MODEL SELECTION AND ESTIMATION IN HIGH DIMENSIONAL SETTINGS

A Thesis
Presented to
The Academic Faculty

by

Rodrigue Ngueyep Tzoumpe

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2015

MODEL SELECTION AND ESTIMATION IN HIGH DIMENSIONAL SETTINGS

Approved by:

Dr. Nicoleta Serban, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. David Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yao Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Pierre Vandekerkhove
Department of Analysis and Applied Mathematics
University Paris-Est

Date Approved: 27 March 2015

To my family and loved ones

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Nicoleta Serban for giving me the opportunity to pursue the PhD under her supervision. Through the years, she provided the guidance and support neccessary for the completion of this work, she also allowed me to pursue my own interests and for that I am extremely grateful. I would also like to express my heartfelt gratitude to my thesis committee: Dr. Pierre Vandekerkhove for our fruitful collaboration and his advice; Dr. Roshan Joseph Vengazhiyil, Dr. David Goldsman and Dr. Yao Xie for insighful comments and questions that helped improve the quality of the thesis.

To my parents Jean-Faustin and Rose Ngueyep, I express my profound love and infinite gratitude for their unconditional support in my academic journey and for the values they instilled in me. I couldn't have achieved this without their trust and love. To my brothers and sisters (Franklin, Michele and Marina) I send my love and thanks for being a source of strength. I will also like to thank the Kouamou's family for their kindness, their generosity and for always welcoming me into their home.

To Arnaud Amadjikpe and Nadia Bokossa, Mamadou Diao and Rose Ndong, Ibrahima and Nefertari Ndiour, Illenin Kondo, and the Nikoue family, I convey my appreciation for their friendship, their support and the shared memories. I would also like to thank Christian Defeu, Stephane Ntwoku and many others for making my time in Atlanta really enjoyable.

I am infinitely grateful to my dear wife Line Francine Kouecheu, whose love, advice and constant encouragement carried me to the finish line.

Last but not least, thanks be to God for his blessings and for his indescribable gifts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Several statistical problems can be described as estimation problems, where the goal is to learn a set of parameters, from some data, by maximizing a criterion. These type of problems are typically encountered in a supervised learning setting, when there is a need to relate an output (or many outputs) to multiple inputs. The relationship between these outputs and these inputs can be complex, and this complexity can be attributed to the high dimensionality of the space containing the inputs and the outputs; the existence of a structural prior knowledge within the inputs or the outputs that if ignored may lead to inefficient estimates of the parameters; and the presence of a non-trivial noise structure in the data. In many scientific fields a strong preference is always granted to parsimonious models, simply because they are easier to use and to explain. In the statistical context parsimonious supervised models can be attained by inducing sparsity in the studied model. The first theme of the thesis is the identification of meaningful relationships between a large number of inputs and a large number of outputs through the use of sparsity inducing methods. The methods proposed that fall under the first theme, will address the high dimensional nature of this problem and the existence of prior structural knowledge in the data. To better understand the challenges raised by these type of problems let's provide some concrete examples.

Time series are data that are collected through time at a constant or non-constant frequency. Due to improvements in data gathering methods and storage technologies, it is now relatively easy to simultaneously keep track of a large number of times

series. So clearly if we are interested in understanding how these time series affect each other, we will have to handle the high dimensionality of the problem that comes from the fact that sometimes there could be more time series than observations per time series. Within the context of time series that are gathered at specific locations in space, it seems natural to also exploit the prior structural knowledge intrinsic to spatio temporal data. Indeed, events that are closer in time are likely to be more influential on the present than events that are further in the past. Similarly from a spatial perspective, time series observed at neighboring sites are likely to influence each other more than time series collected at locations that are farther apart. Better estimates are indeed obtained when the temporal nature of the data is taken into consideration (e.g., see Song (2011)). High dimensional time series data that have spatial characteristics are frequent, for example, in economics we can monitor jointly economics indicator observed in different geographical areas. In meteorology, weather related time series can be measured at discrete recording stations and the goal would be to analyze how and if the meteorological time series measured influence each other.

Many statistical learning problems, with multiple inputs and multiple outputs can be formulated as a multi-responses regression or a multi-category classification problem. For example, in genetics using genetic expression data one can classify the nature of the tumor of cancer patients. In these type of studies, the number of genes profiled is typically much larger than the number of samples available for each genes, since there is a limited number of cancer patients, and the main goal is to discriminate between different type of cancers by using the high-dimensional feature vectors (genes profile). In multi-responses regression, the main interest is to identify a set of shared features that influence the outputs. For example, in the analysis of quarterly healthcare costs at the county or zipcode level in the state of North Carolina, we are interested in identifying predictors that influence all the cost of all (or some) of

the geographical areas. Within this context of multi-responses regression or multi-category classification, challenges mentioned earlier were the presence of nonlinearities between the inputs and the outputs and the existence of prior structural knowledge. The structural knowledge that can be leveraged in this problem, is the fact that some outputs are highly suceptible to share similar inputs. So a common sparsity pattern can be observed across categories or across responses, but the methodology also needs to be flexible to leave the possibility of having inputs that are not shared across all the responses or categories.

The second theme of the thesis deals with a statistical problems where the regression data are believed to belong to two or more distinct unobserved categories. The complexity in the relationship between the predictor and the response lies in the fact that in each category the relationship is different. These models are commonly known as mixture-of-regression models and they have been extensively used in many fields, such as biology, genetics, medicine, economics and engineering, among many other fields. While parametric models such as finite mixture linear regression models remain the most popular techniques in modeling data that exhibit mixture distributions, they are very often not flexible enough to model the nonlinear relationship that exist between the response and predictors and they also assume that the distribution of the noise falls in a known family. To account for these challenges exhibited by modern data, we propose a semiparametric mixture regression model that only assumes the continuity of unknown regression functions and the symmetry of the distribution of the noise.

In the remainder of this introduction we summarize the contributions made in this thesis.

Chapter 2: one of the most commonly used methods for modeling multivariate time series is the Vector Autoregressive Model (VAR). VAR is generally used to identify

lead, lag and contemporaneous relationships describing Granger causality within and between time series. In this chapter, we investigate VAR methodology for analyzing data consisting of multi-layer time series which are spatially interdependent. When modeling VAR relationships for such data, the dependence between time series is both a curse and a blessing. The former because it requires modeling the between time series correlation or the contemporaneous relationships which may be challenging when using likelihood-based methods. The latter because the spatial correlation structure can be used to specify the lead-lag relationships within and between time series, within and between layers. To address these challenges, we propose a $\ell_1 \backslash \ell_2$ regularized likelihood estimation method. The lead, lag and contemporaneous relationships are estimated using a new coordinate descent algorithm that exploits sparsity in the VAR structure, accounts for the spatial dependence and models the error dependence. We assess the performance of the proposed VAR model and compare it with existing methods within a simulation study. We also apply the proposed methodology to a large number of state-level US economic time series.

Chapter 3: in this chapter, we propose a new methodology to tackle the problem of high-dimensional nonparametric learning in the multi-responses or multitask learning setting. We impose sparsity constraints that allow the recovery of the additive functions that are the most influential accross tasks and responses. The methodology instead of applying ℓ_∞ as proposed by Liu et al. (2008), applies a functional $\ell_1 \backslash \ell_2$ norm to each group of additive functions. Each group contains all the additive functions associated with a specific predictor. We derive a novel thresholding condition for the union support recovery in the nonparametric setting. We propose a sparse backfitting based algorithm to solve for the additive functions. Through extensive simulations, we show the superior performance of the methodology.

Chapter 4: Motivated by the analysis of a Positron Emission Tomography (PET)

imaging data considered in Bowen et al. (2012), we introduce a semiparametric topographical mixture model able to capture the characteristics of dichotomous shifted response-type experiments. We propose a pointwise estimation procedure of the proportion and location functions involved in our model. Our estimation procedure is only based on the symmetry of the local noise and does not require any finite moments on the errors (e.g. Cauchy-type errors). We establish under mild conditions minimax properties and asymptotic normality of our estimators. Moreover, Monte Carlo simulations are conducted to examine their finite sample performance. Finally a statistical analysis of the PET imaging data in Bowen et al. (2012) is illustrated for the proposed method.

Chapter 5: In this chapter, we summarize the main contributions of this dissertation and discuss potential direction for future work.

CHAPTER II

LARGE VECTOR AUTO REGRESSION FOR MULTI-LAYER SPATIALLY CORRELATED TIME SERIES

2.1 Introduction

Analyzing multivariate time series is a common statistical problem in several fields, such as economics and environmental sciences. One of the most commonly used methods for modeling multivariate time series is the Vector Autoregressive Model (VAR) introduced by Sims (1980). Generally, VAR has been used to identify Granger causal relationships between variables which vary over time. The primary focus is on the lead and lag effects between time series but often contemporaneous relationships provide additional information about how variables are related to each other over a period of time. In this paper, we investigate VAR methodology for analyzing data consisting of *spatially interdependent multi-layer time series*. Specific examples from various fields are:

• Industrial Economics: multiple time-varying economic indicators such as state level employment rates in the construction industry and the number of building permits issued for new homes observed at the county or even the census tract level within a state or nationally;

• Industrial Engineering: multiple turbines installed at different geographic locations for which time-varying wind speed and generated power are recorded;

• Environmental sciences: multiple measurements observed at different stations as often generated by environmental and climatological studies.

In many of these examples, one layer corresponds to a different measurement or

indicator. Specifically, the observed time series data are

$$Y_{t,k}^{[J]} = Y_{t,s_k^{[J]}}, \text{ with } k \in \{1, \ldots, K_J\} \text{ and } J \in \{1, \ldots, L\}$$

where t is the time unit and L is the number of layers, typically small and in our simulations and application its maximum value is 2. K_J represents the total number of sites in the J^{th} layer and $s_1^{[J]}, \ldots, s_{K_J}^{[J]}$ are the spatial units or locations for the time series in the J^{th} layer. They are recorded as coordinates (latitude and longitude) of each spatial site and are used to map the pair (site number, layer), for example $s_3^{[1]}$ is the latitude and longitude of the 3^{rd} site in layer 1. Each time series $Y_{t,k}^{[J]}$ can be influenced by observations from

• Own lags: $Y_{t-p,k}^{[J]}$ for $p = 1, \ldots, P$, where P is the maximum lag considered in the study.

• Lags of neighboring time series within the same layer J, for example, observations of the time series $Y_{t-p,k'}^{[J]}$ for $p = 1, \ldots, P$ located at a site $s_{k'}^{[J]}$ such that site $s_{k'}^{[J]}$ is close to $s_k^{[J]}$.

• Lags of neighboring time series within layers other than layer J, for example, observations of the time series $Y_{t-p,k^\star}^{[J^\star]}$ in layer $J^\star$ at site $s_{k^\star}^{[J^\star]}$, $J \neq J^\star$ and $s_{k^\star}^{[J^\star]}$ is close to $s_k^{[J]}$.

Therefore, each time series can be influenced by observations within its layer or outside its layer. We assume that for each influential layer, the set of sites that affect a targeted time series is most likely restricted to a close spatial neighborhood of the site of observation of the targeted time series.

To better explain our methodology, we consider the example illustrated in Figure 1. Assume that we have two layers ($L = 2$); the crosses correspond to the sites of time series in layer 1, the circles correspond to the sites of time series in layer 2, and we are interested in predicting the time series at site 21 in layer 1 $\{Y_{t,21}^{[1]}\}$. Based on the proximity of their sites, time series $\{Y_{t,5}^{[1]}\}, \{Y_{t,8}^{[1]}\}, \{Y_{t,15}^{[1]}\}$ in layer 1 contribute to within layer effects (relationships showed by solid lines in figure 1). The time series

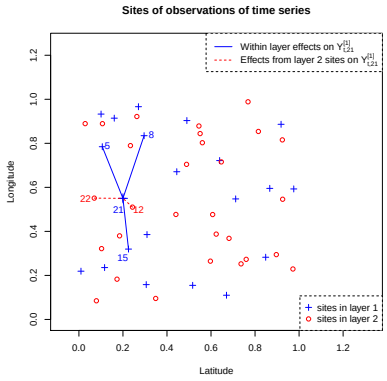


Figure 1: Illustration of spatial representation of time series data with two layers, with targeted site(spatial site 21 in layer 1)

$\{Y^{[2]}_{t,12}\}, \{Y^{[2]}_{t,22}\}$ in layer 2 will be responsible for cross layer effects (dashed lines in figure 1). The anticipated influence of $\{Y^{[2]}_{t,12}\}$ on the response time series $\{Y^{[1]}_{t,21}\}$ will be more important than the influence of $\{Y^{[2]}_{t,22}\}$ on the same response time series, because the coordinates $s^{[2]}_{12}$ of site 12 in layer 2 is closer to the coordinates $s^{[1]}_{21}$ of the target site 21 in layer 1, than the coordinates $s^{[2]}_{22}$ of site 22 in layer 2 are.

In certain settings, the dynamics of a time series can be approximated by a linear function of its own lags and the lags of influential time series. This reduces to a VAR model with complex lead and lag relationships, which often results in a model with a higher dimensionality than the number of observations. On the other hand, only a small number of lead and lag relationships are expected to be significant. Employing methods which account for this sparsity will allow estimation of a high dimensional VAR model. In order to take advantage of the sparsity in the relationships within and between time series, our method uses regularization penalties that are functions of the lags of the time series within the same layer and from different layers. In specifying the regularization penalties, we assume that closer information in time has more relevance. When time series are spatially correlated, we also assume that closer information in space has more relevance. Therefore, the relationships of one targeted time series to other time series are increasingly penalized with higher temporal lags

and at higher spatial distance.

The primary contribution of our paper is a method for identifying lead and lag relationships between a large number of time series that also exhibit strong contemporaneous spatial dependence. Additionally, the method allows for estimating sparse relationships within and between layers of time series. To the best of our knowledge, this is the first and only approach that accounts simultaneously for the large dimensionality of the problem, the temporal dependence among time series, the spatial dependence present in the errors and the layer group effects. The second contribution is an efficient algorithm that can be used to solve the optimization problem in regularized selection approaches for models similar to the one proposed in this paper.

The remainder of the paper is organized as follows. In Section 2.2, we review the literature on model selection with a focus on VAR modeling and we motivate the general approach introduced in this paper. In Section 2.3, we describe the spatial VAR model applied to one layer followed by its extension to a multi-layer setting. Section 2.4 introduces the estimation procedure and explains the computational algorithms used to fit the model. In Section 2.6, we analyze the relationship between state level employment rates in the construction industry and the number of building permits issued for new homes. We conclude with insights in the application of the proposed methodology in Section 2.7. Extensive simulation studies are carried out and their results are presented in web supplements. Additionally, some technical details are also deferred to the Supplemental Material.

2.2 Background and Motivation

The analysis of multivariate time series, and particularly VAR, has been extensively covered in the statistical, computer science and econometrics literature, but most of the existing methods fail to jointly perform model selection and estimation in a high dimensional setting, meaning when the number of time series is large relative

to the sample size. In the context of VAR, variable selection reveals statistically significant relationships within and between time series. Variable selection is critical because a large number of time series implies a large number of potential lead and lag relationships.

One straightforward methodology consists in regressing each time series onto the others separately resulting in multiple regressions, one for each time series. This approach often produces inefficient coefficient estimates due to the large model dimensionality as compared to the sample size of each time series potentially leading to poor forecasting due to overfitting. This challenging aspect has been highlighted in other existing studies (Roecker, 1991; Breiman, 1995).

Alternatively, one could consider variable selection within a multivariate regression model. Variable selection tools based on information criteria have been developed for multivariate regression by Bedrick and Tsai (1994), Fujikoshi and Satoh (1997) among others. Because of the high computational cost, these methods are not used to select the best model among all possible subset structures. Instead, these methods rely on greedy search algorithms, for example, top-down and bottom-up approaches, that are highly unstable, path dependent and suboptimal (Krolzig and Hendry, 2001; Penm and Terrell, 1984). An alternative approach to multivariate regression is to reduce the dimensionality of the predictors - in the VAR context, the predictors consist of lead and lag relationships between time series - using factor analysis. Related work includes reduced-rank regression methods (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998) and the Factor Estimation and Selection (FES) proposed by Yuan et. al (2007). For these methods, because the set of predictors is reduced to a few important principal factors, the interpretation of the Granger causal relationships is difficult. Some papers have proposed to use a bayesian approach for the estimation of multivariate regression, for example Cripps et al. (2005) perform variable selection and covariance selection in multivariate regression models.

The emergence of regularized estimation methods such as the Lasso by Tibshirani (1996) has led to the development of regularized sparse estimation schemes for multivariate regression. For example, Turlach et al. (2005) perform model selection using a L_∞-regularization scheme applied to all the coefficients related to a predictor. Obozinski et al. (2008) apply the $L_1\backslash L_2$ regularization for union support recovery, and Peng et al. (2010) introduce a $L_1\backslash L_2$ penalization method for identification of "master" predictors in a multivariate regression. In a more recent study, Rothman et al. (2010) introduce joint estimation of the regression parameters and the covariance of errors by L_1 regularized log likelihood. But their approach does not apply to time series data as it does not allow for modeling the serial correlation within time series. Song and Bickel (2011) propose to impose lag-dependence in the regularization penalties to estimate a large VAR model. While this method accounts for serial dependence in the data, it doesn't include the effects due to contemporaneous correlation present in the errors. Davis et al. (2012) propose a 2-stage approach for estimating sparse VAR (sVAR) models. Their method uses partial spectral coherence with BIC to select non-zero AR coefficients. But their methodology does not explicitly take into consideration the spatial correlation in the errors.

Although the research studies discussed above are a leap from the more traditional VAR modeling, they are still limited in their application. Particularly, they do not simultaneously select lead, lag and contemporaneous relationships within and between time series. The lead & lag relationship selection performance worsens when we do not account for the spatial correlation in the errors. Moreover, existing approaches do not readily extend to data observed for multiple measurements (e.g. humidity, precipitation and temperature) often called layers (Huang et. al, 2010).

To address these limitations, we use a $L_1\backslash L_2-$ regularized likelihood method to select the temporal lags and spatial sites that influence a targeted time series. Specifically, L_1 regularization is used for selecting individual time series effects while

L_2 regularization is used for selecting entire layers viewed as group effects similar to the sparse group lasso introduced by Friedman et al. (2010). For example, if we are interested in finding the effect of other layers on a time series at a targeted location, the time lag- and spatial distance-weighted regularization associated with the L_2 penalty will perform group selection between the layers. This regularization identifies whether entire layers are not relevant, meaning that all time series in the layer will have no effect on a targeted time series. Since group lasso doesn't yield within group sparsity, to identify the most influential neighborhood for the selected layers, we therefore apply a temporal lag- and spatial distance-weighted L_1- regularization. This penalization approach allows for selection of parsimonious models resulting in efficient parameter estimation and accuracy of time series prediction.

Moreover, to incorporate contemporaneous (spatial) dependence, we propose to use a penalized log-likelihood scheme since it allows the estimation of the covariance matrix of the errors. A similar idea is applied by Rothman et. al (2010) in the context of multivariate regression. Within the multi-layer time series framework, we assume that there is no cross-layer contemporaneous dependence. This assumption allows us to use a divide-and-conquer algorithm to simultaneously solve L optimization problems of smaller size, therefore, reducing the computational effort.

The estimation procedure of our model consists of alternatively solving for the VAR coefficients and solving for the inverse covariance matrix of the errors. To estimate the VAR coefficients, we solve the $L_1 \backslash L_2$- regularization likelihood by providing an algorithm that uses block coordinate descent. To solve for the inverse covariance matrix, we use a spatially weighted graphical lasso method as introduced by Friedman et. al (2008). Details about the model and the estimation algorithm are provided in the next two sections.

2.3 The Model

2.3.1 The VAR model

The model of interest in this paper is the Vector Autoregressive model of order P denoted $VAR(P)$. We assume there are K time series that are centered (no intercept),

$$\mathbf{Y}_t = \boldsymbol{B}_1 \mathbf{Y}_{t-1} + \cdots + \boldsymbol{B}_P \mathbf{Y}_{t-P} + \mathbf{V}_t \tag{1}$$

with time observed on a regular grid where $\mathbf{Y}_t = (Y_{t,1}, \cdots, Y_{t,K})^{'}$ is a $K \times 1$ vector and $Y_{t,k}$ is the observation of the k^{th} time series $\{Y_{t,k}\}$ at time t. $\boldsymbol{B}_p$ is a fixed $(K \times K)$ coefficient matrix for $p = 1, \ldots, P$ and $\mathbf{V}_t = (V_{t,1}, \cdots, V_{t,K})^{'}$ is a $(K \times 1)$ vector of error terms. We assume that the error terms $\mathbf{V}_t$ follow a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$ and that they are independently and identically distributed. We also assume that the VAR is stationary.

The equation in (1) can be expressed as a multivariate regression model

$$\underbrace{\begin{bmatrix} \mathbf{Y}_T^{'} \\ \vdots \\ \mathbf{Y}_P^{'} \end{bmatrix}}_{\mathbb{Y} \in \mathbb{R}^{(T-P)\times K}} = \underbrace{\begin{bmatrix} \mathbf{Y}_{T-1}^{'} & \cdots & \mathbf{Y}_{T-P}^{'} \\ \vdots & \cdots & \vdots \\ \mathbf{Y}_{P-1}^{'} & \cdots & \mathbf{Y}_0^{'} \end{bmatrix}}_{\mathbb{X} \in \mathbb{R}^{(T-P)\times(PK)}} \underbrace{\begin{bmatrix} \boldsymbol{B}_1^{'} \\ \vdots \\ \boldsymbol{B}_P^{'} \end{bmatrix}}_{\mathbb{B} \in \mathbb{R}^{(PK)\times K}} + \underbrace{\begin{bmatrix} \mathbf{V}_T^{'} \\ \vdots \\ \mathbf{V}_P^{'} \end{bmatrix}}_{\mathbb{V} \in \mathbb{R}^{(T-P)\times K}}$$

which is equivalent to

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{V} \tag{2}$$

A common method for the estimation of $\mathbb{B}$ is conditional maximum log-likelihood, where the conditional variables are the lagged time series. The goal is to minimize the negative log-likelihood Gaussian function.

$$g(\Omega, \mathbb{B}) = \mathrm{Tr}\left[\frac{1}{(T-P)} (\mathbb{Y} - \mathbb{X}\mathbb{B})^{'} (\mathbb{Y} - \mathbb{X}\mathbb{B}) \Omega\right] - \log(|\Omega|) \text{ where } \Omega = \boldsymbol{\Sigma}^{-1}. \tag{3}$$

2.3.2 The spatial VAR with one layer

We now assume that each component of an observation at time t, $\mathbf{Y}_t = (Y_{t,1}, \cdots, Y_{t,K})^{'}$, corresponds to a response recorded at each of K different spatial units with coordinates s_k with $k \in \{1, \cdots, K\}$ and $s_i \in \mathbb{R}^2$. We use the notation $\mathbf{Y}_t = (Y_{t,1} = Y_{t,s_1}, \cdots, Y_{t,K} = Y_{t,s_K})^{'}$, where Y_{t,s_k} is the observation of the variable of interest at time t and at spatial unit s_k. In this setting, the precision matrix $\mathbf{\Sigma^{-1}}$of the error terms $\mathbf{V}_t = (V_{t,s_1}, \cdots, V_{t,s_K})$ has a certain degree of sparsity. In particular we assume that time series observed at sites that are far from each other are more likely to have a null entry in the precision matrix. Beyond this assumption, we do not make other structural assumptions such as isotropy or parametric shape. The resulting *one-layer VAR* model becomes

$$\begin{bmatrix} Y_{t,1} \\ \vdots \\ Y_{t,K} \end{bmatrix} = \begin{bmatrix} B_{11}^{(1)} & \cdots & B_{1K}^{(1)} \\ \vdots & \ddots & \vdots \\ B_{K1}^{(1)} & \cdots & B_{KK}^{(1)} \end{bmatrix} \begin{bmatrix} Y_{t-1,1} \\ \vdots \\ Y_{t-1,K} \end{bmatrix} + \cdots + \begin{bmatrix} B_{11}^{(P)} & \cdots & B_{1K}^{(P)} \\ \vdots & \ddots & \vdots \\ B_{K1}^{(P)} & \cdots & B_{KK}^{(P)} \end{bmatrix} \begin{bmatrix} Y_{t-P,1} \\ \vdots \\ Y_{t-P,K} \end{bmatrix} + \begin{bmatrix} V_{t,1} \\ \vdots \\ V_{t,K} \end{bmatrix} \tag{4}$$

For any $p \in \{1, \cdots, P\}$ and any $k, k' \in \{1, \cdots, K\}$, $B_{kk'}^{(p)}$ measures the effect of the observation $Y_{t-p,k'}$ at the spatial location $s_{k'}$ at a past time $t - p$ on the observation $Y_{t,k}$ at the location s_k.

2.3.3 The spatial VAR with multiple layers

The model described in Section 2.3.2 can be generalized to a setting with more than one layer. For instance, a typical geostatistical study involves the joint modeling of two economic indicators, unemployment and house prices, across counties in the US. In such a study, one might arbitrarily set employment rate to be the first layer and the house prices the second layer. More generally, assume that there are L layers, and that for each layer $J \in \{1, \cdots, L\}$, observations are acquired at spatial units $\{s_1^{[J]}, \cdots, s_{K_J}^{[J]}\}$ at discrete times $t \in \{0, \cdots, T\}$. For the layer J, we have K_J time series $\{Y_{t,s_1^{[J]}}^{[J]}\}, \cdots, \{Y_{t,s_{K_J}^{[J]}}^{[J]}\}$. The observation spatial units are not necessarily the same

across layers. Given that the notation in $\{Y_{t,s_k^{[J]}}^{[J]}\}$ involves many forms of indices, we re-express all the time series under the form $Y_t^{[ind]}$ where ind is an index unique to each time series. In what follows, we use the set of indices $D_J = (a_J, a_J + 1 \cdots, b_J)$ where $a_J = \sum_{j=1}^{J-1} K_j + 1$ and $b_J = \sum_{j=1}^{J} K_j$. The time series within the first layer $(J = 1)$ correspond to time series with indices in $D_1 = (1, \cdots, K_1)$ where $a_1 = 1$ and $b_1 = K_1$. The time series within the second layer $(J = 2)$ correspond to time series with indices in $D_2 = (K_1 + 1, \cdots, K_1 + K_2)$ where $a_2 = K_1 + 1$ and $b_2 = K_1 + K_2$. Generally, the vector of time series within the J^{th} layer becomes $\left(Y_t^{[a_J]}, \cdots, Y_t^{[b_J]}\right)$. The total number of time series in the model is $M = \sum_{j=1}^{L} K_j$. We apply the same transformation to the indices associated with the sites, so that, we can interchangeably use $s_1^{[J]}$ and s_{a_J} to denote the site where the first time series $Y_t^{[a_J]}$ in layer J is observed.

When we consider multiple layers the coefficient matrix $\mathbb{B}$ in (2) becomes $\mathbb{B} = [\boldsymbol{B}_{D_1 \cdot}, \cdots, \boldsymbol{B}_{D_L \cdot}]$ where $\boldsymbol{B}_{D_J \cdot}$ represents all the coefficients that affect the observations in layer J:

$$\boldsymbol{B}_{D_J \cdot} = \begin{bmatrix} B_{a_J 1}^{(1)} & \cdots & B_{a_J M}^{(1)} & \cdots & B_{a_J 1}^{(P)} & \cdots & B_{a_J M}^{(P)} \\ \vdots & & \vdots & & \vdots & & \vdots \\ B_{b_J 1}^{(1)} & \cdots & B_{b_J M}^{(1)} & \cdots & B_{b_J 1}^{(P)} & \cdots & B_{b_J M}^{(P)} \end{bmatrix}^T \in \mathbb{R}^{(PM) \times K_J} \tag{5}$$

For any column $\boldsymbol{B}_{i \cdot}$ of $\mathbb{B}$, if $i \in \{a_J, \cdots, b_J\}$ then $\boldsymbol{B}_{i \cdot}$ is a column of the matrix $\boldsymbol{B}_{D_J \cdot}$.

$$\boldsymbol{B}_{i \cdot} = \left(\boldsymbol{B}_{i \cdot}^{(1)'}, \cdots, \boldsymbol{B}_{i \cdot}^{(P)'}\right)' \in \mathbb{R}^{(PM) \times 1}$$

$\boldsymbol{B}_{i \cdot}^{(p)}$ is the sub-column of $\boldsymbol{B}_{i \cdot}$ that contains the coefficients associated with lag order p:

$$\boldsymbol{B}_{i \cdot}^{(p)} = \underbrace{(\underbrace{B_{i1}^{(p)}, \cdots, B_{iK_1}^{(p)}}_{1^{st} \text{ layer effect}}, \cdots, \underbrace{B_{ia_J}^{(p)}, \cdots, B_{ib_J}^{(p)}}_{\text{within layer effect}}, \cdots, , \underbrace{B_{ia_L}^{(p)}, \cdots, B_{ib_L}^{(p)}}_{L^{th} \text{ layer effect}})'}_{\text{Lag p effect}}$$

Next we introduce the estimation method used to estimate the VAR coefficients and the covariance matrix of the errors. We also describe the algorithms used to obtain

these estimates.

2.4 Estimation Algorithm

2.4.1 The methodology

In this section, we introduce the estimation method for one layer data followed by a description of how it extends to multiple layer data.

2.4.1.1 One layer sparse estimation

Spatio-temporal data exhibit statistical features that can be exploited to improve the efficiency of the model parameter estimates. Given the high-dimensional nature of the estimation problem, we need to impose some sparsity inducing constraints on the VAR coefficients $\mathbb{B}$ and potentially on the precision matrix Ω. As assumed in Bańbura et. al (2010) and Song et. al (2011), more recent temporal lags should be more predictive than the more distant lags. The second assumption, usually stated as the First Law of Geography, is that the observations collected at more distant spatial sites should be less influential on the observations collected at the site of interest. Given these constraints and assuming that the precision matrix Ω is known $\Omega = \tilde{\Omega}$, we solve the following optimization problem

$$\hat{\mathbf{B}} = \underset{\mathbb{B}\in\mathbb{R}^{(PK)\times K}}{\operatorname{argmin}} \left[g\left(\tilde{\Omega}, \mathbb{B}\right) + \lambda_1 \sum_{p=1}^{P}\sum_{i=1}^{K}\sum_{k=1}^{K} p^{\alpha} e^{\gamma\|s_i - s_k\|} |B_{ik}^{(p)}| \right] \tag{6}$$

where λ_1 is a penalty parameters and α, γ are lag and distance weight parameters respectively that are always strictly positive. The optimization problem in (31) is convex, since it is the sum of a convex objective function $g(\tilde{\Omega}, \mathbb{B})$ and of a convex penalty. As in Song and Bickel (2011), we account for the lag effect by penalizing more heavily the coefficients associated with observations that are more distant in time. Additionally, we account for the spatial effect by using penalties weighted by a function that depends on the distance function $e^{\gamma\|s_i - s_k\|}$, a similar idea is used in Lozano et. al (2010). For example, if we consider the lagged p time series $\{Y_{t-p}^{[u]}\}$

and $\{Y_{t-p}^{[v]}\}$ influencing the targeted time series $\{Y_t^{[i]}\}$, the penalty on the term $B_{iu}^{[p]}$ is higher than the penalty on the term $B_{iv}^{[p]}$ if $\|s_i - s_u\| > \|s_i - s_v\|$. To account for the lag and spatial effects we can use other penalty functions, for instance, $f(p) = (1 + \log(p))^\alpha$ or $f(p) = \exp(p)^\alpha$ for lag functions, in place of p^α. In this paper, we do not suggest that the penalty functions we chose are optimal. Identifying the optimal functions would considerably increase the number of tuning parameters.

2.4.1.2 Multi-layer sparse estimation

As presented in Section 2.3.3, the J^{th} layer is identified by the index set

$$D_J = \{a_J = \sum_{j=1}^{J-1} K_j + 1, \cdots, b_J = \sum_{j=1}^{J} K_j\}.$$

Let $\boldsymbol{B}_{i\cdot}$ of $\mathbb{B}$ be the column of coefficients corresponding to the time series $\{Y_t^{[i]}\}$ in layer J, meaning $i \in D_J$. The terms in column $\boldsymbol{B}_{i\cdot}$ can be rearranged in the following manner

$$\boldsymbol{B}_{i\cdot} = \{\boldsymbol{B}_{iD_1}, \cdots, \boldsymbol{B}_{iD_l}, \cdots, \boldsymbol{B}_{iD_L}\} \tag{7}$$

Each set of coefficients $\boldsymbol{B}_{iD_l}$ in (7) represents the effect from time series in the l^{th} layer on the time series of interest $\{Y_t^{[i]}\}$. For any layer $l \in \{1, \cdots, L\}$

$$\boldsymbol{B}_{iD_l} = \underbrace{(\overbrace{\underbrace{B_{ia_l}^{(1)}, \cdots, B_{ib_l}^{(1)}}_{1^{st}\text{ lag effect}}}^{\boldsymbol{B}_{iD_l}^{(1)}}, \cdots, \overbrace{\underbrace{B_{ia_l}^{(P)}, \cdots, B_{ib_l}^{(P)}}_{P^{th}\text{lag effect}}}^{\boldsymbol{B}_{iD_l}^{(P)}})}_{l^{th}\text{ Layer effect}} \tag{8}$$

The regularization scheme we propose for the estimation of the column $\boldsymbol{B}_{i\cdot}$ is the following

$$\underbrace{\lambda_1 \sum_{p=1}^{P} \sum_{k=1}^{M} p^\alpha e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}|}_{\text{Within Layer Penalty}} + \underbrace{\lambda_2 \sum_{\substack{l=1 \\ l \neq J}}^{L} \|\boldsymbol{B}_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}}_{\text{Between Layer Penalty}} \tag{9}$$

with $\boldsymbol{B}_{i\cdot}$ such that $i \in D_J$

$$\|\boldsymbol{B}_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} = K_l \begin{bmatrix} B_{iD_l}^{(1)} & \cdots & B_{iD_l}^{(P)} \end{bmatrix} \left(\mathbb{P} \otimes \Delta_l^{[i]} \right) \begin{bmatrix} B_{iD_l}^{(1)} \\ \vdots \\ B_{iD_l}^{(P)} \end{bmatrix} \text{ where}$$

$$\Delta_l^{[i]} = \begin{bmatrix} e^{2\gamma\|s_i - s_{a_l}\|} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{2\gamma\|s_i - s_{b_l}\|} \end{bmatrix} \text{ and } \mathbb{P} = \begin{bmatrix} (1)^{2\alpha} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P^{2\alpha} \end{bmatrix}$$

The first term in (9) is similar to the penalty in (31) for the one layer experiment. This penalty captures the within layer sparsity effect. As explained in Section 2.4.1.1, coefficients corresponding to the lead and lag effects for the time series closer in time and space are less penalized. The second term in (9) uses the group lasso penalty introduced by Yuan et. al (2006). We apply the group penalty to all the layers, except to the J^{th} layer that contains the coefficients $\boldsymbol{B}_{iD_J}$ linked to the time series $\{Y_t^{[i]}\}$ (with $i \in D_J$). By imposing the group sparsity scheme, we will select only the layers that have lag effects on the response time series $\{Y_t^{[i]}\}$. The J^{th} layer, which is not penalized by the group sparsity norm, is always selected, suggesting that we assume that own layer effect is always present.

In (9), we introduce the norm applied to the vector of coefficients $\boldsymbol{B}_{iD_l}$; this norm is used for specifying the between-layer penalty. The matrices $\Delta_l^{[i]}$ and $\mathbb{P}$ in (9) are designed to account for three important statistical features of the data. First, the term K_l quantifies the size of the group since it measures the number of observation spatial units. Consequently, the model applies heavier penalties on layers with more spatial units. Second, the lag and distance weights serve the same purpose as in the one layer case described in (31). Third, a layer with a set of distant sites should be penalized as a group higher than if the sites are nearby. This idea is conveyed through the use of the weight matrix $\tilde{\Delta}_l^{[i]}$.

This penalization scheme induces group-wise and within-layer sparsity. The group-wise penalty allows assessment of between-layer lead and lag relationships. The within-layer penalty will select the influential spatial units within the layers which have a lag influence on the variable of interest $Y_t^{[i]}$. If we assume that the precision matrix is known, then the problem we solve is the following:

$$\underset{\mathbb{B}}{\text{Min}} \left[g(\tilde{\Omega}, \mathbb{B}) + \lambda_1 \sum_{i=1}^{M} \sum_{p=1}^{P} \sum_{k=1}^{M} p^{\alpha} e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}| + \lambda_2 \sum_{i=1}^{M} \sum_{\substack{l=1 \\ l: i \notin D_l}}^{L} \|\boldsymbol{B}_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} \right] \tag{10}$$

Equation (10) is obtained by applying the penalties in (9) to each of the ith columns of the matrix $\mathbb{B}$. The summation of the within layer penalty is over all layers and over all spatial units within the layers. The between layer penalty is applied to all the layers but not to the layer containing the targeted time series.

2.4.1.3 Estimation of the precision matrix with spatial structure

We assume that the spatial covariance matrix is block diagonal, i.e. we only have within layer spatial dependence. Assuming that pairs of time series sampled from distant sites within the same layer are independent, we model each layer covariance using a distance weighted graphical lasso method. This idea was suggested by Friedman et al. (2008). We specify the amount of regularization to depend on the distance between two targeted time series. If we assume that the regression coefficients are known $\left(\mathbb{B} = \tilde{\mathbb{B}}\right)$ the following optimization problem is solved for each layer J.

$$\tilde{\Omega}_J = \underset{\Omega_J \in \mathbb{R}^{K_J \times K_J}:\ \Omega_J \succeq 0}{\text{argmin}} \; g\left(\tilde{\boldsymbol{B}}_{D_J.}, \Omega_J\right) + \lambda_3 \sum_{i=a_J}^{b_J} \sum_{k \neq i} e^{\|s_i - s_k\|} |\Omega_{ik}| \tag{11}$$

With the submatrix, $\boldsymbol{B}_{D_J.} = [\boldsymbol{B}_{a_J.}, \cdots, \boldsymbol{B}_{b_J.}]$ and Ω_J is the precision matrix associated with the J^{th} layer.

The second alternative method one could consider for the estimation of the precision matrix is to use parametric spatial covariance function. This method could be

used in cases where the precision matrix is not sparse.

2.4.1.4 Joint Estimation of the VAR coefficients and the precision matrix

To jointly estimate $\mathbb{B}$ and Ω in a multi-layer and spatial setting, we apply the idea introduced by Rothman et al. (2011), which reduces to penalized likelihood estimation including all regularization schemes introduced for estimation of the VAR coefficients and the covariance matrix. If we have no cross-layer spatial dependence in the errors, $\mathbb{B}$ and Ω are estimated by minimizing the L_1/L_2 regularized negative log-likelihood function g. We can decompose the large optimization problem in L optimization problems of smaller size. The J^{th} layer optimization problem becomes

$$\begin{aligned}\left(\tilde{\boldsymbol{B}}_{D_J\cdot}, \tilde{\Omega}_J\right) = \underset{\substack{\Omega_J \in \mathbb{R}^{K_J \times K_J},\ \Omega_J \succeq 0 \\ \boldsymbol{B}_{D_J\cdot} \in \mathbb{R}^{(PM)\times K_J}}}{\operatorname{argmin}} & g\left(\boldsymbol{B}_{D_J\cdot}, \Omega_J\right) \\ & + \lambda_1 \sum_{i=a_J}^{b_J} \sum_{p=1}^{P} \sum_{k=1}^{M} p^{\alpha} e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}| + \lambda_2 \sum_{i=a_J}^{b_J} \sum_{\substack{l=1 \\ l \neq J}}^{L} \|\boldsymbol{B}_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} \qquad (12) \\ & + \lambda_3 \sum_{i=a_J}^{b_J} \sum_{k \neq i} e^{\|s_i - s_k\|} |\Omega_{ik}|\end{aligned}$$

The problem presented in (12) is not convex, but we can alternatively solve for $B_{D_J\cdot}$ with Ω_J fixed at $\tilde{\Omega}_J$ as in (10), and solve for Ω_J with $\boldsymbol{B}_{D_J\cdot}$ as in (11). In the next section, we introduce the algorithms for solving these two convex optimization problems.

2.4.2 Computational algorithms

The algorithm for the estimation of the VAR coefficients borrows the idea from the block cyclical coordinate descent applied to sparse group lasso in a technical report by Friedman et al. (2010). The algorithm used for estimating the precision matrix is a modified graphical lasso introduced by Friedman et al. (2008).

2.4.2.1 Algorithm for the VAR coefficients

For any layer $J \in \{1, \cdots, L\}$, we define $Y_{D_J} \in \mathbb{R}^{(T-P)\times K_J}$ the set of time series within the J^{th} layer. If we assume that the precision matrix is set at $\tilde{\Omega}_J$, we need to solve a problem similar to problem (10):

$$\begin{aligned}
&\min_{\boldsymbol{B}_{D_J\cdot}} \operatorname{Tr}\left[\frac{1}{T-P}\left(Y_{D_J} - \mathbb{X}\boldsymbol{B}_{D_J\cdot}\right)'\left(Y_{D_J} - \mathbb{X}\boldsymbol{B}_{D_J\cdot}\right)\tilde{\Omega}_J\right] \\
&+ \lambda_1 \sum_{i=a_J}^{b_J}\sum_{p=1}^{P}\sum_{k=1}^{M} p^{\alpha} e^{\gamma\|s_i - s_k\|}|B_{ik}^{(p)}| + \lambda_2 \sum_{i=a_J}^{b_J}\sum_{\substack{l=1 \\ l:\ i\notin D_l}}^{L} \|\boldsymbol{B}_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}
\end{aligned} \tag{13}$$

If we have just one layer the algorithm used is similar to the MRCE of Rothman and Levina (2011). If we have more than one layer the algorithm used is inspired from the sparse group lasso of Friedman et al. (2010). We visit each column of the matrix $\boldsymbol{B}_{D_J}$, and apply a cyclical group coordinate descent procedure to all the coefficients within each column $\boldsymbol{B}_{iD_l}$ associated with layers l such that $l \neq J$. Further, for all the group of coefficients selected in the previous step, we again apply a cyclical coordinate descent to identify the non-null coefficients within the selected groups. The details and the derivations of the algorithm are presented in the supplemental material.

2.4.2.2 Algorithm for the joint estimation of the VAR Coefficients and the precision matrix

The algorithm used to solve problem (12) is the following:

For λ_1 and λ_2

- Set $\hat{B}^{(0)} = 0$ and use graphical lasso to solve L problems (11) $\tilde{\Omega}_J^{(0)} = \tilde{\Omega}_J\left(\hat{B}_{D_J\cdot}^{(0)}\right)$
- For each $l \in 1, \cdots, L$ compute $\hat{B}_{D_l\cdot}^{(m+1)} = \hat{B}_{D_l\cdot}\left(\tilde{\Omega}_l^{(m)}\right)$ by solving problem (10) with algorithm for VAR coefficients.
- Compute $\tilde{\Omega}_J^{(m+1)} = \tilde{\Omega}_J\left(\hat{B}_{D_J\cdot}^{(m+1)}\right)$ by using graphical lasso to solve (11)
- If $\sum_{j,k}\left|\hat{B}^{(m+1)} - \hat{B}^{(m)}\right| < \epsilon\sum_{j,k}\left|\hat{B}_{jk}^{Ridge}\right|$ stop, otherwise start new loop

• $\hat{B}^{Ridge}$ is the solution of the VAR obtained by using a ridge regression for each time series.

2.4.2.3 Selection of tuning parameters

As for any regularization method, achieving a satisfactory performance in terms of model selection and parameter estimation requires proper selection of the penalty parameters $\lambda = (\lambda_1, \lambda_2, \lambda_3)$. Additionally in our method, we need to assess the importance of the distance effects and the lag effects parameters, α and γ respectively. For this, we employ a computationally efficient approach, we use the Bayesian Information Criterion (BIC) introduced by Schwarz (1978) that minimizes

$$BIC = -2 \log L \left(\hat{B}_{\lambda,\alpha,\gamma} \right) + \log(T) df$$

where $\hat{B}_{\lambda,\alpha,\gamma}$ is the estimator associated with the tuning parameters λ, α and γ, $L\left(\hat{B}\right)$ is the maximum likelihood of the VAR model and *df* is the number of degrees of freedom approximated by the number of non-zero estimated parameters. Zhou et al. (2007) finds that if in a regression setting the rank of a design matrix is equal to the number of predictors then the degrees of freedom of the lasso is well approximated by the number of non null coefficients. The BIC criterion allows the determination of the optimal lag for each layer.

The use of BIC for non-convex regularized likelihood is advocated by Bulhmann and Van De Geer (2011) as a simple and computationally convenient method. However, there is no rigorous justification for the use of BIC in the context of regularized non-convex likelihood to date.

For comparison purposes, we also analyze the performance of the lasso and a modified lasso scheme that accounts for the distance between the sites. To select the penalization parameters for these two methods, we use the rolling prediction scheme used by Song and Bickel (2011) for consistency with the existing relevant papers.

2.5 Simulations

2.5.1 Simulations setup

We assess the performance of the method using two simulation experiments. We herein refer to our method as SMTSE (Sparse Multivariate Time Series Estimation). In the first experiment, we assume that the time series are observed for one type of measurement, i.e. one-layer data. In the second experiment, we generate time series from two distinct layers. For each experiment, we evaluate the model selection performance by assessing how well an estimation method captures the sparsity in the lag relationships using metrics such as the True Positive Rate (TPR) and the True Negative Rate (TNR). We measure the estimation performance using the Frobenius norm of the difference matrix between the true VAR matrix and the estimated VAR matrix. In the generative models described below for one and two layers, spatial dependencies are generated by selection of nearest spatial neighbours. This nearest neighbour dependency is not part of the VAR model described in Section 3 of the paper.

One layer simulations. We generate the simulated set of time series as described below:

1. Randomly generate K sites in a $[0,1] \times [0,1]$ square. We use a 2-dimensional uniform distribution to create the site locations.

2. Generate the VAR coefficients

   - Generate 1^{st} own lag coefficient for each time series
     $\forall i \in \{1, \cdots, K\}$, $B_{ii}^{[1]} \sim \text{Uniform}(a, b)$.
   - Randomly select the C_i closest neighbors to the site s_i of time series $\{Y_t^{[i]}\}$
     $C_i \sim \text{Binomial}(T_{neighbors}, P_{neighbors})$ where $T_{neighbors}$ is the maximum possible number of neighbors sites selected and $P_{neighbors}$ is the probability assigned to the binomial distribution.

- Coefficients associated with the C_i closest sites of the targeted site s_i are computed. We denote by S_i the set that contains the index of the C_i closest sites: $\forall j \in S_i$, $B_{ij}^{[1]} = B_{ii}^{[1]} * \exp\left(-\delta \|s_i - s_j\|\right)$. Note that δ is a term used to accentuate the decrease of coefficients associated with time series far from location s_i.

- Generate the coefficients associated with lags greater than 1: $\forall i, j \in \{1, \cdots, K\}^2, B_{ij}^{[l]} = l^{\eta} B_{[ij]}^{[1]}$, with $l > 1$ and $\eta < 1$,.

3. Set the error covariance matrix to: $\Sigma_{ij} = \rho^{\|s_i - s_j\|}$.

4. Simulate K time series of length T from VAR model with VAR coefficients $\mathbb{B}$ and error covariance matrix $\mathbf{\Sigma}$.

Two-layer simulations. To generate the two-layer simulated data, we apply a similar procedure as in the one-layer simulation experiment. Each layer consists of 25 sites. We alternate simulations in which the two layers have an effect on each other and simulations in which only layer one has an effect on layer two. Within-layer effects are always present in all simulations. The covariance matrix for the two layers experiment has a block diagonal structure, with each block defined by $\Sigma_{ij} = \rho^{\|s_i - s_j\|}$

Simulation settings Throughout all simulations we set fixed the following parameters:

• The lower and the upper bounds for the own lag coefficients: $a = -0.5$ and $b = 0.5$.

• The number of sites: $K = 25$.

• The maximum number of influential neighbors for each site: $T_{neighbors} = 5$.

• The probability for the generation of influential neighbors for each site: $P_{neighbors} = 0.8$.

• To reduce the computational cost, the temporal and spatial penalty tuning parameters are set to $\alpha = 1$ and $\gamma = 1$ for all settings.

We vary other parameters including the number of true lags and the variance of the errors. The different simulation settings are:

• **Simulation Settings 1 & 2:** Number of layers $L = 1$, lag order $P = 2$, error covariance level $\rho_1 = 0.1$ (simulation 1), $\rho_2 = 0.7$ (simulation 2). The search for the optimal regularization parameters is performed on the following grid $\lambda_1 = \{1, 10, 20, \cdots, 100\}$, so λ_1 varies by increments of 10, and $\lambda_3 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. After finding the parameter (λ_1, λ_3) that minimize the BIC criterion, we perform a second search on a refined grid around the previous minimum.

• **Simulation Settings 3 & 4:** Number of layers $L = 2$, lag order $P = 1$, error covariance level $\rho_2 = 0.1$ (simulation 3), $\rho_4 = 0.4$ (simulation 4). For the two layers experiments, the regularization parameters are searched in the following set of values, $(\lambda_1, \lambda_2) \in \{1, 10, 20, \cdots, 100\}^2$, and $\lambda_3 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$.

To test the performance of the SMTSE, for each simulation setup, we generate 50 different replications with time series of length $T = 300$. For each simulation setup, we apply the estimation methods assuming 2, 3 or 4 lags for one-layer simulations and 1, 2, 3 or 4 lags for two-layer simulations. We report the following metrics:
$TP = \frac{\#\left[(i,j):\hat{B}_{ij}\neq 0 \text{ and } B_{ij}\neq 0\right]}{\#[(i,j):B_{ij}\neq 0]}$, the true positive rate measuring the ability of a model to capture non null VAR coefficients .
$TN = \frac{\#\left[(i,j):\hat{B}_{ij}= 0 \text{ and } B_{ij}= 0\right]}{\#[(i,j):B_{ij}= 0]}$, the true negative rate measuring the ability of a model to identify null VAR coefficients.
$FE = \sqrt{\sum_{i,j}\left(B_{ij} - \hat{B}_{ij}\right)^2}$, the frobenius norm error measuring the estimation error of the VAR coefficients.

2.5.2 Simulation results

Figures 1 and 2 summarize our findings. In Sub-figures 1(a) and 2(a), we report the true positive rates of SMTSE, of the lasso and of the spatial lasso; in Sub-figures 1(b) and 2(b), we report the true negative rates; and in Sub-figures 1(c) and 2(c), we

report the Frobenius norm. The dark curves are the results obtained for the simulation settings 1 & 3, and the others are for the simulation settings 2 & 4 averaged over the 50 replications. Based on these simulations results, we find:

• SMTSE outperforms the lasso and the spatial lasso for strong and weak error covariance.

• As we increase the number of lags, the true positive rates decrease for all the methods.

• When the error covariance is weak ($\rho = 0.1$), the three methods have similar performances in terms of identification of the non-null VAR coefficients. However, when the level of the error covariance increases ($\rho = 0.7$), our method identifies the non null zero coefficients with much higher accuracy.

• Our method is not significantly sensitive to an increase in the level of the error covariance, this result validating that our estimation procedure improves the efficiency of the VAR coefficients estimates. On the other hand, the Lasso and the Spatial Lasso VAR coefficient estimates are extremely sensitive to the error covariance level since they do not model the covariance structure of the noise.

• In the two-layer setting, whether the error covariance level is strong or weak, our model performs even better (comparatively to Lasso and Spatial Lasso) than in the one layer case. This is because the group penalty excludes many false positives.

• We also study the predictive performance of all the methods considered. For each set of simulations, we use the generative models described above, this time each time series has a length $T = 350$, and we leave 50 points for out-of-sample forecasting. The h-step ahead forecast for time series $\{Y_t^{[i]}\}$ given all the information up to time t ($I(t)$) is $\hat{Y}^{[i]}_{t+h|I(t)}$. The h-step ahead root mean square error for each time series is computed as

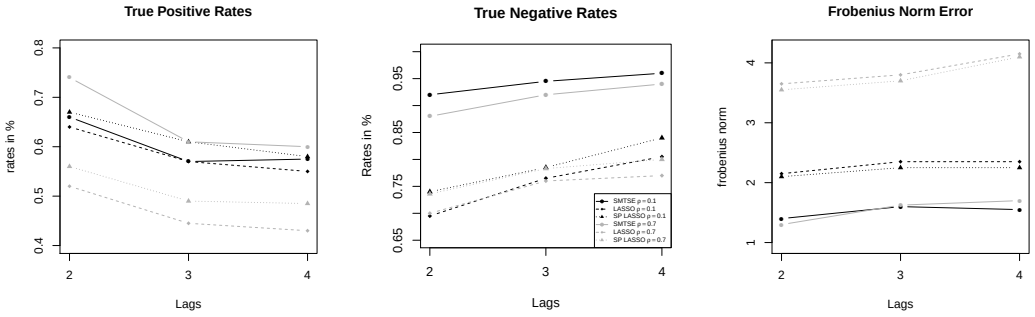
$$\text{RMSE}_i^{(h)} = \left[\frac{1}{(50)} \sum_{t=300-h}^{350-h} \left(\hat{Y}^{[i]}_{t+h|I(t)} - Y^{[i]}_{t+h} \right)^2 \right]^{\frac{1}{2}}.$$

Figures 3 presents the box plots of the accuracy of the out-of-sample prediction measured the root mean squared errors (RMSEs) for the true model, the sparse multivariate time series estimation (SMTSE), the Lasso, the Spatial Lasso (SP LASSO) and the Ordinary Least Squares (OLS). Under all the simulation settings, we observe the SMTSE has a RMSE slightly lower than the RMSE of the Lasso and the RMSE of the Spatial Lasso. The Ordinary Least Squares as expected overfits the model and yields poor out-of-sample forecasts. In Section B of the supplemental material, we report the forecasting performance of all these methods when the number of lags used for estimation is larger than the true number of lags. The proposed method remains competitive when compared to the Lasso and the Spatial Lasso, and the OLS RMSE increases due to overfitting. In Table 1, we report the lag number selected by AIC and BIC criteria under OLS and the lag number selected by our model in average over the 50 replications. We find that the lag is accurately identified using BIC in our method.

- Following the use of OLS + AIC and OLS + BIC introduced by Hsu et al. (2008), we also fitted the simulated models with OLS (results not reported here); as expected this method doesn't introduce sparsity in the VAR coefficient matrices and the Frobenius norm error is on average significantly higher than the values obtained for the other three methods aforementioned. The true positive rate is 1, but the true negative rate is 0.

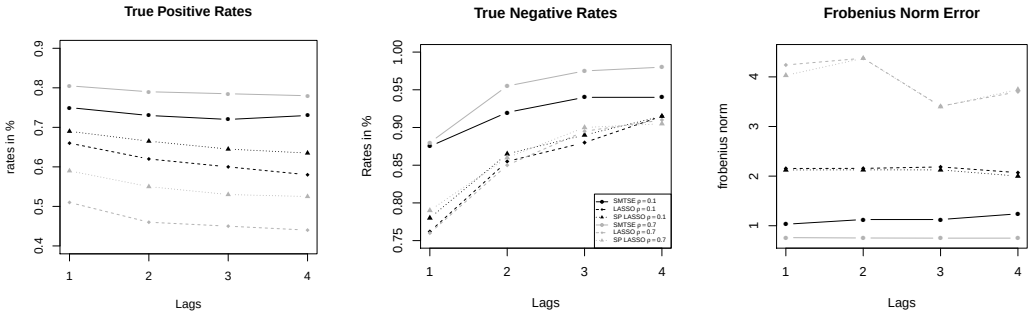
Table 1: Average of the lag selected with OLS + AIC, OLS + BIC and SMTSE for simulated VAR models over 50 replications of simulated data.

Simulation	OLS + AIC	OLS + BIC	SMTSE
1 $VAR(2), \rho = 0.1$	2.00	1.08	2.06
2 $VAR(2), \rho = 0.7$	2.00	1.9	2.08
3 $VAR(1), \rho = 0.1$	1.00	1.00	1.00
4 $VAR(1), \rho = 0.4$	1.00	1.00	1.00



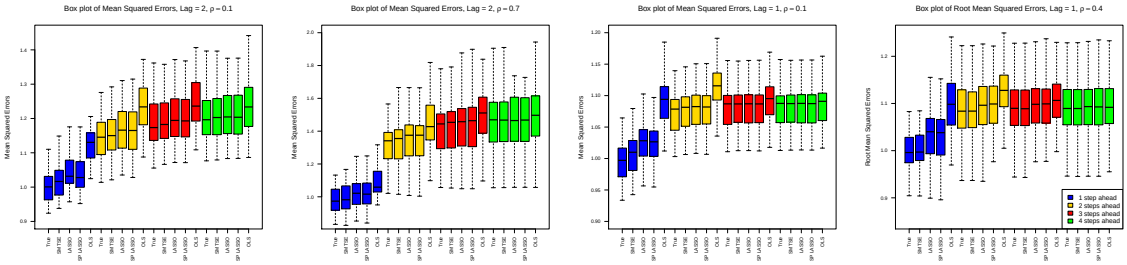
(a) True Positive **(b)** True Negative **(c)** Frobenius Norm Error

Figure 2: Performance metrics for one layer experiment



(a) True Positive **(b)** True Negative **(c)** Frobenius Norm Error

Figure 3: Performance metrics for two layers experiment



(a) Simulation 1 **(b)** Simulation 2 **(c)** Simulation 3 **(d)** Simulation 4

Figure 4: Prediction mean squared error for one and two layers experiment

2.6 Application

In the case study introduced in this section, we consider the time series of construction employment (thousands of persons), and the number of new private housing units authorized by building permits in the United States. Both layers of time series

are observed at the state level and on a monthly basis; they are seasonally adjusted. These time series are collected using the Geographical Economic Data (GEOFRED) of the Federal Reserve Bank of St. Louis. We first removed the time series of certain states because of missing observations. Specifically, the states of Delaware, District of Columbia, Hawaii, Maryland and Nebraska are not present in the employment dataset, the states of South Dakota and Tennessee have 38 consecutive months (January 2008 to February 2011) with missing observations. The states of District of Columbia and Hawaii are not present for the building permits time series. The observations for the two economic measurements span from April 1996 to March 2012; we leave out data from April 2012 to April 2013 for out of sample forecasting. The number of time points is $T = 192$ for a total of 93 time series.

Before applying the three estimation methods, we standardize all time series. To find the optimal set of regularization parameters and the optimal lags, we perform an extensive search over a grid based on the following values $\lambda_1, \lambda_2 \in \{0.1, 10, 20, \cdots, 100\}$, $\lambda_3 \in \{0.01, 0.1, 1, 10\}$; the lags considered for this analysis are $P \in \{1, 2, 3\}$. The values of λ_1 and λ_2 vary by increments of 10, while the parameter λ_3 varies on a log-scale. Additionally, we consider $\alpha \in \{0.1, 1\}$ to accomodate the possibility of a strong or weak temporal decay effect and $\gamma \in \{0.1, 0.01\}$, to scale the spatial distances between the states. The optimal tuning parameters are $\lambda_1 = 30, \lambda_2 = 1, \lambda_3 = 0.01$ for the employment layer, $\lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 0.01$ for the building permit layer , and for both layers the temporal tuning parameter is $\alpha = 1$ and the spatial tuning parameter is $\gamma = 0.01$.

We herein refer to our method as SMTSE(Sparse Multivariate Time Series Estimation).The results from the implementation of the SMTSE are presented in Figure 2(a). The horizontal axis of this figure represents the time series Y_t, while the vertical axis represents the lags. For instance, the first column contains the coefficients affecting the employment rate in the state of Connecticut. The first horizontal black

separation is the first lag effects of employment, the second horizontal black separation is the first lag effects of building permits. The left half (right half) of the matrix contains the coefficients that influence the employment rates (building permits). The vertical black line separates the employment time series from the building permit time series. The states are grouped in 10 economic regions based on their proximity to each other. The grey lines in the resulting coefficient matrix are used to delimit the economic regions.

The SMTSE finds that only one lag is needed for describing the lead and lag relationships among the employment time series, while it suggests a higher degree of persistence for the building permit time series as three lags are selected using BIC. This finding points to the fact that we can select a different number of lags for each layers. This is possible because of the divide-and-conquer approach we adopt.

For our method, the significant VAR coefficients tend to gravitate within the diagonal blocks of each economic region. In contrast, the Lasso (Figure 2(b)) tends to introduce small but non-null AR coefficients for the time series within the two layers. The spatial lasso (Figure 2(c)) is able to eliminate these small but non-null coefficients. Our method and the spatial lasso provide sparser models than the lasso.

For the SMTSE, the own lag effects are positive for all time series. The own lag coefficients for the employment time series are very high while for the building permit time series, the coefficients are smaller although still significantly greater than 0. Moreover, for the building permit time series, the own lag coefficients are slowly decreasing as the number of lags increases implying that these time series are persistent. Additionally, our model uncovers the effects of the building permits on construction employment suggesting that the number of building permits issued for new houses is a leading indicator in the housing industry. Therefore, if the number of building permits increases, it is plausible to expect a rise in the construction employment. Our model also reveals the absence of feedback effect of the employment time series on

the number of building permits issued.

The lasso and the spatial lasso are both unable to provide similar results. The lasso introduces small and noisy estimates for some VAR coefficients and does not explicitly capture the lack of effect of employment on the number of building permits issued. The spatial lasso is able to remove the small and noisy VAR coefficients, but is not able to identify that employment does not lead the building permits issued. The main reason why our method properly identifies the relationships between these two layers lies in the presence of the group penalties that uniformly remove all the coefficients associated with a potential feedback effect from employment time series. In Figure 2(a), we see that the blocks of coefficients that capture the effect of employment on building permit (on the right side of the black vertical line) are all null. But in Figures 2(b) and 2(c), we observe the presence of some coefficients of small magnitudes in these blocks. The coefficient matrices for the lasso, spatial lasso with 1 and 2 lags are presented in the supplemental material. They have a behavior similar to the their counterparts with 3 lags. We also show the coefficient matrices of the OLS with 1 and 2 lags, some of the coefficients in these matrices are very large (order of magnitude of 40 for OLS with 2 lags).

We also report the out-of-sample forecast performance of the three methods. The h-step ahead forecast for time series $\{Y_t^{[i]}\}$ given all the information up to time t ($I(t)$) is $\hat{Y}^{[i]}_{t+h|I(t)}$. The h-step ahead root mean square error for each time series is computed as

$$\text{RMSE}_i^{(h)} = \left[\frac{1}{(50)} \sum_{t=300-h}^{350-h} \left(\hat{Y}^{[i]}_{t+h|I(t)} - Y^{[i]}_{t+h} \right)^2 \right]^{\frac{1}{2}}.$$

The forecast period is from the month of April 2012 to April 2013. We use the RMSE to measure the prediction performance. As seen in Figure 3, the OLS fitted with 2 lags has the worst out-of-sample forecast for all the lag levels, and the OLS with 1 lag has the second worst performance. This poor performance can be explained by the fact that OLS commonly overfits. In contrast, the lasso generates out-of-sample

forecasts that are less accurate than the forecasts resulting from the SMTSE; this is probably due to the presence of a large number of spurious VAR coefficients. The spatial lasso and the lasso have similar forecasting performance. These results imply that simply incorporating spatial distances in the lasso penalty doesn't improve the predictions. But if we also account for the contemporaneous effects through the estimation of the precision matrix, the prediction errors become significantly smaller than the prediction errors associated with the other regularized methods and the ordinary least squares. In Table 1(a), we report the number of non null coefficients in each of the simulated models, we see that the SMTSE yields the second most sparse model and is still able to outperform the other methods in terms of prediction. Table 1(b), shows some typical computational time needed to solve the SMTSE under a very sparse (λ large) and very dense (λ small) settings with **R**, on a 1.80Ghz Intel Xeon Linux computer.

Throughout other experiments not reported here, we found that if we increase the sample size T, the OLS can produce predictions that are more accurate than all the regularized methods including the SMTSE. This can be explained by the fact that the L_2 norm of the prediction error of regularized methods such as the lasso has an upperbound that depends on the inverse of magnitude of the restricted eigenvalue of the matrix $\frac{\mathbf{X^T X}}{\mathbf{n}}$. So the OLS prediction performance could be superior (despite overfitting) to the lasso prediction performance if some compatibility conditions hold for a small restricted eigenvalue (Bülhmann et al., 2011).

2.7 Conclusion

In this paper, we propose a L_1/L_2 penalized likelihood method for estimating large sparse VAR models of time series, which are spatial interdependent. The methodology explicitly accounts for sparsity, group sparsity and spatial contemporaneous correlation among the time series. We also presented algorithms for solving the L_1/L_2

Table 2: Number of non-null coefficients for fitted models and computational time for algorithm under different values for the regularization parameters and with one lag

(a) Number of non-null coefficients

Model	number of non-null coefficients
SMTSE	317
LASSO 1	1845
LASSO 2	1284
LASSO 3	1663
SP LASSO 1	444
SP LASSO 2	289
SP LASSO 3	335
OLS 1	8649
OLS 2	17298

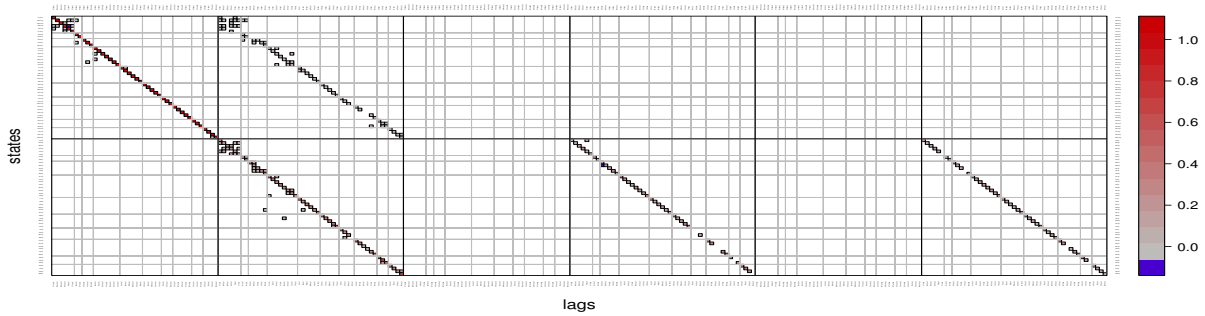
(b) Computational time

Models	λ_1	λ_2	λ_3	α	γ	Time (s)
1	1	1	1	0.1	0.01	49.63
2	1	1	1	1	0.01	48.79
3	1	1	1	0.1	0.1	07.60
4	1	1	1	1	0.1	07.56
5	100	100	10	0.1	0.01	06.41
6	100	100	10	1	0.01	06.20
7	100	100	10	1	0.1	06.17
8	100	100	10	0.1	0.1	06.15

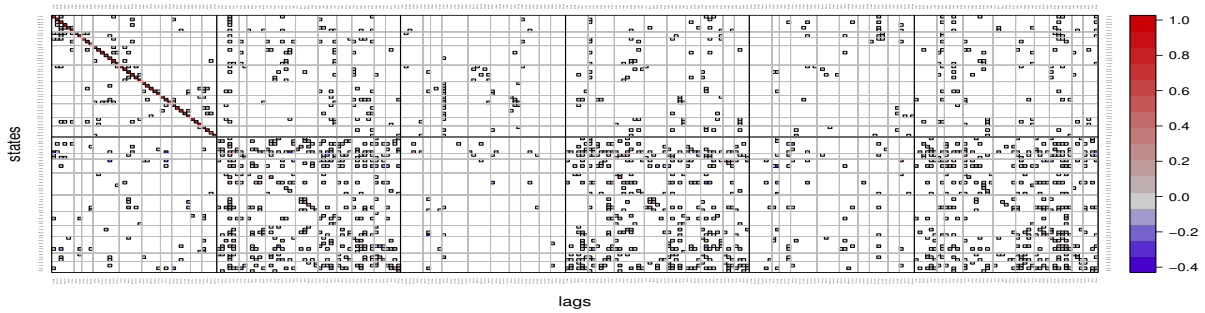
constrained optimization problems obtained after penalizing the VAR coefficients and the error precision matrix.

We performed extensive simulations to evaluate the performance of the proposed method (SMTSE) in comparison with existing approaches. We found that the SMTSE outperforms OLS, lasso and spatial lasso in recovering sparse VAR structures, in estimating the VAR coefficients, and in forecasting future values of the time series (especially, when the time series length is smaller than the number of time series). Importantly, the identification of the sparse VAR structure improves when applying lag and distance weighted penalties to the VAR coefficients, and by penalizing VAR coefficients at the group (specified by layers) level and within groups.

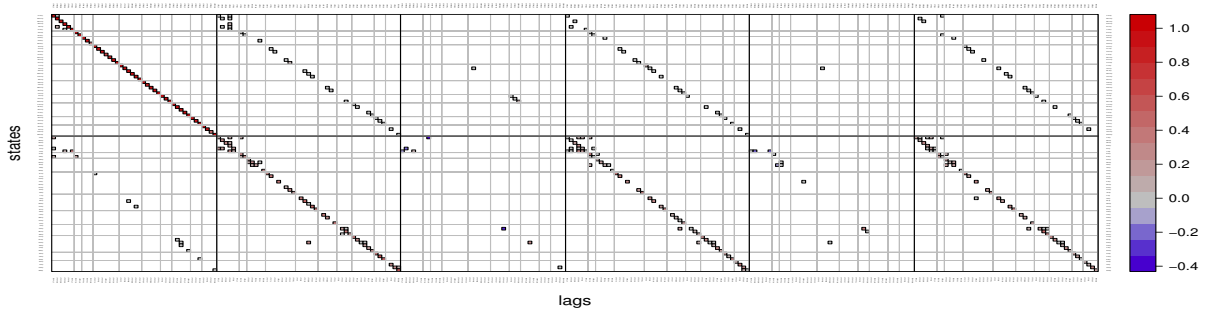
Theoretical properties justifying these results are not presented in this manuscript. Wonyul et al. (2012) consider the joint estimation of a coefficient matrix B and of a precision matrix Σ^{-1} in a multivariate regression problem. They use a doubly penalized joint likelihood with penalties on entries of B and Σ^{-1}. They show the consistency and sparsitency properties of estimates obtained by alternatively solving for B and Σ^{-1}. These theoretical justification could be extended to demonstrate the performance of the proposed method SMTSE.



(a) SMTSE Estimation



(b) LASSO Estimation



(c) Spatial LASSO Estimation

Figure 5: VAR matrix coefficients for employment and building permit time series obtained through (a) SMTSE, (b) LASSO, and (c) Spatial LASSO

Furthermore, the applicability of our method is illustrated by analyzing the relationship between construction employment rates and the number of new private housing units authorized by building permit. The method yields an interpretable model that matches economic intuition.

In both the simulation and application studies covered in this paper, the layers are clearly delineated. One of the reviewers however suggested that the method could

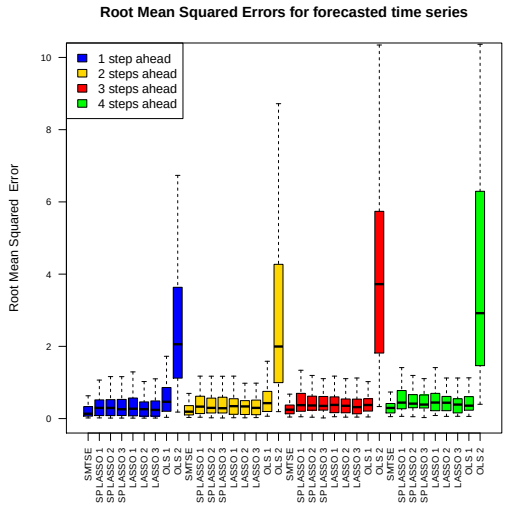


Figure 6: The h-step ahead forecast root mean square error (RMSE) for the Sparse Multivariate Time Series Estimation (SMTSE) method, for Lasso fitted with 1 to 3 lags, Spatial Lasso fitted with 1 to 3 lags, and OLS fitted with 1 to 2 lags. Forecast period T_0 = April 2012 to T_1 = April 2013. From left (1 step ahead RMSE) to right (4 steps ahead RMSE).

be extended to applications with layers that are not necessarily clearly defined. The first step could consist of the estimation of a graphical or cluster model to identify the layers in the data.

CHAPTER III

HIGH DIMENSIONAL MULTIVARIATE ADDITIVE MODELS

3.1 Introduction

The recent improvements of Electronic Health Records (EHR) have lead to the availability of patients level data. These data contain information about a diverse set of events. We are particularly interested in understanding what factors affect the cost of healthcare. To achieve our goal, we use patients level claim data collected in the state of Georgia from 2005 to 2009. Many problems associated with the analysis of cost in healthcare have to deal with certain statistical challenges, that are not directly addressed by existing methodologies. First, the number of possible factors explaining the rising cost of healthcare is relatively large, so it could be useful to apply model selection methods to isolate the most important factors. Second, to be able to make inferences valid within a certain spatial and temporal scale, it is sometimes important to aggregate the patient level data to zipcode or county level and to monthly or quaterly frequency. The aggregations further reduce the sample size of the data and increase the ratio of predictors to sample size and exacerbate the importance of using model selection techniques. The third challenge associated with the statistical analysis of health care cost lies in the presence of nonlinear relationships between claim data and meaningful predictors. In order to address these challenges, it is important to design statistical models that are more complex than model commonly used in this field, e.g: linear regression models. To simultaneously account for the 3 statistical problems cited above, we design a new estimation and model selection method that can be used in high dimensional settings.

Estimation and model selection in a high dimensional setting is a common statistical problem encountered in several fields, including biostatistics, genomics, imaging. The Lasso introduced by Tibshirani et al. (1996) is one of the most popular method for sparse high dimensional estimation and model selection in linear models. Even though linear models are able to capture the most important effects, nonparametric methods such as additive models can provide improvements, by detecting substantial nonlinear effects. Additive models are nonlinear regression functions, where each additive term is a smooth function depending on a single covariate.

In this nonparametric setting, the goal is to perform model selection on the additive terms while also controlling the smoothness of the selected additive components. Lin et al. (2006), Meier (2008) and Ravikumar (2009) , each proposed different methods to solve this problem. None of these approaches are designed to efficiently estimate multi-task sparse additive models simply because they are designed to solve problems with a univariate response. Some methodologies have been proposed to nonparametric multivariate regressions. Liu et al. (2008) adapt the ℓ_∞ regularization method of Turlach (2005), that can perform model selection in the parametric setting to functional model selection. Foygel et al. (2013) proposes a nonparametric reduced rank regression that generalizes reduced rank regression for linear models. Using these methods could lead to selection errors, because once a predictor is selected for a response, it appears as influential for other responses.

Obozinski et al. (2011), designed a parametric regularization method to perform union support recovery for high dimensional multivariate regression. Their method exploits a $\ell_1\backslash\ell_2$-norm that can impose joint sparsity on a group of coefficients. Each group is made of all the coefficients associated with the effect of one covariate on all the responses. In this paper, we generalize this approach to additive regression models. In our study, each group contains all the additive components associated with a covariate. We use the Hilbert spaces $\ell_1\backslash\ell_2$ norm, introduced by Yin et al.

(2012); this norm can induce group sparsity among the additive components.

The contributions of our work include a necessary and sufficient condition for the identification of covariates that are active in at least one of the regression problems, a block coordinate descent algorithm that leads to a sparse backfitting procedure applied to our multitask / multi-responses regression models, a set of extensive simulations that show the superior performance of our methodology and two applications that highlight how the method proposed can be used in different fields. In addition to the analysis of medicaid coast in Georgia, we use the method on a gene microarray dataset to identify biomarkers and to perform tumor classification on 83 cancer patients.

3.2 Background

Let's first introduce some notations, vectors and matrices are denoted by boldfaced letters, hat are added for estimates. If X is a random variable with distribution P_X, and f is a function of x, its $L_2(P_X)$ norm is $\|f\|^2 = \int_{\mathcal{X}} f^2(x) dP_X = \mathbb{E}(f^2)$. For a vector $\mathbf{x} = (x_1, \cdots, x_n)$, the ℓ_2-norm is defined as $\|\mathbf{x}\|_n^2 = \frac{1}{n}\sum_{i=1}^n x_i^2$. For a random variable X_j with $j \in \{1, \cdots, p\}$, $\mathcal{H}_j$ denotes the Hilbert subspace $L_2(P_{X_j})$ of P_{X_j}-measurable functions $f_j(x_j)$ of the single scalar variable X_j with zero mean, $\mathbb{E}[f_j(X_j)] = 0$. The inner product on $\mathcal{H}_j$ is $< f_j, g_j >= \mathbb{E}[f_j(X_j)g_j(X_j)]$, and the associated norm is $\|f_j\| = \sqrt{\mathbb{E}[f_j^2(X_j)]} < \infty$. $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_p$ denotes a Hilbert space of functions of $(x_1, \cdots, x_p)$, that have an additive form $m(\mathbf{x}) = \sum_{j=1}^p f_j(x_j)$, with $f_j \in \mathcal{H}_j$. We consider a K multitask (or multi-response) regression problem. For each regression model, we have a response variable $Y^{(k)}$ and a random vector of covariates $\mathbf{X}^{(k)} = \left(X_1^{(k)}, \cdots, X_p^{(k)}\right)$, with $k \in (1, \cdots, K)$. For each model, we define the set of smooth functions $\mathbf{f}_j(\mathbf{X}_j) = \left(f_j^{(1)}(X_j^{(1)}), \cdots, f_j^{(K)}(X_j^{(K)})\right)$ associated with the covariates $\mathbf{X}_j = \left(X_j^{(1)}, \cdots, X_j^{(K)}\right)$. To simplify our notations, $\forall j \in \{1, \cdots, p\}$ and $k \in \{1, \cdots, K\}$ we will write the function $f_j^{(k)}(X_j^{(k)})$ as $f_j^{(k)}$.

In this section, we give a brief description of existing methods, used to tackle high-dimensional nonparametric regression. These methodologies will be used as a building block for our method. If we consider a random vector $\mathbf{X} = (X_1, \cdots, X_p)$ and a random variable Y. A typical statistical problem is the nonparametric estimation of of nonlinear regression models.

$$Y = m(\mathbf{X}) + \varepsilon, \text{ where } \mathbb{E}(\varepsilon) = 0$$

The data available for this estimation problem are $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, with $i \in \{1, \cdots, n\}$. The regression function is given as the posterior mean of the response variable.

$$m(\mathbf{X}) = m(X_1, \cdots, X_p) = \mathbb{E}[Y|X_1, \cdots, X_p]$$

As presented in Yin et al. (2012), when p = 1, $m(\mathbf{X}) = PY$, where P is a conditional expectation operator $P = \mathbb{E}[.|X]$ used to orthogonally project Y onto a linear space of all measurable functions of X. The function m at point x, is estimated using kernel smoothers.

$$\hat{m}(x) = \sum_{i=1}^{n} s_i(x) y_i = s(x)^T \mathbf{y}$$

where $s_i(x) \propto K_h(|x_i - x|)$ and K_h is a kernel smoother. So for a response vector $\mathbf{y} \in \mathbb{R}^n$, the estimated values are given by $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where $\mathbf{S}$ is the smoother matrix $l_j(x_i)$ with $i, j = 1, \cdots, n$. $\mathbf{S}$ is a natural estimator of P, that will be used in additive models.

Hastie and Tibshirani (1986), proposed additive models as a class of nonlinear regression models.

$$m(X_1, \cdots, X_p) = \alpha + \sum_{j=1}^{p} f_j(X_j)$$

, where f_j are univariate additive components. We will assume without loss of generality that $\alpha = 0$ and that $\mathbb{E}(f_j(X_j)) = 0$. To estimate $m(\mathbf{X})$ in this setting, they

solve the optmization problem below:

$$\underset{\mathbf{f}:f_j\in\mathcal{H}_j}{\text{Min}}\, L(\mathbf{f}) = \frac{1}{2}\mathbb{E}\left[\left(Y - \sum_{j=1}^{p} f_j(X_j)\right)^2\right]$$

They propose to use a backfitting approach to estimate the smooth functions f_j

$$f_j = \mathbb{E}\left[\left(Y - \sum_{l\neq j} f_l\right)|X_j\right] = P_j\left(Y - \sum_{l\neq j} f_l\right)$$

. As stated earlier the natural estimator of P_j, is S_j, so the estimate of each smooth function under this setting is $\hat{f}_j \leftarrow S_j(Y - \sum_{l\neq j} f_l)$. Ravikumar et al. (2009) extends this model to the high-dimensional setting ($p \gg n$) and creates a method labelled SpAM (Sparse Additive Models). They impose sparsity at the function level and thus are able to perform model selection by finding the active additive components. They solve the optimization problem:

$$\underset{\mathbf{f}:f_j\in\mathcal{H}_j}{\text{Min}}\, L(\mathbf{f}) + \lambda\Omega(\mathbf{f}),$$

with $\Omega(\mathbf{f}) = \sum_{j=1}^{p} \|f_j\|$. They use a sparse backfitting algorithm where at each step the smooth function updates are given by:

$$f_j = \left[1 - \frac{\lambda}{\|P_j R_j\|}\right]_+ P_j R_j,$$

where $R_j = Y - \sum_{l\neq j} f_l$ is a partial residual and the operator $[.]_+ = \max(0, .)$. More recently Yin et al. (2012) proposed the GroupSpAM, that adapts the group lasso to the nonparametric setting. They penalize the additive components in a group manner. For a partition $\mathcal{G}$ of $\{1, \cdots, p\}$, they solve the following optimization problem:

$$\underset{\mathbf{f}:f_j\in\mathcal{H}_j}{\text{Min}}\, L(\mathbf{f}) + \lambda\Omega_{\mathbf{group}}(\mathbf{f}),$$

where the penalty $\Omega_{\mathbf{group}}(\mathbf{f})$ is defined as

$$\Omega_{\mathbf{group}}(\mathbf{f}) = \sum_{g\in\mathcal{G}} \sqrt{d_g}\|\mathbf{f}_g\| = \sum_{g\in\mathcal{G}} \sqrt{d_g}\sqrt{\sum_{j\in g} \mathbb{E}\left[f_j^2(X_j)\right]},$$

where $\mathbf{f}_g$ is a set of functions in group g. In their paper they give a block coordinate descent algorithm to update the group of additive functions.

We now introduce our methodology, and exploits the use of the $\ell_2\backslash\ell_1$ functional norm to identify in a multi-task or multivariate regression model, the covariates whose functions are active for at least one of the responses.

3.3 $L_2\backslash L_1$ joint functional sparsity

As stated in the background section, we consider a K multi-task (or a multivariate) regression problem. For each regression problem, we have the following data $\{(\mathbf{x}_i^{(k)}, y_i^{(k)}), i = 1, \cdots, p, k = 1, \cdots, K\}$. we will assume that $n_1 = \cdots = n_K = n$, and that for each model the response variables $Y^{(k)}$ and the covariates $\mathbf{X}^{(k)}$ are standardized. Each model has the form:

$$Y^{(k)} = \alpha^{(k)} + \sum_{j=1}^{p} f_j^{(k)}(X_j^{(k)}) \tag{1}$$

As presented in the case with one response, we assume that $\forall k \in \{1, \cdots, K\} \alpha^{(k)} = 0$ and that $\forall k \in \{1, \cdots, K\}$ and $\forall j \in \{1, \cdots, p\}$ $\mathbb{E}\left[f_j^{(k)}(X_j^{(k)})\right] = 0$. We will find the set of functions $\mathbf{f}_j = \left(f_j^{(1)}, \cdots, f_j^{(K)}\right)$ or $\mathbf{f}^{(k)} = \left(f_1^{(k)}, \cdots, f_p^{(k)}\right)$, by solving the optimization problem:

$$\underset{\mathbf{f}^{(1)}, \cdots, \mathbf{f}^{(K)}}{\text{Min}} \sum_{k=1}^{K} L(\mathbf{f}^{(k)}) + \lambda \Omega_{\ell_{\mathbf{2}}\backslash\ell_{\mathbf{1}}}(\mathbf{f}) \tag{2}$$

With $\mathbf{f} = \left(f^{(1)}, \cdots, f^{(K)}\right)$ and $L(\mathbf{f}^{(k)}) = \frac{1}{2}\mathbb{E}\left[\left(Y^{(k)} - \sum_{j=1}^{p} f_j^{(k)}\right)^2\right]$ and

$$\Omega_{\ell_{\mathbf{2}}\backslash\ell_{\mathbf{1}}}(\mathbf{f}) = \sum_{j=1}^{p} \sqrt{K} \sqrt{\sum_{k=1}^{K} \|f_j^{(k)}\|^2} = \sum_{j=1}^{p} \sqrt{K} \|\mathbf{f}_j\|$$

The penalty $\Omega_{\ell_{\mathbf{2}}\backslash\ell_{\mathbf{1}}}(\mathbf{f})$ in (2), is an extension of the $\ell_2\backslash\ell_1$-norm used for union support recovery in the parametric setting to the nonparametric setting. It also generalizes the SpAM of Ravikumar (2008), in the sense that if $K = 1$, we recover the problem solved in SpAM.

We define $R_j^{(k)} = Y^{(k)} - \sum_{\ell \neq j} f_l^{(k)}$, as the partial residual associated with the k^{th} task and the j^{th} covariate $X_j^{(k)}$. If we consider that all the other covariate functionals are held fixed, we can optimize over the set of functions $\mathbf{f}_j = \left(f_j^{(1)}, \cdots, f_j^{(K)}\right)$ by solving the optimization problem:

$$\hat{f}_j^{(1)}, \cdots, \hat{f}_j^{(K)} = \underset{f_j^{(1)}, \cdots, f_j^{(K)}}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{2} \mathbb{E}\left[\left(R_j^{(k)} - f_j^{(k)}\right)^2\right] + \lambda\sqrt{K}\|\mathbf{f}_j\| \tag{3}$$

Theorem 1: The stationary conditions of the optimization problem (2) with respect to the set of functions $\mathbf{f}_j$ associated with the covariates $\mathbf{X}_j$, is such that:
$\forall k \in \{1, \cdots, K\}$

$$f_j^{(k)} - P_j^{(k)} R_j^{(k)} + \lambda\sqrt{K} u_j^{(k)} = 0 \tag{4}$$

Where

$$\mathbf{u}_j = \begin{cases} \frac{f_j^{(k)}}{\|\mathbf{f}_j\|} & \text{if } \|\mathbf{f}_j\| \neq 0 \text{ for } k \in \{1, \cdots, K\} \\ \mathbf{e}_j \in \mathbb{R}^K & \text{with} \|\mathbf{e}_j\|_2 \leq 1 \text{ when } \|\mathbf{f}_j\| = 0 \end{cases}$$

The proof is given in the appendix.

Theorem 2: The covariates $X_j^{(k)}$ are active through their additive functions $f_j^{(k)}$, for $j = \{1, \cdots, p\}$, $f_j^{(k)} = 0 \ \forall k = \{1, \cdots, K\}$ if and only if

$$\sqrt{\sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)} R_j^{(k)}\right)^2\right]} < \lambda\sqrt{K} \tag{5}$$

We can now provide the algorithms needed to solve the optimization probem in (3)

Algorithm 1: Soft-Thresholding operator Algo1$_\lambda[\hat{R}_j^{(1)}, \cdots, \hat{R}_j^{(K)}, S_j^{(1)}, \cdots, S_j^{(K)}]$

1. *Input* : Smoothing matrices $\mathbf{S}_j^{(k)}$, estimated partial residuals $\hat{\mathbf{R}}_j^{(k)}$, for j fixed and $k \in \{1, \cdots, K\}$, and the regularization parameter λ.

2. Estimate $\mathbf{P}_j^{(k)}\mathbf{R}_j^{(k)} = \mathbb{E}\left[R_j^{(k)}|X_j^{(k)}\right]$ as $\hat{P}_j^{(k)} = S_j^{(k)}\hat{R}_j^{(k)}$ $\forall k \in \{1, \cdots, K\}$

3. Estimate $w_j = \sqrt{\sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)}R_j^{(k)}\right)^2\right]}$ as $\hat{w}_j = \sqrt{\sum_{k=1}^{K} \frac{1}{n}\|\hat{P}_j^{(k)}\|^2}$

4. calculate $\hat{\mathbf{f}}_j$ using $\hat{f}_j^{(k)} = \left[1 - \frac{\lambda\sqrt{K}}{\hat{w}_j}\right]_+ \hat{P}_j^{(k)}$

5. Center all the functions estimated in step 4.
   $\hat{f}_j^{(k)} \leftarrow \hat{f}_j^{(k)} -$ mean$(\hat{f}_j^{(k)})$

6. *Output*: estimated additive functions $\hat{\mathbf{f}}_j = \left(\hat{f}_j^{(1)}, \cdots, \hat{f}_j^{(K)}\right)$

The sparse backfitting applied to solve the problem presented in (3)

Algorithm 2: $L_2\backslash L_1$ Simultaneous Sparse Backfitting

---

1. *Input*: $\left(\mathbf{x}_i^{(k)}, y_i^{(k)}\right)$ with $i = 1, \cdots, n$ and $k = 1, \cdots, K$ and regularization parameter λ

2. Initialize $\hat{f}_j^{(k)} = 0$ and compute smoothers $S_j^{(k)}$ for $j = 1, \cdots, p$ and $k = 1, \cdots, K$

3. Iterate until convergence, for each $j = 1, \cdot, p$

   (a) For each $k = 1, \cdots, K$ compute the partial residuals:
   $\hat{R}_j^{(k)} = Y^{(k)} - \sum_{l \neq j} \hat{f}_j^{(k)}$

   (b) Use the threshold algorithm to find estimates of
   $\mathbf{f}_j = \left(f_j^{(1)}, \cdots, f_j^{(K)}\right)$
   $\hat{f}_j^{(1)}, \cdots, \hat{f}_j^{(K)} \leftarrow \texttt{Algo1}_\lambda[\hat{R}_j^{(1)}, \cdots, \hat{R}_j^{(K)}, S_j^{(1)}, \cdots, S_j^{(K)}]$

4. *Output* : Functions $\hat{\mathbf{f}}_j$ for $j = 1, \cdots, p$

3.4 $L_2\backslash L_1$ and L_1 joint functional sparsity

Assuming that all the tasks share the exact same predictors is not a realistic assumption in many applications. So we propose to induce sparsity within the groups of predictors that appear to be relevant. This translates to the optimization problem below:

$$\underset{\mathbf{f}^{(1)},\cdots,\mathbf{f}^{(K)}}{\text{Min}} \sum_{k=1}^{K} L(\mathbf{f}^{(k)}) + (1-\alpha)\lambda\Omega_{\ell_{\mathbf{2}}\backslash\ell_{\mathbf{1}}}(\mathbf{f}) + \alpha\lambda\Omega_{\ell_{\mathbf{1}}}(\mathbf{f}) \tag{6}$$

where $\Omega_{\ell_{\mathbf{1}}}(\mathbf{f}) = \sum_{j=1}^{p}\sum_{k=1}^{K}\|f_j^{(k)}\|$.

Solving the optimization problem (6) is equivalent to solving the problem for covariate index j while all the others covariates are held constant.

$$\hat{f}_j^{(1)},\cdots,\hat{f}_j^{(K)} = \underset{f_j^{(1)},\cdots,f_j^{(K)}}{\text{argmin}} \sum_{k=1}^{K}\frac{1}{2}\mathbb{E}\left[\left(R_j^{(k)} - f_j^{(k)}\right)^2\right] + (1-\alpha)\lambda\sqrt{K}\|\mathbf{f}_j\| + \alpha\lambda\sum_{k=1}^{K}\|f_j^{(k)}\| \tag{7}$$

Theorem 3: The covariates $X_j^{(k)}$ with $k \in \{1,\cdots,K\}$ are inactive as a group through their additive functions $f_j^{(k)}$, $\mathbf{f}_j = \mathbf{0}$ if and only if

$$\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\left[\left(1 - \frac{\alpha\lambda}{\|P_j^{(k)}R_j^{(k)}\|^2}\right)_+ P_j^{(k)}R_j^{(k)}\right]^2\right]} < (1-\alpha)\lambda\sqrt{K} \tag{8}$$

The proof of this theorem is given in Appendix. An interesting observation can be made about the condition in (8) and the condition introduced in (5). Our new thresholding condition can be seen as a weighted mean of the norms of the projected partial residuals. So if for a covariate $X_j^{(k)}$, $\|P_j^{(k)}R_j^{(k)}\|^2 < \alpha\lambda$, we obtain $\left(1 - \frac{\alpha\lambda}{\|P_j^{(k)}R_j^{(k)}\|^2}\right)_+ = 0$ and the term $P_j^{(k)}R_j^{(k)}$ doesn't contribute to the group condition. So the main difference between our new condition (8) and the condition in (5), lies in the fact that the new condition put an emphasis on the sparsity within the groups.

Theorem 4: For a given index j if the set of covariates $X_j^{(1)}, \cdots, X_j^{(K)}$ is active then a covariate $X_j^{(k)}$ with $k \in \{1, \cdots, K\}$ is inactive through its additive function $f_j^{(k)}$, $f_j^{(k)} = 0$ if and only if

$$\sqrt{\mathbb{E}\left[\left(P_j^{(k)} R_j^{(k)}\right)^2\right]} \leq \alpha\lambda \tag{9}$$

The proof of theorem 4 is also given in Appendix. The condition simply states that the additive function $f_j^{(k)} = 0$ if the functional norm of the partial residual associated with the covariate $X_j^{(k)}$ is less than the threshold $\alpha\lambda$

Algorithm 3: Soft-Thresholding operator Algo3$_\lambda[\hat{R}_j^{(1)}, \cdots, \hat{R}_j^{(K)}, S_j^{(1)}, \cdots, S_j^{(K)}]$

1. *Input* : Smoothing matrices $\mathbf{S}_j^{(k)}$, estimated partial residuals $\hat{\mathbf{R}}_j^{(k)}$, for j fixed and $k \in \{1, \cdots, K\}$, and the regularization parameter λ.
2. Estimate $\mathbf{P}_j^{(k)} \mathbf{R}_j^{(k)} = \mathbb{E}\left[R_j^{(k)} | X_j^{(k)}\right]$ as $\hat{P}_j^{(k)} = S_j^{(k)} \hat{R}_j^{(k)} \ \forall k \in \{1, \cdots, K\}$
3. Estimate $\forall k \in \{1, \cdots, K\}$ $g_j^{(k)} = \left(1 - \frac{\alpha\lambda}{\|P_j^{(k)} R_j^{(k)}\|^2}\right)_+$ as $\hat{g}_j^{(k)} = \left(1 - \frac{\alpha\lambda}{\|\hat{P}_j^{(k)}\|^2}\right)_+$
4. Estimate $h_j = \sqrt{\sum_{k=1}^K \mathbb{E}\left[\left[\left(1 - \frac{\alpha\lambda}{\|P_j^{(k)} R_j^{(k)}\|^2}\right)_+ P_j^{(k)} R_j^{(k)}\right]^2\right]}$ as $\hat{h}_j = \sqrt{\sum_{k=1}^K \frac{1}{n} \|\hat{g}_j^{(k)} \hat{P}_j^{(k)}\|^2}$
5. If $\hat{h}_j < (1 - \alpha)\lambda\sqrt{K}$ set $\hat{\mathbf{f}}_j = \mathbf{0}$
6. Else
   (a) For each $k \in \{1, \cdots, K\}$ if $\|\hat{P}_j^{(k)}\| < \alpha\lambda$ then $\hat{f}_j^{(k)} = 0$
   (b) Else update $\hat{f}_j^{(k)}(i+1) = \frac{\hat{P}_j^{(k)}}{1 + \frac{(1-\alpha)\lambda\sqrt{K}}{\|\hat{\mathbf{f}}_j(i)\|} + \frac{\alpha\lambda}{\|\hat{f}_j^{(k)}(i)\|}}$,
   where i is the i^{th} iteration
7. Center all the estimated functions $\hat{f}_j^{(k)} \leftarrow \hat{f}_j^{(k)} -$ mean$(\hat{f}_j^{(k)})$
8. *Output* : estimated additive functions $\hat{\mathbf{f}}_j = \left(\hat{f}_j^{(1)}, \cdots, \hat{f}_j^{(K)}\right)$

3.5 Simulations

We now investigate the empirical properties of the methodologies proposed. We simulate 100 datasets from a multivariate regression model with $K = 4$ responses, $p = 200$ or 400 covariates. Each covariate X_j is generated as $X_j = \frac{W_j+tU}{1+t}$, with $j \in \{1, \cdots, p\}$ and where $W_1, \cdots, W_p$ and U are i.i.d from Uniform$(-2.5, 2.5)$. The responses are then generated as $Y^{(k)} = \sum_{j=1}^{p} f_j^{(k)}(X_j) + \varepsilon^{(k)}$ where $\varepsilon \sim N(0, \sigma^2)$. The additive functions we used are similar to the functions used in Yin et al. (2012) and Meier et al. (2009). We create training, validation and test datasets of sizes $n_{validation}/n_{train}/n_{test} = (150/150/50)$. We use 10-fold cross validation on the validation datasets to find the optimal regularization parameters. We also compute the Mean Squared Errors (MSE) for each responses by using the test datasets. The correlation between the covariates $X_1, \cdots, X_p$ increases as we increase the value of t. We peform simulations for three possible values of $t = \{0, 1, 2\}$. The component functions used in this simulation are:

We define $z_3 = \frac{x_3+2.5}{5}$

Additive functions of the first response $Y^{(1)}$

$f_1^{(1)}(x_1) = -5\sin(2x_1)$, $f_2^{(1)}(x_2) = x_2^2 + 1.5(x_2 - 1)^2$

$f_3^{(1)}(x_3) = 0.5\sin(2\pi z_3) + \cos(2\pi z_3) + 1.5\sin^2(2\pi z_3) + 2\cos^3(2\pi z_3) + 2.5\sin^3(2\pi z_3)$

Additive functions of the second response $Y^{(2)}$

$f_1^{(2)}(x_1) = x_1^2$, $f_2^{(2)}(x_2) = x_2$, $f_3^{(2)}(x_3) = \frac{4\sin(2\pi z_3)}{2-\sin(2\pi z_3)}$

Additive functions of the third response $Y^{(3)}$

$f_1^{(3)}(x_1) = \frac{2\sin(2\pi x_1)}{2-\sin(2\pi x_1)}$, $f_2^{(3)}(x_2) = 3\sin(\exp(-0.5x_2))$, $f_3^{(3)}(x_3) = -x_3$

Additive functions of the fourth response $Y^{(4)}$

$f_1^{(4)}(x_1) = \exp(-x_1)$, $f_2^{(4)}(x_2) = -0.5\phi(x_2, 0.5, 0.8)$, $f_3^{(4)}(x_3) = -x_3^2$

All the other additive components are $f_j^{(k)}(X_j) = 0$ for any $j \in 4, \cdots, p$ and $k \in \{1, \cdots, 4\}$.

Simulation setup 2: we also consider a second simulation setting in which we induce within group sparsity by setting $f_3^{(1)}(x_3) = f_1^{(2)}(x_1) = f_3^{(3)}(x_3) = f_2^{(4)}(x_2) = \mathbf{0}$.

3.5.1 Simulation Results

In this section, we report the performance of the proposed algorithm on the second simulation setting described.

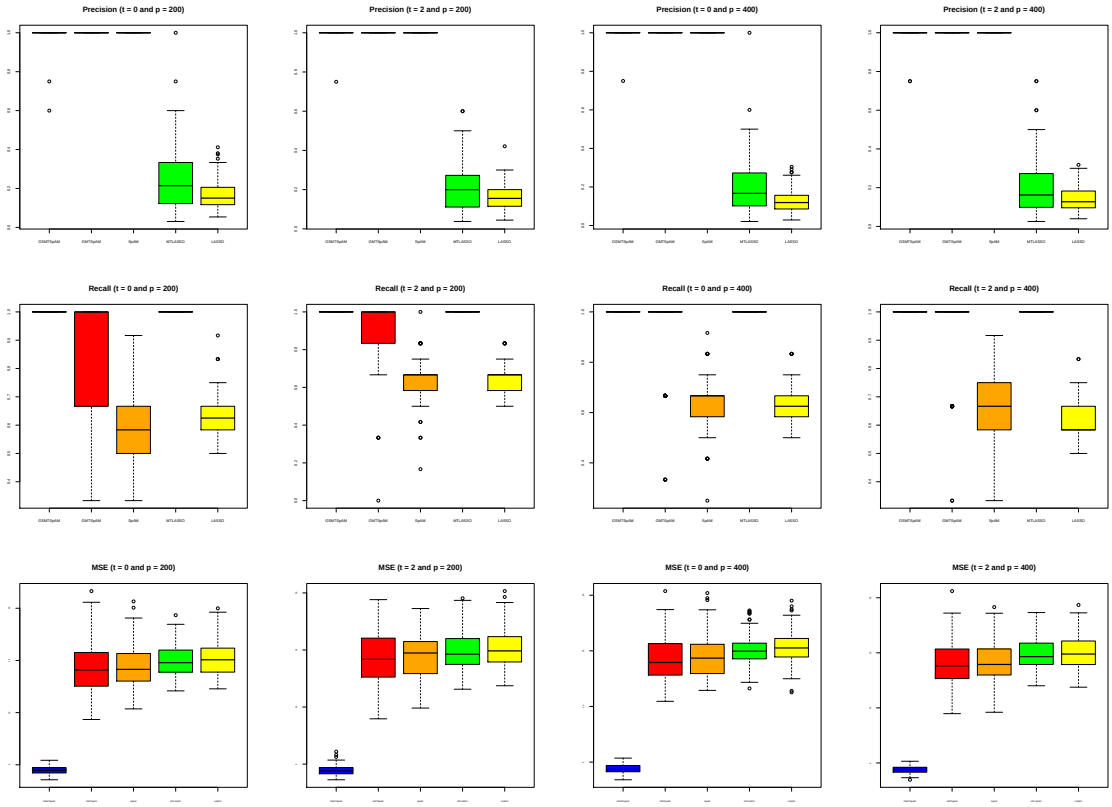


Figure 7: Precision, recall and MSE for GSMTSpAM (blue), GMTSpAM (red), SpAM (orange), MTLASSO (green), LASSO (yellow)

Based on the box plots observed in Figure 8, we find that the fraction of retrieved additive functions that are relevant (precision) is high for the GSMTSpAM and for SpAM, and the method GMTSpAM has a low precision, because it cannot induce within group sparsity. Note that the reduction in precision observe for the GMTSpAM will also be present in the method proposed by Liu et al. (2009), since their method do not account for within group sparsity. We also observe that the precision of

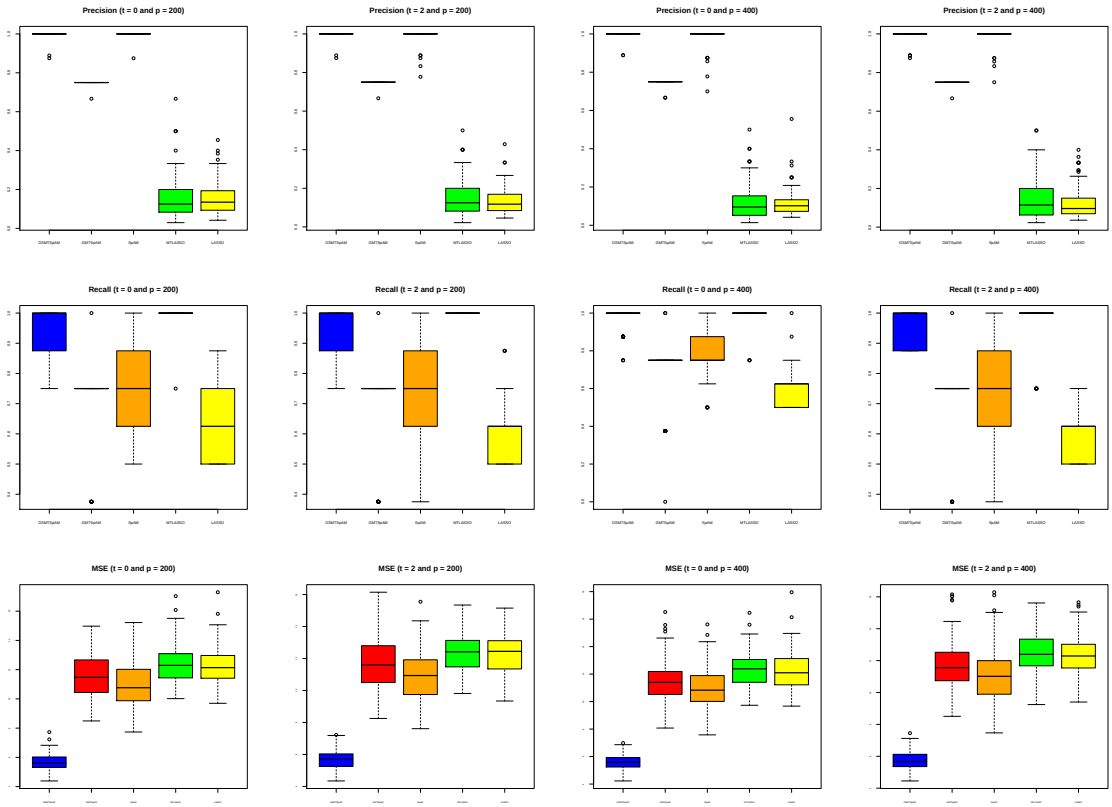


Figure 8: Precision, recall and MSE for GSMTSpAM (blue), GMTSpAM (red), SpAM (orange), MTLASSO (green), LASSO (yellow) for simulation setup 2

Table 3: Comparison of different methods on simulated data. Shown in 4th, 5th are the mean and the standard deviation (in parenthesis) of precisions and recalls. The size of the model and the MSE metrics are shown in the final 2 columns

p	t	method	$\#f_1^{(1)}$	$\#f_1^{(2)}$	$\#f_1^{(3)}$	$\#f_1^{(4)}$	$\#f_2^{(1)}$	$\#f_2^{(2)}$	$\#f_2^{(3)}$	$\#f_2^{(4)}$	$\#f_3^{(1)}$	$\#f_3^{(2)}$	$\#f_3^{(3)}$	$\#f_3^{(4)}$	size (12)
200	0	GSMTSpAM	100	100	100	100	100	100	100	100	100	100	100	100	12.12 (0.89)
		GMTSpAM	100	100	100	100	90	90	90	90	75	75	75	75	10.6 (2.37)
		SpAM	95	95	28	93	94	65	5	24	55	92	78	7	7.26 (1.49)
		MLasso	100	100	100	100	100	100	100	100	100	100	100	100	73.08 (55.16)
		Lasso	25	14	100	100	43	100	16	100	50	100	100	11	53.82 (24.57)
200	2	GSMTSpAM	100	100	100	100	100	100	100	100	100	100	100	100	12.04 (0.4)
		GMTSpAM	98	98	98	98	92	92	92	92	80	80	80	80	10.8 (2.31)
		SpAM	94	90	38	92	93	67	12	18	59	92	91	12	7.58 (1.45)
		MLasso	100	100	100	100	100	100	100	100	100	100	100	100	83.2 (61.07)
		Lasso	33	8	100	100	39	100	17	100	59	100	100	8	54.75 (25.68)
400	0	GSMTSpAM	100	100	100	100	100	100	100	100	100	100	100	100	12.04 (0.4)
		GMTSpAM	100	100	100	100	95	95	95	95	78	78	78	78	10.92 (2.12)
		SpAM	98	90	38	95	96	69	8	16	57	89	88	7	7.51 (1.33)
		MLasso	100	100	100	100	100	100	100	100	100	100	100	100	92.76 (76.20)
		Lasso	24	11	100	100	43	100	12	100	55	100	100	6	71.83 (36.36)
400	0	GSMTSpAM	100	100	100	100	100	100	100	100	100	100	100	100	12.08 (0.56)
		GMTSpAM	100	100	100	100	97	97	97	97	77	77	77	77	10.96 (1.94)
		SpAM	96	87	38	93	96	63	14	30	61	84	90	11	7.63 (1.53)
		MLasso	100	100	100	100	100	100	100	100	100	100	100	100	98.32 (78.31)
		Lasso	21	7	100	100	47	100	18	100	45	100	100	5	63.82 (31.51)

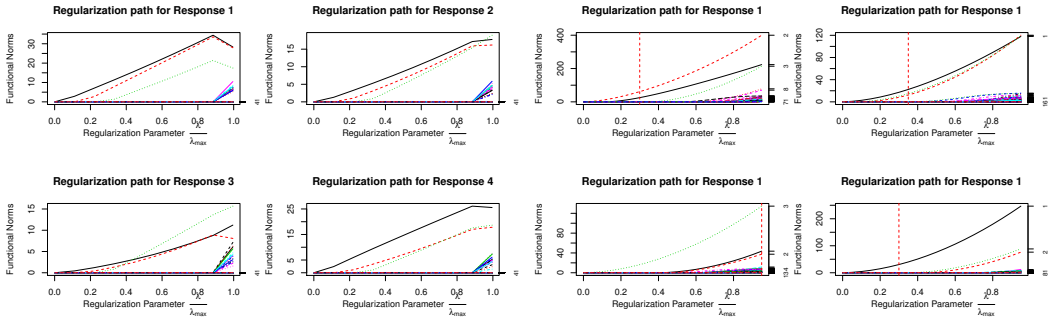
Table 4: Comparison of different methods on simulated data. Shown in 4th, 5th are the mean and the standard deviation (in parenthesis) of precisions and recalls. The size of the model and the MSE metrics are shown in the final 2 columns

p	t	method	$\#f_1^{(1)}$	$\#f_1^{(2)}$	$\#f_1^{(3)}$	$\#f_1^{(4)}$	$\#f_2^{(1)}$	$\#f_2^{(2)}$	$\#f_2^{(3)}$	$\#f_2^{(4)}$	$\#f_3^{(1)}$	$\#f_3^{(2)}$	$\#f_3^{(3)}$	$\#f_3^{(4)}$	size (8)
200	0	GSMTSpAM	100	0	97	100	100	100	71	0	2	100	0	100	7.7 (0.50)
		GMTSpAM	99	99	99	99	89	89	89	89	1	1	1	1	7.56 (1.38)
		SpAM	96	0	87	94	96	73	53	0	0	95	0	8	6.03 (0.99)
		MLasso	100	100	100	100	100	100	100	100	99	99	99	99	76.08 (47.55)
		Lasso	36	6	100	100	45	100	12	7	2	100	5	6	41.64 (21.89)
200	2	GSMTSpAM	100	0	98	100	100	100	70	0	1	100	0	100	7.7 (0.50)
		GMTSpAM	100	100	100	100	82	82	82	82	1	1	1	1	7.32(1.61)
		SpAM	98	0	87	97	96	76	54	0	0	98	0	9	6.21 (1.21)
		MLasso	100	100	100	100	100	100	100	100	100	100	100	100	82 (64.29)
		Lasso	31	9	100	100	43	100	100	7	5	100	1	9	45.31 (22.43)
400	0	GSMTSpAM	100	0	97	100	100	100	78	0	1	100	0	100	7.77 (0.53)
		GMTSpAM	99	99	99	99	85	85	85	85	2	2	2	2	7.44 (1.71)
		SpAM	100	0	86	97	91	76	59	0	0	96	0	3	6.17(1.02)
		MLasso	100	100	100	100	100	100	100	100	97	97	97	97	112.28 (92.59)
		Lasso	28	5	100	100	45	1	8	5	0	100	0	7	51.94 (24.35)
400	0	GSMTSpAM	100	0	100	100	100	100	70	1	3	100	0	100	7.74 (0.50)
		GMTSpAM	100	100	100	100	90	90	90	90	1	1	1	1	7.64 (1.28)
		SpAM	93	0	92	92	94	73	53	0	0	92	0	4	5.99 (1.03)
		MLasso	100	100	100	100	100	100	100	100	94	94	94	94	94.16 (73.11)
		Lasso	24	6	100	100	52	100	6	3	2	100	1	3	52.95 (28.80)

MTLASSO and LASSO are low, suggesting the presence of many falsely selected additive functions. The fraction of relevant instances that are retrieved (recall) is also high for GSMTSpAM,while it is low for SpAM. Since the precision of SpAM is high and its recall is low, we can conclude that SpAM is too conservative and it fails to select some of the true additive functions, while GSTSpAM performs well across metrics. MTLASSO has a high recall but this is simply due to the fact that it selects variables that should not be included in the model. LASSO also tend to have a lower recall than the recall of all the methods illustrated. Last but not least, we also display a boxplot of the Mean Squared Error of the true additive functions when compared to the fitted additive functions. The Mean Squared Error is defined as $MSE = \frac{\sum_{k=1}^{K}\sum_{j=1}^{P}\sum_{i=1}^{N}\left(f_j^{(k)}(x_{ij}^k) - \hat{f}_j^{(k)}(x_{ij}^k)\right)^2}{NPK}$, where the points x_{ij} are randomly generated. We find that the GSMTSpAM yield predictions with smaller errors than all the other methods studied, and naturally the GMTSpAM and SpAM perform better than MTLASSO and LASSO since they account for the nonlinear relationships between the responses and the predictors. In Figure 7 of the web appendix, we display

boxplots of the precision, the recall and the mean squared error (MSE), in the absence of within group sparsity (simulation settings 1). In such a settings the performance of GSMTSpAM remain better than the results of all the studied methods. The GMTSpAM has improved precision and recall, because predictors X_1, X_2, X_3 and X_4 affect all the responses.

In table 4, we report the number of times each function is selected by each of the methods studied and we also show the average size of the models. We see that in the simulation settings 2, the additive functions set to zero are rarely selected by GSMTSpAM and that on average the number of selected additive functions is close to the size of the true model (8 additive functions in simulated settings 2). We clearly see that SpAM is too conservative and select on average 6 additive functions, that MTLASSO selects too many additive functions (model of size greater than 75), and that LASSO selects too many additive functions and yet does not include all the relevant additive functions.



(a) $L_2\backslash L_1$ GSMTSpAM **(b)** SpAM

Figure 9: Regularization Path for the $L_2\backslash L_1$ and L_1 SpAM and SpAM

To further understand the benefits of GSMTSpAM over SpAM, we display in Figure 17 of the web appendix, the full regularization path of the additive functions simulated in settings 1. We are clearly able to see why GSMTSpAM outperforms SpAM in the context of multi-responses regression. The regularization paths of the SpAM associated with the different responses show that relevant additive functions

enter in the model at the same moment non-meaningful predictors enter in the model. This explains why SpAM has a high precision but a low recall. The regularization path of the GSMTSpAM shows that the relevant predictors enter in the model much earlier than the other predictors.

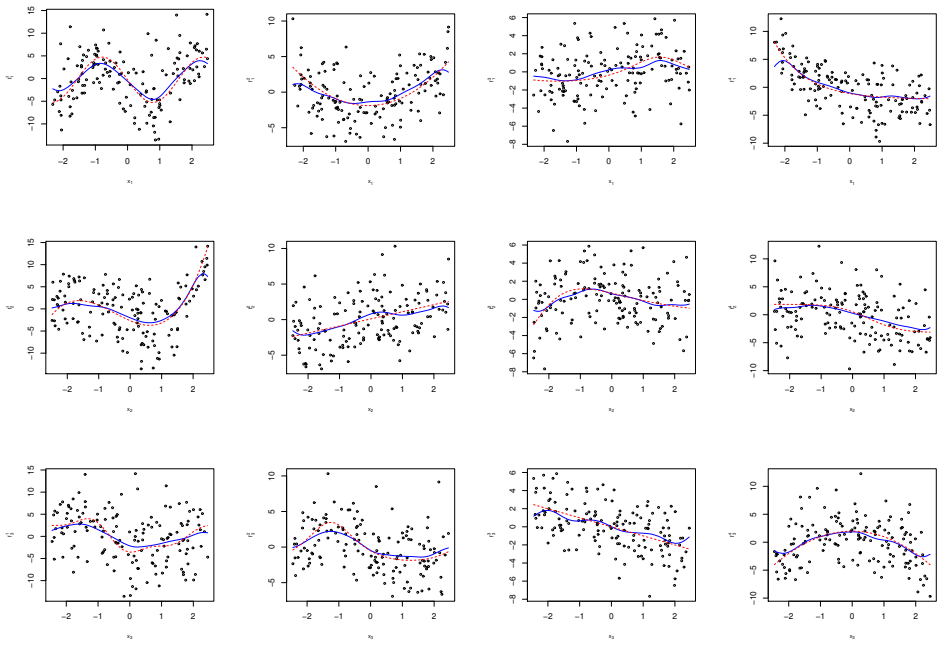


Figure 10: Estimated Additive Functions (solid blue) and true additive functions (dashed red), for one simulation with 150 observations, p = 200 and t = 0

In Figure 10 , we display the simulated functions and the estimated functions with a set of points that show the partial residuals obtained by extracting all the additive functions except the function of interest. The method approximates well the true additive functions.

3.6 Application

3.6.1 Gene Microarray Data of Cancer Patients

In this section, we apply our method on the children cancer data set introduced in Khan et al. (2001). The method is used to classify the small round blue cell tumors (SRBCTs) into 4 categories of cancers, Neuroblastoma (NB), Rhabdomyosarcoma

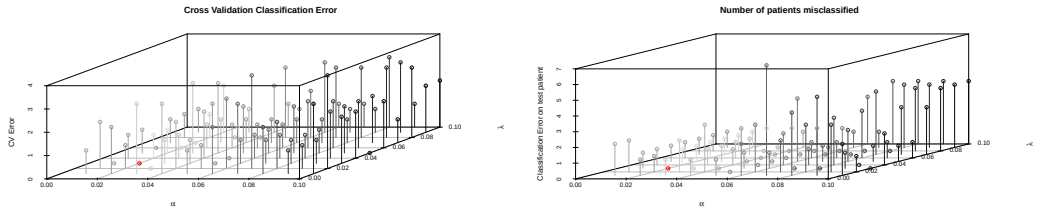
(RMS), Non-Hodgkin Lymphoma (NHL), and the Ewing Family of Tumors (EWS). The data set contains 83 patients, 63 are used for training and 20 are left for testing. For each patient, the expression profile of 2308 genes are measured. This is a benchmark data set that has been used by several groups to compare their classification method against pre existing methodologies. Most of these methods are designed to select a set of important genes that can be informative in classifying tumor types. The existing methodologies can all achieve 100% classification rate on this data set, Khan et al. (2001) applied a neural network approach to find 96 important genes. Tibshirani et al. (2002) developed a method called nearest shruken centroids and were able to achieve a perfect classification rate with 43 genes. Zhang et al. (2008) used a sup-norm support vector machine method to identify 53 genes. Liu et al. (2009) achieved 100% prediction accuracy with only 20 predictors by using a new method called Sparse Multivariate Additive Logistic Regression (SMALR). But when they interpreted the biomarkers selected, they highlighted the fact that some genes identified by their method were not among the genes selected by existing methods. And this non-overlap was not explained in the paper. The Group sparse Multivariate Additive Logistic Regression (GSMALR), we introduced in the paper can achieve 100% prediction accuracy with only 12 genes. Of all the genes selected only 1 gene (810057) doesn't appear in the genes selected by Zhang et al. (2008), and this is due to the fact that they only select among the 100 genes with the highest relevance measure, and only 2 genes do not appear in the genes selected by Tibshirani et al. (2002), (gene 236282 and gene 383188). The gene 810057 is also not included in the list of all genes selected by Liu et al. (2008). In Figure 11, we show a heat map of the selected variables for the 63 patients used to train our model (Figure 11(a)) and the model proposed by Liu et al. (2009) (Figure 11(b)). The y-axis displays the genes and the x-axis is associated with each patient, patients are grouped in 4 different categories, corresponding to different tumors. Both heat maps show four

block structures associated with the four tumor categories. This suggest that the genes selected provide enough information to properly classify the tumor of cancer patients. We see that for our method less genes are needed to achieve the same rate of classification and the model generated is still highly informative.

(a) Genes selected by GSMALR **(b)** Genes selected by SMALR of Liu et al.

Figure 11: Genes Selected by the proposed method and the method of Liu (2009)

In Figure 12 (a), we display the average number of misclassified patients in the training data when the cross validation scheme is performed, and Figure 12 (b) shows the number of patients that are misclassified in the test data set. The points in red represent these values for the optimal regularization parameters $\lambda = 0.03$ and $\alpha = 0.02$ that are selected by cross validation.



(a) Cross validation error **(b)** Test error

Figure 12: Number of misclassified patients for cross validated training sample and for test sample for values of α and λ

3.6.2 Application to Microarray Data from Arabidopsis Thaliana

In this application, we use a microarray dataset first used by (wille et al. 2004), to better understand the biosynthesis of isoprenoid. Isoprenoid is a biological component that plays an important role in some of the vital functions (e.g: photosynthesis

and respiration) of the plant Aribdopsis thaliana. We are interested in uncovering the relationships between genes that belong to distinct genetic pathways. We use the expression profiles of the associated genes to build the genetic networks that regulates the control mechanism for the synthesis of isoprenoids. We will put an emphasis on the determination of the crosstalks between two distinct isoprenoids pathways. The isoprenoid pathways used in the analysis are the Mevalonate pathway and the plastidial pathway. The expression level of the 21 genes in the Mevalonate pathway (MVA) will be used as predictors and the expression level of the 18 genes in the Plastidial pathway (MEP) will be the responses in our estimation model. Each predictor and each response has 118 samples. All the variables are centered and standardized to unit variance. Since we are interested in finding the regulatory network between the pathways, we perform a stability selection analysis. We randomly select 100 observations out of the 118 available samples, and run the GSMTSpAM on the selected subset. We repeat this procedure 100 times, in Figure (13), we display the relationships that are present in at least 90 out of the 100 replications. Some of our results corroborate the findings of Wille et al. (2004) and Lozano et al. (2012), we find for instance that there are no connections that are emanating from genes GGPPS1mt, 3, 4, 8. We also find as in Wille et al. (2004), that there are cross-talk relationships between genes AACT1 and HMGR1 in the MVA pathway and the genes DXR and MECPS in the MEP pathway. Contrary to the results of Wille et al. (2004), we could not identify cross-talks relationship arising from the genes AACT2, HMGS, HMGR2, FPPS2 and MPDC1. Our methodology also uncovers some cross-talk relationships not previously identified by the previous models, such as the multi-level lasso, this could be explained by the fact that the multi-level lasso cannot capture nonlinear relationships between predictors and responses. It would be interesting to investigate if some of the cross-talk relationships identified between genes in the MVA and MEP pathways have a valid biological interpretation. After establishing the performance of

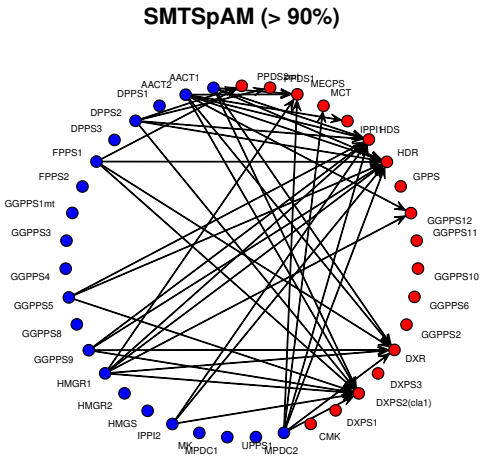


Figure 13: Associations identified between genes from Mevalonate isoprenoid pathway and plastidial pathway using GSMTSpAM

the GSMTSpAM, we apply to the same dataset SpAM, MTSpAM, MTLASSO and LASSO. In Figure (14),We find that the bootstrapped regulatory network will all the possible cross-talk relationships and the bootstrapped graph with the relationships appearing more than 90 times out of 100 are pretty similar when the GSMTSpAM is applied, this suggests that our model is relatively stable. A similar statement cannot be made about SpAM, we find that SpAM introduces a lot of spurious cross-talk relationships and that a limited subset of them are selected more than 90 % of the time.

We are also interested in assessing the predictive ability of our model, we achieve this by providing forecasts for the 18 points that are left out of the training sample in each replication used for the stability selection. Figure (15), shows that the GSMTSpAM performs better than SpAM, LASSO. The MTLASSO yields the best forecast but it also yields an uninterpretable model has illustrated in Figure (16). Note that the LASSO doesn't select any cross-talk relationships more than 90 % of the time.

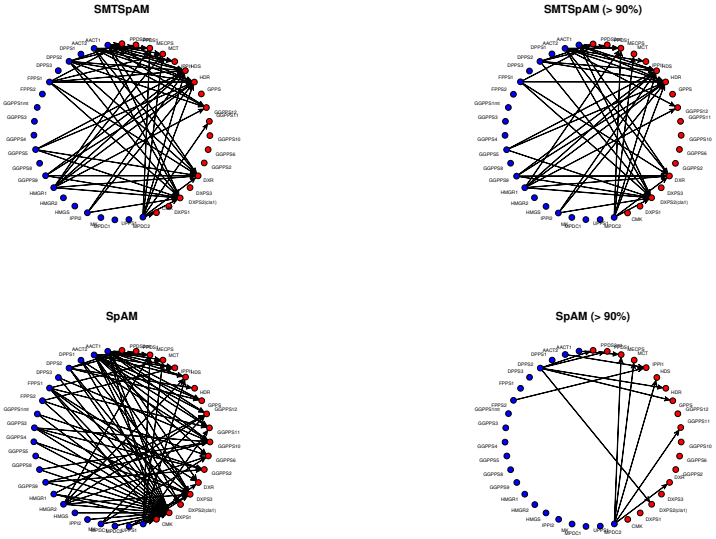


Figure 14: Associations identified between genes from Mevalonate isoprenoid pathway and plastidial pathway by GSMTSpAM and SpAM

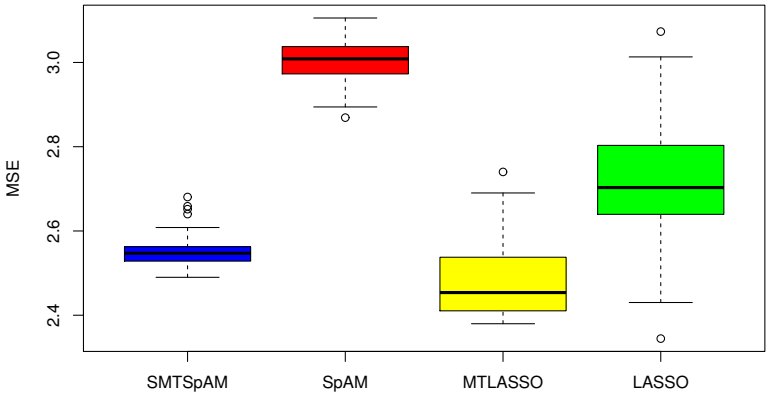


Figure 15: MSE for prediction of MEP

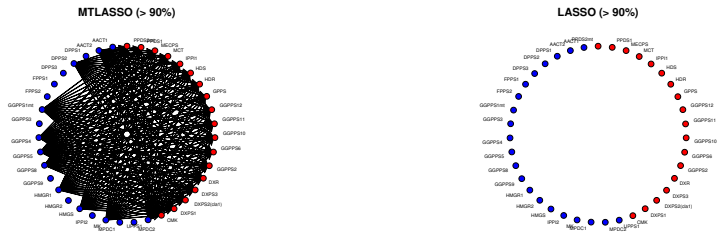


Figure 16: Associations identified between genes from Mevalonate isoprenoid pathway and plastidial pathway by MTLASSO and LASSO

3.6.3 County Level Cost Analysis in North Carolina

In this application, we are interested in identifying the factors that drive the cost of healthcare in North Carolina counties from 2005 to 2009. The objective is also to evaluate the relationship between the systems outcome cost and the relevant determinant of healths that will be identified through our additive model selection scheme. The relationships between the systems outcome cost and these determinants of health, may or may not vary throughout the years. To account for these potential variations, we use as responses in our model the county level cost per medicaid eligible member per month. This metric is simply computed by first summing all the medical claims associated with medicaid eligible patients who reside within the county of interest. After obtaining, the total claims issued for patients living within a county, we divide this quantity by the total number of months of eligibility in the county to obtain the county level cost per member per month. In our model, we will have 5 responses and each response corresponds to the cost per member per month for all the counties in a given year. Given that there are 100 counties in North Carolina, each response will have 100 observations. We will now describe the major determinant of healths that will be used as predictors to explain county level cost variation in the healthcare delivering system.

3.6.3.1 Data Description

The population considered for the analysis consists of patients between the age of 0 to 18 years who lived in the state of North Carolina during the period of 2005 to 2009. To explain the variations observed in the healthcare cost in North Carolina, we use 40 predictors that are grouped in the following categories: demographics, utilization, socio-economic environment, access to care (financial and geographical) and health factors. The demographics variables consists of the percentage of claims who are associated with white patients, non-white patients during the year of interest. The demographics group also contains age related variables such as the percentage of claims who are attributed to patients between the age of 0 and 5 years, between the age of 6 and 14 years, and between the 15 and 18 years. Demographic measures have been extensively used in the literature as a control factor to study the disparities in financial and geographical access to care. The utilization measures included in our analysis are the number of claims per member per months that are associated with inpatient services, outpatient services, other services. Additionally, we also add the number of claims per member per month that are issued after patients are consulted by a physician, at a clinic or for dental services.

The second set of utilization measures used in our analysis are linked to the place where the medical services where provided. These measures are the number of claims per member per month associated with hospitalizations, visits to the medical practitioner office, the emergency rooms, and outpatient hospital. Utilization measures are included in the study since they are directly linked to the cost per member per month. The following papers by Grupp-Phelan et al.(2001), Glynn et al. (2011) and Harrison et al. (2012) have analyzed the relationships between utilization and cost of healthcare systems, the general consensus is that higher utilization of the systems lead to higher cost.

Some socio-economic factors, such as education, economic indicators, crime and

family planning related metrics, have been reported to have an impact on health outcomes. To account for education's impact on health outcomes, we include the county level illiteracy rates,which represents the percentage of the population age 16 and older that lacks basic literacy skills, the percentage of high school graduates or higher and the percentage of Bachelor's degree or higher. The economic factors are the county level per capita income, the unemployment rate and the percentage of household units with a mortgage with housing costs greater than 30% of income in a given county. High housing costs and high unemployment rates can be associated with poor health outcomes. Additionally, since employer-sponsored health insurance is the most prevalent coverage, unemployment can reduce access to health care. We also include crime related variables such as homicide rates, which represents the homicide rate pre 100,000 in a given county and violent crime rates. Family related social variables are the percent of family households with children that are headed by a single parent (male or female householder with no spouse present) and the teen birth rate measured as the number of births per 1000 female population aged 15 tgo 19. There is evidence that teen pregnancy increases the risk of adverse health outcomes for mothers, children and communities. Health factors are directly associated with the cost of health care systems. To measure county level health conditions of the population, we use the percentage of the population age (20 and older) that has a BMI greater or equal to 30 $kg\backslash m^2$ (obesity rates), the diabetes rates, low birth weight defined as the percent of live births for which the infant weighted less than 2500 grams. According to County Health Rankings and Roadmaps, low birth weight are a good predictor of morbidity over the course of a life, and they may help lead to higher utilization of the systems by affected patients. We also use a self reported indicator of health, which the percent of adults reporting in a survey poor or fair health. Nutrition related variables are also added, namely the limited access to healthy foods measured as the percent of population who are low income and do not live close to a grocery

store, and the percent of fast food restaurants within each counties.

The main goal of the study, is also to identify predictors that can be used to intervene and recommend policies that will help reduce the cost per member per month. The determinants of health that can be used for interventions are mainly related to access. Access in our analysis can be interpreted as financial access or geographical access. The proxies used to measure financial access in our analysis are the county level poverty rates and the percent of adults who reported that they could not see the doctor because of cost. The percent of children (under 19) without insurance is also included since lack of health insurance is a barrier to accessing health care. Financial access has been reported to also affect health outcomes ,Doran et al. (2008), but we are interested here in understanding if it affects the cost of care at the county level. The geographical access can be divided in 2 sub groups, availability and accessibility to care. In Gentili et al. (2014), availability is measured as the average travel time to care in each county, while accessibility is the average congestion level in each county. We also add the standard deviation of travel time and congestion, to capture the potential effect of disparities in access to care in the county on cost of care. The first and second moments are computed by looking at travel time and congestion level computed at the census tract level as described by Gentili et al. (2014).

3.6.3.2 Results

We apply our model to identify the most important determinants of health and to estimate the potential nonlinear relationships between the cost of care and the 40 predictors enumerated in the data description. To have a certain level of confidence in the selected variables, we perform a stability selection analysis, that consists in randomly selecting 90 out of the 100 counties as training data, we run the model with regularization parameters in the range $\alpha \in \{0.001, \cdots, 0.1\}$ and $\lambda \in \{10^{(-25)}, \cdots, 1\}$.

For each of the regularization parameters we have 50 values that are equally spaced. For each combination of regularization parameters we fit the model and assess if a predictor is selected, we repeat this procedure 100 times, and report the number of times a predictor is selected. We find that utilization predictors are the most important in driving the cost of care at the county level in North Carolina. In Figure 33 of the appendix, we show that the number of claims per member per month associated with inpatient hospitalization, with miscellaneous services and outpatient and inpatient services are the most influential predictors since they are selected 100 out of 100 times for almost all combinations of the regularization parameters. Other utilization measures such as the preventable hospital stays, the number of claims per member per month associated with dental services, services provided by the physician, visits to the emergency rooms, to the health practitioner office are also meaningful predictors of cost but they are not as influential as the aforementioned utilization measures. For all the relevant utilization measures an increase in number of claims per member per months leads to higher cost of care per member per months.

We do not find strong evidence suggesting that demographics are highly impactful on the cost of care in North Carolina at the county level between 2005 and 2009. Among the socio-economic factors, we find that the percent of housing with high costs and the per capita income are the most important variables, they are however less often included in the model than the most relevant utilization measures.

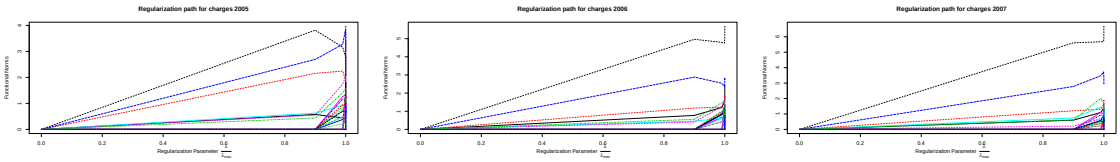
We find evidence that financial and geographical access influenced the cost of care in North Carolina during the years 2005 to 2009. For financial access, we observe the percent of uninsured children and the percent who reported not being able to see a doctor because of cost influenced the cost of care. Figures 35 (u) - (y) in appendix show that the cost of care increases marginally when these measures increase, this increase is mostly noticeable in the lower quantile of the graphs, suggesting that the effect on cost dissipates as the percent of uninsured children within a county

increases (or as the percent of people who could not see the doctor because of cost). For geographical access, we find that only traffic congestion seems to be related to cost of care. In figures 35 (a) - (e) of the appendix, we see that an increase in county level traffic congestion leads to a minor increase in cost of care. So an area of intervention to reduce the cost of care, could be investment that can help reduce congestion in counties of North Carolina.

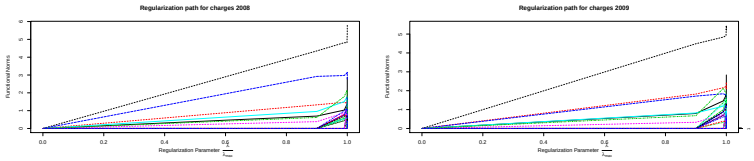
We also perform a cross validation predictive scheme, to have a sense of the optimal values of the parameters α and λ. In figure 36, we see that the values of the parameter α with the best predictions are below 0.1. This suggests that the ℓ_1 penalty is more strongly enforced than the group penalty. This implies that when a predictor affects healthcare cost for one year, it is highly likely that it will affect the cost of care during the other years. So we can conclude that in North Carolina from 2005 to 2009, the relevant county level determinants of health did not change. Figure 17 shows the full regularization paths for the norm of the additive functions associated with the relevant predictors. For all the years, the utilization variables such as the number of claims associated with other services, inpatient services, outpatient services, inpatient hospitalizations are the first variables to enter in the model. Then the average county level congestion, financial access variables and some health indicators then become influential.

3.7 Concluding Remarks

In this chapter, we have presented a new methodology for variable selection when dealing with multivariate additive nonparametric regression or classification. The methodology introduces a new joint penalty, by combining a functional $\ell_2\backslash\ell_1$ norm with a functional ℓ_1 norm applied to additive terms in the multivariate additive regression model. By deriving the subdifferentials of these penalties, we propose a series of backfitting algorithms that can update each additive functions with a closed



(a) Regularization path 2005 **(b)** Regularization path 2006 **(c)** Regularization path 2007



(d) Regularization path 2008 **(e)** Regularization path 2009

Figure 17: Regularization Paths associated with the cost of care for years 2005 to 2009 for $\alpha = 0.005$

form solution. The performance of the algorithms were studied on a series of synthetic data, on a benchmark dataset (gene microarray data of cancer patients) and for the analysis of county level medical costs in the state of North Carolina.

CHAPTER IV

NONPARAMETRIC REGRESSION FOR TOPOGRAPHICAL MIXTURE MODELS WITH SYMMETRIC ERRORS

4.1 Introduction

The model we propose to investigate in this paper is a semiparametric topographical mixture model able to capture the characteristics of dichotomous shifted response-type experiments such as the tumor data in Bowen et al. (2012, Fig. 4). Let suppose that we visit at random the space $\mathbb{R}^d$ ($d \geq 1$) by sampling a sequence of i.i.d. random variables $\mathbf{X}_i$, $i = 1, ..., n$, having common probability distribution function (p.d.f.) $\ell : \mathbb{R}^d \to \mathbb{R}_+$. For each $\mathbf{X}_i$ we observe an output response Y_i whose distribution is a mixture model with probability parameters depending on the design $\mathbf{X}_i$. For simplicity, let us consider first a mixture of two nonlinear regression model:

$$Y_i = W(\mathbf{X}_i)(a(\mathbf{X}_i) + \tilde{\varepsilon}_{1,i}) + (1 - W(\mathbf{X}_i))(b(\mathbf{X}_i) + \tilde{\varepsilon}_{2,i}), \tag{10}$$

where locations are a, $b : \mathbb{R}^d \to \mathbb{R}$, the errors $\{\tilde{\varepsilon}_{1,i}, \tilde{\varepsilon}_{2,i}\}_{i=1,...,n}$ are supposed to be i.i.d with zero-symmetric common p.d.f. f. The mixture in model (10) occurs according to the random variable $W(\mathbf{x})$ at point $\mathbf{x}$, with probability $\pi : \mathbb{R}^d \to (0, 1)$,

$$W(\mathbf{x}) \quad = \quad \begin{cases} 1 & \text{with probability} \quad \pi(\mathbf{x}), \\ 0 & \text{with probability} \quad 1 - \pi(\mathbf{x}). \end{cases}$$

Moreover we assume that, conditionally on the $\mathbf{X}_i$'s, the $\{\tilde{\varepsilon}_{1,i}, \tilde{\varepsilon}_{2,i}\}_i$'s and the $W(\mathbf{X}_i)$'s are independent. Such a model is related to the class of Finite Mixtures of Regression (FMR), see Grün and Leisch (2006) for a good overview. Briefly, statistical inference for the class of parametric FMR model was first considered by Quandt and Ramsey

(1978) who proposed a moment generating function based estimation method. An EM estimating approach was proposed by De Veaux (1989) in the two-component case. Variations of the latter approach were also considered in Jones and McLachlan (1992) and Turner (2000). Hawkins et al. (2001) studied the estimation problem of the number of components in the parametric FMR model using approaches derived from the likelihood equation. In Hurn et al. (2003), the authors investigated a Bayesian approach to estimate the regression coefficients and also proposed an extension of the model in which the number of components is unknown. Zhu and Zhang (2004) established the asymptotic theory for maximum likelihood estimators in parametric FMR models. More recently, Städler et al. (2010) proposed an ℓ_1-penalized method based on a Lasso-type estimator for a high-dimensional FMR model with $d \geq n$. As an alternative to parametric approaches to the estimation of a FMR model, some authors suggested the use of more flexible semiparametric approaches. These approaches can actually be classified into two groups: semiparametric FMR (SFMR) of type I and type II. We say a mixture model is of type I when the mixture probability and location parameters are euclidean, but the mixing distribution is non parametric, whereas a model is of type II when, the other way around, the mixture probability and location are non parametric but the mixing density is known or belongs to a parametric family.

The study of SFMR of type I comes from the seminal work of Hall and Zhou (2003) in which d-variate semiparametric mixture models of random vectors with independent components were considered. These authors proved in particular that, for $d \geq 3$, we can identify a two-component mixture model without parametrizing the distributions of the component random vectors. To the best of our knowledge, Leung and Qin (2006) were the first in estimating a FMR model semiparametrically in that sense. In the two-component case, they studied the case where the components are related by Anderson (1979)'s exponential tilt model. Hunter and Young

(2012) studied the identifiability of an m-component type I SFMR model and numerically investigated a Expectation-Maximization (EM) type algorithm for estimating its parameters. Vandekerkhove (2013) proposed an M-estimation method for a two-component semiparametric mixture of linear regressions with symmetric errors (type I) in which one component is known. Bordes et al. (2013) revisited the same model by establishing new moment-based identifiability results from which they derived explicit $\sqrt{n}$-convergent estimators.

The study of type II SFMR models started with Huang and Yao (2012) who considered a semiparametric linear FMR model with Gaussian noise in which the mixing proportions are possibly covariates-dependent . They established also the asymptotic normality of their local maximum likelihood estimator and investigated a modified EM-type algorithm. Huang et al. (2013) generalized the latter work to nonlinear FMR with possibly covariates-dependent noises. Toshiya (2013) considered a Gaussian FMR model where the joint distribution of the response and the covariate (possibly functional) is itself modeled as a mixture. More recently Montuelle et al. (2013) considered a penalized maximum likelihood approach for Gaussian FMR models with logistic weights.

To improve the flexibility of our FMR model (10) and address the study of models involving design-dependent noises, such as the radiotherapy application from Bowen et al. (2012) displayed below, we will consider a slightly more general model:

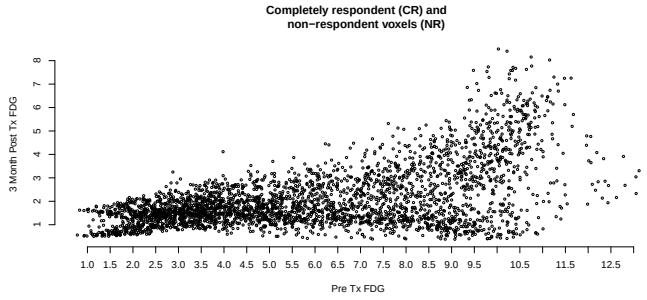


Figure 18: Display of the original PET-radiotherapy data from Bowen et al. (2012)

$$Y_i = W(\mathbf{X}_i)(a(\mathbf{X}_i) + \varepsilon_{1,i}(\mathbf{X}_i)) + (1 - W(\mathbf{X}_i))(b(\mathbf{X}_i) + \varepsilon_{2,i}(\mathbf{X}_i)), \tag{11}$$

such that, given $\{\mathbf{X} = \mathbf{x}\}$, the common p.d.f. of the $\varepsilon_{j,i}(\mathbf{x})$, $j = 1, 2$, denoted $f_{\mathbf{x}}$, is zero-symmetric. We will say that the above model is of type III, i.e. it combines type I and type II properties. Indeed, no parametric assumption is made about the mixing distribution of the errors nor about the mixing proportion and the location parameters, which are possibly design dependent. Our model is still said *semiparametric* because, given $\{\mathbf{X} = \mathbf{x}\}$, the vector $\theta(\mathbf{x}) = (\pi(\mathbf{x},)a(\mathbf{x}), b(\mathbf{x}))$ will be viewed as an Euclidean parameter to be estimated.

Examples of design-point noise dependency.

i) (Topographical scaling) The most natural transformation is probably when considering a topographical scaling of the errors, with $\sigma : \mathbb{R}^d \to \mathbb{R}_+^*$, such that $\varepsilon_{j,i}(\mathbf{X}_i) = \sigma(\mathbf{X}_i)\tilde{\varepsilon}_{j,i}$, $j = 1, 2$, where the $\tilde{\varepsilon}_{j,i}$'s are similar to those involved in (10). The conditional p.d.f of the errors $\varepsilon_{j,i}$ given $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$f_{\mathbf{x}}(y) = \frac{1}{\sigma(\mathbf{x})} f\left(\frac{y}{\sigma(\mathbf{x})}\right), \quad y \in \mathbb{R}. \tag{12}$$

Indeed, if f is zero-symmetric then the errors' distribution inherits trivially the same symmetry property.

ii) (Zero-symmetric varying mixture) Another useful example could be the varying mixing proportion mixture model of r zero-symmetric distributions. For $k = 1, \ldots, r$, we consider proportion functions $\lambda_k : \mathbb{R}^d \to (0, 1)$ with $\sum_{k=1}^r \lambda_k(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^d$. The conditional p.d.f of the errors $\varepsilon_{j,i}$ given $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$f_{\mathbf{x}}(y) = \sum_{k=1}^{r} \lambda_k(\mathbf{x}) f_k(y), \quad y \in \mathbb{R},$$

where the f_k functions are zero-symmetric p.d.f.'s.

iii) (Antithetic location model) Consider a location function $\mu : \mathbb{R}^d \to \mathbb{R}$ and f any arbitrary p.d.f. The conditional p.d.f of the errors $\varepsilon_{j,i}$ given $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$f_{\mathbf{x}}(y) = \frac{1}{2} f(y - \mu(\mathbf{x})) + \frac{1}{2} f(-y + \mu(\mathbf{x})), \quad y \in \mathbb{R},$$

and also results into a zero-symmetric p.d.f.

Note that any combination of the above situations could be considered in model (11) free from specifying any parametric family (provided the resulting zero-symmetry hold). This last remark reveals, in our opinion, the main strength of our model in the sense that it could prove to be a very flexible exploratory tool for the analysis of shifted response-type experiments. Our paper is organized as follows. Section 2 is devoted to identifiability results and a detailed description of our estimation method, while Section 3 is concerned with its asymptotic properties. The finite-sample performance of the proposed estimation method is studied for various scenarios through Monte Carlo experiments in Section 4. In Section 5 we propose to analyze the Positron Emission Tomography (PET) imaging data considered in Bowen et al. (2012). Finally Section 6 is devoted to auxiliary results and main proofs.

4.2 Estimation method

Let us define the joint density of a couple $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, designed from model (11):

$$g(y, \mathbf{x}) = [\pi(\mathbf{x}) f_{\mathbf{x}}(y - a(\mathbf{x})) + (1 - \pi(\mathbf{x})) f_{\mathbf{x}}(y - b(\mathbf{x}))] \ell(\mathbf{x}), \quad (y, \mathbf{x}) \in \mathbb{R}^{d+1}, \quad (13)$$

while the conditional density of Y given $\{\mathbf{X} = \mathbf{x}\}$ (denoted for simplicity $Y/\mathbf{X} = \mathbf{x}$) is

$$g_{\mathbf{x}}(y) = g(y, \mathbf{x})/\ell(\mathbf{x}) = \pi(\mathbf{x}) f_{\mathbf{x}}(y - a(\mathbf{x})) + (1 - \pi(\mathbf{x})) f_{\mathbf{x}}(y - b(\mathbf{x})). \quad (14)$$

We are interested in estimating the parameter $\theta_0 = \theta(\mathbf{x}_0) = (\pi(\mathbf{x}_0), a(\mathbf{x}_0), b(\mathbf{x}_0))$ at some fixed point $\mathbf{x}_0$ belonging to the interior of the support of ℓ ($\ell(\mathbf{x}_0) > 0$), denoted $supp(\ell)$. For simplicity and identifiability matters, we will suppose that θ_0 belongs to the interior of the parametric space $\Xi = [p, P] \times \Delta$, where $0 < p \leq P < 1$ and Δ denotes a compact set of $\mathbb{R}^2 \backslash \{(a, a) : a \in \mathbb{R}\}$.

At fixed $\mathbf{x}_0$, we prove, following Bordes et al. (2006), that identifiability holds up to label switching. Indeed, in [13] authors restricted the set of parameters to $[p, P] \times \Delta$, where $0 < p \leq P < 1/2$. Another way to avoid label switching is to assume $0 < p \leq P < 1$ and $a < b$. In order to have global identifiability of our model, we assume that at some fixed point $\mathbf{x}$ we have $a(\mathbf{x}) < b(\mathbf{x})$ and that functions a and b are differentiable and transversal (i.e. at each crossing point $\mathbf{x}$ where $a(\mathbf{x}) = b(\mathbf{x})$ gradients are different). The rest of this Section is dedicated to identifiability of the model and the estimation procedure.

4.2.1 Mixture of regression models as an inverse problem

We see in formula (14), that the conditional density of Y given $\{\mathbf{X} = \mathbf{x}\}$ can be viewed as a mixture of the errors distribution $f_{\mathbf{x}}$ given $\{\mathbf{X} = \mathbf{x}\}$ with locations $(a(\mathbf{x}), b(\mathbf{x}))$ and mixing proportion $\pi(\mathbf{x})$. Mixture of populations with different locations is a well known inverse problem. Our inversion procedure is done in Fourier domain.

For any function g in $\mathbb{L}_1 \cap \mathbb{L}_2$, let us define its Fourier transform by

$$g^*(u) = \int \exp(iuy) g(y) dy \text{ for all } u \in \mathbb{R}.$$

Here, the estimation method is based on the Fourier transform of the conditional density $g_{\mathbf{x}}(y)$ of $Y/\mathbf{X} = \mathbf{x}$. If the p.d.f. $f_{\mathbf{x}}$ belongs to $\mathbb{L}_1 \cap \mathbb{L}_2$ then so does $g_{\mathbf{x}}$. Denote its Fourier transform by $g^*_{\mathbf{x}}(u)$ for all $u \in \mathbb{R}$. In our model, we observe that

$$g^*_{\mathbf{x}}(u) \quad = \quad \left(\pi(\mathbf{x}) e^{iua(\mathbf{x})} + (1 - \pi(\mathbf{x})) e^{iub(\mathbf{x})}\right) f^*_{\mathbf{x}}(u), \quad u \in \mathbb{R}.$$

Let us denote, for all $t = (\pi, a, b)$ in Ξ and u in $\mathbb{R}$,

$$M(t,u) := \pi e^{iua} + (1-\pi)e^{iub}. \tag{15}$$

Note that $|M(t,u)| \leq 1$ for all $(t,u) \in \Xi \times \mathbb{R}$. Then, we have

$$g^*_{\mathbf{x}}(u) \;\; = \;\; M(\theta(\mathbf{x}),u) f^*_{\mathbf{x}}(u).$$

We introduce for convenience $\omega := \{\omega(1), \omega(2)\}$ a permutation of set $\{1,2\}$, i.e. $\omega \in \{id, s\}$ where $s(1) = 2$ and $s(2) = 1$. For $t = (\pi, a, b)$, we denote $[t]_\omega := t\mathbb{I}_{\omega=id} + (1-\pi, b, a)\mathbb{I}_{\omega=s}$ the parameter affected by a permutation ω of the labels (label 1 corresponding to location a and label 2 corresponding to location b). Let us fix $\mathbf{x}_0 \in supp(\ell)$ such that $\theta(\mathbf{x}_0)$ belongs to the interior of Ξ, denoted $\overset{\circ}{\Xi}$. Noticing that the p.d.f. $f_{\mathbf{x}_0}$ is zero-symmetric we therefore have that $f^*_{\mathbf{x_0}}(u) \in \mathbb{R}$, for all $u \in \mathbb{R}$. If t belongs to Ξ, we prove in the next theorem the *picking* property

$$\Im\left(g^*_{\mathbf{x}_0}(u)\bar{M}(t,u)\right) = 0 \text{ for all } u \in \mathbb{R}, \text{ if and only if } \exists \omega \in \{id, s\} : t = [\theta(\mathbf{x}_0)]_\omega,$$

where $\Im : \mathbb{C} \to \mathbb{R}$ denotes the imaginary part of a complex number and $\bar{M}$ the complex conjugate of M. This result allows us to build a *contrast* function for the parameter $t \in \Xi$:

$$S(t) := S_{\mathbf{x}_0}(t) := \int \Im\left(g^*_{\mathbf{x}_0}(u)\bar{M}(t,u)\right)^2 \ell^2(\mathbf{x}_0) w(u) du. \tag{16}$$

The function $w : \mathbb{R}^d \to \mathbb{R}_+$ is a bounded p.d.f. which helps in computing the integral via Monte-Carlo method and solves integrability issues. We stress the fact that using ℓ^2 instead of ℓ comes from the fact that the contrast estimates a quadratic functional, rather than an expected value.

Remark. The idea of using Fourier transform in order to solve the inverse mixture problem was introduced in Butucea and Vandekerkhove (2014) for density models. In the regression models we deal with the conditional density of $Y/\mathbf{X} = \mathbf{x}_0$ and consider

that it could possibly exist $\mathbf{x}_0 \in supp(\ell)$ such that $\pi(\mathbf{x}_0) = 1/2$ and then $M(\theta(\mathbf{x}_0), u)$ can be 0. This has a major incidence on the definition of the function $S(t)$ where $\bar{M}(t, u)$ appears at the numerator (contrarily to Butucea and Vandekerkhove, [13] where $M(t, u)$ appeared at the denominator). Moreover, smoothing of the information that data bring at a fixed design point $\mathbf{x}_0$ changes dramatically the behavior of the estimators as we shall see later on.

4.2.2 Local and global identifiability

We prove in the following theorem that our model is identifiable (up to a permutation of the labels) and that $S(t)$ defines a contrast on the parametric space Ξ.

Theorem 1 *(Identifiability and contrast property) Consider model (11) provided with $f_{\mathbf{x}}(\cdot) \in \mathbb{L}_2$ for all $\mathbf{x} \in \mathbb{R}^d$. For a fixed point $\mathbf{x}_0$ in the interior of the support of ℓ, we assume that $f_{\mathbf{x}_0}(\cdot)$ is zero-symmetric and that $\theta_0 = \theta(\mathbf{x}_0)$ is an interior point of Ξ. Then we have the following properties:*

i) The Euclidean parameter $\theta_0 = (\pi(\mathbf{x}_0), a(\mathbf{x}_0), b(\mathbf{x}_0))$ is identifiable up to a permutation of the labels when the function $f_{\mathbf{x}_0}(\cdot)$ is uniquely identified.

ii) The function S in (16) is a contrast function, i.e. for all $t \in \Xi$, $S(t) \geq 0$ and $S(t) = 0$ if and only if there exists $\omega \in \{id, s\}$ such that $t = [\theta_0]_\omega$.

Proof. i) The local (for fixed $\mathbf{x}_0$) identifiability of model (14) over Ξ and the set $\mathcal{F}$ of zero-symmetric densities, i.e., using notations involved in (15), for all $(t, t') \in \Xi^2$ and $(f, f') \in \mathcal{F}^2$,

$$M(t, u)f^*(u) = M(t', u)f'^*(u) \Rightarrow \exists \omega \in \{id, s\} : \ t' = t_\omega \text{ and } f = f',$$

is deduced from the proof of Theorem 2.1 in Bordes, Mottelet and Vandekerkhove (2006). The main difference here is that we allow π to lie in $(0, 1)$ whereas in Bordes

et al. (2006) the proportion mixing parameter was constrained to belong to $[0, 1/2)$. This constraint was also an implicit lexicographical ordering to avoid multiple label-permuted mixture representation. When revisiting step by step the proof of the latter theorem, it appears that the condition $\pi \neq 1/2$ is essentially used to avoid spurious model representation when the mixing proportion is allowed to be equal to zero (see discussion of Case 1, top of p. 1223, and the counter-example, p. 1206, in Bordes et al., 2006). When $\pi > 0$, the discussion of equation (37) in Bordes et al. (2006) leads to two obvious solutions $(\pi, a, b) = (\pi', a', b')$ and $(\pi, a, b) = (1 - \pi', b', a')$. To prove that possibly spurious solutions are non-admissible, it suffices to adapt the re-parametrization in (38) of Bordes et al. (2006) to the cases $(a - a', b - b') \neq (0, 0)$ and $(a - b', b - a') \neq (0, 0)$, which basically leads to consider (by symmetry) the following conditions: for $\beta_1 = \pi\pi'$, $\beta_2 = \pi(1 - \pi')$, $\beta_3 = \pi'(1 - \pi)$, $\beta_4 = (1 - \pi)(1 - \pi')$,

- $\beta_3 + \beta_4 = 0 \Leftrightarrow \pi = 1$,
- $\beta_2 + \beta_3 = 0$ and $\beta_4 = 0 \Leftrightarrow \pi = 1$ or $\pi' = 1$,
- $\beta_3 = 0$ and $\beta_4 - \beta_2 = 0 \Leftrightarrow \pi' = 0$ and $\pi = 1/2$, or $\pi = 1$ and $\pi' = 1$,
- $\beta_2 = 0$ and $\beta_4 - \beta_3 = 0 \Leftrightarrow \pi = 0$ and $\pi' = 1/2$, or $\pi' = 1$ and $\pi = 1$.

Note that the above solutions are all non-admissible when $(\pi, \pi') \in (0, 1)^2$. From this remark, we deduce that the Euclidean part of model (14) is also identifiable, up to a permutation of the labels, over *our* parametric space Ξ (including $\pi = 1/2$). To identify now the local noise distribution, we proceed similarly to Step 3 in Bordes et al. (2006). Because for $\omega \in \{id, s\}$

$$M(t, u)f^*(u) = M(t_\omega, u)f'^*(u) = M(t, u)f'^*(u), \quad u \in \mathbb{R},$$

we have to consider the two following cases:

- $\pi \neq 1/2$. Since $|M(t, u)| \geq |1 - 2\pi| > 0$ we deduce $f^*_{\mathbf{x}} = f'^*_{\mathbf{x}}$ and $f_{\mathbf{x}} = f'_{\mathbf{x}}$.

- $\pi = 1/2$. Here it is to be observed that, for t fixed in Ξ, $M(t,u) = 0$ occurs to be null on a countable set of $\mathbb{R}$. Indeed,

$$M(t,u) = 0 \Leftrightarrow au = bu + \pi + 2k\pi,\ k \in \mathbb{Z} \Leftrightarrow u \in \left\{ \frac{\pi + 2k\pi}{a-b},\ k \in \mathbb{Z} \right\}.$$

  Nevertheless this behavior does not affect the identifiability of the noise distribution since we can conclude that the real functions f^* and f'^* coincide over $\mathbb{R}$ except on a countable set of isolated points which is equivalent, by a continuity argument, to the equality over the whole real line.

This concludes the proof of i).

ii) The proof is similar to the proof of Proposition 1 in Butucea and Vandekerkhove (2014), replacing $f^*(\cdot)$ and $g^*(\cdot)$ by $f^*_{\mathbf{x}_0}(\cdot)$ and $g^*_{\mathbf{x}_0}(\cdot)$, respectively, and noticing that $\ell(\mathbf{x}_0)$ is bounded away from zero. ■

Label switching and global identifiability. The label switching phenomenon relies on the fact that the writing of the likelihood of a mixture model is invariant when permuting the label of its components. For example, when considering a k-component mixture model, there exists up to $k!$ mixture representations of the same distribution. To avoid these multiple representations (which obviously affects the estimation methods and their interpretation) there exists many different approaches: i) in the parametric case, Teicher (1963) suggest, for example, to create a lexical ordering on the parametric space, ii) in the Bayesian case, some MCMC-based relabelling algorithms are proposed, see Celeux et al. (2000), Stephens (2000) or Yao and Lindsay (2009), iii) in the two-component semiparametric case, the mixture proportion affected to the first component is constrained to be less than 1/2, see Bordes et al. (2006). In our case, since we plan to estimate the conditional model (14) over a grid of design-points, it would be precisely great to non restrict the proportion mixture

function $\pi(\cdot)$ to be upper-bounded by $1/2$ and also to be able to deal with intersecting curve functions $a(\cdot)$ and $b(\cdot)$. To better understand these situations and propose some practical implementations, we propose now to state, using arguments similar to [37], the global identifiability of our model (13) when $d = 1$. For this purpose, let us introduce the concept of *transversality*.

Definition 2 *Let $\mathbf{x} \in \mathbb{R}$, and let $a(\mathbf{x})$ and $b(\mathbf{x})$ two continuously differentiable real curve-functions. We say that $a(\mathbf{x})$ and $b(\mathbf{x})$ are transversal if $(a(\mathbf{x})-b(\mathbf{x}))^2+\|\dot{a}(\mathbf{x})-\dot{b}(\mathbf{x})\|^2 \neq 0$, for any $x \in \mathbb{R}$, where $\|\cdot\|$ denotes the Euclidean norm.*

The transversality condition imposed on two real curve-functions $a(\mathbf{x})$ and $b(\mathbf{x})$ implies that if $a(\mathbf{x}) = b(\mathbf{x})$, then $\dot{a}(\mathbf{x}) \neq \dot{b}(\mathbf{x})$.

Proposition 1 *Let us suppose that supp(ℓ) is an interval of $\mathbb{R}$ and that $\pi(\mathbf{x}) \in (0, 1)$, respectively $a(\mathbf{x})$ and $b(\mathbf{x})$, is a continuous function, respectively are both differentiable real-functions. If $a(\mathbf{x}_0) < b(\mathbf{x}_0)$ at some fixed point $\mathbf{x}_0$ in the interior of the supp(ℓ) and if $a(\mathbf{x})$ and $b(\mathbf{x})$ are transversal then our model (13) is globally identifiable over supp(ℓ).*

Proof. Let us consider the subset of $\mathbb{R}$

$$\mathcal{E} = \{\mathbf{x}_k : \ a(\mathbf{x}_k) = b(\mathbf{x}_k)\},$$

where the parameter curves intersect . Since parameter curves are transversal, any point in $\mathcal{E}$ is an isolated point. This implies that the set $\mathcal{E} \subset \mathbb{R}$ has no finite accumulation (limit) point and contains at most countably many points. Therefore, without loss of generality, we assume that: $\mathbf{x}_k < \mathbf{x}_{k+1}$ and $(\mathbf{x}_k, \mathbf{x}_{k+1}) \cap \mathcal{E} = \emptyset$, $k \in \mathbb{Z}$. Assume that the conditional model (14) admits another representation, i.e. there exist functions $(\pi', a', b', f'_{\mathbf{x}})$ such that

$$g_{\mathbf{x}}(y) = \pi'(\mathbf{x}) f'_{\mathbf{x}}(y - a'(\mathbf{x})) + (1 - \pi'(\mathbf{x})) f_{\mathbf{x}}(y - b'(\mathbf{x})).$$

We proved in i) of Theorem 1, that for any $\mathbf{x} \notin \mathcal{E}$, model (14) is identifiable, it follows that there exists a permutation $\omega_{\mathbf{x}} := \{\omega_{\mathbf{x}}(1), \omega_{\mathbf{x}}(2)\}$ of set $\{1, 2\}$, i.e. $\omega_{\mathbf{x}} \in \{id, s\}$ where $s(1) = 2$ and $s(2) = 1$, depending on $\mathbf{x}$ such that:

$$\begin{cases} \pi'(\mathbf{x}) = \pi(\mathbf{x}),\ a'(\mathbf{x}) = a(\mathbf{x}),\ b'(\mathbf{x}) = b(\mathbf{x})\ if\ \omega_{\mathbf{x}} = id, \\ \pi'(\mathbf{x}) = 1 - \pi(\mathbf{x}),\ a'(\mathbf{x}) = b(\mathbf{x}),\ b'(\mathbf{x}) = a(\mathbf{x})\ if\ \omega_{\mathbf{x}} = s. \end{cases}$$

(Since the parameter curves $(a(\mathbf{x}), b(\mathbf{x}))$ are continuous and do not intersect on any interval $(\mathbf{x}_k, \mathbf{x}_{k+1})$ the permutation $\omega(\mathbf{x})$ must be constant on the latter interval. In addition, for any $\mathbf{x}_k \in \mathcal{E}$, consider a small interval $(\mathbf{x}_k - \epsilon, \mathbf{x}_k + \epsilon)$ such that $(\mathbf{x}_k - \epsilon, \mathbf{x}_k + \epsilon) \in (\mathbf{x}_{k-1}, \mathbf{x}_{k+1})$. Now, since the parameter curves are transversal, they have different derivatives at $\mathbf{x}_k$, hence the permutation must be constant on the neighborhood $(\mathbf{x}_k - \epsilon, \mathbf{x}_k + \epsilon)$. Indeed, without lack of generality, suppose that $\omega_{\mathbf{x}} = id$ for $\mathbf{x} \in (\mathbf{x}_k, \mathbf{x}_k + \epsilon)$ and $\omega_{\mathbf{x}} = s$ for $\mathbf{x} \in (\mathbf{x}_k - \epsilon, \mathbf{x}_k)$, then the functions a' and b' are non-differentiable anymore since for example:

$$(\dot{a}')^+(\mathbf{x}_k) = \dot{a}(\mathbf{x_k}) \neq \dot{b}(\mathbf{x}_k) = (\dot{a}')^-(\mathbf{x}_k), \tag{17}$$

where $(\dot{a}')^+(\mathbf{x}_k)$ and $(\dot{a}')^-(\mathbf{x}_k)$ denote respectively the right and left side derivative of $a'(\cdot)$ at point $\mathbf{x}_k$. Therefore there exists a permutation ω independent of $\mathbf{x} \in supp(\ell)$ such that

$$\begin{cases} \pi'(\mathbf{x}) = \pi(\mathbf{x}),\ a'(\mathbf{x}) = a(\mathbf{x}),\ b'(\mathbf{x}) = b(\mathbf{x})\ if\ \omega = id, \\ \pi'(\mathbf{x}) = 1 - \pi(\mathbf{x}),\ a'(\mathbf{x}) = b(\mathbf{x}),\ b'(\mathbf{x}) = a(\mathbf{x})\ if\ \omega = s, \end{cases}$$

which concludes the proof of the global identifiability. ■ *Rules under the thumb.* The proof of the above proposition inspires us two practical approaches to handle the label switching problem and lack of identifiability at curve intersection points.

- Label switching. Let us consider, without loss of generality, two nearest neighbors $(\mathbf{x}_1, \mathbf{x}_2)$ over a grid of testing points. Suppose that $a(\mathbf{x}_1)$ and $b(\mathbf{x}_1)$ are identified well separated and (λ, α, β) is a minimizer of $S_{\mathbf{x}_2}(\cdot)$, i.e. $S_{\mathbf{x}_2}(\lambda, \alpha, \beta) = 0$.

Since no big jump is expected by moving from $\mathbf{x}_1$ to $\mathbf{x}_2$, a way to decide which solution is more likely acceptable between $t_1 = (t_{1,i})_{1\leq i\leq 3} = (\lambda, \alpha, \beta)$ and $t_2 = (t_{2,i})_{1\leq i\leq 3} = (1-\lambda, \beta, \alpha)$ could be to select the t with index $r \in \{1, 2\}$ satisfying

$$r = \arg\min_{i\in\{1,2\}} \{|t_{i,2} - a(\mathbf{x}_1)| + |t_{i,3} - b(\mathbf{x}_1)|\}. \tag{18}$$

This approach allows actually to get a sort of prior ordering very similar to the lexicographical ordering proposed by Teicher (1963).

- Crossing point. Let us consider, without loss of generality, three points $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ for which it is known that $a(\mathbf{x}_1) < b(\mathbf{x}_1)$ and $a(\mathbf{x}_3) > b(\mathbf{x}_3)$. If $\mathbf{x}_1$ and $\mathbf{x}_3$ are close enough, we can suspect that $\mathbf{x}_2$ is in the neighborhood of a crossing point, i.e. $a(\mathbf{x}_2) \simeq b(\mathbf{x}_2)$ and decide to linearly interpolate between $\mathbf{x}_1$ and $\mathbf{x}_3$, which for $v = \pi,\ a$ or b leads to

$$v(\mathbf{x}_2) \simeq \frac{v(\mathbf{x}_3) - v(\mathbf{x}_1)}{\mathbf{x}_3 - \mathbf{x}_1}(\mathbf{x_2} - \mathbf{x}_1) + v(\mathbf{x}_1). \tag{19}$$

Note that for v in $\mathcal{C}^k$, $k \geq 1$, we can use an interpolating polynomial of degree k.

Remark. For mixture models with higher number of components, i.e.

$$Y_i = \sum_{j=1}^{J} W_j(\mathbf{X}_i)(\gamma_j(\mathbf{X}_i) + \varepsilon_{j,i}(\mathbf{X}_i)), \quad i = 1, \ldots, n,$$

where $(W_1(\mathbf{x}), ..., W_J(\mathbf{x}))$ are distributed according to a J-components ($J > 2$) multinomial distribution with parameters $(\pi_1(\mathbf{x}), ..., \pi_J(\mathbf{x}))$, and noises $(\varepsilon_{j,i})$, $j = 1, \ldots, J$, i.i.d. according to $f_{\mathbf{x}}$, we assume that there exists a compact set $\Psi \subset]0, 1[^{J-1} \times \mathbb{R}^J$ of parameters $(\pi_1(\mathbf{x}), ..., \pi_{J-1}(\mathbf{x}), \gamma_1(\mathbf{x}), ..., \gamma_J(\mathbf{x}))$ where the model is *identifiable*. Note that the 3-components mixture model has been studied closely in Bordes et al. (2006) and Hunter et al. (2007) where sufficient identifiability conditions were given. The

case where $d > 3$ is more involved for full description and it is still an open question. Identifiability of a location mixture of probability densities was proven in Balabdaoui and Butucea (2014) when the mixing density is a Pólya frequency function. In this setup, if the conditional density of the errors is a symmetric Pólya frequency function, the estimation procedure described hereafter can be adapted over the parameter space Ψ with analogous results.

4.2.3 Estimation procedure

In order to build an estimator of the contrast $S(t)$ defined in (16), a local smoothing has to be performed in order to extract the information that the random design $\mathbf{X}_1, ..., \mathbf{X}_n$ brings to the knowledge of the conditional law of $Y/\mathbf{X} = \mathbf{x}_0$. We use a kernel smoothing approach, but local polynomials or wavelet methods could also be employed. This smoothing is a major difference with respect to the density model considered in Butucea and Vandekerkhove (2014) and all the rates will depend on the smoothing parameter applied to the kernel function.

Estimation of $\theta(\mathbf{x}_0)$. We choose a kernel function $K : \mathbb{R}^d \to \mathbb{R}$ belonging to $\mathbb{L}_1$ and to $\mathbb{L}_4$ and some bandwidth parameter $h > 0$ to be described later on. For $\mathbf{x}_0 \in supp(\ell)$ fixed, we denote

$$\begin{aligned} Z_k(t,u,h) &:= \left(e^{iuY_k}\bar{M}(t,u) - e^{-iuY_k}\bar{M}(t,-u)\right) K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &= \left(e^{iuY_k}M(t,-u) - e^{-iuY_k}M(t,u)\right) K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &= 2 \cdot \Im(e^{iuY_k}M(t,-u))K_h(\mathbf{X}_k - \mathbf{x}_0), \end{aligned} \tag{20}$$

where $K_h(\mathbf{x}) := h^{-d}K(\mathbf{x}/h)$. Indeed, $\bar{M}(t,u) = M(t,-u)$ for all t and u. The empirical contrast of $S(t)$ is defined by

$$S_n(t) = -\frac{1}{4n(n-1)} \sum_{j \neq k, j,k=1}^{n} \int Z_k(t,u,h)Z_j(t,u,h)w(u)du, \tag{21}$$

where $w : \mathbb{R} \to \mathbb{R}_+^*$ is a bounded p.d.f., having a finite moment of order 4, i.e.

$\int u^4 w(u)du < \infty$. From this empirical contrast we then define the estimator

$$\hat{\theta}_n = \arg\inf_{t\in\Theta} S_n(t), \tag{22}$$

of $\theta_0 = \theta(\mathbf{x}_0)$ where the parametric space Θ is now constrained, for unicity of solution, according to a prior knowledge provided by the rule (18). For simplicity we will suppose that at the point of interest $\mathbf{x}_0$ we have $a(\mathbf{x}_0) < b(\mathbf{x}_0)$, which translate into:

$$\Theta = [p, P] \times \Delta_{ord}, \tag{23}$$

where $0 < p \leq P < 1$ and Δ_{ord} denotes a compact set of $\{(a, b) \in \mathbb{R}^2 :\ a < b\}$. We shall study successively the properties of $S_n(t)$ as an estimator of $S(t)$ and deduce consistency and asymptotic normality of $\hat{\theta}_n$ as an estimator of θ_0.

Estimation methodology for $f_{\mathbf{x_0}}$. For the estimation of the local noise density $f_{\mathbf{x}_0}$ we suggest to consider the natural smoothed version of the plug-in density estimate given in Butucea and Vandekerkhove (2013, Section 2.2), under the assumption that $\pi(\mathbf{x}_0) \neq 1/2$.

Let us denote by $\varphi(\mathbf{x}, y) = \ell(\mathbf{x}) f_{\mathbf{x}}(y)$. We plug $\hat{\theta}_n$ in the natural smoothed non-parametric kernel estimator of $\varphi(\mathbf{x}, y)$ deduced from (15), whenever the unknown parameter θ_0 is required. For $\mathbf{x}_0$ fixed, we consider the Fourier transform of $\varphi(\mathbf{x}_0, y)$: $\varphi^*_{\mathbf{x}_0}(u) = \ell(\mathbf{x}_0) f^*_{\mathbf{x}_0}(u) = \ell(\mathbf{x}_0) g^*_{\mathbf{x}_0}(u)/M(\theta_0, u)$. Provided that $\hat{\pi}_n \neq 1/2$, which insures $|M(\hat{\theta}_n, u)| \geq |1 - 2\hat{\pi}_n| \neq 0$, we estimate by

$$\varphi^*_{\mathbf{x}_0,n}(u) = \frac{1}{n}\sum_{k=1}^{n} \frac{Q^*(h_{1,n}u)e^{iuY_k}}{M(\hat{\theta}_n, u)} K_{h_{2,n}}(\mathbf{X_k} - \mathbf{x}_0),$$

where Q is a univariate kernel ($\int Q = 1$ and $Q \in \mathbb{L}_2$) and $(h_{1,n}, h_{2,n})$ are bandwidth parameters properly chosen. Note that $G^*_n(u) := Q^*(h_{1,n}u)/M(\hat{\theta}_n, u)$ is in $\mathbb{L}_1$ and $\mathbb{L}_2$ and has an inverse Fourier transform which we denote by $G_n(u/h_{1,n})/h_{1,n}$. Therefore,

the estimator of $\varphi(\mathbf{x_0}, y)$ is

$$\varphi_n(\mathbf{x}_0, y) = \frac{1}{nh_{1,n}} \sum_{k=1}^{n} G_n\left(\frac{y - Y_k}{h_{1,n}}\right) K_{h_{2,n}}(\mathbf{X_k} - \mathbf{x}_0).$$

Finally the estimator of $f_{\mathbf{x}_0}$ is obtained by considering

$$\hat{f}_{\mathbf{x}_0}(y) = \frac{f_n(y|\mathbf{x}_0)\mathbb{I}_{f_n(y|\mathbf{x}_0)\geq 0}}{\int_{\mathbb{R}} f_n(y|\mathbf{x}_0)\mathbb{I}_{f_n(y|\mathbf{x}_0)\geq 0} dy}, \quad \text{where} \quad f_n(y|\mathbf{x}_0) = \frac{\varphi_n(\mathbf{x}_0, y)}{\ell_n(\mathbf{x}_0)}. \tag{24}$$

where $\ell_n(\mathbf{x}_0) = \frac{1}{n}\sum_{k=1}^{n} K_{h_{2,n}}(\mathbf{X_k} - \mathbf{x}_0)$. The asymptotic properties of this local density estimator are not established yet but we strongly guess that the bandwidth conditions required to prove its convergence and classical convergence rate are similar to those found in the conditional density estimation literature, see Brunel et al. (2010) or Cohen and Le Pennec (2012).

4.3 Performance of the method

We give upper bounds for the mean squared error of $S_n(t)$. We are interested in consistency and asymptotic normality of $\hat{\theta}_n$ and this requires some small amount of smoothness $\alpha > 1$ for the functions π, a and b and for the p.d.f. of the errors. From now on, $\|v\|$ denotes the Euclidean norm of vector v.

We say that a function F is Hölder α-smooth if it belongs to the set of functions $L(\alpha, M)$ with $\alpha = k + \beta > 0$ ($k \in \mathbb{N}$ and $\beta \in (0, 1]$) and $M > 0$, such that F has k bounded derivatives and, for all multi-index $j = (j_1, ..., j_d) \in \mathbb{N}^d$ with $|j| := j_1 + ... + j_d = k$, we have

$$|F^{(j)}(\mathbf{x}) - F^{(j)}(\mathbf{y})| \leq M\|\mathbf{x} - \mathbf{y}\|^{\beta}, \quad (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d.$$

A1. We assume that the functions π, a, b, ℓ belong to $L(\alpha, M)$ with α, $M > 0$.

Remark. We may actually suppose that the functions appearing in our model have different smoothness parameters, but the rate will be governed by the smallest smoothness parameter.

An important consequence of this assumption is that the density ℓ is uniformly bounded by some constant depending only on α and M, i.e. $\sup_{\ell \in L(\alpha,M)} \|\ell\|_\infty < \infty$.

A2. Assume that $f_{\mathbf{x}}(\cdot) \in \mathbb{L}_1 \cap \mathbb{L}_2$ for all $\mathbf{x} \in \mathbb{R}^d$. In addition, we require that there exists a w-integrable function φ such that

$$|f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)| \leq \varphi(u)\|\mathbf{x} - \mathbf{x}'\|^\alpha, \quad (\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d,\ u \in \mathbb{R}.$$

Remark. Note that for the scaling model (12), if f is the $\mathcal{N}(0,1)$ p.d.f. and $\sigma(\cdot)$ is bounded and Hölder α-smooth, we have:

$$|f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)| \leq \frac{u^2}{2}|\sigma^2(\mathbf{x}) - \sigma^2(\mathbf{x}')| \leq C\frac{u^2}{2}\|\mathbf{x} - \mathbf{x}'\|^\alpha.$$

A3. We assume that the kernel K is such that $\int |K| < \infty$, $\int K^4 < \infty$ and that it satisfies also the moment condition

$$\int \|\mathbf{x}\|^\alpha |K(\mathbf{x})| d\mathbf{x} < \infty.$$

A4. The weight function w is a p.d.f. such that

$$\int (u^4 + \varphi(u))w(u)du < \infty.$$

The following results will hold true under the additional assumption on the kernel (see **A3**): $\int \mathbf{x}^j K(\mathbf{x})d\mathbf{x} = 0$, for all j such that $|j| \leq k$.

Proposition 2 *For each $t \in \Theta$ and $\mathbf{x}_0 \in supp(\ell)$ fixed, suppose $\theta_0 \in \overset{\circ}{\Theta}$ and that assumptions* **A1-A4** *hold. Then, the empirical contrast function $S_n(\cdot)$ defined in (21) satisfies*

$$E\left[(S_n(t) - S(t))^2\right] \leq C_1 h^{2\alpha} + C_2 \frac{1}{nh^d},$$

if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, where constants C_1, C_2 depend on Θ, K, w, α and M but are free from n, h, t and $\mathbf{x}_0$.

Theorem 3 (Consistency) *Let suppose that assumptions of Proposition 2 hold. The estimator $\hat{\theta}_n$ defined in (21-22) converges in probability to $\theta(\mathbf{x}_0) = \theta_0$ if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.*

The following theorem establishes the asymptotic normality of the estimator $\hat{\theta}_n$ of θ_0. Recall that $\theta_0 = \theta(\mathbf{x}_0)$ belongs to Θ and that there exists $L_* > 0$ such that $\ell(\mathbf{x}_0) \geq L_*$. We see that the local smoothing with bandwidth $h > 0$ deteriorates the rate of convergence to $\sqrt{nh^d}$ instead of $\sqrt{n}$ for the density model. In the asymptotic variance we will use the following notation:

$$\dot{J}(\theta_0, u) := \Im\left(-\dot{M}(\theta_0, u)\bar{M}(\theta_0, u)\right) f^*_{\mathbf{x}_0}(u)\ell(\mathbf{x}_0), \tag{25}$$

and

$$V(\theta_0, u_1, u_2) := 4 \cdot \int \Im\left(e^{iu_1 y}\bar{M}(\theta_0, u_1)\right) \cdot \Im\left(e^{iu_2 y}\bar{M}(\theta_0, u_2)\right) g_{\mathbf{x}_0}(y)dy, \tag{26}$$

where the function $M(\cdot, \cdot)$ is defined in (15). Note that $\dot{J}(\theta_0, \cdot)$ is uniformly bounded by some constant and that V is well defined for all $(u_1, u_2) \in \mathbb{R} \times \mathbb{R}$ and also uniformly bounded by some constant.

Theorem 4 (Asymptotic normality) *Suppose that assumptions of Proposition 2 hold. The estimator $\hat{\theta}_n$ of θ_0 defined by (21-22), with $h \to 0$ such that $nh^d \to \infty$ and such that $h^{2\alpha+d} = o(n^{-1})$, as $n \to \infty$, is asymptotically normally distributed:*

$$\sqrt{nh^d}(\hat{\theta}_n - \theta_0) \to N(0, \mathcal{S}) \quad \textit{in distribution,}$$

where $\mathcal{S} = \frac{1}{4}\mathcal{I}^{-1}\Sigma\mathcal{I}$, with

$$\mathcal{I} = -\frac{1}{2}\int \dot{J}(\theta_0, u)\dot{J}(\theta_0, u)^\top w(u)du,$$

and

$$\Sigma := \int\int \dot{J}(\theta_0, u_1)\dot{J}^\top(\theta_0, u_2)V(\theta_0, u_1, u_2)w(u_1)w(u_2)du_1du_2,$$

for $\dot{J}$ defined in (25) and V in (26).

The above results show that our estimator of θ_0 behaves like any nonparametric pointwise estimator. This is indeed the case and we provide in the next theorem the best achievable convergence rates uniformly over the large set of functions involved in our model, see assumptions **A1**-**A2**.

Theorem 5 (Minimax rates) *Suppose* **A1**-**A4** *and consider* $\mathbf{x}_0 \in supp(\ell)$ *fixed such that* $\ell(\mathbf{x}_0) \geq L_* > 0$ *for all* $\ell \in L(\alpha, M)$ *and* $\theta_0 = \theta(\mathbf{x}_0) \in \overset{\circ}{\Theta} \setminus \{1/2\}$. *The estimator* $\hat{\theta}_n$ *of* θ_0 *defined by (21-22), with* $h \asymp n^{-1/(2\alpha+d)}$, *as* $n \to \infty$, *is such that*

$$\sup E[\|\hat{\theta}_n - \theta_0\|^2] \leq C n^{-\frac{2\alpha}{2\alpha+d}},$$

where the supremum is taken over all the functions π, a, b, ℓ *and* f^* *checking assumptions* **A1**-**A2**. *Moreover,*

$$\inf_{T_n} \sup E[\|T_n - \theta_0\|^2] \geq c n^{-\frac{2\alpha}{2\alpha+d}},$$

where C, $c > 0$ *depend only on* α, M, Θ, K *and* w, *and the infimum is taken over the set of all the estimators* T_n *(measurable function of the observations* $(X_1, \ldots, X_n)$*) of* θ_0.

Proof hints. Throughout the proofs of the previous results we learn that the estimator $\hat{\theta}_n$ of θ_0, behaves asymptotically as $\dot{S}_n(\theta_0)$ which is a U-statistic with a dominant term whose bias is of order $h^{2\alpha}$ and whose variance is smaller than $C_2(nh^d)^{-1}$. The bias-variance compromise will produce an optimal choice of the bandwidth h of order $n^{-1/(2\alpha+d)}$ and a rate $n^{-\frac{2\alpha}{2\alpha+d}}$. It is the optimal rate for estimating a Hölder α-smooth regression function at a fixed point and the optimality results in the previous theorem are a consequence of the general nonparametric problem, see Stone (1977), Ibragimov and Has'minski (1981) and Tsybakov (2009).

4.4 Practical behaviour

4.4.1 Algorithm

We describe below the initialization scheme and the optimization method used to determine the estimates of the locations $a(\mathbf{x}_k)$, $b(\mathbf{x}_k)$ and the weight functions $\pi(\mathbf{x}_k)$ for a fixed sequence of testing points $\{\mathbf{x}_k,\ k = 1, \ldots, K\}$. To simply differentiate these testing points from the design data points we will allocate specifically the index k for

the numbering of the testing points and the index i for the numbering of the dataset points, i.e. $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$.

Initialization

1. For each design data point $\mathbf{x}_i$, $i = 1, \ldots, n$, fit a kernel regression smoothing $\bar{m}(\mathbf{x}_i)$ with local bandwidth $\bar{h}_{\mathbf{x}_i}$. The `R` package `lokerns`, see Herrmann (2013), can be used.

2. Classify each data point $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ according to: if $y_i > \bar{m}(\mathbf{x}_i)$ classify $(\mathbf{x}_i, y_i)$ in group 1 associated with location $a(\cdot)$, otherwise classify it in group 2 associated with $b(\cdot)$.

3. For each $\mathbf{x}_k$, $k = 1, \ldots, K$, obtain initial value $\bar{a}(\mathbf{x}_k)$, respectively $\bar{b}(\mathbf{x}_k)$, by fitting a kernel regression smoothing based on the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, previously classified in group 1 with local bandwidth $\bar{h}_{1,\mathbf{x}_k}$, respectively in group 2 with local bandwidth $\bar{h}_{2,\mathbf{x}_k}$.

4. Compute the local bandwidth $h_{\mathbf{x}_k} = \min(\bar{h}_{1,\mathbf{x}_k}, \bar{h}_{2,\mathbf{x}_k})$.

5. Fix an arbitrary single value $\bar{\pi}$ for all the $\pi(\mathbf{x}_k)$'s.

Estimation

1. Generate one w-distributed i.i.d sample (U_r), $r = 1, \ldots, N$ dedicated to the pointwize Monte Carlo estimation of $S_n(t)$ defined by:

$$S_n^{MC}(t) = -\frac{1}{4n(n-1)N} \sum_{j \neq k, j,k=1}^{n} \sum_{r=1}^{N} Z_k(t, U_r, h) Z_j(t, U_r, h).$$

   In the Sections 4.2 and 5, we will consider $N = n$ and w the p.d.f. corresponding to the mixture $0.1 \cdot \mathcal{N}(0,1) + 0.9 \cdot \mathcal{U}_{[-2,2]}$.

2. Compute the minimizer $\hat{\theta}(\mathbf{x}_k) = (\hat{\pi}(\mathbf{x}_k), \hat{a}(\mathbf{x}_k), \hat{b}(\mathbf{x}_k))$ of $S_n^{MC}(\cdot)$ evaluated at each point $\mathbf{x}_0 = \mathbf{x}_k$, by using the starting values $(\bar{\pi}, \bar{a}(\mathbf{x}_k), \bar{b}(\mathbf{x}_k))$ and the local bandwidth $h_{\mathbf{x}_k}$.

In our simulations, the above minimization will be deliberately done over a non-constrained space, i.e. generically $\theta(\cdot) \in [0.05, 0.95] \times [A, B]^2$, with $A < B$. Our goal is to analyze experimentally if a performant initialization procedure is able to prevent from spurious phenomenons like the label switching or component merging occurring when $\pi(\mathbf{x}_0)$ is close to 0.5. This kind of information is actually very relevant to interpret correctly some cross-over effects as the one we will observe in Fig. 6 (a). Note that other initialization methods can be figured out. We can for instance use, similarly to Huang et al. (2013), a mixture of polynomial regressions with constant proportions and variances to pick initial values $\bar{a}(x)$ and $\bar{b}(x)$, or the `R` package `flexmix`, see Gruen et al. (2013), that implements a general framework for finite mixture of regression models based on EM-type algorithms (we selected this latter approach for the analysis of radiotherapy application in Section 5).

4.4.2 Simulations

In this section, we propose to measure the performances of our estimator $\hat{\theta}_n(\cdot)$ over a testing sequence $\{\mathbf{x}_k = k/K\}$, $k = 1, \ldots, K = 20$. Given that in the simulation setting the true function $\theta(\cdot)$ is known, we can compute, similarly to Huang et al. (2013), the Root Average Squared Errors (RASE) of our estimator. To this end we generate $M = 100$ datasets $(\mathbf{X}_i^{[z]}, Y_i^{[z]})_{1 \leq i \leq n}$, $z = 1, \ldots, M$ of sizes n= 400, 800, 1200, for each of the scenario described below and, for each scalar parameter $s = a,\ b,\ \pi$, denote by $RASE_s^{[z]}$ the RASE performance associated to the z-th dataset, defined by $RASE_s^{[z]} = (1/K \sum_{k=1}^{K} R_s^{[z]}(k))^{1/2}$, where $R_s^{[z]}(k) = \left(\hat{s}^{[z]}(\mathbf{x}_k) - s(\mathbf{x}_k)\right)^2$, and the empirical RASE by

$$RASE_s = \frac{1}{M} \sum_{z=1}^{M} RASE_s^{[z]}. \tag{27}$$

Let us also define the empirical squared deviation at point $\mathbf{x}_k$ by $\nu_k = \frac{1}{M}\sum_{z=1}^{M} R_s^{[z]}(k)$, and empirical variance of the squared deviation at $\mathbf{x}_k$ by $\sigma_s^2(k) = \frac{1}{M-1}\sum_{z=1}^{M}\left(R_s^{[z]}(k) - \nu_k\right)^2$. From these quantities we deduce the averaged variance of the squared deviations defined by

$$\sigma_s^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_s^2(k). \tag{28}$$

In all the simulation setups, we use the same mixing proportion function $\pi(\cdot)$:

$$\pi(\mathbf{x}) \quad = \quad \frac{\sin(3\pi x) - 1}{15} + 0.4, \quad \mathbf{x} \in [0,1].$$

Gaussian setup (G). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Gaussian topographical scaling model corresponding to (12), i.e. f is the $\mathcal{N}(0,1)$ p.d.f. when the location and scaling functions are

$$a(\mathbf{x}) = 4 - 2\sin(2\pi\mathbf{x}), \quad b(\mathbf{x}) = 1.5\cos(3\pi\mathbf{x}) - 3, \quad \sigma(\mathbf{x}) = 0.9\exp(\mathbf{x}), \quad \mathbf{x} \in [0,1].$$

Student setup (T). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Student distribution with continuous degrees of freedom function denoted $df(\mathbf{x})$. The locations and degrees of freedom functions are

$$a(\mathbf{x}) = 3 - 2\sin(2\pi x), \quad b(\mathbf{x}) = 1.5\cos(3\pi x) - 2, \quad df(\mathbf{x}) = -5x + 8, \quad \mathbf{x} \in [0,1].$$

Laplace setup (L). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Laplace distribution with scaling function $\nu(\mathbf{x})$. The locations and scaling functions are

$$a(\mathbf{x}) = 5 - 3\sin(2\pi\mathbf{x}), \quad b(\mathbf{x}) = 2\cos(3\pi\mathbf{x}) - 4, \quad \nu(\mathbf{x}) = \mathbf{x} + 1, \quad \mathbf{x} \in [0,1].$$

The selected bandwidths, whose mean and standard deviation are reported in Table 4, are obtained at the initialization step and are extracted from the function `lokerns` of the `R`-library `lokern`. This function calculates an estimator of the regression function with an automatically chosen local plugin bandwidth function. The automatically chosen bandwidths are calculated by finding the bandwidths that minimize the

asymptotically optimal mean squared error. To estimate the variance component in the mean squared error this method estimates a functional of a smooth variance function for our heteroscedastic errors.

Comments on Tables 1-3. We report for the simulation setups **(G)**, **(T)** and **(L)** the quantities $RASE_s$ defined in (27), and between parenthesis σ_s^2 defined in (28), for $s = \pi,\ a,\ b$. In these tables, we label our method as NMR-SE (Nonparametric Mixture of Regression with Symmetric Errors). To illustrate the contribution of our method, we compare our results with the RASE obtained by using the local EM-type algorithm proposed by Huang et al. (2013) for Nonparametric Mixture of Regression models with Gaussian noises (method labeled for simplicity NMRG). When the errors of the simulated model are Gaussian, the NMRG estimation should outperform our method, since the NMRG method assumes correctly that the errors are normally distributed, while our method does not make any parametric assumption on the distribution of the errors. When the sample size $n = 400$, the NMRG is more precise than our method, since the $RASE_s$'s and σ_s^2's are both smaller for the NMRG . When we increase the sample size of the simulated datasets to $n = 800,\ 1200$, our method becomes more competitive and yields $RASE_s$'s and σ_s^2's that are lower than those obtained by NMRG . This surprising behavior is probably due to the fact that in model (11) we impose the equality in law of the noises up to a shift parameter, when in the NMRG approach possibly different variances are fitted to each kind of noise, increasing by the way drastically the degrees of freedom of the model to be addressed.

In Tables 2 and 3 we observe that our method has globally smaller $RASE_s$'s and σ_s^2's. This result is not surprising, given that in the estimation methodology of Huang et al. (2013), the distribution of the noise are then completely misspecified under the simulation setups **(T)** and **(L)**. Note however, that when the sample size is small $n = 400$, the NMRG displays better results, which can be explained by the fact that when we generate small size datasets, the points that are supposed to be in the tails

Table 5: Mean and Standard Deviation of RASEs for data with Gaussian Errors

Sample size	Method	$rase_\pi$	$rase_a$	$rase_b$
$n = 400$	NMRG	0.011 (0.015)	0.579 (1.064)	0.228 (0.374)
	NMR-SE	0.007 (0.011)	1.031 (2.061)	0.293 (0.581)
$n = 800$	NMRG	0.010 (0.013)	0.505 (0.986)	0.219 (0.401)
	NMR-SE	0.003 (0.005)	0.492 (0.998)	0.150 (0.269)
$n = 1200$	NMRG	0.009 (0.012)	0.474 (0.892)	0.215 (0.401)
	NMR-SE	0.002 (0.003)	0.311 (0.572)	0.123 (0.264)

of the non-normal distributions are less likely to appear in the dataset. So in that case it can be reasonable to assume that the Gaussian distribution approximates the errors distribution well.

Table 6: Mean and Standard Deviation of RASEs for data with Student Errors

Sample size	Method	$rase_\pi$	$rase_a$	$rase_b$
$n = 400$	NMRG	0.013 (0.018)	0.330 (0.557)	0.135 (0.196)
	NMR-SE	0.010 (0.016)	0.454 (0.932)	0.217 (0.473)
$n = 800$	NMRG	0.011 (0.014)	0.276 (0.530)	0.101 (0.156)
	NMR-SE	0.004 (0.007)	0.192(0.374)	0.175 (0.561)
$n = 1200$	NMRG	0.010 (0.014)	0.216 (0.433)	0.111 (0.165)
	NMR-SE	0.003 (0.005)	0.127 (0.255)	0.053 (0.096)

Table 7: Mean and Standard Deviation of RASEs for data with Laplacian Errors

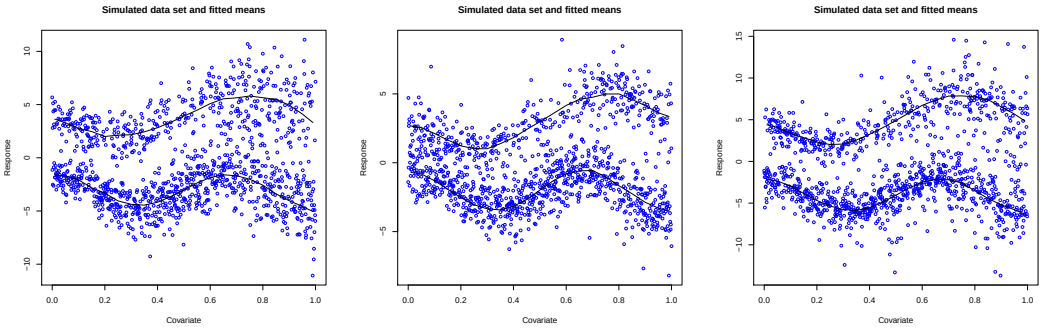
Sample size	Method	$rase_\pi$	$rase_a$	$rase_b$
$n = 400$	NMRG	0.011 (0.014)	0.815 (1.527)	0.323 (0.493)
	NMR-SE	0.007 (0.001)	1.242 (2.420)	0.376 (0.714)
$n = 800$	NMRG	0.010 (0.013)	0.659 (0.192)	0.283 (0.428)
	NMR-SE	0.003 (0.005)	0.489 (0.870)	0.191 (0.398)
$n = 1200$	NMRG	0.009 (0.012)	0.592 (1.072)	0.236 (0.346)
	NMR-SE	0.002 (0.003)	0.308 (0.566)	0.127 (0.2548)

Comments on Figures 1-6. To illustrate the sensitivity of our method and compare it graphically to the NMRG approach we plot in Fig. 1 different samples coming from the setups **(G)**, **(T)**, and **(L)** for $n = 1200$, and in blue lines the corresponding

Table 8: Mean and standard deviation of the `lokerns`-selected Bandwidth.

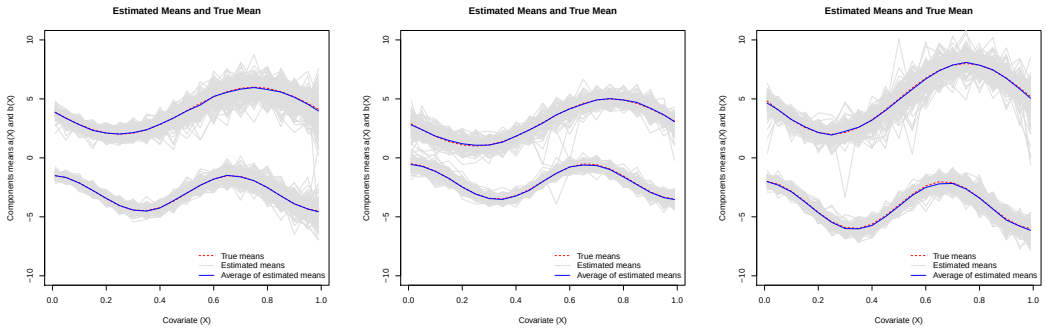
Sample size	Gauss	Student	Laplace
$n = 400$	0.0915 (0.0185)	0.0812 (0.0147)	0.0877(0.0220)
$n = 800$	0.0860(0.0099)	0.0780 (0.0091)	0.0823 (0.0151)
$n = 1200$	0.0813 (0.0072)	0.0743 (0.0061)	0.0791 (0.0122)

true location functions $a(\cdot)$ and $b(\cdot)$. In Fig. 2, respectively Fig. 3, we plot in grey the $M = 200$ segment-line interpolation curves obtained by connecting the points $(\mathbf{x}_k, \hat{s}^{[z]}(\mathbf{x}_k))$, $k = 1, \ldots, K$ where $s(\cdot) = a(\cdot), b(\cdot)$ for the NMRG method, respectively our NMR-SE method. In Fig. 4 and 5 we do the same for $s(\cdot) = \pi(\cdot)$. In Fig. 2-5 the dashed red lines represent the mean curves obtained by connecting the points $(\mathbf{x}_k, \bar{s}(\mathbf{x}_k))$, $k = 1, \ldots, K$ with $\bar{s}(\mathbf{x}_k) = 1/M \sum_{z=1}^{M} \hat{s}^{[z]}(\mathbf{x}_k)$ and $s(\cdot) = a(\cdot)$, $b(\cdot)$ and $\pi(\cdot)$. Let us observe first that the good behavior of the NMR-SE method is confirmed by the small variability of the curves in Fig. 3 and 5 compared to those in Fig. 2 and 4 corresponding to the NMRG method. Secondly it is important to notice that sometimes, since we did not constrained our method to have $\pi \in [p, P]$ with $0 < p < P < 1/2$, we run into some spurious estimation due to label switching or component merging phenomenon.



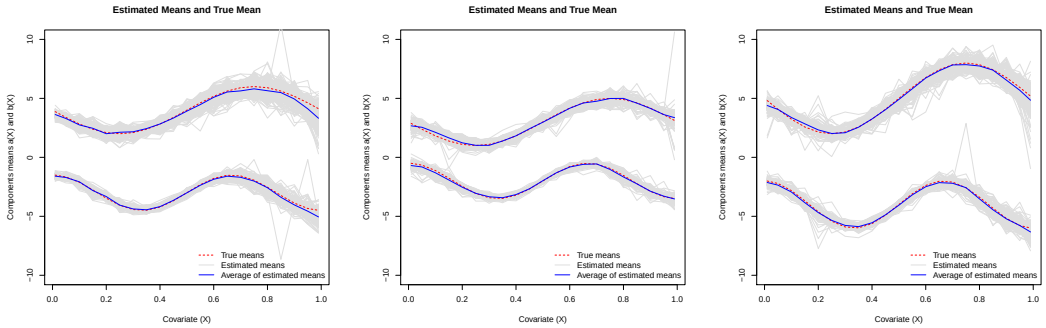
(a) Gaussian distribution **(b)** Student distribution **(c)** Laplace distribution

Figure 19: Examples of simulated datasets with different distribution errors



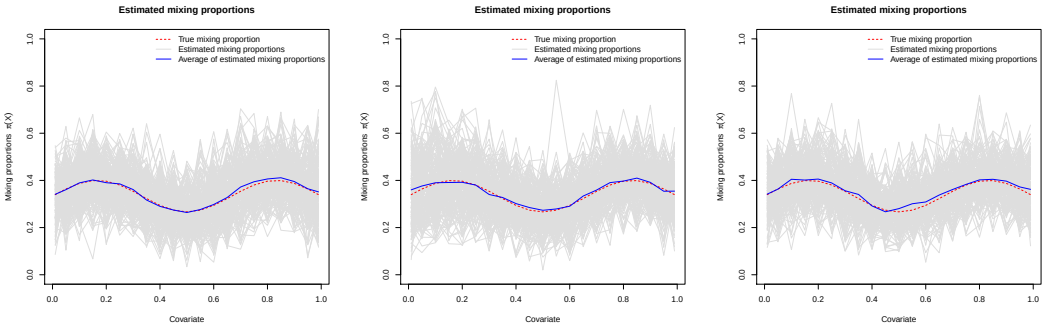
(a) Gaussian distribution **(b)** Student distribution **(c)** Laplace distribution

Figure 20: Mean Curves estimated with NMRG



(a) Gaussian distribution **(b)** Student distribution **(c)** Laplace distribution

Figure 21: Mean Curves estimated with NMR-SE

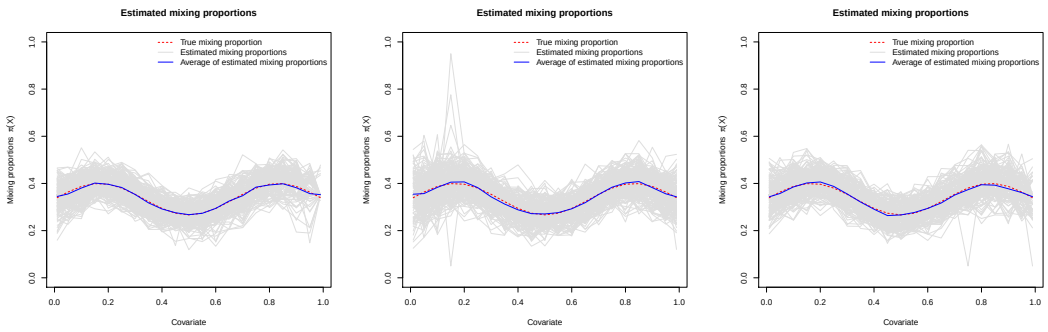


(a) Gaussian distribution **(b)** Student distribution **(c)** Laplace distribution

Figure 22: Mixing proportions estimated with NMRG

4.5 Application in radiotherapy

In this section, we implement the proposed methodology to a dataset obtained from applying Positron Emission Radiotherapy (PET) to a canine patient with locally advanced Sinonasal Neoplasia. These data were provided by Bowen et al. (2012, Fig.



(a) Gaussian distribution **(b)** Student distribution **(c)** Laplace distribution

Figure 23: Mixing proportions curves estimated with NMR-SE

4) who used them to quantify the associations between pre-radiotherapy and post-radiotherapy PETparameters via spatially resolved mixture of linear regressions. Intensity Modulated Radiotherapy is an advanced radiotherapy method that uses computer controlled device to deliver radiation of varying intensities to tumor or smaller areas within the tumor. There is evidence showing that the tumor is not homogeneous in its response to the radiation, and that some regions are more resistant than others. Functional imaging techniques (such as Positron Emission Tomography) can be used to identify the radiotherapy resistant regions within the tumor. For instance, an uptake in PET imaging of follow-up 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG) is empirically linked to a local recurrence of the disease. Bowen et al. (2012), use this approach to construct a prescription function that maps the image intensity values into a local radiation dose that will maximize the probability of a desired clinical outcome. In their manuscript they validate the use of molecular imaging based prescription function against clinical outcome by establishing an association between imaging biomarkers (PET imaging pre-radiotherapy) and regional imaging response to known dosage of therapy (PET imaging post-radiotherapy). The regional imaging response captures the change in imaging signal over an individual image volume element (called a voxel). In our model of interest (11), the pre-radiotherapy PET imaging intensities correspond to the input $\mathbf{X}_i$'s, and the post-radiotherapy PET

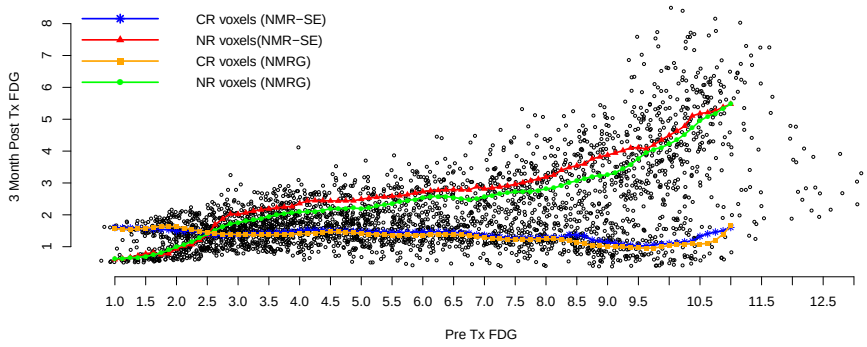
imaging levels are the outputs Y_i's. For many patients, the empirical link between post-treatment PET of FDG (regional imaging response) and pre-treatment PET of FDG (imaging biomarker at baseline) is well captured by a mixture regression model with two components. For a set of voxels with similar pre-treatment PET intensities, the nature of the response to the radiotherapy leads to two groups of voxels. The first group corresponds to voxels that respond well to the radiotherapy, and the second group contains the non-responding voxels. In our model of interest (11), the non-responding voxel group corresponds to the case where $W(\mathbf{X}_i) = 1$. The location parameters of each group appears to change as the pre-radiotherapy imaging intensity $\mathbf{X}_i$ varies. These changes in location are captured in our model by the location functions $a(\cdot)$ or $b(\cdot)$, where $a(\cdot)$, respectively $b(\cdot)$, is the component mean function for the completely responding (CR), respectively non-responding (NR), voxel. Additionally, the proportion of voxels $\pi(\mathbf{X}_i)$ that respond well to treatment depends on the pre-treatment level of the PET, so the mixture model should also account for a mixing proportion that depends on the input $\mathbf{X}_i$. For a given input $\mathbf{x}$, we assume that the intensity level of the completely responding and the non-responding voxel have approximately the same p.d.f. $f_{\mathbf{x}}$ up to a shift parameter, with the topographical scaling structure (12) presented in the Introduction. The variance of the distribution also changes with the level of the covariate (pre-treatment PET FDG). In many cases the variance increases as the intensity of a voxel's PET pre-radiotherapy increases, this is simply due to the fact the responding voxels will have a low post-treatment PET intensity, while the non-responding voxels will not. The aforementioned topographical scaling property, will allow to model this behavior. To obtain initial values for the location curves $a(\cdot)$ and $b(\cdot)$, we first use the `R` package `flexmix`, see Gruen et. al (2013), which allows us to fit defined parametric functions to the mixture. For the mixing proportion function we set a fixed constant value $\bar{\pi}(\mathbf{x}) = 0.4$. The bandwidths are computed according to the methodology described in Section 4.1, except that the

groups are now determined as an output of the `flexmix` package. The behavior of the local bandwidths selected by the `flexmix` package is displayed in Fig. 8.

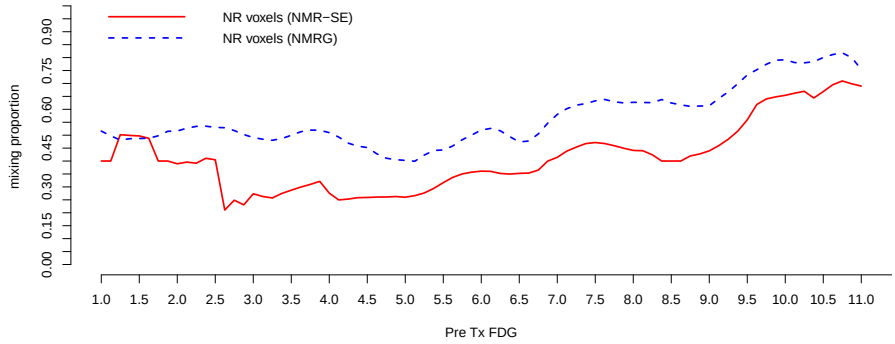
We propose to apply the NMRG and NMR-SE to this dataset. In figure 6(a), we show the PET image response to radiotherapy at 3 months, measured by FDG PET uptake, versus the pre-treatment FDG PET uptake. We also display component means obtained by fitting the NMRG and the NMR-SE. For both methods, we observe that the location functions b(x) corresponding to the completely responding voxels, show little variation across the range of values of pre-treatment FDG PET. NMRG and NMR-SE yield fitted means b(x) that are pretty similar to each other.

The fitted location functions $a(x)$ are associated with the non-responding voxels. For both methods, the estimated component means $a(x)$ increase with the pre-treatment FDG PET uptake. A significant difference between NMR-SE and NMRG lies in the fact that the estimated location function $a(x)$ of NMR-SE is slightly greater than the estimated location function obtained with NMRG. This implies that more voxels will be attributed to the non-responding group when we use NMRG instead of NMR-SE. This is confirmed by the figure (6b), where we display the mixing proportions $\pi(x)$ for each method. As expected, we see that the NMRG yields mixing proportions of non-responding voxels that are larger than the mixing proportions obtained by using our method. The NMRG mixing proportions lies between (40% and 70 %), while the NMR-SE mixing proportions is between (18% and 60%). The NMR-SE mixing proportion of non-responding voxels is less than 40% for this patient when pre-treatment FDG PET uptake is between 2.75 SUV and 6.875 SUV. We can conclude based on the results from our method that the current radiation dose could be appropriate for patients that exhibit pre-treatment FDG PET uptake close to the range aforementioned. On the other hand, NMRG doesn't present a wide range of pre-treatment FDG uptake where the non-responding mixing proportion is less than 50%. We see in addition in Fig. 9 that the conditional distributions, obtained from

formula (24) with $h_{1,n} = h_{2,n} = 0.2$, are about zero-symmetric with reasonably small trimming effect due to $\mathbb{I}_{f_n(y|\mathbf{x}_0)\geq 0}$ in (24) (tiny wave effect on both sides of the main mode). This is a good model validation tool since we are actually able to recover, after local Fourier inversion, the basic symmetry assumption technically made on the distributions of the errors; see for quality comparison other existing (nonconditional) semiparametric inversion density estimates performed on real datasets: Fig. 1-2 (a) in Bordes et al. (2006), Fig. 3 in Butucea and Vandekerkhove (2014), Fig. 5 in Vandekerkhove (2013), or Fig. 2-3 in Bordes et al. (2013).



(a) Scatter of plots of pre-treatment FDG PET vs. post-treatment FDG PET and estimated location functions for the completely respondent and non-respondent voxel subpopulations



(b) Estimated mixing proportions for the completely (CR) and non-respondent (NR) voxel subpopulation

Figure 24: Location and mixing proportion function estimation by using NMR-SE and NMRG methods

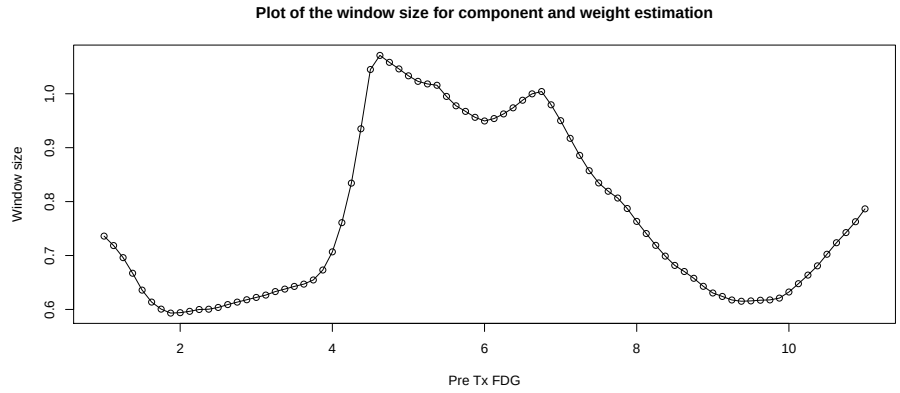


Figure 25: Behavior of the local bandwidths selected by the `flexmix` package in the PET application

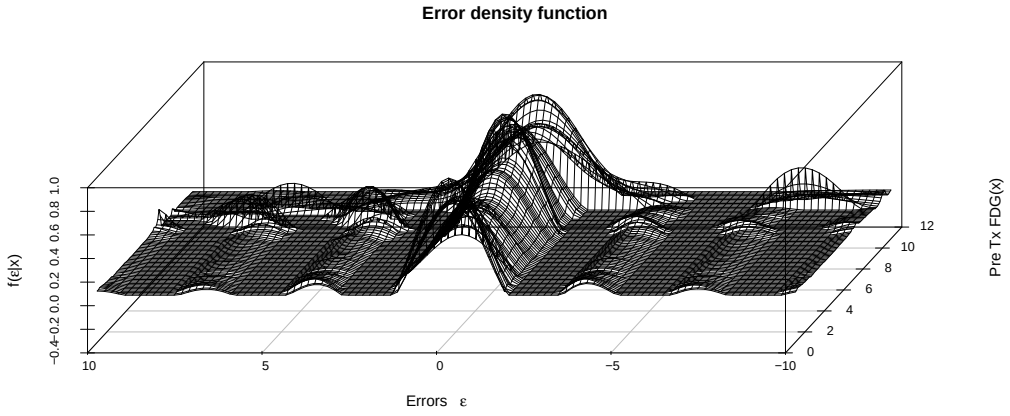


Figure 26: Density Estimates of the errors for the different levels of PET Tx FDG values

CHAPTER V

CONCLUSION

The main topic of this thesis is the analysis of various estimation and model selection methods that can be applied to supervised learning models. The first theme of the thesis dealt with structured sparsity studied from a statistical, algorithmic and applied perspective. The second topic of the thesis is related to nonparametric mixture regression models. We have introduced these problems by describing applications justifying the relevance and usefulness of the models proposed in this thesis.

Our first contribution lies in the use of structured sparsity inducing norms in the context of multivariate time series. We have explained how to leverage prior spatio-temporal information to design sparsity promoting norms that can generate zeros or non-zeros patterns that are optimal from a prediction perspective and that lead to interpretable spatio temporal models. We have proposed an efficient and intuitive algorithm that is based on soft-thresholding of autoregressive parameters and that is also built upon block-coordinate descent procedures. The methodology is applied to a synthetic dataset and is compared to other state of the art regularization methods. To show the usefulness of the proposed method, we forecast a set of state level economic time series using the model selected by optimization algorithm. The VAR time series model obtained matches common economic intuition, since we find that states that are located close to each other tend to influence each other more often.

In the third chapter, we present a functional sparsity inducing methodology that

is suited to fit high dimensional additive multivariate regression or classification models. In this chapter, we also exploits structured sparsity that is implied by the prior information available through group of predictors. Our contribution lies in the ability to select additive functions that are relevant for all responses or categories, but more importantly the optimization model is flexible enough such that an additive function can be selected for a response (or a category) and not be selected for others. To the best of our knowledge, this is the first paper that jointly account for nonlinearities in a multivariate regression (or classification) context and can induce within group sparsity. A new functional block coordinate descent algorithm is developed by using a pertubation of a functional of the objective function and deriving functional subdifferentials of the lagragian constraints present in the optimization problem. By applying the methodology to a benchmark cancer data set, we are able to perfectly classify all the patients to all the cancer categories by using only 12 out of 2308 genes. The state of the art methodology achieve 100% classification rate using at least 20 genes. We also use the model to understand the determinant of health that drive the county level cost of care in North Carolina from 2005 to 2009.

In the final chapter of this thesis, we introduced a semiparametric topographical mixture model that can be applied to characterize nonparametric mixture regression models where the response is dichotomous. The nonparametric nature of the mixture regression model is due to the fact that the locations and proportion functions are nonlinear and depend on a predictor and the density function associated with the errors of the model are only known to be symmetric. We proposed a pointwise contrast based estimation procedure of the proportion and locations functions that rely only on the symmetry of the local noise. A important contribution of the method lies in the fact that it does not impose constraints of finiteness on the moments of the errors. We also studied the asymptotic properties of the estimator by establishing

under mild conditions its minimax properties and its asymptotic normality. We then compared the method to state of the art parametric methods on simulated data and on a Positron Emission Tomography image that can be used for modulating the intensity of radiotherapy treatments of tumors in canine patients.

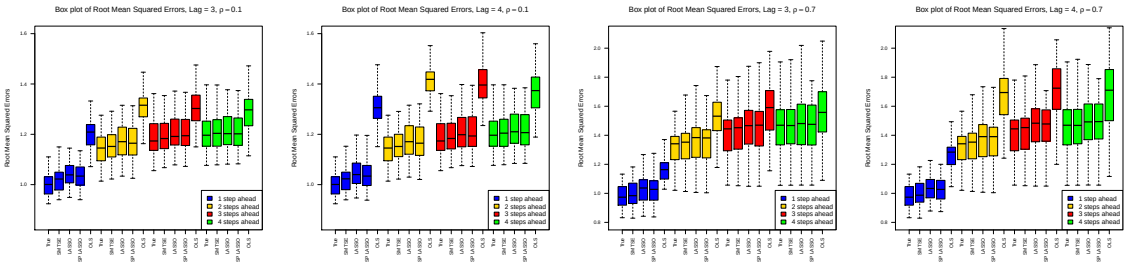
In this thesis, we have limited our application of sparsity inducing norms to supervised learning problems. These methods have also been applied to unsupervised learning problems, for example clustering. We are currently exploring ways to incorporate to semi-supervised learning problems where the data are collected in a spatio-temporal setting. The main idea will be to create sparsity promoting norms that can be used to penalize a maximization likelihood function but they could also be used to create subsets of geographical clusters within which the predictions can be refined and more accurate. By using fusion penalties, we will be able to simultaneously improve predictions and form prediction driven geographical clusters.

APPENDIX A

SUPPLEMENT TO LARGE VECTOR AUTOREGRESSION FOR SPATIALLY CORRELATED TIME SERIES

A.1 Additional prediction results for the simulation Study

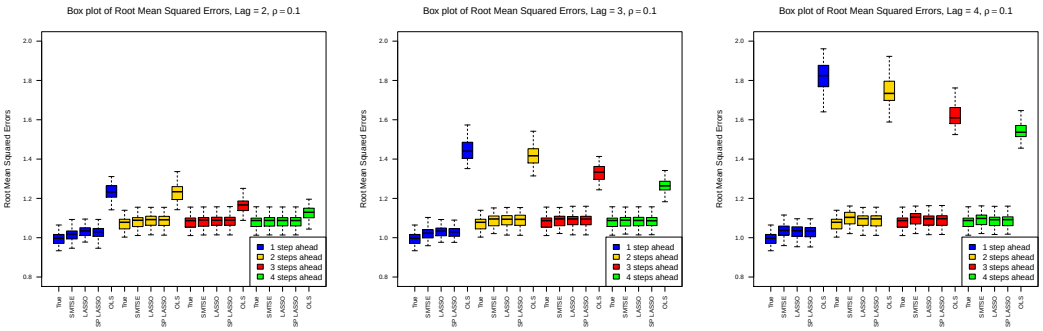
We show additional root mean squared errors (RMSEs) for the simulated models described in section 2.5.1 . For instance, under the simulation setup, we had $\rho = 0.1$ and the number of lags $P = 2$. We presented the results of the prediction performance for simulated models with 2 lags in section 2.5.2. In Figures 27, 28, 29, we show the results when the number of lags used to fit the models is greater than the true number of lags. Under all the simulation settings, the ordinary least squares (OLS) has the worst performance. More importantly as we increase the number of lags, the performance of the OLS worsens because of overfitting. On the other hand, the SMTSE performance is slightly better than the performance of other regularized methods considered in these simulations.



(a) Setting 1 with 3 lags **(b)** Setting 1 with 4 lags **(c)** Setting 2 with 3 lags **(d)** Setting 2 with 4 lags

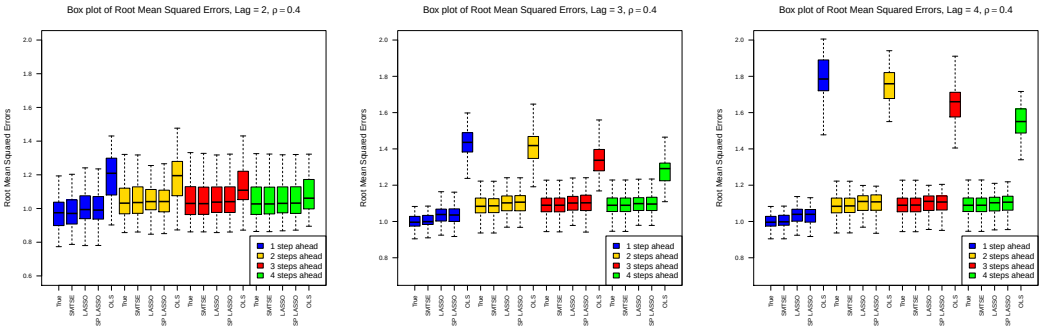
Figure 27: RMSE for one layer simulations under settings 1 and 2.

We also show similar results for the two layers experiments, presented under simulation settings 3 and 4.



(a) Setting 3 with 2 lags **(b)** Setting 3 with 3 lags **(c)** Setting 3 with 4 lags

Figure 28: RMSE for two layer simulations under setting 3.



(a) Setting 4 with 2 lags **(b)** Setting 4 with 3 lags **(c)** Setting 4 with 4 lags

Figure 29: RMSE for two layer simulations under setting 3.

A.2 VAR matrices of case study

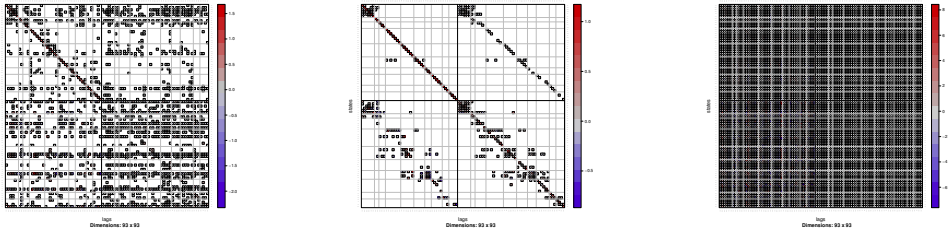
In Figures 30 and 31, we present the plots of the VAR coefficient matrices for the lasso, the spatial lasso and the OLS with 1 and 2 lags.

- The OLS VAR coefficient matrices are dense and lead to models that are not interpretable.

- The lasso selects many non null coefficients with small magnitude and it is unable to provide sparse models. The model resulting from the lasso is not highly interpretable for all the time series. For instance, the lasso is unable to identify the

fact that the employment time series shouldn't affect the number of building permit issued.

• The spatial lasso fitted with one lag yields a model with higher sparsity than the lasso with one lag. The model is also able to capture the spatial effects, since most of the influential coefficients are around the diagonals of the matrices. The spatial lasso with two lags exhibit similar properties and it also captures partially the lack of effect of employment on the number of building permits.



(a) Lasso with 1 lag **(b)** Spatial lasso with 1 lag **(c)** OLS with 1 lag

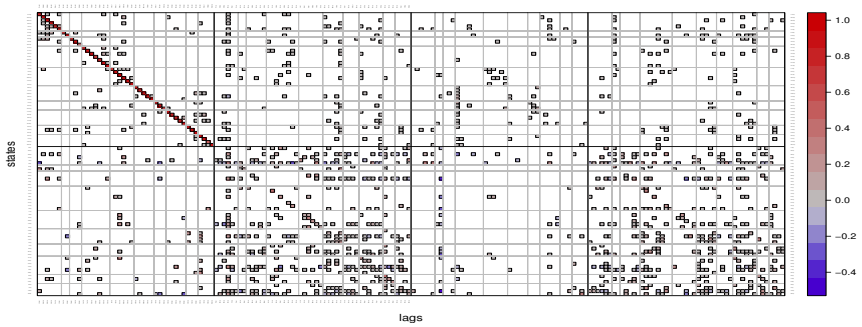
Figure 30: VAR matrix coefficients for employment and building permit time series

A.3 Algorithm for the VAR coefficients

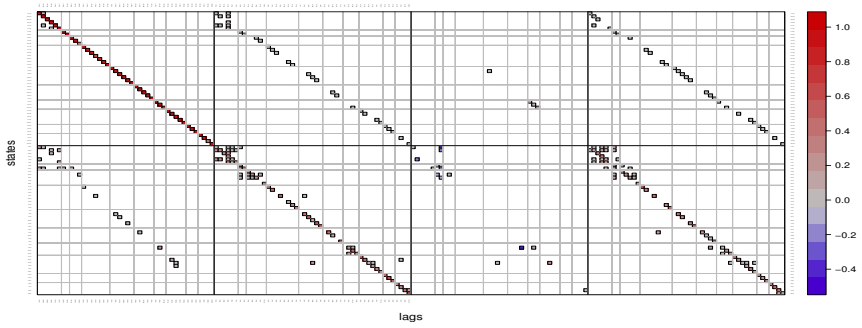
For any layer $J \in \{1, \cdots, L\}$, we define $Y_{D_I} \in \mathbb{R}^{(T-P)\times K_J}$ the set of time series within the J^{th} layer. If we assume that the precision matrix is set at $\tilde{\Omega}_J$, we need to solve a problem below.

$$\begin{aligned} &\min_{B_{D_J.}} \operatorname{Tr}\left[\frac{1}{T-P}\left(Y_{D_J} - \mathbb{X}B_{D_I.}\right)'\left(Y_{D_J} - \mathbb{X}B_{D_I.}\right)\tilde{\Omega}_J\right] + \\ &\lambda_1 \sum_{i=a_J}^{b_J}\sum_{p=1}^{P}\sum_{k=1}^{M} p^{\alpha} e^{\gamma\|s_i - s_k\|}|B_{ik}^{(p)}| + \lambda_2 \sum_{i=a_J}^{b_J}\sum_{\substack{l=1 \\ l:\ i\notin D_l}}^{L} \|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} \end{aligned} \tag{1}$$

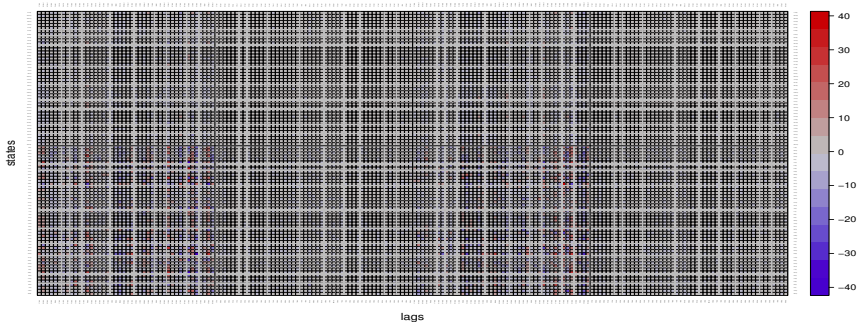
Let $S = \mathbb{X}'\mathbb{X}$ and $H^{[J]} = \mathbb{X}'Y_{D_J}\Omega_J$, For convenience we write $B_{ri} = B_{ij}^{(p)}$ where $(r = (p-1)M + j)$. For each layer we define the sets $A_l = \{a_l, \cdots, b_l, \cdots, (P-1)M + a_l, \cdots, (P-1)M + b_l\}$. For any value $r \in \{1, \cdots, M\}$ we can recover p and j by $p = \lfloor \frac{r}{M} \rfloor + 1$ and $j = (r \mod M) + 1$, If $j \in D_l$ then $r \in A_l$.



(a) Lasso with 2 lags



(b) Spatial lasso with 2 lags



(c) OLS with 2 lags

Figure 31: VAR matrix coefficients for employment and building permit time series

In the following we just write $H^{[J]}$ as H and we assume that for the column of interest $B_{.i.}$ is in the layer J

Algorithm 1:

At the n^{th} iteration, we do the following.

- Set the estimate $\hat{B}^{(n)}_{D_I.} \Leftarrow \hat{B}^{(n-1)}_{D_I.}$

- For each column $B_{i.}$ of B_{D_I}. {

  - For each layer $l \in \{1, \cdots, L\}$ {

    * Test if we only have one layer $(L = 1)$\{

      · For each $r \in A_l$ {

        Compute $U_{ri} = \sum_{q=1}^{PM} \sum_{m=1}^{K_J} S_{rq} \hat{B}_{qm}^{(n)} \tilde{\Omega}_{mi}$

        Update $B_{ri}^{(n)}$ with the minimizer of the function (1)

        along this coordinate direction as in Rothman et. al (2010)

$$\hat{B}_{ri}^{(n)} \Leftarrow \text{sign}\left(\hat{B}_{ri}^{(n)} + \frac{H_{ri} - U_{ri}}{S_{rr}\tilde{\Omega}_{ii}}\right)\left(\left|\hat{B}_{ri}^{(n)} + \frac{H_{ri} - U_{ri}}{S_{rr}\tilde{\Omega}_{ii}}\right| - \frac{\lambda_1 p^{\alpha} e^{\|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}}\right)_+ \quad (2)$$

        with $(x)_+ = \max(0, x)$

      }

    }

    * Else {

      · For each $r \in A_l$ {

        Compute $U_{ri} = \sum_{q=1}^{PM} \sum_{m=1}^{K_J} S_{rq} \hat{B}_{qm}^{(n)} \tilde{\Omega}_{mi}$

        Compute

$$a_{ri} = \left[-U_{ri} + H_{ri} + \sum_{k \in A_l} S_{rk} \hat{B}_{ki}^{(n)} \tilde{\Omega}_{ii}\right]$$

        Compute

$$\hat{t}_r \quad = \quad \begin{cases} \frac{a_{ri}}{\lambda_1 p^{\alpha} e^{\gamma \|s_i - s_j\|}} \text{ if } |\frac{a_{ri}}{\lambda_1 p^{\alpha} e^{\gamma \|s_i - s_j\|}}| \leq 1, \\ \text{sign}\left(\frac{a_{ri}}{\lambda_1 p^{\alpha} e^{\gamma \|s_i - s_j\|}}\right) \text{ if } |\frac{a_{ri}}{\lambda_1 p^{\alpha} e^{\gamma \|s_i - s_j\|}}| > 1 \end{cases}$$

      }

      · Compute $Q(\hat{t})$

$$Q(\hat{t}) = \frac{1}{\lambda_2^2 K_l} \sum_{r \in A_l} \left(\frac{a_{ri}}{p^{\alpha} e^{\gamma \|s_i - s_j\|}} - \lambda_1 \hat{t}_r\right)^2$$

· If $Q(\hat{t}) \leq 1$ set $\hat{B}_{iD_l}^{(n)} \Leftarrow \mathbf{0}$

· Else if $(Q > 1)$ {

For each $r \in A_l$ {

Compute $C_{ri} = H_{ri} - U_{ri} + S_{rr}B_{ri}^{(n)}\tilde{\Omega}_{ii}$

If $|C_{ri}| \leq \lambda_1 p^{\alpha} e^{\|s_i - s_j\|}$ set $\hat{B}_{ri}^{(n)} \Leftarrow 0$
Else If $|C_{ri}| > \lambda_1 p^{\alpha} e^{\|s_i - s_j\|}$ {

$$\hat{B}_{ri}^{(n)} \Leftarrow \underset{Z \in \mathbb{R}}{\operatorname{argmin}} \Theta_1 * Z^2 + \Theta_2 * Z + \lambda_1 \Theta_3 * |Z| + \lambda_2 \sqrt{K_l} \left(\Theta_4 + \Theta_5 * Z^2\right)^{\frac{1}{2}}$$

Where

$$\Theta_1 = S_{rr}\tilde{\Omega}_{ii}$$

$$\Theta_2 = 2 * \left(U_{ri} - H_{ri} - S_{rr}B_{ri}\tilde{\Omega}_{ii}\right) = -2C_{ri}$$

$\Theta_3 = p^{\alpha} e^{\gamma \|s_i - s_j\|}$ where $p = \lfloor \frac{r}{M} \rfloor + 1$, p is the current lag

$$\Theta_4 = \left(\|B_{iD_l}\|^2_{\tilde{\Delta}_l^{[i]}} - p^{2\alpha} e^{2\gamma \|s_i - s_j\|} (B_{ri})^2\right)$$

$$\Theta_5 = p^{2\alpha} e^{2\gamma \|s_i - s_j\|}$$

}

}

}

}

}

- If $\sum_{j,k} \left|\hat{B}_{jk}^{(n)} - \hat{B}_{jk}^{(n-1)}\right| < \epsilon \sum_{j,k} \left|\hat{B}_{jk}^{Ridge}\right|$ stop, otherwise start new loop

A.4 Derivation of the algorithm for the one layer model

For a one layer model with a fixed covariance matrix $\tilde{\Omega}$, we try to minimize the objective function $f(\mathbf{B}, \tilde{\Omega})$.

$$f(\mathbb{B}, \tilde{\Omega}) = Tr\left[(Y - X\mathbb{B})^T (Y - X\mathbb{B})\tilde{\Omega}\right] + \lambda_1 (T - P) \sum_{i=1}^{M} \sum_{p=1}^{P} \sum_{k=1}^{M} p^{\alpha} e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}|$$

Let $S = X^T X$ and $H = X^T Y \tilde{\Omega}$. We also use $B_{ij}^{(p)} = B_{ri}$ with $j = (r \mod M) + 1$ and $p = \lfloor \frac{r}{M} \rfloor + 1$. We also define $U_{ri} = \sum_{q=1}^{PM} \sum_{m=1}^{M} S_{rq} B_{qm} \tilde{\Omega}_{mi}$.

Since we have an l_1-norm we need to derive the directional derivatives associated with each parameter in the VAR model.

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^{+}} = U_{ri} - H_{ri} + \lambda_1 * (T - P) * p^{\alpha} * e^{-\gamma \|s_i - s_j\|} \text{ if } B_{ri} > 0$$

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^{-}} = U_{ri} - H_{ri} - \lambda_1 * (T - P) * p^{\alpha} * e^{-\gamma \|s_i - s_j\|} \text{ if } B_{ri} < 0$$

If we define the current estimate as $B_{ri}^{(n)}$. $B_{ri}^{(n+1)}$ is a optimal if and only if, the directional derivatives $\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}} \geq 0$ in all directions. To update the directional derivatives we can use the formula below.

If the solution is such that $B_{ri} < 0$

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^{+}} = S_{rr} B_{ri}^{(n+1)} \tilde{\Omega}_{ii} - S_{rr} B_{ri}^{(n)} \tilde{\Omega}_{ii} + U_{ri} - H_{ri} - \lambda_1 * (T - P) * p^{\alpha} * e^{-\gamma \|s_i - s_j\|}$$

$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^{+}} > 0$ if and only if $B_{ri}^{(n+1)} > \hat{B}_{ri}^{\star} + \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}}$

where $\hat{B}_{ri}^{\star} = B_{ri}^{(n)} + \frac{H_{ri} - U_{ri}}{S_{rr} \tilde{\Omega}_{ii}}$

If $\hat{B}_{ri}^{\star} + \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} < 0$,we update the B_{ri} by using :

$\hat{B}_{ri}^{(n+1)} = \hat{B}_{ri}^{\star} + \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}}$.

If $\hat{B}^{\star}_{ri} + \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}} \geq 0$ the only solution that satisfies the constraint $B_{ri} < 0$ is $\hat{B}^{(n+1)}_{ri} = 0$.

So we have that

$$\begin{aligned}
\hat{B}^{(n+1)}_{ri} &= \min\left[0, \hat{B}^{\star}_{ri} + \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}}\right] \\
&= \text{Sign}\left(\hat{B}^{\star}_{ri}\right) \max\left[0, |\hat{B}^{\star}_{ri}| - \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}}\right]
\end{aligned}$$

If the solution is such that $B_{ri} < 0$ we use a similar reasoning to show that

$$\begin{aligned}
\hat{B}^{(n+1)}_{ri} &= \min\left[0, \hat{B}^{\star}_{ri} - \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}}\right] \\
&= \text{Sign}\left(\hat{B}^{\star}_{ri}\right) \max\left[0, |\hat{B}^{\star}_{ri}| - \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}}\right]
\end{aligned}$$

So we showed that to update each coefficient in the VAR matrix, we can simply use cyclical coordinate descent $\hat{B}^{(n+1)}_{ri} = \text{Sign}\left(\hat{B}^{\star}_{ri}\right)\left(|\hat{B}^{\star}_{ri}| - \frac{\lambda_1 * (T-P) * p^{\alpha} * e^{\gamma \|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}}\right)_+$

A.5 Derivation of the algorithm for multiple layers model

In a multi-layer setting, with a block diagonal error covariance matrix. If we focus on layer $J \in \{1, \cdots, L\}$, we want to solve the following problem.

$$\min_{B_{D_J.}} f\left(B_{D_J.}, \tilde{\Omega}_J\right) = \operatorname{Tr}\left[\left(Y_{D_J} - \mathbb{X}B_{D_J.}\right)'\left(Y_{D_J} - \mathbb{X}B_{D_J.}\right)\tilde{\Omega}_J\right] +$$
$$\lambda_1 (T-P) \sum_{i=a_J}^{b_J} \sum_{p=1}^{P} \sum_{k=1}^{M} p^{\alpha} e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}| + \lambda_2 (T-P) \sqrt{(K_J)} \sum_{i=a_J}^{b_J} \sum_{\substack{l=1 \\ l:\ i \notin D_l}}^{L} \|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}$$

If we focus on layer $l \in \{1, \cdots, L\}$ and on column i then The subgradient equations are given by:

$$\frac{\partial f\left(B_{D_J.}, \tilde{\Omega}_J\right)}{\partial B_{ri}} = U_{ri} - H_{ri} + \lambda_1 (T-P)\, p^{\alpha} e^{\gamma \|s_i - s_j\|} t_r + \lambda_2 (T-P)\, (K_J)^{\frac{1}{2}} g_r$$

As in supplement A, each equation above corresponds to a VAR coefficient $B_{ij}^{(p)} = B_{ri}$. Where r is such that $j = (r \bmod M) + 1$, $p = \lfloor \frac{r}{M} \rfloor + 1$ and $j \in (a_l, \cdots, b_l)$ with a_l and b_l are defined in subsection 2.3 of the manuscript .
We also have that $r \in A_l = \{a_l, \cdots, b_l, \cdots, (P-1)M + a_l, \cdots, (P-1)M + b_l\}$ $\mathbf{g}$ is the vector that contains all the (g_r) values.

$g_r = \frac{p^{2\alpha} e^{2\gamma \|s_i - s_j\|} B_{ij}^{(p)}}{\|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}}$ if $\|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} \neq 0$
$\|\mathbf{g} W^{-\frac{1}{2}}\| \leq 1$ if $\|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} = 0$
where $W = I_P \otimes \begin{bmatrix} e^{2\gamma \|s_i - s_1^{[l]}\|} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{2\gamma \|s_i - s_{K_l}^{[l]}\|} \end{bmatrix}$ and

$$W^{-\frac{1}{2}} = I_P \otimes \begin{bmatrix} \frac{1}{e^{\gamma \|s_i - s_1^{[l]}\|}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{e^{\gamma \|s_i - s_{K_l}^{[l]}\|}} \end{bmatrix}$$

Additionally, we have that:

$t_r = \text{sign}\,(B_{ri})$ if $B_{ri} \neq 0$

$t_r \in [-1, 1]$ if $B_{ri} = 0$

A necessary and sufficient conditions for $B_{.iD_l} = 0$ is that $\forall r \in A_l$:

$$U_{ri} - H_{ri} - \sum_{k \in A_l} S_{rk} B_{ki} \tilde{\Omega}_{ii} + \lambda_1 (T - P)\, p^{\alpha} e^{\gamma \|s_i - s_j\|} t_r + \lambda_2 (T - P)\, (K_J)^{\frac{1}{2}} g_r = 0$$

has a solution with $\|\mathbf{g} W^{-\frac{1}{2}}\| \leq 1$ and $t_r \in [-1, 1]$.

We define $a_{ri} = [-U_{ri} + H_{ri} + \sum_{k \in A_l} S_{rk} B_{ki} \tilde{\Omega}_{ii}]$, this implies that if the necessary and sufficient condition is satisfied.

$$\forall r \in A_l,\, a_{ri} = \lambda_1 (T - P)\, p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|} t_r + \lambda_2 (T - P)\, p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2} (K_J)^{\frac{1}{2}} g_r$$

We can solve the system of equations resulting from the necessary and sufficient conditions by minimizing.

$$Q(\mathbf{t}) = \|\mathbf{g} W\|_2 = \frac{1}{\lambda_2^2 K_l} \sum_{j=1}^{K_l} \sum_{p=1}^{P} \left[\frac{a_{ri}}{(T - P)\, p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}} - \lambda_1 t_r \right]^2$$

.

We note that the minimum is $J(t) = 0$, it implies that:

$$\frac{\partial Q(t)}{\partial t_r} = -2 \frac{\lambda_1}{\lambda_2^2 K_l} \sum_{j=1}^{K_l} \sum_{p=1}^{P} \left[\frac{a_{ri}}{(T - P)\, p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}} - \lambda_1 t_r \right]$$

$$\frac{\partial Q(t)}{\partial t_r} = 0 \text{ if } t_r = \frac{a_{ri}}{\lambda_1 (T - P)\, p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}}$$

But we need to have $t_r \in [-1, 1]$. So if $\left| \frac{a_{ri}}{\lambda_1 (T-P) p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}} \right| < 1$ then $t_r = \frac{a_{ri}}{\lambda_1 (T-P) p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}}$.

On the other end,

if $\frac{a_{ri}}{\lambda_1(T-P)p^\alpha e^{\gamma\|s_i-s_j^{[l]}\|_2}} > 1$, then $t_r = 1$, and

if $\frac{a_{ri}}{\lambda_1(T-P)p^\alpha e^{\gamma\|s_i-s_j^{[l]}\|_2}} < -1$ then $t_r = -1$. So we have that:

$$\hat{t}_r = \begin{cases} \frac{a_{ri}}{\lambda_1(T-P)p^\alpha e^{\gamma\|s_i-s_j\|_2}} \text{ if } |\frac{a_{ri}}{\lambda_1(T-P)p^\alpha e^{\gamma\|s_i-s_j\|_2}}| \leq 1, \\ \text{sign}\left(\frac{a_{ri}}{\lambda_1(T-P)p^\alpha e^{\gamma\|s_i-s_j\|_2}}\right) \text{ if } |\frac{a_{ri}}{\lambda_1(T-P)p^\alpha e^{\gamma\|s_i-s_j\|_2}}| > 1 \end{cases}$$

Now that $\min J(t)$ is found at $\hat{t}$, we compute $Q\left(\hat{t}\right)$. If $Q\left(\hat{t}\right) < 1$ then the necessary and sufficient conditions stated above is satisfied with $Q\left(\hat{t}\right) = \|\mathbf{g}W\|_2 < 1$, this implies that $B_{iD_l} = 0$.

If $Q\left(\hat{t}\right) > 1$, then some coefficients $B_{ri} \neq 0$ in B_{iD_l}. We have to solve for each coefficients B_{ri} with $r \in A_l$.

If $B_{ri} \geq 0$

$$\frac{\partial f\left(B,\tilde{\Omega}_l\right)}{\partial B_{ri}} = U_{ri} - H_{ri} + \lambda_1 (T-P) p^\alpha e^{\gamma\|s_i-s_j^{[l]}\|_2} + \lambda_2 (T-P) p^{2\alpha} e^{2\gamma\|s_i-s_j^{[l]}\|_2} K_l \frac{B_{ri}}{\|B_{iD_l}\|_2}$$

To update the partial derivative with respect to the B_{ri} coordinate, we use the formula below.

$$\frac{\partial f\left(B,\tilde{\Omega}_l\right)}{\partial B_{ri}} = U_{ri} + S_{rr}\hat{B}_{ri}^{(n+1)}\tilde{\Omega}_{ii} - S_{rr}\hat{B}_{ri}^{(n)}\tilde{\Omega}_{ii} - H_{ri} + \lambda_1 (T-P) p^\alpha e^{\gamma\|s_i-s_j^{[l]}\|_2} + \frac{\lambda_2 (T-P) p^{2\alpha} e^{2\gamma\|s_i-s_j^{[l]}\|_2} K_l}{\|B_{.iD_l}\|_2} \hat{B}_{ri}^{(n+1)}$$

$C_{ri} = H_{ri} - U_{ri} - S_{rr}\hat{B}_{ri}^{(n)}\tilde{\Omega}_{ii}$

$\hat{B}_{ri}^{(n)} = 0$ and $\frac{\partial f\left(B,\tilde{\Omega}_l\right)}{\partial B_{ri}} > 0$ if and only if $C_{ri} < \lambda_1 (T-P) p^\alpha e^{\gamma\|s_i-s_j^{[l]}\|_2}$.

The same reasoning can be applied to the case where $B_{ri} \leq 0$.

This leads to $\hat{B}_{ri}^{(n)} = 0$ and $\frac{\partial f\left(B,\tilde{\Omega}_l\right)}{\partial B_{ri}} > 0$ if and only if $C_{ri} > -\lambda_1 (T-P) p^\alpha e^{\gamma\|s_i-s_j^{[l]}\|_2}$.

So if $|C_{ri}| < \lambda_1 (T-P) p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}$ we set $\hat{B}_{ri}^{(n+1)} = 0$.

if $|C_{ri}| > \lambda_1 (T-P) p^{\alpha} e^{\gamma \|s_i - s_j^{[l]}\|_2}$, we need to minimize the one dimensional function given by.

APPENDIX B

SUPPLEMENT TO HIGH DIMENSIONAL ADDITIVE MODELS

B.1 Connection to Group Lasso

As suggested by Meier et al. (2009), each function $f_j^{(k)}$ can be expressed in cubic B-spline basis with a reasonable number of basis functions.

$$f_j^{(k)}(x) = \sum_{q=1}^{Q} \beta_{j,q}^{(k)} b_{j,q}^{(k)}(x)$$

where $j \in \{1, \cdots, p\}$, $k \in \{1, \cdots, K\}$ and Q is the number of basis functions. $b_{j,q}^{(k)} : \mathbb{R} \leftarrow \mathbb{R}$ is a B-spline basis function used for the j^{th} additive function and the k^{th} response. And where $\boldsymbol{\beta}_j^{(k)} = \left(\beta_{j,1}^{(k)}, \cdots, \beta_{j,Q}^{(k)}\right)$ is a vector of coefficients that uniquely represents $f_j^{(k)}$. We also define the design matrix $\boldsymbol{B}_j^{(k)}$ as the $n \times Q$ design matrix of B-spline basis of the j^{th} predictor for the k^{th} response. The (i, q) entry in the design matrix is given by $B_{j,iq}^{(k)} = b_{j,q}^{(k)}(x_{ij}^{(k)})$, with $i \in \{1, \cdots, n\}$. For each task k, we define the matrix $\boldsymbol{B}^{(k)} = \left[\boldsymbol{B}_1^{(k)}, \cdots, \boldsymbol{B}_p^{(k)}\right]$. The vector of all the responses is given by $\tilde{\boldsymbol{Y}} = \mathbf{vec}(\mathbf{Y})$ and the matrix $\mathbf{B} = \begin{bmatrix} \mathbf{B}^{(1)} & & \\ & \ddots & \\ & & \mathbf{B}^{(k)} \end{bmatrix}$ of dimension nK $\times$ KQP. So the optimization problem is formulated as:

$$\|\tilde{\boldsymbol{Y}} - \mathbf{B}\boldsymbol{\beta}\|_2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j^T \mathbf{M}_j \boldsymbol{\beta}_j} \tag{29}$$

Where $\boldsymbol{\beta}_j = \left(\boldsymbol{\beta}_j^{(1)} \cdots, \boldsymbol{\beta}_j^{(K)}\right) \in \mathbb{R}^{KQ}$ and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{(1)}, \cdots, \boldsymbol{\beta}^{(K)}\right)^T \in \mathbb{R}^{pKQ}$. For any task k, the vector of coefficients associated with the additive functions $\mathbf{f}^{(k)}$ is $\boldsymbol{\beta}^{(k)} =$

$\left(\beta_1^{(k)}, \cdots, \beta_p^{(k)}\right)$. The matrix $\mathbf{M}_j = \begin{bmatrix} \mathbf{M}_j^{(1)} & & \\ & \ddots & \\ & & \mathbf{M}_j^{(K)} \end{bmatrix}$ is used to impose group sparsity on the coefficients associated with functions $\mathbf{f}_j$ and to ensure the smoothness of these functions. Each matrix $\mathbf{M}_j^{(k)} = \mathbf{B}_j^{(k)^T}\mathbf{B}_j^{(k)} + \lambda_2\mathbf{\Theta}_j^{(k)}$, where the Q × Q matrix $\mathbf{\Theta}_j^{(k)}$ contains the inner products of the second derivative of the B-spline basis functions

$$\Theta_{j,mn}^{(k)} = \int b_{j,n}^{(k)''}(x) b_{j,n}^{(k)''}(x) dx$$

with $m, n \in \{1, \cdots, Q\}$. The matrix $\mathbf{M}_j^{(k)}$ can be decomposed to obtain the matrix $\mathbf{\Delta}_j^{(k)}\ \mathbf{M}_j^{(k)} = \mathbf{\Delta}_j^{(k)^T}\mathbf{\Delta}_j^{(k)}$. The coefficients $\boldsymbol{\beta}_j$ can then be transformed to $\tilde{\boldsymbol{\beta}}_j = \mathbf{\Delta}_j\boldsymbol{\beta}_j$, where $\mathbf{\Delta}_j = \begin{bmatrix} \mathbf{\Delta}_j^{(1)} & & \\ & \ddots & \\ & & \mathbf{\Delta}_j^{(K)} \end{bmatrix}$. We also define $\mathbf{\Delta}^{(k)} = \begin{bmatrix} \mathbf{\Delta}_1^{(k)} & & \\ & \ddots & \\ & & \mathbf{\Delta}_p^{(k)} \end{bmatrix}$ and $\mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}^{(1)} & & \\ & \ddots & \\ & & \mathbf{\Delta}^{(K)} \end{bmatrix}$. Using the matrices above, we have that $\tilde{\boldsymbol{\beta}} = \mathbf{\Delta}\boldsymbol{\beta}$ and $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Delta}^{-1}$. The optimization problem in (29) can the be reformulated as a group lasso.

$$\|\tilde{\boldsymbol{Y}} - \tilde{\mathbf{B}}\tilde{\boldsymbol{\beta}}\|_2 + \lambda_1 \sum_{j=1}^{p} \|\tilde{\boldsymbol{\beta}}_j\|_2 \tag{30}$$

B.2 Proof of Theorem 1

The concept in this proof are inspired from the method used by Yin et al. (2012). We consider the loss function associated with the functionals $\mathbf{f}_j$,

$$L(\mathbf{f}_j) = \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K}\left(R_j^{(k)} - f_j^{(k)}(X_j^{(k)})\right)^2\right]$$

We also define a perfubation $\boldsymbol{\mu}_j = \left(\mu_j^{(1)}, \cdots, \mu_j^{(K)}\right)$ where $\forall k = (1, \cdots, K)$, $\mu_j^{(k)} \in \mathcal{H}_j$

$$L(\mathbf{f}_j + \varepsilon\boldsymbol{\mu}_j) = \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K}\left(R_j^{(k)} - f_j^{(k)}(X_j^{(k)}) - \varepsilon\mu_j^{(k)}(X_j^{(k)})\right)^2\right]$$

The approximation of the first order approximation is given by:

$$\begin{aligned} L(\mathbf{f}_j + \varepsilon\boldsymbol{\mu}_j) - L(\mathbf{f}_j) &\approx \varepsilon\mathbb{E}\left[\sum_{k=1}^{K}\mu_j^{(k)}(X_j^{(k)})\left(f_j^{(k)}(X_j^{(k)}) - R_j^{(k)}\right)\right] \\ &= \sum_{k=1}^{K}\mathbb{E}\left[\mu_j^{(k)}(X_j^{(k)})\left(f_j^{(k)}(X_j^{(k)}) - R_j^{(k)}\right)\right] \\ &= \varepsilon\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\mu_j^{(k)}(X_j^{(k)})\left(f_j^{(k)}(X_j^{(k)}) - R_j^{(k)}\right)|X_j^{(k)}\right]\right] \\ &= \varepsilon\sum_{k=1}^{K}\mathbb{E}\left[\left(f_j^{(k)}(X_j^{(k)}) - P_j^{(k)}R_j^{(k)}\right)\mu_j^{(k)}(X_j^{(k)})\right] \\ &= \varepsilon\sum_{k=1}^{K}\left\langle\mu_j^{(k)}(X_j^{(k)}), \left(f_j^{(k)}(X_j^{(k)}) - P_j^{(k)}R_j^{(k)}\right)\right\rangle \end{aligned}$$

The gradient of $L(\mathbf{f}_j)$ is then:

$$\bigtriangledown L(\mathbf{f}_j) = \left[f_j^{(k)} - P_j^{(k)}R_j^{(k)}\right]_{j=1,\cdots,p}$$

This leads to the stationary condition presented in theorem 1

$$f_j^{(k)} - P_j^{(k)}R_j^{(k)} + \lambda\sqrt{K}u_j^{(k)} = 0$$

B.3 Proof of Theorem 2

Proof. We will first show that the condition above is necessary.

if $\forall k \in \{1,\cdots,K\}\ f_j^{(k)} = 0$, then the stationary condition (3) becomes $P_j^{(k)}R_j^{(k)} = \lambda\sqrt{K}u_j^{(k)}$

$$\begin{aligned} \sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\left(P_j^{(k)}R_j^{(k)}\right)^2\right]} &= \lambda\sqrt{K}\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[(u_j^{(k)})^2\right]} \\ &= \lambda\sqrt{K}\|\mathbf{e}_j\|_2 \\ &\leq \lambda\sqrt{K} \end{aligned}$$

We now prove that the condition is sufficient. if $\exists k \in \{1,\cdots,K\}$ such that $f_j^{(k)} \neq 0$ the stationary condition (3) becomes

$$f_j^{(k)} - P_j^{(k)}R_j^{(k)} + \lambda\sqrt{K}\frac{f_j^{(k)}}{\|\mathbf{f}_j\|} = 0,\ \forall k \in \{1,\cdots,K\}$$

We define the vector of conditional expectation operators $\mathbf{H}_j$ and the vector of partial residuals $\mathbf{R}_j$

$$\mathbf{H}_j = \begin{bmatrix} P_j^{(1)} \\ \vdots \\ P_j^{(K)} \end{bmatrix} \text{ and } \mathbf{R}_j = \begin{bmatrix} R_j^{(1)} \\ \vdots \\ R_j^{(K)} \end{bmatrix}, \mathbf{I}_K \text{ is an identity matrix of size K}$$

So for covariates $\mathbf{X}_j = \left(X_j^{(1)}, \cdots, X_j^{(K)}\right)$, we have

$$\mathbf{H}_j \mathbf{R}_j = \mathbf{f}_j + \lambda\sqrt{K}\frac{\mathbf{f}_j}{\|\mathbf{f}_j\|}$$

So this implies that

$$\begin{aligned} \|\mathbf{H}_j\mathbf{R}_j\|^2 &= \sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)}R_j^{(k)}\right)^2\right] \\ &= \left\|\left(\mathbf{I}_K + \frac{\lambda\sqrt{K}}{\|\mathbf{f}_j\|}\right)\mathbf{f}_j\right\|^2 \\ &= \|\mathbf{f}_j\|^2 + \lambda^2 K + \lambda\sqrt{K}\|\mathbf{f}_j\| \end{aligned}$$

since $\exists k$, such that $f_j^{(k)} \neq 0$ we know that $\|\mathbf{f}_j\| > 0$, so this implies that

$$\sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)}R_j^{(k)}\right)^2\right] \geq \lambda\sqrt{K}$$

We now derive some steps that will be needed to update the estimated additive components $f_j^{(k)}$. For each set of covariates $\mathbf{X}_j$, we define w_j
such that $w_j^2 = \sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)}R_j^{(k)}\right)^2\right]$, which implies that $w_j = \|\mathbf{f}_j\| + \lambda\sqrt{K}$.
If $\exists k \in \{1, \cdots, K\}$ such that $f_j^{(k)} \neq 0$ then

$$\begin{aligned} f_j^{(k)} &= \frac{\|\mathbf{f}_j\|}{\|\mathbf{f}_j\| + \lambda\sqrt{K}} P_j^{(k)}R_j^{(k)} \\ &= \left[1 - \frac{\lambda\sqrt{K}}{w_j}\right] P_j^{(k)}R_j^{(k)} \end{aligned}$$

if $\forall k \in \{1, \cdots, K\}$, $f_j^{(k)} = 0$ then $w_j = \lambda\sqrt{K}$ So the additive components are updated as follows

$$f_j^{(k)} = \left[1 - \frac{\lambda\sqrt{K}}{w_j}\right]_+ P_j^{(k)}R_j^{(k)} \tag{31}$$

B.4 Proof of Theorem 3

The stationary condition for the problem introduced in (6)

$$f_j^{(k)} - P_j^{(k)} R_j^{(k)} + (1-\alpha)\lambda\sqrt{K} u_j^{(k)} + \alpha\lambda v_j^{(k)} = 0$$

Where

$$\mathbf{u}_j = \begin{cases} \frac{f_j^{(k)}}{\|\mathbf{f}_j\|} & \text{if } \|\mathbf{f}_j\| \neq 0 \text{ for } k \in \{1, \cdots, K\} \\ \mathbf{e}_j \in \mathcal{H}_j^K & \text{with} \|\mathbf{e}_j\|_2 \leq 1 \text{ when } \|\mathbf{f}_j\| = 0 \end{cases}$$

and

$$v_j^{(k)} = \begin{cases} \frac{f_j^{(k)}}{\|f_j^{(k)}\|} & \text{if } \|f_j^{(k)}\| \neq 0 \text{ for } k \in \{1, \cdots, K\} \\ o_j^{(k)} \in \mathcal{H}_j & \text{with} \|o_j^{(k)}\| \leq 1 \text{ if } \|f_j^{(k)}\| = 0 \end{cases}$$

So for each $k \in \{1, \cdots, K\}$

$$u_j^{(k)} = \frac{1}{(1-\alpha)\lambda\sqrt{K}} \left[P_j^{(k)} R_j^{(k)} - \alpha\lambda v_j^{(k)} \right]$$

If $\mathbf{f}_j = \mathbf{0}$ then $\|\mathbf{e}_j\| \leq 1$

$$J(\mathbf{v}_j) = \|\mathbf{u}_j\|^2 = \frac{1}{(1-\alpha)^2\lambda^2 K} \sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)} R_j^{(k)} - \alpha\lambda v_j^{(k)} \right)^2 \right]$$

Solving the systems of equation given by the stationary conditions is equivalent to minimizing the norm of the set of functions $\|\mathbf{u}_j\|$. If we take the functional derivative of $\|\mathbf{u}_j\|^2$ with respect to $\mathbf{v}_j$, we get:

$$\partial J(\mathbf{v}_j) \propto \frac{1}{(1-\alpha)^2\lambda^2} \sum_{k=1}^{K} \mathbb{E}\left[\left(P_j^{(k)} R_j^{(k)} - \alpha\lambda v_j^{(k)} \right) \eta_j^{(k)} \right]$$

where $\partial v_j^{(k)} = \eta_j^{(k)}$

if $f_j^{(k)} = 0$ then $\|v_j^{(k)}\| \leq 1$, if $\|P_j^{(k)} R_j^{(k)}\| \leq \alpha\lambda$ then the minimum of $J(\mathbf{v}_j)$ is reached for $v_j^{(k)} = \frac{P_j^{(k)} R_j^{(k)}}{\alpha\lambda}$

If $\|P_j^{(k)} R_j^{(k)}\| \geq \alpha\lambda$ then the minimum is reached for $v_j^{(k)} = \frac{P_j^{(k)} R_j^{(k)}}{\|P_j^{(k)} R_j^{(k)}\|}$ since $\|v_j^{(k)}\|$ has to be less or equal to 1.

So we obtain that:

$$v_j^{(k)} = \begin{cases} \frac{P_j^{(k)} R_j^{(k)}}{\alpha\lambda} & \text{if } \|P_j^{(k)} R_j^{(k)}\| \leq \alpha\lambda \text{ for } k \in \{1, \cdots, K\} \\ \frac{P_j^{(k)} R_j^{(k)}}{\|P_j^{(k)} R_j^{(k)}\|} & \text{if } \|P_j^{(k)} R_j^{(k)}\| \geq \alpha\lambda \end{cases}$$

If we plug $v_j^{(k)}$ in the expression of $u_j^{(k)}$, we find

$$u_j^{(k)} = \frac{1}{(1-\alpha)\lambda\sqrt{K}} \left[1 - \frac{\alpha\lambda}{\|P_j^{(k)} R_j^{(k)}\|}\right]_+ P_j^{(k)} R_j^{(k)}$$

We have that $\mathbf{f}_j = 0$ if and only if

$$\begin{aligned} \|\mathbf{u}_j\| &\leq 1 \\ \sqrt{\sum_{k=1}^{K} \mathbb{E}\left[\left[\left(1 - \frac{\alpha\lambda}{\|P_j^{(k)} R_j^{(k)}\|}\right)_+ P_j^{(k)} R_j^{(k)}\right]^2\right]} &\leq (1-\alpha)\lambda\sqrt{K} \end{aligned}$$

B.5 Proof of Theorem 4

If $\mathbf{f}_j \neq \mathbf{0}$ then $\exists k \in \{1, \cdots, K\}$ such that $f_j^{(k)} \neq 0$ and the stationary condition associated with the functions $\mathbf{f}_j$ are

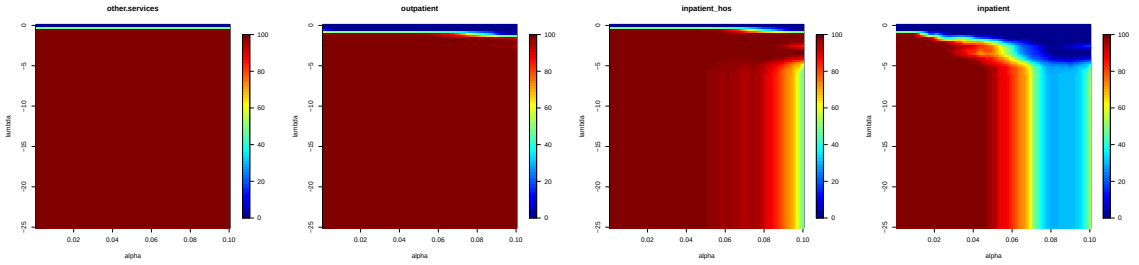
$$P_j^{(k)} R_j^{(k)} + (1-\alpha)\lambda\sqrt{K} \frac{f_j^{(k)}}{\|\mathbf{f}_j\|} + \alpha\lambda v_j^{(k)} = 0$$

if $f_j^{(k)} = 0$ then $\|P_j^{(k)} R_j^{(k)}\| \leq \alpha\lambda$

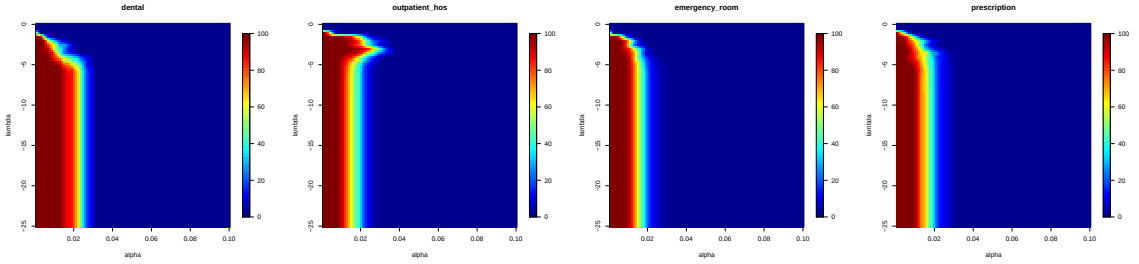
Now we prove that if $\|P_j^{(k)} R_j^{(k)}\| \leq \alpha\lambda$ then $f_j^{(k)} = 0$.

If $\exists k \in 1, \cdots, K$ such that $f_j^{(k)} \neq 0$ then

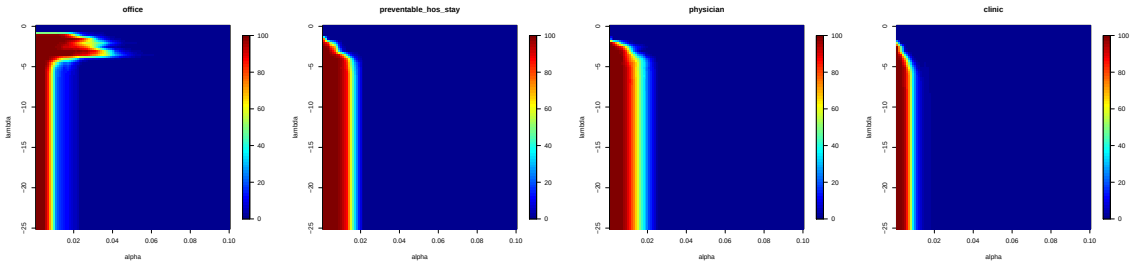
$$\begin{aligned} P_j^{(k)} R_j^{(k)} &= f_j^{(k)} + \lambda(1-\alpha)\sqrt{K} \frac{f_j^{(k)}}{\|\mathbf{f}_j\|} + \alpha\lambda \frac{f_j^{(k)}}{\|f_j^{(k)}\|} \\ \|P_j^{(k)} R_j^{(k)}\|^2 &= \left(1 + \frac{\alpha\lambda}{\|f_j^{(k)}\|} + \frac{(1-\alpha)\lambda\sqrt{K}}{\|\mathbf{f}_j\|}\right)^2 \|f_j^{(k)}\|^2 \\ \|P_j^{(k)} R_j^{(k)}\| &\geq \alpha\lambda \end{aligned}$$



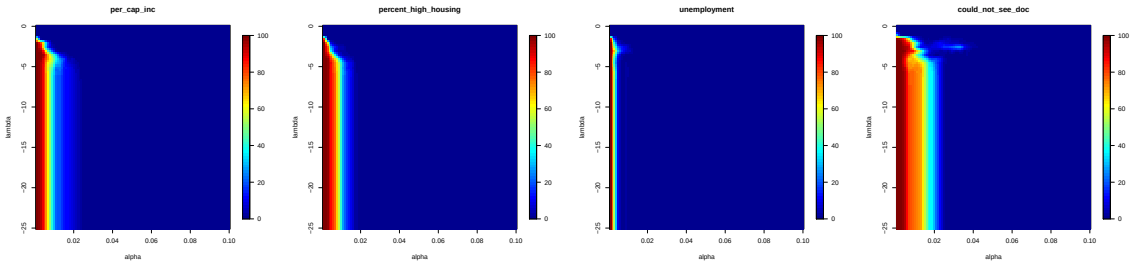
(a) Other services **(b)** outpatient services **(c)** Inpatient hospital **(d)** Inpatient services



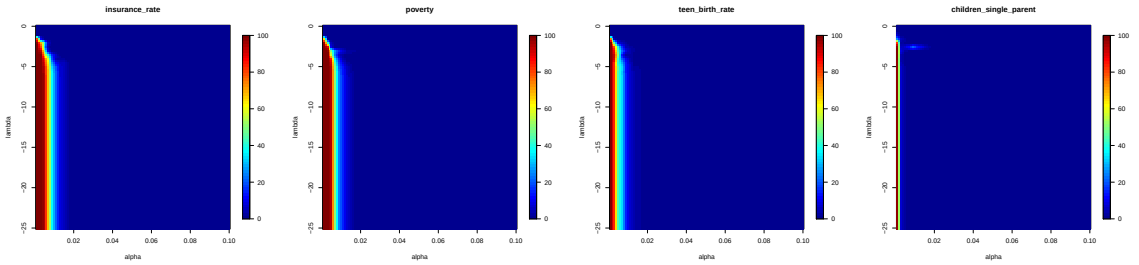
(e) Dental services **(f)** outpatient hospital **(g)** Emergency room **(h)** Prescriptions



(i) Office **(j)** Prescriptions **(k)** Physician **(l)** Clinic

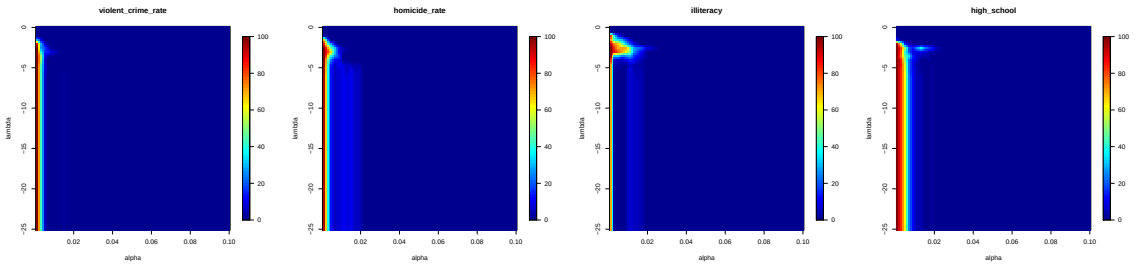


(m) Per Capita Income **(n)** Pct High Housing **(o)** Unemployment **(p)** couldn't see doctor

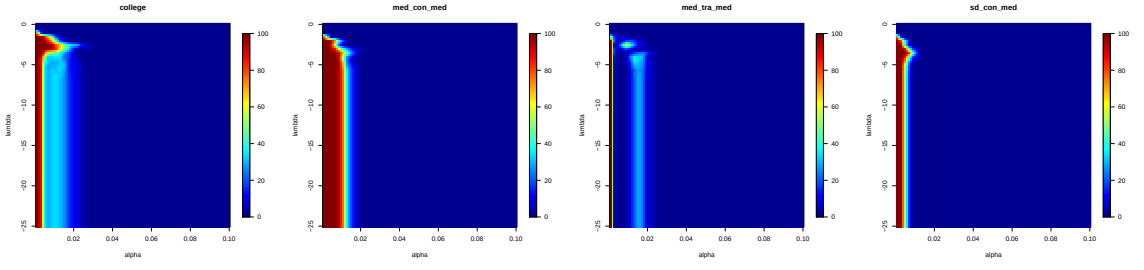


(q) Insurance rates **(r)** Poverty rates **(s)** Teen birth rate **(t)** single parent

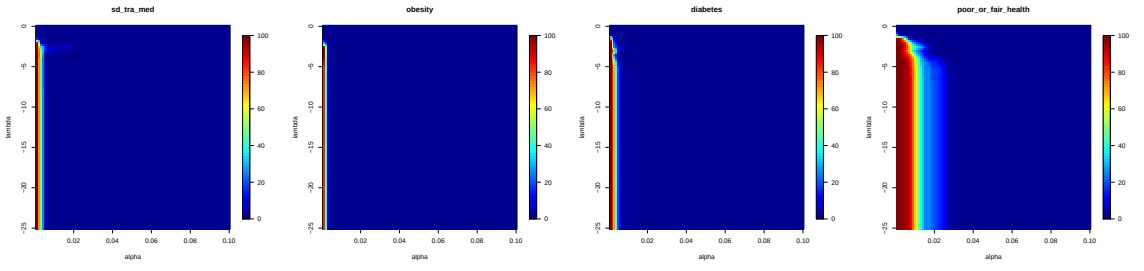
Figure 32: Stability Selection Plots for the Predictors Used in Medicaid Cost Analysis



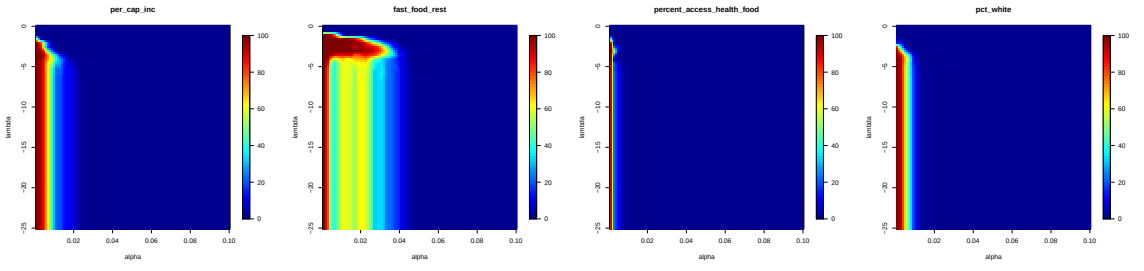
(a) Violent crime **(b)** Homicide rate **(c)** Illiteracy **(d)** high school



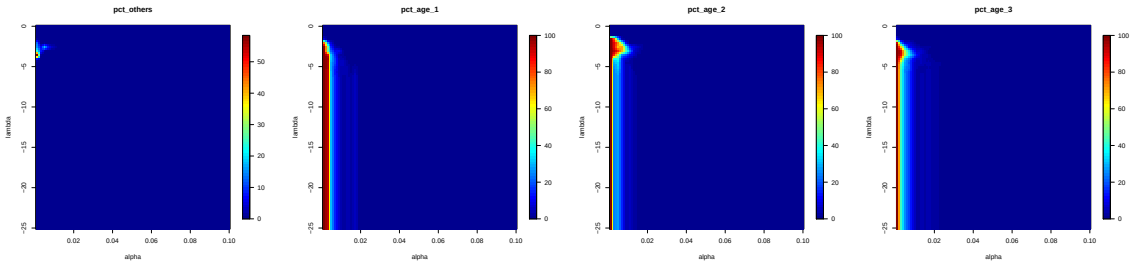
(e) College **(f)** Congestion **(g)** Travel time **(h)** Sd congestion



(i) Sd travel time **(j)** Obesity rates **(k)** Diabetes rates **(l)** poor health

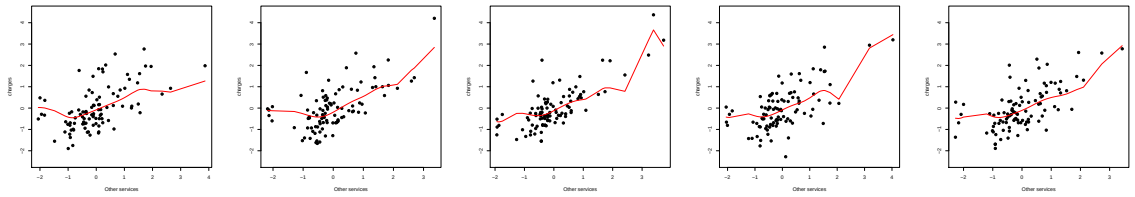


(m) Per Capita Income **(n)** Pct fast food rest **(o)** access healthy food **(p)** pct white

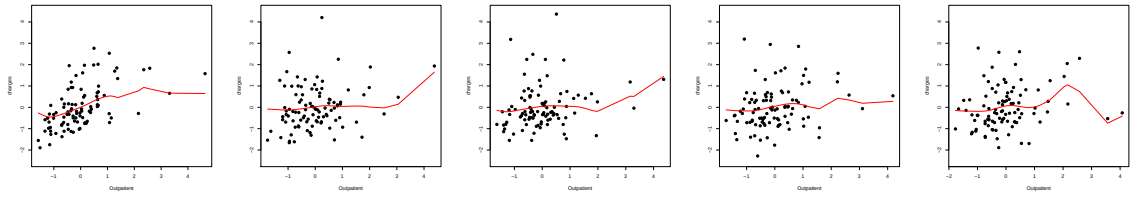


(q) pct non-white **(r)** Pct age (0-5) **(s)** Pct age (6 - 14) **(t)** Pct age (15 - 18)

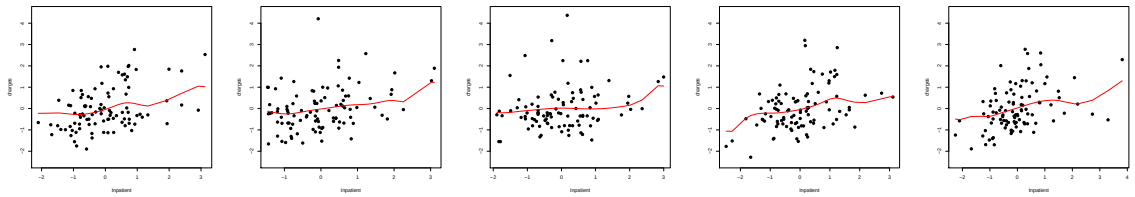
Figure 33: Stability Selection Plots for the Predictors Used in Medicaid Cost Analysis (Continued)



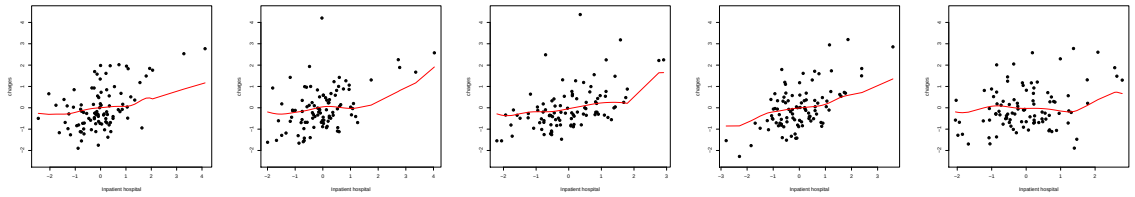
(a) Other services 2005 **(b)** Other services 2006 **(c)** Other services 2007 **(d)** Other services 2008 **(e)** Other services 2009



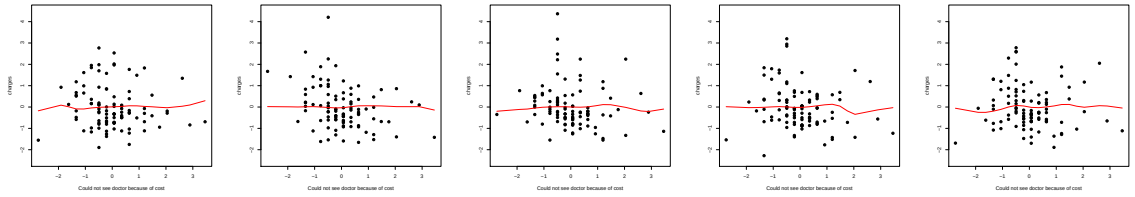
(f) Outpatient services 2005 **(g)** Outpatient services 2006 **(h)** Outpatient services 2007 **(i)** Outpatient services 2008 **(j)** Outpatient services 2009



(k) Inpatient services 2005 **(l)** Inpatient services 2006 **(m)** Inpatient services 2007 **(n)** Inpatient services 2008 **(o)** Inpatient services 2009

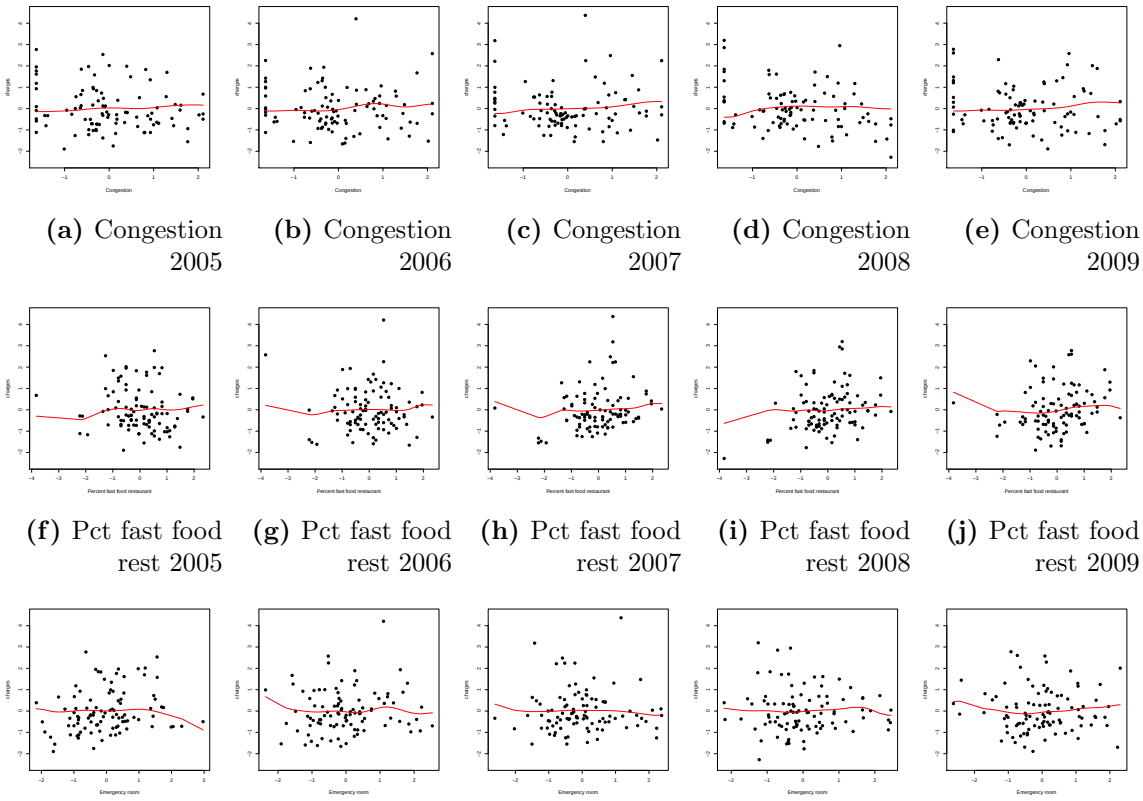


(p) Inpatient hospital 2005 **(q)** Inpatient hospital 2006 **(r)** Inpatient hospital 2007 **(s)** Inpatient hospital 2008 **(t)** Inpatient hospital 2009



(u) Couldn't see doctor 2005 **(v)** Couldn't see doctor 2006 **(w)** Couldn't see doctor 2007 **(x)** Couldn't see doctor 2008 **(y)** Couldn't see doctor 2009

Figure 34: Plots of some determinant of healths against the charges and the estimated additive functions



(a) Congestion 2005 **(b)** Congestion 2006 **(c)** Congestion 2007 **(d)** Congestion 2008 **(e)** Congestion 2009

(f) Pct fast food rest 2005 **(g)** Pct fast food rest 2006 **(h)** Pct fast food rest 2007 **(i)** Pct fast food rest 2008 **(j)** Pct fast food rest 2009

(k) Emergency room 2005 **(l)** Emergency room 2006 **(m)** Emergency room 2007 **(n)** Emergency room 2008 **(o)** Emergency room 2009

Figure 35: Plots of some determinant of healths against the charges and the estimated additive functions (Continued)

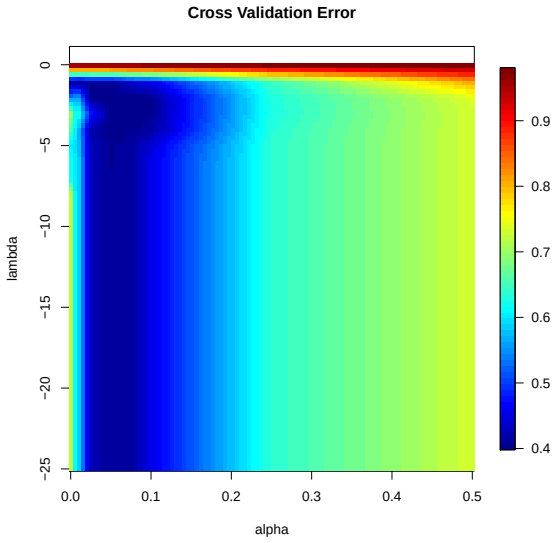


Figure 36: Cross validation error for Medical Cost Prediction for 2005 to 2009

APPENDIX C

SUPPLEMENT TO SEMIPARAMETRIC TOPOGRAPHICAL MIXTURE MODELS WITH SYMMETRIC ERRORS

Let us denote by $\|\cdot\|$ the Euclidean norm of a vector and by $\|\cdot\|_2$ the Frobenius norm of any squared matrix. Recall the definition of Z_k in (20) and let $J(t,u,h) := E[Z_1(t,u,h)]$. Let $\dot{Z}_k$ and $\dot{J}$ denote respectively the gradient of Z_k and J with respect to their first argument t.

Lemma 1 *Under assumption* **A1** *we have:*

i) For all $(u,h) \in \mathbb{R} \times \mathbb{R}_+^*$ *and any* $k = 1, ..., n$,

$$\sup_{t\in\Theta} |Z_k(t,u,h)| \leq \frac{2\|K\|_\infty}{h^d}, \quad \sup_{t\in\Theta} |J(t,u,h)| \leq 2\|\ell\|_\infty \cdot \int |K|.$$

ii) For all $(u,h) \in \mathbb{R} \times \mathbb{R}_+^*$ *and any* $k = 1, ..., n$,

$$\sup_{t\in\Theta} \|\dot{Z}_k(t,u,h)\| \leq 4(1+|u|)\frac{\|K\|_\infty}{h^d}, \quad \sup_{t\in\Theta} \|\dot{J}(t,u,h)\| \leq 4(1+|u|)\|\ell\|_\infty \cdot \int |K|.$$

iii) For all $(u,h) \in \mathbb{R} \times \mathbb{R}_+^*$ *and any* $k = 1, ..., n$,

$$\begin{aligned} \sup_{t\in\Theta} \|\ddot{Z}_k(t,u,h)\|_2 &\leq C(1+|u|+u^2)\frac{\|K\|_\infty}{h^d}, \\ \sup_{t\in\Theta} \|\ddot{J}_k(t,u,h)\|_2 &\leq C(1+|u|+u^2)\|\ell\|_\infty \cdot \int |K|, \end{aligned}$$

for some constant $C > 0$.

Proof of Lemma 1. i) It is easy to see, from $|M(t,u)| \leq 1$, that

$$|Z_k(t,u,h)| \leq 2|K_h(\mathbf{X}_k - \mathbf{x}_0)| \leq 2\frac{\|K\|_\infty}{h^d},$$

and that

$$|J(t,u,h)| \leq 2\left|\int \Im\left(g^*_{\mathbf{x}}(u)\bar{M}(t,u)\right) K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})d\mathbf{x})\right| \leq 2\|\ell\|_\infty . \int |K|.$$

ii) We note that

$$\begin{aligned}\dot{Z}_k(t,u,h) &= \left\{ e^{iuY_k} \begin{pmatrix} e^{-iu\alpha} - e^{-iu\beta} \\ -iu\pi e^{-iu\alpha} \\ -iu(1-\pi)e^{-iu\beta} \end{pmatrix} \right. \\ &\quad \left. -e^{-iuY_k} \begin{pmatrix} e^{iu\alpha} - e^{iu\beta} \\ iu\pi e^{iu\alpha} \\ iu(1-\pi)e^{iu\beta} \end{pmatrix} \right\} K_h(\mathbf{X}_k - \mathbf{x}_0),\end{aligned}$$

and that

$$\begin{aligned}E[\dot{Z}_k(t,u,h)] = \dot{J}_k(t,u,h) &= \int \left\{ g_{\mathbf{x}^*}(u) \begin{pmatrix} e^{-iu\alpha} - e^{-iu\beta} \\ -iu\pi e^{-iu\alpha} \\ -iu(1-\pi)e^{-iu\beta} \end{pmatrix} \right. \\ &\quad \left. -g_{\mathbf{x}^*}(-u) \begin{pmatrix} e^{iu\alpha} - e^{iu\beta} \\ iu\pi e^{iu\alpha} \\ iu(1-\pi)e^{iu\beta} \end{pmatrix} \right\} K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})d\mathbf{x}.\end{aligned}$$

We thus have

$$\begin{aligned}\|\dot{Z}_k(t,u,h)\| &= \left\| e^{iuY_k}\dot{M}(t,-u) - e^{-iuY_k}\dot{M}(t,u) \right\| K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &\leq \left(2\left(2^2 + P^2u^2 + (1-p)^2u^2\right)\right)^{1/2} K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &\leq 4(1+|u|)\frac{\|K\|_\infty}{h^d},\end{aligned}$$

and

$$\begin{aligned}\|\dot{J}_k(t,u,h)\| &= \int \left\| g^*_{\mathbf{x}}(u)\dot{M}(t,-u) - g^*_{\mathbf{x}}(-u)\dot{M}(t,u) \right\| |K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})|d\mathbf{x} \\ &\leq \left(2\left(2^2 + P^2u^2 + (1-p)^2u^2\right)\right)^{1/2} \int |K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})|d\mathbf{x} \\ &\leq 4(1+|u|)\|\ell\|_\infty . \int |K|.\end{aligned}$$

iii) Formula of $\ddot{M}(t,u)$ being tedious, we shortly write that

$$\ddot{Z}_k(t,u,h) \;=\; \left\{e^{iuY_k}\ddot{M}(t,-u) - e^{-iuY_k}\ddot{M}(t,u)\right\} K_h(\mathbf{X}_k - \mathbf{x}_0),$$

and deduce our bound from the above expression using arguments similar to i) and ii). ■

Lemma 2 *i) For all $(t,t') \in \Theta^2$, there exists a constant $C_1 > 0$ such that*

$$|S_n(t) - S_n(t')| \leq C_1 \|t - t'\| \sum_{j\neq k,j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)}.$$

ii) For all $(t,t') \in \Theta^2$, there exists a constant $C_2 > 0$ such that

$$\|\ddot{S}_n(t) - \ddot{S}_n(t')\|_2 \leq C_2 \|t - t'\| \sum_{j\neq k,j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)}.$$

iii) There exists some constants C_1, $C_2 > 0$ depending on Θ, α, M, K such that

$$E\left[\left(\sum_{j\neq k,j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} - \ell^2(\mathbf{x}_0)\right)^2\right] \leq C_1 h^{2\alpha} + \frac{C_2}{nh^d},$$

as $h \to 0$ and $nh^d \to \infty$.

Proof. i) By a first order Taylor expansion we have

$$S_n(t) - S_n(t') = -\frac{1}{2n(n-1)} \int (t-t')^\top \sum_{j\neq k,j,k=1}^{n} \dot{Z}_k(t_u,u,h) Z_j(t_u,u,h) w(u) du,$$

where for all $u \in \mathbb{R}$, t_u lies in the line segment with extremities t and t'. Therefore, according to calculations made in the proofs of Lemma 1 i) and ii), we obtain

$$|S_n(t) - S_n(t')| \leq \|t - t'\| \int_{\mathbb{R}} 4(1+|u|)w(u)du \left| \sum_{j\neq k,j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} \right|,$$

which ends the proof of i) by using assumption **A4**.

ii) Let recall first that

$$\ddot{S}_n(t) = \frac{-1}{2n(n-1)} \sum_{k\neq j} \int \left[\ddot{Z}_k(t,u,h) Z_j(t,u,h) + \dot{Z}_k(t,u,h)\dot{Z}_j(t,u)^\top\right] w(u) du.$$

We shall bound from above as follows

$$\begin{aligned}\|\ddot{S}_n(t,u)-\ddot{S}_n(t',u)\|_2 \;\leq\; & \frac{1}{2n(n-1)}\sum_{k\neq j}\Big\{\Big\|\int(\ddot{Z}_k(t,u,h)-\ddot{Z}_k(t',u,h))Z_j(t,u)w(u)du\Big\|_2 \\ & +\Big\|\int\ddot{Z}_k(t',u,h)(Z_j(t,u,h)-Z_j(t',u,h))w(u)du\Big\|_2 \\ & +\Big\|\int\dot{Z}_k(t,u,h)(\dot{Z}_j(t,u,h)-\dot{Z}_j(t',u,h))^\top w(u)du\Big\|_2 \\ & +\Big\|\int(\dot{Z}_k(t,u,h)-\dot{Z}_k(t',u,h))\dot{Z}_j(t',u,h)^\top w(u)du\Big\|_2\Big\}.\end{aligned}$$

For each term in the previous sum, we use Taylor expansion and upper-bounds similar to those developed in the proof of Lemma 1, and get

$$\begin{aligned}&\Big\|\ddot{S}_n(t,u)-\ddot{S}_n(t',u)\Big\|_2 \\ \leq\; & \|t-t'\|\,C\int(1+|u|+u^2+|u|^3)w(u)du\left|\sum_{j\neq k,j,k=1}^{n}\frac{K_h(\mathbf{X}_k-\mathbf{x}_0)K_h(\mathbf{X}_j-\mathbf{x}_0)}{n(n-1)}\right|,\end{aligned}$$

for some constant $C>0$, which finishes the proof by using assumption **A4**.

iii) The proof is a consequence of Proposition 2 hereafter. ■

Proof of Proposition 2. We shall bound from above the mean square error by the usual decomposition into squared bias plus variance.

Note that

$$E[S_n(t)] \;=\; -\frac{1}{4}\int(E[Z_1(t,u,h)])^2w(u)du$$

as $(Y_i,\mathbf{X}_i)$, $i=1,...,n$ are independent. Moreover,

$$\begin{aligned}E[Z_1(t,u,h)] &= \int\int\left(e^{iuy}M(t,-u)-e^{-iuy}M(t,u)\right)K_h(\mathbf{x}-\mathbf{x}_0)g(y,\mathbf{x})dyd\mathbf{x} \\ &= \int\left(\int\left(e^{iuy}M(t,-u)-e^{-iuy}M(t,u)\right)g_{\mathbf{x}}(y)dy\right)\ell(\mathbf{x})K_h(\mathbf{x}-\mathbf{x}_0)d\mathbf{x} \\ &= \int\left(g^*_{\mathbf{x}}(u)M(t,-u)-g^*_{\mathbf{x}}(-u)M(t,u)\right)\ell(\mathbf{x})K_h(\mathbf{x}-\mathbf{x}_0)d\mathbf{x}.\end{aligned}$$

Let us denote by $L(\mathbf{x},t,u):=g^*_{\mathbf{x}}(u)M(t,-u)-g^*_{\mathbf{x}}(-u)M(t,u)$, which is further equal to

$$L(\mathbf{x},t,u)=2i\cdot\Im\left(g^*_{\mathbf{x}}(u)M(t,-u)\right)=2i\cdot\Im\left(M(\theta(\mathbf{x}),u)M(t,-u)\right)f^*_{\mathbf{x}}(u).$$

We can write $E[Z_1(t,u,h)] = [(L(\cdot,t,u)\ell) \star K_h](\mathbf{x}_0)$, where $\star$ denotes the convolution product. The bias of $S_n(t)$ is bounded from above as follows:

$$\begin{aligned} |E[S_n(t)] - S(t)| &= \frac{1}{4}\left| \int \left([(L(\cdot,t,u)\ell) \star K_h]^2(\mathbf{x}_0) - L^2(\mathbf{x}_0,t,u)\ell^2(\mathbf{x}_0) \right) w(u)du \right| \\ &\leq \frac{1}{4}\int |[(L(\cdot,t,u)\ell) \star K_h](\mathbf{x}_0) - L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)| \\ &\quad \cdot |[(L(\cdot,t,u)\ell) \star K_h](\mathbf{x}_0) + L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)| \, w(u)du. \end{aligned}$$

Now

$$|L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)| \leq 2\|\ell\|_\infty \leq 2C,$$

as $\|\ell\|_\infty$ is further bounded by a constant $C = C(\alpha, M)$ depending only on α, $M > 0$, uniformly over $\ell \in L(\alpha, M)$ (see remark following condition **A1**). We also have

$$\begin{aligned} E[Z_1(t,u,h)] = |[(L(\cdot,t,u)\ell) \star K_h](\mathbf{x}_0)| &\leq \int |L(\mathbf{x},t,u)|l(\mathbf{x})|K|_h(\mathbf{x}-\mathbf{x}_0)d\mathbf{x} \\ &\leq 2C\int |K|. \end{aligned} \tag{32}$$

Moreover, for all $u \in \mathbb{R}$,

$$\begin{aligned} & |[(L(\cdot,t,u)\ell) \star K_h](\mathbf{x}_0) - L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)| \\ \leq & \int |L(\mathbf{x}+\mathbf{x}_0,t,u)\ell(\mathbf{x}+\mathbf{x}_0) - L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)| \cdot |K|_h(\mathbf{x})d\mathbf{x} \\ \leq & \; c(|u| + \varphi(u)) \int \|\mathbf{x}\|^\alpha \cdot |K|_h(\mathbf{x})d\mathbf{x} \leq c \cdot h^\alpha (|u| + \varphi(u)) \int \|\mathbf{x}\|^\alpha \cdot |K|(\mathbf{x})d\mathbf{x}, \end{aligned}$$

under our assumptions **A1-A4**. Indeed, that implies that $L(\cdot,t,u)\ell(\cdot)$ is Hölder α-smooth for all $(t,u) \in \Theta \times \mathbb{R}$, with some constant $c > 0$, see Lemma 3. Therefore we get

$$|E[S_n(t)] - S(t)| \leq 2C(1 + \int |K|)\, c \left(\int \|\mathbf{x}\|^\alpha \cdot |K|(\mathbf{x})d\mathbf{x} \right) \cdot \left(\int |u| w(u)du \right) \cdot h^\alpha.$$

Similarly to $S_n(t)$ variance decomposition, we write

$$\begin{aligned}
& S_n(t) - E[S_n(t)] \\
= & \frac{-1}{4n(n-1)} \sum_{j \neq k} \left(\int (Z_j(t,u,h) Z_k(t,u,h) - E^2[Z_1(t,u,h)]) w(u) du \right) \\
= & \frac{-1}{2n} \sum_j \int (Z_j(t,u,h) - E[Z_1(t,u,h)]) E[Z_1(t,u,h)] w(u) du \\
& + \frac{-1}{4n(n-1)} \sum_{j \neq k} \left(\int (Z_j(t,u,h) - E[Z_1(t,u,h)])(Z_k(t,u,h) - E[Z_1(t,u,h)]) w(u) du \right) \\
= & T_1 + T_2, \text{ say.}
\end{aligned}$$

Terms in T_1 and T_2 are uncorrelated and thus $Var(S_n(t)) = Var(T_1) + Var(T_2)$. On the one hand,

$$\begin{aligned}
Var(T_1) &= \frac{1}{4n} Var \left(\int (Z_1(t,u,h) - E[Z_1(t,u,h)]) E[Z_1(t,u,h)] w(u) du \right) \\
&= \frac{1}{4n} E \left[\left| \int (Z_1(t,u,h) - E[Z_1(t,u,h)]) E[Z_1(t,u,h)] w(u) du \right|^2 \right] \\
&\leq \frac{1}{4n} E \left[\int |Z_1(t,u,h) - E[Z_1(t,u,h)]|^2 w(u) du \right] \int |E[Z_1(t,u,h)]|^2 w(u) du,
\end{aligned}$$

according to Cauchy-Schwarz inequality. Now we use (32) and obtain

$$Var(T_2) \leq \frac{1}{4n} \left(2C \int |K| \right)^2 \int E[|Z_1(t,u,h)|^2] w(u) du.$$

We have,

$$\begin{aligned}
E[|Z_1(t,u,h)|^2] &= E \left[E \left[\left| 2i \cdot \Im \left(e^{iuY} M(t,-u) \right) \right|^2 \middle| \mathbf{X} \right] (K_h(\mathbf{X} - \mathbf{x}_0))^2 \right] \\
&= 4E \left[\left| \Im \left(g^*_{\mathbf{X}}(u) M(t,-u) \right) \right|^2 (K_h(\mathbf{X} - \mathbf{x}_0))^2 \right] \\
&\leq 4 \int \frac{1}{h^{2d}} K^2 \left(\frac{\mathbf{x} - \mathbf{x}_0}{h} \right) \ell(\mathbf{x}) d\mathbf{x} \\
&\leq 4C \frac{\int K^2}{h^d}.
\end{aligned}$$

Therefore,

$$Var(T_1) \leq 4C^3 \frac{(\int |K|)^2 \int K^2}{nh^d}, \tag{33}$$

for all $t \in \Theta$, $h > 0$.

On the other hand,

$$\begin{aligned} Var(T_2) &= \frac{1}{16n(n-1)} E\left[\left|\int (Z_1(t,u,h) - E[Z_1(t,u,h)])(Z_2(t,u,h) - E[Z_2(t,u,h)])w(u)du\right|^2\right] \\ &\leq \frac{1}{16n(n-1)} E\left[\int |Z_1(t,u,h) - E[Z_1(t,u,h)]|^2 |Z_2(t,u,h) - E[Z_2(t,u,h)]|^2 w(u)du\right] \\ &\leq \frac{1}{16n(n-1)} \int E^2[|Z_1(t,u,h)|^2] w(u)du \leq \frac{1}{16n(n-1)} \left(\frac{2C\int K^2}{h^d}\right)^2 \\ &= \frac{C^2(\int K^2)^2}{4n(n-1)h^{2d}}, \end{aligned}$$

which is clearly a $o((nh^d)^{-1})$ and concludes the proof. ■

Lemma 3 (Smoothness of $L(\mathbf{x},t,u)\ell(\mathbf{x})$) *Assume* **A1**-**A4**. *There exists a constant $C > 0$, such that for all $(\mathbf{x},\mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d$ and all $(t,u) \in \Theta \times \mathbb{R}$:*

$$|L(\mathbf{x},t,u)\ell(\mathbf{x}) - L(\mathbf{x}',t,u)\ell(\mathbf{x}')| \leq C(|u| + \varphi(u))\|\mathbf{x} - \mathbf{x}'\|^\alpha.$$

Proof. For $t = (\pi, a, b) \in \Theta$, and $(\mathbf{x}, u) \in \mathbb{R}^d \times \mathbb{R}$ we write

$$L(\mathbf{x},t,u)\ell(\mathbf{x}) = f^*_{\mathbf{x}}(u)\ell(\mathbf{x})\mathcal{T}(\mathbf{x},t,u), \text{ and } \mathcal{T}(\mathbf{x},t,u) := \sum_{i=1}^4 \mathcal{T}_i(\mathbf{x},t,u)$$

where

$$\begin{aligned} \mathcal{T}_1(\mathbf{x},t,u) &= \pi(\mathbf{x})\pi \sin[u(a(\mathbf{x}) - a)], \\ \mathcal{T}_2(\mathbf{x},t,u) &= \pi(\mathbf{x})(1-\pi) \sin[u(a(\mathbf{x}) - b)], \\ \mathcal{T}_3(\mathbf{x},t,u) &= (1-\pi(\mathbf{x}))\pi \sin[u(b(\mathbf{x}) - a)], \\ \mathcal{T}_4(\mathbf{x},t,u) &= (1-\pi(\mathbf{x}))(1-\pi) \sin[u(b(\mathbf{x}) - b)]. \end{aligned}$$

For all $(\mathbf{x},\mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d$ we have

$$\begin{aligned} &|L(\mathbf{x},t,u)\ell(\mathbf{x}) - L(\mathbf{x}',t,u)\ell(\mathbf{x}')| \\ &\leq 2|f^*_{\mathbf{x}}(u)\ell(\mathbf{x})||\mathcal{T}(\mathbf{x},t,u) - \mathcal{T}(\mathbf{x}',t,u)| + 2|\mathcal{T}(\mathbf{x}',t,u)||f^*_{\mathbf{x}}(u)\ell(\mathbf{x}) - f^*_{\mathbf{x}'}(u)\ell(\mathbf{x}')| \\ &\leq 2\|\ell\|_\infty |\mathcal{T}(\mathbf{x},t,u) - \mathcal{T}(\mathbf{x}',t,u)| + 2|f^*_{\mathbf{x}}(u)\ell(\mathbf{x}) - f^*_{\mathbf{x}'}(u)\ell(\mathbf{x}')|. \end{aligned}$$

Let us now show the α-smooth Hölder property of $\mathcal{T}_1$, the proof for the other $\mathcal{T}_i$'s being completely similar. For all $(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d$

$$\begin{aligned} |\mathcal{T}_1(\mathbf{x}, t, u) - \mathcal{T}_1(\mathbf{x}', t, u)| &\leq |\sin[u(a(\mathbf{x}) - a)] - \sin[u(a(\mathbf{x}') - a)]| + |\pi(\mathbf{x}) - \pi(\mathbf{x}')| \\ &\leq |u||(a(\mathbf{x}) - a(\mathbf{x}')]| + |\pi(\mathbf{x}) - \pi(\mathbf{x}')| \\ &\leq M|u|\|\mathbf{x} - \mathbf{x}'\|^\alpha + M\|\mathbf{x} - \mathbf{x}'\|^\alpha. \end{aligned}$$

On the other hand we have

$$\begin{aligned} |f^*_{\mathbf{x}}(u)\ell(\mathbf{x}) - f^*_{\mathbf{x}'}\ell(\mathbf{x}')| &\leq |\ell(\mathbf{x}) - \ell(\mathbf{x}')| + \|\ell\|_\infty |f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)|, \\ &\leq (M + \|\ell\|_\infty \varphi(u))\|\mathbf{x} - \mathbf{x}'\|^\alpha, \end{aligned}$$

which concludes the proof. ■

Proof of Theorem 3. Our method is based on a consistency proof for mininum contrast estimators by Dacunha-Castelle and Duflo (1993, pp.94–96). Let us consider a countable dense set D in Θ, then $\inf_{t\in\Theta} S_n(t) = \inf_{t\in D} S_n(t)$, is a measurable random variable. We define in addition the random variable

$$W(n, \xi) = \sup\left\{|S_n(t) - S_n(t')|;\ (t, t') \in D^2,\ \|t - t'\| \leq \xi\right\},$$

and recall that $S(\theta_0) = 0$. Let us consider a non-empty open ball B_* centered on θ_0 such that S is bounded from below by a positive real number 2ε on $\Theta\backslash B_*$. Let us consider a sequence $(\xi_p)_{p\geq 1}$ decreasing to zero, and take p such that there exists a covering of $\Theta\backslash B_*$ by a finite number κ of balls $(B_i)_{1\leq i\leq\kappa}$ with centers $t_i \in \Theta$, $i = 1, \ldots, \kappa$, and radius less than ξ_p. Then, for all $t \in B_i$, we have

$$S_n(t) \geq S_n(t_i) - |S_n(t) - S_n(t_i)| \geq S_n(t_i) - \sup_{t\in B_i} |S_n(t) - S_n(t_i)|,$$

which leads to

$$\inf_{t\in\Theta\backslash B_*} S_n(t) \geq \inf_{1\leq i\leq\kappa} S_n(t_i) - W(n, \xi_p).$$

As a consequence we have the following events inclusions

$$\begin{aligned}\left\{\hat{\theta}_n \notin B_*\right\} &\subseteq \left\{\inf_{t\in\Theta\setminus B_*} S_n(t) < \inf_{t\in B_*} S_n(t) < S_n(\theta_0)\right\} \\ &\subseteq \left\{\inf_{1\le i\le\kappa} S_n(t_i) - W(n,\xi_p) < S_n(\theta_0)\right\} \\ &\subseteq \{W(n,\xi_p) > \varepsilon\} \cup \left\{\inf_{1\le i\le\kappa} (S_n(t_i) - S_n(\theta_0)) \le \varepsilon\right\}.\end{aligned}$$

In addition we have

$$\begin{aligned}&P\left(\inf_{1\le i\le\kappa} (S_n(t_i) - S_n(\theta_0)) \le \varepsilon\right) \\ &\le 1 - \prod_{i=1}^{\kappa} (1 - [P(|S_n(t_i) - S(t_i)| \ge \varepsilon) + P(|S_n(\theta_0) - S(\theta_0)| \ge \varepsilon)]),\end{aligned}$$

where, according to Proposition 2, the last two terms in the right hand side of the above inequality vanish to zero if $h^d n \to \infty$ and $h \to 0$ as $n \to \infty$. To conclude we use Lemma 2 and notice that, for all $(t, t') \in \Theta^2$, we have

$$\begin{aligned}&|S_n(t) - S_n(t')| \\ &\le \frac{C\|t-t'\|}{n(n-1)} \left| \sum_{j\ne k, j,k=1}^{n} K_h(\mathbf{X}_k - \mathbf{x}_0) K_h(\mathbf{X}_j - \mathbf{x}_0) \right| \\ &\le C\|t-t'\|\ell^2(\mathbf{x}_0) + C\|t-t'\| \left| \sum_{j\ne k, j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0) K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} - \ell^2(\mathbf{x}_0) \right| \quad (34)\end{aligned}$$

We deduce from above that

$$\begin{aligned}P(W(n,\xi_p) > \varepsilon) &\le P\left(C\xi_p\ell^2(\mathbf{x}_0) > \frac{\varepsilon}{2}\right) \\ &\quad + \left(\frac{2C\xi_p}{\varepsilon}\right)^2 E\left[\left(\sum_{j\ne k, j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0) K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} - \ell^2(\mathbf{x}_0)\right)^2\right],\end{aligned}$$

where the last term in the right hand side is of order $(nh^d)^{-1} + h^{2\alpha}$ and tends to 0 by our assumption on h. Since for p sufficiently large we have $C\xi_p\ell^2(\mathbf{x}_0) < \varepsilon/2$ and thus $P\left(C\xi_p\ell^2(\mathbf{x}_0) > \varepsilon/2\right) = 0$, this concludes the proof of the consistency in probability of $\hat{\theta}_n$ when $nh^d \to \infty$ and $h \to 0$ as $n \to \infty$. ■

Proof of Theorem 4. By a Taylor expansion of $\dot{S}_n$ around θ_0, we have

$$0 = \dot{S}_n(\hat{\theta}_n) = \dot{S}_n(\theta_0) + \ddot{S}_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0),$$

where $\bar{\theta}_n$ lies in the line segment with extremities $\hat{\theta}_n$ and θ_0.

Let us study the behaviour of

$$\dot{S}_n(\theta_0) = \frac{-1}{2n(n-1)} \sum_{j \neq k} \int \dot{Z}_k(\theta_0, u, h) Z_j(\theta_0, u, h) w(u) du,$$

where $\dot{Z}_k$ denotes the gradient of Z_k with respect to the first argument. Recall that $\theta_0 = \theta(\mathbf{x}_0) = (\pi(\mathbf{x}_0), a(\mathbf{x}_0), b(\mathbf{x}_0))$ and therefore

$$J(t, u, h) = E[Z_1(t, u, h)] = 2i \int \Im \left(M(\theta(\mathbf{x}), u) M(t, -u) \right) f^*_{\mathbf{x}}(u) \ell(\mathbf{x}) K_h(\mathbf{x} - \mathbf{x}_0) d\mathbf{x},$$

satisfies $J(\theta_0, u, h) \to 0$ as $h \to 0$. Indeed, the last integral may be equal to 0 if the set $\{\mathbf{x} : \theta(\mathbf{x}) = \theta(\mathbf{x}_0)\}$ has Lebesgue measure 0, or tends (by uniform continuity in $\mathbf{x}$ of the integrand) to

$$2i \Im \left(M(\theta(\mathbf{x}_0), u) M(\theta(\mathbf{x}_0), -u) \right) f^*_{\mathbf{x}_0}(u) \ell(\mathbf{x}_0) = 0.$$

Moreover,

$$\dot{Z}_k(t, u, h) = \Im \left(\dot{M}(t, -u) e^{iuY_k} \right) K_h(\mathbf{X}_k - \mathbf{x}_0).$$

Denote $\dot{J}(t, u, h) = E[\dot{Z}_k(t, u, h)]$ and observe that

$$\dot{J}(t, u, h) = \int \Im \left(\dot{M}(t, -u) M(\theta(\mathbf{x}), u) f^*_{\mathbf{x}}(u) \right) K_h(\mathbf{x} - \mathbf{x}_0) \ell(\mathbf{x}) d\mathbf{x}.$$

Then, we decompose $\dot{S}_n(\theta_0)$ as follows

$$\begin{aligned}
& \dot{S}_n(\theta_0) \\
= \; & \frac{-1}{2n(n-1)} \sum_{j \neq k} \int \left(\dot{Z}_k(\theta_0, u, h) - \dot{J}(\theta_0, u, h) \right) \left(Z_j(\theta_0, u, h) - E[Z_j(\theta_0, u, h)] \right) w(u) du \\
& - \frac{1}{2n} \sum_{j=1}^{n} \int \dot{J}(\theta_0, u, h) (Z_j(\theta_0, u, h) - E[Z_j(\theta_0, u, h)]) w(u) du \\
:= \; & -\frac{1}{2} (A_n(h) + B_n(h)), \qquad (35)
\end{aligned}$$

where terms in $A_n(h)$ and $B_n(h)$ are uncorrelated. On the one hand, we use a multivariate Central Limit Theorem for independent random variables taking values in a

Hilbert space, following Kandelaki and Sozanov (1964) or Gikhman and Skorokhod (2004, Theorem 4, page 396). This will give us the limit behavior of the term

$$B_n(h) = \frac{1}{n}\sum_{j=1}^{n} U_j(h), \quad U_j(h) := \int \dot{J}(\theta_0, u, h)(Z_j(\theta_0, u, h) - E[Z_j(\theta_0, u, h)])w(u)du.$$

The random variables $U_j(h)$, $j = 1, ..., n$ are independent, centered, but their common law depend on n via h. Our goal is to show that

$$nh^d Var(B_n(h)) = \sum_{j=1}^{n} Var\left(\sqrt{\frac{h^d}{n}}U_j(h)\right) \to \Sigma, \quad \text{as } n \to \infty \tag{36}$$

and that

$$\sum_{j=1}^{n} E\left[\left\|\sqrt{\frac{h^d}{n}}U_j(h)\right\|^4\right] = \frac{h^{2d}}{n}E[\|U_1(h)\|^4] \to 0, \quad \text{as } n \to \infty. \tag{37}$$

Indeed, (37) implies the Lindeberg's condition in Kandelaki and Sozanov (1964):

$$\sum_{j=1}^{n} E\left[\left\|\sqrt{\frac{h^d}{n}}U_j(h)\right|^2 \cdot \mathbb{I}_{\left\|\sqrt{h^d/n}U_j(h)\right\| \geq \varepsilon}\right] \to 0, \quad \text{as } n \to \infty, \text{ for any } \varepsilon > 0.$$

On the other hand, we prove that

$$\sqrt{nh^d}A_n(h) \to 0, \text{ in probability, as } n \to \infty, \tag{38}$$

stating that $\sqrt{nh^d}A_n(h)$ negligible term and that, as a consequence, the limiting behavior of $\sqrt{nh^d}\dot{S}_n(\theta_0)$ is only driven by $\sqrt{nh^d}B_n(h)$. This will end the proof of the theorem.

Let us prove (36) and (37). Note that $nh^d Var(B_n(h)) = h^d Var(U_1(h))$ and that

$$\begin{aligned} & Var(U_1(h)) \\ = & \int\int \dot{J}(\theta_0, u_1, h)\dot{J}^\top(\theta_0, u_2, h)Cov(Z_1(\theta_0, u_1, h), Z_1(\theta_0, u_2, h))w(u_1)w(u_2)du_1du_2. \end{aligned}$$

Similarly to Proposition 2, by uniform continuity in $\mathbf{x}$ of the integrand in $\dot{J}$, we get

$$\lim_{h\to 0} \dot{J}(\theta_0, u, h) := \dot{J}(\theta_0, u).$$

See that $\|\dot{J}(\theta_0,u)\| \leq 2(1+|u|)\|\ell\|_\infty$ and that the latter upper bound is integrable with respect to the measure $w(u)du$ by assumption on w. It remains to study:

$$\begin{aligned} & Cov(Z_1(\theta_0,u_1,h), Z_1(\theta_0,u_2,h)) \\ = \; & E\left[Z_1(\theta_0,u_1,h)Z_1(\theta_0,u_2,h)\right] - E\left[Z_1(\theta_0,u_1,h)\right] E\left[Z_1(\theta_0,u_2,h)\right]. \end{aligned}$$

From (32) we deduce that

$$h^d |E\left[Z_1(\theta_0,u_1,h)\right] E\left[Z_1(\theta_0,u_2,h)\right]| \leq h^d \left(2C\int |K|\right)^2 \to 0,$$

when $h \to 0$ as $n \to \infty$. We also have

$$\begin{aligned} & h^d E\left[Z_1(\theta_0,u_1,h)Z_1(\theta_0,u_2,h)\right] \\ = \; & 4\cdot \int\int \Im\left(e^{iu_1 y} M(\theta_0,-u_1)\right) \Im\left(e^{iu_2 y} M(\theta_0,-u_2)\right) \frac{1}{h^d} K^2(\frac{\mathbf{x}-\mathbf{x}_0}{h}) g(y,\mathbf{x}) dy d\mathbf{x} \\ = \; & 4\cdot \int \Im\left(e^{iu_1 y} M(\theta_0,-u_1)\right) \cdot \Im\left(e^{iu_2 y} M(\theta_0,-u_2)\right) g(y,\mathbf{x}_0) dy (\int K^2)(1+o(1)) \\ = \; & 4\cdot \int \Im\left(e^{iu_1 y} M(\theta_0,-u_1)\right) \cdot \Im\left(e^{iu_2 y} M(\theta_0,-u_2)\right) g_{\mathbf{x}_0}(y) dy \cdot \ell(\mathbf{x}_0) (\int K^2)(1+o(1)), \end{aligned}$$

as $h \to 0$. See also that we can write

$$\begin{aligned} V(\theta_0,u_1,u_2) \; := \; & \int \left(e^{iu_1 y} M(\theta_0,-u_1) - e^{-iu_1 y} M(\theta_0,u_1)\right) \\ & \cdot \left(e^{iu_2 y} M(\theta_0,-u_2) - e^{-iu_2 y} M(\theta_0,u_2)\right) g_{\mathbf{x}_0}(y) dy \\ = \; & M(\theta_0,u_1+u_2) M(\theta_0,-u_1) M(\theta_0,-u_2) f^*_{\mathbf{x}_0}(u_1+u_2) \\ & - M(\theta_0,u_1-u_2) M(\theta_0,-u_1) M(\theta_0,u_2) f^*_{\mathbf{x}_0}(u_1-u_2) \\ & - M(\theta_0,-u_1+u_2) M(\theta_0,u_1) M(\theta_0,-u_2) f^*_{\mathbf{x}_0}(-u_1+u_2) \\ & + M(\theta_0,-u_1-u_2) M(\theta_0,u_1) M(\theta_0,u_2) f^*_{\mathbf{x}_0}(-u_1-u_2) \end{aligned}$$

and this is a bounded function with respect to u_1 and u_2. Therefore

$$h^d Var(U_1(h)) \to \int\int \dot{J}(\theta_0,u_1)\dot{J}^\top(\theta_0,u_2) V(\theta_0,u_1,u_2) w(u_1) w(u_2) du_1 du_2 =: \Sigma,$$

as $h \to 0$. This proves (36).

Now, denote by $v^{(k)}$ the k-th coordinate of a vector v and use Jensen inequality to see that

$$\begin{aligned} E[\|U_1(h)\|^4] &\leq 3\left(E[(U_1^{(1)}(h))^4] + E[(U_1^{(2)}(h))^4] + E[(U_1^{(3)}(h))^4]\right) \\ &\leq 3\sum_{k=1}^{3} E\left[\left(\int \dot{J}^{(k)}(\theta_0,u,h)(Z_1(\theta_0,u,h) - E[Z_1(\theta_0,u,h)])w(u)du\right)^4\right] \\ &\leq 3\sum_{k=1}^{3}\int |\dot{J}^{(k)}(\theta_0,u,h)|^4 E\left[|Z_1(\theta_0,u,h)|^4\right] w(u)du. \end{aligned}$$

We have $|\dot{J}^{(k)}(\theta_0,u,h)| \leq 4(1+|u|)(\int |K|)\|\ell\|_\infty$ by Lemma 1 and

$$\begin{aligned} E\left[|Z_1(\theta_0,u,h)|^4\right] &= \int\int 4\left|\Im\left(e^{iuy}M(\theta_0,-u)\right)\right|^4 \frac{1}{h^{4d}}K^4\left(\frac{\mathbf{x}-\mathbf{x}_0}{h}\right) g(y,\mathbf{x})dyd\mathbf{x} \\ &\leq \frac{4}{h^{3d}}\int \frac{1}{h^d}K^4\left(\frac{\mathbf{x}-\mathbf{x}_0}{h}\right)\ell(\mathbf{x})d\mathbf{x} \\ &\leq \frac{O(1)}{h^{3d}}\left(\int K^4\right)\|\ell\|_\infty, \end{aligned}$$

as $h \to 0$. Therefore,

$$\frac{h^{2d}}{n}E[\|U_1(h)\|^4] \leq \frac{O(1)}{nh^d}\int |K| \cdot \int K^4 \cdot \int (1+|u|)^4 w(u)du = o(1),$$

as $n \to \infty$ and $h \to 0$ such that $nh^d \to \infty$. This proves (37).

To prove (38), we notice that $A_n(h)$ defined in (35) can be treated similarly to T_1 in (33). By this remark, we easily prove that $Var(A_n) = o\left((nh^d)^{-1}\right)$ which insure the wanted result.

Let us prove that

$$\ddot{S}_n(\theta_n) \longrightarrow \mathcal{I}(\theta_0), \text{ in probability, as } n \to \infty,$$

where $\mathcal{I} = \mathcal{I}(\theta_0) = -\frac{1}{2}\int \dot{J}(\theta_0,u)\dot{J}^\top(\theta_0,u)w(u)du$, and $\dot{J}(\theta_0,u)$ is defined in (25). We start by writing the triangular inequality

$$\|\ddot{S}_n(\theta_n) - \mathcal{I}\| \leq \|\ddot{S}_n(\theta_n) - \ddot{S}_n(\theta_0)\| + \|\ddot{S}_n(\theta_0) - E(\ddot{S}_n(\theta_0))\| + \|E(\ddot{S}_n(\theta_0)) - \mathcal{I}\|.$$

Then using upper bounds similar to (34) slighly adapted to $\ddot{S}_n$ instead of S_n and the convergence in probability of $\hat{\theta}_n$ towards θ_0 established in Theorem 3, we have that $\|\ddot{S}_n(\theta_n) - \ddot{S}_n(\theta_0)\| \to 0$ in probability as $n \to \infty$. By writting

$$E(\ddot{S}_n(\theta_0)) = -\frac{1}{2}\int \left(\ddot{J}(\theta_0,u,h)J(\theta_0,u,h) + \dot{J}(\theta_0,u,h)\dot{J}(\theta_0,u,h)^\top\right) w(u)du$$

and noticing, according to Bochner's Lemma, that $J(\theta_0,u,h) \to 0$ and $\dot{J}(\theta_0,u,h) \to \dot{J}(\theta_0,u)$ as $h \to 0$, we have, according to the Lebesgue's theorem, that $E[\ddot{S}_n(\theta_0)]$ tends to $\mathcal{I}$ as $h \to 0$. Finally we decompose $-2n(n-1)(\ddot{S}_n(\theta_0) - E[\ddot{S}_n(\theta_0)]) = \sum_{l=1}^{3}(D_{1,l} + D_{2,l})$ where

$$\begin{aligned}
D_{1,1} &= \sum_{k\neq j}\int (\ddot{Z}_k(\theta_0,u,h) - \ddot{J}(\theta_0,u,h))(Z_j(\theta,u,h) - J(\theta_0,u,h))w(u)du \\
D_{1,2} &= (n-1)\sum_{k}\int (\ddot{Z}_k(\theta_0,u,h) - \ddot{J}(\theta_0,u,h))J(\theta_0,u,h)w(u)du \\
D_{1,3} &= (n-1)\sum_{j}\int \ddot{J}(\theta_0,u,h)(Z_j(\theta,u,h) - J(\theta_0,u,h))w(u)du,
\end{aligned}$$

and

$$\begin{aligned}
D_{2,1} &= \sum_{k\neq j}\int (\dot{Z}_k(\theta_0,u,h) - \dot{J}(\theta_0,u,h))(\dot{Z}_j(\theta,u,h) - \dot{J}(\theta_0,u,h))^\top w(u)du \\
D_{2,2} &= (n-1)\sum_{k}\int (\dot{Z}_k(\theta_0,u,h) - \dot{J}(\theta_0,u,h))J(\theta_0,u,h)^\top w(u)du \\
D_{2,3} &= (n-1)\sum_{j}\int \dot{J}(\theta_0,u,h)(Z_j(\theta,u,h) - J(\theta_0,u,h))^\top w(u)du.
\end{aligned}$$

Noticing that terms $D_{i,3}$, $i = 1,2$, respectively $D_{i,j,}$, $i = 1,2$ and $j = 2,3$, can be treated as T_1 respectively T_2 in the proof of Proposition 2, we obtain

$$Var\left(\ddot{S}_n(\theta_0)\right) = O\left(\frac{1}{nh^d}\right),$$

which concludes the proof. ■

REFERENCES

[1] Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika*, 17–26.

[2] Anderson, T. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, **22**: 327-351.

[3] Balabdaoui, F. and Butucea, C. (2014) On location mixtures with Pólya frequency components. *Statist. Probab. Letters*, in press.

[4] Bańbura, M., Giannone, D., Reichlin, L. (2010). Large bayesian vector autoregressions. *Journal of Applied Econometrics*, **25**: 71-92.

[5] Bedrick, E. and Tsai, C. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**: 226-231.

[6] Bordes, L., Kojadinovic, I. and Vandekerkhove, P. (2013) Semiparametric estimation of a two-component mixture of linear regressions in which one component is known. *Electr. J. Statist.*, 2603-2644.

[7] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232.

[8] Bowen, R. S., Chappell R. J., Bentzen S. M., Deveau, M. A., Forrest L. J., and Jeraj, R. (2012). Spatially resolved regression analysis of pre-treatment FDG, FLT and Cu-ATSM PET from post-treatment FDG PET: an exploratory study. *Radiother. Oncol.* **105**, 41–48.

[9] Breiman, L. (1995) Better subset regression using the nonnegative garrote, *Technometrics*, **4**:373-384.

[10] Brüggeman, R., Lütkepol, H. (2001). Lag selection in subset VAR models with an application to U.S. monetary system. In: *Friedmann,R., Knüppel, L., Lütkepol, H. (Eds.), Econometric Studies- A festschrift in honour of Joachim Frohn. LIT, Münster*,pp.107-128.

[11] Brunel E., Comte F. and Lacour, C. (2010) Minimax estimation of the conditional cumulative distribution function under random censorship. *Sankhya Series A*, **72**, 293-330.

[12] Bülhmann P. and Sarah Van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. *Heidelberg: Springer Verlag.*

[13] Butucea, C. and Vandekerkhove, P. (2014). Semiparametric mixtures of symmetric distributions. *Scand. J. Statist.*, **41**, p. 227-239.

[14] Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.*, **95**, 957–970.

[15] Cohen, S. and Le Pennec, E. (2012). Conditional Density Estimation by Penalized Likelihood Model Selection and Applications. URL `http://arxiv.org/abs/1103.2021`.

[16] Dacunha-Castelle, D. and Duflo, M. (1983). *Probabilités et Statistique 2. Problèmes à temps mobile.* Masson, Paris.

[17] De Luna, X., and Genton, M.G. (2005). Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica*, **15**: 547-568.

[18] De Veaux, R. D. (1989). Mixtures of linear regressions. *Comput. Statist. Data Analyis*, **8**, 227–245.

[19] Doran, T., Fullwood, C., Kontopantelis, E., and Reeves, D. (2008). Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. *The Lancet*, **372**(9640), 728-736.

[20] Foygel, R., Horrell, M., Drton, M., Lafferty, J. (2012). Nonparametric Reduced Rank Regression *arXiv:1301.1919,*

[21] Friedman,J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*,**9**: 432-441.

[22] Friedman,J., Hastie, T., Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso, *Technical Report*

[23] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression, *Biometrika*, **84**: 707-716.

[24] Gentili, M., Serban, N., and Swann, J. (2014). Spatial Access to Pediatric Primary Care: A study of disparities across multiple states in the US. *DECISION MODELS for SMARTER CITIES*

[25] Gikhman, I. and Skorokhod, A. (2004). *The theory of stochastic processes. I* Springer-Verlag, Berlin.

[26] Glynn, L. G., Valderas, J. M., Healy, P., Burke, E., Newell, J., Gillespie, P., and Murphy, A. W. (2011). The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Family practice,* **28**(5), 516-523.

[27] Gruen, B., Leisch, F., and Sarkar, D. (2013) *flexmix: Flexible Mixture Modeling.* URL http://CRAN.R-project.org/package=flexmix. R package version 2.3-11.

[28] Grün, B. and Leisch, F. (2006) *Fitting finite mixtures of linear regression models with varying and fixed effects in R.* In A. Rizzi and M. Vichi, editors, Compstat 2006, Proceedings in Computational Statistics, 853–860.

[29] Grupp-Phelan, J., Lozano, P., and Fishman, P. (2001). Health care utilization and cost in children with asthma and selected comorbidities. *Journal of Asthma,* **38**(4), 363-373.

[30] Hall, P., and Zhou, X-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31**, 201–224.

[31] Harrison, P. L., Pope, J. E., Coberley, C. R., and Rula, E. Y. (2012). Evaluation of the relationship between individual well-being and future health care utilization and cost. *Population health management*, **15**(6), 325-330.

[32] Hastie, T. and Tibshirani, R., (1986). Generalized Additive Models. *Statistical Science,* **1**(3), 297 - 318.

[33] Hawkins, D. S., Allen, D. M. and Stomber, A. J. (2001). Determining the number of components in mixtures of linear models. *Computational Statistics and Data Analysis*, **38**, 15–48.

[34] Herrmann E. (2013). *lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth,* 2013. URL http://CRAN.R-project.org/package=lokern. R package version 1.1-4.

[35] Hsu, N.J., Hung, H.L. and Chang V.M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, **52**: 3645-3657.

[36] Huang, H., Hsu, N., Theobald D.M., Breidt F.J. (2010) Spatial lasso with the application to GIS model selection, *Journal of Computational and Graphical Statistics*, **19**(4): 963-983.

[37] Huang, M., Li, R. and Wang, S. (2013). Nonparametric mixture of regression models. *J. Amer. Statist. Soc.* **108**, 229–241.

[38] Huang, M. and Yao, W. (2012). Mixture of Regression Models with Varying Mixing Proportions: A Semiparametric Approach. *J. Amer. Statist. Assoc.* **107**, 711-724.

[39] Hunter, D. R. and Young, D. S. (2012) Semiparametric mixtures of regressions. *J. Nonparam. Statist.* **24**, 19-38.

[40] Hunter, D. R., Wang, S. and Hettmanspeger, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251.

[41] Hurn, M., Justel, A. and Robert, C. P. (2003). Estimating mixtures of regressions. *J. Comput. Graph. Statist.* **12**, 1–25.

[42] Ibragimov, I. A. and Has'minski, R. Z. (1981). *Statistical estimation. Asymptotic theory.* Applications of Mathematics. Springer-Verlag, New York-Berlin.

[43] Izenman A.J. (1975). Reduced-rank regression for the multivariate linear regression model. *Journal of Multivariate Analysis*, **5**(2): 248-264.

[44] Jones, P. N. and McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Australian J. Statist.* **34**, 233–240.

[45] Kandelaki, N. P., and Sozanov, V. V. (1964). On a central limit theorem for random elements with values in Hilbert space. *Theory Probab. Appl.* **71** 38–46.

[46] Khan, J., Wei, J. S., Ringner, M., Saa, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**, 673 - 679.

[47] Krolzig, H.-M., Hendry, D.F. (2001). Computer automation of general linear regression model selection procedures, *Journal of Economic Dynamics and Control*, **25**: 831-866.

[48] Leung, D. H-Y., and Qin, J. (2006). Semi-parametric inference in a bivariate (multivariate) mixture model. *Statistica Sinica*, **16**, 153–163.

[49] Lin, Y. and Zhang, H. H., (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 2272 - 2297.

[50] Liu, H., Lafferty, J. and Wasserman, L., (2009). Nonparametric Regression and classification with joint sparsity constraints. In *Advances in neural information processing systems*, 969-976.

[51] Lozano, A., Li, H., Niculescu-Mizil, A., Liu, Y. Perlich C., Hosking, J., Abe, N. (2009). Spatial-temporal causal modeling for climate change attribution, *IBM T.J. Watson Research Center*.

[52] Lozano, A., and Grzegorz S. (2012) "Multi-level Lasso for Sparse Multi-task Regression." *ICML. 2012*.

[53] Lütkepohl, H. (2005). New introduction to multiple time series analysis, *Heidelberg: Springer Verlag*.

[54] Meier, L., Van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, **37**(6B), 3779-3821.

[55] Montuelle, L., Le Pennec, E., and Cohen, S. (2013). Gaussian Mixture Regression model with logistic weights, a penalized maximum likelihood approach. URL `http://arxiv.org/pdf/1304.2696v1.pdf`.

[56] Obonzinski, G., Wainwright, M.J. and Jordan M.I., (2008). Union support recoveryin high-dimensional multivariate regression. *Communication, Control and Computing,* 46^{th} *Annual Allerton Conference* : 21-26.

[57] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D., Pollack, J., Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, **4**: 53-77.

[58] Penm, J.H.W., Terrell, R.D., (1982). On the recursive fitting of the subset autoregressions. *Journal of Time Series Analysis* **3**: 969 - 983.

[59] Penm, J.H.W., Terrell, R.D., (1984). Multivariate subset autoregressive modeling with zero constraints for detecting causality. *Journal of Econometrics*, **24**: 311 - 330.

[60] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2009). SpAM: Sparse additive models. *Journal of the Royal Statistical Society,* Series B , **71**(5), 1009-1030.

[61] Reinsel, G., Velu, R. (1998). Multivariate reduced-rank regression: Theory and applications, *New York: Springer*

[62] Quandt, R. and Ramsey, J. (1978). Estimating mixtures of normal distributions and switching regression. *J. Amer. Statist. Assoc.* **73**, 730–738.

[63] Roecker,E. (1991). Predicition error and its estimation for subset selected models, *Technometrics*, **33** (4): 459-468.

[64] Rothman, A., Levina, E., Zhu, J., (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19**(4): 947-962.

[65] Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.

[66] Sims C.A. (1980). Macroeconomics and reality. *Econometrica*. **48**(1): 1-48.

[67] Song S., Bickel P. (2011). Large vector autoregressions, arXiv:1106.3915v1.

[68] N. Städler, N., Bühlmann, P. — and van de Geer, S. (2010). ℓ_1-penalization for mixture of regression models. *Test*, **19**, 209–256.

[69] Stephens, M. (2000). Dealing with label switching in mixture models. *J. Roy. Statist. Soc. Ser. B*, **62**, 795-809.

[70] Stone, C. J. (1977) Consistent nonparametric regression. With discussion and a reply by the author. *Ann. Statist.* **5**, 595-645.

[71] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society*, series B, **58**(1): 267-288.

[72] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U.S.A.* **99**, 6567 - 6572.

[73] Toshiya, H. (2013). Mixture regression for observational data, with application to functional regression models. URL `http://arxiv.org/abs/1307.0170`.

[74] Tsybakov, A. B. (2009) *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York.

[75] Turlach B.A., Venables, W.N., and Wright, S.J. (2005). Simultaneous variable selecion. *Technometrics*, **47**(3): 349-363.

[76] Turner, R. (2000). Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Applied Statistics*. **49**, 371–384.

[77] Turner, R. (2011). Mixreg: Functions to fit mixtures of regressions. `http://CRAN.R-project.org/package=mixreg. R package version 0.0-4`.

[78] Vandekerkhove, P. (2013). Estimation of a semiparametric mixture of regressions model. *J. Nonparam. Statist.*, 25, 181-208.

[79] Wille, Anja, et al. Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol* 5.11 (2004): R92.

[80] Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *J. Amer. Statist. Assoc.*, **104**, 758-767.

[81] Yin, J., Chen X. and Xing P. E., (2012). Group Sparse Additive Models. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 871-878

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

[83] Yuan,M., Ekici, A., Lu, Z., and Monteiro, R., (2007). Dimension reduction and coeffcient estimation in multivariate linear regression. *Journal of Royal Statistical Society*, series B, **69**(3): 329-346.

[84] Zhang, H. H., Liu, Y., Wu, Y. and Zhu, J. (2008). Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electronic Journal of Statistics* **2**, 149 - 1167.

[85] Zhu, H. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *J. Roy. Statist. Soc. Ser. B*, **66**, 3–16.