Sparse functional regression models: minimax rates and contamination

Wei Xiong

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

ABSTRACT

Sparse functional regression models: minimax rates and contamination

Wei Xiong

In functional linear regression and functional generalized linear regression models, the effect of the predictor function is usually assumed to be spread across the index space. In this dissertation we consider the sparse functional linear model and the sparse functional generalized linear models (GLM), where the impact of the predictor process on the response is only via its value at one point in the index space, defined as the sensitive point. We are particularly interested in estimating the sensitive point. The minimax rate of convergence for estimating the parameters in sparse functional linear regression is derived. It is shown that the optimal rate for estimating the sensitive point depends on the roughness of the predictor function, which is quantified by a "generalized Hurst exponent". The least squares estimator (LSE) is shown to attain the optimal rate. Also, a lower bound is given on the minimax risk of estimating the parameters in sparse functional GLM, which also depends on the generalized Hurst exponent of the predictor process. The order of the minimax lower bound is the same as that of the weak convergence rate of the maximum likelihood estimator (MLE), given that the functional predictor behaves like a Brownian motion.

Another problem we consider in this dissertation is a contaminated sparse functional generalized model, where the sensitive point is prone to subject-specific random contaminations that are likely to occur in applications. A numerical approach to estimating the sensitive point in this setting was proposed based on the Monte Carlo expectation-maximization (MCEM) algorithm. It is shown that when contaminations are present, the rate for estimating the sensitive point is reduced to the parametric rate from the faster rate achieved by the MLE in the contamination-free scenario.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to Dr. Ian McKeague, who is my dissertation advisor and a great professor, for his huge help and support on the dissertation throughout my time at Columbia University. I could not have succeeded without his generous assistance and supervision. I thank Dr. Martin A. Lindquist for his collaborative work with Dr. McKeague that motivated my dissertation topic and for offering the data set and valuable advice on my research. I also thank Dr. R. Todd Ogden, Dr. Melanie M. Wall, Dr. Min Qian and Dr. Yang Feng for generously serving on my dissertation committee.

To my family

Chapter 1

Introduction

The term "functional data" was coined by Ramsay and Dalzell (1991) for curve and image data. With the help of advances in technology, especially in accurate instruments, it is possible for scientists to record measurements in an almost continuous fashion. As a result, functional data commonly arise in a wide variety of applied contexts in chemometrics, physical science, and biomedical studies, as described in Ramsay and Silverman (2002). One example would be the angles in the sagittal plane formed by the hip and by the knee as children go through a gait cycle (Olshen et al., 1989). Such change in the data collection technique gives rise to new statistical challenges. At first glance, it seems natural to consider the analysis of functional data as a multivariate problem. However, due to the large number of variables recorded, one is faced with "the curse of high-dimensionality" (Bellman, 1957) that causes intensive computation and numerical instability. Furthermore, treating functional data as multivariate vectors ignores the correlations between adjacent measurements, especially if they are recorded in a temporal order. Traditionally, the field of longitudinal data and correlated data analysis deals with such situation. However, functional data tend to be much more densely measured than common longitudinal data and it could be difficult to apply typical methodologies employed in longitudinal data analysis, e.g. generalized estimating equations, to extract meaningful information from functional data. Time series analysis usually handles measurements that are recorded closely in time. But it generally requires certain distributional assumptions, such as second moment stationarity, that may not be satisfied by functional data.

Consequently, unique functional data analysis techniques have been developed by treating a functional datum as an element in the continuous functional space, instead of a random vector in a finite dimensional space. There has been a vast literature on functional data analysis in which a wide rage of methodologies have been proposed. Some statisticians treat functional data as realizations of smooth random functions, and the observed data on discrete points are interpolated (if no measurement error is assumed) or smoothed (if measurement error is present) by various techniques, such as cubic splines, smoothing splines and kernel smoothers. In other situations, functional data are considered as realizations of random processes which do not have to be smooth. Time series data that are observed on a densely grid of time points can be one example. Ramsay and Silverman (2002) described in extensive detail about smoothing functional data, functional principal components analysis, and different models involving functional data.

Among these techniques, functional linear regression (FLR) and generalized functional linear models have received considerable attention due to their simplicity and interpretability. Based on the type of dependent variables and independent variables, functional linear models can be classified into several sub-classes: a functional response and a scalar independent variable; a scalar response and a functional independent variable; a functional response and a functional independent variable. The problem of estimating the slope function when the response is scalar and the independent variable is functional has received particular attention. The magnitude of the slope function indicates the amount of impact the functional predictor has on the response. Therefore it is meaningful to develop estimators that are both accurate and interpretable. The slope function estimation problem for functional generalized linear models have also been discussed by several authors, which is generally an extension of the estimation theory for functional linear models. Estimators that achieve optimal rates have been developed based on functional PCA techniques.

Most of the functional linear regression and functional generalized linear models literature assume that the impact of the predictor process is spread across the index space. In certain applications this assumption might not hold, as we will see in the examples presented below. Meanwhile, a large part of the literature assume smooth functional data and apply certain types of smoothing techniques to the observed process. The extent of

smoothing is determined by the smoothing parameter, which is typically chosen to optimize the predictive power of the model rather than for the sake of estimation. In some cases, the impact of the functional predictor is sparse in the index space and it is of main interest to estimate the sensitive points at which the values of the functional independent variable are associated with the response. In this setting, it may be improper to smooth the functional data as we may lose essential information on the loci of the sensitive points. Lindquist and McKeague (2009) and McKeague and Sen (2010) proposed sparse functional linear regression and sparse functional generalized linear models, assuming the functional predictor is a (fractional) Brownian-like process, and developed a least squares estimator and a maximum likelihood estimator for the two models, respectively. They showed that these estimators are consistent and fast-converging. However, it was not clear whether they are rate-optimal.

Optimal rates of convergence have been an important topic in estimation theories. In particular, the problem of minimax estimation has been studied extensively in the past, in that it offers a reasonable criterion to assess the optimality of estimators. From a theoretical point of view, minimax estimators achieve optimality uniformly across the parameter space and thus avoid the illusion of super-efficiency. Many maximum likelihood estimators are shown to be minimax. From an application perspective, the minimax criterion selects estimators that are both accurate and robust. A large part of this dissertation is dedicated to establishing the minimax rates for estimating the parameters in the sparse functional linear regression model and the sparse functional GLM, and showing that the LSE and the MLE for these two models, respectively, could be rate-optimal.

A complication posed by practical issues is the contamination of the sensitive point by random errors. Such concern is raised by fMRI studies presented below. We propose the contaminated sparse functional GLM to accommodate this situation and construct a Monte-Carlo EM based procedure to estimate the model parameters. Such procedure bypass the complicated form of the predictor trajectories and the prohibitive computational cost of direct optimization. It is shown in several simulation studies and a real life data analysis that the proposed procedure generates desirable results.

1.1 Motivating examples

Various studies in the biomedical field involve the analysis of functional data. Some of them are intrinsically smooth curve data. Others are more similar to stochastic processes. We are mainly motivated by the latter and will present two examples below. One arises from chromosome-wide gene expression profiles. The other example comes from an fMRI study that investigates brain responses to anxiety (Lindquist et al., 2007; Lindquist, 2008).

1.1.1 Gene expression

Our first motivation arises from gene expression studies that measure the activity of numerous genes simultaneously. In these studies, it is of interest to identify genes whose expression behavior is correlated with clinical outcomes. For example, Emilsson et al. (2008) studied gene expression levels at over 24,000 loci in samples of adipose tissue to detect genes associated with body mass index and other obesity-related outcomes. Gruvberger-Saal (2004) used gene expression profiles from the specimen of breast cancer tumors to predict estrogen receptor protein concentration, an important prognostic marker for breast tumors; see also Buness et al. (2007).

In the gene expression literature, statistical methods have been proposed to detect differentially-expressed genes. Most of them are based on multiple testing procedures, which may ignore the correlation structure of the predictor process. See, for example, Dudoit and van der Laan (2008) and Salas-Gonzalez et al. (2009). We might also use the gene expression profile across an entire chromosome as functional predictors, and the scalar clinical outcomes as response variables. Lindquist and McKeague (2009) and McKeague and Sen (2010) proposed the sparse functional regression models, where the functional predictor is associated with the response only through its value at one point in the index space, called the sensitive point. Under this setting, the effective genes can be viewed as sensitive points, and sparse functional regression models can serve as an ideal tool for estimating the location of the influential gene

It is known that gene expression profiles display fractal behavior, i.e. self-similarity over a range of scales. In fact, fractals often arise when spatiotemporal patterns at higher levels

emerge from localized interactions and selection processes acting at lower levels, as is the case of gene expression activity. Moreover, recent discovery (Lieberman-Aiden et al., 2009) shows that chromosomes are folded as "fractal globules" which can easily unfold during gene activation, which further helps explain the fractal behavior of gene expression profiles. The fractal behavior implies nice properties of the least squares estimator for the sparse functional linear model, as pointed out by McKeague and Sen (2010).

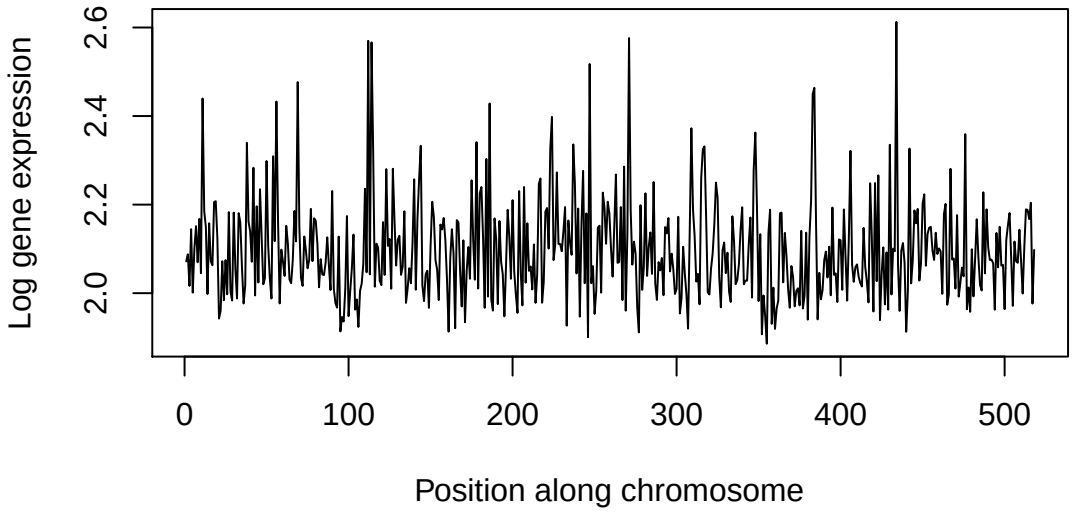


Figure 1.1: Log gene expression at 518 loci along chromosome 17 in tissue from a breast cancer patient.

1.1.2 fMRI study

Functional data frequently arise in brain imaging studies (Tian, 2010; Aston et al., 2006). Modern functional brain imaging techniques, such as PET (positron emission tomography), fMRI (functional magnetic resonance imaging), EEG (electro-encephalography) and MEG (magneto-encephalography), have been used to measure different aspects of brain activity at discrete time points during an experiment using different principles. These measurements,

called time courses, can be treated as functions of time. Functional data analysis has been applied to brain image data for dimension reduction (or feature extraction), spatial classification in fMRI studies, and the inverse problem in MEG studies.

It is of special interest to estimate the timing of psychological activity onset. In the fMRI context, multi-subject change-point estimation has been employed to estimate the onset times of brain activity (Lindquist et al., 2007; Robinson et al., 2010). But this technique only makes use of the information contained in the fMRI time courses and does not exploit the association between the time courses and the clinical outcomes. Sparse functional regression models are better suited for the estimation to make use of the information contained in both the clinical outcomes and the time courses, which can be viewed as functional predictors and the onset time as the sensitive point.

The data set we will use in this dissertation was described in Lindquist et al. (2007). 25 participants were scanned with BOLD fMRI at 3 T (GE, Milwaukee, WI). They were classified as resilient or non-resilient according to a written test with 13 being resilient and 12 non-resilient. Each of them performed a 7-minute anxiety-provoking speech preparation task. The design was an off-on-off design, with the anxiety-provoking period occurring between lower-anxiety resting periods. Participants were informed that they were to be given 2 minutes to prepare a 7-minute speech, topic of which would be revealed to them during scanning. After the start of fMRI acquisition, there was 2 minutes of resting baseline. At the end of this period, subjects viewed an instruction slide for 15 seconds that described the speech topic. After 2 minutes of silent preparation, another instruction screen appeared for 15 seconds that informed subjects that they would not have to give the speech. An additional 2-minute period of resting baseline followed, completing the functional run. Images were acquired every 2 seconds throughout the course of the run.

A series of 215 fMRI images were acquired during the 7-minute speech preparation task. The brain activity may differ from baseline in a short period of time in response to a stimulus. This onset time point of brain activity is of particular interest since the signal intensities at this point are mostly associated with the clinical outcomes. When the onset time is unknown, one may consider using the entire time course on a voxel as a functional datum to predict anxiety levels. We present in Figure 1.3 the trajectories of the image

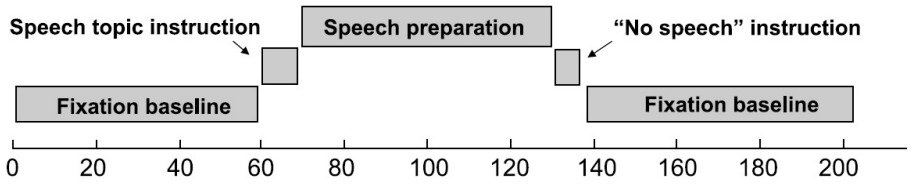


Figure 1.2: A schematic of the experimental task design for the fMRI study, from Lindquist et al. (2007). The design was an off-on-off design, with the anxiety-provoking period occurring between lower-anxiety resting periods.

signals from the ventromedial prefrontal cortex, a region known to be related to anxiety, for a resilient and a non-resilient participant.

We can show that the second moment of the increment of the fMRI time courses has an exponential rate of decay when the increment in time shrinks to zero. This behavior will be characterized by the generalized Hurst exponent (GHE) defined in (3.2.1). Here, we calculate the squared differences of the signal processes, for each lag from 1 to 100, of the 25 participants and take the average to approximate the mean squared increments in (3.2.1). The average squared increments are plotted against the size of the lags, both in log scale, in Figure 1.4. Half the slope of the fitted line, $H = 0.198$, can serve as a crude estimate of the generalized Hurst exponent. The reader is referred to Qian (2004) and Feder (1988) for more standard estimation methods.

1.2 New questions raised about sparse functional regression models

The first question we consider in this dissertation is: what is the optimal way of such estimation using the sparse functional regression models? Are there better estimators than the MLE and the LSE, proposed in Lindquist and McKeague (2009) and McKeague and Sen (2010), respectively? It would be of interest to find out the optimal rates for such estimation. Also the functional predictor X is assume to be a (fractional) Brownian motion,

at least in the neighborhood of the true sensitive point. This condition is difficult to verify in application.

The second question is: how do we deal with the situation where the sensitive point might be contaminated by subject-specific errors? D'Esposito et al. (2003) showed that fMRI signals may be affected by aging, pathology and other disorders. the sparse functional regression models assume a fixed sensitive point. We might want to have a model that accounts for random sensitive points caused by such contamination.

1.3 Our contributions

In the first part of this dissertation, we will address the problem of the optimal rates for estimating the parameters in the sparse functional linear model. We will relax the conditions on X and derive the minimax rates under the milder conditions. Specifically, we assume the existence of a "generalized Hurst exponent" of X that characterizes its local behaviors. Intuitively, it requires the second moment of an increment of X to converge to zero at an exponential rate as the increment vanishes. If X has a generalized Hurst exponent $H \in (0, 1]$, and other mild conditions are satisfied, we will show that the least squares estimator $\hat{\eta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n)$ is the minimax estimator in the mean squared error sense, with component-wise rates $n^{1/2}$, $n^{1/2}$, and $n^{1/(2H)}$ respectively. Since H can be viewed as a measure of the roughness of X, the convergence rate for estimating θ is determined by the smoothness of the predictor process.

The second part of the dissertation discusses the optimal rates of convergence for estimating the parameters in the sparse functional generalized linear models. By assuming again the existence of a generalized Hurst exponent, we establish a lower bound on the minimax risk. We will show that this lower bound is of the same order as that of the weak convergence rate of the maximum likelihood estimator established in Lindquist and McKeague (2009) under the assumption that X behaves like a two-sided Brownian motion.

The third part of the dissertation proposes an extension of the sparse functional GLM to incorporate random contaminations of sensitive points. We propose a Monte Carlo EM algorithm that computes the maximum likelihood estimator of the mode of the contaminated

sensitive point's pdf as well as the regression coefficients. The numerical properties of the proposed algorithm are tested in several simulations. It is shown that the convergence rate of the mode estimator is the parametric rate $n^{1/2}$, when the contamination is present with a smooth density, in contrast to the faster rate $n^{1/(2H)}$ in the non-contamination setting.

1.4 Structure of this dissertation

Chapter 2 will give a brief review on the functional data analysis literature, especially on functional linear regression and functional generalized linear regression. We will then introduce the definition of minimaxity and discuss the common methods and recent work in the field of minimax estimation. In Chapter 3, we establish the minimax rates for estimating the parameters in sparse functional linear regression. Sections 3.4 and 3.5 present the main results on the lower and upper bounds for the minimax risk, respectively. Detailed proof of the lemmas and theorems is given in Section 3.6. In Chapter 4, we establish a lower bound on the minimax risk of estimating the parameters in sparse functional *generalized* linear regression. It is shown in Section 4.3 that the lower bound is of the same asymptotic order as that of the lower bound for minimax risk estimating the parameters in sparse functional linear model. Detailed proof is given in Section 4.4. In Chapter 5, we give the results of five simulation studies. The first two studies evaluate the performance of the LSE for sparse functional linear regression and compare its mean squared error (MSE) to that of the lasso and the FLR estimator. The third and fourth studies perform the similar procedures on the MLE for spares functional GLM. The last simulation test the performance of the sparse functional linear model as a working model when the data are generated from a FLR model with a spike-shaped regression function. In Chapter 6, we consider the situation where the sensitive point in the sparse functional GLM model is contaminated by random subject-specific errors. A computational solution to estimating the parameters is derived based on EM algorithm and Monte Carlo approximations. Section 6.4 describes the details of the estimation procedure. The asymptotic properties of the proposed estimator are given in Section 6.3. We present the results of four simulation studies in Chapter 7. Furthermore an application to the fMRI data is presented in Chapter 8. In Chapter 9 we conclude the

important findings in previous chapters and discuss directions for future research.

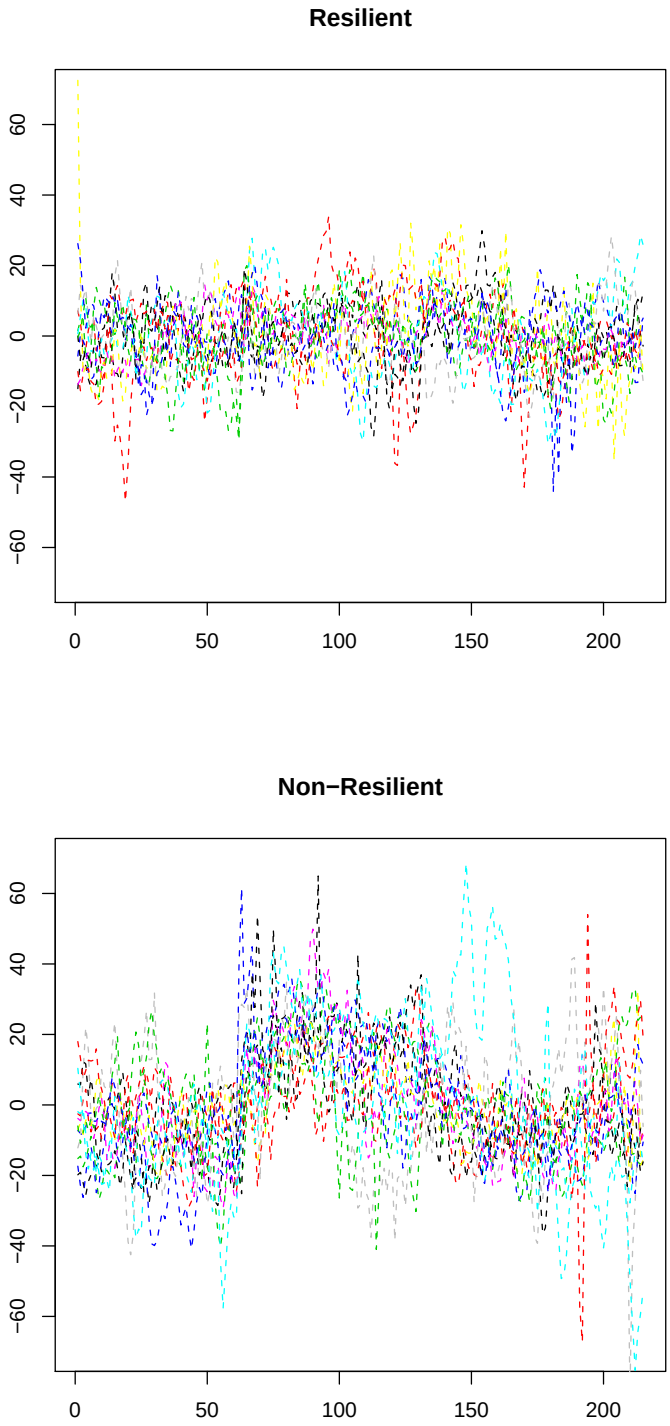


Figure 1.3: The fMRI signal over the ventromedial prefrontal cortex in reaction to an anxiety-provoking taske for resilient and non-resilient subjects.

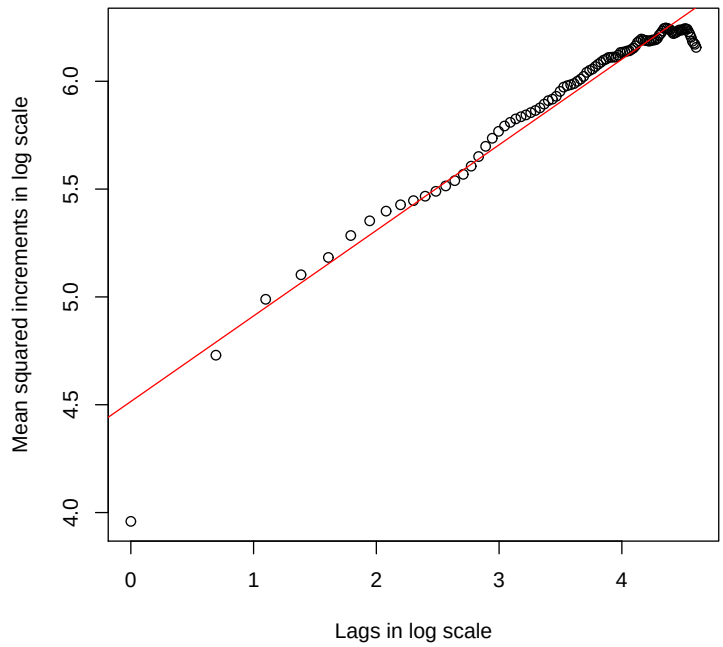


Figure 1.4: Average squared increments of fMRI time courses against time lags, both in logarithm scale. The red line is the fitted linear regression line.

Chapter 2

Literature review

2.1 Functional Data Analysis

2.1.1 Smoothing and Regularization

Conceptually, functional data are thought of as sample paths of a continuous-time stochastic process. A graphic illustration would be a collection of curves over the parameter space of the stochastic process. Although the observed trajectories are often rough and fluctuating, in many applications of functional data analysis, there is scientific reason to believe that the true trajectory is a smooth function and is observed with random errors.

In practice, almost all measurements of continuous-time processes are made on discrete grid of the parameter space. A sample of functional data is typically denoted as $(t_{ij}, y_{ij}), j = 1, \ldots, n_i$, where t_{ij} is usually a time point but can also represent spatial location or other parameter space index. To recover the underlying smooth function, various smoothing techniques have been employed. There is an extensive literature on nonparametric smoothing; see for example Eubank (1988); Fan and Gijbels (1996); Ruppert and Wand (1994). In the functional data analysis context, the smooth function is usually represented by a linear combination of basis functions. Choices of basis systems include Fourier bases for periodic data, B-spline bases for non-periodic data and wavelet bases where derivatives are not required Ramsay and Silverman (2002).

When a series of basis functions is selected, the functional data are fitted to the bases

according to a certain criterion such as the least squares loss

$$\text{SMSSE}(\mathbf{y}|\mathbf{c}) = \sum_{j=1}^{n}[y_j - \sum_{k}^{K} c_k\phi_k(t_j)]^2.$$

where $\phi_k, k = 1, \ldots, K$ are the basis functions. A matrix expression is $\text{SMSSE}(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c})$). It is often the case that the observations on different time points are not independent and the above model does not apply. To account for correlation between measurements, the regression is carried out by weighted least squares $\text{SMSSE}(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{y} - \mathbf{\Phi c})$).

Since any continuous function can be arbitrarily approximated by sufficiently many bases, the number of the bases used is a smoothing parameter controlling the roughness of the fitted curve. However, this control is discontinuous as we can only tune the degree of smoothing by adding or removing one basis term. A more powerful option for smoothing discrete functional data would be *the roughness penalty*. The quantity of a function's roughness is typically measured by its integrated squared second derivative

$$\text{PEN}_2(x) = \int [D^2x(s)]^2 ds,$$

while more general roughness penalties are proposed in Ramsay and Silverman (2002) by means of an mth order differential operator L. The *penalized* residual sum of squares will then be

$$\text{PENSSE}_\lambda(x|\mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]'\mathbf{W}[\mathbf{y} - x(\mathbf{t})] + \lambda \times \text{PEN}_2(x),$$

where x is the fitted curve. de Boor (2001) proved that $\text{PENSSE}_\lambda(x|\mathbf{y})$ is minimized by a cubic spline with knots at the data points t_j. Thus we can choose to expand $x(t)$ with respect to a spline basis

$$x(t) = \sum_{k}^{K} c_k\phi_k(t) = \mathbf{c}'\phi(\mathbf{t}).$$

The roughness penalty term can be expressed in matrix form in terms of the inner products of the derivatives of the basis functions. The estimated coefficient vector $\mathbf{c}$ is then computed by matrix algebra. The tuning parameter is generally chosen by the cross-validation (CV) or generalized cross-validation (GCV) method.

2.1.2 Functional Principal Component Analysis

After smoothing the raw functional data, some preliminary steps of registering and displaying the curves are usually taken. To further explore their variational patterns and better understand the the variance-covariance structure, the principal component analysis (PCA) from classical multivariate statistics is extended to the case of random functions.

There are many ways to define functional principal components. The most common one is via the Karhunen–Loéve decomposition of a random function. Suppose X is a square-integrable random function defined on an interval $\mathcal{I}$. Let $\eta = E(X)$, the mean function of X. The variance-covariance function of X is a bivariate function $K(u,v) = E[\{X(u) - \eta(u)\}\{X(v) - \eta(v)\}]$, which can be viewed as an operator on the space of square-integrable functions from $\mathcal{I}$ to the real line: if $\psi \in L^2(\mathcal{I})$, then $K\psi(u) = \int_{\mathcal{I}} K(u,v)\psi(v)dv$.

Similar to the variance-covariance matrix in multivariate statistics, the variance-covariance function can be decomposed into its eigenvalues and eigenfunctions

$$K(u,v) = \sum_{j=1}^{\infty} \theta_j \psi_j(u)\psi_j(v),$$

where θ_j and ψ_j are obtained by solving the equation

$$K\psi(u) = \theta\psi(u).$$

Then the Karhunen–Loéve expansion of X is given by

$$X(u) = \eta(u) + \sum_{j=1}^{\infty} \xi_j \psi_j(u),$$

where the random coefficients $\xi_1, \xi_2, \ldots$ are defined as $\xi_j = \int_{\mathcal{I}}(X - \eta)\psi_j$. They have zero means and are uncorrelated. Their variances are given by $\theta_j = E(\xi_j^2)$.

Given the smoothed observed functional data $\mathcal{X}$, the functional principal components are obtained via the empirical variance-covariance function

$$\hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}[X_i(u) - \bar{X}(u)][X_i(v) - \bar{X}(v)],$$

where $\bar{X} = n^{-1}\sum_i X_i$. The approximate eigenvalues and eigenfunctions satisfy

$$\hat{K}(u,v) = \sum_{j=1}^{\infty} \hat{\theta}_j \hat{\psi}_j(u)\hat{\psi}_j(v),$$

where $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \cdots \geq 0$. So the eigenfunctions are sorted in a descending order according to the percentage of variability in the data $\mathcal{X}$ they can explain. The signs of the approximate eigenfunctions are chosen so that $\int_{\mathcal{I}} \psi_j \hat{\psi}_j \geq 0$.

By discretizing the empirical variance-covariance function and solving the eigen-equations, we obtain approximate eigenvalues and discrete approximate eigenfunctions. Any convenient interpolation method can then be employed to obtain the continuous approximate eigenfunctions. The eigenfunctions $\hat{\psi}_1, \hat{\psi}_2, \ldots$ form a complete orthonormal basis of the square-integrable function space. Therefore, for any give function $b \in L^2(\mathcal{I})$ we can have the expansions

$$X_i - \eta_i = \sum_{j=1}^{\infty} \xi_{ij} \hat{\psi}_j, \tag{2.1.1}$$

$$b = \sum_{j=1}^{\infty} b_j \hat{\psi}_j \tag{2.1.2}$$

where $\xi_{i1}, \xi_{i2}, \ldots$ and $b_1, b_2 \ldots$ are random functions of the data $\mathcal{X}$ and thus random variables. When we truncate the expansions to obtain a lower dimensional approximation of the functional data, it is assured by the order of the order of the principal components that the majority of variations in X_i, hence most of the information contained in the data, is preserved in the truncated series.

Functional principal component analysis (FPCA) proves a powerful tool to understanding the features of curve data and has become an important part of functional data analysis. Studies of FPCA include Rice and Silverman (1991), Silverman (1996), Cardot (2000), James et al. (2000), Hall and Hosseini-Nasab (2006) and Peng and Paul (2009). Yao et al. (2005) applied FPCA to longitudinal data analysis. Aguilera et al. (1999a) and Aguilera et al. (1999b) used a weighted FPCA to forecast a continuous time series. Kneip and Utikal (2001) explored testing differences in a set of density function curves using FPCA. Viviani et al. (2005) used FPCA to analyze fMRI images of human brain areas scanned along time. We will talk more about the application of FPCA to functional linear regression and functional generalized linear regression in subsequent sections.

2.1.3 Functional Linear Regression

Having explored the variability of a functional variable, we want to further investigate how its variation explains, or is explained by, variations of other variables. The classical linear model is the first to be extended to the functional context. Here we only consider the case with functional predictors and scalar response, i.e. we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the X_i's are a random sample of a stochastic process X defined on a compact interval $\mathcal{I}$, and the Y_i's are copies of a random variable Y that satisfies the functional linear regression model,

$$Y = a + \int_{\mathcal{I}} b\ X + \epsilon. \tag{2.1.3}$$

Here, a is a constant scalar intercept in the linear model, b is the slope function which belongs to $L^2(\mathcal{I})$, and the error ϵ is also a scalar. Model (2.1.3) has wide applications to various practical problems. The main interest usually focuses on estimating b. Since b is a function rather than a scalar, the knowledge of where it takes large or small values can be very helpful for understanding how the functional explanatory variable interplays with the outcome variable.

Like obtaining random functions from observed functional data, estimation of the slope function b is an infinite-dimensional problem. Without any constraint on b, we could choose $\hat{a}$ and $\hat{b}$ to reduce the residual sum of squares to zero and perfectly predict the response variable. The resulting slope function would be very ragged and hard to interpret. Hence certain smoothing is needed to regulate the slope estimator. Like the ones used to smooth X_i, the regularization methods usually consist of the truncate basis approach and the roughness penalty approach.

Another justification of regularization is by viewing the estimation of b as an ill-posed inverse problem. We can write (2.1.3) as

$$Y_i - \mu = \int_{\mathcal{I}} b\ (X_i - x) + \epsilon_i,$$

where $x = E(X_i)$ and $\mu = E(Y_i) = a + \int bx$. Then if we denote $g(u) = \mathbf{E}[\{Y(u) - \mu(u)\}\{X(u) - x(u)\}]$, it follows from Fubini's theorem that

$$Kb = g. \tag{2.1.4}$$

Solving the normal equation (2.1.4) involves the inversion of operator K. But since it is a compact linear operator on the infinite dimensional space $L^2(\mathcal{I})$, K does not have a bounded inverse. This calls for further constraint on b.

The regularization method, like the one used to smooth X_i, generally takes one of two possible forms. The first method expand b in a series of basis functions and then truncate the expansion at the pth term, where p is chosen large enough the capture the features in b but small enough to avoid overfitting. Then the expansion coefficients can be estimated via ordinary least squares. Typically, the functional predictors and the slope function are expressed with respect to the functional principal components, as in (2.1.1) and (2.1.2). The second method uses the roughness penalty on the least squares loss function to shrink the variability in b. A common choice of roughness penalty is the integrated squared derivative $\int b^{(m)}(t)^2 dt$ and usually $m = 2$. Then the estimate of b is the function that minimizes

$$\sum_{i=1}^{n}(Y_i - a - \int b(t)\ X_i(t)dt)^2 + \lambda \int b^{(m)}(t)^2 dt$$

for a chosen $\lambda > 0$.

Hall and Horowitz (2007) considered the principal component method in functional linear models in detail and gave the minimax convergence rates of estimators for the slope function. They showed that the minimax rate of convergence for estimating the slope function $\beta(\cdot)$ in terms of the mean integrated squared error is determined by the smoothness of both β and the covariance kernel of the predictor process. We will give a more detailed review of their work in Section 2.3.

Crambes et al. (2009) took the roughness penalty method and considered smoothing splines estimators for model (2.1.3). They assume X_i are observed at p equidistant points $t_1, \ldots, t_p \in \mathcal{I}$. Then the estimator for b is determined by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \bar{Y} - \frac{1}{p}\sum_{j=1}^{p} b(t_j)(X_i(t_j) - \bar{X}(t_j))\right]^2 + \rho\left[\frac{1}{p}\sum_{j=1}^{p}\pi_b^2(t_j) + \int_0^1 (b^{(m)}(t))^2\, dt\right],$$

where $\pi_b(t)$ is the minimizer of $\sum_{j=1}^{p}(b(t_j) - \pi_b(t_j))^2$ among all polynomials with degree $m-1$, i.e. $\pi_b(\mathbf{t}) = \mathbf{P}_m b(\mathbf{t}), \mathbf{t} = (t_1, \ldots, t_p)'$, where $\mathbf{P}_m$ is the projection matrix projecting into the m-dimensional linear space of all (discretized) polynomials of degree $m-1$.

The solution $\hat{b}$ is a natural spline of order $2m$ with knots at $t_1, \ldots, t_p$. The extra term $\frac{1}{p}\sum_{j=1}^{p} \pi_b^2(t_j)$ ensures the existence of a unique solution by adding the nonsingular projection matrix $\mathbf{P}_m$ into the ridge-regression-type estimator of the spline coefficients. The authors gave a closed form of the solution and considered the convergence rates of the estimators with respect to L^2 semi-norms induced by the covariance operator K, $\|u\|_K^2 = \langle Ku, u\rangle$ with $\langle u, v\rangle = \int_{\mathcal{I}} u(t)v(t)\,dt$. They derived optimal rates of convergence in the sense that the smoothing spline estimator is minimax with convergence rates

$$\|\hat{b} - b\|_{K_{n,p}}^2 = O_p(n^{-(2m+2q+1)/(2m+2q+2)}),$$

assuming b is m-times continuously differentiable and general conditions on X. The value of q quantifies the rate of decrease $\sum_{j=k+1}^{\infty} \theta_j = O(k^{-2q})$ and $\|\cdot\|_{K_{n,p}}$ is the discretized empirical semi-norm

$$\|\mathbf{u}\|_{K_{n,p}}^2 \triangleq \frac{1}{p}\mathbf{u}^\tau \left(\frac{1}{np}\mathbf{X}^\tau\mathbf{X}\right)\mathbf{u}.$$

The purpose of using the L^2 semi-norms induced by K is to focus on the convergence rates of the prediction error rather than estimation error. Cai and Hall (2006) investigated the rates of convergence on the error $a + \langle b, x\rangle - \hat{a} - \langle \hat{b}, x\rangle$ where x is a fixed function. For a random function X_{n+1}, Crambes et al. (2009) showed the rates of convergence on the prediction error is determined by $\|\hat{b} - b\|_K^2$:

$$\mathbf{E}\left[\left(\hat{a} + \int_{\mathcal{I}} \hat{b}(t)X_{n+1}(t)\,dt - a - \int_{\mathcal{I}} b(t)X_{n+1}(t)\,dt\right)^2 \middle| \hat{a}, \hat{b}\right] = \|\hat{b} - b\|_K^2 + O_p(n^{-1}).$$

The convergence of $\hat{b}$ with respect to $\|\cdot\|_K^2$ is very different from the convergence under the usual L^2-norm $\|\cdot\|^2$. In fact, under the general conditions on X in Crambes et al. (2009), it can only be shown that $\|\hat{b} - b\|^2$ is bounded in probability. Additional conditions, such as (2.3.4) – (2.3.6), have to be assumed to derive stronger results on $\|\hat{b} - b\|^2$.

Apart from estimation and prediction, the interpretability of the slope function b is another challenging issue in functional linear models. The magnitude of the absolute values of b only provide a vague and qualitative sense of how the functional predictor can influence the response. To this end, James et al. (2009) proposed a method called "Functional Linear Regression That's Interpretable" (FLiRTI). They divide the interval $\mathcal{I}$ into a fine grid of points and assumed one or more of the slope function's derivatives are sparse, i.e. $b^{(d)}(t) = 0$

over large regions of t for one or more values of $d = 0, 1, 2, \ldots$. They then used variable selection methods such as the LASSO and Dantzig selector to fit the model. By choosing d appropriately, FLiRTI is flexible enough to deal with a large range of situations and produce interpretable estimates of b.

2.1.4 Functional Generalized Linear Models

While functional linear models are widely applicable, they may be too restrictive for situations where Y_i are non-Gaussian. In the same spirit as in classical multivariate analysis, James (2002) gave a functional analogy to generalized linear models (GLM) described in McCullagh and Nelder (1989). He assumed the distribution of Y belongs to the exponential family. The linear predictor in GLM is replaced by an integral in the functional generalized linear model:

$$g(\mu) = a + \int X(t)b(t)\,dt, \tag{2.1.5}$$

where $\mu = E(Y)$ and g is the link function. X was then expressed by natural cubic splines with random coefficients: $X(t) = \mathbf{s}(t)^\tau\gamma$, $\gamma \sim N(\mu_\gamma, \Gamma)$. Here $\mathbf{s}(t)$ represents the q-dimensional spline basis at time t, γ the q-dimensional spline coefficients for the predictor, and μ_γ and Γ the mean and variance of the γ's. The observed predictor function $x(t)$ was further assumed to be $x(t) = X(t) + e(t)$, where $e(t)$ is a zero-mean stationary Gaussian error process. Let $\mathbf{x}_i$ and $\mathbf{e}_i$ be the vectors of observations and measurement errors for individual i at its observation time points $t_{i1}, \ldots, t_{in_i}$ and let $S_i = (\mathbf{s}(t_{i1}), \ldots, \mathbf{s}(t_{in_i}))$ be the corresponding spline basis matrix, then the previous model can be written as

$$\begin{aligned} g(\mu_i) &= a + b_1\gamma_i, \qquad \gamma \sim N(\mu_\gamma, \Gamma) \\ \mathbf{x}_i &= S_i\gamma_i + \mathbf{e}_i, \qquad \mathbf{e}_i \sim N(0, \sigma_x^2 I) \end{aligned}$$

where $b_1 = \int b(t)\mathbf{s}(t)\,dt$. Since the spline coefficients can be viewed as missing data, one can employ the EM algorithm to estimate the model.

Müller and Stadtmüller (2005) proposed an alternative framework for generalized functional linear models. Their model is less an extension of the classical generalized linear model introduced by McCullagh and Nelder (1989) than and extension of the quasi-likelihood method of Wedderburn (1974) in that they do not assume the distribution of the response

need to be a member of the exponential family. They only assume the quasi-likelihood model

$$Y_i = g\left(a + \int b(t)X_i(t)dw(t)\right) + e_i, \qquad i = 1, \ldots, n,$$

where e is the random error with zero mean and a variance component structure depending on $\eta = a + \int b(t)X_i(t)dw(t)$. The authors expanded X and β with respect to an orthonormal basis of the function space $L^2(dw)$ in a similar way to (2.1.1) and (2.1.2), and truncate them at p. The truncated linear predictor η is then $\eta_i = a + \sum_{j=1}^{p} b_j\xi_{ij} = \sum_{j=0}^{p} b_j\xi_{ij}$, where $b_0 = a$, $\xi_{i0} = 1$, b_j are the basis coefficients of b for $j \geq 1$, and ξ_{ij} are the jth basis coefficients of X_i for $j \geq 1$. As p grows to infinity, this quantity approaches η with arbitrary accuracy. The estimating equation for $\hat{b}$ is then constructed as

$$U(b) = \sum_{i=1}^{n}(Y_i - \mu_i)g'(\eta_i)\xi_i/\sigma^2(\mu_i) = 0$$

where $\xi_i^T = (\xi_{i0}, \ldots, \xi_{ip})$ and $U(b)$ is the vector-valued score function. The equation is solved by iterated weighted least squares. The authors gave the asymptotic properties of the slope estimator with respect to the L^2-norm induced by the generalized autocovariance operator kernel

$$G(s,t) = \mathbf{E}\left(\frac{g'(\eta)^2}{\sigma^2(\mu)}X(s)X(t)\right).$$

Assuming some technical conditions on $p = p_n$ and the decay speed of the truncated tail, it was shown that

$$\frac{\int\int(\hat{b}(s) - b(s))(\hat{b}(t) - b(t))G(s,t)dw(s)dw(t) - (p_n + 1)}{\sqrt{2(p_n + 1)}} \rightarrow_d N(0,1) \qquad \text{as } n \rightarrow \infty.$$

Escabias et al. (2005) used a functional PCA approach to generalized linear models. They approximate the sample paths with a finite number of FPCA's and use the component scores as covariates in the logistic model. They address the issue of multicollinearity by conducting another PCA of the design matrix. Another important article is Cardot and Sarda (2006), who expressed only the slope function but not the predictor function in terms of B-splines. They then estimated the spline coefficients using penalized likelihood function with the usual penalty on the integrated squared derivative of the slope function. Asymptotic properties of the estimator were also considered under the covariance kernel induced L^2-norm.

Dou et al. (2010) extended the results of Hall and Horowitz (2007) using the principal component method in functional generalized linear models and gave the minimax risk of estimating the slope function. They showed that the minimax rate is again determined by the smoothness of both β and the covariance kernel of the predictor process. We will give a more detailed review of their work in Section 2.3.

James and Silverman (2005) made further extension by integrating generalized linear models, generalized additive models (GAM) and projection pursuit regression (PPR) into one procedure, "functional adaptive model estimation" (FAME), to handle functional predictors. The model is given by

$$g(\mu_i) = a + \sum_{k=1}^{r} f_k \left(\int X_i(t) b_k(t)\, dt \right).$$

Here both f_k and b_k are unknown functions and estimated in the fitting procedure and r is arbitrary. The model extends the standard projection pursuit regression by adding a link function to handle non-Gaussian or categorical response and replace the linear predictor with an integral of $X(t)b_k(t)$. This allows for a great deal of flexibility and is thus more general than GLM and GAM.

The fitting procedure is carried out by expanding X, b_k and f_k with respect to cubic splines and maximizing a penalized log likelihood via an iterative approach. Specifically, the algorithm starts with $r = 1$, fixing a and f_1, and fits a smooth b_1 using the penalty regularization method. b_1 is then fixed and a and f_1 are fitted by any GAM package. The procedure iterates between these steps until the penalized likelihood converges. Then f_1 and b_1 are fixed and the procedure is repeated for the $r = 2$ model subject to zero correlation between $\int X_i(t)b_1(t)\, dt$ and $\int X_i(t)b_2(t)\, dt$. The nested models grow until r reaches the preset maximum value. The authors proved that the estimates for the intercept and the expansion coefficients have $\sqrt{n}$-rates of convergence.

2.2 Sparse functional models

All the aforementioned methods assume that the influence of the functional predictors is spread over the entire time interval $\mathcal{I}$ or a continuous region of it. This might not be the case in some applications. Consider the gene expression case presented in Section 1.1.1,

for instance. Only a few genes are expected to be associated with the clinical outcome and thus the impact of the expression profile is sparse across the chromosome. Another example is the fMRI study mentioned in Section 1.1.2, where the main interest is to find a time interval that most clearly distinguishes between resilient and non-resilient individuals. D'Esposito, Deouell and Gazzaley (2003) showed there are scientific reasons to believe there are only a small number of time points in the fMRI time course, at which the brain activity is associated with the anxiety levels. In this case the impact of brain activity may not be captured by the integral used in tradition functional regression.

In fact, in fMRI studies, estimation of the precise timing of the underlying psychological activity is critical for many data analyses (Lindquist et al., 2007; Robinson et al., 2010) and is of main interest. Sometimes the onset time is assumed known *a priori*. However, in many areas of psychological inquiry, such as Examples include studies of drug uptake, emotional states or experiments with sustained stimulus, it is hard to specify this information in advance. In the work of Lindquist et al. (2007) and Lindquist et al. (2008), a Hierarchical exponentially weighted moving average (HEWMA) method was proposed to estimate the onset times of psychological activities. A drawback of this estimation procedure is that the change points were assumed to be fixed across subjects. In Robinson et al. (2010), the conditions were relaxed to assume that the change points for each subject are random, and a maximum likelihood procedure was developed for estimating the change points.

However, these methods are both based on multi-subject change point estimation approaches. For one thing, the sensitive point of interest is not necessarily a change-point in the random processes, e.g. cancer-related genes in the gene expression profiles. For another, when a scalar outcome is observed besides the functional data, we would like to exploit the association between the response variable Y and the functional predictor X and make use of this information to estimate the sensitive point, i.e. we want to estimate the point at which the value of X is mostly related to Y. The sparse functional regression models, proposed in Lindquist and McKeague (2009) and McKeague and Sen (2010), suit this problem very well. Below we review the sparse functional regression models that were proposed specifically to capture these point impacts.

2.2.1 Sparse functional linear regression

Motivated by the gene expression profile data, McKeague and Sen (2010) proposed a sparse functional linear model

$$Y = \alpha + \beta X(\theta) + \varepsilon. \tag{2.2.1}$$

to estimate the sensitive points in the index set of the predictor processes and give confidence intervals to the estimations. The intercept α and the slope β are both scalars and the sensitive point θ is the parameter of main interest. Y is a continuous response and $X = \{X(t), t \in [0,1]\}$ is a continuous stochastic process. ε is a mean-zero error term and is independent of X.

The authors proposed a least squares estimator (LSE) $\hat{\eta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n)$ defined by

$$(\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n) = \arg\min_{\alpha,\beta,\theta} \sum_{i=1}^{n} [Y_i - \alpha - \beta X_i(\theta)]^2. \tag{2.2.2}$$

McKeague and Sen (2010) made use of the fractal behavior of gene expression profiles reported in Lieberman-Aiden et al. (2009) by assuming the increments of X are locally *fractional* Brownian motion (fBm) with Hurst exponent $0 < H < 1$. Under additional conditions that θ_0 is an interior point of $[0,1]$, β_0 is nonzero, and the error ε has finite moment of order greater than 2, they showed that $\hat{\alpha}$ and $\hat{\beta}$ have the usual parametric convergence rate, $\sqrt{n}$, and $\hat{\theta}_n$ has a rate $n^{1/(2H)}$ of convergence. Confidence intervals for θ_0 were further constructed using parametric bootstrap, since the limiting distribution of $\hat{\theta}_n$ depends on the Hurst exponent H, which is unlikely to be known in practice. The authors also investigated the misspecified case where the data are generated partially or completely from a standard functional linear model.

2.2.2 Sparse functional generalized linear regression

To estimate the brain activity onset time in the fMRI study, Lindquist and McKeague (2009) proposed a point-impact functional logistic model

$$\text{logit}[P(Y = 1|X)] = \alpha + \beta X(\theta). \tag{2.2.3}$$

Here the Y is a binary outcome, representing the anxiety levels, $X = \{X(t), t \in [0,1]\}$ is a continuous stochastic process, representing the image signal process, and $\theta \in [0,1]$ is the

onset time point. Under the assumption that X is a two-sided Brownian motion around θ_0, the authors derived the asymptotic properties of the maximum likelihood estimators (MLE) of (α, β, θ) for three different cases, i.e. prospective sampling, retrospective sampling and generalized linear models. Specifically, they showed that the MLE $\hat{\alpha}_n$ and $\hat{\beta}_n$ are $\sqrt{n}$-rate and the MLE $\hat{\theta}_n$ has rate n. They further derived the explicit limiting distribution of the MLEs and constructed Wald-type confidence intervals.

To assess the performance of the sparse functional GLM, the authors compared it to the functional logistic regression model, which the authors refer to as the *functional-impact* (FI) model, and the LASSO with simulated and real life data. The sparse functional GLM was found to produce more accurate and interpretable results, while the functional logistic regression was shown to have a tendency to over-smooth the estimate of the regression function when there is a point-impact. The authors attribute this phenomenon to the roughness penalty on the regression function since the smoothing parameter is usually chosen by cross-validation in order to optimize the predictive performance of the model. The sparse functional GLM is also more interpretable than the lasso path diagram since it provides confidence intervals around sensitive time points selected by lasso.

2.3 Minimax estimation

In this dissertation we will focus on developing the optimal rates for estimating the parameters in sparse functional linear regression and sparse functional GLM with the minimax criterion. In this section, we will give a brief review on the literature of minimax problems, especially those in the functional data analysis field.

Rates of convergence have always been an essential topic in the asymptotic statistical literature. Different definitions have been proposed to describe estimators with a "best" rate of convergence. In many cases it is required that a best estimator not only achieves a best rate at a fixed model, but also at models close to one particular model of interest. In other words the estimator not only converges point-wisely to one model, but also converges uniformly at the same rate in a neighborhood of this model. This requirement is formalized as the definition of minimax estimators. Intuitively, this definition considers an estimator's

worst performance in a neighborhood of a particular parameter value. The reason to require a rate of convergence to hold uniformly lies in the fact that it excludes superefficient estimators, which take advantage of only point-wise limit behaviors. See Pollard (2010) for a more detailed introduction to minimax problems.

Consider a family of statistical models $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ defined on some fixed probability space $(\Omega, \mathcal{F})$. In the parametric estimation setting, the *Cramér-Rao* inequality gives a lower bound on the variance of any estimator of η under regularity conditions. An unbiased estimator in a regular model that achieves the Cramér-Rao lower bound is called *Fisher efficient*. It seems that the Fisher efficiency gives a guideline to finding the "best" estimator. Indeed, under regularity conditions, maximum likelihood estimators can be proved to be asymptotically unbiased and efficient and they are used a wide range of applications. However, the major pitfall of the Fisher efficiency lies in the superefficient points, at which the Cramér-Rao lower bound is violated and estimators exist that are asymptotically more efficient than any asymptotically Fisher efficient estimator. See Korostelev and Korosteleva (2011) for more details.

To avoid such pitfall, the idea of minimax risk was developed. Given a nonnegative loss function on Ξ^2, $L(t, \eta)$, the risk of an estimator $\hat{\eta}$ is defined by the expected loss $PL(\hat{\eta}, \eta(P))$. A commonly used loss function is the quadratic loss, $L(t, \eta) = |t - \eta|^2$. An estimator $\hat{\eta}^*$ is called minimax if its maximum risk does not exceed that of any other estimator $\hat{\eta}$

$$\sup_{P \in \mathcal{P}} R(\eta, \hat{\eta}^*, L) \leq \sup_{P \in \mathcal{P}} R(\eta, \hat{\eta}, L), \tag{2.3.1}$$

where

$$R(\eta, \hat{\eta}, L) = \mathbb{E}_\eta L(\hat{\eta}, \eta(P)).$$

The minimax criterion is closely related to the Bayesian criterion. Assume there is a prior density of η, $\pi(\eta)$, defined on Ξ, which reflects the knowledge about the parameter before any observation. The *Bayes risk* of $\hat{\eta}$ is

$$\beta(\hat{\eta}, L, \pi) = \int_\Xi R(\eta, \hat{\eta}, L)\ \pi(\eta)\ d\eta.$$

An estimator is called the *Bayes estimator* if it minimizes the Bayes risk. It turns out that if a Bayes estimator $\hat{\theta}$ has constant risk, i.e. $\pi(\{\eta \in \Xi : R(\eta, \hat{\eta}, L) = \sup_{P \in \mathcal{P}} R(\eta, \hat{\eta}, L)\}) = 1$,

then it is also a minimax estimator. However, it could be difficult to find a Bayes estimator with constant risk.

In addition to the Bayes criterion, several alternative approaches are available to derive minimax convergence rates. One class of methods are based on hypothesis testing arguments, which relate the minimax lower bound to the affinity to the total variation distance between the null and the alternative hypotheses. Below we will review three of these methods that are most widely used. We will also summarize some important work on minimax estimation in the functional data analysis field. In addition, in view of the change point estimation techniques employed in Lindquist et al. (2007) and Robinson et al. (2010), we will review the literature on minimax change-point estimation too. We will see that, the problems of minimax rates for FLR and change-point estimation are mostly nonparametric problems, and the optimal rates typically depend on the smoothness of the predictor process (for FLR) and the underlying process (for change-point). However, in the sparse functional regression case, the parameter space is finite-dimensional and the predictor process may not be smooth or differentiable. In this case, the approach to deriving the minimax rates might be very different.

2.3.1 Hypothesis-testing based approaches

In many minimax problems, the upper bound on the minimax risk is given by a specific estimator, as can be seen in Section 2.3.1. Some important tools for establishing the minimax lower bound are hypothesis-testing based approaches, which have been widely employed in the problem of minimax estimation. Among them, the most popular ones might be Le Cam, Assouad and Fano's methods. The first method deals with two sets of hypotheses, while the Assouad and Fano's methods deal with multiple hypotheses indexed by the vertices of a hypercube and those of a simplex, respectively. We will briefly review these three methods. The reader is referred to Yu (1997) for more details.

Define the total variation affinity between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ as

$$\|\mathbb{P} \wedge \mathbb{Q}\| = \inf_{f_0+f_1\geq 1}\{\mathbb{P}f_0 + \mathbb{Q}f_1\},$$

where the infimum runs over nonnegative measurable functions satisfying the inequality pointwise. Let $\mathscr{G}$ be the σ-field on which $\mathbb{P}$ and $\mathbb{Q}$ are defined. Then the affinity is closely

related to their total variation distance

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} = \sup_{A \in \mathscr{G}} \{\mathbb{P}(A) - \mathbb{Q}(A)\}. \tag{2.3.2}$$

In fact, it can be shown that

$$\|\mathbb{P} \wedge \mathbb{Q}\| = 1 - \|\mathbb{P} - \mathbb{Q}\|_{TV}. \tag{2.3.3}$$

Le Cam (1973) relates the testing problem of two sets of hypotheses to the total variation distance between the convex hulls of the two classes of hypotheses (probability measures). Intuitively, if the testing between these two sets is to be powerful, then their convex hulls should be well separated. Since estimators also imply tests between subsets of the parameter space, Le Cam's bound also provides a lower bound for the accuracy of an estimator.

Lemma 2.3.1. *Let $\mathcal{P}$ be a family of probability measures and $\eta(P)$ is the parameter of interest taking values in a pseudo-metric space $(\mathcal{D}, d)$. Let $\hat{\eta}$ be an estimator of $\eta(P)$. Suppose D_1 and D_2 are two subsets of $\mathcal{D}$ and let $c = \inf\{d(s_1, s_2),\ s_1 \in D_1, s_2 \in D_2\}$. Suppose also $\mathcal{P}_1$ and $\mathcal{P}_2$ are the subsets of $\mathcal{P}$ corresponding to D_1 and D_2 respectively. Denote by $co(\mathcal{P})$ the convex hull of $\mathcal{P}$. Then*

$$\sup_{P \in \mathcal{P}} E_P d(\hat{\eta}, \eta) \geq \frac{1}{2} c \sup_{P_i \in co(\mathcal{P}_i)} \|P_1 \wedge P_2\|.$$

In many cases, two simple hypotheses are sufficient for deriving sharp minimax lower bounds. However, in other situations, it may help obtain better lower bounds to consider the convex hulls of the $\mathcal{P}_\rangle$, because the supremum of the total variation over the convex hulls can be much larger than the supremum over the simple hypotheses themselves. See also Donoho and Liu (1991) for example.

Assouad's lemma obtain a minimax lower bound based on a class of 2^m hypotheses indexed by vertices of a m-dimensional hypercube. We will present the form of Assouad's lemma given by Devroye (1987), which emphasize the decomposability of the (pseudo) distance d into a sum of m (pseudo) distances, which correspond to m estimation subproblems. Each subproblem is like testing the hypotheses indexed by neighboring vertices on the hypercube along the direction determined by the particular subproblem, and the argument used in Le Cam's method can be applied to each of the subproblems.

Lemma 2.3.2. *Let $m \geq 1$ be an integer and let $\mathcal{F}_m = \{P_\tau : \tau \in \{-1,1\}^m\}$ contain 2^m probability measures. Write $\tau \sim \tau'$ if τ and τ' differ in only one coordinate, and write $\tau \sim_j \tau'$ when that coordinate is the jth. Suppose that there are m pseudo-distances on $\mathcal{D}$ such that for any $x, y \in \mathcal{D}$*

$$d(x,y) = \sum_{j=1}^{m} d_j(x,y),$$

and further that, if $\tau \sim_j \tau'$,

$$d_j(\eta(P_\tau), \eta(P_{\tau'})) \geq \alpha_m.$$

Then

$$\max_{P_\tau \in \mathcal{F}_m} E_\tau d(\hat{\eta}, \eta(P_\tau)) \geq m \cdot \frac{\alpha_m}{2} \min\{\|P_\tau \wedge P_{\tau'}\| : \tau \sim \tau'\}.$$

The relation $\tau \sim \tau'$ can be written in terms of the Hamming distance W as $W(\tau, \tau') = 1$, where

$$W(\tau, \tau') = \frac{1}{2} \sum_{j=1}^{m} |\tau_j - \tau'_j|,$$

is the number of places where τ and τ' differ.

Devroye (1987) also gives a generalized Fano's lemma in the setting that $\eta(P)$ is the density of P and d is the L^1 norm. The lemma presented below is a slightly stronger version of Fano's lemma given in Han and Verdú (1994) with less involved proof than those in the statistics literature, which is based on information theory concepts and Fano's original inequality.

Lemma 2.3.3. *Let $\tau \geq 2$ be an integer and let $\mathcal{M}_\tau \subset \mathcal{P}$ contain r probability measures indexed by $j = 1, 2, \ldots, r$ such that for all $j \neq j'$*

$$d(\eta(P_\tau), \eta(P_{\tau'})) \geq \alpha_r,$$

and

$$K(P_j, P_{j'}) = \int \log(P_j / P_{j'}) dP_j \leq \beta_\tau.$$

Then

$$\max_j E_j d(\hat{\eta}, \eta(P_\tau)) \geq \frac{\alpha_j}{2}\left(1 - \frac{\beta_r + \log 2}{\log r}\right).$$

As noted in Birgé (1986), "[Fano's Lemma] is in a sense more general because it applies in more general situations. It could also replace Assouad's Lemma in almost any practical cases ...". Compared with Le Cam's method, however, Fano's method does not deal with the case of two simple hypotheses since $r = 2$ the lower bound it gives is non-positive.

The examples in the literature suggest that, Lecam's method often works well when a real functional is estimated, but it can be challenging to find the appropriate two sets of hypotheses it requires. On the other hand, the other two lemmas seem to give the optimal rates when the entire function is being estimated, with Assouad's lemma seeming easier to use and therefore more popular than Fano's.

2.3.2 Minimax estimation for FLR and functional GLM

In the functional data analysis literature, several authors investigated the minimax estimation problem for the FLR model (2.1.3) (Cardot and Johannes, 2010; Cai and Hall, 2006; Hall and Horowitz, 2007). In particular, Hall and Horowitz (2007) showed that the minimax rates of convergence for estimating the slope function under the mean integrated squared error are determined by the smoothness of both the slope function and the covariance kernel of the predictor process.

Hall and Horowitz took the principal component approach by solving the normal equation (2.1.4) using expansions (2.1.1) and (2.1.2). The PCA-based estimator of b is given by

$$\hat{b} = \sum_{j=1}^{m} \hat{b}_j \hat{\psi}_j, \quad \hat{b}_j = \hat{\theta}_j^{-1} \hat{g}_j, \quad \hat{g}_j = \int \hat{g} \hat{\psi}_j$$

and $\hat{g}(u) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})\{X_i(u) - \bar{X}(u)\}$. Here the truncation point m is a smoothing parameter, and ψ and θ are defined as before.

To give the optimal rates of convergence, Hall and Horowitz make the following assumptions: X has finite fourth moment; $E(\psi_j^4) \leq C\theta_j^2$ for all j; and the errors ϵ_i are identically distributed with zero mean and variance not exceeding C;

$$\theta_j - \theta_{j+1} \geq C^{-1} j^{-\alpha-1}, \text{ for } j \geq 1 \tag{2.3.4}$$

and

$$|b_j| \leq Cj^{-\beta}, \tag{2.3.5}$$

$$\alpha > 1, \ \frac{1}{2}\,\alpha + 1 < \beta. \tag{2.3.6}$$

Condition (2.3.4) prevents the spacings between adjacent order statistics from being too small. It also follows that θ_j must not be less than a constant multiple of $j^{-\alpha}$. Conditions (2.3.5) and (2.3.6) basically require that the target function b is sufficiently smooth relative to the covariance kernel K and the expansion coefficients b_j do not decrease too fast.

By choosing the tuning parameter m such that

$$m \asymp n^{1/(\alpha+2\beta)}, \tag{2.3.7}$$

the authors gave the minimax convergence rate of estimators for the slope function and showed that the estimator based on PCA attains this rate under the above assumptions. Specifically, it was shown that

$$\int_{\mathcal{I}} (\hat{b} - b)^2 = O_p(n^{-(2\beta-1)/(\alpha+2\beta)}) \tag{2.3.8}$$

uniformly on $\mathcal{F}(C, \alpha, \beta)$, the set of distributions F of (X, Y) that satisfy (2.3.4)–(2.3.6). The values of α and β basically measure the smoothness of X and b, and condition (2.3.6) link the smoothness together.

Dou et al. (2010) extended this result to the functional generalized linear model setting and provided a minimax estimator of the slope function. They made assumptions analogous to the assumptions made by Hall and Horowitz (2007) and used the same bandwidth choice. They proposed a finite-dimensional approximation of the maximum likelihood estimator and proved that it has convergence rate $\rho_n = n^{(1-2\beta)/(\alpha+2\beta)}$. A variation on Assouad's Lemma was applied to deriving the minimax lower bound (Yu, 1997; van der Vaart, 1998). The Assouad's Lemma gives lower bound for the minimax risk over a class of 2^m probability measures based on testing hypotheses indexed by vertices of a m-dimensional hypercube, which in turn is a lower bound for the minimax risk over the entire parameter space. The authors showed that the lower bound is also of order $n^{(1-2\beta)/(\alpha+2\beta)}$. Therefore the approximated MLE achieves the optimal rates.

An alternative approach to deriving the minimax rates is to establish statistical equivalence between the model of interest and a model whose minimax estimator is already known, in the sense that Le Cam's metric (Le Cam, 1986; Le Cam and Yang, 1990) for the distance between the two models converges to zero as n goes to infinity. Such asymptotic equivalence will imply that any minimax procedure in one problem will automatically yield the corresponding procedure in the other with equal optimal rates.

Brown and Low (1996) provided an important result on asymptotic equivalence of nonparametric regression problems and white noise with drift problems. Specifically, consider two models

$$Y_{ni} = f(x_{ni}) + \sigma(x_{ni})\varepsilon_{ni}, \quad \varepsilon_{ni} \overset{\text{i.i.d.}}{\sim} N(0,1),\ i = 1, \ldots, n \tag{2.3.9}$$

and

$$dZ_t^{(n)} = f(t)dt + \lambda(t)dB_t/\sqrt{n}, \tag{2.3.10}$$

where B_t is a standard Brownian motion. The parameter space Θ consists of a possibly large set of choices of f. Here the predictors are given by

$$x_{ni} = H^{-1}(i/(n+1)), \quad i = 1, \ldots, n,$$

in a deterministic scheme, where H is an increasing c.d.f., and

$$X_{ni} \sim H \quad \text{i.i.d.} i = 1, \ldots, n,$$

in a random scheme.

Assuming on a compact interval I

$$\left| \frac{\partial}{\partial t} \log \sigma(t) \right| \leq C_1, t \in I$$

for some $C_1 < \infty$,

$$\sup\{|f(t)| : t \in I, f \in \Theta\} = B < \infty,$$

$$H'(t) = h(t) > 0 \text{ a.e. on } I,$$

and some uniform smoothness condition on f, the authors showed that (2.3.9) and (2.3.10) are asymptotically equivalent in Le Cam's sense, under either deterministic or random scheme, with $\lambda^2(t) = \sigma^2(t)/h(t)$.

Meister (2011) showed that the functional linear regression model (2.1.3), written as $Y = \langle X, \phi \rangle + \varepsilon$ where ϕ is the regression function and $\langle \cdot, \cdot \rangle$ denotes the $L_2([0,1])$-inner product, is equivalent to a white noise model with drift

$$dY(t) = [\Gamma^{1/2}\phi](t)dt + n^{-1/2}\sigma dW(t)$$

where $\Gamma^{1/2}$ is the square root of the covariance operator of X, defined by $\Gamma^{1/2}\Gamma^{1/2} = \Gamma$ and $\Gamma f = \int EX(\cdot)X(t)f(t)dt$. Such equivalence, combined with the results in Cavalier and Tsybakov (2002), gave sharp minimax constants in the FLR model.

2.3.3 Minimax estimation of change points

Change-point and singularity detection is often essential to signal processing in the fields of economics, medicine and physical science, since they may contain important information with scientific significance. For example, in pattern recognition, discontinuities of the image signal intensity function may indicate the location of the edge of an object. Overviews of the area and references can be found in Carlstein et al. (1994) and Korostelëv and Tsybakov (1993)

Due to the lack of knowledge of the underlying function, the problem of change-point estimation is usually considered in a nonparametric framework. The cases where the observations are direct are extensively studied. The simplest case is that of a single jump of a function that are assumed to satisfy some smoothness condition otherwise. One approach to solving this problem is to exploit the the result of Brown and Low (1996) mentioned earlier, derive minimax estimators from white noise models and then apply it in nonparametric regression setting. Korostelëv (1987) took this approach and showed that, in the Gaussian white noise model (2.3.10), assuming the mean function $f(\cdot)$ is finite on the [0,1] interval, has a unique jump at an interior time point, and is Lipschitz elsewhere, then the optimal rate for estimating the change-point is n^{-1}. Another approach is based on certain kernel estimators and the analysis of their differences (Yin, 1988; Müller, 1992; Hall and Titterington, 1992; Wu and Chu, 1993). Wang (1995) gave a closely related result in wavelets.

Raimondo (1998) extended the problem of estimating a change-point to estimating a "cusp" of an arbitrary order. Raimondo assumed the underlying signal f is observed at

discrete time points subject to additive noises. f is smooth except at one point, θ, where f is "α-discontinuous" in a Hölder sense. He claimed that the asymptotic minimax rate for estimating θ is $r_n = n^{-1/(1+2\alpha)}$.

In addition, change-point problems based on indirect observations have also received substantial attention. Neumann developed a minimax estimation method in the setting of ill-posed statistical inverse problems. He assumed that the observations are i.i.d. samples of (X,Y), which satisfies

$$Y = X + \xi,$$

where X is a random variable with unknown probability density f and ξ is the error term independent of X with known probability density K. It was also assumed that f has a discontinuous jump at θ and satisfies a Lipschitz condition elsewhere. The author proposed an estimator based on the difference of one-sided deconvoluting kernel estimators. It was shown that the minimax rate of this estimator is $n^{-1/(\beta+3/2)}$ if $\beta \geq 1/2$, and $n^{-1/(1+2\beta)}$ if $\beta < 1/2$, where β is the degree of ill-posedness of the inverse problem, i.e. the tails of the characteristic function $\hat{K}(\omega)$ of ξ decay at rate $|\omega|^{-\beta}, \beta > 0$.

Goldenshluger et al. (2006) took the white noise approach under the indirect and noisy observation setting

$$dY(x) = (\mathbf{K}f)(x)dx + n^{1/2}dW(x), \qquad x \in \mathbb{R}, \tag{2.3.11}$$

where $W(\cdot)$ is the standard two-sided Brownian motion that corresponds to the noise in data. $\mathbf{K}$ is the convolution operator, modeling the indirectness of the observation, that is defined by

$$(\mathbf{K}f)(x) = \int_{-\infty}^{\infty} K(x-y)f(y)dy$$

where $K \in L_1(\mathbb{R})$ and $f \in L_2(\mathbb{R})$. It was shown that estimating θ in Raimondo's model is equivalent to estimating θ in (2.3.11) when K is the Green's function of a linear differential operator of integer β, and there is a discrepancy between the rates of convergence obtained by Raimondo and by Neumann. Goldenshluger *et al.* showed that the faster rate obtained by Neumann, can indeed be attained and they are optimal for the white noise model (2.3.11). In particular, the authors showed that if f is mth order differentiable except at the change-point and bounded for all x, then the minimax rate for estimating θ is

$\min\{n^{-(m+1)/(2m+2\beta+1),-1/(2\beta+1)}\}$, provided the Fourier transform $\hat{K}$ of K decreases at the rate $|\omega|^{-\beta}, \beta > 0$.

Chapter 3

Optimal rates of convergence for the sparse functional linear model

We have seen in Section 2.3 that a considerable part of the functional data analysis literature has been focused on the problem of minimax estimation for the functional linear model (2.1.3), where the slope function $\beta(\cdot)$ is the parameter of interest. To our knowledge, however, the problem of minimax estimation for the sparse functional linear model has yet to be studied. We may not directly use the results from the FLR model because the slope function β has infinite dimensions, therefore its estimation problem is closely related to nonparametric minimax estimation, while the sparse functional linear model (2.2.1) is a parametric model. Therefore, finding the minimax estimators for parameters in the sparse functional linear regression, particularly the sensitive point θ, is expected to involve different techniques and arguments. In fact, it has not been studied yet to our knowledge.

In this chapter, we aim to resolve this problem and derive minimax convergence rates for estimating the parameters in the sparse functional linear regression model. Section 3.1 specifies the model and describes the calculation procedure to obtain the minimax estimator. We will use milder conditions on the predictor process than those assumed in McKeague and Sen (2010). In particular, we define a "generalized Hurst exponent" of the functional predictor and discuss its properties in Section 3.2. Section 3.3 presents the entire list of conditions and gives an example that meets these requirements. Sections 3.4 and 3.5 present

the main results on the minimax risk for estimating parameters in sparse functional linear regression. Section 3.6 gives the complete proofs to the lemmas and theorems.

3.1 Model specification and estimation

We shall assume the data consist of independent, identically distributed pairs $\{(X_i, Y_i), i = 1, \ldots, n\}$, which are i.i.d. replicates of (X, Y), where Y is a scalar response and X is a stochastic process. For example in the fMRI study, X_i could be the image signal process at a particular voxel from the ith patient. In the gene expression study, X_i could be the chromosome-wise expression profile from the ith patient. Suppose the index space of X is a compact interval in the real line, which we will take to be $[0, 1]$ without loss of generality. Recall that the sparse functional linear regression model (2.2.1) is given by

$$Y = \alpha + \beta X(\theta) + \varepsilon.$$

The intercept α and the slope β are both scalars and the sensitive point θ is the parameter of main interest. ε is a mean-zero error term and is independent of X. We shall see that the specification of ε is fairly general and it is unnecessary to assume a parametric model for it. Therefore, the sparse functional linear regression model is indexed by the parameters $\eta = (\alpha, \beta, \theta)$.

The least squares estimator of η is given in (2.2.2). Since the paths of X_i do not have a definite functional form, which could be nondifferentiable, the object function in (2.2.2) could be non-convex and the LSE is obtained via a profile estimate procedure. Specifically, for each fixed θ, a profile estimate of (α, β) is given by

$$(\hat{\alpha}_n(\theta), \hat{\beta}_n(\theta)) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{n} [Y_i - \alpha - \beta X_i(\theta)]^2. \tag{3.1.1}$$

Then an estimate of θ is given by

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} [Y_i - \hat{\alpha}(\theta) - \hat{\beta}(\theta) X_i(\theta)]^2. \tag{3.1.2}$$

Finally $\hat{\eta}_n$ is obtained from

$$(\hat{\alpha}_n(\hat{\theta}_n), \hat{\beta}_n(\hat{\theta}_n), \hat{\theta}_n). \tag{3.1.3}$$

In practice, X_i are observed on discrete points and the profile estimate 3.1.1 has close form solutions, so this estimate procedure is tractable.

3.2 Generalized Hurst exponent

To derive the minimax rates, we make certain assumptions on the second moment structure of X. These conditions are more general than assuming that X is a (fractional) Brownian motion. The most important one among them is that X has a *generalized Hurst exponent* $H > 0$ that satisfies

$$E|X(\theta_1) - X(\theta_2)|^2 \asymp |\theta_1 - \theta_2|^{2H}, \quad \forall \theta_1, \theta_2 \in [0, 1]. \tag{3.2.1}$$

The symbol $\asymp$ here, and in the sequel, means bounded from above and below up to (positive) constants, which can depend on the covariance structure of X and the parameter space Ξ. This condition requires that the second moment of an increment of X converges to zero at an exponential rate as the increment vanishes. Intuitively, H describes the local smoothness of the covariance of X. For fixed θ_1 and θ_2, the smaller H is, the larger the increment is likely to be, which means X is likely to have stronger fluctuations and its trajectories are rougher.

Example 3.2.1. *(Fractional Brownian motion) A (standard) fractional Brownian motion (fBm) with Hurst exponent $H \in (0, 1]$ is a Gaussian process $B_H = \{B_H(t), t \in \mathbb{R}\}$ having continuous sample paths, mean zero, and covariance function*

$$Cov\{B_H(t), B_H(s)\} = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t - s|^{2H}\right). \tag{3.2.2}$$

Based on its mean and covariance functions, it can be seen that $E|X(\theta_1) - X(\theta_2)|^2 = |\theta_1 - \theta_2|^{2H}$. Therefore, the generalized Hurst exponent of X reduces to its Hurst exponent.

Some properties of the generalized Hurst exponent are described in the following proposition. The proof of the proposition can be found in Section 3.6.

Proposition 3.2.2. *Suppose X has generalized Hurst exponent H, then*

(a) H is unique;

(b) if X is a Gaussian process, then its trajectories are Lipschitz of any order $\alpha < H$, in the sense that

$$|X(t) - X(s)| \leq \xi |t - s|^{\alpha} \quad \forall t, s \in [0, 1] \tag{3.2.3}$$

almost surely, where ξ has moments of all orders;

(c) $H \leq 1$.

Remark 3.2.3. *The proposition shows that the generalized Hurst exponent is well-defined. Also, the existence of a Gaussian process's generalized Hurst exponent implies Lipschitz continuity of its trajectories. Note that such property is different from the smooth trajectory assumptions commonly made in the functional linear regression literature, since the order of the Lipschitz continuity is limited by $H \leq 1$, and thus the trajectories might not be differentiable.*

3.3 Conditions

The sparse functional linear regression model (2.2.1) is indexed by $\eta = (\alpha, \beta, \theta) \in \Xi$, with the true parameters denoted as

$$\eta_0 = (\alpha_0, \beta_0, \theta_0).$$

We assume the parameter space $\Xi = [-b, b\ ] \times \{\beta \in \mathbb{R} : a \leq |\beta| \leq b\} \times (0, 1)$. Here $0 < a < b < \infty$ are fixed constants. The boundedness of the parameter space is crucial to obtaining the minimax bounds. For example, $|\beta_0|$ is bounded away from 0 in order to ensure identifiability of the parameters and avoid irregularity of the estimators. It is common to assume bounded parameter spaces in the minimax estimation literature. See Cai et al. (2010) and Cai and Jin (2010) for example.

In addition, the following assumptions are made. We need the first two conditions to derive the minimax lower bound and all six conditions for the minimax upper bound. Let $\lesssim$ mean that the left side is bounded above by a (positive) constant times the right side. Define $\gtrsim$ similarly.

(A1) X has a generalized Hurst exponent $H \in (0, 1]$.

(A2) $E[\sup_{\theta\in(0,1)} |X^p(\theta)|] < \infty$ for every $p \geq 1$, and $\inf_{\theta\in(0,1)} E|X(\theta)|^2 > 0$.

(A3) The trajectories of X are Lipschitz, i.e. (3.2.3) holds, for all $\alpha < H$ almost surely.

(A4) For any $q > 0$, $E[\sup_{|t-s|<\delta} |X(t) - X(s)|^q] \lesssim \delta^{Hq}$ for $\delta > 0$.

(A5) $|E_\eta[X(\theta)(X(\theta_1) - X(\theta))]| / |\theta_1 - \theta|^H \to 0$ uniformly w.r.t. $\eta = (\alpha, \beta, \theta) \in \Xi$ as $\theta_1 \to \theta$.

(A6) $E|\varepsilon|^p < \infty$ for all $p > 0$.

The assumptions on X are milder than assuming it is fBm. In fact, We can verify that a fBm satisfies conditions **(A1)** – **(A5)**. **(A1)** is trivial to prove; **(A2)** follows from Theorem 2.1 of Berman (1985); from Proposition 3.2.2, we know that condition **(A3)** holds since X has Gaussian increments and a Hurst exponent H; the validity of condition **(A4)** follows from Theorem 1.1 of Novikov and Valkeila (1999); recalling the covariance function of fBm we can show that **(A5)** holds if $|\theta_0|$ is bounded away from 0. These milder conditions allow us to consider predictor functions among a broader class of stochastic processes.

3.4 Minimax lower bound for sparse functional linear regression

In this section, we establish a lower bound on the minimax risk of estimating the parameters in sparse functional linear regression. As noted in Section 2.3.1, Le Cam's method is a widely used technique to derive minimax lower bounds for parametric estimation problems, by relating the problem of hypothesis testing to the total variation affinity between the null and the alternative distributions. We will make use of this method in our theory.

One challenge posed by the sparse functional linear model is to calculate the total variation affinity between two joint distributions of (X, Y), $\mathbb{P}_{n,1}$ and $\mathbb{P}_{n,2}$, since X is infinite-dimensional. However, noticing that the distribution of X does not involve the unknown parameters, we are able to adapt Le Cam's lemma to our case by exploiting the tower property of conditional expectations. The following lemma is a direct consequence of Le Cam's method, which involves calculating the affinity between two conditional distributions of Y given X, $\mathbb{Q}_{n,1}$ and $\mathbb{Q}_{n,2}$, rather than between two joint distributions of (X, Y). Let

$\mathbb{P}_{n,X}$ denote the marginal distribution of $(X_1, \ldots, X_n)$. A detailed description about the notation can be found in Section .

Lemma 3.4.1. *Let $\tilde{\eta}$ be any estimator of η based on a sample from a distribution in the collection $\{\mathbb{P}_\eta, \eta \in \Xi\}$. Let L be a loss function and $c(\eta_1, \eta_2) = \inf_{\eta \in \Xi}\{L(\eta, \eta_1) + L(\eta, \eta_2)\}$ then*

$$\sup_{\eta \in \Xi} \mathbb{E}_\eta L(\tilde{\eta}, \eta) \geq \frac{1}{2}\ c(\eta_1, \eta_2)\mathbb{P}_{n,X}\ \|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\|.$$

We will consider the minimax risk in the mean squared error sense. Theorem 2.1 in McKeague and Sen (2010) suggests the rates for estimating α, β and θ may differ. Therefore we apply the squared error loss function to each component of η: define $L_\theta(\eta, \eta_0) = |\theta - \theta_0|^2$ and L_α and L_β are defined similarly. With the help of the previous lemma, we can choose η_1 and η_2 sufficiently close while still bounding the total variational affinity away from zero, and show that for any estimator $\tilde{\eta}_n$,

$$\begin{aligned}
&\sup_{\eta \in \Xi} \mathbb{E}_\eta\ L_\alpha(\tilde{\eta}_n, \eta) \gtrsim n^{-1}, \\
&\sup_{\eta \in \Xi} \mathbb{E}_\eta\ L_\beta(\tilde{\eta}_n, \eta) \gtrsim n^{-1}, \\
&\sup_{\eta \in \Xi} \mathbb{E}_\eta\ L_\theta(\tilde{\eta}_n, \eta) \gtrsim n^{-1/H},
\end{aligned}$$

which immediately implies the following theorem.

Theorem 3.4.2. *Suppose conditions **(A1)** and **(A2)** hold, then the minimax risk of estimating η over Ξ satisfies*

$$\begin{aligned}
&\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_\eta\ L_\alpha(\tilde{\eta}_n, \eta) \geq C_1 n^{-1}, \\
&\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_\eta\ L_\beta(\tilde{\eta}_n, \eta) \geq C_1 n^{-1},\ \textit{and} \\
&\inf_{\tilde{\eta}_n, \eta} \sup_{\eta \in \Xi} \mathbb{E}_\eta\ L_\theta(\tilde{\eta}_n, \eta) \geq C_1 n^{-1/H},
\end{aligned}$$

where the supremums are taken over the parameter space Ξ and the infimums are taken over any estimator of the form $\tilde{\eta}_n = (\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\theta}_n)$, of η. $C_1 > 0$ only depends on the second moment structure of X and the parameter space Ξ.

A detailed proof to Theorem 3.4.2 can be found in Section 3.6.

3.5 Minimax upper bound for sparse functional linear regression

In this section, we establish the minimax upper bound for the sparse FLR model by deriving the second moment convergence rate of the least squares estimator. The rate is shown to be of the same order of the minimax lower bound previously derived. Thus the LSE attains the optimal rate and the minimax lower bound is rate-sharp.

M-estimation theory is a popular technique to establish the convergence rates and limiting distributions of estimators obtained from optimizing an object function. In fact, if we denote by $\mathbb{P}_n$ the empirical measure, $m_\eta(Y,X) = [Y - \alpha - \beta X(\theta)^2]$, and $\mathbb{M}_n(\eta) = \mathbb{P}_n m_\eta = \frac{1}{n}\sum_{i=1}^n [Y_i - \alpha - \beta X_i(\theta)]^2$, then the LSE

$$\hat{\eta}_n = \arg\min_\eta \mathbb{M}_n(\eta).$$

However, the typical strategy stated in Theorem 3.2.5 of van der Vaart and Wellner (1996) only provides the weak convergence rate of the M-estimator. A recent result from Nishiyama (2010) extended Theorem 3.2.5 of van der Vaart and Wellner (1996) and offers an approach to obtaining the moment convergence rate of any order $p \geq 1$. However, establishing the minimax upper bound requires a moment convergence rate that is valid uniformly across the parameter space. Therefore we further extend the result of Nishiyama (2010) in the following.

Define

$$\begin{aligned}
\mathbb{M}(\eta) &= Pm_\eta = E[\alpha_0 + \beta_0 X(\theta_0) + \varepsilon - \alpha - \beta X(\theta)]^2 \\
&= E[(\alpha_0 - \alpha) + (\beta_0 X(\theta_0) - \beta X(\theta))]^2 + \sigma^2 \\
&= (\alpha_0 - \alpha)^2 + E[(\beta_0 X(\theta_0) - \beta X(\theta))^2] + \sigma^2 \\
&= (\alpha_0 - \alpha)^2 + \sigma^2 + (\beta_0 - \beta)^2 E[X^2(\theta_0)] + \beta^2 E[X(\theta_0) - X(\theta)]^2 \\
&\quad + 2\beta(\beta_0 - \beta) E[X(\theta_0)(X(\theta_0) - X(\theta))].
\end{aligned}$$

The third equation follows from the independence between X and ε and the fact that $E(\varepsilon) = 0$. The fourth equation follows from the fact that $E(X(\theta)) = 0,\ \forall\theta \in [0,1]$. Our result is based on the following lemma, which is a direct consequence of Theorem 1

in Nishiyama (2010). It gives a moment convergence rate, r_n, that is uniform over the parameter space.

Lemma 3.5.1. *Let $\mathbb{M}_n$ be stochastic processes indexed by a semimetric space (Ξ, d) and $\mathbb{M} : \Xi \mapsto \mathbb{R}$ a deterministic function such that for a constant $\epsilon > 0$,*

$$\mathbb{M}(\eta) - \mathbb{M}(\eta_0) \leq -\epsilon d(\eta, \eta_0)^2, \quad \forall \eta \in \Xi. \tag{3.5.1}$$

Suppose that there exists functions ϕ_n such that $\delta \mapsto \delta^{-\alpha}\phi_n(\delta)$ is non-increasing for some $\alpha < 2$ (not depending on n) and that for every $p \geq 1$ there exists a constant $C_p > 0$ such that for every $\delta > 0$

$$\left(E_{\eta_0} \left| \sup_{d(\eta,\eta_0)<\delta} |(\mathbb{M}_n - \mathbb{M})(\eta) - (\mathbb{M}_n - \mathbb{M})(\eta_0)| \right|^p \right)^{1/p} \leq C_p \frac{\phi_n(\delta)}{\sqrt{n}}. \tag{3.5.2}$$

Let $r_n^2 \phi_n(1/r_n) \leq \sqrt{n}$ for every n. If ϵ and C_p are both independent of η_0, and the sequence η_n satisfies $\mathbb{M}_n(\eta_n) \geq \mathbb{M}_n(\eta_0) - r_n^{-2}$, then

$$\sup_{\eta_0 \in \Xi} \sup_n E_{\eta_0} |r_n d(\eta_n, \eta_0)|^p < \infty$$

for every $p \geq 1$.

Define $d(\eta, \eta_0) = \max\{|\alpha - \alpha_0|, |\beta - \beta_0|, |\theta - \theta_0|^H\}$. With the help of conditions **(A1)** – **(A6)**, we are able to prove that the ϵ and C_p in the previous lemma exist and do not depend on η_0. Hence the LSE $\hat{\eta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n)$ satisfies

$$\sup_{\eta_0 \in \Xi} \sup_n E_{\eta_0} |r_n d(\hat{\eta}_n, \eta_0)|^2 < \infty$$

which translates to,

$$\begin{aligned}
&\sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \; L_\alpha(\hat{\eta}_n, \eta) \lesssim n^{-1}, \\
&\sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \; L_\beta(\hat{\eta}_n, \eta) \lesssim n^{-1}, \\
&\sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \; L_\theta(\hat{\eta}_n, \eta) \lesssim n^{-1/H}.
\end{aligned}$$

It follows that

Theorem 3.5.2. *If conditions* ***(A1)*** *–* ***(A6)*** *are satisfied, then*

$$\inf_{\tilde{\eta}_n} \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \; L_\alpha(\tilde{\eta}_n, \eta_0) \lesssim n^{-1},$$
$$\inf_{\tilde{\eta}_n} \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \; L_\beta(\tilde{\eta}_n, \eta_0) \lesssim n^{-1}, \; \textit{and}$$
$$\inf_{\tilde{\eta}_n} \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \; L_\theta(\tilde{\eta}_n, \eta_0) \lesssim n^{-1/H},$$

where the supremums are taken over the parameter space for η and the infimums are taken over any estimator of η.

A detailed proof is given in Section 3.6. It can be seen that the minimax upper bound given by the LSE is of the same order as the minimax lower bound. Therefore the LSE attains the optimal rates, and we have established the minimax rates for estimating the parameters in the sparse functional linear model.

Corollary 3.5.3. *If conditions* ***(A1)*** *–* ***(A6)*** *are satisfied, then*

$$\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_{\eta} \; L_\alpha(\tilde{\eta}_n, \eta) \asymp n^{-1}, \tag{3.5.3}$$
$$\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_{\eta} \; L_\beta(\tilde{\eta}_n, \eta) \asymp n^{-1}, \; \textit{and} \tag{3.5.4}$$
$$\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_{\eta} \; L_\theta(\tilde{\eta}_n, \eta) \asymp n^{-1/H}. \tag{3.5.5}$$

It is implied that the minimax rate for estimating θ_0 is at most $n^{1/(2H)}$, which is faster or equal to the usual parametric rate $n^{1/2}$. Also the rate for estimating θ_0 increases as H decreases. An intuitive explanation is that a smaller H indicates a rougher X, thus making it easier to distinguish the sensitive point from the rest of the index space.

This result is in contrast to that of Hall and Horowitz (2007), where the estimator of the slope function β has faster convergence when β and the covariance function of X is smoother. This is not surprising because when the impact of the predictor function is spread across the index space, the smoothness of its paths will enable us to "borrow" information from the observations in the adjacent neighborhood. On the contrary, when the impact is sparse in the index space, the roughness of the predictor function's trajectories makes it easier to identify the sensitive point.

3.6 Proofs

3.6.1 Preliminaries

We will view the predictor function $X = \{X(t), t \in [0,1]\}$ as a random element $X : (\Omega, \mathcal{F}, P) \to (\mathcal{X}, \mathscr{A})$, where $(\Omega, \mathcal{F}, P)$ is a probability space, $\mathcal{X} \equiv \mathbb{R}^{[0,1]}$ is the set of all real-valued functions on $[0,1]$, and $\mathscr{A}$ is the smallest σ-field on $\mathcal{X}$ with respect to which all the coordinate functions of the form

$$\pi_t(x) \mapsto x(t), \quad \forall x \in \mathbb{R}^{[0,1]}, t \in [0,1]$$

are measurable. Let the scalar response Y be a random variable $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathscr{B})$, where $\mathscr{B}$ is the Borel σ-field on the real line $\mathbb{R}$. Let P_X and P_Y be the distribution induced by X and Y on $(\mathcal{X}, \mathscr{A})$ and $(\mathbb{R}, \mathscr{B})$ respectively. Define their joint distribution P_{XY} on $\mathcal{X} \times \mathbb{R}$ by

$$P_{XY}(A \times B) = P(X^{-1}(A) \cap Y^{-1}(B)), \quad \forall A \in (A), B \in \mathscr{B}.$$

It follows from Theorem 33.3 in Durrett (1996) that there exists a regular conditional distribution $P_{Y|X}(\cdot,\cdot)$: $(\mathcal{X}, \mathscr{B}) \to [0,1]$, such that for $x \in \mathcal{X}$ a.e. P_X, $P_{Y|X}(x,\cdot)$ is a probability measure on $\mathscr{B}$, and for any $B \in \mathscr{B}$, $P_{Y|X}(\cdot, B)$ is $\mathscr{A}$-measurable and

$$P_{XY}(A \times B) = \int_A P_{Y|X}(x, B)\, dP_X(x) \quad \text{for all} \quad A \in \mathscr{A} \quad \text{and} \quad B \in \mathscr{B}.$$

Let the probability space be $(\Omega, \mathcal{F}, P)$. We will denote P_η as the probability distribution of (X, Y), P_X as the marginal distribution of X (a probability measure on $\mathcal{X} = \mathbb{R}^{[0,1]}$), and Q_λ as the conditional distribution of Y given X, where $\lambda = \alpha + \beta X(\theta)$. The existence of Q_λ is guaranteed by the existence of a regular conditional distribution of a random variable given a random element. Given X, Q_λ is the Gaussian measure $N(\lambda, \sigma^2)$ on $\mathbb{R}$. Their empirical counterparts are $\mathbb{P}_{n,\eta} = \otimes_{i \le n} P_{\eta,i}$, $\mathbb{P}_{n,X} = P_X^n$ and $\mathbb{Q}_{n,\eta,X_1,\ldots,X_n} = \otimes_{i \le n} Q_{\lambda_i}$, respectively. By the property of regular conditional distributions, $\mathbb{P}_{n,\eta}$ can be rewritten as an iterated expectation,

$$\mathbb{P}_{n,\eta} = \mathbb{P}_{n,X} \mathbb{Q}_{n,\eta,X_1,\ldots,X_n}.$$

We will abbreviate $\mathbb{P}_{n,\eta_j}$ to $\mathbb{P}_{n,j}$ and $\mathbb{Q}_{n,\eta_j,X_1,\ldots,X_n}$ to $\mathbb{Q}_{n,j}$ respectively.

To avoid measurability problems we will always use outer expectation/probability, and denote them by E and P. To obtain sharper constants in the minimax bounds, we also define

$$C_l = \sup \{C > 0 : E|X(\theta_1) - X(\theta_2)|^2 \geq C|\theta_1 - \theta_2|^{2H}, \text{ for all } \theta_1, \theta_2 \in [0,1]\}, \tag{3.6.1}$$

$$C_u = \inf \{C > 0 : E|X(\theta_1) - X(\theta_2)|^2 \leq C|\theta_1 - \theta_2|^{2H}, \text{ for all } \theta_1, \theta_2 \in [0,1]\}, \tag{3.6.2}$$

and assume they are both attained and nonzero. Finally, let $\sup_{\theta\in(0,1)} E|X(\theta)|^2 \equiv K$ and $\inf_{\theta\in(0,1)} E|X(\theta)|^2 \equiv \rho$.

3.6.2 Proofs of the properties of GHE and the extended Le Cam's lemma

Proof of Proposition 3.2.2: (a) If there exist $H_1, H_2 > 0$ that both satisfy (3.2.1), then there exist constants $C_{l,1}, C_{l,2}, C_{u,1}, C_{u,2} > 0$ such that for any $\theta_1, \theta_2 \in [0,1]$,

$$C_{l,1}|\theta_1 - \theta_2|^{2H_1} \leq E|X(\theta_1) - X(\theta_2)|^2 \leq C_{u,1}|\theta_1 - \theta_2|^{2H_1}$$

and

$$C_{l,2}|\theta_1 - \theta_2|^{2H_2} \leq E|X(\theta_1) - X(\theta_2)|^2 \leq C_{u,2}|\theta_1 - \theta_2|^{2H_2}.$$

It follows that

$$|\theta_1 - \theta_2|^{2(H_1 - H_2)} \leq C_{u,2}/C_{l,1}, \quad \text{for any } \theta_1, \theta_2 \in [0,1],$$

which implies that $H_1 \geq H_2$. Similarly we can show that $H_1 \leq H_2$. Thus $H_1 = H_2$.

(b) From the definition of H, there exists $L > 0$ such that

$$E|X(t) - X(s)|^2 \leq L|t - s|^{2H}$$

for all $t, s \in [0,1]$. Because X has Gaussian increments, for every $p > 0$, there exists some C_p such that

$$E|X(t) - X(s)|^{2p} \leq C_p L|t - s|^{2Hp} \quad \forall t, s \in [0,1].$$

It follows from Theorem 2.1 in Revuz and Yor (1999) that X has a modification $\tilde{X}$ whose paths are Hölder continuous of order $\alpha < (2Hp - 1)/(2p) = H - 1/(2p)$. Since p can be

chosen arbitrarily large, this is true for all $\alpha < H$.

(c) Let X_G be a Gaussian process such that for every $0 \leq s < t \leq 1$,

$$E|X_G(t) - X_G(s)|^2 = E|X(t) - X(s)|^2.$$

From (b), we know that there exists $1 < \alpha < H$ such that the paths of X_G are almost surely Hölder continuous of order α. Since any Hölder continuous function of order greater than 1 must be constant, X_G is constant almost surely. Therefore

$$E|X(t) - X(s)|^2 = E|X_G(t) - X_G(s)|^2 = 0,$$

which contradicts with (3.2.1). ∎

Proof of Lemma 3.4.1: Abbreviate $c(\eta_1, \eta_2)$ to c. Given X, the pair of functions $f_j = L(\tilde{\eta}, \eta_j)/c, j = 1, 2$ are included amongst the pairs in (2.3.1) that define the affinity between $\mathbb{Q}_{n,1}$ and $\mathbb{Q}_{n,2}$. Thus

$$\begin{aligned} 2\sup_{\eta\in\Xi} \mathbb{E}_\eta L(\tilde{\eta}, \eta) &\geq \mathbb{E}_{\eta_1} L(\tilde{\eta}, \eta_1) + \mathbb{E}_{\eta_2} L(\tilde{\eta}, \eta_2) \\ &= c(\mathbb{P}_{n,1} f_1 + \mathbb{P}_{n,2} f_2) \\ &= c\mathbb{P}_{n,X}(\mathbb{Q}_{n,1} f_1 + \mathbb{Q}_{n,2} f_2) \\ &\geq c\mathbb{P}_{n,X} \, \|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\|. \end{aligned}$$

∎

3.6.3 Proof of the minimax lower bound

Proof of Theorem 3.4.2: Select $\eta_1 = (\alpha_1, \beta_1, \theta_1)$ and $\eta_2 = (\alpha_2, \beta_2, \theta_2)$ such that $|\alpha_2 - \alpha_1| = cm^{-1}$, $|\beta_2 - \beta_1| = cm^{-1}$ and $|\theta_2 - \theta_1| = ck^{-1}$. Here $c > 0$ is a constant depending only on the second moment structure of X and the parameter space Ξ, while m and k go to infinity as $n \to \infty$. Then Lemma 3.4.1 implies

$$\begin{aligned} &\sup_{\eta\in\Xi} \mathbb{E}_\eta L_\alpha(\tilde{\eta}, \eta) \geq \frac{1}{4} c^2 m^{-2} \|\mathbb{P}_{n,1} \wedge \mathbb{P}_{n,2}\|, \\ &\sup_{\eta\in\Xi} \mathbb{E}_\eta L_\beta(\tilde{\eta}, \eta) \geq \frac{1}{4} c^2 m^{-2} \|\mathbb{P}_{n,1} \wedge \mathbb{P}_{n,2}\|, \text{ and} \\ &\sup_{\eta\in\Xi} \mathbb{E}_\eta L_\theta(\tilde{\eta}, \eta) \geq \frac{1}{4} c^2 k^{-2} \|\mathbb{P}_{n,1} \wedge \mathbb{P}_{n,2}\|. \end{aligned} \tag{3.6.3}$$

We want to show that $\|\mathbb{P}_{n,1} \wedge \mathbb{P}_{n,2}\|$ is bounded away from 0 as n goes to infinity. Recalling (2.3.3), it suffices to show that $\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV}$ is bounded away from 1. Note that given $\mathbb{X} = (X_1, \ldots, X_n)$, $\mathbb{Q}_{n,j}$ are multivariate Gaussian distributions, thus the subset of $\mathbb{R}^n$ that achieves the supremum in (2.3.2) defining $\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV}$ is

$$\begin{aligned}
A &= \{y \in \mathbb{R}^n : \mathbb{Q}_{n,1}(y) \geq \mathbb{Q}_{n,2}(y)\} \\
&= \left\{y \in \mathbb{R}^n : \frac{d\mathbb{Q}_{n,1}}{d\mathbb{Q}_{n,2}}(y) \geq 1\right\} \\
&= \left\{y \in \mathbb{R}^n : \frac{\prod_{i=1}^n \left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \lambda_i^1)^2\right]\right\}}{\prod_{i=1}^n \left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \lambda_i^2)^2\right]\right\}}\right\} \\
&= \left\{y \in \mathbb{R}^n : \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n ((y_i - \lambda_i^1)^2 - (y_i - \lambda_i^2)^2)\right]\right) \geq 1\right\} \\
&= \left\{y \in \mathbb{R}^n : \sum_{i=1}^n [(y_i - \lambda_i^1)^2 - (y_i - \lambda_i^2)^2] \leq 0\right\} \\
&= \left\{y \in \mathbb{R}^n : \sum_{i=1}^n [(\lambda^2 - \lambda^1)(y_i - \lambda_i^1) \leq \frac{1}{2}\sum_{i=1}^n (\lambda_i^1 - \lambda_i^2)^2]\right\},
\end{aligned}$$

where $\lambda_i^j = \alpha_j + \beta_j X_i(\theta_j)$. Define $\Lambda^j = (\lambda_1^j, \ldots, \lambda_2^j)$, $\upsilon = (\Lambda^2 - \Lambda^1)/|\Lambda^1 - \Lambda^2|$ and $\tau = |\Lambda^1 - \Lambda^2|/(2\sigma)$. Then

$$\begin{aligned}
A &= \{y \in \mathbb{R}^n : (\Lambda^2 - \Lambda^1)'(y - \Lambda^1) \leq \frac{1}{2}|\Lambda^1 - \Lambda^2|^2\} \\
&= \{y \in \mathbb{R}^n : \upsilon'(y/\sigma - \Lambda^1/\sigma) \leq \tau\},
\end{aligned}$$

and give $\mathbb{X}$, $\upsilon'(y/\sigma - \Lambda^1/\sigma)$ follows $N(0,1)$ under $\mathbb{Q}_{n,1}$ and $N(2\tau, 1)$ under $\mathbb{Q}_{n,2}$. Thus

$$\begin{aligned}
\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV} &= \mathbb{P}_{n,X}(\mathbb{Q}_{n,1}A - \mathbb{Q}_{n,2}A) \\
&= \mathbb{P}_{n,X}[P(N(0,1) \leq \tau) - P(N(2\tau,1) \leq \tau)] \\
&= \mathbb{P}_{n,X}[P(|N(0,1)| \leq \tau)] \\
&= \mathbb{P}_{n,X}[\Phi(\tau) - 1/2] \\
&\leq \frac{2}{\sqrt{2\pi}}\mathbb{P}_{n,X}[\tau],
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function. It follows from Cauchy–Schwarz inequality that

$$\begin{aligned}\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1}-\mathbb{Q}_{n,2}\|_{TV} &\leq \frac{2}{\sqrt{2\pi}}\left[\mathbb{P}_{n,X}\left[\tau^2\right]\right]^{1/2}\\ &= \left[\frac{1}{2\pi\sigma^2}\mathbb{P}_{n,X}|\Lambda^1-\Lambda^2|^2\right]^{1/2}.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{P}_{n,X}|\Lambda^1-\Lambda^2|^2 &= \mathbb{P}_{n,X}\sum_{i=1}^{n}\left[(\alpha_1+\beta_1X_i(\theta_1)-\alpha_2-\beta_2X_i(\theta_2))^2\right]\\ &\leq \mathbb{P}_{n,X}\sum_{i=1}^{n}[2|\alpha_1-\alpha_2|^2+\\ &\qquad 2|\beta_1X_i(\theta_1)-\beta_2X_i(\theta_1)+\beta_2X_i(\theta_1)-\beta_2X_i(\theta_2)|^2]\\ &\leq \mathbb{P}_{n,X}\sum_{i=1}^{n}[2|\alpha_1-\alpha_2|^2+4|X_i^2(\theta_1)|\cdot|\beta_1-\beta_2|^2+\\ &\qquad 4|\beta_2|^2\cdot|X_i(\theta_1)-X_i(\theta_2)|^2],\end{aligned}$$

The two inequalities come from the fact that $|a+b|^2\leq 2|a|^2+2|b|^2$, for all $a,b\in\mathbb{R}$. By condition **(A1)**, recalling the definition (3.6.2), we have

$$|X_i(\theta_1)-X_i(\theta_2)|^2\leq C_u|\theta_1-\theta_2|^{2H}.$$

By condition **(A2)**,

$$\sup_{\theta\in(0,1)}E|X^2(\theta)|\equiv K<\infty.$$

Recalling the upper bound on $|\beta|$ in the parameter space Ξ, we have

$$\begin{aligned}\mathbb{P}_{n,X}|\Lambda^1-\Lambda^2|^2 &\leq n\left[2|\alpha_1-\alpha_2|^2+4K\cdot|\beta_1-\beta_2|^2+4b^2\cdot C_u|\theta_1-\theta_2|^{2H}\right]\\ &\leq n\cdot(2c^2m^{-2}+4Kc^2m^{-2}+4C_ub^2c^{2H}k^{-2H}).\end{aligned}$$

If we choose $m=n^{1/2}$, $k=n^{1/(2H)}$ and $0<c\leq\min\{(\pi\sigma^2/(2+4K))^{1/2},(\pi\sigma^2/(4C_ub^2))^{1/(2H)}\}$, then

$$\begin{aligned}\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1}\wedge\mathbb{Q}_{n,2}\| &= 1-\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1}-\mathbb{Q}_{n,2}\|_{TV}\\ &\geq 1-\left[\frac{1}{2\pi\sigma^2}\ n\cdot(2c^2n^{-1}+4Kc^2n^{-1}+4C_ub^2c^{2H}n^{-1})\right]^{1/2}\\ &= 1-\left[\left((1+2K)c^2+2C_ub^2c^{2H}\right)/(\pi\sigma^2)\right]^{1/2}>0.\end{aligned}$$

Denoting $C_1 = c^2\left(1 - \left[\left((1+2K)c^2 + 2C_u b^2 c^{2H}\right)/(\pi\sigma^2)\right]^{1/2}\right)/4$ and recalling (3.6.3), we have proved Theorem 3.4.2. ∎

3.6.4 Proof of the minimax upper bound

We first prove Lemma 3.5.1 which is a consequence of Theorem 1 in Nishiyama (2010). It is an extension of Theorem 3.2.5 in van der Vaart and Wellner (1996) and, instead of providing the rate of weak convergence at one point, gives a moment convergence rate that is uniform over the parameter space.

Proof of Lemma 3.5.1: Choose $\gamma \geq 1$ such that $\alpha - 2 + \gamma^{-1} < 0$. For each n, we set $S_{j,n} = \{\eta : 2^{j-1} < r_n d(\eta, \eta_0) \leq 2^j\}$. Note that

$$\mathbb{M}(\eta) - \mathbb{M}(\eta_0) \leq -\epsilon d(\eta, \eta_0)^2 \leq -\epsilon \frac{2^{2j-2}}{r_n^2}, \quad \forall \eta \in S_{j,n}.$$

Choose j_0 such that $\frac{\epsilon}{2} 2^{2j_0-2} \geq 1$. Then, for all $j \geq j_0$ it holds that $\epsilon 2^{2j-2} - 1 \geq \frac{\epsilon}{2} 2^{2j_0-2}$. Now we have

$$\begin{aligned} E|r_n d(\hat{\eta}_n, \eta_0)|^p &\leq 2^{(j_0-1)p} P(r_n d(\hat{\eta}_n, \eta_0) \leq 2^{j_0-1}) + \sum_{j=j_0}^{\infty} 2^{jp} P(\hat{\eta}_n \in S_{j,n}) \\ &\leq 2^{(j_0-1)p} + \sum_{j=j_0}^{\infty} 2^{jp} P\left(\sup_{\eta \in S_{j,n}} (\mathbb{M}(\eta) - \mathbb{M}(\eta_0)) \geq -r_n^{-2}\right). \end{aligned}$$

We denote $W_n = \mathbb{M}(\eta) - \mathbb{M}(\eta_0)$. The second term on the right hand side is bounded by

$$\begin{aligned}
& \sum_{j=j_0}^{\infty} 2^{jp} P\left(\sup_{\eta \in S_{j,n}} (W_n(\eta) - W_n(\eta_0)) \geq -r_n^{-2} + \epsilon \frac{2^{2j-2}}{r_n^2} \right) \\
\leq & \sum_{j=j_0}^{\infty} 2^{jp} P\left(\sup_{\eta \in S_{j,n}} (W_n(\eta) - W_n(\eta_0)) \geq \frac{\frac{\epsilon}{2} 2^{2j-2}}{r_n^2} \right) \\
\leq & \sum_{j=j_0}^{\infty} 2^{jp} P\left(\sup_{\eta \in S_{j,n}} (W_n(\eta) - W_n(\eta_0)) \geq \left| \frac{\frac{\epsilon}{2} 2^{2j-2}}{r_n^2} \right|^{\eta p} \right) \\
\leq & \sum_{j=j_0}^{\infty} 2^{jp} \left| \frac{r_n^2}{\frac{\epsilon}{2} 2^{2j-2}} \frac{C_p \phi_n(2^j/r_n)}{\sqrt{n}} \right|^{\eta p} \\
\leq & \sum_{j=j_0}^{\infty} 2^{jp} \left| \frac{r_n^2}{\frac{\epsilon}{2} 2^{2j-2}} \frac{C_p 2^{j\alpha} \phi_n(1/r_n)}{\sqrt{n}} \right|^{\eta p} \\
\leq & \sum_{j=j_0}^{\infty} 2^{jp} \left| C_p \frac{2^{j\alpha}}{\frac{\epsilon}{2} 2^{2j-2}} \right|^{\eta p} \\
= & \sum_{j=j_0}^{\infty} \left| C_p \frac{2^{j(\alpha-2+\eta^{-1})}}{\frac{\epsilon}{8}} \right|^{\eta p}
\end{aligned}$$

Since $\alpha - 2 + \gamma^{-1} < 0$, this series is finite. Since j_0 only depends on ϵ, and ϵ and C_p are independent of η_0, the bound on $E|r_n d(\hat{\eta}_n, \eta_0)|^p$ is universal with respect to $\eta_0 \in \Xi$. ∎

Proof of Theorem 3.5.2: We are going to establish the minimax upper bound by showing that the LSE attains the minimax lower bound. The following proof approaches this problem in two steps: step 1, establish the consistency of the least squares estimator; and step 2, derive the rate of convergence of the mean squared errors of the LSE.

Consistency. We will show that inequality (3.5.1) holds. By conditions **(A1)** and **(A5)**,

$$\begin{aligned}
\mathbb{M}(\eta) - \mathbb{M}(\eta_0) &= (\alpha_0 - \alpha)^2 + (\beta_0 - \beta)^2 E[X^2(\theta_0)] + \beta^2 E[X(\theta_0) - X(\theta)]^2 \\
&\quad + 2\beta(\beta_0 - \beta) E[X(\theta_0)(X(\theta_0) - X(\theta))] \\
&\geq (\alpha_0 - \alpha)^2 + (\beta_0 - \beta)^2 \rho + \beta^2 C_l |\theta_0 - \theta|^{2H} \\
&\quad + 2\beta(\beta_0 - \beta) E[X(\theta_0)(X(\theta_0) - X(\theta))].
\end{aligned}$$

By condition **(A5)**, we can choose $\epsilon_1 < \frac{1}{2}\min\{\frac{\rho}{2b}, \frac{aC_l}{2}\}$ such that

$$\begin{aligned} 2\beta(\beta_0-\beta)E[X(\theta_0)(X(\theta_0)-X(\theta))] &\geq -2|\beta||\beta_0-\beta|\epsilon_1|\theta_0-\theta|^H \\ &\geq -|\beta|\epsilon_1\{|\beta_0-\beta|^2+|\theta_0-\theta|^{2H}\}. \end{aligned}$$

Then (3.5.1) is satisfied with $\epsilon = \max\{1, \rho/2, C_l b^2/2\}$. It follows that $\mathbb{M}(\eta)$ has a unique and well-separated minimizer at η_0.

By Theorem 3.2.3 (i) of van der Vaart and Wellner (1996) it suffices to show that $\mathbb{M}_n \xrightarrow{P} \mathbb{M}$ uniformly in $\Xi = [-b, b\,] \times \{\beta \in \mathbb{R} : a \leq |\beta| \leq b\} \times (0,1)$. For the uniform convergence, we only need to show that the class $\mathcal{F} = \{m_\eta : \eta \in \Xi\}$ is P-Glivenko Cantelli (P-GC). Note that almost all trajectories of X are Lipschitz of any order strictly less than H by condition **(A3)**. It is shown below that m_η is also Lipschitz in η. Thus the bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, L^1P)$ is finite and $\mathcal{F}$ is P-GC, by Theorems 2.7.11 and 2.4.1 of van der Vaart and Wellner (1996).

Rate of convergence. We will use Lemma 3.5.1 to give a second moment convergence rate of the LSE that is uniform over the parameter space. It is further shown that the LSE attains the optimal rates in the mean squared error sense. This establishes the minimax upper bound.

Since we have already proved (3.5.1), it is enough to prove (3.5.2). Let $\mathcal{M}_\delta \doteq \{m_\eta - m_{\eta_0} : \tilde{d}(\eta, \eta_0) < \delta\}$, where $\delta \in (0, 1]$. Because

$$\begin{aligned} m_\eta(X,Y) - m_{\eta_0}(X,Y) &= [Y-\alpha-\beta X(\theta)]^2 - [Y-\alpha_0-\beta_0 X(\theta_0)]^2 \\ &= (\alpha^2-\alpha_0^2) + \beta^2[X^2(\theta)-X^2(\theta_0)] + (\beta^2-\beta_0^2)X^2(\theta_0) \\ &\quad - 2Y(\alpha-\alpha_0) - 2\beta Y[X(\theta)-X(\theta_0)] - 2(\beta-\beta_0)YX(\theta_0) \\ &\quad + 2\alpha\beta[X(\theta)-X(\theta_0)] + 2\alpha(\beta-\beta_0)X(\theta_0) \\ &\quad + 2\beta_0(\alpha-\alpha_0)X(\theta_0) \\ &\leq |\alpha+\alpha_0||\alpha-\alpha_0| + \beta^2|X^2(\theta)-X^2(\theta_0)| \\ &\quad + X^2(\theta_0)|\beta+\beta_0||\beta-\beta_0| + 2|Y||\alpha-\alpha_0| \\ &\quad + 2|Y||\beta||X(\theta)-X(\theta_0)| + 2|Y||X(\theta_0)||\beta-\beta_0| \\ &\quad + 2|\alpha||\beta||X(\theta)-X(\theta_0)| + 2|\alpha||X(\theta_0)||\beta-\beta_0| \\ &\quad + 2|\beta_0||X(\theta_0)||\alpha-\alpha_0|, \end{aligned}$$

$\mathcal{M}_\delta$ has envelope

$$\begin{aligned} M_\delta(X,Y) \equiv\ & 2b\delta + b^2 \sup_{|t-s|^H<\delta} |X^2(t) - X^2(s)| \\ & + 2b \sup_{\theta_0} |X^2(\theta_0)|\delta + 2\sup_{\eta_0}|Y|\delta \\ & + 2b \sup_{\eta_0}|Y| \sup_{|t-s|^H<\delta} |X(t) - X(s)| \\ & + 2\sup_{\theta_0}|X(\theta_0)| \sup_{\eta_0}|Y|\delta + 2b^2 \sup_{|t-s|^H<\delta} |X(t) - X(s)| \\ & + 2b \sup_{\theta_0}|X(\theta_0)|\delta + 2b\sup_{\theta_0}|X(\theta_0)|\delta. \end{aligned} \tag{3.6.4}$$

By the boundedness of the parameter space Ξ,

$$|\alpha + \alpha_0| \le 2b, \quad |\beta + \beta_0| \le 2b.$$

By conditions **(A2)**, we have

$$E\sup_{\theta_0}|X(\theta_0)| < \infty, \text{ and } E\sup_{\theta_0}|X^2(\theta_0)| < \infty.$$

By condition **(A4)**,

$$\sup_{|t-s|^H<\delta} |X(t) - X(s)| \lesssim \delta.$$

By condition **(A6)**,

$$E\sup_{\eta_0}|Y| \le E\sup_{\eta_0}(|\alpha_0| + |\beta_0||X_(\theta_0)| + |\varepsilon|) < \infty.$$

Together with Hölder's inequality, we can show that all nine terms in (3.6.4) have the pth moment bounded by δ^p up to a constant independent of η_0, and thus $EM_\delta^p \lesssim \delta^p$.

Next we prove that m_η is "Lipschitz in parameter". Without loss of generality, for simplicity of notations, we assume that $\alpha = 0$ and $\beta = 1$. Noting that $m_\theta(X,Y) = (Y - X(\theta))^2$, we then have

$$\begin{aligned} |m_{\theta_1} - m_{\theta_2}| &= [Y - X(\theta_1) + Y - X(\theta_2)][X(\theta_2) - X(\theta_1)] \\ &\le 2(\sup_\theta |X(\theta)| + \sup_{\theta_0}|Y|)|X(\theta_1) - X(\theta_2)| \\ &\le L|\theta_1 - \theta_2|^\alpha, \end{aligned}$$

where $L = 2(\sup_\theta |X(\theta)| + \sup_{\theta_0}|Y|)\xi$. The second inequality follows from condition **(A3)**. By conditions **(A2)** and **(A6)**, L has moments of all orders. Consequently that the bracketing entropy integral $J_{[\ ]}(1, \mathcal{M}_\delta, L^2(P)$ is uniformly bounded as a function of $\delta \in (0,1]$,

see van der Vaart and Wellner (1996), p. 294. Using Theorem 2.14.2 of van der Vaart and Wellner (1996), we have

$$\| \, \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \|_{P,1} \lesssim J_{[\,]}(1, \mathcal{M}_\delta, L^2(P))(EM_\delta^2)^{1/2} \lesssim \delta$$

for all $\delta \in (0,1]$, where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ and the constants in the inequalities are universal. By Theorem 2.14.5 of van der Vaart and Wellner (1996), for all $p > 2$ and large n

$$\begin{aligned} \| \, \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \|_{P,p} &\lesssim \| \, \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \|_{P,1} + n^{-1/2+1/p} \, \|M_\delta\|_{P,p} \\ &\lesssim \delta + n^{-1/2+1/p} \, |EM_\delta^p|^{1/p} \\ &\lesssim \delta. \end{aligned}$$

The constants in the inequalities $\lesssim$ depend only on the value of p. By Lyapounov's inequality, (3.5.2) is also true for $1 \leq p \leq 2$. It follows from the previous lemma that the LSE converges in second moment uniformly w.r.t. η_0, i.e.

$$\sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \, L_s(\hat{\eta}, \eta_0) \lesssim n^{-1}, s = \alpha, \beta, \text{ and } \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \, L_\theta(\hat{\eta}, \eta_0) \lesssim n^{-1/H}.$$

It follows that

$$\begin{aligned} &\inf_{\tilde{\eta}} \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \, L_\alpha(\tilde{\eta}, \eta_0) \lesssim n^{-1}, \\ &\inf_{\tilde{\eta}} \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \, L_\beta(\tilde{\eta}, \eta_0) \lesssim n^{-1}, \text{ and} \\ &\inf_{\tilde{\eta}} \sup_{\eta_0 \in \Xi} \mathbb{E}_{\eta_0} \, L_\theta(\tilde{\eta}, \eta_0) \lesssim n^{-1/H}. \end{aligned}$$

∎

Chapter 4

Minimax lower bound for the sparse functional GLM

Generalized linear models provide a powerful tool for relating predictor variables to continuous or categorical responses McCullagh and Nelder (1989). It has long been a topic of interest to consider a functional predictor in the generalized linear regression model. Like functional linear regression models, functional generalized linear regression models are able to take into account the information contained in an entire curve when predicting a scalar response, while the latter can cope with situations where the outcome variable is not necessarily continuously distributed. See Chapter 2 for examples of functional generalized linear regression.

Here we consider the sparse functional generalized linear models. First proposed in Lindquist and McKeague (2009), the models were shown to have substantial value in practice. The authors gave two examples of applications using the functional logistic regression model, a special case of sparse functional GLM with binary outcomes. One example involves the gene expression profile data introduced in Section 1.1.1, with genome-wise microarray expression levels as the predictor and diagnosis of breast cancer as the binary outcome. The other example applies the sparse functional GLM model to the fMRI data introduced in Section 1.1.2, where time courses of fMRI signals at one voxel are the predictor functions and the resilient/non-resilient anxiety status is the binary outcome. It was shown that sparse

functional GLM is able to detect sensitive points with potential scientific meanings. The authors constructed a maximum likelihood estimator for the model and showed that the MLE $\hat{\eta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n)$ converges to a non-degenerate limiting distribution at component-wise rates $n^{1/2}$, $n^{1/2}$ and n, under the assumption that X is a two-sided Brownian motion in the neighborhood of the sensitive point θ_0. It remains unclear whether the MLE is rate-optimal.

In this chapter, we extend the result from the previous chapter and establish a lower bound on the minimax risk for estimating the sensitive point and the regression coefficients under milder conditions. It is shown that the MLE's weak convergence rate is of the same order as that of the minimax lower bound, which suggests the rate-optimality of the MLE. Section 4.1 specifies the sparse functional GLM and describes the maximum likelihood estimation procedure. Section 4.2 gives the list of conditions needed for the minimax lower bound. Section 4.3 presents the main result of this chapter, the minimax lower bound, and compares it to the weak convergence rate of the MLE. Finally, Section 4.4 gives the complete proof to the theorem.

4.1 Model specification and estimation

Suppose the data consist of independent, identically distributed pairs $\{(X_i, Y_i), i = 1, \ldots, n\}$ that are replicates of (X, Y). Again X is a stochastic process indexed by [0,1]. Here Y is a scalar response that could be non-Gaussian. Specifically we assume that conditional on X,

$$Y|X \sim Q_\lambda, \tag{4.1.1}$$

where $\{Q_\lambda : \lambda \in \mathbb{R}\}$ is an exponential family of probability measures with densities

$$dQ_\lambda/dQ_0 = \exp(\lambda y - \psi(\lambda)),$$

and

$$\lambda = \alpha + \beta X(\theta).$$

Here θ is again the sensitive point at which the value of X is associated with Y. The intercept α and the slope β are both scalars. The model is indexed by $\eta = (\alpha, \beta, \theta)$.

The maximum likelihood estimator of η is given by

$$\begin{aligned}\hat{\eta}_n &= (\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n) = \arg\max_{\eta} \mathbb{P}_n m_\eta \\ &= \arg\max_{\eta} \sum_{i=1}^{n} \{Y_i[\alpha + \beta X_i(\theta)] - \psi[\alpha + \beta X_i(\theta)]\}, \end{aligned} \tag{4.1.2}$$

where $\mathbb{P}_n$ is the empirical measure, and

$$m_\eta = Y[\alpha + \beta X(\theta)] - \psi[\alpha + \beta X(\theta)]$$

is the log-likelihood function. Again, the likelihood function is maximized via a profile estimate procedure. Specifically, for each fixed θ, a profile estimate of (α, β) is given by

$$(\hat{\alpha}_n(\theta), \hat{\beta}_n(\theta)) = \arg\max_{(\alpha,\beta)} \sum_{i=1}^{n} \{Y_i[\alpha + \beta X_i(\theta)] - \psi[\alpha + \beta X_i(\theta)]\}. \tag{4.1.3}$$

Then an estimate of θ is given by

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \{Y_i[\alpha + \beta X_i(\theta)] - \psi[\alpha + \beta X_i(\theta)]\}. \tag{4.1.4}$$

Finally $\hat{\eta}_n$ is obtained from

$$(\hat{\alpha}_n(\hat{\theta}_n), \hat{\beta}_n(\hat{\theta}_n), \hat{\theta}_n). \tag{4.1.5}$$

In practice, the profile estimate 4.1.3 can be obtained by standard software solving for GLM. These procedures are usually fast and efficient, so the entire procedure is tractable despite the large number of points on which observations are made.

4.2 Conditions

The sparse functional generalized linear regression model is indexed by $\eta = (\alpha, \beta, \theta) \in \Xi$. Again, we assume the parameter space

$$\Xi = [-b, b\,] \times \{\beta \in \mathbb{R} : a \leq |\beta| \leq b\} \times (0, 1).$$

Here $0 < a < b < \infty$ are constants. Also, the following assumptions on X are made to derive the minimax lower bound.

(A1) X has a generalized Hurst exponent $H \in (0, 1]$.

(A2) $\sup_{\theta\in(0,1)} E|X^2(\theta)| \equiv K < \infty$.

We also make the following assumptions on ψ:

(B1) There exists an increasing real function G on $\mathbb{R}^+$ such that

$$|\psi^{(3)}(\lambda + h)| \leq \psi^{(2)}(\lambda)G(|h|) \quad \forall\lambda \text{ and } h.$$

Without loss of generality we assume $G(0) \geq 1$.

(B2) For each $\epsilon > 0$ there exists a finite constant C_ϵ for which $\psi^{(2)}(\lambda) \leq C_\epsilon \exp(\epsilon\lambda^2)$ for all $\lambda \in \mathbb{R}$. Equivalently, $\psi^{(2)}(\lambda) \leq \exp(o(\lambda^2))$ as $|\lambda| \to \infty$.

As shown in Dou et al. (2010), conditions **(B1)** and **(B2)** on the ψ function imply that

$$h^2(Q_\lambda, Q_{\lambda+\delta}) \leq \delta^2\psi^{(2)}(\lambda)(1 + |\delta|)G(|\delta|) \quad \forall\lambda, \delta \in \mathbb{R}. \tag{4.2.1}$$

Here $h(P, Q)$ denotes the Hellinger distance between two probability measure P and Q. If both P and Q are dominated by some measure μ, with densities p and q, then $h^2(P, Q) = \mu(\sqrt{p} - \sqrt{q})^2$. The total variation distance is bounded by the Hellinger distance,

$$\|P - Q\|_{TV} \leq h(P, Q). \tag{4.2.2}$$

For product measures we use the bound

$$h^2(\otimes_{i\leq n}P_i, \otimes_{i\leq n}Q_i) \leq \Sigma_{i\leq n}h^2(P_i, Q_i). \tag{4.2.3}$$

4.3 Minimax lower bound for sparse functional GLM

In this section we will derive the lower bound on the minimax risk of estimating the parameters in sparse functional GLM. We will again use Lemma (3.4.1) to associate the minimax lower bound with the total variation affinity between two conditional distributions, $\|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\|$, that correspond to two simple hypotheses. The generality of exponential families, in particular the functional form of $\psi(\cdot)$, makes it more complicated to bound the total variation affinity than the similar procedure in Chapter 3. We approach the problem by first bounding the affinity by the Hellinger distance between $\mathbb{Q}_{n,1}$ and $\mathbb{Q}_{n,2}$, using

inequality (4.2.2), and then bounding the Hellinger distance using inequality (4.2.1). The following theorem gives the minimax lower bound. Details of the proof can be found in Section 4.4.

Theorem 4.3.1. *Suppose conditions **(A1)**,**(A2)**, **(B1)** and **(B2)** hold, then the minimax risk of estimating η over Ξ satisfies*

$$\begin{aligned}
&\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_\eta \; L_\alpha(\tilde{\eta}_n, \eta) \geq C_2 n^{-1}, \\
&\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_\eta \; L_\beta(\tilde{\eta}_n, \eta) \geq C_2 n^{-1}, \text{ and} \\
&\inf_{\tilde{\eta}_n} \sup_{\eta \in \Xi} \mathbb{E}_\eta \; L_\theta(\tilde{\eta}_n, \eta) \geq C_2 n^{-1/H},
\end{aligned}$$

where the supremums are taken over the parameter space Ξ and the infimums are taken over any estimator of the form $\tilde{\eta}_n = (\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\theta}_n)$, of η. $C_2 > 0$ only depends on the second moment structure of X and the parameter space Ξ.

The theorem shows that the minimax rate for estimating θ_0 is at most $n^{1/(2H)}$, which is faster or equal to the usual parametric rate $n^{1/2}$. The rougher the trajectories of X are, the quicker can $\hat{\theta}_n$ converge to θ_0. This result is in consistency with the minimax lower bound for sparse functional linear regression derived in Section 3.4, but in contrast to the result in Dou et al. (2010), where the convergence rate of the estimator of the slope function β is faster when the covariance function of X and β is smoother. This is not surprising because when the impact of the predictor function is spread across the index space, the smoothness of its paths will enable us to "borrow" information from the observations in the adjacent neighborhood. On the contrary, when the impact is sparse in the index space, the roughness of the predictor function's trajectories makes it easier to identify the sensitive point.

4.4 Proof

We use the same notations as in Section 3.6, except that here Q_λ stands for a density from an exponential family rather than a Gaussian distribution. In addition, when we want to indicate that a bound involving constants $c, C, C_1, \ldots$ holds uniformly over all models indexed by a set of parameters Ξ, we write $c(\Xi), C(\Xi), C_1(\Xi), \ldots$.

4.4.1 Proof of the minimax lower bound

Proof of Theorem 4.3.1: Select $\eta_1 = (\alpha_1, \beta_1, \theta_1)$ and $\eta_2 = (\alpha_2, \beta_2, \theta_2)$ such that $|\alpha_2 - \alpha_1| = cm^{-1}, |\beta_2 - \beta_1| = cm^{-1}$ and $|\theta_2 - \theta_1| = ck^{-1}$. Here $c > 0$ is a constant while m and k go to infinity as $n \to \infty$. Then Lemma 3.4.1 implies that

$$\begin{aligned} \sup_{\eta \in \Xi} \mathbb{E}_\eta L_\theta(\tilde{\eta}, \eta) &\geq \frac{1}{4} c^2 k^{-2} \mathbb{P}_{n,X} \|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\|, \text{ and} \\ \sup_{\eta \in \Xi} \mathbb{E}_\eta L_s(\tilde{\eta}, \eta) &\geq \frac{1}{4} c^2 m^{-2} \mathbb{P}_{n,X} \|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\|, \ s = \alpha, \beta. \end{aligned} \tag{4.4.1}$$

We want to show that $\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\|$ is bounded away from 0 as n goes to infinity. It suffices to show that $\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV}$ is bounded away from 1. Define $\lambda_i^j = \alpha_j + \beta_j X_i(\theta_j), i = 1, \ldots, n, j = 1, 2$. Then we have

$$\begin{aligned} \mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV} &\leq \mathbb{P}_{n,X}\left[\Sigma_{i \leq n} h^2(Q_{\lambda_i^1}, Q_{\lambda_i^2})\right]^{1/2} \\ &\leq \left[\mathbb{P}_{n,X}\Sigma_{i \leq n} h^2(Q_{\lambda_i^1}, Q_{\lambda_i^2})\right]^{1/2}. \end{aligned}$$

The first inequality follows from (4.2.2) and (4.2.3). The second one uses the Cauchy–Schwarz inequality. By inequality (4.2.1),

$$h^2(Q_{\lambda_i^1}, Q_{\lambda_i^2}) \leq C(\Xi)|\lambda_i^1 - \lambda_i^2|^2.$$

Therefore,

$$\begin{aligned} [\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV}]^2 &\leq C(\Xi)\mathbb{P}_{n,X}\Sigma_{i \leq n}\left[(\alpha_1 + \beta_1 X_i(\theta_1) - \alpha_2 - \beta_2 X_i(\theta_2))^2\right] \\ &\leq C(\Xi)\mathbb{P}_{n,X}\Sigma_{i \leq n}[2|\alpha_1 - \alpha_2|^2 + 4|X_i(\theta_1)|^2 \cdot |\beta_1 - \beta_2|^2 + \\ &\quad + 4|\beta_2|^2 \cdot |X_i(\theta_1) - X_i(\theta_2)|^2] \\ &\leq C(\Xi)n\left[2|\alpha_1 - \alpha_2|^2 + 4K \cdot |\beta_1 - \beta_2|^2 + 4b^2 \cdot C_u|\theta_1 - \theta_2|^{2H}\right] \\ &\leq C(\Xi)n \cdot (2c^2m^{-2} + 4Kc^2m^{-2} + 4C_u b^2 c^{2H} k^{-2H}). \end{aligned}$$

The second inequality uses twice the fact that $|a + b|^2 \leq 2|a|^2 + 2|b|^2$, for all $a, b \in \mathbb{R}$. The third inequality follows from conditions **(A1)** and **(A2)**, and the upper bound on $|\beta|$ in the parameter space Ξ. If we choose $m = n^{1/2}$, $k = n^{1/(2H)}$ and $0 < c \leq \min\{(\pi\sigma^2/(2+4K))^{1/2},$

$(\pi\sigma^2/(4C_u b^2))^{1/(2H)}\}/C(\Xi)$, then

$$\begin{aligned}
\mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} \wedge \mathbb{Q}_{n,2}\| &= 1 - \mathbb{P}_{n,X}\|\mathbb{Q}_{n,1} - \mathbb{Q}_{n,2}\|_{TV} \\
&\geq 1 - \left[C(\Xi)\ n \cdot (2c^2 n^{-1} + 4Kc^2 n^{-1} + 4C_u b^2 c^{2H} n^{-1})\right]^{1/2} \\
&= 1 - \left[C(\Xi)\left((1+2K)c^2 + 2C_u b^2 c^{2H}\right)/(\pi\sigma^2)\right]^{1/2} > 0.
\end{aligned}$$

Denoting $C_1 = c^2\left(1 - \left[C(\Xi)\left((1+2K)c^2 + 2C_u b^2 c^{2H}\right)/(\pi\sigma^2)\right]^{1/2}\right)/4$ and recalling (4.4.1), we have proved Theorem 3.4.2. ∎

Chapter 5

Simulation studies for sparse functional regression models

In this chapter we report the results of five simulation studies. They assess the finite-sample performance of the least squares estimator and the maximum likelihood estimator for sparse functional linear regression and sparse functional GLM, respectively, and compare them to other estimators in terms of the mean squared errors. The first simulation illustrates the behavior of the LSE for sparse functional linear regression with different sample sizes and different values of H. We want to see if the lower bound derived in Theorem 3.4.2 is of the same order as the upper bound given by the LSE. The second simulation compares the performance of the least squares estimator with two other estimators, using different sample sizes and different values of H. One of the other estimators is derived from the lasso and the other from the commonly used functional linear regression model (2.1.3). The third and fourth simulations examines the MSE of the MLE for sparse functional GLM and compare it to that of the lasso based estimator and the functional GLM-based estimator, respectively. The last one considers the case where the data are generated from a functional linear model with spike-shaped regression functions, and compares the performance of the LSE to that of the lasso- and FLR-based estimators

5.1 LSE for the sparse functional linear model

In this simulation we generate pairs $(X_i, Y_i), i = 1, \ldots, n$ from the sparse functional linear regression model (2.2.1) and evaluate the performance of the LSEs, $\hat{\beta}_n$ and $\hat{\theta}_n$, with varying sample sizes and different values of H. We want to see if their MSEs indeed converge at rate n and $n^{1/H}$, i.e.

$$n \cdot \mathbb{E}|\hat{\beta}_n - \beta_0|^2 = O_p(1) \quad \text{and} \tag{5.1.1}$$

$$n^{1/H} \cdot \mathbb{E}|\hat{\theta}_n - \theta_0|^2 = O_p(1). \tag{5.1.2}$$

Because a fBm is convenient to simulate and it satisfies all the conditions in Chapter 3, we generate fBms as predictor processes X_i, using the `R` 2.11.1 function `fbmSim` in the library `fArma`, on a uniform grid of 200 points over the [0,1] interval. To capture the asymptotic behavior of $\hat{\eta}_n$, three different sample sizes $n = 30, 50$, and 100 are considered. Because the minimax rates for $\hat{\theta}_n$ depend on the Hurst exponent H, we also consider 50 values of the Hurst exponent H, equally spaced in (0,1). The scalar responses Y_i are generated from the sparse functional linear regression model (2.2.1) with $\alpha_0 = 0, \beta_0 = 1, \theta_0 = 0.5$ and $\sigma = 0.3$.

Least squares estimators are fitted for each of the 1000 simulated samples. The MSEs of $\hat{\beta}$ and $\hat{\theta}$ are approximated by an average of the squared error loss. To evaluate the convergence rates of the estimators, the MSEs are multiplied by n and $n^{1/H}$, respectively, and plotted against H (Figures 5.1 and 5.2). To remove the Monte Carlo errors, we smoothed the curves using the lowess method (span = 1/3). When H gets very small, the resolution of the simulated trajectories may not be adequate, so only the results for $H > 0.2$ are displayed. The dashed line is the constant C_1 defined in Section 3.6.3, as a function of H, with $b = \beta_0 = 1$ and $c = \min\{(\pi\sigma^2/(2+4K))^{1/2}, (\pi\sigma^2/(4C_u b^2))^{1/(2H)}\}$. The variance-covariance structure of fBMs implies that $C_l = C_u = K = 1$. A closer look at C_1, given in Figure 5.3, shows that it is positive.

It can be seen first that the MSEs of $\hat{\beta}$ and $\hat{\theta}$ are indeed above the given lower bound, which verifies the validity of our result. Secondly, the estimators converge quickly as n reaches 100, indicated by the fact that the MSE curves corresponding to $n = 50$ and $n = 100$ are already very close. The third observation is that, as H approaches 1, the MSEs increase substantially. This is not surprising since the larger the H, the smoother the paths

of X, and the harder it is to identify θ_0 and estimate β_0. The increasing rescaled MSE of $\hat{\theta}_n$ as H gets very small is caused by the rescaling, rather than inadequate resolution, as we have also tried doubling the number of grid points for generating X and obtained similar results. Finally, we see that the constant C_1 is of a much smaller scale than the MSEs. This may be caused by the choice of X being fBM, which is a very special case compared to the general class of random functions we allow in the theorem. Also the choice of $c \leq \min\{(\pi\sigma^2/(2+4K))^{1/2}, (\pi\sigma^2/(4C_u b^2))^{1/(2H)}\}$ might be a strict requirement. It may be possible to find a bigger c that still makes C_1 positive.

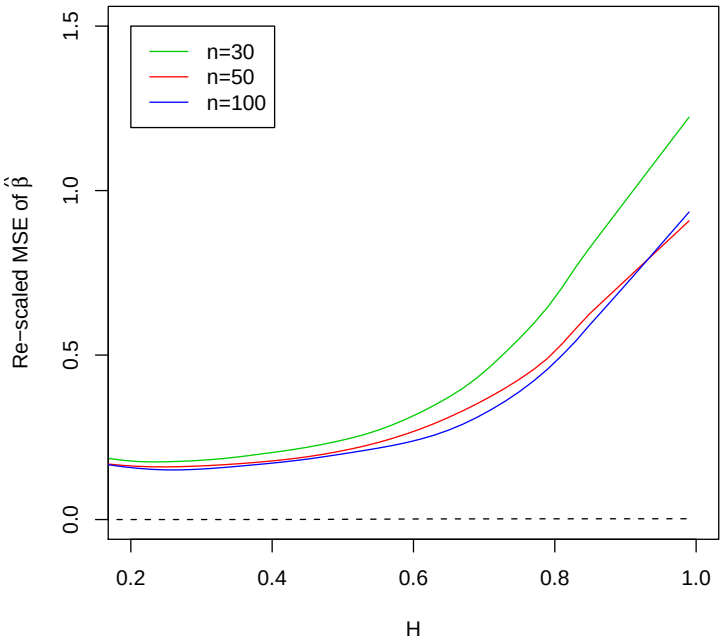


Figure 5.1: Empirical MSEs of $\hat{\beta}_n$ for sparse functional linear model, multiplied by n. The dashed line is C_1 changing with H. The MSEs are greater than the constant, indicating the minimax lower bound is valid.

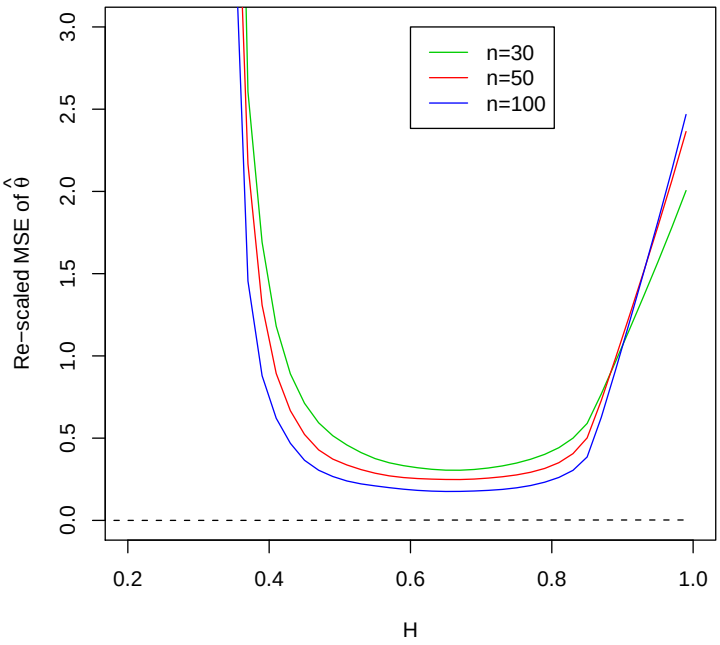


Figure 5.2: Empirical MSEs of $\hat{\theta}_n$ for sparse functional linear model, multiplied by $n^{1/H}$. The dashed line is C_1 changing with H. The MSEs are greater than the constant, indicating the minimax lower bound is valid.

5.2 Comparison of the LSE to the lasso and the FLR estimators

We next compare the LSE with two other estimators derived from the lasso and the estimator of the slope function in the FLR model. Since there has not been any estimator proposed for the sparse functional linear model other than the LSE, to our knowledge, we choose to compare it with two alternatives that most naturally come to mind. The lasso has been used for variable selection extensively. Since in reality the functional predictor X is observed on discrete points, estimating the sensitive point is similar to a variable selection problem. The functional linear regression model, as we have introduced in Chapter 2, is commonly used to associate functional predictors to scalar responses. We want to see if it

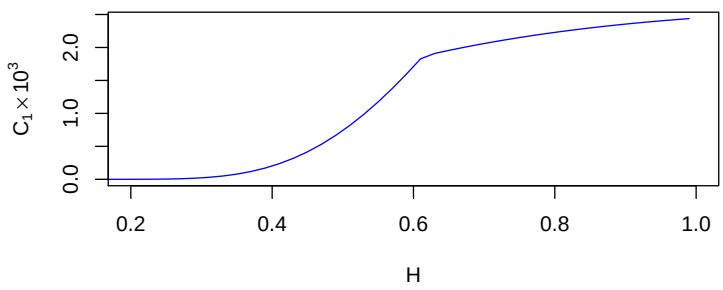


Figure 5.3: C_1 given in Theorem 3.4.2, as a function of H, in units of 10^{-3}. The constant is indeed positive but not sharp enough, compared to the MSEs of the estimates, which could be the result of the choice of X or the choice of c.

still works well in the situation where the impact of the functional explanatory variable is via its value at a sensitive point.

The data are generated in the same way as in Section 5.1. For the lasso, we view the values of X at the discrete 200 points as 200 predictor variables. The initial selection of the lasso is used as the estimate of θ. If there are multiple initial selections, the smallest point in (0,1) is used. For the FLR estimator, we use (2.1.3) as the working model and use the maximizer of the estimated $\beta(\cdot)$ as the estimate of θ.

To fit the lasso, we use the coordinate descent algorithm implemented in the `R` package `glmnet` (Friedman et al., 2010). For the FLR estimator, we use the `R` 2.11.1 function `fglm` in the `MFDF` package, which implements the procedure proposed by Dou et al. (2010). The predictor function $X(\cdot)$ is expressed with a B-spline basis of order 4 (piecewise cubic), with the uniform grid of observation times used as the knots. β is estimated using functional principle component analysis, and its roughness is controlled by how many functional PCAs are used, chosen to minimize the integrated squared error loss. Three sample sizes, $n = 30, 50, 100$ are again considered. The MSE of the three estimators are multiplied by $n^{1/H}$, lowess-smoothed (span = 1/3) and plotted against H (Figures 5.4, 5.5, and 5.6). In the labels we use "sparse FLR" to stand for the sparse functional linear model and "FLR" for the functional linear regression.

We can see from the graphs that the MSEs of the lasso-based estimators are almost identical with the least squares estimates in terms of their mean squared errors, while the MSE of the maximizer of $\beta(\cdot)$ is much higher as expected, since the data are generated from the sparse functional linear model and the functional linear model is misspecified. This does not contradict with the result of Hall and Horowitz (2007), since the minimax rate in (2.3.8) was given in the squared integral loss function, which reflects the risk of estimating the entire curve of the slope function instead of a sensitive point. The results show that, by assuming the impact of the functional predictor is spread across the interval, the classic FLR model cannot well estimate the sensitive point if the data were generated from the sparse functional GLM. Another implication is that, although the lasso is a variable selection procedure by design, it can be used to estimate the sensitive point in practice, where the continuous predictor process is almost always measured at discrete points. Since the coordinate descent algorithm is fast and efficient, it could be a much cheaper alternative to the LSE in terms of computational cost. On the other hand, the sparse functional linear model can be used for doing inference on θ_0 with its theoretical properties.

5.3 MLE for the sparse functional GLM

In this simulation we generate pairs $(X_i, Y_i), i = 1, \ldots, n$ from the sparse functional logistic model (2.2.3) and evaluate the performance of the MLEs, $\hat{\beta}_n$ and $\hat{\theta}_n$, with varying sample sizes and different values of H. Here we want to see if the lower bound in Theorem 4.3.1 is valid, i.e.

$$n \cdot \mathbb{E}|\hat{\beta}_n - \beta_0|^2 \geq C_2 \quad \text{and} \tag{5.3.1}$$

$$n^{1/H} \cdot \mathbb{E}|\hat{\theta}_n - \theta_0|^2 \geq C_2. \tag{5.3.2}$$

For the logistic distribution,

$$\psi(\lambda) = \log(1 + e^{\lambda}),$$

and

$$\psi^{(2)}(\lambda) = \frac{e^{\lambda}}{(1 + e^{\lambda})^2} < 1.$$

So we can take C_2 to be a positive constant. We choose $C_2 = 0.01$.

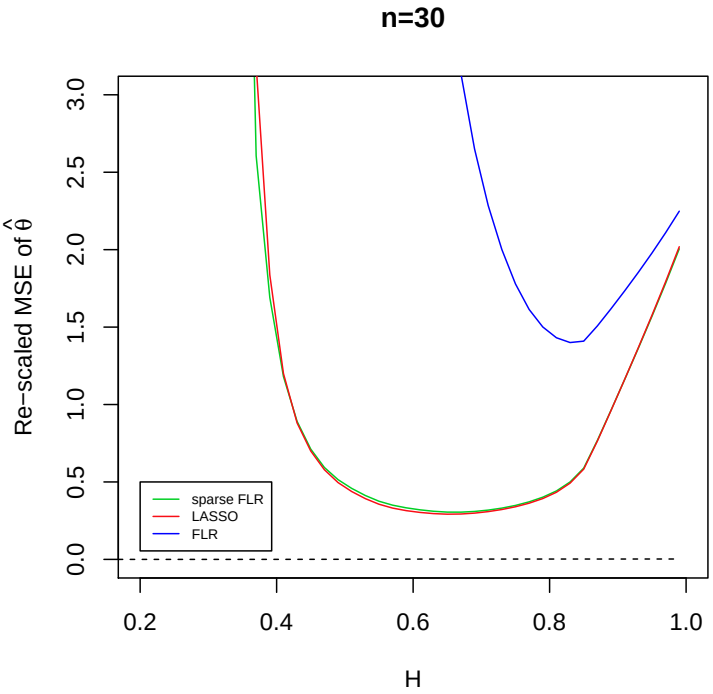


Figure 5.4: Empirical MSEs of $\hat{\theta}_n$ from the sparse functional linear model, the lasso and the functional GLM, multiplied by $n^{1/H}$, $n = 30$. The lasso and the sparse functional GLM have similar performance, but the functional linear model has higher MSEs.

Again fBms are used as predictor processes X_i, generated on a uniform grid of 200 points over the [0,1] interval for three different sample sizes $n = 30, 50$, and 100. 50 values of the Hurst exponent H equally spaced in (0,1) are considered. The scalar responses Y_i are generated from the sparse functional logistic model (2.2.3) with $\alpha_0 = 0, \beta_0 = 1$, and $\theta_0 = 0.5$.

Maximum likelihood estimators are fitted for each of the 1000 simulated samples. The MSEs of $\hat{\beta}$ and $\hat{\theta}$ are approximated by the average of the squared error loss, multiplied by n and $n^{1/H}$, respectively, and plotted against H (Figure 5.7). To remove the Monte Carlo errors, we smoothed the MSE plots using the lowess method (span = 1/3). When H gets close to 0 or 1, the behavior of the estimators may become irregular, so only the results for $0.2 < H < 0.8$ are displayed.

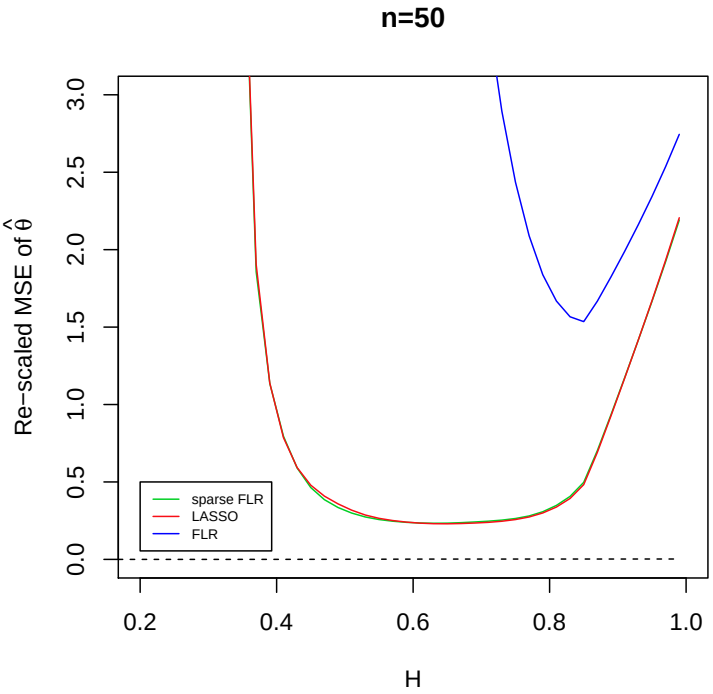


Figure 5.5: Empirical MSEs of $\hat{\theta}_n$ from the sparse functional linear model, the lasso and the functional GLM, multiplied by $n^{1/H}$, $n = 50$. The lasso and the sparse functional linear model have similar performance, but the functional linear model has higher MSEs.

It can be seen that the MSEs of $\hat{\beta}$ and $\hat{\theta}$ are indeed above the lower bound (dashed lines), therefore the asymptotic order of the minimax lower bound is valid. Also can be seen is that the MLEs converge slower than the LSEs for the sparse functional linear regression model, since there is a much larger difference in the re-scaled MSEs when the sample size is 50 and 100 than in the sparse functional linear regression case. Furthermore, the magnitude of the re-scaled MSEs is much larger than that in the sparse functional linear regression case, which also indicates a slower convergence of the MLEs. Finally, as H gets larger, the rescaled MSE of $\hat{\beta}_n$ increases as in the sparse functional linear model case, but the rescaled MSE of $\hat{\theta}_n$ is decreasing, which is again caused by the rescaling.

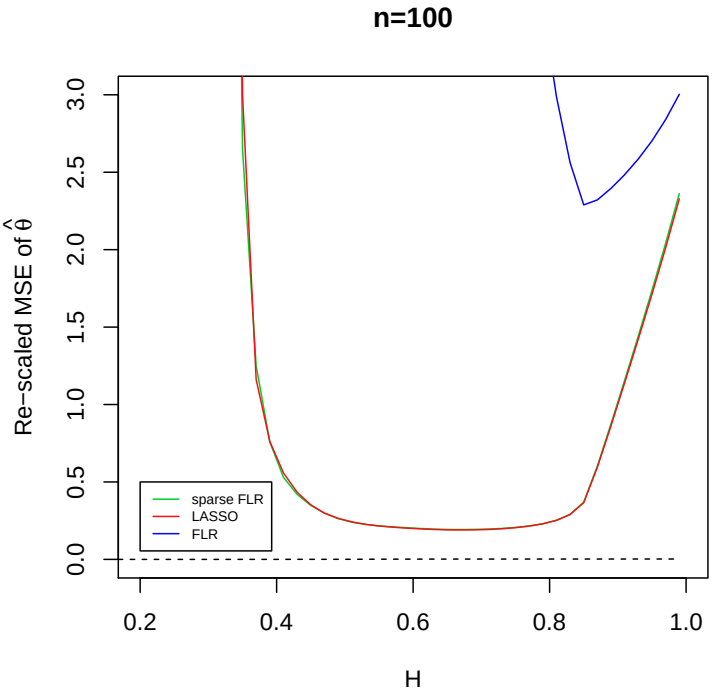


Figure 5.6: Empirical MSEs of $\hat{\theta}_n$ from the sparse functional linear model, the lasso and the functional GLM, multiplied by $n^{1/H}$, $n = 100$. The lasso and the sparse functional linear model have similar performance, but the functional linear model has higher MSEs.

5.4 Comparison of the MLE to the lasso and the functional GLM estimators

We next compare the MLE with two other estimators derived from the lasso and the functional logistic regression, which is a special case of the functional generalized linear regression estimator proposed by Dou et al. (2010). The lasso-based and functional GLM-based estimators are obtained in a similar way to the approach in Section 5.2. Only here we use the logit link function and the binomial outcome distribution. Three sample sizes, $n = 30, 50, 100$ are again considered. The MSE of the three estimators are re-scaled, lowess-smoothed (span = 1/3) and plotted against H (Figures 5.8 and 5.9). In the labels we use "sparse FGLR" to stand for the sparse functional GLM and "FGLR" for the functional

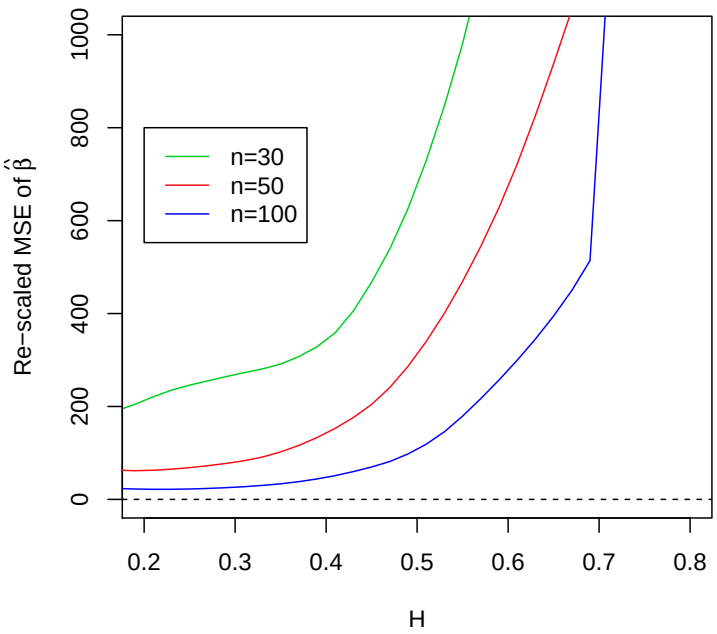
GLM.

Our results show that, just like the sparse functional linear model, the estimates based on the lasso is almost identical with the least squares estimates in terms of their mean squared errors, while the functional GLM estimate is outperformed by the others. This is more obvious when n becomes larger. The functional GLM cannot well estimate the sensitive point if the data were generated from the sparse functional GLM. It is also suggested that the lasso-based estimator is a reasonable alternative to the MLE when estimating the parameters in sparse functional GLM.

5.5 Misspecification by a functional linear model

In this simulation, we generate data from the functional linear model (2.1.3) with a spike-shaped regression function, and treat the sparse functional linear model (2.2.1) as the working model. We will compare the performance of the LSE to that of the lasso- and FLR-based estimators. The spike-shaped regression function $\beta(t)$ is taken as a Gaussian pdf centered at $t = 0.5$. We consider two separate standard deviations for the spike function, $\sigma = 0.01$ and 0.03, respectively. In each case, we specify the sample size $n = 40$, $\alpha_0 = 0$, and the error standard deviation $\sigma_0 = 0.3$. Again, functional Brownian motions are generated as the functional predictors, with a series of Hurst exponents equally spaced in (0,1).

Figure 5.10 shows the results of fitting the sparse functional model, along with the lasso- and FLR-based estimators. It can be seen that if the slope function is more spread out, the error of estimating β_0 in the sparse functional linear model is higher, since the assumption of the sparse functional linear model is more violated. In each case, the estimates based on the sparse functional linear model and the lasso still outperform the ones based on the FLR model. This suggests that, although the sparse functional model is misspecified, it may still be a better choice to capture the spike-shaped feature of $\beta(t)$. As shown in the simulation from Lindquist and McKeague (2009), the functional linear model wrongly suggests that the influence of the predictor is substantial over the whole time course, thus might lose information when estimating the center of the spike.



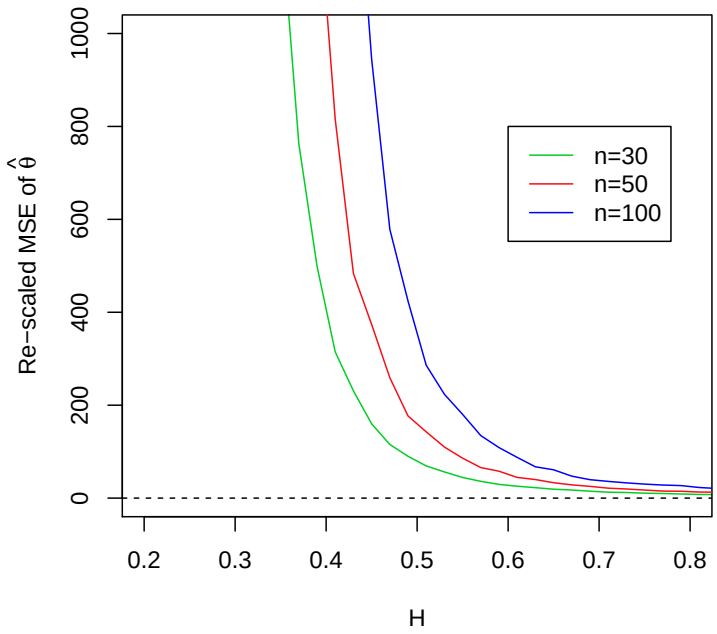


Figure 5.7: Empirical MSEs of $\hat{\beta}_n$ and $\hat{\theta}$ for sparse functional GLM, multiplied by n and $n^{1/H}$, respectively.

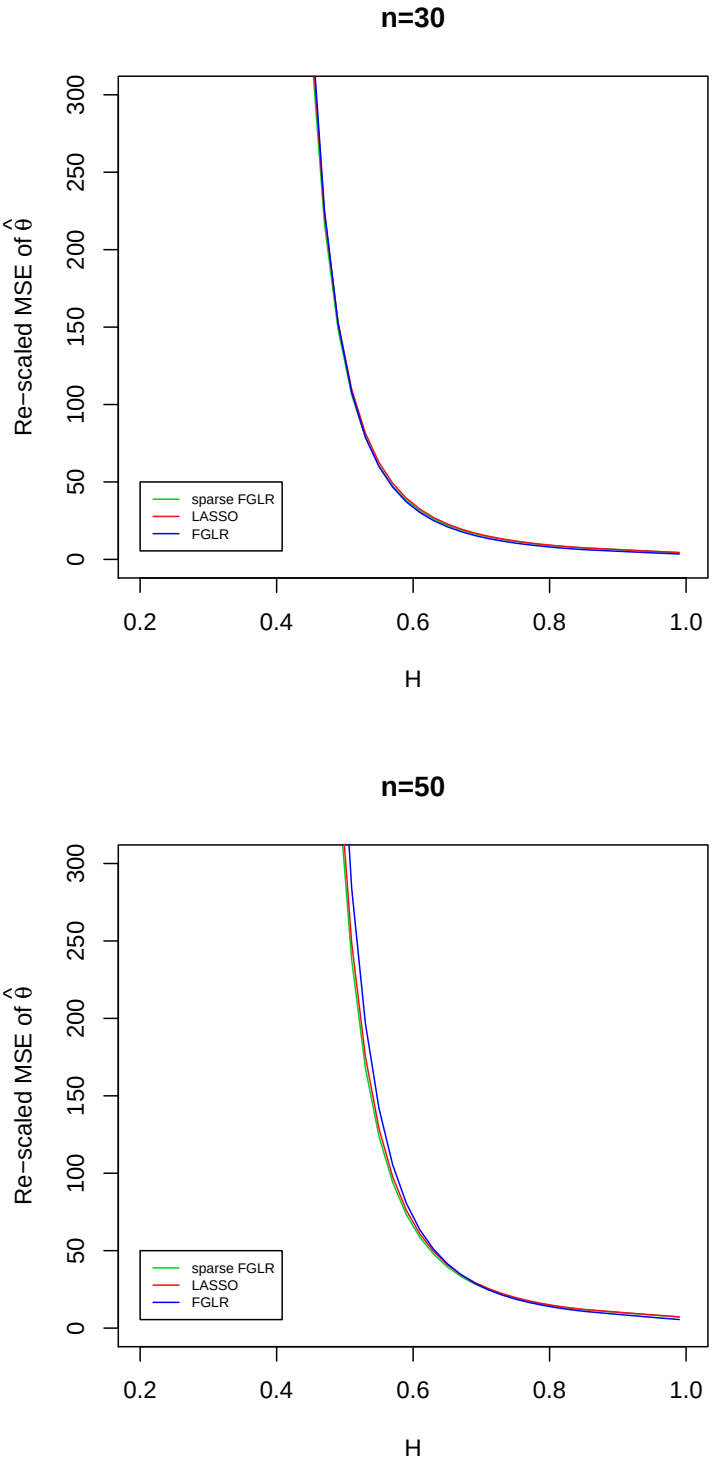


Figure 5.8: Empirical MSEs of $\hat{\theta}_n$ from sparse functional GLM, the lasso and functional GLM, multiplied by $n^{1/H}$, $n = 30$ and 50. The lasso and the sparse functional linear model have similar performance, but the functional GLM has higher MSEs.

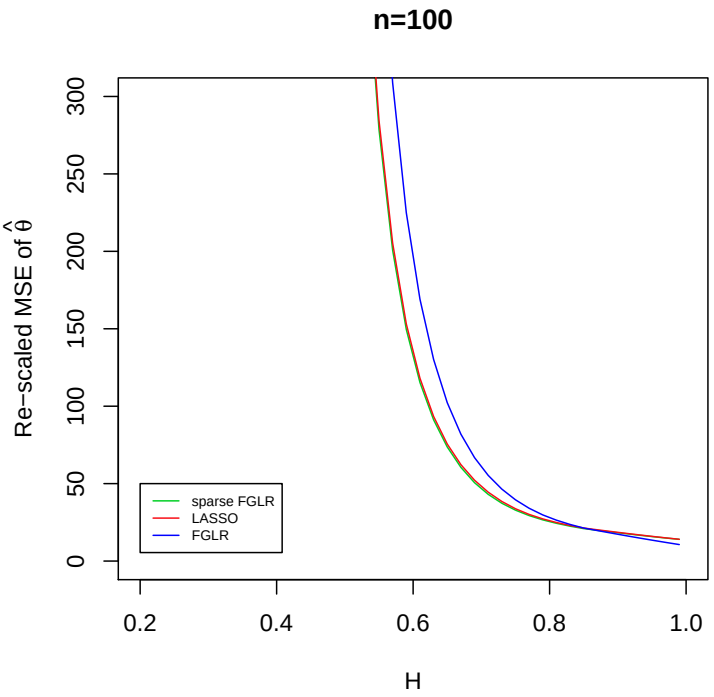


Figure 5.9: Empirical MSEs of $\hat{\theta}_n$ from sparse functional GLM, the lasso and functional GLM, multiplied by $n^{1/H}$, $n = 100$. The lasso and the sparse functional GLM have similar performance, but the functional GLM has higher MSEs.

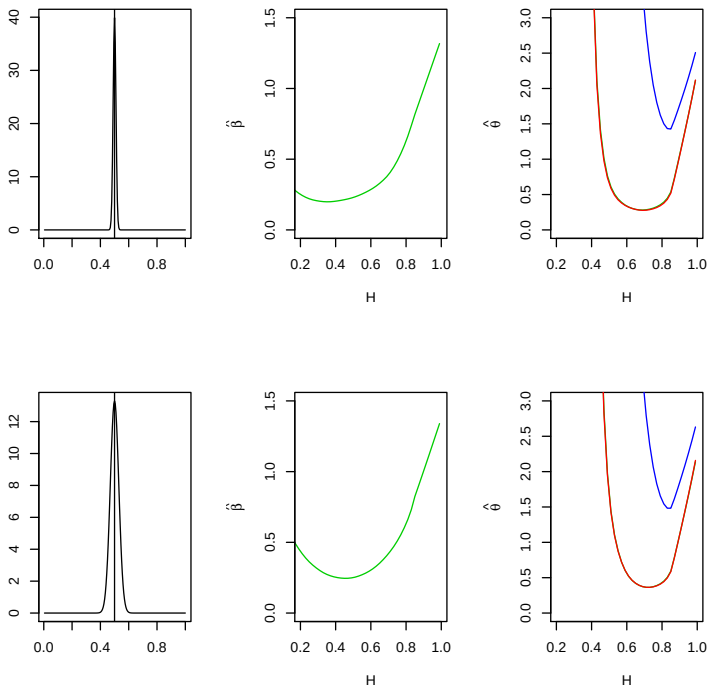


Figure 5.10: The regression function $\beta(t)$ is taken as two separate Gaussian pdfs centered at $t = 0.5$, with standard deviations 0.01 and 0.03, respectively (first column). The smoothed MSEs of the estimated scalar slope $\hat{\beta}_n$ in the sparse functional linear model (second column), multiplied by n. The smoothed MSEs of the estimated $\hat{\theta}_n$, based on the LSE (green), the lasso-based estimates (red), and the FLR-based estimates (blue), multiplied by $n^{1/H}$ (third column). The panels can be compared to Figures 5.3 and 5.4.

Chapter 6

Contaminated sparse functional GLM

We have seen in Chapter 4 that sparse functional GLM is a convenient tool to estimate sensitive points at which functional predictors impact non-Gaussian scalar outcomes. One limitation of the sparse functional GLM, however, is that it assumes a fixed sensitive point. In reality this might not be the case and the sensitive point could be random and prone to contamination. For example, the timing of psychological activities reflected by fMRI signals might be affected by personal aging, disorders and pathology, such as cerebrovascular diseases (D'Esposito et al., 2003).

In face of this complication, we extend the sparse functional GLM to a contaminated sparse functional GLM in this chapter to allow for random sensitive time point. We will construct an estimating procedure based on a Monte Carlo EM algorithm. We will evaluate the performance of the proposed estimator in several simulation studies and a real data analysis. It is also shown that the maximum marginal likelihood estimator converges at the parametric rate, $n^{1/2}$, if the contamination distribution is smooth, in contrast to the faster $n^{1/(2H)}$ rate of the MLE for the sparse functional GLM.

The rest of the chapter is organized as follows. Section 6.1 describes the motivation to proposing the contaminated sparse functional GLM in detail. Section 6.2 specifies the model structure and assumptions. We also discuss the connection between this model and

the latent variable models, especially generalized linear mixed models (GLMM). Section 6.4 describes the proposed Monte-Carlo EM estimation procedure and discusses the numerical issues that could occur in reality. Section 6.3 presents the asymptotic properties of the MLE. Section 6.5 gives the detailed proofs to the theory. In Chapter 7, we will present the results of four simulation studies that illustrate the finite-sample behaviors of the MCEM estimator. The proposed method is applied to a real data analysis in Chapter 8.

6.1 Motivation

Functional magnetic resonance imaging (fMRI) measures neuronal activity indirectly, through the blood-oxygen-level-dependent (BOLD) signal. The BOLD signal depends on neurovascular coupling – the processes by which neural activity influences the haemodynamic properties of the surrounding vasculature. As pointed out by D'Esposito et al. (2003), there is empirical evidence that these mechanisms might be altered in normal ageing and disease. So, interpretation of BOLD fMRI studies of individuals with different ages or pathology might be more challenging than is commonly acknowledged.

For example, in one fMRI study, the severe extra-cranial carotid stenosis in a patient without MRI evidence of an infarct led to neurovascular uncoupling that presented as a negative BOLD signal response during performance of a simple motor task. Both the level and the onset time of the BOLD sinal in response to a finger-tapping task in motor cortex on the side carotid stenosis have been altered by the effect of cerebrovascular pathology. Therefore, the onset time of brain activity indicated by the BOLD signal in the fMRI study presented in Section 1.1.2 could be susceptible to subject-specific effects. It is necessary to consider a model that incorporates random sensitive points.

6.2 Model specification

To meet the previously mentioned practical need and deal with randomly distributed sensitive points, we propose the *contaminated* sparse functional GLM. We assume the data consist of independent, identically distributed pairs (X_i, Y_i), $i = 1, \ldots, n$, which are i.i.d. replicates of (X, Y), where Y is a scalar response that could be non-Gaussian, and X is

a stochastic process on [0,1]. The contaminated sparse functional GLM is structured as follows.

$$Y|X \sim Q_\lambda, \quad \lambda = \alpha + \beta X(\tau), \tag{6.2.1}$$

$$\tau \sim \theta + W, \text{ given } \theta + W \in [0,1]. \tag{6.2.2}$$

Here Y is a scalar response, X is a functional predictor, and Q_λ is a distribution from the exponential family defined as before. θ is the contamination-free sensitive point that we want to estimate. W is a random contamination that is independent of X, with a known density $p_W(\cdot)$ that is smooth, unimodal and symmetric at 0. An example of such density would be the Gaussian density. Since the contaminated sensitive point τ cannot be outside of [0,1], we assume its distribution is the same as $\theta + W$ truncated to [0,1]. Thus the conditional density of Y given X and τ only depends on α and β, denoted as $p_{\alpha,\beta}(Y|X,\tau)$, and the density of τ only depends on θ, denoted as $p_\theta(\tau)$, and the contaminated sparse functional GLM is indexed by $\eta = (\alpha, \beta, \theta)$. From the previous assumptions, we have

$$p_{\alpha,\beta}(y|X,\tau) = \exp(\lambda y - \psi(\lambda)), \tag{6.2.3}$$

$$p_\theta(x) = \frac{p_W(x-\theta) \cdot \mathbf{1}_{[0,1]}(x)}{\int_0^1 p_W(x-\theta)\ dx}. \tag{6.2.4}$$

Inspired by the estimation procedure for the sparse functional GLM, we propose the maximum likelihood estimator of η in the contaminated sparse functional GLM, given by

$$\hat{\eta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n) = \arg\max_\eta \mathbb{M}_n(\alpha, \beta, \theta), \tag{6.2.5}$$

where the log likelihood $\mathbb{M}_n(\alpha, \beta, \theta) = \mathbb{P}_n m(\alpha, \beta, \theta)$,

$$m(\alpha, \beta, \theta) = \log \int_0^1 p_{\alpha,\beta}(Y|X,\tau) p_\theta(\tau) d\tau, \tag{6.2.6}$$

and $\mathbb{P}_n$ is the empirical distribution of the data on (X, Y).

6.2.1 Connection to generalized latent variable models

We notice that the contaminated sparse functional GLM has some similarity to latent variable models, since the contaminated sensitive points $\tau_i, i = 1, \ldots, n$ are unobservable in reality. In fact, generalized latent variable models have long been established and applied

to multiple aspects of health sciences, for example repeated measures, measurement error and multilevel modeling (Skrondal and Rabe-Hesketh, 2003; Huber et al., 2004; Stefanski, 2000). A comprehensive survey can be found in Skrondal and Rabe-Hesketh (2004).

Depending on the context, latent variables can be defined in different ways. In general, they are random variables whose realizations are hidden from us. As Skrondal and Rabe-Hesketh (2004) presented, latent variables can be used to describe multiple phenomena, such as 'true' variables measured with error, hypothetical constructs, unobserved heterogeneity, missing data, counterfactuals or potential outcomes, and latent responses underlying categorical variables. In particular, measurement error models combined with regression models can be used to avoid diluted regression effects when a covariate has been measured with error. It is well-known that the naive approach to estimate the slope in a simple linear regression when the predictor is measured with error produces biased estimates Stefanski (2000).

Another class of widely used models is mixed effects models or multilevel regression models. Multilevel data arise when units are nested in clusters, for example siblings in the same family. Repeated measurements taken on the same subject can be viewed as clustered data too. The units belonging to the same cluster share the same cluster-specific influences, but these influences cannot all be modeled as covariates in that we often have limited knowledge regarding relevant covariates and our data set may furthermore lack information on these covariates. As a result there is cluster-level unobserved heterogeneity leading to correlation between responses for units in the same cluster, after conditioning on covariates. Unobserved heterogeneity is modeled by including random effects in a multilevel regression model.

In addition to linear regression models, where the outcome variable is assumed to be continuous and often Gaussian, generalized latent variable models have been employed to cope with non-Gaussian variables, including dichotomous, grouped, censored, ordinal, unordered polytomous or nominal, pairwise comparisons, rankings or permutations, counts, and durations or survival responses. For example, generalized linear mixed models combines

mixed effects models with generalized linear models to incorporate non-Gaussian responses:

$$E(Y|\nu) = \mu,$$
$$g(\mu) = \nu = X'\beta + \sum_{l=2}^{L} z_l'\zeta_l,$$

where $g(\cdot)$ is a link function, β is the vector of fixed effects, and $z^{(l)'}$ is an M_l-dimensional vector of explanatory variables with random coefficients $\zeta^{(l)}$ at level l.

It is easy to see the similarity between the generalized linear mixed model to our proposed model (6.2.1), since they are both derived from the GLM framework and have random variables in the linear predictor part. However, in our case the latent variable is an argument of a *random* function, while in the latent variable literature, the relation between the response and the latent variable is characterized by a *fixed* functional form. This complicates the theoretical and numerical studies of the contaminated sparse functional GLM. For example, we may not be able to directly use the Newton-Raphson type of optimization scheme to obtain the maximum likelihood, since we do not know the functional form of the predictor trajectories.

6.3 Asymptotics

Ideally we would want to derive the asymptotic distribution of the MCEM estimators. However, it is not straightforward to deal with such an approximate MLE in a multilevel model. In fact, Hall et al. (2011) obtained, for the first time, the precise asymptotic distribution of Gaussian variational approximation estimators for a simple Poisson mixed model. Therefore, in this section we consider the asymptotics of the exact MLE, given in (6.2.5), which the MCEM algorithm is supposed to converge to.

For the sake of simplicity, we will fix α and β, and treat m and $\mathbb{M} = Pm$ as functions of just θ. It can be shown that $\hat{\alpha}_n$ and $\hat{\beta}_n$ converge at $\sqrt{n}$-rate. To derive the asymptotics of $\hat{\theta}_n$, we assume that the contamination density satisfies the following conditions

(C1) $p_W(\cdot)$ is smooth, unimodal and symmetric at 0.

(C2) $\sup_{x\in(0,1)} \left|\frac{d}{dx} p_W(x)\right| < \infty$, and $\inf_{x\in(0,1)} |p_W(x)| > 0$.

An example that satisfies the above conditions is $N(\theta, \sigma^2)$ truncated to [0,1]. The following theorem gives the large-sample distribution of $\hat{\theta}_n$. Here θ_0 denotes the true value of θ.

Theorem 6.3.1. *If (6.2.1) and (6.2.2) hold, $0 < \theta_0 < 1$, $\beta \neq 0$, and conditions **(C1)** and **(C2)** are valid, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, I_{\theta_0}^{-1}).$$

Here $I_{\theta_0} = P_{\theta_0} \dot{m}_{\theta_0} \dot{m}'_{\theta_0}$ is the Fisher information and $\dot{m}_{\theta_0}$ is the score function.

Remark 6.3.2. *In the proof we will see that, it only requires conditions **(C1)** and **(C2)** to establish the consistency and the convergence rate of the MLE. Only the limiting distribution involves the density of W. This indicates that using a Gaussian contamination as the working model will still give a consistent estimator that converges at rate $n^{1/2}$.*

The convergence rate of the MLE does not depend on X and is slower than the MLE of θ in the sparse functional GLM. An intuitive explanation would be that the contamination "smooths out" the local irregularities of the predictor process and makes the estimation problem a regular maximum likelihood estimation problem.

6.4 Numerical procedure

In this section, we devise a numerical algorithm to obtain the MLE given in (6.2.5). From the previous section we know that the choice of $p_W(\cdot)$ does not affect the consistency and the convergence rate of the MLE. Therefore, we will assume that $W \sim N(0, \sigma_c^2)$ with know variance σ_c^2 for computational purposes. This implies that $\tau \sim N(\theta, \sigma_c^2)$ truncated to $[0, 1]$ and

$$p_\theta(\tau) = \frac{\frac{1}{\sigma_c}\phi(\frac{\tau-\theta}{\sigma_c})}{\Phi(\frac{1-\theta}{\sigma_c}) - \Phi(\frac{-\theta}{\sigma_c})}, \tag{6.4.1}$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ being the density and cumulative distribution function of standard normal, respectively.

Since model (6.2.1) contains unobservable random variables τ, we need to maximize the marginal likelihood that involves integration over the τ_i's,

$$L(\eta|\mathbf{Y}, \mathbf{X}) = \iint \prod_{i=1}^{n} p_{\alpha,\beta}(Y_i|X_i, \tau_i) p_\theta(\tau_i)\, \mathrm{d}\tau_1 \ldots \mathrm{d}\tau_n. \tag{6.4.2}$$

Direct maximization of (6.4.2) in close form seems prohibitive. A possible method is stochastic optimization. Specifically, we replace the marginal likelihood (6.4.2) with

$$\begin{aligned} L(\alpha,\beta,\theta|Y,X) &= \iint \prod_{i=1}^{n} p(Y_i|X_i,\tau_i)p(\tau_i|X_i)\,\mathrm{d}\tau_1\ldots\mathrm{d}\tau_n \\ &\approx 1/M\sum_{j=1}^{M}\prod_{i=1}^{n} p(Y_i|X_i,\tau_i^{(j)}), \end{aligned} \tag{6.4.3}$$

where $\tau_i^{(j)}$ are generated from $p_\theta(\cdot)$. However, the derivatives of this function with respect to η might still be difficult to calculate and the gradient-descent type of optimization method might be unstable. We might also maximize over θ by profiling out (α,β) but this method is computationally intensive.

A common technique to bypass the integrals and fit the hierarchical model is the Expectation–Maximization algorithm. It has been implemented in various studies involving latent variables and missing data (Dempster et al., 1977; Sammel et al., 1997). The algorithm consists of two steps: the expectation step (E-step), where the conditional expectation of the full log-likelihood given the observed data and the current estimate is calculated, and the maximization step (M-step), where the conditional expectation is optimized and the maximizer is used as the updated estimate. The conditional expectation of the full log-likelihood given (X_i,Y_i) and the current estimate $\eta^{(t)}$ is given by

$$\begin{aligned} Q(\eta|\eta^{(t)}) &\equiv \mathrm{E}[\log L(\eta)|\mathbf{Y},\mathbf{X},\eta^{(t)}] \\ &= \mathrm{E}\left[\sum_{i=1}^{n}[\log p(Y_i|X_i,\tau_i)+\log p(\tau_i|X_i)]\middle|\mathbf{Y},\mathbf{X},\eta^{(t)}\right] \\ &= \mathrm{E}\left[\sum_{i=1}^{n}[Y_i\lambda_i-\psi(\lambda_i)+\log p_\theta(\tau_i)]\middle|\mathbf{Y},\mathbf{X},\eta^{(t)}\right] \\ &= \sum_{i=1}^{n}\{Y_i[\alpha+\beta\mathrm{E}[X_i(\tau_i)|X_i,Y_i,\eta^{(t)}]]-\mathrm{E}[\psi(\alpha+\beta X_i(\tau_i))|X_i,Y_i,\eta^{(t)}]\} \\ &\quad + \mathrm{E}\sum_{i=1}^{n}[\log(p_\theta(\tau_i))|X_i,Y_i,\eta^{(t)}] \end{aligned} \tag{6.4.4}$$

Since we do not have the explicit functional form of the realized trajectories of X, the computation of the above expectations is analytically intractable. McCulloch (1997) proposed a Monte Carlo EM algorithm for estimating the parameters in generalized linear

mixed models, where the conditional expectation was replaced with a Monte Carlo average. This method was proposed to overcome the nonlinearity of ψ, but it can also be employed in our case. Specifically, Monte Carlo samples of the missing data are generated from the conditional distribution of τ_i given (X_i, Y_i), and the conditional expectation of the log-likelihood $Q(\eta|\eta^{(t)})$ is approximated by

$$\begin{aligned}\tilde{Q}_M(\eta|\eta^{(t)}) &= \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{M}\{Y_i[\alpha+\beta X_i(\tau_i^{(j)})]-\psi(\alpha+\beta X_i(\tau_i^{(j)}))\} \\ &\quad + \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{M}\{\log[p_\theta(\tau_i^{(j)})]\}. \end{aligned} \tag{6.4.5}$$

By Bayes theorem, the conditional density of τ can be written as

$$p(\tau \in \mathrm{d}t|Y=y, X, \eta^{(t)}) \propto p(Y=y|\tau \in \mathrm{d}t, X, \eta^{(t)})p(\tau \in \mathrm{d}t|X, \eta^{(t)}), \tag{6.4.6}$$

where $\eta^{(t)}$ is the parameters estimated from the tth EM iteration. The two conditional densities on the right side of (6.4.6) can be obtained from assumptions (6.2.1) and (6.2.2). This enables us to construct a Metropolis-Hastings (M-H) algorithm and generate M samples $\tau_i^{(1)}, \ldots, \tau_i^{(M)}$ given the ith observation (X_i, Y_i) and the current estimate $\eta^{(t)}$.

Note that we can maximize the first term on the right side of (6.4.5) over (α, β), and the second term over θ separately. The first term is up to a multiplicative constant the log-likelihood function as if we observed data $(X_i^{(j)}, Y_i^{(j)})$, where $X_i^{(j)} = X_i(\tau_i^{(j)})$ and $Y_i^{(j)} = Y_i$, and $(X_i^{(j)}, Y_i^{(j)})$ were related via the working model: $p(Y|X) = \exp((\alpha+\beta X)Y - \psi(\alpha+\beta X))$. Then we can use any standard software for fitting logistic models. The second term can be maximized using a Newton-Raphson type of algorithm. And the maximizer of $\tilde{Q}_M(\eta|\eta^{(t)})$ is the updated estimate $\eta^{(t+1)} = (\alpha^{(t+1)}, \beta^{(t+1)}, \theta^{(t+1)})$.

In summary, the proposed MCEM estimating procedure consists of the following steps.

1. Select initial estimates $\eta^{(0)} = (\alpha^{(0)}, \beta^{(0)}, \theta^{(0)})$.

2. Generate Monte-Carlo samples $\tau_i^{(1)}, \ldots, \tau_i^{(M)}$ from the conditional distribution (6.4.6) for each $i = 1, \ldots, n$ using Metropolis–Hastings algorithm.

3. E-step: calculate the approximate conditional distribution (6.4.5).

4. M-step: maximize (6.4.5) over η and take the maximizer as the updated estimate of η. Repeat steps 2 to 4 until convergence.

6.4.1 Practical issues in implementation

In this procedure we use the Gaussian distribution $N(0, \sigma_c^2)$ as the working model for the contamination W. In fact we can use any distribution that has a smooth density with mode zero. We will show in Section 6.3 that the choice is irrelevant to the consistency and the convergence rate of the proposed estimator.

It is well-known that the MCEM estimator is not deterministic due to the Monte-Carlo errors, and the Monte-Carlo sample size should be automatically increased after iterations in the MCEM algorithm, otherwise the updated estimates do not converge but fluctuate around the truth. Booth and Hobert (1999) constructed a sandwich variance estimate for the maximizer at each approximate E-step. Various methods have been proposed to reduce the burden of increasing Monte-Carlo sample sizes. See for example Levine and Casella (2001), Caffo et al. (2005) and Zipunnikov and Booth (2006). In our simulation studies presented in Chapter 7, we find the fluctuation tolerable without increasing the Monte Carlo sample size. So we skip this procedure to save computation time.

6.5 Proofs

Proof to Theorem 6.3.1: We will use the strategy based on M-estimation theory (see Chapter 3.2 in van der Vaart and Wellner (1996)). We will first establish the identifiability of model (6.2.1) and the consistency of $\hat{\theta}_n$, and then derive its convergence rate and limiting distribution.

Consistency. From (6.2.4) and conditions **(C1)** and **(C2)**, we can show that

1. $p_\theta(\cdot)$ is a smooth, unimodal (with mode θ) density supported by $[0, 1]$.
2. $\sup_{\theta,\tau\in(0,1)} \left|\frac{\partial}{\partial\theta} p_\theta(\tau)\right| \doteq U < \infty$, and $\inf_{\theta,\tau\in(0,1)} |p_\theta(\tau)| \doteq L > 0$.

First we prove that $\mathbb{M}(\theta)$ has a unique maximum at θ_0. Since

$$\frac{\partial p_\theta}{\partial \theta}|_{\theta_0} = 0,$$

and $|\frac{\partial}{\partial\theta}p_\theta|$ is bounded, it follows from Dominated Convergence Theorem that

$$\begin{aligned}\mathbb{M}'(\theta)|_{\theta_0} &= P\frac{\partial m_\theta}{\partial\theta}|_{\theta_0} \\ &= P\frac{\int_0^1 p(Y|X,\tau)\frac{\partial}{\partial\theta}p_\theta(\tau)|_{\theta_0}\mathrm{d}\tau}{\int_0^1 p(Y|X,\tau)p_\theta(\tau)|_{\theta_0}\mathrm{d}\tau} = 0.\end{aligned}$$

Similarly, since

$$\frac{\partial^2 p_\theta}{\partial\theta^2} < 0,$$

for all $\theta \in (0,1)$ such that $\theta \neq \theta_0$,

$$\begin{aligned}\mathbb{M}''(\theta) &= P\frac{\partial^2 m_\theta}{\partial\theta^2} \\ &= P\frac{\int_0^1 p(Y|X,\tau)\frac{\partial^2}{\partial\theta^2}p_\theta(\tau)d\tau\int_0^1 p(Y|X,\tau)p_\theta(\tau)\mathrm{d}\tau - [\int_0^1 p(Y|X,\tau)\frac{\partial}{\partial\theta}p_\theta(\tau)d\tau]^2}{[\int_0^1 p(Y|X,\tau)p_\theta(\tau)\mathrm{d}\tau]^2} < 0.\end{aligned}$$

Therefore θ_0 is the unique maximizer of $\mathbb{M}(\theta)$ and model (6.2.1) is identifiable. It also follows that θ_0 is well-separated in the sense that

$$\mathbb{M}(\theta_0) > \sup_{\theta\notin G}\mathbb{M}(\theta),$$

for every open set G that contains θ_0.

By condition **(C2)**,

$$\dot{m}(X,Y) \equiv \sup_{\theta\in(0,1)}\left|\frac{\int_0^1 p(Y|X,\tau)\frac{\partial}{\partial\theta}p_\theta(\tau)\mathrm{d}\tau}{\int_0^1 p(Y|X,\tau)p_\theta(\tau)\mathrm{d}\tau}\right| < U/L < \infty. \tag{6.5.1}$$

Therefore m_θ is Lipschitz in (0,1):

$$|m_{\theta_1}(X,Y) - m_{\theta_2}(X,Y)| \leq \dot{m}(X,Y)|\theta_1 - \theta_2|, \tag{6.5.2}$$

with $P|\dot{m}| < \infty$. Hence $\mathcal{F} \equiv \{m_\theta(X,Y) : \theta \in (0,1)\}$ is P-Glivenko-Cantelli, according to Example 19.7 in van der Vaart (1998). So we have $\mathbb{M}_n = \mathbb{P}_n m_\theta \longrightarrow \mathbb{M}$ uniformly on $(0,1)$ a.s. and by Theorem 3.2.3 of van der Vaart and Wellner (1996), $\hat{\theta}_n \longrightarrow \theta_0$ in probability.

Rate of convergence and limiting distribution. Define $d(\cdot,\cdot)$ as the Euclidean distance. Note that $\mathbb{M}'(\theta_0) = 0$ and $\mathbb{M}''(\theta_0) < 0$, so

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim d^2(\theta,\theta_0)$$

for all θ in (0,1). The class of functions $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}$ is Lipschitz based on (6.5.2), and $\dot{m}(X, Y)$ has a finite second moment based on (6.5.1). From Example 3.2.22 in van der Vaart and Wellner (1996),

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0} + o_p(1). \tag{6.5.3}$$

Here $-V = -\mathbb{M}''(\theta_0) = I_{\theta_0}$ is the Fisher information, $\dot{m}_{\theta_0}$ is the score function, and $\mathbb{G}_n\dot{m}_{\theta_0}$ is asymptotically zero-mean normal with variance $I_{\theta_0}^{-1}$. It follows that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, I_{\theta_0}^{-1}).$$

∎

Chapter 7

Simulation studies for the contaminated sparse functional GLM

In this chapter we present the results of four simulation studies. They assess the performance of the MCEM algorithm, the finite-sample behavior of the MLE for contaminated sparse functional GLM, and compare it to the sparse functional GLM estimator under various scenarios. We will restrict attention to contaminated sparse functional logistic regression. The first simulation evaluates the convergence of the MCEM algorithm. The second simulation illustrates the behavior of the estimators of α_0, β_0 and θ_0 in repeated application of the proposed method. The third simulation studies the relationship between the contaminated sparse functional GLM and the sparse functional GLM. The fourth one is designed to explore the impact of H on the estimates.

7.1 Simulation model description

This section describes the simulation model we use for the investigations of finite-sample performance of the proposed estimator. In application we frequently encounter random samples $(X_i, Y_i), i = 1, \ldots, n$, where Y_i are binary outcomes and X_i are functional predictors. Take the gene expression study and the fMRI study, described in Sections 1.1.1 and 1.1.2

respectively, for example. Therefore, we will use a contaminated sparse functional logistic model as illustration throughout the simulations, which is given by

$$\begin{aligned} &\text{logit}[P(Y=1|X,\tau)] = \alpha + \beta X(\tau), \\ &P(\tau \le t) = P(\theta + \ W \le t | \theta + \ W \in [0,1]), \text{ and} \\ &W \sim N(0, \sigma_c^2). \end{aligned} \tag{7.1.1}$$

Here θ is the sensitive point of main interest, α and β are scalar intercept and slope, respectively, and W is a Gaussian contamination. In our simulations, we directly generate τ from a truncated normal distribution with the corresponding mean, standard deviation and cut-off points.

7.2 Convergence of MCEM

The first simulation concerns the convergence of the MCEM algorithm. A sample of size $n = 40$ are generated from the contaminated sparse functional logistic model with $\alpha_0 = 1$, $\beta_0 = 3$, $\theta_0 = 0.7$ and $\sigma_c = 0.05$. The predictor processes X_i are generated as Brownian motions over a uniform grid of 201 using the `R` 2.11.1 function `fbmSim` in the library `fArma`. We restrict θ to this grid. The randome sensitive points τ_i are generated from a truncated normal distribution using the `R` 2.11.1 function `rtnorm` in the library `msm`. We assume $\sigma_c = 0.05$ is known.

E-step. The conditional expectation of the log-likelihood is approximated by an average over Monte-Carlo samples of the unobserved τ_i (6.4.4). To generated τ_i from its conditional distribution given (X_i, Y_i), we use a Metropolis-Hastings algorithm, which is a Markov Chain Monte Carlo (MCMC) sampling technique. It first draws a random sample from a proposal distribution, and then update the current Markov chain with the random sample by a probability based on a likelihood ratio. It is shown that the Markov chain's equilibrium distribution is the desired distribution we want to draw samples from (Hastings, 1970).

We choose $N(0, 0.1)$ as the proposal distribution in the M-H algorithm. The MCMC chain is run for 500 iterations and the first 200 samples are used as burn-in. Figure 7.2 is the graph of the Gelman-Rubin's R statistic based on five independent Markov chains. The blue

dotted line is the critical value for convergence. It suggests that the chains converge after the 200-step burn-in. Therefore, in the MCEM algorithm, we will run the M-H algorithm for 500 iterations and use the last 300 steps as the Monte Carlo sample.

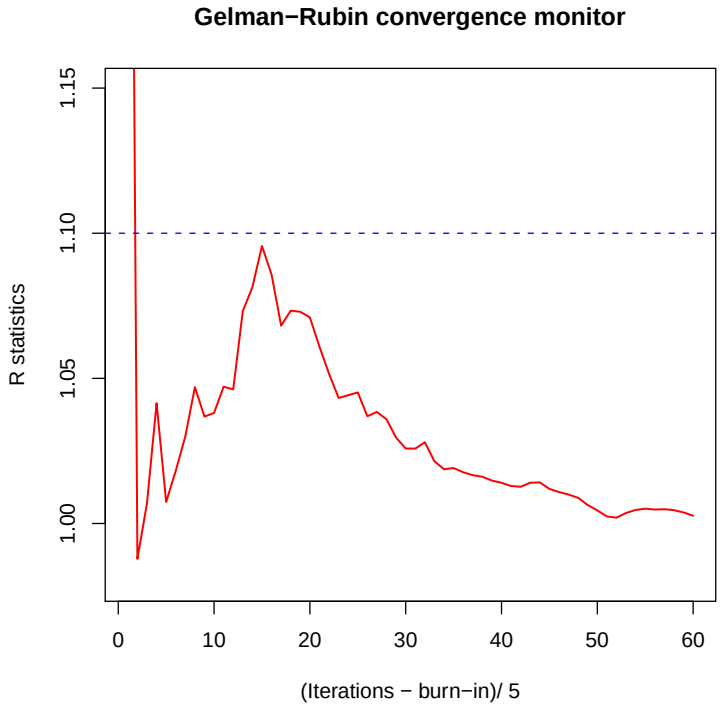


Figure 7.1: Convergence of the Metropolis-Hastings algorithm, evaluated by Gelman-Rubin's R statistic, which is s based a comparison of within-chain and between-chain variances, similar to a classical analysis of variance. The blue dotted line is the critical value.

M-step. To maximize the target function (6.4.5), we use the "L-BFGS-B" method in the `R` function `optim` to maximize the second term on the right side of (6.4.5) and update the estimate of θ, and use the `R` function `glm` to maximize the first term and update the estimates of α and β.

We set the initial values of the estimates as $\alpha_1 = 1.5$, $\beta_1 = 2$ and $\theta_1 = 0.5$ and run the MCEM algorithm for 1000 steps. Figure 7.2 shows the updated $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\theta}$ in each step. As we previously mentioned, the MCEM estimates do not converge to a deterministic

limit, but rather fluctuate randomly about some stationary point with Monte Carlo noise. It can be seen that the algorithm reaches an equilibrium after 200 steps. In the following we will always run 500 steps, which takes about 220 seconds, and take the average of the last 100 steps as the final estimation. Table 7.2 lists the means and standard deviations of the last 100 updates for different values of α, β and θ. We find the randomness of the MCEM estimates under tolerance.

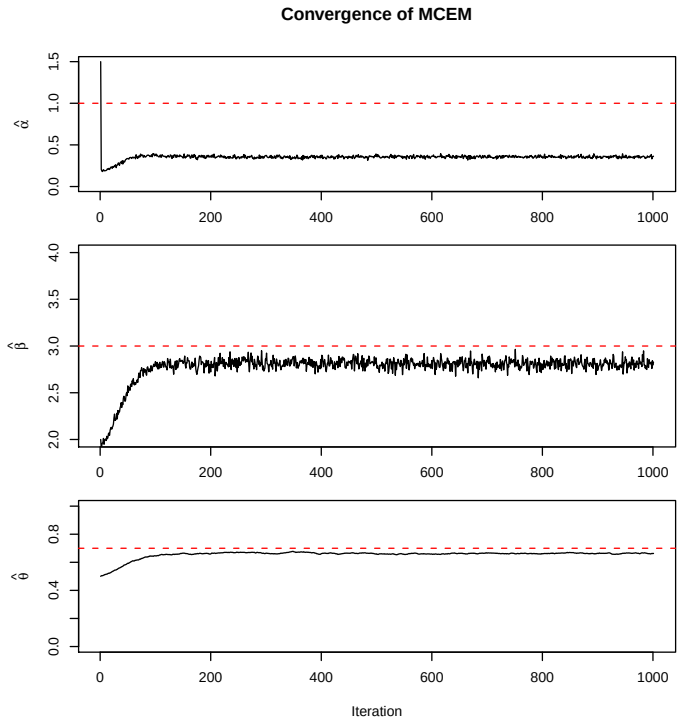


Figure 7.2: Convergence of the MCEM algorithm. The MCEM estimate updates (black solid lines) fluctuates about the MLEs, which should be close to the true parameters (red dotted lines).

7.3 Distribution of the MLE

In this simulation, we investigates the finite-sample distributions of the MLE. We generate 1000 samples from the contaminated sparse functional logistic model, with $n = 40$, $\alpha_0 = 1$,

Table 7.1: Means and standard deviations of the last 100 updates in a 500-iteration MCEM algorithm under different true parameters.

True parameters†			Mean (Std)		
α_0	β_0	θ_0	$\hat{\alpha}_n$	$\hat{\beta}_n$	$\hat{\theta}_n$
0	1.5	0.5	0.609 (0.016)	3.226 (0.081)	0.427 (0.002)
		0.7	-0.282 (0.018)	2.023 (0.025)	0.713 (0.027)
0	3	0.5	0.492 (0.020)	3.769 (0.075)	0.477 (0.002)
		0.7	0.443 (0.024)	4.625 (0.136)	0.533 (0.004)
1	1.5	0.5	0.501 (0.003)	0.853 (0.011)	0.523 (0.003)
		0.7	0.721 (0.002)	0.467 (0.018)	0.546 (0.004)
1	3	0.5	0.600 (0.012)	2.892 (0.038)	0.529 (0.006)
		0.7	0.356 (0.013)	2.823 (0.045)	0.664 (0.003)

†: We assume $\sigma_c = 0.05$ is known.

$\beta_0 = 3$, $\theta_0 = 0.7$ and $\sigma_c = 0.05$. The MCEM procedure is applied to these simulated data sets and maximum likelihood estimates are generated. We removed 21 cases where $\hat{\beta}_n$ is greater than 15, considering them the results of non-convergence or local maximum.

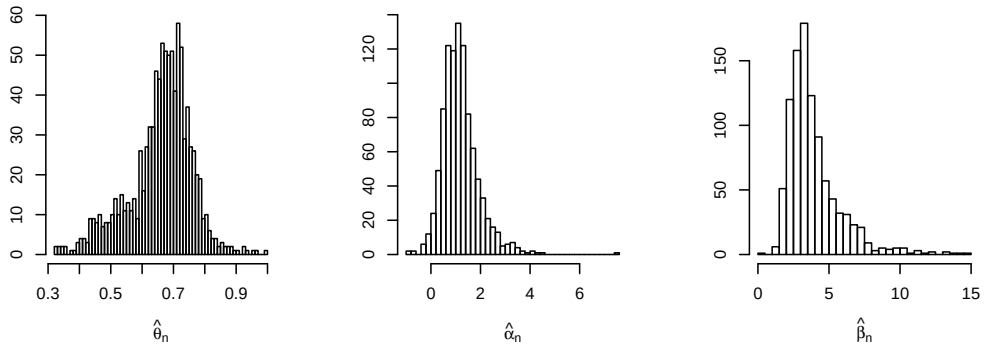


Figure 7.3: Histograms of estimates of $\hat{\theta}_n$ (left), $\hat{\alpha}_n$ (center), $\hat{\beta}_n$ (right) from 1000 replications, with truth $\theta_0 = 0.7$, $\alpha_0 = 1$, and $\beta_0 = 3$.

The histograms of $\hat{\alpha}_n$, $\hat{\beta}_n$ and $\hat{\theta}_n$ are displayed in Figure 7.3. Their means (standard deviations) are 1.188 (0.741), 3.922 (1.874) and 0.660 (0.100), respectively. Our simulation results indicate that the MLEs are consistent and efficient. Note that the convergence rate of $\hat{\theta}_n$ is comparable to those of $\hat{\alpha}_n$ and $\hat{\beta}_n$, which is in contrast to the faster convergence orate of $\hat{\theta}_n$, as described in Lindquist and McKeague (2009). This is not surprising, as we have shown in Chapter 6 that the maximum likelihood estimator of θ converges at the parametric rate $n^{1/2}$, the same rate as those of $\hat{\alpha}_n$ and $\hat{\beta}_n$, as a result of the contamination's smoothing effect.

7.4 Compare MCEM to the sparse functional GLM

In this simulation, we explore the relationship between the sparse functional GLM and the contaminated sparse functional GLM. The data are generated from the contaminated sparse logistic model with $\alpha_0 = 1$, $\beta_0 = 3$, $\theta_0 = 0.7$ and $\sigma_c = 0.01$. Different sample sizes are considered. We apply both the MCEM method and the sparse functional GLM as working models to 1000 replications. Again we remove those cases where $\hat{\beta}_n > 15$. The results are summarized in Table 7.2.

The first observation is that, as the sample size increases, the means of both estimators approach the truth, suggesting that both estimators are consistent. Another observation is that, the estimates of θ_0 have smaller standard deviation in the contaminated sparse functional GLM than the sparse functional GLM, suggesting the former could be more efficient. Finally, although the data are generated with contamination, the sparse functional GLM still has decent performance, indicating that it is robust against contamination.

7.5 Dependence on the Hurst exponent

In the previous simulations, we have assumed the predictor X to be a Brownian motion. The contaminated sparse functional GLM, however, did not make this restriction. In fact, as shown in Section 6.3, the consistency and rate of convergence of the MLEs do not depend on X. In this simulation, we want to test the validity of this conclusion. Therefore, we generate fractional Brownian motions with three different Hurst exponent values, $H =$

Table 7.2: Means and standard deviations of the MCEM estimators and the sparse functional GLM estimators, based on 1000 replications.

Sample size	$\hat{\alpha}_{n,CPI}$	$\hat{\beta}_{n,CPI}$	$\hat{\theta}_{n,CPI}$	$\hat{\alpha}_{n,PI}$	$\hat{\beta}_{n,PI}$	$\hat{\theta}_{n,PI}$
$n = 30$	1.245 (0.973)	4.208 (2.203)	0.639 (0.118)	1.387 (1.120)	4.704 (2.470)	0.686 (0.153)
$n = 40$	1.188 (0.741)	3.922 (1.874)	0.660 (0.100)	1.247 (0.816)	4.116 (2.041)	0.700 (0.133)
$n = 50$	1.145 (0.673)	3.693 (1.550)	0.655 (0.096)	1.184 (0.712)	3.756 (1.528)	0.705 (0.108)
$n = 80$	1.088 (0.452)	3.460 (1.126)	0.675 (0.071)	1.076 (0.412)	3.352 (0.990)	0.709 (0.079)
$n = 100$	1.073 (0.381)	3.314 (0.880)	0.684 (0.061)	1.058 (0.365)	3.196 (0.776)	0.705 (0.072)
$n = 120$	1.063 (0.335)	3.271 (0.761)	0.685 (0.054)	1.041 (0.319)	3.136 (0.654)	0.705 (0.058)

CPI: MCEM estimators using the contaminated sparse GLM.

PI: maximum likelihood estimators using the sparse GLM.

$0.3, 0.5$ (corresponding to Brownian motions) and 0.7, as the predictor process, and evaluate the performance of the propose maximum likelihood estimators. The true parameters are again taken as $\alpha_0 = 1$, $\beta_0 = 3$, $\theta_0 = 0.7$ and $\sigma_c = 0.05$. The histograms of $\hat{\theta}_n$ and $\hat{\beta}_n$ based on 500 samples are given in Figure 7.4.

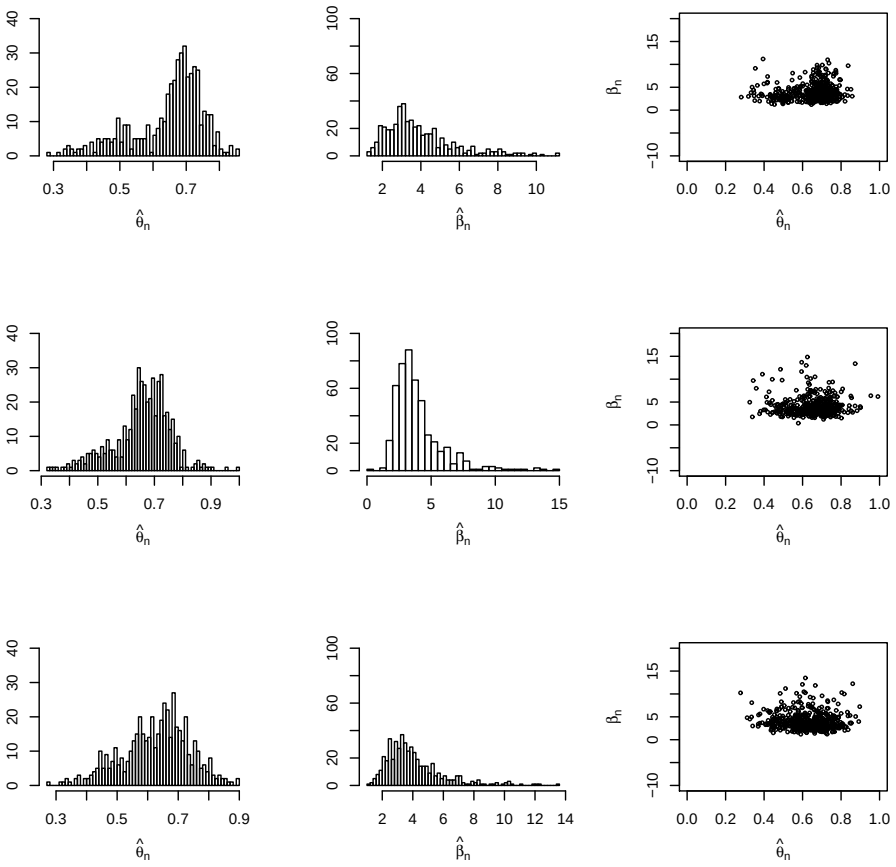


Figure 7.4: Histograms and scatter plots of $\hat{\theta}_n$ and $\hat{\beta}_n$ for H = 0.3 (top row), H = 0.5 (middle row), and H = 0.7 (bottom row), based on 500 samples of size n = 40. The estimation accuracy does not depend on the Hurst exponent H.

The histograms show that, the distribution of $\hat{\theta}_n$ and $\hat{\beta}_n$ are independent of the Hurst exponent of X. We again see that the convergence rate of $\hat{\theta}_n$ is comparable to that of $\hat{\beta}_n$, supporting the theory that they both converge at $\sqrt{n}$-rate. Also presented in Figure 7.4 are

scatter plots of $\hat{\beta}_n$ against $\hat{\theta}_n$, which show that $\hat{\theta}_n$ and $\hat{\beta}_n$ are asymptotically independent.

Chapter 8

Application to the fMRI data

In this chapter, we will apply the contaminated sparse functional GLM, proposed in Chapter 6, to the fMRI data introduced in Section 1.1.2 to illustrate the proposed MCEM method. We will not apply it to the gene expression data since it may be ungrounded to assume random positions of the genes that are mostly related to the outcome. There are scientific reasons, however, to believe that the onset time point of brain activity in fMRI studies can be random and prone to subject-specific errors. Below we will first describe the data set, and then provide the model information and the estimation procedure employed in the analysis of this data set. Finally, we present the interpretation and discussion of the results.

8.1 Description of the fMRI data

In fMRI studies, estimation of the precise timing of the underlying psychological activity is critical for many data analyses (Lindquist et al., 2007; Robinson et al., 2010). In Robinson et al. (2010), particularly, a multi-subject change point estimation procedure was proposed to allow for random onset times of psychological activities. However, this method cannot incorporate additional information contained in the observed response variables that are associated with the functional predictors via the onset time points. The contaminated sparse functional GLM proposed in this article overcomes this problem.

Here we will consider the data set described in Lindquist et al. (2007). In this study, 25 participants were scanned with BOLD fMRI at 3 T (GE, Milwaukee, WI), of whom 13

were classified as resilient and 12 were classified as non-resilient according to a written test. Each of them performed a 7-minute anxiety-provoking speech preparation task. The design was an off-on-off design, with the anxiety-provoking period occurring between lower-anxiety resting periods. During the task, 215 fMRI images were acquired. For calibration purposes, we further edit the data set by taking off the individual mean over the first resting period from the entire time course, as it is only the relative change in signal that is important. We re-scale the time period to [0,1], in which the 215 observation time points are equally spaced.

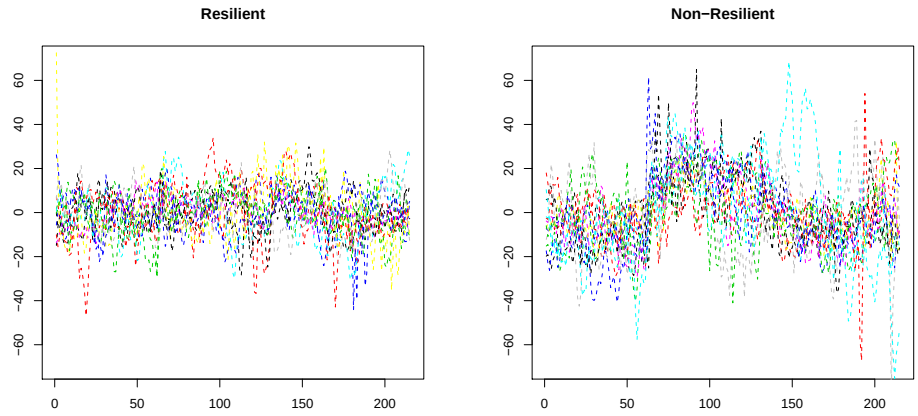


Figure 8.1: The fMRI signal over the ventromedial prefrontal cortex in reaction to an anxiety-provoking taske for resilient (left) and non-resilient (right) subjects.

Our main interest is to estimate the onset time of brain activity at which the intensity of the fMRI signal best differentiates the resilient and the nonresilient participants. As pointed out by D'Esposito et al. (2003), there is empirical evidence that the BOLD fMRI signals might be altered in normal ageing and disease. Therefore, the onset time of brain activity indicated by the BOLD signal in this fMRI study could be susceptible to subject-specific effects.

8.2 Model specification and parameter estimation

We apply the contaminated sparse logistic model (7.1.1) to the data from the fMRI study for anxiety-levels. Here Y represents the anxiety level with $Y = 1$ standing for resilient and $Y = 0$ for nonresilient. X represents the fMRI signal time course from the ventromedial prefrontal cortex, a region known to be related to anxiety. The sensitive point θ represents the mean time point at which the human brain respond to stimulus. The contamination error W represents the advance and delay of such time point due to personal reasons.

We will estimate the onset time as well as the regression coefficients, under various assumptions about the contamination. Specifically, when there is no contamination assumed, we will use the sparse functional GLM for estimation. And when we assume the existence of contamination, we use the contaminated sparse functional GLM and specify different values of the standard deviation of the contamination, since in the MCEM algorithm it is assumed known. The starting values of the MCEM procedures are taken as the estimates from the sparse functional GLM approach, which assumes no contamination. The MCEM algorithm is run for 100 steps with visual checking of convergence. The average over the last 10 steps is taken to be the final estimate. We used three different values of σ_c: 0.02, 0.05 and 0.1.

8.3 Results

Table 8.3 summarizes the results. The sensitive time point estimated when $\sigma_c = 0.02$ is the 85th time point, which is 30 seconds into the anxiety-provoking period of the task. This is slightly different from the result when assuming no contamination and using the sparse functional logistic model. Also, the sparse functional logistic model is fairly robust to contamination, with a 5% noise level leading to a 0.5% shift in the estimate of θ_0.

We can see from the table that the absolute values of $\hat{\alpha}_n$ and $\hat{\beta}_n$ decrease as σ_c increases. In other words, the estimated effect is attenuated when higher contamination levels are assumed. An analogous situation in the measurement error literature is the attenuation

Table 8.1: Application to fMRI data: the maximum likelihood estimates of (α, β, θ) in the contaminated sparse functional logistic model using the MCEM procedure.

	MCEM estimates		
Err. σ_c †	$\hat{\alpha}_n$	$\hat{\beta}_n$	$\hat{\theta}_n$
0	4.335	-0.425	0.391
0.02	3.983	-0.378	0.392
0.05	3.448	-0.320	0.396
0.1	2.794	-0.271	0.422

†: The error standard deviation σ_c controls the magnitude of the contamination. Since it is unknown in reality, we repeat the estimate procedure for different σ_c's.

bias in the simple linear regression:

$$Y = \alpha + \beta X^* + \varepsilon,$$
$$X = X^* + \eta,$$

where X is the observed scalar predictor and η is the measurement error. When the measurement error η is ignored and the regression coefficients are estimated using the original approach, the absolute value of the estimated β tends to be smaller than the truth, i.e. there is an "attenuation bias". This is in contrast to the bias we have seen in the above example. An intuitive explanation to this phenomenon is that, when no contamination is assumed, the sparse functional logistic regression essentially selects the θ on which X is mostly associated with Y, reflected by a larger $\hat{\beta}_n$, and when contamination is assumed, the estimate $\hat{\beta}_n$ represents the association between Y and a mixture of the values of X on a neighborhood of θ_0, which is very likely to be weaker than the former.

Chapter 9

Conclusions

In this chapter, we make conclusions and discuss the advantages of the proposed method as well as some of its limitations. In the first section, we summarize the results obtained in the previous chapters, comment on the derived minimax bounds for sparse functional regression models, and discuss the theoretical and computational features of the proposed MCEM estimator for contaminated sparse functional GLM. In the second section, we point out a few possible directions for future work on these topics.

9.1 Key findings

This dissertation addresses two questions. One is the optimal rates for estimating the parameters in the sparse functional linear regression model and the sparse functional generalized linear regression model, proposed in McKeague and Sen (2010) and Lindquist and McKeague (2009), respectively. We have established a minimax lower bound for the sparse functional linear model, by applying a variation of Le Cam's method and bounding the total variation affinity between two simple hypotheses. We also derived a minimax upper bound for the sparse functional linear model, exploiting a result from Nishiyama (2010) on the moment convergence of M-estimators and establishing the second moment convergence rate of the least squares estimator. It was shown that the minimax upper bound is of the same asymptotic order as that of the minimax lower bound, which implies that the least squares estimator attains the optimal rate for estimating the parameters in the sparse functional

linear regression model. It was also shown that the estimators for the regression coefficients converge at the parametric rate, while the estimator for the sensitive point converges at a possibly faster rate that depends on the roughness of the predictor process, which is quantified by a "generalized Hurst exponent" that we proposed.

In a similar way, we obtained a minimax lower bound for the sparse functional GLM, bounding the affinity between two simple hypotheses by their Hellinger distance, and then using an inequality that linearizes this distance. It was seen that this lower bound has the same asymptotic rate as the minimax rate for the sparse functional linear regression model. It can also be seen that this rate is the same as the weak convergence rate of the maximum likelihood estimator for the sparse functional GLM derived in Lindquist and McKeague (2009). One limitation to our study is that it is not straightforward to replicate the argument in the sparse functional linear regression case and derive a second moment convergence rate of the MLE that hold uniformly over the parameter space. Thus the minimax upper bound for the sparse functional GLM remains to be established.

Another problem we addressed is to extend the sparse functional GLM to the contaminated sensitive point settings. This extension is motivated by a complication in fMRI studies, where the BOLD signals tend to be affected by personal aging, disorders and pathology (D'Esposito et al., 2003). In this scenario, the sensitive point is likely to be contaminated by random errors and the sparse functional GLM fails to formulate such situation by assuming a universal θ. We proposed a contaminated sparse functional GLM to allow for random sensitive points, and constructed a numerical estimating procedure for the parameters in the model. The procedure is based on a Monte Carlo EM algorithm, which calculates the conditional expectation in the E-step by Monte Carlo approximation. It was shown that the proposed estimator is consistent and converges to the truth at rate $n^{1/2}$. The MCEM was tested in several simulation studies and a real fMRI study. The results show that the estimation procedure has reasonable performance for practical use.

A drawback of the contaminated sparse functional GLM is its intensive computation. While directly maximizing the likelihood function is time consuming, the MCEM algorithm is also costly because of its slow convergence. A possible solution might be to combine the two methods, i.e. to run the MCEM algorithm for a few iterations and use the results as

starting points of direct optimization. Another limitation of the estimating procedure is the specification of the contamination variance. Although we have shown that the maximum likelihood estimators are consistent asymptotically in spite of the magnitude of the contamination, in the simulations we can see that the regression coefficient estimators still decrease as the error variance increases. One possible solution is a data-driven procedure that estimates the variance in advance, and then plug it into the MCEM procedure. A good starting point is the multi-subject change point estimation technique described in Robinson et al. (2010). Another possibility is to obtain the variance of the contamination from other sources in scientific research, for example Handwerker et al. (2004) and Menz et al. (2006).

9.2 Topics for future research

It is of interest to derive a minimax upper bound for the sparse functional GLM. Although the maximum likelihood estimator has been shown to have a weak convergence rate of the same order as the lower bound, the upper bound on the convergence rate in the minimax sense has yet to be established. If we want to replicate the approach used in the spares functional linear regression case, one difficulty might be to establish the quadratic approximation (3.5.1), since if we Taylor expand the target function $\mathbb{M}(\eta)$ around η_0, the non-linearity of $\psi^{(2)}(\cdot)$ might complicate the situation and make it difficult to obtain a universal constant ϵ in (3.5.1).

An alternative approach to deriving the minimax rates might be to establish statistical equivalence between the sparse functional regression models and white noise models, in the sense that Le Cam's metric (Le Cam, 1986; Le Cam and Yang, 1990) for the distance between the two models converges to zero as n goes to infinity. It is implied from the asymptotic equivalence that any minimax procedure in one problem will automatically yield the corresponding procedure in the other with equal optimal rates. Such equivalence has been established for deriving the optimal rates for linear functional estimation, nonparametric regression and functional linear regression. See for example, Brown and Low (1996), Cai and Low (2004) and Meister (2011).

In particular, Meister (2011) showed that the functional linear regression model (2.1.3),

written as $Y = \langle X, \phi \rangle + \varepsilon$ where ϕ is the regression function and $\langle \cdot, \cdot \rangle$ denotes the $L_2([0,1])$-inner product, is equivalent to a white noise model with drift

$$dY(t) = [\Gamma^{1/2}\phi](t)dt + n^{-1/2}\sigma dW(t)$$

where $W(t)$ denotes a standard Wiener process defined on the interval [0,1] and $\Gamma^{1/2}$ is the unique positive definite symmetric square root of the covariance operator of X, defined by $\Gamma^{1/2}\Gamma^{1/2} = \Gamma$ and $\Gamma f = \int EX(\cdot)X(t)f(t)dt$ for any $f \in L_2[0,1]$. Such equivalence, combined with the results in Cavalier and Tsybakov (2002), gave sharp minimax constants in the FLR model. It would be of interest to investigate the equivalence between the sparse functional linear model and a white noise model.

We may also consider the equivalence between the sparse functional GLM and a white noise model. Grama and Nussbaum (1998) extended the work of Brown and Low (1996) and established the equivalence between nonparametric generalized models and white noise models. Specifically, suppose at points $t_i = i/n, i = 1, \ldots, n$, we observe independent r.v.'s Y_i, which follow a distribution from an exponential family Q_λ with parameters $\lambda_i = f(t_i) \in \Lambda$, where $f : [0,1] \to \Lambda$ is an unknown function belonging to a smoothness class Σ, then it was shown that this model is equivalent to a white noise model

$$dY_t^n = \Gamma(f(t)) + \frac{1}{\sqrt{n}}dW_t, \quad t \in [0,1],$$

where $\Gamma(\lambda) : \Lambda \to R$ is a function such that $\Gamma'(\lambda) = I(\lambda)^{1/2}$ and $I(\lambda)$ is the Fisher information in the local exponential family Q_λ. It would be of interest to extend this result to the sparse functional GLM case and establish an equivalence to a white noise model.

From the application example described in Chapter 8, we can see that the estimation of parameters in the sparse functional GLM is moderately sensitive to the choice of the contamination level σ_c. So it may be helpful to estimate σ_c in advance. One way is to estimate it from other sources of scientific research. In fact, estimation of hemodynamic response functions (HRF) is often an integral part of event-related fMRI analyses, and variations of HRFs across individuals and brain regions have been recently studied (Handwerker et al., 2004; Menz et al., 2006). From these results, we might be able to know the variance of the onset times *a priori*. Alternatively, we may also use multi-subject change point estimation

technique described in Robinson et al. (2010) to estimate the standard deviation of the change points in the fMRI time courses and use is as a substitute for σ_c.

Bibliography

Aguilera, A. M., F. A. Ocaña, and M. J. Valderrama (1999a). Forecasting time series by functional PCA. Discussion of several weighted approaches. *Comput. Statist. 14*(3), 443–467.

Aguilera, A. M., F. A. Ocaña, and M. J. Valderrama (1999b). Forecasting with unequally spaced data by a functional principal component approach. *Test 8*(1), 233–253.

Aston, J. A. D., F. E. Turkheimer, and M. Brett (2006). Hbm functional imaging analysis contest data analysis in wavelet space. *Human Brain Mapping 27*(5), 372–379.

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Berman, S. M. (1985). The maximum of a Gaussian process with nonconstant variance. *Ann. Inst. H. Poincaré Probab. Statist. 21*(4), 383–391.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields 71*(2), 271–291.

Booth, J. G. and J. P. Hobert (1999). Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlow EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodology) 61*(1), 265–285.

Brown, L. D. and M. G. Low (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist. 24*(6), 2384–2398.

Buness, A., R. Kuner, M. Ruschhaupt, A. Poustka, H. Sultmann, and A. Tresch (2007, September). Identification of aberrant chromosomal regions from gene expression microarray studies applied to human breast cancer. *Bioinformatics 23*(17), 2273–2280.

Caffo, B. S., W. Jank, and G. L. Jones (2005). Ascent-based Monte Carlo expectation-maximization. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*(2), 235–251.

Cai, T. T. and P. Hall (2006). Prediction in functional linear regression. *Ann. Statist. 34*(5), 2159–2179.

Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist. 38*(1), 100–145.

Cai, T. T. and M. G. Low (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist. 32*(2), 552–576.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist. 38*(4), 2118–2144.

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparametr. Statist. 12*(4), 503–538.

Cardot, H. and J. Johannes (2010). Thresholding projection estimators in functional linear models. *J. Multivariate Anal. 101*(2), 395–408.

Cardot, H. and P. Sarda (2006). Linear regression models for functional data. In *The art of semiparametrics*, Contrib. Statist., pp. 49–66. Physica-Verlag/Springer, Heidelberg.

Carlstein, E., H. Müller, and D. Siegmund (Eds.) (1994). *Change-point Problems*, Number 23 in IMS Monograph. Institute of Mathematical Statistics, Hayward, CA.

Cavalier, L. and A. Tsybakov (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields 123*(3), 323–354.

Chesher, A. (1991). The effect of measurement error. *Biometrika 78*(3), 451–462.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist. 37*(1), 35–72.

de Boor, C. (2001). *A practical guide to splines* (Revised ed.), Volume 27 of *Applied Mathematical Sciences*. New York: Springer-Verlag.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B 39*(1), 1–38. With discussion.

D'Esposito, M., L. Y. Deouell, and A. Gazzaley (2003, November). Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nature reviews. Neuroscience 4*(11), 863–872.

Devroye, L. (1987). *A course in density estimation*, Volume 14 of *Progress in Probability and Statistics*. Boston, MA: Birkhäuser Boston Inc.

Donoho, D. L. and R. C. Liu (1991). Geometrizing rates of convergence. II, III. *Ann. Statist. 19*(2), 633–667, 668–701.

Dou, W., D. Pollard, and H. Zhou (2010). Functional regression for general exponential families. Unpublished manuscript.

Dudoit, S. and M. J. van der Laan (2008). *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. New York: Springer.

Durrett, R. (1996). *Probability: theory and examples* (Second ed.). Belmont, CA: Duxbury Press.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, and et al. (2008). Genetics of gene expression and its effect on disease. *Nature 452*(7186), 423–428.

Escabias, M., A. M. Aguilera, and M. J. Valderrama (2005). Modeling environmental data by functional principal component logistic regression. *Environmetrics 16*(1), 95–107.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*, Volume 90 of *Statistics: Textbooks and Monographs*. New York: Marcel Dekker Inc.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*, Volume 66 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Feder, J. (1988). *Fractals*. Physics of Solids and Liquids. New York: Plenum Press. With a foreword by Benoit B. Mandelbrot.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010, 2). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Goldenshluger, A., A. Tsybakov, and A. Zeevi (2006). Optimal change-point estimation from indirect observations. *Ann. Statist. 34*(1), 350–372.

Grama, I. and M. Nussbaum (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields 111*(2), 167–214.

Gruvberger-Saal, S. K. E. A. (2004). Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles. *Molecular Cancer Therapeutics 3*(2), 161–168.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist. 35*(1), 70–91.

Hall, P. and M. Hosseini-Nasab (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol. 68*(1), 109–126.

Hall, P., T. Pham, M. P. Wand, and W. S. S. J. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *Ann. Statist. 39*(5), 2502–2532.

Hall, P. and D. M. Titterington (1992). Edge-preserving and peak-preserving smoothing. *Technometrics 34*(4), 429–440.

Han, T. S. and S. Verdú (1994). Generalizing the Fano inequality. *IEEE Trans. Inform. Theory 40*(4), 1247–1251.

Handwerker, D. A., J. M. Ollinger, and M. D'Esposito (2004). Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage 21*(4), 1639–1651.

Hastings, W. K. (1970). Monte Carlow sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Huber, P., E. Ronchetti, and M.-P. Victoria-Feser (2004). Estimation of generalized linear latent variable models. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*(4), 893–908.

James, G. M. (2002). Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64*(3), 411–432.

James, G. M., T. J. Hastie, and C. A. Sugar (2000). Principal component models for sparse functional data. *Biometrika 87*(3), 587–602.

James, G. M. and B. W. Silverman (2005). Functional adaptive model estimation. *J. Amer. Statist. Assoc. 100*(470), 565–576.

James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *Ann. Statist. 37*(5A), 2083–2108.

Kneip, A. and K. J. Utikal (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc. 96*(454), 519–542. With comments and a rejoinder by the authors.

Korostelev, A. and O. Korosteleva (2011). *Mathematical statistics*, Volume 119 of *Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society. Asymptotic minimax theory.

Korostelëv, A. P. (1987). Minimax estimation of a discontinuous signal. *Teor. Veroyatnost. i Primenen. 32*(4), 796–799.

Korostelëv, A. P. and A. B. Tsybakov (1993). *Minimax theory of image reconstruction*, Volume 82 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist. 1*, 38–53.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. New York: Springer-Verlag.

Le Cam, L. and G. L. Yang (1990). *Asymptotics in statistics: some basic concepts*. Springer Series in Statistics. Springer-Verlag.

Levine, R. A. and G. Casella (2001). Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Statist. 10*(3), 422–439.

Lieberman-Aiden, E., N. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. Lajoie, P. Sabo, M. Dorschner, R. Sandstrom, B. Bernstein, M. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. Mirny, E. Lander, and J. Dekker (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science 326*(5), 289–93.

Lieberman-Aiden, E., N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, and et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science 326*(5950), 289–293.

Lindquist, M., J. Meng Loh, L. Atlas, and T. Wager (2008). Modeling the hemodynamic response function in fmri: Efficiency, bias and mis-modeling. *Neuroimage*.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci. 23*(4), 439–464.

Lindquist, M. A. and I. W. McKeague (2009). Logistic regression with Brownian-like predictors. *J. Amer. Statist. Assoc. 104*(488), 1575–1585.

Lindquist, M. A., C. Waugh, and T. D. Wager (2007). Modeling state-related fmri activity using change-point theory. *NeuroImage 35*(3), 1125–1141.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc. 92*(437), 162–170.

McKeague, I. W. and B. Sen (2010). Fractals with point impact in functional linear regression. *Ann. Statist. 38*(4), 2559–2586.

Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist. 39*(3), 1471–1495.

Menz, M. M., J. Neumann, K. Müller, and S. Zysset (2006). Variability of the bold response over time: an examination of within-session differences. *NeuroImage 32*(3), 1185–1194.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist. 20*(2), 737–761.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *Ann. Statist. 33*(2), 774–805.

Nishiyama, Y. (2010). Moment convergence of M-estimators. *Stat. Neerl. 64*(4), 505–507.

Novikov, A. and E. Valkeila (1999). On some maximal inequalities for fractional Brownian motions. *Statist. Probab. Lett. 44*(1), 47–54.

Olshen, R. A., E. N. Biden, M. P. Wyatt, and D. H. Sutherland (1989). Gait analysis and the bootstrap. *Ann. Statist. 17*(4), 1419–1440.

Peng, J. and D. Paul (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Comput. Graph. Statist. 18*(4), 995–1015.

Pollard, D. (2010). Asymptopia. Unpublished manuscript.

Qian, B. (2004). Hurst exponent and financial market predictability. *Proceedings of The 2nd IASTED international conference on financial engineering and applications*, 203¨C209.

Radchenko, P. (2008). Mixed-rates asymptotics. *Ann. Statist. 36*(1), 287–309.

Raimondo, M. (1998). Minimax estimation of sharp change points. *Ann. Statist. 26*(4), 1379–1397.

Ramsay, J. O. and C. J. Dalzell (1991). Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B 53*(3), 539–572. With discussion and a reply by the authors.

Ramsay, J. O. and B. W. Silverman (2002). *Applied functional data analysis: methods and case studies*. Springer Series in Statistics. New York: Springer-Verlag.

Revuz, D. and M. Yor (1999). *Continuous martingales and Brownian motion* (Third ed.), Volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Berlin: Springer-Verlag.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B 53*(1), 233–243.

Robinson, L. F., T. D. Wager, and M. A. Lindquist (2010). Change point estimation in multi-subject fmri studies. *NeuroImage 49*(2), 1581–92.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Ann. Statist. 22*(3), 1346–1370.

Salas-Gonzalez, D., E. E. Kuruoglu, and D. P. Ruiz (2009). Modelling and assessing differential gene expression using the alpha stable distribution. *Int. J. Biostat. 5*, Art. 16, 23.

Sammel, M. D., L. M. Ryan, and J. M. Legler (1997). Latent Variable Models for Mixed Discrete and Continuous Outcomes. *Journal of the Royal Statistical Society. Series B (Methodological) 59*(3).

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist. 24*(1), 1–24.

Skrondal, A. and S. Rabe-Hesketh (2003). Some applications of generalized linear latent and mixed models in epidemiology: repeated measures, measurement error and multilevel modeling. *Public Health 13*(2), 265–278.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized latent variable modeling*. Interdisciplinary Statistics. Chapman & Hall/CRC, Boca Raton, FL. Multilevel, longitudinal, and structural equation models.

Stefanski, L. A. (2000). Measurement error models. *J. Amer. Statist. Assoc. 95*(452), 1353–1358.

Tian, T. S. (2010). Functional data analysis in brain imaging studies. *Frontiers in psychology 1*(October), 11.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. New York: Springer-Verlag. With applications to statistics.

Viviani, R., G. Grön, and M. Spitzer (2005). Functional principal component analysis of fmri data. *Human Brain Mapping 24*(2), 109–129.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika 82*(2), 385–397.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika 61*, 439–447.

Wu, J. S. and C. K. Chu (1993). Kernel-type estimators of jump points and values of a regression function. *Ann. Statist. 21*(3), 1545–1566.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc. 100*(470), 577–590.

Yin, Y. Q. (1988). Detection of the number, locations and magnitudes of jumps. *Comm. Statist. Stochastic Models 4*(3), 445–455.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. New York: Springer.

Zipunnikov, V. V. and J. G. Booth (2006). Monte Carlow em for generalized linear mixed models using randomized spherical radial integration. *Statistics*, 1–22.