Generalized Kernel Regularized Least Squares Estimator with Parametric Error Covariance

Justin Dang* and Aman Ullah†

Abstract

A two-step estimator of a nonparametric regression function via KRLS with parametric error covariance is proposed. 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. An empirical example is illustrated with estimating an airline cost function with heteroskedastic errors. The derivatives are evaluated, and the average partial effects of the inputs are determined in the applications.

*Department of Economics, University of California, Riverside. Email: jdang015@ucr.edu
†Department of Economics, University of California, Riverside. Email: aman.ullah@ucr.edu

1 Introduction

Peter Schmidt has made many seminal contributions in advancing the statistical inference methods and their applications in time series, cross section, and panel data econometrics in general (Schmidt, 1976a) and, in particular, in the areas of dynamic econometric models, estimation and testing of cross-sectional and panel data models, crime and justice models (Schmidt and Witte, 1984), survival models (Schmidt and Witte, 1988). His fundamental and innovative contributions on the econometrics of stochastic frontier production/cost models have made significant impact on the generations of econometricians (e.g., Schmidt (1976b), Aigner et al. (1977), Amsler et al. (2017), Amsler et al. (2019)). Also, he has contributed many influential papers on developing efficient procedures involving the generalized least squares (GLS) method (see Guilkey and Schmidt (1973), Schmidt (1977), Arabmazar and Schmidt (1981), Ahu and Schmidt (1995)) among others. These were for the parametric models, whereas here we consider the nonparametric models.

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 paper, we will focus on KRLS; this method is also known as Kernel Ridge Regression (KRR) in the machine learning literature and is the kernelized version of the simple ridge regression to allow for nonlinearities in the model.

In this paper, 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 Hainmueller and Hazlett (2014). However, Hainmueller and Hazlett (2014) only consider 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 De Brabanter et al. (2011), allowing

for heteroskedastic errors. Although not directly stated, the LSSVM model in De Brabanter et al. (2011) is equivalent to KRR/KRLS when an intercept term is included in the model. Following Hainmueller and Hazlett (2014), we will use KRLS without an intercept. Although De Brabanter et al. (2011) allow 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 paper, 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. 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 paper is as follows: Section 2 discusses the model framework and the GKRLS estimator, Section 3, Section 4, and Section 5 show the finite sample properties, asymptotic properties, and partial effects and derivatives of the GKRLS estimator, respectively, Section 6 runs through a simulation example, Section 7 illustrates an empirical example with heteroskedastic errors, and Section 8 concludes the paper.

2 Generalized KRLS Estimator

Consider the nonparametric regression model:

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

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

$$\mathbb{E}[U_i U_j] = \omega_{ij}(\theta_0) \text{ for some } \theta_0 \in \mathbb{R}^p, i, j = 1, \ldots, n. \tag{2}$$

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

2.1 Naive KRLS Estimator

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

$$m(\mathbf{x}_0) = \sum_{i=1}^{n} c_i K_\sigma(\mathbf{x}_i, \mathbf{x}_0), \tag{3}$$

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,

$$K_\sigma(\mathbf{x}_i, \mathbf{x}_0) = \mathrm{e}^{-\frac{1}{\sigma^2}||\mathbf{x}_i - \mathbf{x}_0||^2}. \tag{4}$$

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 (Hainmueller and Hazlett, 2014), 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

$$\arg\min_{\mathbf{c}} (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c})^\top (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c}) + \lambda \mathbf{c}^\top \mathbf{K}_\sigma \mathbf{c}, \tag{5}$$

where $\mathbf{y}$ is the vector of training data corresponding to the dependent variable, $\mathbf{K}_\sigma$ 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

$$\widehat{\mathbf{c}}_1 = (\mathbf{K}_{\sigma_1} + \lambda_1 \mathbf{I})^{-1}\mathbf{y}. \tag{6}$$

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

$$\widehat{m}_1(\mathbf{x}_0) = \sum_{i=1}^{n} \widehat{c}_{1,i} K_{\sigma_1}(\mathbf{x}_i, \mathbf{x}_0). \tag{7}$$

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. (1) by P^{-1}, where $P^{-1} = P^{-1}(\theta)$ and we condense the notation and the dependence on θ is implied.

$$P^{-1}\mathbf{y} = P^{-1}\mathbf{m} + P^{-1}\mathbf{U}. \tag{8}$$

The transformed error term, $P^{-1}U$ has mean $\mathbf{0}$ 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

$$\arg\min_{\mathbf{c}} (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c})^\top \Omega^{-1} (\mathbf{y} - \mathbf{K}_\sigma \mathbf{c}) + \lambda \mathbf{c}^\top \mathbf{K}_\sigma \mathbf{c}. \tag{9}$$

The solution for vector is

$$\hat{\mathbf{c}}_2 = (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2\mathbf{I})^{-1}\Omega^{-1}\mathbf{y} \tag{10}$$

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 $\widehat{\Omega} = \widehat{\Omega}(\hat{\theta})$.

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

$$\widehat{m}_2(\mathbf{x}_0) = \sum_{i=1}^{n} \widehat{c}_{2,i} K_{\sigma_2}(\mathbf{x}_i, \mathbf{x}_0) \tag{11}$$

The two step procedure is outlined below

1. Estimate Eq. (1) by naive KRLS from Eq. (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. (8) by KRLS under the transformed variables as in Eq. (9) and Eq. (11). Denote these estimates as GKRLS.

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.

3.1 Estimation of Bias and Variance

In this subsection, we estimate the bias and variance of the two step estimator. Following, De Brabanter et al. (2011), notice that the GKRLS estimator is a linear smoother.

Defintion 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*

$$\widehat{m}(\mathbf{x}_0) = \sum_{i=1}^{n} l_i(\mathbf{x}_0) Y_i, \tag{12}$$

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

For in sample data, Eq. (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 paper, 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 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. (1).*

Assumption 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 1, Assumption 1, and Assumption 2, the conditional mean and variance can be obtained by the following theorem.

Theorem 1. *The GKRLS estimator in Eq. (11) is*

$$\begin{aligned} \widehat{m}_2(\mathbf{x}_0) &= \sum_{i=1}^{n} l_i(\mathbf{x}_0) Y_i \\ &= L(\mathbf{x}_0)^\top \mathbf{y}, \end{aligned} \tag{13}$$

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

$$L(\mathbf{x}_0) = \left[K_{\sigma_2, \mathbf{x}_0}^{*\top} (\Omega^{-1} \mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \right]^\top, \tag{14}$$

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. (1)*, has conditional mean*

$$\mathbb{E}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \mathbf{m} \tag{15}$$

and conditional variance

$$\mathrm{Var}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \Omega L(\mathbf{x}_0). \tag{16}$$

Proof: see Appendix A.

From Theorem 1, the conditional bias can be written as

$$\begin{aligned} \mathrm{Bias}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] &= \mathbb{E}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}] - m(\mathbf{x}_0) \\ &= L(\mathbf{x}_0)^\top \mathbf{m} - m(\mathbf{x}_0) \end{aligned} \tag{17}$$

Following De Brabanter et al. (2011), we will estimate the conditional bias and variance by the following:

Theorem 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*

$$\widehat{\mathrm{Bias}}[\widehat{m}_2(\mathbf{x}_2)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \widehat{\mathbf{m}}_2 - \widehat{m}_2(\mathbf{x}_0) \tag{18}$$

and

$$\widehat{\mathrm{Var}}[\widehat{m}_2(\mathbf{x}_0)|X = \mathbf{x}_0] = L(\mathbf{x}_0)^\top \widehat{\Omega} L(\mathbf{x}_0). \tag{19}$$

Proof: See Appendix B.

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. (18) and Eq. (19), except that the linear smoother vector $L(\mathbf{x}_0)$ will be different. Let

$$L_1(\mathbf{x}_0) = \left[K^{*\top}_{\sigma_1,\mathbf{x}_0}(\mathbf{K}_{\sigma_1} + \lambda_1 \mathbf{I})^{-1}\right]^\top \tag{20}$$

be the smoother vector for KRLS. Then, Eq. (7) can be rewritten as

$$\widehat{m}_1(\mathbf{x}_0) = L_1(\mathbf{x}_0)^\top \mathbf{y}. \tag{21}$$

Using Theorem 1 and Theorem 2 and applying them to the KRLS estimator, the estimated conditional bias and variance of KRLS are

$$\widehat{\text{Bias}}[\widehat{m}_1(\mathbf{x}_0)|X = \mathbf{x}_0] = L_1(\mathbf{x}_0)^\top \widehat{\mathbf{m}}_1 - \widehat{m}_1(\mathbf{x}_0) \tag{22}$$

$$\widehat{\text{Var}}[\widehat{m}_1(\mathbf{x}_0)|X = \mathbf{x}_0] = L_1(\mathbf{x}_0)^\top \widehat{\Omega} L_1(\mathbf{x}_0), \tag{23}$$

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

$$\widehat{\mathbf{m}}_{2,c} = \widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}], \tag{24}$$

then

$$\lim_{n\to\infty} \text{Var}[\widehat{\mathbf{m}}_{2,\text{c}}|\mathbf{X}] = 0 \tag{25}$$

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.

The estimated conditional bias from Eq. (18) and conditional variance from Eq. (19) can be used to construct pointwise confidence intervals. Asymptotic normality of the proposed estimator is given via the central limit theorem.

Theorem 4. *Under Assumptions 1 to 3,* $\widehat{\mathbf{m}}_2$ *is asymptotically normal by the central limit theorem:*

$$\widehat{\mathbf{m}}_2 - \text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] - \mathbf{m} \xrightarrow{d} N(\mathbf{0}, \text{Var}[\widehat{\mathbf{m}}_2|\mathbf{X}]), \tag{26}$$

where $\text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] = \mathbf{L}\mathbf{m} - \mathbf{m}$ *and* $\text{Var}[\widehat{\mathbf{m}}_2|\mathbf{X}] = \mathbf{L}\Omega\mathbf{L}^\top$.
Proof: See Appendix D.

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. (18) and (19) as estimates and can conduct approximate bias corrected $100(1-\alpha)\%$ pointwise confidence intervals from Theorem 4 as

$$\widehat{m}_2(\mathbf{x}_i) - \widehat{\text{Bias}}[\widehat{m}_2(\mathbf{x}_i)|X = \mathbf{x}_i] \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}[\widehat{m}_2(\mathbf{x}_i)|X = \mathbf{x}_i]} \tag{27}$$

for all i. Furthermore, 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.

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

$$\begin{aligned}\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) &= \sum_{i=1}^{n} \frac{\partial K_{\sigma_2}(\mathbf{x}_i, \mathbf{x}_0)}{\partial \mathbf{x}_0^{(r)}} \widehat{c}_{2,i} \\ &= \frac{2}{\sigma_2^2} \sum_{i=1}^{n} \mathrm{e}^{-\frac{1}{\sigma_2^2}||\mathbf{x}_i - \mathbf{x}_0||^2} (\mathbf{x}_i^{(r)} - \mathbf{x}_0^{(r)}) \widehat{c}_{2,i},\end{aligned} \tag{28}$$

using the RBF kernel in Eq. (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 5. *The GKRLS derivative estimator in Eq.* (28) *with the RBF kernel in Eq.* (4) *can be rewritten as*

$$\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) = S_r(\mathbf{x}_0)^\top \mathbf{y}, \tag{29}$$

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*

$$S_r(\mathbf{x}_0) = \left[K_{\sigma_2, \mathbf{x}_0}^{*\top} \Delta_r (\Omega^{-1} \mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \right]^\top \tag{30}$$

is the smoother vector for the first partial derivative with respect to the r*th variable. Then, the conditional mean of the GKRLS derivative estimator is*

$$\mathbb{E}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) | X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \mathbf{m} \tag{31}$$

and conditional variance is

$$\operatorname{Var}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0) | X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \Omega S_r(\mathbf{x}_0). \tag{32}$$

Proof: see Appendix E.

Using Theorem 5, the conditional bias and variance can be estimated as follows

Theorem 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.* (28) *are obtained by*

$$\widehat{\text{Bias}}[\widehat{m}^{(1)}_{2,r}(\mathbf{x}_0)|X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\mathbf{m}} - \widehat{m}^{(1)}_{2,r}(\mathbf{x}_0) \tag{33}$$

and

$$\widehat{\text{Var}}[\widehat{m}^{(1)}_{2,r}(\mathbf{x}_0)|X = \mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\Omega} S_r(\mathbf{x}_0). \tag{34}$$

Proof: See Appendix F.

The average partial derivative with respect to the rth variable is

$$\widehat{m}^{(1)}_{avg,r} = \frac{1}{n'} \sum_{j=1}^{n'} \widehat{m}^{(1)}_{2,r}(\mathbf{x}_{0,j}) \tag{35}$$

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

$$\text{Bias}[\widehat{m}^{(1)}_{avg,r}|X] = \frac{1}{n'} \boldsymbol{\iota}_{n'}^\top \mathbf{S}_{0,r} \mathbf{m} - \frac{1}{n'} \boldsymbol{\iota}_{n'}^\top \mathbf{m}^{(1)}_{0,r} \tag{36}$$

and

$$\text{Var}[\widehat{m}^{(1)}_{avg,r}|X] = \frac{1}{n'^2} \boldsymbol{\iota}_{n'}^\top \mathbf{S}_{0,r} \Omega \mathbf{S}_{0,r}^\top \boldsymbol{\iota}_{n'}, \tag{37}$$

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 $\mathbf{m}^{(1)}_{0,r}$ is the $n' \times 1$ vector of derivatives evaluated at each $\mathbf{x}_{0,j}, j = 1, \ldots, n'$.

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. (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:

$$m(X) = \sin(X) \tag{38}$$

and the derivative is given by

$$m^{(1)}(X) = \cos(X) \tag{39}$$

For the error terms, we consider two cases. First, 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 $N(0, 1)$. 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 De Brabanter et al. (2011). 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. De Brabanter et al. (2011) only consider 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 the bias corrected results are reported below in Figure 1, Figure 2, and in Table 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.

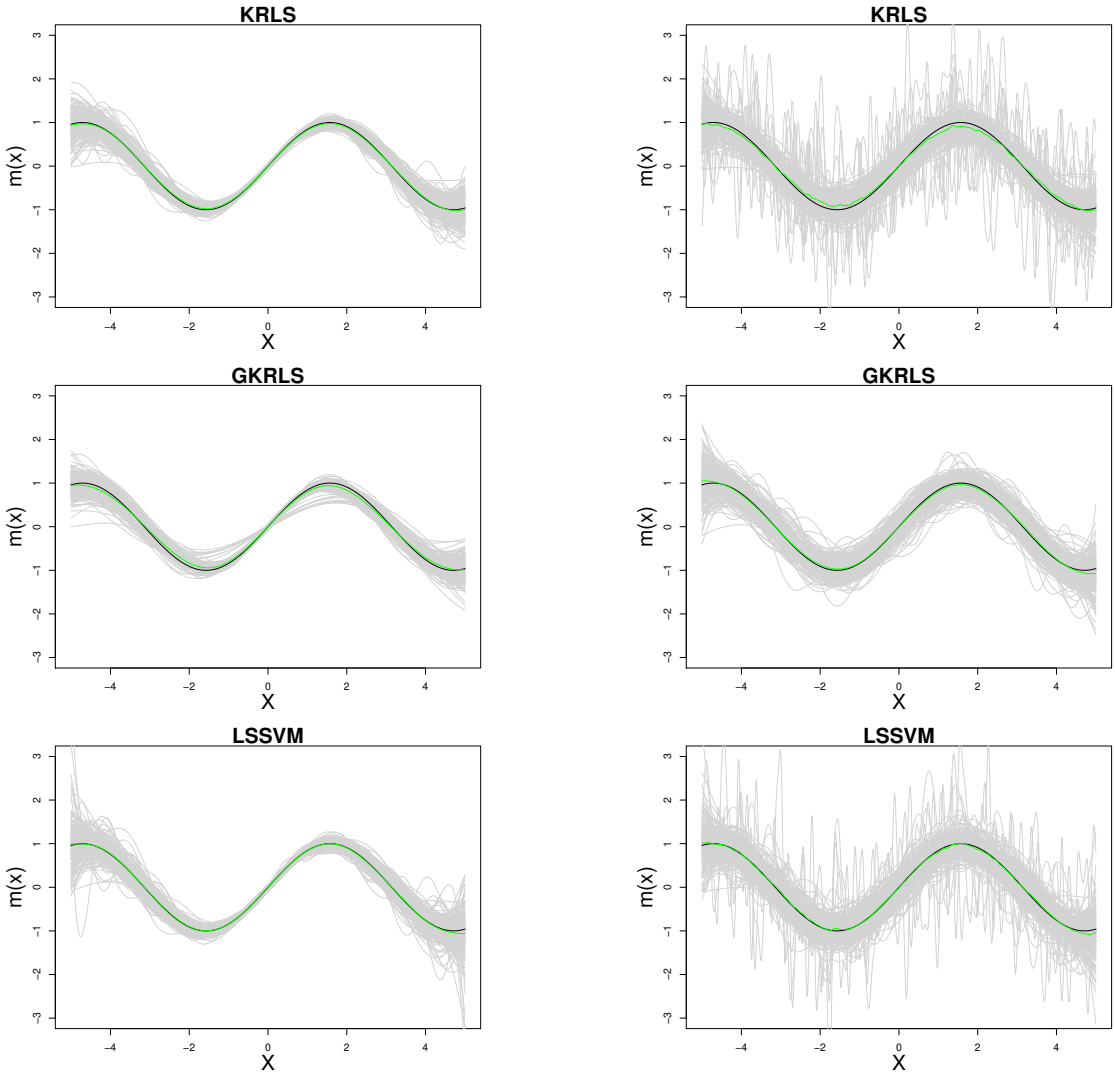

Figure 1: *The left (right) plots show the estimated bias corrected predictions under heteroskedastic errors (AR(2) errors). The top, middle, and bottom plots refer to the KRLS, GKRLS, and LSSVM estimators. The grey curves show the bias corrected predicted values from all simulations evaluated at the 500 evaluation data points spanning from -5 to 5. The green curves denote the average bias corrected predictions at each evaluation point across all simulations, and the black curve represents the true regression function in Eq.* (38).

[1]The following R packages were used for conducting simulations: Borchers (2021), Hyndman and Khandakar (2008), and McLeod et al. (2007).

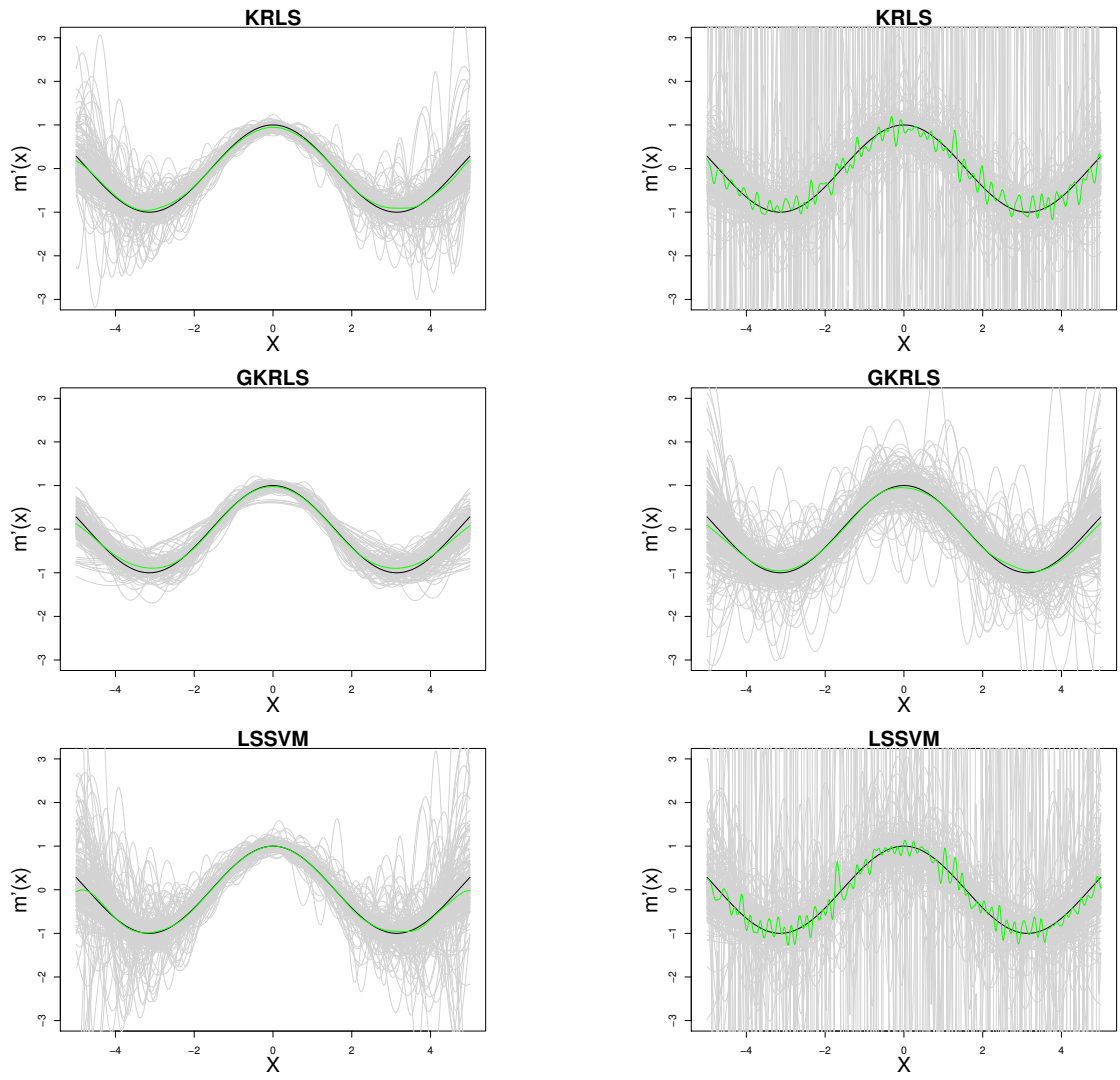

Figure 2: *The left (right) plots show the estimated bias corrected derivatives under heteroskedastic errors (AR(2) errors). The top, middle, and bottom plots refer to the KRLS, GKRLS, and LSSVM estimators. The grey curves show the bias corrected predicted values from all simulations evaluated at the 500 evaluation data points spanning from -5 to 5. The green curves denote the average bias corrected predicted derivative estimates at each evaluation point across all simulations, and the black curve represents the true derivative of the regression function in Eq.* (39).

Figure 1 shows simulation results under Eq. (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 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 2 shows simulation results for the derivative given in Eq. (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 compared 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 1 and Figure 2.

Table 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 1 are bias corrected and 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

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 1: *The table reports the bias, variance, and MSE for KRLS, GKRLS, and LSSVM estimators under Eq.* (38)*, Eq.* (39)*, and the cases of heteroskedastic and AR(2) errors. All reported estimates are bias corrected and are averaged across all simulations.*

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 2: *The table reports the bias, variance, and MSE for GKRLS estimator under Eq.* (38)*, Eq.* (39)*, and the cases of heteroskedastic and AR(2) errors for different sample sizes,* $n = 100, 200, 400$*. All reported estimates are biased corrected and are averaged across all simulations. All hyperparameters are fixed and set to 1.*

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.

Table 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 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.

7 Application

We implement an empirical application from the U.S. airline industry with heteroskedastic errors.[2] For the data set, we set 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

$$MSE = \frac{1}{n'} \sum_{j=1}^{n'} (\widehat{m}(\mathbf{x}_{0,j}) - y_j)^2, \tag{40}$$

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]

$$MSE_{boot} = \frac{1}{B} \frac{1}{n'} \sum_{b=1}^{B} \sum_{j=1}^{n'} (\widehat{m}_b(\mathbf{x}_{0,j}) - y_j)^2 \tag{41}$$

$$MSE_{boot,deriv} = \frac{1}{B} \frac{1}{n'} \sum_{b=1}^{B} \sum_{j=1}^{n'} (\widehat{m}_b^{(1)}(\mathbf{x}_{0,j}) - \widehat{m}_{avg}^{(1)}(\mathbf{x}_{0,j}))^2, \tag{42}$$

[2]The data for the application is from Greene (2018) and can be downloaded at https://pages.stern.nyu.edu/ wgreene/Text/Edition7/tablelist8new.htm

[3]The R package by Davison and Hinkley (1997) was used to obtain the bootstrap samples.

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 $\widehat{m}_{avg}^{(1)}(\cdot)$ is the simple average of $f = 1, \ldots, 4$ models (KRLS, GKRLS, LSSVM, and OLS):

$$\widehat{m}_{avg}^{(1)}(\mathbf{x}_{0,j}) = \frac{1}{4}\sum_{f=1}^{4} \widehat{m}_f^{(1)}(\mathbf{x}_{0,j}). \tag{43}$$

7.1 U.S. Airline Industry

We obtain the data on the efficiency in production of airline services from Greene (2018). To model heteroskedasticity we estimate GKRLS for the following:

$$\log C_{it} = m(\log Q_{it}, \log P_{it}) + U_{it}, \tag{44}$$

$$\omega_{it} = \exp(\gamma_1 + \gamma_2 Loadfactor_{it}), \tag{45}$$

where C_{it} is the total cost, Q_{it} is output, and P_{it} is the price of fuel, and $Loadfactor$ is 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 bias corrected results for the estimated regression function and its derivative for GKRLS and KRLS models in Figure 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

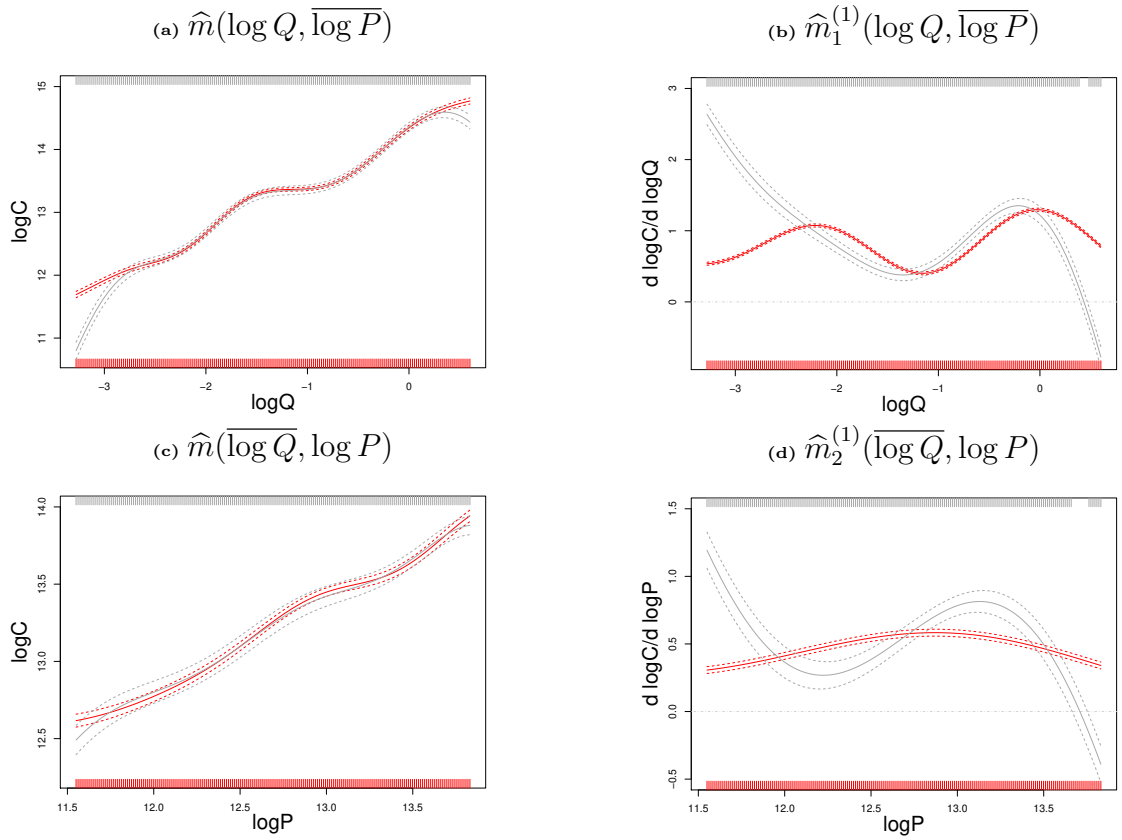

Figure 3: *The left (right) plots show the estimated bias corrected predictions of the regression function (derivative) for 200 evenly spaced points across the support for each independent variable. The predictions are plotted for* $\log Q$ *and* $\log P$*, holding all else fixed at their medians. The red (grey) curves correspond to the bias corrected predictions made by GKRLS (KRLS). The dashed lines are the bias corrected pointwise 95% confidence intervals.*

partial derivatives being positive almost everywhere. Accounting for heteroskedasticity in the estimation of the regression function, GKRLS is somewhat smoother 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.

The bias corrected average partial derivatives and corresponding standard errors are reported in Table 3. These averages are calculated by training each model on the 70 obser-

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 3: *Bias corrected 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 average partial derivative with respect to each regressor. The White standard errors are reported for the OLS model.*

vations 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 5 are used in our calculations. The reported estimates for GKRLS and KRLS correspond to the derivative plots in Figure 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 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

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

[4]The R package by Zeileis (2006) was used to obtain the White heteroskedastic standard erorrs.

Then, the partial derivatives are

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

$$\log P : \beta_3$$

Table 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.

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: *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. All reported estimates are bias corrected.*

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 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 5. For the partial derivative with respect to output, 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 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 Hsiao et al. (2007) 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 5: *The bootstrapped MSEs for the GKRLS, KRLS, LSSVM, and OLS partial derivatives are reported. The rows represent the MSE estimates of the partial derivative with respect to each regressor. All reported estimates are bias corrected.*

8 Conclusion

Overall, this paper 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. An empirical example is illustrated with estimating an airline cost function with heteroskedastic errors. The derivatives are evaluated, and the average partial effects of the inputs are determined in the application. In the empirical exercise, GKRLS shows different regression

function and derivative function estimates and is more efficient than KRLS.

Compliance with Ethical Standards

This research received no funding. Justin Dang declares that he has no conflict of interest. Aman Ullah declares that he has no conflict of interest. This article does not contain any studies with human participants performed by any of the authors.

References

Ahu, S. C. and Schmidt, P. "A separability result for gmm estimation, with applications to gls prediction and conditional moment tests." *Econometric Reviews*, 14(1):19–34, 1995. doi:10.1080/07474939508800301.

Aigner, D., Lovell, C., and Schmidt, P. "Formulation and estimation of stochastic frontier production function models." *Journal of Econometrics*, 6(1):21–37, 1977.

Amsler, C., Prokhorov, A., and Schmidt, P. "Endogenous environmental variables in stochastic frontier models." *Journal of Econometrics*, 199(2):131–140, 2017. ISSN 0304-4076. doi:https://doi.org/10.1016/j.jeconom.2017.05.005.

Amsler, C., Schmidt, P., and Tsay, W.-J. "Evaluating the cdf of the distribution of the stochastic frontier composed error." *Journal of Productivity Analysis*, 52(1-3):29–35, 2019. doi:10.1007/s11123-019-00554-9.

Arabmazar, A. and Schmidt, P. "Further evidence on the robustness of the tobit estimator to heteroskedasticity." *Journal of Econometrics*, 17(2):253–258, 1981. ISSN 0304-4076. doi:https://doi.org/10.1016/0304-4076(81)90029-4.

Borchers, H. W. *pracma: Practical Numerical Math Functions*, 2021. R package version 2.3.3.

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

De Brabanter, K., De Brabanter, J., Suykens, J. A. K., and De Moor, B. "Approximate confidence and prediction intervals for least squares support vector regression." *IEEE Transactions on Neural Networks*, 22(1):110–120, 2011. doi:10.1109/TNN.2010.2087769.

Greene, W. *Econometric Analysis*. Pearson, 2018. ISBN 9780134461366.

Guilkey, D. K. and Schmidt, P. "Estimation of seemingly unrelated regressions with vector autoregressive errors." *Journal of the American Statistical Association*, 68(343):642–647, 1973. ISSN 01621459.

Hainmueller, J. and Hazlett, C. "Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach." *Political Analysis*, 22(2):143–168, 2014. ISSN 10471987, 14764989.

Hsiao, C., Li, Q., and Racine, J. "A consistent model specification test with mixed discrete and continuous data." *Journal of Econometrics*, 140(2):802–826, 2007.

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

McLeod, A. I., Yu, H., and Krougly, Z. "Algorithms for linear time series analysis: With r package." *Journal of Statistical Software*, 23(5), 2007.

Schmidt, P. *Econometrics*. Marcel Dekker, Inc., New York, 1976a.

Schmidt, P. "On the Statistical Estimation of Parametric Frontier Production Functions." *The Review of Economics and Statistics*, 58(2):238–239, 1976b.

Schmidt, P. "Estimation of seemingly unrelated regressions with unequal numbers of observations." *Journal of Econometrics*, 5(3):365–377, 1977. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(77)90045-8.

Schmidt, P. and Witte, A. D. *An Economic Analysis of Crime and Justice*. Academic Press, New York, 1984.

Schmidt, P. and Witte, A. D. *Predicting Recidivism Using Survival Models*. Springer-Verlag, New York, 1988.

White, H. *Asymptotic Theory for Econometricians*. Economic Theory, Econometrics, and Mathematical Economics. Emerald Group Publishing Limited, 2001. ISBN 9780127466521.

Zeileis, A. "Object-oriented computation of sandwich estimators." *Journal of Statistical Software*, 16(9):1–16, 2006. doi:10.18637/jss.v016.i09.

Appendices

A Proof of Theorem 1

First, we note that the GKRLS estimator is a linear smoother by substituting Eq. (10) into Eq. (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

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

B Proof of Theorem 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

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

De Brabanter et al. (2011) 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 Proof of Theorem 3

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

D Proof of Theorem 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 $\text{Bias}[\widehat{\mathbf{m}}_2|\mathbf{X}] = \mathbf{Lm} - \mathbf{m}$,

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

Note that $\text{E}[\mathbf{Lu}|\mathbf{X}] = \mathbf{0}$ and $\text{Var}[\mathbf{Lu}|\mathbf{X}] = \mathbf{L}\Omega\mathbf{L}^\top$. 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

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

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.[5] 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 White (2001), $\{Z_t, \mathcal{F}_t\}$ 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

[5] We follow the proof similar to the case of dependent identially distributed observations provided by White (2001).

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

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

E Proof of Theorem 5

First, we note that the GKRLS derivative estimator is a linear smoother by substituting Eq. (10) into Eq. (28),

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

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

$$S_r(\mathbf{x}_0) = \left[K^{*\top}_{\sigma_2,\mathbf{x}_0} \Delta_r (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \right]^\top \tag{46}$$

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

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

and

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

F Proof of Theorem 6

The bias of the GKRLS derivative estimator in Eq. (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 the GKLRS derivative estimator can be estimated by

$$\widehat{\text{Var}}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\Omega} S_r(\mathbf{x}_0) \tag{47}$$