Kernel Ridge-Type Shrinkage Estimators in Partially Linear Regression Models with Correlated Errors

S. Ejaz Ahmed , Ersin Yilmaz * , Dursun Aydın

Posted Date: 8 April 2025

doi: 10.20944/preprints202504.0689.v1

Keywords: Shrinkage estimation; partially linear models; multicollinearity; ridge-type kernel smoother; parameter selection

Article

Kernel Ridge-Type Shrinkage Estimators in Partially Linear Regression Models with Correlated Errors

S. Ejaz Ahmed [1], Ersin Yilmaz [2,*,†] and Dursun Aydin [3,†]

[1] Brock University, Department of Mathematics and Statistics, St. Catharines, Ontario, L2S 3A

[2] Aalto University, Department of Computer Science, Konemiehentie 2, 02150 Espoo, Finland

[3] Department of Mathematics, University of Wisconsin Oshkosh, USA

* Correspondence: yilmazersin13@hotmail.com

† Mugla Sitki Kocman University, Faculty of Science, Department of Statistics, Mugla, Turkey, 4800.

Abstract: This paper introduces ridge-type kernel smoothing estimators for partially linear time-series models that employ shrinkage estimation to handle autoregressive errors and severe multicollinearity in the parametric component. By combining a generalized ridge penalty with kernel smoothing, the proposed estimators solve inflated variances arising from linear dependencies among predictors, while also accounting for autocorrelation. Four well-known selection criteria—Generalized Cross Validation (GCV), Improved Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC), and Risk Estimation via Classical Pilots (RECP)—are used to optimally choose both the bandwidth and shrinkage parameters. We provide closed-form expressions for these estimators, establish their asymptotic properties, and present a risk-based analysis that highlights the benefits of ordinary and positive-part shrinkage extensions. Simulation studies confirm that the introduced shrinkage approaches outperforms standard methods when predictors are strongly correlated, with this advantage growing as sample sizes increase. An application to airline delay time-series data further illustrates the efficacy and practical interpretability of the introduced methodology.

Keywords: shrinkage estimation; partially linear models; multicollinearity; ridge-type kernel smoother; parameter selection

MSC: 62G08; 62J07; 62M10

1. Introduction

In many time-series applications, one must balance interpretability and flexibility in modeling. Partially linear regression models achieve this by expressing the response as a parametric linear function of some covariates alongside a nonparametric smooth function of an additional variable. Specifically, we consider the following model setup:

$$y_t = \mathbf{x}_t^T \boldsymbol{\beta} + f(z_t) + \varepsilon_t, \ t = 1, \dots, n \tag{1.1}$$

where y_t denotes stationary response at time t, and its prediction depends on the p-dimensional vector of explanatory variables $\mathbf{x}_t = (x_{t1}, \dots, x_{tp})$ with $x_t = (x_1, \dots, x_n)'$ and an extra univariate variable z_t. Also, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown p-dimensional parameter vector to be estimated, $f(.)$ is an unknown smooth function, and $\varepsilon_i's$ are the error terms.

Unlike classical linear models, (1.1) offers the interpretability of $\boldsymbol{\beta}$ while retaining the capacity to model complex, nonlinear effects through $f(.)$ ([14]; Robinson, 1998). However, two mojor issues often arise:

1. Multicollinearity among $\mathbf{x}_t$. When the columns of $\mathbf{x}_t$ are nearly linearly dependent, ordinary least squares (OLS) or unpenalized nonparametric estimators can have extremely high variances [10].

 doi:10.20944/preprints202504.0689.v1

2. Autocorrelation or dependence among ε_t, as is common in time-series or longitudinal settings. We capture this by a first-order autoregressive process:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,\ |\rho| < 1,\ u_t \sim i.i.d.N(0,\sigma^2),\ t = 1,2,\dots,n \qquad (1.2)$$

ensuring that ε_t is stationarily dependent over time ([7,15]). Also, to simplify the notational illustration, a matrix–vector form of the model (1.1) can be stated as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \qquad (1.3)$$

where $\mathbf{y} = (y_1,\dots,y_n)$, $\mathbf{X} = (\mathbf{x}_1,\dots,\mathbf{x}_n)'$, $\mathbf{f} = \big(f(z_1),\dots,f(z_n)\big)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1,\dots,\varepsilon_n)$. The main goal is to estimate the unknown parameter vector $\boldsymbol{\beta}$, the smooth function $f(t)$ and the mean vector $\boldsymbol{\mu} = \mathbf{X\beta} + \mathbf{f}$ based on the observations from the data set $\{y_t,\ \mathbf{x}_t, z_t\}$. Partially linear models enable easier interpretation of the effect of each variable and owing to the "curse of dimensionality", these models can be preferred to a completely nonparametric regression model. Specifically, partially linear models are more practical than the classical linear model because they combine both parametric part presented numerically and nonparametric part displayed graphically.

Eventhough partially linear models have specific advantages, these models typically assume uncorrelated errors ([6,8,14]). In real-world time-series or panel data, ignoring (1.2) can degrade efficiency and produce biased inferences ([11,19]). Moreover, multicollinearity among $\mathbf{x}_t$ amplifies these problems, inflating variances to the point that $\boldsymbol{\beta}$ estimates become unreliable ([13,20]).

To mitigate these issues, we adopt ridge-type kernel smoothing for the parametric component. In essence, a shrinkage penalty is added so that the linear portion $\boldsymbol{\beta}$ is "shrunk" toward zero to stabilize estimation. Mathematically, for instance, a ridge-type kernel (RTK) estimator of $\boldsymbol{\beta}$ can be expressed (after suitable data transformation) as in (2.3) which has been extensively studied by [17]. This ridge-inspired framework stems from Stein-type shrinkage theory, which shows that biased estimators can dominate unbiased ones, particularly in moderate-to-high dimensions or under strong correlation ([1,2]). Indeed, Stein's paradox reveals that for $p \geq 3$ shrinking an estimator can lower the overall mean squared error despite introducing bias, a phenomenon extensively explored by [1,3]).

When errors follow (1.2), we further incorporate transformations to handle the autocovariance structure, yielding generalized ridge-type kernel (GRTK) estimators that optimally exploit the correlation $\boldsymbol{\rho}$. These GRTK methods align with earlier biased estimation approaches for linear or partially linear models with correlated errors ([13,20,21]). By merging the Stein-type shrinkage logic with kernel smoothing, one obtains robust estimators of both $\boldsymbol{\beta}$ and $f(.)$ that effectively address multicollinearity and autocorrelation in a unified fashion.

In this paper, we formalize and extend such modified ridge-type kernel smoothing estimators for partially linear time-series models. We focus on:

- The precise data transformation to accommodate (1.2).
- The interplay between shrinkage k and the bandwidth λ for kernel smoothing.
- Selecting these tuning parameters (k,λ) using multiple criteria, including GCV, AICc, BIC, and RECP.

This framework is motivated by [3] broader research on shrinkage-based variable selection and penalty methods, which shows that penalized procedures (including ridge, lasso, and Stein-like estimators) can vastly reduce variance in semiparametric regressions—particularly with correlated data or limited samples. Consequently, the main goal is to fill a gap in semiparametric regression by uniting autoregressive error modeling (1.2), handling with multicollinearity and Stein-type shrinkage in kernel-based estimation.

The rest of the paper is organized as follows. Section 2 explains how we fit the autoregressive structure (1.2) and handle collinearity. Section 3 provides the mathematical formulations of our GRTK shrinkage estimators for both $\boldsymbol{\beta}$ and the smooth function $f(.)$, including asymptotic properties. Section 4 details the procedures for choosing k and bandwidth λ. Section 5 illustrates,

 doi:10.20944/preprints202504.0689.v1

through extensive simulation and an airline delay time-series dataset, how the proposed methodology outperforms traditional methods under multicollinearity and autocorrelation. Finally, Section 6 concludes with a discussion of key findings and future research directions.

2. Fitting the Model Error Structure

Assume that $y_t = (y_1, \dots, y_n)'$ is a realization from a stationary time series described by model (1.1) with autoregressive error terms satisfying the following assumption.

Assumption 2.1: The error ε_t is a stationary dependent sequence with $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma_\varepsilon^2$, and $E|\varepsilon_t|^{2+\delta} < \infty$ for some constant δ. Letting $\boldsymbol{\varepsilon_t} = (\varepsilon_1, \dots, \varepsilon_n)$ and $Var(\boldsymbol{\varepsilon_t}) = E(\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_t}') = \sigma^2\boldsymbol{R}$, where $\boldsymbol{R}$ is a $(n \times n)$ positive definite symmetric matrix.

Here, Assumption 2.1 is standard in time-series analysis because it allows for modeling of stationary errors with finite variance. Such an assumption is common in many time-series models where the error term's statistical properties remain constant over time. Note that the autoregressive errors follow an $n-$dimensional multivariate normal distribution with a mean zero and stationary $(n \times n)$ covariance matrix $\boldsymbol{\Sigma}$. We may also write this expression in the equivalent form $\boldsymbol{\varepsilon_t} = (\varepsilon_1, \dots, \varepsilon_n) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{R}$ is a $n \times n$ covariance matrix, given by

$$\boldsymbol{R} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix} \tag{2.1a}$$

with the inverse

$$\boldsymbol{R^{-1}} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & -\rho & 1 \end{bmatrix} \tag{2.1b}$$

where ρ denotes the correlation between ε_t and ε_{t-1}, as defined before. It should be emphasized that **once** we have an estimate $\widehat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$ and an estimator $\hat{f}$ of the unknown smooth function f, the parameter ρ can be estimated by the residuals from the semiparametric regression $e_t = y_t - \left(\mathbf{x}_t\widehat{\boldsymbol{\beta}} + \hat{f}(z_t)\right)$, defined as

$$\hat{\rho} = \frac{Cov\,(e_t, e_{t-1})}{\sqrt{Var(e_t)}\sqrt{Var(e_{t-1})}} = \frac{\sum_{t=2}^{n} e_t\, e_{t-1}}{\sum_{t=1}^{n} e_t^2} \tag{2.2}$$

There are several methods to estimate $\boldsymbol{\beta}$ and f. In this work, we adopt the ridge type kernel (*RTK*) method discussed by [17], generalizes partial kernel method proposed by [14]. Specifically, the *RTK* estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the form

$$\widehat{\boldsymbol{\beta}}_{RTK} = \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\tilde{\mathbf{y}} \tag{2.3}$$

where $k > 0$ is a shrinkage parameter, $\mathbf{I}$ is an $(p \times p)$ identity matrix, $\tilde{x}_t = x_t - \sum_{i=1}^{n} W_{\lambda t}(z_i)x_i = \left\{(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X} = \widetilde{\mathbf{X}}\right\}$, $\tilde{y}_t = y_t - \sum_{i=1}^{n} W_{\lambda t}(z_i)y_i = \{(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y} = \tilde{\mathbf{y}}\}$, and $\mathbf{W}_\lambda$ is a kernel smoother matrix of weights satisfying the satisfying Assumption (2.3), given by

$$W_{\lambda t}(z) = K\left(\frac{z_t - z}{\lambda}\right) \Big/ \sum_{t=1}^{n} K\left(\frac{z_t - z}{\lambda}\right) = \mathbf{W}_\lambda \tag{2.4}$$

where λ is a bandwidth parameter to selected, and $K(.)$ is kernel function defined in Remark 2.2. For example, $K(.)$ could be Gaussian, Epanechnikov, etc. Then, an estimator $\hat{f}$ of the nonparametric part $f(z)$ is obtained by

$$\hat{\mathbf{f}}_{RTK} = \sum_{i=1}^{n} W_{\lambda t}(z)\left(y_t - \mathbf{x}_t\hat{\beta}_{RTK}\right) = \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{RTK}\right) \tag{2.5}$$

 doi:10.20944/preprints202504.0689.v1

Hence, the autoregressive coefficient $\hat{\rho}$ arises from (2.2) and the noise of the AR(1) is obtained by $\hat{u}_t = e_t - \hat{\rho}e_{t-1}$. Note that since the errors in the model (1.1) are serially correlated, the estimators defined in (2.3) and (2.5) is not asymptotically efficient. To improve efficiency, we use kernel ridge type weighted estimation based on transforming data $\tilde{\mathbf{y}}$.

3. Generalized Ridge Type Kernel Estimation

The generalized ridge type kernel (*GRTK*) estimator of the parametric components in the model (1.1) is constructed by combining the partial kernel method with a suitable transformation to account for ρ in the AR(1) errors. We will then extend this GRTK estimator to incorporate shrinkage estimation, analyze its asymptotic properties, and discuss its statistical characteristics. For simplicity, assume the true parameter vector $\boldsymbol{\beta}$ and true error covariance matrix $\mathbf{R}$ are given. From the $E(\varepsilon_t) = 0$, one can see that $f(z_t) = y_t - \mathbf{x}_t\boldsymbol{\beta}$, $t = 1, \dots, n$. Thus, for a given $\boldsymbol{\beta}$, the natural estimator of the nonparametric component $f(.)$ is

$$\hat{\mathbf{f}} = \sum_{t=1}^{n} W_{\lambda t}(z)\left(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}\right) = \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{3.1}$$

where $\mathbf{W}_\lambda$ is a kernel smoother matrix, as defined in (2.4). Since $\mathbf{R}$ defined in Assumption (2.1) is positive definite, there exists a $(n \times n) -$ dimensional matrix $\mathbf{P}$ such that $\mathbf{P}'\mathbf{P} = \mathbf{R}^{-1}$. We first assume that the matrix $\mathbf{R}$ and, hence, $\mathbf{P}$ are known. Then, we fit the error structure by the following residuals:

$$e_t = y_t - \mathbf{x}_t\boldsymbol{\beta} - \hat{f} = \tilde{y}_t - \tilde{\mathbf{x}}_t\boldsymbol{\beta}, 1 \le t \le n \tag{3.2}$$

Here, as defined in (2.3), $\tilde{y}_t$ and $\tilde{\mathbf{x}}$ are locally centered residuals of y_t and $\mathbf{x}$, respectively. By considering these partial residuals, the *GRTK* estimator of the vector $\boldsymbol{\beta}$ is obtained by minimizing the following weighted least squares (WLS) equation.

$$\begin{aligned} WLS(\boldsymbol{\beta}) &= \left(\mathbf{P}(\tilde{y}_t - \tilde{\mathbf{x}}_t\boldsymbol{\beta})\right)'\left(\mathbf{P}(\tilde{y}_t - \tilde{\mathbf{x}}_t\boldsymbol{\beta})\right) + k\boldsymbol{\beta}'\boldsymbol{\beta} \\ &= \left\|\mathbf{P}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)\right\|_2^2 + k\|\boldsymbol{\beta}\|_2^2 \end{aligned} \tag{3.3}$$

where $\|.\|_2$ denotes the Euclidean norm, $k > 0$ is a shrinkage parameter to be selected by a selection method, $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$ and $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}$ are the centered residuals, as defined in (2.3). Also, $\mathbf{W}_\lambda$ is a kernel smoothing matrix based on a bandwidth parameter $\lambda > 0$ to be selected.

Algebraically, after some operations, the solution of the equation $WLS(\boldsymbol{\beta})$ yields the following *GRTK* estimator:

$$\hat{\boldsymbol{\beta}}_{GRTK}(k) = \left(\tilde{\mathbf{X}}'\mathbf{R}^{-1}\tilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\tilde{\mathbf{X}}'\mathbf{R}^{-1}\tilde{\mathbf{y}} \tag{3.4}$$

Replacing $\boldsymbol{\beta}$ in the (3.1) with the $\hat{\boldsymbol{\beta}}_{GRTK}(k)$ in (3.4) to produce an *GRTK* estimator of the nonparametric component in the partially linear model $f(.)$ given by

$$\hat{\mathbf{f}}_{GRTK} = \sum_{t=1}^{n} W_{\lambda t}(z)\left(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}_{GRTK}(k)\right) = \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GRTK}(k)\right) \tag{3.5}$$

This section details the construction of the Generalized Ridge-Type Kernel (GRTK) estimator for the parametric component in model (1.1). This is achieved by combining the partial kernel method with a suitable transformation that accounts for ρ in the AR(1) errors. We will then extend this GRTK estimator to incorporate shrinkage estimation, analyze its asymptotic properties, and discuss its statistical characteristics.

3.1. Shrinkage Estimators with GRTK Estimators

While the GRTK estimator in (3.4) addresses multicollinearity and autocorrelation simultaneously, further variance reduction is possible via Stein-type shrinkage. By combining the full

 doi:10.20944/preprints202504.0689.v1

model GRTK estimator with a more parsimonious submodel estimator, we can develop Stein-type shrinkage estimators that may achieve lower risk under certain conditions. This approach follows the framework developed by [3] and extends it to the partially linear time-series context. To develop the shrinkage estimators, we consider two models:

1. *Full Model (FM)*: The complete model with all predictors $(p = p_1 + p_2)$ where $p_1 < p$ is the number of significant coefficients and p_2 denotes the number of sparse coefficients, estimated using GRTK as in equation (3.4).

2. *Submodel (SM)*: A reduced model with dimension $(p_1 < p)$, selected via Bayesian information criterion (BIC) to minimize model selection, yielding $\widehat{\boldsymbol{\beta}}_{SM}$.

Let $\widehat{\boldsymbol{\beta}}_{FM}$ be the parametric estimate from the ridge-type kernel smoothing (GRTK) on all predictors as full model $\widehat{\boldsymbol{\beta}}_{GRTK}(k)$ from equation (3.4) and let $\widehat{\boldsymbol{\beta}}_{SM}$ be the corresponding estimate when only a submodel of predictors is retained using BIC. Accordingly, ordinary shrinkage estimator $(\widehat{\boldsymbol{\beta}}_S)$ can be give as follows:

$$\widehat{\boldsymbol{\beta}}_S = \widehat{\boldsymbol{\beta}}_{FM} - \left[(p_2 - 2)/\hat{T}\right](\widehat{\boldsymbol{\beta}}_{FM} - \widehat{\boldsymbol{\beta}}_{SM}), \tag{3.6}$$

where $\hat{T}$ is a "distance measure" given below, and $(p_2 - 2)/\hat{T}$ parallels the classical shrinkage factor. Intuitively, we are pulling $\widehat{\boldsymbol{\beta}}_{FM}$ toward $\widehat{\boldsymbol{\beta}}_{SM}$, thus introducing some bias but potentially reducing variance and risk. The distance is given by:

$$\hat{T} = n\hat{\sigma}^{-2}\widehat{\boldsymbol{\beta}}_2'\left[\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{U}}_1\widetilde{\mathbf{X}}_2\right]\widehat{\boldsymbol{\beta}}_2\,,$$

where $\widehat{\boldsymbol{\beta}}_2'$ is the portion of $\widehat{\boldsymbol{\beta}}_{FM}$ (or $\widehat{\boldsymbol{\beta}}_{SM}$) corresponding to the submodel's parametric indices. $\widetilde{\mathbf{X}}_2$ is the part of the design matrix associated with those sparse indices. $\widetilde{\mathbf{U}}_1 = \mathbf{I}_n - \widetilde{\mathbf{X}}_1\left[\widetilde{\mathbf{X}}_1'\widetilde{\mathbf{X}}_1 + k\mathbf{I}_{p_1}\right]^{-1}\widetilde{\mathbf{X}}_1'$ is a partial residual projection, removing the other block of parametric covariates from the fit. $\hat{\sigma}^2$ is given in (3.17) but note that $\mathbf{H}_{\lambda,k}$ is chosen SM-based partial kernel fit (see [21] for details).

To prevent over-shrinking when $(p_2 - 2)/\hat{T}$ is large or negative, we define a positive-part version:

$$\widehat{\boldsymbol{\beta}}_{PS} = \widehat{\boldsymbol{\beta}}_{FM} - \left[\frac{p_2-2}{\hat{T}}\right]^{+}(\widehat{\boldsymbol{\beta}}_{FM} - \widehat{\boldsymbol{\beta}}_{SM}), \tag{3.7}$$

where $(x)^+ = \max\{0, x\}$. Thus, if $(p_2 - 2)/\hat{T}$ is negative, we do no shrinkage; if it is positive, we shrink as in (3.6).

Following a partially linear kernel-smoothing approach (GRTK), the nonparametric function $f(.)$ can be recomputed (or simply computed once) based on these final shrunk parametric estimates using smoothing matrix $\mathbf{W}_\lambda$:

$$\hat{\mathbf{f}}_S = \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_S), \tag{3.8a}$$

and

$$\hat{\mathbf{f}}_{PS} = \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{PS}), \tag{3.8b}$$

Hence, each final semiparametric estimators become $(\widehat{\boldsymbol{\beta}}_S, \hat{\mathbf{f}}_s), (\widehat{\boldsymbol{\beta}}_{PS}, \hat{\mathbf{f}}_{Ps})$.

3.2. Asymptotic Distribution of the Estimators

This subsection presents the assumptions and theorems necessary to analyze the asymptotic properties of the proposed $\widehat{\boldsymbol{\beta}}_{GRTK}(k)$ and $\hat{f}_{GRTK}(k)$ estimators. Understanding the asymptotic distribution is crucial for statistical inference. In this context, we introduce the following assumptions that can be easily satisfied.

Assumption 3.1. As in setting of the semiparametric regression model, covariates x_i and z_i are related via the following nonparametric regression model.

 doi:10.20944/preprints202504.0689.v1

$$x_{ij} = g_j(z_i) + \eta_{ij}\,, (1 \le i \le n, 1 \le j \le p) \tag{a}$$

where $\eta_{ij} = (\eta_{i1}, \dots, \eta_{ip})'$ are real sequence satisfying.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \eta_i \eta_i' = \boldsymbol{B} \tag{b}$$

and

$$max_{1\le j\le n} \| \sum_{i=1}^{n} W_{\lambda i}(z_i)\eta_i \| = O(n^{-1/6}) \tag{c}$$

where $\mathbf{B}$ is a $(p \times p)$ positive definite matrix and $\|.\|$ indicates the Euclidean norm.

Assumption 3.2. The functions $g(.)$ and $f_j(.), 1 \le j \le k,$ satisfy a Lipschitz continuous of order 1 on data domain.

Assumption 3.3. The weight functions $W_{\lambda i}$ satisfy these conditions:

(i.) $max_{1\le i\le n} \sum_{j=1}^{n} W_{\lambda i}(z_j) = O(1)$

(ii.) $max_{1\le i,j\le n} W_{\lambda i}(z_j) = O(n^{-2/3})$

(iii.) $max_{1\le j\le n} \sum_{i=1}^{n} W_{\lambda i}(z_j) I\left(|z_i - z_j| > O(n^{-1/3})\right) = O(n^{-1/3})$, where $I(A)$ is the indicator function of a set A.

(iv.) $tr(W_\lambda' W_\lambda) = tr(W_\lambda) = O(\lambda^{-1})$, where $tr(A)$ denotes the trace of a square matrix A.

Remark 3.1: The η_{ij} stated in (a) of Assumption 3.1 behaves like as a zero mean uncorrelated sequence of stationary random variables of independent of ε_t. If the covariates x_i and z_i are the observations of the independent and identically distributed random variables, then $g_j(z)$ can be considered as $E(x_{ij}|z_i) = g_j(z_i)$, $E(\eta_i \eta_i') = \boldsymbol{B}$, and $\eta_{ij} = \left(x_{ij} - g_j(z_i)\right)$. See [14] for more details. In this case, (b) holds with probability 1 according to the law of large numbers, and (c) is provided by Lemma 1 in [16].

Remark 3.2: If $K(x)$ is a kernel function, then $K(x) = K\left(\frac{x}{\lambda}\right)$ is also a kernel function based on a positive bandwidth parameter λ, and it satisfies the following properties:

(a) $\int K(x)dx = 1$, (b) $\int xK(x)dx = 0$,
(c) $\mu_2 = \int x^2 K(x)dx < \infty$, (d) $K(x) \ge 0$ for all x, and $K(x) = K(-x)$ (3.9)

These properties show that a kernel function needs to be symmetric and continuous probability density function with mean zero and constant variance. Note that the bandwidth parameter λ should be chosen optimally by a selection criterion in kernel estimation or smoothing. For example, a large λ provides an extremely smooth curve or estimate, while a small λ produces a wiggly function curve.

Theorem 3.1: Suppose that the Assumption 2.1 and Assumptions 3.1–3.3 hold, and assume that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \eta_i \boldsymbol{R}^{-1} \eta_i' = \boldsymbol{V}$ is a non-singular covariance matrix with $\eta_i = (\eta_1, \dots, \eta_n)'$. Then, as $n \to \infty$, we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{GRTK}(k) - \boldsymbol{\beta}\right) \xrightarrow{D} N(0, \sigma^2 \boldsymbol{V}^{-1}) \text{ and } \sqrt{n}(\hat{\rho} - \rho) \xrightarrow{D} N\left(0, n^{-1}(1-\rho^2)\right)$$

where $\xrightarrow{D}$ denotes convergence in distribution. See [18] for the proof of the Theorem 3.1.

The asymptotic normality established in Theorem 3.1 allows for valid statistical inference based on the GRTK estimator. Intuitively, this result holds because the transformed data approach effectively accounts for autocorrelation, while the kernel smoothing effectively separates the nonparametric component, allowing the parametric component to be estimated consistently with a regular $\sqrt{(n)}$ convergence rate.

Theorem 3.2: If the Assumptions 2.1 and 3.1–3.3 hold, then it hold

 doi:10.20944/preprints202504.0689.v1

$$max_{1\le i\le n}\left|\hat{f}_{GRTK}(z_i)-f(z_i)\right| = O\left(n^{-1/3}\log(n)\right) \text{ a.s.}$$

See [18] for proof of the Theorem 3.2. The asymptotic normality and convergence results established in Theorems 3.1 and 3.2 provide a theoretical foundation for the GRTK estimators. In the following section, we analyze additional statistical properties, such as bias and variance.

This convergence rate for the nonparametric component is optimal in the sense that no estimator can achieve a faster uniform convergence rate without additional structural assumptions. The logarithmic factor arises from the need to establish uniform rather than pointwise convergence.

Note that Theorem 3.2 express that $\hat{f}_{GRTK}$ reaches the optimal strong convergence rate. Also, a consistent estimator of the asymptotic covariance matrix $\boldsymbol{V}$ is required for statistical inference based on $\widehat{\boldsymbol{\beta}}_{GRTK}(k)$. The estimate of covariance matrix $\boldsymbol{V}$ can be obtained as

$$\widehat{\boldsymbol{V}} = \frac{1}{n}\left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}}\right) \tag{3.10}$$

where $\widehat{\mathbf{V}}$ is the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{GRTK}(k)-\boldsymbol{\beta}\right)$. Regarding the shrinkage estimator's asymptotic properties for the currently considered low-dimensional ($p<n$) settings, we outline the key asymptotic results. Let $p_2 \ge 3$ be the size of the submodel portion. Suppose β_2 is subject to a local alternatives framework:

$$\beta_{2n} = \frac{1}{\sqrt{n}}\vartheta, \text{ for fixed } \vartheta \in \mathbb{R}^{p_2},$$

So that $\beta_{2n}\to 0$ as $n\to\infty$. Additionally, let β_{1n} denotes the complementary block (nonzero coefficients) of dimension p_1. Under standard regularity conditions ([4,21]) we can claim that:

- Both $\widehat{\boldsymbol{\beta}}_S$ and $\widehat{\boldsymbol{\beta}}_{PS}$ remain consistent for $\boldsymbol{\beta}$ as a point of consistency
- By construction, these estimators introduce shrinkage-based bias in exchange for variance reduction. The bias depends on $(p_2-2)/\hat{T}$ or $\left[\frac{(p_2-2)}{\hat{T}}\right]^+$ and and generally involves certain noncentral χ^2-based expectations.
- Regarding the asymptotic quadratic bias (AQDB), asymptotic covariance and asymptotic distributional risk (ADR), similar expansions hold showing how $\widehat{\boldsymbol{\beta}}_S$ and $\widehat{\boldsymbol{\beta}}_{PS}$ incorporate both ridge (FM) plus Stein corrections.

Exact closed-form expressions for the bias, covariance, and risk match those in [21] or [3], with minor notational changes for $\hat{T}$ and the partial kernel smoothing. Hence, we omit rewriting them here. However, for completeness, detailed bias and risk expansions for these estimators appear in Appendix B. The main takeaway is that $\widehat{\boldsymbol{\beta}}_{PS}$ typically achieves the smallest asymptotic distributional risk among the four choices (FM, SM, ordinary Stein, positive-part Stein) whenever $(p_2-2)/\hat{T}$ would become negative or large, thus showing the advantage of positive-part truncation in a partially linear framework.

3.3. Statistical Properties of the Estimators

Here, we detail the statistical properties of the GRTK estimator, including its bias, variance, and expected value. These properties help to characterize the estimator's behavior and quality. As defined in (3.4), when $k=0$, the *GRTK* estimators of the parametric and nonparametric components reduce to ordinary kernel smoothing (*KS*) estimators for a partially linear model with correlated error. They can be defined, respectively, as follows:

$$\widehat{\boldsymbol{\beta}}_{KS} = \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{y}} \quad \text{and} \quad \hat{\mathbf{f}}_{KS} = \mathbf{W}_\lambda\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{KS}\right) \tag{3.11}$$

Using the abbreviation

$$G_k = \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}}+k\mathbf{I}\right)^{-1}$$

Expected value, bias and variance of the estimator $\widehat{\boldsymbol{\beta}}_{GRTK}$ can be defined, respectively, as

$$E\left[\widehat{\boldsymbol{\beta}}_{GRTK}(k)\right] = G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}}\,E[\boldsymbol{\beta}] + G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}}$$

 doi:10.20944/preprints202504.0689.v1

$$= \boldsymbol{\beta} - k\, G_k\boldsymbol{\beta} + G_k\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}}$$

(3.12)

$$Bias\left(\hat{\boldsymbol{\beta}}_{GRTK}(k)\right) - \boldsymbol{\beta} = G_k\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} - k\, G_k\boldsymbol{\beta} \tag{3.13}$$

$$Var\left(\hat{\boldsymbol{\beta}}_{GRTK}(k)\right) = \sigma^2 G_k\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I}-\mathbf{W}_\lambda)^2\boldsymbol{R}^{-1}\tilde{\mathbf{X}}\; G_k \tag{3.14}$$

The implementation details of Equations (3.12) - (3.14) are given in Appendix A1.

Clearly, $E\left[\hat{\boldsymbol{\beta}}_{GRTK}(k)\right] \neq \boldsymbol{\beta}$ for any > 0 . Hence, the *GRTK* estimator is a biased estimator. From the expression above it is clear that the expectation of the *GRTK* estimator vanishes as k tends to infinity

$$\lim_{k\to\infty}\hat{\boldsymbol{\beta}}_{GRTK}(k) = \left(\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} = 0$$

Hence, all coefficients of the parametric component are shrunken towards zero as the ridge (or shrinkage) parameter increases.

As shown in the (3.14), although the smoothing method provides the estimates of the components in the model, they do not directly provide an estimate of the variance of the error terms (i.e. σ^2). In a general partially linear model, the estimate of variance σ^2 can be found by the residual sum of squares

$$\begin{aligned} RSS = (\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}}) &= \left(\mathbf{y} - \left(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}}\right)\right)'\left(\mathbf{y} - \left(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}}\right)\right) \\ &= \left(\mathbf{y} - \mathbf{H}_{\lambda,\boldsymbol{k}}\mathbf{y}\right)'\left(\mathbf{y} - \mathbf{H}_{\lambda,\boldsymbol{k}}\mathbf{y}\right) = \left\|(\mathbf{I}-\mathbf{H}_{\lambda,k})\mathbf{y}\right\|_2^2 \end{aligned} \tag{3.15}$$

where $\mathbf{H}_{\lambda,k} = \mathbf{W}_\lambda + (\mathbf{I}-\mathbf{W}_\lambda)^2\mathbf{X}G_k\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}$ is a hat matrix which depends on a smoothing parameter $\lambda > 0$ and a shrinkage parameter $k > 0$ for partially linear regression model. Note that the fitted values of the model defined in the (1.1) is obtained by this matrix $\mathbf{H}_{\lambda,k}$, given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \left[\mathbf{W}_\lambda + (\mathbf{I}-\mathbf{W}_\lambda)^2\mathbf{X}G_k\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\right]\mathbf{y} = \mathbf{H}_{\lambda,k}\mathbf{y} \tag{3.16}$$

where $G_k = \left(\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}$, as defined in equations (3.12-3.14). The implementation details of Equation (3.16) is given in Appendix A2. Thus, similar to ordinary least squares regression, estimation of the error variance can be stated as

$$\hat{\sigma}^2 = \frac{\left\|(\mathbf{I}-\mathbf{H}_{\lambda,k})\mathbf{y}\right\|_2^2}{tr(\mathrm{I}-\mathbf{H}_{\lambda,k})^2} = \frac{RSS}{n-p} \tag{3.17}$$

where $tr(\mathrm{I}-\mathbf{H}_{\lambda,k})^2 = n - tr\left[2\mathbf{H}_{\lambda,k} - \left(\mathbf{H}_{\lambda,k}\right)'\left(\mathbf{H}_{\lambda,k}\right)\right] = n-p$ denotes the residual degrees of freedom (see, for example, [5]). It appears from the denominator of equation (3.17) that the degrees of freedom for RSS can also be expressed as the sample size minus the number of estimated parameters in the model.

3.4. Assessing the Risk and Efficiency

To further evaluate the estimators, we now consider measures of risk and efficiency. As discussed earlier, $Var(\hat{\boldsymbol{\beta}})$ and $Bias(\hat{\boldsymbol{\beta}})$ important criteria for assessing estimator quality. This section introduces the Mean Squared Deviation (MSD) and quadratic risk function to compare the performance of different estimators. In general, the ill-effect of $Bias(\hat{\boldsymbol{\beta}})$ is known as information loss, and this loss can be measured by a function stated as the mean squared deviation (*MSD*). Note that the *MSD* is a risk function corresponding to the expected value of the squared error loss for an estimator. In terms of a squared error loss, the MSD can be defined as a matrix consisting of the sum of the variance-covariance matrix and the squared bias. While the MSD matrix provides comprehensive information about estimation quality, comparing matrices directly is challenging. Therefore, we derive a scalar measure by taking the trace of the MSD matrix, which represents the sum of mean squared errors across all coefficient estimates:

 doi:10.20944/preprints202504.0689.v1

$$MSD(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) = Var(\widehat{\boldsymbol{\beta}}) + Bias(\widehat{\boldsymbol{\beta}})^2 \tag{3.18}$$

The equation (3.18) gives detailed information about the quality of an estimator. In addition to the MSD matrix, the expected loss, called the scalar-valued version, can also be used to compare different estimators. For convenience, we will work with the scalar valued mean squared deviation.

Definition 3.1: The quadratic risk function of an estimator $\widehat{\boldsymbol{\beta}}$ of the vector $\boldsymbol{\beta}$ is defined as the scaler valued version of the mean squared deviation matrix (*SMDE*), given by

$$R(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}, \boldsymbol{V}) = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{V}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) = tr\left(\boldsymbol{V}\left(MSD(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\right)\right) = SMSD(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$$

where $\boldsymbol{V}$ is a $(p \times p)$ symmetric and non-negative definite matrix. Based on the above risk function, we can define the following criterion to compare estimators.

Definition 3.2: Let the vectors $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ be the two competing estimators of a parameter vector $\boldsymbol{\beta}$. If the difference of the matrices MSD is non-negative definite, it can be said that the estimator $\widehat{\boldsymbol{\beta}}_2$ is superior to estimator $\widehat{\boldsymbol{\beta}}_1$, and is given by

$$\Delta(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2) = MSD(\widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) - MSD(\widehat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}) \geq 0$$

Accordingly, we get the following result.

Theorem 3.2 ([22]): Let the vectors $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ be the two different estimators of a parameter vector $\boldsymbol{\beta}$**.** Therefore, the following two expressions are equivalent

(i.) $R(\widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}, \boldsymbol{V}) - R(\widehat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}, \boldsymbol{V}) \geq 0$ for all non-negative definite matrices $\boldsymbol{V}$

(ii.) $MSD(\widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) - MSD(\widehat{\boldsymbol{\beta}}_2, \boldsymbol{\beta})$ is a non-negative definite matrix.

The results of Theorem 3.2 denotes that $\widehat{\boldsymbol{\beta}}_2$ has a smaller $MSD(\widehat{\boldsymbol{\beta}}_2, \boldsymbol{\beta})$ than the estimator $\widehat{\boldsymbol{\beta}}_1$ if and only if the $R(\widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}, \boldsymbol{V})$ of $\widehat{\boldsymbol{\beta}}_2$ averaging over every quadratic risk function is less than that of the estimator $\widehat{\boldsymbol{\beta}}_1$. Thus, the superiority of $\widehat{\boldsymbol{\beta}}_2$ over $\widehat{\boldsymbol{\beta}}_1$ can be observed by comparing the matrices MSD.

Substituting equations (3.13) and (3.14) in equation (3.18), the matrix MSD of the proposed estimator $\widehat{\boldsymbol{\beta}}_{GRTK}(k)$ is obtained as

$$\begin{aligned} MSD(\widehat{\boldsymbol{\beta}}_{GRTK}(k), \boldsymbol{\beta}) &= Var(\widehat{\boldsymbol{\beta}}_{GRTK}(k)) + Bias(\widehat{\boldsymbol{\beta}}_{GRTK}(k))^2 \\ &= G_k\left(\sigma^2\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)^2\,\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1} + \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} - k\,\boldsymbol{\beta}\right)^2\right)G_k \end{aligned}$$

Furthermore, considering Definition 3.1, the quadratic risk function for $\widehat{\boldsymbol{\beta}}_{GRTK}(k)$ can be stated as follows:

$$\begin{aligned} SMSD(\widehat{\boldsymbol{\beta}}_{GRTK}(k), \boldsymbol{\beta}) &= R(\widehat{\boldsymbol{\beta}}_{GRTK}(k), \boldsymbol{\beta}, \boldsymbol{V}) \\ &= tr\left[\boldsymbol{V}\left(MSD(\widehat{\boldsymbol{\beta}}_{GRTK}(k), \boldsymbol{\beta})\right)\right] \\ &= tr\left[MSD(\widehat{\boldsymbol{\beta}}_{GRTK}(k), \boldsymbol{\beta})\right] \\ &= tr\Big[G_k\Big(\sigma^2\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)^2\,\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1} \\ &\qquad + \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} - k\,\boldsymbol{\beta}\right)'\left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} - k\,\boldsymbol{\beta}\right)\Big)G_k\Big] \end{aligned}$$

The theoretical properties established in this section demonstrate that GRTK estimators effectively balance bias and variance while accounting for both multicollinearity and error autocorrelation. However, the practical performance of these estimators depends critically on the selection of appropriate tuning parameters (λ, k). In the following section, we address this challenge by introducing several parameter selection criteria.

4. Choosing the Penalty Parameters

This section focuses on selecting the bandwidth parameter λ and the shrinkage parameter k, both of which are crucial components of the generalized weighted least squares equation (3.3). The goal is to determine the optimal values for λ and k. To achieve this, we employ parameter selection

 doi:10.20944/preprints202504.0689.v1

methods. The parameters λ and k are chosen by minimizing specific criteria. The most widely used selection criteria are summarized below:

Generalized cross- validation: The optimal bandwidth λ and shrinkage parameter k can be determined by implementing the GCV score function:

$$(\hat{\lambda}, \hat{k}) = \underset{\lambda>0,\ k>0}{\operatorname{argmin}}\ \boldsymbol{GCV}(\lambda, k) = \underset{\lambda>0,\ k>0}{\operatorname{argmin}} \left(\frac{n^{-1}\left\|(\mathbf{I}-\mathbf{H}_{\lambda,k})\mathbf{y}\right\|^2}{\left[n^{-1}tr(\mathbf{I}-\mathbf{H}_{\lambda,k})\right]^2} \right) \tag{4.1}$$

where $\mathbf{H}_{\lambda,k} = \mathbf{W}_\lambda + (\mathbf{I}-\mathbf{W}_\lambda)^2\mathbf{X}G_k\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}$ with $G_k = \left(\tilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}$, as is defined in the equations (3.15) and (3.16), is the smoother matrix based on the parameters λ and k.

Improved Akaike information criterion: To eliminate overfitting when the sample size is relatively small, [9] proposed AIC_c, an improved version of the Classical Akaike Information Criterion

$$\boldsymbol{AIC_c}(\lambda, k) = \underset{\lambda>0,\ k>0}{\operatorname{argmin}} \left\{ log\left(\frac{\left\|(\mathbf{I}-\mathbf{H}_{\lambda,k})\mathbf{y}\right\|^2}{n-tr(\mathbf{H}_{\lambda,k})} \right) + 1 + \frac{2\{tr(\mathbf{H}_{\lambda,k})+1\}}{n-tr(\mathbf{H}_{\lambda,k})-2} \right\} \tag{4.2}$$

Bayesian Information Criterion: Schwarz's criterion, also known as the BIC is another statistical measure for selection of the penalty parameters. The BIC criterion is expressed as

$$\boldsymbol{BIC}(\lambda, k) = \underset{\lambda>0,\ k>0}{\operatorname{argmin}} \left(\frac{\left\|(\mathbf{I}-\mathbf{H}_{\lambda,k})\mathbf{y}\right\|^2}{n-tr(\mathbf{H}_{\lambda,k})} \right) + \left(\frac{log(n)}{n} \right) tr(\mathbf{H}_{\lambda,k}) \tag{4.3}$$

Risk estimation using classical pilots (RECP): The key idea now is to estimate the risk $R(\mathbf{y}, \hat{\mathbf{y}})$ by plugging-in pilot estimates of $\mathbf{y}$ and σ^2 into (4.4), and choose the λ and k that minimizes the criterion (see [12])

$$R(\mathbf{y}, \hat{\mathbf{y}}) = \underset{\lambda>0,\ k>0}{\operatorname{argmin}} \frac{1}{n-tr(\mathbf{H}_{\lambda,k})} \left\{ \left\|(\mathbf{H}_{\lambda,k}-\mathbf{I})\mathbf{y}\right\|^2 + \sigma^2 tr(\mathbf{H}_{\lambda,k}{}'\mathbf{H}_{\lambda,k}) \right\} \tag{4.4}$$

Each of these four criteria offers a valid approach for parameter selection; however, they have different characteristics in practice. GCV generally provides a good balance between bias and variance across different sample sizes. AIC_c is generally more suitable for smaller samples, where overfitting is a concern. BIC tends to produce more parsimonious models and is often preferred when the true model is believed to be sparse. RECP often provides good fits for the nonparametric component but can be more sensitive to the correlation structure. The simulation study in the next section provides further guidance on criterion selection under different data scenarios.

5. Numerical Examples

5.1. Simulation Study

This section presents a Monte Carlo simulation study designed to compare the estimation performances of the introduced kernel-type ridge estimator (GRTK) and shrinkage estimators for model (1.1) in the presence of correlated errors. To generate parametric predictors, we use a sparse model with $p = 20$. To introduce multicollinearity, covariates are generated with a specific level of collinearity, denoted as δ, the following equation is used:

$$x_{ij} = \sqrt{(1-\delta^2)}d_{ij} + \delta d_{ip}, i = 1, \dots, n,\ \ j = 1, \dots, p \tag{5.1}$$

where p is the number of predictors, d_{ij}'s generated as $d_{ij} \sim N(\mu_d = 0, \sigma_d^2 = 1)$, and $\delta = (0.95, 0.99)$ denotes the two correlation levels between the predictors of parametric component. In this context, simulated data sets are generated from the following model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + f(z_t) + \varepsilon_t,\ \ t = 1,2, \dots, n, \tag{5.2}$$

as defined in (1.1). In given model (5.2), $\boldsymbol{\beta} = \left(-1,\ 1.5,\ 2,\ 5, \mathbf{0}_{p_2}\right)'$, $\mathbf{x}_t'$ is constructed by equation (5.1); the regression function $f(z_t) = 6z_t \sin(z_t)^2$ and $z_t = 5(t-0.5)/n$; the error terms ε_t are generated using a first order autoregressive process (that is, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$) with $\rho = 0.5$ and $u_t \sim NIID(0, \sigma_u^2 =$

 doi:10.20944/preprints202504.0689.v1

1). Using data generation procedure given by equations (5.1) and (5.2), we consider the three different sample sizes that are $n = 50, 150$ and 300 to investigate the performance of the introduced estimators for low, medium and large samples, and each we use 1000 repetition for each simulation combination. The simulation results are provided in the following figures and tables. In addition, Figure 1 shows the generated data to clarify the data generation procedure better.

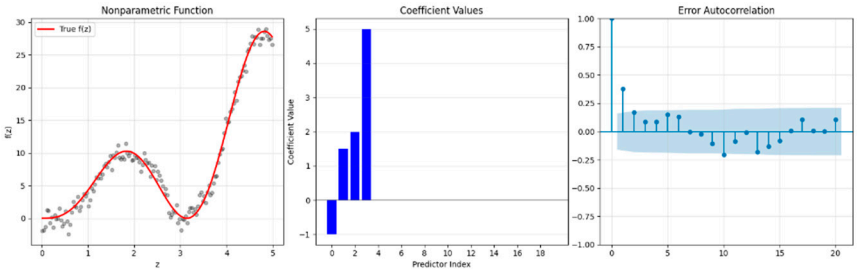

Figure 1. Simulated data with true $f(z_t)$, sparse coefficients with zero and non-zero ones and autocorrelation function of error terms for $n = 150, \delta = 0.95.$.

From a computational perspective, the GRTK estimation procedure involves several steps: (1) initial parameter estimation to obtain residuals, (2) estimation of the autocorrelation parameter ρ, (3) data transformation, (4) parameter selection for (λ, k), and (5) final estimation. The most computationally intensive step is typically the selection of optimal (λ, k) pairs, requiring a two-dimensional grid search. In our implementation, the GCV-based estimator was computationally most efficient, followed by BIC, AICc, and RECP where RECP involves a pilot variance estimation, and it is more costly than others.

Before presenting the results, as mentioned in the sections above, selection of bandwidth of kernel smoother (λ) and the shrinkage (or ridge) parameter (k) has a critical importance due to its effect on the estimation accuracy. Therefore, for different simulation configurations, optimal pairs of $(\hat{\lambda}, \hat{k})$ are presented in Table 1 and Figure 2. Hence, it is possible to see how these parameters are affected by the multicollinearity level between predictors (δ) and sample size (n). When examining Table 1, the chosen bandwidth and shrinkage parameters, along with the specified four criteria, are evident for all possible simulation configurations. From the values, it can be observed that the values of the pair $(\hat{\lambda}, \hat{k})$ increase as the correlation level rises, potentially negatively impacting estimation quality in terms of smoothing. On the other hand, the increase in $\hat{k}$, particularly when multicollinearity is high, is an expected behavior aimed at avoiding indefinability in the variance-covariance matrix. For large sample sizes, the criteria tend to select lower $\hat{\lambda}$, but higher "$\hat{k}$" values as a general tendency. Similar selection behavior is observed when $\delta = 0.99$. We take the behavior of $(\hat{\lambda}, \hat{k})$ obtained in Table 1 and follow similar roadmap for the shrinkage estimators that are obtained rely on the GRTK estimator.

 doi:10.20944/preprints202504.0689.v1

Table 1. Optimal $(\hat{\lambda}, \hat{k})$ pairs selected by AIC_c, BIC and $RECP$ critera for different simulation configurations for GRTK.

		$\delta = 0.95$				$\delta = 0.99$			
n		GCV	AIC_c	BIC	$RECP$	GCV	AIC_c	BIC	$RECP$
50	$\hat{\lambda}$	0.029	0.059	0.038	0.055	0.039	0.112	0.048	0.054
	$\hat{k}$	0.553	1.368	0.881	0.806	1.053	1.599	1.348	0.686
150	$\hat{\lambda}$	0.015	0.031	0.019	0.039	0.017	0.031	0.018	0.075
	$\hat{k}$	1.155	1.016	0.933	0.976	1.084	1.067	1.144	0.914
300	$\hat{\lambda}$	0.021	0.027	0.022	0.027	0.011	0.014	0.013	0.046
	$\hat{k}$	1.323	1.323	1.283	0.717	1.059	1.078	1.067	1.005

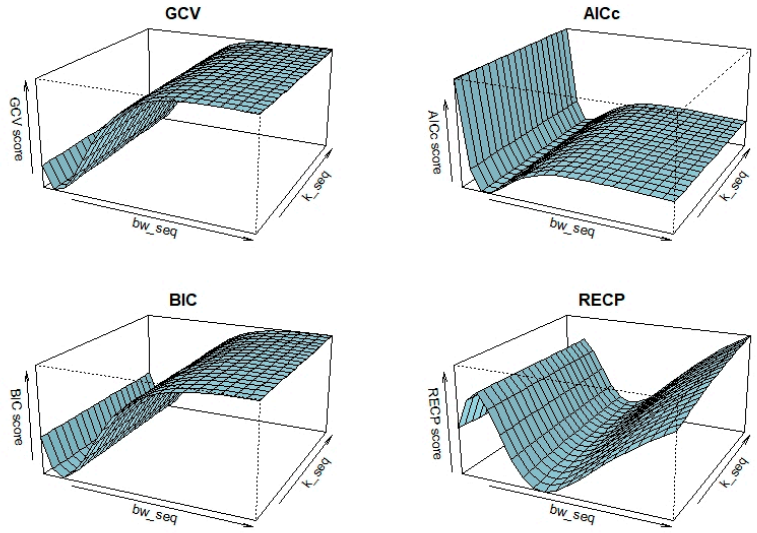

Figure 2. Selection of $(\hat{\lambda}, \hat{k})$ by AIC_c, BIC and $RECP$ under certain conditions: $n = 50, \delta = 0.95$.

Additionally, Figure 1 illustrates the selection procedure of each criterion for different possible values of bandwidth and shrinkage parameters simultaneously. It can be concluded from this that GCV, AIC_c, and BIC exhibit closer patterns than $RECP$. This difference is evidently sourced from its risk estimation procedure with the pilot variance shown in (4.4). Accordingly, $RECP$-based estimator mostly present different performances under different conditions.

After determination of the optimal tuning parameters $(\hat{\lambda}, \hat{k})$, time-series models are estimated and performance of the estimated regression coefficients of parametric component $(\widehat{\boldsymbol{\beta}}_{GRTK}, \widehat{\boldsymbol{\beta}}_{S}, \widehat{\boldsymbol{\beta}}_{PS})$ based on the selection criteria are obtained. In this sense, bias, variance and $SMSD$ values of $(\widehat{\boldsymbol{\beta}}_{GRTK}, \widehat{\boldsymbol{\beta}}_{S}, \widehat{\boldsymbol{\beta}}_{PS})$ are presented in Table 2. In addition, 3D figure is given to observe the effect of both sample size and correlation level on the quality of the estimates in terms of parametric component.

Table 2 reports the numerical performance of the *baseline* Generalized Ridge-Type Kernel $(\widehat{\boldsymbol{\beta}}_{GRTK})$ estimator under various sample sizes (n) and collinearity levels (ρ). Looking at the bias, variance, and aggregated measures such as SMSD, one sees that high correlation $(\rho = 0.99)$ generally inflates estimation error compared to moderate correlation $(\rho = 0.95)$. Conversely, larger sample sizes $(n = 150$ and $n = 300)$ mitigate these issues, resulting in smaller bias, variance, and SMSD. These findings confirm the paper's theoretical claims that ridge-type kernel smoothing can handle partially linear models with correlated errors more robustly as n grows. Nevertheless, because the paper

 doi:10.20944/preprints202504.0689.v1

works with *multiple* ($p \geq 2$) predictors, the variance inflation is still non-trivial, suggesting that *additional* shrinkage mechanisms beyond baseline GRTK might yield further improvements.

Table 2. Outcomes of simulations with bias, variance and SMSD scores for $\widehat{\boldsymbol{\beta}}_{GRTK}$.

		$\delta = 0.95$				$\delta = 0.99$			
n	metric	GCV	AIC_c	BIC	$RECP$	GCV	AIC_c	BIC	$RECP$
50	$Bias$	1.952	2.699	1.984	1.872	3.805	5.679	3.773	3.714
	Var	2.899	2.710	2.860	2.955	2.650	0.000	2.737	2.887
	$SMSD$	6.709	9.992	6.798	6.458	17.131	32.250	16.976	16.682
150	$Bias$	1.101	5.679	1.164	1.124	2.802	5.679	2.860	2.693
	Var	1.645	0.000	1.509	1.531	2.446	0.000	2.172	2.433
	$SMSD$	2.857	32.250	2.864	2.794	10.300	32.250	10.352	9.688
300	$Bias$	0.618	0.595	0.587	0.553	1.768	1.647	1.893	1.745
	Var	0.906	0.885	0.876	0.886	2.570	2.698	2.148	2.290
	$SMSD$	1.288	1.238	1.220	1.191	5.697	5.411	5.730	5.336

Table 3 summarizes the performance of the *shrinkage* estimator (notated as $\widehat{\boldsymbol{\beta}}_S$), which further regularizes the GRTK approach by incorporating Stein-type shrinkage toward a submodel. Relative to Table 2's baseline GRTK results, the shrinkage estimator consistently displays smaller bias–variance trade-offs across most combinations of ρ and n. In particular, when the predictors are highly collinear, the $\widehat{\boldsymbol{\beta}}_S$ estimator's additional penalization proves beneficial by reducing the inflated variance typically caused by near-linear dependencies among regressors. Hence, these results validate one of the paper's key arguments: that combining kernel smoothing with shrinkage can outperform a pure ridge-type kernel approach in high-dimensional or strongly correlated settings, especially for moderate sample sizes.

Table 3. Outcomes of simulations with bias, variance and SMSD scores for $\widehat{\boldsymbol{\beta}}_{PS}$.

			$\delta = 0.95$				$\delta = 0.99$			
n	metric	(p_1, p_2)	GCV	AIC_c	BIC	$RECP$	GCV	AIC_c	BIC	$RECP$
50	$Bias$	(3.0, 17.0)	1.943	2.70	1.98	1.86	3.58	5.68	3.54	3.48
	Var		2.857	2.71	2.82	2.91	2.36	0.00	2.42	2.52
	$SMSD$		6.633	9.98	6.72	6.38	15.17	32.25	14.99	14.60
150	$Bias$	(3.0, 17.0)	1.100	5.68	1.16	1.12	2.79	5.68	2.85	2.68
	Var		1.643	0.00	1.51	1.53	2.42	0.00	2.14	2.39
	$SMSD$		2.854	32.25	2.86	2.79	10.20	32.25	10.24	9.56
300	$Bias$	(3.2, 16.8)	0.618	0.59	0.59	0.55	1.77	1.65	1.89	1.74
	Var		0.906	0.88	0.88	0.89	2.57	2.70	2.14	2.28
	$SMSD$		1.288	1.24	1.22	1.19	5.69	5.41	5.72	5.32

Table 4 presents the outcomes for the *positive-part Stein* ($\widehat{\boldsymbol{\beta}}_{PS}$) shrinkage version, which modifies the ordinary Stein estimator by truncating the shrinkage factor at zero. According to the paper's premise, PS is designed to avoid "overshrinking" when the shrinkage factor becomes negative and is thus expected to give the strongest performance among the three estimators, particularly in the given scenario with high collinearity. Indeed, Table 4's bias, variance, and SMSD values are generally on par with—or superior to—those reported in Tables 2 and 3. When $\rho = 0.99$ is especially large, the PS estimator still manages to keep both the bias and variance relatively contained, highlighting that

 doi:10.20944/preprints202504.0689.v1

positive-part shrinkage delivers robust protection against severe collinearity. As a result, these findings support the paper's claim that the PS shrinkage mechanism provides an appealing balance of bias and variance under challenging data conditions, outperforming both baseline GRTK and standard Stein shrinkage (S) in many of the tested simulation settings.

Table 4. Outcomes of simulations with bias, variance and SMSD scores for $\widehat{\boldsymbol{\beta}}_S$.

			$\delta = 0.95$				$\delta = 0.99$			
n	metric	(p_1, p_2)	GCV	AIC_c	BIC	$RECP$	GCV	AIC_c	BIC	$RECP$
	$Bias$		1.943	2.698	1.976	1.863	3.580	5.679	3.545	3.477
50	Var	(3.0, 17.0)	2.857	2.707	2.819	2.911	2.358	0.010	2.421	2.517
	$SMSD$		6.633	9.984	6.723	6.381	15.174	32.250	14.986	14.604
	$Bias$		1.100	5.679	1.163	1.123	2.791	5.679	2.846	2.679
150	Var	(3.0, 17.0)	1.643	0.010	1.507	1.529	2.415	0.010	2.137	2.387
	$SMSD$		2.854	32.250	2.861	2.790	10.204	32.250	10.237	9.562
	$Bias$		0.618	0.595	0.587	0.552	1.767	1.647	1.891	1.744
300	Var	(3.2, 16.8)	0.906	0.885	0.876	0.885	2.566	2.698	2.142	2.283
	$SMSD$		1.288	1.238	1.220	1.191	5.690	5.411	5.719	5.324

Overall, the results across Tables 2–4 demonstrate that each proposed estimator—baseline GRTK ($\widehat{\boldsymbol{\beta}}_{GRTK}$), shrinkage ($\widehat{\boldsymbol{\beta}}_S$), and positive-part Stein shrinkage ($\widehat{\boldsymbol{\beta}}_{PS}$)- responds predictably to varying sample sizes and degrees of multicollinearity. Larger samples result in smaller bias, variance, and SMSD values, confirming the benefits of increased information for parameter estimation. Simultaneously, strong correlation among predictors inflates these metrics, illustrating the challenge posed by multicollinearity. Transitioning from GRTK to the two shrinkage-based estimators generally yields progressively better performance, particularly in higher-dimensional settings (where $n > p$) with $p = 20$. In particular, the positive-part Stein approach ($\widehat{\boldsymbol{\beta}}_{PS}$) often delivers the most stable results, indicating that its careful control of shrinkage is advantageous for mitigating both variance inflation and overshrinking.

Figure 3 visually compares the distributions of the estimated coefficients for $\widehat{\boldsymbol{\beta}}_{GRTK}$, $\widehat{\boldsymbol{\beta}}_S$, and $\widehat{\boldsymbol{\beta}}_{PS}$ via boxplots under different sample sizes and correlation levels, complementing the numerical findings in Tables 2–4. In each panel, the baseline $\widehat{\boldsymbol{\beta}}_{GRTK}$ estimator often deviates from the true coefficient value and zero for the bias, but it remains competitive with shrinkage estimators in terms of variance, especially when the data exhibit high collinearity, reflecting the variance inflation reported in Table 2. By contrast, the two shrinkage methods—$\widehat{\boldsymbol{\beta}}_S$ (Table 3) and $\widehat{\boldsymbol{\beta}}_{PS}$ (Table 4)—show boxplots that are close to real values of the coefficients, indicating smaller biases and more stable estimates. Particularly under higher correlation ($\rho = 0.99$) and moderate sample sizes, $\widehat{\boldsymbol{\beta}}_{PS}$ tends to yield less spread than the other estimators, consistent with the idea that positive-part Stein shrinkage further controls overshrinking while tackling multicollinearity. Overall, these graphical patterns support the tables' numeric trends, emphasizing how adding shrinkage to a kernel-based estimator improves both bias reduction and variance stabilization in partially linear models.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2025 doi:10.20944/preprints202504.0689.v1

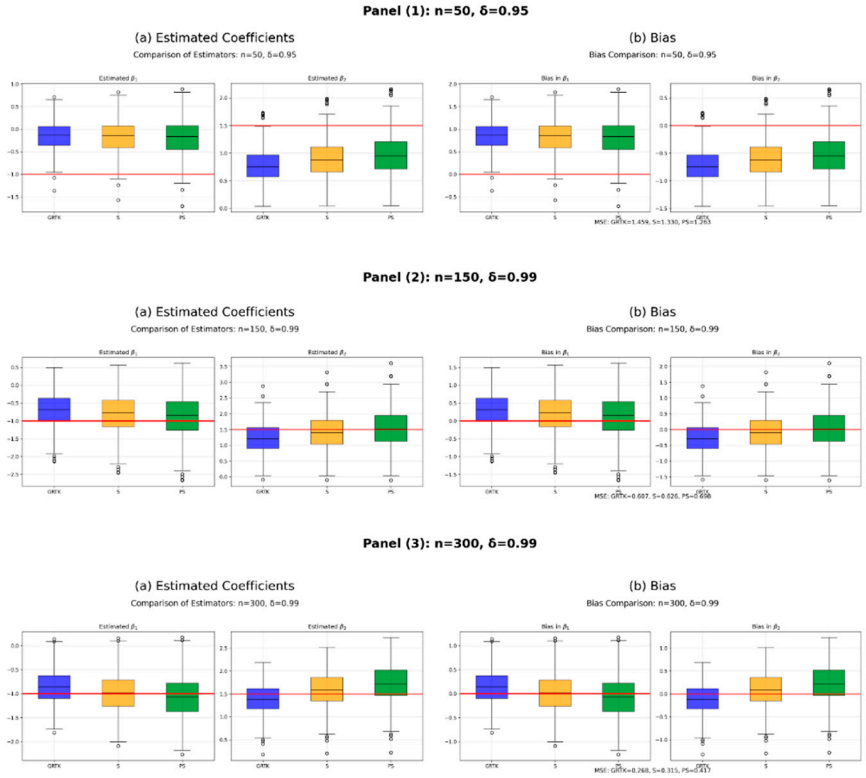

Figure 3. Boxplots for bias and estimated regression coefficients that obtained based on four criteria GCV, AIC_c, BIC and $RECP$ under certain conditions.

Table 5 reports the mean squared errors (MSE) for estimating the nonparametric component $f(z)$ under various combinations of sample size (n) and correlation level ρ. Reflecting patterns observed in the parametric estimates, a larger ρ typically increases MSE values, while growing the sample size reduces them. Notably, the table shows that the proposed $\hat{\mathbf{f}}_{GRTK}$ and its shrinkage versions ($\hat{\mathbf{f}}_S$ and $\hat{\mathbf{f}}_{PS}$) all perform better as n becomes larger; this aligns with the theoretical expectation that more data stabilize both kernel smoothing and shrinkage procedures in a partially linear context.

Figure 4 further illustrates these differences in estimating $f(z)$ by depicting the fitted nonparametric curves for selected simulation settings. Panels (a)–(b) highlight how increasing ρ can distort the estimated function, leading to higher MSE—an effect that is more pronounced for smaller samples. Meanwhile, panels (c)–(d) show that when nnn grows, each estimator recovers the true nonlinear pattern more closely, supporting the numerical results in Table 5. Taken together, Table 5 and Figure 4 confirm that although multicollinearity and smaller sample sizes can hamper nonparametric estimation, the combination of kernel smoothing with shrinkage remains robust and delivers reasonable reconstructions of $f(z)$ in challenging time-series settings.

Table 5. MSE values for the estimated nonparametric component $\hat{\mathbf{f}}_{GRTK}$ for all simulation combinations.

		GRTK				S				PS			
n	δ	GCV	AIC_c	BIC	$RECP$	GCV	AIC_c	BIC	$RECP$	GCV	AIC_c	BIC	$RECP$
50	0.95	3.668	**3.562**	3.747	3.758	1.267	1.396	1.267	**1.086**	1.872	1.996	**1.826**	1.958
	0.99	**3.819**	4.209	3.908	4.272	**1.345**	2.762	1.555	2.499	2.281	2.109	**2.014**	2.660
150	0.95	**3.178**	3.329	3.353	3.367	1.277	1.213	1.183	**0.968**	**1.353**	1.451	1.420	1.461
	0.99	**2.956**	3.448	3.396	3.444	1.355	1.585	1.346	**0.871**	**1.192**	1.583	1.338	1.412
300	0.95	**1.932**	2.266	2.025	2.264	0.802	**0.726**	0.861	0.947	0.937	0.878	**0.759**	0.957
	0.99	2.277	2.245	**2.228**	2.319	**1.077**	1.245	1.218	1.319	0.977	0.945	**0.928**	1.319

 doi:10.20944/preprints202504.0689.v1

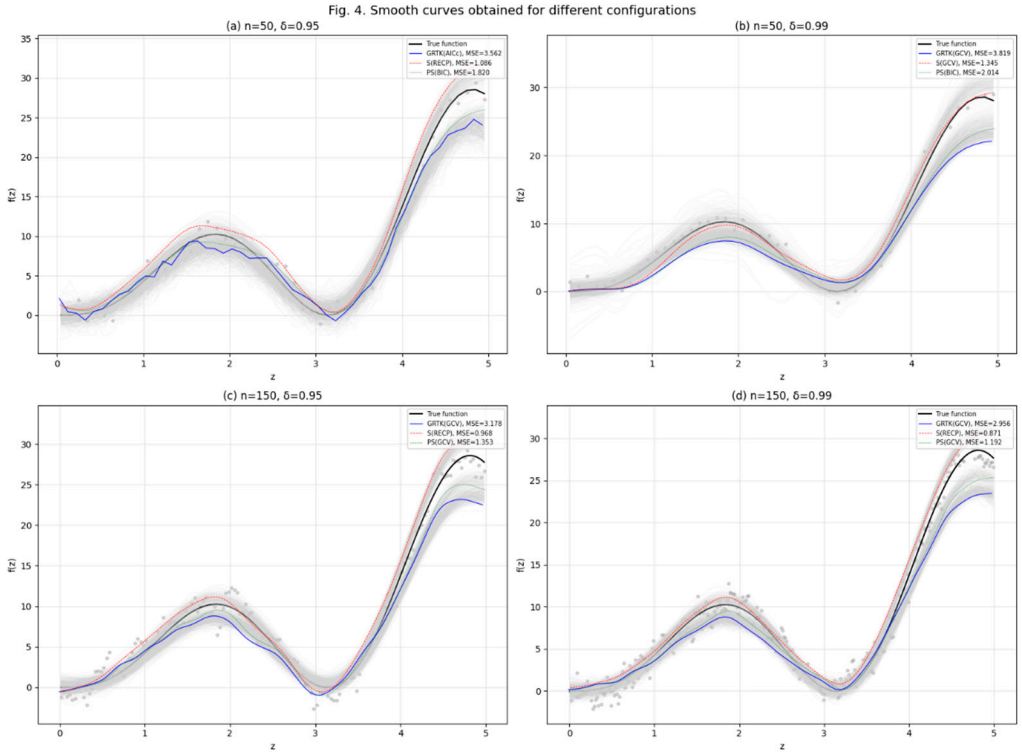

Figure 4. Smooth curves obtained for different configurations to show effect of the correlation level (δ) in panels (a)-(b) and the effect of the sample size (n) in panels (a)-(c).

5.2. Real Data Example: Airline Delay Dataset

In this section, we consider the airline delay dataset to show the performance of the modified GRTK estimators. The Bureau of Transportation Statistics (BTS), a division of the U.S. Department of Transportation (DOT), monitors the punctuality of domestic flights conducted by major air carriers. Commencing in June 2003, BTS initiated the collection of comprehensive information as a daily time series regarding the reasons behind flight delays. Based on this collected data, we obtain the partially linear time-series model with GRTK estimators along with shrinkage estimators and measure the comparative performance according to the four selection criteria for both shrinkage (k) and bandwidth (λ) parameters. The dataset involves originally 1048576 data points which is almost impossible to process for the analysis. In this paper, we extracted a representative consecutive series from the data with $n = 300$ data points. Notice that we do not try to solve a specific data-driven problem but showing the merits of the introduced semiparametric estimators. Therefore dataset with $n = 300$ is enough to represent both the modelling procedure and its alignment with the simulation settings. To construct the semiparametric time-series model, we consider the following six variables as explanatory variables for the parametric component of the model that are departure time $(DepTime)$, CRS departure time $(CRSDepTime)$, actual elapsed time $(ActualElapsedTime)$, CRS elapsed time $(CRSElapsedTime)$, departure delay $(DepDelay)$ and distance $(Distance)$.The air time of the aircrafts $(AirTime)$ is determined as a nonparametric variable for the model. The reason for that can be seen in Figure 5 with the hypothetical curve which can be counted as evidence for the nonlinear relationship between the response variable arrival delays $(ArrDelay)$ and the $AirTime$ variable. Accordingly, the model to be estimated is given by:

$$ArrDelay_t = \sum_{j=1}^{p=6} x_{jt}\,\beta_j + f(AirTime_t) + \varepsilon_t, t = 1, \dots, 300 \qquad (5.3)$$

where x_{jt}'s denote the six explanatory covariates defined above, $\beta_j's$ are the corresponding regression coefficients and ε_t's are the autocorrelated error terms as shown in (1.2). Autocorrelation parameter is estimated as $\rho = 0.01$ for difference of $ArrDelay_t$. In Figure 5, autocorrelation functions are provided for both original (non-stationary) and differenced-$ArrDelay_t$ series.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2025 doi:10.20944/preprints202504.0689.v1

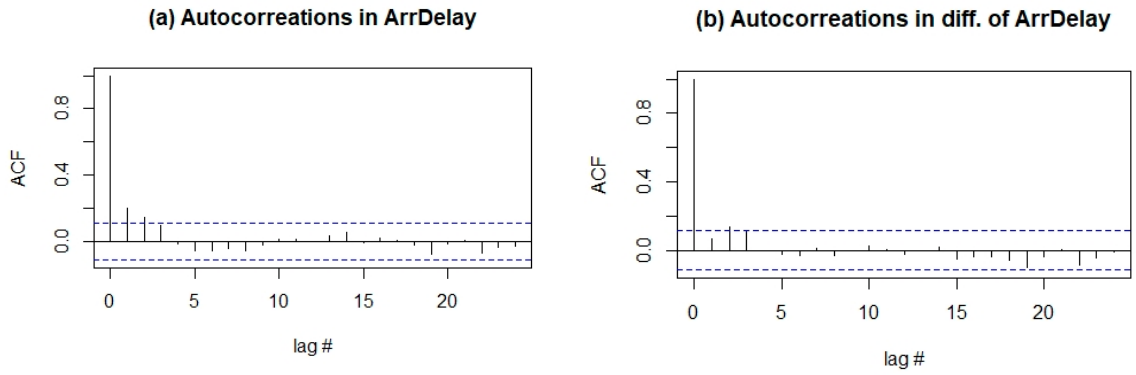

Figure 5. Autocorrelation functions (acf) for the model error in (5.3): Panel (a) acf plot of original response variable; panel (b) acf plot for difference of the $ArrDelay$ variable.

Figure 6, each panel includes important information about the data to describe the time-series and its properties. In panel (a), variance inflation factors (VIFs) of the predictors in the parametric component of the model reveal that five out of six variables have VIF values greater than 10, indicating a significant multicollinearity problem that needs to be addressed. In panel (b), the linear relationship between predictors and the response variable $ArrDelay$ can be observed, which is also evident in panel (c), depicting pairs of variables with correlation levels. Finally, in panel (d), the relationship between $ArrDelay$ and the nonparametric covariate $AirTime$ is illustrated with a hypothetical curve.

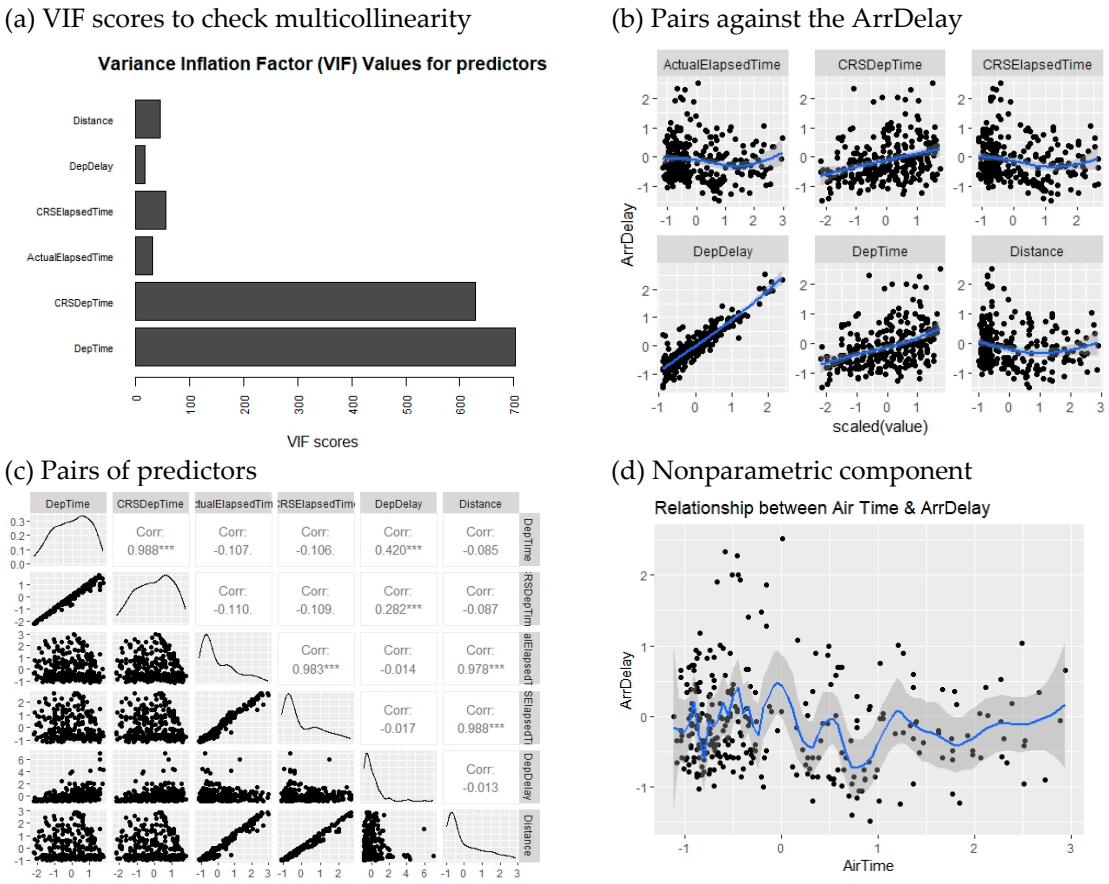

Figure 6. Informative plots for the Airline Delay dataset.

In the simulation study, 3D plots are generated to illustrate the selection of bandwidth (λ) and shrinkage parameter (k) for the four criteria, as shown in Figure 7. According to process in Figure 7,

 doi:10.20944/preprints202504.0689.v1

selected pairs of $(\hat{\lambda}, \hat{k})$ for the criteria are as follows: $(\hat{\lambda}_{GCV} = 1.357, \hat{k}_{GCV} = 1.289)$, $(\hat{\lambda}_{AIC_c} = 1.242, \hat{k}_{AIC_c} = 2)$, $(\hat{\lambda}_{BIC} = 1.128, \hat{k}_{BIC} = 2)$ and $(\hat{\lambda}_{RECP} = 0.10, \hat{k}_{RECP} = 0.01)$. After obtaining the optimal tuning parameters, the performances of the GRTK estimators are presented in Table 4. This table includes the model variance, MSE of the nonparametric component estimate, and overall variance of the regression coefficient, calculated by $tr(Var(\widehat{\boldsymbol{\beta}}))$, where $Var(\widehat{\boldsymbol{\beta}})$ is shown in (3.11). The best scores are indicated with bold colors in Table 4. According to the results, consistent with the findings from the simulation study, for $n = 300$, we observe closer performances.

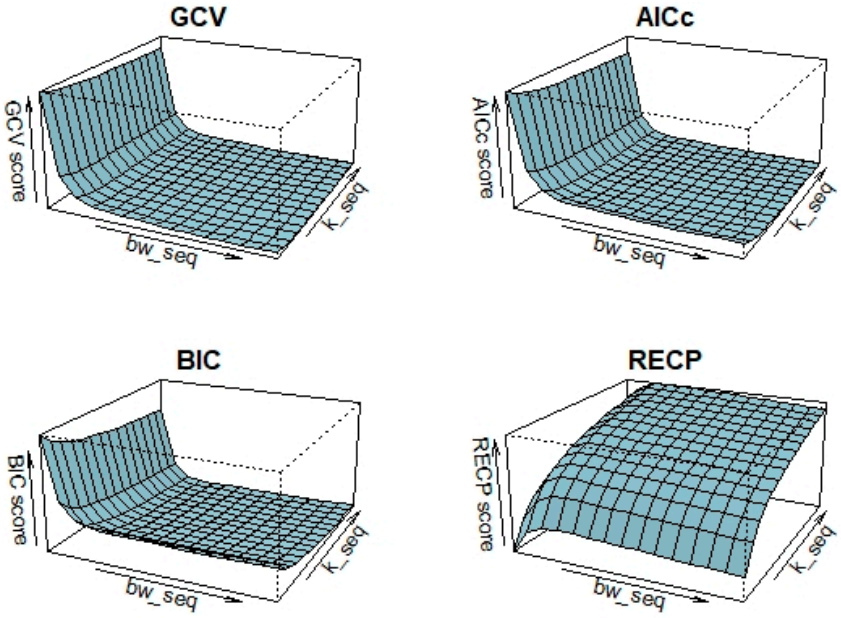

Figure 7. Selection of $(\hat{\lambda}, \hat{k})$ based on the four criteria GCV, AIC_c, BIC and $RECP$ for airline delay dataset.

Simultaneously, the $RECP$-based estimator shows good performance, closely followed by the GCV-based estimator. AIC_c and BIC-based estimators also exhibit considerable performances and are close to the other two. Additionally, to assess the statistical significance of the estimated regression coefficients, Student t-test statistics and their p-values are presented in Table 5. Furthermore, the estimated models are tested using the F-test, and p-values are provided. The results indicate that all estimated models based on the four criteria are statistically significant ($p < 0.05$). However, $\hat{\beta}_{DepTime}$, $\hat{\beta}_{CRSDepTime}$, and $\hat{\beta}_{Distance}$ do not make a significant contribution to the estimated model, despite the overall significance of the models. In conclusion, it can be stated that all criteria yield meaningful models based on their parameter selection procedures.

Table 6. Performance of the GRTK estimators for parametric and nonparametric components of the model.

Measure/Criterion	GCV	AIC_c	BIC	$RECP$
σ^2_{GRTK}	**0.952**	0.953	0.953	0.965
$MSE(\hat{f}_{GRTK})$	**0.932**	0.936	0.936	0.935
$Var(\hat{\beta}_{GRTK})$	0.270	**0.246**	**0.246**	0.31
σ^2_S	**0.712**	0.809	0.741	0.825
$MSE(\hat{f}_S)$	**0.821**	0.836	0.831	0.857
$Var(\hat{\beta}_S)$	**0.210**	0.225	0.226	0.280
σ^2_{PS}	0.705	0.713	**0.673**	0.711
$MSE(\hat{f}_{PS})$	**0.735**	0.736	**0.735**	0.735
$Var(\hat{\beta}_{PS})$	**0.172**	0.211	0.183	0.191

Bold color indicates the best scores.

 doi:10.20944/preprints202504.0689.v1

Table 6 presents the final performance measures of the GRTK-based estimators (including their shrinkage variants) for the real airline delay dataset, focusing on both parametric and nonparametric components. These metrics—covering model variance, the MSE of the nonparametric estimate, and the overall variance of the regression coefficients—show that all four selection criteria ($GCV, AICc, BIC, RECP$) yield closely comparable results, mirroring the pattern observed in the simulation study for moderate-to-large sample sizes. In particular, RECP exhibits a slight edge in some measures, yet the other criteria remain competitive. This consistency with the simulation outcomes underscores two key conclusions: (1) the newly introduced GRTK framework and its shrinkage modifications successfully mitigate variance inflation in the presence of multicollinearity; and (2) once an adequately sized dataset is available, different penalty-parameter selection methods converge toward similarly effective solutions in partially linear time-series models.

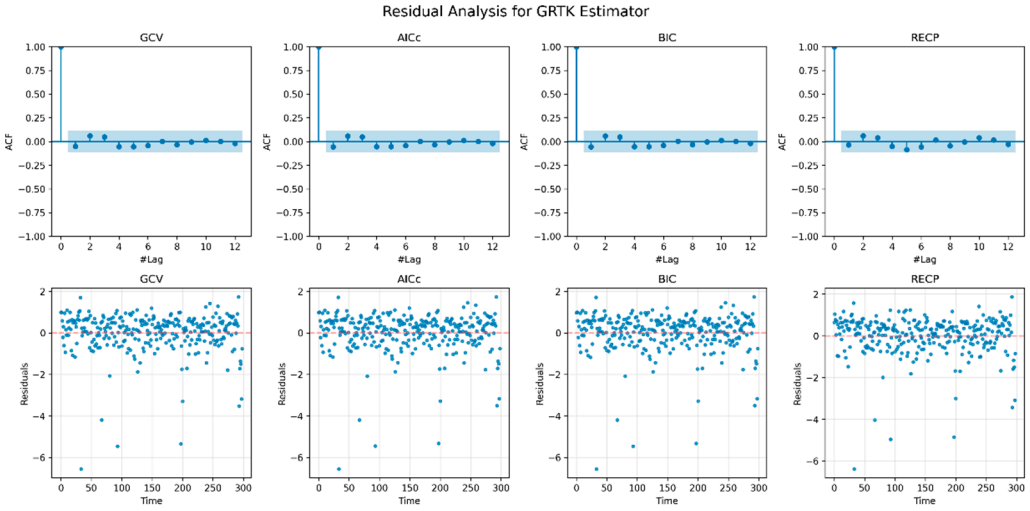

Figure 8. Residual analysis for the airline delay dataset. In the plots at the top of the figure show the acf plots of residuals obtained for the estimated models based on the corresponding criterion. The plots at the bottom show the scatterplots of the residuals around $x = 0$.

Figure 8 provides a residual analysis for each of the four GRTK-based models (including shrinkage versions) applied to the airline delay dataset. The top row displays the autocorrelation functions (ACFs) of the residuals, indicating whether significant time dependence remains after fitting. In all four cases, autocorrelation appears modest and tapers off quickly, suggesting that the estimated models have captured most of the serial dependence in the data. The bottom row shows residuals scattered around zero without any pronounced pattern, hinting at an absence of systematic bias or heteroscedasticity. Consequently, the visual diagnostic supports the conclusion that, regardless of the specific selection criterion (GCV, AICc, BIC, RECP), the introduced GRTK estimators adequately address the combined challenges of autocorrelation and multicollinearity in this real-world time-series setting.

Regarding the fitted curves, it presents the estimated curves for the nonparametric component $f(z)$ under the proposed GRTK framework, incorporating both ordinary Stein shrinkage (S) and positive-part Stein shrinkage (PS), as well as the baseline GRTK. Each panel corresponds to one of the four selection criteria—GCV, AICc, BIC, and RECP—which choose optimal pairs $(\hat{\lambda}, \hat{k})$ to govern bandwidth ($\hat{\lambda}$) and shrinkage ($\hat{k}$). Mathematically, all four criteria minimize a penalized objective function (e.g., GCV, AICc, etc.) to balance bias (from a potentially large $\hat{\lambda}$) and variance (from insufficient shrinkage). As seen in the figure, different choices of $(\hat{\lambda}, \hat{k})$ yield subtle variations in the smoothness and curvature of$f(z)$. For example, AIC_c typically enforces a slightly larger bandwidth to address overfitting risks in moderate samples, resulting in a smoother curve, whereas BIC can pick a smaller λ or a larger k, leading to more parsimonious fits with sharper inflection points. GCV

 doi:10.20944/preprints202504.0689.v1

often strikes a middle ground, and RECP sometimes selects a relatively large bandwidth or smaller shrinkage factor if its pilot-based risk estimation deems this configuration optimal.

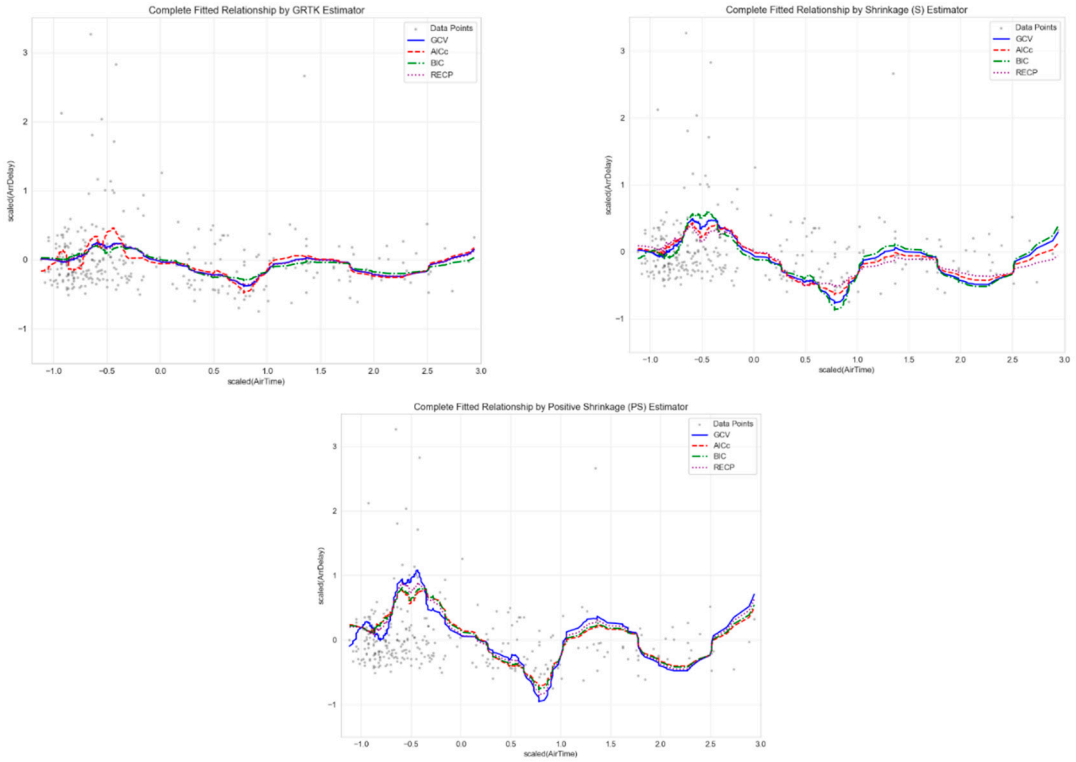

Figure 9. Fitted curves obtained based on the four criteria.

Beyond these distinctions, the figure also highlights the effect of adding shrinkage to the baseline GRTK estimates. Visually, the PS-based curves often appear slightly more adaptive in regions where GRTK or S might be overly penalized, reflecting a more balanced bias–variance trade-off. Despite these finer differences, all curves capture the primary nonlinear trends and fall within reasonable bounds, reinforcing the broader conclusion that combining kernel smoothing with either Stein or positive-part Stein shrinkage effectively handles both autocorrelation and multicollinearity in semiparametric time-series models.

6. Conclusions

This paper introduces and analyzes modified ridge-type kernel smoothing estimators tailored for semiparametric time-series models with multicollinearity in the parametric component. By combining generalized ridge-type kernel (GRTK) methodology with shrinkage and positive-part Stein shrinkage versions, we address both near-linear dependencies among regressors and autocorrelation in the error structure. We also explore how to optimally select the bandwidth (λ) and shrinkage parameter (k) using four widely used criteria: GCV, AIC_c, BIC, and $RECP$.

From the detailed simulation study and the real-world airline delay dataset, the following conclusions can be drawn:

- The GCV-based GRTK estimator effectively balances bias and variance for both parametric and nonparametric components. In simulations, it consistently provides stable estimates for linear coefficients $\boldsymbol{\beta}$ and suitably smooth fits for the nonlinear function $f(z)$.
- All four criteria show instability in small samples, as expected. However, in medium and large samples, the proposed estimators achieve more reliable performance, with reduced bias, variance, and SMSD. The GRTK-based approach effectively manages the autoregressive nature of the errors, highlighting the importance of accounting for correlation in time-series data.

 doi:10.20944/preprints202504.0689.v1

- Shrinkage estimators, especially positive-part Stein, excel at mitigating variance inflation and overshrinking under strong multicollinearity. They outperform standard methods when predictors are strongly correlated, particularly in larger samples.
- Airline delay data results align with simulations. All selection methods yield comparable models, demonstrating their ability to balance penalty tuning. This advantage becomes most pronounced in moderate and large samples, and GCV often provides a straightforward yet effective method for setting (λ, k).

Building on the insights gained here, several cases remain open for further research such as extending to non-stationary time-series data, developing locally adaptive bandwidth selection, incorporating alternative variable selection techniques for high-dimensional data, investigating efficient algorithms for parameter selection.

As a result, the proposed estimators fill an important gap in semiparametric modeling by jointly tackling multicollinearity and autocorrelation, thereby yielding more robust and interpretable results for partially linear models in time-series contexts.

Author Contributions: Conceptualization, S.E.A. and D.A.; methodology, D.A. and E.Y.; software, E.Y.; validation, S.E.A., D.A.; formal analysis, E.Y.; investigation, E.Y.; resources, S.E.A.; data curation, E.Y.; writing—original draft preparation, S.E.A. and E.Y.; writing—review and editing, S.E.A.; visualization, E.Y.; supervision, S.E.A. and D.A.; project administration, S.E.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: For simulation studies R function can be accessed from the link: https://github.com/yilmazersin13/Partially-linear-time-series-model and for the real data example AirDelay dataset is publicly available in Kaggle platform and accessed by the link: https://www.kaggle.com/datasets/undersc0re/flight-delay-and-causes.

Acknowledgments: The research of Professor S. Ejaz Ahmed was supported by the Natural Sciences and the Engineering Research Council (NSERC) of Canada.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Derivations of GRTK Estimators

Appendix A1. Derivations of Equations (3.12 - 3.14)

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{GRTK}(k) &= \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{y}} \\ &= \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\left((\mathbf{I}-\mathbf{W}_\lambda)\mathbf{y}\right) \\ &= G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\left((\mathbf{I}-\mathbf{W}_\lambda)(\mathbf{X}\boldsymbol{\beta}+\mathbf{f}+\boldsymbol{\varepsilon})\right) \\ &= G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\left(\widetilde{\mathbf{X}}\boldsymbol{\beta}+\tilde{\mathbf{f}}+(\mathbf{I}-\mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right) \\ &= G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}}\boldsymbol{\beta} + G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} + G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I}-\mathbf{W}_\lambda)\boldsymbol{\varepsilon}\end{aligned}$$

Accordingly, equation (3.12) can be expressed as:

$$\begin{aligned}E\left(\widehat{\boldsymbol{\beta}}_{GRTK}\right) &= G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}}\,E(\boldsymbol{\beta}) + G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} \\ &= \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I} - k\mathbf{I})\boldsymbol{\beta} \\ &\quad + \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} \\ &= \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\left[\left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)\boldsymbol{\beta} - k\mathbf{I}\boldsymbol{\beta}\right] \\ &\quad + \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} \\ &= \left(\mathbf{I} - k\left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\right)\boldsymbol{\beta} + \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}} \\ &= \boldsymbol{\beta} - k\left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\boldsymbol{\beta} + \left(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}\right)^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{f}}\end{aligned}$$

Equivalently, from (3.9), we obtain the bias as in equation (3.13):

 doi:10.20944/preprints202504.0689.v1

$$E\left(\widehat{\boldsymbol{\beta}}_R(k)\right) = E\left(\left(\mathbf{I}_p + k(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\right)^{-1}\widehat{\boldsymbol{\beta}}\right)$$
$$= E\left[\left(\mathbf{I}_p + k(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\right)^{-1}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)\mathbf{y}\right]$$
$$= (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{X}}'\tilde{\mathbf{f}})$$

Also, the variance of an estimator $\widehat{\boldsymbol{\beta}}_R(k)$ given in (3.14) can be derived as follows:

$$Var\left(\left(\widehat{\boldsymbol{\beta}}_R(k)\right)\right) = E\left[\left\{\left(\widehat{\boldsymbol{\beta}}_R(k)\right) - E\left(\widehat{\boldsymbol{\beta}}_R(k)\right)\right\}\left\{\left(\widehat{\boldsymbol{\beta}}_R(k)\right) - E\left(\widehat{\boldsymbol{\beta}}_R(k)\right)\right\}'\right]$$

$$= E\begin{Bmatrix}\left[(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\boldsymbol{\beta} + (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'\tilde{\mathbf{f}} + (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right] \\ -\left[(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\boldsymbol{\beta} + (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'\tilde{\mathbf{f}}\right]\end{Bmatrix}^2$$
$$= E\left\{\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right)\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right)'\right\}$$
$$= E\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)^2\right)\left((\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'\right)E(\boldsymbol{\varepsilon}^2)$$
$$= \sigma^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + k\mathbf{I}_p)^{-1}\widetilde{\mathbf{X}}$$

As a result, it can be expressed as the following result with abbreviation as in (3.14):

$$Var\left(\left(\widehat{\boldsymbol{\beta}}_R(k)\right)\right) = \sigma^2\mathbf{G}_k\widetilde{\mathbf{X}}'(\mathbf{I}_n - \mathbf{W}_\lambda)^2\widetilde{\mathbf{X}}\mathbf{G}_k$$

Hence, derivation has been completed.

Appendix A2. The Proof of the Equation (3.16)

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \mathbf{X}(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I})^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\tilde{\mathbf{y}} + \mathbf{W}_\lambda(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$
$$= \mathbf{X}(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I})^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y} + \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}(\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I})^{-1}\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}\right)$$
$$= \mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y} + \mathbf{W}_\lambda(\mathbf{y} - \mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y})$$
$$= \mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y} + \mathbf{W}_\lambda\mathbf{y} - \mathbf{W}_\lambda\mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)$$
$$= \mathbf{W}_\lambda\mathbf{y} + \mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y} - \mathbf{W}_\lambda\mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$$
$$= \mathbf{W}_\lambda\mathbf{y} + (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$$
$$= [\mathbf{W}_\lambda + (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)]\mathbf{y}$$

Thus, the hat matrix is written as follows:

$$\mathbf{H}_{\lambda,\boldsymbol{k}} = \mathbf{W}_\lambda + (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}G_k\widetilde{\mathbf{X}}'\boldsymbol{R}^{-1}(\mathbf{I} - \mathbf{W}_\lambda)$$

Appendix B. Asymptotic Supplement: Bias, AQDB, and ADR

This appendix provides a treatment of the asymptotic bias, Asymptotic Quadratic Distributional Bias (AQDB), and Asymptotic Distributional Risk (ADR) for the GRTK-based estimators introduced in Section 3. Specifically, we focus on:

1. The baseline GRTK (FM) estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{GRTK}}(k) = \widehat{\boldsymbol{\beta}}_{\mathrm{FM}}$ from equation (3.4),
2. The ordinary Stein shrinkage estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$ from equation (3.6), and
3. The positive-part Stein shrinkage estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{PS}}$ from equation (3.7).

with covariance matrix $\mathbf{R} = \mathrm{Cov}(\boldsymbol{\varepsilon})$ capturing possible autocorrelation (e.g., AR(1)). Throughout, we assume $\mathbf{R}$ is known (or replaced with a consistent estimator $\widehat{\mathbf{R}}$, which suffices for the same asymptotic properties). We also assume standard regularity conditions on the kernel smoothing for $f(\cdot)$ (bandwidth $\lambda \to 0$ and smoothness conditions as $n \to \infty$). Under these conditions, all estimators in Section 3 are consistent for $\boldsymbol{\beta}$, and we can derive their bias, variance, and risk expansions.

 doi:10.20944/preprints202504.0689.v1

Appendix B1. Baseline GRTK (FM) Estimator

Recall that the full-model GRTK estimator may be written as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} = \left(\widetilde{\mathbf{X}}'\mathbf{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\widetilde{\mathbf{X}}^{\top}\mathbf{R}^{-1}\tilde{\mathbf{y}}$$

where $\tilde{\mathbf{y}} = (\mathbf{I}_n - \mathbf{W}_\lambda)\mathbf{y}$ is the partially centered response (subtracting off the estimated nonparametric component $\hat{f}(\cdot)$), and $\mathbf{X}$ is the $n \times p$ design matrix for the parametric covariates. Let us Define

$$\mathbf{L} = \left(\widetilde{\mathbf{X}}'\mathbf{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\widetilde{\mathbf{X}}^{\top}\mathbf{R}^{-1}$$

Hence $\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} = \mathbf{L}\tilde{\mathbf{y}}$. If $\tilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ (the decomposition after partial kernel smoothing of $f(\cdot)$, then

$$\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} = \mathbf{L}\widetilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{L}\tilde{\varepsilon}$$

One can show that $\mathbf{L}\widetilde{\mathbf{X}}$ is effectively $\mathbf{I}_p - k\left(\widetilde{\mathbf{X}}^{\top}\mathbf{R}^{-1}\widetilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}$, capturing the ridge shrinkage on the parametric coefficients. Hence,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} = \left[\mathbf{I}_{\mathrm{p}} - k\left(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + \mathrm{k}\mathbf{I}_{\mathrm{p}}\right)^{-1}\right]\boldsymbol{\beta} + \mathbf{L}\tilde{\boldsymbol{\varepsilon}}.$$

Exact Bias. Taking expectation (condition on $\mathbf{X}$) and noting $E[\tilde{\varepsilon}] \approx \mathbf{0}$ for large n, we get

$$\begin{aligned}\mathrm{Bias}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}\right) &= \mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}\right] - \boldsymbol{\beta} \\ &\approx -k\left(\widetilde{\mathbf{X}}'\mathbf{R}^{-1}\widetilde{\mathbf{X}} + \mathrm{k}\mathbf{I}_{\mathrm{p}}\right)^{-1}\boldsymbol{\beta}\end{aligned}$$

plus small $o(1)$ terms from the smoothing residual. Since $\|\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\| = O(n)$ in typical regressions and k may vanish with n (or remain bounded), this bias goes to zero as $n \to \infty$. The asymptotic behavior of this bias term depends crucially on how k scales with n. We can distinguish several cases:

(i) If $k = O(1)$ (remains constant or bounded as n→∞):Since $\|\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\| = O(n)$, we have $(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + k\mathbf{I})^{-1} = \mathrm{O}(1/\mathrm{n})$, and the bias term becomes $k \cdot O(1/n) \cdot \boldsymbol{\beta} = O\left(\frac{1}{n}\right)$, which converges to zero at rate $1/n$.

(ii) If $k = O(n^{\alpha})$ for some $0 < \alpha < 1$: The bias term becomes $O(n^{\alpha-1})$, which still converges to zero when $\alpha < 1$, but at a slower rate than in case (i).

(iii) If $k = O(n)$: The bias may not vanish asymptotically, potentially compromising consistency.

Therefore, for consistency of the $\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}$ estimator, we require that $k(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + k\mathbf{I})^{-1} \to 0$ as $n \to \infty$, which is satisfied when $k = o(n)$. This asymptotic requirement aligns with our parameter selection methods in Section 4, which implicitly balance the bias-variance tradeoff by choosing appropriate k values. Hence the GRTK estimator is consistent and has $O(k)$ bias in finite n, shrinking $\boldsymbol{\beta}$ towards $\mathbf{0}$. In summary:

$$\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}\right) = O(k) \to \mathbf{0} \text{ as } n \to \infty$$

Appendix B2. Shrinkage Estimators

Denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{SM}}$ the submodel estimator, obtained by fitting the same GRTK procedure but only to a subset of $p_1 < p$ regressors using BIC criterion. In practice, we set the omitted $(p - \mathrm{p}_1)$ coefficients to zero. Then the ordinary Stein-type shrinkage estimator is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S}} = \widehat{\boldsymbol{\beta}}_{\mathrm{SM}} + \hat{\theta}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}}\right)$$

where $\hat{\theta} \in [0,1]$ is a shrinkage factor determined from the data (discussed in Subsection B3). If $\hat{\theta} = 1$, no shrinkage is applied (we use the full model); if $\hat{\theta} = 0$, we revert to the submodel; for $0 < \hat{\theta} < 1$, we form a weighted compromise. The positive-part shrinkage estimator is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{PS}} = \widehat{\boldsymbol{\beta}}_{\mathrm{SM}} + max\{0, \hat{\theta}\}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}}\right)$$

so that negative values of $\hat{k}$ (which might arise due to sampling error) are replaced by 0 to prevent over-shrinkage. Regarding the bias of shrinkage estimators,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S}} - \boldsymbol{\beta} = \left(\widehat{\boldsymbol{\beta}}_{\mathrm{SM}} - \boldsymbol{\beta}\right) + \hat{\theta}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}}\right)$$

note that $\widehat{\boldsymbol{\beta}}_{\mathrm{SM}}$ itself has some bias (particularly if the omitted $(p - p_1)$ coefficients are not truly zero), while $\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}$ has the $O(\theta)$ ridge bias but includes all parameters. Let $\left(\widehat{\boldsymbol{\beta}}_{\mathrm{SM}}\right) = \mathbf{b}_{\mathrm{SM}}$ and $\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}\right) = \mathbf{b}_{\mathrm{FM}}$.

 doi:10.20944/preprints202504.0689.v1

A first-order approximation, assuming $\hat{\theta}$ is not strongly correlated with the random errors in $(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}})$, gives

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}_{\mathrm{S}}] - \boldsymbol{\beta} \approx [1 - \mathrm{E}[\hat{\theta}]](\mathbf{b}_{\mathrm{SM}}) + \mathrm{E}[\hat{\theta}](\mathbf{b}_{\mathrm{FM}})$$

Hence the bias of $\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$ is approximately a linear mixture of submodel bias and fullmodel bias. In large n, we typically find $\mathbf{b}_{\mathrm{FM}} = O(\theta)$ small, and $\mathbf{b}_{\mathrm{SM}}$ is either $O(\theta)$ (when the submodel is correct) or (if some omitted coefficients are actually nonzero) has a bigger $O(1)$ or $O(n^{-1/2})$ term. Thus, $\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$ has a bias that is smaller than the submodel's in cases where $E[\hat{\theta}] > 0$.

Positive-Part Variant. For $\widehat{\boldsymbol{\beta}}_{\mathrm{PS}}$, the same expansion holds but with $\hat{\theta}$ replaced by $\max\{\hat{\theta}, 0\}$. This cannot exceed $E[\hat{\theta}]$, so

$$(\widehat{\boldsymbol{\beta}}_{\mathrm{PS}}) \leq (\widehat{\boldsymbol{\beta}}_{\mathrm{S}})$$

elementwise (in a typical risk comparison sense). That is, the positive-part always improves or equals the Stein shrinkage in terms of bias magnitude, since it disallows negative $\hat{\theta}$.

As given in Section 3, the distance measure T quantifies how far $\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}$ is from the submodel $\widehat{\boldsymbol{\beta}}_{\mathrm{SM}}$. A convenient choice is the squared Mahalanobis distance:

$$T = (\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}})^{\top}[\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}})]^{-1}(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}} - \widehat{\boldsymbol{\beta}}_{\mathrm{SM}}).$$

Under the "null" submodel assumption (that the omitted parameters are actually zero), T follows approximately a χ^2 distribution with $\nu = (p - p_2)$ degrees of freedom, possibly noncentral if the omitted effects are not truly zero. The main text (see eqn. (3.6)) uses T in the formula for the shrinkage factor:

$$\hat{\theta} = 1 - \frac{\nu - 2}{T}$$

where $\nu = (p - p_1)$. If $T < \nu - 2, \hat{\theta}$ goes negative; the positive-part approach sets $\hat{\theta}_+ = \max(0, \hat{\theta})$. Asymptotic Distribution of T. Under standard conditions (normal or asymptotically normal errors, large n):

$$T \xrightarrow{d} \chi^2_{\nu}(\Lambda),$$

where Λ is a noncentrality parameter linked to how large the omitted true coefficients are. In a local-alternatives framework with $\boldsymbol{\beta}_2 = \boldsymbol{\delta}/\sqrt{n}$ for the omitted block, Λ remains finite as $n \to \infty$. If $\Lambda = 0$ (null case), $E[T] \approx \nu$ and $(T) \approx 2\nu$. If $\Lambda > 0, T$ tends to be $> \nu$, so $\hat{\theta} \approx 1$. This adaptivity is what drives the shrinkage phenomenon.

- If T is large ($\gg \nu$), we suspect omitted coefficients are nonzero, so $\hat{\theta} \approx 1$ retains the full-model estimate.
- If T is near $\nu, \hat{\theta}$is moderate $(0 < \hat{\theta} < 1)$, partially shrinking FM toward SM.
- If $T < \nu - 2$, then $\hat{\theta}$ is negative, which is clipped to 0 by the positive-part rule.

Appendix B3. Asymptotic Quadratic Bias (AQDB) and Distributional Risk (ADR)

We next assess each estimator's risk and define the Asymptotic Quadratic Distributional Bias (AQDB). For an estimator $\widehat{\boldsymbol{\beta}}$, define

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) = E[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = Tr[\mathrm{Var}(\widehat{\boldsymbol{\beta}})] + \|\mathrm{Bias}(\widehat{\boldsymbol{\beta}})\|^2$$

In the local alternatives setting (where the omitted block $\boldsymbol{\beta}_2 = \boldsymbol{\delta}/\sqrt{n}$), we often look at $n\mathrm{MSE}(\widehat{\boldsymbol{\beta}})$ as $n \to \infty$, which splits into an asymptotic variance component plus an asymptotic (squared) bias:

$$\lim_{n\to\infty} nE[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2] = \mathrm{Tr}(\boldsymbol{\Sigma}_{\hat{\beta}}) + \|\boldsymbol{\mu}_{\hat{\beta}}\|^2$$

Here $\boldsymbol{\Sigma}_{\hat{\beta}} = \lim_{n\to\infty} \mathrm{Var}(\sqrt{n}\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{\mu}_{\hat{\beta}} = \lim_{n\to\infty} \sqrt{n}(E[\widehat{\boldsymbol{\beta}}] - \boldsymbol{\beta})$. We denote:

$$\underbrace{\mathrm{Tr}(\boldsymbol{\Sigma}_{\hat{\beta}})}_{\text{Asymptotic Variance}} + \underbrace{\|\boldsymbol{\mu}_{\hat{\beta}}\|^2}_{\text{AQDB}} = \mathrm{ADR}(\widehat{\boldsymbol{\beta}}),$$

where the Asymptotic Quadratic Distributional Bias (AQDB) is $\|\boldsymbol{\mu}_{\hat{\beta}}\|^2$ and the Asymptotic Distributional Risk (ADR) is their sum.

- Full-Model (FM): has negligible bias (so $\boldsymbol{\mu}_{\mathrm{FM}} = 0$), but a higher variance from estimating all p coefficients. Hence $\mathrm{ADR}(\widehat{\boldsymbol{\beta}}_{\mathrm{FM}}) = \mathrm{Tr}(\boldsymbol{\Sigma}_{\mathrm{FM}})$, typically $> \mathrm{Tr}(\boldsymbol{\Sigma}_{\mathrm{SM}})$.

 doi:10.20944/preprints202504.0689.v1

- Submodel (SM): has lower variance (only p_1 parameters) but possibly large bias if $\boldsymbol{\beta}_2 \neq \mathbf{0}$, giving $\boldsymbol{\mu}_{\mathrm{SM}} \neq \mathbf{0}$ and a big $\|\boldsymbol{\mu}_{\mathrm{SM}}\|^2$. Specifically, if the omitted block is $\boldsymbol{\delta}/\sqrt{n}$, then $\boldsymbol{\mu}_{\mathrm{SM}} \approx (0, \boldsymbol{\delta})$ so $\mathrm{AQDB(SM)} = \|\boldsymbol{\delta}\|^2$.
- Stein (S): interpolates. Variance is $< \mathrm{Var(FM)}$ but $> \mathrm{Var(SM)}$. Bias is less than the SM's if omitted effects are actually nonzero, but not zero: $(1 - E[\hat{\theta}])$ times $\boldsymbol{\delta}$ roughly, so $\mathrm{AQDB(S)} \approx (1 - E[\hat{\theta}])^2 \|\boldsymbol{\delta}\|^2$.
- Positive-Part Stein (PS): ensures no negative shrinkage, so $\mathrm{AQDB(PS)} \approx \left(1 - E\left[\hat{\theta}_+\right]\right)^2 \|\boldsymbol{\delta}\|^2$ with $E\left[\hat{\theta}_+\right] \geq E[\hat{\theta}]$. Consequently, $\mathrm{AQDB(PS)} \leq \mathrm{AQDB(S)}$. Moreover, $\mathrm{Var(PS)} \leq \mathrm{Var(S)}$ in typical settings, so $\widehat{\boldsymbol{\beta}}_{\mathrm{PS}}$ dominates $\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$ in ADR.

Thus, positive-part shrinkage is usually recommended when $p \geq 3$ or under moderate correlation/collinearity, as it yields the lowest or nearly lowest ADR across different parameter regimes. This is consistent with the numerical results in Section 5 of the main text and references such as [21] and [3].

References

1. Ahmed, S.E. *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*; Springer: New York, USA, **2014**.
2. Ahmed, S.E. *Penalty, Shrinkage and Pretest Strategies in Statistical Modeling*. In Frontiers in Statistics; Springer: Cham, Switzerland, **2018**.
3. Ahmed, S.E.; Ahmed, F.; Yüzbaşı, B. *Post-Shrinkage Strategies in Statistical and Machine Learning for High Dimensional Data*; CRC Press: Boca Raton, USA, **2023**.
4. Ahmed, S.E.; Nicol, C.J. An application of shrinkage estimation to the nonlinear regression model. *Comput. Stat. Data Anal*. **2012**, 56(11), 3309–3321.
5. Aydın, D.; Yılmaz, E.; Chamidah, N.; Lestari, B.; Budiantara, I.N. Right-censored nonparametric regression with measurement error. *Metrika* **2025**, 88(2), 183–214.
6. Aydin, D.; Yilmaz, E. Modified estimators in semiparametric regression models with right-censored data. *J. Stat. Comput. Simul*. **2018**, 88(8), 1470–1498.
7. Gao, J. Asymptotic theory for partly linear models. *Commun. Stat.–Theory Methods* **1995**, 24(8), 1985–2009.
8. Green, P.J.; Silverman, B.W. *Nonparametric Regression and Generalized Linear Model*; Chapman & Hall: London, UK, **1994**.
9. Hurvich, C.M.; Simonoff, J.S.; Tasi, C.L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc. B* **1998**, 60, 271–293.
10. Hu, H. Ridge Estimation of a Semiparametric Regression Model. *J. Comput. Appl. Math.* **2005**, 176(1), 215–222.
11. Kazemi, M.; Shahsvani, D.; Arashi, M.; Rodrigues, P.C. Identification for partially linear regression model with autoregressive errors. *J. Stat. Comput. Simul.* **2021**, 91(7), 1441–1454.
12. Lee, T.C.M. Smoothing parameter selection for smoothing splines: a simulation study. *Comput. Stat. Data Anal.* **2003**, 42, 139–148.
13. Özkale, M.R. A jackknifed ridge estimator in the linear regression model with heteroscedastic or correlated errors. *Stat. Probab. Lett.* **2008**, 78(18), 3159–3169.
14. Speckman, P. Kernel smoothing in partially linear model. *J. R. Statist. Soc. B* **1988**, 50, 413–436.
15. Schick, A. Efficient estimation in a semiparametric additive regression model with autoregressive errors. *Stoch. Process. Their Appl.* **1996**, 61(2), 339–361.
16. Shi, J.; Lau, T.S. Empirical likelihood for partially linear models. *J. Multivar. Anal.* **2000**, 72(1), 132–148.
17. Yılmaz, E.; Yuzbasi, B.; Aydin, D. Choice of smoothing parameter for kernel type ridge estimators in semiparametric regression models. *REVSTAT–Stat. J.* **2021**, 19(1), 47–69.
18. You, J.; Chen, G. Semiparametric generalized least squares estimation in partially linear regression models with correlated errors. *J. Statist. Plan. Infer.* **2007**, 137(1), 117–132.
19. You, J.; Zhou, Y. Empirical likelihood for semiparametric varying-coefficient partially linear regression models. *Stat. Probab. Lett.* **2006**, 76(4), 412–422.

 doi:10.20944/preprints202504.0689.v1

20. Yüzbaşı, B.; Ahmed, S.E. Shrinkage and penalized estimation in semi-parametric models with multicollinear data. *J. Stat. Comput. Simul.* **2016**, 86(17), 3543–3561.
21. Yüzbaşı, B.; Ahmed, S.E.; Aydın, D. Ridge-type pretest and shrinkage estimations in partially linear models. *Stat. Pap.* **2020**, 61, 869–898.
22. Theobald, C.M. Generalizations of mean square error applied to ridge regression. *J. R. Statist. Soc. B* **1974**, 36(1), 103–106.