

## Supplemental Material for “Exact Inference for Transformed Large-Scale Varying Coefficient Models with Applications”

This supplemental material summarizes more technical details to support the finding from the main content. Appendix A gives the details of a restriction of using sufficient statistics and Poisson regression for migration studies in big data computing. Appendices B - E detailed the proof of Theorems 1 to 4. Appendix F shares more information of the motivation dataset. Appendix G give more simulation results.

### Appendix A: The sufficient statistics for Poisson Linear Regression and its limitation in big data computing

The Poisson linear regression model is widely used for modeling count data, including migration data. For illustrating the characteristics of its sufficient statistics, suppose there is a given dataset  $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ , where  $y_i$  denotes the count response variable and  $\mathbf{x}_i$  represents the vector of predictor variables for the  $i$ -th observation, the model assumes that  $y_i$  follows a Poisson distribution with a mean parameter  $\lambda_i$  such that:  $y_i \mid \mathbf{x}_i \sim \text{Poisson}(\lambda_i)$  with  $\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ , where  $\boldsymbol{\beta}$  is the vector of regression coefficients. The likelihood function for the entire dataset can be expressed as the product of the individual likelihoods  $L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^n P(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) = \prod_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$ , where  $\lambda_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$ . The corresponding log-likelihood function is given by:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left( y_i \mathbf{x}_i^T \boldsymbol{\beta} - e^{\mathbf{x}_i^T \boldsymbol{\beta}} - \log(y_i!) \right).$$

From the likelihood function, the sufficient statistics for the regression coefficients  $\boldsymbol{\beta}$  are

$$\left( \sum_{i=1}^n y_i \mathbf{x}_i, \mathbf{X} \right).$$

The term  $\sum_{i=1}^n y_i \mathbf{x}_i$  encapsulates the information from the response variable  $y_i$ , while the matrix  $\mathbf{X}$  includes the covariate information necessary for estimating  $\boldsymbol{\beta}$ . To utilize these sufficient statistics, it remains necessary to read all data points  $\mathbf{X}$ . This requirement poses significant computational challenges, especially with large datasets. Consequently, developing efficient computational methods to handle and process large-scale data is crucial for the practical application of Poisson regression models in extensive datasets.

### Appendix B: Proof of Theorem 1

Given the Bayesian linear model

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon$$

with  $\epsilon \sim N(0, \sigma^2 \mathbf{I})$  and  $\boldsymbol{\beta} \sim N(0, \mathbf{P}_{\lambda^*}^{-1})$ , the posterior distribution of  $\boldsymbol{\beta}$  given data is proportional to

$$\begin{aligned} & -\frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{P}_{\lambda^*} \boldsymbol{\alpha} - \frac{1}{2\sigma^2} (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\alpha})^\top (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\alpha}) \\ \propto & -\frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{P}_{\lambda^*} \boldsymbol{\alpha} - \frac{1}{2\sigma^2} \left( \boldsymbol{\alpha}^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \boldsymbol{\alpha} - 2\mathbf{y}^\top \tilde{\mathbf{X}} \boldsymbol{\alpha} + \mathbf{y}^\top \mathbf{y} \right) \\ \propto & -\frac{1}{2} (\boldsymbol{\alpha} - \sigma^{-2} (\sigma^{-2} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \mathbf{P}_{\lambda^*})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y})^\top (\sigma^{-2} (\sigma^{-2} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \mathbf{P}_{\lambda^*})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y} \\ & \text{(by completing the square)} \end{aligned}$$

This implies the posterior ditribution is a multivariate normal distribution with mean

$$-\sigma^{-2}(\sigma^{-2}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \mathbf{P}_{\lambda^*})^{-1}\tilde{\mathbf{X}}^T\mathbf{y} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \mathbf{P}_{\lambda})^{-1}\tilde{\mathbf{X}}^T\mathbf{y},$$

which is exactly the coefficient estimator (2) from the P-spline method, and the posterior variance covariance matrix is  $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \mathbf{P}_{\lambda}$ . This completes the proof of Theorem 1.

## Appendix C: Proof of Theorem 2

Consider the bayesian linear model first

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon$$

with  $\epsilon \sim N(0, \sigma^2\mathbf{I})$  whose likelihood function is proportional (up to a constant) to

$$\begin{aligned} & \exp\left(-\frac{1}{2}(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta})\right) \\ &= \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\tilde{\mathbf{X}}^T\mathbf{y} + g(\mathbf{y})^Tg(\mathbf{y})\right]\right) \\ &= \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}^T\left(\sum_{k=1}^K\sum_{\ell=1}^{n_k}\tilde{\mathbf{x}}_{k,\ell}^T\tilde{\mathbf{x}}_{k,\ell}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\left(\sum_{k=1}^K\sum_{\ell=1}^{n_k}\tilde{\mathbf{x}}_{k,\ell}g(y_{k,\ell})\right) + \sum_{k=1}^K\sum_{\ell=1}^{n_k}g(y_{k,\ell})^2\right]\right) \end{aligned}$$

The sufficient statistics for  $\boldsymbol{\beta}$  obtained from the likelihood function are

$$\boldsymbol{\Gamma}_k = \left(a_k = \sum_{\ell=1}^{n_k} g(y_{k,\ell})^2, \mathbf{b}_k = \sum_{\ell=1}^{n_k} g(y_{k,\ell})\tilde{\mathbf{x}}_{k,\ell}, \mathbf{C}_k = \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell}^T\tilde{\mathbf{x}}_{k,\ell}\right)$$

With these notation, the regression coefficient estimator (3) of the main paper can be expressed as

$$\hat{\boldsymbol{\beta}} = \left(\sum_k \mathbf{C}_k + \mathbf{P}_{\lambda}\right)^{-1} \left(\sum_k \mathbf{b}_k\right),$$

which is exactly the  $\hat{\boldsymbol{\alpha}}$  in (3) of the main paper. The average of square residuals served as the estimator for the variance can be expressed as

$$\hat{\sigma}^2 = \frac{\sum_k a_k - (\sum_k \mathbf{b}_k)^T (\sum_k \mathbf{C}_k)^{-1} (\sum_k \mathbf{b}_k)}{n}$$

Therefore, if the sufficient statistics  $\boldsymbol{\Gamma}_k$  are recorded for each subset  $\mathbf{D}_k$ , we can exactly recover the estimators  $\boldsymbol{\alpha}$  and  $\sigma^2$  as specified in the theorem. This completes the proof.

## Appendix D: Proof of Theorem 3

Recall that with a local linear expression on each varying coefficient function of model (1) of the main paper expanded at  $\mathbf{h}_0$ , we can express all the unknown regression coefficient as a vector  $\boldsymbol{\beta}$  and the associated model matrix is denoted by  $\tilde{\mathbf{X}}$ . Under the notation with implementing the weighted least square criterion for local polynomial methods for deriving the optimizer of  $\boldsymbol{\beta}$ , the resulting estimator can be expressed as

$$\hat{\beta} = (\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{W} g(\mathbf{y}), \quad (1)$$

where  $\mathbf{W}$  is a diagonal matrix, which can be expressed as  $\text{diag}(K_\delta(\|\mathbf{h}_{111} - \mathbf{h}_0\|), \dots, K_\delta(\|\mathbf{h}_{1J1} - \mathbf{h}_0\|), \dots, K_\delta(\|\mathbf{h}_{IJ1} - \mathbf{h}_0\|), \dots, K_\delta(\|\mathbf{h}_{11T} - \mathbf{h}_0\|), \dots, K_\delta(\|\mathbf{h}_{1JT} - \mathbf{h}_0\|), \dots, K_\delta(\|\mathbf{h}_{IJT} - \mathbf{h}_0\|))$ .

Suppose we partition the dataset into  $K$  subsets denoted by  $\{\mathbf{D}_k\}_{k=1}^K = \{y_{k,\ell}, \tilde{\mathbf{x}}_{k,\ell}\}_{\ell=1}^{n_k}$ , where  $n_k$  is the sample size of the  $k$ -th sub-dataset. Under the Bayesian linear model (8) of the main paper, the likelihood function is proportional to

$$\begin{aligned} & \exp\left(-\frac{1}{2}(g(\mathbf{y}) - \tilde{\mathbf{X}}\beta)^T W(g(\mathbf{y}) - \tilde{\mathbf{X}}\beta)\right) \\ = & \exp\left(-\frac{1}{2}\left[\beta^T \tilde{\mathbf{X}}^T W \tilde{\mathbf{X}} \beta - 2\beta^T \tilde{\mathbf{X}}^T W g(\mathbf{y}) + g(\mathbf{y})^T W g(\mathbf{y})\right]\right) \\ = & \exp\left(-\frac{1}{2}\left[\beta^T \left(\sum_{k=1}^K \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell}^T W_k \tilde{\mathbf{x}}_{k,\ell}\right) \beta - 2\beta^T \left(\sum_{k=1}^K \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell}^T W_k \mathbf{y}_k\right) + \sum_{k=1}^K \sum_{\ell=1}^{n_k} g(\mathbf{y}_k)^T W_k g(\mathbf{y}_k)\right]\right) \end{aligned}$$

This implies the sufficient statistics for estimating  $\beta$  from the likelihood function are

$$(a_k^{LL} = \sum_{\ell=1}^{n_k} g(\mathbf{y}_k)^T W_k g(\mathbf{y}_k), \mathbf{b}^{LL}_k = \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell}^T W_k \mathbf{y}_k, \mathbf{C}_k^{LL} = \sum_{\ell=1}^{n_k} \tilde{\mathbf{x}}_{k,\ell}^T W_k \tilde{\mathbf{x}}_{k,\ell}) \quad (2)$$

and the average of the squares of residuals, a consistent estimator of  $\sigma^2$  in model (1) of the main paper can be expressed as

$$\hat{\sigma}^2 = \frac{\sum_k a_k^{LL} - (\sum_k \mathbf{b}_k^{LL})^T (\sum_k \mathbf{C}_k^{LL})^{-1} (\sum_k \mathbf{b}_k^{LL})}{n}$$

Therefore, if the sufficient statistics are recorded for each subset  $\mathbf{D}_k$ , we can exactly recover the estimators  $\alpha$  and  $\sigma^2$  as specified in the theorem. This completes the proof of Theorem 3.

## Appendix E: Proof of Theorem 4

Denote the linear smoother matrix from the proposed estimator (3) of the main paper by  $\mathbf{S}$ , i.e.,  $\mathbf{S}g(\mathbf{y}) = \tilde{\mathbf{X}}^T \hat{\alpha} = \tilde{\mathbf{X}}^T (\sum_k \mathbf{C}_k + \mathbf{P}_\lambda)^{-1} (\sum_k \mathbf{b}_k)$ . Let  $s_{vw}$  be  $(v, w)$ -th element of the smoother matrix  $\mathbf{S}$ . When we delete the  $v$ -th column, the  $v$ -th row now sums to  $1 - s_{vv}$ . Renormalizing by this factor, we can express the prediction value evaluated at  $v$ -th data point trained by the dataset whose  $v$ -th data point is deleted by  $g(\hat{y}_{-v}) = \frac{1}{1-s_{vv}} \sum_{\substack{w=1 \\ w \neq v}}^n s_{vw} g(y_w)$ . Note that the original predicted value (use whole dataset without deleting  $v$ -th data point) is  $g(\hat{y}_v) = \sum_{w=1}^n s_{vw} g(y_w)$ . Multiplying by  $1 - s_{ii}$  on  $g(\hat{y}_{-v})$  and rearranging terms yields:

$$\begin{aligned} g(\hat{y}_{-v}) &= \sum_{\substack{w=1 \\ w \neq v}}^n s_{vw} g(y_w) + s_{vv} g(\hat{y}_{-v}) \\ &= \sum_{w=1}^n s_{vw} g(y_w) + s_{vv} g(\hat{y}_{-v}) - s_{vv} g(y_v) \\ &= g(\hat{y}_v) + s_{vv} g(\hat{y}_{-v}) - s_{vv} g(y_v) \end{aligned}$$

From which we conclude

$$g(y_v) - g(\hat{y}_{-v}) = g(y_v) - g(\hat{y}_v) + s_{vv}(g(y_v) - g(\hat{y}_{-v})), \text{ which implies}$$

$$g(y_v) - g(\hat{y}_{-v}) = \frac{g(y_v) - g(\hat{y}_v)}{1 - s_{vv}}$$

This equation shows that the leave-one-out residual can be computed from the residuals obtained without leaving one out and the diagonal elements  $s_{vv}$  of the smoother matrix  $\mathbf{S}$ . Thus, the cross-validation criterion becomes

$$\text{cv}_\lambda = \frac{1}{n} \sum_{v=1}^n \left( \frac{g(y_v) - g(\hat{y}_v)}{1 - s_{vv}} \right)^2$$

Note that  $g(\hat{y}_v)$  and  $s_{vv}$  for all  $v$  are dependent on  $\lambda$ .

For Generalized cross-validation, we replace  $s_{vv}$  in the denominator with their average denoted by  $\text{tr}(\mathbf{S})/n$ , where  $\text{tr}(\mathbf{S})$  is the trace operator taking the summation of the diagonal matrix for matrix  $\mathbf{S}$

$$\text{gcv}_\lambda = \frac{1}{n} \sum_{v=1}^n \left( \frac{g(y_v) - g(\hat{y}_v)}{1 - \text{tr}(\mathbf{S})/n} \right)^2$$

Thus,  $\text{gcv}_\lambda$  is a simple function of the average squared residual and this completed the proof of Theorem 4.

## Appendix F: More Information about the Motivation Dataset

Table 1 summarizes the variables we considered in the motivation dataset. The table also presents whether the variables are significant important or not by using the method described in section 4.3.

In addition to the general correlation values listed in Table 1, we further examine the correlation values related to migration patterns. For clearer illustration, Figure 1 provides a visual representation where the upper panel displays the number of migration pairs originating from each state, and the lower panel shows the number of migration pairs moving out of each state. This factor is closely related to the significant influence of the distance between the origins and destinations of migration pairs. From the figure, it is evident that outbound migration patterns are strongly associated with migration counts—darker (lighter) blue areas in Figure A often correspond to darker (lighter) blue areas in Figure B. However, the strength of this relationship varies across states. This observation motivates the application of the proposed varying coefficient (VC) model to the migration data for deeper insights, as further detailed in Section 5.

## Appendix G: Additional Simulation Studies

In addition to the simulation studies from Model 1 in the main paper, this section presents additional simulations to evaluate the robustness of the proposed method. Specifically, we consider three simulation settings generated from the model

$$Y = \sin(2U) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \varepsilon :$$

Variable Description	Significance	Variable Description	Significance
Number of Flood Claims (O)	Yes	Number of Flood Claims (D)	Yes
Near the Coastal Area (O)	Yes	Near the Coastal Area (D)	No
Having Hurricane in the past 6 months (O)	Yes	Having Hurricane in the past 6 months (D)	No
Population (O)	No	Population (D)	Yes
Percentage of Personal Residence (O)	Yes	Percentage of Personal Residence (D)	Yes
Percentage of Hispanic People (O)	No	Percentage of Hispanic People (D)	No
Percentage of Black People (O)	No	Percentage of Black People (D)	No
Percentage of White People (O)	No	Percentage of White People (D)	No
Longitude (O)	No	Longitude (D)	No
Latitude (O)	Yes	Latitude (D)	Yes
The distance between origin and destination	Yes		

Table 1: Summary table of variable descriptions with their significance for the motivation migration dataset: (O) indicate the variable is for origin and (D) is for destination.

**Setting 2-1:** The covariates  $U, Z_1, Z_2$ , and  $Z_3$  are generated as follows.  $U, Z_1$ , and  $Z_2$  are jointly normally distributed with mean 0, variance 1, and pairwise correlation coefficients of 0.5. The binary covariate  $Z_3$  is independent of  $U, Z_1$ , and  $Z_2$ , taking the value 1 with probability 0.4 and 0 with probability 0.6. The model parameters are set such that  $\beta_1 = 2$ , while  $\beta_2$  and  $\beta_3$  are both set to  $\theta$ , with  $\theta = 0$ .

**Setting 2-2:** This setting is identical to Setting 2-1, except that  $\theta = 0.5$ , introducing correlation into the model.

**Setting 2-3:** This setting is the same as Setting 2-1, but the error term  $\varepsilon$  follows a mixture normal distribution given by

$$\varepsilon \sim \frac{2}{3}N\left(0, \frac{1}{2}\right) + \frac{1}{3}N(0, 2),$$

introducing non-i.i.d. and non-normal error terms.

These settings are intentionally designed to deviate from the assumptions of the proposed method, creating more challenging scenarios to assess its robustness. Setting 2-1 explores the method's performance when the covariates include non-varying coefficients. Setting 2-2 examines the effect of correlated inputs on model accuracy. Setting 2-3 investigates the impact of non-i.i.d. and non-normal error terms. By analyzing these scenarios, we aim to determine whether the proposed method can maintain its performance under varying and less ideal conditions.

The simulation is repeated 100 times and the results are summarized in Table 2. From the table, we observed that the proposed method demonstrates robust performance in this example, even under more challenging conditions. The running times are similar to those observed in Example 1, as the scale of the dataset and computational complexity are comparable. However, the prediction errors, measured using RMSPE, are larger in this example due to the presence

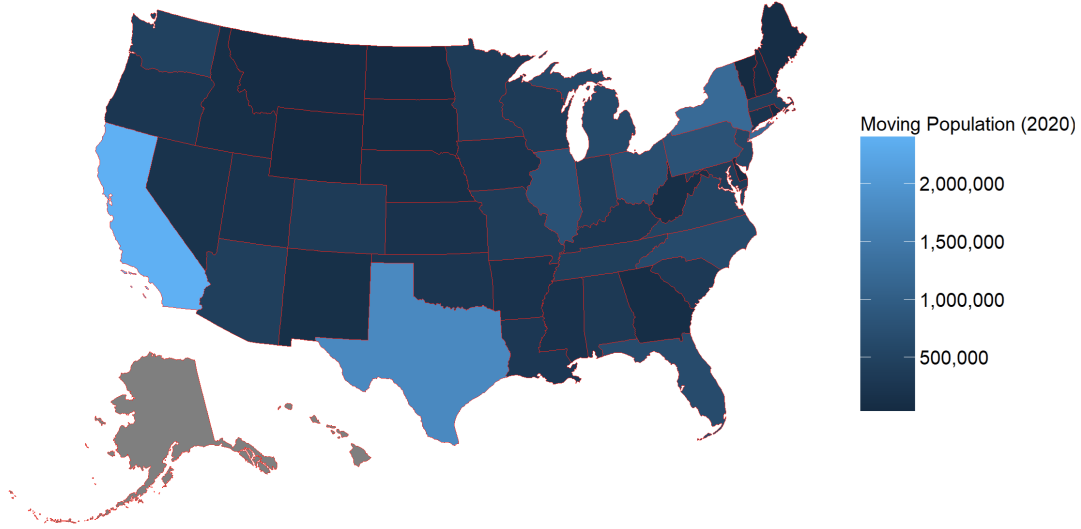
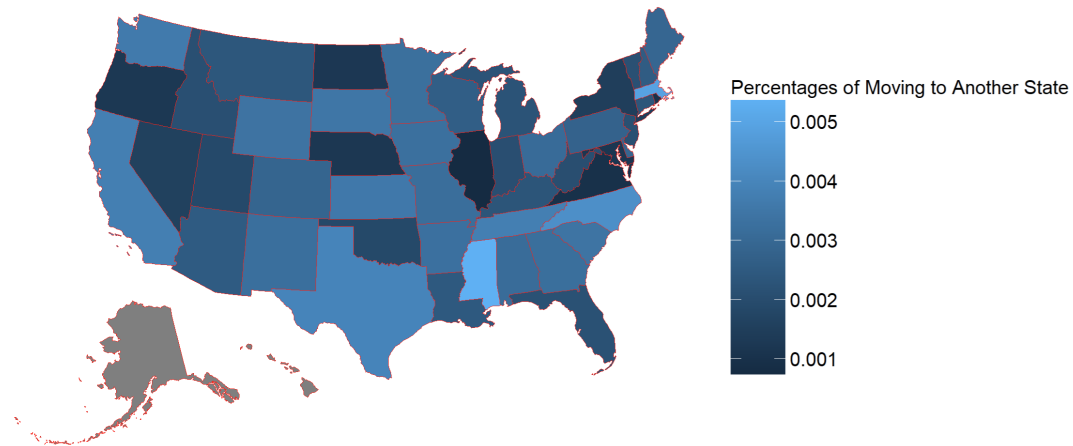
**A****B**

Figure 1: Caption

of more complex error terms, as discussed in the setup. Despite these additional challenges, the proposed method consistently outperforms or remains comparable to other competing methods in both time efficiency and prediction accuracy. This highlights the method's ability to maintain reliable performance while effectively balancing computational cost and predictive power, even in the face of increased model complexity.

Methods	TAVCM	Linear Model	Theorem 3
Case 2-I: Sample Size $10^6$ (Mean (sd))			
Time	1.375 (0.097)	0.001 (0.001)	1.362 (0.102)
RMSPE	0.052 (0.002)	1.231 (0.237)	0.054 (0.002)
Case 2-II: Sample Size $10^7$ (Mean (sd))			
Time	2.516 (0.126)	0.002 (0.005)	2.532 (0.157)
RMSPE	0.050 (0.001)	1.132 (0.228)	0.051 (0.001)
Case 2-III: Sample Size $10^8$ (Mean (sd))			
Time	4.128 (0.209)	0.002 (0.003)	4.177 (0.127)
RMSPE	0.051 (0.001)	1.037 (0.216)	0.050 (0.001)
Case 3-I: Sample Size $10^6$ (Mean (sd))			
Time	1.421 (0.091)	0.001 (0.001)	1.419 (0.113)
RMSPE	0.083 (0.005)	1.872 (0.237)	0.081 (0.004)
Case 3-II: Sample Size $10^7$ (Mean (sd))			
Time	2.512 (0.102)	0.002 (0.002)	2.434 (0.101)
RMSPE	0.079 (0.004)	1.872 (0.212)	0.076 (0.004)
Case 3-III: Sample Size $10^8$ (Mean (sd))			
Time	4.325 (0.105)	0.003 (0.002)	4.332 (0.103)
RMSPE	0.078 (0.004)	1.912 (0.201)	0.079 (0.004)
Case 4-I: Sample Size $10^6$ (Mean (sd))			
Time	1.532 (0.094)	0.002 (0.001)	1.487 (0.115)
RMSPE	0.109 (0.015)	2.203 (0.289)	0.103 (0.011)
Case 4-II: Sample Size $10^7$ (Mean (sd))			
Time	2.623 (0.108)	0.003 (0.002)	2.521 (0.104)
RMSPE	0.103 (0.013)	2.242 (0.271)	0.103 (0.012)
Case 4-III: Sample Size $10^8$ (Mean (sd))			
Time	4.537 (0.112)	0.004 (0.002)	4.412 (0.108)
RMSPE	0.102 (0.012)	2.291 (0.273)	0.101 (0.010)

Table 2: Summary of numerical comparison results in terms of root mean square prediction errors (RMSPE) and the fitting time of the proposed methods for cases 2 and 3