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 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 KRLS regression and in the second step, another KRLS regression 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 KRLS in both heteroskedastic errors and autocorrelated errors cases. An empirical example is illustrated with estimating an airline cost function under a random effects model with heteroskedastic and correlated errors. The derivatives are evaluated, and the average partial effects of the inputs are determined in the application.

Keywords: Nonparametric estimator; Semiparametric models; Machine Learning; Kernel Regularized Least Squares; Covariance; Heteroskedasticity; Serial Correlation.

JEL Classification: C, C01, C1, C13, C14, C5, C51, C52.

*Department of Economics, University of San Diego. Email: justindang@sandiego.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 unbiased 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), al-

lowing for heteroskedastic errors. Although not directly stated, the LSSVM estimator 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 estimators 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 KRLS and in the second step, we estimate a model by KRLS 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 for a random effects model with heteroskedastic and correlated 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 $\mathbb{E}[U_i|X_1, \ldots, X_n] = \mathbb{E}[U_i|\mathbf{X}] = 0$, where $\mathbf{X} = (X_1, \ldots, X_n)^\top$ and

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

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 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 can be user specified but in this paper we only consider the RBF kernel in Eq. (4). The kernel function's hyperparameter σ and the regularization parameter λ can also be user specified or can be found via cross validation. The subscript of one denotes the 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 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 KRLS estimator. In practice, cross validation can be used for obtaining estimates for both hyperparameters. 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 $\widehat{\Omega} = \widehat{\Omega}(\hat{\theta})$.[1]

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 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.

[1] $\widehat{\Omega}$ can be thought of as a working covariance matrix since the parametric functional form may be subject to misspecification. One method to avoid misspecification is to estimate Ω nonparametrically. For example, under heteroskedasticity, one can estiamate Ω by a semiparametric KRLS estimator of the conditional variance (Dang and Ullah, 2022). Other solutions may be explored as future work.

2.3 Selection of Hyperparameters

Throughout this paper, we focus on the RBF kernel in Eq. (4), which contains the hyperparameter σ_1 (and σ_2). Since these parameters appear as squared in the RBF kernel in Eq. (4), we can instead search for the hyperparameters σ_1^2 and σ_2^2. The selection of the hyperparameters $\lambda_1, \lambda_2, \sigma_1^2$, and σ_2^2 is selected via leave one out cross validation (LOOCV). However, prior to cross validation, it is common in penalized methods to scale the data to have mean of 0 and standard deviation of 1. This way, the penalty parameters λ_1 and λ_2 do not depend on the scale of the data or the magnitude of the coefficients. Note that the scaling of the data does not affect the interpretations of predictions and marginal effects since the estimates can be translated back to their original scale and location.

For the hyperparameters, σ_1^2 and σ_2^2, Hainmueller and Hazlett (2014) suggest setting $\sigma^2 = q$, the number of regressors. Therefore, in items 1 and 2 in the two step procedure, $\sigma_1^2 = q$ and $\sigma_2^2 = q$. Then, only the penalty hyperparameters λ_1 and λ_2 need to be chosen. λ_1 is chosen via LOOCV in item 1 of the two step procedure using Eq. (5). λ_2 is then chosen via LOOCV in item 2 of the two step procedure using Eq. (9). If one wishes to also search for σ_1^2 and σ_2^2, one would perform LOOCV to find λ_1 and σ_1^2 simultaneously in item 1 using Eq. (5) and then perform another LOOCV to find λ_2 and σ_2^2 simultaneously in item 2 of the two step procedure using Eq. (9).

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. $\mathbb{E}[U_i|\mathbf{X}] = 0$ *and* $\mathbb{E}[U_i U_j|\mathbf{X}] = \omega_{ij}(\theta)$ *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,

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

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

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

and the bias corrected GKRLS estimator is $\sqrt{n}$*-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:*

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

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

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

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*

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

and

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

Proof: See Appendix F.

The average partial derivative with respect to the rth variable is

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

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

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

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 $\mathbf{x}_{0,j}, j = 1, \ldots, n'$.

5.1 First Differences for Binary Independent Variables

Unlike for the continuous case, partial effects for binary independent variables should be interpreted as and estimated by first differences. That is, the estimated effect of going from $x^{(b)} = 0$ to $x^{(b)} = 1$ can be determined by

$$\begin{aligned}\widehat{m}_{FD_b}(\mathbf{x}_0) &= \widehat{m}(x^{(b)} = 1, \mathbf{x}_0) - \widehat{m}(x^{(b)} = 0, \mathbf{x}_0) \\ &= L_{FD_b}(\mathbf{x}_0)^\top \mathbf{y}\end{aligned} \tag{38}$$

where $\widehat{m}_{FD_b}(\cdot)$ is the first difference estimator for the bth binary independent variable, $x^{(b)}$ is a binary variable that takes the values 0 or 1, $\mathbf{x}_0$ is the $(q-1)\times 1$ vector of the other independent variables evaluated at some test observation, and $L_{FD_b}(\mathbf{x}_0) \equiv L(x^{(b)} = 1, \mathbf{x}_0) - L(x^{(b)} = 0, \mathbf{x}_0)$ is the first difference smoother vector. The conditional bias and variance of the first difference GKRLS estimator in Eq. (38) are shown in the following theorem.

Theorem 7. *Using Theorems 1 and 2, the conditional bias and variance for the GKRLS first difference estimator in Eq.* (38) *are obtained by*

$$\text{Bias}[\widehat{m}_{FD_b}(\mathbf{x}_0)|X = \mathbf{x}_0] = L_{FD_b}(\mathbf{x}_0)^\top \mathbf{m} - m_{FD_b}(\mathbf{x}_0) \tag{39}$$

and

$$\text{Var}[\widehat{m}_{FD_b}(\mathbf{x}_0)|X=\mathbf{x}_0] = L_{FD_b}(\mathbf{x}_0)^\top \Omega L_{FD_b}(\mathbf{x}_0), \tag{40}$$

where $m_{FD_b}(\mathbf{x}_0) = m(x^{(b)}=1,\mathbf{x}_0) - m(x^{(b)}=0,\mathbf{x}_0)$.

Proof: See Appendix G

Then, the conditional bias and variance can be estimated as follows:

$$\widehat{\text{Bias}}[\widehat{m}_{FD_b}(\mathbf{x}_0)|X=\mathbf{x}_0] = L_{FD_b}(\mathbf{x}_0)^\top \widehat{\mathbf{m}} - \widehat{m}_{FD_b}(\mathbf{x}_0) \tag{41}$$

$$\widehat{\text{Var}}[\widehat{m}_{FD_b}(\mathbf{x}_0)|X=\mathbf{x}_0] = L_{FD_b}(\mathbf{x}_0)^\top \widehat{\Omega} L_{FD_b}(\mathbf{x}_0). \tag{42}$$

Note that Eq. (38) provides the pointwise first difference estimates. If one is interested in the average partial effect of going from $x^{(b)}=0$ to $x^{(b)}=1$, the following average first difference GKRLS estimator would be used.

$$\widehat{m}_{\overline{FD},b} = \frac{1}{n'}\sum_{j=1}^{n'} \widehat{m}_{FD_b}(\mathbf{x}_{0,j}) \tag{43}$$

This average partial effect is similar to the continuous case and can be compared to traditional parametric partial effects as in the case of least squares coefficients. The conditional bias and variance of the average first difference GKRLS estimator in Eq. (43) are:

$$\text{Bias}[\widehat{m}_{\overline{FD}_b}(\mathbf{x}_0)|X=\mathbf{x}_0] = \frac{1}{n'}\boldsymbol{\iota}_{n'}^\top \mathbf{L}_{FD_{0,b}}\mathbf{m} - \frac{1}{n'}\boldsymbol{\iota}_{n'}^\top \mathbf{m}_{FD_{0,b}} \tag{44}$$

$$\text{Var}[\widehat{m}_{\overline{FD}_b}|X=\mathbf{x}_0] = \frac{1}{n'^2}\boldsymbol{\iota}_{n'}^\top \mathbf{L}_{FD_{0,b}}\Omega\mathbf{L}_{FD_{0,b}}^\top, \tag{45}$$

where $\mathbf{L}_{FD_{0,b}}$ is the $n' \times n$ smoother matrix with the jth row as $L_{FD_b}(\mathbf{x}_{0,j}), j=1,\ldots,n'$, and $\mathbf{m}_{FD_{0,b}}$ is the $n' \times 1$ vector of first differences evaluated at each $\mathbf{x}_{0,j}, j=1,\ldots,n'$. The conditional bias and variance of the average first difference estimator can be estimated using Eqs. (41) and (42).

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 three independent variables X that is generated from

$$\begin{aligned} X_1 &\sim Bern(0.5) \\ X_2 &\sim N(0, 1) \\ X_3 &\sim U(-1, 1). \end{aligned} \tag{46}$$

The specification for m is:

$$m(X_i) = 5 - 2X_{i,1} + \sin(X_{i,2}) + 3X_{i,3} \tag{47}$$

and the partial derivatives with respect to each independent variable are given by

$$\begin{aligned} m_1^{(1)}(X_i) &= -2 \\ m_2^{(1)}(X_i) &= \cos(X_{i,2}) \\ m_3^{(1)}(X_i) &= 3 \end{aligned} \tag{48}$$

For the error terms, we consider two cases.

$$\begin{aligned} U_i &= 0.7U_{i-1} + V_i \\ V_i &\sim N(0, 5^2) \end{aligned} \tag{49}$$

and

$$U_i \sim N\left(0, \exp(X_{i,1} + 0.2X_{i,2} - 0.3X_{i,3})\right) \tag{50}$$

First, in Eq. (49), U_i is generated by an AR(1) process. Second, U_i is heteroskedastic but independent of each other with $\text{Var}[U_i|X_i] = \exp(X_{i,1} + 0.2X_{i,2} - 0.3X_{i,3})$.

In addition to the proposed estimator, we compare four other nonparametric estimators: the KRLS estimator (KRLS), Local Polynomial (LP) estimator with degree zero, Random Forest (RF), and Support Vector Machine (SVM). The KRLS estimator is used as a comparison to GKRLS to show the magnitude of the efficiency loss from ignoring the information in the error covariance matrix. The KRLS, LP, RF, and SVM estimators do not utilize the covariance matrix in estimating the regression function and excludes heteroskedasticity or autocorrelation of the errors. For the GKRLS and KRLS estimators, we set $\sigma_1^2 = \sigma_2^2 = 3$, the number of independent variables in this example, and implement leave one out cross validation to select the hyperparameters, λ_1 and λ_2.[2] The variance function under the heteroskedastic case is estimated by least squares from the regression of the log residuals on X. Taking the exponential would give the predicted variance estimates. Under the case of AR(1) errors, the covariance function is estimated from an AR(1) model. We run 200 simulations for each of the two cases and the bias corrected results are reported below in Table 1.[3] To evaluate the estimators, mean squared error is used as the main criterion, where we also investigate the bias and variance. To compare results, all estimators are evaluated from 300 data points generated from Eqs. (46) and (47).

Table 1 displays the evaluations, including bias, variance, and MSE of the estimators for the regression function under both error cases. Note that the GKRLS and KRLS estimates in Table 1 are bias corrected. All estimates are averaged across all simulations. Estimates based on GKRLS seem to exhibit similar finite sample bias as KRLS, and there is an obvious

[2]The hyperparameters of the LP, RF, and SVM estimators are chosen by their default methods in their respective R packages.

[3]The following R packages were used for conducting simulations: Borchers (2021), Hyndman and Khandakar (2008), McLeod et al. (2007), Boos and Nychka (2022), Hayfield and Racine (2008), Liaw and Wiener (2002), and Meyer et al. (2022).

Simulation Evaluation for $m(\mathbf{x}_0)$

		MSE	Variance	Bias
Autocor. Errors	GKRLS	2.8562	1.6311	0.0140
	KRLS	2.9767	2.3835	-0.0094
	LP	3.4623	3.0822	-0.0112
	RF	3.8442	3.5013	0.0205
	SVM	5.7663	5.6482	0.0263
Heterosk. Errors	GKRLS	0.2287	0.1702	0.0103
	KRLS	0.2366	0.1766	-0.0148
	LP	0.2696	0.1958	0.0055
	RF	0.5917	0.1372	0.0178
	SVM	0.2632	0.2105	-0.0001

Table 1: *The table reports the bias, variance, and MSE of GKRLS, KRLS, LP, RF, and SVM estimators for the regression function $m(\mathbf{x}_0)$ under the cases of heteroskedastic and AR(1) errors generated from Eqs.* (46)*,* (47)*,* (49) *and* (50)*. The GKRLS and KRLS estimates are bias corrected. All estimates are averaged across all simulations.*

reduction in the variability with smaller variance of the proposed estimator relative to KRLS. Note that GKRLS estimation provides a 31.6% and a 3.6% decrease in the variance for estimating the regression function for the autocorrelated and heteroskedastic errors, relative to KRLS. With smaller variance, GKRLS also has a smaller MSE, making GKRLS superior to KRLS. Compared to the other nonparametric estimators, LP, RF, and SVM, the GKRLS estimator outperforms the others in terms of MSE and is the preferred method in the presence of heteroskedasticity or autocorrelation.

Table 2 displays the evaluations, including bias, variance, and MSE of the bias GKRLS and KRLS estimators for the partial derivatives of the regression function with respect to each of the independent variables under both error cases.[4] Since X_1 is discrete, the partial derivative is estimated by first differences. Similar to the regression estimates, for both heteroskedastic and AR(1) errors, the variability from estimating the derivative is reduced by GKRLS estimation relative to KRLS estimation. In addition, the efficiency gain in estimating

[4]The derivatives are not reported for LP, RF, and SVM since derivative estimation for RF and SVM methods are uncommon. The derivative estimates for LP can be obtained but in this simulation the GKRLS estimator is superior with respect to MSE.

Simulation Evaluation for $m_r^{(1)}(\mathbf{x}_0)$

		GKRLS			KRLS		
		MSE	Variance	Bias	MSE	Variance	Bias
Autocor. Errors	X_1	1.1708	0.4092	0.8239	2.1013	1.7419	0.5017
	X_2	0.3800	0.0887	-0.3567	0.7745	0.5502	-0.2700
	X_3	5.2002	0.3361	-2.0737	5.5494	1.6599	-1.7282
Heterosk. Errors	X_1	0.3290	0.2835	0.0950	0.3291	0.2922	0.0914
	X_2	0.2414	0.1695	-0.0421	0.2524	0.1718	-0.0534
	X_3	2.0529	0.5746	-0.7904	2.1461	0.5876	-0.8218

Table 2: *The table reports the bias, variance, and MSE of the bias corrected GKRLS and KRLS estimators and the cases of heteroskedastic and AR(1) errors for the derivative of the regression function $m_r^{(1)}(\mathbf{x}_0)$ generated from Eqs.* (46) *to* (50)*. Each row represents the MSE, variance, and bias of the partial derivative estimates with respect to X_r, $r = 1, 2, 3$. All estimates are averaged across all simulations.*

both the regression and the derivative seems to be more evident in the AR(1) 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 and therefore MSE is clearly evident in Tables 1 and 2, making the GKRLS the preferred estimator.

Table 3 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, we set $\sigma_1^2 = \sigma_2^2 = 3$ and the hyperparameters λ_1 and λ_2 are found by LOOCV. For the regression function and the derivative and for both error covariance structures, the squared bias, variance, and MSE all decrease as the sample size increases, which implies that the GKRLS estimator is consistent in this simulation exercise.

Simulation Results for Consistency of GKRLS

		Autocor. Errors			Heterosk. Errors		
		MSE	Variance	Bias^2	MSE	Variance	Bias^2
$m(\mathbf{x}_0)$	$n = 100$	4.9665	2.8562	1.6112	0.4113	0.2287	0.1309
	$n = 200$	2.7170	1.6311	0.8786	0.3012	0.1702	0.0993
	$n = 400$	2.2496	1.2251	0.7326	0.1101	0.0585	0.0316
$m_1^{(1)}(\mathbf{x}_0)$	$n = 100$	2.3091	0.5590	1.7501	0.5880	0.5196	0.0683
	$n = 200$	1.1708	0.4092	0.7615	0.3290	0.2835	0.0455
	$n = 400$	0.6992	0.2647	0.4345	0.1964	0.1695	0.0269
$m_2^{(1)}(\mathbf{x}_0)$	$n = 100$	0.4614	0.1164	0.3449	0.3751	0.2702	0.1049
	$n = 200$	0.3800	0.0887	0.2913	0.2414	0.1695	0.0719
	$n = 400$	0.2962	0.0715	0.2247	0.1601	0.1063	0.0539
$m_3^{(1)}(\mathbf{x}_0)$	$n = 100$	6.6704	0.4951	6.1753	2.8633	0.8853	1.9780
	$n = 200$	5.2002	0.3361	4.8641	2.0529	0.5746	1.4783
	$n = 400$	4.4179	0.2261	4.1918	1.5181	0.3793	1.1388

Table 3: *The table reports the bias, variance, and MSE of the GKRLS estimator for both the regression function and the partial derivatives and for the cases of heteroskedastic and AR(1) errors generated from Eqs.* (46) *to* (50) *for different sample sizes,* $n = 100, 200, 400$. *All reported estimates are biased corrected and are averaged across all simulations. The kernel hyperparameters are set as* $\sigma_1^2 = \sigma_2^2 = 3$ *and the hyperparameters* λ_1 *and* λ_2 *are found by LOOCV.*

7 Application

We implement an empirical application from the U.S. airline industry with heteroskedastic and autocorrelated errors using a panel of 6 firms over 15 years.[5] For the data set, we set aside a portion of the data for training and the other for testing. We estimate the model with four methods, GKRLS, KRLS, LP, and Generalized Least Squares (GLS), and compare their results in terms of mean squared error (MSE). To evaluate the out of sample performance of each method, the predicted out of sample MSEs are computed as follows

$$MSE_e = \frac{1}{n'T} \sum_{i=1}^{n'} \sum_{t=1}^{T} (y_{0,it} - \widehat{m}_e(\mathbf{x}_{0,it}))^2 \tag{51}$$

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

where MSE_e is the mean squared error for the e^{th} estimator and n' is the number of observations in the testing data set and $j = 1, \ldots, n'$. In this empirical exercise, $n' = 1$ and $T = 15$ since we leave out the first firm as a test set. To assess the estimated average derivatives, we use the bootstrap to calculate the MSEs for the average partial effects. We report the bootstrapped MSEs for the average derivative by the following.[6]

$$MSE_{e,r} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{m}^{(1)}_{avg,e,r,b} - \frac{1}{4}\sum_{e}\widehat{m}^{(1)}_{avg,e,r}\right)^2 \tag{52}$$

where B is the number of bootstraps with $b = 1, \ldots, B$, $\widehat{m}^{(1)}_{avg,e,r,b}(\cdot)$ is the b^{th} bootstrapped average partial first derivative with respect to the r^{th} variable for the e^{th} estimator, and $\frac{1}{4}\sum_e \widehat{m}^{(1)}_{avg,e,r}$ is the simple average of the average partial first derivatives with respect to the r^{th} variable from the four estimators (GLS, GKRLS, KRLS, and LP):

$$\begin{aligned} \widehat{m}^{(1)}_{avg,e,r} &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{m}^{(1)}_{e,r}(\mathbf{x}_{it}), \\ e &= \{\text{GLS, GKRLS, KRLS, LP}\} \end{aligned} \tag{53}$$

7.1 U.S. Airline Industry

We obtain the data on the efficiency in production of airline services from Greene (2018). Since the data are a panel of 6 firms for 15 years, we consider the one way random effects model:

$$\log C_{it} = m(\log Q_{it}, \log P_{it}) + \alpha_i + \varepsilon_{it}, \tag{54}$$

where the dependent variable $Y_{it} = \log C_{it}$ is the logarithm of total cost, the independent variables $X_{it} = (\log Q_{it}, \log P_{it})^\top$ are the logarithms of output and the price of fuel, respectively, α_i is the firm specific effect, and ε_{it} is the idiosyncratic error term. In this empirical

[6]The R package by Callaway (2022) was used to obtain the bootstrap samples.

setting, we assume $\mathbb{E}[\varepsilon_{it}|\mathbf{X}] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\mathbf{X}] = \sigma_{\varepsilon_i}^2$, $\mathbb{E}[\alpha_i|\mathbf{X}] = 0$, $\mathbb{E}[\alpha_i^2|\mathbf{X}] = \sigma_{\alpha_i}^2$, $\mathbb{E}[\varepsilon_{it}\alpha_j|\mathbf{X}] = 0$ for all i, t, j, $\mathbb{E}[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}] = 0$ if $t \neq s$ or $i \neq j$, and $\mathbb{E}[\alpha_i\alpha_j|\mathbf{X}] = 0$ if $i \neq j$. Consider the composite error term $U_{it} \equiv \alpha_i + \varepsilon_{it}$. Then, the model in Eq. (54) can be rewritten as

$$\log C_{it} = m(\log Q_{it}, \log P_{it}) + U_{it}, \tag{55}$$

In Eq. (55), the independent variables are strictly exogenous to the composite error term, $\mathbb{E}[U_{it}|\mathbf{X}] = 0$. The variance of the composite error term is $\mathbb{E}[U_{it}^2|\mathbf{X}] = \sigma_{\alpha_i}^2 + \sigma_{\varepsilon_i}^2$. Therefore, in this empirical example, we allow for firm specific heteroskedasticity. In other words, the variance of the error terms are not constant across firms, but are constant over time for each firm. Since there is a time component, we allow an individual firm to be correlated across time but not with other firms, that is, $\mathbb{E}[U_{it}U_{is}|\mathbf{X}] = \sigma_{\alpha_i}^2$, $t \neq s$ and $\mathbb{E}[U_{it}U_{js}|\mathbf{X}] = 0$ for all t and s if $i \neq j$. Note that the correlation across time can be different for every firm. Therefore, in this empirical framework, we allow the error terms to be heteroskedastic across firms and correlated across time.

To estimate Eq. (55) by GKRLS and KRLS in the framework set up in this paper, we can write the model in matrix notation. Consider

$$\mathbf{y} = \mathbf{m} + \mathbf{U}, \tag{56}$$

where $\mathbf{y}$ is the $nT \times 1$ vector of $\log C_{it}$, $\mathbf{m}$ is the $nT \times 1$ vector of the regression function $m(X_{it})$, and $\mathbf{U}$ is the $nT \times 1$ vector of U_{it}, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Then, the $nT \times nT$ error covariance matrix Ω is

$$\Omega = \text{Var}[\mathbf{U}|\mathbf{X}] = \text{diag}(\Sigma_1, \ldots, \Sigma_n), \tag{57}$$

where $\Sigma_i = \sigma_{\varepsilon_i}^2 \mathbf{I}_T + \sigma_{\alpha_i}^2 \boldsymbol{\iota}_T \boldsymbol{\iota}_T^\top, i = 1, \ldots, n$ has dimension $T \times T$, $\mathbf{I}_T$ is a $T \times T$ identity matrix and $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones. To use the GKRLS estimator in this empirical framework,

we first estimate Eq. (55) or Eq. (56) by KRLS and obtain the residuals, denoted by $\widehat{u}_{it}$. To estimate the error covariance matrix Ω, the variances of the firm specific error and the idiosyncratic error, $\sigma^2_{\alpha_i}$ and $\sigma^2_{\varepsilon_i}$ need to be estimated. Consider the following consistent estimators using time averages,

$$\widehat{\sigma}^2_{U_i} = \frac{1}{T}\widehat{\mathbf{u}}_i^\top \widehat{\mathbf{u}}_i \tag{58}$$

$$\widehat{\sigma}^2_{\alpha_i} = \frac{1}{T(T-1)/2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \widehat{u}_{it}\widehat{u}_{is} \tag{59}$$

$$\widehat{\sigma}^2_{\varepsilon_i} = \widehat{\sigma}^2_{U_i} - \widehat{\sigma}^2_{\alpha_i}, \tag{60}$$

where $\widehat{\mathbf{u}}_i$ is the $T \times 1$ vector of residuals for the i^{th} firm. Now, plugging these estimates in for Ω, the GKRLS estimator can be estimated as in the previous sections. For further details, please see Appendix H.

With regards to the other comparable estimators, the KRLS and LP estimators are used to estimate Eq. (55) or Eq. (56) ignoring the heteroskedasticity and correlation in the composite error, $\mathbf{U}$. Note that the KRLS estimator uses the error covariance matrix in the variances and standard errors but does not use the error covariance in estimating the regression function. Lastly, the GLS estimator is used as a parametric benchmark to compare to the standard random effects panel data model.[7]

The data contain 90 observations of 6 firms for 15 years, from 1970-1984. We split the data into two parts, where the first 15 observations, which corresponds to the first firm, are used as testing data and 75 observations, which corresponds to the last five firms, are set as training data to evaluate out of sample performance. So, for the training data, $i = 1, \ldots, 5$ and $t = 1, \ldots, 15$, with a total of 75 observations for training. For the GKRLS and KRLS estimators, all hyperparameters are chosen via LOOCV.[8]

[7]The R package by Croissant and Millo (2008) was used to obtained the Random Effects GLS estimator.

[8]For the LP estimator, cross validation is used to select the hyperparameters. The local constant estimator is used, although one can use the local linear estimator, which gives similar results to that of the local constant.

Average Partial Derivatives for Airline Data

	log(Q)	log(P)
GLS	0.8436	0.4188
	(0.0311)	(0.0181)
GKRLS	0.8130	0.4247
	(0.0034)	(0.0082)
KRLS	0.8248	0.4581
	(0.016)	(0.0457)
LP	0.5885	0.2260
	(0.0276)	(0.0138)

Table 4: *Bias corrected average partial derivatives and their standard errors in parentheses are reported for GLS, GKRLS, KRLS, and LP estimators. The columns represent the estimates of the average partial derivative with respect to each regressor.*

The bias corrected average partial derivatives and corresponding standard errors are reported in Table 4. These averages are calculated by training each estimator on the five firms with 75 observations in the training data set. The estimates are bias corrected and the results from Section 5 are used in our calculations. All estimators display positive and significant relationships between cost and each of the regressors, output and price, with their average partial derivatives being positive. The elasticity with respect to output ranges from 0.5885 to 0.8436 and with respect to price ranges from 0.2260 to 0.4581. More specifically, for the GKRLS estimator, a 10% increase in output would increase the total cost by an average of 8.13% and a 10% increase in fuel price would increase the total cost by an average of 4.25% holding all else fixed. Comparing the GKRLS and KRLS methods, the estimates of the average partial derivatives are similar but the standard errors are significantly reduced for GKRLS for both output and fuel price, implying a gain in efficiency. Therefore, using the information and the structure of the error covariance in Eq. (57) in estimated the regression function allows GKRLS to provide more robust estimates of the average partial effects of each independent variable compared to KRLS.

Table 4 shows that the GLS estimator overestimates the elasticity with respect to output and underestimates the elasticity with respect to fuel price compared to those of GKRLS.

The LP estimator appears to provide different average partial effect estimates compared to the rest of the estimators. One possible explanation is that the bandwidths may not be the most optimal since data-driven bandwidth selection methods (e.g., cross validation) fail when there is correlation in the errors (De Brabanter et al., 2018). Since the data is panel structured, there is correlation across time, making bandwidth selection for LP estimators difficult. The LP estimates are from the local constant estimator; however, the local linear estimator provides similar estimates of the average partial effects to those of the local constant estimator. Nevertheless, the LP average partial effects of each variable are positive and significant, which are consistent with the other methods. Furthermore, GKRLS provides similar average partial effects with respect to output and price but is more efficient in terms of smaller standard errors relative to the other considered estimators.

MSEs for Airline Data

	MSE	$\text{MSE}_{\log Q}$	$\text{MSE}_{\log P}$
GLS	0.0106	0.0042	0.0018
GKRLS	0.0091	0.0030	0.00001
KRLS	0.0306	0.0031	0.0024
LP	0.0191	0.2900	0.0867

Table 5: *The MSEs are reported for the GLS, GKRLS, KRLS, and LP, estimators. The first column are the out of sample MSEs calculated by Eq.* (51) *and the second and third columns are the bootstrapped MSEs for the average partial derivatives calculated by Eq.* (52)*. The GKRLS and KRLS estimates are bias corrected.*

To assess the estimators in terms of out of sample performance, we calculate the MSEs using the 15 observations in the testing data set. Table 5 reports MSEs for the four considered estimators. The first column reports the out of sample MSEs using the 15 observations from the first firm. Out of all the considered estimators, the GKRLS estimator outperforms the others in terms of MSE. In other words, the GKRLS estimator can be seen as the superior method in estimating the regression function in this empirical example. The bootstrapped MSEs for the average partial derivatives, calculated by Eq. (52), are reported in the second and third columns of Table 1. For both the average partial derivatives with respect to

output and price, GKRLS produces the lowest MSE, outperforming the other estimators. In addition, since GKRLS incorporates the error covariance structure, efficiency is gained and therefore reductions in MSEs are made relative to KRLS. Overall, GKRLS is considered to be the best method in terms of MSE for estimating both the airline cost function and the average partial effects with respect to output and price.

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 under a random effects model with heteroskedastic and correlated errors. The average derivatives are evaluated, and the average partial effects of the inputs are determined in the application. In the empirical exercise, GKRLS is more efficient compared to KRLS and is the most preferred method for estimating the airline cost function and its average partial derivatives in terms of MSE.

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.

Boos, D. D. and Nychka, D. *Rlab: Functions and Datasets Required for ST370 Class*, 2022. R package version 4.0.

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

Callaway, B. *BMisc: Miscellaneous Functions for Panel Data, Quantiles, and PrintingResults*, 2022. R package version 1.4.5.

Croissant, Y. and Millo, G. "Panel data econometrics in R: The plm package." *Journal of Statistical Software*, 27(2):1–43, 2008. doi:10.18637/jss.v027.i02.

Dang, J. and Ullah, A. "Machine-learning-based semiparametric time series conditional variance: Estimation and forecasting." *Journal of Risk and Financial Management*, 15(1), 2022. ISSN 1911-8074. doi:10.3390/jrfm15010038.

De Brabanter, K., Cao, F., Gijbels, I., and Opsomer, J. "Local polynomial regression with correlated errors in random design and unknown correlation structure." *Biometrika*, 105(3):681–690, 2018. ISSN 0006-3444. doi:10.1093/biomet/asy025.

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.

Hayfield, T. and Racine, J. S. "Nonparametric econometrics: The np package." *Journal of Statistical Software*, 27(5):1–32, 2008. doi:10.18637/jss.v027.i05.

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

Liaw, A. and Wiener, M. "Classification and regression by randomforest." *R News*, 2(3):18–22, 2002.

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

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. *e1071: Misc Functions of the Department of Statistics, ProbabilityTheory Group (Formerly: E1071), TU Wien*, 2022. R package version 1.7-12.

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.

Appendices

A Proof of Theorem 1

First, we note that the GKRLS estimator is a linear smoother by substituting Eq. (10) into Eq. (11)

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

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

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

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

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

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

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

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 $d_i, i = 1, \ldots, n$ 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.

We also provide an alternative proof of consistency. Consider the GKRLS coefficient estimator of $\mathbf{c}$ in Eq. (10):

$$\begin{aligned}
\widehat{c}_2 &= (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2\mathbf{I})^{-1}\Omega^{-1}\mathbf{y} \\
&= (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2\mathbf{I})^{-1}\Omega^{-1}\left(\mathbf{K}_{\sigma_2}\mathbf{c} + \mathbf{u}\right) \\
&= \left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2} + \frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\frac{1}{n}\Omega^{-1}\left(\mathbf{K}_{\sigma_2}\mathbf{c} + \mathbf{u}\right) \\
&= \left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2} + \frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2}\right)\mathbf{c} + \left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2} + \frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\left(\frac{1}{n}\Omega^{-1}\right)\mathbf{u}
\end{aligned}$$

Again, since we consider the bias corrected estimator, $\widehat{\mathbf{m}}_{2,c}$, we can focus on the conditional variance. However, below we also show that the non-bias corrected estimator has zero conditional bias in the limit. Taking the conditional bias of $\widehat{\mathbf{c}}_2$:

$$\text{Bias}[\widehat{\mathbf{c}}_2|\mathbf{X}] = \left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2} + \frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2}\right)\mathbf{c} - \mathbf{c},$$

where the strict exogeneity assumption $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ is used. Furthermore, if we assume λ_2 is fixed or does not grow as fast as n and $\left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2}\right) \to \mathbf{Q}$, a positive definite matrix with finite elements, when $n \to \infty$, then $\text{Bias}[\widehat{\mathbf{c}}_2|\mathbf{X}] \to \mathbf{0}$ as $n \to \infty$.

Taking the conditional variance of $\widehat{\mathbf{c}}_2$:

$$\text{Var}[\widehat{\mathbf{c}}_2|\mathbf{X}] = \frac{1}{n}\left(\frac{\Omega^{-1}\mathbf{K}_{\sigma_2}}{n} + \frac{\lambda_2\mathbf{I}}{n}\right)^{-1}\left(\frac{\Omega^{-1}}{n}\right)\left[\left(\frac{\Omega^{-1}\mathbf{K}_{\sigma_2}}{n} + \frac{\lambda_2\mathbf{I}}{n}\right)^{-1}\right]^{\top}.$$

Again, we assume that λ_2 is fixed or does not grow as fast as n and $\left(\frac{1}{n}\Omega^{-1}\mathbf{K}_{\sigma_2}\right) \to \mathbf{Q}$, a positive definite matrix with finite elements. Furthermore, if we assume that $\left(\frac{1}{n}\Omega^{-1}\right) \to \mathbf{Q}_\Omega$, a matrix with finite elements when $n \to \infty$, then $\text{Var}[\widehat{\mathbf{c}}_2|\mathbf{X}] \to \mathbf{0}$ as $n \to \infty$. Therefore, $\underset{n\to\infty}{\text{plim}}\, \widehat{\mathbf{c}}_2 = \mathbf{c}$.

Now, consider the GKRLS estimator $\widehat{\mathbf{m}}_2 = \mathbf{K}_{\sigma_2}\widehat{\mathbf{c}}_2$. Then,

$$\begin{aligned}
\underset{n\to\infty}{\text{plim}}\, \widehat{\mathbf{m}}_2 &= \mathbf{K}_{\sigma_2}\left(\underset{n\to\infty}{\text{plim}}\, \widehat{\mathbf{c}}_2\right) \\
&= \mathbf{K}_{\sigma_2}\mathbf{c} \\
&= \mathbf{m},
\end{aligned}$$

proving consistency of $\widehat{\mathbf{m}}_2$. Note that since the variance is $O(1/n)$, $\widehat{\mathbf{m}}_2$ is $\sqrt{n}$-consistent.

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

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

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 heterogeneously 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

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

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

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

The following results will be for the case of autocorrelated errors, where observations are dependent and identically distributed.[9] Define $\mathbf{L}_n \equiv \mathbf{K}_{\sigma_2}\left(\frac{\Omega^{-1}\mathbf{K}_{\sigma_2}+\lambda_2\mathbf{I}}{n}\right)^{-1}\Omega^{-1}$ and $L_n(\mathbf{X}_t)$

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

as the $t-$th row of $\mathbf{L}_n$. 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) $\{L_n(X_{thi})u_{th}, \mathcal{F}_t\}$ is an adapted mixingale of size -1, $h = 1, \ldots, p, i = 1, \ldots, n$; (b) $\mathbb{E}|L_n(X_{thi})u_{th}|^2 < \infty, h = 1, \ldots, p, i = 1, \ldots, n$; (c) $\mathbf{V}_n \equiv \text{Var}\left(\frac{1}{\sqrt{n}}\mathbf{L}_n\mathbf{u}\right)$ is uniformly positive definite; (iv) $\mathbb{E}|L_n(X_{thi})|^2 < \infty, h = 1, \ldots, p, i = 1, \ldots, n$; (v) $\lim_{n\to\infty} L_n(\mathbf{X}_t) = L(X_t)$ and $\lim_{n\to\infty} \mathbf{L}_n = \mathbf{L}$.

Consider $n^{-1/2}\sum_{t=1}^n \boldsymbol{\lambda}^\top \mathbf{V}^{-1/2} L_n(\mathbf{X}_t)u_t$, 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

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

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

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

since for Δ sufficiently large, $\mathbb{E}|L_n(X_{thi})u_{th}|^2 < \Delta < \infty$ 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

just given, we can write

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

Applying Minkowski's inequality, it follows that

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

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

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

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 definite 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, $n^{-1/2}\sum_{t=1}^{n}\boldsymbol{\lambda}^\top\mathbf{V}^{-1/2}L_n(\mathbf{X}_t)u_t \xrightarrow{d} N(0,1)$. Since this holds for every $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^\top\boldsymbol{\lambda} = 1$, it follows from Cramér-Wold Theorem, that $n^{-1/2}\mathbf{V}^{-1/2}\sum_{t=1}^{n}L_n(\mathbf{X}_t)u_t \xrightarrow{d}$

$N(\mathbf{0}, \mathbf{I})$. Hence, $\sqrt{n}\mathbf{L}\mathbf{u} \xrightarrow{d} N(\mathbf{0}, \mathbf{L}\Omega\mathbf{L}^\top)$ and it then follows that

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

E Proof of Theorem 5

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

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

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

$$S_r(\mathbf{x}_0) = \left[K_{\sigma_2,\mathbf{x}_0}^{*\top} \Delta_r (\Omega^{-1}\mathbf{K}_{\sigma_2} + \lambda_2 \mathbf{I})^{-1} \Omega^{-1} \right]^\top \tag{61}$$

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

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

and

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

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 GKRLS derivative estimator can be estimated by

$$\widehat{\text{Var}}[\widehat{m}_{2,r}^{(1)}(\mathbf{x}_0)|X=\mathbf{x}_0] = S_r(\mathbf{x}_0)^\top \widehat{\Omega} S_r(\mathbf{x}_0) \tag{62}$$

G Proof of Theorem 7

The conditional bias of the GKRLS first difference estimator in Eq. (38) is

$$\begin{aligned}
&\text{Bias}[\widehat{m}_{FD_b}(\mathbf{x}_0)|X=\mathbf{x}_0] \\
&\quad = L(x^{(b)}=1,\mathbf{x}_0)^\top \mathbf{m} - m(x^{(b)}=1,\mathbf{x}_0) - \left[L(x^{(b)}=0,\mathbf{x}_0)^\top \mathbf{m} - m(x^{(b)}=0,\mathbf{x}_0)\right] \\
&\quad = \left[L(x^{(b)}=1,\mathbf{x}_0) - L(x^{(b)}=0,\mathbf{x}_0)\right]^\top \mathbf{m} - \left[m(x^{(b)}=1,\mathbf{x}_0) - m(x^{(b)}=0,\mathbf{x}_0)\right] \\
&\quad = L_{FD_b}(\mathbf{x}_0)^\top \mathbf{m} - m_{FD_b}(\mathbf{x}_0),
\end{aligned}$$

where $m_{FD_b}(\mathbf{x}_0) = m(x^{(b)}=1,\mathbf{x}_0) - m(x^{(b)}=0,\mathbf{x}_0)$ is the true first difference of m with respect to the bth variable and $L_{FD_b}(\mathbf{x}_0) = L(x^{(b)}=1,\mathbf{x}_0) - L(x^{(b)}=0,\mathbf{x}_0)$ is the first difference smoother vector.

The conditional variance of the GKRLS first difference estimator in Eq. (38) is

$$\begin{aligned}
&\text{Var}[\widehat{m}_{FD_b}(\mathbf{x}_0)|X=\mathbf{x}_0] \\
&\quad = L(x^{(b)}=1,\mathbf{x}_0)^\top \Omega L(x^{(b)}=1,\mathbf{x}_0) + L(x^{(b)}=0,\mathbf{x}_0)^\top \Omega L(x^{(b)}=0,\mathbf{x}_0) \\
&\qquad - L(x^{(b)}=1,\mathbf{x}_0)^\top \Omega L(x^{(b)}=0,\mathbf{x}_0) - L(x^{(b)}=0,\mathbf{x}_0)^\top \Omega L(x^{(b)}=1,\mathbf{x}_0) \\
&\quad = \left[L(x^{(b)}=1,\mathbf{x}_0) - L(x^{(b)}=0,\mathbf{x}_0)\right]^\top \Omega \left[L(x^{(b)}=1,\mathbf{x}_0) - L(x^{(b)}=0,\mathbf{x}_0)\right] \\
&\quad = L_{FD_b}(\mathbf{x}_0)^\top \Omega L_{FD_b}(\mathbf{x_0}).
\end{aligned}$$

H A Random Effects Model for Airline Data used in Section 7

Consider the following random effects model for an airline cost function:

$$Y_{it} = m(X_{it}) + \alpha_i + \varepsilon_{it},$$

$Y_{it} = \log C_{it}$, $X_{it} = (\log Q_{it}, \log P_{it})^\top$, α_i is the firm specific effect, and ε_{it} is the idiosyncratic error term. In this empirical setting, we assume

$$\mathbb{E}[\varepsilon_{it}|\mathbf{X}] = 0$$

$$\mathbb{E}[\varepsilon_{it}^2|\mathbf{X}] = \sigma_{\varepsilon_i}^2$$

$$\mathbb{E}[\alpha_i|\mathbf{X}] = 0$$

$$\mathbb{E}[\alpha_i^2|\mathbf{X}] = \sigma_{\alpha_i}^2$$

$$\mathbb{E}[\varepsilon_{it}\alpha_j|\mathbf{X}] = 0 \text{ for all } i, t, j$$

$$\mathbb{E}[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}] = 0 \text{ if } t \neq s \text{ or } i \neq j$$

$$\mathbb{E}[\alpha_i\alpha_j|\mathbf{X}] = 0 \text{ if } i \neq j$$

Consider the composite error term $U_{it} \equiv \alpha_i + \varepsilon_{it}$. Then, the model with the composite error term is

$$Y_{it} = m(X_{it}) + U_{it}$$

Note that the independent variables are strictly exogenous; the regressors are mean independent of each error term and therefore of the composite error term:

$$\begin{aligned}\mathbb{E}[U_{it}|\mathbf{X}] &= \mathbb{E}[\alpha_i|\mathbf{X}] + \mathbb{E}[\varepsilon_{it}|\mathbf{X}] \\ &= 0.\end{aligned}$$

In this framework, we allow for the errors to be heteroskedastic and correlated across time. The variance of the composite error term is

$$\begin{aligned}\mathbb{E}[U_{it}^2|\mathbf{X}] &= \mathbb{E}[\alpha_i^2|\mathbf{X}] + \mathbb{E}[\varepsilon_{it}^2|\mathbf{X}] + 2\mathbb{E}[\alpha_i\varepsilon_{it}|\mathbf{X}] \\ &= \sigma_{\alpha_i}^2 + \sigma_{\varepsilon_i}^2,\end{aligned}$$

where $\mathbb{E}[\alpha_i \varepsilon_{it} | \mathbf{X}] = 0$ by assumption. The covariance of the composite errors is

$$\begin{aligned}
\mathbb{E}[U_{it} U_{is} | \mathbf{X}] &= \mathbb{E}[(\alpha_i + \varepsilon_{it})(\alpha_i + \varepsilon_{is}) | \mathbf{X}] \\
&= \mathbb{E}[\alpha_i^2 | \mathbf{X}] \\
&= \sigma_{\alpha_i}^2 \text{ for } t \neq s
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[U_{it} U_{js} | \mathbf{X}] &= \mathbb{E}[(\alpha_i + \varepsilon_{it})(\alpha_j + \varepsilon_{js}) | \mathbf{X}] \\
&= 0 \text{ for all } t \text{ and } s \text{ if } i \neq j \ .
\end{aligned}$$

Therefore, this framework allows for heteroskedasticity with respect to firms and correlation across time and the correlation across time can be firm specific.

Define the $T \times 1$ vector of errors for firm i as $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})^\top$, $i = 1, \ldots, n$, where we stack the errors over time for each firm. Then define the $T \times T$ error covariance matrix for each firm, Σ_i, as

$$\begin{aligned}
\Sigma_i &= \mathbb{E}[\mathbf{u}_i \mathbf{u}_i^\top | \mathbf{X}] \\
&= \sigma_{\alpha_i}^2 \boldsymbol{\iota}_T \boldsymbol{\iota}_T^\top + \sigma_{\varepsilon_i}^2 \mathbf{I}_T \\
&= \begin{pmatrix}
\sigma_{\alpha_i}^2 + \sigma_{\varepsilon_i}^2 & \sigma_{\alpha_i}^2 & \cdots & \sigma_{\alpha_i}^2 \\
\sigma_{\alpha_i}^2 & \sigma_{\alpha_i}^2 + \sigma_{\varepsilon_i}^2 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \sigma_{\alpha_i}^2 \\
\sigma_{\alpha_i}^2 & \cdots & \sigma_{\alpha_i}^2 & \sigma_{\alpha_i}^2 + \sigma_{\varepsilon_i}^2
\end{pmatrix}.
\end{aligned}$$

Therefore, the $nT \times nT$ error covariance matrix Ω is block diagonal as

$$\begin{aligned}\Omega &= \operatorname{diag}(\Sigma_1, \ldots, \Sigma_n) \\ &= \begin{pmatrix} \Sigma_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \Sigma_n \end{pmatrix}\end{aligned}$$

To estimate the random effects model of airline cost by GKRLS, first, we follow item 1 of the two step procedure outlined in Section 2. To get a consistent estimate of the error covariance matrix Ω, we can estimate the error variances using the residuals from the first step as

$$\begin{aligned}\widehat{\sigma}^2_{U_i} &= \frac{1}{T}\widehat{\mathbf{u}}_i^\top \widehat{\mathbf{u}}_i \\ \widehat{\sigma}^2_{\alpha_i} &= \frac{1}{T(T-1)/2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \widehat{u}_{it}\widehat{u}_{is} \\ \widehat{\sigma}^2_{\varepsilon_i} &= \widehat{\sigma}^2_{U_i} - \widehat{\sigma}^2_{\alpha_i}.\end{aligned}$$

Since averages are used to estimate the variances and by the law of large numbers $\widehat{\sigma}^2_{\alpha_i}$ and $\widehat{\sigma}^2_{\varepsilon_i}$ are consistent estimators of $\sigma^2_{\alpha_i}$ and $\sigma^2_{\varepsilon_i}$. Then, using these estimates for the error covariance, we follow item 2 of the two step procedure to get GKRLS estimates of the cost function.

In order to apply the asymptotic results established in Section 4, we must have $nT \to \infty$. Then, consistency and asymptotic normality of the GKRLS estimator under the random effects model discussed in Section 7 can be applied. In addition, since time averages are used to estimate the variances, we also must have $T \to \infty$. $T \to \infty$ is needed to apply the law of large numbers to get consistent estimates of $\sigma^2_{\alpha_i}$ and $\sigma^2_{\varepsilon_i}$. Since we assume that $T \to \infty$, it must be that $nT \to \infty$, and applying Theorems 3 and 4, the GKRLS estimator is consistent and asymptotically normal.