Random Forest-Based Estimation of Partial Linear Single-Index Model in Longitudinal Data

Hussein Jabbar Bayyoodh and Mohammed Sadeq Aldouri
Department of Statistics, College of Administration and Economics, University of Baghdad, 10071 Baghdad, Iraq
husseenjb@gmail.com, dr_aldouri@coadec.uobaghdad.edu.iq

Keywords: Partial Linear Single-Index Model, Longitudinal Data, Random Forest (RF), Local Polynomial, Semiparametric, Two-Stage.

Abstract: This paper discusses the estimation of a partial linear single-index model (PLSIM) for longitudinal data. It proposes a hybrid Random Forest-based estimator complemented by statistical regularization to ensure parameter stability and accurate retrieval of the nonlinear component. The proposed approach is based on estimating $g(\cdot)$ via Random Forests with a conservative selection of the number of trees with a 1-SE rule and subject-wise K-fold cross-validation, estimating the trend index β, and estimating the coefficients of the linear component θ. On balanced simulation data of sizes $N = [50,100,15]$ and a real function $g(u) = \sin(u)$, the hybrid estimator showed high accuracy in retrieving $g(\cdot)$ across the sample-supported domain, with a coefficient of determination of $R_9^2 \approx 0.96 - 0.98$ and a decreasing mean square error with increasing size, while the overall model performance stabilized at $R^2 \approx 0.90 - 0.92$ and $MSE \approx 0.10 - 0.13$. θ biases appeared small across all scenarios, while the β estimate maintained functional stability reflected in a strong visual match between the real and estimated, with systematically reduced marginal deviations compared to the conventional two-stage estimator, which showed greater sensitivity to the bootstrap parameter and index error and higher overall MSE at the larger sample. The results demonstrate that combining RF with statistical methods provides a practical and accurate path for estimating longitudinal PLSIM models, with straightforward applicability and limited parameter tuning. The study suggests potential future improvements in performance by expanding the framework to account for heteroscedasticity and random effects.

1 INTRODUCTION

In recent years, the partial single-index semiparametric model (PLSIM) has emerged as a framework that combines the simplicity of a linear component with the flexibility of a nonlinear component formulated via a single linear index to avoid the "curse of dimensionality." This balances interpretability and statistical efficiency. This framework was generalized to the generalized case (GPLSIM), paving the way for the development of early semiparametric estimation algorithms. At the independent data level (i.i.d.), [1] introduced a "back-fitting" algorithm, but it often requires smoothing reduction and exhibits numerical instability; Therefore, alternatives emerged, such as the local curve estimation penalized by splines [2] and the minimum mean-variance-average (MAVE) method of [3], which avoid over-smoothing and unify the loss criterion between components. Later, [4] proposed a "profile least squares" method, which achieves semiparametric efficiency and provides fit tests and parameter selection (such as SCAD) within the PLSIM framework.

From another perspective, [5] developed two-order estimators that adopt local smoothing and constrained index vector estimation. They showed that imposing the identification constraint in the estimation equation improves the variance of the estimator for the index coefficients without affecting the coefficients of the linear part. Moving to longitudinal data with within-individual correlation, [6] formulated a longitudinal model of the formula and developed a semi-parametric generalized estimating equations (SGEE) framework suitable for balanced and unbalanced (dense and loose) cases, with variance-correlation decomposition for estimating the action matrices.

They demonstrated that SGEE estimators are intuitively more efficient than PULS estimators when the weights are chosen as the inverse of the error covariance matrix. They also showed that the convergence properties differ substantially between dense and loose cases.

In a parallel vein, [7] presented a three-stage approach that combines kernel parametric regression to estimate the link function with GEE for the two-part coefficients. They demonstrated that the semi-parametric information bounds of the parametric components are reached and that the efficiency of the function estimator is improved when the covariance is properly characterized. The research has been extended to more complex cases: [8] addressed the issue of measurement error and presented consistent estimators within the PLSIMeM model with a proof of convergent modularity, emphasizing the need to address error to avoid significant bias. [9] studied a panel version with fixed effects and proposed removing individual effects using the dummy variables method with SMAVE to obtain consistent estimators for the two components and coefficients.

The semi-parametric efficiency framework for parametric components in PLSIM was formulated by [10] by calculating efficiency frontiers based on the Severini and Tripathi method, which allows for evaluating the performance of multiple estimators relative to the theoretical bound.

Building on this literature, my study adopts a PLSIM model with longitudinal data, employs Random Forests to estimate the link function structure and deal with high nonlinearity and interactions, and combines this with statistical methods (such as GEE/SGEE and kernel/parametric regularization) to improve the accuracy, efficiency, of inference compared to traditional approaches.

2 MODEL AND METHOD ESTIMATION

2.1 Model

The partial linear single-index model in longitudinal data is written in the form [11]:

$$Y_{it} = Z_{it}^T\theta + g(X_{it}^T\beta) + \varepsilon_{it} \quad (1)$$
$$, i = 1, \dots, N \ , t = 1, \dots, T$$

Where Y_{it} represents the response variable of individual i at time t, X_{it} and Z_{it} are explanatory variables of the linear and non-linear parts, respectively, with dimensions R^q and R^p, (θ, β) are unknown the parameter vectors of the variables to be estimated with dimension $R^q$and R^p, $\| \beta \| = 1$ ($\|\cdot\|$ representing the Euclidean Norm), and the first element must be positive. $g(\cdot)$ is an unknown link function and ε_{it} are random error and $E(\varepsilon_{it} \setminus X_{it}, Z_{it}) = 0$

2.2 Methods Estimation

In this section, the estimation methodologies used to estimate the partial linear single-index model (PLSIM) parameters in the context of longitudinal data will be presented and detailed. Two different approaches have been adopted, reflecting two contrasting approaches in estimation philosophy.

The first relies on machine learning algorithms, represented by the Random Forests method, which focuses on predictive power and the ability to handle complex, nonlinear relationships in the data. The second approach relies on classical statistical techniques, represented by the Two-Step Method, which is based on estimating parameters in a stepwise manner, combining the linear structure of the model with the nonparametric estimation of the nonlinear function.

2.2.1 Random Forest Method

In this section, we briefly describe the standard Random Forest algorithm based on Classification and Regression Decision Trees (CART). A Random Forest is a cluster of decision trees, each constructed from a bootstrap version of the training data. Each tree is grown according to the principle of recursive splitting; construction starts from the root node and the node-splitting procedure is repeated until certain stopping rules are met.

The basic guiding principle of node splitting is to minimize the impurity of the response variable within each node of the tree. Impurity is often measured by the Gini index when the response is categorical, or by variance when the response is quantitative. In a binary decision tree such as CART, the splitting procedure consists of selecting the splitting variable and determining the splitting rule [12].

The growth of each decision tree terminates if the nodes to be split are indeed pure (i.e., all samples within the node belong to the same class or share a single response value), or if certain pre-specified stopping rules (such as a minimum sample size constraint) are met. The nodes in the final layer of the tree are called leaves and are used to predict new observations.

To perform prediction using a Random Forest, an observation is passed through each decision tree in the forest. In each constructed tree, the observation follows partitioning rules until it settles on a leaf that produces a prediction of its class or response value, depending on whether the task is classification or regression. The final prediction for the observation is then formed either by majority voting (in classification) or by averaging (in regression), based on the results of all the trees in the forest. Because the Random Forest algorithm uses bootstrap sampling to construct each tree, some observations are excluded when constructing a particular tree. By treating these out-of-bag (OOB) observations as observations to be predicted, an estimate of the prediction error of the constructed forest can be obtained.

The Random Forest algorithm has been used in many fields, including genetic epidemiology, bioinformatics, and precision medicine. Its predictive power stems from the aggregation of a large number of weak learners. Performance is excellent, especially when the correlations between trees in the forest are low.

In addition, a variable importance measure can be obtained for each predictor, which measures its relevance to the prediction. This makes it possible to select variables in high-dimensional data based on the variable importance measure. However, a drawback of the random forest is its poor interpretability; unlike a single decision tree, the forest output is difficult to interpret [13].

The partial linear single-index model in longitudinal data is written in the form:

$$Y_{it} = Z_{it}^T\theta + g(X_{it}^T\beta) + \varepsilon_{it}.$$

Here we aim to estimate: the linear coefficients θ, the index parameters β, and the link function $g(\cdot)$. The basic idea is to estimate g in a flexible nonparametric way using random forests on the single index $S_{it} = X_{it}^T\beta$, updating θ by the generalized estimating equations (GEE) method, and updating β by an external optimization method that takes into account the definition of the indicator. The partial residuals are defined by the formula:

$$R_{it} = Y_{it} - Z_{it}^T\theta. \quad (2)$$

We estimate g via a single-entry random forest regression on the pairs(S_{it}, R_{it}). To detect longitudinal correlation, we use cluster bootstrap when building each tree, and measure the out-of-bag error at the individual level. We obtain the approximant, which is the random forest estimator:

$$E[R|S = s] = \hat{g}(s). \quad (3)$$

Then the parameter θ is updated via generalized estimation equations (GEE) and the parameter β is updated by minimizing a loss function based on the performance of the Random Forest (out-of-bag error at the individual level) while keeping the constraint $\| \beta \| = 1$. In practice, a numerical optimization without derivatives is used for each evaluation of β. The Random Forest is retrained and a cluster OOB error is calculated.

2.2.2 Two Stage Method

In the framework of the partial linear single index model (PLSIM), [5]presented a two-stage non-iterative estimation methodology that addresses the difficulty of simultaneous estimation of θ and β by assuming dimensionality reduction of the variable Z through the relationship $Z = \phi(X^T\beta_z) + \eta$ with η being independent of X. The correlated part of Z is removed to estimate θ efficiently with partial regression/profile, and then β is recovered with an estimation equation that exploits the constraint $\| \beta \| = 1$, achieving marginal efficiency higher than common approaches. [14] extended the idea to the case of longitudinal data with a two-stage, iterative estimator that combines local polynomial smoothing with bias-corrected generalized estimating equations (GEE), under standard identification assumptions.

First stage algorithm:

1) First, create a regression model Z and X that is regressed to obtain $\hat{\beta}_z$ an estimate of β_z.
2) Use local smoothing estimation to obtain $\hat{\eta}_{it} = Z_{it} - \varphi\left(X_{it}^T\hat{\beta}_z\right)$ also $Y_{it} = \hat{\eta}_{it}\theta_0 + h\left(X_{it}^T\beta_0 + X_{it}^T\hat{\beta}_z\right) + \varepsilon_{it}$ where η and X are independent of each other
3) For Y and η, a linear regression is performed to obtain $\hat{\theta}_0$ an initial estimate of θ_0.
4) Create a regression model $Y_{it} - Z_{it}^T\hat{\theta}_z$ and X is regressed to obtain $\hat{\beta}_0$. An initial estimate of β_0.
5) Using local polynomial smoothing estimation, the possible initial estimators $\hat{g}(\cdot)$ and $\hat{\acute{g}}(\cdot)$ of the continuum function $g(\cdot)$ and its first derivative $\acute{g}()$ are obtained.

$$\hat{g}(u) = \hat{g}(u;\theta,\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{nit}\,(u;\hat{\beta}_0)\left(Y_{it} - Z_{it}^T\hat{\theta}_z\right), \quad (4)$$

$$\hat{\acute{g}}(u) = \hat{\acute{g}}(u;\theta,\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{W}_{nit}\,(u;\hat{\beta}_0)\left(Y_{it} - Z_{it}^T\hat{\theta}_z\right). \quad (5)$$

Second stage algorithm:

1) Use step 4 to obtain the initial estimate of $\hat{\beta}_0$, which is obtained by solving the generalized estimation equation to correct the bias $\hat{\theta}$ The initial estimate of θ_0

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\{Y_{it} - Z_{it}^T\theta - \hat{g}(X_{it}^T\beta)\}\{Z_{it} - \hat{E}(Z_{it} \setminus X_{it}^T\beta)\} = 0. \quad (6)$$

2) Use the updated initial estimate $\hat{\theta}$ of θ to form a new residual value $Y_{it} - Z_{it}^T\hat{\theta}$, which is then obtained by solving the generalized estimation equation for bias correction as follows: Initial estimate $\hat{\beta}$ of β (7)

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\{Y_{it} - Z_{it}^T\theta - \hat{g}(X_{it}^T\beta)\}\dot{g}(X_{it}^T\beta)\{X_{it} - \hat{E}(X_{it} \setminus X_{it}^T\beta)\} = 0.$$

3) Use the updated estimates of $\hat{\theta}$ and $\hat{\beta}$ in steps 6 and 7 to update the estimates of $\hat{g}$ and $\dot{\hat{g}}$, following the steps in step 5.

The first stage uses linear regression and local linear smoothing estimation to obtain initial estimates of the unknown parameters, the connection functions, their derivatives, and their residuals. In the second stage, the final estimates of the unknown parameters and connection functions are obtained using the generalized estimation equation to correct for bias.

3 EVALUATION METRICS

Comparing estimation methods is a fundamental step in statistical research to select the most appropriate methods. A wide range of evaluation metrics exists; some focus on comparing parameter estimates, while others focus on measuring the performance of the model as a whole. In general, the choice of appropriate metrics depends on the nature of the data and the objectives of the statistical analysis. In this paper, the following evaluation criteria will be used to assess the quality of the estimation and the overall performance of the model:

$$MSE = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \hat{Y}_{it}\right)^2}{N * T},$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \hat{Y}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \hat{Y}_{it}\right)^2},$$

$$MSE\left(g(u)\right) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(g(X_{it}^T\beta) - \hat{g}(X_{it}^T\beta)\right)^2}{N * T},$$

$$R^2\left(g(u)\right) = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(g(X_{it}^T\beta) - \hat{g}(X_{it}^T\beta)\right)^2}{N * T}.$$

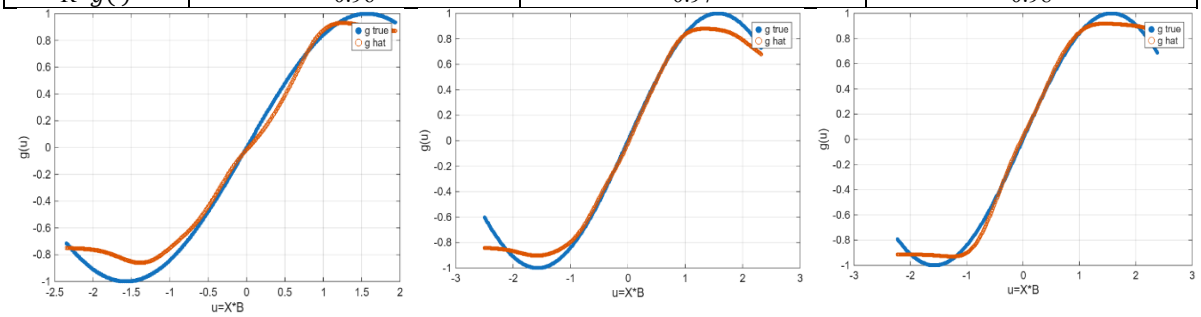
4 SIMULATIONS

The simulation is based on a balanced longitudinal panel of $N = [50\ 100\ 150]$ individuals with each individual observed across $T = 5$ time periods (total NT) and generates data for a partial linear single index model of the form $Y_{it} = Z_{it}^T\theta + g(X_{it}^T\beta) + \varepsilon_{it}$ where $g(\cdot)$ is an unknown nonlinear function and the linear part represents the coefficients of Z. The explanatory variables are first generated as independent standard vectors $X_{it} \sim N(0, I_q)$, $Z_{it} \sim N(0, I_p)$ With dimensions $p = 3$, $q = 2$ Index direction is fixed by normalization $\beta = \frac{1}{\sqrt{3}}[1, -1, 1]^T$ The coefficients of the linear part are proven θ = [0.5, -0.8] ^{T}.A real nonlinear function is defined as $g(u) = \sin(u)$ With index $u = X_{it}^T\beta$ Time dependence within the individual is introduced via a first-order AR(1) autoregressive error process with $\rho = 0.3$ and standard deviation $\sigma = 0.3$ Generated the beginning $\varepsilon_{i1} \sim N(0, \sigma^2)$ then $\varepsilon_i t = \rho * \varepsilon_i(t - 1) + \eta_i t$ where $\eta_i t \sim N(0, \sigma^2)$ With innovations independent across individuals and time periods, independent of (X, Z), and independent of paths between individuals, with this structure, the contribution of the linear component $Z_{it}^T\theta$ and the contribution of the nonlinear component $\boldsymbol{g}(u) = \sin(u)$ are added to the temporal noise to produce the response Y_{it} Also, the Gaussian kernel function was used to estimate the smoothing functions with the two-stage method, and the rule of thumb method was adopted to choose the smoothing parameter.

As reported in Table 1 that the Random Forest-based estimator recovers the nonlinear structure of $g(\cdot)$ very efficiently across the three sample sizes. The coefficient of determination for g remains very high between $N = 100$ and $N = 150$, with a clear improvement in the error for g when moving from $N = 100$ to $N = 150$ ($MSEg$: $0.014 \rightarrow 0.007$). In the scatter plots (Fig. 1), we observe a near-perfect overlap between the true and estimated values in the middle range of $u = X_{it}^T\beta$. This pattern is consistent with the rule used to select the number of trees, which reduces the variance and leaves a small amount of smoothing bias at the boundary, as highlighted by the slight differences between the two curves.

Table 1: Bias and MSEs of the estimates of the parameters and link function using the random forest estimation method.

Parameter	N=50		N=100		N=150	
	bias	MSE	bias	MSE	bias	MSE
θ_1	-0.006904	0.000048	0.034732	0.001206	0.028998	0.000841
θ_2	0.017612	0.000310	-0.001497	0.000002	0.023634	0.000559
β_1	-0.097507	0.009508	0.002478	0.000006	0.069430	0.004821
β_2	0.030754	0.000946	0.043754	0.001914	0.044561	0.001986
β_3	0.108932	0.011866	0.038340	0.001470	-0.031631	0.001000
MSE $g(\cdot)$	0.014395		0.009880		0.007315	
MSE MODEL	0.1284		0.1051		0.1023	
R^2 MODEL	0.90		0.92		0.92	
R^2 $g(\cdot)$	0.96		0.97		0.98	

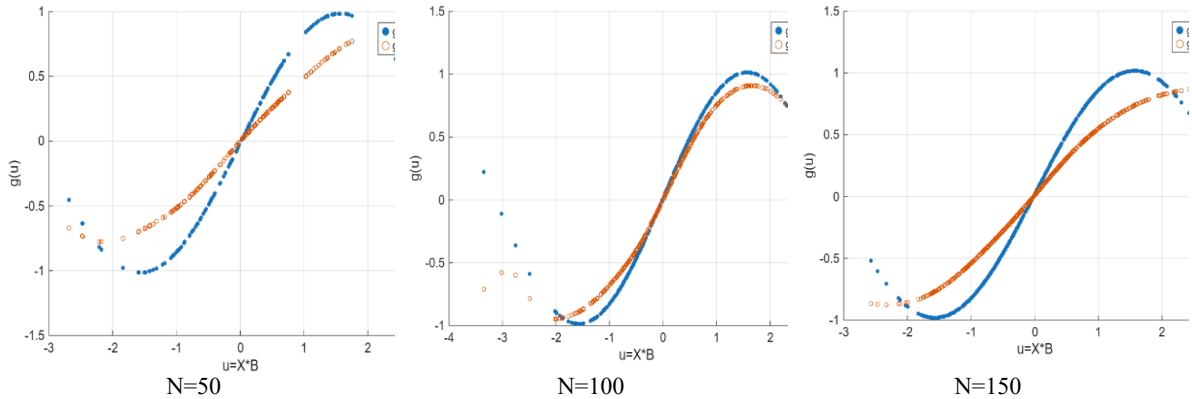


N=50 N=100 N=150

Figure 1: Scatter diagram of true link function and estimated link function using random forest estimation method.

Table 2: Bias and MSE of the estimates of the parameters and link function using two-stage estimation method.

Parameter	N=50		N=100		N=150	
	bias	MSE	bias	MSE	bias	MSE
θ_1	-0.0471	0.0022	0.0447	0.0020	-0.0046	0.00002
θ_2	-0.0503	0.0025	-0.0019	0.000003	0.0337	0.0011
β_1	0.0407	0.0017	-0.1414	0.0200	-0.3047	0.0928
β_2	-0.1841	0.0339	-0.0718	0.0052	0.2480	0.0615
β_3	-0.3819	0.1458	0.0460	0.0021	0.3267	0.1067
MSE $g(\cdot)$	0.048372		0.010492		0.042853	
MSE MODEL	0.315793		0.123913		0.235832	
R^2 MODEL	0.78		0.91		0.83	
R^2 $g(\cdot)$	0.85		0.97		0.89	



N=50 N=100 N=150

Figure 2: Scatter diagram of true link function and estimated link function using two-stage estimation method.

At the linear level, the biases of θ and their squared errors remain very small across all sizes; the recorded values of θ_1, θ_2 are within 10^{-3} or less, with limited non-monotonic fluctuation due to Monte-Carlo fluctuations and nothing more for the index trend β, the results show a pattern expected in single-index models: a significant improvement in one component (here β_3: $MSE = 0.0119$ at $N = 50$ to 0.0010 at $N = 150$), and stability or slight fluctuation in the other components (β_1, β_2) due to the sensitivity of the β update to the derivative of $\hat{g}$ and the density variation along the u axis.

Looking at the overall model performance, the mean squared error decreases from 0.128 at $N = 50$ to $0.105 - 0.102$ at $N = 100 - 150$, while the overall R^2 stabilizes around $0.90/0.92$. This behavior makes sense because the irreducible noise component (with $AR(1)$ and $\sigma = 0.3$) sets a theoretical lower bound on performance even with larger samples, while the improved g estimate and the linear component bring R^2 close to the theoretical value predicted from the analysis of variance. Visually, the three panels show that the overlap in the center tightens with N, and that the tail differences shrink but do not completely disappear due to limited statistical support there, which explains the small remaining difference in MSEg and overall R^2 between the different sizes.

In the two-stage estimator results for the function $g()$, the scattered series in the graphs appear almost identical in the middle range of $u = X_{it}^T\beta$ at N = 100, but show a vertical contraction and slight skew in the tails at N = 50, and especially at N = 150, where $\hat{g}$ tends to be more linear and less curved than true. This marginal gap is not random but is due to the index error caused by the biases of β; a small horizontal shift in u translates into a shape deviation of the estimated function, even if θ is accurately estimated (Fig. 2). Therefore, the R^2 of g is generally high but less stable than in the forest case, decreasing from 0.97 at N=100 to 0.89 at N=150 and associated with a high MSEg (from 0.048 to 0.043 and then to approximately 0.043 with slight variations), reflecting precisely the loss of accuracy at the margins rather than in the core. At the level of linear coefficients, the estimate of θ remains generally good with small biases; however, the non-monotonic oscillations among the three magnitudes (a clear improvement in θ_1 at N = 150 versus a relative decline in θ_2) indicate the sensitivity of the regression stage to the presence of AR (1) time correlation and also to the interference of the linear part with the g level when it is not perfectly aligned. The direction of the index β is most affected: at N = 50, the errors are moderate and some of its components improve at N = 100, but at N = 150, the biases of all components are amplified and the MSE jumps (e.g., β_1: 0.0017 → 0.0200 → 0.0928), a well-known pattern in two-stage estimators when weights based on the derivative of $\hat{g}$ are used (Table 2). As N increases, points on the fringes multiply, showing greater instability in β. The index error pushes points from a high-curvature domain to a low-curvature domain, and the kernel function responds with an over- smoothing. The impact on the overall model performance is clear: the MSE MODEL fluctuates between 0.316 at N=50 and 0.124 at N=100, rising again to 0.236 at N=150, and the overall R^2 drops from 0.91 to 0.83 in the larger sample. This "decline" with the increase in R^2 is not so much an irrational as it is evidence of inconsistency.

5 CONCLUSIONS

This paper aims to estimate a longitudinal partial linear single index model with AR (1) time errors, and to highlight the effect of algorithm design on the quality of retrieval of the nonlinear function $g(\cdot)$ and the accuracy of the linear parameters θ and the trend of the index θ. The results show that the Random Forest estimator, with the number of trees chosen according to the 1-SE rule, provides high-quality retrieval of the function $g(\cdot)$ across the three sample sizes. R_g^2 remained within the range of 0.96–0.98, and the mean square error of the function decreased significantly as N increased. The overall performance of the model stabilized around $R^2 \approx 0.90/0.92$ and $MSE \approx 0.10/0.13$. This accuracy is explained by the bias-variance balance achieved by the 1-SE rule, which reduces the variance of the estimation of θ and limits the propagation of the error to a fraction of g. In contrast, although the two-stage estimator maintained good performance at the core (especially at N=100), it turned out to be more sensitive to global band and index error; the quality of g visually deteriorated at the margins and its overall metrics fluctuated as N increased (overall R^2 falling to about 0.83 and MSE rising to about 0.24 at N=150). The increasing biases of β with larger size indicate that weights based on the derivative of $\hat{g}$ reduce information in low-slope regions and amplify the effect of leverage points at the extremities, which translates into a systematic marginal contraction of $\hat{g}$ even with a reasonable estimate of θ. In conclusion, the Random Forests and Time Correlation Evaluation-based approach provides an accurate path

to recovering nonlinear structure in longitudinal PLSIM models, outperforming the two-stage estimator in our setting in terms of stability and fit, with a clear roadmap for future methodological improvements focusing on local adaptation, removing bias in β, and strengthening theoretical safeguards under more realistic scenarios.

REFERENCES

[1] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand, "Generalized partially linear single-index models," J. Am. Stat. Assoc., vol. 92, no. 438, p. 477, 1997, doi: 10.2307/2965697.

[2] D. Ruppert, M. P. Wand, and R. J. Carroll, Semiparametric Regression, Cambridge University Press, 2003, doi: 10.1017/cbo9780511755453.

[3] Y. Xia and W. Härdle, "Semi-parametric estimation of partially linear single-index models," J. Multivar. Anal., vol. 97, no. 5, pp. 1162-1184, 2006, doi: 10.1016/j.jmva.2005.11.005.

[4] H. Liang, X. Liu, R. Li, and C. L. Tsai, "Estimation and testing for partially linear single-index models," Ann. Stat., vol. 38, no. 6, pp. 3811-3836, 2010, doi: 10.1214/10-AOS835.

[5] J. L. Wang, L. Xue, L. Zhu, and Y. S. Chong, "Estimation for a partial-linear single-index model," Ann. Stat., vol. 38, no. 1, pp. 246-274, 2010, doi: 10.1214/09-AOS712.

[6] J. Chen, D. Li, H. Liang, and S. Wang, "Semiparametric GEE analysis in partially linear single-index models for longitudinal data," Ann. Stat., vol. 43, no. 4, pp. 1682-1715, 2015, doi: 10.1214/15-AOS1320.

[7] Q. Cai and S. Wang, "Efficient estimation in partially linear single-index models for longitudinal data," Scand. J. Stat., vol. 46, no. 1, pp. 116-141, 2019, doi: 10.1111/sjos.12340.

[8] H. Liang and N. Wang, "Partially linear single-index measurement error models," Stat. Sin., vol. 15, no. 1, pp. 99-116, 2005.

[9] J. Chen, J. Gao, and D. Li, "Estimation in partially linear single-index panel data models with fixed effects," J. Bus. Econ. Stat., vol. 31, no. 3, pp. 315-330, 2013, doi: 10.1080/07350015.2013.775093.

[10] T. Chen and T. Parker, "Semiparametric efficiency for partially linear single-index regression models," J. Multivar. Anal., vol. 130, pp. 376-386, 2014, doi: 10.1016/j.jmva.2014.06.006.

[11] S. Ma, H. Liang, and C. L. Tsai, "Partially linear single index models for repeated measurements," J. Multivar. Anal., vol. 130, pp. 354-375, 2014, doi: 10.1016/j.jmva.2014.06.011.

[12] L. Capitaine, R. Genuer, and R. Thiébaut, "Random forests for high-dimensional longitudinal data," Stat. Methods Med. Res., vol. 30, no. 1, pp. 166-184, 2021, doi: 10.1177/0962280220946080.

[13] E. H. Young and R. D. Shah, "ROSE random forests for robust semiparametric efficient estimation," 2024, [Online]. Available: http://arxiv.org/abs/2410.03471.

[14] C. Chang, "Research on two-stage estimation of partially linear single-index model with longitudinal data," Acad. J. Sci. Technol., vol. 5, no. 1, pp. 112-115, 2023.