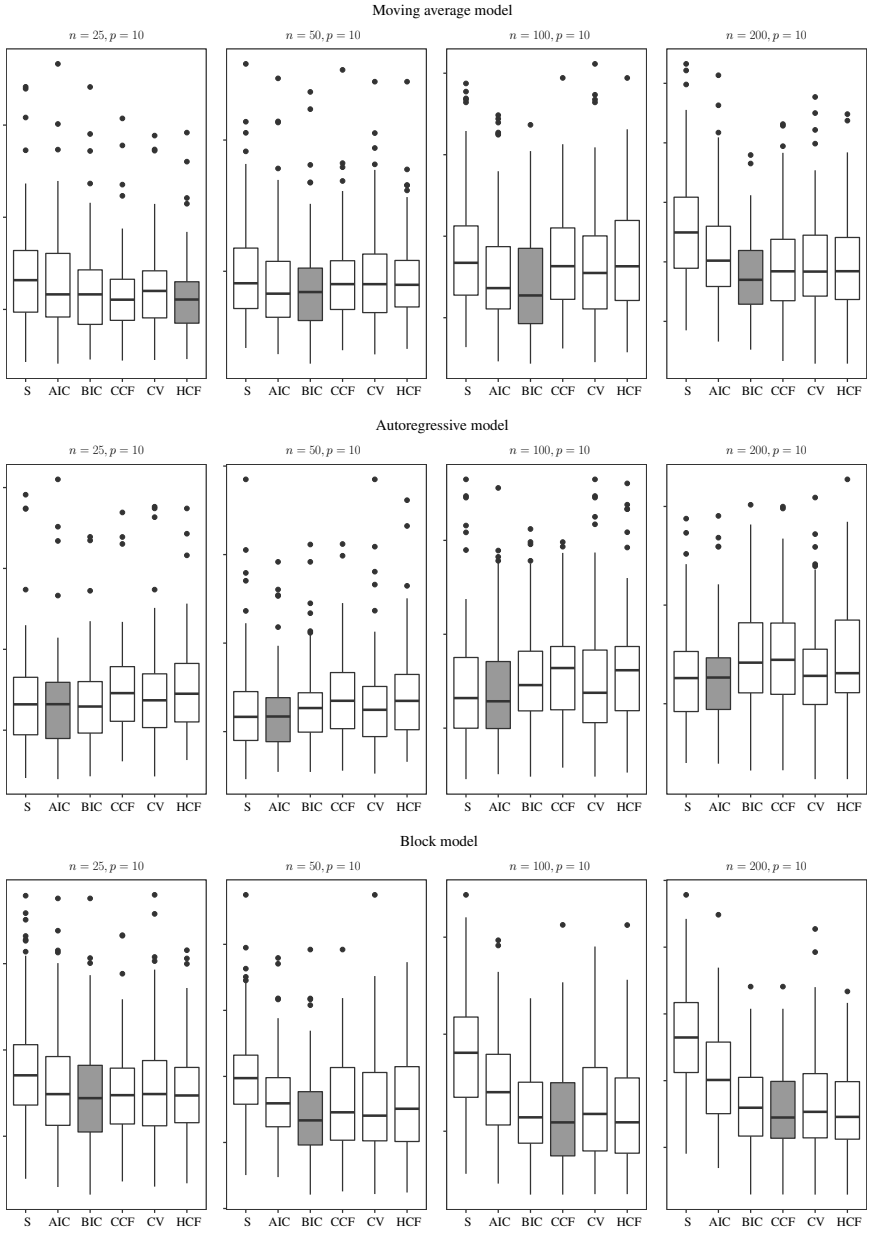

Figure C.1: Frobenius loss for the sample covariance matrix (“S"), the COMET by AIC (“AIC"), BIC (“BIC") and Qiu and Liyanage (2019) (“CCF"), the hard thresholding by cross-validation (“CV") and Qiu and Liyanage (2019) (“HCF") when $p = 10$. The estimator with grey box shows the lowest Frobenius loss on average.

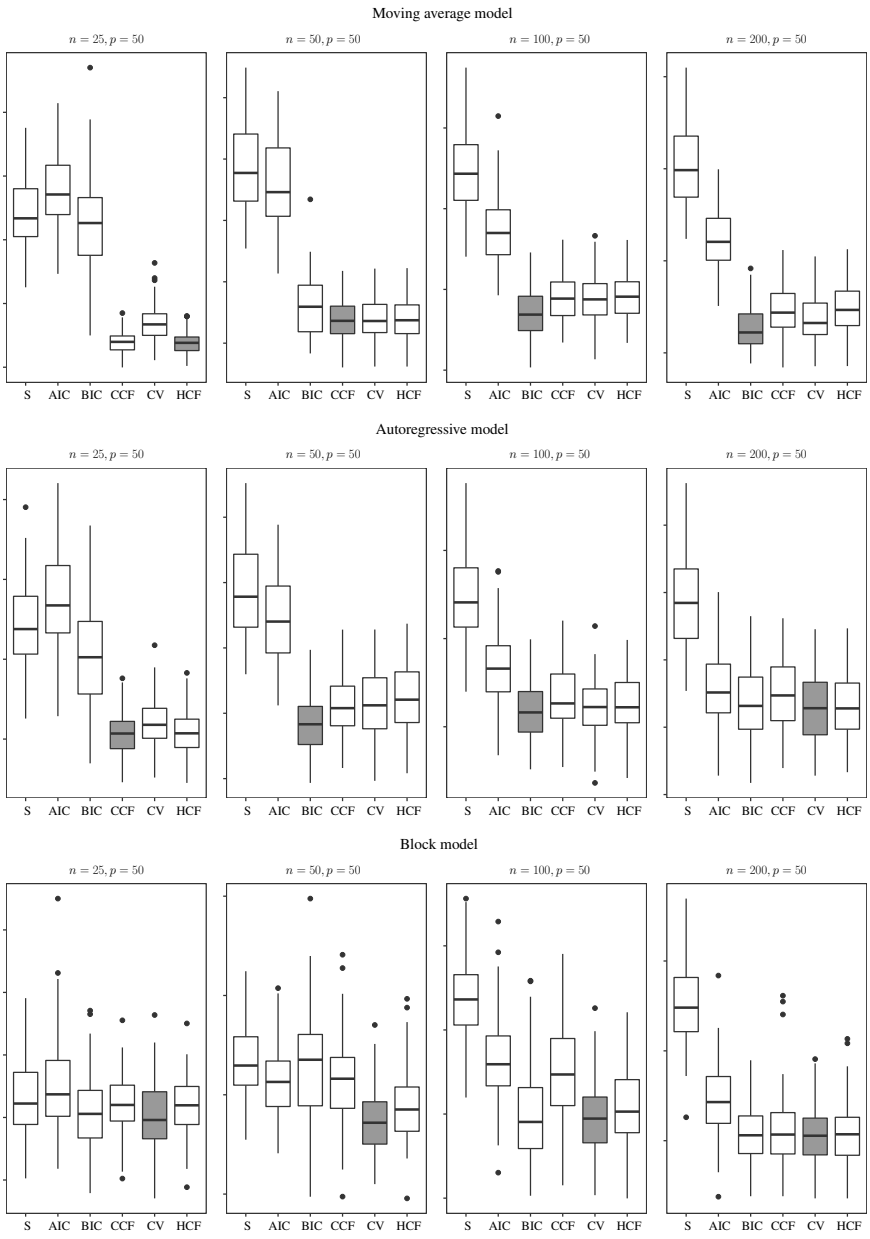

Figure C.2: Frobenius loss for the sample covariance matrix (“S"), the COMET by AIC (“AIC"), BIC (“BIC") and Qiu and Liyanage (2019) (“CCF"), the hard thresholding by cross-validation (“CV") and Qiu and Liyanage (2019) (“HCF") when $p = 50$. The estimator with grey box shows the lowest Frobenius loss on average.

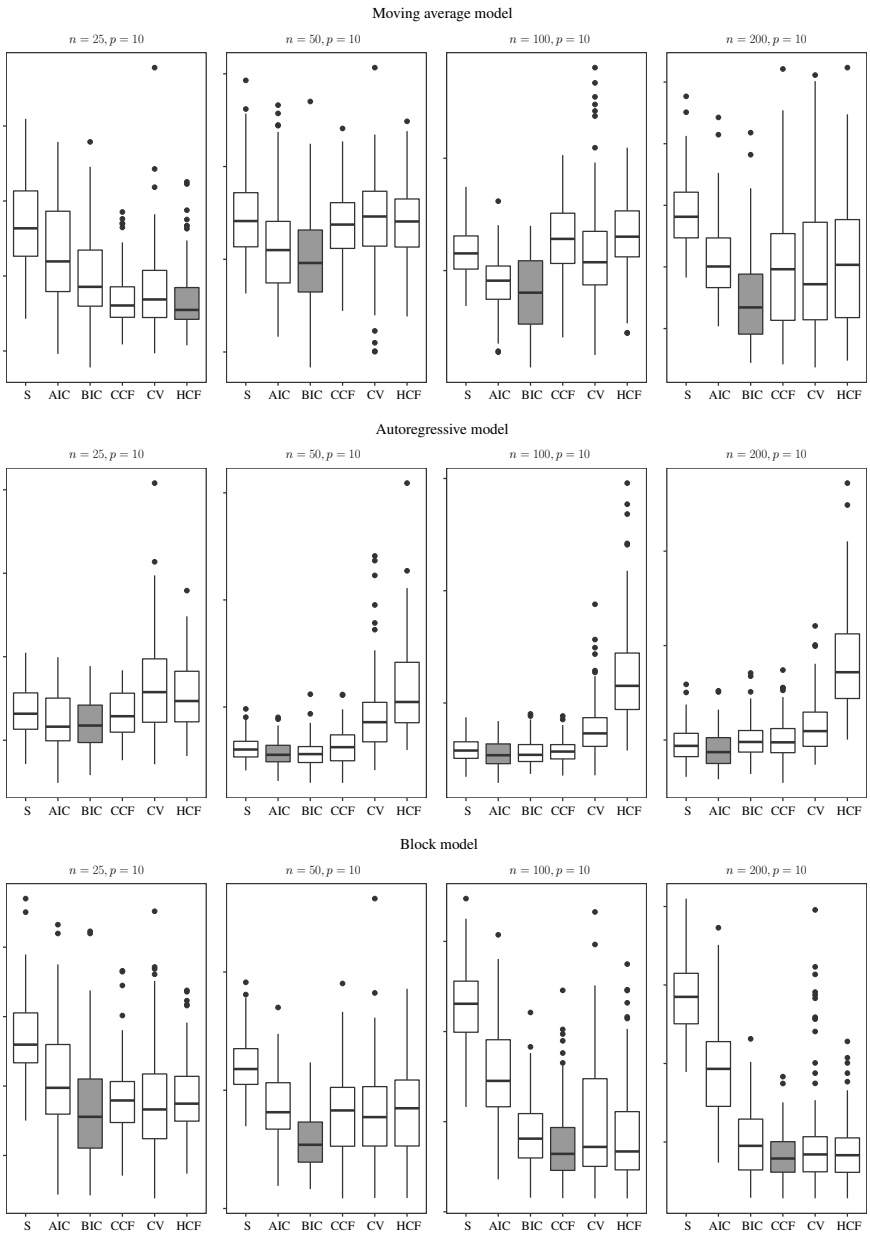

Figure C.3: Entropy loss for the sample covariance matrix (“S"), the COMET by AIC (“AIC"), BIC (“BIC") and Qiu and Liyanage (2019) (“CCF"), the hard thresholding by cross-validation (“CV") and Qiu and Liyanage (2019) (“HCF") when $p = 10$. The estimator with grey box shows the lowest Frobenius loss on average.

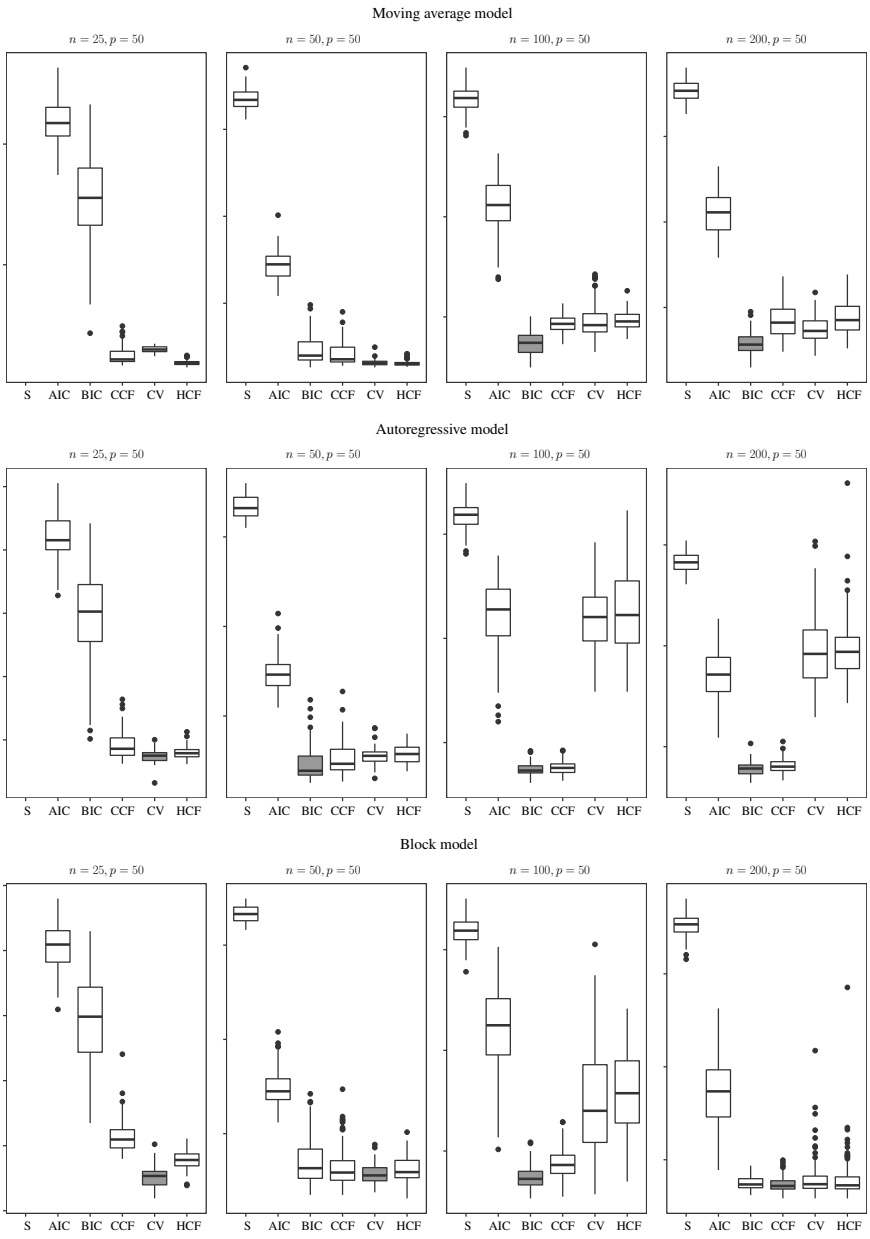

Figure C.4: Entropy loss for the sample covariance matrix ("S"), the COMET by AIC ("AIC"), BIC ("BIC") and Qiu and Liyanage (2019) ("CCF"), the hard thresholding by cross-validation ("CV") and Qiu and Liyanage (2019) ("HCF") when $p = 50$. The estimator with grey box shows the lowest Frobenius loss on average.

n	p	moving average model				block model			
		AIC	BIC	CV	CF	AIC	BIC	CV	CF
25	10	**0.61**/0.25	0.43/0.11	0.19/0.05	0.08/**0.01**	**0.91**/0.25	0.81/0.11	0.59/0.08	0.31/**0.01**
	50	**0.91**/0.75	0.73/0.41	0.07/0.01	0.02/**0.00**	**0.99**/0.74	0.91/0.31	0.53/0.04	0.31/**0.01**
50	10	**0.78**/0.21	0.59/0.07	0.34/0.07	0.24/**0.01**	**0.99**/0.20	0.95/0.05	0.83/0.05	0.70/**0.01**
	50	**0.92**/0.46	0.49/0.04	0.12/0.00	0.10/**0.00**	**1.00**/0.47	0.73/0.01	0.94/0.05	0.81/**0.01**
100	10	**0.95**/0.19	0.85/0.04	0.78/0.15	0.58/**0.01**	**1.00**/0.17	1.00/0.02	0.99/0.03	0.98/**0.01**
	50	**0.97**/0.27	0.79/0.03	0.47/0.00	0.40/**0.00**	**1.00**/0.28	0.99/0.01	1.00/0.02	0.97/**0.00**
200	10	**1.00**/0.17	0.98/0.02	0.97/0.10	0.90/**0.00**	**1.00**/0.17	1.00/0.01	1.00/0.03	1.00/**0.00**
	50	**1.00**/0.21	0.97/0.01	0.93/0.01	0.80/**0.00**	**1.00**/0.15	1.00/0.00	1.00/0.00	1.00/**0.00**

Table C.1: True positive rate (left) / false positive rate (right) under the moving average model and the block model. The autoregressive model was not compared since there is no zero entry in the covariance matrix. AIC and BIC were used for the COMET. Cross-validation (CV) and the closed-form threshold (CF) were used for the hard thresholding. The estimator with the highest true positive rate or the lowest false positive rate is shown in bold.

n	p	moving average		autoregressive		block	
		CV	CF	CV	CF	CV	CF
25	10	4	0	32	33	11	2
	50	76	2	98	64	100	100
50	10	0	0	10	27	0	0
	50	0	0	98	95	100	100
100	10	0	0	1	2	0	0
	50	0	0	87	83	61	69
200	10	0	0	0	0	0	0
	50	0	0	1	1	0	1

Table C.2: Percentage of non-positive definite hard thresholding estimators. Cross-validation (CV) and the closed-form threshold (CF) were used for selecting the threshold parameter.

C.2 Additional Analysis Results for PREDICT-HD

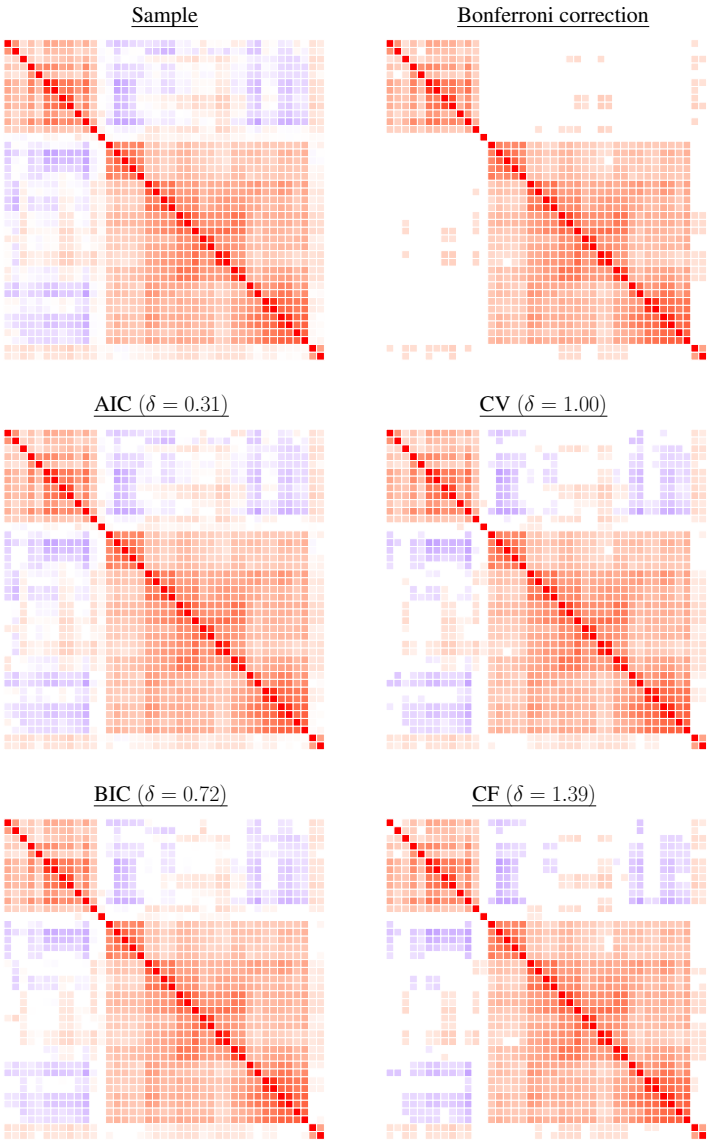

Figure C.5: Heatmaps of the correlations for the sample covariance matrix, the covariance matrix with Bonferroni correction and thresholding estimators; AIC, COMET with AIC-threshold; BIC, COMET with BIC-threshold; CV, hard thresholding with cross-validation; CF, hard thresholding with closed-from threshold. The covariance matrix with Bonferroni correction and both hard thresholding estimators were not positive definite. Positive correlations are shown in red and negative correlations are shown in blue. Zero correlations are shown in white. δ represents the adaptive threshold selected.