UNIVERSITY OF CALIFORNIA
RIVERSIDE

Machine Learning Estimation of Nonparametric Econometric Models and Marginal Effects

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Justin Dang

June 2022

Dissertation Committee:

Dr. Aman Ullah, Chairperson
Dr. Tae-Hwy Lee
Dr. Gloria Gonzalez-Rivera

The Dissertation of Justin Dang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my family, friends, loved ones, and advisor, without whose help, I would not have been here.

To my parents and brother for all the support.

ABSTRACT OF THE DISSERTATION

Machine Learning Estimation of Nonparametric Econometric Models and Marginal Effects

by

Justin Dang

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2022
Dr. Aman Ullah, Chairperson

Nowadays, with advanced technology, it is easier to obtain data like never before. With more available data, comes new information that economists can extract to uncover relationships between economic variables. By using new state of the art machine learning algorithms and techniques that can handle data efficiently and can identify trends and patterns easily, we can help solve economic problems, theoretically and empirically. The primary goal of this dissertation is to help bridge the gap between machine learning and econometrics. With powerful machine learning models that exhibit great predictive ability, it would be useful to further explore machine learning methods and add them to our econometrics toolbox. In addition, we wish to extend these models to incorporate problems often faced in econometric models, including partial effects estimation using first derivatives, evaluating concavity of various economic functions using second derivatives, and allowing for heteroskedastic and autocorrelated errors in an econometric model. These issues are clearly often faced in economics, but not so much in machine learning. To incorporate machine learning techniques, machine learning estimation of nonparametric models

and marginal effects are established throughout the dissertation. A derivative estimation procedure of smoothing weighted difference quotients based on random forest is proposed. The procedure of smoothing weighted difference quotients based on random forest is then used to estimate second derivatives. Lastly, a generalized framework for Kernel Regularized Least Squares that incorporates information in the error covariance when estimating the regression function is proposed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many machine learning models can approximate any function we wish to estimate. We focus on machine learning regression techniques that are nonparametric in the sense that they provide data based specification of unknown nonparametric functions. These nonparametric methods can estimate functions based on what the data says and they do not rely on a pre-specified functional form of the data. Throughout the chapters, we focus on two main nonparametric regression methods in the machine learning literature: Kernel Regularized Least Squares (KRLS) and Local Random Forest (LRF). However, the literature on these models are mainly focused on the regression (conditional mean) functions, but lack results on partial derivatives, including marginal effects and concavity of functions, and on issues related to dependent or heteroskedastic observations, which are of some primary interests to economists. Estimating the partial marginal effects involves estimating derivatives of the estimated regression function, which allows economists to better interpret models to see how a change in a variable will affect the outcome. Allowing for non independent and

identically distributed observations allow for a more flexible setting often seen in economic data.

Nowadays, with advanced technology, it is easier to obtain data like never before. With more available data, comes new information that economists can extract to uncover relationships between economic variables. By using new state of the art machine learning algorithms and techniques that can handle data efficiently and can identify trends and patterns easily, we can help solve economic problems, theoretically and empirically. Some of the techniques used in machine learning are similar to those of econometric ones especially those that focus on regression analysis where input data is mapped to some outcome variable. With the help of technology, new and more data can be easily stored. With machine learning methods that can handle data efficiently and can identify trends and patterns easily, economists can uncover relationships between variables. Since machine learning is a subfield of computer science, machine learning methods deal with different data and solve different problems. For example, machine learning data often solve prediction tasks such as face detection: if we were given data on certain images, can we determine if the image contains an image of a face? Other examples include predicting what shows Netflix thinks we like based on what shows we previously watched and predicting products bought on Amazon based on our search history and similar products we purchased.

On the other hand, although economists also deal with prediction tasks, we are also interested in "how" or "why" a variable is important in an economic relationship. As an example, most people go to school to get a higher education to be able to get a higher wage. The economic question is how much more will a person make by going to school for

one more year of schooling? This alludes to the importance of the partial marginal effect of an independent variable on the dependent variable. In other words, economists care about derivatives. Derivative estimation in machine learning is very limited and one goal of this dissertation aims to further expand on estimating derivatives in a machine learning framework. Various machine learning models are considered "black boxes" and with the help of derivatives, we can gain a better understanding of the relationships between variables defined in a model.

Clearly, derivative estimation plays a major role in economics. Derivatives help provide interpretability of the relationships between an independent variable and a dependent variable. In the classical case of estimating a linear parametric model, to find the partial effect of a given variable, the derivative of the estimated regression equation is usually taken. For example, to estimate the marginal propensity to consume (MPC), which is the proportion of extra income that is spent on consumption, we would estimate a function of income, savings, and other determinants and the derivative of the estimated consumption function with respect to income is taken.

However, for certain data based nonparametric or machine learning models, there may not be an analytical form of the estimated regression equation or the derivative and as a result, the partial effect may not be estimated. As an example, the random forest estimator, or any tree based estimator, of the regression function does not have an explicit analytical form of the derivative nor it is smooth. The random forest estimator produces step-wise regression functions that are not analytically differentiable, which is a huge hindrance in determining the partial marginal effect. Chapter 2 proposes a derivative estimator that

is completely data driven to estimate the first derivative. The estimator uses difference quotients, based on a variant of random forests, that are smoothed. Incorporating random forests in estimating derivatives help add more interpretability of forest-based models to explain the relationships between variables via marginal effects. Asymptotic properties of the estimator are established, and the performance of the estimator is addressed in both simulation and in an empirical application of evaluating the partial effects of education and age on income.

In addition to the first derivative, economists also care about the second derivative, which help determine the curvature of the estimated function and how the first derivative changes due to a change in the independent variable. One example of the use of second derivatives include estimating the concavity of a the earnings function. It is often the case that earnings increases slower in later years that in earlier years. In this case, the earnings function would be concave and the first derivative would be positive, while the second derivative is negative. Other examples include determining if a production function exhibits diminishing marginal returns and profit maximization, where if the second derivative is negative, then maximum profit is obtained. Therefore, second derivatives play a major role in many economic settings. Chapter 3 extends on the methods of Chapter 2 by proposing a nonparametric second derivative estimator based on random forest and difference quotients. By estimating derivatives in a random forest framework, more interpretability is added to forest-based models where concavity of estimated random forest-based regression functions can be determined via second derivative estimation. Asymptotic properties are discussed in detail. The second derivative estimator is shown to outperform other methods in simulation

and an empirical exercise is done by evaluating the concavity of a power plant production function.

In other times, various machine learning models make assumptions about the error term, and that they are independent and identically distributed. In various economic data, including time series, the errors are autocorrelated, clearly breaking this assumption. The questions we try to answer in Chapter 4 are can we allow for the errors to be autocorrelated or heteroskedastic? If we allow for errors in this fashion, can we somehow use the information in the error covariance to enhance the regression estimates. Chapter 4 proposes a two-step estimator of a nonparametric regression function via KRLS with parametric error covariance. The naive KRLS, not considering any information in the error covariance, is improved by incorporating a parametric error covariance, allowing for both heteroskedasticity and autocorrelation, in estimating the regression function. A two step procedure is used, where in the first step, the parametric error covariance is estimated from the residuals obtained by a naive regression and in the second step, a KRLS model based on transformed variables from the error covariance is estimated. Theoretical results including bias, variance, and asymptotics are derived. Simulation results show that the proposed estimator outperforms the naive KRLS in both heteroskedastic errors and autocorrelated errors cases. Two empirical examples are illustrated with estimating an airline cost function with heteroskedastic errors and with estimating a money demand equation with autocorrelated errors. The derivatives are evaluated, and the average partial effects of the inputs are determined in these applications.

The primary goal of this disseratation is to help bridge the gap between machine learning and econometrics. With powerful machine learning models that exhibit great predictive ability, it would be useful to further explore machine learning methods and add them to our econometrics toolbox. In addition, we wish to extend these models to incorporate problems often faced in econometric models. In Chapter 2, we propose a derivative estimation procedure of smoothing weighted difference quotients based on random forest. Chapter 3 extends the procedure of smoothing weighted difference quotients based on random forest to estimate second derivatives. Lastly, Chapter 4 provides a generalized framework for KRLS that incorporates information in the error covariance when estimating the regression function. Chapter 5 concludes the dissertation with possible extensions for future research.

Chapter 2

Smoothed Nonparametric Derivative Estimation using Random Forest Based Weighted Difference Quotients

2.1 Introduction

Derivative estimation plays a major role in economics. Derivatives help provide interpretability of the relationships between an independent variable and a dependent variable, for example, changing X by one unit, how much will be the change on Y, holding all else fixed? This is known as the partial marginal effect (first derivative) of X on Y. An economic example of the need of derivatives include estimating the marginal propensity

to consume (MPC), the proportion of extra income that is spent on consumption. In this case, consumption, as a function of income, savings, and other determinants, is estimated. To determine the MPC, the partial derivative of this estimated equation is taken. Another economic example is estimating the effect of schooling or experience on earnings. In labor economics, the relationship between education, experience, and earnings is widely studied. It is often the case that people with more education and experience have higher earnings than those with less education and experience. To quantify how much more a person would earn with more education or experience, the derivative of an earnings regression is estimated to determine the returns to schooling or experience. Clearly, estimating derivatives is vital in understanding and solving economic problems.

After estimating a linear parametric model, to find the partial effect of a given variable, the derivative of the estimated regression equation is usually taken.[1] However, for certain data based nonparametric or machine learning models, there may not be an analytical form of the estimated regression equation and as a result, the partial effect may not be estimated. As an example, the derivative of the machine learning based random forest estimator of the regression function does not have an explicit analytical form nor it is smooth. Forest based regression models produce step-wise regression functions that are not analytically differentiable, which is a huge hindrance in determining the partial marginal effect. Machine learning models are very flexible and can estimate any function well; however, some of these models are considered to be "black boxes." To alleviate this issue, we

[1] In the classical case of Ordinary Least Squares, the derivative of the estimated linear regression function is a constant. However, a constant marginal effect may be too restrictive and the derivative may vary across the space of X. To deal with this, instead of specifying a parametric form of the regression function, resulting in a parametric form of the derivative, both the regression function and the derivative are estimated in a data driven way using nonparametric methods in this chapter.

estimate derivatives of a foreset based model, which allows for a better understanding of the underlying relationship between the data and can be more interpretable in terms of how a change in a variable will change the outcome variable.

In econometrics, the partial marginal effect can be estimated by taking the derivative of the estimated regression function. However, such a marginal effect estimation obtained from the linear or non-linear parametric regression model is well known to be biased and inconsistent unless it is from a correctly specified model, which is rare. When the model is from a data based nonparametric kernel regression, the estimates of derivatives are obtained by the local polynomial regression method in [15], which may not be very smooth. Also, see the estimation of derivatives from spline regression from [32] and [47] . Recently, in the data based machine learning regression area, some interest in estimating derivatives has emerged. For example, [17] estimates derivatives by introducing a machine learning model using logistic function, called boosted smooth transition regression trees (BooST). However, if a logistic specification is not correct then the proposed estimator of derivatives may become statistically inconsistent.

Some other methods have also developed, in which data based difference quotients (DQ) have been utilized to estimate derivatives, where the derivative is determined by taking the ratio of differences between two data points. For example, [43] uses symmetric DQ to run a locally weighted linear regression where the estimate of the derivative is just the intercept term under an equispaced design, where data points are equally spaced along the support of the independent variable. Also, see [4] and [28] papers for the estimation of derivatives using DQ method under the equispaced design. [30] and [31] then apply this

DQ method under the random design, where the regressor X is no longer restricted to be equally spaced apart. Such a procedure is denoted as DQSmooth. The results based on DQSmooth produce noisy estimates of the derivative, and local polynomial regression is then used to smooth the derivatives in attempt to reduce the variance. However, the derivative estimation procedure proposed by [30] has some shortcomings. First, their estimator uses observed data on the dependent variable Y, whereas we propose to consider estimated values of Y, in order to further reduce the variance of the derivative estimator, based on a variant of the machine learning estimator, like random forest. Second, their method only considers the scalar X variable whereas we extend the estimator and its derivatives to include multivariate X, which is commonly used in econometrics.

Random forests procedure [5] is a popular method for estimating nonparametric regression estimation for predictions. However, a drawback of random forests is that they are unable to capture any smoothness in the estimated regression surface. This is most likely why derivative estimation based on random forests is limited, which is due to the nonsmooth-like nature of the estimator. In an effort to address the issue of the inability to fit smooth signals, [18] uses random forests as an adaptive kernel method, where random forests are incorporated in a local polynomial regression framework with a ridge penalty, denoted as Local Linear Forests. Since [18] considers a local linear regression, the derivative can be estimated by the coefficient of the first order derivative of the local linear regression. However, [18] does not focus on estimating the derivative in their paper and instead only estimates the regression function. Instead of using local linear regression, any degree polynomial may be desired, especially if higher order derivatives need to be estimated. This

procedure will be denoted as Local Random Forests (LRF).[2] However, we show in simulation that the derivative obtained by LRF are not only very noisy but they also tend to zero, and thus, LRF is a poor estimator of the derivative.

Under above scenarios, in this chapter, we propose to estimate derivatives using a procedure, denoted as DQSmoothLRF, where difference quotients are first obtained using predictions from LRF. Then, they are smoothed through a local polynomial kernel regression. The proposed derivative estimator contributes to both the nonparametric derivative estimation literature and random forest literature by providing more interpretability of forest based models to explain the relationships between variables via marginal effects estimation. It is shown that the proposed derivative estimator improves substantially relative to LRF considered by [18].

The rest of the chapter is as follows: Section 2.2 goes through the procedure for estimating the first derivative and its properties, Section 2.3 extends the estimator to the multivariate case, Section 2.4 briefly discusses the LRF estimator, Section 2.5 displays the simulation results in comparison to the benchmark estimators, Section 2.6 walks through an emprical example of evaluating the partial effects of education and age on income, and Section 2.7 concludes the chapter.

2.2 First Order Derivative Estimation

Consider the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are independently and identically distributed sampled from a population (X, Y), where $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. Under

[2]Local Random Forest is simply an extension to Local Linear Forests [18], where instead of local linear regression, local polynomial regression of any degree can be used.

the random design, X is a random variable generated from some unkown density f and distribution F. Consider the model

$$Y_i = m(X_i) + e_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $m(X) = \mathbb{E}[Y|X = x]$ is the conditional mean function. Assume that $\mathbb{E}[e|X = x] = 0$ and $\text{Var}[e|X = x] = \sigma_e^2 < \infty$. Now, consider a special case of X where X is standard uniformly distributed. That is, let $X = U \sim \mathcal{U}(0, 1)$, where $\mathcal{U}(0, 1)$ is the uniform distribution between 0 and 1. Now, let $U \sim \mathcal{U}(0, 1)$ and consider the same model but in the special case where $X \sim \mathcal{U}(0, 1)$

$$Y_i = r(U_i) + e_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where $r(u) = \mathbb{E}[Y|U = u]$, $\mathbb{E}[e|U = u] = 0$, and $\text{Var}(e|U = u) = \sigma_e^2 < \infty$. Assume that the bivariate data (U, Y) is ordered with respect to U.

2.2.1 Weighted Difference Quotients

Using the weighted combination of symmetric difference quotients around the ith point from [28], the noisy derivative estimator proposed by [30] under a random design is

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} \omega_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{i+j} - U_{i-j}} \right), \tag{2.3}$$

for $k+1 \leq i \leq n-k$ and hence $k \leq (n-1)/2$, where $\omega_{i,j}$ sum to one for $j = 1, \ldots, k$ and $\widehat{Y}_i^{(1)}$ is an estimator of the first derivative. The term noisy is used to differentiate this derivative estimator and the smoothed derivative estimator later in this section. By minimizing the variance of Eq. (2.3), [30] shows that the optimal weights are

$$\omega_{i,j} = \frac{(U_{i+j} - U_{i-j})^2}{\sum_{l=1}^{k} (U_{i+l} - U_{i-l})^2}, \quad j = 1, \ldots, k, \tag{2.4}$$

where the weights sum to one across $j = 1, \ldots, k$. This estimator is only for the interior points, so Eq. (2.3) as well as the weights need to be modified for the boundaries, with observation $1 < i < k + 1$ being in the left boundary and $n - k < i < n$ being in the right boundary. Note that this estimator will not produce estimates for $i = 1, n$, and these two observations can be ignored.

For this chapter, instead of using observed values of Y for Y_{i+j} and Y_{i-j}, we propose using estimated values from a Local Random Forest (LRF) model, a machine learning (ML) model. First, we estimate Eq. (2.2) by LRF, and produce its fitted values as

$$\widehat{Y}_{i,LRF} = \widehat{r}(U_i), \quad i = 1, \ldots, n. \tag{2.5}$$

In the papers previously mentioned about derivative estimation, all estimate the derivative via the difference quotient using observed values of Y. By estimating the relationship between U and Y first, this allows us to try to pick up the signal from the data before taking the derivative, extend the procedure to the multivariate case, and provide more interpretability of forest based models. Therefore, the proposed noisy derivative estimator is

$$\widehat{Y}^{(1)}_{i,LRF} = \sum_{j=1}^{k} w_{i,j} \frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_{i-j})}{U_{i+j} - U_{i-j}}, \tag{2.6}$$

where $\widehat{Y}^{(1)}_{i,LRF}$ denotes the derivative estimator based on LRF predictions, $\widehat{r}(\cdot)$. By minimizing the variance of Eq. (2.6), the optimal weights are obtained in Prop. 2.1. Let $\mathbb{U} = (U_{i-j}, \ldots, U_{i+j})$ for $i > j,\ i + j \leq n$, and $j = 1, \ldots, k$.

Proposition 2.1 *Under the model in Eq.* (2.2) *and for interior data points,* $k + 1 \leq i \leq n - k$*, minimizing the variance of Eq.* (2.6) *subject to* $\sum_{j=1}^{k} w_{i,j} = 1$ *gives the optimal*

weights (minimum conditional variance) as

$$w_{i,j} = \frac{(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})}, \quad j = 1, \ldots, k, \tag{2.7}$$

where $\sigma^2_{\widehat{r},i\pm j} \equiv \mathrm{Var}[\widehat{r}(U_{i\pm j})|\mathbb{U}]$*, the variance of LRF estimator,* $\widehat{r}(U_{i\pm j})$*, for observation* $i \pm j$*.*

Proof: see Appendix A.1.

We can see that the optimal weights in Eq. (2.7) are similar to that of Eq. (2.4), however these weights depend on the variance of the LRF estimator, $\sigma^2_{\widehat{r},i\pm j}$. If we make the assumption that the variance of the estimator is the same for $i > j,\ i + j \leq n$, and $j = 1, \ldots, k$, then the optimal weights are the same as Eq. (2.4) in [30].

2.2.2 Asymptotic Properties of the Noisy Derivative Estimator

First, notice that the difference $U_i - U_j$ is simply the difference of uniform order statistics, where

$$U_i - U_j \sim \mathrm{Beta}(i - j, n - i + j + 1) \quad \text{for } i > j \tag{2.8}$$

by [10].

Lemma 2.1 *Define* $U \overset{iid}{\sim} \mathcal{U}(0,1)$ *and sort the random variables in order of magnitude such that* $U_1 < \cdots < U_n$*. Then,*

$$\begin{aligned} U_{i+j} - U_{i-j} &= \frac{2j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right) \\ U_{i+j} - U_i &= \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right) \\ U_i - U_{i-j} &= \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right). \end{aligned}$$

Proof: see [30].

This result, combined with Prop. 2.1 leads to the bias and variance of the first order derivative estimator.

Theorem 2.1 *Under the model in Eq.* (2.2) *and assume* $r(\cdot)$ *is twice continuously differentiable on* $[0,1]$, $k \to \infty$ *as* $n \to \infty$, *and under the assumptions of Theorem 1 in [18], with* $\omega \leq 0.2$ *and subsamples of size* s *with* $s = n^{\beta}$, *for*

$$\beta_{min} := 1 - \left(1 + \frac{d}{1.56\pi} \frac{\log(\omega^{-1})}{\log((1-\omega)^{-1})}\right) < \beta < 1,$$

$\mathrm{Var}[\hat{r}(\cdot)] = O(n^{-(1-\beta)})$, *where* ω *is the minimum fraction of parent observations into each child node,* π *is the minimum probability that a variable is split, and* d *is the number of regressors. Then, from the optimal weights obtained in Prop. 2.1, under the uniform random design on the interval* $[0,1]$, *the conditional absolute bias and conditional variance of the proposed derivative estimator in Eq.* (2.6) *for the interior data points* $k+1 \leq i \leq n-k$ *are*

$$\left|\mathrm{bias}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}]\right| \leq \sup_{u\in[0,1]} \left|r^{(2)}(u)\right| \frac{3k(k+1)}{4(n+1)(2k+1)} + o_p\left(n^{-1}k\right) \tag{2.9}$$

$$\mathrm{Var}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] \leq \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} + o_p(n^{(1+\beta)}k^{-3}). \tag{2.10}$$

Proof: see Appendix A.2.

From Theorem 2.1, the pointwise consistency of the derivative estimator can be obtained.

Corollary 2.1 *Under the assumptions of Theorem 2.1,* $k \to \infty$ *as* $n \to \infty$ *such that* $n^{(1+\beta)}k^{-3} \to 0$ *and* $n^{-1}k \to 0$. *Then, using the weights in Prop. 2.1, for any* $\varepsilon > 0$ *and for* $k+1 \leq i \leq n-k$,

$$P\left(\left|\widehat{Y}^{(1)}_{i,LRF} - r^{(1)}(U_i)\right| \geq \varepsilon\right) \to 0 \tag{2.11}$$

Proof: see Appendix A.3.

The parameter k, the number of symmetric differences around the ith data point, depicts the bias-variance tradeoff; larger k increases bias but decreases variance. Therefore, k is chosen by minimizing the asymptotic upper bound of the coniditonal mean integrated squared error (MISE).

Corollary 2.2 *Under the assumptions of Theorem 2.1, the optimal k that is chosen by minimizing the asymptotic upper bound of the conditional MISE is*

$$k_{opt} = \underset{k \in \mathbb{N}^+ \setminus \{0\}}{\arg\min} \left\{ \left(\mathcal{B} \frac{3k(k+1)}{4(n+1)(2k+1)} \right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} \right\}, \tag{2.12}$$

where $\mathcal{B} \equiv \sup_{u \in [0,1]} \left| r^{(2)}(u) \right|$.

Proof: See Appendix A.4.

To find the optimal number of symmetric difference quotients, $\mathcal{B}$ needs to be estimated, which can be done by a local cubic polynomial regression. A grid search for k or any optimization solver can then be used to find the optimal value of k.

Remark 2.1 *Taking the first order condition of Eq.* (2.12)*, we will not get an analytical solution for k_{opt}. However, if we retain the higher order terms, we can get an approximation for k_{opt}*

$$\hat{k}_{opt} = \lfloor 2^{4/5} \widehat{\mathcal{B}}^{-2/5} n^{(3+\beta)/5} \rfloor \tag{2.13}$$

given an estimate for $\mathcal{B}$.

2.2.3 Boundary Correction

So far, we have only discussed points in the interior. In order to reduce the variance, slight modifications to the estimator is needed at the boundaries. Similar to the

boundary corrections made in [4], [8],and [30], the observations lying at the left boundary with index $1 < i < k+1$, the modified weighted difference estimator is

$$\widehat{Y}^{(1)}_{i,LRF} = \sum_{j=1}^{k(i)} w_{i,j} \left(\frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_{i-j})}{U_{i+j} - U_{i-j}} \right) + \sum_{j=k(i)+1}^{k} w_{i,j} \left(\frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_i)}{U_{i+j} - U_i} \right) \tag{2.14}$$

with weights

$$w_{i,j} = \begin{cases} \frac{(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})}{\sum_{l=1}^{k(i)}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})+\sum_{l=k(i)+1}^{k}(U_{i+l}-U_i)^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i})}, & 1 \le j \le k(i) \\ \frac{(U_{i+j}-U_i)^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i})}{\sum_{l=1}^{k(i)}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})+\sum_{l=k(i)+1}^{k}(U_{i+l}-U_i)^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i})}, & k(i) < j \le k \end{cases} \tag{2.15}$$

for $k(i) = i - 1$. Here, the weights are similar to those from the interior in that the weights are standardized by the inverse variances of the LRF estimator. The observations at the right boundary can be estimated in a similar fashion for $k(i) = n - i$.

2.2.4 Smoothing Weighted Difference Quotients

The first order derivative estimators Eq. (2.6) and Eq. (2.14) depend on the variance of the LRF estimator; forest based estimators are very noisy and therefore may affect the variance of the derivative estimator. Another issue of the weighted difference quotient estimator is that it cannot be evaluated at any arbitrary test point, and can only be evaluated for points in the training sample. Therefore, [30] and [31] propose smoothing the estimator via local polynomial regression.

First, observe that the first order derivative estimator in Eq. (2.6), along with their modified boundary correction estimators, will create a new variable for observations $i = 2, \ldots, n-1$. Then, consider the new model,

$$\widehat{Y}^{(1)}_{LRF} = r^{(1)}(U) + \tilde{e}, \tag{2.16}$$

where the first derivative estimator, $\widehat{Y}^{(1)}_{LRF}$, is some unknown first derivative function, $r^{(1)}(\cdot)$, of U, with error, $\tilde{e}$. By construction of the first derivative estimator, the errors will be correlated and assume that $\mathbb{E}[\tilde{e}|U] = 0$, $\text{Cov}[\tilde{e}_i, \tilde{e}_j|U_i, U_j] = \sigma^2_{\tilde{e}}\rho_n(U_i - U_j)$ for $i \neq j$, and $\sigma^2_{\tilde{e}} < \infty$. The correlation function, $\rho_n(\cdot)$ goes to zero as $n \to \infty$ and must satisfy $\rho_n(0) = 1$, $\rho_n(u) = \rho_n(-u)$, and $-1 \leq \rho_n(u) \leq 1$, assumed in [3] and [30]. The goal is then to enhance the noisy derivative estimator $\widehat{Y}^{(1)}_{LRF}$ by nonparametric smoothing. However, the correlation in the errors will affect the bandwidth selection for any nonparametric smoothing and to counteract these effects, [3] proposes using a bimodal kernel K such that $K(0) = 0$ and showed that under mild assumptions, using such a kernel will remove any effects of the correlation on the bandwidth selection without the need to estimate the correlation structure. We will denote the bimodal kernel as $\bar{K}$:

$$\bar{K}(u) = (2/\sqrt{\pi})u^2 \exp(-u^2). \tag{2.17}$$

Next, we fit a local polynomial regression of $\widehat{Y}^{(1)}_{LRF}$ on U. The local polynomial regression estimator of degree p for a given test observation u_0 is provided by minimizing

$$\min_{\beta_L \in \mathbb{R}} \sum_{i=1}^{n} \left\{ \widehat{Y}^{(1)}_{i,LRF} - \sum_{L=0}^{p} \beta_L \left(U_i - u_0\right)^L \right\}^2 K\left(\frac{U_i - u_0}{h}\right) \tag{2.18}$$

where β_L are the solutions to the weighted least squares problem and $K(\cdot)$ is a kernel function. The $(L+1)$th order derivative $r^{(L+1)}(u_0)$ for $L = 0, 1, \ldots, p$ is estimated by $\widehat{r}^{(L+1)}(u_0) = L!\widehat{\beta}_L$. In matrix notation, the solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{U}^\top \mathbf{W} \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{W} \widehat{\mathbf{y}}^{(1)}, \tag{2.19}$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^\top$ is the $(p+1) \times 1$ vector of solutions to the minimization problem, $\mathbf{W} = \text{diag}(K((U_i - u_0)/h))$ is the $(n-2) \times (n-2)$ diagonal matrix of weights based on a

specified kernel function, $\widehat{\mathbf{y}}^{(1)} = (\widehat{Y}^{(1)}_{2,LRF}, \ldots, \widehat{Y}^{(1)}_{n-1,LRF})$ is the $(n-2)\times 1$ vector of the first derivative estimators, and

$$\mathbf{U} = \begin{pmatrix} 1 & (U_2 - u_0) & \cdots & (U_2 - u_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (U_{n-1} - u_0) & \cdots & (U_{n-1} - u_0)^p \end{pmatrix},$$

the $(n-2)\times(p+1)$ centered regression matrix. Therefore, the smoothed first order derivative estimator is

$$\widehat{r}^{(1)}(u_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_1^\top (\mathbf{U}^\top \mathbf{W} \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{W} \widehat{\mathbf{y}}^{(1)} \tag{2.20}$$

where $\boldsymbol{\epsilon}_i$ is a column vector that picks out the ith element.

2.2.5 Asymptotic Properties of the Smoothed Derivative Estimator

Next, we discuss some asymptotic results of the final smoothed derivative estimator in Eq. (2.20). The following theorem states the upper bound of the conditional bias and variance of $\widehat{r}^{(1)}(\cdot)$.

Theorem 2.2 *Under the assumptions in Theorem 2 of [30] and in Theorem 2.1, $k \to \infty$ as $n \to \infty$, and the weights given in Prop. 2.1, the conditional bias and variance of Eq.* (2.20) *for p odd is*

$$\begin{aligned} \text{Bias}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] &\leq \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + \mathcal{B} \frac{3k(k+1)}{4(n+1)(2k+1)} \tilde{c}_p \right] \{1 + o_p(1)\} \\ &= \left[\left(\int t^{p+1} K_0^*(t) dt \right) \frac{1}{(p+1)!} r^{p+2}(u_0) h^{p+1} \right. \\ &\quad \left. + \mathcal{B} \frac{3k(k+1)}{4(n+1)(2k+1)} \left(\int K_0^*(t) dt \right) \right] \{1 + o_p(1)\} \end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}[\widehat{r}^1(u_0)|\tilde{\mathbb{U}}] &\leq \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\frac{1+\rho_c}{h(n-2k)}\boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}\boldsymbol{\epsilon}_1\{1+o_p(1)\},\\ &= \int K_0^{*2}(t)dt\frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\frac{1+\rho_c}{h(n-2k)}\{1+o_p(1)\}\end{aligned}$$

where $\mathcal{B} = \sup_{u\in[0,1]}|r^{(2)}(u)|$, $\mathbf{S} = (\mu_{i+j})_{0\leq i,j\leq p}$ *with* $\mu_j = \int u^j K(u)du$, $\mathbf{S}^* = (\nu_{i+j})_{0\leq i,j\leq p}$ *with* $\nu_j = \int u^j K^2(u)du$, $c_p = (\mu_{p+1},\ldots,\mu_{2p+1})^\top$, $\tilde{c}_p = (\mu_0,\mu_1,\ldots,\mu_p)^\top$, $\boldsymbol{\epsilon}_1 = (1,0,\ldots,0)^\top$, *and the equivalent kernel* $K_0^*(t) = \boldsymbol{\epsilon}_1^\top\mathbf{S}^{-1}(1,t,\ldots,t^p)^\top K(t)$.

Proof: see Appendix A.5.

With smoothing using local polynomial regression, the bandwidth h needs to be estimated. Following [30], we find k and h as follows: k is found by optimizing AMISE in Cor. 2.2 and the bandwidth h is then estimated by using the bimodal kernel $\bar{K}$ in Eq. (2.17), denoted as $\widehat{h}_b$ by cross validation. Then, $\widehat{h}_b$ can be used as a pilot bandwidth which can be related to the bandwidth, $\widehat{h}$, of the usual unimodal kernel, such as the gaussian kernel,

$$K(u) = 1/\sqrt{2\pi}\exp(-u^2/2). \tag{2.21}$$

[4] shows the relationship between the bimodal and unimodal bandwidth,

$$\widehat{h} = 1.01431\widehat{h}_b, \tag{2.22}$$

for local cubic regression using a gaussian kernel.[3] Therefore, after fitting a local cubic regression of $\widehat{Y}_{LRF}^{(1)}$ on U with bimodal kernel, $\bar{K}(\cdot)$, and bandwidth, $\widehat{h}_b$, we refit a local cubic regression with unimodal kernel, $K(\cdot)$ and bandwidth, $\hat{h}$, defined in Eq. (2.22).

[3]For pth degree local regression and for different kernel functions, please see [3].

From Theorem 2.2, for p odd, the pointwise consistency follows.

Corollary 2.3 *Under the assumptions of Theorem 2.1 and Theorem 2.2,* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$, $k \to \infty$ *as* $n \to \infty$ *such that* $n^{-1}k \to 0$ *and* $n^{\beta}k^{-3}h^{-1} \to 0$. *Then, for the weights given in Prop. 2.1, for any* $\varepsilon > 0$,

$$P(|\widehat{r}^{(1)}(u_0) - r^{(1)}(u_0)| \geq \epsilon) \to 0. \tag{2.23}$$

Proof: See Appendix A.6.

2.2.6 Generalizing to Arbitrary Distributions

In general, X may not follow a standard uniform distribution. To generalize X with unknown distribution F, [30] suggests using probability integral transform (PIT)

$$F(X) \sim \mathcal{U}(0,1). \tag{2.24}$$

By using the PIT, we know that the transformed data is $(F(X_1), Y_1), \ldots, (F(X_n), Y_n)$ which has the same distribution as $(U_1, Y_1), \ldots, (U_n, Y_n)$. Then, the same procedure can be used under the transformed data set. Notice, however, that the derivatives, $\widehat{r}^{(1)}(U_i)$ are in the transformed space. Now, to get the derivative in the original space, the chain rule is used

$$\frac{dm(X)}{dX} = \frac{dr(U)}{dU}\frac{dU}{dX} = f(X)\frac{dr(U)}{dU} \tag{2.25}$$

where $m(X) = r(F(X))$. Since the distribution and density, F and f, are unknown, kernel estimators can be used to estimate the distribution and density, with plug-in bandwidths. As a result, the full smoothed first order derivative estimator in the original space is

$$\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U). \tag{2.26}$$

To obtain the estimate of the derivative in the original space, simply multiply the smoothed derivative in the uniform space by the estimate of the density. To summarize, the full algorithm is described in Algorithm 2.1.

Note that this procedure gives the pointwise derivatives of Y with respect to X, which may vary across the space of X. To summarize the overall derivative, we can take the sample average of the derivative estimates to estimate the average derivative, that is

$$\widehat{m}^{(1)}_{avg} = \frac{1}{n'} \sum_{i=1}^{n'} \widehat{m}^{(1)}(X_i) \tag{2.27}$$

This result may be useful in estimating the global partial effect of a variable or even in comparison to the partial effect given by an OLS coefficient.

Algorithm 2.1 Smoothed Nonparametric Derivative Estimation using Weighted Difference Quotients based on LRF

procedure DQSmoothLRF(training independent variable X, training dependent variable Y, number of symmetric difference quotients k, degree of local polynomial $p = 3$, grid of bandwidths to be considered h_{grid}, tunable hyperparameters of LRF model θ)

1: $U \leftarrow \widehat{F}(X)$

▷ kernel cumulative distribution function estimation

2: $\widehat{r}(U) \leftarrow$ regress Y on U by a LRF model with tunable hyperparameters θ

3: $\widehat{Y}_{LRF}^{(1)} \leftarrow$ difference quotient in Eq. (2.6)

4: $\widehat{h}_b \leftarrow$ bandwidth selection from local polynomial regression of $\widehat{Y}_{LRF}^{(1)}$ on U

▷ with degree p, kernel $\bar{K}$ in Eq. (2.17), and searched over bandwidths h_{grid}

5: $\widehat{h} \leftarrow 1.01431\widehat{h}_b$ (for local cubic regression)

▷ bandwidth relationship between bimodal and unimodal gaussian kernels

6: $\widehat{r}^{(1)}(U) \leftarrow$ local polynomial regression

▷ with degree p, unimodal gaussian kernel K, and bandwidth $\widehat{h}$

7: $\widehat{m}^{(1)}(X) \leftarrow \widehat{f}(X)\widehat{r}^{(1)}(U)$

▷ back transform into original space as in Eq. (2.26) after kernel density estimation of X

2.3 Extension to the Multivariate Case

It is very rare to have models with univariate X in economics. In this section we try to extend the procedure to the multivariate case. Suppose we have the model

$$Y_i = m(X_{i,1}, \ldots, X_{i,d}) + e_i, \quad i = 1, \ldots, n, \tag{2.28}$$

where d is the number of independent variables. For now, suppose all of the regressors are standard uniformly distributed. Then, consider the regression of the form

$$Y_i = r(U_{i,1}, \ldots, U_{i,d}) + e_i. \tag{2.29}$$

Then, we follow the same steps as before. First, we estimate $r(\cdot)$ by LRF. Now, the first partial derivative with respect to the sth variable is given by the weighted difference quotient

$$\widehat{Y}^{(1)}_{i,s,LRF} = \sum_{j=1}^{k_s} w_{i,j} \frac{\widehat{r}(U_{i+j,s}, \bar{U}) - \widehat{r}(U_{i-j,s}, \bar{U})}{U_{i+j,s} - U_{i-j,s}}, \tag{2.30}$$

where $\bar{U}$ contains the other $d-1$ variables evaluated at their medians. Note that we order the data with respect to the sth variable and that variables can have different number of symmetric difference quotients denoted by k_s. We can then use this estimate of the first partial derivative as the dependent variable in a local polynomial regression on $U_1, \ldots, U_d$,

$$\widehat{Y}^{(1)}_{s,LRF} = r^{(1)}_s(U_1, \ldots, U_d) + \tilde{e}, \tag{2.31}$$

where the subscript s denotes the derivative with respect to the sth variable. Since the errors are correlated, we use the multivariate bimodal kernel,

$$\bar{K}(\mathbf{u}) = (2d/\pi^{\frac{d}{2}})||\mathbf{u}||^2 \exp(-||\mathbf{u}||^2). \tag{2.32}$$

Then, we can correct the bandwidth using the relationship between bimodal and unimodal kernel functions as in [4] for each bandwidth to get an estimate of the partial derivative, $\widehat{r}_s^{(1)}(U_s, \bar{U})$.

Now, suppose that X_m for $m = 1 \ldots, d$ are not all standard uniformly distributed. First, we can estimate the joint CDF function of all independent variables, $F_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$. To obtain the marginal CDFs of a specified variable, we take the limits in the arguments of the joint CDF of the other variables, $F_{X_s}(x_s) = F_{X_1,\ldots,X_d}(+\infty, \ldots, x_s, \ldots, +\infty)$. Then, we have for each regressor,

$$F_{X_m}(X_m) \sim \mathcal{U}(0,1), \quad m = 1, \ldots, d. \tag{2.33}$$

Therefore, the new data $(F_{X_1}(X_{1,1}), \ldots, F_{X_d}(X_{1,d}), Y_1), \ldots, (F_{X_1}(X_{n,1}), \ldots, F_{X_d}(X_{n,d}), Y_n)$, has the same distribution as $(U_{1,1}, \ldots, U_{1,d}, Y_1), \ldots, (U_{n,1}, \ldots, U_{n,d}, Y_n)$.

$$\begin{aligned} Y_i &= r(F_{X_1}(X_{i,1}), \ldots, F_{X_d}(X_{i,d})) + e_i \\ &= r(U_{i,1}, \ldots, U_{i,d}) + e_i, \end{aligned} \tag{2.34}$$

which is the case where we have all standard uniform variables as regressors. To get derivatives in the original space,

$$\frac{\partial m(X_s, \bar{X})}{\partial X_s} = \frac{\partial r(U_s, \bar{U})}{\partial U_s} \frac{\partial U_s}{\partial X_s} = f_{X_s}(X_s) \frac{\partial r(U_s, \bar{U})}{\partial X_s} \tag{2.35}$$

Using a multivariate kernel density estimator for $f_{X_1,\ldots,X_d}$, and marginalizing for the *sth* variable, the final smooth partial derivative with respect to the sth variable is

$$\widehat{m}_s^{(1)}(X_s, \bar{X}) = \widehat{f}_{X_s}(X_s) \widehat{r}_s^{(1)}(U_s, \bar{U}), \tag{2.36}$$

where $\widehat{f}_{X_s}(X_s) = \sum_{x_1} \cdots \sum_{x_{s-1}}, \sum_{x_{s+1}} \cdots \sum_{x_d} \widehat{f}_{X_1,\ldots,X_d}(x_1, \ldots, x_s, \ldots, x_d)$. Note that $\bar{X}$ and $\bar{U}$ mean we are holding the other variables fixed (at their medians or some other fixed

constant). So, we evaluate the derivatives holding the other variables fixed at their medians in the X space and transform these fixed values in the U space when fitting a regression function and smoothing the derivatives.

2.4 Local Random Forest

This section briefly discusses the Local (Linear) Random Forest (LRF) estimator. In what follows, assume the i.i.d data $(\mathbf{x}_i, y_i)$, for $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and assume the model in Eq. (2.1), either for the univariate case ($d = 1$) or the multivariate case ($d > 1$). LRF uses a local polynomial regression with $p = 1$ (local linear) with forest based weights instead of kernel based weights [18]. The objective function includes a weighted quadratic loss with L_2 regularization,

$$\arg\min_{\delta_q} \sum_{i=1}^{n} \left(y_i - \sum_{q=0}^{p} \delta_q^T (\mathbf{x}_i - \mathbf{x}_0)^q \right)^2 a_i(\mathbf{x}_i, \mathbf{x}_0) + \lambda ||\delta_1||^2 \tag{2.37}$$

where λ is the regularization strength parameter and $a_i(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbf{1}\{x_i \in L_b(\mathbf{x}_0)\}}{|L_b(\mathbf{x}_0)|}$ is the weight function. Here, B is the number of bootstrap replications, $\mathbf{1}\{\cdot\}$ is the indicator function, and $L_b(\mathbf{x}_0)$ denotes the the leaf of the b-th bootstrapped tree that contains the testing point $\mathbf{x}_0$. δ_q, for $q = 0, 1$ denotes the conditional mean function and its derivative evaluated at $\mathbf{x}_0$. The minimization problem is the same as in Eq. (2.18), except for the weight function and the regularization term. Using similar notation as before, the minimization problem can be expressed in matrix form as

$$\arg\min_{\boldsymbol{\delta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\delta})^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\delta}) + \lambda \boldsymbol{\delta}^\top \mathbf{J} \boldsymbol{\delta}, \tag{2.38}$$

where $\mathbf{J}$ is the identity matrix with the first diagonal element as zero, $\mathbf{X}$ is the centered regression matrix with the first column being ones, $\mathbf{W}$ is the diagonal matrix of weights $a(\cdot)$, $\mathbf{y}$ is the vector of the dependent variable, and $\boldsymbol{\delta}$ is the gradient vector with the first element being the estimate for the mean regression function. Then, the solution is

$$\widehat{\boldsymbol{\delta}} = (\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{J})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{y} \tag{2.39}$$

and the resulting prediction function evaluated at a test point $\mathbf{x}_0$ is

$$\widehat{m}(\mathbf{x}_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\delta}}. \tag{2.40}$$

To find λ, a random forest model is trained first and the weights $a_i(\cdot)$ are obtained from the forest. Then, the regularization parameter λ can be found by cross validation. LRF can then be used in step 3 of Algorithm 2.1, with transformed regressor, U.

As a reference, the first derivative can be obtained from LRF.

$$\widehat{m}^{(1)}(\mathbf{x}_0) = \boldsymbol{\epsilon}_2^\top \widehat{\boldsymbol{\delta}} \tag{2.41}$$

However, as stated in Section 2.1, the derivative based on LRF will be noisy and tend to zero due to the ridge penalty. Even when there is no ridge penaly, $\lambda = 0$, the derivative will still have high variance. This is due to the nature of the high variance in random forests. In the following section, we will show that there is a huge improvement when the derivative based on LRF is compared to the proposed estimator, DQSmoothLRF.

2.5 Simulations

This section shows derivative estimation based on the proposed estimator, DQSmoothLRF as well as other estimators. To compare results, we consider the following data

generating process (DGP), which is from [31] with sample size $n = 700$ and $e \sim N(0, 0.2^2)$. Since we know the true function, the analytical expression of the derivative is also known.

$$m(X) = \cos(2\pi X)^2 \quad \text{for } X \sim \text{beta}(2, 2) \tag{2.42}$$

$$m^{(1)}(X) = -4\pi \cos(2\pi X) \sin(2\pi X). \tag{2.43}$$

For all simulations, we estimate the density f and distribution F by the `R` package `ks` [14]. The parameter k, the number of symmetric difference quotients, is estimated by Cor. 2.2. For the weights $w_{i,j}$ based on the variances of the LRF estimator in Eq. (2.7), we assume that the variances are constant for the k points round i, so that the weights collapse to Eq. (2.4). Therefore, plots and results show the derivative estimators based on the weights obtained in Eq. (2.4). We do this so that results under the proposed model can be easily compared to the benchmark model, the model proposed by [30], and since the simulated errors are homoskedastic, this assumption may be reasonable.

We show estimates of the first derivative considering four different models, (1) DQSmooth, the benchmark model of the derivative based on difference quotients proposed by [30], (2) DQSmoothLRF, the proposed estimator based on LRF estimates of Y, (3) LocCubic, a local cubic regression, a common regression technique to estimate derivatives in the nonparametric literature, and (4), LRF, the model proposed by [18], a benchmark model of the derivative based on random forests. For the last estimator, LRF, the original paper by [18] focuses on estimating the conditional mean function, not its derivatives. The paper also considers only a local linear approach. For these simulations, we consider a local cubic with $\lambda = 0$ for the LRF estimator. The reason for zero ridge penalty, is that since the ridge parameter, λ, penalizes the curvature of the function, it will force the derivative

estimates toward zero, although the conditional mean function may not be flat. All local polynomial regressions are estimated using `locpol` package [6]. When estimating LRF based models, we use the `grf` package [42]. To assess the models, we use mean squared error (MSE) and mean absolute error (MAE), defined as

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}\left(\widehat{m}^{(1)}(X_j) - m^{(1)}(X_j)\right)^2 \tag{2.44}$$

$$\text{MAE} = \frac{1}{m}\sum_{j=1}^{m}\left|\widehat{m}^{(1)}(X_j) - m^{(1)}(X_j)\right|. \tag{2.45}$$

We evaluate all models at 500 evenly spaced points from 0.05 to 0.95, where $m = 500$..

Results for the first derivative are plotted in Figure 2.1. The grey curves depict one estimated first derivative function for a simulation run, where all simulations are plotted. The solid colored curves represent the average of all simulations at each of the 500 evenly spaced points of X from 0.05 to 0.95 and the black curve represents the true derivative we wish to estimate. As seen from the figure, all models on average seem to estimate the true first derivative accurately. However, the difference is notable due to the variance of the models, which is depicted by how far away on average the grey curves are from their solid curves. By first glance, both DQsmooth and DQSmoothLRF seem to have lower variance and are more accurate in the sense of lower variability. In addition, DQSmoothLRF also appears to have slightly smaller variability in estimating the first derivative compared to DQSmooth. This is a direct result of using estimated values of Y by a LRF, where the signal is picked up from the noise, instead of using raw values of Y, which is the case for DQSmooth. The LocCubic and LRF models seem to have larger variability than the DQSmooth models. Notice the extremely large variability in the LRF estimator for the first

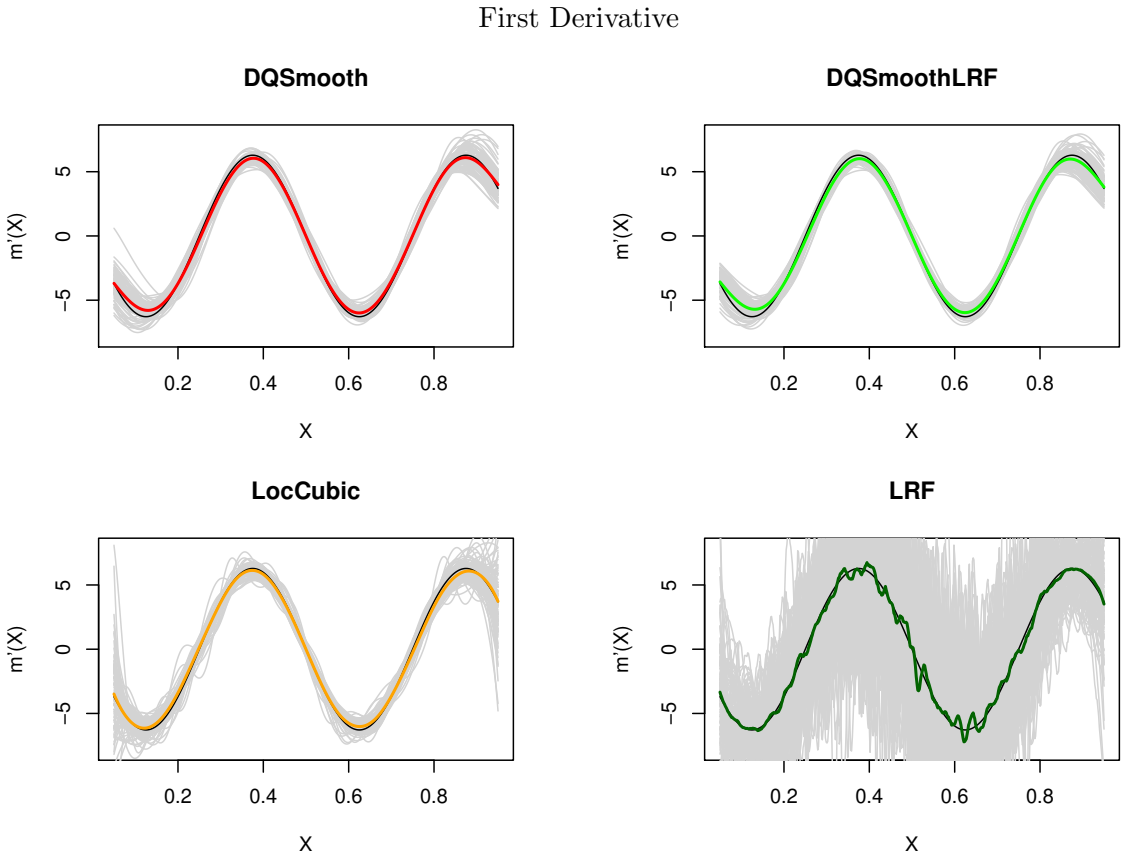

Figure 2.1: *Each plot shows the estimates for the first derivative for DQSmooth [30], DQSmoothLRF (the proposed estimator), LocCubic, and LRF estimators. Note that the first derivative estimator based on LRF is for* $\lambda = 0$*. The grey curves in each plot depict estimates of the first derivative for one simulation, where 100 simulations are plotted. The solid colored curves represent the mean predicted values across all simulations. The solid black curve represents the true first derivative. All estimators are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

derivative; this justifies the need to enhance derivative estimates based on LRF.

Results for the simulations are shown in Table 2.1, where the bias, variance, MSE, and MAE are reported for the first derivative across the four models under consideration. First, the mean bias from DQSmooth and DQSmoothLRF estimators are roughly the same indicating that the bias of derivative estimates for DQSmoothLRF is similar to that of the bias for DQSmooth. However, improvement on DQSmooth is shown through the variance of

the DQSmoothLRF estimator, with a 20% reduction in the variance relative to the variance of DQSmooth. Although the LocCubic model has lower absolute mean bias compared to all models, the variance is over double of those of the DQSmooth models. The LRF estimator performs the worst and we can see a significant improvement in estimates for the first derivative estimator when using DQSmoothLRF. Lastly, DQSmoothLRF performs the best in terms of both MSE and MAE compared to all models for the first derivative estimation. Overall, in these simulations, we have shown that DQSmoothLRF outperforms all other estimators in terms of variance, MSE, and MAE.

Model Assessment

First Derivative	Bias	Variance	MSE	MAE
DQSmooth	0.0168	0.2618	0.3194	0.4221
DQSmoothLRF	0.0121	0.2085	0.2943	0.4180
LocCubic	-0.0045	0.4437	0.4789	0.4565
LRF	-0.0209	12.9508	13.1182	2.7767

Table 2.1: *The top and bottom panel show the bias, variance, MSE, and MAE for the first derivative, comparing the four models, DQSmooth [30], DQSmoothLRF (the proposed estimator), LocCubic, and LRF. All estimates are averaged across all simulations. Note that the results based on LRF is for* $\lambda = 0$*. All models are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

2.6 Empirical Application: Partial Effects of Education and Age on Income

In this section, we apply the DQSmoothLRF procedure to a labor economic example, where the relationshiop between earnings, education, age, and sex is analyzed. Since the DQSmoothLRF procedure can estimate partial derivatives for continuous variables only, we will evaluate the partial effects of education and age on income, holding other variables fixed. That is, suppose the relationship between the variables come from the following model.

$$income = m(educ, age) + e, \tag{2.46}$$

where *income* is income, *educ* is education level in years, and *age* is the age of the individual. On average, higher levels of income tend to be associated with higher levels of education. As a result, the partial effect of education on income is expected to be positive, on average. Another common finding is that income tends to increase slower in age, implying that the partial effect changes over the course of an individual's earning years. Lastly, sex is a common determinant of one's income, in which females on average tend to have lower income than males. In this exercise, to account for sex, we will examine the partial effects of age and education on an individual's income separately for females and males. That is, we will estimate two models and estimate the partial effects: one for females and one for males. In this example, we will determine if the partial effects are consistent with the literature and in addition if the partial effects differ for males and females.

As the benchmark model of income, let $m(\cdot)$ have the following linear specificaiton,

$$m(educ, age) = \beta_1 + \beta_2 age + \beta_3 age^2 + \beta_4 educ + \beta_5 age \times educ, \tag{2.47}$$

which will be estimated by Ordinary Least Squares (OLS). From this specification, *income* is quadratic in *age*, and *educ* is interacted with *age*, which allows for heterogeneous partial effects of education for different ages. The interaction term between *age* and *educ* also allows for heterogeneous partial effects of age for different levels of education. The linear regression function, specified by Eq. (2.47) gives the following partial effects with respect to education and age, holding all other factors fixed.

$$m^{(1)}_{educ}(educ, \overline{age}) = \beta_4 + \beta_5 \overline{age} \tag{2.48}$$

$$m^{(1)}_{age}(\overline{educ}, age) = \beta_2 + 2\beta_3 age + \beta_5 \overline{educ}, \tag{2.49}$$

where other variables are held fixed at their medians, or some constant. For this specification, the partial effect of education is constant, holding age fixed. On the other hand, the partial effect of age is not constant across the support of age, holding education fixed.

The data analyzed in this example come from an unbalanced panel of 7,293 households from the German Socioeconomic Panel data set.[4] In the data set, the variables include income, age, gender, education, and other sociodemographic variables. In this data set, $female = 1$ for females and 0 for males, *age* is age in years, *educ* denotes years of schooling, and *income* represents the household nominal monthly net income in German marks/1,000,000. For this empirical exercise, we will use 500 random observations from a cross section of 4,481 observations for the year 1988. From the 500 randomly chosen

[4]The data used are obtained from [38] and can be obtained from `http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/`.

observations, 250 are males and 250 are females. We will estimate the regression function specified in Eq. (2.47) and the nonparametric form and its derivatives in Eq. (2.46) using the DQSmoothLRF procedure for each subsample.

Figure 2.2 shows the pointwise marginal effects of education (Figure 2.2a) and age (Figure 2.2b) on income, holding other factors fixed. In each plot, the marginal effects are plotted for males and females in red and blue lines, respectively. Each partial derivative is evaluated for 200 evenly spaced points across the support of the variable of interest. In Figure 2.2a, the variable *age* is set to be constant at the median, $\overline{age} = 45$, and in Figure 2.2b, the variable *educ* is set to be constat at the median, $\overline{educ} = 10.5$.[5] The partial effect of education on income is positive throughout most of the entire support of education, and the largest effect occurs when an individual has around 11 years of education. At around this level of education, individuals would soon complete their high school degree and the effect of increasing education by one more year, or finishing their high school education, is the largest effect compared to any other level of education. The partial effect of education for females (blue curve) is greater than that for males (red curve) in the earlier levels of education. This would indicate that there is some interaction effect between education and sex of the individual, and that the partial effect of education on income depends on sex.

The partial effect of age is positive in the beginning of the early years of an individual and negative in the later years. Since the partial derivative crosses the zero line at around age 45 to 49, this implies that income increases slower in age and income is parabolic in age. For males, the partial derivative crosses the zero line at 49, suggesting that

[5]The medians are calculated from the sample of 500 chosen observations.

(a) $\widehat{m}^{(1)}(educ, \overline{age})$ **(b)** $\widehat{m}^{(1)}(\overline{educ}, age)$

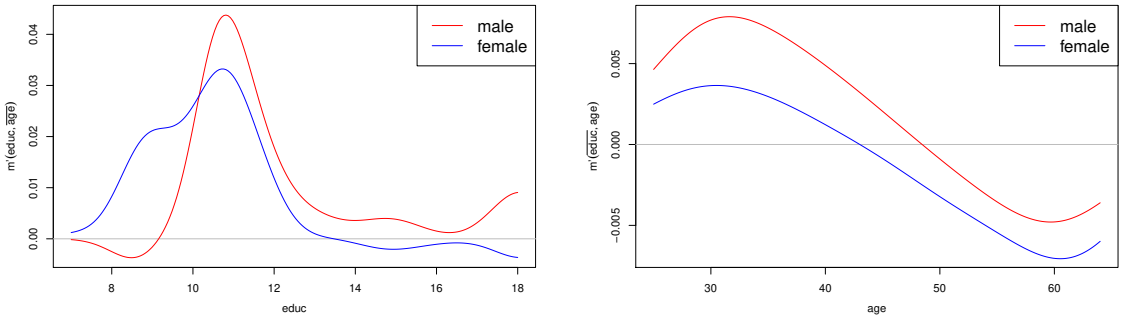

Figure 2.2: *The figures report the pointwise marginal effects of education (Figure 2.2a) and age (Figure 2.2a) on income, holding all else fixed. The red and blue curves are the partial effects for males and females, respectively. For the continuous variables,* $\overline{age}$ *and* $\overline{educ}$*, they are set at their medians.*

males make the most income at 49, holding all other factors fixed. For females, the partial derivative crosses the zero line a little earlier at age 45, implying that females make the most income at 45, holding all other factors fixed. Furthermore, the partial effect is larger for males compared to females in their earlier years and less negative for males compared to females in their later years. Since the partial effect of age on income is clearly different for females and males, the partial effect of age depends on sex.

The average partial effects of age and education on earnings for DQSmoothLRF and OLS are reported in Table 2.2. For fixed variables, the variable *age* and *educ* are fixed at their respective medians. These estimates are simply the averages of the pointwise derivatives displayed in Figure 2.2. Looking at the partial derivatives, holding all else fixed at their medians, the partial effect of education is 0.0091 for males and 0.0085 for females, estimated by DQSmoothLRF compared to 0.0126 for males and 0.0149 for females,

Average Partial Effects of Education and Age on Income

	Education		Age	
	DQSmoothLRF	OLS	DQSmoothLRF	OLS
Male	0.0091	0.0126	0.0018	0.0013
Female	0.0085	0.0149	-0.0012	-0.0026

Table 2.2: *The table reports the average partial effect of age and education on earnings, holding all other factors fixed at their medians, for DQSmoothLRF and for the derivatives estimated by OLS in Eq.* (2.48) *and Eq.* (2.49)*. The average is taken across 200 evenly spaced points of the support of each independent variable.*

estimated by OLS. To get a better understanding of the partial effects, we can look at how large the partial effect is in comparison to the median of income in the sample, which is 0.33. Therefore, since the median income in the data is 0.33, the partial effects suggest that an additional year of eduation is associated witht a change in predicted income of about 2.76% (i.e. 0.0091/0.33) for for males and 2.56%, estimated by DQSmoothLRF and 3.82% for females and 4.52% for females estimated by OLS. In this case, OLS estimates report a higher average partial effect whereas DQSmoothLRF estimates say the opposite. Some studies including [13] find that returns to schooling (i.e. the partial effect of education) is higher for women than for men. A possible explanation is that education has a double effect on females' earnings, where education increases females' skills and productivity and reduces the gap in male and female earnings attributable to factors such as discrimination. Although the the average partial effect of education for males is larger than females, when estimated by DQSmoothLRF, the pointwise partial derivative plot with respect to education

in Figure 2.2a shows that the partial effect is larger for females than males, but only in earlier levels of education.

The average partial effect of age on income is positive for males when estimated by DQSmoothLRF and OLS, holding everything else fixed at their medians, although DQSmoothLRF estimates a larger positive average partial effect for males. However, the average partial effect of age is negative for females, with DQSmoothLRF reporting a less severe average effect. Recall that the pointwise partial effects of age on income displayed in Figure 2.2b are positive in earlier years and negative in later years for both males and females. When taking the average, the overall average partial effect of age is positive for males. However, for females, the average partial effect is negative. This would imply that the partial effect is much larger for males in earlier years and smaller in later years relative to females. Similar to the pointwise partial effects analysis, the average partial effects are clearly different for females and males.

Overall, in this empirical application, we estimate the partial effects of education and age on income. We find that the partial effect of education is positive and that the effect is larger for females than males, but only in lower levels of education. Moreover, the partial effect of age is positive in the earlier years of life and negative in the later years and that the effect is larger in the earlier years and less severe in later years for males relative to females. In addition to the pointwise partial derivatives, the average partial derivatives also show that the partial effects of education and age depend on sex.

2.7 Conclusion

Overall, derivatives can help economists find the partial marginal effect of a variable. In this chapter, we propose a method, DQSmoothLRF, that smooths random forest based difference quotients to estimate first derivatives. We improve on the original method in [30] by using estimated values of the dependent variable from LRF in forming difference quotients, instead of using the dependent variable itself, in hopes of reducing variance and by including multiple variables in the model, instead of the simple univariate case. Improvement is also made in comparision to derivatives estimated by LRF in [18], where derivatives of the regression equation were not even focused on and in providing better interpretation for forest based models by evaluating derivatives derived from random forests. We have shown in simulation that the proposed estimator outperforms the benchmark ones as well as a popular method of estimating derivatives in economics, local polynomial regression; a reduction in variance, MSE, and MAE are all evident when using the proposed estimator. Lastly, we provide an empirical example using household income data to evaluate the partial effects of education and age and find that these partial effects differ for males and females.

Chapter 3

Smoothed Nonparametric Second Derivative Estimation using Random Forest Based Weighted Difference Quotients

3.1 Introduction

At times, economists are interested in the second derivative, which help depict the curvature of a function and how the slope changes due to a change in the independent variable. For example, with second derivatives, we can determine whether the earnings function is concave or convex or see if a production function exhibits diminishing marginal returns. In another economic instance where second derivatives are needed, in profit maximization,

if the second derivative is negative, then maximum profit is obtained. Clearly, second order derivatives are vital in various economic applications.

In econometrics, second derivatives can be estimated by taking the second derivative of the estimated regression function. However, such estimation obtained from the linear or non-linear parametric regression model is well known to be biased and inconsistent unless it is from a correctly specified model, which is rare. When the model is from a data based nonparametric kernel regression, the estimates of derivatives are obtained by the local polynomial regression method in [15], which may not be very smooth.

Recently, in the data based machine learning regression area, some interest in estimating derivatives has emerged. For example, data based difference quotients (DQ) have been utilized to estimate derivatives, where the derivative is determined by taking the ratio of differences between two data points. See Section 2.1 for more information about DQ method. To estimate the second derivative by DQ, the difference of the first order DQ is taken, i.e. second order difference quotients. To further enhance DQ, [30] applies local polynomial regression to smooth the second derivatives estimated by DQ in attempt to reduce the variance. However, the derivative estimation procedure proposed by [30] has some shortcomings. First, their estimator uses observed data on the dependent variable Y, whereas we propose to consider estimated values of Y, in order to further reduce the variance of the derivative estimator, based on a variant of the machine learning estimator, like random forest. Second, their method only considers the scalar X variable whereas we extend the estimator and its second derivatives to include multivariate X. This chapter essentiallly extends Chapter 2 to second derivatives.

Random forests procedure [5] is a popular method for estimating nonparametric regression estimation for predictions. However, a drawback of random forests is that they are unable to capture any smoothness in the estimated regression surface. This is most likely why second derivative estimation based on random forests is limited, which is due to the nonsmooth-like nature of the estimator. In an effort to address the issue of the inability to fit smooth signals, [18] uses random forests as an adaptive kernel method, where random forests are incorporated in a local polynomial regression framework with a ridge penalty, denoted as Local Linear Forests, although derivatives are not thoroughly explored. Instead of using local linear regression, any degree polynomial may be desired, especially if higher order derivatives need to be estimated, such as in this case where we study the second derivative. The second derivative can be estimated by the coefficient of the second order derivative of the local polynomial regression.This procedure will be denoted as Local Random Forests (LRF).[1] However, we show in simulation that the second derivative obtained by LRF are very noisy, and thus, LRF is a poor estimator of the second derivative.

Under above scenarios, in this chapter, we propose to estimate second derivatives using a procedure, denoted as DQ2SmoothLRF, where second order difference quotients are first obtained using predictions from LRF. Then, they are smoothed through a local polynomial kernel regression. The proposed second derivative estimator contributes to both the nonparametric derivative estimation literature and random forest literature by providing more interpretability of forest based models to explain the relationships between variables via second order derivative estimation. It is shown that the proposed second

[1]Local Random Forest is simply an extension to Local Linear Forests [18], where instead of local linear regression, local polynomial regression of any degree can be used.

derivative estimator improves substantially relative to LRF considered by [18] and to the second derivative procedure proposed by [30].

The rest of the paper is as follows: Section 3.2 discusses second derivative estimation and its properties, Section 3.3 briefly discusses the LRF estimator, Section 3.4 displays the simulation results in comparison to the benchmark estimators, Section 3.5 walks through an emprical example of evaluating the concavity of a Chilean power plant production function, and Section 3.6 concludes the paper.

3.2 Second Order Derivative Estimation

Consider the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are independently and identically distributed sampled from a population (X, Y), where $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. Under the random design, X is a random variable generated from some unkown density f and distribution F. Consider the model in Eq. (2.1), where $m(X) = \mathbb{E}[Y|X = x]$ is the conditional mean function. Assume that $\mathbb{E}[e|X = x] = 0$ and $\text{Var}[e|X = x] = \sigma_e^2 < \infty$. Similar to the first derivative, consider a special case of X where X is standard uniformly distributed. That is, let $X = U \sim \mathcal{U}(0, 1)$, where $\mathcal{U}(0, 1)$ is the uniform distribution between 0 and 1. Assume that the bivariate data (U, Y) is ordered with respect to U.

The second order weighted difference quotient proposed by [30] is

$$\widehat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{Y_{i+j+k_1} - Y_{i+j}}{U_{i+j+k_1} - U_{i+j}} - \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{i-j-k_1} - U_{i-j}}\right)}{U_{i+j+k_1} - U_{i+j} - U_{i-j-k_1} - U_{i-j}}, \tag{3.1}$$

with weights

$$w_{i,j,2} = \frac{(2j + k_1)^2}{\sum_{j=1}^{k_2}(2j + k_1)^2}, \tag{3.2}$$

where k_1 and k_2 are positive integers that represent the number of first and second order difference quotients about observation i such that the weights $w_{i,j,2}$ sum to one across j. Note that the estimator is for observations in the interior for $k_1+k_2+1 \leq i \leq n-k_1-k_2$. Also note that [30] chooses the second order derivative weights to be proportional to the inverse of the conditional variance of each quotient. Similar to the first order derivative, we replace Y with estimates of Y using the LRF estimator, denoted by $\widehat{r}$. Define ${}^{+}\widehat{Y}^{(1)}_{i+j,LRF} = \frac{\widehat{r}(U_{i+j+k_1})-\widehat{r}(U_{i+j})}{U_{i+j+k_1}-U_{i+j}}$ and ${}^{-}\widehat{Y}^{(1)}_{i-j,LRF} = \frac{\widehat{r}(U_{i-j-k_1})-\widehat{r}(U_{i-j})}{U_{i-j-k_1}-U_{i-j}}$. This leads to the following proposition.

Proposition 3.1 *Under the assumptions of Theorem 2.1 and $r(\cdot)$ is three times differentiable, the estimator for the second derivative based on symmetric difference quotients is*

$$\widehat{Y}^{(2)}_{i,LRF} = \sum_{j=1}^{k_2} w_{i,j,2} \frac{{}^{+}\widehat{Y}^{(1)}_{i+j,LRF} - {}^{-}\widehat{Y}^{(1)}_{i-j,LRF}}{C_{i,j,k_1}}, \tag{3.3}$$

where $C_{i,j,k_1} = (U_{i+j+k_1} - U_{i+j} - U_{i-j-k_1} - U_{i-j})/2$ is chosen such that each individual quotient $\frac{{}^{+}\widehat{Y}^{(1)}_{i+j,ML} - {}^{-}\widehat{Y}^{(1)}_{i-j,ML}}{C_{i,j,k_1}}$ is an asymptotically unbiased estimator of the second order derivative $r^{(2)}(\cdot)$ for $j = 1, \ldots, k_2$ and where each weight is selected to be proportional to the inverse of the conditional variance of each quotient:

$$w_{i,j} = \frac{\frac{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})^2}{{}^{+}V_{i+j}+{}^{-}V_{i-j}}}{\sum_{j=1}^{k_2} \frac{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})^2}{{}^{+}V_{i+j}+{}^{-}V_{i-j}}}, \tag{3.4}$$

with ${}^{+}V_{i+j} \equiv \frac{\sigma^2_{\widehat{r},i+j+k_1}+\sigma^2_{\widehat{r},i+j}}{(U_{i+j+k_1}-U_{i+j})^2}$ and ${}^{-}V_{i-j} \equiv \frac{\sigma^2_{\widehat{r},i-j-k_1}+\sigma^2_{\widehat{r},i-j}}{(U_{i-j-k_1}-U_{i-j})^2}$.

Proof: See Appendix B.1.

Applying Lemma 2.1, and using the leading order of the weight, the weight can be approximated by

$$w_{i,j,2} = \frac{(2j+k_1)^2/(\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j})}{\sum_{j=1}^{k_2}(2j+k_1)^2/(\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j})}. \tag{3.5}$$

Notice that if the variances are approximately the same for the $2 \cdot (k_1 + k_2)$ observations around i, the the proposed weights will be equal to that of [30] in Eq. (3.2).

Theorem 3.1 *Assume the model in Eq.* (2.2)*, r is three times continuously differentiable on* $[0, 1]$*, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$, and the assumptions of Theorem 1 in [18], with* $\mathrm{Var}[\hat{r}(\cdot)] = O(n^{-(1-\beta)})$*. Then, from the optimal weights obtained in Eq.* (3.5)*, under the uniform random design on the interval* $[0, 1]$*, the conditional absolute bias and conditional variance of the proposed second derivative estimator in Eq.* (3.3) *for the interior data points* $k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2$ *are*

$$\left| \mathrm{Bias}[\hat{Y}^{(2)}_{i,LRF} | \tilde{\mathbb{U}}] \right| \leq \frac{\sup_{u \in [0,1]} |r^{(3)}(u)|}{n+1} \times \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\} \tag{3.6}$$

$$\mathrm{Var}[\widehat{Y}^{(2)}_{i,LRF} | \mathbb{U}] \leq \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\}. \tag{3.7}$$

Proof: see Appendix B.2.

From Theorem 3.1, the pointwise consistency of $\hat{Y}^{(2)}_{i,LRF}$ can be obtained.

Corollary 3.1 *Under the assumptions of Theorem 3.1, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^{3+\beta}k_1^{-2}k_2^{-3} \to 0$, and $n^{3+\beta}k_1^{-4}k_2^{-1} \to 0$. Then, using the weights in Eq.* (3.5)*, for any $\varepsilon > 0$ and for $k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2$,*

$$P\left(\left| \widehat{Y}^{(2)}_{i,LRF} - r^{(2)}(U_i) \right| \geq \varepsilon \right) \to 0 \tag{3.8}$$

Proof: see Appendix B.3.

From Theorem 3.1, the number of difference quotients k_1 and k_2 play a role in the bias variance tradeoff. Similar to the symmetric difference quotients for the first derivative, the higher k_1 and k_2 are, the higher the bias but the lower the variance and vice versa. The following corollary chooses the numbers of symmetric difference quotients for the second derivative considering the bias variance tradeoff.

Corollary 3.2 *Under the assumptions of Theorem 3.1, the optimal k_1 and k_2 that is chosen by minimizing the asymptotic upper bound of the conditional MISE is*

$$(k_1, k_2)_{opt} = \underset{k \in \mathbb{N}^+ \setminus \{0\}}{\arg\min} \left\{ \left(\frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \right)^2 + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2 \sum_{j=1}^{k_2}(2j+k_1)^2} \right\} \tag{3.9}$$

where $\mathcal{B}_2 \equiv \sup_{u \in [0,1]} \left| r^{(3)}(u) \right|$.

Proof: See Appendix B.4.

The quantity $\mathcal{B}_2$ can be estimated by a local polynomial regression of order $p = 4$ to obtain an estimate of the third derivative of r. The optimal value pair $(k_1, k_2)_{opt}$ can be obtained using a grid search or any optimization method. Points on the left boundary, $i < k_1 + k_2 + 1$, and points on the right boundary, $i > n - k_1 - k_2$, need to be adjusted and can be done in a similar analysis for the noisy first derivative estimator in Section 2.2.3.

3.2.1 Smoothing Second Order Weighted Difference Quotients

The second order derivative estimator Eq. (3.3) depends on the variance of the LRF estimator; forest based estimators are very noisy and therefore may affect the variance of the derivative estimator. Another issue of the weighted difference quotient estimator is

that it cannot be evaluated at any arbitrary test point, and can only be evaluated for points in the training sample. In order to smooth the second order weighted difference quotients, a local polynomial regression of the second order derivative estimates on U. First, rewriting Eq. (3.1) as

$$\begin{aligned}\widehat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{i,j,2}\frac{\left(\frac{r(U_{i+j+k_1})-r(U_{i+j})}{U_{i+j+k_1}-U_{i+j}} - \frac{r(U_{i-j-k_1})-r(U_{i-j})}{U_{i-j-k_1}-U_{i-j}}\right)}{U_{i+j+k_1}-U_{i+j}-U_{i-j-k_1}-U_{i-j}} \\ + 2\sum_{j=1}^{k_2} w_{i,j,2}\frac{\left(\frac{e_{i+j+k_1}-e_{i+j}}{U_{i+j+k_1}-U_{i+j}} - \frac{e_{i-j-k_1}-e_{i-j}}{U_{i-j-k_1}-U_{i-j}}\right)}{U_{i+j+k_1}-U_{i+j}-U_{i-j-k_1}-U_{i-j}},\end{aligned} \tag{3.10}$$

where the first term is denoted as the true second derivative, $r^{(2)}(U)$ and the last term is the new error term, $\acute{e}$,

$$\widehat{Y}_i^{(2)} = r^{(2)}(U) + \acute{e}. \tag{3.11}$$

As in the case for the first derivative, a bimodal kernel, K such that $K(0) = 0$, is used to counteract the effect of the correlated errors on bandwidth selection. Instead of using the second order weighted difference quotient in Eq. (3.1), we propose using Eq. (3.3) with weights Eq. (3.5) for the estimate of the second derivative,

$$\widehat{Y}_{i,LRF}^{(2)} = r^{(2)}(U) + \acute{e}. \tag{3.12}$$

Now, a local polynomial regression is estimated for the model Eq. (3.12), where $\widehat{Y}_{i,LRF}^{(1)}$ is replaced by $\widehat{Y}_{i,LRF}^{(2)}$ in Eq. (2.18) is minimized. However, by construction of the second derivative estimator, the errors will be correlated and assume that $\mathbb{E}[\tilde{e}|U] = 0$, $\mathrm{Cov}[\tilde{e}_i, \tilde{e}_j|U_i, U_j] = \sigma_{\tilde{e}}^2\rho_n(U_i - U_j)$ for $i \neq j$, and $\sigma_{\tilde{e}}^2 < \infty$. The correlation function, $\rho_n(\cdot)$ goes to zero as $n \to \infty$ and must satisfy $\rho_n(0) = 1$, $\rho_n(u) = \rho_n(-u)$, and $-1 \leq \rho_n(u) \leq 1$, assumed in [3] and [30]. The goal is then to enhance the noisy derivative estimator $\widehat{Y}_{LRF}^{(2)}$ by nonparametric smoothing. However, the correlation in the errors will affect the bandwidth selection for any

nonparametric smoothing and to counteract these effects, we will use the bimodal kernel in Eq. (2.17) to remove any effects of the correlation on the bandwidth selection without the need to estimate the correlation structure. To start, we fit a local polynomial regression of $\widehat{Y}^{(2)}_{LRF}$ on U. That is, the local polynomial regression estimator for the second derivative of degree p for a given test observation u_0 is provided by minimizing

$$\min_{\beta_L \in \mathbb{R}} \sum_{i=1}^{n} \left\{ \widehat{Y}^{(2)}_{i,LRF} - \sum_{L=0}^{p} \beta_L \left(U_i - u_0\right)^L \right\}^2 K\left(\frac{U_i - u_0}{h}\right) \tag{3.13}$$

where β_L are the solutions to the weighted least squares problem and $K(\cdot)$ is a kernel function. The $(L+2)$th order derivative $r^{(L+2)}(u_0)$ for $L = 0, 1, \ldots, p$ is estimated by $\widehat{r}^{(L+2)}(u_0) = L!\widehat{\beta}_L$. Following [30], the error term $\acute{e}$ satisfies $\mathbb{E}[\acute{e}|U] = 0$ and $\mathrm{Cov}[\acute{e}_i, \acute{e}_j|U_i, U_j] = \sigma^2_{\acute{e}}\rho'_n(U_i - U_j)$. Then, the solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{U}^\top \mathbf{W}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{W}\widehat{\mathbf{y}}^{(\mathbf{2})}, \tag{3.14}$$

where $\mathbf{U}$ is the centered regression matrix with the first column being ones, $\mathbf{W}$ is the diagonal matrix of kernel weights, and $\widehat{\mathbf{y}}^{(2)}$ is the vector of second order weighted difference quotients in Eq. (3.3). Therefore, the smoothed second order derivative estimator is given by

$$\widehat{r}^{(2)}(u_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_1^\top (\mathbf{U}^\top \mathbf{W}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{W}\widehat{\mathbf{y}}^{(\mathbf{2})} \tag{3.15}$$

With smoothing by local polynomial regression, the bandwidth h needs to be estimated. Following Chapter 2, we find $k = (k_1, k_2)$ and h as follows: k is found by optimizing AMISE in Eq. (3.9) and the bandwidth h is then estimated by using the bimodal kernel $\bar{K}$ in Eq. (2.17), denoted as $\widehat{h}_b$ by cross validation. Then, $\widehat{h}_b$ can be used as a pilot bandwidth which can be related to the bandwidth, $\widehat{h}$, of the usual unimodal kernel in

Eq. (2.21). Therefore, after fitting a local cubic regression of $\widehat{Y}^{(2)}_{LRF}$ on U with bimodal kernel, $\bar{K}(\cdot)$, and bandwidth, $\widehat{h}_b$, we refit a local cubic regression of $\widehat{Y}^{(2)}_{LRF}$ on U with unimodal kernel, $K(\cdot)$ and bandwidth, $\hat{h}$, defined in Eq. (2.22).

3.2.2 Generalizing to Arbitrary Distributions

Similar to the first derivative, to generalize to any unknown distribution for X, we again use PIT to transform the variables X to U. Since, $m(X) = r(F(X))$, the second derivative of m with respect to X is

$$\frac{d^2m(X)}{dX^2} = \frac{d}{dX}\left(\frac{dr(U)}{dU}\frac{dU}{dX}\right) = \frac{d}{dX}\Big(f(X)r^{(1)}(U)\Big) = f^{(1)}(X)r^{(1)}(U) + f(X)r^{(2)}(U), \tag{3.16}$$

where $f(X)$ is the density of X, $f^{(1)}(X)$ is the first derivative of the density of X, $r^{(1)}(X)$ is the first derivative of $r(\cdot)$ with respect to X and $r^2(X)$ is the second derivative of $r(\cdot)$ with respect to X. $f(X)$ and $f^{(1)}(X)$ can be estimated by kernel density and kernel derivative density estimators with plug in bandwidths and $r^{(1)}(\cdot)$ can be estimated by Eq. (2.20) and $r^{(2)}(\cdot)$ can be estimated by Eq. (3.15). To summarize, the full algorithm is described in Algorithm 3.1.

Note that this procedure gives the pointwise second derivatives of Y with respect to X, which may vary across the space of X. To summarize the overall derivative, we can take the sample average of the second derivative estimates to estimate the average second derivative, that is

$$\widehat{m}^{(2)}_{avg} = \frac{1}{n'}\sum_{i=1}^{n'} \widehat{m}^{(2)}(X_i). \tag{3.17}$$

Algorithm 3.1 Smoothed Nonparametric Second Derivative Estimation using Weighted Difference Quotients based on LRF

procedure DQ2SmoothLRF(training independent variable X, training dependent variable Y, number of symmetric difference quotients $k = (k_1, k_2)$, degree of local polynomial $p = 3$, grid of bandwidths to be considered h_{grid}, tunable hyperparameters of LRF model θ)

1: $U \leftarrow \widehat{F}(X)$

▷ kernel cumulative distribution function estimation

2: $\widehat{r}(U) \leftarrow$ regress Y on U by a LRF model with tunable hyperparameters θ

3: $\widehat{Y}^{(2)}_{LRF} \leftarrow$ difference quotient in Eq. (3.3) with k_1 and k_2

4: $\widehat{h}_b \leftarrow$ bandwidth selection from local polynomial regression of $\widehat{Y}^{(2)}_{LRF}$ on U

▷ with degree p, kernel $\bar{K}$ in Eq. (2.17), and searched over bandwidths h_{grid}

5: $\widehat{h} \leftarrow 1.01431\widehat{h}_b$ (for local cubic regression)

▷ bandwidth relationship between bimodal and unimodal gaussian kernels

6: $\widehat{r}^{(2)}(U) \leftarrow$ local polynomial regression

▷ with degree p, unimodal gaussian kernel K, and bandwidth $\widehat{h}$

7: $\widehat{r}^{(1)}(U) \leftarrow$ Algorithm 2.1 with k_1

8: $\widehat{m}^{(2)}(X) \leftarrow \widehat{f}^{(1)}(X)\widehat{r}^{(1)}(U) + \widehat{f}(X)\widehat{r}^{(2)}(U)$

▷ back transform into original space as in Eq. (3.16) after kernel density and derivative estimation of X

3.2.3 Extension to the Multivariate Case

In this section we try to extend the procedure to the multivariate case. Suppose we have the model

$$Y_i = m(X_{i,1}, \ldots, X_{i,d}) + e_i, \quad i = 1, \ldots, n, \tag{3.18}$$

where d is the number of independent variables. For now, suppose all of the regressors are standard uniformly distributed. Then, consider the regression of the form

$$Y_i = r(U_{i,1}, \ldots, U_{i,d}) + e_i. \tag{3.19}$$

Then, we follow the same steps as before. First, we estimate $r(\cdot)$ by LRF. Now, the second partial derivative with respect to the sth variable is given by the second order weighted difference quotient

$$\widehat{Y}^{(2)}_{i,s,LRF} = \sum_{j=1}^{k_{2,s}} w_{i,j,2} \frac{{}^{+}\widehat{Y}^{(1)}_{i+j,s,LRF} - {}^{-}\widehat{Y}^{(1)}_{i-j,s,LRF}}{C_{i,j,k_1}}, \tag{3.20}$$

where $C_{i,j,k_1,s} = (U_{i+j+k_{1,s},s} - U_{i+j,s} - U_{i-j-k_{1,s},s} - U_{i-j,s})/2$, ${}^{+}\widehat{Y}^{(1)}_{i+j,s,LRF} = \frac{\widehat{r}(U_{i+j+k_{1,s},s},\bar{U}) - \widehat{r}(U_{i+j,s},\bar{U})}{U_{i+j+k_{1,s},s} - U_{i+j,s}}$, and ${}^{-}\widehat{Y}^{(1)}_{i-j,s,LRF} = \frac{\widehat{r}(U_{i-j-k_{1,s},s},\bar{U}) - \widehat{r}(U_{i-j,s},\bar{U})}{U_{i-j-k_{1,s},s} - U_{i-j,s}}$ where $\bar{U}$ contains the other $d-1$ variables evaluated at their medians. Note that we order the data with respect to the sth variable and that variables can have different number of symmetric difference quotients denoted by $k_{2,s}$. We can then use this estimate of the second partial derivative as the dependent variable in a local polynomial regression on $U_1, \ldots, U_d$,

$$\widehat{Y}^{(2)}_{s,LRF} = r^{(2)}_s(U_1, \ldots, U_d) + \tilde{e}, \tag{3.21}$$

where the subscript s denotes the derivative with respect to the sth variable. Since the errors are correlated, we use the multivariate bimodal kernel in Eq. (2.32). Then, we can correct

the bandwidth using the relationship between bimodal and unimodal kernel functions as in [4] for each bandwidth to get an estimate of the second partial derivative, $\widehat{r}_s^{(2)}(U_s, \bar{U})$.

Now, suppose that X_m for $m = 1 \ldots, d$ are not all standard uniformly distributed. First, we can estimate the joint CDF function of all independent variables, $F_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$. To obtain the marginal CDFs of a specified variable, we take the limits in the arguments of the joint CDF of the other variables, $F_{X_s}(x_s) = F_{X_1,\ldots,X_d}(+\infty, \ldots, x_s, \ldots, +\infty)$. Then, we have for each regressor,

$$F_{X_m}(X_m) \sim \mathcal{U}(0,1), \quad m = 1, \ldots, d. \tag{3.22}$$

Therefore, the new data $(F_{X_1}(X_{1,1}), \ldots, F_{X_d}(X_{1,d}), Y_1), \ldots, (F_{X_1}(X_{n,1}), \ldots, F_{X_d}(X_{n,d}), Y_n)$, has the same distribution as $(U_{1,1}, \ldots, U_{1,d}, Y_1), \ldots, (U_{n,1}, \ldots, U_{n,d}, Y_n)$.

$$\begin{aligned} Y_i &= r(F_{X_1}(X_{i,1}), \ldots, F_{X_d}(X_{i,d})) + e_i \\ &= r(U_{i,1}, \ldots, U_{i,d}) + e_i, \end{aligned} \tag{3.23}$$

which is the case where we have all standard uniform variables as regressors. To get second derivatives in the original space, we can estimate the following

$$m_s^{(2)}(X_s, \bar{X}) = f_{X_s}^{(1)}(X_s) r^{(1)}(U_s, \bar{U}) + f_{X_s}(X_s) r^{(2)}(U_s, \bar{U}). \tag{3.24}$$

3.3 Local Random Forest

This section briefly discusses the Local (Linear) Random Forest (LRF) estimator. In what follows, assume the i.i.d data $(\mathbf{x}_i, y_i)$, for $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and assume the model in Eq. (2.1), either for the univariate case ($d = 1$) or the multivariate case ($d > 1$). LRF uses a local polynomial regression with $p = 1$ (local linear) with forest based weights

instead of kernel based weights [18]. The objective function includes a weighted quadratic loss with L_2 regularization, specified in Eq. (2.37). Note that δ_q, for $q = 0, 1, 2$ denotes the conditional mean function, its derivative evaluated at $\mathbf{x}_0$, and its second derivative evaluated at $\mathbf{x}_0$. Recall that the solution vector is $\widehat{\boldsymbol{\delta}} = (\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{J})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{y}$ and the resulting prediction function evaluated at a test point $\mathbf{x}_0$ is $\widehat{m}(\mathbf{x}_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\delta}}$. And the second derivative can be obtained from LRF by

$$\widehat{m}^{(2)}(\mathbf{x}_0) = \boldsymbol{\epsilon}_3^\top \widehat{\boldsymbol{\delta}}. \tag{3.25}$$

However, as stated in Section 3.1, the second derivative based on LRF will be noisy due to the nature of the nonsmooth estimated regression function, ultimately resulting in the second derivative having high variance. In the following section, we will show that there is a huge improvement when the second derivative based on LRF is compared to the proposed estimator, DQ2SmoothLRF.

To find hyperparameter, λ, a random forest model is trained first and the weights $a_i(\cdot)$ are obtained from the forest. Then, the regularization parameter λ can be found by cross validation. LRF can then be used in step 3 of Algorithm 3.1, with transformed regressor, U.

3.4 Simulations

This section shows derivative estimation based on the proposed estimator as well as other estimators. To compare results, we consider the following data generating process (DGP) from Section 2.5. Given the true function for m, the second derivative is

$$m^{(2)}(X) = 8\pi^2(\sin(2\pi X)^2 - \cos(2\pi X)^2). \tag{3.26}$$

Then, we follow the procedure outlined in Algorithm 3.1, where the numbers of symmetric difference quotients for the first and second derivative, k_1 and k_2, are estimated by Cor. 3.2. Similar to the first derivative, for the weights $w_{i,j,2}$ in Eq. (3.5), we assume that the variances are constant for the k_1 and k_2 points round i, so that the weights do not depend on the variances of the LRF estimator.

For all simulations, we estimate the density f and distribution F by the `R` package `ks` [14]. The parameters k_1 and k_2, the numbers of symmetric difference quotients, is estimated by Cor. 3.2. For the weights $w_{i,j}$ based on the variances of the LRF estimator in Eq. (3.5), we assume that the variances are constant for the points round i. We do this so that results under the proposed model can be easily compared to the benchmark model, the model proposed by [30], and since the simulated errors are homoskedastic, this assumption may be reasonable.

We show estimates of the second derivatives considering four different models, (1) DQ2Smooth, the benchmark model of the second derivative based on difference quotients proposed by [30], (2) DQ2SmoothLRF, the proposed estimator based on LRF estimates of Y, (3) LocCubic, a local cubic regression, a common regression technique to estimate derivatives in the nonparametric literature, and (4), LRF, the model proposed by [18], a benchmark model of the second derivative based on random forests. For the last estimator, LRF, the original paper by [18] focuses on estimating the conditional mean function, not its derivatives. The paper also considers only a local linear approach. For these simulations, we consider a local cubic with $\lambda = 0$ for the LRF estimator. The reason for zero ridge penalty, is that since the ridge parameter, λ, penalizes the curvature of the function, it will

force the derivative estimates toward zero, although the conditional mean function may not be flat. All local polynomial regressions are estimated using `locpol` package [6]. When estimating LRF based models, we use the `grf` package [42]. To assess the models, we use mean squared error (MSE) and mean absolute error (MAE), defined in Eq. (2.44) and . We evaluate all models at 500 evenly spaced points from 0.05 to 0.95, where $m = 500$.

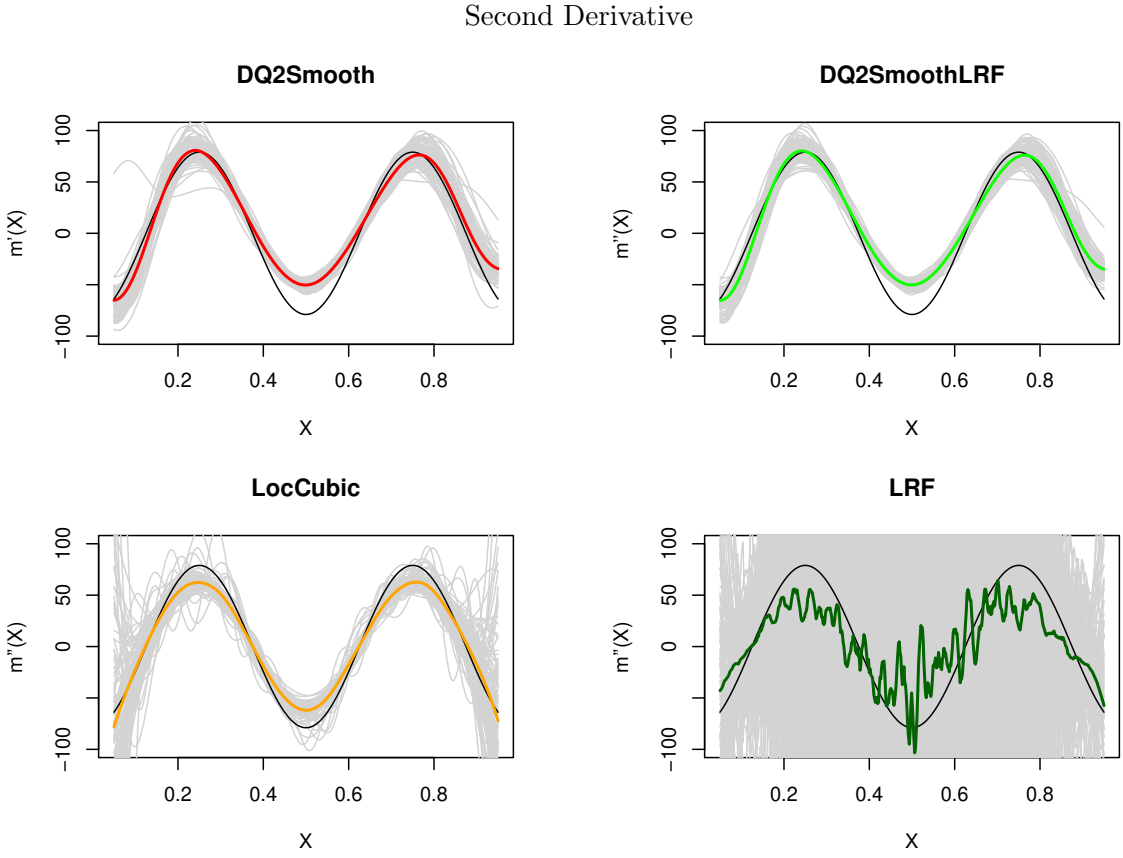

Figure 3.1: *Each plot shows the estimates for the second derivative for DQ2Smooth [30], DQ2SmoothLRF (the proposed estimator), LocCubic, and LRF estimators. Note that the second derivative estimator based on LRF is for $\lambda = 0$. The grey curves in each plot depict estimates of the second derivative for one simulation, where 100 simulations are plotted. The solid colored curves represent the mean predicted values across all simulations. The solid black curve represents the true second derivative. All estimators are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

Results for the second derivative are plotted in Figure 3.1. The colored curves are analagous to those from the first derivative plot in Figure 2.1. Here, it appears that that most estimators roughly get the overall shape of the second derivative. Similar to the first derivative, the first noticeable difference is the variance of each estimator, where both DQ2Smooth and DQ2SmoothLRF estimators have the smallest variance. Although LocCubic seems to have smaller bias, the variance is significantly larger than the DQ2Smooth procedures. In the last case, LRF has extremely large variance, and there is much improvement on estimating the second derivative based on DQ2SmoothLRF compared to LRF alone.

Results for the simulations are shown in Table 3.1, where the bias, variance, MSE, and MAE are reported for both the first and second derivative across the four models under consideration. First, the mean bias from DQ2Smooth and DQ2SmoothLRF estimators are roughly the same indicating that the bias of second derivative estimates for DQ2SmoothLRF is similar to that of the bias for DQ2Smooth. However, improvement on DQ2Smooth is shown through the variance of the DQ2SmoothLRF estimator, with a 32% reduction in the variance relative to the variance of DQ2Smooth. Although the LocCubic model has lower absolute mean bias compared to all models, the variance is over double of those of the DQ2Smooth models. The LRF estimator performs the worst and we can see a significant improvement in estimates for the second derivative estimator when using DQ2SmoothLRF. Lastly, DQ2SmoothLRF performs the best in terms of both MSE and MAE compared to all models for both the first derivative and second derivative estimations. Overall, in these simulations, we have shown that DQ2SmoothLRF has the smallest variance, MSE,

and MAE, implying that the proposed procedure is the preferred method of estimating the second derivative compared to others.

Model Assessment for Second Derivative

	Bias	Variance	MSE	MAE
DQ2Smooth	5.7984	86.0250	259.8858	12.6552
DQ2SmoothLRF	5.4323	58.7255	231.5355	12.1516
LocCubic	-1.9564	631.8972	747.2560	15.8320
LRF	-4.1653	18,042.3700	18,892.7700	96.5819

Table 3.1: *The top and bottom panel show the bias, variance, MSE, and MAE for the second derivative respectively, comparing the four models, DQ2Smooth [30], DQ2SmoothLRF (the proposed estimator), LocCubic, and LRF. All estimates are averaged across all simulations. Note that the results based on LRF is for* $\lambda = 0$*. All models are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

3.5 Empirical Application: Convex Technology

A conventional assumption of a production possibility set or technology is convexity [29]. One way to check the convexity assumption is to evaluate the curvature of the production function, known as the criterion of quasi-concavity. [39] shows that the the law of diminishing marginal productivity in at least one input and quasi-concavity are violated, indicating nonconvexities in agricultural technology. [29] tests the convexity assumption and show that cost functions determined by convex technology are heavily downward bi-

ased compared to those determined by nonconvex technology. In this paper, we will check the curvature and quasi-concavity criterion of the production function by estimating its second derivative by the proposed method.

We will use Chilean hydro-electric power generation plants [1]. To avoid any technical change over time, we single out the year 1997 for 16 power plants with monthly data, providing 188 observations.[2] The data contain one output, electricity generated (q), prices, and quantities of three inputs: capital (k), water (w), and labor (l).[3]

First, the production function can be modeled as an unknown function of the three inputs.

$$q = m(k, w, l) + e. \tag{3.27}$$

We wish to evaluate the curvature of the production function to test quasi-concavity by estimating its second derivative in each of the three arguments. First, we transform the inputs as in Section 2.3 and Section 3.2.3 to rewrite the model as

$$q = r(U_k, U_w, U_l) + e_i, \tag{3.28}$$

where U_k, U_w, and U_l are the uniformly transformed variables of capital, water, and labor, respectively. Figure 3.2 depicts the first (left three plots) and second (right three plots) derivatives of the production function in Eq. (3.27). The first derivatives with respect to capital $\widehat{m}^{(1)}(k, \bar{w}, \bar{l})$, water $\widehat{m}^{(1)}(\bar{k}, w, \bar{l})$, and labor $\widehat{m}^{(1)}(\bar{k}, \bar{w}, l)$ are evaluated at 200 evenly spaced points across the sample space of each input while holding the other inputs fixed at their medians. The second derivatives are evaluated analogously.

[2]The data can be found from the Journal of Applied Econometrics Data Archive. Note that there are four missing observations for the 16 power plants during the year 1997.

[3]Further details about the data can be found in [1].

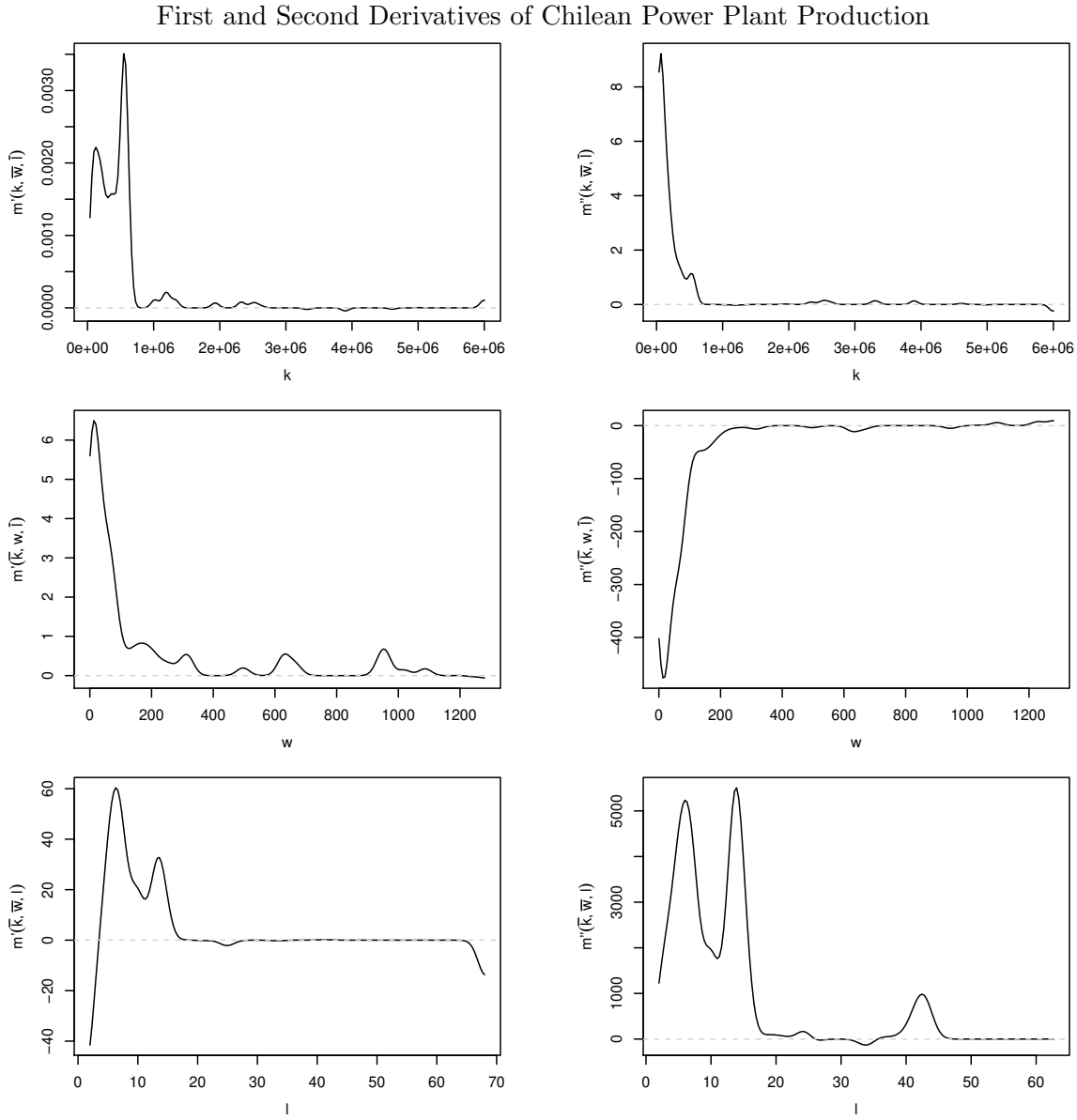

Figure 3.2: *The three figures on the left and right are the first and second derivatives, respectively, estimated by DQSmoothLRF and DQ2SmoothLRF. All estimates are evaluated across 200 evenly spaced points for each of the three inputs, while holding the other two inputs fixed at their medians.*

From Figure 3.2, for capital, the first derivative is positive when capital is small, indicating that electricity output is increasing in capital. The second derivative is positive

around this support, indicating that the slope of the production function with respect to capital is increasing, which implies that this part of the production function is convex. As a result the level set will be nonconvex, violating the quasi-concavity criterion and convex technology assumption. The bottom plots of Figure 3.2, referring to the input labor, seem to also violate these assumptions, where the production function is convex in parts of the support for labor. The middle plots show that production is increasing in the input for water and the second derivative is negative in the start of the support for water. Therefore, production appears to be quasi-concave in water. Overall, electricity production is convex in at least two of its inputs and does not produce convex technology, which would further confirm the nonconvex results obtained in [29].

Average Derivatives of Chilean Power Plant Production

	First Derivative		Second Derivative	
	OLS	DQSmoothLRF	OLS	DQ2SmoothLRF
Capital	0.00001	0.0002	0	0.3264
Water	0.2176	0.5296	0	-30.0765
Labor	0.2815	4.5731	0	64.1013

Table 3.2: *Average partial first and second derivatives with respect to each of the inputs, capital, water, and labor, are reported. The partial derivatives are evaluated at 200 evenly spaced points across the support of each regressor, where each derivative is held constant for other factors. The reported results are the average of the evaluated 200 derivative points. The median is reported as the estimate for the second derivative for DQ2SmoothLRF due to some outliers in the tail of the second derivative.*

Table 3.2 shows the estimated average first and second derivatives of the production function with respect to the input variables, capital, water, and labor. Average derivatives for OLS are also reported as reference, where a naive linear regression function is estimated. The first derivatives, partial marginal effects, will be constant, and as a result, the second derivative will be zero. All inputs have a positive average partial effect and the production function is increasing in the three inputs. However, when evaluating the first derivative, the average partial effect is underestimated compared to the average effect estimated by DQ2SmoothLRF for all of the inputs. The second derivative with respect to water is negative on average, implying that the estimated production function is concave in the input water. However, the other two inputs have a positive average second derivative, implying that the estimated production function is convex in the inputs capital and labor. Furthermore, the results from Table 3.2 regarding the positive second derivatives with respect to capital and labor further justify that the electricity production function breaks the convex technology assumption.

3.6 Conclusion

Overall, second derivatives help economists to check convexity assumptions by evaluating the curvature of a production function (second derivative). In this paper, we propose a method, DQ2SmoothLRF, that smooths random forest based difference quotients to estimate first and second derivatives. We improve on the original method in [30] by using estimated values of the dependent variable from LRF in forming difference quotients, instead of using the dependent variable itself, in hopes of reducing variance and by including

multiple variables in the model, instead of the simple univariate case. Improvement is also made in comparision to second derivatives estimated by LRF in [18], where derivatives of the regression equation were not even focused on and in providing better interpretation for forest based models by evaluating second derivatives derived from random forests. We have shown in simulation that the proposed estimator outperforms the benchmark ones as well as a popular method of estimating derivatives in economics, local polynomial regression; a reduction in variance, MSE, and MAE are all evident when using the proposed estimator. Lastly, we provide an empirical example using Chilean hydro-electric power generation plants data to assess the curvature of the production function in order to check the convex technology assumption assumed in the literature, in which we found that this assumption is broken for this dataset.

Chapter 4

Generalized Kernel Regularized Least Squares Estimator with Parametric Error Covariance

4.1 Introduction

Nonparametric regression function estimators are useful econometric tools. Common methods to estimate a regression function are kernel based methods, such as Kernel Regularized Least Squares (KRLS), Support Vector Machines (SVM), Local Polynomial Regression, etc. However, in order to avoid overfitting the data, some type of regularization, lasso or ridge, is generally used. In this chapter, we will focus on KRLS; this method is also known as Kernel Ridge Regression (KRR) and is the kernelized version of the simple ridge regression to allow for nonlinearities in the model.

In this chapter, we establish fitting a nonparametric regression function via KRLS under a general parametric error covariance. Some theoretical results, including pointwise marginal effects, unbiasedness, consistency and asymptotic normality, on KRLS are found in [20]. However, [20] only considers errors to be homoskedastic and that the estimator is ubiased for estimating the postpenalization function, not for the true underlying function. Confidence interval estimates for Least Squares Support Vector Machine (LSSVM) are discussed in [12], allowing for heteroskedastic errors. Although not directly stated, the LSSVM model in [12] is equivalent to KRR/KRLS when an intercept term is included in the model. Following [20], we will use KRLS without an intercept. Although [12] allows for heteroskedastic errors, none of the papers mentioned thus far discuss incorporating the error covariance in estimating the regression function itself, making these type of models inefficient. In this chapter, we focus on making KRLS more efficient by incorporating a parametric error covariance, allowing for both heteroskedasticity and autocorrelation, in estimating the regression function. We use a two step procedure where in the first step, we estimate the parametric error covariance from the residuals obtained by naive KRLS and in the second step, we estimate a KRLS model based on transformed variables based on the error covariance. The proposed method is similar to that of [41], but instead of KRLS, local linear regression is used. We also provide estimating derivatives based on the two step procedure, allowing us to determine the partial effects of the regressors on the dependent variable.

The structure of this chapter is as follows: Section 4.2 discusses the model framework and the GKRLS estimator, Section 4.3, Section 4.4, and Section 4.5 show the fi-

nite sample properties, asymptotic properties, and partial effects and derivatives of the GKRLS estimator, respectively, Section 4.6 runs through a simulation example, Section 4.7 illustrates two empirical examples with heteroskedastic and autocorrelated errors, and Section 4.8 concludes the chapter.

4.2 Generalized KRLS Estimator

Consider the nonparametric regression model:

$$Y_i = m(X_i) + U_i, \quad i = 1, \ldots, n, \tag{4.1}$$

where X_i is a $q \times 1$ vector of exogenous regressors, and U_i is the error term such that $\mathbb{E}[U_i] = 0$ and

$$\mathbb{E}[U_i U_j] = \omega_{ij}(\theta_0) \text{ for some } \theta_0 \in \mathbb{R}^p, i, j = 1, \ldots, n. \tag{4.2}$$

In this framework, we allow the error covariance to be parametric, where the errors can be autocorrelated or non-identically distributed across observations.

4.2.1 Naive KRLS Estimator

For KRLS, the function $m(\cdot)$ can be approximated by some function in the space of functions constituted by

$$m(\mathbf{x}_0) = \sum_{i=1}^{n} c_i K_\sigma(\mathbf{x}_i, \mathbf{x}_0), \tag{4.3}$$

for some test observation $\mathbf{x}_0$ and where $c_i,\ i = 1, \ldots, n$ are the parameters of interest, which can be thought of as the weights of the kernel functions $K_\sigma(\cdot)$. The subscript of the kernel function, $K_\sigma(\cdot)$, indicates that the kernel depends on the bandwidth parameter, σ.

We will use the Radial Basis Function (RBF) kernel,

$$K_\sigma(\mathbf{x}_i, \mathbf{x}_0) = \mathrm{e}^{-\frac{1}{\sigma^2}||\mathbf{x}_i - \mathbf{x}_0||^2}. \tag{4.4}$$

Notice that the RBF kernel is very similar to the Gaussian kernel, in that it does not have the normalizing term out in front and that σ is proportional to the bandwidth h in the Gaussian kernel often used in nonparametric local polynomial regression. This functional form is justified by a regularized least squares problem with a feature mapping function that maps $\mathbf{x}$ into a higher dimension [20], where this derivation of KRLS is also known as Kernel Ridge Regression (KRR). Overall, KRLS uses a quadratic loss with a weighted L_2-regularization. Then, in matrix notation, the minimization problem is

$$\arg\min_{\mathbf{c}} (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c})^\top (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c}) + \lambda \mathbf{c}^\top \mathbf{K}_\sigma \mathbf{c}, \tag{4.5}$$

where $\mathbf{y}$ is the vector of training data corresponding to the dependent variable and $\mathbf{K}_\sigma$ is the kernel matrix, with $K_{\sigma,i,j} = K_\sigma(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j = 1, \ldots, n$ and $\mathbf{c}$ is the vector of coefficients that is optimized over. The solution to this minimization problem is

$$\widehat{\mathbf{c}}_1 = (\mathbf{K}_{\sigma_1} + \lambda_1 \mathbf{I})^{-1} \mathbf{y}. \tag{4.6}$$

The kernel function and its parameters are user specified but can be found via cross validation along with the regularization parameter λ. The subscript of one denotes the naive KRLS estimator, or the first stage estimation. Finally, predictions for KRLS can be made by

$$\widehat{m}_1(\mathbf{x}_0) = \sum_{i=1}^{n} \widehat{c}_{1,i} K_{\sigma_1}(\mathbf{x}_i, \mathbf{x}_0), \tag{4.7}$$

4.2.2 An Efficient KRLS Estimator

The naive KRLS estimator, $\widehat{m}_1(\cdot)$ does not take into consideration any information in the error covariance structure and therefore is inefficient. As a result, consider the $n \times n$ error covariance matrix, $\Omega(\theta)$, where $\omega_{ij}(\theta)$ denotes the (i,j)th element. Assume that $\Omega(\theta) = P(\theta)P(\theta)'$ for some square matrix $P(\theta)$ and let $p_{ij}(\theta)$ and $v_{ij}(\theta)$ denote the (i,j)th element of $P(\theta)$ and $P(\theta)^{-1}$. Let $\mathbf{m} \equiv (m(X_1), \ldots, m(X_n))'$ and $\mathbf{U} \equiv (U_1, \ldots, U_n)'$. Now, premultiply the model in Eq. (4.1) by P^{-1}, where $P^{-1} = P^{-1}(\theta)$ and we condense the notation and the dependence on θ is implied.

$$P^{-1}\mathbf{y} = P^{-1}\mathbf{m} + P^{-1}\mathbf{U}. \tag{4.8}$$

The transformed error term, $P^{-1}U$ has mean 0 and covariance matrix as the identity matrix. Therefore, we consider a regression of $P^{-1}\mathbf{y}$ on $P^{-1}\mathbf{m}$. This simply re-scales the variables by the inverse of their square root of their variances. Since $\mathbf{m} = \mathbf{K}_\sigma \mathbf{c}$, the quadratic loss function with L_2 regularization under the transformed variables is

$$\arg\min_{\mathbf{c}} (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c})^\top \Omega^{-1} (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c}) + \lambda \mathbf{c}^\top \mathbf{K}_\sigma \mathbf{c}. \tag{4.9}$$

The solution for vector is

$$\hat{\mathbf{c}}_2 = (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \mathbf{y} \tag{4.10}$$

Note that the solution obtained depends on the bandwidth parameter σ_2 and ridge parameter λ_2, which can be different than the hyperparameters used in the Naive KRLS estimator. In practice, cross validation can be used for obtaining estimates for both hyperparametrs. Here, it is assumed that Ω is known if θ is known. However, if θ is unknown, it can be estimated consistently and Ω can be replaced by $\Omega = \Omega(\hat{\theta})$.

Furthermore, predictions for the generalized KRLS estimator can be made by

$$\widehat{m}_2(\mathbf{x}_0) = \sum_{i=1}^{n} \widehat{c}_{2,i} K_{\sigma_2}(\mathbf{x}_i, \mathbf{x}_0) \tag{4.11}$$

The two step procedure is outlined below

1. Estimate Eq. (4.1) by Naive KRLS from Eq. (4.7) with bandwidth parameter, σ_1 and ridge parameter, λ_1. Obtain the residuals which can then be used to get a consistent estimate for Ω.

2. Estimate Eq. (4.8) by KRLS under the transformed variables as in Eq. (4.9) and Eq. (4.11). Denote these estimates as GKRLS.

4.3 Finite Sample Properties

In this section, finite sample properties of both KRLS and GKRLS estimators, including the estimation procedures of bias and variance, are discussed in detail.

4.3.1 Estimation of Bias and Variance

In this subsection, we estimate the bias and variance of the two step estimator. Following, [12], notice that the GKRLS estimator is a linear smoother.

Defintion 4.1 *An estimator $\widehat{m}$ of m is a linear smoother if, for each $\mathbf{x}_0 \in \mathbb{R}^q$, there exists a vector $L(\mathbf{x}_0) = (l_1(\mathbf{x}_0), \ldots, l_n(\mathbf{x}_0))^\top \in \mathbb{R}^n$ such that*

$$\widehat{m}(\mathbf{x}_0) = \sum_{i=1}^{n} l_i(\mathbf{x}_0) Y_i, \tag{4.12}$$

where $\widehat{m}(\cdot) : \mathbb{R}^q \to \mathbb{R}$.

For in sample data, Eq. (4.12) can be written in matrix form as $\widehat{\mathbf{m}} = \mathbf{L}\mathbf{y}$, where $\widehat{\mathbf{m}} = (\widehat{m}(X_1), \ldots, \widehat{m}(X_n))^\top \in \mathbb{R}^n$ and $\mathbf{L} = (l(X_1)^\top, \ldots, l(X_n)^\top)^\top \in \mathbb{R}^{n \times n}$, where $\mathbf{L}_{ij} = l_j(X_i)$. The ith row of $\mathbf{L}$ show the weights given to each Y_i in estimating $\widehat{m}(X_i)$. For the rest of the chapter, we will denote $\widehat{m}_2(\cdot)$ as the prediction made by GKRLS for a single observation and $\widehat{\mathbf{m}}_2$ as the $n \times 1$ vector of predictions made for the training data.

To obtain the bias and variance of the GKRLS estimator, we assume the following:

Assumption 4.1 *The regression function $m(\cdot)$ to be estimated falls in the space of functions represented by $m(\mathbf{x}_0) = \sum_{i=1}^n c_i K_\sigma(\mathbf{x}_i, \mathbf{x}_0)$ and assume the model in Eq.* (4.1)*.*

Assumption 4.2 $\mathbb{E}[U_i] = 0$ *and* $\mathbb{E}[U_i U_j] = \omega_{ij}(\theta)$ *for some* $\theta \in \mathbb{R}^p, i, j = 1, \ldots, n$

Using Definition 4.1, Assumption 4.1, and Assumption 4.2, the conditional mean and variance can be obtained by the following theorem.

Theorem 4.1 *The GKRLS estimator in Eq.* (4.11) *is*

$$\begin{aligned} \widehat{m}_2(\mathbf{x}_0) &= \sum_{i=1}^n l_i(\mathbf{x}_0) Y_i \\ &= L(\mathbf{x}_0)^\top \mathbf{y}, \end{aligned} \tag{4.13}$$

and $L(\mathbf{x}_0) = (l_1(\mathbf{x}_0), \ldots, l_n(\mathbf{x}_0))^\top$ is the smoother vector,

$$L(\mathbf{x}_0) = \left[K^{*\top}_{\sigma_2, \mathbf{x}_0} (\Omega^{-1} \mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \right]^\top, \tag{4.14}$$

with $K^*_{\sigma_2, \mathbf{x}_0} = (K_{\sigma_2}(\mathbf{x}_1, \mathbf{x}_0), \ldots, K_{\sigma_2}(\mathbf{x}_n, \mathbf{x}_0))^\top$ the kernel vector evaluated at point $\mathbf{x}_0$.

Then, the estimator, under model Eq. (4.1)*, has conditional mean*

$$\mathbb{E}[\widehat{m}_2(\mathbf{x}_0) | X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \mathbf{m} \tag{4.15}$$

and conditional variance

$$\text{Var}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \Omega L(\mathbf{x}_0). \tag{4.16}$$

Proof: see Appendix C.1.

From Theorem 4.1, the conditional bias can be written as

$$\begin{aligned} \text{Bias}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] &= \mathbb{E}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}] - m(\mathbf{x}_0) \\ &= L(\mathbf{x}_0)^\top \mathbf{m} - m(\mathbf{x}_0) \end{aligned} \tag{4.17}$$

Following [12], we will estimate the conditional bias and variance by the following:

Theorem 4.2 *Let $L(\mathbf{x}_0)$ be the smoother vector evaluated at $\mathbf{x}_0$ and let $\widehat{\mathbf{m}}_2 = (\widehat{m}_2(\mathbf{x}_1), \ldots, \widehat{m}_2(\mathbf{x}_n))^\top$ be the in sample GKRLS predictions. For a consistent estimator of the covariance matrix such that $\widehat{\Omega} \to \Omega$, the estimated conditional bias and variance for GKRLS are obtained by*

$$\widehat{\text{Bias}}[\widehat{m}_2(\mathbf{x}_2)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \widehat{\mathbf{m}}_2 - \widehat{m}_2(\mathbf{x}_0) \tag{4.18}$$

and

$$\widehat{\text{Var}}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \widehat{\Omega} L(\mathbf{x}_0). \tag{4.19}$$

Proof: See Appendix C.2.

4.3.2 Bias and Variance of KRLS

First, note that the KRLS estimator is also a linear smoother, so the bias and the variance take the same form as in Eq. (4.18) and Eq. (4.19), except that the linear smoother

vector $L(\mathbf{x}_0)$ will be different. Let

$$L_1(\mathbf{x}_0) = \left[K^{*\top}_{\sigma_1,\mathbf{x}_0}(\mathbf{K}_{\sigma_1} + \lambda_1 \mathbf{I})^{-1}\right]^\top \tag{4.20}$$

be the smoother vector for KRLS. Then, Eq. (4.7) can be rewritten as

$$\widehat{m}_1(\mathbf{x}_0) = L_1(\mathbf{x}_0)^\top \mathbf{y}. \tag{4.21}$$

Using Theorem 4.1 and Theorem 4.2 and applying them to the KRLS estimator, the conditional bias and variance of KRLS are

$$\widehat{\text{Bias}}[\widehat{m}_1(\mathbf{x}_0)|X = \mathbf{x}_0] = L_1(\mathbf{x}_0)^\top \widehat{\mathbf{m}}_1 - \widehat{m}_1(\mathbf{x}_0) \tag{4.22}$$

$$\widehat{\text{Var}}[\widehat{m}_1(\mathbf{x}_0)|X = \mathbf{x}_0] = L_1(\mathbf{x}_0)^\top \widehat{\Omega} L_1(\mathbf{x}_0), \tag{4.23}$$

where $\widehat{\mathbf{m}}_1$ is the $n \times 1$ vector of fitted values for KRLS. Note that the estimate of the covariance matrix, Ω, will be the same for both KRLS and GKRLS.

4.4 Asymptotic Properties

The asymptotic properties of GKRLS, including consistency, asymptotic normality, and bias corrected confidence intervals are covered in this section. To obtain consistency of the GKRLS estimator, we also assume:

Assumption 4.3 *Let $\lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$ and as $n \to \infty$, for singular values of $\mathbf{L}P$ given by d_i, $\sum_{i=1}^n d_i^2$ grows slower than n once $n > M$ for some $M < \infty$.*

Theorem 4.3 *Under Assumptions 1-3, and let the bias corrected fitted values be denoted by*

$$\widehat{\mathbf{m}}_{2,c} = \widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}], \tag{4.24}$$

then

$$\lim_{n\to\infty} \text{Var}[\widehat{\mathbf{m}}_{2,\text{c}}|\mathbf{X}] = 0 \tag{4.25}$$

and the bias corrected GKRLS estimator is consistent with $\operatorname{plim}_{n\to\infty} \widehat{m}_{c,n}(\mathbf{x}_i) = m(\mathbf{x}_i)$ *for all* i*.*

Proof: See Appendix C.3.

The estimated conditional bias from Eq. (4.18) and conditional variance from Eq. (4.19) can be used to construce pointwise confidence intervals. Asymptotic normality of the proposed estimator is given via the central limit theorem.

Theorem 4.4 *Under Assumptions Assumptions 4.1 to 4.3,* $\widehat{\mathbf{m}}_2$ *is asymptotically normal by the central limit theorem:*

$$\widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] - \mathbf{m} \xrightarrow{d} N(\mathbf{0}, \text{Var}[\widehat{\mathbf{m}}_2|\mathbf{X}]). \tag{4.26}$$

Proof: See Appendix C.4.

Since GKRLS is a biased estimator for m, we need to adjust the pointwise confidence intervals to allow for bias. Since the exact conditional bias and variance are unknown, we can use Eqs. (4.18) and (4.19) as estimates and can conduct approximate bias corrected $100(1-\alpha)\%$ pointwise confidence intervals from Theorem 4.4 as

$$\widehat{m}_2(\mathbf{x}_i) - \widehat{\text{Bias}}[\widehat{m}_2(\mathbf{x}_i)|X = \mathbf{x}_i] \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}[\widehat{m}_2(\mathbf{x}_i)|X = \mathbf{x}_i]} \tag{4.27}$$

for all i. To test the significance of the estimated regression function at an observation point, we can use the bias corrected confidence interval to see if 0 is in the interval.

4.5 Partial Effects and Derivatives

We also derive an estimator for pointwise partial derivatives with respect to a certain variable $\mathbf{x}^{(r)}$. The partial derivative of the GKRLS estimator, $\widehat{m}_2(\mathbf{x}_0)$ with respect to the rth variable is

$$\begin{aligned}\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) &= \sum_{i=1}^{n} \frac{\partial K_{\sigma_2}(\mathbf{x}_i, \mathbf{x}_0)}{\partial \mathbf{x}_0^{(r)}} \widehat{c}_{2,i} \\ &= \frac{2}{\sigma_2^2} \sum_{i=1}^{n} \mathrm{e}^{-\frac{1}{\sigma_2^2}||\mathbf{x}_i - \mathbf{x}_0||^2} (\mathbf{x}_i^{(r)} - \mathbf{x}_0^{(r)}) \widehat{c}_{2,i},\end{aligned} \tag{4.28}$$

using the RBF kernel in Eq. (4.4) and where $\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) \equiv \frac{\partial \widehat{m}_2(\mathbf{x}_0)}{\partial \mathbf{x}^{(r)}}$. To find the conditional bias and variance of the derivative estimator, we use the following:

Theorem 4.5 *The GKRLS derivative estimator in Eq.* (4.28) *with the RBF kernel in Eq.* (4.4) *can be rewritten as*

$$\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) = S_r(\mathbf{x}_0)^\top \mathbf{y}, \tag{4.29}$$

where $\Delta_r \equiv \frac{2}{\sigma_2^2} \operatorname{diag}(\mathbf{x}_1^{(r)} - \mathbf{x}_0^{(r)}, \ldots, \mathbf{x}_n^{(r)} - \mathbf{x}_0^{(r)})$ *is a* $n \times n$ *diagonal matrix, and*

$$S_r(\mathbf{x}_0) = \left[K_{\sigma_2, \mathbf{x}_0}^{*\top} \Delta_r (\Omega^{-1} \mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \right]^\top \tag{4.30}$$

is the smoother vector for the first partial derivative with respect to the r*th variable.*

$$\mathbb{E}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) | X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \mathbf{m} \tag{4.31}$$

and conditional variance

$$\operatorname{Var}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) | X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \Omega S_r(\mathbf{x}_0). \tag{4.32}$$

Proof: see Appendix C.5.

Using Theorem 4.5, the conditional bias and variance can be estimated as follows

Theorem 4.6 *Let $S_r(\mathbf{x}_0)$ be the smoother vector for the partial derivative evaluated at $\mathbf{x}_0$ and let $\widehat{\mathbf{m}}_2 = (\widehat{m}_2(\mathbf{x}_1), \ldots, \widehat{m}_2(\mathbf{x}_n))^\top$ be the in sample GKRLS predictions. For a consistent estimator of the covariance matrix such that $\widehat{\Omega} \to \Omega$, the estimated conditional bias and variance for GKRLS derivative estimator in Eq.* (4.28) *are obtained by*

$$\widehat{\text{Bias}}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\mathbf{m}} - \widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) \tag{4.33}$$

and

$$\widehat{\text{Var}}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\Omega} S_r(\mathbf{x}_0). \tag{4.34}$$

Proof: See Appendix C.6.

The average partial derivative with respect to the rth variable is

$$\widehat{m}_{avg,r}^{(1)} = \frac{1}{n'} \sum_{j=1}^{n'} \widehat{m}_{2,r}^{(1)}(\mathbf{x}_{0,j}) \tag{4.35}$$

The bias and variance of the average partial derivative estimator is given by

$$\text{Bias}[\widehat{m}_{avg,r}^{(1)}|X] = \frac{1}{n'} \boldsymbol{\iota}_{n'}^\top \mathbf{S}_{0,r} \mathbf{m} - \frac{1}{n'} \boldsymbol{\iota}_{n'}^\top \mathbf{m}_{0,r}^{(1)} \tag{4.36}$$

and

$$\text{Var}[\widehat{m}_{avg,r}^{(1)}|X] = \frac{1}{n'^2} \boldsymbol{\iota}_{n'}^\top \mathbf{S}_{0,r} \Omega \mathbf{S}_{0,r}^\top \boldsymbol{\iota}_{n'}, \tag{4.37}$$

where n' is the number of observations in the testing set, $\boldsymbol{\iota}_{n'}$ is a $n' \times 1$ vector of ones, $\mathbf{S}_{0,r}$ is the $n' \times n$ smoother matrix with the jth row as $S_r(\mathbf{x}_{0,j}), j = 1, \ldots, n'$, and $\mathbf{m}_{0,r}^{(1)}$ is the $n' \times 1$ vector of derivatives evaluated at each $\mathbf{x}_{0,j}, j = 1, \ldots, n'$.

4.6 Simulations

We conduct simulations that show the performance with respect to gaining efficiency of the proposed generalized KRLS estimator. Consider the data generating process from Eq. (4.1):

$$Y_i = m(X_i) + U_i, \quad i = 1, \ldots, n. \tag{1}$$

We consider the sample size of $n = 200$ and univariate X that is generated from $Unif[-5, 5]$. The specification for m is:

$$m(X) = \sin(X) \tag{4.38}$$

and the derivative is given by

$$m^{(1)}(X) = \cos(X) \tag{4.39}$$

For the error terms, we consider two cases. Following [41], U_i is generated by an AR(2) process, where $U_i = 0.5U_{i-1} - 0.4U_{i-2} + \varepsilon_i$ and ε_i are iid $N(0, 1)$. Also, following [33] and [41], in the case of the AR(2) errors, the first two diagonal elements in the square root matrix (P) of the covariance matrix (Ω) of $\mathbf{U}$ are unique from the others. In the second case, U_i are heteroskedastic but independent of each other, where $U_i = \sqrt{0.05X_i^2 + 0.01}\varepsilon_i$ with $\varepsilon_i \sim N(0, 1), i = 1, \ldots, n$.

In addition to the proposed estimator, we compare two other models: the naive KRLS estimator (KRLS) and the LSSVM proposed by [12]. The naive estimator is used as a comparison to show the magnitude of the efficiency loss from ignoring the information in the error covariance matrix. [12] only considers heteroskedasticity in the LSSVM model, not allowing for autocorrelation in the errors. In addition, LSSVM does not utilize the

covariance matrix in estimating the regression function. For all models, we implement leave one out cross validation to select the hyperparameters. The variance function under the heteroskedastic case is estimated by nonlinear least squares by obtaining the estimated coefficients (a, b, c) in $a + \log(bX^2 + c)$. Taking the exponential would give the predicted variance estimates. Under the case of AR(2) errors, the covariance function is estimated from an AR(2) model. We run 200 simulations for each of the two cases and results are reported below in Figure 4.1, Figure 4.2, and in Table 4.1.[1] To evaluate the models, mean squared error is used as the main criterion, where we also investigate the bias and variance of the estimators. To compare results, all models are evaluated from 500 evenly spaced points from -5 to 5.

Figure 4.1 shows simulation results under Eq. (4.38). All simulation estimates for KRLS and GKRLS are plotted in grey and the averages across all simulations are plotted as green curves. For both heteroskedastic and AR(2) errors, the variability, depicted as how far or spread out the grey curves are from their average in green, is reduced as we move from the top two plots, where KRLS regressions are estimated, to the middle two plots, where GKRLS regressions are estimated. However, under the case of heteroskedastic errors, the GKRLS estimates appear to be slightly more biased relative to the KRLS model. This may be a result of the bias and variance tradeoff where an addition of small bias for a larger reduction in the variance can lead to an overall better fit and estimator, in terms of mean squared error. On the other hand, under the case of AR(2) errors, estimates based on GKRLS seem to exhibit the same finite sample bias as KRLS, and there is an obvious

[1]The following R packages were used for conducting simulations: [2], [27], and [34].

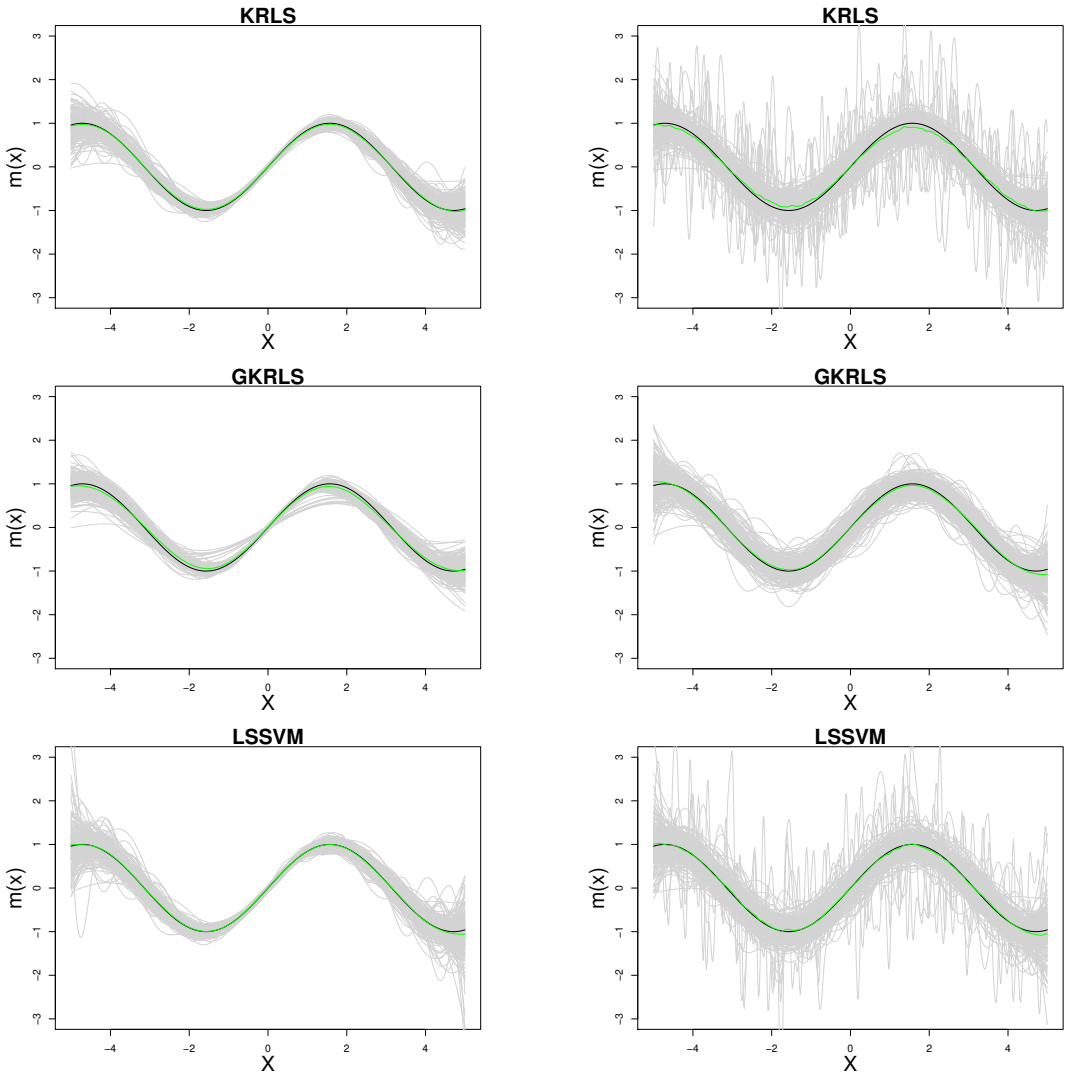

Figure 4.1: *The left (right) plots show the estimated predictions under heteroskedastic errors (AR(2) errors). The top, middle, and bottom plots refer to the KRLS, GKRLS, and LSSVM estimators. The grey curves show the predicted values from all simulations evaluated at the 500 evaluation data points spanning from -5 to 5. The green curves denote the average predictions at each evaluation point across all simulations, and the black curve represents the true regression function in Eq.* (4.38).

reduction in the variability of the proposed estimator relative to KRLS. LSSVM estimates are also plotted as a comparison and they show similar results to KRLS.

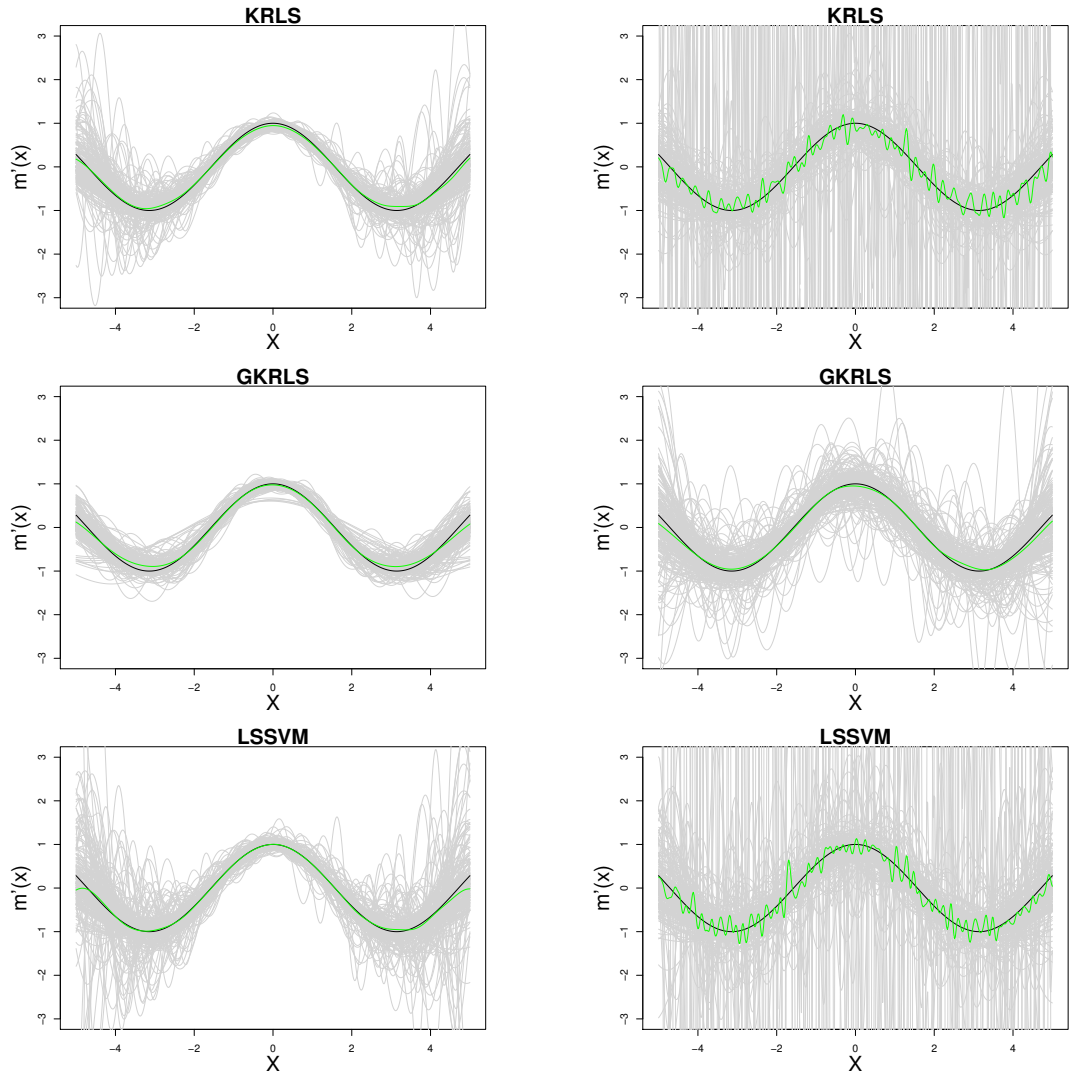

Figure 4.2: *The left (right) plots show the estimated derivatives under heteroskedastic errors (AR(2) errors). The top, middle, and bottom plots refer to the KRLS, GKRLS, and LSSVM estimators. The grey curves show the predicted values from all simulations evaluated at the 500 evaluation data points spanning from -5 to 5. The green curves denote the average predicted derivative estimates at each evaluation point across all simulations, and the black curve represents the true derivative of the regression function in Eq.* (4.39).

Figure 4.2 shows simulation results for the derivative given in Eq. (4.39). Similar to the regression estimates, for both heteroskedastic and AR(2) errors, the variability from

estimating the derivative is reduced from GKRLS estimation, as seen as the two middle plots, relative to KRLS estimation, as seen as the top two plots. In addition, the efficiency gain in estimating both the regression and the derivative seems to be more evident in the AR(2) case comared to the heteroskedastic case. A possible explanation for this is that the covariance matrix contains more information in the off-diagonal elements compared to the diagonal covariance matrix in the heteroskedastic case. Overall, when estimating the regression function and its derivative for this simulation example, the reduction in variance is clearly evident in Figure 4.1 and Figure 4.2.

Table 4.1 displays the evaluations, including bias, variance, and MSE of the estimators for both error cases and for both the regression function and the derivative. Note that all estimates in Table 4.1 are averaged across all simulations. For both error covariance structures, GKRLS estimates of the regression function have the smallest average bias in absolute terms. Furthermore, GKRLS has the lowest variance, and therefore lowest MSE, making GKRLS the preferred method. Note that GKRLS estimation provides a 21.9% and 26.3% decrease in the variance for estimating the regression function for the heteroskedastic errors and autocorrelated errors, relative to KRLS. When estimating the derivative, the reduction in variance is substantial. GKRLS estimation of the derivative provides a 66.4% and 96.2% decrease in the variance for heteroskedastic and autocorrelated errors, relative to KRLS. Note that LSSVM provide similar estimates to KRLS. Moreover, for both regression and derivative function estimations, GKRLS is the preferred method and variance reduction is significant.

Model Simulation Evaluation

			Bias	Variance	MSE
$m(\cdot)$	Heteroskedastic Errors	KRLS	-0.0019	0.0210	0.0213
		GKRLS	-0.0017	0.0164	0.0188
		LSSVM	-0.0019	0.0308	0.0308
	Autocorrelated Errors	KRLS	-0.0033	0.0730	0.0769
		GKRLS	-0.0030	0.0538	0.0548
		LSSVM	-0.0030	0.0828	0.0835
$m^{(1)}(\cdot)$	Heteroskedastic Errors	KRLS	-0.0017	0.0860	0.0881
		GKRLS	0.0064	0.0289	0.0328
		LSSVM	-0.0145	0.3120	0.3150
	Autocorrelated Errors	KRLS	-0.0112	5.4506	5.4848
		GKRLS	-0.0150	0.2082	0.2111
		LSSVM	-0.0136	5.8460	5.8783

Table 4.1: *The table reports the bias, variance, and MSE for KRLS, GKRLS, and LSSVM estimators under Eq.* (4.38)*, Eq.* (4.39)*, and the cases of heteroskedastic and AR(2) errors. All reported estimates are averaged across all simulations.*

Table 4.2 shows the simulation results for the consistency of GKRLS. The bias, variance, and MSE are reported for sample sizes of $n = 100, 200, 400$. In this example, in order to see the effect of increasing the sample size, all hyperparameters are fixed and set to

Simulation Results for Consistency of GKRLS

		Heteroskedastic Errors			Autocorrelated Errors		
		Bias	Variance	MSE	Bias	Variance	MSE
$m(\cdot)$	$n = 100$	0.0005	0.0725	0.0730	0.0010	0.1675	0.1685
	$n = 200$	0.0002	0.0369	0.0371	0.0005	0.0867	0.0872
	$n = 400$	0.0001	0.0194	0.0194	0.0002	0.0456	0.0458
$m^{(1)}(\cdot)$	$n = 100$	0.0060	0.5135	0.5195	0.0082	0.7626	0.7708
	$n = 200$	0.0022	0.3264	0.3286	0.0053	0.4712	0.4766
	$n = 400$	0.0017	0.2082	0.2099	0.0023	0.2815	0.2838

Table 4.2: *The table reports the bias, variance, and MSE for GKRLS estimator under Eq.* (4.38)*, Eq.* (4.39)*, and the cases of heteroskedastic and AR(2) errors for different sample sizes,* $n = 100, 200, 400$*. All reported estimates are averaged across all simulations. All hyperparameters are fixed and set to 1.*

1. For the regression function and the derivative and for both error covariance structures, the bias, variance, and MSE all decrease as the sample size increases, which implies that the GKRLS estimator is consistent in this simulation exercise.

4.7 Applications

We implement two empirical applications: U.S. airline industry with heteroskedastic errors and money demand equation with correlated errors.[2] For both data sets we set

[2]The data for both applications are from [19] and can be downloaded at https://pages.stern.nyu.edu/ wgreene/Text/Edition7/tablelist8new.htm

aside a portion of the data for training and the other for testing. We estimate four models, GKRLS, KRLS, LSSVM, and OLS, and compare their results in terms of mean squared error (MSE). To evaluate the out of sample performance of each model, the predicted out of sample MSEs are computed as follows

$$MSE = \frac{1}{n'}\sum_{j=1}^{n'}(\widehat{m}(\mathbf{x}_{0,j}) - y_j)^2, \tag{4.40}$$

where n' is the number of observations in the testing data set and $j = 1, \ldots, n'$. The in sample MSEs are also reported for the training data. To assess the estimated derivatives, we use the bootstrap to calculate the out of sample MSEs. We report the bootstrapped MSEs for the regression function and its derivative by the following.[3]

$$MSE_{boot} = \frac{1}{B}\frac{1}{n'}\sum_{b=1}^{B}\sum_{j=1}^{n'}(\widehat{m}_b(\mathbf{x}_{0,j}) - y_j)^2 \tag{4.41}$$

$$MSE_{boot,deriv} = \frac{1}{B}\frac{1}{n'}\sum_{b=1}^{B}\sum_{j=1}^{n'}(\widehat{m}_b^{(1)}(\mathbf{x}_{0,j}) - \widehat{m}_{avg}^{(1)}(\mathbf{x}_{0,j}))^2, \tag{4.42}$$

where B is the number of bootstraps with $b = 1, \ldots, B$, $\widehat{m}_b(\cdot)$ and $\widehat{m}_b^{(1)}(\cdot)$ are the bth bootstrapped estimated regression function and its first derivative respectively, and $\widehat{m}^{(1)}(\cdot)$ is the simple average of $f = 1, \ldots, 4$ models (KRLS, GKRLS, LSSVM, and OLS):

$$\widehat{m}_{avg}^{(1)}(\mathbf{x}_{0,j}) = \frac{1}{4}\sum_{f=1}^{4}\widehat{m}_f^{(1)}(\mathbf{x}_{0,j}). \tag{4.43}$$

[3]The R package by [11] was used to obtain the bootstrap samples.

4.7.1 U.S. Airline Industry

We obtain the data on the efficiency in production of airline services from [19]. To model heteroskedasticity we estimate GKRLS for the following:

$$\log C_{it} = m(\log Q_{it}, \log P_{it}) + U_{it}, \tag{4.44}$$

$$\omega_{it} = \exp(\gamma_1 + \gamma_2 Loadfactor_{it}), \tag{4.45}$$

where C_{it} is the total cost, Q_{it} is output, and P_{it} is the price of fuel, and $Loadfactor$, the average capacity utilization of the fleet. The data contain 90 observations of 6 firms for 15 years, from 1970-1984. For simplicity, we pool all of the data for estimation and assume that $Loadfactor$ appears in the variance of the error term. We randomly split the data into two parts, where 70 observations are used as training data and 20 observations are set as testing data to evaluate out of sample performance. For the GKRLS, KRLS, and LSSVM models, all hyperparameters are chosen via cross validation.

We plot the results for the estimated regression function and its derivative for GKRLS and KRLS models in Figure 4.3. For visual purposes, we train the data on the 70 observations in the training data set and evaluate both models with 200 evenly spaced points across the support of each regressor while holding the other variables fixed at their medians. The solid (dashed) curves in red and grey depict the bias corrected point estimates (pointwise 95% confidence interval) for GKRLS and KRLS respectively. Both models seem to display a positive relationship between cost and each of the regressors, output and price, with their partial derivatives being positive almost everywhere. Accounting for the heteroskedasticity in the estimation of the regression function, GKRLS is somewhat smoother

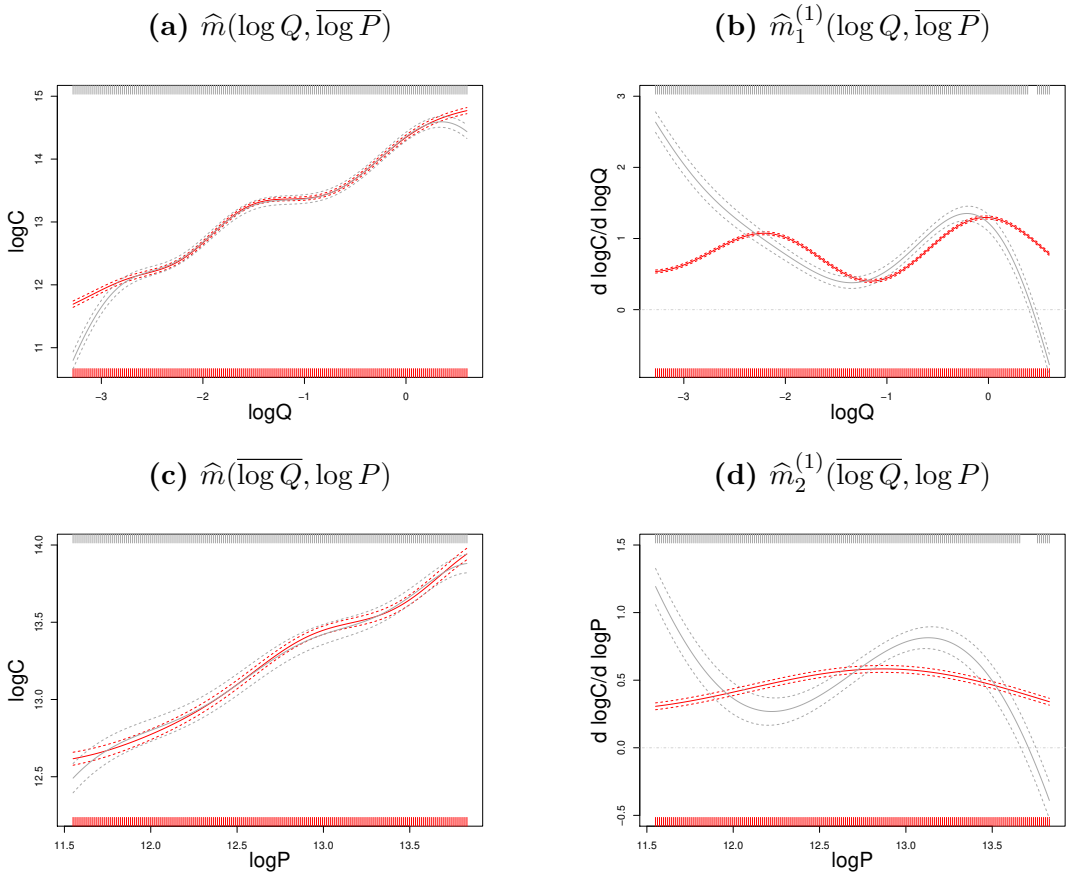

Figure 4.3: *The left (right) plots show the estimated bias corrected predictions of the regression function (derivative) for 200 evenly spaced points across the support for each independent variable. The predictions are plotted for* $\log Q$ *and* $\log P$*, holding all else fixed at their medians. The red (grey) curves correspond to the predictions made by GKRLS (KRLS). The dashed lines are the pointwise confidence intervals.*

for both the estimation of the regression function and its partial derivatives. In addition, the confidence intervals are slightly smaller that that of KRLS, implying that there is an efficiency gain in the GKRLS estimates. The red (grey) tick marks indicate the significance of the estimated regression function and its derivative evaluated at each testing observation for GKRLS (KRLS), where we check to see if zero lies within the interval. Both models are significant almost everywhere in the support for the regression functions and derivatives.

Average Partial Derivative Estimates
for Airline Data

	logQ	logP
GKRLS	0.8495	0.4756
	(0.011)	(0.013)
KRLS	0.9786	0.5129
	(0.0244)	(0.0248)
LSSVM	0.8614	0.6503
	(0.0165)	(0.0106)
OLS	0.9347	0.4167
	(0.0522)	(0.0195)

Table 4.3: *Average partial derivatives and their standard errors in paranetheses are reported for GKRLS, KRLS, LSSVM, and OLS models. The columns represent the estimates of the partial derivative with respect to each regressor. The White standard errors are reported for the OLS model.*

The average partial derivatives and corresponding standard errors are reported in Table 4.3. These averages are calculated by training each model on the 70 observations in the training data set and evaluating all model derivatives with 200 evenly spaced points across the support of each regressor while holding the other variables fixed at their medians. The estimates are bias corrected and the results from Section 4.5 are used in our calculations. The reported estimates for GKRLS and KRLS correspond to the derivative plots in

Figure 4.3. The White heteroskedastic standard errors are reported for the OLS model.[4] Comparing GKRLS and KRLS, the estimates of the partial derivative are similar but the standard errors are significantly reduced for GKRLS, where we see a gain in efficiency, as we have confirmed from Figure 4.3. The partial derivative estimates for LSSVM are similar to those for KRLS but is more efficient. Assuming that GKRLS is the correct model, KRLS and LSSVM would underestimate the elasticity with respect to output and overestimate the elasticity with respect to price. For OLS, we estimate the following

$$\log C_{it} = \beta_0 + \beta_1 \log Q_{it} + \beta_2 \log^2 Q_{it} + \beta_3 \log P_{it} + \varepsilon.$$

Then the partial derivatives are

$$\log Q : \beta_1 + 2\beta_2 \log Q_{it}$$

$$\log P : \beta_3$$

Table 4.3 shows that the OLS model overestimates the elasticity with respect to output and underestimates the elasticty with respect to price compared to those of GKRLS.

To assess the models in terms of out of sample performance, we calculate the MSEs using the 20 observations in the testing data set. Table 4.4 reports MSEs for the four considered models. The first and second rows report the out of sample MSEs using the 20 observations and the bootstrap respectively. The last row reports the in sample MSEs. Considering the nonparametric models, GKLRS, KRLS, and LSSVM, the GKRLS estimator outperforms the others in terms of MSE. The bootstrapped MSEs for the partial derivatives are reported in Table 4.5. For the partial derivative with respect to output,

[4]The R package by [45] was used to obtain the White heteroskedastic standard erorrs.

MSEs for Ailrine Data

	GKRLS	KRLS	LSSVM	OLS
Out of Sample	0.0064	0.0145	0.0129	0.0193
Boot Out of Sample	0.0150	0.0565	0.0268	0.0203
In Sample	0.0016	0.0027	0.0032	0.0175

Table 4.4: *The MSEs are reported for GKRLS, KRLS, LSSVM, and OLS models. The first and second rows are the out of sample MSE and the bootstrapped MSE for the 20 observations in the testing set. The third row is the in sample MSE for the observations in the training set.*

GKRLS produces the lowest MSE, outperforming the other models. Considering only the nonparametric models, the smallest MSE is the one obtained by GKRLS for the derivative with respect to price. However, OLS has the lowest overall MSE for the derivative with respect to price. Looking at the plots with respect to price in Figure 4.3, the GKRLS estimator (red curve) appears to produce a somewhat linear function in price, holding output fixed. We conducted a test for correct specification [26] of a linear model of the cost function in terms of price and failed to reject the null at the 5% level, indicating that cost may be linear in price for this particular data set. This reason is one justification as to why OLS performs the best in terms of the lowest bootstrapped MSE for the derivative with respect to price, since the cost function may be in fact linear with respect to price but not output.

Bootstrapped Partial Derivative MSEs
for Airline Data

	$\log Q$	$\log P$
GKRLS	0.1057	0.0488
KRLS	0.4199	0.3130
LSSVM	0.3745	0.2561
OLS	0.1195	0.0259

Table 4.5: *The bootstrapped MSEs for the GKRLS, KRLS, LSSVM, and OLS partial derivatives are reported. The rows represent the MSEs estimates of the partial derivative with respect to each regressor.*

4.7.2 Money Demand Equation

The data for the money demand equation is obtained from [19] and we consider the following model:

$$\log(M1/CPI)_t = m(\log GDP_t, \log TBILL_t) + u_t \tag{4.46}$$

where $M1$ is nominal money stock, CPI is consumer price index, and $TBILL$ is the quaterly average of month end 90 day treasury bill rate. The errors are assumed to be correlated and an ARMA(1,2) is used to model the residuals. The data contain quarterly data from 1950 to 2000, with 204 total observations. The first 188 observations are used as training data and the last 16 observations are used as testing data to evaluate out of sample performance. For all machine learning models, all hyperparameters are chosen via cross validation.

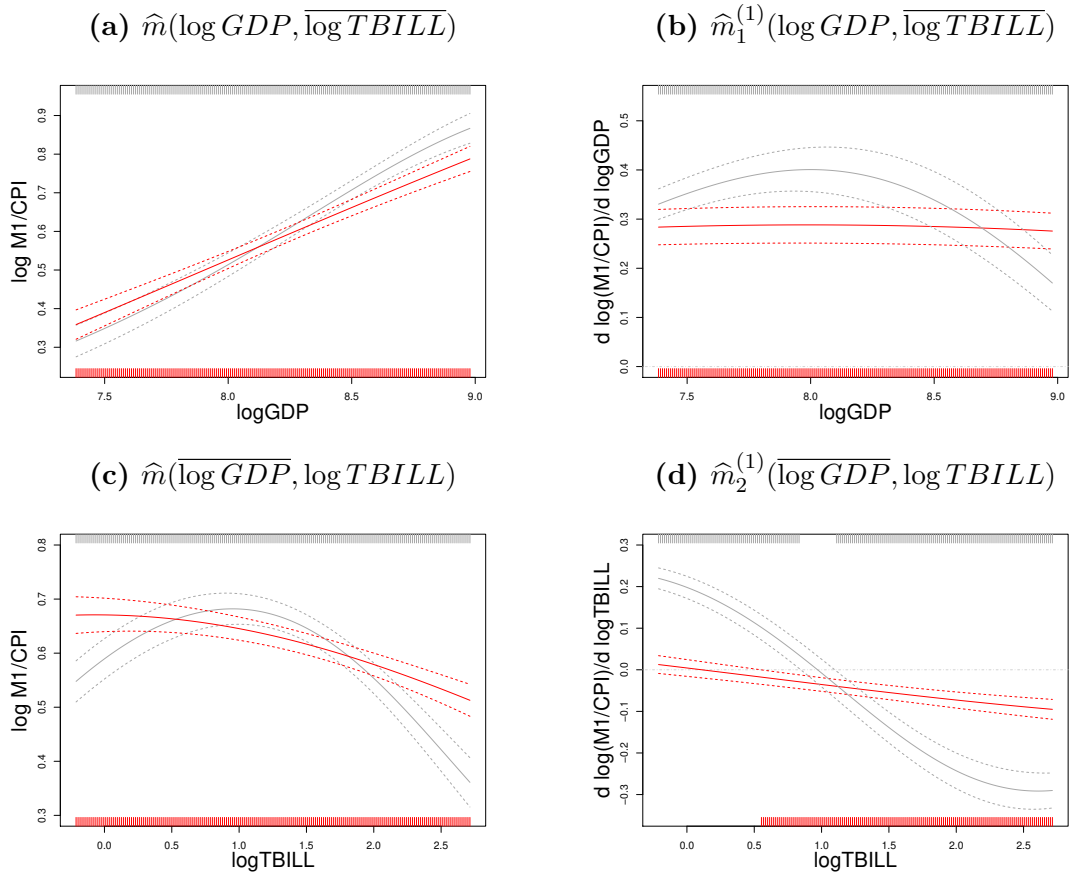

Figure 4.4: *The left (right) plots show the estimated bias corrected predictions of the regression function (derivative) for 200 evenly spaced points across the support for each independent variable. The predictions are plotted for* $\log GDP$*, and* $\log TBILL$*, holding all else fixed at their medians. The red (grey) curves correspond to the predictions made by GKRLS (KRLS). The dashed lines are the pointwise confidence intervals.*

Similar to the empirical example using the airline data, we plot the results for the estimated regression function and its derivatives for GKRLS and KRLS models in Figure 4.4. In this example, we train the data on the 188 observations in the training data set and evaluate both models with 200 evenly spaced points across the support of each regressor while holding the other variables fixed at their medians. The solid (dashed) curves in red and grey depict the bias corrected point estimates (pointwise 95% confidence

Average Partial Derivative Estimates
for Money Demand Data

	logGDP	logTBILL
GKRLS	0.2850	-0.0439
	(0.0188)	(0.0082)
KRLS	0.3426	-0.0618
	(0.0221)	(0.0106)
LSSVM	0.2327	-0.0451
	(0.0076)	(0.0041)
OLS	0.3996	-0.1477
	(0.0957)	(0.0627)

Table 4.6: *Average partial derivatives and their standard errors in paranetheses are reported for GKRLS, KRLS, LSSVM, and OLS models. The columns represent the estimates of the partial derivative with respect to each regressor. The Newey-West standard errors are reported for the OLS model.*

interval) for GKRLS and KRLS respectively. Looking at plots (a) and (b) of Figure 4.4, there seems to be a positive relationship between GDP and money demand. We can see that GKRLS estimates are also smoother compared to KRLS. GKRLS estimates of the regression function appears to be linear and hence the derivative to be constant so GDP may have a linear relationship with money demand in logarithms. The partial derivative with respect to GDP for GKRLS is constant, positive, and significant for the entire support of $\log GDP$. For treasury bill rate (plots (c) and (d) of Figure 4.4), GKRLS appears to

show a negative relationship between treasury bill rate and money demand. However, KRLS appears to have a positive relationship in the beginning of the support of $\log TBILL$ and negative towards the end of the support, holding all else fixed. Looking at the derivative with respect to $\log TBILL$, the derivative calculated by GKRLS is negative and significant for almost the entire support. However, the derivative calculated by KRLS is positive and significant for most of the first half of the support and negative and significant at the end of the support. Basic economic theory tells us that higher treasury bill rates should correspond to lower demand for money. As a result, the GKRLS model would be a more reasonable model to accept. The difference in the two models could be the case that KRLS does not consider the correlation in the errors when estimating the regression function and therefore its derivatives. After taking the information from the residuals in the first step, GKRLS can adjust for the correlation in the estimated function and its derivative to get more sensible results.

MSEs for Money Demand Data

	GKRLS	KRLS	LSSVM	OLS
Out of Sample	0.0012	0.0095	0.0221	0.0137
Boot Out of Sample	0.0052	0.0086	0.0154	0.0128
In Sample	0.0045	0.0022	0.0012	0.0033

Table 4.7: *The MSEs are reported for GKRLS, KRLS, LSSVM, and OLS models. The first and second rows are the out of sample MSE and the bootstrapped MSE for the 16 observations in the testing set. The third row is the in sample MSE for the observations in the training set.*

Table 4.6 shows the average partial derivatives and standard errors with respect to each regressor for the four models evaluated at 200 evenly spaced points.[5] The estimates reported for GKRLS and KRLS correspond to the derivative plots in Figure 4.4. We can see that the estimates for KRLS and GKRLS differ significantly, where KRLS derivative estimates show larger elasticities with respect to GDP and treasury bill rate, implying that KRLS may overestimate the elasticities. Although GKRLS partial derivative estimates are slightly smaller relative to KRLS, GKRLS standard errors are smaller, making GKRLS more efficient. LSSVM partial derivative estimates are very similar to those of GKRLS. Note that although the average partial derivative estimates of GKRLS and LSSVM may be very similar, the pointwise derivatives are not similar.

To assess the models in terms of out of sample performance, we calculate the MSEs using the last 16 observations in the testing data set. Table 4.7 reports MSEs for the four considered models. The first and second rows report the out of sample MSEs using the 16 observations and the bootstrap respectively. The last row reports the in sample MSEs. In terms of out of sample performance, GKRLS achieves the lowest MSE and outperforms the other considered models. The bootstrapped MSEs for the partial derivatives are reported in Table 4.8. For both regressors, the MSEs are the lowest for GKRLS, with GKRLS outperforming the other models.

[5]The R package by [45] was used to obtain the Newey-West standard erorrs.

Bootstrapped Partial Derivative MSEs
for Money Demand Data

	$\log GDP$	$\log TBILL$
GKRLS	0.0104	0.0021
KRLS	0.0279	0.0049
LSSVM	0.0154	0.0238
OLS	0.0168	0.0040

Table 4.8: *The bootstrapped MSEs for the GKRLS, KRLS, LSSVM, and OLS partial derivatives are reported. The rows represent the estimates of the partial derivative with respect to each regressor.*

4.8 Conclusion

Overall, this chapter proposes a nonparametric regression function estimator via KRLS under a general parametric error covariance. The two step procedure allows for heteroskedastic and serially correlated errors, where in the first step, KRLS is used to estimate the regression function and the parametric error covariance, and in the second step, KRLS is used to estimate the regression function using the information in the error covariance. The method improves efficiency in the regression estimates as well as the partial effects estimates compared to standard KRLS. The conditional bias and variance, pointwise marginal effects, consistency, and asymptotic normality of GKRLS are provided. Simulations show that there are improvements in variance and MSE reduction when considering GKRLS relative to KRLS. Two empirical examples are illustrated with estimating an airline cost

function with heteroskedastic errors and with estimating a money demand equation with autocorrelated errors. The derivatives are evaluated, and the average partial effects of the inputs are determined in these applications. In both empirical exercises, GKRLS shows different regression function and derivative function estimates and is more efficient than KRLS.

Chapter 5

Conclusions

We have made three main contributions in bridging the gap between machine learning and econometrics. First, we provide an alternative way of estimating first derivatives of models using weighted difference quotients based on machine learning regression models, allowing economists to analyze the partial marginal effect of a variable and therefore interpret the underlying mechanisms of the model. The procedure is most impactful for models that produce noisy regression functions such as LRF, where direct derivative estimates are not smooth and are not analytically available. Second, we extend the derivative procedure based on smoothing random forest based difference quotients to second derivatives. Second derivatives are often useful in economics to determine the concavity of various functions, e.g.'s production function and earnings function. Lastly, information held in the error covariance is used in estimating a nonparametric regression function via KRLS is used. It is shown that ignoring the information in the error structure can lead to misleading results and incorrect interpretations of the regression function. Therefore, incorporting the error

structure in estimating regressions allow for more efficient and robust results for both the regression function and its derivatives.

In terms of future work, any machine learning estimator can be used in the estimation of derivatives, established in Chapters 2 and 3. However, one estimator that comes in mind is the k-Nearest Neighbor (kNN) estimator. Just like forest-based models, kNN is a powerful nonparametric estimator that produces non-smooth regression functions. Therefore, obtaining marginal effects is non-trivial; however, under the frameworks in Chapters 2 and 3, kNN based derivatives can be estimated. Another possible route is to incorporate endogeneity in estimating the derivatives. In both derivative chapters, the regressors are assumed to be exogenous. It is often the case that many economic variables are considered endogenous and therefore will cause partial effects to be biased. In the derivative chapters, the assumption of exogeneity may be relaxed and more robust derivatives can be estimated.

Another possible area of research is to further expand on the generalized machine learning techniques. A parametric error covariance is used in Chapter 4; however, to make the procedure fully nonparametric, KRLS can be used to estimate the error covariance. Not only can this method generalize to KRLS, but it can also generalize to other machine learning methods such as random forest, support vector machine, kNN, neural networks, etc. In all of these machine learning methods, none of the information in the error structure is used. If there is information that is not exploited, then our estimates may not be as efficient. Therefore, it may be beneficial to incorporate the error structure in estimating the regression function to better enhance our predictions.

In [9], a new combined semiparametric estimator of the conditional variance that takes the product of a parametric estimator and a nonparametric estimator based on KRLS is proposed. In this paper, KRLS is used to estimate the nonparametric component. The paper discusses how to estimate the semiparametric estimator using real data and how to use the estimator to make forecasts for the conditional variance. Simulations are conducted to show the dominance of the proposed estimator in terms of mean squared error. An empirical application using S&P 500 daily returns is analyzed, and the semiparametric estimator effectively forecasts future volatility. The paper can be extended to include theoretical properties of the semiparametric estimator. In addition, extensions to the multivariate conditional variance and covariance can also be estimated, which are studied in the parametrics econometrics through multivariate ARCH and GARCH models. Lastly, instead of KRLS, any other machine learning technique may be used in estimating the nonparametric component.

Overall, it is an exciting time with the advancement of technology and the amount of economic data we have easily accessible. With the help of machine learning, a lot of data based and nonparametric approaches can be used to help solve economic problems. I hope to further continue this area of research.

Bibliography

[1] Scott E. Atkinson and Jeffrey H. Dorfman. Feasible estimation of firm-specific allocative inefficiency through bayesian numerical methods. *Journal of Applied Econometrics*, 24(4):675–697, 2009.

[2] Hans W. Borchers. *pracma: Practical Numerical Math Functions*, 2021. R package version 2.3.3.

[3] K. Brabanter, F. Cao, I. Gijbels, and J. Opsomer. Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, 105:681–690, 09 2018.

[4] Kris Brabanter, Jos De Brabanter, Bart De Moor, and I. Gijbels. Derivative estimation with local polynomial fitting. *The Journal of Machine Learning Research*, 14:281–301, 01 2013.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Jorge Luis Ojeda Cabrera. *locpol: Kernel Local Polynomial Regression*, 2018. R package version 0.7-0.

[7] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[8] Richard Charnigo, Benjamin Hall, and Cidambi Srinivasan. A generalized cp criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

[9] Justin Dang and Aman Ullah. Machine-learning-based semiparametric time series conditional variance: Estimation and forecasting. *Journal of Risk and Financial Management*, 15(1), 2022.

[10] H.A. David and H.N. Nagaraja. *Order Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1970.

[11] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997. ISBN 0-521-57391-2.

[12] K. De Brabanter, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks*, 22(1):110–120, 2011.

[13] Christopher Dougherty. Why are the returns to schooling higher for women than for men? *The Journal of Human Resources*, 40(4):969–988, 2005.

[14] Tarn Duong. *ks: Kernel Smoothing*, 2020. R package version 1.11.7.

[15] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1996.

[16] Jeremy Ferwerda, Jens Hainmueller, and Chad J. Hazlett. Kernel-based regularized least squares in R (KRLS) and Stata (krls). *Journal of Statistical Software*, 79(3):1–26, 2017.

[17] Yuri Fonseca, Marcelo Medeiros, Gabriel Vasconcelos, and Alvaro Veiga. BooST: Boosting Smooth Trees for Partial Effect Estimation in Nonlinear Regressions. Papers 1808.03698, arXiv.org, August 2018.

[18] R. Friedberg, Julie Tibshirani, S. Athey, and S. Wager. Local linear forests. *ArXiv*, abs/1807.11408, 2018.

[19] W.H. Greene. *Econometric Analysis*. Pearson, 2018.

[20] Jens Hainmueller and Chad Hazlett. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168, 2014.

[21] Peter Hall, J. W. Kay, and D. M. Titterington. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[23] Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

[24] Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

[25] Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

[26] Cheng Hsiao, Qi Li, and Jeffrey Racine. A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics*, 140(2):802–826, 2007.

[27] Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22, 2008.

[28] Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2 edition, 2008.

[29] Kristiaan Kerstens and Ignace Van de Woestyne. Cost functions are nonconvex in the outputs when the technology is nonconvex: convexification is not harmless. *Annals of Operations Research*, 305:81–106, 2021.

[30] Yu Liu and Kris De Brabanter. Smoothed nonparametric derivative estimation using weighted difference quotients. *Journal of Machine Learning Research*, 21(65):1–45, 2020.

[31] Yu Liu and Kris De Brabanter. Derivative estimation in random design. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3445–3454. Curran Associates, Inc., 2018.

[32] Shujie Ma, Jeffrey Racine, and Aman Ullah. Nonparametric Estimation of Marginal Effects in Regression-spline Random Effects Models. Working Papers 201920, University of California at Riverside, Department of Economics, September 2019.

[33] Carlos Martins-Filho and Feng Yao. Nonparametric regression estimation with general parametric error covariance. *Journal of Multivariate Analysis*, 100(3):309 – 333, 2009.

[34] A. I. McLeod, Hao Yu, and Zinovi Krougly. Algorithms for linear time series analysis: With r package. *Journal of Statistical Software*, 23(5), 2007.

[35] Jacob Mincer. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, Inc, 1974.

[36] Kevin M. Murphy and Finis Welch. Empirical age-earnings profiles. *Journal of Labor Economics*, 8(2):202–229, 1990.

[37] Adrian Pagan and Aman Ullah. *Nonparametric Econometrics*. Cambridge University Press, 1999.

[38] Regina T. Riphahn, Achim Wambach, and Andreas Million. Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics*, 18:387–405, 2003.

[39] Johannes Sauer. Economic theory and econometric practice: Parametric efficiency analysis. *Empirical Economics*, 31(4):1061–1087, 2006.

[40] A.N. Shiryaev. *Probability*. Springer, New York, 2nd edition, 1996.

[41] Liangjun Su, Aman Ullah, and Yun Wang. Nonparametric regression estimation with general parametric error covariance: a more efficient two-step estimator. *Empirical Economics*, 45(2):1009–1024, 2013.

[42] Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, and Marvin Wright. *grf: Generalized Random Forest*, 2020. R package version 1.2.0.

[43] WenWu Wang and Lu Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16(81):2617–2641, 2015.

[44] H. White. *Asymptotic Theory for Econometricians*. Economic Theory, Econometrics, and Mathematical Economics. Emerald Group Publishing Limited, 2001.

[45] Achim Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006.

[46] Achim Zeileis and Torsten Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3):7–10, 2002.

[47] Zhou and D. A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108, 2000.

Appendix A

Chapter 2 Appendix

A.1 Proof of Prop. 2.1

The conditioal variance of Eq. (2.6) is

$$\begin{aligned}
\mathrm{Var}[\widehat{Y}^{(1)}_{i,ML}|\mathbb{U}] &= \mathrm{Var}\left[\sum_{j=1}^{k} w_{i,j}\frac{\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})}{U_{i+j}-U_{i-j}}\middle|\mathbb{U}\right] \\
&= \left(1-\sum_{j=2}^{k} w_{i,j}\right)^2 \frac{\mathrm{Var}[\widehat{r}(U_{i+1})|\mathbb{U}]+\mathrm{Var}[\widehat{r}(U_{i-1})|\mathbb{U}]}{(U_{i+1}-U_{i-1})^2} \\
&\quad + \sum_{j=2}^{k} w_{i,j}^2 \frac{\mathrm{Var}[\widehat{r}(U_{i+j})|\mathbb{U}]+\mathrm{Var}[\widehat{r}(U_{i-j})|\mathbb{U}]}{(U_{i+j}-U_{i-j})^2} \\
&= \left(1-\sum_{j=2}^{k} w_{i,j}\right)^2 \frac{\sigma^2_{\widehat{r},i+1}+\sigma^2_{\widehat{r},i-1}}{(U_{i+1}-U_{i-1})^2} + \sum_{j=2}^{k} w_{i,j}^2 \frac{\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j}}{(U_{i+j}-U_{i-j})^2}.
\end{aligned}$$

For all $j = 1, \ldots, k$, take the partial derivative with respect to $w_{i,j}$ and set it to zero results in

$$w_{i,j} = w_{i,1}\frac{(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})}{(U_{i+1}-U_{i-1})^2/(\sigma^2_{\widehat{r},i+1}+\sigma^2_{\widehat{r},i-1})},$$

which shows the relationship between $w_{i,1}$ and $w_{i,j}$. Since $\sum_{j=1}^{k} w_{i,j} = 1$,

$$\sum_{j=1}^{k} w_{i,j} = \frac{w_{i,1}}{(U_{i+1} - U_{i-1})^2/(\sigma^2_{\widehat{r},i+1} + \sigma^2_{\widehat{r},i-1})} \sum_{j=1}^{k} (U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j}) = 1$$

Substituting for $w_{i,1}$ gives

$$\frac{w_{i,j}}{(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})} \sum_{j=1}^{k} (U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j}) = 1,$$

proving the proposition.

A.2 Proof of Theorem 2.1

Consider the taylor expansions for $\widehat{r}(U_{i+j})$ and $\widehat{r}(U_{i-j})$ in the neighborhood of U_i

$$\widehat{r}(U_{i+j}) = \widehat{r}(U_i) + (U_{i+j} - U_i)\widehat{r}^{(1)}(U_i) + \frac{(U_{i+j} - U_i)^2}{2}\widehat{r}^{(2)}(\zeta_{i,i+j})$$
$$\widehat{r}(U_{i-j}) = \widehat{r}(U_i) + (U_{i-j} - U_i)\widehat{r}^{(1)}(U_i) + \frac{(U_{i-j} - U_i)^2}{2}\widehat{r}^{(2)}(\zeta_{i-j,i}),$$

where $\zeta_{i,i+j} \in]U_i, U_{i+j}[$ and $\zeta_{i-j,i} \in]U_{i-j}, U_i[$. Then, using Lemma 2.1 and Prop. 2.1, the absolute conditional bias is

$$\begin{aligned}
\left| \text{Bias}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] \right| &= \left| \mathbb{E}\left[\sum_{j=1}^{k} w_{i,j} \frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_{i-j})}{U_{i+j} - U_{i-j}} \middle| \mathbb{U} \right] - r^{(1)}(U_i) \right| \\
&= \left| \sum_{j=1}^{k} w_{i,j} \frac{r(U_{i+j}) + \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - r(U_{i-j}) - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}} \right. \\
&\qquad \left. - r^{(1)}(U_i) \right| \\
&= \left| \sum_{j=1}^{k} w_{i,j} \left[\frac{r(U_{i+j}) - r(U_{i-j})}{U_{i+j} - U_{i-j}} + \frac{\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}} \right] \right. \\
&\qquad \left. - r^{(1)}(U_i) \right| \\
&= \left| \sum_{j=1}^{k} w_{i,j} \left[\frac{r(U_i) + r^{(1)}(U_i)(U_{i+j} - U_i) + \frac{1}{2} r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2}{U_{i+j} - U_{i-j}} \right. \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{r(U_i) + r^{(1)}(U_i)(U_{i-j} - U_i) + \frac{1}{2}r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2}{U_{i+j} - U_{i-j}} \\
&\quad + \frac{\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}}\Bigg] - r^{(1)}(U_i)\Bigg| \\
&= \left|\frac{1}{2}\sum_{j=1}^{k} w_{i,j}\left[\frac{r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2 - r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2}{U_{i+j} - U_{i-j}}\right]\right. \\
&\quad \left. + \sum_{j=1}^{k} w_{i,j}\left[\frac{\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}}\right]\right|,
\end{aligned}
$$

where $\zeta_{i,i+j} \in]U_i, U_{i+j}[$ and $\zeta_{i-j,i} \in]U_{i-j}, U_i[$. Now, substitute the ith weight and use $\sigma^2_{\widehat{r},i} = O(n^{-(1-\beta)})$ for all i.

$$
\begin{aligned}
w_{i,j} &= \frac{(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})} \\
&= \frac{(U_{i+j} - U_{i-j})^2}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2}
\end{aligned}
$$

$$
\begin{aligned}
\left|\text{Bias}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}]\right| &= \left|\frac{\frac{1}{2}\sum_{j=1}^{k}(U_{i+j} - U_{i-j})(r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2 - r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2)}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2}\right. \\
&\quad \left. + \frac{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})(\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}])}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2}\right| \\
&= \frac{1}{2}\left|\frac{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})(r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2 - r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2)}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2}\right|
\end{aligned}
$$

The last equality holds since we have subsamples of size s with $s = n^\beta$, which allows the errors of the forests to be variance-dominated. Then, the absolute conditional bias is bounded by the same bound provided by [30] for $k \to \infty$ as $n \to \infty$.

From Prop. 2.1, the conditional variance is

$$\begin{aligned}
\mathrm{Var}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] &= \mathrm{Var}\left[\sum_{j=1}^{k} w_{i,j}\frac{\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})}{U_{i+j}-U_{i-j}}\Big|\mathbb{U}\right] \\
&= \mathrm{Var}\left[\sum_{j=1}^{k}\left\{\frac{(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})}{\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})}\frac{\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})}{U_{i+j}-U_{i-j}}\right\}\Big|\mathbb{U}\right] \\
&= \mathrm{Var}\left[\frac{\sum_{j=1}^{k}[(U_{i+j}-U_{i-j})/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})][\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})]}{\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})}\Big|\mathbb{U}\right] \\
&= \frac{\sum_{j=1}^{k}[(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})^2]\,\mathrm{Var}[\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})|\mathbb{U}]}{\left(\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})\right)^2} \\
&= \frac{\sum_{j=1}^{k}[(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})^2][(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})]}{\left(\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})\right)^2} \\
&= \sum_{l=1}^{k}\frac{\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l}}{(U_{i+l}-U_{i-l})^2}.
\end{aligned}$$

Then, using Lemma 2.1 and $\sigma^2_{\widehat{r},i} = O(n^{-(1-\beta)})$ for all i, the conditional variance is bounded above by

$$\begin{aligned}
\mathrm{Var}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] &= \sum_{l=1}^{k}\frac{\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l}}{(U_{i+l}-U_{i-l})^2} \\
&\le 2n^{-(1-\beta)}\sum_{l=1}^{k}\frac{1}{(U_{i+l}-U_{i-l})^2} \\
&= 2n^{-(1-\beta)}\frac{1}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\{1+o_p(1)\}} \\
&= \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\{1+o_p(1)\}
\end{aligned}$$

for $k \to \infty$ as $n \to \infty$.

A.3 Proof of Cor. 2.1

As $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^{(1+\beta)}k^{-3}$ from Theorem 2.1, the upperbound of the conditional bias and conditional variance tend to zero. Therefore,

$$\lim_{n\to\infty} \text{MSE}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] = 0.$$

Use Chebyshev's inequality to complete the proof.

A.4 Proof of Cor. 2.2

Using the bias-variance decomposition of means squared error (MSE), the MSE is bounded above by

$$\text{MSE}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] \leq \left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}.$$

Then, the conditional mean integrated squared error (MISE), is

$$\begin{aligned}
\text{MISE}[\widehat{Y}^{(1)}_{ML}|\mathbb{U}] &= \mathbb{E}\int_0^1 \left(\widehat{Y}^{(1)}_{LRF}(U) - r^{(1)}(U)|\mathbb{U}\right)^2 dU \\
&= \int_0^1 \mathbb{E}\left(\widehat{Y}^{(1)}_{LRF}(U) - r^{(1)}(U)|\mathbb{U}\right)^2 dU \\
&\leq \left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} + o_p(n^{-2}k^2 + n^{(1+\beta)}k^{-3})
\end{aligned}$$

Therefore, the asymptotic conditional MISE (AMISE) is

$$\text{AMISE}[\widehat{Y}^{(1)}_{LRF}|\mathbb{U}] \leq \left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}.$$

A.5 Proof of Theorem 2.2

Conditional bias:

$$\begin{aligned}
\text{Bias}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] &= \mathbb{E}[\widehat{r}^{(1)}(u_0)] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \mathbb{E}[\widehat{\mathbf{y}}^{(1)}|\widetilde{\mathbb{U}}] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \left(\begin{bmatrix} r^{(1)}(U_{k+1}) \\ \vdots \\ r^{(1)}(U_{n-k}) \end{bmatrix} + \begin{bmatrix} \text{Bias}[\widehat{Y}^{(1)}_{k+1,LRF}|\mathbb{U}] \\ \vdots \\ \text{Bias}[\widehat{Y}^{(1)}_{n-k,LRF}|\mathbb{U}] \end{bmatrix} \right) - r^{(1)}(u_0) \\
&= \left\{ \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \begin{bmatrix} r^{(1)}(U_{k+1}) \\ \vdots \\ r^{(1)}(U_{n-k}) \end{bmatrix} - r^{(1)}(u_0) \right\} \\
&\qquad + \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \begin{bmatrix} \text{Bias}[\widehat{Y}^{(1)}_{k+1,LRF}|\mathbb{U}] \\ \vdots \\ \text{Bias}[\widehat{Y}^{(1)}_{n-k,LRF}|\mathbb{U}] \end{bmatrix}
\end{aligned}$$

For p odd, from Theorem 3.1 in [15], the first term is

$$\boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \begin{bmatrix} r^{(1)}(U_{k+1}) \\ \vdots \\ r^{(1)}(U_{n-k}) \end{bmatrix} - r^{(1)}(u_0) = \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + o_p(h^{p+1}),$$

where $c_p = (\mu_{p+1}, \ldots, \mu_{p+1})^\top$, $\mu_j = \int u^j K(u)du$, and $\mathbf{S} = (\mu_{i+j})_{0 \le i,j \le p}$. From Theorem 2.1, for $k \to \infty$ as $n \to \infty$, the second term is

$$\boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \begin{bmatrix} \text{Bias}[\widehat{Y}^{(1)}_{k+1,LRF}|\mathbb{U}] \\ \vdots \\ \text{Bias}[\widehat{Y}^{(1)}_{n-k,LRF}|\mathbb{U}] \end{bmatrix} \le \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} \mathbf{U}^\top \mathbf{W} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\times \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \{1 + o_p(1)\}$$

$$\le \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \tilde{c}_p \{1 + o_p(1)\},$$

and since from [30], it is shown that

$$\mathbf{S}_{n-2k} = \mathbf{U}^\top \mathbf{W} \mathbf{U}$$

$$= (n-2k) f(u_0) H \mathbf{S} H \{1 + o_p(1)\},$$

where $H = \text{diag}\{1, h, \ldots, h^p\}$, and

$$\mathbf{U}^\top \mathbf{W} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = (n-2k) f(u_0) H \tilde{c}_p \{1 + o_p(1)\} 1,$$

where $\tilde{c}_p = (\mu_0, \mu_1, \ldots, \mu_p)^\top$. Finally, the conditional bias of the smoothed derivative estimator is bounded by

$$\text{Bias}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] \le \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} \right.$$

$$\left. + \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \right] \{1 + o_p(1)\}$$

Conditional Variance:

When $k \to \infty$ as $n \to \infty$, the conditional variance from Theorem 2.1 is

$$\mathrm{Var}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] = \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\{1+o_p(1)\}$$

and from Theorem 1 of [3],

$$\begin{aligned}\mathrm{Var}[\widehat{r}^{(1)}(u_0)|\tilde{\mathbb{U}}] &= \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1}(\mathbf{U}^\top \mathbf{W}\,\mathrm{Var}[\widehat{\mathbf{Y}}^{(1)}|\widetilde{\mathbb{U}}]\mathbf{W}\mathbf{U})\mathbf{S}_{n-2k}^{-1}\boldsymbol{\epsilon}_1 \\ &\leq \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\frac{1+f(u_0)\rho_c}{h(n-2k)f(u_0)}\boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}\boldsymbol{\epsilon}_1\{1+o_p(1)\},\end{aligned}$$

where $\lim_{n\to\infty} n\int \rho_n(x)dx = \rho_c$ and $\mathbf{S}^* = (\nu_{i+j})_{0\leq i,j\leq p}$ with $\nu_j = \int u^j K^2(u)du$. For p odd, from Theorem 3.1 of [15],

$$\begin{aligned}\int K_0^*(t)dt &= \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1}\left(\int K(t)dt, \quad \int tK(t)dt \quad \ldots, \quad \int t^p K(t)dt\right)^\top \\ &= \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1}\tilde{c}_p.\end{aligned}$$

Similarly,

$$\int t^{p+1}K_0^*(t)dt = \boldsymbol{\epsilon}_1\mathbf{S}^{-1}c_p, \quad \int K_0^{*2}(t)dt = \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}{-}1\boldsymbol{\epsilon}_1.$$

A.6 Proof of Cor. 2.3

For $h \to 0$, $nh \to \infty$ and $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^\beta k^{-3}h^{-1} \to 0$, then from Theorem 2.2, the upperbound of the conditional bias and conditional variance go to zero. Then,

$$\lim_{n\to\infty} \mathrm{MSE}[\widehat{r}^{(1)}(u_0)|\tilde{\mathbb{U}}] = 0$$

and use Chebyshev's inequality to complete the proof.

A.7 More Simulation Studies

The table below shows simulations, under the same DGP in Section 2.5, for two other candidate estimators that are not discussed in the paper. The first is SmoothLRF, where the a LRF is trained on the data, and the fitted values are obtained, $\widehat{y}_{LRF}$. Then a local polynomial regression is used on the new dataset $(x, \widehat{y}_{LRF})$ and the first derivative can be obtained from the second element of the gradient vector. The second is DoubleLocCubic where a local cubic regression is trained on the data and the first derivative is obtained, denoted by $\widehat{y}^{(1)}_{LocCubic}$. Then another local cubic regression is used on the new dataset, $(x, \widehat{y}^{(1)}_{LocCubic})$ and the first derivative can be otbained from the first element of the gradient vector.

For the proposed estimator, DQSmoothLRF, and local cubic regression, LocCub, the results are the same as those in Section 2.5 and are here for comparison. Overall, DQSmoothLRF still outperforms the other estimators, where SmoothLRF does worse in reducing variance, MSE, and MAE. However, the derivative estimated by SmoothLRF does significantly better than the derivative estimated by LRF. DoubleLocCubic does improve upon LocCub, showing a reduction in variance, MSE, and MAE. Overall, DQSmoothLRF seems to be robust in estimating the first derivative, even to other candidate estimators.

Simulations of Other Candidate Estimators

	Bias	Variance	MSE	MAE
DQSmoothLRF	0.0121	0.2085	0.2943	0.4180
SmoothLRF	0.0011	0.3457	0.3699	0.4395
LocCubic	-0.0045	0.4437	0.4789	0.4565
DoubleLocCubic	-0.0046	0.4312	0.4667	0.4503

Table A.1: *The table shows bias, variance, MSE, and MAE for the first derivative, comparing four models, DQSmoothLRF (the proposed estimator), LocCubic, and DoubleLocCubic. All estimates are averaged across all simulations. All models are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

Appendix B

Chapter 3 Appendix

B.1 Proof of Prop. 3.1

Under the assumptions of Prop. 3.1 and using Lemma 2.1,

$$
\begin{aligned}
\mathbb{E}[{}^{+}\widehat{Y}^{(1)}_{i+j,LRF} -{}^{-} \widehat{Y}^{(1)}_{i-j,LRF}|\mathbb{U}] &= \frac{r(U_{i+j+k_1}) + \text{Bias}[\widehat{r}(U_{i+j+k_1})|\mathbb{U}] - r(U_{i+j}) - \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}]}{U_{i+j+k_1} - U_{i+j}} \\
&\quad - \frac{r(U_{i-j-k_1}) + \text{Bias}[\widehat{r}(U_{i-j-k_1})|\mathbb{U}] - r(U_{i-j}) - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i-j-k_1} - U_{i-j}} \\
&= r^{(1)}(U_{i+j}) + \frac{1}{2}r^{(2)}(U_{i+j})(U_{i+j+k_1} - U_{i+j})(1 + o_p(1)) \\
&\quad - r^{(1)}(U_{i-j}) + \frac{1}{2}r^{(2)}(U_{i-j})(U_{i-j-k_1} - U_{i-j})(1 + o_p(1)) \\
&\quad + \frac{\text{Bias}[\widehat{r}(U_{i+j+k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}]}{U_{i+j+k_1} - U_{i+j}} \\
&\quad - \frac{\text{Bias}[\widehat{r}(U_{i-j-k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i-j-k_1} - U_{i-j}} \\
&= \frac{1}{2}r^{(2)}(U_i)(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})(1 + o_p(1)) \\
&\quad + \frac{\text{Bias}[\widehat{r}(U_{i+j+k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}]}{U_{i+j+k_1} - U_{i+j}}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\text{Bias}[\widehat{r}(U_{i-j-k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i-j-k_1} - U_{i-j}} \\
&= \frac{1}{2} r^{(2)}(U_i)(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})(1 + o_p(1)),
\end{aligned}
$$

where the subsample rate β is chosen such that $\beta_{min} < \beta < 1$, the LRF estimator is asymptotically unbiased. The weight for observation i is selected to be proportional to the inverse of the conditional variance of each quotient,

$$
\begin{aligned}
w_{i,j,2} &= \frac{1/\operatorname{Var}\left[\left.\frac{+\widehat{Y}^{(1)}_{i+j,LRF} - -\widehat{Y}^{(1)}_{i-j,LRF}}{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})}\right|\mathbb{U}\right]}{\sum_{j=1}^{k_2} 1/\operatorname{Var}\left[\left.\frac{+\widehat{Y}^{(1)}_{i+j,LRF} - -\widehat{Y}^{(1)}_{i-j,LRF}}{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})}\right|\mathbb{U}\right]} \\
&= \frac{1/\left[\frac{\sigma^2_{\widehat{r},i+j+k_1}+\sigma^2_{\widehat{r},i+j}}{(U_{i+j+k_1}+U_{i+j})^2} + \frac{\sigma^2_{\widehat{r},i-j-k_1}+\sigma^2_{\widehat{r},i-j}}{(U_{i-j-k_1}+U_{i-j})^2}}{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})^2}\right]}{\sum_{j=1}^{k_2} 1/\left[\frac{\sigma^2_{\widehat{r},i+j+k_1}+\sigma^2_{\widehat{r},i+j}}{(U_{i+j+k_1}+U_{i+j})^2} + \frac{\sigma^2_{\widehat{r},i-j-k_1}+\sigma^2_{\widehat{r},i-j}}{(U_{i-j-k_1}+U_{i-j})^2}}{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})^2}\right]},
\end{aligned}
$$

where $\sigma^2_{\widehat{r}}$ is the variance of the LRF estimator.

B.2 Proof of Theorem 3.1

Since r is three times continuously differentiable on the interval $[0, 1]$, consider the Taylor expansions for $r(U_{i+j+k_1})$ and $r(U_{i-j-k_1})$ in the neighborhood of U_{i+j} and U_{i-j},

$$
\begin{aligned}
r(U_{i+j+k_1}) &= r(U_{i+j}) + \sum_{q=1}^{2} \frac{1}{q!}(U_{i+j+k_1} - U_{i+j})^q r^{(q)}(U_{i+j}) \\
&\quad + \frac{(U_{i+j+k_1} - U_{i+j})^3}{6} r^{(3)}(\zeta_{i+j,i+j+k_1}) \\
r(U_{i-j-k_1}) &= r(U_{i-j}) + \sum_{q=1}^{2} \frac{1}{q!}(U_{i-j-k_1} - U_{i-j})^q r^{(q)}(U_{i-j}) \\
&\quad + \frac{(U_{i-j-k_1} - U_{i-j})^3}{6} r^{(3)}(\zeta_{i-j-k_1,i-j}),
\end{aligned}
$$

where $\zeta_{i+j,i+j+k_1} \in]U_{i+j}, U_{i+j+k_1}[$ and $\zeta_{i-j-k_1,i-j} \in]U_{i-j-k_1}, U_{i-j}[$. Also, consider the Taylor expansions for $r^{(1)}(U_{i+j})$, $r^{(1)}(U_{i-j})$, $r^{(2)}(U_{i+j})$, and $r^{(2)}(U_{i-j})$ in the neighborhood of U_i

$$r^{(1)}(U_{i+j}) = r^{(1)}(U_i) + (U_{i+j} - U_i)r^{(2)}(U_i) + \frac{(U_{i+j} - U_i)^2}{2} r^{(3)}(\zeta_{i,i+j})$$
$$r^{(1)}(U_{i-j}) = r^{(1)}(U_i) + (U_{i-j} - U_i)r^{(2)}(U_i) + \frac{(U_{i-j} - U_i)^2}{2} r^{(3)}(\zeta_{i-j,i})$$
$$r^{(2)}(U_{i+j}) = r^{(2)}(U_i) + (U_{i+j} - U_i)r^{(3)}(\zeta'_{i,i+j})$$
$$r^{(2)}(U_{i-j}) = r^{(2)}(U_i) + (U_{i-j} - U_i)r^{(3)}(\zeta'_{i-j,i})$$

where $\zeta_{i,i+j} \in]U_i, U_{i+j}[$, $\zeta_{i-j,i} \in]U_{i-j}, U_i[$, $\zeta'_{i,i+j} \in]U_i, U_{i+j}[$, and $\zeta'_{i-j,i} \in]U_{i-j}, U_i[$. The absolute conditional bias of $\hat{Y}^{(2)}_{i,LRF}$ is

$$\begin{aligned}
\left| \text{Bias}[\hat{Y}^{(2)}_{i,LRF} | \tilde{\mathbb{U}}] \right| &= \left| \mathbb{E}[\hat{Y}^{(2)}_{i,LRF}] - r^{(2)}(U_i) \right| \\
&= \left| \mathbb{E}\left[2 \sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{\widehat{r}(U_{i-j-k_1}) - \widehat{r}(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right] - r^{(2)}(U_i) \right| \\
&= \left| 2 \sum_{j=1}^{k_2} \left\{ w_{i,j,2} \frac{\left(\frac{r(U_{i+j+k_1}) - r(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{r(U_{i-j-k_1}) - r(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right.\right. \\
&\quad \left.\left. + \frac{\left(\frac{\text{Bias}[\widehat{r}(U_{i+j+k_1})] - \text{Bias}[\widehat{r}(U_{i+j})]}{U_{i+j+k_1} - U_{i+j}} - \frac{\text{Bias}[\widehat{r}(U_{i-j-k_1})] - \text{Bias}[\widehat{r}(U_{i-j})]}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right\} - r^{(2)}(U_i) \right| \\
&= \left| 2 \sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{i+j+k_1}) - r(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{r(U_{i-j-k_1}) - r(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} - r^{(2)}(U_i) \right| \\
&\leq \sup_{u \in [0,1]} |r^{(3)}(u)| \left(\sum_{j=1}^{k_2} w_{i,j,2} \frac{(U_{i+j} - U_i)^2 + (U_{i-j} - U_i)^2}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right. \\
&\quad + \sum_{j=1}^{k_2} w_{i,j,2} \frac{(U_{i+j} - U_i)(U_{i+j+k_1} - U_{i+j}) + (U_{i-j} - U_i)(U_{i-j-k_1} - U_{i-j})}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \\
&\quad \left. + \sum_{j=1}^{k_2} w_{i,j,2} \frac{\frac{1}{3}(U_{i+j+k_1} - U_{i+j})^2 + \frac{1}{3}(U_{i-j-k_1} - U_{i-j})^2}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right)
\end{aligned}$$

where the last equality holds for the subsample rate β such that $\beta_{min} < \beta < 1$. Then, using the weights in Eq. (3.5), Lemma 2.1, and $\sigma_{\widehat{r}} = O(n^{-(1-\beta)})$, where $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ gives

$$\left| \text{Bias}[\hat{Y}^{(2)}_{i,LRF} | \tilde{\mathbb{U}}] \right| \leq \frac{\sup_{u \in [0,1]} |r^{(3)}(u)|}{n+1} \times \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\}$$

The variance of the second derivative estimator, $\hat{Y}^{(2)}_{i,LRF}$, is

$$\begin{aligned}
\text{Var}[\hat{Y}^{(2)}_{i,LRF} | \tilde{\mathbb{U}}] = \text{Cov} \Bigg[& 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{\widehat{r}(U_{i-j-k_1}) - \widehat{r}(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}}, \\
& 2\sum_{l=1}^{k_2} w_{i,l,2} \frac{\left(\frac{\widehat{r}(U_{i+l+k_1}) - \widehat{r}(U_{i+l})}{U_{i+l+k_1} - U_{i+l}} - \frac{\widehat{r}(U_{i-l-k_1}) - \widehat{r}(U_{i-l})}{U_{i-l-k_1} - U_{i-l}} \right)}{U_{i+l+k_1} + U_{i+l} - U_{i-l-k_1} - U_{i-l}} \Bigg] \\
= 4 & \sum_{j=1}^{k_2} \sum_{l=1}^{k_2} \Bigg\{ w_{i,j,2} w_{i,l,2} \left(\frac{1}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right) \\
& \times \left(\frac{1}{U_{i+l+k_1} + U_{i+l} - U_{i-l-k_1} - U_{i-l}} \right) \\
& \times \Bigg(\frac{\text{Cov}[\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j}), \widehat{r}(U_{i+l+k_1}) - \widehat{r}(U_{i+l})]}{(U_{i+j+k_1}) - U_{i+j})((U_{i+l+k_1}) - U_{i+l})} \\
& - \frac{\text{Cov}[\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j}), \widehat{r}(U_{i-l-k_1}) - \widehat{r}(U_{i-l})]}{(U_{i+j+k_1}) - U_{i+j})((U_{i-l-k_1}) - U_{i-l})} \\
& - \frac{\text{Cov}[\widehat{r}(U_{i-j-k_1}) - \widehat{r}(U_{i-j}), \widehat{r}(U_{i+l+k_1}) - \widehat{r}(U_{i+l})]}{(U_{i-j-k_1}) - U_{i-j})((U_{i+l+k_1}) - U_{i+l})} \\
& + \frac{\text{Cov}[\widehat{r}(U_{i-j-k_1}) - \widehat{r}(U_{i-j}), \widehat{r}(U_{i-l-k_1}) - \widehat{r}(U_{i-l})]}{(U_{i-j-k_1}) - U_{i-j})((U_{i-l-k_1}) - U_{i-l})} \Bigg) \Bigg\},
\end{aligned}$$

where the first covariance is

$$\begin{aligned}\mathrm{Cov}[\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j}), \widehat{r}(U_{i+l+k_1}) - \widehat{r}(U_{i+l})] &= \mathrm{Cov}[\widehat{r}(U_{i+j+k_1}), \widehat{r}(U_{i+l+k_1})] \\ &\quad - \mathrm{Cov}[\widehat{r}(U_{i+j}), \widehat{r}(U_{i+l+k_1})] \\ &\quad - \mathrm{Cov}[\widehat{r}(U_{i+j+k_1}), \widehat{r}(U_{i+l})] \\ &\quad + \mathrm{Cov}[\widehat{r}(U_{i+j}), \widehat{r}(U_{i+l})]\end{aligned}$$

The first and fourth covariances are $\sigma^2_{\widehat{r},i+j+k_1}$ and $\sigma^2_{\widehat{r},i+j}$ respectively when $j = l$, the second covariance is $\sigma^2_{\widehat{r},i+l+k_1}$ when $j = l + k_1$, and the third covariance is $\sigma^2_{\widehat{r},i+l}$ when $j + k_1 = l$. The other covariances can be obtained in a similar fashion. Now, using the weights in Eq. (3.5), $\sigma^2_{\widehat{r}} = O(n^{-(1-\beta)})$, and Lemma 2.1, where $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$,

$$\begin{aligned}\mathrm{Var}[\hat{Y}^{(2)}_{i,LRF}|\tilde{\mathbb{U}}] = \frac{4}{n^{1-\beta}} \sum_{j=1}^{k_2} \Bigg\{ &\frac{w_{i,j,2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2} \\ &\times \left(\frac{2}{(U_{i+j+k_1} - U_{i+j})^2} + \frac{2}{(U_{i-j-k_1} - U_{i-j})^2} \right) \Bigg\} \\ - \frac{4}{n^{1-\beta}} \sum_{j=1}^{k_2-k_1} \Bigg\{ &w_{i,j,2} w_{i,j+k_1,2} \left(\frac{1}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right) \\ &\times \left(\frac{1}{U_{i+j+2k_1} + U_{i+j+k_1} - U_{i-j-2k_1} - U_{i-j-k_1}} \right) \\ &\times \left(\frac{1}{(U_{i+j+k_1} - U_{i+j})(U_{i+j+2k_1} - U_{i+j+k_1})} \right. \\ &\qquad \left. + \frac{1}{(U_{i-j-k_1} - U_{i-j})(U_{i-j-2k_1} - U_{i-j-k_1})} \right) \Bigg\} \\ - \frac{4}{n^{1-\beta}} \sum_{j=1+k_1}^{k_2} \Bigg\{ &w_{i,j,2} w_{i,j-k_1,2} \left(\frac{1}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right) \\ &\times \left(\frac{1}{U_{i+j} + U_{i+j-k_1} - U_{i-j} - U_{i-j+k_1}} \right) \\ &\times \left(\frac{1}{(U_{i+j+k_1} - U_{i+j})(U_{i+j} - U_{i+j-k_1})} \right.\end{aligned}$$

$$+\frac{1}{(U_{i-j-k_1}-U_{i-j})(U_{i-j}-U_{i-j+k_1})}\Bigg)\Bigg\}$$

$$\leq \frac{4}{n^{1-\beta}}\sum_{j=1}^{k_2}\frac{w_{i,j,2}}{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})^2}$$

$$\times\left(\frac{2}{(U_{i+j+k_1}-U_{i+j})^2}+\frac{2}{(U_{i-j-k_1}-U_{i-j})^2}\right)$$

$$=\frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}.$$

B.3 Proof of Cor. 3.1

For $k_1\to\infty$ and $k_2\to\infty$ as $n\to\infty$ and from Theorem 3.1,

$$\left|\text{Bias}[\hat{Y}_{i,LRF}^{(2)}|\tilde{\mathbb{U}}]\right| \leq \frac{\sup_{u\in[0,1]}|r^{(3)}(u)|}{n+1}$$

$$\times\frac{2\sum_{j=1}^{k_2}j^3+3k_1\sum_{j=1}^{k_2}j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2}j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2}j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2}j}\{1+o_p(1)\}$$

$$=O_p\left(\max\left\{\frac{k_1}{n},\frac{k_2}{n}\right\}\right)$$

and

$$\text{Var}[\widehat{Y}_{i,LRF}^{(2)}|\mathbb{U}]\leq\frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}$$

$$=O_p\left(\max\left\{\frac{n^{3+\beta}}{k_1^2k_2^3},\frac{n^{3+\beta}}{k_1^4k_2}\right\}\right)$$

Then, for $k_1\to\infty$ and $k_2\to\infty$ as $n\to\infty$ such that $n^{-1}k_1\to 0$, $n^{-1}k_2\to 0$, $n^{3+\beta}k_1^{-2}k_2^{-3}\to 0$, and $n^{3+\beta}k_1^{-4}k_2^{-1}\to 0$, the conditional bias and conditional variance tend to zero. Therefore,

$$\lim_{n\to\infty}\text{MSE}[\widehat{Y}_i^{(2)}|\tilde{\mathbb{U}}]=0.$$

Use Chebyshev's inequality to complete the proof.

B.4 Proof of Cor. 3.2

Using the bias-variance decomposition of means squared error (MSE), the MSE is bounded above by

$$\begin{aligned}\mathrm{MSE}[\widehat{Y}^{(1)}_{i,LRF}|\tilde{\mathbb{U}}] \leq & \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 \{1+o_p(1)\} \\ & + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}\end{aligned}$$

Then, the conditional mean integrated squared error (MISE), is

$$\begin{aligned}\mathrm{MISE}[\widehat{Y}^{(2)}_{ML}|\tilde{\mathbb{U}}] &= \mathbb{E}\int_0^1 \left(\widehat{Y}^{(2)}_{LRF}(U) - r^{(2)}(U)|\mathbb{U}\right)^2 dU \\ &= \int_0^1 \mathbb{E}\left(\widehat{Y}^{(2)}_{LRF}(U) - r^{(2)}(U)|\mathbb{U}\right)^2 dU \\ &\leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 \{1+o_p(1)\} \\ &\quad + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}\end{aligned}$$

Therefore, the asymptotic conditional MISE (AMISE) is

$$\begin{aligned}\mathrm{AMISE}[\widehat{Y}^{(2)}_{LRF}|\tilde{\mathbb{U}}] \leq & \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 \\ & + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}.\end{aligned}$$

Appendix C

Chapter 4 Appendix

C.1 Proof of Theorem 4.1

First, we note that the GKRLS estimator is a linear smoother by substituting Eq. (4.10) into Eq. (4.11)

$$\begin{aligned}
\widehat{m}_2(\mathbf{x}_0) &= \sum_{i=1}^{n} \widehat{c}_{2,i} K_{\sigma_2}(\mathbf{x}_i, \mathbf{x}_0) \\
&= K_{\sigma_2,\mathbf{x}_0}^{*\top} \widehat{\mathbf{c}}_2 \\
&= K_{\sigma_2,\mathbf{x}_0}^{*\top} (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \mathbf{y} \\
&= L(\mathbf{x}_0)^{\top} \mathbf{y},
\end{aligned}$$

where $L(\mathbf{x}_0) = \left[K_{\sigma_2,\mathbf{x}_0}^{*\top}(\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2\mathbf{I})^{-1}\Omega^{-1}\right]^{\top}$ and $K_{\sigma_2,\mathbf{x}_0}^{*} = (K_{\sigma_2}(\mathbf{x}_1, \mathbf{x}_0), \ldots, K_{\sigma_2}(\mathbf{x}_n, \mathbf{x}_0))^{\top}$ the kernel vector evaluated at point $\mathbf{x}_0$.

Then, the conditional mean and variance of GKRLS can be derived as follows

$$\begin{aligned}\mathbb{E}[\widehat{m}_2|X=\mathbf{x}_0] &= L(\mathbf{x}_0)^\top \mathbb{E}[\mathbf{y}|\mathbf{X}] \\ &= L(\mathbf{x}_0)^\top \mathbf{m}\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}[\widehat{m}_2(\mathbf{x}_0)|X=\mathbf{x}_0] &= L(\mathbf{x}_0)^\top \operatorname{Var}[\mathbf{y}|\mathbf{X}]\, L(\mathbf{x}_0) \\ &= L(\mathbf{x}_0)^\top \Omega L(\mathbf{x}_0).\end{aligned}$$

C.2 Proof of Theorem 4.2

The exact bias for GKRLS for the training data is given by

$$\mathbb{E}[\widehat{\mathbf{m}}_2|X=\mathbf{x}] - \mathbf{m} = (\mathbf{L}-\mathbf{I})\mathbf{m},$$

and observe that the residuals are obtained by

$$\begin{aligned}\widehat{\mathbf{u}}_2 &= \mathbf{y} - \widehat{\mathbf{m}}_2 \\ &= \mathbf{y} - \mathbf{L}\mathbf{y} \\ &= (\mathbf{I}-\mathbf{L})\mathbf{y}.\end{aligned}$$

And the expectation of the residuals is given by

$$\begin{aligned}\mathbb{E}[\widehat{\mathbf{u}}_2|X=\mathbf{x}] &= \mathbf{m} - \mathbf{L}\mathbf{m} \\ &= -\operatorname{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}].\end{aligned}$$

[12] suggests estimating the conditional bias by smoothing the negative residuals

$$\begin{aligned}\widehat{\text{Bias}}[\widehat{\mathbf{m}}_2|\mathbf{X}] &= -\mathbf{L}\widehat{\mathbf{u}}_2 \\ &= -\mathbf{L}(\mathbf{I}-\mathbf{L})\mathbf{y} \\ &= (\mathbf{L}-\mathbf{I})\widehat{\mathbf{m}}_2.\end{aligned}$$

Therefore, the conditional bias can be estimated at any point $\mathbf{x}_0$ by

$$\widehat{\text{Bias}}[\widehat{m}_2(\mathbf{x}_0)|X=\mathbf{x}_0] = L(\mathbf{x}_0)^\top \widehat{\mathbf{m}} - \widehat{m}_2(\mathbf{x}_0)$$

For the conditional variance, we assume that the error covariance matrix $\Omega = \Omega(\theta)$ can be consistently estimated by $\widehat{\Omega} = \widehat{\Omega}(\widehat{\theta})$. Then, using a consistent estimator of the error covariance matrix, the conditional variance of GKLRS can be estimated by

$$\widehat{\text{Var}}[\widehat{m}_2(\mathbf{x}_0)|X=\mathbf{x}_0] = L(\mathbf{x}_0)^\top \widehat{\Omega} L(\mathbf{x}_0).$$

C.3 Proof of Theorem 4.3

Since the bias corrected fitted values, $\widehat{\mathbf{m}}_c$, have zero conditional bias, we can focus on the conditional variance. From Theorem 4.1, the conditional variance of the GKRLS estimator is

$$\begin{aligned}\text{Var}[\widehat{\mathbf{m}}_2|\mathbf{X}] &= \mathbf{L}\Omega\mathbf{L}^\top \\ &= \mathbf{L}PP^\top\mathbf{L}^\top \\ &= \mathbf{L}P(\mathbf{L}P)^\top \\ &= \mathbf{A}\mathbf{A}^\top,\end{aligned}$$

where $\mathbf{A} \equiv \mathbf{L}P$. Consider the singular value decomposition of $\mathbf{A}$, where $\mathbf{D}$, $\mathbf{U}$, $\mathbf{V}$ are the singular values, left singular vectors, and right singular vectors respectively.

$$\begin{aligned}
\text{Var}[\widehat{\mathbf{m}}_2|\mathbf{X}] &= \mathbf{A}\mathbf{A}^\top \\
&= \mathbf{U}\mathbf{D}\mathbf{V}(\mathbf{U}\mathbf{D}\mathbf{V})^\top \\
&= \mathbf{U}\mathbf{D}^2\mathbf{U}^\top \\
&= \mathbf{U}\begin{pmatrix} d_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_n^2 \end{pmatrix}\mathbf{U}^\top,
\end{aligned}$$

where $d_i, i = 1, \dots, n$ denotes the ith diagonal element of $\mathbf{D}$, i.e. the ith singular value of $\mathbf{L}P$. To examine the sum of the variances of $\widehat{\mathbf{m}}_2$, the trace of the variance matrix is evaluated.

$$\begin{aligned}
\text{tr}(\text{Var}[\widehat{\mathbf{m}}_2|\mathbf{X}]) &= \text{tr}(\mathbf{U}\mathbf{D}^2\mathbf{U}^\top) \\
&= \text{tr}(\mathbf{D}^2\mathbf{U}^\top\mathbf{U}) \\
&= \text{tr}(\mathbf{D}^2) \\
&= \sum_i^n d_i^2.
\end{aligned}$$

For large enough n, $\text{tr}(\mathbf{D}^2)$ slows in growth and converges to some constant, M, and the average variance of $\widehat{m}(\mathbf{x}_i)$ is $\frac{1}{n}\sum_{i=1}^n d_i^2$. Recall that d_i^2 denotes the ith squared singular value of $\mathbf{L}P$ and is proportional to the variance explained by a given singular vector of $\mathbf{L}P$. Given the construction of $\mathbf{L}P$, the columns of this product matrix can be thought of as weights of the data, scaled by the standard deviation of the error term. Therefore, the number of large singular values will grow initially with n but the number of important dimensions or

singular values will start to grow slowly with n. As a result, the average variance of $\widehat{m}(\mathbf{x}_i)$, which is $\frac{1}{n}\sum_{i=1}^{n} d_i^2$, shrinks to zero as $n \to \infty$. Since the average variance shrinks to zero, then each individual variance must approach zero as n becomes large.

C.4 Proof of Theorem 4.4

Consider the difference between the bias corrected fitted values and the true values, $\widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] - \mathbf{m}$, where $\text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] = \mathbf{L}\mathbf{m} - \mathbf{m}$,

$$\widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] - \mathbf{m} = \mathbf{L}\mathbf{u}$$

Note that $\text{E}[\mathbf{L}\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $\text{Var}[\mathbf{L}\mathbf{u}|\mathbf{X}] = \mathbf{L}\Omega\mathbf{L}^\top$. The following results will be for the case of heteroskedastic errors, where observations are independent and heterogenesously distributed. Consider the individual variances for each observation,

$$\text{Var}[L(\mathbf{x}_i)u_i|\mathbf{X}] = L(\mathbf{x}_i)^\top \Omega L(\mathbf{x}_i)$$

and let s_n^2 be the sum of the variances,

$$s_n^2 = \sum_{i=1}^{n} L(\mathbf{x}_i)^\top \Omega L(\mathbf{x}_i).$$

As long as the sum is not dominated by any particular term and if $L(\mathbf{x}_i)u_i$ are independent vectors distributed with mean $\mathbf{0}$ and variance $L(\mathbf{x}_i)^\top \Omega L(\mathbf{x}_i) < \infty$ and $s_n^2 \to \infty$ as $n \to \infty$, then

$$\mathbf{L}\mathbf{u} \xrightarrow{d} N(\mathbf{0}, \mathbf{L}\Omega\mathbf{L}^\top),$$

by Lindeberg-Feller central limit theorem. It then follows that

$$\widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] - \mathbf{m} \xrightarrow{d} N(\mathbf{0}, \mathbf{L}\Omega\mathbf{L}^\top).$$

The following results will be for the case of autocorrelated errors, where observations are dependent and identically distributed.[1] Given (i) $Y_t = m(\mathbf{X}_t) + u_t, t = 1, 2, \ldots$; (ii) $\{(\mathbf{X}_t, u_t)\}$ is a stationary ergodic sequence; (iii) (a) $\{L(X_{thi})u_{th}, \mathcal{F}_t\}$ is an adapted mixingale of size -1, $h = 1, \ldots, p, i = 1, \ldots, n$; (b) $\mathbb{E}|L(X_{thi})u_{th}|^2 < \infty, h = 1, \ldots, p, i = 1, \ldots, n$; (c) $\mathbf{V}_n \equiv \text{Var}(\sum_{t=1}^n L(\mathbf{X}_t)u_t)$ is uniformly positive definite; (iv) $\mathbb{E}|L(X_{thi})|^2 < \infty, h = 1, \ldots, p, i = 1, \ldots, n$.

Consider $\sum_{t=1}^n \boldsymbol{\lambda}^\top \mathbf{V}^{-1/2} L(\mathbf{X}_t)u_t$, where $\mathbf{V}$ is any finite positive definite matrix. By Theorem 3.35 of [44], $\{Z_t, \mathcal{F}_t\}$ is an adapted stochastic sequence because Z_t is measurable with respect to $\mathcal{F}_t$. To see that $\mathbb{E}(Z_t^2) < \infty$, note that we can write

$$\begin{aligned} Z_t &= \boldsymbol{\lambda}^\top \mathbf{V}^{-1/2} L(\mathbf{X}_t)u_t \\ &= \sum_{h=1}^p \boldsymbol{\lambda}^\top \mathbf{V}^{-1/2} L(\mathbf{X}_{th})u_{th} \\ &= \sum_{h=1}^p \sum_{i=1}^n \tilde{\lambda}_i L(X_{thi})u_{th}, \end{aligned}$$

where $\tilde{\lambda}_i$ is the ith element of the $n \times 1$ vector $\tilde{\boldsymbol{\lambda}} \equiv \mathbf{V}^{-1/2}\boldsymbol{\lambda}$. By definition of $\boldsymbol{\lambda}$ and $\mathbf{V}$, there exists $\Delta < \infty$ such that $|\tilde{\lambda}_i| < \Delta$ for all i. It follows from Minkowski's inequality that

$$\begin{aligned} \mathbb{E}(Z_t^2) &\leq \left[\sum_{h=1}^p \sum_{i=1}^n \left(\mathbb{E}|\tilde{\lambda}_i L(X_{thi})u_{th}|^2\right)^{1/2}\right]^2 \\ &\leq \left[\Delta \sum_{h=1}^p \sum_{i=1}^n \left(\mathbb{E}|L(X_{thi})u_{th}|^2\right)^{1/2}\right]^2 \\ &\leq [\Delta p n \Delta^{1/2}]^2 \leq \infty, \end{aligned}$$

since for Δ sufficiently large, $\mathbb{E}|L(X_{thi})u_{th}|^2 < \Delta < \infty$ given $(iii.b)$ and the stationarity assumption. Next, we show $\{Z_t, \mathcal{F}_t\}$ is a mixingale of size -1. Using the expression for Z_t

[1] We follow the proof similar to the case of dependent identially distributed observations provided by [44].

just given, we can write

$$\begin{aligned}\mathbb{E}([\mathbb{E}(Z_0|\mathcal{F}_{-m})]^2) &= \mathbb{E}\left(\left[\mathbb{E}\left(\sum_{h=1}^{p}\sum_{i=1}^{n}\tilde{\lambda}_i L(X_{0hi})u_{0h}|\mathcal{F}_{-m}\right)\right]^2\right) \\ &= \mathbb{E}\left(\left[\sum_{h=1}^{p}\sum_{i=1}^{n}\mathbb{E}\left(\tilde{\lambda}_i L(X_{0hi})u_{0h}|\mathcal{F}_{-m}\right)\right]^2\right).\end{aligned}$$

Applying Minkowski's inequality, it follows that

$$\begin{aligned}\mathbb{E}([\mathbb{E}(Z_0|\mathcal{F}_{-m})]^2) &\leq \left[\sum_{h=1}^{p}\sum_{i=1}^{n}\left(\mathbb{E}\left[\mathbb{E}\left(\tilde{\lambda}_i L(X_{0hi})u_{0h}|\mathcal{F}_{-m}\right)^2\right]\right)^{1/2}\right]^2 \\ &\leq \left[\Delta\sum_{h=1}^{p}\sum_{i=1}^{n}\left(\mathbb{E}\left[\mathbb{E}(L(X_{0hi})u_{0h}|\mathcal{F}_{-m})^2\right]\right)^{1/2}\right]^2 \\ &\leq \left[\Delta\sum_{h=1}^{p}\sum_{i=1}^{n}c_{0hi}\gamma_{mhi}\right]^2 \\ &\leq [\Delta pn\bar{c}_0\bar{\gamma}_m]^2,\end{aligned}$$

where $\bar{c}_0 = \max_{h,i} c_{0hi} < \infty$ and $\bar{\gamma}_m = \max_{h,i}\gamma_{mhi}$ is of size -1. Thus, $\{Z_t, \mathcal{F}_t\}$ is a mixingale of size -1. Note that

$$\begin{aligned}\text{Var}(n\bar{Z}_n) &= \text{Var}\left(\sum_{t=1}^{n}\boldsymbol{\lambda}^\top \boldsymbol{V}^{-1/2}L(\mathbf{X}_t)u_t\right) \\ &= \boldsymbol{\lambda}^\top\mathbf{V}^{-1/2}\mathbf{V}_n\mathbf{V}^{-1/2}\boldsymbol{\lambda} \to \bar{\sigma}^2 < \infty.\end{aligned}$$

Hence $\mathbf{V}_n$ converges to a finite matrix. Set $\mathbf{V} = \lim_{n\to\infty}\mathbf{V}_n = \mathbf{L}\Omega\mathbf{L}^\top$ which is positive definitie given (iii.c). Then, $\bar{\sigma}^2 = \boldsymbol{\lambda}^\top\mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2}\boldsymbol{\lambda} = 1$. Then by the martingale central limit theorem, $\sum_{t=1}^{n}\boldsymbol{\lambda}^\top\mathbf{V}^{-1/2}L(\mathbf{X}_t)u_t \xrightarrow{d} N(0,1)$. Since this holds for every $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^\top\boldsymbol{\lambda} = 1$, it follows from Cramér-Wold Theorem, that $\mathbf{V}^{-1/2}\sum_{t=1}^{n}L(\mathbf{X}_t)u_t \xrightarrow{d} N(\mathbf{0},\mathbf{I})$. Hence, $\mathbf{Lu} \xrightarrow{d} N(\mathbf{0},\mathbf{L}\Omega\mathbf{L}^\top)$ and it then follows that

$$\widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] - \mathbf{m} \xrightarrow{d} N(\mathbf{0},\mathbf{L}\Omega\mathbf{L}^\top).$$

C.5 Proof of Theorem 4.5

First, we note that the GKRLS derivative estimator is a linear smoother by substituting Eq. (4.7) and Eq. (4.10) into Eq. (4.28),

$$\begin{aligned}
\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) &= \frac{2}{\sigma_2^2}\sum_{i=1}^{n} \mathrm{e}^{-\frac{1}{\sigma_2^2}||\mathbf{x}_i-\mathbf{x}_0||^2}(\mathbf{x}_i^{(r)}-\mathbf{x}_0^{(r)})\widehat{c}_{2,i} \\
&= K_{\sigma_2,\mathbf{x}_0}^{*\top}\Delta_r\widehat{\mathbf{c}}_2 \\
&= K_{\sigma_2,\mathbf{x}_0}^{*\top}\Delta_r(\Omega^{-1}\mathbf{K}_{\sigma_2}+\lambda_2\mathbf{I})^{-1}\Omega^{-1}\mathbf{y} \\
&= S_r(\mathbf{x}_0)^\top\mathbf{y},
\end{aligned}$$

where $\Delta_r \equiv \frac{2}{\sigma_2^2}\,\mathrm{diag}(\mathbf{x}_1^{(r)}-\mathbf{x}_0^{(r)},\ldots,\mathbf{x}_n^{(r)}-\mathbf{x}_0^{(r)})$ is a $n\times n$ diagonal matrix and

$$S_r(\mathbf{x}_0) = \left[K_{\sigma_2,\mathbf{x}_0}^{*\top}\Delta_r(\Omega^{-1}\mathbf{K}_{\sigma_2}+\lambda_2\mathbf{I})^{-1}\Omega^{-1}\right]^\top \tag{C.1}$$

is the smoother vector for the first partial derivative with respect to the rth variable. Then, the conditional mean and variance of the GKRLS can be derived as follows

$$\begin{aligned}
\mathbb{E}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] &= S_r(\mathbf{x}_0)^\top\mathbb{E}[\mathbf{y}|\mathbf{X}] \\
&= S_r(\mathbf{x}_0)^\top\mathbf{m}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] &= S_r(\mathbf{x}_0)^\top\,\mathrm{Var}[\mathbf{y}|\mathbf{X}]\,S_r(\mathbf{x}_0) \\
&= S_r(\mathbf{x}_0)^\top\Omega S_r(\mathbf{x}_0).
\end{aligned}$$

C.6 Proof of Theorem 4.6

The bias of the GKRLS derivative estimator in Eq. (4.28)

$$\begin{aligned}\text{Bias}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] &= S_r(\mathbf{x}_0)^\top \mathbb{E}[\mathbf{y}|\mathbf{X}] - m_r^{(1)}(\mathbf{x}_0) \\ &= S_r(\mathbf{x}_0)^\top \mathbf{m} - m_r^{(1)}(\mathbf{x}_0),\end{aligned}$$

where $m_r^{(1)}(\mathbf{x}_0)$ is the true first partial derivative of m with respect to the rth variable. Since this quantity as well as $\mathbf{m}$ is unknown, we estimate both to calculate the conditional bias.

$$\widehat{\text{Bias}}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\mathbf{m}}_2 - \widehat{m}_{2,r}^{(1)}(\mathbf{x}_0),$$

where $\widehat{\mathbf{m}}_2$ is the $n \times 1$ vector of in sample GKRLS predictions of $\mathbf{m}$ and $\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)$ is the estimated GKRLS derivative prediction evaluated at point $\mathbf{x}_0$.

For the conditional variance, we assume that the error covariance matrix $\Omega = \Omega(\theta)$ can be consistently estimated by $\widehat{\Omega} = \widehat{\Omega}(\widehat{\theta})$. Then, using a consistent estimator of the error covariance matrix, the conditional variance of GKLRS can be estimated by

$$\widehat{\text{Var}}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\Omega} S_r(\mathbf{x}_0) \tag{C.2}$$