Varying Coefficient Model via Adaptive Spline Fitting

Xufei Wang
Two Sigma Investments, LP *
Bo Jiang
Two Sigma Investments, LP and
Jun S. Liu
Department of Statistics, Harvard University

October 2, 2023

Abstract

The varying coefficient model is a potent dimension reduction tool for nonparametric modeling and has received extensive attention from researchers. Most existing methods for fitting this model utilize polynomial splines with equidistant knots and treat the number of knots as a hyperparameter. However, imposing equidistant knots tends to be overly rigid, and systematically determining the optimal number of knots is also challenging. In this article, we address these challenges by employing polynomial splines with adaptively selected and predictor-specific knots to fit the varying coefficients in the model. We propose an efficient dynamic programming algorithm to find the optimal solution. Numerical results demonstrate that our new method achieves significantly smaller mean squared errors for coefficient estimations compared to the equidistant spline fitting method. An implementation of our method in R is available at https://github.com/wangxf0106/vcmasf. Proofs of the theorems are provided in the online supplementary materials.

Keywords: Mean squared error; Splines and knots; Varying coefficient model.

*The views expressed herein are the authors alone and are not necessarily the views of Two Sigma Investments, LP, or any of its affiliates.

1 Introduction

The accurate estimation of the relationship between a response variable and multiple predictor variables is a fundamental challenge in statistical machine learning and various scientific applications. Linear regression, among parametric models, is a simple yet powerful approach. However, its linearity assumption is often violated in real-world applications, limiting its effectiveness.

Nonparametric models, on the other hand, do not assume any specific relationship between the response and predictors, allowing for greater flexibility in modeling nonlinear relationships. However, fitting nonparametric models requires imposing local smoothness conditions, typically achieved using specific kernels or spline basis functions. This is necessary to avoid the overfitting issue, where the model becomes too complex and performs well on the training data but poorly on new, unseen data. Unfortunately, this general strategy is susceptible to the curse of dimensionality, especially when dealing with high-dimensional datasets. In such cases, these methods become ineffective in capturing the true relationship and computationally expensive.

Addressing these challenges is essential to develop robust and efficient models capable of handling complex data relationships. Researchers are actively exploring novel techniques and algorithms to overcome the limitations of linear regression's linearity assumptions and the curse of dimensionality in nonparametric models.

The varying coefficient model (Hastie and Tibshirani, 1993) serves as a bridge between linear and nonparametric models, offering an appealing compromise between simplicity and flexibility. In this class of models, regression coefficients are not fixed constants; instead, they vary as a function of certain conditioners, resulting in a more flexible approach due to the infinite dimensionality of the corresponding parameter space. Varying coefficient modeling presents a powerful strategy to address the curse of dimensionality, setting it apart from standard nonparametric approaches. Additionally, it inherits advantages from linear models, such as simplicity and interpretability.

A typical setup for the varying coefficient model is as follows: given the response variable $y \in \mathbf{R}$ and predictors $X = (x_1, \ldots, x_p)^\top \in \mathbf{R}^p$, the model assumes the relationship: $y =$

$\sum_{j=1}^{p} \beta_j(u)x_j + \epsilon$, where u is the conditional random variable, typically represented as a scalar. This modeling approach has found diverse applications across various data types, including longitudinal data (Huang et al., 2004; Tang and Cheng, 2012), functional data (Zhang and Wang, 2014; Hu et al., 2019), spatial data (Wang and Sun, 2019; Finley and Banerjee, 2020), and can be naturally extended to address different types of time series data (Huang and Shen, 2004; Lin et al., 2019).

There are three major approaches to estimating the coefficients $\beta_j(u)$ $(j = 1, \ldots, p)$. One widely acknowledged approach is the smoothing spline method proposed by Hastie and Tibshirani (1993), with recent follow-up work using P-spline by Jullion et al. (2009). Another approach, proposed by Fan and Zhang (1999) and Fan and Zhang (2000), is the predictor-specific kernel method for coefficient estimation. This method involves local linear smoothing to model function $\beta_j(u)$ in the first step, followed by applying local cubic smoothing on the residuals in the second step. A recent adaptation of their work is an adaptive estimator by Chen et al. (2015).

The third approach involves approximating the coefficient functions using a basis expansion, such as polynomial B-splines. This method has gained popularity due to its simplicity in estimation and inference, along with good theoretical properties. Compared to smoothing spline and kernel methods, the polynomial spline method with a finite number of knots strikes a balance between model flexibility and interpretability. Huang et al. (2002), Huang and Shen (2004), and Huang et al. (2004) utilized a set of polynomial estimators, assuming equidistant knots and choosing the number of knots such that the bias terms become asymptotically negligible to ensure local asymptotic normality.

Most polynomial spline approaches involve optimizing a set of finite-dimensional classes of functions, such as the space of polynomial B-splines with L equally spaced knots. However, if the real turning points of the coefficients are not equidistant, using equally spaced knots requires selecting a sufficiently large L to capture the resolution of the coefficients accurately. In practice, determining the value of L involves a parameter search process, alongside the estimation of other parameters. This comparison is necessary to identify the optimal fixed number of knots. Selecting too few knots might overlook the high-frequency

information of $\beta_j(u)$, whereas selecting too many knots could lead to overfitting in regions where the coefficients barely change. Moreover, when the number of predictors is very large, and possibly even exceeds the sample size, dealing with the issue of variable selection can further complicate the matter.

In this paper, we propose two adaptive algorithms for fitting piece-wise linear functions with automatically selected turning points for the univariate conditioner variable u. These algorithms offer significant advantages over existing methods, as they can automatically determine the optimal positions of knots, which model the turning points of the true coefficients. We demonstrate that our methods select knots that are almost surely the true change points when the coefficients are piece-wise linear in u. Additionally, we show that the residual variance of the fitted model converges to the true data variance. To address high-dimensional settings, we combine the knots selection algorithms with the adaptive group LASSO method for variable selection, inspired by the idea of Wei et al. (2011) who applies the adaptive group LASSO to basis expansions of predictors.

In our simulation studies, we illustrate that the new adaptive method achieves smaller mean squared errors (MSEs) for estimating coefficients compared to available methods and also improves variable selection performance. Finally, we apply the method to two real datasets: (a) a COVID-19 infection dataset for the state of New York, where we observe that the association between environmental factors and COVID-19 infected cases varies over time; (b) the Boston Housing data (Harrison and Rubinfeld, 1978), where we investigate how factors affecting housing prices vary in effect along with the educational level of the location.

2 Methods and theory for adaptive spline fitting

2.1 Knots selection for polynomial spline

In varying coefficient models, each coefficient $\beta_j(u)$ is a function of the conditional variable u, which we estimate by fitting a polynomial spline on u. In this paper, we assume that u is a univariate variable. Let $X_i = (x_{i,1}, \ldots, x_{i,p})^\top \in \mathbf{R}^p$, u_i, and y_i denote the ith observations

of the predictor vector, the conditional variable, and the response variable, respectively, respectively, for $i = 1, \ldots, n$.

We suppose that the knots are common to all coefficients and located at $d_1 < \ldots < d_L$, with the corresponding B-splines of order D denoted as $B_k(u)$ $(k = 1, \ldots, D + L)$, which are piece-wise polynomials of degree $D - 1$. Each varying coefficient can be represented as $\beta_j(u) = \sum_{k=1}^{D+L} h_{j,k} B_k(u)$, where the coefficients $h_{j,k}$ are estimated by minimizing the following sum of squared errors:

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} x_{i,j} \sum_{k=1}^{D+L} h_{j,k} B_k(u_i) \right\}^2 . \tag{1}$$

In previous work, the knots for polynomial splines were typically chosen as equidistant quantiles of u and were the same for all predictors. While the approach is computationally straightforward, the knots chosen in this way cannot adequately reflect the varying smoothness between and within the coefficients. To address this limitation, we propose an adaptive knot selection approach where the knots can be interpreted as turning points of the coefficients.

For knots $d_1 < \ldots < d_L$, we define the segmentation scheme $S = \{s_1, \ldots, s_{L+1}\}$ for the observed samples ordered by u, where $s_\ell = \{i \mid d_{\ell-1} < u_i \leq d_\ell\}$, with $d_0 = -\infty$ and $d_{L+1} = \max\{u\}$. If the true coefficients $\beta(u_i) = (\beta_1(u_i), \ldots, \beta_p(u_i))^\top \in \mathbf{R}^p$ form a linear function of u within each segment s, i.e., $\beta(u_i) = a_s + b_s u_i$ for $a_s, b_s \in \mathbf{R}^p$, then the observed response satisfies:

$$y_i = a_s^\top X_i + b_s^\top (u_i X_i) + \epsilon_i, \ \epsilon_i \sim N(0, \sigma_s^2). \tag{2}$$

Thus, the coefficients can be estimated by maximizing the log-likelihood function, which is equivalent to minimizing the loss function:

$$\text{Loss}(S) = \sum_{s \in S} |s| \log \hat{\sigma}_s^2, \tag{3}$$

where $|s|$ denotes the number of data points in segmentation s, and $\hat{\sigma}_s^2$ is the residual variance obtained by regressing y_i over $(x_{i,1}, \ldots, x_{i,p}, u_i x_{i,1}, \ldots u_i x_{i,p})^\top$ for $i \in s$. We also use $|S|$ to denote the number of segments in S (which is equal to $L + 1$).

Because any almost-everywhere continuous function can be approximated by piece-wise linear functions, we can employ the estimation framework in (2) and (3). Since Loss(S)

in (3) always decreases as we break one of its segments arbitrarily into two, we need to penalize the number of segments $|S|$ while ensuring that the number of data points $|s|$ within a segment s is greater than a lower bound $m_s = n^\alpha$ $(0 < \alpha < 1)$:

$$\text{Loss}(S, \lambda_0) = \sum_{s \in S} |s| \log \hat{\sigma}_s^2 + \lambda_0 |S| \log(n). \tag{4}$$

Here, $\lambda_0 > 0$ represents the penalty strength. The optimal segmentation scheme is the one that minimizes the penalized loss function (4), and the corresponding knots are referred to as the selected knots. When λ_0 is very large, this strategy tends to select no knots, whereas when λ_0 approaches 0, it can select as many knots as $\lfloor n^{1-\alpha} \rfloor - 1$. We determine the optimal λ_0 by minimizing the Bayesian information criterion (Schwarz et al., 1978) of the fitted model.

Given a particular λ_0, let $L(\lambda_0)$ be the number of knots finally proposed, and the resulting fitted polynomial spline model with these knots is denoted as $\hat{f}(X, u)$. Then, we have

$$\text{BIC}(\lambda_0) = n \log \left[\frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \hat{f}(X_i, u_i) \Big)^2 \right] + p \left(L(\lambda_0) + D \right) \log(n). \tag{5}$$

The optimal λ_0 is determined by searching over a grid to minimize $\text{BIC}(\lambda_0)$. We refer to this procedure as the global adaptive knots selection strategy, as it assumes that all the coefficient functions have the same set of knots. In Section 3.2, we will discuss how to allow each coefficient function to have its own set of knots.

Here we only use the piece-wise linear model (2) and loss function (4) for knots selection, but we will fit the varying coefficients with B-splines derived from the resulting knots via minimizing (1). In this way, the fitted varying coefficients are smooth functions and the smoothness is determined by the order of the splines. This method is referred to as the *global adaptive spline fitting* throughout the paper.

2.2 Theoretical properties of the selected knots

The proposed method is invariant to the marginal distribution of u. Without loss of generality, we assume that u follows a uniform distribution on the interval $[0, 1]$. Below we introduce Definition 1 and Condition 1. Definition 1 characterizes a turning point as a

local maximum or minimum of $\beta_j(u)$. Under Condition 1, we show in Theorem 1 that the adaptive knots selection approach can almost surely detect all the turning points of $\beta(u)$.

Definition 1. *We call* $0 < t_1 < \ldots < t_T < 1$ $(T < \infty)$ *the turning points of* $\beta(u)$ *for* $u \in (0,1)$*, if, for any* $t_{\tau-1} < u_1 < u_2 < t_\tau < u_3 < u_4 < t_{\tau+1}$ $(\tau = 1, \ldots, T)$*,*

$$[\beta_j(u_1) - \beta_j(u_2)][\beta_j(u_3) - \beta_j(u_4)] < 0,$$

for some index j*, where* $t_0 = 0$ *and* $t_{T+1} = 1$*.*

Condition 1. *Let* $0 < t_1 < \ldots < t_T < 1$ $(T < \infty)$ *be the turning points of* $\beta(u)$ *for* $u \in (0,1)$*; and set* $t_0 = 0$ *and* $t_{T+1} = 1$*. For each turning point* t_τ $(\tau = 1, \ldots, T)$ *and coefficient function* $\beta_j(u)$ $(j = 1, \ldots, p)$*, there exist constants* $\phi > 0$ *and* $\chi \geq 0$ *such that:*

$$\left| \frac{\mathrm{cov}(\beta_j(u), u \mid u \in I)}{\mathrm{var}(u \mid u \in I)} \right| \geq \phi |I|^\chi, \tag{6}$$

where $I = (v, t_\tau]$ *for* $v \in (t_{\tau-1}, t_\tau)$ *or* $I = (t_\tau, v]$ *for* $v \in (t_\tau, t_{\tau+1})$*.*

Inequality (6) implies that the varying coefficient functions $\beta_j(u)$ cannot be too flat within each segment between two adjacent turning points. For example, if k is the smallest integer such that the k-th derivative at t_τ, $D_u^k \beta_j(u)|_{u=t_\tau} \neq 0$, then we can do a local polynomial approximation to $\beta_j(u)$ in I, which leads to that $\chi = k - 1$.

Theorem 1. *Suppose* $y = \beta(u)^\top X + \epsilon$*, where* $u \sim \mathrm{Unif}(0,1)$*,* $\epsilon \sim N(0, \sigma^2)$*,* $X \in \mathbf{R}^p$ *is a bounded vector,* $\beta(u) \in \mathbf{R}^p$ *is a bounded continuous function, and* X*,* u*,* ϵ *are mutually independent. Let* $0 < t_1 < \ldots < t_T < 1$ *be the turning points of* $\beta(u)$ *defined by Definition 1, which satisfies Condition 1. Suppose* $d_1 < \ldots < d_L$ *are the selected knots with* $m_s = n^\alpha$ *(where* $\frac{4\chi+8}{4\chi+9} \leq \alpha < 1$*). For each turning point* t_τ*,* $\min_{l=1}^L |d_l - t_\tau|$ *is its closest distance with the selected knots. Then for any* $0 < \gamma < 1 - \alpha$ *and* $\lambda_0 > 0$*, the closest distance for each turning point* t_τ *will converge to 0, and the number of selected knots is greater than* T*. In other words,*

$$\mathrm{pr}\left(L \geq T,\ \max_{\tau=1}^{T} \min_{l=1}^{L} |d_l - t_\tau| < n^{-\gamma} \right) \to 1,\ n \to \infty.$$

Details of the proof are provided in Section 1 of the supplementary materials. Although the adaptive spline fitting method is motivated by a piece-wise linear model, Theorem 1 demonstrates that, with probability approaching 1, we can accurately detect all the turning points for general varying coefficient functions. Consequently, the selected knots are likely to form a superset of the real turning points, especially when λ_0 is small. We tune λ_0 using BIC (5) to find the optimal set for the given data.

If the underlying varying coefficients are piece-wise linear and not necessarily continuous, Definition 2 provides further characterization of a change point for $\beta(u)$. In this context, a change point refers to a point where the linear function varies. It is important to note that for piece-wise linear function $\beta(u)$, a change point may not be a turning point as it can represent a connecting point of two lines with slopes of the same signs (thus neither a local maximum nor a minimum). In Theorem 2, we demonstrate that the adaptive knots selection method can almost surely discover the change points of $\beta(u)$ without false positive selection.

Definition 2. *We refer to* $0 < c_1 < \ldots < c_T < 1$ $(T < \infty)$ *as the change points of* $\beta(u)$ *for* $u \in (0,1)$ *if the coefficient* $\beta(u) = (\beta_1(u), \ldots, \beta_p(u))^\top$ *can be almost surely defined as*

$$\beta(u) = a_\tau + ub_\tau, \ c_{\tau-1} < u \leq c_\tau, \ \tau = 1, \ldots, T+1,$$

where $c_0 = 0$, $c_{T+1} = 1$ *and* $a_\tau, b_\tau \in \mathbf{R}^p$. *Moreover, for* $\tau = 1, \ldots, T$, *we have*

$$(a_\tau - a_{\tau+1})^\top (a_\tau - a_{\tau+1}) + (b_\tau - b_{\tau+1})^\top (b_\tau - b_{\tau+1}) > 0.$$

Theorem 2. *Suppose* (X, y, u) *follows the same assumptions as in Theorem 1, except that* $\beta(u)$ *is a piece-wise linear function of* u *and not necessarily continuous. Let* $0 < c_1 < \ldots < c_T < 1$ *be the change points defined in Definition 2, and let* $d_1 < \ldots < d_L$ *be the selected knots with* $m_s = n^\alpha$ $(\frac{8}{9} \leq \alpha < 1)$. *Similarly, for each change point* c_τ, $\min_{l=1}^{L} |d_l - c_\tau|$ *is its closest distance with the selected knots. Then for* $0 < \gamma < 1 - \alpha$ *and* $\lambda_0 > 120p(T+2)^2$, *the closest distance for each change point* c_τ *will converge to 0, and the number of selected knots will be exactly* T. *In other words,*

$$\Pr\left(L = T, \ \max_{\tau=1}^{T} |d_\tau - c_\tau| < n^{-\gamma}\right) \to 1, \ n \to \infty.$$

Details of the proof are provided in Section 2.1 of the supplementary materials. The theorem demonstrates that if the varying coefficient function is piece-wise linear, the method can discover all the change points with almost 100% accuracy. The corollary below is a special case of Theorem 2 when the true coefficient $\beta(u)$ is 0.

Corollary 1. *Suppose x is a bounded univariate independent variable, $u \sim \text{Unif}(0,1)$, the response $y \sim N(0,\sigma^2)$, and x, u, y are mutually independent. Then, for $\lambda_0 > 120$, with probability greater than $1 - 3n^{8+2(1-\alpha)-\lambda_0/12}$, the adaptive knots selection method will not select any knots for $m_s = n^\alpha$ ($\frac{8}{9} \leq \alpha < 1$).*

2.3 Theoretical properties of the residual variance

In Theorem 3, we demonstrate that when the true underlying varying coefficients are continuous and bounded, the residual variance of the varying coefficient model fitted by the optimal segmentation scheme converges to the true error variance. It is important to note that the piece-wise linear coefficients are fitted within each segment individually and may not be continuous at the knots. In Theorem 4, we will show that when the true coefficients are continuous piece-wise linear, we can refit the coefficient with an order-2 polynomial spline, and the residual variance will still converge to the true error variance.

Theorem 3. *Suppose $y = \beta(u)^\top X + \epsilon$, where $u \sim \text{Unif}(0,1)$, $\epsilon \sim N(0,\sigma^2)$, X is bounded, X, u, ϵ are mutually independent. Additionally, $\beta(u)$ is a bounded continuous function differentiable almost everywhere with a bounded first derivative. For $m_s = n^\alpha$ ($\frac{2}{3} \leq \alpha < 1$) and a given $\lambda_0 > 0$, let S^* be the optimal segmentation scheme obtained by minimizing (3), and $\hat{\beta}(u)$ be the fitted piece-wise linear coefficient. Then, the corresponding residual variance $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(y_i - \hat{\beta}(u_i)^\top X_i\right)^2$ converges to σ^2 in probability.*

Details of the proof are provided in Section 3 of the supplementary materials. When the true underlying coefficients $\beta(u)$ are continuous piece-wise linear functions, we can generalize the conclusion of Theorem 3 by replacing the fitted coefficients with the polynomial spline of order 2 (a continuous piece-wise linear function), and the residual variance will still converge to the ground truth data variance.

Theorem 4. *Suppose* (X, y, u) *follows the same assumption as in Theorem 3, except that* $\beta(u)$ *is a continuous piece-wise linear function of* u*. For* $m_s = n^\alpha$ $(\frac{8}{9} \leq \alpha < 1)$*, let* $\tilde{\beta}(u)$ *be the fitted piece-wise linear coefficients from (1) with knots from the optimal segmentation scheme. Then the corresponding residual variance* $\tilde{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\left(y_i - \tilde{\beta}(u_i)^\top X_i\right)^2$ *converges to* σ^2 *in probability when* $\lambda_0 > 120p(T+2)^2$*.*

The proof is available in Section 4 of the supplementary materials. This theorem guarantees that replacing the piece-wise linear coefficients with splines does not harm the performance. Moreover, it makes more sense to provide an estimated coefficient with the same continuity as the underlying coefficients, and that's why we propose to refit the coefficients with polynomial splines.

3 Efficient computation in low and high dimensions

3.1 Dynamic programming for adaptive knots selection

The brute force algorithm to compute the optimal knots has a computational complexity of $O(2^n)$ and is impractical for large n. As presented in Algorithm 1, a dynamic programming algorithm with a computational complexity of order $O(n^2)$ can be implemented to find the optimal solution exactly, which is a significant improvement in efficiency. If we further assume that the knots can only be chosen from a predetermined set $\mathcal{M}$, such as $\mathcal{M} = \{u_{(m)} : \ m = \lceil j\sqrt{n}\rceil, j = 1, \ldots, \lfloor\sqrt{n}\rfloor - 1\}$, the computational complexity can be further reduced to $O(|\mathcal{M}|^2)$. This reduction in complexity is particularly useful when dealing with large datasets. It's worth noting that the algorithm presented in Section 2.4 of Wang et al. (2017) is a special case with $x = u$.

When the algorithm is run with a grid of λ_0, we repeat Steps 2 and 3 for all the λ_0's, and return the final model with the minimum BIC(λ_0).

3.2 Predictor specific adaptive knots selection

The global adaptive knots selection method described in Section 2.1 assumes a common set of knot locations for all coefficient functions, similar to most existing methods for

Algorithm 1 Dynamic Programming for Optimal Knots Selection

1. Data preparation: Arrange the data $(X_i, y_i, u_i)_{i=1}^n$ in ascending order of u_i, without loss of generality, such that $u_1 < u_2 < \cdots < u_n$.

2. Forward recursion to find the minimum loss:

   (a) Set $m_s = \lceil n^\alpha \rceil$ be the smallest segment size and define $\lambda = \lambda_0 \log(n)$.

   (b) Initialize a $2 \times n$ array $\{(\text{Loss}_i, \text{Prev}_i)^\top\}_{i=1}^n$ with $\text{Loss}_0 = \text{Prev}_0 = 0$.

   (c) For i ranging from m_s to n, performing the following recursive updates:

$$\text{Loss}_i = \min_{i' \in I_i}(\text{Loss}_{i'-1} + \ell_{i':i} + \lambda), \quad \text{Prev}_i = \arg\min_{i' \in I_i}(\text{Loss}_{i'-1} + \ell_{i':i} + \lambda),$$

   where $I_i = \{1\} \cup \{m_s + 1, \ldots, i - m_s + 1\}$ and $\ell_{i':i} = (i - i' + 1) \log \hat{\sigma}^2_{i':i}$ where $\hat{\sigma}^2_{i':i}$ is the residual variance of regressing y_{i_0} on $(X_{i_0}, u_{i_0} X_{i_0})$ $(i_0 = i', \ldots, i)$.

3. Backward tracing to find knots: Let $\text{P} = \text{Prev}_n$. If $\text{P} = 1$, no knot is needed, and the process ends; otherwise, add $0.5(u_{\text{P}-1} + u_{\text{P}})$ as a new knot and update $\text{P} = \text{Prev}_{\text{P}}$; repeat the process until $\text{P} = 1$.

polynomial spline fitting. However, in some cases, different coefficients may have varying degrees of smoothness relative to u, making it preferable to have a different set of knots for each predictor. To address this, we propose a predictor-specific adaptive spline fitting algorithm on top of the global knot selection. Suppose the fitted model for the global adaptive spline fitting is $\hat{f}(X, u) = \sum_{j=1}^p \hat{\beta}_j(u) x_j$, where $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbf{R}^{n \times p}$ represents the predictor matrix and $\mathbf{x}_j$ is its jth column. Additionally, let $\mathbf{y}$ be the response vector, and $\mathbf{u}$ be the conditioner vector. As shown in Algorithm 2, we update the knots for each coefficient function of predictor $\mathbf{x}_j$ by using the same knots selection procedure with the residual vector:

$$\mathbf{r}_{-j} = \mathbf{y} - \sum_{\ell \neq j} \hat{\beta}_\ell(\mathbf{u}) \circ \mathbf{x}_\ell. \tag{7}$$

as the response and $\mathbf{x}_j$ as the only predictor, where $\hat{\beta}_\ell(\mathbf{u}) \circ \mathbf{x}_\ell$ represents the element-wise product. We then check if the updated knots lead to an improved BIC value for the model

and repeat this step until no further improvement is achieved.

Algorithm 2 Predictor-specific Knots Selection

1. Run Algorithm 1 for the full data $(\mathbf{X}, \mathbf{y}, \mathbf{u})$ to obtain the fitted model $\hat{f}(X, u)$ and compute its BIC with (5).

2. Update knots: For j ranging from 1 to p,

   (a) Compute residual $\mathbf{r}_{-j}$ using Equation (7), and run Algorithm 1 with the data $(\mathbf{x}_j, \mathbf{r}_{-j}, \mathbf{u})$ to obtain the new fitted model $\hat{f}^j(X, u)$. Then, compute its BIC_j.

   (b) If $\min \text{BIC}_j < \text{BIC}$ and $j^* = \arg\min \text{BIC}_j$, update the current model by setting $\hat{f}(X, u) = \hat{f}^{j^*}(X, u)$; otherwise do nothing.

3. Repeat Step 2 until no BIC improvements, return $\hat{f}(X, u)$.

In Step 2 of the algorithm, when computing the BIC for the updated model, each predictor may have a different number of knots, so the term $p\,(L(\lambda_0) + D)\log(n)$ in (5) is replaced by $\left(\sum_{j=1}^{p} L_j(\lambda_0) + pD\right)\log(n)$, where $L_j(\lambda_0)$ is the number of knots for predictor $\mathbf{x}_j$. An alternative approach to model the heterogeneity among the coefficients is to replace the initial model in Step 1 with $\hat{f}(X, u) = 0$, and continue with the algorithm. However, starting from the global model is preferred because fitting to the residual instead of the original response minimizes the mean squared error (MSE) more efficiently. Section 4.1 demonstrates that the predictor-specific knots can further reduce the MSE for the fitted coefficients compared with the global knot selection approach.

3.3 Knots selection in sparse high-dimensional problems

When dealing with a large number of predictors, and when only a small subset of predictors have non-zero varying coefficients, we perform variable selection for all the predictors. For each predictor, we first conduct marginal knots selection and fitting by running Algorithm 1 on data $(\mathbf{x}_j, \mathbf{y}, \mathbf{u})$ and obtaining the B-spline functions $\{B_{j,k}(u)\}_{k=1}^{L_j+D}$, where L_j is the number of knots and D is the order of the B-splines. Next, we apply the variable selection

method proposed by Wei et al. (2011), which is a generalization of the group and adaptive LASSO methods (Yuan and Lin, 2006; Zou, 2006). The group LASSO tends to over-select variables, and the adaptive group LASSO was introduced as a remedy (Wei et al., 2011). In their original algorithm, the knots for each predictor are chosen as equidistant quantiles and are not predictor-specific. However, our Algorithm 3 allows for more flexible and optimal knot selections, improving the variable selection performance in high-dimensional settings.

Algorithm 3 Variable Selection for Fitting Varying Coefficients

1. Select knots for each predictor: Run Algorithm 1 for each predictor $\mathbf{x}_j$, and obtain the B-splines coefficient functions $\left\{\{B_{j,k}(u)\}_{k=1}^{L_j+D}\right\}_{j=1}^{p}$, where L_j is the number of knots for $\mathbf{x}_j$ and D is the order of splines.

2. Run group LASSO for variable selection under the following loss function

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{i,j}\sum_{k=1}^{L_j+D} h_{j,k}B_{j,k}(u_i)\right)^2 + \lambda_1\sum_{j=1}^{p}(h_j^\top R_j h_j)^{1/2},\ \lambda_1 > 0,$$

   where $h_j = (h_{j,1},\ldots,h_{j,L_j+D})^\top$ and R_j is the kernel matrix whose (k_1,k_2) element is $E[B_{j,k_1}(u)B_{j,k_2}(u)]$. Denote the fitted coefficients as $\tilde{h}_{j,k}$.

3. Run adaptive group LASSO for the updated loss function

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{i,j}\sum_{k=1}^{L_j+D} h_{j,k}B_{j,k}(u_i)\right)^2 + \lambda_2\sum_{j=1}^{p}\omega_j(h_j^\top R_j h_j)^{1/2},\ \lambda_2 > 0,$$

   with weight $\omega_k = \infty$ if $\tilde{h}_j^\top R_j \tilde{h}_j = 0$; and $\omega_k = (\tilde{h}_j^\top R_j \tilde{h}_j)^{-1/2}$ otherwise. The fitted coefficients are $\hat{h}_{j,k}$ and the selected variables are those with $\hat{h}_j^\top R_j \hat{h}_j \neq 0$.

Corollary 1 guarantees that Step 1 of Algorithm 3 will likely select zero knots for predictors that are independent of the response variable y. This is beneficial because it helps avoid overfitting and reduces unnecessary computational burden for predictors that do not have varying coefficients. In Steps 2 and 3, the tuning parameter λ_1 and λ_2 are chosen by minimizing the Bayesian Information Criterion (BIC) BIC (5) for the fitted model. In this process, the degrees of freedom are computed with only the selected predictors, ensuring a

fair comparison and providing a more accurate measure of model complexity. Simulation studies in Section 4.2 demonstrate that, with the adaptive knots selection in Step 1, Algorithm 3 shows superior performance in selecting the correct predictors with a reasonable number of samples compared to existing methods. This highlights the effectiveness of our proposed approach in high-dimensional settings with varying coefficient models.

4 Empirical studies

4.1 Simulation study for adaptive spline fitting

The simulation example is adapted from the work of Tang and Cheng (2012). In this example, we compare the performance of the global and predictor-specific adaptive spline fitting approaches, along with the equidistant spline fitting approach and the kernel method implemented in the `tvReg` package by Casas and Fernandez-Casal (2019). The simulation model has a longitudinal structure, commonly encountered in biomedical applications. Each simulation involves n individuals, and each individual has a scheduled time set of $0, 1, \ldots, 19$ to generate observations. However, a scheduled time can be skipped with a probability of 0.6, leading to no observations being generated at that time point. For non-skipped scheduled times, the real observed time is obtained by adding a random disturbance from a uniform distribution $\text{Unif}(0, 1)$ to the scheduled time.

The time-dependent predictors are denoted as $X(u) = (x_1(u), x_2(u), x_3(u), x_4(u))^\top$, where:

$$x_1(u) = 1, \quad x_2(u) \sim \text{Bern}(0.6), \quad x_3(u) \sim \text{Unif}(0.1u, 2 + 0.1u),$$
$$x_4(u) \mid x_3(u) \sim N\left(0, \frac{1 + x_3(u)}{2 + x_3(u)}\right).$$

The response is $y_i(u_{i,q}) = \sum_{j=1}^{4} \beta_j(u_{i,q}) x_{i,j}(u_{i,q}) + \epsilon_i(u_{i,q})$ for individual i at time $u_{i,q}$, with

$$\beta_1(u) = 1 + 3.5\sin(u - 3), \qquad \beta_2(u) = 2 - 5\cos(0.75u - 0.25),$$
$$\beta_3(u) = 4 - 0.04(u - 12)^2, \qquad \beta_4(u) = 1 + 0.125u + 4.6\,(1 - 0.1u)^3.$$

The random error $\epsilon_i(u_{i,q})$'s are independent of the predictors, independent between different individuals, and positively correlated within the same individual. This correlation structure

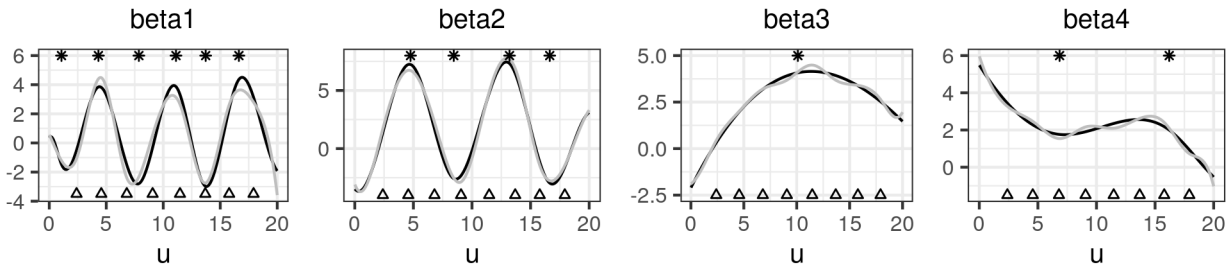

Figure 1: The true coefficients and fitted coefficients using equidistant and predictor-specific knots. Black line: true coefficients; triangle: equidistant knots; grey lines: fitted coefficients with equidistant knots; stars: predictor-specific knots; dotted lines: fitted coefficients with predictor-specific knots.

is common in longitudinal data, where observations within the same individual tend to be more similar due to repeated measurements over time. More precisely,

$$\epsilon_i(u_{i,q}) = v_i(u_{i,q}) + e_i(u_{i,q}),$$

where $e_i(u_{i,q}) \overset{i.i.d.}{\sim} N(0,4)$, and $v_i(u_{i,q}) \sim N(0,4)$ with correlation structure

$$\text{cor}\left(v_{i_1}(u_{i_1,q_1}), v_{i_2}(u_{i_2,q_2})\right) = I(i_1 = i_2)\exp(-|u_{i_1,q_1} - u_{i_2,q_2}|).$$

Figure 1 displays the true coefficients and the fitted coefficients by the equidistant and predictor-specific spline fitting methods for an example with $n = 200$, along with the selected knots. It is worth noting that the number of knots for the equidistant fitting approach is also chosen by minimizing the model's Bayesian Information Criterion (BIC) (5). The figure illustrates that the fitted coefficients obtained by the predictor-specific method are smoother than those obtained by the equidistant method, particularly for the less volatile coefficients $\beta_3(u)$ and $\beta_4(u)$. This difference is attributed to the predictor-specific method's utilization of only 1 or 2 knots for these two coefficients, which better captures their underlying smoothness, while the equidistant method employs 8 knots.

We compare the four methods based on the mean squared errors (MSEs) of their estimated coefficients, which are calculated as follows:

$$\text{MSE}_j = \frac{1}{N}\sum_{i=1}^{n}\sum_{q=1}^{n_i}\frac{\left(\hat{\beta}_j(u_{i,q}) - \beta_j(u_{i,q})\right)^2}{\text{range}(\beta_j)^2}, \tag{8}$$

	$\text{MSE}_1 \times 1e2$	$\text{MSE}_2 \times 1e2$	$\text{MSE}_3 \times 1e2$	$\text{MSE}_4 \times 1e2$
equidistant	2.53 (1.14)	0.26 (0.10)	0.64 (0.29)	0.09 (0.04)
global	2.24 (1.12)	0.32 (0.12)	0.55 (0.26)	0.08 (0.04)
predictor-specific	**1.30** (0.83)	**0.23** (0.10)	**0.28** (0.21)	**0.04** (0.03)
kernel	2.80 (0.92)	0.38 (0.11)	0.81 (0.26)	0.14 (0.04)

Table 1: MSE_j for the equidistant, global, and predictor-specific adaptive spline fitting methods, compared with the kernel method.

where $\beta_j(u_{i,q})$ and $\hat{\beta}_j(u_{i,q})$ are the true and estimated coefficients at time point $u_{i,q}$ for individual i, respectively. n_i represents the total number of observations for individual i, and $N = \sum_{i=1}^{n} n_i$. Additionally, we define the range of the true coefficients as

$$\text{range}(\beta_j) = \max_{0<u<20} \beta_j(u) - \min_{0<u<20} \beta_j(u).$$

The simulation is conducted with $n = 200$ for 1000 repetitions. For the adaptive spline methods, we consider only the knots from $\frac{m}{\lceil\sqrt{N}\rceil}$ quantiles of $u_{i,q}$, where m ranges from 1 to $\lfloor\sqrt{N}\rfloor - 1$. Table 1 presents the average MSE_j for the proposed global and predictor-specific methods, along with the equidistant spline fitting and the kernel methods. The predictor-specific method shows the smallest MSE_j for all four coefficients, demonstrating its superior performance compared to the other three methods. Additionally, the equidistant method selects on average 6.9 knots, the global adaptive method selects an average of 6.1 global knots for all predictors, while the predictor-specific method selects fewer knots on average: 5.8 knots for $x_1(u)$, 4.2 for $x_2(u)$, 1.5 for $x_3(u)$ and 0.8 for $x_4(u)$. This outcome aligns with expectations, as $\beta_3(u)$ and $\beta_4(u)$ are less volatile than $\beta_1(u)$ and $\beta_2(u)$.

4.2 Simulation study for variable selection

We use the simulation example from Wei et al. (2011) to compare the performance of our method with the one using adaptive group LASSO and equidistant knots. Similar to the previous subsection, there are n individuals, and each has a scheduled time set $\{0, 1, \ldots, 29\}$ to generate observations and a skipping probability of 0.6. For each non-skipped scheduled

time, the observed time is the scheduled time plus a random disturbance generated from Unif$(0,1)$. We construct $p = 500$ time-dependent predictors as follows:

$$\begin{aligned} x_1(u) &\sim \text{Unif}\,(0.05 + 0.1u, 2.05 + 0.1u) \\ x_j(u) \mid x_1(u) &\sim N\left(0, \frac{1 + x_1(u)}{2 + x_1(u)}\right),\ j = 2, \ldots, 5, \\ x_6(u) &\sim N\left(3\exp\{(u + 0.5)/30\}, 1\right), \\ x_j(u) &\overset{i.i.d.}{\sim} N(0, 4),\ j = 7, \ldots, 500. \end{aligned}$$

The same individual's predictors $x_j(u)$ $(j = 7, \ldots, 500)$ are correlated with

$$\text{cor}\left(x_j(u_1), x_j(u_2)\right) = \exp(-|u_1 - u_2|).$$

The response for individual i at observed time $u_{i,q}$ is

$$y_i(u_{i,q}) = \sum_{j=1}^{6} \beta_j(u_{i,q}) x_{i,j}(u_{i,q}) + \epsilon_i(u_{i,q}).$$

The time-varying coefficients $\beta_j(u)$ $(j = 1, \ldots, 6)$ are

$$\begin{aligned} \beta_1(u) &= 15 + 20\sin\{\pi(u + 0.5)/15\}, \quad \beta_2(u) = 15 + 20\cos\{\pi(u + 0.5)/15\}, \\ \beta_3(u) &= 2 - 3\sin\{\pi(u - 24.5)/15\}, \quad \beta_4(u) = 2 - 3\cos\{\pi(u - 24.5)/15\}, \\ \beta_5(u) &= 6 - 0.2(u + 0.5)^2, \quad \beta_6(u) = -4 + 5 * 10^{-4}(19.5 - u)^3. \end{aligned}$$

The random error $\epsilon_i(u_{i,q})$ is independent of the predictors and follows the same distribution as that in Section 4.1.

We simulate cases with $n = 50, 100, 200$ and replicate each set 200 times. Three metrics are considered: the average number of selected variables, the percentage of cases when there is no false negative, and the percentage of cases when there is no false positive or negative. A comparison of our method with the variable selection method using equidistant knots (Wei et al., 2011) is summarized in Table 2, demonstrating that our method significantly outperforms the method of Wei et al. (2011) without predictor-specific knots selection.

equidistant knots

	# selected variables	% no false negative	% no false positive or negative
$n = 50$	7.04	72	68
$n = 100$	6.21	87	84
$n = 200$	6.13	99	93

adaptive selected knots

	# selected variables	% no false negative	% no false positive or negative
$n = 50$	5.96	96.50	96.50
$n = 100$	6.00	100	100
$n = 200$	6.00	100	100

Table 2: Variable selection performance for adaptive group LASSO with and without predictor-dependent knots selection.

5 Applications

5.1 Environmental factors and COVID-19

The data set we investigated contains daily measurements of meteorological data and air quality data in 7 counties of the state of New York between March 1, 2020, and September 30, 2021. The meteorological data were obtained from the National Oceanic and Atmospheric Administration Regional Climate Centers, Northeast Regional Climate Center at Cornell University: http://www.nrcc.cornell.edu. The daily data are based on the average of the hourly measurements of several stations in each county and include records of five meteorological components: temperature (in Fahrenheit), dew point (in Fahrenheit), wind speed (in miles per hour), precipitation (in inches), and humidity (in percentage). The air quality data were obtained from the Environmental Protection Agency: https://www.epa.gov. The data contain daily records of two major air quality components: the fine particles with an aerodynamic diameter of 2.5μm or less, denoted as $PM_{2.5}$ (in $\mu g/m^3$), and ozone (also measured in $\mu g/m^3$).

The main objective of the study is to understand the association between the meteoro-

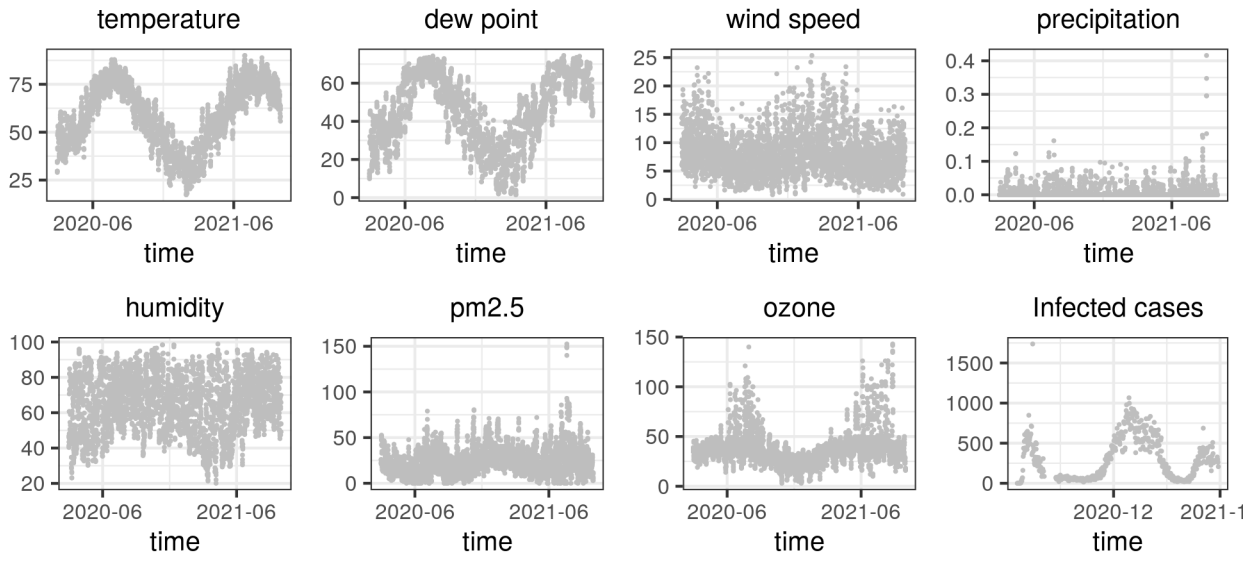

Figure 2: Environmental measurements and COVID-19 infected cases in New York County, NY.

logical measurements, together with pollutant levels, and the number of COVID-19 infected cases. COVID-19 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2. The study aims to examine whether this association varies over time. The daily infected records were retrieved from the official website of the Department of Health, New York State: https://data.ny.gov. To remove the variation of recorded cases between weekdays and weekends, the study considers the weekly average infected cases, which are calculated as the average between each day and the following 6 days. During the analysis, it is also observed that the temperature factor and dew point factor were highly correlated. Consequently, the dew point factor was removed when fitting the model. Figure 2 shows scatter plots of daily infected cases in New York County and the 7 environmental components over time, providing an initial visualization of the data.

To address the issue of a right-skewed distribution, we take the logarithmic transformation of the weekly averaged infected cases, denoted as y, effectively removing the right tail. We then proceed to fit a varying coefficient model with the following predictors: $x_1 = 1$ as intercept, x_2 as temperature, x_3 as wind speed, x_4 as precipitation, x_5 as humidity, x_6 as $PM_{2.5}$ and x_7 as ozone. The time variable u serves as the conditioner for our model. To

ensure comparability between each $\beta_j(u)$, all predictors except the constant are normalized before fitting. Furthermore, we apply a data clipping procedure to limit values within the range of -3 to 3, efficiently removing outliers. The varying coefficient model can be expressed as follows:

$$y_i(u_{i,q}) = \sum_{j=1}^{7} \beta_j(u_{i,q}) x_{i,j}(u_{i,q}) + \epsilon(u_{i,q}), \tag{9}$$

where $u_{i,q}$ is the qth recorded time for the ith county, and $y_i(u_{i,q})$ and $x_{i,j}(u_{i,q})$ are the corresponding records for county i at time $u_{i,q}$. The term $\epsilon(u_{i,q}) \overset{i.i.d.}{\sim} N(0, \sigma^2)$ represents the error term in the model, with each error term independently and identically distributed with mean 0 and variance σ^2.

We apply both the equidistant and the proposed predictor-specific adaptive spline fitting methods to fit the data. The resulting fitted coefficient functions $\beta_j(u)$ for each predictor are shown in Figure 3 for both methods. The figures indicate that there is a strong time effect on each coefficient function. For instance, the intercept exhibits several peaks corresponding to the initial outbreak and the delta variant outbreak. Additionally, rapid changes in coefficients are observed around March 2020, likely due to the early stages of the outbreak when the number of infected cases was underestimated due to fewer tests being conducted.

Moreover, the coefficient curves reveal that the most influential predictor is temperature. For most of the period, the coefficient is negative, indicating a negative association between high temperature and virus transmission. This observation is consistent with the findings in the study by Notari (2021), which suggests that COVID-19 spread is slower at high temperatures. The analysis also demonstrates the time-varying nature of the coefficient.

Furthermore, the fitted coefficients obtained using the predictor-specific knots are less volatile compared to those obtained using the equidistant knots, particularly for the predictors of temperature, wind speed, precipitation, humidity, and $PM_{2.5}$. This suggests that the predictor-specific knots provide a more stable and accurate representation of the coefficient functions over time.

The rolling window approach is used to evaluate the predictability of the proposed method. We use a training size of at least 1 year and a rolling window of 1 week. For each date u after March 1, 2021, we fit two models, one with equidistant knots and another

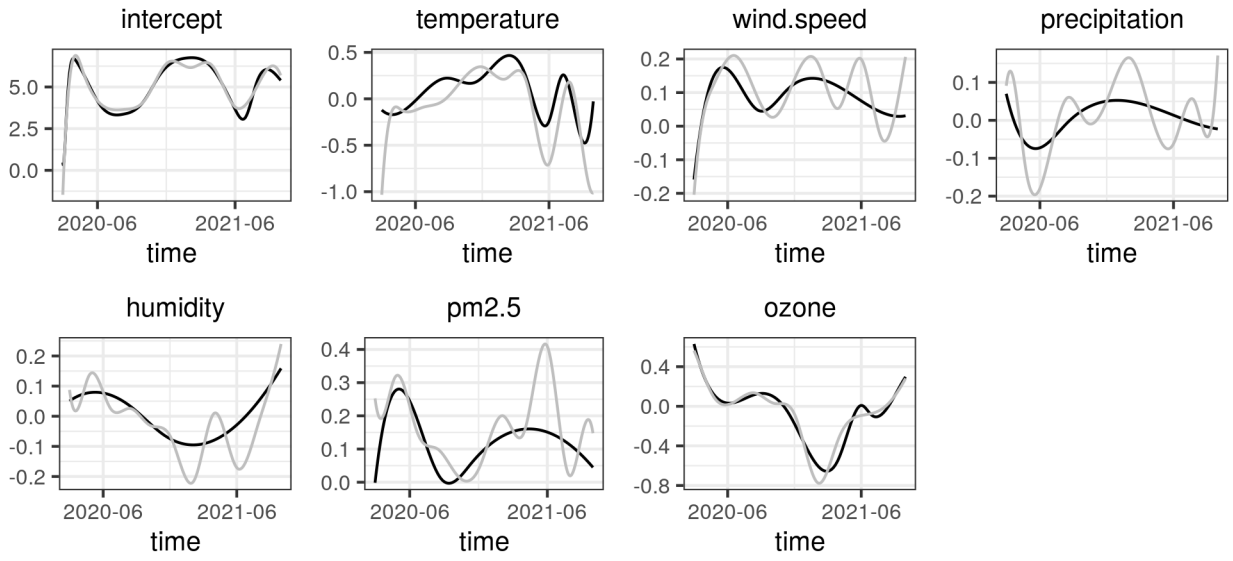

Figure 3: Black lines: fitted coefficients with predictor-specific knots; grey lines: fitted coefficients with equidistant knots.

with predictor-specific knots, using data $\{X_i(u_{i,q}), y_i(u_{i,q})\}$ for $u_{i,q} < u$, and then predict $y_i(u_{i,q})$ for $u \leq u_{i,q} < u + 7$. The root mean squared error (RMSE) for the model with equidistant knots is 0.829, whereas the RMSE for the model with predictor-specific knots fitting is 0.716. This indicates that the predictor-specific knots provide a more accurate prediction of the weekly average infected cases compared to the equidistant knots method.

Environmental factors may not have an immediate effect on the number of recorded COVID-19 cases due to the incubation period of 2-14 days before the onset of symptoms and the additional time required for test results to be available (2-7 days). To study whether there are lagging effects between the predictor and response variables, we fit a varying coefficient model with predictor-specific knots for each time lag ν. In this approach, we use data $y_i(u_{i,q} + \nu)$ and $X_i(u_{i,q})$, to fit model (9) similarly as in Fan and Zhang (1999). Figure 4 shows the residual root mean squared error (RMSE) for each time lag, revealing a minimum at the 13-day lag. This indicates that the predictors at day u are most predictive for infected numbers between day $u + 13$ and $u + 19$, which aligns with the incubation period and the time it takes to receive test results. In other words, the environmental factors have the most significant impact on the number of recorded COVID-19 cases about

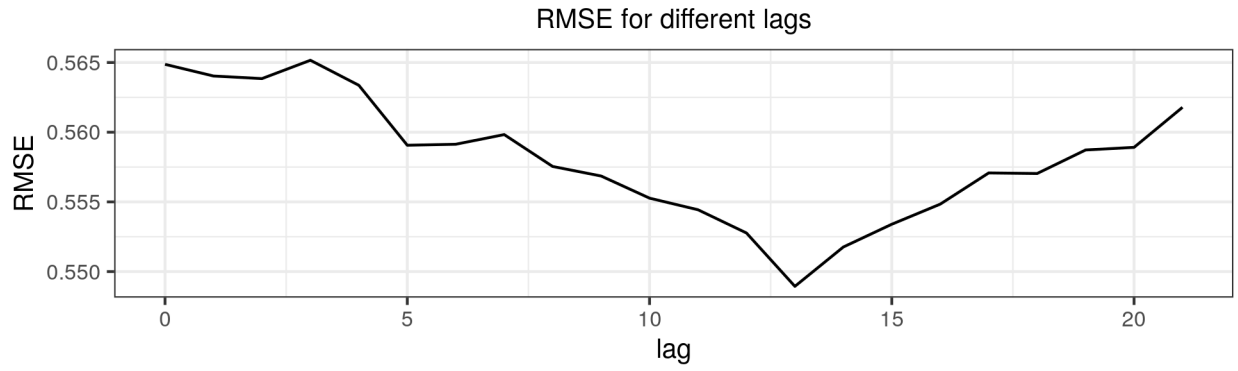

Figure 4: Root mean squared error for the logarithm of weekly averaged infected cases, with lag $\nu = 0, \ldots, 21$ days.

12 days after their measurements. This finding highlights the importance of considering time lags when studying the association between environmental factors and the number of COVID-19 cases.

5.2 Boston housing data

We use the Boston housing price data from Harrison and Rubinfeld (1978), consisting of $n = 506$ observations for the census districts of the Boston metropolitan area. The data is available in the R-package `lmbench`. Following Wang and Xia (2009) and Hu and Xia (2012), the response variable is `medv`, representing the medium value of owner-occupied homes in 1,000 USD. The conditioner is `lstat` , defined as a linear combination of the proportion of adults with high school education or above and the proportion of male workers classified as laborers. We use the following predictors: `int` (the intercept), `crim` (per capita crime rate by town), `rm` (average number of rooms per dwelling), `ptratio` (pupil-teacher ratio by town), `nox` (nitric oxides concentration parts per 10 million), `tax` (full-value property-tax rate per 10,000 USD), and `age` (proportion of owner-occupied units built prior to 1940).

To prepare the data for analysis, we transform the conditioner `lstat` so that its marginal distribution is Unif$(0, 1)$. We also apply a logarithm transformation to the response variable`medv`. Additionally, we standardize the predictors (except for the intercept) such that

the marginal distribution is standard normal. Since some of the predictors are highly correlated with the conditioner, we perform separate regressions of each predictor against the transformed `lstat`, and use the normalized residuals in the subsequent analysis. This process ensures that the predictors are not overly influenced by the correlation with the conditioner, allowing us to study their individual effects on the response variable.

We apply the predictor-specific varying coefficient linear model to predict the response using the residualized predictors, with the transformed `lstat` as the conditioner. Figure 5 displays the fitted coefficients as a function of the conditioner, along with the 95% confidence intervals represented by dotted lines. The dashed lines represent the x-axis. The confidence interval is computed by conditioning the selected knots for each predictor.

The results reveal the conditioner-varying effects of most predictors. The intercept exhibits significant variation with `lstat`, indicating that the housing price is negatively and almost linearly impacted by `lstat`. The coefficient for `rm` is generally positive, indicating that houses with more rooms tend to have higher prices. However, the impact of `rm` becomes less significant as `lstat` increases, suggesting that the number of rooms may not be as crucial a factor in areas with high `lstat` . Variable `crim` is highly correlated with `lstat`. After removing the influence of `lstat`, the residualized `crim` shows interesting fluctuations, ranging from insignificant to positive and then to negative effects. This behavior might be attributed to a confounding effect with other unutilized variables, such as the location's convenience and attractiveness for tourists. Overall, the predictor-specific varying coefficient linear model provides valuable insights into the conditioner-varying effects of the predictors on housing prices, considering the complex relationships between the variables. The confidence intervals obtained by conditioning on the selected knots offer a comprehensive understanding of the varying effects for different conditions of `lstat` .

We further assess the predictive performance of both the simple linear model and the varying coefficient model using 10-fold cross-validation. For the linear model, we incorporate all the predictors employed in the varying coefficient model along with the conditioner `lstat`. After transforming the MSE back to the original scale, we obtain a value of 23.52 for the simple linear model, while the varying coefficient model yields a lower MSE of 20.51.

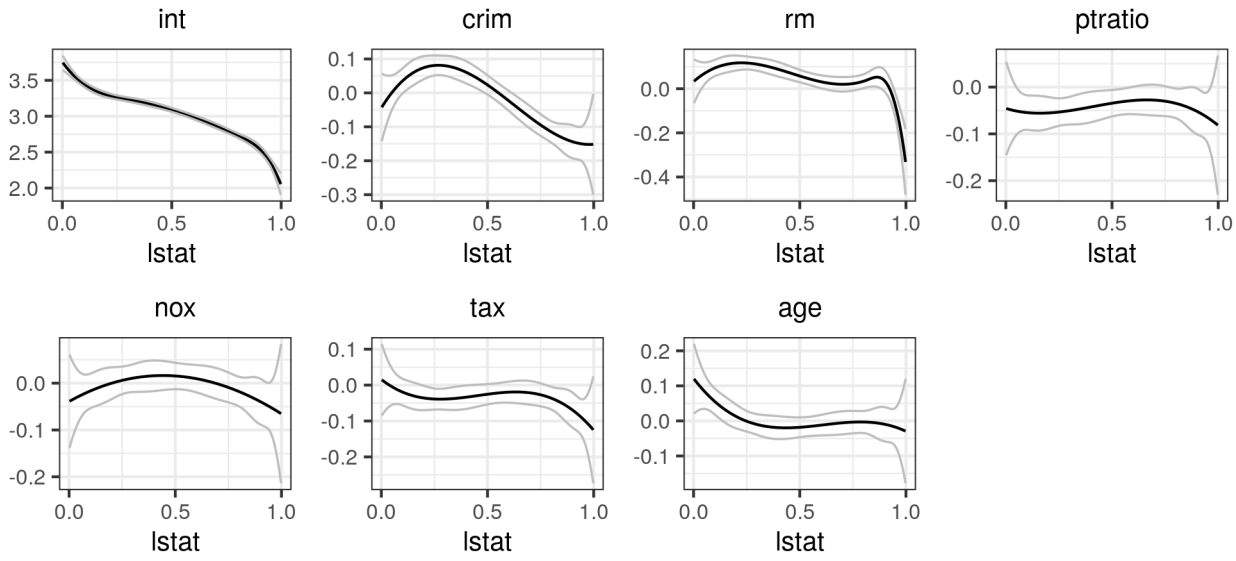

Figure 5: Fitted coefficient for Boston Housing data, with conditioner as the transformed `lstat`.

This confirms that the varying coefficient model is more suitable and provides better predictions for this dataset compared to the simple linear model.

6 Discussion

In this paper, we have introduced three algorithms for fitting varying coefficient models with adaptive polynomial splines and conducting variable selection in high dimensions. The first algorithm is a global approach that selects knots using a recursive method, assuming the same set of knots for all the coefficient functions. On the other hand, the second algorithm is a predictor-specific approach, allowing each predictor to have its own set of knots. This is achieved by iteratively applying the global knot selection algorithm to each predictor. Finally, the third algorithm is designed for variable selection and utilizes an adaptive group LASSO method to select important predictors, taking advantage of the predictor-specific knots selection approach. Together, these algorithms provide flexible and efficient methods for fitting varying coefficient models with adaptive polynomial splines and performing variable selection, making them suitable for high-dimensional datasets.

The coefficients modeled by polynomial splines with a finite number of non-regularly positioned knots offer increased flexibility and interpretability compared to standard splines with equidistant knot placements. Simulation studies demonstrate that both the global and predictor-specific algorithms outperform commonly used kernel methods and the equidistant spline fitting method in terms of mean squared errors (MSEs), with the predictor-specific algorithm achieving the best performance. To efficiently find the optimal knot locations, we have introduced a fast dynamic programming algorithm with a computational complexity of no more than $O(n^2)$, which can be further reduced to $O(n)$ if we allow the resolution of the knot locations to be $O(\sqrt{n})$. Overall, the proposed algorithms provide effective and practical solutions for fitting varying coefficient models with adaptive polynomial splines and conducting variable selection, offering improved flexibility and accuracy in modeling complex relationships between predictors and responses.

Throughout the article, we assume that the conditioner variable u is univariate. However, the proposed predictor-specific spline approach can be easily extended to cases where each coefficient $\beta_j(u)$ has its own univariate conditioner variable u. Nonetheless, it remains a challenging task to generalize the proposed method to multi-dimensional conditioners and to model correlated errors.

For researchers and practitioners interested in applying the proposed algorithms, we have developed an R package that implements these methods. The package is available at https://github.com/wangxf0106/vcmasf and contains comprehensive instructions on how to use the software effectively. It provides a user-friendly interface for fitting varying coefficient models with adaptive polynomial splines and conducting variable selection, making it accessible to a wide range of users in diverse fields of research and application.

Acknowledgement

We would like to express our sincere gratitude to the National Oceanic and Atmospheric Administration Regional Climate Centers, especially the Northeast Regional Climate Center at Cornell University, for their generous cooperation and sharing of the meteorological data used in this study. We are also thankful to the Environmental Protection Agency and

the Department of Health, New York State, for providing access to the air quality data and daily infected records, which were invaluable for our research.

We acknowledge the support received from the NSF grant DMS-2015411 and DMS-1903139, which played a crucial role in funding and supporting this research project. Their financial assistance has greatly contributed to the successful completion of this work. There are no competing interests to declare.

SUPPLEMENTARY MATERIAL

Supplementary Proofs: Proofs for Theorem 1 - 4. (proofs.pdf)

R-package `tvReg`: R-package `tvReg` containing code to perform the equidistant, global and predictor-specific spline fitting methods. The package also contains all datasets used as examples in the article. (vcmasf.zip)

References

Casas, I. and Fernandez-Casal, R. (2019), "tvreg: Time-varying coefficient linear regression for single and multi-equations in r," *SSRN*, URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3363526.

Chen, Y., Wang, Q., and Yao, W. (2015), "Adaptive estimation for varying coefficient models," *Journal of Multivariate Analysis*, 137, 17–31, URL https://www.sciencedirect.com/science/article/pii/S0047259X15000305?via%3Dihub.

Fan, J. and Zhang, W. (1999), "Statistical estimation in varying coefficient models," *The Annals of Statistics*, 27, 1491 – 1518, URL https://doi.org/10.1214/aos/1017939139.

— (2000), "Simultaneous confidence bands and hypothesis testing in varying-coefficient models," *Scandinavian Journal of Statistics*, 27, 715–731, URL http://www.jstor.org/stable/4616637.

Finley, A. O. and Banerjee, S. (2020), "Bayesian spatially varying coefficient models in the spbayes r package," *Environmental Modelling & Software*, 125, 104608, URL https://doi.org/10.1016/j.envsoft.2019.104608.

Harrison, D. and Rubinfeld, D. L. (1978), "Hedonic housing prices and the demand for clean air," *Journal of environmental economics and management*, 5, 81–102, URL https://www.sciencedirect.com/science/article/abs/pii/0095069678900062?via%3Dihub.

Hastie, T. and Tibshirani, R. (1993), "Varying-coefficient models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 757–796, URL http://www.jstor.org/stable/2345993.

Hu, L., Huang, T., and You, J. (2019), "Estimation and identification of a varying-coefficient additive model for locally stationary processes," *Journal of the American Statistical Association*, 114, 1191–1204, URL https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1482753.

Hu, T. and Xia, Y. (2012), "Adaptive semi-varying coefficient model selection," *Statistica Sinica*, 575–599, URL https://www.jstor.org/stable/24310026.

Huang, J. Z. and Shen, H. (2004), "Functional coefficient regression models for non-linear time series: A polynomial spline approach," *Scandinavian Journal of Statistics*, 31, 515–534, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2004.00404.x.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-coefficient models and basis function approximations for the analysis of repeated measurements," *Bimoetrika*, 89, 111–128, URL https://academic.oup.com/biomet/article-abstract/89/1/111/221461.

— (2004), "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," *Statistica Sinica*, 763–788, URL https://www.jstor.org/stable/24307415.

Jullion, A., Lambert, P., Beck, B., and Vandenhende, F. (2009), "Pharmacokinetic parameters estimation using adaptive bayesian p-splines models," *Pharm Stat*, 8, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.336.

Lin, H., Yang, B., Zhou, L., Yip, P. S., Chen, Y.-Y., and Liang, H. (2019), "Global kernel estimator and test of varying-coefficient autoregressive model," *Canadian Journal of Statistics*, 47, 487–519, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11510.

Notari, A. (2021), "Temperature dependence of covid-19 transmission," *Science of The Total Environment*, 763, 144390, URL https://www.sciencedirect.com/science/article/pii/S0048969720379213.

Schwarz, G. et al. (1978), "Estimating the dimension of a model," *The annals of statistics*, 6, 461–464, URL https://www.jstor.org/stable/2958889.

Tang, Q. and Cheng, L. (2012), "Componentwise b-spline estimation for varying coefficient models with longitudinal data," *Statistical Papers*, 53, 629–652, URL https://link.springer.com/article/10.1007/s00362-011-0369-2.

Wang, H. and Xia, Y. (2009), "Shrinkage estimation of the varying coefficient model," *Journal of the American Statistical Association*, 104, 747–757, URL https://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0138.

Wang, W. and Sun, Y. (2019), "Penalized local polynomial regression for spatial data," *Biometrics*, 75, 1179–1190, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13077.

Wang, X., Jiang, B., and Liu, J. (2017), "Generalized r-squared for detecting dependence," *Biometrika*, 104, 129–139, URL https://academic.oup.com/biomet/article/104/1/129/3045032.

Wei, F., Huang, J., and Li, H. (2011), "Variable selection and estimation in high-dimensional varying-coefficient models," *Statistica Sinica*, 21, 1515, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3902862/.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67, URL https://academic.oup.com/jrsssb/article/68/1/49/7110631.

Zhang, X. and Wang, J.-L. (2014), "Varying-coefficient additive models for functional data," *Biometrika*, 102, 15–32, URL https://doi.org/10.1093/biomet/asu053.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429, URL https://www.tandfonline.com/doi/abs/10.1198/016214506000000735.