MULTIVARIATE FUNCTIONAL PREDICTOR SELECTION

by

Ali Mahzarnia

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2021

Approved by:

Dr. Jun Song

Dr. Michael Grabchak

Dr. Jiancheng Jiang

Dr. Sara M. Levens

ABSTRACT

ALI MAHZARNIA. Multivariate functional predictor selection. (Under the direction of DR. JUN SONG)

We propose methods for functional predictor selection and the estimation of smooth functional coefficients simultaneously in a scalar-on-function regression problem under a high-dimensional multivariate functional data setting. In particular, we develop two methods for functional group-sparse regression under a generic Hilbert space of infinite dimension. We show the convergence of algorithms and the consistency of the estimation and the selection (oracle property) under infinite-dimensional Hilbert spaces. Simulation studies show the effectiveness of the methods in both the selection and the estimation of functional coefficients. The applications to functional magnetic resonance imaging (fMRI) reveal the human brain regions related to ADHD and IQ. In addition, we apply the proposed methods to an econometric data set to find the related functional covariates to GDP of a country. To extend the results, we propose numerical algorithms for more complex models, such as nonlinear (via RKHS), logistic, sparse function–on–function, and standardization the results of the sparse scalar–on–function models before we list the applications of these extensions to the brain image data analysis.

DEDICATION

I must admit that two persons I have met at the UNC Charlotte Graduate School are the most brilliant people I have had a chance to visit personally: Dr. Julie Goodliffe and Dean Tom Reynolds. I strongly hope and will do all I can to see that UNC Charlotte becomes a full research university one day as they are working very hard every day to see the same thing happens. I would like to thank my adviser very much for his guidance through my study, and show my gratitude to the committee members.

I believe in the freedom and liberty of a human being above any ideology and religion that I know. Hence, I appreciate that I live in a period in human history when the United States of America exists as a country. A free country that offers equal opportunity and dignity, disregarding individuals' nationalities, religions, backgrounds, family wealth, social connections, etc. I am thankful for this country because I know that I would have never been able to do what I love the most—doing research in science, mathematics, and statistics—had I not immigrated to this country for my Ph.D. degree. I do not necessarily believe that God does not exist, nor do I believe that it does, but if it does, I hope that God bless the United States of America forever.

Most of all, I would like to dedicate this thesis to my parents, who have been more than a father and a mother to me, Nemat and Nahid. No matter where they are and what they do, they are always with me in my heart.

ACKNOWLEDGEMENTS

I want to acknowledge the Graduate Assistant Support Program (GASP) grants by UNC Charlotte and the Graduate School funding throughout my doctoral studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAL Automated anatomical labeling.

ADHD Attention-Deficit/Hyperactivity Disorder.

ADMM Alternating direction method of multipliers.

AIC Akaike information criterion.

BIC Bayesian information criterion.

BOLD Blood-oxygen-level-dependent.

CPU Central processing Unit.

CS Cauchy Schwartz.

CT Computed tomography.

DTI Diffusion Tensor Imaging.

DWI Diffusion–weighted imaging.

EEG Electroencephalogram.

ERP Event-related potential.

FDA Functional data analysis.

FDG Fludeoxyglucose.

fMRI Functional magnetic resonance imaging.

fNIRS Functional Near–infrared spectroscopy.

GASP Graduate Assistant Support Program.

GB Gigabyte.

GDP Gross domestic product.

GMD Groupwise-majorization-descent.

IQ Intelligence quotient.

KKT Karush–Kuhn–Tucker.

LAD Least absolute deviation.

LASSO Least absolute shrinkage and selection operator.

MEG Magnetoencephalography.

MFG Multivariate functional group.

MFG-EN Multivariate functional group Elastic Net.

MFG-LASSO Multivariate functional group LASSO.

MRI Magnetic resonance imaging.

MRS Magnetic resonance spectroscopy.

MSE Mean square error.

NIR Near–infrared.

NIRS Near–infrared spectroscopy.

NMR Nuclear magnetic resonance.

OLS Ordinary least squares.

PET Positron emission tomography.

QM Quadratic majorization.

RAM Random access memory.

RBF Radial basis function.

RKHS Reproducing kernel Hilbert space.

RMSE Root-mean-square error.

ROI Regions of interest.

SCAD Smoothly clipped absolute deviation.

SO Proton density.

SPECT Single–photon emission computed tomography.

SPM Statistical Parametric Mapping.

UNCC University of North Carolina at Charlotte.

PREFACE

In the past decades, functional data analysis (FDA) has received significant attention in which an entire function is an observation. (1) introduced a general framework of FDA and many other researchers investigated the estimation and inference methods of functional data. See (2), (3), (4), and (5). More recently, FDA has been extended to multivariate functional data that can deal with multiple functions as a single observation. See (6; 7). However, the sparseness of functional predictors in the multivariate model has not been studied well compared to the univariate case. Hence, we aim to develop theories and algorithms for the sparse functional regression methods with functional predictor selection when we have scalar data as response values and high-dimensional multivariate functional data as predictors.

Under the multivariate setting, numerous sparse models have been studied with the introduction of ℓ_1-penalty. Least absolute shrinkage and selection operator (LASSO) introduces a penalty term to the least square cost function, which performs both variable selection and shrinkage (8). The LASSO-type penalty, such as the Elastic Net (9), the smoothly clipped absolute deviation (SCAD) (10), their modifications (the adaptive LASSO (11) and the adaptive Elastic Net (12)) are developed to overcome the lack of theoretical support and the practical limitations of the LASSO, such as the saturation. These methods were developed to overcome the challenges and enjoy asymptotic properties when the sample size increases, such as estimation and selection consistency, also known as the oracle property.

Recently, the sparse models have been extended to the functional data. Initially, a majority of the literature seeks the sparseness of the time domain. Examples include (13) and related articles on univariate functional data and (14) multivariate functional data. On the other hand, (15) proposed a model considering the sparseness in the functional predictors under the multivariate functional data setting. In particular, they introduced a model based on the least absolute deviation (LAD) and the group

LASSO in the presence of outliers in functional predictors and responses. Its numerical examples and data application show the effectiveness in practice, but theoretical properties and detailed algorithms have not been explored. To this end, we develop methods for the scalar-on-function regression model, which allows sparseness of the functional predictors and the simultaneous estimation of the smooth functional coefficients. To implement it with the actual data, we derive two algorithms for each of the optimization problems. Finally, we show both the functional predictor selection consistency and the estimation consistency.

One motivating example for the proposed methods is the application to functional magnetic resonance imaging (fMRI). The dataset consists of the functional signals of the brain activities measured by blood-oxygen-level-dependent (BOLD), which detects hemodynamic changes based on the metabolic demands followed by neural activities. There are pre-specified regions of the brain, and the BOLD signals associated with multiple voxels in each region are integrated into one signal for that region. Thus, the fMRI data are considered to be multivariate functional data in which each functional predictor represents the signals from a region of the brain. In section 8.1, we regress the ADHD index to the regional BOLD activities of the fMRI of the human subjects. There are 116 regions of the brain in the data, and the proposed methods reduce the regions to 41 regions with significantly lower errors than the linear functional regression. Figure 1displays the regions of the brain's atlas that are identified by the proposed method. It shows that the methods simplify the data analysis and provide clear representation while keeping the crucial information. The study shows an urgent need for new approaches in the fields of medical and life sciences and other related areas. The following quote from (16) further motivates to study the applications of the sparse multivariate functional regression in fMRI.

Think of the challenge of the fMRI with the analogous situation one would have if, when

flying over a city at night, an attempt is made to determine the city activities in detail by simply observing where the lights are on. The information is extremely sparse, but with time, specific inferences can be drawn. – Peter A. Bandettini, *fMRI*, 2020

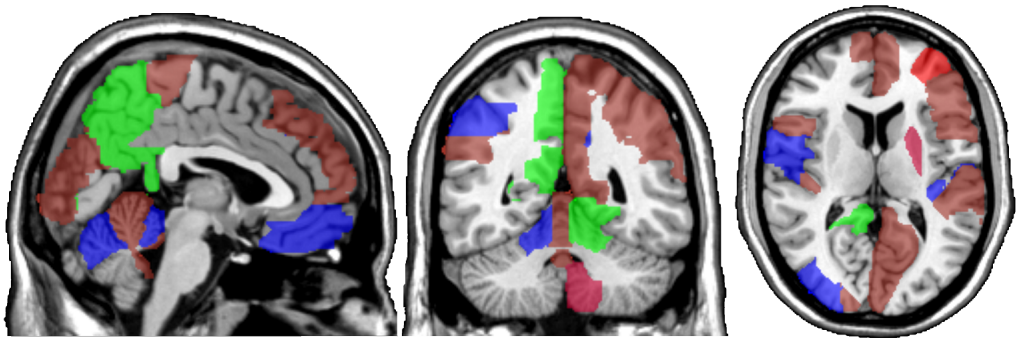


Figure 1: The regions of interest, the BOLD activities of which correlate the most with the ADHD score variability in a sample of subjects and achieve the lowest prediction error. The regions associated with ADHD are colored red, those associated with ADHD Hyper/Impulsive are blue, and those associated with ADHD Inattentive are colored green.

The rest of the thesis is organized as follows. In chapter 1, we introduce a brief history of brain imaging and a brief history of statistical methods. In chapter 2, we illustrate the proposed methods' general framework and the notations used in this thesis. In chapter 3, we describe the model and the optimization problem that we consider. Then, we develop an explicit solution to the optimization problem and illustrate a detailed procedure using alternating direction method of multipliers (ADMM) in chapter4. We also derive another algorithm, called groupwise-majorization-descent (GMD), along with the strong rule for faster computation in chapter 5. In chapter 6, we develop asymptotic results, including the consistency of the proposed methods and the oracle property. In chapter 7, we show the effectiveness of the methods by conducting simulation studies. In chapter 8, we apply the methods to a resting–state fMRI dataset and an econometric data set. In chapter 9, we explore extensions in different directions; we propose numerical algorithms for more complex models, such

as nonlinear (via RKHS), logistic, function–on–function, and standardization of the final results. The standardization of the sparse scalar–on–function models can be done by estimation of the standard deviations of the norms of the functional coefficients. We list the applications of these extensions to the brain image data analysis in this chapter as well. Concluding discussions are made in chapter 10. Finally, the appendix includes all of the proofs, a 3–D displays of ROIs, the list of regions of the brain associated with ADHD and the IQ scores, and the list of variables and countries of the econometric data. We created an `R` package `MFSGrp` for the computation, and it is available at https://github.com/Ali-Mahzarnia/MFSGrp.

CHAPTER 1: A BRIEF HISTORY OF BRAIN IMAGING AND STATISTICAL METHODS

In the first section of this chapter, we first briefly explain common knowledge of fMRI with a brief history. In the second section, we will discuss the classical statistical methods.

1.1 Functional magnetic resonance imaging (fMRI)

Most of the information in this section is common knowledge and can be found on many resources such as (16).

Before brain imaging, researchers studied behaviors associated with brain injury in patients. This approach was used in 1861 to determine the corresponding brain regions to the ability of word production, in 1874 for the ability to understand sentences, and 1909 for the visual cortex. In 1848, such a method was used to detect the brain regions associated with the personality of an individual who survived an accident on the train rail. The patient injured his frontal lobe.

In 1895, the X-ray was invented, and it was partially used for brain imaging; however, it could mainly detect bone structure instead of the soft brain tissues. In 1961, Computed Tomography (CT), which is based on X-ray, was introduced. It produces volumetric images. In 1970 the Magnetic resonance imaging (MRI) was introduced. In 2003, the inventor of MRI won the Nobel prize for slice selection and echo planner imaging, which are non-invasive high-speed MRI-based imaging techniques and detect soft tissues. The MRI detects tissue type as well as lesions (injured tissue). It can detect white matter, gray matter, fat, Cerebrospinal Fluid (CFS), tumor, trauma (injury), hemorrhage (internal bleeding), fiber tracts connections, iron concentration,

blood flow, elasticity, and most importantly, BOLD. In addition, there are tissue–specific MRI parameters that MRI machines control based on the target (anatomic, contrast):

- $T1$: Longitudinal relaxation.
- $T2$, $T2^*$: Transitive relaxation.
- Proton density (SO).
- Flow velocity.
- Diffusion coefficients.
- Magnetic susceptibility.

Diffusion-weighted imaging (DWI) is an MRI derivative that was invented in 1980. The word diffusion refers to random thermal motion. Another MRI derivative is Diffusion Tensor Imaging (DTI) that was developed in 1990, as well as Tractography which detects white matter and connectivity.

Before fMRI invention, there used to be other methods for brain imaging. Positron emission tomography (PET) is an invasive method, and it provides spatial labels. It has a higher resolution than the other methods but lower resolution than fMRI. It is impossible to repeat it because it is invasive, i.e., radioactive materials are injected or consumed and must be inside the human body.

Near-infrared (NIR) was used in 1970 to image the hemodynamic activities of the human brain. Brain is semi-transparnet to NIR. Spectroscopy is the study between electromagnetic fields and matters. NIR spectroscopy (NIRS) is a non-invasive method to detect oxygen concentration beneath the skull. The oxygen of the blood has a light absorption feature. This method can be geared toward detecting hemoglobin and total blood concentration. It is wearable with a bed–size producer; therefore, it is movable.

Following NIRS, functional NIRS (fNIRS) was introduced. There are similarities between fMRI signals and fNIRS. However, unlike fMRI, fNIRS cannot detect the activities deeper than 2-3 cm inside the skull, and due to the distance between the skull and the equipment, it has a lower resolution than fMRI. However, fNIRS requires much smaller equipment than fMRI.

In 1924 electroencephalogram (EEG) was introduced. Compared to fMRI, it is inexpensive, and it has a lower resolution. In 1938, it was widely accepted and used in the medical world. It catches the electrical activities of the brain, while it is a non-invasive method. The detectors are electrodes placed on the skull that capture the local saturation of electrical activities of the neurons. It is believed that neurons depolarize to transfer information or activity to one another and then depolarize to have the potential power for the subsequent depolarization. Hence, such an imaging method catches these interactions. EEG is used to diagnose brain death, coma, epilepsy, sleep disorder, and behavioral researches. However, it has limited spatial resolution and limited certainty due to ambiguous localization. In addition, other electrical sources can easily cast noise on this method's output and estimation. Event-related potential (ERP) is a derivative of EEG. It measures the average EEG time-locked when a stimulus is present or when the subject does a task. It is used in cognitive science.

Magnetoencephalography (MEG) measures magnetic fields on the scalp. Compared to fMRI, it is expensive, and it has a lower resolution. Its advantage over EEG is that it is not affected by the inhomogeneous electrical conductivity of the brain. Hence, it is more precise in localization of the brain activity due to low distortion. It records the parallel magnetic fields to the scalp's surface, while EEG records the perpendicular electrical. In 1980, the factories developed MEG devices with 3000 sensors that cover the skull. They are mobile, so they can be used in the task–based brain imaging experiments when subjects move naturally. One of the uses of this method is in

pathology before brain surgery.

The history of brain imaging started with imaging animal subjects' brains by an invasive method. In 1880, Angelo Mosso published a book based on an experiment in which a tilting bed would move toward the head when a stimulus was present. It proved the blood flow of the brain when its regions are active. In 1945, invasive brain imaging methods were used to map the blood flow in the human brain. Notably, in 1960 and 1970, Xenon inhalation was a newly invented method to enter radioactive into the blood and track its path with the help of a scintillator (Luminescence) detector. This approach helped to draw the first functional image of the human brain. In 1980, the study of regional brain change of flow started by scientists. This study was helpful to detect the activities such as speaking, reading, and more and their associated brain regions.

Magnetic resonance spectroscopy (MRS) is another non-invasive method that detects nuclear magnetic resonance (NMR). In chemistry, it is used to detect molecules of an element in solid or liquid materials. MRI detects the abundance of protons in the water, while MRS can be geared toward detecting compounds based on their unique resonant spectroscopy. It can detect many different chemical elements; among them, only hydrogen and phosphorus are present in the human body. It was first used on a mouse head. It is sensitive to a magnetic field which makes the results noisy. It is used in disease detection such as Cancer, Alzheimer's, Parkinson's, and epilepsy.

A chemical substance, Radioligand, was used in 1980 in brain research through an invasive method. Two methods were associated with such a substance: Single-photon emission computed tomography (SPECT) and Positron emission tomography (PET). The first substance for PET was H_2O^{15}. Fludeoxyglucose (FDG) could trace radioactive sugar positron emission via PET. Sugar has a much longer half-life (110 minutes) than the other substances, which makes it suitable for these methods.

In 1991, the first fMRI results were introduced. Before that, magnetic resonance

imaging (MRI) technology was available worldwide. In 1996, enough equipment was available such as high-speed gradient, time–series, echo–planner imaging for fMRI to become the most popular research brain imaging method. fMRI research produces 5,000 papers per year on average. There are 60,000 relevant papers since 1992.

fMRI technique determines activities through time by measuring Blood oxygen level-dependent (BOLD). fMRI has two paradigms: resting-state fMRI and task fMRI, which is done by the presence of a stimulus. However, it is the most dominant brain research methodology, fMRI is still considered cartography (drawing map) because we cannot look at the actual neural activities so far. Instead, we look at the regions to analyze local activities. Besides, fMRI data is a multi-subject database. Due to the noise, individual variants, and unavailable real-time inferences techniques, fMRI has only a few clinical applications. fMRI technique measures Hemodynamic changes via BOLD.

1.2 Classical statistical models

This section is based on a comparison performed by many researchers. As an example, read (17). Proofs at the end of the section are borrowed from different papers that are referenced. Next, we consider and compare different multivariate linear regression sparse models. The advantages and disadvantages of such models are listed. Finally, various penalties are examined for high-dimensional linear regression models.

1.2.1 Setup

As per notation in this section, consider a set of identically and independently distributed (i.i.d) $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ for $i = 1, \ldots, n$. The equation, thus, assumes that the relation between x_i and y_i is linear as follows:

$$y_i = x_i^T \beta + \epsilon_i, \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ is the unknown coefficient vector, and ϵ_i are mean zero random errors. For linear regression, we can ignore the intercept. Centering covariates and responses before running regression would compute the coefficient's vector in the same way as if they had not been centered. However, if centered, the means can be added back to the estimated model's equation to make up the estimated intercept. This fact is not valid- in generalized linear regressions models such as logistic, which is a type of log–likelihood regression. In a matrix form, we have:

$$y = X\beta + \epsilon, \tag{1.2}$$

where y is column vector of responses y_i, X is the design matrix with ith rows x_i^T, and ϵ is column vector of random errors ϵ_i. Take $X_1, \ldots, X_p$ column vectors of X.

1.2.2 Assumptions

There are a few rather strong assumptions to be made.

- $E(y_i|x_i)$ is a linear function of x_i.
- x_i are fixed and ϵ_i are i.i.d. This situation is equivalent to $\epsilon_i = y_i - E(y|x_i)$ to be i.i.d mean zero Gaussian distributed random variable. As a result, homoscedasticity holds.

1.2.3 Least square

The sample level least square coefficients are solutions to the following optimization problem:

$$\min_{\beta} E_n(\mathbf{Y} - \mathbf{X}\beta)^2, \tag{1.3}$$

where $(\mathbf{X}, \mathbf{Y})$ are population versions of (x_i, y_i), and $E_n(\cdot)$ is the empirical mean operator. This problem is equivalent to:

$$\min_{\beta} \|y - X\beta\|_2^2. \tag{1.4}$$

The solution exists and is unique if $rank(X) = p$. This situation is equivalent to a scenario where predictors X_j are linearly independent. In this case, the solution is $\hat{\beta} = (X^T X)^{-1} X^T y$. Fitted values are the projection of y onto the column space of X.

- Advantages:
  - A closed-form solution exists if the above assumption regarding $rank(X)$ is justified in the data. Thus, there is no need to resort to iterative algorithms to approximate the solution.
  - No matter if the solution is or is not unique, fitted values are.
- Disadvantages:
  - If $p > n$, there are multiple solutions (i.e. if $rank(X) < p$).
  - If the solution is not unique, prediction values based on new data can differ from one solution to another. Then, interpretation is meaningless for out-of-sample predictions.

1.2.4 Sparsity

To motivate high dimensional regression models, consider estimating β whose $j_1, \ldots, j_q$-th components are zero. Support of the β denoted by $Supp(\beta)$ is the complement set of $\{j_1, \ldots, j_q\}$ with respect to the super set $\{1, \ldots, p\}$.

1.2.5 Oracle estimator

Oracle estimator is least square estimator as if one knows $O = Supp(\beta)$ before estimating. Least square estimation, in this case, will be

$$\beta = \begin{cases} \hat{\beta}_O = (X_O^T X_O)^{-1} X_O^T y \\ \hat{\beta}_{-O} = 0, \end{cases}$$

where X_O is column of X with indices contained in O, $\hat{\beta}_O$ is elements of $\hat{\beta}$ with indices inside O, and $\hat{\beta}_{-O}$ its complement.

Having oracle properties for an estimator of β usually refers to having root n consistency and selection consistency (identifying the support correctly), asymptotic normality, or combinations of these properties. It means that the estimator would be as good as if we knew the true support and ran least square on the true active set. Thus, a valuable estimator has oracle properties without paying the price of testing all combinations of covariates. Some models' estimators have other features besides selection consistency, while even oracle estimator does not have them. Estimation shrinkage is one of these features.

1.2.6 Regularization: The penalty method

To perform a penalty method on the regression problem, one would start with fixing (a) regularization parameter(s) such as λ. Suppose that we are interested in the least square solution(s) of the problem (1.4) that are in a C_λ, (usually a convex) subset of $\mathbb{R}^p$. If none of the elements of C_λ is a solution to the least square problem, we seek elements of it that minimize the least square problem the most. This approach is the same as finding C_λ's elements with minimum distance to the least square's solutions and equivalent to projecting the least square solution(s) onto C_λ. To measure how close a solution to the least square regression is to the elements of C_λ, we can use a distance function induced by the usual norm in $\mathbb{R}^p$ such as Euclidean.

This method of measuring the distance allows us to aim for a solution that overcomes the above problem- being close to elements of C_λ-and may slightly deviate from the goal of the least square problem, which is to minimize the problem (1.4).

This technique is equivalent to adding a (usually convex) penalty term $P_\lambda(\cdot)$ applied on the vector β to the optimization problem (1.4). We can apply different penalty functions and different regularization parameters for each element of the coefficient vector β by $P^j_{\lambda_j}(\cdot)$ for $j = 1, \ldots, p$. However, to complete the regularization, this setup requires a net search of p parameters λ_j (simultaneously) at the computational level. Hence, it is computationally expensive, given that each λ_j takes a fairly long computational time, especially if the penalized objective function does not have a closed–form solution and requires an iterative algorithm to approximate its solution numerically. Thus, we set all of the λ_j as a unique regularization parameter λ, and all penalty functions the same as $P_\lambda(\cdot)$.

Mostly but not always, the regularization technique introduces a bias to the estimation in return for a variance reductionâan estimation's variance reduction results in lower prediction error. In addition, there are various penalty terms with different features; some lead the structure of the solution(s) to have asymptotic properties, a closed-form solution, different rates of convergences, uniqueness, saturation, or a combination of these features.

The main reason for regularizing the least square problem is often to reduce out–of–sample error which is equivalent to shrinking the variance of the estimation. It can be done through an estimation shrinkage or the sparsity (i.e., setting part of the estimated coefficient vector as zero or removing extraneous covariates). It can be shown that the in-sample error is always lower with a higher number of predictors. Thus, regularization must be done based on a criterion representing the out-of-sample error in the training data. The training data can be divided into two parts: one is used to estimate the unknown coefficients, while the other is treated as the testing data

set. There are various cross-validation methods such as the k–fold or leave one out to partition the data.

1.2.7 Relation between C_λ and $P_\lambda(\cdot)$

As pointed out, solving optimization least square problem (1.4) with constraint C_λ is often equivalent to solving an unconstrained but penalized least square problem. Often, the constraint is defined by a convex function $Q(\cdot)$ such that $C_\lambda = \{\beta \in \mathbb{R}^p | Q(\beta) \leq k_\lambda\}$. It can be shown that the least square problem with such a constraint is equivalent to the unconstrained penalized least square problem where the penalty term $P_\lambda(\cdot) = \lambda Q(\cdot)$ is applied to β and added to the least square objective function. In other words, by making some assumptions about the data, (X, y), it can be demonstrated that for every λ, there is one and only one constant k_λ such that the solutions to the optimization least square problem constrained by C_λ- with the definition above-is the same as the solutions to the penalized least square problem with penalty $P_\lambda(\cdot)$. Relaxing such assumptions, this conjecture holds not precisely but approximately.

1.2.8 Regularization parameter

A regularization parameter can be found at the computational level through a line-search with the minimum mean square error (MSE) criteria among all tested λs. The following are some of the widely used cross-validation methods.

- The k–fold method divides the training data into k divisions. For each fixed λ, we run the regression on one of the k-1 partitions of the kth divisions. It then uses estimated coefficients to predict based on the rest of the data and computes the MSE. The process will be completed when this is done for all of the k combinations of $k-1$ partitions. The result is the mean of all k computed MSEs. In the next step, the fixed λ would change to a different amount, taken from the grid points of the line search. A λ with the minimum mean of MSEs

would be chosen. Finally, a regression would be run on the whole training data set with the chosen λ to compute the unknown coefficients.

- The leave–one–out cross–validation method is essentially the same as the $n-1$ fold cross–validation. It leaves one data point out of the whole training data in each step and treats it as a test set. It then computes MSE. After repeating for all n data points, it takes the average of all such MSEs. Although it can be computationally expensive to compute and find the best λ, it has a closed form solution for some penalty forms. In some cases, reasonably accurate estimation is available instead of a generalized cross-validation's closed form solution of λ. The existence of the closed-form or its approximation makes it preferable over an often regular size of k (such as 10) in the k–fold cross-validation method.

1.2.9 Penalty terms

As stated, some penalized optimization problems do not have a closed–form solution and must be approximated via iterative algorithms. In addition, some of these problems solve for sparse solutions. The active set or $Supp(\hat{\beta})$ is a set of indices that are involved in a regression's estimation at the time the algorithm runs.
The solution path (regularization path) is the estimated coefficient values as a function of regularization parameter: $\hat{\beta}(\lambda)$ for $\lambda \in [0, \infty)$. In some cases, the closed form of the solution path can be derived, and in others, it cannot.

1.2.9.1 Best subset selection

We can set $C_\lambda = \{\beta \in \mathbb{R}^p; \|\beta\|_0 \leq k_\lambda\}$, where $\|.\|_0$ counts the number of non-zero elements of a vector, and k_λ is a natural number that depends on λ. We can take C_λ as the space of constraint of the least square problem (1.4). This is equivalent to solving an unconstrained but penalized least square problem with penalty term $P_\lambda(\cdot) = \lambda\|.\|_0$ (applied on vector β). The solution to each regularization parameter is basically a least square solution for some active estimated coefficients and zero for

others.

- Advantages:
  - The estimator is unbiased.
  - The solution is sparse.
  - The selection consistency holds for this estimator.
  - If X is orthogonal, which is a strong assumption about a data set, a closed-form solution exists: $H_{\sqrt{2\lambda}}(X^T y)$, where $H_t(\cdot)$ is the hard–treshholding function at level t.
- Disadvantages:
  - C_λ is not convex, thus algorithmically, the convex feature does not apply; a local minimum is not global, as opposed to the situation for a convex optimization problem.
  - It is computationally expensive. More precisely, 2^p separate regressions must be run. This is because the selection process and removal of the covariates are not linear: if a regressor is selected at the current step where a set of regressors produce the lowest MSE, there is no explicit instruction to keep pr remove such regressors in the following steps. Thus, all possible subsets of regressors must be tested for the final comparison.
  - It behaves discontinuously with respect to the response values: when y changes, so does the active set-unknown coefficients can be set to be non-zero and estimated with a different set of response values y. As a result, the estimated coefficients jump discontinuously.
  - The above limitation leads to a high estimation's variance, which is the main reason for a high prediction error.

- Although it has selection consistency, this estimator does not shrink the estimated coefficients.
- The solution path cannot be derived mathematically.

1.2.9.2 Stepwise regression

Referred to (18) and (19) that it was appeared for the first time, the forward and the backward selection can be seen as the best subset selection except for the process of selection is sequentially and linear. It starts from an empty active set and gradually enters new variables into the active set (forward) or from a complete set of variables in the active set and removes them one by one (backward). Winner stays: if a new regressor decreases out–of–sample MSE in the cross–validation, it stays in the following steps; otherwise, it will be removed. There are different error criteria, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC). These criteria are equivalent to selecting a new variable that maximizes the absolute correlation with the residual from that of the regression in the last step and entering it into the active set. A modification of this method in signal processing is the orthogonal matching pursuit that enters a new variable into the active set if the variable maximizes the inner product with the residual. The final result is not as good as the best subset selection because the active set is updating gradually throughout the steps, while the best subset selection, at each step, finds the best subset over all possible active sets of that step's size. However, the computation is faster than the best subset selection. (20) and (21) showed a modification of this algorithm called infinitesimal stagewise, the s solution path of which can be as good as that of the LASSO.

1.2.9.3 Stagewise regression

It is similar to the forward stepwise regression. The difference lies in the initial estimation and the iterative algorithm. In order to perform the stagewise regression,

we can start with a zero estimate for all coefficients. Then we find the coordinate(s) that maximizes the absolute normal equation, which is the derivative of ℓ_2 norm, with the currently estimated coefficients. This approach is equivalent to finding the coordinate(s) such that the correlation between the covariate and the current residual is the maximum that it can be. We can update that specific coordinate(s) of estimated coefficients with a gradient descent–like procedure. In this procedure-gradient decent- a term is added by a factor of learning rate. This term has the sign of the normal equation with the currently estimated coefficients multiplied by a factor of the error of the coordinates. We can stop when the difference is insignificant. This procedure takes longer than the stepwise and shorter than the best subset selection. If we replace the sign function with its input in the stagewise update algorithm, it is called ϵ–boosting, gradient boosting, or least squares boosting regression.

1.2.9.4 Ridge regression

We can set $C_\lambda = \{\beta \in \mathbb{R}^p; \|\beta\|_2^2 \leq k_\lambda\}$, where $\|.\|_2$ is ℓ_2 norm, and k_λ is a non-negative number that depends on λ. Here again, C_λ is the space of constraint of the least square problem (1.4). This is equivalent to solving an unconstrained but penalized least square problem with a penalty term $P_\lambda(\cdot) = \lambda\|.\|_2^2$ (applied on vector β). To simplify the closed form solution, we can take $P_\lambda(\cdot) = \frac{\lambda}{2}\|.\|_2^2$. The solution for a fixed λ is $(X^TX + \lambda I)^{-1}X^Ty$, where I is identity matrix. The regularization parameter λ can be found through a cross–validations method.

- Advantages:
  - The penalty term is convex, so the convex problems' feature holds; local minima is global.
  - A closed–form solution exists and is unique.
  - The term $(X^TX + \lambda I)$ in the closed-form is always a non-singular term, and so the existence of the solution does not depend on the $rank(X)$ unlike

in the least square.

- The estimation is performed with a shrinkage feature that can reduce estimation's variance and out-of-sample error.

- Disadvantages:

  - The solution is not sparse because the selection is not performed through this regression.

  - This estimator is asymptotically biased.

1.2.9.5 LASSO regression

Suppose that the constraint space of the least square problem (1.4) is $C_\lambda = \{\beta \in \mathbb{R}^p; \|\beta\|_1 \leq k_\lambda\}$, where $\|.\|_1$ is ℓ_1 norm, and k_λ is a non-negative number that depends on λ. This is equivalent to solving an unconstrained but penalized least square problem with a penalty term $P_\lambda(\cdot) = \lambda\|.\|_1$ (applied on vector β). In this case, the penalty term is not differentiable around the zero vector and the behavior of the subdifferential must be considered that would lead to a treshholding rule.
We can fix λ. The equi-correlation index set $I = \{j \in \{1, \ldots, p\} : |X_j^T(y - X\hat{\beta})| = \lambda\}$ is the active set of LASSO- $Supp(\hat{\beta}_{LASSO})$-that is if $j \notin I$ then $\hat{\beta}_j = 0$. Thus, LASSO solution for a fixed λ is:

$$\beta = \begin{cases} \hat{\beta}_I = (X_I^T X_I)^{-1}(X_I^T y - \lambda \text{sign}(\hat{\beta}_I)) \\ \hat{\beta}_{-I} = 0. \end{cases}$$

Of course, this form of the solution is not practical because $sing(\hat{\beta}_I)$ is unknown. Nevertheless, it helps develop the theoretical results.

- Advantages:

  - The penalty term is convex, so the convex problem feature holds; local minima is global.

- If $X^T X$ is non-singular C_λ is strictly convex, and then the solution is unique.
- Besides performing coefficient shrinkage, it selects covariates, while it is not as computationally expensive as the best subset selection.
- Fitted values $X\hat{\beta}$ are unique.
- Sings of coefficients are unique in different solutions if the solution is not unique.
- Because the fitted values and the sub-differential is unique, the active set for a fixed λ is also unique. I contains the indices of the LASSO solutions that obtain the smallest ℓ_2 norm when plugged into the objective function. These solutions are in the limiting neighborhood of the Elastic Net solution path when ℓ_2 penalty is zero.
- Proven by (22), under a general condition, LASSO has a unique solution. The condition is justified if X has a general position: a finite-dimensional space with a dimension less than the rank of X includes at most the dimensions of positive and negative signs of columns of X ($\pm X_j$).
- The above condition is satisfied if X_{ij} has an (absolutely) continuous joint distribution (with respect to the measure). This assumption makes the Theorem(22) useful when such a condition can be justified for a data set.
- LASSO is a combination of performing the least square on the active set and ridge–type shrinkage on the estimation. Shrinkage is performed when columns of X are orthogonal.
- If X is not orthogonal, shrinkage goes in the wrong direction and enlarges the coefficients instead. However, even in this case, ℓ_1 norm of the coefficients of the LASSO estimation is always less than or equal to the that of the least square estimation on the LASSO active set.

– LASSO estimation is a continuous function of y: estimations gradually go to zero or move away from it when changing response values. The best subset selection is discontinuous, which leads to a high estimation's variance and out-of-sample prediction's error. This feature is because LASSO's fitted values are non-expansive in response values, nor are they Lipschitz continuous with a constant 1 or smaller.

– The solution path is a piecewise linear function of λ, which is continuous. Thus, if the solution is computed for grid points of λs, the whole path can be estimated with an interpolation. The support set with respect to λ, or $I(\lambda) = Supp(\hat{\beta}(\lambda))$ changes. In other words, grid points of λ can be picked between a large value of λ that forces all estimated coefficients to be zero and a small value of λ for which none of the coefficients are estimated as zero. To choose grid points, we only pick λ values at which $|I(\lambda)|$ changes by one unit. Then use these grids to estimate the regularization parameter. This grid makes the cross-validation part of the algorithms that approximate the solution as simple as a step-wise regression in practice. The proof can be found in (17) as well as in (23), (24) and (20).

– Using the above feature, the least angle regression algorithm is developed. However, this algorithm does not allow coefficients to pass below 0. Thus it is not precisely solving for LASSO. See (20).

– We suppose that the following mild assumptions hold: Gaussianity or sub-Gaussianity of the errors, homoscedasticity, and the design matrix X is fixed. An oracle inequality holds even if the population model is not linear. In other words, the distance between the fitted values under the population model and the solution(s) of LASSO divided by the sample size is less than or equal to a decaying term with the rate $\sqrt{n}$ in probability plus a constant term. "In probability" refers to the fact that the chance that the inequality

holds is (at least) $1 - \delta$ for a fixed δ, where the decaying term goes from infinity to a constant that decays at rate $\sqrt{n}$ when δ moves from 0 to 1. This equation shows that the distance between the best subset selection solution and the LASSO solution have the same situation-a decaying upper bound without the constant term. This inequality is referred to as the oracle inequality.

– Proven by (25), LASSO can have faster convergence results than stated in the disadvantage part-below. We can assume the assumptions in the previous paragraph. Furthermore, we can assume that X has a compatibility condition with respect to the true support, then LASSO can have an in–sample risk converging at a rate of n similar to that of the best subset selection. It can also be shown that the ℓ_1 distance of the estimated values and the true values of the coefficients, or error bound in ℓ_1 norm, converges to zero in probability at a rate of $\sqrt{n}$.

– Instead of compatibility condition, we assume that the design matrix justifies the restricted eigenvalue condition. In this case, the error bound in ℓ_2 norm converges to zero in probability at the rate of n.

– Proven by (26) through the primal-dual witness method, LASSO has a support recovery feature. That is, if X_I has full rank, the minimum eigenvalue condition holds, and the mutual incoherence condition is satisfied on top of the above basic assumptions regarding errors' distribution and structure of X. Then, it can be shown that the active set of the solution(s) is precisely the same as the true active set with a high probability for some regularization parameters.

– Shown by (27) and justified by the worst-case of orthogonal X in (28), the upper bound of the minimax prediction error decays in probability at a rate of n. While under some eigenvalue restriction assumption, the upper

bound would be equal to the minimax prediction error's lower bound with a probability of at least half.

- Disadvantages:
  - The LASSO estimator is biased.
  - C_λ is not strictly convex if $X^T X$ is singular (which is the case when $p > n$).
  - The solution is not unique.
  - The closed form solution does not exist.
  - Iterative algorithms are used to approximate the solution(s) numerically.
  - Due to the non-uniqueness of the solution, the out-of-sample prediction is not well-defined.
  - Saturation: While the general position of X guarantees uniqueness, it has a drawback. The general position implies that X_I has a full rank where $I = Supp(\hat{\beta}_{\text{LASSO}})$. Then, $|I|$ is less than or equal to the rank of the design matrix. Therefore, the number of sparse coefficients in the solution cannot be greater than the sample size. Thus, LASSO cannot achieve sparsity more than the sample size under the general position's assumption, which guarantees uniqueness.
  - Although in theory, LASSO performs a ridge–type shrinkage while running the least square regression on the active set, the ridge shrinkage of LASSO can go in a wrong direction due to a possible high correlation between variables in the active set. Then it can enlarge the estimation instead.
  - We can take the in-sample risk's bound as the expected value of the ℓ_2 distance between the fitted values and the true response values, divided by the sample size. We can suppose some mild assumptions hold: such as Gaussianity or sub-Gaussianity of errors, homoscedasticity, and suppose

X is fixed. Then, it can be readily shown that the in-sample risk's bound with respect to the sample size decreases slower than the risk bound of the best subset selection. Precisely one decreases $\sqrt{n}$ faster than the other.

- Consider the out-of-sample or predictive risk in similar reasoning to the above paragraph. On top of the above assumptions, we assume that a new data point is independent of the design matrix but with the same unknown distribution. (29) shows a similar result for the in-sample risk of the LASSO. The bound is again slower than the bound of the least square regression by a factor of $\sqrt{n}$.
- (30) and (31) show that the oracle property (defined below in the adaptive LASSO) does not hold.

1.2.9.6 Relaxed LASSO

In order to perform the relaxed LASSO, we start with running the LASSO. We then store its solution's active set as well as the regularization parameter. Next, we run another LASSO regression only on the stored active set while scaling the previous regularization parameter by a new regularization parameter which is a number less than 1. If this scale is zero, it is precisely running the least square regression on the active set after the LASSO. (32) invented such a method to overcome the problematic situation where the LASSO fails to shrink and instead enlarges the coefficients' estimations.

1.2.9.7 Adaptive LASSO

To reduce the bias in LASSO estimation, adaptive LASSO was introduced by (31) that penalizes the larger estimated coefficients-in terms of magnitude- heavier. It uses a weight term for each variable, which is reciprocal power to a (second) regularization parameter of the magnitude of an initial estimation. Initial estimation can be least square, ridge, or LASSO. It is proven that if the initial estimation is $\sqrt{n}$ con-

sistent, adaptive LASSO enjoys the oracle property: The selection of the active set happens correctly with a probability converging to one. At the same time, asymptotic normality holds with the covariance of the oracle estimator.

1.2.9.8 Elastic net

Elastic Net is a combination of LASSO and ridge regression with two regularization parameters for each penalty. The motivation behind (33)'s works were to overcome the strict convexity of LASSO (which causes non-uniqueness) and saturation. It also pulls variables with similar effects (strong correlation) to the same direction in or out of the final active set.

1.2.9.9 Non-convex penalties with good theoretical properties

The least square optimization problem can be written with non-convex penalties such as bridge or power penalties: ℓ_α norms for an $\alpha < 1$. The SCAD and the MC+ are two common non-convex penalties that obtain oracle properties. However, local minima are not global, and in general cases, an algorithm to compute the solutions is not developed; the developed theory under some mild assumptions gives an algorithm to approximate the solutions.

1.2.10 Group penalties

Suppose one would want to run a high–dimensional regression and select covariates, and perhaps performs shrinkage as well. We can assume they take all of the powers of each variable from one to ten and would want to know if any of these powers have a significant role in the regression. In this situation, each variable and its powers can be treated as a group. Penalties can be written with respect to these groups instead of single coordinates. There are other scenarios where a group penalty can help the investigation, such as functional regression that we will explore.

Grouping can be performed on most of the above penalty terms, such as LASSO, Scad, and Elastic Net. In addition, some of the above modifications such as adaptive

LASSO can also be performed via group penalties.

1.2.10.1 Group LASSO

Group LASSO was introduced by (34) and (35). We take $\beta = ((\beta^1)^T, \ldots, (\beta^G)^T)$ where β^g are partitions associated with groups of variables.

We suppose that the convex constraint space of the least square problem (1.4) is $C_\lambda = \{\beta \in \mathbb{R}^p; \sum_1^G \sqrt{p_g}\|\beta^g\|_2 \leq k_\lambda\}$, where $\|.\|_2$ is the ℓ_2 norm, k_λ is a non-negative number that depends on λ, and p_g is the dimension of β^g. This is equivalent to solving an unconstrained but penalized least square problem with a penalty term $P_\lambda(\beta) = \lambda \sum_1^G \sqrt{p_g}\|\beta^g\|_2$. We note that in this situation, X is divided to blocks X^j that are matrices of n by p_j for $j = 1, \ldots, G$.

Multi-task learning is a similar method where grouping is based on different response values from different multiple regression problems.

One of the main advantages is that it can be shown that the adaptive group LASSO enjoys oracle property similar to the adaptive LASSO. See (36).

The first asymptotic property of the group LASSO was proven by (37). Furthermore, it is shown that the group LASSO estimator under not-too–strong assumptions is consistent both in terms of selection and estimation in the ℓ_2 norm.

1.2.10.2 Theoretical results: notations, assumptions, and the model

Since the regularization parameter λ depends on the sample size, we denote the regularization parameter by λ_n.

The following three assumptions would be repeated in the upcoming Theorems, where they would be referred to as the three basic assumptions.

Populations' assumptions on the joint distribution of $(\mathbf{X}, \mathbf{Y})$.

- The fourth moments of X, and Y are finite. $E(\mathbf{Y}^4) < \infty$, and $E(\|\mathbf{X}\|^4) < \infty$.
- The covariance matrix of $\mathbf{X}$ is non-singular and is invertible $\Sigma_{\mathbf{XX}} = E(\mathbf{X}\mathbf{X}^T) - E(\mathbf{X})E(\mathbf{X}^T)$.

- For any minimizer of $E(\mathbf{Y}-\mathbf{X}\beta)^2$ with respect to β, $E((\mathbf{Y}-\mathbf{X}\beta)^2|X)$ is almost surely positive, and greater than $\sigma_{\min} > 0$.

In the last assumption, we are not trying to fix the conditional variance. Instead, it intuitively states that given the covariates, even using the best linear predictors to explain the variability in the response value, there is a strictly positive variability left at the population level. The second assumption guarantees uniqueness of the population problem's solution $\beta = \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}$ where $\Sigma_{\mathbf{XY}} = E(\mathbf{X}\mathbf{Y}^T) - E(\mathbf{X})E(\mathbf{Y})$. By centering, the constant is removed from the model. Thus, the theoretical results are only stated for β- without the constant part. However, the same results for the constant in the model would follow immediately after the results for β.

Denote the population error by $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\beta$.

Consider $\hat{\beta}$, the solution to the optimization problem:

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda_n \sum_{1}^{G} w_j \|\beta^g\|_2, \tag{1.5}$$

where w_j are fixed scalar weights. This problem is equivalent to:

$$\min_{\beta} \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{XY}\beta + \frac{1}{2}\beta^T\hat{\Sigma}_{XX}\beta + \lambda_n \sum_{1}^{G} w_j \|\beta^g\|_2. \tag{1.6}$$

If the second assumption holds, with a probability tending to 1, the solution to (1.6) exists uniquely.

We define the true support set (active set) of the coefficients with $Supp(\beta)$. It is a subset of indices $j = 1, \ldots, G$ of blocks of β such as β^j (with size p_j) such that $\beta^j \neq \overrightarrow{0}$. We define the $Supp(\hat{\beta})$ similarly. Using sub–differential of (1.5), the support

$I = Supp(\hat{\beta})$ is such that for any $j \in I$, and $i \notin I$

$$\begin{cases} \|\hat{\Sigma}_{X^i X}\beta - \hat{\Sigma}_{X^i Y}\| \leq \lambda_n w_i \\ \hat{\Sigma}_{X^j X}\beta - \hat{\Sigma}_{X^j Y} = -\frac{\lambda_n w_j}{\|\beta^j\|}\beta^j. \end{cases} \tag{1.7}$$

The following assumptions limit most of the population covariance norm only between the truly active variables $\mathbf{X}^O$ which is a partition of random variable of $\mathbf{X}$ associated with $O = Supp(\beta)$. Take $D = diag(w_j/\|\beta^j\|)$. For any $j \in O^c$:

Assumption 1. $\|\Sigma_{\mathbf{X}^j\mathbf{X}^I}\Sigma_{\mathbf{X}^O\mathbf{X}^O}D\beta^O\| < w_j$

Assumption 2. $\|\Sigma_{\mathbf{X}^j\mathbf{X}^I}\Sigma_{\mathbf{X}^O\mathbf{X}^O}D\beta^O\| \leq w_j$

The first assumption is stronger. It is a necessary and a sufficient condition for the group LASSO to have a consistent solution path proven by (38). It leads to unbiasedness in ℓ_2 norm (consistency) as well as in ℓ_0 norm-selection consistency.

It is shown in (39) and (40) that if the dimension of the blocks is one, these conditions are tightly similar to those of the LASSO consistency.

The notation $f(n) = \omega(g(n))$ means $\lim_{n\longrightarrow\infty} |\frac{f(n)}{g(n)}| = \infty$.

1.2.10.3 Theoretical results

Theorem 1. *Suppose that the tree basic assumptions are hold as well as the strong assumption 1. For any sequence of λ_n such that $\lambda_n = o(1)$ and $\lambda_n\sqrt{n} = \omega(1)$, a solution to (1.5) $\hat{\beta}_n \xrightarrow{p} \beta$, and $I = Supp(\hat{\beta}_n) \xrightarrow{p} Supp(\beta) = O$.*

Theorem 2. *Suppose that the tree basic assumptions are hold. If there is a sequence of λ_n such that solution to (1.5) $\hat{\beta}_n \xrightarrow{p} \beta$ and $I \xrightarrow{p} O$ then the weak assumption (2) must hold.*

In general, M-estimation theory states that if the sample–level of an objective function converges uniformly to the population level of it, its optimizer will converge to

the population's optimizer in probability. The uniform convergence of such an objective function would be guaranteed by the Uniform Law of Large number that holds if the function is continuous in β^O, the parameter space is compact, and some technical conditions on the measurability holds. More details can be found in the following article. The asymptotic result of the M-estimation of Theorem 1 by (41) and (30) are used. Here is a sketch of the asymptotic consistency proof of M-estimation in terms of maximization.

Lemma 1. *Suppose $\hat{\theta}_n$ maximizes M_n which is the sample level objective function of M that is minimized by θ_0. Assume that:*

- $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$ *which is uniform convergence.*
- $\forall \epsilon > 0$ *we have* $\sup\{M(\theta) | d(\theta, \theta_0) > \epsilon\} \leq M(\theta_0)$. *This is weaker than the uniqueness of θ_0. This is so called wide-spread maximum.*
- $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$.

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Note that all three above assumptions would be justified if Θ is compact, M is continuous, and θ_0 is unique. The first two of these stronger alternative assumptions leads to uniform convergence by The Uniform Law of Large Number.

Proof. By the second assumption for any $\epsilon > 0$, there is a $\delta > 0$ such that

$$P(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \leq P(M(\theta_0) - M(\hat{\theta}_n) \geq \delta) \tag{1.8}$$

Inside the right hand side of the probability add and subtract two terms $M_n(\theta_0)$ and $M_n(\hat{\theta}_n)$. Finally, by the feature of the union and the rule of inclusion (ignoring the subtraction of intersections), the above probability is less than or equal to the

following summation of three probabilities:

$$P(M(\theta_0) - M_n(\theta_0) > \delta/3) + P(M_n(\theta_0) - M_n(\hat{\theta}_n) > \delta/3) + P(M_n(\hat{\theta}_n) - M(\hat{\theta}_n) > \delta/3). \tag{1.9}$$

The first and the third probability terms go to zero by the first assumption. Similarly, the third assumption implies that the middle probability goes to zero. □

The following Lemma helps to prove Theorem 1. It states that under the three above basic assumptions, if the $Supp(\beta)$ is known and is used accordingly, then $\hat{\beta}_n \xrightarrow{p} \beta$ for any $\lambda_n = o(1)$.

Lemma 2. *Suppose that the three basic assumptions hold. First, denote $O = Supp(\beta)$ and X^O as the blocks of X associated with groups indexed in O. A minimizer of*

$$\begin{aligned} &\min_{\beta^O} \frac{1}{2n}\|y - X^O\beta^O\|_2^2 + \lambda_n \sum_{g\in O} \|\beta^g\|_2 = \\ &\min_{\beta^O} \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{X^OY}\beta^O + \frac{1}{2}(\beta^O)^T\hat{\Sigma}_{X^OX^O}\beta^O + \lambda_n \sum_{g\in O} w_j\|\beta^g\|_2. \end{aligned} \tag{1.10}$$

converges to β^O in probability if $\lambda_n \to 0$.

Proof. If $\lambda_n \xrightarrow{n\to\infty} 0$, the objective function (1.10) converges to an objective function of β^O such as:

$$\frac{1}{2}\Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{X}^O\mathbf{Y}}\beta^O + \frac{1}{2}(\beta^O)^T\Sigma_{\mathbf{X}^O\mathbf{X}^O}\beta^O + \lambda_n \sum_{g\in O} w_j\|\beta^g\|_2. \tag{1.11}$$

The global minimum of such an objective function is uniquely the vector of true coefficients β^O. This is because by the basic assumption 2, $\Sigma_{\mathbf{X}^O\mathbf{X}^O}$ is positive definite. Thus, the function is quadratic form and strictly convex. Using the results of M-estimation, the proof is complete. □

We take a minimizer of (1.10). First, we denote such vector with $\hat{\beta}^O$. Later, we

extend it with zero on coordinate blocks associated to O^C to have a vector with the exact size of β.

Proof of Theorem 1.

Proof. Lemma 2 shows that $\hat{\beta}^O$ is an ℓ_2 consistent estimator of β. We show that the probability that $\hat{\beta}^O$ is optimal for (1.6) tends to one. As a reminder, because of the second basic assumption that $\Sigma_{\mathbf{XX}}$ is non-singular, solution of (1.6) exists uniquely with probability tending to 1. We need to show that $\hat{\beta}^O$ satisfies two conditions in (1.7) with probability tending to one. By definition of $\hat{\beta}^O$, it must satisfy the second condition of (1.7). The first one needs to be established.

A simple algebra shows:

$$\hat{\Sigma}_{XY} = \hat{\Sigma}_{XY}\beta + \hat{\Sigma}_{X\epsilon}, \tag{1.12}$$

where $\epsilon = Y - X\beta$. By the third basic assumption $\Sigma_{\boldsymbol{\epsilon X}} = 0$. Thus,

$$\hat{\Sigma}_{XY} = (\Sigma_{\mathbf{XY}} + O_p(n^{-1/2}))\beta + O_p(n^{-1/2}). \tag{1.13}$$

This is because of the convergence of the empirical covariances to population covariances (41) that can be applied due to the first basic assumption–the fourth population moments are bounded. We can change the notation to β^O, which has a smaller vector size–limited to the true nonzero partitions of true coefficients. A simple algebra shows:

$$\hat{\Sigma}_{XY} = \Sigma_{\mathbf{X}\mathbf{X}^O}\beta^O + O_p(n^{-1/2}). \tag{1.14}$$

Recall that $\hat{\beta}^O$ is consistent and bounded in probability. Thus, we have:

$$\hat{\Sigma}_{XY} - \hat{\Sigma}_{XX^O}\hat{\beta}^O = \Sigma_{\mathbf{X}\mathbf{X}^O}(\beta^O - \hat{\beta}^O) + O_p(n^{-1/2}). \tag{1.15}$$

By removing appropriate rows from each side, we modify the previous equation (1.15) to the following two equations:

$$\hat{\Sigma}_{X^O Y} - \hat{\Sigma}_{X^O X^O}\hat{\beta}^O = \Sigma_{\mathbf{X}^O \mathbf{X}^O}(\beta^O - \hat{\beta}^O) + O_p(n^{-1/2}). \tag{1.16}$$

$$\hat{\Sigma}_{X^C Y} - \hat{\Sigma}_{X^C X^O}\hat{\beta}^O = \Sigma_{\mathbf{X}^C \mathbf{X}^O}(\beta^O - \hat{\beta}^O) + O_p(n^{-1/2}), \tag{1.17}$$

where $C = O^C$ is the complement of $Supp(\beta)$. The second condition of (1.7) which is already satisfied according to the definition of $\hat{\beta}^O$ is

$$\hat{\Sigma}_{X^O X^O}\hat{\beta}^O - \hat{\Sigma}_{X^O Y} = -\lambda_n D^O \hat{\beta}^O \tag{1.18}$$

where $D^O = Diag(\frac{w_j}{\|\hat{\beta}^j\|})$ for $j \in O$. Multiplying equation (1.18) by $\Sigma^{-1}_{\mathbf{X}^O \mathbf{X}^O}$ we have:

$$\Sigma^{-1}_{X^O X^O}(\hat{\Sigma}_{X^O X^O}\hat{\beta}^O - \hat{\Sigma}_{X^O Y}) = -\lambda_n \Sigma^{-1}_{X^O X^O} D^O \hat{\beta}^O \tag{1.19}$$

Using (1.16) and (1.18):

$$\hat{\beta}^O - \beta^O = -\lambda_n (\Sigma_{\mathbf{X}^O \mathbf{X}^O})^{-1} D^O \hat{\beta}^O + O_p(n^{-1/2}). \tag{1.20}$$

Then, we use (1.17) and (1.20):

$$\hat{\Sigma}_{X^C Y} - \hat{\Sigma}_{X^C X^O}\hat{\beta}^O = \lambda_n \Sigma_{\mathbf{X}^C \mathbf{X}^O}(\Sigma_{\mathbf{X}^O \mathbf{X}^O})^{-1} D^O \hat{\beta}^O + O_p(n^{-1/2}). \tag{1.21}$$

In particular, for any $i \in C$:

$$\hat{\Sigma}_{X^i Y} - \hat{\Sigma}_{X^i X^O}\hat{\beta}^O = \lambda_n \Sigma_{\mathbf{X}^i \mathbf{X}^O}(\Sigma_{\mathbf{X}^O \mathbf{X}^O})^{-1} D^O \hat{\beta}^O + O_p(n^{-1/2}). \tag{1.22}$$

We can divide both sides of (1.22) by $w_i \lambda_n$ before taking the limit. Left hand side

$\frac{1}{\lambda_n w_i}(\hat{\Sigma}_{X^iY} - \hat{\Sigma}_{X^iX^O}\hat{\beta}^O)$ would converge to the following in probability:

$$\frac{1}{w_i}\Sigma_{\mathbf{X}^i\mathbf{X}^O}(\Sigma_{\mathbf{X}^O\mathbf{X}^O})^{-1}D\beta^O \tag{1.23}$$

where D is defined in assumption 1. This is because of the consistency of $\hat{\beta}^O$ and $\frac{O_p(n^{-1/2})}{\lambda_n} = o_p(1)$ due to the fact that fact that $\lambda_n\sqrt{n} = \omega(n)$. We use assumption 1 which states that the norm of (1.23) is strictly less than one-divide both sides by w_j. Thus, the probability of $\hat{\beta}^O$ being feasible for (1.6) is:

$$\lim_{n\to\infty} \mathbb{P}\{\forall i \in O^C, \|\frac{1}{\lambda_n w_i}(\hat{\Sigma}_{X^iY} - \hat{\Sigma}_{X^iX^O}\hat{\beta}^O)\| \leq 1\} = 1. \tag{1.24}$$

Because if $\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \mathbb{P}(B_n) = 1$, by contradiction it can be proven that $\lim_{n\to\infty} \mathbb{P}(A_n \cap B_n) = 1$. The fact from the assumption 1 that for each $i \in C$, $\lim_{n\to\infty} \mathbb{P}\{\|\frac{1}{\lambda_n w_i}(\hat{\Sigma}_{X^iY} - \hat{\Sigma}_{X^iX^O}\hat{\beta}^O)\| \leq 1\} = 1$ was used. □

Note that this Theorem is proved $\forall i \in O^C$, which is the set of indices such that $\hat{\beta}^i = 0$. It is the support of $\hat{\beta}^O$ by construction. Thus, the optimality condition related to the sub-differential holds for such a vector.

1.2.10.4 Logistic group LASSO

A group penalty can be applied to regression models other than the least square. An example is the generalized linear model and, in particular logistic regression. When penalizing this regression model, it becomes a penalized likelihood logistic regression. Applications arise when the response values are binary in a classification problem. It is shown in (42) that the group LASSO estimate is consistent in this case.

1.2.10.5 Theoretical results: notations, assumptions, and the problem

We assume that $\mathbf{Y}$ follows a Bernoulli distribution. We denote $\mathbb{P}(\mathbf{Y} = 1|X) = \mathbf{P}$. We suppose that log of odds is linear in X: $log(\mathbf{P}/(1-\mathbf{P})) = \alpha + \mathbf{X}\beta$. $\mathbf{P}$ can be written as function of β and $\mathbf{X}$. Denote it by $\mathbf{P}_\beta(\mathbf{X})$. This is equivalent to

setting $\mathbf{P}_\beta(\mathbf{X}) = S(\alpha + \mathbf{X}\beta)$ where $S(\cdot)$ is the Sigmoid function. As opposed to the least square regression, it is not possible to remove the intercept by centering the populating variables $(\mathbf{X}, \mathbf{Y})$ or their sample level counterparts. In this set of notations, we consider the groups as: $\beta = (\alpha, (\beta^1)^T, \ldots, (\beta^G)^T)$. In the sample level, the group penalized problem is to find the minimizer of the following objective function:

$$\min_{\beta} -l(\beta) + \lambda_n \sum_{1}^{G} w(p_g)\|\beta^g\|_2, \tag{1.25}$$

where $l(\beta)$ is the log-likelihood and $w(p_g) = \sqrt{p_g}$. The scalars $w(p_g) = \sqrt{p_g}$ re-scales the penalty term: the larger the partition the heavier the penalization on them. Intuitively, the larger the partition size, the more variables or function of them are in that partition. Thus, it must contribute more to the correlation. As a result it must be penalized more and this procedure would still give a reasonable estimate. We note that that the intercept is not penalized. The log–likelihood loss can be written in alternative forms. One of these forms is as follows:

$$\sum_{i=1}^{n} y_i(\eta_\beta(x_i) - \log(1 + \exp(\eta_\beta(x_i))), \tag{1.26}$$

where the link function $\eta_\beta(x_i)$ is $\alpha + x_i^T\beta$. Because β is grouped, we use the following notation. We consider $\eta_\beta(x_i) = \alpha + \sum_{g=1}^{G}(x_i^g)^T\beta^g$, where x_i^g is a partition of vector x_i associated with the group g.

Lemma 3. *If* $0 < \sum_{i=1}^{n} y_i < n$ *the minimum in the optimization problem (1.25) is attained.*

Proof. Let $\beta^1, \ldots, \beta^G$ be fixed. Intercept α is the minimizer of

$$g(\alpha) = -\sum_{i=1}^{n}[y_i(\alpha + c_i)] - log(1 + \exp(\alpha + c_i))], \tag{1.27}$$

where $c_i = \sum_{g=1}^{G} (x_i^G)^T \beta^G$. The derivative of $g(\cdot)$ is:

$$g'(\alpha) = -\sum_{i=1}^{n} \{y_i - \frac{\exp(\alpha + c_i)}{1 + 1 + \exp(c_i)}\} \tag{1.28}$$

We have $\lim_{\alpha \to \infty} g'(\alpha) = n - \sum_{i=1}^{n} y_i$ that is assumed to be positive in this lemma. Also, $\lim_{\alpha \to -\infty} g'(\alpha) = -\sum_{i=1}^{n} y_i$ is assumed to be negative. In addition, $g'(\cdot)$ is continuous and strictly increasing. The second derivative is $g''(\alpha) = \sum_{i=1}^{n} \{S(\alpha + c_i)(1 - S(\alpha + c_i))\}$ where $S(\cdot)$ is the Sigmoid function that is bounded between zero and one. Therefore, the second derivative is strictly positive. Thus, by the mean value Theorem there must be a unique $\alpha^* \in \mathbb{R}$ such that $g'(\alpha) = 0$. This value-α-is a function of fixed $\beta^1, \ldots, \beta^G$. We denote such a function with $\alpha^*(\beta^1, \ldots, \beta^G)$. We replace α in (1.25) with $\alpha^*(\beta^1, \ldots, \beta^G)$. This lets us eliminate the intercept in Lemma regarding obtaining the minimizer. The optimization problem (1.25) with such a notation can be written as an optimization problem with new sets of variables $\beta^1, \ldots, \beta^G$ and without a penalty term. Instead of the penalty term, the optimization problem is under constraint $\sum_{1}^{G} w(p_g) \|\beta^g\|_2 < t$, for some $t > 0$. This is an optimization problem of a continuous function over a compact and convex set. Thus, by the theory of duality, the minimum is attained. □

If X is of full rank, the minimizer of (1.25) is unique; otherwise, the minimizer is an element of a convex set. This convex set is such that all of its elements minimize the objective function with the same objective value. This assumption at the population level would be equivalent to the second population assumption in the theoretical results of the group LASSO section 1.2.10.2. If such an assumption holds for the population-the covariance matrix is non-singular- the solution is unique for the population level optimization problem of the logistic regression model as well.

1.2.10.6 Consistency

We take the negative log–likelihood as the loss function. We define the function $\gamma_\beta(\cdot,\cdot)$ as follows.

$$\gamma_\beta(x,y) = -(y\eta_\beta(x) - \log(1+\exp\{\eta_\beta(x)\}).$$

We define the population (or theoretical) risk as $R(\beta) = \mathbb{E}[\gamma(\mathbf{X},\mathbf{Y})]$. It follows that the empirical counterpart of the population risk is $R_n(\beta) = \mathbb{E}[\gamma(\mathbf{X},\mathbf{Y})] = \sum_{i=1}^{n} \frac{1}{n}\eta_\beta(x_i,y_i)$. For now, we remove n from λ_n and divide the optimization problem (1.25) by the sample size:

$$\hat{\beta}_\lambda = \min_\beta \frac{-l(\beta)}{n} + \frac{\lambda}{n}\sum_1^G w(p_g)\|\beta^g\|_2 = R_n(\beta) + \frac{\lambda}{n}\sum_1^G w(p_g)\|\beta^g\|_2. \tag{1.29}$$

As it can be seen, with such a notation, the penalized empirical risk is the same as the optimization problem (1.25) divided by the sample size. We consider $\beta_0 = \arg\min_\beta R(\beta)$. If the model is well–specified, the distribution of $\mathbf{Y}$ is as it is supposed to be in the population before taking a sample. Thus, if the model is well–specified we have; $\mathbf{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{P}_{\beta_0}(\mathbf{X})$. We use the following distance function to measure the distance between the estimation and true value.

$$d^2(\eta_{\hat{\beta}},\eta_{\beta_0}) = \mathbb{E}[|\eta_{\hat{\beta}_\lambda}(\mathbf{X}) - \eta_{\beta_0}(\mathbf{X})|^2]. \tag{1.30}$$

The following three assumptions are necessary for the next theoretical results.

Assumption 3. *For some constant* $0 < \epsilon \leq 0.5$*:*

$$\epsilon \leq \mathbf{P}_{\beta_0}(\mathbf{X}) \leq 1-\epsilon.$$

Assumption 4. *The covariance matrix of* $\mathbf{X}$ *is non-singular and invertible:* $\Sigma_{\mathbf{XX}} =$

$E(\mathbf{X}\mathbf{X}^T)$. We denote the smallest eigenvalue of it by ν^2.

Assumption 5. *We normalize X^g so that $\mathbf{E}[X^g(X^g)^T] = I_{p_g}$ for $g = 1, \ldots, G$. In addition, we assume that there is a constant L_n such that:*

$$\max_X \max_g((X^g)^T X^g) \leq nL_n^2. \tag{1.31}$$

The smallest-fastest-possible rate of L_n is $O(1/n)$ due to the normalization. Intuitively, the expected value of the normalized matrix is the identity where the trace of it is p_g. Thus, the maximum of the summation over squared of elements on average (: divided by n) cannot approach zero faster than a constant divided by n ($O(1/n)$). If some covariates are categorical then $L_n = O(1/n)$. This means that the probability of at least one incident different from others is bounded away from 0 and 1.

We denote N_0 as the number of nonzero elements of β_0. There exist global constants $C1, C2, C3, C4$ and constants c_1, c_2 depending on ϵ and ν and $\max_g(p_g)$ such that when two following conditions are satisfied:

Assumption 6. $C_1(1 + N_0^2)L_n^2 \log(G) \leq c_1$ *and* $C_1 \log(G) \leq \lambda \leq c_1/(1 + N_0^2)L_n^2$,

then it can be shown that:

$$\mathbb{P}\{d^2(\eta_{\hat{\beta}}, \eta_{\beta_0}) \geq c_2 \frac{(1 + N_0)\lambda}{n}\} \leq C_2\{\log(n) \exp(-\frac{\lambda}{C_3}) + \exp(-\frac{1}{C_4 L_n^2})\} \tag{1.32}$$

Proof. The argument of this result is similar to one in (43). In this paper, a hing loss function is considered instead of logistic, which can be extended to a logistic model. The reason is that the only thing about the loss function used in the proof is its Lipschitz property. By the assumption 3, the logistic loss function is quadratic near its (set of) minimizer(s). In the above paper's proof, there is a difficulty due to the unknownness of the marginal behavior of the hinge loss. However, due to the quadratic behavior of the logistic, it is not an issue here. The group LASSO problem

is reduced to ℓ_1 penalty similar to that of LASSO if each $p_g = 1$. The extension of the consistency results from the LASSO to group LASSO is possible if $\max_g p_g$ does not depend on n. Furthermore, the group LASSO uses a normalization of partitions of the design matrix. To extend the consistency results from LASSO to group LASSO, it is necessary to show that such normalization is uniformly close (or converging) to the theoretical normalization in the population. It means that it must be shown that the empirical and the theoretical eigenvalues of the design matrix per group cannot be too far uniformly over the groups. □

The equation (1.32) can complete the asymptotic consistency result by assuming that ϵ, ν, and $\max_g(p_g)$ are fixed and that $G \gg \log(n)$. We take $\lambda \sim \log(G)$. We suppose that $N_0 = O(1)$ and $L_n^2 = O(1/\log(G))$ which means it is $\leq O(1/log(log(n)))$. This bound is larger than $O(1/n)$. Thus, this assumption is

$$O(1/n) \leq L_n^2 \leq O(1/\log(\log(n))).$$

Then, we have

$$d^2(\eta_{\hat{\beta}}, \eta_{\beta_0}) = O_P(\log(G)/n),$$

which is the parametric rate. This is because if the inequality inside $\mathbb{P}$ is divided by $\log(G)/n$, the limit of the right hand side when $n \to \infty$ would go to zero which is the definition of $d^2(\eta_{\hat{\beta}}, \eta_{\beta_0})/(\log(G)/n) = O_P(1)$. The limit of the right hand side is readily seen to be zero due to above assumptions. If $L_n = O(1/n)$, the maximal growth of N_0 is $O(\sqrt{n/\log(G)})$. When N_0 is exactly of such order:

$$d^2(\eta_{\hat{\beta}}, \eta_{\beta_0}) = O_P(\sqrt{\log(G)/n}).$$

This result can be proven by replacing N_0 by number of best approximations of η_{β_0}. This approximation balances estimate's and approximation's error. Such results can

be proven for a Gaussian regression as well.

CHAPTER 2: FUNCTIONAL REGRESSION

2.1 Preliminary and notation

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let T_j be a compact set in $\mathbb{R}^{d_j}$ for $j = 1, \ldots, p$. Let $\mathcal{H}^1, \ldots, \mathcal{H}^p$ be separable Hilbert spaces of functions from T_j to $\mathbb{R}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}^j}$. Let $\mathcal{H} = \mathcal{H}^1 \times \cdots \times \mathcal{H}^p$ be endowed with the inner product

$$\langle f, g \rangle_{\mathcal{H}} = \langle f^1, g^1 \rangle_{\mathcal{H}^1} + \ldots + \langle f^p, g^p \rangle_{\mathcal{H}^p},$$

for any $f = (f^1, \ldots, f^p)^\mathsf{T} \in \mathcal{H}$, and $g = (g^1, \ldots, g^p)^\mathsf{T} \in \mathcal{H}$. Then, $\mathcal{H}$ is also a separable Hilbert space. Let $X : \Omega \to \mathcal{H}$ be a measurable function with respect to $\mathcal{F}_X/\mathcal{B}_X$ where $\mathcal{B}_X$ is the Borel σ-field generated by open sets in $\mathcal{H}$.

Let X be a random element in $\mathcal{H}$. If $E\|X\|_{\mathcal{H}} < \infty$; then, the linear functional $f \mapsto E\langle f, X \rangle_{\mathcal{H}}$ is bounded. By Riesz's representation Theorem, there is a unique element in $\mathcal{H}$, say μ_X, such that $\langle \mu_X, f \rangle_{\mathcal{H}} = E\langle f, X \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$. See (44). We call μ_X the mean element of X or expectation of X. If we can further assume $E\|X\|^2_{\mathcal{H}} < \infty$, the operator $\mathcal{H} \to \mathcal{H}$,

$$\Gamma_{XX} = E[\{X - E(X)\} \otimes \{X - E(X)\}], \tag{2.1}$$

exists and is a Hilbert-Schmidt operator, where $\otimes$ indicates a tensor product computed in a way that for $x, y, z \in \mathcal{H}$, $(x \otimes y)(z) = \langle y, z \rangle_{\mathcal{H}} x$. See (45).

Let Y be a random element in $\mathcal{H}_Y$. Subsequently, we can define the covariance

operator between X and Y by

$$\Gamma_{YX} = E[\{Y - E(Y)\} \otimes \{X - E(X)\}],$$

which maps from $\mathcal{H}$ to $\mathcal{H}_Y$. Γ_{XY} can be similarly defined. For convenience, throughout this thesis, we assume that $E(X) = 0$ and $E(Y) = 0$ without loss of generality. Hence, the regression model is

$$Y = \langle X, \beta \rangle_{\mathcal{H}} + \epsilon,$$

where $\beta \in \mathcal{H}$ is the unknown coefficient function, and ϵ is an error term which is a mean zero random variable and independent of X. Consider Y as a scalar random variable. We can rewrite $\beta(\cdot) = (\beta^1(\cdot), \ldots, \beta^p(\cdot))$ and

$$\langle X, \beta \rangle_{\mathcal{H}} = \textstyle\sum_{j=1}^{p} \langle X^j, \beta^j \rangle_{\mathcal{H}^j}.$$

2.2 Sample level

We observe n random copies from the model denoted by $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we observe X_i^j on $\{t_{i1}^j, \ldots, t_{ia_i}^j\}$ for each $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We assume that $\mathcal{H}^j$ is spanned by a finite given set of basis functions, $\mathcal{B} = \{b_1, \ldots, b_m\}$. For example: Fourier, B-spline, or other basis. The coordinate representation of each elements with respect to this basis is as follows. For any $f \in \mathcal{H}^j$, there exist a unique vector $a \in \mathbb{R}^m$ such that $f(\cdot) = \sum_{k=1}^{m} a_k b_k^j(\cdot)$. We call the vector a, the coordinate of f and denote it $[f]_{\mathcal{B}^j}$. We assume that $\mathcal{H}^j$ has L_2-inner product with respect to the Lebesgue measure,

$$\langle f, g \rangle_{\mathcal{H}^j} = \int_{T_j} f(t)g(t)dt, \quad \text{for any } f, g \in \mathcal{H}^j.$$

Let G^j be a $m \times m$ matrix whose (i,k)-th entry is $\langle b_i, b_k \rangle_{\mathcal{H}^j} = \int_{T_j} b_i(t) b_k(t) dt$. This is called the graham matrix. Let G be a $mp \times mp$ block-diagonal matrix with blocks G^j. The inner product with respect to this set of finite basis is:

$$\langle f, g \rangle_{\mathcal{H}} = [f]_{\mathcal{B}}^{\mathsf{T}} G [g]_{\mathcal{B}}, \quad \text{for any } f, g \in \mathcal{H}.$$

2.3 Functional OLS and functional ridge

We consider $[X_{1:n}]_{\mathcal{B}}$ to be a n by $m*p$ matrix constructed by stacking the coefficients of each functional covariates. Its first row from column 1 to m is $[X_1^1]$. We take $[\beta]$ as mp vector, the first m elements of which are $[\beta^1]$. We denote $Q = I - n^{-1} 1_n 1_n^{\mathsf{T}}$ and $[\tilde{X}_{1:n}]_{\mathcal{B}} = [X_{1:n}]_{\mathcal{B}} Q$—the centered design matrix. The sample–level matrix form of the linear model with such a notation is:

$$Y = [\tilde{X}_{1:n}]_{\mathcal{B}} G [\beta] + \epsilon$$

The least square estimation $[\hat{\beta}] = \arg\min_{\beta} \frac{1}{2} E_n (Y - \langle X, \beta \rangle_{\mathcal{H}})^2$ is:

$$\arg\min_{\beta} \frac{1}{2n} \| Y - [\tilde{X}_{1:n}]_{\mathcal{B}} G [\beta] \|_2^2 = ([\tilde{X}_{1:n}]_{\mathcal{B}}^T [\tilde{X}_{1:n}]_{\mathcal{B}} G)^{-1} ([\tilde{X}_{1:n}]_{\mathcal{B}}^T Y).$$

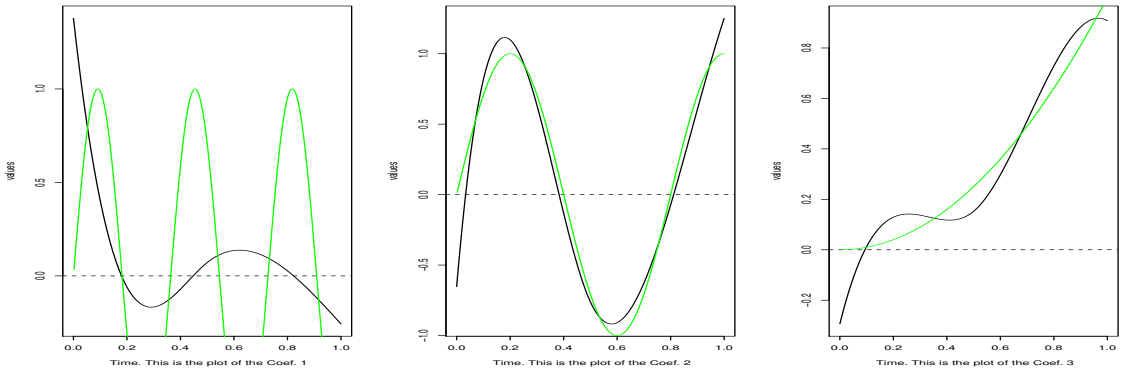
Similarly, the ridge estimation $[\hat{\beta}] = \arg\min_{\beta} \frac{1}{2} \| Y - [X_{1:n}]_{\mathcal{B}} G [\beta] \|_2^2 + \frac{\lambda}{2} \langle \beta, \beta \rangle_{\mathcal{H}}$ is:

$$([\tilde{X}_{1:n}]_{\mathcal{B}}^T [\tilde{X}_{1:n}]_{\mathcal{B}} G + \lambda I)^{-1} ([\tilde{X}_{1:n}]_{\mathcal{B}}^T Y).$$

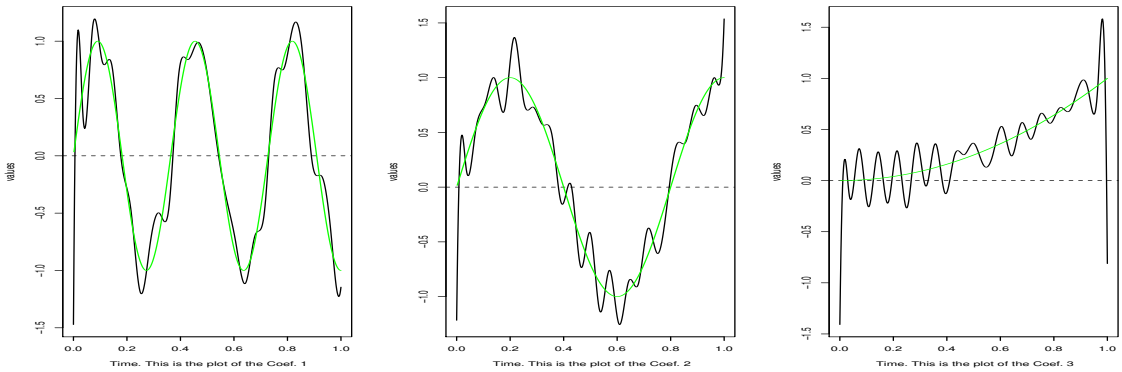
2.4 Penalizing the curvature

We suppose that a simulation is run with the following settings. The time period is $T = [0, 1]$, and the unknown coefficients are $\beta^1(t) = \sin(\frac{11\pi t}{2})$, $\beta^2(t) = \sin(\frac{5\pi t}{2})$, $\beta^3(t) = t^2$. The number of covariates are $p = 3$ and they are Brownian motions

$X^j(t_i^*) = \sum_{k=1}^{i} N_k^j$, where $1 \leq i \leq 500$, $N_k^j \sim N(0,1)$. We take $n = 200$ of such a simulation model. We use %80 of the data as the train set and the %20 rest as the test set. The population model is $Y =< X, \beta > +2.2 + \epsilon$ where $\epsilon \sim N(0, 0.05)$. Figure 2.1 shows the difference of OLS estimations for the curve in two scenarios- $m = 5$ B-spline basis and $m = 31$ basis. When the number of basis is not adequate, the estimations of the true coefficients- the green curves- are well only for true curves that are not too complicated and wiggly: In the first three figures, we note that the second and the third curves that are fairly simple are estimated well enough. However, for the coefficients that are not so simple such as β^1, this number of basis is not enough, as seen in the left panel of the first three figures. On the other hand, if there are too many basis used for functional conversion—$m = 31$ basis—the first coefficient is estimated well. However, the estimation of the second and the third is too wiggly—the second and the third figures from left in the second rows of figures. As a result, the RMSE for a higher number of basis is lower, while the curve estimations for simple curves are poor. Therefore, we need a new method to overcome curvature over-fitting while using too many basis.



(a) $m = 5$ B–spline OLS. Root mean squared error (RMSE) = 0.81



(b) $m = 31$ B–spline. RMSE = 0.184

Figure 2.1: OLS with 5, and 31 basis for comparison.

2.4.1 Curvature penalty

We can achieve smoother estimation even with a high number of basis by penalizing the second derivative of the final estimated curve. The sample level least square objective function with the curvature penalty is as following:

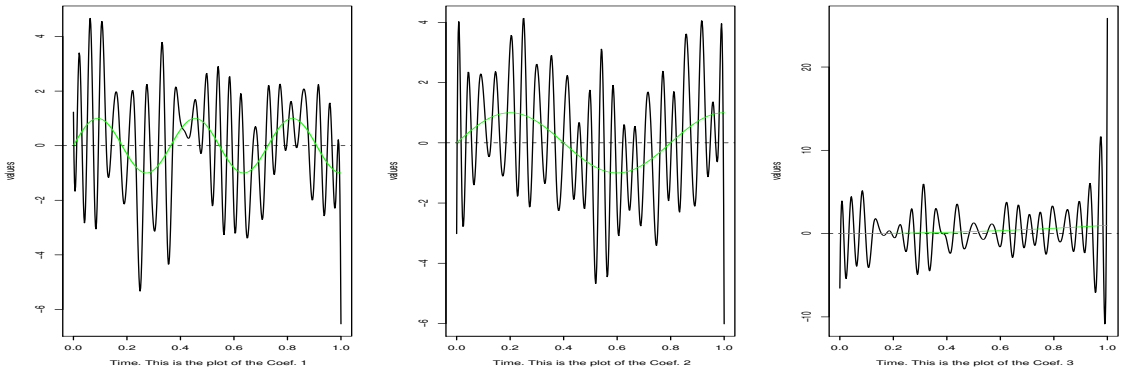
$$\arg\min_{\beta} \frac{1}{2n} \|Y - [X_{1:n}]_{\mathcal{B}} G[\beta]\|_2^2 + \frac{\lambda_{\text{der}}}{2} \|\beta''\|_{\mathcal{H}}^2,$$

where $\|\beta''\|_{\mathcal{H}^j}^2 = [\beta]^T G''[\beta]$, and G'' is a block diagonal matrix with blocks G''^j that are $m \times m$ matrix. The (i,k)-th entry of G''^j is $\langle b_i'', b_k'' \rangle_{\mathcal{H}^j} = \int_{T_j} b_i''(t) b_k''(t) dt$ which is the inner product of the second derivative of the basis elements. The closed-form

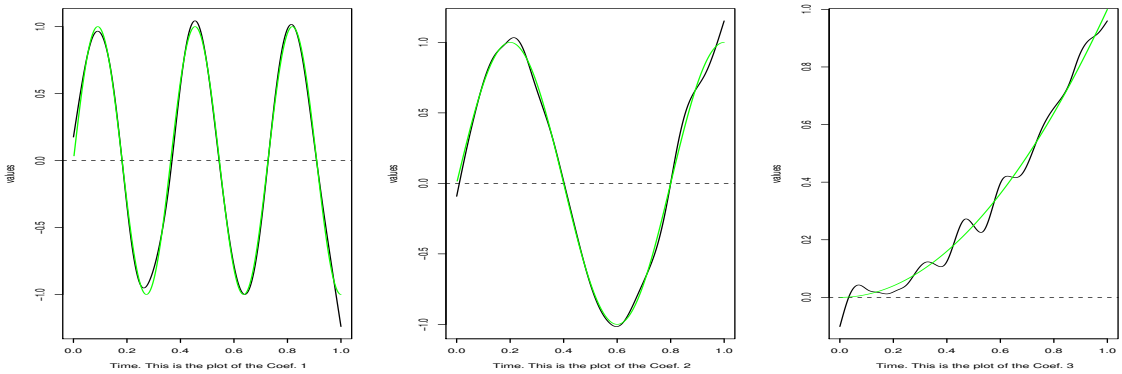
solution to this objective function is:

$$\hat{\beta}(\lambda_{\text{der}}) = (G[\tilde{X}_{1:n}]_{\mathcal{B}}^{T}[\tilde{X}_{1:n}]_{\mathcal{B}}G + \lambda_{\text{der}}G'')^{-1}(G[\tilde{X}_{1:n}]_{\mathcal{B}}^{T}Y).$$

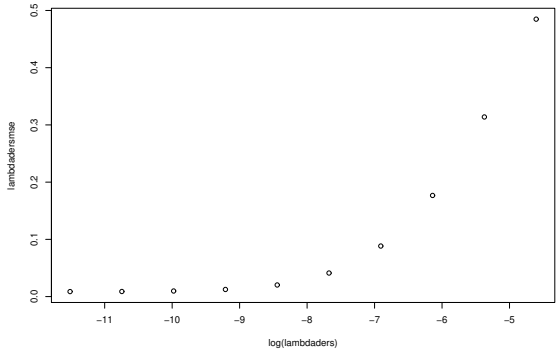
We regularize λ_{der} with a k–fold cross-validation on the train set. Figure 2.2 shows the difference of OLS when penalizing the second curvature with large number of basis, $m = 51$ B–splines. Although there are too many basis the curve estimations for the second and the third coefficients are smooth when OLS is equipped with a curvature penalty. In addition, the RMSE result of the OLS regression with the curvature penalty is the best among all four situations in figures 2.1 and 2.2.



(a) $m = 51$ B–spline OLS. RMSE = 0.48



(b) $m = 51$ with penalizing the second derivative of estimations. RMSE= 0.168



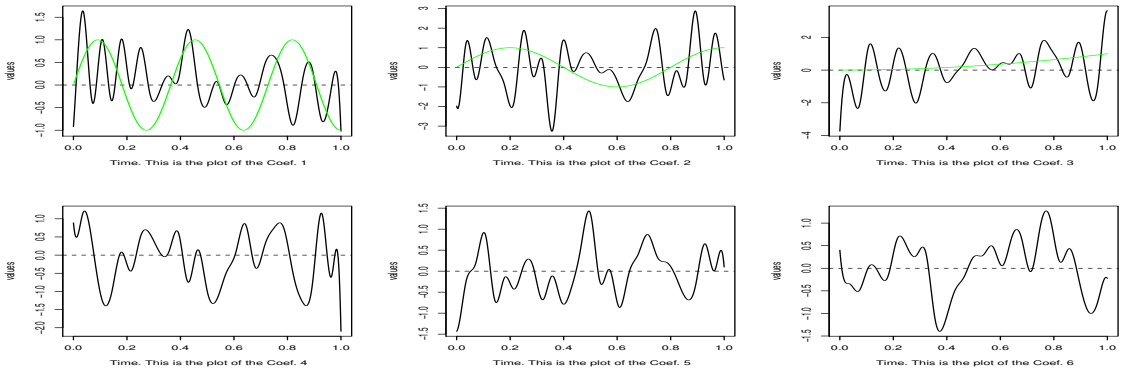
(c) Cross-validation and λ_{der} tuning

Figure 2.2: OLS estimation without and with the second derivative penalty for comparison. The figure at the bottom is the cross-validation results for λ_{der}.

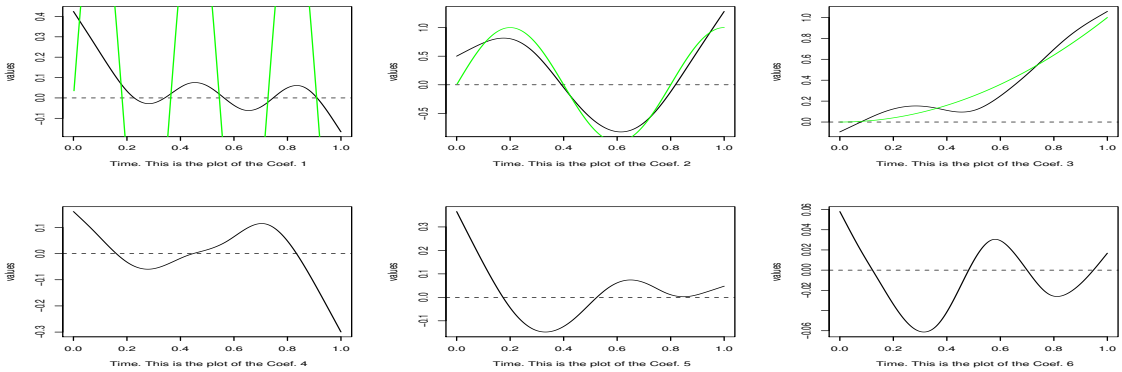
2.5 Penalizing the curvature while the population is sparse

Before proposing the sparse functional regression models, we consider a sparse simulation model to test the effect of the curvature penalty when applied to the OLS regression model. We consider $p = 35$ number of covariates. The true coefficients are $\beta^1(t) = \sin(\frac{11\pi t}{2})$, $\beta^2(t) = \sin(\frac{5\pi t}{2})$, $\beta^3(t) = t^2$, and $\beta^4(t) = \ldots = \beta^{35}(t) = 0$. The population is sparse and the active set is $\{1, 2, 3\}$. The population model is $Y =< X^1, \beta^1 > + < X^2, \beta^2 > + < X^3, \beta^3 > +2.2 + \epsilon$ where $\epsilon \sim N(0, 0.05)$.

Figure 2.3 shows that when penalizing the curvature, OLS works better in terms of RMSE and curve estimations. Except for β^1, the rest of the coefficients are estimated well with the curvature penalty in the second panel. Conversely, non-regularized functional OLS in the first panel performs poorly. Hence, we will keep the curvature penalty when we introduce the sparse functional regression models.



(a) $m = 31$ B–spline OLS. RMSE $= 7.149$



(b) $m = 31$ B–spline OLS while Penalizing the curvature. RMSE= 2.041

Figure 2.3: Comparison of OLS with and without curvature penalty when the population is sparse.

CHAPTER 3: SPARSE FUNCTIONAL MODEL: MODEL DESCRIPTION

We are interested in the situations where the predictors are multivariate functions, but only a few functional predictors affect the response. i.e., a random variable Y and random functions $X^j \in \mathcal{H}^j$ have the following relation,

$$Y = \sum_{j \in J} \langle X^j, \beta^j \rangle_{\mathcal{H}^j} + \epsilon, \tag{3.1}$$

where $J \subseteq \{1, \ldots, p\}$ is an unknown active set of indices involved in this regression model, and ϵ is a mean zero error term that is independent of X.

Assume that we have a random sample of size n from the model (3.1). To estimate β and the active set J, we consider the following objective function.

$$L(\beta; \lambda_{1n}) = \frac{1}{2} E_n (Y - \langle X, \beta \rangle_{\mathcal{H}})^2 + \lambda_{1n} \textstyle\sum_{j=1}^{p} \|\beta^j\|_{\mathcal{H}^j}, \quad \beta \in \mathcal{H}, \tag{3.2}$$

where E_n is the expectation with the empirical distribution. We added the group-LASSO type penalty so that each group includes one functional component in the infinite–dimensional Hilbert space, $\mathcal{H}^j$, $j = 1, \ldots, p$. Note that the norm in the penalty term is L_2-norm which makes the objective function convex. In addition, we propose an alternative objective function to gain a more stable solution path.

$$L(\beta; \lambda_{1n}, \lambda_{2n}) = \frac{1}{2} E_n (Y - \langle X, \beta \rangle_{\mathcal{H}})^2 + \lambda_{1n} \textstyle\sum_{j=1}^{p} \|\beta^j\|_{\mathcal{H}^j} + \lambda_{2n} \sum_{j=1}^{p} \|\beta^j\|_{\mathcal{H}^j}^2, \quad \beta \in \mathcal{H}, \tag{3.3}$$

The quadratic term allows us to have a stable solution path and encourages further grouping effects. It is similar to the Elastic Net proposed by (9), but it is different in that the norm in the first penalty term uses L_2-norm, and both the two penalties are applied group-wisely. The group-wise second penalty also gives us a substantial computational advantage.

Furthermore, we also consider the smoothing penalty of the functional coefficients $\beta \in \mathcal{H}$ by adding the term, $\lambda_{3n}\|\beta''\|^2_{\mathcal{H}}$ to the objective functions, (3.2) and (3.3). It allows us to estimate smooth functional coefficients and to select the functional predictors simultaneously. In addition, it provides a better interpretation of the functional coefficients in this linear functional regression model.

CHAPTER 4: ESTIMATION: ADMM

In this section, we develop the algorithm for solving the optimization problems introduced in the section 3 via the alternating direction method of multipliers (ADMM), popularly used in a general convex optimization problem. See (46). Consider the following optimization problem.

$$\begin{aligned} \arg\min_{\beta,\gamma} \quad & f(\beta) + g(\gamma) \\ \text{s.t.} \quad & \beta - \gamma = 0, \end{aligned} \tag{4.1}$$

where γ is duplicate variable in $\mathcal{H}$, $f(\beta) = \frac{1}{2}E_n(Y - \langle X, \beta\rangle_{\mathcal{H}})^2$, and $g(\gamma) = \lambda \sum_{j=1}^{p} \|\gamma^j\|_{\mathcal{H}^j}$. Blocks γ^j are associated with their counterparts' blocks β^j. The augmented Lagrangian with its parameter $\rho > 0$ is

$$L_\rho(\beta, \gamma, \eta) = f(\beta) + g(\gamma) + \langle \eta, \beta - \gamma\rangle_{\mathcal{H}} + \frac{\rho}{2}\|\beta - \gamma\|_{\mathcal{H}}^2, \tag{4.2}$$

where the Lagrangian multiplier is $\eta \in \mathcal{H}$. The ADMM update rules are

$$\begin{aligned} \beta^{\text{new}} &:= \arg\min_{\beta} L_\rho(\beta, \gamma, \eta) \\ \gamma^{\text{new}} &:= \arg\min_{\gamma} L_\rho(\beta^{\text{new}}, \gamma, \eta) \\ \eta^{\text{new}} &:= \eta + \rho(\beta^{\text{new}} - \gamma^{\text{new}}). \end{aligned} \tag{4.3}$$

For computational convenience, it is a usual practice to consider the scaled dual parameter of the ADMM. Let $u = \frac{1}{\rho}\eta$. It is straightforward to verify that the update

rules (4.3) with a scaled dual parameter are equivalent to

$$\begin{aligned}\beta^{\text{new}} &:= \arg\min_{\beta}\left(f(\beta)+\frac{\rho}{2}\|\beta-\gamma+U\|_{\mathcal{H}}^2\right)\\ \gamma^{\text{new}} &:= \arg\min_{\gamma}\left(g(\gamma)+\frac{\rho}{2}\|\beta^{\text{new}}-\gamma+U\|_{\mathcal{H}}^2\right)\\ U^{\text{new}} &:= U+\beta^{\text{new}}-\gamma^{\text{new}}.\end{aligned} \tag{4.4}$$

4.0.1 Coordinate representation of functional data

Our method is based on the basis-expansion approach to the functional data. Suppose that we have n random copies from the model (3.1) denoted by $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we observe X_i^j on $\{t_{i1}^j, \ldots, t_{ia_i^j}^j\}$ for each $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

At the sample level, we assume that $\mathcal{H}^j$ is spanned by a given set of basis functions, $\mathcal{B}^j = \{b_1^j, \ldots, b_{m_j}^j\}$. Thus, for any $f \in \mathcal{H}^j$, there exists a unique vector $a \in \mathbb{R}^{m_j}$ such that $f(\cdot) = \sum_{k=1}^{m_j} a_k b_k^j(\cdot)$. We call the vector a, the coordinate of f and denote it $[f]_{\mathcal{B}^j}$. We also assume that $\mathcal{H}^j$ is constructed with the L_2-inner product with respect to the Lebesgue measure,

$$\langle f, g\rangle_{\mathcal{H}^j} = \int_{T_j} f(t)g(t)dt, \quad \text{for any } f, g \in \mathcal{H}^j.$$

Let G^j be $m_j \times m_j$ matrix whose (i,k)-th entry is $\langle b_i^j, b_k^j\rangle_{\mathcal{H}^j} = \int_{T_j} b_i^j(t)b_k^j(t)dt$, and let G be $M \times M$ block-diagonal matrix whose j-th block is G^j where $M = \sum_{j=1}^p m_j$. Consequently, for any $f, g \in \mathcal{H}$,

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{j=1}^p \sum_{i=1}^{m_j}\sum_{k=1}^{m_j}([f^j]_{\mathcal{B}^j})_i([g^j]_{\mathcal{B}^j})_k\langle b_i^j, b_k^j\rangle_{\mathcal{H}^j} = \sum_{j=1}^p [f^j]_{\mathcal{B}^j}^{\mathsf{T}} G^j [g^j]_{\mathcal{B}^j} = [f]_{\mathcal{B}}^{\mathsf{T}} G [g]_{\mathcal{B}},$$

where $[f]_{\mathcal{B}}, [g]_{\mathcal{B}}$ are the $\mathbb{R}^M$-dimensional vectors obtained by stacking $[f^j]_{\mathcal{B}^j}$ and $[g^j]_{\mathcal{B}^j}$ respectively. We use the basis-expansion approach for each functional covariate X_i^j for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, which is also used in (47; 48). Without loss of

generality, we assume $m = m_1 = \cdots = m_p$ and $M = pm$.

Suppose that A is a linear operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ in which the basis for $\mathcal{H}_1$ is $\mathcal{B} = \{b_1, \ldots, b_m\}$ and the basis for $\mathcal{H}_2$ is $\mathcal{C} = \{c_1, \ldots, c_k\}$. Then, we define the coordinate representation of the operator A to be $k \times m$ matrix, say ${}_{\mathcal{C}}[A]_{\mathcal{B}}$, whose (i, j)-th entry is $([Ab_j]_{\mathcal{C}})_i$. It can be easily shown that ${}_{\mathcal{C}}[Ax]_{\mathcal{B}} = {}_{\mathcal{C}}[A]_{\mathcal{B}}[x]_{\mathcal{B}}$ for any $x \in \mathcal{H}_1$. For notational convenience, if the basis system is obvious in the context, we remove the subscripts of the coordinate representation throughout this thesis. The following lemma provides a further simplification for easy computations.

Lemma 4. *Let $Q = I - n^{-1}1_n1_n^{\mathsf{T}}$. Let $[X_{1:n}]_{\mathcal{B}}$ be the $pm \times n$ matrix, the k-th column of which is $[X_k]_{\mathcal{B}}$. Then*

$$ {}_{\mathcal{B}}[\hat{\Gamma}_{XX}]_{\mathcal{B}} = n^{-1}[X_{1:n}]_{\mathcal{B}}Q[X_{1:n}]_{\mathcal{B}}^{\mathsf{T}}G = n^{-1}[\tilde{X}_{1:n}]_{\mathcal{B}}[\tilde{X}_{1:n}]_{\mathcal{B}}^{\mathsf{T}}G, $$

where $[\tilde{X}_{1:n}]_{\mathcal{B}} = [X_{1:n}]_{\mathcal{B}}Q$. In addition, let Y be the n-dimensional vector, the elements of which are the observations $Y_1, \ldots, Y_n$. Then

$$ [\hat{\Gamma}_{YX}] = n^{-1}Y^{\mathsf{T}}[\tilde{X}_{1:n}]_{\mathcal{B}}^{\mathsf{T}}G. $$

4.0.2 Orthogonalization

To achieve computational efficiency, we orthonormalize the basis system via Karhunen-Loève expansion of the covariance operator of each of the functional predictors. For each $j = 1, \ldots, p$, define Γ_{jj} to be the covariance operator of X^j. Consequently, we have the following lemma.

Lemma 5. *Let $(\lambda_1^j, v_1^j), \ldots, (\lambda_m^j, v_m^j)$ be the pairs of eigenvalues and vectors of $(G^j)^{1/2}[\tilde{X}_{1:n}^j]_{\mathcal{B}^j}[\tilde{X}_{1:n}^j]_{\mathcal{B}^j}^{\mathsf{T}}(G^j)^{1/2}$ with $\lambda_1^j \geq \ldots \geq \lambda_m^j$, and let $[\phi_k^j]_{\mathcal{B}^j} = (G^j)^{-1/2}v_k^j$ for $k =$*

$1, \ldots, m$. *Then, the Karhunen-Loève expansion of* $\hat{\Gamma}_{jj}$ *is*

$$\hat{\Gamma}_{jj} = \sum_{k=1}^{m} \lambda_k^j \phi_k^j \otimes \phi_k^j.$$

Define a $m \times m$ matrix

$$\Phi^j = \left([\phi_1^j]_{\mathcal{B}^j} \quad \cdots \quad [\phi_m^j]_{\mathcal{B}^j} \right).$$

Since ϕ_m^j's are the eigenfunctions of a self-adjoint operator, they are orthonormal. Thus, for any $x \in \mathcal{H}^j$,

$$\begin{aligned} x(\cdot) &= \sum_{k=1}^{m} \langle x, \phi_k^j \rangle_{\mathcal{H}^j} \phi_k^j(\cdot) \\ &= \sum_{k=1}^{m} [x]_{\mathcal{B}^j}^{\mathsf{T}} G^j [\phi_k^j]_{\mathcal{B}^j} \phi_k^j(\cdot). \end{aligned}$$

Define $\mathcal{C}^j = \{\phi_1^j, \ldots, \phi_m^j\}$ to be the new basis system for $\mathcal{H}^j$. Then, we have

$$[X_i^j]_{\mathcal{C}^j} = (\Phi^j)^{\mathsf{T}} G^j [X_i^j]_{\mathcal{B}^j}, \quad i = 1, \ldots, n, j = 1, \ldots, p.$$

We assume that the coordinate of $\mathcal{H}$ is based on the orthonormal basis system throughout this section. Thus,

$$[\hat{\Gamma}_{XX}] = \mathrm{diag}(\lambda_1^1, \ldots, \lambda_m^1, \ldots, \lambda_1^p, \ldots, \lambda_m^p), \quad [X_{1:n}] = \mathrm{diag}(\Phi_1^{\mathsf{T}}, \ldots, \Phi_p^{\mathsf{T}}) G [X_{1:n}]_{\mathcal{B}},$$

and $\langle f, g \rangle_{\mathcal{H}} = [f]^{\mathsf{T}}[g]$ for any $f, g \in \mathcal{H}$.

4.0.3 Estimation

Using the representation, we can express the optimization (4.4) as follows.

$$
\begin{aligned}
[\beta^{\text{new}}] &:= \arg\min_{\beta\in\mathcal{H}} \Big(f(\beta) + \frac{\rho}{2}([\beta] - [\gamma] + [U])^{\mathsf{T}}([\beta] - [\gamma] + [U]) \Big) \\
[\gamma^{\text{new}}] &:= \arg\min_{\gamma\in\mathcal{H}} \Big(g(\gamma) + \frac{\rho}{2}([\beta^{\text{new}}] - [\gamma] + [U])^{\mathsf{T}}([\beta^{\text{new}}] - [\gamma] + [U]) \Big) \\
[U^{\text{new}}] &:= [U] + [\beta^{\text{new}}] - [\gamma^{\text{new}}],
\end{aligned}
\tag{4.5}
$$

where

$$
\begin{aligned}
f(\beta) &= \frac{1}{2}E_n(Y - \langle X, \beta\rangle_{\mathcal{H}})^2 = (2n)^{-1}\textstyle\sum_{i=1}^{n}\{Y_i^2 - 2(Y_i \otimes X_i)\beta + \langle\beta, (X_i \otimes X_i)\beta\rangle_{\mathcal{H}}\} \\
&= \frac{1}{2}\hat{\sigma}_{YY} - \hat{\Gamma}_{YX}\beta + \frac{1}{2}\langle\beta, \hat{\Gamma}_{XX}\beta\rangle_{\mathcal{H}} = \frac{1}{2}\hat{\sigma}_{YY} - [\hat{\Gamma}_{YX}][\beta] + \frac{1}{2}[\beta]^{\mathsf{T}}[\hat{\Gamma}_{XX}][\beta],
\end{aligned}
$$

and $g(\gamma) = \lambda\sum_{j=1}^{p}\|\gamma^j\|_{\mathcal{H}^j} = \lambda\sum_{j=1}^{p}\sqrt{[\gamma^j]^{\mathsf{T}}[\gamma^j]}$.

Under the finite-dimensional representation of the functional elements in $\mathcal{H}$, one can see that the optimization in (4.5) is a convex optimization problem.

Theorem 3. *The solution to the optimization problem (3.2) can be achieved by iterating over the following update rules.*

$$
\begin{aligned}
[\beta^{new}] &= ([\tilde{X}_{1:n}][\tilde{X}_{1:n}]^{\mathsf{T}} + n\rho I_M)^{-1}([\tilde{X}_{1:n}]Y + n\rho([\gamma] - [U])) \\
[(\gamma^j)^{new}] &= S^{\mathcal{H}^j}_{\frac{\lambda}{\rho}}([(\beta^j)^{new}] + [U^j]) \qquad\qquad j = 1\dots p \\
[U^{new}] &= [U] + [\beta^{new}] - [\gamma^{new}],
\end{aligned}
\tag{4.6}
$$

where $[\gamma^j]$, $[U^j]$ are corresponding blocks to $[\beta^j]$, and $S^{\mathcal{H}^j}_{\lambda}(h) = 1_{\{\|h\|_{\mathcal{H}^j} > \lambda\}}\left(1 - \frac{\lambda}{\|h\|_{\mathcal{H}^j}}\right)_+ h$ for $h \in \mathcal{H}^j$.

If we do not consider orthogonalization, Theorem 3 would contain element G^j in the updates. In this case, the proof of numerical convergence of the update rules is slightly different from that of (46). However, due to the orthogonalization, the

proof of the numerical convergence of the updates in Theorem 3 to the solution of the optimization problem (3.2) is identical to that of the ADMM in (46). Hence, it is omitted.

4.0.4 Different penalty terms

In this section, we investigate the different penalty terms in two directions: one for the functional predictor selection and the other for the smooth coefficient functions β.

4.0.4.1 Multivariate functional group elastic net

LASSO does not provide a unique solution. To achieve uniqueness and overcome the saturation property, the Elastic Net penalty has been introduced by combining the ℓ_1-norm and ℓ_2-norm by (9) for the multivariate data. Functional data are intrinsically infinite-dimensional objects. Thus, we propose a multivariate functional-version optimization problem for the Elastic Net penalty by grouping each functional predictor as follows.

$$\frac{1}{2}E_n(Y - \langle X, \beta\rangle_{\mathcal{H}})^2 + \lambda(1-\alpha)\textstyle\sum_{j=1}^{p}\|\beta^j\|_{\mathcal{H}^j} + \alpha\lambda\sum_{j=1}^{p}\|\beta^j\|_{\mathcal{H}^j}^2, \tag{4.7}$$

where $\alpha \in [0, 1]$ and $\lambda > 0$ are the tuning parameters.

This optimization problem still follows the structure of the ADMM algorithm in (4.1) with $g(\gamma) = \lambda(1-\alpha)\sum_{j=1}^{p}\|\gamma^j\|_{\mathcal{H}^j} + \alpha\lambda\sum_{j=1}^{p}\|\gamma^j\|_{\mathcal{H}^j}^2$. It can be easily shown that the only difference from the original version is the γ-update in Theorem 3. Hence, we have the following update rules.

Theorem 4. *The solution to the optimization problem (4.7) can be achieved by iter-*

ating over the following update rules.

$$
\begin{aligned}
&[\beta^{new}] = ([\tilde{X}_{1:n}][\tilde{X}_{1:n}]^{\mathsf{T}} + n\rho I_M)^{-1}([\tilde{X}_{1:n}]Y + n\rho([\gamma] - [U])) \\
&[(\gamma^j)^{new}] = \frac{\rho}{\rho + 2\alpha\lambda} S^{\mathcal{H}^j}_{\frac{\lambda(1-\alpha)}{\rho}}([(\beta^j)^{new}] + [U^j]) \qquad j = 1 \dots p \qquad (4.8) \\
&[U^{new}] = [U] + [\beta^{new}] - [\gamma^{new}].
\end{aligned}
$$

Regularization parameters can be adjusted through a net search cross–validation.

4.0.4.2 The smoothness of functional coefficients β

According to the simulation, the previous algorithm provides a wiggly estimation of functional coefficients β most of the time. It might be fine if we are only interested in the prediction; however, it is not the case because we consider the linear functional regression. We propose an algorithm that controls the roughness of β simultaneously to avoid the over-fitting problems and obtain smooth functional coefficients. In particular, we impose the penalty on the curvature of the coefficients by adding $\frac{\lambda_{\text{der}}}{2}\|\beta''\|^2_{\mathcal{H}}$ to the objective function (4.4). We include this term in $f(\cdot)$ function in the ADMM structure. Finally, the first update rule (4.6) in Theorem 3 becomes

$$
[\beta^{\text{new}}] := ([\tilde{X}_{1:n}][\tilde{X}_{1:n}]^{\mathsf{T}} + n\rho I_M + \lambda_{\text{der}} G'')^{-1}([\tilde{X}_{1:n}]Y + n\rho([\gamma] - [U])), \qquad (4.9)
$$

where G'' is a block-diagonal matrix whose j-th block matrix is
$((G^j)'')_{ik} = \int_{T^j} (\phi_i^j)''(t)(\phi_k^j)''(t)dt = \langle (\phi_i^j)'', (\phi_k^j)'' \rangle_{\mathcal{H}^j}$ for $i, k = 1, \dots, m$, $j = 1 \dots, p$.

For each j, $(G^j)''$ can be derived from the second derivative Gram matrix for the original basis, say $(B^j)''$, where $((B^j)'')_{ik} = \int_{T^j} (b_i^j)''(t)(b_k^j)''(t)dt = \langle (b_i^j)'', (b_k^j)'' \rangle_{\mathcal{H}^j}$. Note that

$$
[\phi_i^j]_{\mathcal{B}^j} = (G^j)^{-1}((\Phi^j)^{-1})^{\mathsf{T}}[\phi_i^j]_{\mathcal{C}^j} = (G^j)^{-1}((\Phi^j)^{-1})^{\mathsf{T}} e_i,
$$

where e_i is i-th standard basis in $\mathbb{R}^m$. Then,

$$\begin{aligned}\langle(\phi_i^j)'', (\phi_k^j)''\rangle_{\mathcal{H}^j} &= \langle\textstyle\sum_{\ell=1}^m([\phi_i^j]_{\mathcal{B}^j})_\ell(b_\ell^j)'', \sum_{\ell=1}^m([\phi_k^j]_{\mathcal{B}^j})_\ell(b_\ell^j)''\rangle_{\mathcal{H}^j} \\ &= e_i^\mathsf{T}(\Phi^j)^{-1}(G^j)^{-1}(B^j)''(G^j)^{-1}(\Phi^j)^{-1})^\mathsf{T}e_j.\end{aligned}$$

4.0.4.3 Tuning

The initial values for γ and U are zero, and the initial β is the ridge regression estimation in the first update rule (4.6). Then, we set the augmented parameter or the step size, ρ, to be 1 and stay the same through the algorithm. The different values of ρ only change the optimal λ values on the grid or optimal $(1-\alpha)\lambda$ on the net. The larger the ρ, the smaller the optimized regularization parameter of the soft threshold operator. In some practices of augmented Lagrangian, it is possible to choose a small step size and gradually increase it to 1 in each iteration. It is also stated in (46) why $\rho = 1$ is a suitable choice in the ADMM algorithm.

I use the k-fold cross–validation for choosing the mixing parameter α, the regularization parameter of the second derivative penalty λ_{der}, and the main regularization parameter λ. In particular, for each α and each λ_{der} on the net, we search for the optimal λ. To pick the initial λ, we first find the ridge estimation β with parameter $\rho = 1$. We then compute the norm of each group of functional coefficients, $\|\beta^k\|$. Note that in the second update of Theorem 3, the soft threshold operator would eliminate all blocks if λ is slightly higher than the maximum of these norms. On the other hand, this update would keep all the coefficients if λ is slightly lower than the smallest norm. Therefore, a reasonable procedure is to design a grid of λs between a number slightly lower than the minimum norm of the blocks and a number slightly higher than the maximum norm of these block coefficients.

CHAPTER 5: Estimation: GMD

This section derives the groupwise-majorization-descent (GMD) algorithm for solving the objective functions in chapter 3. Unlike the ADMM, this algorithm is geared toward the objective function with group-wise penalty terms. Motivated by (49), we derive the GMD algorithm under our setting. In addition, we do not force the basis functions to be orthogonal, which allows us to have more flexibility. Thus, we use the coordinate system based on the original basis $\mathcal{B}$ without orthogonalization throughout this section.

5.1 Algorithm

The MFG-Elastic Net problem without the orthogonalization is

$$\arg\min_{\beta} \frac{1}{2}\|Y - [\tilde{X}_{1:n}]^{\mathsf{T}} G[\beta]\|_2^2 + \frac{\lambda_{\text{der}}}{2}[\beta'']^T G[\beta''] + \lambda(1-\alpha)\textstyle\sum_{j=1}^{p}\|\beta^j\|_{\mathcal{H}^j} + \alpha\lambda\sum_{j=1}^{p}\|\beta^j\|_{\mathcal{H}^j}^2, \tag{5.1}$$

where the coordinates are associated with the original basis $\mathcal{B}$. This optimization problem and the following derived algorithm include the steps that also solve for the MFG-LASSO ($\alpha = 0$) and the ridge regression ($\alpha = 1$). In equation (5.1), we remove n for computational convenience. It will be adjusted when we seek the λ_{der} and λ in the grid construction. We define the loss function as follows.

$$L(\beta) = \frac{1}{2}\|Y - [\tilde{X}_{1:n}]^{\mathsf{T}} G[\beta]\|_2^2 + \frac{\lambda_{\text{der}}}{2}[\beta'']^T G[\beta'']. \tag{5.2}$$

Consequently, the objective function (5.1) is $L(\beta) + g(\beta)$ where
$g(\beta) = \lambda(1-\alpha)\sum_{j=1}^{p}\sqrt{[\beta^j]^{\mathsf{T}} G^j [\beta^j]} + \alpha\lambda\sum_{j=1}^{p}[\beta^j]^{\mathsf{T}} G^j [\beta^j]$.

Lemma 6. *The loss function (5.2) satisfies the quadratic majorization (QM) condition with* $H = G[\tilde{X}_{1:n}]^{\mathsf{T}}[\tilde{X}_{1:n}]G + \lambda_{der}B''$. *In other words, for any* $\beta, \beta^* \in \mathcal{H}$,

$$L(\beta) \leq L(\beta^*) + ([\beta] - [\beta^*])\nabla L(\beta^*) + \frac{1}{2}([\beta] - [\beta^*])^{\mathsf{T}}H([\beta] - [\beta^*]), \tag{5.3}$$

where,

$$\nabla L(\beta^*|D) = G[\tilde{X}_{1:n}]([\tilde{X}_{1:n}]^{\mathsf{T}}G[\beta] - Y) + \lambda_{der}B''[\beta^*], \tag{5.4}$$

where $|D$ *refers to condition given data, or given the design matrix.*

Let $U = -\nabla L(\beta^*)$. In addition to Lemma 6, it is straightforward to see that if $\beta \neq \beta^*$, we have the strict inequality,

$$L(\beta|D) < L(\beta^*|D) - ([\beta] - [\beta^*])^T U(\beta^*) + \frac{1}{2}([\beta] - [\beta^*])^{\mathsf{T}}H([\beta] - [\beta^*]). \tag{5.5}$$

Thus, it leads to the strict descent property of the updating algorithm. Let β^* be the current solution to the optimization problem and β be the next update. Assume that we update the β for each functional predictor $j = 1, \ldots, p$. In other words, $[\beta] - [\beta^*]$ has a form of $(0, \ldots, 0, [\beta^j] - [(\beta^*)^j], 0, \ldots, 0)^{\mathsf{T}}$, which leads to simplification of the objective function in the new optimization problem. Let U^j be the sub-vector of U with the indices $(m(j-1)+1, \ldots, mj)$. Let H^j be the j-th block diagonal matrix of H. Then, (5.3) is

$$\begin{aligned} L(\beta) &\leq L(\beta^*) - ([\beta^j] - [(\beta^*)^j])U^j + \frac{1}{2}([\beta^j] - [(\beta^*)^j])^{\mathsf{T}}H^j([\beta^j] - [(\beta^*)^j]) \\ &\leq L(\beta^*) - ([\beta^j] - [(\beta^*)^j])U^j + \frac{1}{2}\gamma_j([\beta^j] - [(\beta^*)^j])^{\mathsf{T}}([\beta^j] - [(\beta^*)^j]), \end{aligned}$$

where γ_j is a value slightly larger than the largest eigenvalue of H^j, which further relaxes the upper bound. In practice, we take $\gamma_j = (1 + \epsilon^*)\eta_j$ with $\epsilon^* = 10^{-6}$ where

η_j is the largest eigenvalue of H^j. Finally, the update rule for β^j is the solution to the following optimization problem.

$$\arg\min_{\beta^j \in \mathcal{H}^j} -([\beta^j] - [(\beta^*)^j])U^j + \frac{1}{2}\gamma_j([\beta^j] - [(\beta^*)^j])^{\mathsf{T}}([\beta^j] - [(\beta^*)^j]) + g^j(\beta), \quad (5.6)$$

where g^j is the j-th term of $g(\cdot)$. We have a closed-form solution to this problem using a similar trick of Lemma 15.

$$[\beta^j]^{(\text{new})} = \frac{1}{2\alpha\lambda + \gamma_j} S^{\mathcal{H}^j}_{\lambda(1-\alpha)}(U^j + \gamma_j[\beta^j]^{(\text{old})}), \quad j = 1, \ldots, p, \quad (5.7)$$

where $U^j = -\nabla L(\beta^{j\,(\text{old})})$ and $\nabla L(\beta) = G[\tilde{X}_{1:n}]([\tilde{X}_{1:n}]^{\mathsf{T}} G[\beta] - Y) + \lambda_{\text{der}} B''[\beta]$.

5.1.1 Tuning parameter selection

While iterating over this GMD update rule, we can reduce the computational burden more efficiently during the tuning parameter selection with the strong rule technique. See (50).

Step 1. (Initialization) Given $\alpha \in (0, 1)$, the largest λ in the grid points is the smallest value of λ such that all its associated coefficients are zero. In particular, using the KKT condition (see Lemma 15), the largest λ in the grid points is

$$\lambda^{(1)} = \max_j \frac{\|U^j(0)\|}{1-\alpha}.$$

Therefore, the initial β is zero. Then, the smallest λ of the grid points is set to be a certain small number to include all the functional predictors, usually a fraction of the largest λ value of the grid. The process of searching for the optimal λ starts with the largest value of the grid points and moves backward to the smallest value.

Step 2. (Iteration) At $\lambda^{(k)}$, we add j-th functional predictor to the active set if it

satisfies the strong rule condition,

$$\|U^j([\beta^j(\lambda^{(k)})])\| > (2\lambda^{(k+1)} - \lambda^{(k)})(1-\alpha),$$

for $j = 1, \ldots, p$. Subsequently, we update β with these reduced predictors by iterating the update rule (5.7) until numerical convergence. The stopping criteria for this iterative process can be chosen the absolute or relative. Next, to make sure that the strong rule does not leave out some of the worthy coefficients, we check the KKT condition on the rest of the blocks of the current solution,

$$\|U^j([\beta^j_{update}(\lambda^{(k+1)})])\| < \lambda^{(k+1)}(1-\alpha),$$

where $\beta^j_{update}(\lambda^{(k+1)})$ is the updated β^j when the iterative GMD algorithm hits the stopping criteria on the result of the strong rule screening. If j-th functional coefficient violates the KKT condition, we add it to the active set and update β using (5.7). This process of checking the KKT condition and updating continues until no functional coefficient violates the KKT condition. We store the solution of the final updated value to $\beta^j(\lambda^{(k+1)})$. We use $\beta^j(\lambda^{(k+1)})$ to repeat *(Step 2)* for the next value of λ (warm start).

5.2 Comparison of the two estimation methods

It is worth mentioning that the strong rule does not allow the main regularization for λ to be computed in parallel because of the warm start, i.e., We search for λ sequentially. However, the main computational cost is paid in this regularization. The strong rule allows the algorithm to enjoy predictor screening, which leads to a cost-effective computation by storing and computing on a smaller vector size. On the other hand, the strong rule does not seem valid for the ADMM algorithm because there are two objective functions involved in this algorithm. Hence, it is possible to

tune the regularization parameters in parallel via ADMM.

CHAPTER 6: ASYMPTOTIC RESULTS

In this section, we derive the consistency of the multivariate functional group LASSO (MFG-LASSO) when functions are fully observed. It is worth mentioning that the second derivative penalty term in the loss function has zero regularization parameter when the number of time points and the number of basis are infinity, i.e., fully observable. Therefore, the asymptotic properties of such a model are not considered with the curvature penalty for the asymptotic properties when we assume fully observable functional covariates. In particular, the consistency breaks down to the selection consistency and the estimation consistency, which is known as the oracle property.

We first illustrate the consistency of the operators used in the estimation procedure. Since the implementation in the section 4.0.1 is based on the method of moments estimate, the following lemma is an immediate result from the functional-version of the central limit Theorem in a separable Hilbert space. See (45).

Lemma 7. *If* $E\|X\|_{\mathcal{H}}^4 < \infty$ *and* $EY^4 < \infty$, *then*

1. $\sqrt{n}(\hat{\Gamma}_{XX} - \Gamma_{XX}) \xrightarrow{\mathscr{D}} N(0, \Sigma_{XX})$,
2. $\sqrt{n}(\hat{\Gamma}_{YX} - \Gamma_{YX}) \xrightarrow{\mathscr{D}} N(0, \Sigma_{YX})$,
3. $\sqrt{n}(\hat{\Gamma}_{YY} - \Gamma_{YY}) \xrightarrow{\mathscr{D}} N(0, \Sigma_{YY})$,

where $\Sigma_{XX} = E[\{(X - EX) \otimes (X - EX) - \Gamma_{XX}\} \otimes \{(X - EX) \otimes (X - EX) - \Gamma_{XX}\}]$ *and* Σ_{YX}, Σ_{YY} *are similarly defined.*

Now, we limit our index to J, the true active set of the population functional coefficient β. For convenience, we use the notation for truncated-version by the

superscript J such that $\beta^J = (\beta^j : j \in J) \in \mathcal{H}^J$.

Lemma 8. *In addition to the assumptions in Lemma 7, assume that for any j, there exists $g^j \in H^j$ such that $\beta^j = \Gamma^{1/2}_{X^jX^j}(g^j)$. This means each β^j is in the range of $\Gamma^{1/2}_{X^jX^j}$. Consider β^J_n as a minimizer of*

$$\frac{1}{2}E_n[(Y - \langle X^J, \beta^J\rangle)^2] + \lambda_n \sum_{j\in J} \|\beta^j\|_{\mathcal{H}^j}. \tag{6.1}$$

If $\lambda_n \to 0$ and $\lambda_n\sqrt{n} \to \infty$, then $\|\beta^J_n - \beta^J\|_{\mathcal{H}}$ converges to zero in probability, slightly slower than $\sqrt{\lambda_n} + \lambda_n^{-1}n^{-1/2}$.

The above lemma illustrates that if we know the true functional predictors, the solution to the optimization problem (3.2) achieves consistency. Let $M_n(\cdot)$ be the objective function in (6.1). Then,

$$M_n(\beta) = \frac{1}{2}\hat{\Gamma}_{YY} - \hat{\Gamma}_{YX^J}\beta + \frac{1}{2}\langle\beta, \hat{\Gamma}_{X^JX^J}\beta\rangle + \lambda_n \sum_{j\in J} \|\beta^j\|_{\mathcal{H}}. \tag{6.2}$$

Note that (6.2) is asymptotically strictly convex as long as we can assume that $\Gamma_{X^JX^J}$ is a positive-definite operator. Similarly, the original objective function (3.2) also has a unique solution if we can assume that Γ_{XX} exists and is positive definite. Finally, using Lemma 8 as a bridge, we prove the consistency of the proposed estimate in the following Theorem.

Theorem 5. *Assume that*

1. *The fourth moments of X and Y are bounded.*
2. *For any j, there exists $g^j \in H^j$ such that $\beta^j = \Gamma^{1/2}_{X^jX^j}(g^j)$.*
3. *In the population, we have such a condition that,*

$$\max_{i\in J^c} \|\Gamma^{1/2}_{X^iX^i} C_{X^iX^J} C^{-1}_{X^JX^J} diag((\cdot)/\|\beta^j\|_{\mathcal{H}})(g^J)\|_{\mathcal{H}^J} < 1,$$

where $C_{X^iX^J}$ and $C_{X^JX^J}$ are the correlation operators defined in (51)- $\Gamma_{X^iX^J} = \Gamma^{1/2}_{X^iX^i} C_{X^iX^J} \Gamma^{1/2}_{X^JX^J}$.

Then, the multivariate functional group LASSO estimate satisfies the following.

1. *Let $\hat{\beta}$ be the solution minimizing (3.2), and $\hat{J} = \{j; \hat{\beta}^j \neq 0\}$ be the estimated active set. Then, $P(\hat{J} = J)$ converges to 1.*

2. *$\|\hat{\beta} - \beta\|_{\mathcal{H}} \to 0$ in probability if λ_n approaches zero slower than the rate of $n^{-1/2}$.*

Assumption 1 is commonly used in the condition for the functional central limit Theorem. In addition, such an assumption guarantees the decay of the eigenvalues of the covariance of X. Assumption 2 states that the functional coefficients β^j lie in the support of the functional predictor X, which means that we restrict the potential range of β to be in the range of Σ_{XX}. Assumption 3 is a modified version of the necessary condition for the LASSO to be consistent that is derived in (11). In fact, this assumption states intuitively that the correlation between zero covariates and all nonzero covariates is bounded by an upper bound, so the active covariates do not drag or pull the indices of non-active covariates in the final active set when sample size grows. This assumption will be used in the proof of selection consistency.

The rate of convergence is at most $O_p(\sqrt{\lambda_n} + n^{-1/2}\lambda_n^{-1})$. This is the upper bound of the rate of estimation convergence in Lemma 8 when the true active set is known and indices are limited to it.

It is worth mentioning that the natural rivals of the proposed models, such as group sparse regression models (group LASSO and group Elastic Net) without basis transformation, do not provide a smooth estimation. In addition, they are extremely slow to estimate due to a large number of time points in the data; hence, in the following two chapters (simulation and application), we do not include them for comparison with the proposed methods.

In addition, the choice of the number of basis in simulation and application-the next two chapters-is based on the second derivative penalty and cross-validation. The curvature penalty and smooth estimation would assure us that we do not have too many basis because the curve estimation results stay the same at some number of basis when increasing the number of basis-the population coefficients can be expanded by a finite number of basis. Hence, we only need to find the smallest number of basis that would work the best. To do that, we increase the number of basis until there is no further enhancement in terms of the in-sample prediction error. For example, in the application chapter, we used $m = 31$ basis because any number of basis fewer than that has a larger in-sample prediction error, and every basis larger than that up to even $m = 110$ has almost the same in-sample prediction error. Naturally, we tend to use the fewest number of basis between 31 and 110 to reduce the complexity of the final model.

CHAPTER 7: SIMULATION STUDIES

In this section, we investigate the performance of the proposed method for scalar on functional penalized regressions through a simulation study. Consider $T = [0, 1]$ with a hundred observed time points equally-spaced, $\{t_1, \ldots, t_{100}\}$. Suppose that there are $p = 19$ random functional covariates, X^j, for $j = 1, \ldots, 19$, observed on a hundred time points equally-spaced in $T = [0, 1]$, say $\{t_1, \ldots, t_{100}\}$. For $i = 1, \ldots, n$, we first generate $X_i = (X_i^1, \ldots, X_i^p)$ on 500 time points, $\{t_1^*, \ldots, t_{500}^*\}$, where X_i^j is from a form of the Brownian motion,

$$X^j(t_i^*) = \sum_{k=1}^{i} N_k^j,$$

where $1 \leq k \leq 500$, $N_k^j \sim N(0, 1)$. We generate the response values following the model

$$Y = \langle X^1, \beta^1 \rangle + \langle X^2, \beta^2 \rangle + \langle X^3, \beta^3 \rangle + \sigma\epsilon,$$

where $\epsilon \sim N(0, 1)$, $\beta^1(t) = \sin(\frac{3\pi t}{2})$, $\beta^2(t) = \sin(\frac{5\pi t}{2})$, and $\beta^3(t) = t^2$ that are elements of $\mathcal{H}^j$ for $j = 1, 2, 3$. Therefore, there are three functional predictors out of 19 in the population active set, $J = \{1, 2, 3\}$. We drop 400 observed time points so that the remaining 100 time points are equally spaced over $[0, 1]$. To compute the inner product with more accuracy, we used 500 points in Riemann sum approximation of the inner product integrals before dropping the 400 time points.

To investigate the method thoroughly, we applied different numbers of observations (100, 200, 500) and different standard deviations for the residual term $\sigma = 0.01, 0.1, 1$. In each sample, we divide the observations into two sets for training and test sets (80% for the training set and 20% for the test set). We measure the root mean squared

error (RMSE) of the prediction for the response values of the test set. In addition, we measure the number of functional predictors that are chosen correctly. More specifically, we count the correctly identified functional predictors in the population active set, the size of which is 3, and in the population inactive set, the size of which is 16, while predicting the test response values. With a cross-validation on the number of basis between 5 and 110, and the prediction error criteria, we choose $m = 21$ B-spline basis functions to convert the observed values to functional objects and coordinate representations. The second derivative penalty would guarantee that we do not overfit the curve estimations -after some number of basis, the curve estimations remain the same. We use 5-fold cross–validation to tune the regularization parameters on a net.

In each scenario, we generate 100 samples, compute the percentages of correctly selected functional predictors that are tabulated in Table 7.1, and compute the mean and standard deviation of the test RMSE in Table 7.2. Furthermore, we compare the sparse methods and the scalar on functional ordinary least square method (OLS), ridge regression, and the oracle procedure in which only the functional predictors in the population active set are used in the OLS. For the sparse models, we apply the multivariate functional group LASSO (MFG-LASSO) and the MFG-Elastic Net (MFG-EN). The two algorithms, GMD with the strong rule and ADMM, provide similar results while the GMD algorithm is much faster on serial systems and ADMM is faster on parallel computational systems. Thus, we show the results using the GMD and strong rule algorithm in this thesis.

From Table 7.1, we can see that the MFG-sparse methods effectively select the correct functional predictors. It also shows consistency in an empirical way. In particular, they always select the active set correctly even with a large noise, but the selection performances of eliminating the inactive set predictors are poor with a small sample or large noise. The MFG-EN tends to choose more functional predictors than others. It is an expected result since the MFG-EN penalty includes the quadratic

Table 7.1: Percentages of correct selection in the test set under various simulation scenarios. In each case, 100 random samples are used. In each sample, we count the correctly identified functional predictors for the active set of the size three and the inactive set of the size 16. Then, we compute the overall percentage out of 100 samples.

Parameters		Selection	Methods			
σ	n		OLS	Ridge	MFG-LASSO	MFG-EN
0.01	100	Inactive	0	0	76	66
		Active	100	100	100	100
	200	Inactive	0	0	93	88
		Active	100	100	100	100
	500	Inactive	0	0	100	99
		Active	100	100	100	100
0.1	100	Inactive	0	0	73	64
		Active	100	100	100	100
	200	Inactive	0	0	92	86
		Active	100	100	100	100
	500	Inactive	0	0	100	99
		Active	100	100	100	100
1	100	Inactive	0	0	25	21
		Active	100	100	100	100
	200	Inactive	0	0	29	24
		Active	100	100	100	100
	500	Inactive	0	0	51	44
		Active	100	100	100	100

term, giving more stability but choosing more predictors. Because the oracle estimator assumes that the true active and inactive sets are known before OLS is run on the sample with the indices of the true active set, it always hits 100 percent for selection of active and inactive incises in this table; hence, we do not display this estimator in this table.

Table 7.2 illustrates the estimation performance using the test RMSE. The overall behavior of the methods in terms of prediction errors is similar to that of the selection performance. As the sample size grows, the RMSEs are closer to the oracle estimator's, and their standard deviations decrease. Compared to the OLS, the sparse methods outperform when there are not enough observations, or the functions are noisy. The OLS performs slightly better than the sparse methods with large

Table 7.2: Average test RMSE of different methods under different simulation scenarios. In each case, 100 random samples are used to compute the mean and standard deviation with parentheses.

Parameters		Methods				
σ	n	OLS	Ridge	MFG-LASSO	MFG-EN	Oracle
0.01	100	1.57 (0.47)	2.41 (0.54)	1.01 (0.55)	1.02 (0.55)	0.9 (0.61)
	200	0.7 (0.45)	1.22 (0.35)	0.75 (0.43)	0.76 (0.43)	0.66 (0.47)
	500	0.48 (0.3)	0.72 (0.22)	0.56 (0.26)	0.57 (0.26)	0.47 (0.31)
0.1	100	1.6 (0.47)	2.41 (0.55)	1.02 (0.55)	1.03 (0.54)	0.91 (0.6)
	200	0.73 (0.44)	1.22 (0.35)	0.76 (0.43)	0.77 (0.42)	0.67 (0.47)
	500	0.5 (0.29)	0.73 (0.22)	0.58 (0.26)	0.58 (0.26)	0.49 (0.3)
1	100	2.99 (0.65)	2.82 (0.57)	1.64 (0.44)	1.67 (0.45)	1.5 (0.44)
	200	1.95 (0.32)	1.8 (0.31)	1.37 (0.31)	1.38 (0.31)	1.32 (0.31)
	500	1.38 (0.19)	1.37 (0.17)	1.21 (0.17)	1.21 (0.17)	1.18 (0.18)

enough n and small noises. However, the standard errors of the OLS RMSE are larger than that of the MFG-methods. The ridge method is worse than the OLS with the small noise, but it is better than the OLS with the large noise. Overall, the sparse methods, MFG-LASSO and MFG-EN, perform the best in general because their results are very close to the oracle estimations. Considering that the sparse methods use much fewer functional predictors, the simulation results illustrate the remarkable effectiveness of the proposed methods in reducing both the model complexity and the prediction error.

Figure 7.1 shows the estimated functional coefficients $\hat{\beta}^1(\cdot), \ldots, \hat{\beta}^6(\cdot)$ from the MFG-LASSO in a hundred simulation samples when $n = 100$, $\sigma = 1$, the worst performance case. It must be mentioned that the estimations are individually smooth (for each of the 100 simulations) as they should be because of the curvature penalty. How-

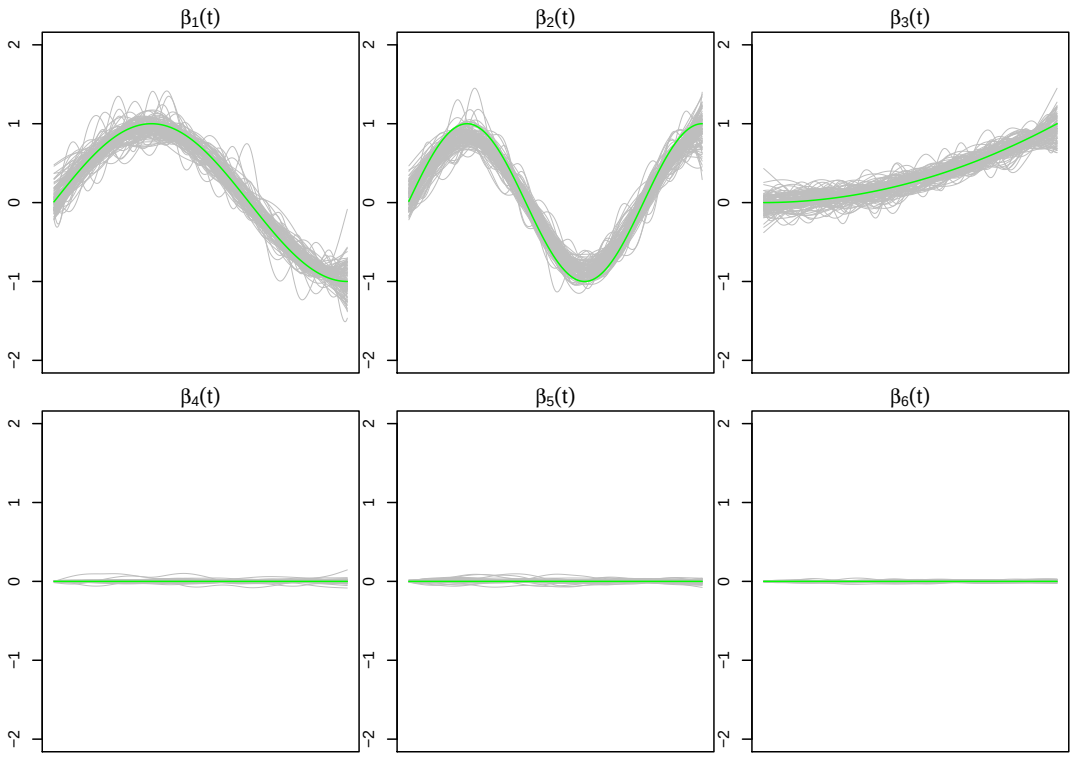


Figure 7.1: This figure displays the estimated functional coefficients by the MFG-LASSO from a hundred simulated data sets when $n = 100$, $\sigma = 1$. The green curves are the true coefficient curves, and the grey curves are the estimated coefficients. The estimated curves for the remaining of the coefficients from the seventh to the nineteenth are very similar to the fourth, fifth , and sixth functions (inactive coefficients) displayed in this figure.

ever, the estimated curves for 100 samples are highly variant due to the large noise. Thus, the curves do not look smooth when they are displayed in a single figure. The green curves are the true functions, and the rest of the curves are the estimations. Figure 7.2 shows the results when $n = 500$, $\sigma = 0.01$, the best performance case.

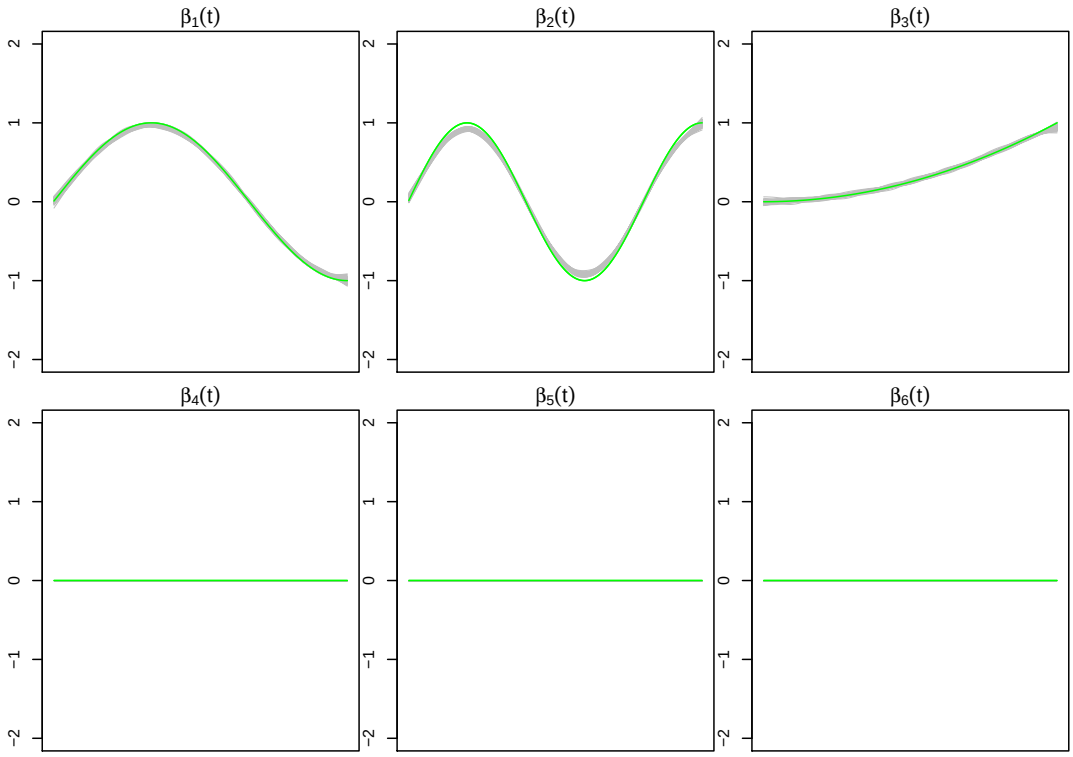


Figure 7.2: This figure displays the estimated functional coefficients by the MFG-LASSO from a hundred simulated data sets when $n = 500$, $\sigma = 0.01$. The green curves are the true coefficient curves, and the grey curves are the estimated coefficients. The estimated curves for the remaining of the coefficients from the seventh to the nineteenth are very similar to the fourth, fifth , and sixth functions (inactive coefficients) displayed in this figure.

CHAPTER 8: APPLICATIONS

In this chapter, the proposed methods are applied to fMRI and econometric data.

8.1 Applications to fMRI

We apply the proposed methods to a human brain fMRI data set collected by the New York University. This data set is part of the ADHD-200 resting-state fMRI and anatomical datasets. The parent project is 1000 Functional Connectomes Project. The fMRI machine measures the BOLD-contrast activities of the brain during a 430 seconds period of time. To extract the time courses, 172 equally-spaced signal values were recorded as the observed points within the 430 seconds period of time. Before the analysis, the automated anatomical labeling (AAL) (52) was applied to the raw fMRI data by averaging the BOLD activities of the clusters of voxels in $p = 116$ regions of the brain, the regions of interest (ROI). This procedure is called masking, clustering the voxels by regions and averaging the time series signals within the region. The data consists of between five to seven brain resting–state fMRI records taken from 290 human subjects. We randomly chose two brain images from each human subject and cleaned the data by removing missing response values. We choose different response values in each regression analysis, such as the subjects' intelligence quotient (IQ) scores, verbal IQ, performance IQ, attention deficit hyperactivity disorder (ADHD) index, ADHD Inattentive, and ADHD Hyper/Impulsive. Then, we split the data by 80% for the training set and 20% for the test set. Using cross-validation on the number of basis between 10 and 110 and the prediction error criteria, we choose $m = 31$ Fourier basis functions in the function approximation procedure.

Table 8.1 describes the test RMSE and the sparsity of the regression models. The

results show that the scalar on function OLS does not work in that the RMSE is higher than the standard deviation of the response values in the test set. The ridge regression has a significantly lower RMSE while it does not select functional covariates. The MFG-LASSO eliminates more than half of the brain regions except for the performance IQ, while its RMSE is slightly higher than the MFG-EN in most cases. The MFG-EN performs the best in terms of the RMSE while it selects more functional predictors than the MFG-LASSO. It is worth mentioning that when we change the proportion of the train and test data set to 90% and 10%, the ratio $\frac{RMSE}{\hat{\sigma}_{\mathrm{Y}_{\mathrm{test}}}}$ decreases significantly for sparse regressions; however, to be consistent with the simulations, we keep the 80% to 20% proportions for the train and test sets.

At the time of writing, no research study uses the same data. However, there are articles on predicting the IQ score based on human brain measurements. (53) predicted IQ score based on structural magnetic resonance imaging (MRI). In order to predict the IQ score, they use two methods: Principal component analysis on gray matter volume of each voxel and Atlas-based grey matter volume while adjusting for the brain size in both methods. The reported RMSE with 90% to 10% train to test proportions in this study is 13.07 at its best, while the standard deviation of the IQ scores in the whole sample including the test set is $\hat{\sigma}_Y = 12.94$. Nevertheless, the MFG-LASSO provides an RMSE of 6.32, and the MFG-EN provides 5.91. In addition, to the higher accuracy, the proposed methods have much less complexity of the model. (53) selects more than $20,000$ principal features among all of the features associated with $556,694$ voxels in the data. Meanwhile, the proposed methods use 53 functional predictors for MFG-LASSO and 106 functional predictors for MFG-EN. In each functional predictor, we use 172 time points in the raw data. Therefore, the proposed methods have obvious advantages in reducing the model complexity and achieving higher accuracy. Running one regression analysis with the proposed methods using the GMD/Strong Rule is on average around two to three minutes

Table 8.1: The results of applying the proposed methods to the fMRI data when predicting the IQ and ADHD scores.

Response value	Method	RMSE	Zero curves of 116 ROI
Y=IQ score	Least square	19.01	0
Range: $73 - 142$	Ridge	5.98	0
$\hat{\sigma}_{\mathrm{Y_{test}}} = 13.45$	MFG-LASSO	6.32	63
	MFG-EN	5.91	10
Y= Verbal IQ	Least square	23.03	0
Range: $65 - 143$	Ridge	7.02	0
$\hat{\sigma}_{\mathrm{Y_{test}}} = 13.25$	MFG-LASSO	6.98	68
	MFG-EN	6.44	16
Y= Performance IQ	Least square	19.69	0
Range: $72 - 137$	Ridge	6.27	0
$\hat{\sigma}_{\mathrm{Y_{test}}} = 13.89$	MFG-LASSO	6.79	40
	MFG-EN	6.06	10
Y=ADHD Index	Least square	28.86	0
Range: $40 - 99$	Ridge	8.18	0
$\hat{\sigma}_{\mathrm{Y_{test}}} = 15.22$	MFG-LASSO	8.49	75
	MFG-EN	8.06	25
Y=ADHD Inattentive	Least square	27.81	0
Range: $40 - 90$	Ridge	8.40	0
$\hat{\sigma}_{\mathrm{Y_{test}}} = 15.30$	MFG-LASSO	9.21	75
	MFG-EN	8.67	30
Y=ADHD Hyper/Impulsive	Least square	26.47	0
Range: $41 - 90$	Ridge	7.66	0
$\hat{\sigma}_{\mathrm{Y_{test}}} = 14.66$	MFG-LASSO	8.42	60
	MFG-EN	8.54	52

on a dual Core-i7 CPU with 16 GB memory, while the mentioned article claims an equivalent computation of 36,000 hours using two CPU kernels and 5 GB RAM. In addition, there is another research study, (54). In this article, the RMSE does not get any better than around 14 while data is from a combination of resting–state and task fMRI, and the sparse method uses voxels' functional connectivities (Pearson correlation between BOLD time series signals) as the input features.

In figure 8.1 and figure 8.2, we display the regions associated with the estimated active sets for IQ and ADHD by the MFG-LASSO, respectively. The final active sets of the algorithms were extracted and matched with the AAL's atlas, where each of the regions has a label. The regions were manually entered into the WFU picked atlas (55) tool of the SPM-12 ran on MATLAB 2020b to produce mask.nii files. The mask files were imported on MRIcron software to produce the multi-slice images.

The active sets cover the regions associated with IQ in (56) , such as the cerebello-parietal component and the frontal component. It is mentioned in the paper that the parietal and the frontal regions are strongly associated with intelligence by maintaining a connection with the cerebellum and the temporal regions. The shaded areas cover the ones mentioned in (57) as well. We provide the name of the regions associated with these active sets in the appendix.

Interestingly, ADHD and IQ share a large proportion of common active sets. For instance, when MFG-LASSO is applied, they overlap in 35 ROIs where the size of active sets are 53 and 41 for IQ and ADHD, respectively. On the other hand, the ROIs that are associated with ADHD but not with IQ are the middle and superior frontal, the Parahippocampal, the inferior parietal, and the superior temporal pole gyri. In addition, the ratio of the number of right hemisphere regions to the left ones associated with IQ is significantly greater than that of ADHD.

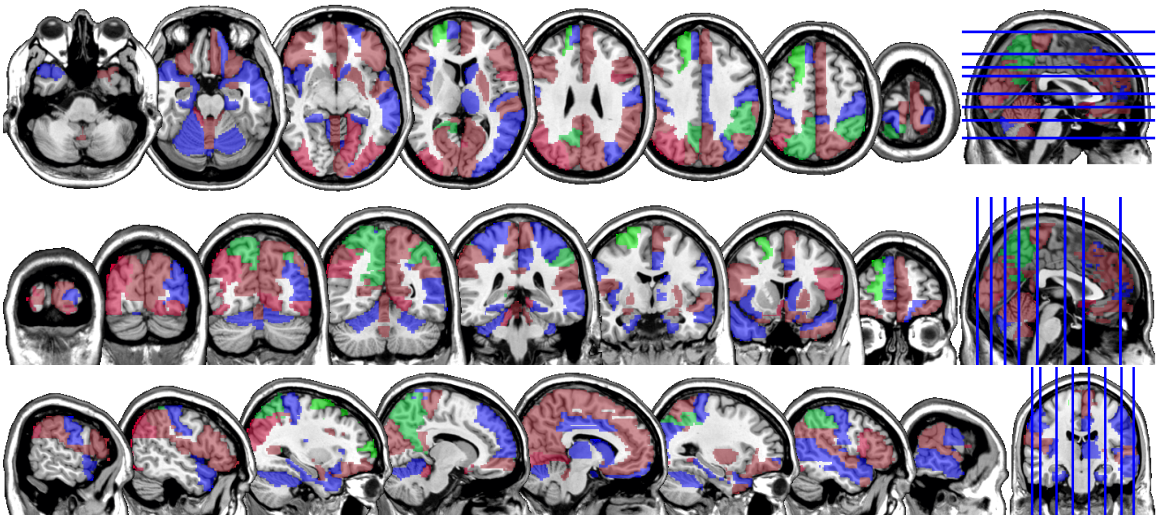


Figure 8.1: The multi-slice display (Axial, Coronal, Sagittal) of the regions of interest, the BOLD activities of which achieves the lowest prediction error and correlate the most with the IQ score variability in the sample when the MFG-LASSO is used. The regions associated with the IQ score are colored red, those associated with the performance IQ are blue, and those associated with the verbal IQ are colored green.

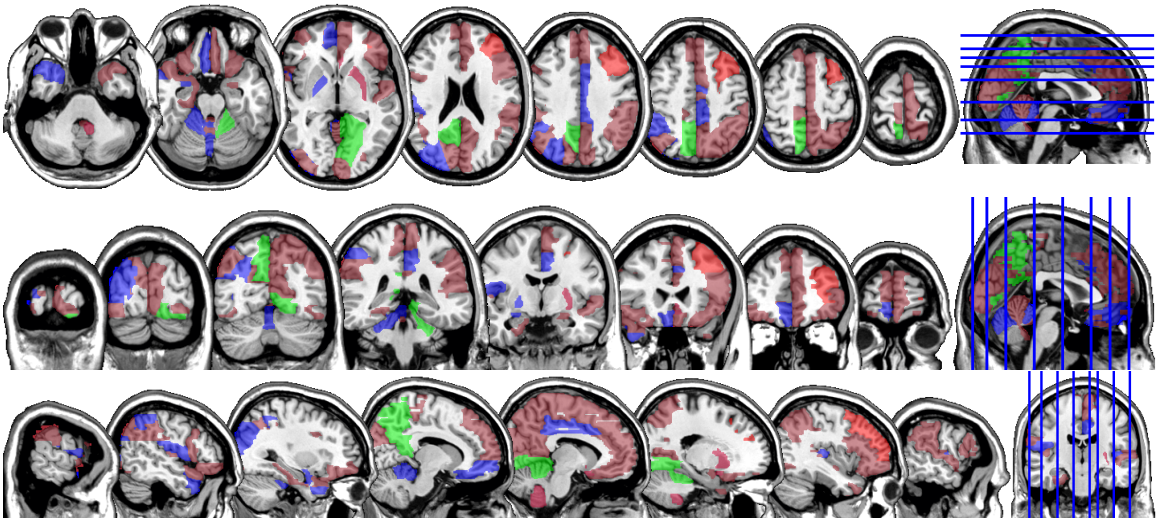


Figure 8.2: The multi-slice display (Axial, Coronal, Sagittal) of the regions of interest, the BOLD activities of which achieves the lowest prediction error and correlate the most with the ADHD score variability in the sample when the MFG-LASSO is used. The regions associated with the ADHD score are colored red, those associated with the ADHD Hyper/Impulsive are blue, and those associated with the ADHD Inattentive score are colored green.

Table 8.2: The results of applying the proposed methods to the econometric data while predicting Per Capita GDP.

Response value	Method	RMSE	Zero curves of 111 variables
Y=GDP	Least square	422312610.3	0
Range: $260 - 85135$	Ridge	8582.6	0
$\hat{\sigma}_{\mathrm{Y}_{\mathrm{test}}} = 20258.1$	Group LASSO	6998.8	95
	Group Elastic Net	7448.2	46

8.2 Application to econometric: Per capita GDP

In this section, we report the results of the proposed methods when they were applied to an econometric data set. We choose $p = 111$ functional variables, the list of which is in the appendix section. These variables are annually recorded from $1995 - 2018$ at 24 equally spaced time points for $n = 137$ countries; therefore, we consider these time-series signals as the observed points of the functional covariates. The list of the countries is mentioned in the appendix as well. These countries' per capita Gross Domestic Product (GDP) at 2019 are taken as the response values. The records of "Per Capita GDP in US Dollars" are from United Nations Statistics Division extracted in December 2020. In order to convert the time series, $m = 13$ Fourier bases are used after a cross–validation on the number of basis. The train and the test data sets are randomly chosen with a proportion of eighty to twenty percent. The results are reported in the table 8.2. The results show a significant proportion of variability in the response values of the test set is explained by its linear relationship with only 16 functional variables. We list the functional predictors selected by the functional group LASSO method. In addition, we categorize the signs of the estimated coefficient curves if they are entirely below or above the x-axis with "negative" and "positive" labels. This sign can be interpreted as the overall linear functional (through the time) contribution of the variable to the response variability. We denote $(+)$ and $(-)$ for the curves that show increasing or decreasing behaviors. This behavior, along with the sign of the curve, can be interpreted as the direction

of the linear functional contribution of the variable within time. For instance, the increasing percentage of the rural population through time correlates negatively with the per capita GDP of a country. Figures 8.3 display these curves where the period of time is scaled to the range $(0, 1)$.

- Negative curves: [2] Rural population (+), [10] Age dependency ratio (young) (+), [60] Contributing family workers (male) (−), [66] Employment in agriculture (female) (−), [96] Unemployment (total), [97] Unemployment (youth female), [99] Unemployment (youth total), [101] Vulnerable employment (male), [102] Vulnerable employment (total) (−).

- Positive curves: [3] Urban population (−), [58] Survival to age 65 (male) (−), [73] Employment in services (male) (+), [90] Ratio of female to male labor force participation rate (−).

- None: [7] Immunization (DPT), [78] Employment to population ratio (young male), [110] Inflation (GDP deflator).

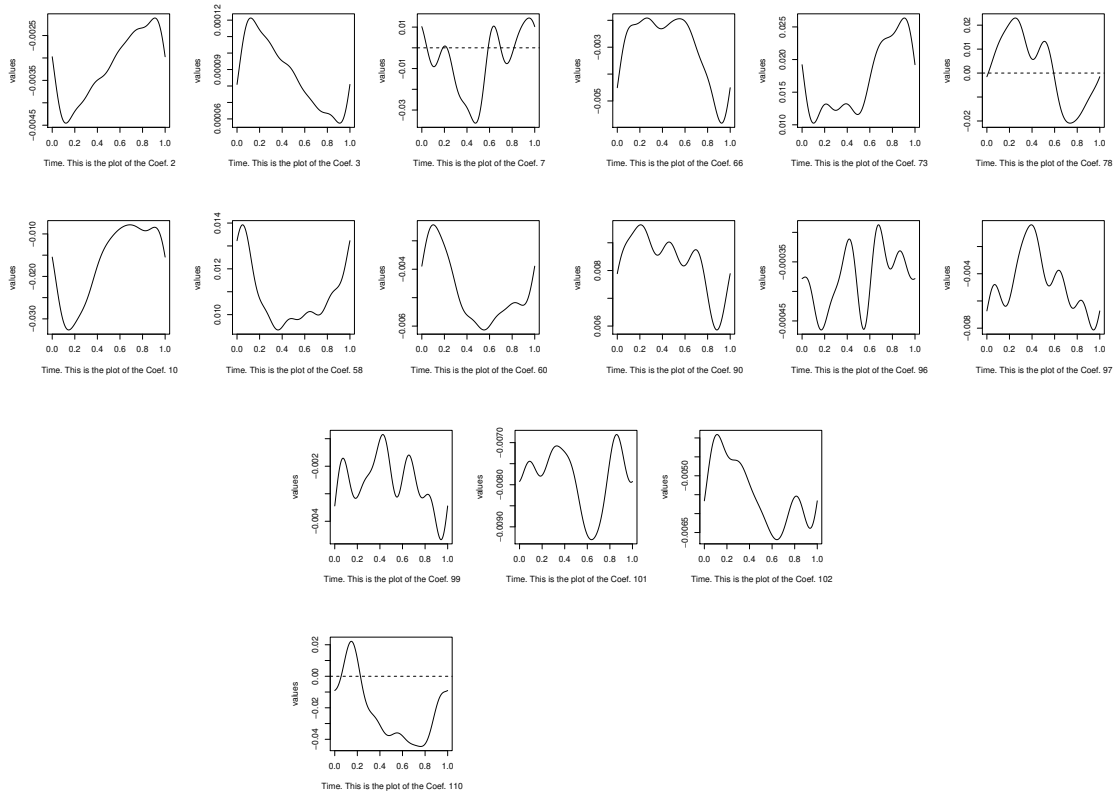


Figure 8.3: The estimated curves of coefficients among $p = 111$, selected by the group LASSO when predicting per capita GDP.

CHAPTER 9: FUTURE DEVELOPMENTS

In this chapter, we explain some of the future developments.

9.1 Sparse nonlinear scalar-on-function regression and predictor selection

We can assume that we have scalar responses and the predictors are multivariate functional data. i.e., Y is a random variable and X is a random element in $\mathcal{H}^1 \times \ldots \times \mathcal{H}^p$. Then, we consider that only a few sets of functional predictors are related to Y. The model complexity is too high since it includes higher-order interaction terms between functional predictors. Thus, we can suppose that the functional space for f, $\{g : \mathcal{H} \to \mathbb{R}\}$ has an additive structure in the following way.

Assumption 7. *There exists $f : \mathcal{H} \to \mathbb{R}$ and a set $\mathscr{A} = \{a_1, \ldots, a_d\} \subseteq \{1, \ldots, p\}$ such that*

$$Y = f_1(X^{a_1}) + \ldots + f_d(X^{a_d}) + \epsilon, \tag{9.1}$$

where ϵ is a mean-zero random variable that is independent of X.

We can further modify the model to be a non-parametric one.

Assumption 8. *There exists $f : \mathcal{H} \to \mathbb{R}$, $g : \mathcal{H}_Y \to \mathbb{R}$ and a set $\mathscr{A} = \{a_1, \ldots, a_d\} \subseteq \{1, \ldots, p\}$ such that*

$$g(Y) = f_1(X^{a_1}) + \ldots + f_d(X^{a_d}) + \epsilon, \tag{9.2}$$

where $\mathcal{H}_Y$ is the space for Y. ϵ is a mean-zero random variable that is independent of X.

For each $j = 1, \ldots, p$, define $\mathfrak{M}^j$ to be the reproducing kernel Hilbert space (RKHS) over $\mathcal{H}^j$ with a reproducing kernel κ^j. If we choose a Gaussian radial basis function (RBF) as the reproducing kernel, it can be shown that $\mathfrak{M}^j$ is a dense subset in $L_2(\mathcal{H}^j)$, which is a rich enough space to search for the $f^j(\cdot)$.

Lemma 9. *Let $\mathfrak{M}^j$ be a reproducing kernel Hilbert space (RKHS) with κ^j for $j = 1, \ldots, p$. Define $\mathfrak{M} = \mathfrak{M}^1 \times \cdots \times \mathfrak{M}^p$ with the inner product,*

$$\langle f, g \rangle_{\mathfrak{M}} = \textstyle\sum_{j=1}^{p} \langle f^j, g^j \rangle_{\mathfrak{M}^j}, \quad f, g \in \mathfrak{M}$$

Then $\mathfrak{M}$ is also a RKHS with the reproducing kernel, $\kappa(f, g) = \sum_{j=1}^{p} \kappa^j(f^j, g^j)$.

Note that, in addition to Lemma 9, $\mathfrak{M}$ is a dense subset of $L_2(\mathcal{H})$ if we use the Gaussian RBF as the reproducing kernel.

The Sample level loss function with the sparse penalty is

$$L_n(f; \lambda_n, X, Y) = E_n(Y - f(X))^2 + \lambda_n \textstyle\sum_{j=1}^{p} \|f^j\|_{\mathfrak{M}^j}. \tag{9.3}$$

Given a random element $X \in \mathcal{H}$, $\kappa(\cdot, X)$ is another random element in $\mathfrak{M}$. We define the covariance operator for $\kappa(\cdot, X)$ by

$$\Gamma_{XX} = E[\{\kappa(\cdot, X) - E\kappa(\cdot, X)\} \otimes \{\kappa(\cdot, X) - E\kappa(\cdot, X)\}].$$

Note that this is equivalent to assuming that Γ_{XX} is a unique element satisfying

$$\langle f, \Gamma_{XX} g \rangle_{\mathfrak{M}} = \mathrm{cov}(f(X), g(X)),$$

for any $f, g \in \mathfrak{M}$.

9.1.1 Nonlinear functional sparse group LASSO

Let $(X_1^1(\cdot),\ldots,X_1^p(\cdot)),\ldots,(X_n^1(\cdot),\ldots,X_n^p(\cdot))$ be the n-random copies of $X \in \mathcal{H} = \mathcal{H}^1 \times \cdots \times \mathcal{H}^p$. Define $\mathfrak{M}_n^j$ to be the RKHS over $\{X_1^j,\ldots,X_n^j\} \subseteq \mathcal{H}^j$ with a reproducing kernel $\kappa^j(\cdot,\cdot)$, for $j = 1,\ldots,p$. In other words, $\mathfrak{M}_n^j = \text{span}(\{\kappa^j(\cdot, X_i^j) : i = 1,\ldots,n\})$ with the inner product,

$$\langle \kappa^j(\cdot, X_i^j), \kappa^j(\cdot, X_k^j)\rangle_{\mathfrak{M}^j} = \kappa^j(X_i^j, X_i^k).$$

Define $\mathfrak{M}_n = \bigoplus_j \mathfrak{M}_n^j$. Then, $\mathfrak{M}_n^j$ and $\mathfrak{M}_n$ are the subspaces of $\mathfrak{M}^j$ and $\mathfrak{M}$ respectively. The following Theorem justifies the sample-level estimation.

Theorem 6. *Let $\hat{f} \in \mathfrak{M}$ be the solution to the following optimization problem,*

$$\begin{aligned} &\textit{minimize} \quad E_n(Y - f(X))^2 + \lambda_n \textstyle\sum_{j=1}^p \|f^j\|_{\mathfrak{M}^j}, \\ &\textit{subject to} \quad f \in \mathfrak{M}. \end{aligned} \tag{9.4}$$

$\hat{f}$ is the solution to optimization problem over the restricted space, $\mathfrak{M}_n$,

$$\begin{aligned} &\textit{minimize} \quad E_n(Y - f(X))^2 + \lambda_n \textstyle\sum_{j=1}^p \|f^j\|_{\mathfrak{M}^j}, \\ &\textit{subject to} \quad f \in \mathfrak{M}_n. \end{aligned}$$

Proof: Let $\mathfrak{M}_n^\perp \subseteq \mathfrak{M}$ be the orthogonal complement of $\mathfrak{M}_n$. For any $f \in \mathfrak{M}$, there exist $f_n \in \mathfrak{M}_n$ and $f_n^\perp \in \mathfrak{M}_n^\perp$ such that $f = f_n + f_n^\perp$. Then, (9.4) is

$$\begin{aligned} &\textstyle\sum_{i=1}^n (Y_i - \langle f_n + f_n^\perp, \kappa(\cdot, X_i)\rangle_{\mathfrak{M}})^2 + \lambda_n \sum_{j=1}^p (\|(f_n)^j\|_{\mathfrak{M}^j} + \|(f_n^\perp)^j\|_{\mathfrak{M}^j}) \\ &= \textstyle\sum_{i=1}^n (Y_i - \langle f_n, \kappa(\cdot, X_i)\rangle_{\mathfrak{M}})^2 + \sum_{j=1}^p (\|(f_n)^j\|_{\mathfrak{M}^j} + \|(f_n^\perp)^j\|_{\mathfrak{M}^j}) \\ &\geq \textstyle\sum_{i=1}^n (Y_i - \langle f_n, \kappa(\cdot, X_i)\rangle_{\mathfrak{M}})^2 + \sum_{j=1}^p (\|(f_n)^j\|_{\mathfrak{M}^j}. \end{aligned}$$

Thus, the minimizer of the last equation over $f_n \in \mathfrak{M}_n$ is the minimizer of the original optimization problem over $f \in \mathfrak{M}$. □

9.1.2 Implementation and Coordinate representation of Hilbertian element

For the first level Hilbert space, suppose that we have n random copies from the model (9.2) denoted by $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we observe X_i^j on $\{t_{i1}^j, \ldots, t_{ia_i}^j\}$ for each $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

At the sample level, we assume that $\mathcal{H}^j$ is spanned by a given set of basis functions, $\mathcal{B}^j = \{b_1^j, \ldots, b_{m_j}^j\}$. Thus, for any $f \in \mathcal{H}^j$, there exist a vector in $a \in \mathbb{R}^{m_j}$ uniquely such that $f(\cdot) = \sum_{k=1}^{m_j} a_k b_k^j(\cdot)$. We call the vector a the coordinate of f and denote it by $[f]_{\mathcal{B}^j}$. We also assume that $\mathcal{H}^j$ is constructed with the L_2-inner product with respect to the Lebesgue measure,

$$\langle f, g \rangle_{\mathcal{H}^j} = \int_{T_j} f(t)g(t)dt, \quad \text{for any } f, g \in \mathcal{H}^j.$$

Let $G_{\mathcal{B}^j}^j$ be $m_j \times m_j$ matrix whose (i,k)-th entry is $\langle b_i^j, b_k^j \rangle_{\mathcal{H}^j} = \int_{T_j} b_i^j(t) g_k^j(t)dt$. Then for any $f^j, g^j \in \mathcal{H}^j$,

$$\langle f^j, g^j \rangle_{\mathcal{H}^j} = [f^j]_{\mathcal{B}^j}^{\mathsf{T}} G_{\mathcal{B}^j}^j [g^j]_{\mathcal{B}^j}.$$

We use the basis-expansion approach for each functional covariate in the first level Hilbert space, X_i^j for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, which was also used in (47; 48). Without loss of generality, we assume $m = m_1 = \cdots = m_p$ and $M = pm$.

For $j = 1, \ldots, p$ consider the second level (reproducing) Hilbert space with respect to Gaussian radial basis functions $\kappa^j(X_i^j, X_l^j) = exp(-\gamma^j \|X_i^j - X_l^j\|_{\mathcal{H}^j}^2)$ for $i, l = 1 \ldots n$. We need to first compute $\langle X_i^j, X_l^j \rangle_{\mathcal{H}^j}$ for all possible values of $i, l = 1, \ldots, n$ as described above in order to compute the $\|X_i^j - X_l^j\|_{\mathcal{H}^j}^2$.

Given a vector of $\Gamma = (\gamma^1, \ldots, \gamma^p)$, the sample level problem in the second level Hilbert space is defined as follows. Next, we define the loss function:

$$L([f]) = \frac{1}{2}\|Y - G[f]\|_2^2, \tag{9.5}$$

where $[f] \in \mathbb{R}^{np}$ and $G \in \mathbb{R}^{n \times np}$ has vertical blocks $G^j \in \mathbb{R}^{n \times n}$ for each $j = 1, \ldots, p$ that are $G^j = (G)_{\{i,(l-1)p+1:lp\}} = \kappa^j(X_i^j, X_l^j)$ for $i, l = 1, \ldots, n$. The sample level optimization problem 9.4 is $\arg\min_{[f]} L([f]) + g(f)$ where $g(f) = \lambda \sum_{j=1}^p \sqrt{[f^j]^\mathsf{T} G^j [f^j]} + \alpha \sum_{j=1}^p [f^j]^\mathsf{T} G^j [f^j]$.

Lemma 10. *The loss function (9.5) satisfies the quadratic majorization (QM) condition with $H = G^\mathsf{T} G$. In other words, for any $f, f^* \in \mathcal{H}$,*

$$L([f]) \leq L([f^*]) + ([f] - [f^*])\nabla L([f^*]) + \frac{1}{2}([f] - [f^*])^\mathsf{T} H([f] - [f^*]), \tag{9.6}$$

where $\nabla L(f^*|D) = G^\mathsf{T}(G[f]^* - Y)$.

9.1.2.1 Non-linear group majorization decent: The first algorithm

Lemma 11. ***Gradient Decent Threshold****: Suppose that a positive definite matrix $\mathbb{G} \in \mathbf{R}^{n \times n}$, vector $y \in \mathbf{R}^n$, and constants $\lambda, \alpha \in \mathbb{R}$ are known. Consider the objective function:*

$$h(x) = \frac{1}{2}\|x - y\|^2 + \lambda\alpha x^\mathsf{T}\mathbb{G}x + \lambda(1-\alpha)\sqrt{x^\mathsf{T}\mathbb{G}x}. \tag{9.7}$$

If $\sqrt{y^\mathsf{T}\mathbb{G}y} < \lambda(1-\alpha)$ then the minimizer of $h(\cdot)$ is $0 \in \mathbb{R}^n$. Otherwise, if $\sqrt{y^\mathsf{T}\mathbb{G}y} > \lambda(1-\alpha)$, the minimizer of $h(\cdot)$ can be approximated via gradient decent by initializing x and updating $x_{new} = x - \nabla h(x)$ where $\nabla h(x) = (I + 2\alpha\lambda\mathbb{G} + \frac{\lambda(1-\alpha)}{\sqrt{x\mathbb{G}x}}\mathbb{G})x - y$.

Proof. The sub–gradient of $h(\cdot)$ is

$$\nabla h(x) = x - y + 2\alpha\lambda\mathbb{G}x + \lambda(1-\alpha)s,$$

where $s \in \mathbb{R}^n$ is the sub–differntial of $(x^T\mathbb{G}x)^{\frac{1}{2}}$. By Cauchy–Schwartz (CS) inequality:

$$s = \begin{cases} \frac{\mathbb{G}x}{(x^T\mathbb{G}x)^{\frac{1}{2}}} & x \neq 0 \\ \{z \in \mathbb{R}^n | (z^T\mathbb{G}z) \leq 1\} & x = 0 \end{cases} \tag{9.8}$$

Therefore, $x = 0 \in \mathbb{R}^n$ if and only if $0 \in \mathbb{R}^n$ is in the sub–differntial of $h(\cdot)$. This is equivalent to the treshholding rule that $x = 0$ if $\sqrt{y^\mathsf{T}\mathbb{G}y} < \lambda(1-\alpha)$. On the other hand, if $\sqrt{y^\mathsf{T}\mathbb{G}y} > \lambda(1-\alpha)$, $h(\cdot)$ is a convex problem. Hence, its solution can be approximated via gradient decent. □

Note that in the above Lemma, the stopping criteria can be based on the distance of the x updates in $\mathbb{R}^n$ and the absolute difference of the objective function $h(\cdot)$ within updates. Let f^* be the current solution to the optimization problem and f be the next update. Assume that we update the f for $j = 1, \ldots, p$. In other words, $[f] - [f^*]$ has a form of $(0, \ldots, 0, [f^j] - [(f^*)^j], 0, \ldots, 0)^\mathsf{T}$, which leads to simplification of the objective function of the new optimization problem. Let $U = -\nabla L(f^*)$ and U^j be the sub-vector of U with the indices $(n(j-1)+1, \ldots, nj)$. Let H^j be the j-th block diagonal matrix of H in lemma 10. Then, the problem is

$$\begin{aligned} L([f]) &\leq L([f^*]) - ([f^j] - [(f^*)^j])U^j + \frac{1}{2}([f^j] - [(f^*)^j])^\mathsf{T} H^j ([f^j] - [(f^*)^j]) \\ &\leq L([f^*]) - ([f^j] - [(f^*)^j])U^j + \frac{1}{2}\eta_j([f^j] - [(f^*)^j])^\mathsf{T}([f^j] - [(f^*)^j]), \end{aligned}$$

where η_j is a value slightly larger than the largest eigenvalue of H^j, which further relaxes the upper bound. In practice, we take $\eta_j = (1+\epsilon^*)\zeta_j$ with $\epsilon^* = 10^{-6}$ where ζ_j is the largest eigenvalue of H^j. Finally, the update rule for f^j is the solution to the

following optimization problem:

$$\arg\min_{f^j \in \mathcal{H}^j} -([f^j] - [(f^*)^j])U^j + \frac{1}{2}\eta_j([f^j] - [(f^*)^j])^{\mathsf{T}}([f^j] - [(f^*)^j]) + g^j(f), \quad (9.9)$$

where g^j is the j-th term of $g(\cdot)$. We use a similar trick of Lemma 11. Consider $U^j = -\nabla L([f^j]^{(\text{old})})$, $\nabla L(f^*|D) = G(G^{\mathsf{T}}[f] - Y)$, and $e = U^j + \gamma_j[f^j]^{(\text{old})}$. If $\sqrt{e^{\mathsf{T}}Ge} < \lambda(1-\alpha)$, $[f^j]^{(\text{new})} = 0$. Otherwise, if $\sqrt{e^{\mathsf{T}}Ge} > \lambda(1-\alpha)$, $[f^j]^{(\text{new})}$ can be approximated as following. First initialize $[f^j]$, then until convergence $[f^j]_{(\text{update})} = [f^j] - (I - \frac{\lambda}{\sqrt{[f^j]G[f^j]}})[f^j] + e$. The convergence criteria can be a combinations of the absolute difference of the objective function 9.9 at each step and the euclidean distance of $[f^j]_{(\text{update})}$ with $[f^j]$. After convergence, update $[f^j]^{(\text{new})} = [f^j]_{(\text{update})}$.

Initialization;

Compute $H = G^{\mathsf{T}}G$, and η_j;

while *the stopping criteria does not meet for* $[f]$ **do**

  for $j = 1, \ldots, p$;

  Compute $U = -G^{\mathsf{T}}(G[f] - Y)$, and $e = U^j + \eta_j[f^j]$;

  if $e^T G^j e < (\lambda(1-\alpha)/\eta_j)^2$ **then**

    $[f^j] = 0 \in \mathbb{R}^n$;

  else

    $[f^j] = e$;

    while *the stopping criteria does not meet for* $[f^j]$ **do**

      $[f^j]_{(\text{update})} = [f^j] - (I + 2\alpha\lambda\eta_j G^j + \frac{\lambda G^j}{\sqrt{[f^j]^T G^j [f^j]}})[f^j] + e$

    end

  end

end

Algorithm 1: Nonlinear GMD

9.1.2.2 Alternative algorithm

The following lemma help to develop a second algorithm.

Lemma 12. *Take $x, y \in \mathbb{R}^m$ where y is known.*

$$\arg\min_x \left(\frac{1}{2}\|x-y\|^2 + \lambda\|x\| \right) = S_\lambda(y), \tag{9.10}$$

where $S_\lambda(y) := 1_{\{\|y\|>\lambda\}} \left(1 - \frac{\lambda}{\|y\|}\right)_+ y$ is the block soft threshold operator in real space.

Proof of Lemma 12. Observe that

$$\arg\min_x \left(\frac{1}{2}(x-y)^\mathsf{T}(x-y) + \lambda\|x\| \right) = \arg\min_x \left(\frac{1}{2}(x^\mathsf{T}x - 2x^\mathsf{T}y) + \lambda\|x\| \right).$$

To satisfy the Karush-Kuhn-Tucker (KKT) stability condition, the derivative of the above objective function with respect to x must be equal to zero. If the derivative does not exist, subdifferential must include zero. The derivative is $x - y + \lambda s_x$, where s_x is the subdifferential of $\|x\|$ at x.

If $x \neq 0$, $s_x = x/\|x\|$ and the KKT condition gives

$$x(1 + \lambda/\|x\|) = y.$$

Compute the $\|y\|$ in preceding equation and solve for $\|x\|$. Plugging it back into the equation gives us,

$$x = (1 - \lambda/\|y\|)y.$$

The condition $x \neq 0$ is equivalent to $\|y\| > \lambda$. On the other hand, $x = 0$ is equivalent to $0 \in -y + \lambda s_x$, or $y \in \lambda s_x$. In this case $s_x = \{z \in \mathbb{R}^m | \|z\| \leq 1\}$. Therefore, $\|y\|^2 \leq \lambda^2$ which completes the proof. □

Lemma 13. *For $x, y \in \mathbb{R}^m$ where y is known and constants a, b:*

$$\arg\min_x \left(\frac{1}{2}(x-y)^\mathsf{T}(x-y) + a(x^\mathsf{T}x)^{\frac{1}{2}} + \frac{b}{2}x^\mathsf{T}x \right) = \frac{1}{b+1}S_a(y).$$

Proof of Lemma 13.

The proof is similar to the one in lemma 12. The only difference is the derivative of the objective function. It is: $x - y + as + bx$, where s is the sub–differential. The rest of the proof is straight forward. If $x \neq 0$ we see that $x(1 + b + \frac{a}{\|x\|}) = y$. Taking norm $\|.\|$ from both sides, solving for $\|x\|$, and plugging it back we have $x = (\frac{1}{1+b})(1 - \frac{a}{\|y\|})y$. Note that this is only possible when $\|x\| > 0$, which means $\|y\| > a$. If $x = 0$, it results in $0 \in -Gy + as$, or $Gy \in as$. Thus, since in this case $s = \{[Z] | Z \in \mathcal{H}^l \,\&\, \|Z\|_{\mathcal{H}^l} \leq 1\}$, $\|Gy\| \leq a$, which completes the derivation above.

□

The quadratic form in $g(\cdot)$ allows the following inequality:

$$g^j(f) < \lambda\sqrt{\omega_j}\textstyle\sum_{j=1}^p \sqrt{[f^j]^\mathsf{T}[f^j]} + \alpha\omega_j \sum_{j=1}^p [f^j]^\mathsf{T}[f^j],$$

where ω_j are slightly larger than the maximum eigenvalues of G^js. Hence, it is possible to extend the inequality 9.9 to:

$$\begin{aligned} 1 - ([f^j] - [(f^*)^j])U^j + \frac{1}{2}\eta_j([f^j] - [(f^*)^j])^\mathsf{T}([f^j] - [(f^*)^j]) + g^j(f) < \\ -([f^j] - [(f^*)^j])U^j + \frac{1}{2}\eta_j([f^j] - [(f^*)^j])^\mathsf{T}([f^j] - [(f^*)^j]) \\ + \lambda\sqrt{\omega_j}\textstyle\sum_{j=1}^p \sqrt{[f^j]^\mathsf{T}[f^j]} + \alpha\omega_j \sum_{j=1}^p [f^j]^\mathsf{T}[f^j]. \end{aligned}$$

This algorithm finds the minimizer of the right-hand side of such inequality.

$$\arg\min_{f^j \in \mathcal{H}^j} -([f^j]-[(f^*)^j])U^j + \frac{1}{2}\eta_j([f^j]-[(f^*)^j])^{\mathsf{T}}([f^j]-[(f^*)^j]) + \lambda\sqrt{\omega_j}\sum_{j=1}^{p}\sqrt{[f^j]^{\mathsf{T}}[f^j]} + \alpha\omega_j\sum_{j=1}^{p}[f^j]^{\mathsf{T}}[f^j] \tag{9.11}$$

Similar to the Lemma 13, the optimizer has a closed form solution:

$$[f^j]^{(\text{update})} = \frac{1}{2\alpha\omega_j + \eta_j} S_{\lambda\sqrt{\omega_j}}(U^j + \eta_j[f^j]^{(\text{old})}), \quad j = 1, \ldots, p, \tag{9.12}$$

where $U^j = -\nabla L([f^j]^{(\text{old})})$.

The numerical convergence of such an algorithm is proven similar to that of the first one due to the strictly descending property. However, the advantage of the second algorithm over the first one is that although it requires more iterations for overall group updates to converge, it has a closed-form solution to update each of the groups. In other words, the inner loop in the first algorithm is removed; therefore, when p is large, it is expected that the alternative algorithm converges faster.

9.1.2.3 Strong rule

While iterating one of the two above algorithms, we can reduce the computational burden more efficiently based on the tuning parameter with the Strong Rule technique. See (50).

Step 1. (Initialization) Given $\alpha \in (0, 1)$, we search for the smallest value of λ such that all coefficients are zero. Using the KKT condition (see Lemma 12), this value which is the largest among the grid points, is

$$\lambda^{(1)} = \max_j \frac{\|U^j(0)\|}{(1-\alpha)\sqrt{\omega_j}}.$$

Therefore, the initial f is zero. Then the smallest λ of the grid points are set to be

a certain small number to include all functional predictors. This process starts from the largest value in the grid point and moves backward to the smallest value.

Step 2. (Iteration) At $\lambda^{(k)}$, we add the j-th functional predictor to the active set if it satisfies the Strong Rule condition,

$$\|U^j([f^j(\lambda^{(k)})])\| > (2\lambda^{(k+1)} - \lambda^{(k)})(1-\alpha)\sqrt{\omega_j},$$

for $j = 1, \ldots, p$. Then, we update f with this reduced predictors using one of the two algorithms in previous sections. Next, in order to make sure that the Strong Rule does not leave out some of the worthy coefficients, we check the KKT condition on the rest of the blocks of the current solution,

$$\|U^j([f^j_{update}(\lambda^{(k+1)})])\| < \lambda^{(k+1)}(1-\alpha)\sqrt{\omega_j},$$

where $f^j_{update}(\lambda^{(k+1)})$ is the updated f^j from the Strong Rule screening. If the j-th functional coefficient violates the KKT condition, we add it to the active set and update f using the algorithms. This process of checking the KKT condition and updating continues until there is no functional coefficient that violates the KKT condition when we store the solution of the final update as the $f^j(\lambda^{(k+1)})$.

9.1.3 Prediction

In algorithm 1, $[f]$ is computed which is the estimation of the coordinate representation of $f \in \mathfrak{M}$ with respect to basis $\kappa(\cdot, X_1), \ldots \kappa(\cdot, X_n)$ that are elements in the RKHS at the population level. Let new observations $X^* = (X^*_1, \ldots, X^*_{ntst})$ be in the test or validation set for $Y^* = (Y^*_1, \ldots, Y^*_{ntst})$. Then, the estimation of $f(X^*_i)$ is a linear combination of $\kappa(X^*_i, X_1), \ldots, \kappa(X^*_i, X_n)$ for $i = 1, \ldots, ntst$. In other words, to use the computed $[f]$ in the prediction of new response values, we must first compute the coordinate representation of the new observed covariates in

the first level Hilbert space as explained in the subsection 9.1.2, then, compute a matrix $G^* \in \mathbb{R}^{ntst \times np}$ that has vertical blocks $G^j \in \mathbb{R}^{ntst \times n}$ for each $j = 1, \ldots, p$ that are $G^j = (G)_{\{i,(l-1)p+1:lp\}} = \kappa^j((X_i^j)^*, X_l^j)$ for $i = 1, \ldots, ntst$ and $l = 1, \ldots, n$. The prediction is $\hat{Y}^* = G^*[f]$.

9.1.4 Γ tuning

The hessian of G with respect to Γ is a diagonal matrix with strictly positive values. Hence G is convex with respect to Γ. In addition, $G^T G$ is convex in Γ, and its inverse is bounded; hence, the minimizer of 5.2 is convex in Γ. The G^* and the norm operation are convex, so the out-of-sample error is convex in Γ if G is of full rank. If the model is correctly specified as in 9.1, the non-linearity of f^j is not affected by the inactive functional predictors. Therefore, we can rely on coordinate descent to optimize for coordinates of Γ based on the cross-validation with the non-penalized model. In addition, we can randomly set the initial states of the coordinate descent on the net to converge to the solution. This optimization problem is strictly convex with respect to individual γ^j; hence, a reasonable point estimate to determine the boundaries of the net grids can be found by validating via a simple non-penalized least square estimation for individual f^j.

Instead of a fixed net, we can use the partial derivative of the in-sample mean squared error with respect to γ^j and use it in a gradient coordinate descent method. This method is similar to coordinate descent, while to minimize with respect to each coordinate, we employ this partial derivative instead of optimizing over a grid. The in-sample mean squared error is strictly convex with respect to individual γ^j if G is or is not of full rank.

$$\frac{\partial}{\partial \gamma^j} \|Y - G_\Gamma[f]\|_2^2 = 2 * (Y - G_\Gamma[f])^T G_\Gamma'^j[f^j], \tag{9.13}$$

where $G_\Gamma'^j \in \mathbb{R}^{n \times n}$ with elements $(G_\Gamma'^j)_{\{i,l\}} = \|X_w^j - X_l^j\|_{\mathcal{H}}^2 \kappa^j(X_i^j, X_l^j)$ for $i, l = 1, \ldots, n$.

Hence, on each coordinate of the Γ we move toward the minimizer by:

$$\gamma^j_{\text{update}} = \gamma^j - \alpha \frac{\partial}{\partial \gamma^j} \|Y - G_\Gamma[f]\|_2^2, \tag{9.14}$$

where α is the learning rate. It is worth mentioning that only a part of G must be updated: G^j after γ_j updates. After stopping, we would do it with the next coordinate until all coordinates are minimized. Then, after updating all coordinates, we check the in-sample mean squared error update with the last time that all coordinates were updated. The stopping criteria are when these two errors are close enough, or when $\|\Gamma_{update} - \Gamma\|^2$ is small enough

9.2 Sparse logistic scalar-on-function regression

Logistic regression can be seen as a modification of functional linear regression model and a particular case of penalized likelihood regression used to analyse binary dependent variable $\mathbf{Y} \in \{0, 1\}$. The population model is assumed to be:

$$\log(\frac{P}{1-P}) = \alpha + \langle X, \beta \rangle_{\mathcal{H}}, \tag{9.15}$$

where $P = P(\mathbf{Y} = 1|X)$, and $\alpha \in \mathbf{R}$. This can be also written as

$$P = S(\alpha + \langle X, \beta \rangle_{\mathcal{H}}), \tag{9.16}$$

where $S(\cdot)$ is one dimensional Sigmoid function $S(z) = (1 + e^{-z})^{-1}$.

Thus, $\mathbf{Y}$ assumed to have Bernoulli distribution with parameter $S(\alpha + \langle X, \beta \rangle_{\mathcal{H}})$. The joint likelihood of α and β given a random sample $(X_1, \mathbf{Y}_1), \ldots (X_n, \mathbf{Y}_n)$ is:

$$L(\alpha, \beta) = \Pi_{i=1}^n S(\alpha + \langle X_i, \beta \rangle_{\mathcal{H}})^{\mathbf{Y}_i} (1 - S(\alpha + \langle X_i, \beta \rangle_{\mathcal{H}}))^{1-\mathbf{Y}_i}, \tag{9.17}$$

Hence, the log–likelihood is:

$$l(\alpha,\beta)=\sum_{i=1}^{n}\mathbf{Y}_i\log(S(\alpha+\langle X_i,\beta\rangle_{\mathcal{H}}))+(1-\mathbf{Y}_i)\log(1-S(\alpha+\langle X_i,\beta\rangle_{\mathcal{H}})). \quad (9.18)$$

We can look at $\mathbf{L}(\alpha,\beta):=-l(\alpha,\beta)$ as the cost function of this model. This means in the original optimization problem $\mathbf{f}(\cdot)$ is substituted. The sample level of gradient is:

$$\nabla L(\alpha,[\beta])=-\sum_{i=1}^{n}\{y_i\log(S(\alpha+[x_i]\mathcal{G}[\beta]))+(1-y_i)\log(1-S(\alpha+[x_i]\mathcal{G}[\beta]))\}.$$

Consequently, the objective function is $L([\beta])+g(\beta)$ where
$g(\beta)=\lambda(1-\alpha)\sum_{j=1}^{p}\sqrt{[\beta^j]^{\mathsf{T}}G^j[\beta^j]}+\alpha\lambda\sum_{j=1}^{p}[\beta^j]^{\mathsf{T}}G^j[\beta^j]$.

9.2.1 First algorithm

The following Lemma shows that a similar iteration rule to that of (5.7) with the Strong Rule can numerically solve for the Functional Logistic model solution.

Lemma 14. *The loss function (9.17) satisfies the quadratic majorization (QM) condition with* $H=1/4(G[\tilde{X}_{1:n}]^{\mathsf{T}}[\tilde{X}_{1:n}]G)+\lambda_{der}B''$. *In other words, for any* $\beta,\beta^*\in\mathcal{H}$,

$$L([\beta])\leq L([\beta^*])+([\beta]-[\beta^*])\nabla L([\beta^*])+\frac{1}{2}([\beta]-[\beta^*])^{\mathsf{T}}H([\beta]-[\beta^*]). \quad (9.19)$$

The proof is similar to that of (58). Thus, the logistic optimization problem can be numerically solved by iterating on the following update rule.

$$[\beta^j]^{(\text{new})}=\frac{1}{2\alpha\lambda+\gamma_j}S^{\mathcal{H}^j}_{\lambda(1-\alpha)}(U^j+\gamma_j[\beta^j]^{(\text{old})}),\quad j=1,\ldots,p, \quad (9.20)$$

where $U^j=-\nabla L([\beta^j]^{(\text{old})})$.

9.2.2 Second algorithm

We can consider $\mathbf{f}(\alpha, \beta) := -l(\alpha, \beta)$ as the cost function of this model, and find its minimizer. This means that in the original optimization problem (4.1), $\mathbf{f}(\cdot)$ is substituted. The sample version of Lagrangian problem, $f(\cdot)$ would be replaced by

$$f(\alpha, [\beta]) = -\sum_{i=1}^{n} \{y_i \log(S(\alpha + [x_i]\mathcal{G}[\beta])) + (1 - y_i) \log(1 - S(\alpha + [x_i]\mathcal{G}[\beta]))\}.$$

For convenience, we denote the vector $b = (\alpha, [\beta]^T)^T$, block matrix $\mathcal{F} = \begin{bmatrix} 1 & 0 \\ 0 & \mathcal{G} \end{bmatrix}$, and vector $q_i = [1, [x_i]]$. With the above notation, the loss function is:

$$f(b) = -\sum_{i=1}^{n} \{y_i \log(S(q_i \mathcal{F} b)) + (1 - y_i) \log(1 - S(q_i \mathcal{F} b))\}.$$

$[\beta]$–update in scaled ADMM procedure would be an α and a $[\beta]$–update simultaneously or a b–update. Note that we would not penalize α.

Proposition 1. *An α–update and $[\beta]$–update for the optimization problem (3.2) in the Logistic regression case using gradient descent method is :*

$$\alpha^{new} := \alpha - aD^{\alpha}([X], Y, \mathcal{G}, \rho, \alpha, [\beta]) \tag{9.21}$$

$$[\beta^{new}] := [\beta] - bD([X], Y, \mathcal{G}, \rho, \alpha, [\beta], [\gamma] - [U]), \tag{9.22}$$

where a and b are appropriate learning rates, and

$$D^{\alpha}([X], Y, \mathcal{G}, \rho, \alpha, [\beta]) := \sum_{i=1}^{n} \{S(\alpha + [x_i]\mathcal{G}[\beta]) - y_i)[x_i]^T\}$$

$$D([X], Y, \mathcal{G}, \rho, \alpha, [\beta], \theta) := \mathcal{G}\left(\sum_{i=1}^{n} \{S(\alpha + [x_i]\mathcal{G}[\beta]) - y_i)[x_i]^T\} + \rho[\beta] - \rho\theta\right)$$

until $\|[\beta^{new}] - [\beta]\|_2^2$ and $|\alpha^{k+1} - \alpha^k|$ are smaller than desired algorithm's thresholds,

or justify the relative criterion instead-specially in the case of applying line search backtracking to find the optimal learning rates.

Proof. The convergence of above algorithm is guaranteed by the fact that he input of $-\log(S(\cdot))$ and $-\log(1-S(\cdot))$ (which are already convex) is linear (affine function) in α and $[\beta]$, thus $f(\alpha,[\beta])$ is convex.
For $[\beta]$-update we have:

$$(\alpha^{\text{new}},[\beta^{\text{new}}]) := \arg\min_{\alpha,[\beta]} \Big(f(\alpha,[\beta]) + \frac{\rho}{2}([\beta]-[\gamma]+[U])^T \mathcal{G}([\beta]-[\gamma]+[U]) \Big)$$

It is straightforward to differentiate with respect to $\alpha,[\beta]$ which completes the derivation of a gradient descent algorithm. □

It is well known that gradient descent is an appropriate start for a convex optimization problem; however, after multiple steps, the updated values can fall into a neighborhood of the solution and continue toward it slowly. On the other hand, if the current value is in a closed neighborhood of the solution, the Newton method-Hessian matrix- converges faster. A combination of Gradient descent to approach a neighborhood of the solution, then a Newton method that converges quickly is suggested. Denote $Q = (q_1^T,\ldots,q_n^T)^T$.

Proposition 2. *A gradient descent b–update for the optimization problem (3.2) in a Logistic regression is:*

$$b^{new} := b - aD(Q,Y,\mathcal{F},\rho,b^k), \tag{9.23}$$

with

$$D(Q,Y,\mathcal{F},\rho,b) = \mathcal{F}\{Q^T(S(Q\mathcal{F}b)-Y)+\rho(b-\theta)\},$$

where a is appropriate learning rate, $S(Q\mathcal{F}b)$ is element–wise Sigmoid function, and $\theta = (\alpha, ([\gamma^k] - [U^k])^T)^T$. The iterations continue until $\|b^{new} - b\|_2^2$ is smaller than the desired algorithm's threshold. Alternatively they continue until relative stopping criteria is justified. The relative stopping criteria is a more reasonable approach in case of line search backtracking in order to update learning rates.

On the other hand, a b–update with the Newton method is

$$b^{new} := b - a' H(Q, Y, \mathcal{F}, \rho, b^k)^{-1} D(Q, Y, \mathcal{F}, \rho, b) \tag{9.24}$$

where a' is the appropriate learning rate and

$$H(Q, Y, \mathcal{F}, \rho, b) = \mathcal{F}\{Q^T(diag(S(Q\mathcal{F}b))(1 - S(Q\mathcal{F}b))Q\mathcal{F} + \rho I_{mp+1}\}.$$

Note that $diag(S(Q\mathcal{F}b))$ is a diagonal matrix of element-wise applying Sigmoid function on $Q\mathcal{F}b$.

Proof. The convergence of the algorithm is guaranteed by the fact that input of $-\log(S(\cdot))$ and $-\log(1 - S(\cdot))$ (which are already convex functions) are linear (affine function) in b, thus, $f(b)$ is convex. We have:

$$b := \arg\min_b \left(f(b) + \frac{\rho}{2}(b - \theta)^T \mathcal{F}(b - \theta) \right).$$

It is straightforward to differentiate twice with respect to b, which completes the derivation of the gradient descent and Newton algorithm above. □

9.2.3 fMRI applications

The application of the sparse functional logistic regression can be found in the fMRI experiments. Analyzing the resting-state fMRI data for a diseased classification is one of them. For example: Alzheimer +1, Healthy -1 is a binary variable. Through such

an analysis, we can select the ROIs associated with such a disease. In addition to this application, we can use such a model in the following paradigm in a blocked designed analysis of a task fMRI experiment. If stimuli are presented for all subjects at the same time, we can analyze the associated ROI activities when the stimulus is on as +1, and otherwise -1. Then, it is possible to select the ROIs associated with such stimuli.

While the code implantation and the estimation consistency can be straightforward, it seems that the selection consistency is not already developed for the classical multivariate version of such a model, so verifying the oracle property of the functional version would be a challenge.

9.3 Sparse function-on-function regression

We can consider the situations where the response value is a function and the predictors are multivariate functions, but only a few functional predictors affect the response. i.e., a random function Y and random functions $X^j \in \mathcal{H}^j$ have the following relation,

$$Y = \sum_{j \in \mathscr{A}} \beta^j(X^j) + \epsilon, \tag{9.25}$$

where $\mathscr{A} \subseteq \{1, \ldots, p\}$ is an unknown active set of indices involved in this regression model.

Assume that we have a random sample of size n from the model (9.25). Then, we propose the following objective function to estimate the unknown operator β and the active set $\mathscr{A}$.

$$L(\beta; \lambda_{1n}) = \frac{1}{2} E_n(Y - \beta(X))^2 + \lambda_{1n} \textstyle\sum_{j=1}^{p} \|\beta^j\|_F, \quad \beta \in \mathcal{H}. \tag{9.26}$$

9.3.1 Iterative algorithm

The following can be an algorithm to solve the problem.

$$[\beta^j]^{(\text{new})} = \frac{1}{2\alpha\lambda + \gamma_j} S^{\mathcal{H}^j}_{\lambda(1-\alpha)}(U^j + \gamma_j[\beta^j]^{(\text{old})}), \quad j = 1, \ldots, p. \tag{9.27}$$

where $U^j = -\nabla L([\beta^j]^{(\text{old})})$ and the norm is Frobenius norm and soft treshholding rule is with respect to this norm.

9.3.2 fMRI application

An important application can be analyzing random event-related designs in a task fMRI experiment. For example, the response value can be taken as a binary time series of stimuli that are on or off randomly for each individual. Through such an analysis, we can select ROIs that are associated with the stimulus.

9.4 Standardization

Suppose the scalar on function penalized regression (3.2) through ADMM. Consider Theorem 3. The following algorithm estimates the standard deviations of the norm of the coefficient curves in the final active set.

9.4.1 Algorithm

The only update that would result in different norms for each functional coefficient is the second update, γ–update. Hence we can estimate the standard deviation of the norm of $[\beta]$ and finally threshold it in the γ^j update for each j. The third update would affect all functional coefficients' standard deviations with the same amount; hence, we ignore it. Denote:

$$[\beta_i^{\text{new}}] = ([\tilde{X}_i][\tilde{X}_i]^{\mathsf{T}} + n\rho I_M)^{-1}([\tilde{X}_i]Y_i + n\rho([\gamma] - [U])) \qquad i = 1, \cdots, n \tag{9.28}$$

For each functional coefficient $j = 1, \cdots, p$, we can estimate the standard deviation of $\|\beta^j\|$ based on variation of $\|\beta_i^j\|$ for $i = 1, \cdots, n$.

9.4.2 fMRI Applications

Such estimations for the standard deviations of the norm of the estimated functions in the active set can be used to rank the output of the penalized group LASSO regression with respect to the ratio of the norms of estimations and their standard deviations. This can be used to rank the importance of final ROIs or voxels. On the other hand, it can be used to remove the estimated curves that are not already removed through the sparse regression but have significant noise and weak signal. Except for theoretical verification of such a method, there is not much challenge in the implementation and code.

CHAPTER 10: CONCLUSION

We propose new methods for scalar-on-function regression with the functional predictor selection and the estimation of smooth coefficient functions when the predictors are multivariate functional data. We derive the algorithm for the implementation and develop the consistency of the methods by showing its oracle property. The simulation and real data application show the effectiveness of the methods with the superior performance of the proposed penalized methods over the functional regression model with the OLS. Furthermore, the proposed methods provide higher accuracy and low complexity of the model in the fMRI study. It shows that there is an urgent need in the fields of medical sciences and other related areas.

The manuscript also has a potential impact on the field of statistical research for more advanced sparse functional models. Considering that there is not enough investigation into sparse modeling of multivariate functional data, the computation algorithm derived in this thesis will pave the way to develop other novel sparse methods. In addition, the methods can be extended to the nonlinear regression model via the reproducing kernel Hilbert space (RKHS). Since the theoretical justification is constructed under the infinite-dimensional setting, the extension on the RKHS can adopt the results from this thesis. Furthermore, the proposed methods are based on groups such that a single functional predictor forms a group. Hence, it can be easily extended to the sparse models where multiple functional predictors form a group. For example, instead of averaging out fMRI signals of voxels over the brain regions, we would keep the original data and apply the MFG methods with groups formed by each region's voxels activities. Then, we might figure out a new foundation that has been removed in the masking procedure.

In addition, extensions of the proposed methods can be applied to a wide range of research areas. For example, extending the result to binary response values can have applications in block design fMRI experiments where a stimulus status is on or off for all subjects simultaneously. This model can then select ROIs or voxels associated with the stimulus. Furthermore, such an extension can be used to classify the ROI or voxels associated with a disease in a case-control study. Standardizing the results by estimating the standard deviation of the norm of the estimated coefficient curves can lead to a rank analysis of the ROI or voxels in the final active set of the sparse models. Such a rank analysis determines the importance of each ROI or voxel in the final active set and reveals the weak signal and large noise curves. Aside from these two potential extensions and their fMRI applications, extension to functional response values can have an essential application in event-related design task fMRI experiment data analysis where response values are a binary time series of a stimulus status that is randomly on or off for each subject in time.

REFERENCES

[1] J. Ramsay and B. Silverman, *Functional Data Analysis, 2nd Ed.* Springer-Verlag, 2005.

[2] F. Yao, H. G. Müller, and J. Wang, "Functional data analysis for sparse longitudinal data," *Journal of American Statistical Association*, vol. 100, pp. 577–590, 2005.

[3] F. Yao, H. G. Müller, and J. Wang, "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, vol. 33, pp. 2873–2903, 2005.

[4] L. Horváth and P. Kokoszka, *Inference for Functional Data with Applications*. Springer, 2012.

[5] J. Wang, J. Chiou, and H.-G. Müller, "Functional data analysis," *Annual Review of Statistics and Its Application*, vol. 3, pp. 257–295, 2016.

[6] J. Chiou, Y. Yang, and Y. Chen, "Multivariate functional linear regression and prediction," *Journal of Multivariate Analysis*, vol. 146, pp. 301 – 312, 2016.

[7] C. Happ and S. Greven, "Multivariate functional principal component analysis for data observed on different (dimensional) domains," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 649–659, 2018.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal Royal Statistics Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, vol. B, no. 1, p. 301â320, 2005.

[10] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[11] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[12] H. Zou and H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Annals of Statistics*, vol. 37, no. 4, p. 1733â1751, 2009.

[13] G. M. James, J. Wang, J. Zhu, *et al.*, "Functional linear regression thatâs interpretable," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2083–2108, 2009.

[14] R. Blanquero, E. Carrizosa, A. Jiménez-Cordero, and B. Martín-Barragán, "Variable selection in classification for multivariate functional data," *Information Sciences*, vol. 481, pp. 445–462, 2019.

[15] J. Pannu and N. Billor, "Robust group-lasso for functional regression model," *Communications in Statistics - Simulation and Computation*, vol. 46, no. 5, pp. 3356–3374, 2017.

[16] P. A. Bandettini, *fMRI*. MIT Press, 2020.

[17] R. Tibshirani and L. Wasserman, "Sparsity, the lasso, and friends," *Statistical Machine Learning*, Spring 2017.

[18] M. Efroymson, "Stepwise regression: a backward and forward look," *Eastern Regional Meetings of the Institute of Mathematical Statistics*, 1966.

[19] N. Draper and H. Smith, *Applied Regression Analysis*. Wiley, 1966.

[20] B. Efron, T. Hastie, J. I., and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, p. 407â499, 2004.

[21] R. J. Tibshirani, "A general framework for fast stagewise algortihms," *Journal of Machine Learning Research*, vol. 16, p. 2543â2588, 2015.

[22] R. J. Tibshirani, "The lasso problem and uniqueness," *Electronic Journal of Statistics*, vol. 7, pp. 1456–1490, 2013.

[23] M. Osborne, B. Presnell, and B. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, p. 2389â404, 2000.

[24] M. Osborne, B. Presnell, and B. Turlach, "On the lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, p. 319â33, 2000.

[25] S. van de Geer and P. Buhlmann, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, p. 1360â1392, 2009.

[26] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using 'l1-constrained quadratic programming (lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, p. 2183â2202, 2009.

[27] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over 'q-ballâs," *IEEE Transactions on Information Theory*, vol. 57, no. 10, p. 6976â6994, 2011.

[28] D. Foster and E. George, "The risk inflation criterion for multiple regression," *The Annals of Statistics*, vol. 22, no. 4, p. 1947â1975, 1994.

[29] S. Chatterjee, "Assumptionless consistency of the lasso," *arXiv*, p. 1303.5817, 2013.

[30] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *The Annals of Statistics*, vol. 28, no. 5, p. 1356â1378, 2000.

[31] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, p. 1418â1429, 2006.

[32] N. Meinshausen, "Relaxed lasso," *Computational Statistics Data Analysis*, vol. 52, p. 374â393, 2007.

[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, vol. 67, no. 2, p. 301â320, 2005.

[34] S. Bakin, "Adaptive regression and model selection in data mining problems," *PhD thesis*, 1999.

[35] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society*, vol. Series B 68, no. 1, p. 49â67, 2006.

[36] C. Zhang and Y. Xiang, "On the oracle property of adaptive group lasso in high-dimensional linear models," *Stat Papers*, vol. 57, p. 249â265, 2016.

[37] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[38] M. Yuan and Y. Lin, "On the non-negative garrotte estimator," *Journal of The Royal Statistical Society*, vol. Series B 69, no. 2, p. 143â161, 2007.

[39] P. Zhao and B. Yu., "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, p. 2541â2563, 2006.

[40] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, p. 1418â1429, 2006.

[41] A. W. V. der Vaart, *Asymptotic Statistics*. Cambridge Univ. Press, 1998.

[42] S. v. d. G. Lukas Meier and P. BÃŒhlmann, "The group lasso for logistic regression," *Journal R. Statist. Soc.*, vol. B, no. 70, p. 53â71, 2008.

[43] L. Meier, S. van de Geer, and P. Buhlmann., "The group lasso for logistic regression," *Technical Report*, vol. 131, p. 1418â1429, 2006.

[44] J. B. Conway, *A Course in Functional Analysis, Second Edition*. Springer, 1990.

[45] T. Hsing and R. Eubank, *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

[46] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[47] J. Song and B. Li, "Nonlinear and additive principal component analysis for functional data," *Journal of Multivariate Analysis*, vol. 181, 2021.

[48] B. Li and J. Song, "Dimension reduction for functional data based on weak conditional moments," *submitted*, 2018.

[49] Y. Yang and H. Zou, "A fast unified algorithm for solving group-lasso penalize learning problems," *Statistics and Computing*, vol. 25, no. 6, pp. 1129–1141, 2015.

[50] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society*, vol. 74, no. 2, pp. 245–266, 2012.

[51] C. Baker, "Joint measures and cross-covariance operators," *Transactions of the American Mathematical Society*, vol. 186, p. 273â289, 1973.

[52] N. Tzourio-Mazoyer and et al., "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain.," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[53] K. Hilger and et al., "Predicting intelligence from brain gray matter volume," *Brain Structure and Function*, vol. 225, no. 273-89, p. 2111â2129, 2020.

[54] L. Xiao, J. Stephen, and et al., "A manifold regularized multi-task learning model for iq prediction from two fmri paradigms," *IEEE Transactions on Biomedical Engineering*, vol. 67, 2020.

[55] J. Maldjian and et al., "An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fmri data sets," *Neuroimage .*, vol. 19, no. 3, pp. 1233–1239, 2003.

[56] Y. B. Yoon and et al., "Brain structural networks associated with intelligence and visuomotor ability," *Frontiers in Human Neuroscience*, vol. 7, no. 1, p. 44, 2017.

[57] N. Goriounova and H. Mansvelder, "Genes, cells and brain areas of intelligence," *Frontiers in Human Neuroscience*, vol. 13, p. 44, 2019.

[58] Y. Yang and H. Zou, "A fast unified algorithm for solving group-lasso penalized learning problems," *Statistics and Computing*, vol. 25, no. 6, 2014.

[59] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[60] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *Annals of statistics*, pp. 1356–1378, 2000.

APPENDIX A: PROOFS

Proof of Lemma 4 The representation of $[\hat{\Gamma}_{XX}]$ can be shown by the relation between the two following equations.

$$\langle f, \hat{\Gamma}_{XX} g\rangle_{\mathcal{H}} = [f]_{\mathcal{B}}^{\mathsf{T}} G[X_{1:n}]_{\mathcal{B}} Q[\hat{\Gamma}_{XX}]_{\mathcal{B}}[g]_{\mathcal{B}} = E_n(\langle f, X - E_n X\rangle_{\mathcal{H}}\langle g, X - E_n X\rangle_{\mathcal{H}}),$$
$$\langle f, \hat{\Gamma}_{XX} g\rangle_{\mathcal{H}} = [f]_{\mathcal{B}}^{\mathsf{T}}[\hat{\Gamma}_{XX}][g]_{\mathcal{B}},$$

for any $f, g \in \mathcal{H}$. The second equation can be shown as following. For any $\beta \in \mathcal{H}$,

$$\begin{aligned}\hat{\Gamma}_{YX}\beta &= E_n\{(Y - E_n Y) \otimes (X - E_n X)\}\beta = E_n\{(Y - E_n Y)\langle X - E_n X, \beta\rangle_{\mathcal{H}}\} \\ &= E_n\{(Y - E_n Y)[X - E_n X]^{\mathsf{T}} G[\beta]\}.\end{aligned}$$

We can also see that $\hat{\Gamma}_{XY} = n^{-1}[\tilde{X}_{1:n} Y]$. □

Lemma 15. *Take $x, y \in \mathbb{R}^m$ where y is known.*

$$\arg\min_x \left(\frac{1}{2}\|x - y\|^2 + \lambda\|x\|\right) = S_\lambda(y), \tag{A.1}$$

where $S_\lambda(y) := 1_{\{\|y\|>\lambda\}}\left(1 - \frac{\lambda}{\|y\|}\right)_+ y$ is the block soft threshold operator in real space.

Proof of Lemma 15. Observe that

$$\arg\min_x \left(\frac{1}{2}(x - y)^{\mathsf{T}}(x - y) + \lambda\|x\|\right) = \arg\min_x \left(\frac{1}{2}(x^{\mathsf{T}} x - 2x^{\mathsf{T}} y) + \lambda\|x\|\right).$$

To satisfy the KarushâKuhnâTucker (KKT) stability condition, the derivative of the above objective function with respect to x must be equal to zero. If the derivative does not exist, the subdifferential must include zero. The derivative is $x - y + \lambda s_x$, where s_x is the subdifferential of $\|x\|$ at x.

If $x \neq 0$, $s_x = x/\|x\|$ and the KKT condition gives

$$x(1 + \lambda/\|x\|) = y.$$

Compute the $\|y\|$ in the preceding equation and solve for $\|x\|$. Plugging it back into the equation gives us,

$$x = (1 - \lambda/\|y\|)y.$$

The condition $x \neq 0$ is equivalent to $\|y\| > \lambda$. On the other hand, $x = 0$ is equivalent to $0 \in -y + \lambda s_x$, or $y \in \lambda s_x$. In this case $s_x = \{z \in \mathbb{R}^m | \|z\| \leq 1\}$. Therefore, $\|y\|^2 \leq \lambda^2$ which completes the proof. □

Proof of Theorem 3.

1) β-update.

Consider the objective function for β in (4.5). After removing the constant terms with respect to β, with the help of Lemma 4, we have

$$\begin{aligned}
[\beta^{\text{new}}] &:= \arg\min_{\beta} \left(f(\beta) + \frac{\rho}{2}([\beta] - [\gamma] + [U])^{\mathsf{T}}([\beta] - [\gamma] + [U]) \right) \\
&= \arg\min_{\beta} \left(\frac{1}{2n}([\beta]^{\mathsf{T}}[\tilde{X}_{1:n}][\tilde{X}_{1:n}]^{\mathsf{T}}[\beta] - 2[\beta]^{\mathsf{T}}[\tilde{X}_{1:n}]Y) + \frac{\rho}{2}\{[\beta]^{\mathsf{T}}[\beta] - 2[\beta]^{\mathsf{T}}([\gamma] - [U])\} \right).
\end{aligned}$$

Differentiate with respect to β, and set the derivative equal to zero to satisfy the KKT conditions. The result is:

$$n^{-1}[\tilde{X}_{1:n}][\tilde{X}_{1:n}]^{\mathsf{T}}[\beta] - n^{-1}[\tilde{X}_{1:n}]Y + \rho([\beta] - ([\gamma] - [U])) = 0.$$

Solve for β, which completes the derivation. Note that the result is similar to the functional ridge regression.

2) γ-update.

Similarly, if we remove the constant terms with respect to γ and expand the objective function for γ, we have

$$[\gamma^{\text{new}}] := \arg\min_{\gamma} \left(g(\gamma) + \frac{\rho}{2}([\beta^{\text{new}}] - [\gamma] + [U])^{\mathsf{T}}([\beta^{\text{new}}] - [\gamma] + [U]) \right) =$$
$$\arg\min_{\gamma} \left(\sum_{j=1}^{p} \{ \lambda([\gamma^j]^{\mathsf{T}}[\gamma^j])^{\frac{1}{2}} + \frac{\rho}{2}([\gamma^j] - ([(\beta^j)^{\text{new}}] + [U^j]))^{\mathsf{T}}([\gamma^j] - ([(\beta^j)^{\text{new}}] + [U^j])) \} \right).$$

Note that the objective function is now additive, which allows us to optimize γ for each γ^j, $j = 1 \ldots, p$. Thus, the above optimization is equivalent to

$$[(\gamma^j)^{\text{new}}] := \arg\min_{\gamma^j} \left(\lambda([\gamma^j]^{\mathsf{T}}[\gamma^j])^{\frac{1}{2}} + \frac{\rho}{2}([\gamma^j] - ([(\beta^j)^{\text{new}}] + [U^j]))^{\mathsf{T}}([\gamma^j] - ([(\beta^j)^{\text{new}}] + [U^j])) \right),$$

for $j = 1, \ldots p$. Applying Lemma 15 completes the proof. □

Lemma 16. *For $x, y \in \mathbb{R}^m$ where y is known and a, b are constants*

$$\arg\min_{x} \left(\frac{1}{2}(x-y)^{\mathsf{T}}(x-y) + a(x^{\mathsf{T}}x)^{\frac{1}{2}} + \frac{b}{2}x^{\mathsf{T}}x \right) = \frac{1}{b+1} S_a(y).$$

Proof of Lemma 16.

The proof is similar to that of lemma 15. The only difference is the derivative of the objective function. It is $x - y + as + bx$, where s is the subdifferential. The rest of the proof is straightforward. If $x \neq 0$ we see that $x(1+b+\frac{a}{\|x\|}) = y$. Taking norm $\|.\|$ from both sides, solving for $\|x\|$, and plugging it back, we would have $x = (\frac{1}{1+b})(1 - \frac{a}{\|y\|})y$. Note that this is only possible when $\|x\| > 0$, which means $\|y\| > a$. If $x = 0$, it results in $0 \in -y + as$, or $y \in as$. Since in this case $s = \{[Z] | Z \in \mathbb{R}^m \& \|Z\| \leq 1\}$, $\|y\| \leq a$, which completes the derivation above. □

Proof of Theorem 4. The proof is a direct result of the combination of Theorem 3

and Lemma 16. □

Lemma 17. *Assume that Γ_{XX} is a positive definite operator and when n approaches infinity, λ_n approaches zero slower than the rate at which $\sqrt{n}$ approaches infinity. Then, $\|(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\Gamma_{XX} - (\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX}\|_{\mathcal{H}} = O_p(\lambda_n^{-1}n^{-1/2})$, and $\|(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\hat{\Gamma}_{XX} - (\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX}\|_{\mathcal{H}} = O_p(\lambda_n^{-1}n^{-1/2})$, where $\|\cdot\|_{\mathcal{H}}$ is the operator norm.*

Proof of Lemma 17. Note that $\Gamma_{XX}(\Gamma_{XX} + \lambda_n I) = I - \lambda_n(\Gamma_{XX} + \lambda_n I)^{-1}$ and $(\hat{\Gamma}_{XX} + \lambda_n I)\hat{\Gamma}_{XX} = I - \lambda_n(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}$. Therefore,

$$(\hat{\Gamma}_{XX} + \lambda_n I)^{-1} - (\Gamma_{XX} + \lambda_n I)^{-1} = (\hat{\Gamma}_{XX} + \lambda_n I)^{-1}(\Gamma_{XX} - \hat{\Gamma}_{XX})(\Gamma_{XX} + \lambda_n I)^{-1}.$$

To be specific, if we add and subtract $\lambda_n(\hat{\Gamma}_{XX}+\lambda_n I)^{-1}(\Gamma_{XX}+\lambda_n I)^{-1}$ in the left-hand side of the above equation, we can easily derive the right-hand side of the equation. In addition, we have

$$\begin{aligned}&(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\Gamma_{XX} - (\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX} \\ &= (\hat{\Gamma}_{XX} + \lambda_n I)^{-1}(\Gamma_{XX} - \hat{\Gamma}_{XX})(\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX}.\end{aligned} \tag{A.2}$$

Note that $(\hat{\Gamma}_{XX} + \lambda_n I)^{-1} = (\Gamma_{XX} + O_p(n^{-1/2}) + \lambda_n I)^{-1}$ by Lemma 7. Thus, its norm is $\|(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\|_{\mathcal{H}} = O_p(\lambda_n^{-1})$. By Lemma 7, $\|(\Gamma_{XX} - \hat{\Gamma}_{XX})\|_{\mathcal{H}} = O_p(n^{-1/2})$. The norm of product of the last two parentheses is bounded by 1. Hence, $\|(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\Gamma_{XX} - (\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX}\|_{\mathcal{H}} = O_p(\lambda_n^{-1}n^{-1/2})$.

For the second convergence rate, note that

$$
\begin{aligned}
&(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\hat{\Gamma}_{XX} - (\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\Gamma_{XX} \\
&= (\hat{\Gamma}_{XX} + \lambda_n I)^{-1}(\hat{\Gamma}_{XX} - \Gamma_{XX}) = O_p(\lambda_n^{-1} n^{-1/2}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\|(\hat{\Gamma}_{XX}+\lambda_n I)^{-1}\hat{\Gamma}_{XX} - (\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX}\|_{\mathcal{H}} \\
&\leq \|(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\hat{\Gamma}_{XX} - (\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\Gamma_{XX}\|_{\mathcal{H}} \\
&+ \|(\hat{\Gamma}_{XX} + \lambda_n I)^{-1}\Gamma_{XX} - (\Gamma_{XX} + \lambda_n I)^{-1}\Gamma_{XX}\|_{\mathcal{H}} \\
&= O_p(\lambda_n^{-1} n^{-1/2}).
\end{aligned}
$$

□

Proof of Lemma 8.

The following proof is similar to the proof mentioned in (37) which considers a different penalty term that is square of the group LASSO penalty. Then, they proved the consistency by stating that the solution path of the group LASSO will be the same. Instead, we consider a different optimization problem $\tilde{M}_n(\cdot)$ proposed below, which directly leads to the consistency of multivariate functional group LASSO.

Denote $\tilde{\beta}_n^J$ as the unique minimizer of the following objective function.

$$
\tilde{M}_n(\alpha) = \frac{1}{2}\hat{\Gamma}_{YY} - \hat{\Gamma}_{YX^J}(\alpha) + \frac{1}{2}\langle \alpha, \hat{\Gamma}_{X^J X^J}(\alpha)\rangle + \frac{\lambda_n}{2}\sum_{j\in J}\frac{\|\alpha^j\|_{\mathcal{H}^j}^2}{\|\beta^j\|_{\mathcal{H}^j}}, \quad \alpha \in \mathcal{H},
$$

where β^j is the j-th functional component of β^J in the population model. $\tilde{\beta}_n^J$ has a closed-form solution similar to the solution of a functional predictor ridge regression

$$
\tilde{\beta}_n^J = (\hat{\Gamma}_{X^J X^J} + \lambda_n D)^{-1}(\hat{\Gamma}_{X^J Y}),
$$

where D is a diagonal operator, $diag((\cdot)/\|\beta^j\|)$. We can replace $\hat{\Gamma}_{X^J Y}$ by the following expression, after adding and subtracting $\hat{\Gamma}_{X^J X^J}(\beta^J)$.

$$\tilde{\beta}_n^J = (\hat{\Gamma}_{X^J X^J} + \lambda_n D)^{-1}(\hat{\Gamma}_{X^J X^J}\beta^J + \hat{\Gamma}_{X\epsilon}), \tag{A.3}$$

where $\hat{\Gamma}_{X\epsilon}$ is the empirical covariance operator between observed functional data X and the population error, $\epsilon = Y - \langle X, \beta\rangle = Y - \langle X^J, \beta^J\rangle$. D is a self-adjoint operator, and $\|\beta^j\|_{\mathcal{H}} \neq 0$ for all $j \in J$ by the definition of the population active set J. This means there are positive constants $D_{\min} = 1/\max_{j\in J}\|\beta^j\|_{\mathcal{H}}$ and $D_{\max} = 1/\min_{j\in J}\|\beta^j\|_{\mathcal{H}}$ such that $D_{\max} I \succcurlyeq D \succcurlyeq D_{\min} I$. The closed-form solution (A.3) can be broken down into multiple terms. One of the terms is

$$(\hat{\Gamma}_{X^J X^J} + \lambda_n D)^{-1}(\hat{\Gamma}_{X\epsilon}). \tag{A.4}$$

Applying the same technique in the proof of Lemma 17 and using the result of Lemma 7, we can see that $\|\hat{\Gamma}_{X^J X^J} + \lambda_n D^{-1}\|_{\mathcal{H}} \leq D_{\min}^{-1}\lambda_n^{-1}$, and

$$(\hat{\Gamma}_{X^J X^J} + \lambda_n D)^{-1}(\hat{\Gamma}_{X\epsilon}) = O_p(n^{-1/2}\lambda_n^{-1}).$$

Hence, we have

$$\begin{aligned}
\tilde{\beta}_n^J - \beta^J &= (\hat{\Gamma}_{X^J X^J} + \lambda_n D)^{-1}(\hat{\Gamma}_{X^J X^J}\beta^J + \hat{\Gamma}_{X\epsilon}) - \beta^J \\
&= (\hat{\Gamma}_{X^J X^J} + \lambda_n D)^{-1}(\hat{\Gamma}_{X^J X^J}\beta^J) - (\Gamma_{X^J X^J} + \lambda_n D)^{-1}\Gamma_{X^J X^J}\beta^J \\
&+ (\Gamma_{X^J X^J} + \lambda_n D)^{-1}\Gamma_{X^J X^J}\beta^J - \beta^J + O_p(n^{-1/2}\lambda_n^{-1})
\end{aligned} \tag{A.5}$$

The first two terms of the last equation in (A.5) is $O_p(n^{-1/2}\lambda_n^{-1})$ by Lemma 17. By using $(\Gamma_{X^J X^J} + \lambda_n D)^{-1}\Gamma_{X^J X^J} = I - \lambda_n(\Gamma_{X^J X^J} + \lambda_n D)^{-1} D$, we can simplify the third

and fourth terms of (A.5) as

$$(\Gamma_{X^J X^J} + \lambda_n D)^{-1}\Gamma_{X^J X^J}\beta^J - \beta^J = (-\lambda_n(\Gamma_{X^J X^J} + \lambda_n D)^{-1}D)\beta^J. \tag{A.6}$$

Consequently, we have

$$\tilde{\beta}_n^J - \beta^J = (-\lambda_n(\Gamma_{X^J X^J} + \lambda_n D)^{-1}D)\beta^J + O_p(n^{-1/2}\lambda_n^{-1}). \tag{A.7}$$

Now, we show the norm of $\lambda_n(\Gamma_{X^J X^J}+\lambda_n D)^{-1}D$ is $O_p(\sqrt{\lambda_n}+n^{-1/2}\lambda_n^{-1})$. Let $h^J \in \mathcal{H}^J$ be the element in the assumption such that $\beta^J = \Gamma^{1/2}_{X^J X^J} h^J$. Then,

$$\begin{aligned}
&\|\lambda_n(\Gamma_{X^J X^J} + \lambda_n D)^{-1}D\beta^J\|^2_{\mathcal{H}^J} \\
&= \lambda_n^2\langle\beta^J, D(\Gamma_{X^J X^J} + \lambda_n D)^{-2}D\beta^J\rangle_{\mathcal{H}^J} \\
&\leq \lambda_n^2 D^2_{\max}\langle\beta^J, (\Gamma_{X^J X^J} + \lambda_n D_{\min} I)^{-2}\beta^J\rangle_{\mathcal{H}^J} \\
&\leq \lambda_n D^2_{\max} D^{-1}_{\min}\langle\beta^J, (\Gamma_{X^J X^J} + \lambda_n D_{\min} I)^{-1}\beta^J\rangle_{\mathcal{H}^J} \\
&= \lambda_n D^2_{\max} D^{-1}_{\min}\langle\Gamma^{1/2}_{X^J X^J} h^J, (\Gamma_{X^J X^J} + \lambda_n D_{\min} I)^{-1}\Gamma^{1/2}_{X^J X^J} h^J\rangle_{\mathcal{H}^J} \\
&\leq \lambda_n D^2_{\max} D^{-1}_{\min}\|h^J\|^2_{\mathcal{H}}.
\end{aligned}$$

The third line of the above equation is valid because $\|\Gamma_{X^J X^J}+\lambda_n D_{\min} I\|_{\mathcal{H}^J} \geq \lambda_n D_{\min}$. Combining the results above, we have

$$\|\tilde{\beta}_n^J - \beta^J\|_{\mathcal{H}} = O_p(\sqrt{\lambda_n} + n^{-1/2}\lambda_n^{-1}).$$

Now, let's compare $\tilde{\beta}_n^J$ and β_n^J where β_n^J is the solution to the optimization problem of $M_n(\alpha)$. Consider the following equation.

$$M_n(\alpha) - \tilde{M}_n(\alpha) = \lambda_n \sum_{j\in J}\left(\|\alpha^j\|_{\mathcal{H}^j} - \frac{\|\alpha^j\|^2_{\mathcal{H}^j}}{2\|\beta^j\|_{\mathcal{H}^j}}\right). \tag{A.8}$$

The partial Fréchet derivative of the equation (A.8) with respect to α^i for an $i \in J$ is

$$D_{\alpha^i}(M_n(\alpha) - \tilde{M}_n(\alpha)) = \lambda_n \left(\frac{\langle \alpha^i, \cdot \rangle_{\mathcal{H}^i}}{\|\alpha^i\|_{\mathcal{H}^i}} - \frac{\langle \alpha^i, \cdot \rangle_{\mathcal{H}^i}}{\|\beta^i\|_{\mathcal{H}^i}} \right). \tag{A.9}$$

Since β^J are nonzero, (A.9) is continuously differentiable around β^J, and $D_{\alpha^i}\tilde{M}_n(\tilde{\beta}_n^J)) = 0$, we have

$$\|D_{\alpha^i} M_n(\tilde{\beta}_n^J)) - 0\| = \lambda_n \left\| \frac{\langle \tilde{\beta}_n^i, \cdot \rangle_{\mathcal{H}^i}}{\|\tilde{\beta}_n^i\|_{\mathcal{H}^i}} - \frac{\langle \tilde{\beta}_n^i, \cdot \rangle_{\mathcal{H}^i}}{\|\beta^i\|_{\mathcal{H}^i}} \right\|,$$

where the $\|\cdot\|$ is the operator norm. In addition, since $\beta^i \neq 0$ for $i \in J$, it can be easily shown that

$$\|D_{\alpha^i} M_n(\tilde{\beta}_n^J)) - 0\|_{\mathcal{H}^i} \leq C\lambda_n \|\beta^J - \tilde{\beta}_n^J\|_{\mathcal{H}^J},$$

for some constant $C > 0$. Thus, we have

$$\|D_{\alpha^i} M_n(\tilde{\beta}_n^J))\|_{\mathcal{H}^i} = \lambda_n O_p(\lambda_n^{1/2} + n^{-1/2}\lambda_n^{-1}). \tag{A.10}$$

Now, since M_n is strictly convex near the true β^J, its second-order Fréchet derivative has a lower bound. Consequently, we have

$$M_n(\alpha^J) \geq M_n(\tilde{\beta}_n^J) + \langle D_{\alpha^J} M_n(\tilde{\beta}_n^J), (\alpha^J - \tilde{\beta}_n^J) \rangle_{\mathcal{H}^J} + C'\lambda_n \|\alpha^J - \tilde{\beta}_n^J\|_{\mathcal{H}^J}^2,$$

for some $C' > 0$. Suppose that α^J is near $\tilde{\beta}_n^J$ and let $\eta_n = \|\alpha^J - \tilde{\beta}_n^J\|_{\mathcal{H}^J}^2$ which tends to zero. Subsequently, we can rewrite the lower bound such that

$$M_n(\alpha^J) \geq M_n(\tilde{\beta}_n^J) + \eta_n \lambda_n O_p(\sqrt{\lambda}_n + n^{-1/2}\lambda_n^{-1}) + C'\lambda_n \eta_n^2, \tag{A.11}$$

If the last term is tending to zero slower than the second term, we can conclude

that all minima of $M_n(\cdot)$ are inside the ball $\{\alpha^J : \|\alpha^J - \tilde{\beta}_n^J\|_{\mathcal{H}^J}^2 < \eta\}$ with probability tending to one. This is because $M_n(\cdot)$, on the edge of the ball, takes values greater the ones inside the ball. i.e., the global minimum of $M_n(\cdot)$ is at most η_n away from $\tilde{\beta}_n^J$. Thus, the necessary condition for the proof is $\eta_n \lambda_n^{3/2} = o(\lambda_n \eta_n^2)$ and $n^{-1/2}\eta_n = o(\lambda_n \eta_n^2)$. Alltogether, we have the consistency results if η_n converges to zero slower than $\lambda_n^{1/2} + n^{-1/2}\lambda_n^{-1}$. □

Proof of Theorem 5. We rewrite the multivariate functional group LASSO objective function (3.2) as,

$$\hat{M}_n(\alpha) = \frac{1}{2}\hat{\Gamma}_{YY} - \hat{\Gamma}_{YX}\alpha + \frac{1}{2}\langle \alpha, \hat{\Gamma}_{XX}\alpha \rangle_{\mathcal{H}} + \lambda_n \sum_{j=1}^{p} \|\alpha^j\|_{\mathcal{H}^j}.$$

Denote a minimizer of $\hat{M}_n(\cdot)$ by $\hat{\beta}_n$. Since it is a convex function, it has a unique minimizer. In addition, if λ_n goes to zero, the objective function converges to the regression problem without the penalty whose unique minimizer is β. Thus, it is easy to see that $\hat{J} = \{j : \hat{\beta}_n^j(\cdot) \neq 0\}$ converges to J via the M-estimation theory. See (59) and (60).

Now, we extend β_n^J in Lemma 8 with zero functions as β_n^i for $i \in J^c$, name it $\beta_n \in \mathcal{H}$. Note that, it is a consistent estimator of β by Lemma 8. Since both of the $M_n(\cdot)$ and $\hat{M}_n(\cdot)$ have unique minimizers and the β_n is a consistent estimator of β, the consistency of $\hat{\beta}_n$ can be shown, if we can show that β_n satisfies the optimal conditions for $\hat{M}_n(\cdot)$ with a probability tending to one. The (asymptotically) optimal conditions of $\hat{M}_n(\cdot)$ are

$$\begin{cases} \|\hat{\Gamma}_{X^i X}\alpha - \hat{\Gamma}_{X^i Y}\|_{\mathcal{H}^i} \leq \lambda_n & i \notin J \\ \langle \hat{\Gamma}_{X^j X}\alpha, \cdot \rangle_{\mathcal{H}^j} - \hat{\Gamma}_{YX^j}(\cdot) = -\frac{\lambda_n}{\|\alpha^j\|_{\mathcal{H}^j}} \langle \alpha^j, \cdot \rangle_{\mathcal{H}^j} & j \in J. \end{cases}$$

The second equation is immediately satisfied with $\alpha = \beta_n$, since it satisfies the KKT

condition for $M_n(\cdot)$. We focus on the above inequality of the optimal condition. The first derivative condition for minimizing $M_n(\cdot)$ implies that β_n^J should justify the following equation.

$$-\hat{\Gamma}_{YX^J}(\cdot) + \langle \hat{\Gamma}_{X^JX^J}\beta_n^J, \cdot\rangle_{\mathcal{H}^J} + \lambda_n \sum_{j\in J} \frac{\langle \beta_n^j, \cdot\rangle_{\mathcal{H}^j}}{\|\beta_n^j\|_{\mathcal{H}^j}} = 0.$$

Define D_n be an operator from $\mathcal{H}^J$ to $\mathcal{H}^J$ such that $D_n(\alpha^J) = diag(\alpha^j/\|\beta_n^j\|)$ for $j \in J$. We rewrite the above equation as

$$-\hat{\Gamma}_{YX^J}(\cdot) + \langle (\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)\beta_n^J, \cdot\rangle_{\mathcal{H}^J} = 0.$$

In addition, note that

$$\hat{\Gamma}_{YX^J}(\cdot) = \langle \hat{\Gamma}_{X^JY}, \cdot\rangle_{\mathcal{H}^J} = \langle \hat{\Gamma}_{X^JX^J}\beta^J + \hat{\Gamma}_{X\epsilon}, \cdot\rangle_{\mathcal{H}^J}.$$

Thus, we have

$$\langle \beta_n^J, \cdot\rangle = \langle (\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}(\hat{\Gamma}_{X^JX^J}\beta^J + \hat{\Gamma}_{X^J\epsilon}), \cdot\rangle_{\mathcal{H}^J}.$$

Furthermore, by using a similar technique used in (A.6),

$$(\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}\hat{\Gamma}_{X^JX^J}\beta^J = \beta^J - (\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}\lambda_n D_n\beta^J.$$

Thus, for an $i \in J^c$:

$$\begin{aligned}\hat{\Gamma}_{X^iY} - \hat{\Gamma}_{X^iX^J}\beta_n^J &= \hat{\Gamma}_{X^iY} - \hat{\Gamma}_{X^iX^J}\beta^J + \lambda_n\hat{\Gamma}_{X^iX^J}(\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}D_n\beta^J \\ &\quad - \hat{\Gamma}_{X^iX^J}(\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}\hat{\Gamma}_{X^J\epsilon} \\ &= \lambda_n\hat{\Gamma}_{X^iX^J}(\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}D_n\beta^J + \hat{\Gamma}_{X^i\epsilon} \\ &\quad - \hat{\Gamma}_{X^iX^J}(\hat{\Gamma}_{X^JX^J} + \lambda_n D_n)^{-1}\hat{\Gamma}_{X^J\epsilon},\end{aligned}$$

by using the fact that $\hat{\Gamma}_{X^iY} - \hat{\Gamma}_{X^iX^J}(\beta^J) = \hat{\Gamma}_{X^i\epsilon}$. At this point, the formulation has a similar form, derived in Theorem 11 of (37). Furthermore, Lemma 8 satisfies the condition necessary to derive the rest of the proof so that they can be derived similarly. □

APPENDIX B: LISTS AND 3D DISPLAY

Figure B.1 displays the regions of interests associated with the active sets similar to that of figures 8.1 and 8.2 but in three dimensions. The colors match those of figures 8.1 and 8.2.

List of regions of interests: The following are the lists of the regions of interest of the human brain used in the application section 8.1. The atlas labels of the human brain and full names can be found at Atlas Label.

The list of the regions of interest associated with the active set of MFG-LASSO when the response value is IQ score:

"Frontal–Mid–Orb–L", "Frontal–Mid–Orb–R", "Frontal–Inf–Oper–L", "Frontal–Inf–Oper–R","Frontal–Inf–Tri–L", "Frontal–Inf–Tri–R", "Frontal–Inf–Orb–L","Frontal–Inf–Orb–R", "Rolandic–Oper–R", "Supp–Motor–Area–L", "Olfactory–L", "Olfactory–R", "Frontal–Sup–Medial–L","Frontal–Med–Orb–L", "Frontal–Med–Orb–R", "Rectus–L", "Cingulum–Ant–L", "Cingulum–Post–L", "Cingulum–Post–R", "Amygdala–L", "Amygdala–R", "Calcarine–L", "Calcarine–R", "Cuneus–L", "Cuneus–R", "Lingual–L", "Occipital–Sup–L", "Occipital–Sup–R", "Occipital–Mid–R", "Occipital–Inf–L", "Occipital–Inf–R", "Parietal–Sup–L", "Parietal–Inf–R", "SupraMarginal–L", "SupraMarginal–R", "Angular–L", "Angular–R", "Precuneus–L", "Paracentral–Lobule–L", "Paracentral–Lobule–R","Putamen–L", "Pallidum–R", "Heschl–R", "Temporal–Sup–L", "Temporal–Pole–Mid–L", "Cerebellum–3–L", "Cerebellum–3–R", "Vermis–1–2", "Vermis–3", "Vermis–4–5", "Vermis–6", "Vermis–9", "Vermis–10".

The list of the regions of interest associated with the active set of MFG-LASSO when the response value is Verbal IQ:

"Frontal–Sup–R", "Frontal–Mid–Orb–L","Frontal–Mid–Orb–R", "Frontal–Inf–Oper–R","Frontal–Inf–Tri–L","Frontal–Inf–Tri–R", "Frontal–Inf–Orb–L","Frontal–Inf–Orb–R","Rolandic–Oper–R", "Supp–Motor–Area–L","Olfactory–L",

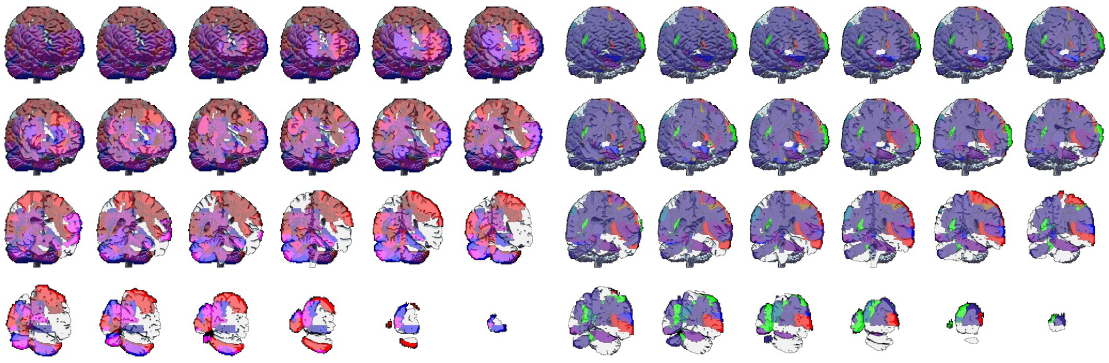


Figure B.1: Three-dimensional display of figures 8.1 and 8.2. The results when the ADHD score is the response value are in the right panel and the results when the IQ score is the response value are in the left panel.

"Frontal–Sup–Medial–L" "Frontal–Med–Orb–L","Frontal–Med–Orb–R","Rectus–L", "Cingulum–Ant–L", "Cingulum–Post–L", "Cingulum–Post–R", "Amygdala–L", "Amygdala–R", "Calcarine–L", "Calcarine–R", "Cuneus–L", "Cuneus–R", "Occipital–Sup–L", "Parietal–Sup–L", "Parietal–Sup–R", "Parietal–Inf–L", "Parietal–Inf–R", "SupraMarginal–L", "SupraMarginal–R", "Angular–L", "Precuneus–L", "Precuneus–R", "Paracentral–Lobule–L","Paracentral–Lobule–R" "Putamen–L", "Pallidum–R", "Heschl–R", "Temporal–Sup–L", "Temporal–Pole–Mid–L","Cerebellum–3–L", "Vermis–1–2", "Vermis–3", "Vermis–4–5", "Vermis–6", "Vermis–9", "Vermis–10", .

The list of the regions of interest associated with the active set of MFG-LASSO when the response value is Performance IQ:

"Frontal–Sup–Orb–L","Frontal–Mid–Orb–L","Frontal–Mid–Orb–R", "Frontal–Inf–Oper–L","Frontal–Inf–Oper–R","Frontal–Inf–Tri–L", "Frontal–Inf–Tri–R","Frontal–Inf–Orb–L","Frontal–Inf–Orb–R", "Rolandic–Oper–R", "Supp–Motor–Area–L","Olfactory–L", "Olfactory–R", "Frontal–Sup–Medial–L","Frontal–Sup–Medial–R" "Frontal–Med–Orb–L","Frontal–Med–Orb–R","Rectus–L", "Insula–R", "Cingulum–Ant–L", "Cingulum–Mid–L", "Cingulum–Post–L", "Cingulum–Post–R", "ParaHippocampal–L", "ParaHippocampal–R","Amygdala–L", "Amygdala–R", "Calcarine–L", "Calcarine–R", "Cuneus–L", "Cuneus–R", "Lingual–L", "Occipital–Sup–L", "Occipital–Sup–R", "Occipital–Mid–L", "Occipital–Mid–R", "Occipital–Inf–L", "Occipital–Inf–R", "Postcentral–L", "Postcentral–R", "Parietal–Sup–L", "Parietal–Sup–R", "Parietal–Inf–L", "Parietal–Inf–R", "SupraMarginal–L", "SupraMarginal–R", "Angular–L", "Angular–R", "Precuneus–L", "Precuneus–R", "Paracentral–Lobule–L" "Paracentral–Lobule–R","Caudate–L", "Putamen–L", "Pallidum–R", "Thalamus–L", "Heschl–L", "Heschl–R", "Temporal–Sup–L", "Temporal–Pole–Sup–L", "Temporal–Pole–Sup–R","Temporal–Mid–L", "Temporal–Pole–Mid–L", "Temporal–Pole–Mid–R","Cerebellum–3–L", "Cerebellum–3–R", "Cerebellum–4–5–R", "Cerebellum–6–L", "Cerebellum–6–R", "Vermis–1–2", "Vermis–3", "Vermis–4–5", "Vermis–6", "Vermis–7", "Vermis–9", "Vermis–10".

The list of the regions of interest associated with the active set of MFG-LASSO when the response value is ADHD score:

"Frontal–Mid–L", "Frontal–Mid–Orb–L","Frontal–Mid–Orb–R", "Frontal–Inf–Oper–L","Frontal–Inf–Oper–R","Frontal–Inf–Tri–L", "Frontal–Inf–Orb–L","Frontal–Inf–Orb–R","Supp–Motor–Area–L", "Olfactory–L", "Frontal–Sup–Medial–L","Frontal–Sup–Medial–R" "Frontal–Med–Orb–L","Rectus–L", "Cingulum–Ant–L", "Cingulum–Post–L", "ParaHippocampal–R","Amygdala–L", "Calcarine–L", "Cuneus–L", "Cuneus–R", "Occipital–Inf–L", "Occipital–Inf–R", "Parietal–Sup–L", "Parietal–Inf–L", "SupraMarginal–L", "SupraMarginal–R", "Angular–L", "Angular–R", "Precuneus–L", "Paracentral–Lobule–L","Paracentral–Lobule–R","Putamen–L", "Heschl–R", "Temporal–Sup–L", "Temporal–Pole–Sup–R", "Temporal–Pole–Mid–L","Cerebellum–9–L", "Vermis–1–2", "Vermis–4–5", "Vermis–10".

The list of the regions of interest associated with the active set of MFG-LASSO when the response value is ADHD Inattentive:

"Frontal–Mid–Orb–L","Frontal–Mid–Orb–R","Frontal–Inf–Oper–L", "Frontal–Inf–Oper–R","Frontal–Inf–Tri–L","Frontal–Inf–Orb–L", "Frontal–Inf–Orb–R","Supp–Motor–Area–L","Frontal–Sup–Medial–L" "Frontal–Sup–Medial–R","Frontal–Med–Orb–L","Rectus–L", "Cingulum–Ant–L", "Cingulum–Post–L", "Cingulum–Post–R", "ParaHippocampal–R","Amygdala–L", "Calcarine–L", "Cuneus–L", "Cuneus–R", "Lingual–L", "Occipital–Inf–L", "Occipital–Inf–R", "Parietal–Sup–L", "Parietal–Inf–L", "SupraMarginal–L", "SupraMarginal–R", "Angular–L", "Angular–R", "Precuneus–L", "Precuneus–R", "Paracentral–Lobule–L","Paracentral–Lobule–R" "Heschl–R", "Temporal–Sup–L", "Temporal–Pole–Sup–R", "Temporal–

Pole–Mid–L","Cerebellum–4–5–R", "Vermis–1–2", "Vermis–4–5", "Vermis–10".

The list of the regions of interest associated with the active set of MFG-LASSO when the response value is ADHD Hyper/Impulsive:
"Frontal–Mid–Orb–L","Frontal–Mid–Orb–R","Frontal–Inf–Oper–L", "Frontal–Inf–Oper–R","Frontal–Inf–Tri–L","Frontal–Inf–Orb–L", "Frontal–Inf–Orb–R","Rolandic–Oper–R", "Supp–Motor–Area–L", "Olfactory–L", "Frontal–Sup–Medial–L","Frontal–Sup–Medial–R" "Frontal–Med–Orb–L","Frontal–Med–Orb–R","Rectus–L", "Rectus–R", "Cingulum–Ant–L", "Cingulum–Mid–L", "Cingulum–Post–L", "ParaHippocampal–R","Amygdala–L", "Amygdala–R", "Calcarine–L", "Cuneus–L", "Cuneus–R", "Occipital–Sup–R", "Occipital–Mid–R", "Occipital–Inf–L", "Occipital–Inf–R", "Parietal–Sup–L", "Parietal–Inf–L", "Parietal–Inf–R", "SupraMarginal–L", "SupraMarginal–R", "Angular–L", "Angular–R", "Putamen–L", "Pallidum–R", "Heschl–L", "Heschl–R", "Temporal–Sup–L", "Temporal–Pole–Sup–R", "Temporal–Pole–Mid–L","Temporal–Pole–Mid–R","Cerebellum–3–R", "Cerebellum–4–5–R", "Cerebellum–9–L", "Vermis–1–2", "Vermis–3", "Vermis–4–5", "Vermis–6", "Vermis–7", "Vermis–10".

The list of the regions that are associated with IQ but not with ADHD by the MFG-LASSO:
"Frontal–Inf–Tri–R","Rolandic–Oper–R","Olfactory–R", "Frontal–Med–Orb–R","Cingulum–Post–R", "Amygdala–R", "Calcarine–R", "Lingual–L", "Occipital–Sup–L","Occipital–Sup–R", "Occipital–Mid–R","Parietal–Inf–R","Pallidum–R", "Cerebellum–3–L", "Cerebellum–3–R", "Vermis–3", "Vermis–6", "Vermis–9",

The list of the regions that are associated with ADHD but not with IQ by the MFG-LASSO:
"Frontal–Mid–L", "Frontal–Sup–Medial–R","ParaHippocampal–R","Parietal–Inf–L", "Temporal–Pole–Sup–R","Cerebellum–9–L".

List of variables and countries in the econometric data: The following are the lists of the countries and functional covariates used in the application section 8.2.

List of functional covariates in the econometric data:
[1] Population growth (annual %), [2] Rural population (% of total population), [3] Urban population (% of total population), [4] Urban population growth (annual %), [5] Rural population growth (annual %), [6] Adjusted savings: education expenditure (% of GNI), [7] Immunization, DPT (% of children ages 12-23 months), [8] Age dependency ratio (% of working-age population), [9] Age dependency ratio, old (% of working-age population), [10] Age dependency ratio, young (% of working-age population), [11] Immunization, measles (% of children ages 12-23 months), [12] Population ages 00-04, female (% of female population), [13] Population ages 00-04, male (% of male population), [14] Population ages 0-14 (% of total population), [15] Population ages 0-14, female (% of female population), [16] Population ages 0-14, male (% of male population), [17] Population ages 05-09, female (% of female population), [18] Population ages 05-09, male (% of male population), [19] Population ages 10-14, female (% of female population), [20] Population ages 10-14, male (% of male population), [21] Population ages 15-19, female (% of female population), [22] Population ages 15-19, male (% of male population), [23] Population ages 15-64 (% of total population), [24] Population ages 15-64, female (% of female population), [25] Population ages 15-64, male (% of male population), [26] Population ages 20-24, female (% of female population), [27] Population ages 20-24, male (% of male population), [28] Population ages 25-29, female (% of female population), [29] Population ages 25-29, male (% of male population),

[30] Population ages 30-34, female (% of female population), [31] Population ages 30-34, male (% of male population), [32] Population ages 35-39, female (% of female population), [33] Population ages 35-39, male (% of male population), [34] Population ages 40-44, female (% of female population), [35] Population ages 40-44, male (% of male population), [36] Population ages 45-49, female (% of female population), [37] Population ages 45-49, male (% of male population), [38] Population ages 50-54, female (% of female population), [39] Population ages 50-54, male (% of male population), [40] Population ages 55-59, female (% of female population), [41] Population ages 55-59, male (% of male population), [42] Population ages 60-64, female (% of female population), [43] Population ages 60-64, male (% of male population), [44] Population ages 65 and above (% of total population), [45] Population ages 65 and above, female (% of female population), [46] Population ages 65 and above, male (% of male population), [47] Population ages 65-69, female (% of female population), [48] Population ages 65-69, male (% of male population), [49] Population ages 70-74, female (% of female population), [50] Population ages 70-74, male (% of male population), [51] Population ages 75-79, female (% of female population), [52] Population ages 75-79, male (% of male population), [53] Population ages 80 and above, female (% of female population), [54] Population ages 80 and above, male (% of male population), [55] Population, female (% of total population), [56] Population, male (% of total population), [57] Survival to age 65, female (% of cohort), [58] Survival to age 65, male (% of cohort), [59] Contributing family workers, female (% of female employment) (modeled ILO estimate), [60] Contributing family workers, male (% of male employment) (modeled ILO estimate), [61] Contributing family workers, total (% of total employment) (modeled ILO estimate), [62] Employers, female (% of female employment) (modeled ILO estimate), [63] Employers, male (% of male employment) (modeled ILO estimate), [64] Employers, total (% of total employment) (modeled ILO estimate), [65] Employment in agriculture (% of total employment) (modeled ILO estimate), [66] Employment in agriculture, female (% of female employment) (modeled ILO estimate), [67] Employment in agriculture, male (% of male employment) (modeled ILO estimate), [68] Employment in industry (% of total employment) (modeled ILO estimate), [69] Employment in industry, female (% of female employment) (modeled ILO estimate), [70] Employment in industry, male (% of male employment) (modeled ILO estimate), [71] Employment in services (% of total employment) (modeled ILO estimate), [72] Employment in services, female (% of female employment) (modeled ILO estimate), [73] Employment in services, male (% of male employment) (modeled ILO estimate), [74] Employment to population ratio, 15+, female (%) (modeled ILO estimate), [75] Employment to population ratio, 15+, male (%) (modeled ILO estimate), [76] Employment to population ratio, 15+, total (%) (modeled ILO estimate), [77] Employment to population ratio, ages 15-24, female (%) (modeled ILO estimate), [78] Employment to population ratio, ages 15-24, male (%) (modeled ILO estimate), [79] Employment to population ratio, ages 15-24, total (%) (modeled ILO estimate), [80] Labor force participation rate for ages 15-24, female (%) (modeled ILO estimate), [81] Labor force participation rate for ages 15-24, male (%) (modeled ILO estimate), [82] Labor force participation rate for ages 15-24, total (%) (modeled ILO estimate), [83] Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate), [84] Labor force participation rate, female (% of female population ages 15-64) (modeled ILO estimate), [85] Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate), [86] Labor force participation rate, male (% of male population ages 15-64) (modeled ILO estimate), [87] Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate), [88] Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate), [89] Labor force, female (% of total labor force), [90] Ratio of female to male labor force participation rate (%) (modeled ILO estimate), [91] Self-employed, female (% of female employment) (modeled ILO estimate), [92] Self-employed, male (% of male employment) (modeled ILO estimate), [93] Self-employed, total (% of total employment) (modeled ILO estimate), [94] Unemployment, female (% of female labor force) (modeled ILO

estimate), [95] Unemployment, male (% of male labor force) (modeled ILO estimate), [96] Unemployment, total (% of total labor force) (modeled ILO estimate), [97] Unemployment, youth female (% of female labor force ages 15-24) (modeled ILO estimate), [98] Unemployment, youth male (% of male labor force ages 15-24) (modeled ILO estimate), [99] Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate), [100] Vulnerable employment, female (% of female employment) (modeled ILO estimate), [101] Vulnerable employment, male (% of male employment) (modeled ILO estimate), [102] Vulnerable employment, total (% of total employment) (modeled ILO estimate), [103] Wage and salaried workers, female (% of female employment) (modeled ILO estimate), [104] Wage and salaried workers, male (% of male employment) (modeled ILO estimate), [105] Wage and salaried workers, total (% of total employment) (modeled ILO estimate), [106] Merchandise trade (% of GDP), [107] GDP growth (annual %), [108] Inflation, GDP deflator (annual %), [109] GDP per capita growth (annual %), [110] Inflation, GDP deflator: linked series (annual %), [111] Adjusted savings: carbon dioxide damage (% of GNI).

List of couturiers in the econometric data:

"Albania", "Algeria", "Angola", "Argentina", "Armenia", "Australia", "Austria", "Azerbaijan", "The Bahamas", "Bahrain", "Bangladesh", "Barbados", "Belarus", "Belize", "Benin", "Bhutan", "Bolivia", "Botswana", "Brazil", "Brunei", "Bulgaria", "Burundi", "Cabo Verde", "Cambodia", "Cameroon", "Canada", "Central African Republic", "Chad", "Chile", "China", "Colombia", "Comoros", "Dem. Rep. Congo", "Congo", "Costa Rica", "Cyprus", "Czech Republic", "Denmark", "Dominican Republic", "Ecuador", "Egypt", "El Salvador", "Eswatini", "Ethiopia", "Fiji", "Finland", "France", "Gabon", "The Gambia", "Georgia", "Germany", "Ghana", "Greece", "Guatemala", "Guinea", "Guinea-Bissau", "Guyana", "Haiti", "Honduras", "Hungary", "Iceland", "India", "Indonesia", "Ireland", "Israel", "Italy", "Jamaica", "Japan", "Jordan", "Kazakhstan", "Kenya", "Korea", "Lebanon", "Lesotho", "Madagascar", "Malawi", "Malaysia", "Mali", "Malta", "Mauritania", "Mauritius", "Mexico", "Mongolia", "Morocco", "Mozambique", "Namibia", "Nepal", "Netherlands", "New Zealand", "Nicaragua", "Niger", "Nigeria", "North Macedonia", "Norway", "Oman", "Pakistan", "Panama", "Papua New Guinea", "Paraguay", "Peru", "Philippines", "Poland", "Portugal", "Romania", "Russia", "Rwanda", "Saudi Arabia", "Senegal", "Slovak Republic", "Slovenia", "Solomon Islands", "South Africa", "Spain", "Sri Lanka", "St. Lucia", "St. Vincent and the Grenadines", "Sudan", "Sweden", "Switzerland", "Tajikistan", "Tanzania", "Thailand", "Togo", "Tonga", "Trinidad and Tobago", "Tunisia", "Turkey", "Turkmenistan", "Uganda", "Ukraine", "United States", "Uruguay", "Uzbekistan", "Vanuatu", "Vietnam", "Zambia", "Zimbabwe".