mathematics

MDPI

Article

INLA Estimation of Semi-Variable Coefficient Spatial Lag Model—Analysis of PM2.5 Influencing Factors in the Context of Urbanization in China

Qiong Pang and Xijian Hu *

College of Mathematics and System Science, Xinjiang University, Urumqi 830046, China; 107552100631@stu.xju.edu.cn
* Correspondence: xijianhu@xju.edu.cn; Tel.: +86-130-7990-0717

Abstract: The Semi-variable Coefficient Spatial Lag Model (SVC-SLM) not only addresses the "dimension disaster" associated with the Varying Coefficient Spatial Lag Model(VC-SLM), but also overcomes the non-linear problem of the variable coefficient, and fully explores the hidden information of the model. In this paper, INLA is firstly used to estimate the parameters of (SVC-SLM) by using B-spline to deal with the non-parametric terms, and the comparative experimental results show that the INLA algorithm is much better than MCMCINLA in terms of both time efficiency and estimation accuracy. For the problem of identifying the constant coefficient terms in the SVC-SLM, the bootstrap test is given based on the residuals. Taking the PM2.5 data of 31 provinces in mainland China from 2015 to 2020 as an empirical example, parametric, non-parametric, and semi-parametric perspectives establish three models of Spatial Lag Model (SLM), VC-SLM, SVC-SLM, which explore the relationship between the covariate factors and the level of urbanization as well as their impacts on the concentration of PM2.5 in the context of increasing urbanization; among the three models, the SVC-SLM has the smallest values of DIC and WAIC, indicating that the SVC-SLM is optimal.

Keywords: semi-variable coefficient spatial lag model; INLA; bootstrap; PM2.5; urbanization

MSC: 62H11

Citation: Pang, Q.; Hu, X. INLA Estimation of Semi-Variable Coefficient Spatial Lag Model—Analysis of PM2.5 Influencing Factors in the Context of Urbanization in China. *Mathematics* **2024**, *12*, 953. https://doi.org/10.3390/math12070953

Academic Editor: Vesna Rajić

Received: 12 February 2024
Revised: 16 March 2024
Accepted: 20 March 2024
Published: 23 March 2024

1. Introduction

The correct choice of an appropriate parametric model can lead to accurate inferences, whereas serious errors in model selection can lead to misleading results, as Robinson (1988) emphasized in his study [1]. Although non-parametric modeling approaches are usually more robust and may be less accurate, a balanced approach can be adopted by using a semi-parametric variable coefficient regression model, which combines the advantages of both parametric and non-parametric models with a high degree of flexibility. It allows some of the coefficients of the model to be parametric while capturing the non-linear relationship between the variable coefficients, thus better adapting to the complexity of the data. It is one of the most effective models developed in recent years to explain the non-linearities and linearities between variables, and with its good explanatory power, it has received a wealth of research and applications in the fields of economics, geography and ecology, among others.

The spatial Econometric Model aims to study the effects of spatial correlation, dependence and spatial patterns on economic phenomena, and has made considerable progress in recent decades. The Spatial Lag Model (SLM), as one of the most important models in Spatial Econometrics, has undergone continuous development and improvement. It has evolved from the classical parametric form to non-parametric and semi-parametric forms to meet various spatial data analysis needs.

SLM is a classical parametric framework within the realm of spatial econometric models, which was first introduced by Cliff (1970) [2]. Cliff's pioneering work ingeniously extended the concept of autocorrelation into the spatial domain. The general expression of SLM is $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \sum_{j=1}^{n} \boldsymbol{x}_j\beta_j + \boldsymbol{\varepsilon}$; the coefficients of $\boldsymbol{x}_j$ are all linear coefficients for β_j. Although SLM has been proven effective in analyzing spatial correlation features, its model typically assumes a linear relationship between the response variable and covariates, and this assumption of a linear coefficient can introduce bias. Moreover, real-world data often exhibit complex non-linear patterns, making it challenging to accurately capture parameter estimates within the framework of parametric assumptions. In such scenarios, the extension to non-parametric or semi-parametric perspectives of the SLM model becomes crucial.

To overcome the linear constraints of SLM and better reflect the non-linear relationship in the actual data, the assumption of a linear relationship between covariates and response variables was relaxed on the basis of SLM. By introducing the variable coefficient term, the Varying Coefficient Spatial Lag Model (VC-SLM) was constructed. In the VC-SLM, the variable coefficient is set as a function of a variable, which can be one of the covariables or some other indicator. This paper focuses on the coefficients of the explanatory variables as a function of the change about a certain variable, the expression is: $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \sum_{j=1}^{n} \boldsymbol{x}_j\boldsymbol{\alpha}_j(\boldsymbol{U}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ responds to the response covariate, $\boldsymbol{x}_j$ and $\boldsymbol{U}$ are covariates, n is the number of variables, and N is the number of samples. Here, $\mathbf{ff_j}(\mathbf{U})$ represents the variable coefficient terms, denoting an unknown non-parametric function on the covariate $\mathbf{U}$, making the VC-SLM non-parametric at this stage. Because the non-parametric has the limitation of "dimension disaster ", there is not much literature on this research. Li et al. (2018) [3] crafted a generalized moment estimation technique tailored for VC-SLM, subsequently applying it to analyze the estimation of China's total factor productivity growth rate. Teng et al. (2023) [4] proceeded to prove the VC-SLM using the MCMCINLA method and effectively incorporated it into empirical studies.

In the VC-SLM, all the coefficients are defined as variable coefficients, meaning they all change with the change in a certain variable. However, when a portion of the $\boldsymbol{\alpha_j(U)}$ in the VC-SLM is constant, such as $\boldsymbol{\alpha_j(U)} = \beta_j, 1 \leq j < n$, the variable coefficient is transformed into a semi-variable coefficient, also called a partial variable coefficient, it corresponds to the semi-parametric model examined in this paper, known as the Semi-Variable Coefficient Spatial Lag Model (SVC-SLM). When all coefficients become constant, that is, $\boldsymbol{\alpha_j(U)} = \beta_j, 1 \leq j \leq n$, the model degenerates into a parametric model, also known as a linear model, specifically the SLM.

The SVC-SLM is a special instance of the model of VC-SLM, where a portion of the variable coefficients are treated as constants. This characteristic endows it with remarkable flexibility and dimensionality reduction capabilities. Currently, there is a large body of theory and applications surrounding this model. Su et al. (2010) [5] is the earliest literature that combines the SLM with a semi-parameter; however, it initially applied the kernel estimation method to the non-parametric components, and then uses the cross-section fitted great likelihood estimation method to obtain the estimation of the model parameters, which is quite computationally intensive due to solving a large system of equations; Su (2012) [6], building upon Su and Jin (2010) [5], proposed a more generalized model allowing for heteroskedasticity and spatial correlation in the errors, employing a two-step estimation method for GMM estimates; nonetheless, it struggled to seamlessly integrate non-parametric and semi-parametric approaches for estimation. Li et al. (2013) [7] proposed a brand new class of SVC-SLM, and derived the cross-sectional maximum likelihood estimation for the model; however, its estimation method only performs better under small sample conditions. Hoshino (2018) [8] constructed a semi-parametric GMM estimation and used it for the study of crime data. Despite their contributions, all of the above estimation methods for the SVC-SLM have an obvious shortcoming of being computationally intensive and time-consuming.

As for the variable coefficient models, the choice of a suitable method for approximating non-parametric components holds paramount significance. Scholars have diligently explored various non-parametric estimation techniques, including kernel estimation [5], local polynomial estimation [9], nearest-neighbor estimation, spline estimation, and penalized spline [10]. Among these methods, B-spline estimation stands out for several compelling reasons. It requires fewer parameters, primarily concerning the selection of the number of nodes, and it displays insensitivity to the choice of nodes. Furthermore, in comparison to local polynomial regression estimation, it eliminates the need to select window widths, resulting in faster computation and greater stability. Therefore, by employing B-spline to approximate the variable coefficient function, the non-parametric component of the model can be transformed into a linear form, allowing for the derivation of a "linear SLM".

In terms of variable coefficient selection in semivariate coefficient models, Guo et al. (2012) [11] adopted an empirical approach of selecting non-parametric components directly from the SLM by identifying non-significant variables whose confidence intervals of the estimates include zero. However, with the advancement of a more robust and widely accepted bootstrap test for variable coefficient selection, this empirical intuitive method of selecting constant coefficients appeared to have no theoretical basis. Li et al. (2016) [12] advocated the use of residual-based bootstrapping to assess whether parametric components in partially linear spatial autoregressive models satisfy specific linearity constraints. Furthermore, Du et al. (2021) [13] applied bootstrap techniques to focus on the coefficient functions in variable coefficient models. With higher statistical rigor and reliability than empirical selection methods, these bootstrap tests have become essential tools for making informed decisions about variable coefficients in such models.

Integrated Nested Laplace Approximation (INLA) is an algorithm proposed by Rue et al. (2009) [14] for approximate Bayesian inference, which is particularly suitable for high-dimensional, complex, or large-scale data analysis. INLA is able to efficiently compute Markov Chain Monte Carlo (MCMC) samples without the need for Bayesian inference results, so it strikes a good balance between computational speed and accuracy, making the estimation and inference of complex models more feasible. This has led to the rapid development of INLA algorithms in fields such as epidemiology [15], tourism [16], and ecology [17]. Research on INLA and spatial econometric modeling has only gradually emerged in the last decade, with Bivand et al. (2014) [18] using INLA for inference for spatial econometric models, but for the case where the covariates are linearly related to the response variable. Subsequently, Gómez-Rubio et al. (2021) [19] introduced a new concept of latent class in spatial econometric modeling and demonstrated that the basic spatial econometric model conforms to the basic INLA framework, and is finally used to empirically compare other algorithms highlighting the INLA advantages. Although Teng [4] proved that the non-parametric SLM conforms to the INLA framework, the MCMCINLA algorithm was ultimately used for parameter estimation, which has not yet overcome the fact that INLA currently supports only parametric forms of spatial econometric models and is also more time-consuming than INLA. At present, there are no scholars who estimate non-parametric spatial econometric models or even semi-parametric spatial econometric models with INLA.

After discussing the background of the models and algorithms, the focus turns to the empirical level. Numerous prior methods have been employed to investigate the determinants of PM2.5 concentration. Some scholars have conducted in-depth studies on the relationship between PM2.5 concentration and urbanization in China by constructing linear spatial econometric models, and have examined this association from various aspects. For example, Liu et al. (2022) [20] examined the causal relationship between urbanization and PM2.5 through an empirical study in China; Yang et al. (2020) [21] used the Spatial Durbin Model to demonstrate that socio-economic factors such as population density have a positive influence on PM2.5 concentration, with covariates including the level of urbanization, industrial activities, vehicular emissions. In the study of urbanization and PM2.5, Chou et al. (2020) [22] explored in depth the spillover effect of population urbanization

on PM2.5 concentration. Among the influencing factors of PM2.5, Lai et al. (2022) [23] showed that PM2.5 concentration was negatively correlated with meteorological factors such as precipitation; Wang et al. (2021) [24] highlighted the significant negative impacts of environmental regulations on PM2.5 pollution; Gao, et al. (2018) [25] explored the impacts of domestic, industrial and motor vehicle exhaust emissions on PM2.5. In essence, the current research on PM2.5 mainly adopts a parametric perspective, but in practice, the linear assumption between the factors affecting PM2.5 concentration may not be able to fully elucidate the changes in the spatial distribution of PM2.5 concentration.

In recent years, as China's urbanization has accelerated, it has undoubtedly driven economic and social advancement. However, it has also brought to the forefront a pressing environmental issue—the proliferation of haze and the associated problem of PM2.5 pollution [26]. Recognizing the severity of this challenge, the Chinese government has elevated haze management to a national strategic level. The Chinese State Council has issued critical directives such as the Action Plan for Prevention and Control of Air Pollution and the Three-Year Action Plan for Winning the Battle for the Blue Sky. These plans emphasize PM2.5 reduction as a pivotal component of comprehensive pollution prevention and control efforts.

This paper is the first to use INLA to estimate the SVC-SLM, extending the studies of Gómez-Rubio (2021) [19], Teng (2023) [4], Su and Jin [5]. In the algorithm aspect, the INLA is used for the first time to estimate the SLM in the non-parametric and semi-parametric perspectives, which fills the gap where INLA cannot estimate the parametric SLM; through the simulation with different sample sizes, positive and negative autocorrelation, periodic and non-periodic variable coefficient functions and the comparison with MCMCINLA algorithm, the advantages of INLA algorithm in terms of short time consumption and high accuracy are highlighted. In the model testing aspect, a bootstrap test under the INLA algorithm is given for the problem of identifying the constant coefficient terms in the SVC-SLM, and the simulations are set up to highlight the validity of the bootstrap test and the efficacy of the test of the estimated statistics. Finally, drawing on non-parametric and semi-parametric perspectives with urbanization as a distinctive explanatory variable, we explore the spatial determinants of PM2.5 concentrations in 31 regions across mainland China from 2015 to 2020. Through model comparison, hidden information is further unearthed, highlighting the necessity of applying the SVC-SLM model and the efficiency of the INLA algorithm in estimation.

The rest of this paper is as follows: In Section 2, data sources and preprocessing of variables affecting PM2.5 concentration are given. The estimation of SVC-SLM based on the INLA algorithm is carried out from four aspects: the construction of SVC-SLM, proving whether the SVC-SLM satisfies the GMRF structure, the steps of the INLA algorithm for SVC-SLM and the bootstrap test. Section 3 is the numerical simulation of the INLA algorithm, which includes three parts: simulating SVC-SLM estimation based on the INLA algorithm, performing a comparative experiment with the MCMCINLA algorithm, and simulating the bootstrap test based on INLA to distinguish the constant coefficient of SVC-SLM estimation. Section 4, utilizes SLM, VC-SLM and SVC-SLM models to compare and analyze the influencing factors of PM2.5 data to prove the rationality of the proposed INLA algorithm and bootstrap test. Section 5 gives some summary results. Finally, Section 6 puts forward the suggestion and prospect of the paper.

2. Materials and Methods

2.1. Data Sources and Preprocessing

We selected data from the annual averages of the 31 provinces in mainland China, excluding Taiwan, Hong Kong, and Macau, for the years 2015 to 2020 as our study dataset. This dataset allowed us to analyze the spatial distribution characteristics of haze pollution.

The previous literature [20–24] had investigated various factors influencing PM2.5 from social, economic, meteorological, and other perspectives, in summary: the causal relationship between urbanization and PM2.5 [20], socio-economic factors such as population

density and industrial structure affecting PM2.5 [21], the impact of different levels of urbanization, industrial activities, and vehicular emissions on PM2.5 [22] , meteorological factors including temperature, humidity, and wind speed affecting PM2.5 [23], and the influence of environmental factors on PM2.5 concentrations [24]. Therefore, seven covariates are finally selected, namely, urbanization rate, GDP per capita, per capita annual local financial expenditure on environmental protection, industrial emissions, domestic emissions, motor vehicle emissions, and local financial expenditure on environmental protection, and average annual precipitation. The selection of the variables covered five perspectives, including social, economic, and human activities, as well as environmental protection and natural meteorology. Table 1 gives the source of the data, the name of the variable of the indicator, the domains in which the variables were selected, and the description of the variables.

To address heteroskedasticity, all explanatory variables were log-transformed.

Table 1. Description of Data Indicators.

Variable	Variable Selection Angle	Indicator Definitions	Units	Data Sources
ln_PM2.5	haze concentration	Annual Average PM2.5 Concentration	micrograms/cubic meter/year	Columbia University Center for Social and Economic Research and Data
ln_Urban	societies	Average Annual Urbanization Rate	%	China Statistical Yearbook
ln_GDP	economics	Gross Domestic Product	billion yuan/year	China Statistical Yearbook
ln_Industry	human activity	Annual Average Industrial Exhaust Emissions	Ten thousand tons/year	China Environmental Statistical Yearbook
ln_Life	human activity	Annual Average Life Exhaust Emission	Ten thousand tons/year	China Environmental Statistical Yearbook
ln_Car	human activity	Annual Average Motor Vehicle Exhaust Emissions	Ten thousand tons/year	China Environmental Statistical Yearbook
ln_Environment	environmental protection	Per Capita Annual Local Financial Expenditure on Environmental Protection	yuan/person/year	China Environmental Statistical Yearbook
ln_Rain	natural environment	Average Annual Rainfall	millimeters/ year	European Union and European Centre for Medium-Range Weather Forecasts

Spatial correlation analysis is a necessary step preceding spatial measurement analysis. Global spatial autocorrelation measures the overall distribution of observed objects, and the degree of correlation is often assessed using the global Moran's I index [27]. The Moran's I index value falls within the range of $[-1, 1]$. A value greater than 0 indicates a spatial positive correlation between PM2.5 concentration and its influencing factors in the region. When Moran's I index value is greater than 0, it signifies a spatially positive correlation between regions for PM2.5 concentration and its influencing factors. Conversely, a value less than 0 indicates a spatially negative correlation, while a value equal to 0 suggests the absence of spatial correlation between regions.

The results of the global spatial autocorrelation test for each year and for the 6 years of 2015–2020 as a whole are given in Table 2. The results show that the Moran's I index of PM2.5 concentration for each year and for 2015–2020 as a whole is always significant at the 5% significance level, with a p-value much less than 0.05, indicating that haze pollution is

spatially correlated and the Moran's I value of 0.5080 for 2015–2020 suggests that there is a spatial positive correlation of PM2.5 concentration among provincial regions. Consequently, it is both logical and imperative to employ spatial econometric models that encompass spatial effects.

Table 2. Global spatial autocorrelation test for provincial-level PM2.5 in China, 2015–2020.

Year	Moran I Statistic	*p*-Value
2015	0.5374	4.57×10^{-7}
2016	0.4945	2.69×10^{-6}
2017	0.5011	1.91×10^{-6}
2018	0.3974	1.07×10^{-4}
2019	0.4345	2.84×10^{-5}
2020	0.3288	9.00×10^{-4}
2015–2020	0.5080	2.20×10^{-16}

2.2. SVC-SLM Estimation Based on INLA

2.2.1. Model Construction

For the PM2.5 influencing factors in this study, there are as many as seven, although the advantage of the non-parametric regression model lies in its ability to adapt without needing to preset the specific form of the regression function, it often encounters the "dimensionality catastrophe" problem in practical applications. At such times, the model of SVC-SLM is more capable of explaining the actual problem [7]:

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{\alpha}_j(\boldsymbol{U}) + \sum_{k=p+1}^{n} \boldsymbol{x}_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{W}$ is the spatial weight matrix between regions, given that the focus is on the 31 provincial districts in mainland China, excluding Taiwan, Hong Kong, and Macau, a ROOK-type neighborhood weight matrix is selected for analysis, under this matrix two regions are considered to be adjoining on the condition that they are adjacent on the boundary, irrespective of whether they are adjacent on the corners. The spatial lag term, denoted by ρ in $\rho \boldsymbol{W} \boldsymbol{y}$, ensures $|\rho| < 1$ to reflect the spatial dependence of the response variable $\boldsymbol{y}$, so ρ is called the spatial autocorrelation coefficient. In this empirical study, $\boldsymbol{y}$ represents PM2.5 data sampled from 31 provinces over 6 years, resulting in $N = 186$. Here, n represents the number of covariates, p represents the number of variables whose coefficients are variable coefficients, and correspondingly, $n - p$ is the number of variables whose coefficients are constant coefficients, in the model of SVC-SLM we studied, $n = 6$. $\boldsymbol{x}_j$ and $\boldsymbol{U}$ are the covariates, $\boldsymbol{U}$ denotes urbanization in the empirical evidence, β_j is the constant coefficient, and $\boldsymbol{\alpha}_j(\boldsymbol{U})$ is the varying coefficient. $\boldsymbol{\varepsilon}$ is the error term, which can be either Gaussian or non-Gaussian, and in order to draw more general conclusions, the main assumption in this paper is to obey the Gaussian distribution with zero mean and diagonal covariance matrix $\sigma^2 \boldsymbol{I}_N$, $\boldsymbol{I}_N$ is the identity matrix of order N.

2.2.2. Proof SVC-SLM of GMRF Structure

The INLA algorithm is a fast computational method provided by Rue et al. [14] for the standard generalized linear models of Gaussian Markov Random Fields GMRF, which includes Hidden Gaussian Random Field models [28].

Model (1) contains variable coefficient term $\boldsymbol{\alpha}_j(\boldsymbol{U})$ and is not a simple linear model; Gómez-Rubioa [19] pointed out that INLA can deal with spatial econometric models with random effects potential with linear predictors, in order to realize the ability to use INLA for parameter estimation, firstly, the variable coefficients $\boldsymbol{\alpha}_j(\boldsymbol{U})$ in the Model (1) are converted into a linear form using the B-spline transform, thus transforming the SVC-SLM into a "linear SLM".

For the variable coefficients $\boldsymbol{\alpha_j(U)}$ in Model (1), refer to Teng [4] using B-spline processing. Assuming that each variable coefficient $\boldsymbol{\alpha_j(U)}$ in Model (1) is smooth, there exists a set of basis functions ${B_d}^{(j)}$ and constants ${b_d}^{(j)}$ such that $\boldsymbol{\alpha_j(U)} \approx \sum_{d=1}^{h} {B_d}^{(j)} {b_d}^{(j)} = \boldsymbol{B^{(j)} b^{(j)}}$, which are expanded into a matrix form as follows:

$$\boldsymbol{\alpha_j(U)} \approx \boldsymbol{B^{(j)} b^{(j)}} = B_1^{(j)} b_1^{(j)} + B_2^{(j)} b_2^{(j)} + \ldots + B_h^{(j)} b_h^{(j)} \tag{2}$$

at this juncture, $\boldsymbol{B^{(j)}}$ represents a B-spline basis function, constituting a $p \times h$-dimensional design matrix, and $\boldsymbol{b^{(j)}}$ denotes a spline basis, forming an h-dimensional vector. Let $\boldsymbol{Z_j = x_j B^{(j)}}$, so Model (1) is rewritten as:

$$\begin{aligned} \boldsymbol{y} &= \rho \boldsymbol{W y} + \textstyle\sum_{j=1}^{p} \boldsymbol{x_j \alpha_j(U)} + \sum_{k=p+1}^{n} \boldsymbol{x_k} \beta_k + \boldsymbol{\varepsilon} \\ &= (\boldsymbol{I_N} - \rho \boldsymbol{W})^{-1} (\textstyle\sum_{j=1}^{p} Z_j \boldsymbol{b^{(j)}} + \sum_{k=p+1}^{n} \boldsymbol{x_k} \beta_k + \boldsymbol{\varepsilon}). \\ &= (\boldsymbol{I_N} - \rho \boldsymbol{W})^{-1} (\boldsymbol{Z b} + \boldsymbol{X \beta} + \boldsymbol{\varepsilon}) \end{aligned} \tag{3}$$

where the variable $\boldsymbol{V} = (\boldsymbol{Z}, \boldsymbol{X})$ with $\boldsymbol{Z} = \sum_{j=1}^{p} Z_j$ and $\boldsymbol{X} = \sum_{k=p+1}^{n} \boldsymbol{x_k}$, thus the coefficients are changed to $\boldsymbol{\psi} = (\boldsymbol{b}, \boldsymbol{\beta})$, accordingly, $\boldsymbol{b} = \sum_{j=1}^{p} \boldsymbol{b^{(j)}}$ and $\boldsymbol{\beta} = \sum_{k=p+1}^{n} \beta_k$. Rewrite the Model (3) as:

$$\boldsymbol{y} = (\boldsymbol{I_N} - \rho \boldsymbol{W})^{-1} (\boldsymbol{V \psi} + \boldsymbol{\varepsilon}) \tag{4}$$

The Model (4) is a general form of SLM, but it cannot be directly fitted to INLA yet, we need to construct potential classes to conform to the INLA framework. Model (4) is written as a new potential SLM given in literature [19]:

$$\begin{cases} \boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\zeta} \\ \boldsymbol{\eta} = (\boldsymbol{I_N} - \rho \boldsymbol{W})^{-1} (\boldsymbol{V \psi} + \boldsymbol{\varepsilon}) \end{cases} \tag{5}$$

where $\boldsymbol{\zeta}$ is the small error of $\boldsymbol{y}$ added to Model (5).

$\boldsymbol{\psi}$ has a Gaussian prior with zero mean and precision matrix $\boldsymbol{Q}$; the potential effects are already defined in Model (5), so $\boldsymbol{Q}$ is fixed [19], and $\boldsymbol{\varepsilon}$ obeys a Gaussian distribution with zero mean and precision matrix $\tau \boldsymbol{I_N}$, where τ is the precision parameter. Based on the previous analyses, $(\boldsymbol{\eta}, \boldsymbol{\psi})$ is a GMRF with zero mean and precision matrix $\boldsymbol{K}$ (structure below), and therefore, the Model (1) conforms to the INLA framework. (Only the main results are shown here, see Appendix A for the exact proof procedure).

$$\begin{aligned} \boldsymbol{K} &= \begin{pmatrix} \boldsymbol{P} & -\boldsymbol{P}(\boldsymbol{I_N} - \rho \boldsymbol{W})^{-1} \boldsymbol{V} \\ -\boldsymbol{V}'(\boldsymbol{I_N} - \rho \boldsymbol{W})^{-1} \boldsymbol{P} & \boldsymbol{Q} + \tau \boldsymbol{V}' \boldsymbol{V} \end{pmatrix} \\ &= \begin{pmatrix} \tau(\boldsymbol{I_N} - \rho \boldsymbol{W}')(\boldsymbol{I_N} - \rho \boldsymbol{W}) & -\tau(\boldsymbol{I_N} - \rho \boldsymbol{W}')\boldsymbol{V} \\ -\tau \boldsymbol{V}'(\boldsymbol{I_N} - \rho \boldsymbol{W}) & \boldsymbol{Q} + \tau \boldsymbol{V}' \boldsymbol{V} \end{pmatrix} \end{aligned}$$

2.2.3. Steps of the INLA Algorithm for SVC-SLM

The INLA algorithm is essentially a fast computational method used to estimate the posterior distribution of parameters.

In the potential Model (5), after the B-spline basis expansion, the prior distribution is first assigned to the set of coefficients $\boldsymbol{\psi}$, and the spatial autocorrelation parameter ρ, so the set of hyper-parameters to be estimated by the model is $\boldsymbol{\Theta} = \{\rho, \boldsymbol{\psi}\} = \{\rho, \boldsymbol{b}, \boldsymbol{\beta}\}$. The main purpose of INLA is to obtain the joint conditional posterior distribution for $\boldsymbol{\eta}$ and $\boldsymbol{\Theta}$ [14]. The INLA algorithm estimates the marginal posterior distribution of $\boldsymbol{\eta}$ and $\boldsymbol{\Theta}$ in three steps: estimate $\pi(\boldsymbol{\Theta}|\boldsymbol{y})$, estimate $\pi(\Theta_j \mid \boldsymbol{y})$, and estimate $\pi(\eta_j \mid \boldsymbol{\Theta}, \boldsymbol{y})$. The specific INLA estimation process is shown in Figure 1.

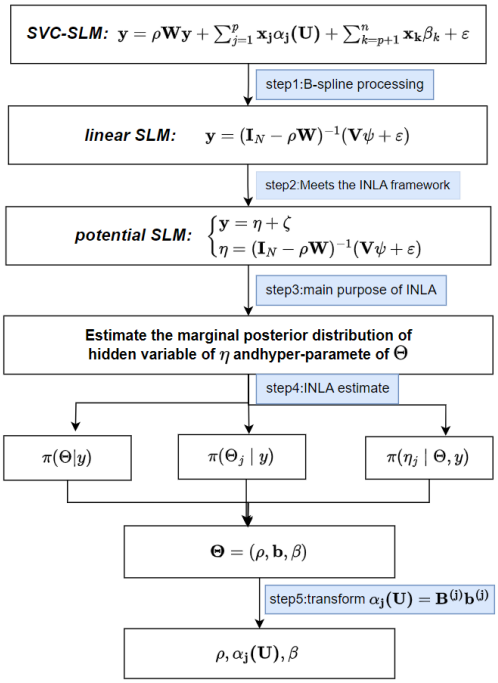

Figure 1. Flow chart of SVC-SLM for INLA algorithm.

As shown in Figure 1, the whole algorithmic process is carried out in five main steps:

1. From the SVC-SLM of Model (1) B-spline is processed into a linear SLM, i.e., Model (4).
2. Section 2.2.2 has demonstrated that SVC-SLM satisfies the INLA framework, suggesting the Model (1) is suitable for processing using the INLA methodology.
3. For the three main steps of the INLA method, INLA first requires different prior distributions to describe the different parameters in the model, including the coefficient vector $\boldsymbol{\psi}$, the spatial autocorrelation parameter ρ, and the precision of the error term τ. By default, the coefficient vector $\boldsymbol{\psi}$ is distributed with a multivariate Gaussian distribution with zero mean and precision matrix $\boldsymbol{Q}$ (which must be specified), The code for setting the parameters of $\boldsymbol{\psi}$ is :

$$betaprec < -0.0001$$

$$Q.beta < -Diagonal(n = ncol(mmatrix), betaprec)$$

τ conforms to a log gamma distribution, and in order to control the spatial autocorrelation parameter of ρ in the $(0,1)$ interval, set the prior distributions of τ and ρ with $hyper$:

$$hyper < -list(prec = list(prior = \text{“}loggamma\text{”}, param = c(0.01, 0.01)),\\ rho = list(initial = 0, prior = \text{“}logitbeta\text{”}, param = c(1,1)))$$

After the parameter setting, the a posteriori estimation of the regression coefficients can be obtained by using the $inla()$ function, the key code for the process is as follows:

$$inla(y \sim -1 + f(idx, model = \text{“}slm\text{”},\\ args.slm = list(rho.min = rho.min, rho.max = rho.max,\\ W, X, Q.beta), hyper), data, \ldots)$$

here, $rho.min$ and $rho.max$ are determined by the spatial weighting matrix of $\boldsymbol{W}$, which are the minimum and maximum eigenroots of $\boldsymbol{W}$, respectively; $data$ represents the complete data set, W is the spatial weight matrix as above, X stands for $\boldsymbol{V}$ in Model (5), i.e., the set of all variables.

4. The a posteriori estimate of $\boldsymbol{\Theta} = \{\rho, \boldsymbol{b}, \boldsymbol{\beta}\}$ can be obtained using the $summary()$ function; note, that the $\boldsymbol{b}$ obtained directly at this point is an estimate on $\boldsymbol{b}^{(j)}$.
5. Finally, it is also necessary to multiply $\boldsymbol{b}^{(j)}$ with the design matrix $\boldsymbol{B}^{(j)}$, and the final value of the variation coefficient of $\boldsymbol{\alpha}_j(\boldsymbol{U})$, can be obtained by using Model (2) with $\boldsymbol{\alpha}_j(\boldsymbol{U}) = \boldsymbol{B}^{(j)}\boldsymbol{b}^{(j)}$. In the end, INLA successfully estimated all coefficients, denoted as $\boldsymbol{\Theta} = \{\rho, \boldsymbol{\alpha}_j(\boldsymbol{U}), \boldsymbol{\beta}\}$.

2.2.4. Bootstrap Test for Constant Coefficients

In the context of the SVC-SLM, it becomes crucial to contemplate whether the regression coefficients vary based on certain variables. This essentially involves testing whether specific regression coefficients remain constant. In other words, it involves scrutinizing whether the coefficient function $\boldsymbol{\alpha}_j(\boldsymbol{U})$ within an SVC-SLM remains invariant and equal to β_j.

To test whether some of the coefficients of the independent variables vary with some covariate $\boldsymbol{U}$ is crucial for the identification of the constant coefficients of an SVC-SLM, hence, the following assumptions:

$$\begin{cases} H_0 : \boldsymbol{\alpha}_j(\boldsymbol{U}) = \beta_k, k = p+1, \ldots, n \\ H_1 : \text{ all } \boldsymbol{\alpha}_j(\boldsymbol{U})(j = 1, \ldots, n) \text{ changing with } \boldsymbol{U} \end{cases}$$

For this hypothesis, in the null hypothesis H_0 , the corresponding model is SVC-SLM, which is the focus of this paper, and the INLA method proposed in this paper can be used to fit it. The model under the alternative hypothesis H_1 is VC-SLM, the specific model is Model (6), and the definition of Model (6) is basically the same as that of Model (1), where the coefficients are all variable coefficients $\boldsymbol{\alpha}_j(\boldsymbol{U})$.

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{\alpha}_j(\boldsymbol{U}) + \boldsymbol{\varepsilon} \tag{6}$$

Since Teng [4] proved that the VC-SLM conforms to the INLA framework, INLA is also used to estimate the VC-SLM under H_1 . The selection of window width draws on the suggestion of Fan et al. (2005) [29], that is, the same window width is used to fit the null hypothesis and the alternative hypothesis, so that their logarithmic likelihood is comparable. In the actual implementation, the window width selected by the model under the alternative hypothesis H_1 is selected as the window parameter for model fitting.

The statistic of the bootstrap test selects the generalized likelihood ratio statistic [30], which is an important metric to measure the goodness of fit difference between the original model and the alternative model. Constructing the generalized likelihood ratio statistic as T is expressed as follows:

$$T = \frac{RSS_{H_0} - RSS_{H_1}}{RSS_{H_1}} \tag{7}$$

here, RSS_{H_0} and RSS_{H_1} are the residual sum of squares obtained after the calculation for the null hypothesis and the alternative hypothesis, respectively.

The p-value estimated by the bootstrap test is:

$$p = p_{H_0}(T \geq t) \tag{8}$$

where $p_{H_0}(\bullet)$ represents the probability calculated under the null hypothesis H_0 , and t is the observed value of the statistic T. α is the given significance level, if $\rho < \alpha$, null hypothesis H_0 is rejected; otherwise, the opposite.

See Appendix B for the detailed bootstrap process.

3. Simulation of the INLA Algorithm

3.1. Simulation of the SVC-SLM Using INLA

According to the "Voronoi subdivision method" proposed by Stakhovych et al. (2009) [31] for generating ROOK-type spatial weight matrices for the simulated data, SVC-SLM refers to the following model:

$$\boldsymbol{y} = (\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}(\boldsymbol{x_1}\beta_1 + \boldsymbol{x_2}\beta_2 + \boldsymbol{x_3}\boldsymbol{\alpha_1}(\boldsymbol{U}) + \boldsymbol{x_4}\boldsymbol{\alpha_2}(\boldsymbol{U}) + \boldsymbol{\varepsilon}) \tag{9}$$

Let the covariate be $\boldsymbol{x_1} \sim U[-1,1], \boldsymbol{x_2} \sim U[-3,2], \boldsymbol{x_3} \sim U[-3,3], \boldsymbol{x_4} \sim U[0,2], \boldsymbol{U} \sim U[-1,1]$ and the standard deviation be $\sigma = 0.1$.

To be more relevant, the two variable coefficient functions are set as periodic and non-periodic functions, respectively:

$$\boldsymbol{\alpha_1}(\boldsymbol{U}) = 0.5 \times e^{(\boldsymbol{U}^2 + 3 \times \boldsymbol{U})}, \boldsymbol{\alpha_2}(\boldsymbol{U}) = 2 \times \sin(3 \times \Pi \times \boldsymbol{U})$$

In order to compare the effect of different parameters on the estimation effect, there are the following settings.

1. The sample values N are a small sample 30, medium sample 100 and large sample 400;
2. $\beta_1 = 0.4, \beta_2 = 0.6$;
3. $\rho = -0.9, -0.5 - 0.1, 0.1, 0.5, 0.9$.

A total of 18 sets of simulation experiments with different scenarios were carried out, each group was simulated 100 times to evaluate the estimation performance and robustness of the algorithm under different sample sizes and different degrees of positive and negative autocorrelation. The estimates presented in Tables 3 and 4 are averages based on the results of 100 times of INLA simulations.

The mean square error (MSE) and the deviation information criterion (DIC) are selected as test indexes for model and parameter fitting, and the smaller the value of MSE and DIC, the better the estimation performance. Note, that in order to fit the model, the following simulation experiments set the variance of Gaussian likelihood to the same fixed small value for the same sample size (by setting the log precision to 15), so this results in the same DIC value for the same sample size in the same model, this setup refers to Gómez-Rubio et al. [19]. These complex calculations are performed using the R-INLA (http://www.r-inla.org, accessed on 1 March 2022) R4.2.3 package.

According to the results in Tables 3 and 4, it can be seen that INLA can provide robust parameter estimation results under various circumstances. The results show that for three different samples and six different correlations, with the larger the sample size and the larger the spatial autocorrelation contained in the model, the MSE of the estimation results of parameter ρ, the estimation results of constant-coefficient β_1, β_2, and the estimation results of variable coefficient $\boldsymbol{\alpha_1}(\boldsymbol{U})$, $\boldsymbol{\alpha_2}(\boldsymbol{U})$ all become smaller and smaller, indicating that the fitting effect is becoming better and better. In addition, the DIC value decreased with the increase in the sample size, indicating that the model fitting effect became better and better with the increase in the sample size.

Figures 2 and 3 show the fitting curves of the coefficient function $\boldsymbol{\alpha_1}(\boldsymbol{U})$, $\boldsymbol{\alpha_2}(\boldsymbol{U})$ of the forward autocorrelation variable coefficient term simulated by each example, respectively. The red dotted line is the true curve, the black solid line is the fitting curve, and the gray band is the 95% confidence interval. In the figure, CaseA, CaseB, and CaseC represent the cases when $N = 30$, $N = 100$, and $N = 400$, respectively. (Since similar conclusions are reached about negative correlations, only coefficient estimates for positive correlations are given here).

Comparing Figure 2 with Figure 3, it is found that no matter a small sample or large sample, the fitting graph of non-periodic function $\boldsymbol{\alpha_1}(\boldsymbol{U})$ is better than that of periodic function $\boldsymbol{\alpha_2}(\boldsymbol{U})$, whether it is the width of the confidence interval or the coincidence between the real curve and the fitting curve.

Table 3. ρ estimation results, MSE and DIC of six simulation cases of SVC-SLM.

N	ρ	$\tilde{\rho}$	MSE_ρ	MSE_y	DIC
30	−0.9	−0.8707	2.94×10^{-5}	4.81×10^{-13}	−334.86
	−0.5	−0.4850	7.47×10^{-6}	3.98×10^{-13}	−334.86
	−0.1	−0.1040	5.56×10^{-7}	1.01×10^{-13}	−334.86
	0.1	0.0890	4.07×10^{-6}	2.79×10^{-13}	−334.86
	0.5	0.4842	8.30×10^{-6}	1.64×10^{-13}	−334.86
	0.9	0.8922	2.04×10^{-6}	9.93×10^{-13}	−334.86
100	−0.9	−0.8981	4.09×10^{-8}	3.69×10^{-12}	−1116.21
	−0.5	−0.4980	3.70×10^{-8}	3.33×10^{-12}	−1116.21
	−0.1	−0.0991	8.54×10^{-9}	3.20×10^{-12}	−1116.21
	0.1	0.1006	3.49×10^{-9}	3.17×10^{-12}	−1116.21
	0.5	0.4997	8.96×10^{-10}	3.31×10^{-13}	−1116.21
	0.9	0.8997	7.02×10^{-10}	3.73×10^{-12}	−1116.21
400	−0.9	−0.8999	8.31×10^{-11}	9.39×10^{-12}	−4464.85
	−0.5	−0.5002	6.45×10^{-11}	8.60×10^{-12}	−4464.85
	−0.1	−0.1034	6.07×10^{-12}	5.07×10^{-12}	−4464.85
	0.1	0.0994	7.78×10^{-10}	8.09×10^{-12}	−4464.85
	0.5	0.4993	1.09×10^{-9}	8.42×10^{-12}	−4464.85
	0.9	0.8997	2.90×10^{-10}	9.47×10^{-12}	−4464.85

Table 4. Estimation results of constant and variable coefficients for six simulation examples of SVC-SLM.

N	ρ	$\tilde{\beta_1}$	$\tilde{\beta_2}$	MSE_{β_1}	MSE_{β_2}	MSE_{α_1}	MSE_{α_2}
30	−0.9	0.7306	0.5925	3.65×10^{-3}	2.02×10^{-6}	6.58×10^{-2}	3.45×10^{-2}
	−0.5	0.6931	0.6228	2.86×10^{-3}	1.74×10^{-5}	3.66×10^{-2}	3.10×10^{-1}
	−0.1	0.6379	0.6524	1.89×10^{-4}	9.15×10^{-5}	2.34×10^{-2}	2.93×10^{-1}
	0.1	0.6105	0.6642	1.48×10^{-3}	1.37×10^{-4}	2.24×10^{-2}	2.92×10^{-1}
	0.5	0.5720	0.6756	9.86×10^{-4}	1.90×10^{-4}	2.34×10^{-2}	2.94×10^{-1}
	0.9	0.5346	0.6650	6.04×10^{-4}	1.41×10^{-4}	2.44×10^{-2}	2.93×10^{-1}
100	−0.9	0.4325	0.6073	1.06×10^{-5}	5.35×10^{-7}	8.16×10^{-4}	4.30×10^{-3}
	−0.5	0.4322	0.6074	1.04×10^{-5}	5.41×10^{-7}	8.12×10^{-4}	4.25×10^{-3}
	−0.1	0.4317	0.6073	1.01×10^{-5}	5.36×10^{-7}	7.91×10^{-4}	4.16×10^{-3}
	0.1	0.4315	0.6073	9.93×10^{-6}	5.32×10^{-7}	7.80×10^{-4}	4.12×10^{-3}
	0.5	0.4311	0.6072	9.65×10^{-6}	5.18×10^{-7}	7.56×10^{-4}	4.06×10^{-3}
	0.9	0.4307	0.6072	9.44×10^{-6}	5.16×10^{-7}	7.43×10^{-4}	4.02×10^{-3}
400	−0.9	0.4059	0.5981	8.60×10^{-8}	9.17×10^{-9}	1.98×10^{-4}	9.52×10^{-4}
	−0.5	0.4062	0.5980	9.48×10^{-8}	9.89×10^{-9}	1.98×10^{-4}	9.48×10^{-4}
	−0.9	0.4059	0.5981	8.60×10^{-8}	9.17×10^{-9}	1.98×10^{-4}	9.52×10^{-4}
	0.1	0.4066	0.5979	1.07×10^{-7}	1.06×10^{-8}	1.95×10^{-4}	9.37×10^{-4}
	0.5	0.4067	0.5979	1.12×10^{-7}	1.08×10^{-8}	1.92×10^{-4}	9.31×10^{-4}
	0.9	0.4067	0.5979	1.12×10^{-7}	1.15×10^{-8}	1.92×10^{-4}	9.26×10^{-4}

In general, the confidence interval of the coefficient function of the variable coefficient term is relatively wider in the case of small samples, and the fluctuation between the boundary and trough and peak is larger. With the increase in the sample size, the estimation accuracy of the coefficient function of the variable coefficient term is effectively improved in the case of medium and large samples, starting from the sample $N = 100$. The overall curve fitting of both non-periodic function $\boldsymbol{\alpha_1(U)}$ and periodic function $\boldsymbol{\alpha_2(U)}$ is almost close to the true value, and the estimation effect is quite good, which further shows that B-spline can deal with the problem of variable coefficients very well.

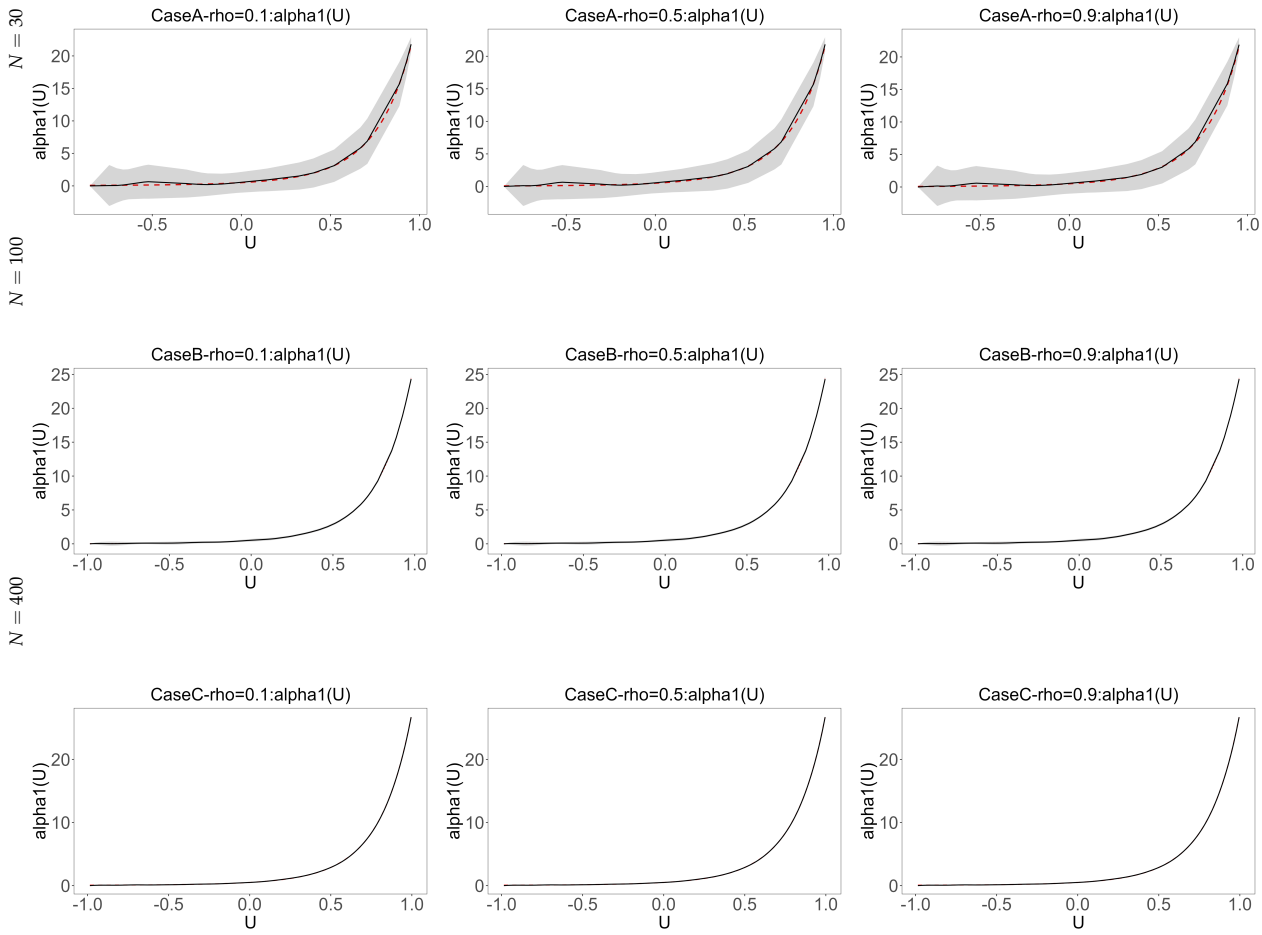

Figure 2. Nonperiodic function $\alpha_1(U)$ fit for 6 simulation cases.

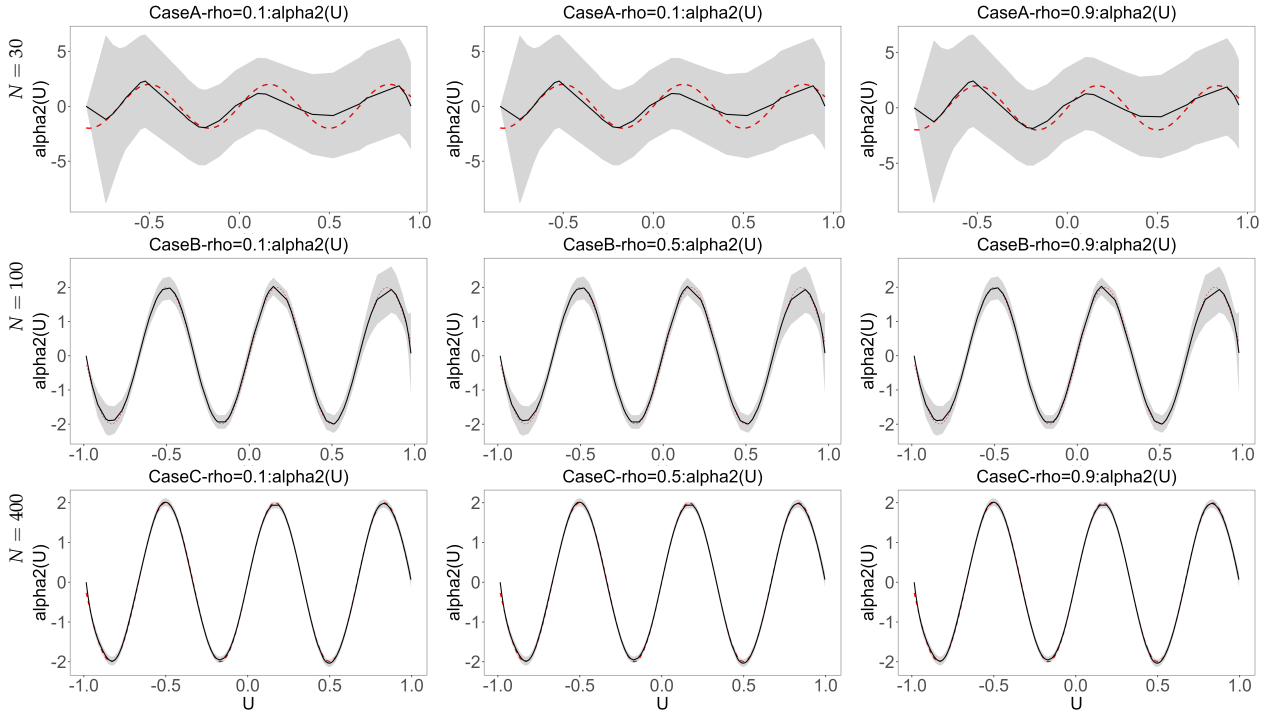

Figure 3. Periodic function $\alpha_2(U)$ fit for 6 simulation cases.

3.2. Comparison of INLA and MCMCINLA Algorithm

In order to highlight the accuracy of the SVC-SLM estimation of the INLA algorithm, the simulation comparison of the MCMCINLA algorithm is given here. The prerequisite for the use of the MCMCINLA algorithm is to satisfy the basic framework of INLA [4], and it has been proved in Section 2.2.2 and Appendix A that the SVC-SLM conforms to the basic framework of INLA, which indicates that the SVC-SLM is also suitable for the MCMCINLA algorithm.

So as to minimize the impact of sample size and spatial autocorrelation on parameter estimation, $N = 400$ and the spatial autocorrelation coefficient $\rho = 0.5$ are set here. Set the burn-in simulation 20 times, burn-in rejection to keep one of the five iterations, and the final total of 80 iterations of the simulation analysis.

Based on the findings presented in Tables 5 and 6, it is apparent that INLA outperforms MCMCINLA in terms of both estimation time and accuracy. The estimation time required for MCMCINLA is approximately 425 times longer than that of INLA, and for the test index of parameter fitting, the MSE of MCMCINLA is higher than the MSE of INLA. Moreover, the operation of INLA is simple and easy, while the adjustment process of MCMCINLA is relatively cumbersome and requires manual adjustment of the sampling parameters such as step size. In summary, after a thorough comparison and analysis of the two algorithms, it becomes evident that the INLA algorithm excels in terms of both time efficiency and estimation accuracy.

Table 5. Estimation results of each parameter of SVC-SLM under INLA/MCMCINLA algorithm and time.

Algorithm	$\tilde{\rho}$	$\tilde{\beta_1}$	$\tilde{\beta_2}$	Time
INLA	0.4993	0.4067	0.5979	9.96 [1]
MCMCINLA	0.4983	0.3852	0.6041	4234.32 [1]

[1] The unit of time is second.

Table 6. MSE for the estimation of each parameter of the SVC-SLM under the INLA/MCMCINLA algorithm.

Algorithm	MSE_ρ	MSE_{β_1}	MSE_{β_2}	$MSE_{\alpha_1(u)}$	$MSE_{\alpha_2(u)}$	MSE_y
INLA	1.09×10^{-9}	1.12×10^{-7}	1.08×10^{-8}	1.92×10^{-4}	9.31×10^{-4}	8.42×10^{-12}
MCMCINLA	7.23×10^{-9}	5.47×10^{-7}	4.20×10^{-8}	3.95×10^{-4}	1.06×10^{-3}	9.48×10^{-12}

3.3. Bootstrap Simulation Experiment

Aimed at verifying the practicability of the bootstrap test for the INLA algorithm, the corresponding bootstrap simulation of Model (9) is given as follows. See Appendix B for the specific bootstrap verification process.

In order to be more appropriate to the actual situation, the two variable coefficient functions are set as periodic function and aperiodic function, respectively, so as to explore the study of variable coefficient in different cases by the INLA algorithm. The standard deviation of ε is $\sigma = 1$, and the value range of each covariable and the setting of its coefficient are shown in Table 7.

Table 7. Bootstrap tested the value range and coefficient setting of each covariable of the model.

Concomitant Variable	Variable Value	Coefficients of Covariates	Setting of Coefficient
x_1	$U(0,3)$	$\alpha_1(u)$	$0.2e^{(u^2+3u)}$
x_2	$U(0,3)$	$\alpha_2(u)$	$0.3\sin(3\Pi u)$
x_3	$U(0,3)$	$\beta_1(u)$	$0.4 + c\sin(3\Pi u)$
x_4	$U(0,3)$	$\beta_2(u)$	$0.6 + cu(3-4u)$
u	$U(0,1)$	-	-

Table 7 of c is a constant, and the efficacy of the test is evaluated by the value taken for c. When $c = 0$, it means that H_0 is true and $c \neq 0$ means that H_1 is true, and the deviation between H_0 and H_1 increases with the absolute value of c.

To examine the influence of error distribution on the test performance, two different error distributions with distinct characteristics are considered, such as $N(0,1), U(-4,4)$. The null hypothesis at this point is $H_0 : \boldsymbol{\beta_1(U)} = \beta_1$ and $\boldsymbol{\beta_2(U)} = \beta_2$.

For each simulation, according to the Appendix B and Model (A7), $b = 500$ bootstrap samples were taken from each repetition to calculate the p-value analyzed in terms of both the size of the test and the efficacy of the test, respectively, and the experiment is repeated $K = 100$ times.

(1) The Validity of the Test

Make the coefficient function $\boldsymbol{\beta_1(U)}$ and $\boldsymbol{\beta_2(U)}$ for $c = 0$, namely, H_0 is true and $\beta_1(U)$=0.4, $\beta_2(U)$=0.6, then in the original hypothesis and alternative hypothesis and each setting calculate the frequency of 100 repetitions less than a given level of significance (i.e., the rejection of H_0); from the results of Table 8 it can be seen when H_0 is true in all the experimental settings, regardless of whether it is a small sample of $N = 30$ or a large sample of $N = 100$, the rate of rejection of the null hypothesis under the null hypothesis is also reasonably close to the $\alpha = 0.05$ significance level, and there is no significant difference in the simulation results regardless of whether the error distribution is normal or uniform. All these demonstrate the effectiveness of the bootstrap approach to zero distribution.

(2) Test Statistic Efficacy

To evaluate the efficacy of the test statistics, the article considers scenarios where $c \neq 0$ namely H_1 is true, assuming that all coefficients are variable coefficients. In this case, the values of c in the coefficient function $\boldsymbol{\beta_1(U)}$ and $\boldsymbol{\beta_2(U)}$ were set to $0.4 + c\sin(3\Pi\boldsymbol{U})$ and $0.6 + c\boldsymbol{U}(3 - 4\boldsymbol{U})$. As the sample size increased or c increased, the deviation between the alternative hypothesis and the null hypothesis led to a gradual increase in the p-value for the test, converging toward a value of 1.

The results presented in Table 8 demonstrate that under different error distributions and two distinct values of spatial lag coefficients ρ, the resulting probabilities do not exhibit significant differences. This indicates that the bootstrap test method employed exhibits strong test efficacy and robustness, even in the face of variations in the degree of spatial autocorrelation and error distribution.

Table 8. Probability of bootstrap test rejecting H_0 at significance level $\alpha = 0.05$.

C	ρ	N	$N(0,1)$	$U(-4,4)$
0	0.3	30	0.00	0.02
		100	0.02	0.05
	0.5	30	0.00	0.01
		100	0.04	0.04
0.3	0.3	30	0.97	0.97
		100	0.98	1.00
	0.5	30	0.92	0.97
		100	0.98	0.98
0.6	0.3	30	0.95	1.00
		100	0.97	0.98
	0.5	30	0.97	0.96
		100	0.99	0.98

4. Analysis of Spatial Influencing Factors of PM2.5

In order to reflect the superiority and practical value of the SVC-SLM studied in this paper and the INLA method studied in this paper, the SLM, VC-SLM and SVC-SLM models are discussed from three perspectives of parameter, non-parameter and semi-parameter under the background of continuous improvement of urbanization level, the relationship between covariate factors and urbanization level and their impact on PM2.5 concentration.

4.1. Three Model Construction

The variable PM2.5 in Table 1 is selected as the response variable of $\boldsymbol{y}$, and the remaining seven variables of ln_Urban, ln_GDP, ln_Industry, ln_Life, ln_Car, ln_Rain and ln_Environmen are selected as covariables $\boldsymbol{x_j}$, $j = 1, 2, \ldots, 7$ to construct the SLM. See Section 2.1 for specific data preprocessing.

$$\begin{aligned} ln\_PM2.5 = (\boldsymbol{I}_N - \rho\boldsymbol{W})^{-1}(\beta_1 ln\_Urban + \beta_2 ln\_GDP + \beta_3 ln\_Industry \\ + \beta_4 ln\_Life + \beta_5 ln\_Car + \beta_6 ln\_Environment + \beta_7 ln\_Rain + \boldsymbol{\varepsilon}) \end{aligned} \tag{10}$$

where the $\boldsymbol{W}$ is a space weight matrix of ROOK-type in $N \times N$ dimensions, $N = 186$ representing a total of 186 sample data from 31 provinces in mainland China, excluding Taiwan, Hong Kong, and Macau, covering the six years from 2015 to 2020.

Based on Model (10), relaxed the assumption of a linear relationship between covariates and the response variable and introduced variable coefficient terms $\boldsymbol{\alpha_j(U)}$. $\boldsymbol{U}$ represents the urbanization indicator, i.e., ln_Urban is taken as $\boldsymbol{U}$ in the variable coefficient function, and PM2.5 is still taken as response variable $\boldsymbol{y}$; the remaining six variables served as covariates, where $j = 1, 2, \ldots, 6$. A model of the VC-SLM model is constructed to further explore the relationship between various influencing factors and PM2.5 under the background of urbanization. The VC-SLM is denoted as:

$$\begin{aligned} ln\_PM2.5 = (\boldsymbol{I}_N - \rho\boldsymbol{W})^{-1}(\boldsymbol{\alpha_1(U)} ln\_GDP + \boldsymbol{\alpha_2(U)} ln\_Industry \\ + \boldsymbol{\alpha_3(U)} ln\_Life + \alpha_4(\boldsymbol{U}) ln\_Car \\ + \boldsymbol{\alpha_5(U)} ln\_Environment + \boldsymbol{\alpha_6(U)} ln\_Rain + \boldsymbol{\varepsilon}) \end{aligned} \tag{11}$$

The premise of constructing SVC-SLM is to use the INLA estimation method and bootstrap test method to identify whether the Model (11) has constant coefficients. Table 9 conducts a series of tests on different null hypotheses, and the model corresponding to H_1 in this case is Model (11). Correspondingly, the null hypothesis and alternative hypothesis in Table 9 are:

$$\begin{cases} H_0 : \boldsymbol{\alpha_j(U)} = \beta_k, \quad k = p+1, \ldots, 6 \\ H_1 : \text{ all } \boldsymbol{\alpha_j(U)} \text{ changing with } \boldsymbol{U} \end{cases}$$

As mentioned in Section 3.3, the bootstrap test is applicable to such kinds of tests. Similarly, the hypothesis corresponding to Table 9 can be substituted into the bootstrap test process in Appendix B to obtain the corresponding p-value of the trip. $b = 500$ bootstrap samples were extracted in 100 repetitions to calculate the p-value. Table 9 shows the p-value of the test for bootstrap.

Table 9. p-Values for different null hypotheses $\boldsymbol{\alpha_j(U)} = \beta_j$, j = 1, 2,. . . , 6 and their bootstrap tests at significance level $\alpha = 0.05$.

Null Hypothesis	Covariates	p-Value
$\boldsymbol{\alpha_1(U)} = \beta_1$	ln_GDP	0.35
$\boldsymbol{\alpha_2(U)} = \beta_2$	ln_Industry	0.81
$\boldsymbol{\alpha_3(U)} = \beta_3$	ln_Life	0.00
$\boldsymbol{\alpha_4(U)} = \beta_4$	ln_Car	0.00
$\boldsymbol{\alpha_5(U)} = \beta_5$	ln_Environment	0.00
$\boldsymbol{\alpha_6(U)} = \beta_6$	ln_Rain	0.85
$\boldsymbol{\alpha_j(U)} = \beta_j, j = 1, 2, 6$	ln_GDP, ln_Industry, ln_Rain	1.00

As can be seen from Table 9, the p-values of the bootstrap test of ln_GDP, ln_Industry and ln_Rain are all significantly greater than the significance level $\alpha = 0.05$, so the null hypothesis of these three covariables cannot be rejected, which indicates that their regression coefficient is a constant and is not affected by covariable $\boldsymbol{U}$. In addition, the results in the last row of Table 9 show that the p-value of the bootstrap test for adding ln_Life,

ln_Car and ln_Environment to the null hypothesis is also greater than the significance level $\alpha = 0.05$, which indicates that the hypothesis accepts the null hypothesis. That is the three regression coefficients (ln_Life, ln_Car, and ln_Environment) all change with the change in the covariable $\boldsymbol{U}$.

Finally, an additional hypothesis is further verified, that is, whether the three remaining regression coefficients ln_Life, ln_Car and ln_Environment change with the change of covariable $\boldsymbol{U}$ at the same time. In this case, the alternative hypothesis changes to the last row of Table 9. The null hypothesis is changed to add one or more of the remaining coefficients of $\boldsymbol{\alpha_3(U)}$, $\boldsymbol{\alpha_4(U)}$, and $\boldsymbol{\alpha_5(U)}$ to the null hypothesis in the last row of Table 9. The p-value of the results is between 0.00 and 0.03, indicating that the coefficients $\boldsymbol{\alpha_3(U)}$, $\boldsymbol{\alpha_4(U)}$, $\boldsymbol{\alpha_5(U)}$ do change simultaneously with the covariable $\boldsymbol{U}$.

So the SVC-SLM is:

$$\begin{aligned} ln\_PM2.5 = (\boldsymbol{I}_N - \rho\boldsymbol{W})^{-1}(&\boldsymbol{\alpha_3(U)}ln\_Life + \boldsymbol{\alpha_4(U)}ln\_Car \\ &+ \boldsymbol{\alpha_5(U)}ln\_Environment + \beta_1 ln\_GDP \\ &+ \beta_2 ln\_Industry + \beta_6 ln\_Rain + \boldsymbol{\varepsilon}) \end{aligned} \tag{12}$$

4.2. Model Selection

Since DIC [32] is often used to evaluate the fitting effect and model complexity of different Bayesian models, and the widely applicable information criterion (WAIC) can approach the results of Bayesian cross-validation without being affected by parameterization, therefore, in this study, DIC and WAIC were used to evaluate the relative advantages and disadvantages of the three models.

The smaller the DIC and WAIC values, the better the prediction ability and generalization ability of the model, that is, the model has a better fitting performance on the unseen data. Table 10 shows the fitting DIC and WAIC values of the three models.

Table 10. The fitting evaluation results of the three models.

Variable	SLM	VC-SLM	SVC-SLM
DIC	−1082.24	−1058.70	−1114.87
WAIC	−1106.40	−1099.96	−1127.99

As shown in Table 10, the DIC and WAIC values of SVC-SLM are the smallest, which is smaller than models of SLM and VC-SLM. This verifies the correctness that SVC-SLM is the best model to fit the data in this paper, even though the values of DIC and WAIC for the SLM are smaller than for the VC-SLM, indicating that it is inappropriate to use VC-SLM to improve on the basis of SLM. The applicability of SVC-SLM in this paper is further explained.

4.3. Analysis of Influencing Factors of PM2.5 by Three Models

The INLA algorithm is used to fit the Models (10)–(12). The obtained estimates of spatial autocorrelation parameters and the posterior estimates of regression coefficients of each influencing factor are shown in Table 11. The autocorrelation charts of the three models are shown in Figure 4, the regression coefficients of each influencing factor of Model (11) are shown in Figure 5, and the regression coefficients of some influencing factors of Model (12) is shown in Figure 6.

According to Figure 4 and Table 11, the spatial autocorrelation of PM2.5 in SLM, VC-SLM and SVC-SLM models are all positive. They are 0.8008 (95%CI: 0.6610, 0.9050), 0.6858 (95%CI: 0.5941, 0.7678) and 0.5499 (95%CI: 0.4041, 0.6781), respectively, indicating that there is a significant spatial positive correlation between PM2.5 concentrations in the three model provinces. To a certain extent, PM2.5 in one province will affect the transmission of PM2.5 concentrations in the surrounding provinces. Further investigation shows that the spatial autocorrelation parameter of ρ in the model (range from 0.8008 to 0.5499) and (range

from 0.6858 to 0.5499) gradually decrease as SLM to SVC-SLM or VC-SLM to SVC-SLM. This phenomenon may be attributed to the fact that in the context of the semi-parametric model of SVC-SLM avoids assuming simple linear relationships between variables (as in SLM) and avoids making arbitrary non-parametric assumptions about all parameters (as in VC-SLM), instead adopting a more balanced approach that helps prevent overestimation. The spatial autocorrelation parameter of ρ in SVC-SLM can be estimated more accurately.

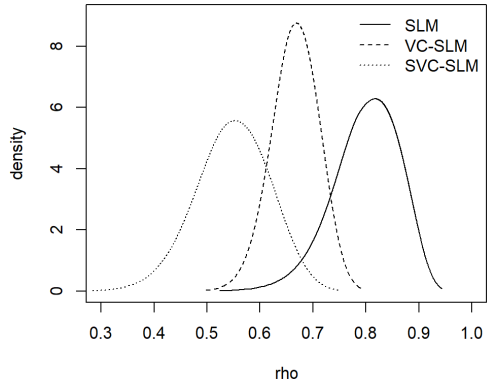

Figure 4. Density plots of autocorrelation coefficients of the three models.

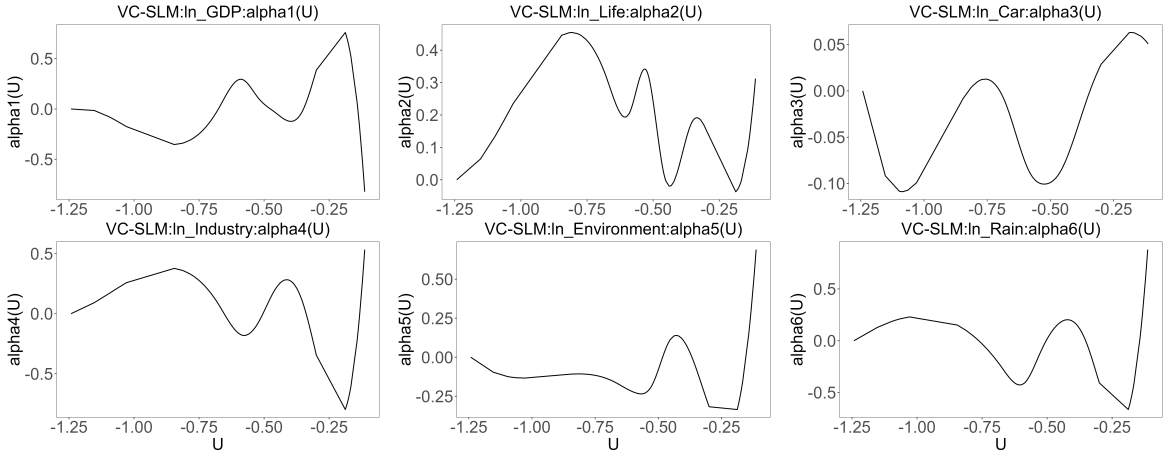

Figure 5. Figure of regression variable coefficient $\boldsymbol{\alpha_j(U)}, j = 1, 2, ..., 6$ of VC-SLM.

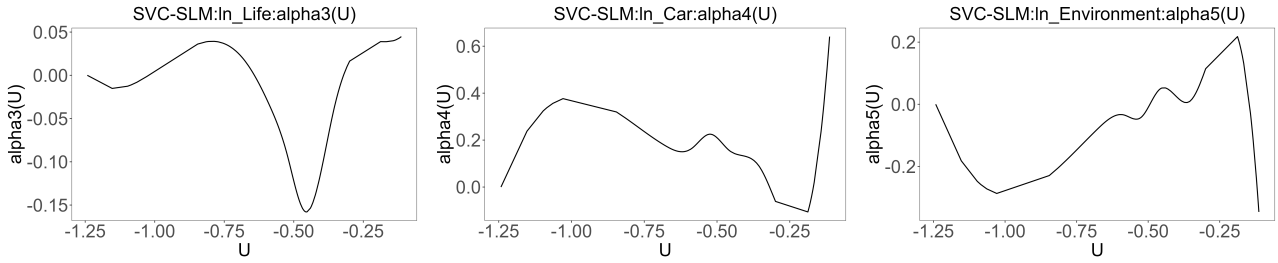

Figure 6. Figure of the regression variable coefficient $\boldsymbol{\alpha_j(U)}, j = 3, 4, 5$ of SVC-SLM.

The corresponding results of VC-SLM and SVC-SLM in Table 11 are Models (11) and (12), respectively. The values of all variable coefficients of VC-SLM and part of the variable coefficients of SVC-SLM in Table 11 are calculated to obtain the mean value.

Table 11. Posterior estimation of covariate regression coefficients under three models.

Variable	SLM	VC-SLM	SVC-SLM
ρ	0.8008 (0.6610, 0.9050)	0.6858 (0.5941, 0.7678)	0.5499 (0.4041, 0.6781)
ln_Urban	0.1036 (−0.1643, 0.3731)	–	–
ln_GDP	0.2574 (0.1604, 0.3551)	0.0495 (−0.5375, 0.6361)	0.0241 (−0.0438, 0.0917)
ln_Industry	0.1320 (0.0767, 0.1872)	0.2085 (−0.1317, 0.5484)	0.1578 (0.09561, 0.2196)
ln_Life	−0.0202 (−0.0569, 0.0163)	−0.0488 (−0.2755, 0.1777)	−0.0445 (−0.2238, 0.1347)
ln_Car	−0.2022 (−0.3309, −0.0731)	0.0244 (−0.7439, 0.7881)	0.1821 (−0.1413, 0.5052)
ln_Environment	−0.1956 (−0.2671, −0.1230)	−0.0807 (−0.5116, 0.3499)	−0.0378(−0.2655, 0.1896)
ln_Rain	−0.2887 (−0.3927, −0.1851)	−0.1257 (−0.7930, 0.5410)	−0.1161 (−0.1933, −0.0390)

4.3.1. The Linear Influencing Factors of PM2.5 under SLM

The results of SLM in Table 11 show that the 95% confidence interval of the linear coefficient of ln_Urban and ln_Life both contain 0, indicating that the influence of these two factors on PM2.5 concentration is not statistically significant. However, it is worth noting that ln_Life has a significant negative impact on PM2.5 concentration, which contradicts common sense and the findings of Li et al. (2023) [33]. This uncertainty is to be studied in the model later in this paper. ln_GDP and ln_Industry have significant positive effects on regional PM2.5 concentration, with a posterior mean of 0.2574 and 0.1320, respectively. Among them, ln_GDP has the strongest promoting effect on PM2.5 concentration. This may be due to the fact that regions with high GDP tend to have more industrial production and economic activity, as well as higher levels of car ownership and usage, leading to increased emissions and particulate matter production, thereby raising PM2.5 concentrations. Both ln_Environment and ln_Rain have negative effects on PM2.5, indicating that Environmental treatment and rainfall have the same significant impact on air pollution.

4.3.2. The Influencing Factors of PM2.5 of VC-SLM in the Background of Urbanization

Similarly, the INLA algorithm was used to estimate the Model (11). Surprisingly, when all the influencing factors are affected by urbanization, ln_Life has a promoting effect on PM2.5 concentration, which is consistent with the findings of common sense and Li (2023) [33]. This suggests that previous simple linear assumptions about the effect of ln_Life on PM2.5 concentrations may not be appropriate. In addition, under the scenario that all influencing factors are affected by urbanization, ln_Industry has replaced ln_GDP as the most significant positive influencing factor for provincial PM2.5 concentration, while ln_Rain is still the strongest inhibiting factor for provincial PM2.5 concentration. According to Figure 5, it is possible to further observe the changing trend of PM2.5 concentration and all factors under the scenario of intensified inter-provincial urbanization in China. The specific analysis of each factor of the VC-SLM is not conducted here. The influence of each factor of the SVC-SLM on PM2.5 concentration will be introduced in detail next.

4.3.3. The Influencing Factors of PM2.5 of SVC-SLM in the Background of Urbanization

As can be seen from the bootstrap test results in Table 9, the regression coefficients $\boldsymbol{\alpha_3(U)}$,$\boldsymbol{\alpha_4(U)}$,$\boldsymbol{\alpha_5(U)}$ of the three covariables ln_Life, ln_Car and ln_Environment in Model (12) change with the change of covariable $\boldsymbol{U}$. It should be emphasized that in Model (10), ln_Car and ln_Environment have significant impacts on PM2.5. However, bootstrap test results show that these two variables should be regarded as non-parametric components. This finding confirms that the previous method of directly selecting non-significant variables (that is, coefficient confidence intervals containing 0) as non-parametric components from "linear SLM" is too subjective and slightly lacking in theoretical basis. This emphasizes the importance and necessity of bootstrap testing of constant coefficients in semi-parametric models.

According to the results in the last column of Table 9, SVC-SLM regression results are written as follows:

$$\begin{aligned} ln\_PM2.5 = (\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}(&0.0241 ln\_GDP + 0.1578 ln\_Industry \\ &- 0.0445 ln\_Life + 0.1821 ln\_Car \\ &- 0.0378 ln\_Environment - 0.1161 ln\_Rain + \boldsymbol{\varepsilon}) \end{aligned} \tag{13}$$

The Model (13) has three linear influences on the concentration of PM2.5, they are ln_GDP, ln_Industry and ln_Rain. ln_Rain is still the factor that exerts the greatest inhibitory effect on air pollution concentration, with a posterior mean of −0.1161. ln_Industry shows a significant linear contribution to the PM2.5. The 95% confidence interval for ln_GDP contains 0, indicating that GDP has no effect on PM2.5.

The Model (13) of non-linear effects on PM2.5 concentration are ln_Life, ln_Car and ln_Environment. Figure 6 shows the regression coefficient changes of the three influencing factors of the Model (12) with increasing urbanization.

As shown in Figure 6, the influence of three non-linear influencing factors on PM2.5 is a curve of change, which once again confirms that there is no direct linear relationship between these three influencing factors and PM2.5, as confirmed by the bootstrap test. Next, there will be an in-depth study on the changes in PM2.5 concentration and these three factors under the scenario of increasing urbanization at the provincial level in China.

In the analysis of ln_Life in the early stage of urbanization, ln_Life had a mild inhibitory effect on PM2.5, which may be related to the relatively low exhaust gas emissions in the early stage of urbanization development. This weak inhibitory effect may be attributed to increased environmental awareness and the initial implementation of policies.

As for ln_Car, the overall impact pattern of ln_Car on PM2.5 concentration is similar to "N-shaped". In other words, with the advancement of urbanization, the impact of life and motor vehicle emissions on PM2.5 concentration is as follows: promoting in the initial stage, inhibiting in the middle stage, and transiting from inhibiting to positive in the later stage. Here, is a possible explanation: (1) In the initial stage, with the advancement of urbanization, the urban population grows and economic activities expand. This has led to an increase in the number of motor vehicles and industrial output, resulting in increased emissions of motor vehicle exhaust and industrial exhaust gases. Therefore, this contributes to the increase in PM2.5 concentration, which has a positive impact. This is consistent with the results of Chen et al. (2023) [34] that industrial, urban, and automobile exhaust gases play a role in promoting PM 2.5. (2) In the middle stage of urbanization, social concern about environmental pollution may, over time, prompt the government to take measures to control motor vehicle exhaust and industrial exhaust emissions. These measures could include restricting vehicle access, promoting the use of clean energy and enforcing stricter emissions standards. (3) In the later period, economic growth and energy demand are expected to increase as urbanization continues. The ability to implement policies may also fluctuate. Therefore, the inhibiting effect of motor vehicle and industrial emissions on PM2.5 concentration gradually wanes and eventually becomes a promoting effect.

In the analysis of ln_Environment, the chart of the impact of ln_Environment on PM2.5 concentration seems to be an inverted "N-shape", which is opposite to the chart of ln_Car in a certain sense, which is in line with practical significance. As Figure 6 shows, in the initial phase of increased urbanization, the effects of investment in environmental governance accumulate over time. This gradual accumulation enhances their suppressive effect on pollutant emissions, thus restraining the rise in PM2.5 concentrations until they peak. However, as the impact of governance investment becomes apparent, it will eventually approach a certain limit. At this critical juncture, even if investments in environmental governance persist, their lasting impact is limited because governance measures have reached a certain level of effectiveness. In addition, technological advances and policy adjustments may introduce new governance effects at specific stages, making the inhibitory effect weaker. In the process of advancing urbanization, population and economic growth will lead to

the emergence of new sources of pollution. These new sources may weaken the inhibitory effect or even reverse it to a boosting effect. However, with the continued development of urbanization, the government has sought to better control pollution and improve air quality by increasing investment in environmental governance. Such ongoing environmental governance will continue to have an inhibitory effect on PM2.5 concentrations [35].

5. Result

In this paper, the INLA algorithm is used for the first time to estimate and empirically analyze the SVC-SLM model. The specific conclusions are as follows:

(1) In simulation experiments, a comparison with the MCMCINLA algorithm reveals that the INLA algorithm not only outperforms in terms of parameter fitting accuracy but also has a faster computational speed, fully demonstrating the feasibility of the INLA algorithm.

(2) In the empirical analysis of the influencing factors of PM2.5 concentration, this study comprehensively analyzes the three models of SLM, VC-SLM, and SVC-SLM from three perspectives: parametric, non-parametric, and semi-parametric. In terms of model selection, the SVC-SLM exhibits the smallest values of DIC and WAIC, highlighting its applicability in this study.

(3) In Section 4.3.3 of the empirical study, directly selecting non-significant variables (that is, coefficient confidence intervals containing 0) as non-parametric components from linear SLM models is too subjective and slightly lacking in theoretical basis. This emphasizes the importance of the SVC-SLM from a semi-parameter perspective, utilizing bootstrap tests to determine the significance and necessity of which regression coefficients are constant and which are variable. Therefore, employing bootstrap tests in this paper is deemed rational and prudent.

(4) Empirical research results indicate that rainfall has become the most crucial factor in reducing PM2.5 concentration.

(5) The results of the empirical study show that when the linear SLM transitions to the SVC-SLM, i.e., in the context of accelerating urbanization, the inhibitory effect of investment in environmental governance on PM2.5 concentrations diminishes. The posterior mean value moves from -0.1956 to -0.0378. On the contrary, the effect of motor vehicle tailpipe emissions on PM2.5 concentrations undergoes a transition from inhibition to facilitation. The a posteriori mean value moved from -0.2022 to 0.1821.

In summary, the results verify the applicability of the INLA algorithm and SVC-SLM model from both theoretical and empirical aspects, which highlights the significance of this study.

6. Discussion

(1) The empirical research findings offer valuable insights for relevant government departments to strengthen policies in areas such as environmental governance investment, changes in motor vehicle emissions, and rainfall suppression. For instance, in controlling automotive exhaust emissions, it is advisable to consider implementing stricter emission standards, requiring new and existing vehicles to reduce emission levels. Exploring incentives or mandates for the use of electric or hydrogen fuel cell vehicles, along with improving the accessibility and efficiency of public transportation systems, can contribute to reducing individual car usage. In terms of environmental protection investment, increasing government funding for air pollution control and encouraging private sector participation in environmental projects, especially in clean energy initiatives, could be an effective strategy. Additionally, given the impact of rainfall and wind direction on PM2.5 concentration, addressing meteorological conditions is crucial for maintaining air quality. Developing early warning systems and emergency measures related to meteorology would be beneficial.

(2) For this study, there are still several areas that could be further explored. Firstly, the part of the variable coefficient term studied in this paper is $\boldsymbol{\alpha}_j(\boldsymbol{U})$, which reflects

the dynamic variations of $\boldsymbol{y}$ concerning a covariable $\boldsymbol{U}$. Future research could consider partial coefficient terms as, reflecting the dynamic variations of $\boldsymbol{\alpha}_j(\boldsymbol{T})$ concerning time $\boldsymbol{T}$. Secondly, this paper focuses on longitudinal data analysis, and future research could extend it to spatial panel data, adjusting the coefficient terms to explore deeper insights into the changes in $\boldsymbol{\alpha}_j(\boldsymbol{u}, \boldsymbol{v})$ based on spatial geographical locations.

Author Contributions: Conceptualization, Q.P. and X.H.; methodology, Q.P. and X.H.; data curation, Q.P.; resources, Q.P.; software, Q.P.; validation, Q.P. and X.H.; formal analysis, Q.P.; investigation, Q.P.; writing—original draft preparation, Q.P.; writing—review and editing, Q.P. and X.H.; visualization, X.H.; supervision, X.H.; project administration, X.H.; funding acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by the National Natural Science Foundation of China, grant number 11961065.

Data Availability Statement: All the data used in this paper can be obtained from the Columbia University Center for Social and Economic Research and Data (https://sedac.ciesin.columbia.edu/, accessed on 31 December 2020), the China Statistical Yearbook (http://www.stats.gov.cn, accessed on 31 December 2020), the China Environmental Statistical Yearbook (https://data.cnki.net/yearBook/single?id=N2023070120, accessed on 31 December 2020) and the European Union and European Centre for Medium-Range Weather Forecasts (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land-monthly-means?tab=overview, accessed on 31 December 2020).

Acknowledgments: We are grateful to the editor, associated editor, and three referees for their valuable suggestions and comments that greatly improved the article.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Appendix A

Appendix A is proof of the GMRF structure of the SVC-SLM. The SVC-SLM in this paper is:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{\alpha}_j(\boldsymbol{U}) + \sum_{k=p+1}^{n} \boldsymbol{x}_k \beta_k + \boldsymbol{\varepsilon} \tag{A1}$$

Model (A1) is written as a new potential SLM model as:

$$\begin{cases} \boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\zeta} \\ \boldsymbol{\eta} = (\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}(\boldsymbol{V}\boldsymbol{\psi} + \boldsymbol{\varepsilon}) \end{cases} \tag{A2}$$

Since inside R-INLA is the joint distribution of $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$, which is $[\boldsymbol{\eta}, \boldsymbol{\psi}]$. So the goal is to prove $[\boldsymbol{\eta}, \boldsymbol{\psi}]$ is the GMRF as a sparse precision matrix to make it conform to the INLA framework. According to Bayes' theorem, we first calculate the conditional distribution of $\boldsymbol{\eta}$ over $\boldsymbol{\psi}$ as $[\boldsymbol{\eta}|\boldsymbol{\psi}]$.

Assuming that the joint distribution is Gaussian, and therefore, the conditional distribution $[\boldsymbol{\eta}|\boldsymbol{\psi}]$ is also Gaussian, the expectation and variance of $[\boldsymbol{\eta}|\boldsymbol{\psi}]$ are expressed as follows:

$$\boldsymbol{E} = E(\boldsymbol{\eta} \mid \boldsymbol{\psi}) = (\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}\boldsymbol{V}\boldsymbol{\psi}$$

and

$$\begin{aligned} \boldsymbol{D} &= \mathrm{var}(\boldsymbol{\eta} \mid \boldsymbol{\psi}) = \mathrm{var}((\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}\boldsymbol{V}\boldsymbol{\psi} + (\boldsymbol{I}_N - \rho W)^{-1}\boldsymbol{\varepsilon} \mid \boldsymbol{\psi}) \\ &= (\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}\frac{1}{\tau}\boldsymbol{I}_N((\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1})' \\ &= \frac{1}{\tau}(\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1}((\boldsymbol{I}_N - \rho \boldsymbol{W})^{-1})' \end{aligned}$$

The precision of $[\boldsymbol{\eta}|\boldsymbol{\psi}]$ can be expressed as $\boldsymbol{P}$:

$$\boldsymbol{P} = prec[\boldsymbol{\eta} \mid \boldsymbol{\psi}] = \frac{1}{\boldsymbol{D}} = \tau(\boldsymbol{I}_N - \rho \boldsymbol{W}')(\boldsymbol{I}_N - \rho \boldsymbol{W})$$

The precision matrix $\boldsymbol{P}$ is symmetric and sparse, so that the joint distribution of $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$ is

$$\begin{aligned}[\boldsymbol{\eta}, \boldsymbol{\psi}] = [\boldsymbol{\eta}|\boldsymbol{\psi}][\boldsymbol{\psi}] &\propto \exp\{-\frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{E})'\boldsymbol{P}(\boldsymbol{\eta} - \boldsymbol{E})\} \exp\{-\frac{1}{2}\boldsymbol{\psi}'\boldsymbol{Q}\boldsymbol{\psi}\} \\ &= \exp\{-\frac{1}{2}(\boldsymbol{\eta}'\boldsymbol{P}\boldsymbol{\eta} - \boldsymbol{\eta}'\boldsymbol{P}\boldsymbol{E} - \boldsymbol{E}'\boldsymbol{P}\boldsymbol{\eta} + \boldsymbol{E}'\boldsymbol{P}\boldsymbol{E} + \boldsymbol{\psi}'\boldsymbol{Q}\boldsymbol{\psi})\} \\ &= \exp\{-\frac{1}{2}(\boldsymbol{\eta}, \boldsymbol{\psi})'\boldsymbol{K}(\boldsymbol{\eta}, \boldsymbol{\psi})\}\end{aligned}$$

where $\boldsymbol{K}$ is the precision matrix, and its matrix structure is

$$\begin{aligned}\boldsymbol{K} &= \begin{pmatrix} \boldsymbol{P} & -\boldsymbol{P}(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{V} \\ -\boldsymbol{V}'(\boldsymbol{I}_N - \rho\boldsymbol{W})^{-1}\boldsymbol{P} & \boldsymbol{Q} + \tau\boldsymbol{V}'\boldsymbol{V} \end{pmatrix} \\ &= \begin{pmatrix} \tau(\boldsymbol{I}_N - \rho\boldsymbol{W}')(\boldsymbol{I}_N - \rho\boldsymbol{W}) & -\tau(\boldsymbol{I}_N - \rho\boldsymbol{W}')\boldsymbol{V} \\ -\tau\boldsymbol{V}'(\boldsymbol{I}_N - \rho\boldsymbol{W}) & \boldsymbol{Q} + \tau\boldsymbol{V}'\boldsymbol{V} \end{pmatrix}\end{aligned} \tag{A3}$$

Note, that in order to obtain the result of Model (A3), the following formula was used:

$$\begin{aligned}\boldsymbol{\eta}\boldsymbol{P}\boldsymbol{E} &= \boldsymbol{\eta}'\tau(\boldsymbol{I}_N - \rho\boldsymbol{W}')(\boldsymbol{I}_N - \rho\boldsymbol{W})(\boldsymbol{I}_N - \rho\boldsymbol{W})^{-1}\boldsymbol{V}\boldsymbol{\psi} \\ &= \tau\boldsymbol{\eta}'(\boldsymbol{I}_N - \rho\boldsymbol{W}')\boldsymbol{V}\boldsymbol{\psi}\end{aligned}$$

$$\boldsymbol{E}'\boldsymbol{P}\boldsymbol{\eta} = (\boldsymbol{\eta}'\boldsymbol{P}\boldsymbol{E})' = \tau\boldsymbol{\psi}'\boldsymbol{V}'(\boldsymbol{I}_N - \rho\boldsymbol{W})\boldsymbol{\eta}$$

and

$$\begin{aligned}\boldsymbol{E}'\boldsymbol{P}\boldsymbol{E} &= \tau\boldsymbol{\psi}'\boldsymbol{V}'(\boldsymbol{I}_N - \rho\boldsymbol{W}')^{-1}(\boldsymbol{I}_N - \rho\boldsymbol{W}')(\boldsymbol{I}_N - \rho\boldsymbol{W})(\boldsymbol{I}_N - \rho\boldsymbol{W})^{-1}\boldsymbol{V}\boldsymbol{\psi} \\ &= \tau\boldsymbol{\psi}'\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\psi}\end{aligned}$$

So that,

$$E(\boldsymbol{\eta}, \boldsymbol{\psi}) = 0 \tag{A4}$$

Therefore, $(\boldsymbol{\eta}, \boldsymbol{\psi})$ is a GMRF with a mean of 0 and a precision matrix of $\boldsymbol{K}$, where matrix $\boldsymbol{K}$ is a highly sparse block matrix, so Model (A1) conforms to the INLA framework.

Appendix B

As mentioned in Section 2.2.4, for the SVC-SLM, the question of common interest is whether some of the coefficients can be regarded as constants, hence the following bootstrap hypothesis:

$$\begin{cases} H_0 : \boldsymbol{\alpha}_j(\boldsymbol{U}) = \beta_k, k = p+1, \ldots, n \\ H_1 : \text{ all } \boldsymbol{\alpha}_j(\boldsymbol{U})(j = 1, \ldots, n) \text{ changing with } \boldsymbol{U} \end{cases}$$

For this hypothesis, the corresponding model under the original hypothesis H_0 is SVC-SLM. The specific form is:

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \sum_{j=1}^{p} \boldsymbol{x}_j\boldsymbol{\alpha}_j(\boldsymbol{U}) + \sum_{k=p+1}^{n} \boldsymbol{x}_k\beta_k + \boldsymbol{\varepsilon} \tag{A5}$$

Alternative hypothesis H_1 corresponds to VC-SLM:

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \sum_{j=1}^{n} \boldsymbol{x}_j\boldsymbol{\alpha}_j(\boldsymbol{U}) + \boldsymbol{\varepsilon} \tag{A6}$$

The detailed bootstrap procedure is as follows:

1. Using the initial values $\{y_i; x_{ij}, x_{ik}; U_i\}(i = 1, 2, \ldots, N; j = 1, 2, \ldots, p; k = p+1, \ldots, n; p \leq n)$ and the determined window width based on H_1. Fitting the VC-SLM model under the alternative assumption H_1, i.e., estimating Model (A6) using the INLA algorithm, yields estimated parameters that are the lag coefficient $\tilde{\rho}$ and variable coefficients of $\tilde{\boldsymbol{\alpha}}_j(\boldsymbol{U})$, resulting in the residual vector $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_N)'$, and compute the sum of squared residuals RSS_{H_1} ;

2. Based on the original data and the window width determined in step 1, fit the variable coefficient model under the null hypothesis H_0 . With INLA estimating the Model (A5) to obtain the lag coefficient $\hat{\rho}$, the variable coefficient of $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{U})$ and constant-coefficient $\hat{\beta}_k$ are calculated. Then, calculate the sum of squared residuals RSS_{H_0} and substitute it into Model (7) to obtain the observed value T for this statistic.
3. Centering the residual vector $\tilde{\boldsymbol{\varepsilon}}$ obtained in step 1 to obtain $\tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{c}} = (\tilde{\varepsilon}_{1c}, \tilde{\varepsilon}_{2c}, ..., \tilde{\varepsilon}_{Nc})'$, where $\tilde{\varepsilon}_{ic} = \tilde{\varepsilon}_i - \frac{1}{N}\sum_{i=1}^{N}\tilde{\varepsilon}_i, i = 1, 2, ..., N$;
4. Sampling back from the centered residual vector to obtain a new residual $\boldsymbol{\varepsilon}^* = (\varepsilon_1{}^*, \varepsilon_2{}^*, \ldots, \varepsilon_N{}^*)'$, generating a new observation $\boldsymbol{y}^*$ of the response variable, specifically, $\boldsymbol{y}^* = \hat{\rho}\boldsymbol{W}\boldsymbol{y}^* + \sum_{j=1}^{p}\boldsymbol{x}_j\hat{\boldsymbol{\alpha}}_j(\boldsymbol{U}) + \sum_{k=p+1}^{n}\boldsymbol{x}_k\hat{\beta}_k + \varepsilon^*$, and thus generating bootstrap data $\{y_i*; x_{ij}, x_{ik}; U_i\}(i = 1, 2, \ldots, N; j = 1, 2, \ldots, p; k = p+1, \ldots, n; p \leq n)$;
5. Utilizing the generated bootstrap data, refit the models corresponding to the null and alternative hypotheses, calculating the sum of squared residuals $RSS_{H_0}{}^*$ and $RSS_{H_1}{}^*$ at this point. Substitute these values into Model (7) to compute the bootstrap observed value T^* for this statistic;
6. Repeat step 4 and step 5 a total of K times to obtain the b bootstrap observation of the test statistic T, which we denote by $t_1^*, t_2^*, \ldots, t_K^*$, Then the p-value estimate of the bootstrap test, i.e., the p-value in Model (8) is estimated as:

$$\tilde{p} = \frac{1}{K}\sum_{i=1}^{K} I(t_i^* \geq t) \tag{A7}$$

where $I(\bullet)$ is the schematic function, and t is the observed value of the statistic T.

References

1. Robinson, P.M. Root-N-consistent semiparametric regression. *Econom. J. Econom. Soc.* **1988**, *56*, 931–954. [CrossRef]
2. Cliff, A.D.; Ord, K. Spatial autocorrelation: A review of existing and new measures with applications. *Econ. Geogr.* **1970**, *46* (Suppl. 1), 269–292. [CrossRef]
3. Li, K.M.; Chen, J.B. Generalized moment estimation for non-parametric spatial lag models. *J. Appl. Math. Univ. Ser. A* **2018**, *33*, 140–156. (In Chinese)
4. Teng, J.; Ding, S.; Zhang, H.; Hu, X. MCMCINLA estimation of varying coefficient spatial lag model—A study of China's economic development in the context of population aging. *PLoS ONE* **2023**, *18*, e0279504. [CrossRef]
5. Su, L.; Jin, S.; Zhang, H.; Hu, X. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *J. Econom.* **2010**, *157*, 18–33. [CrossRef]
6. Su, L. Semiparametric GMM estimation of spatial autoregressive models. *J. Econom.* **2012**, *167*, 543–560. [CrossRef]
7. Li, K.M.; Chen, J.B. Maximum likelihood estimation of cross-section for space hysteresis models with variable coefficients of semi-parameter. *Quant. Econ. Tech. Econ. Res.* **2013**, *30*, 85–98. (In Chinese)
8. Hoshino, T. Semiparametric spatial autoregressive models with endogenous regressors: With an application to crime data. *J. Bus. Econ. Stat.* **2018**, *36*, 160–172. [CrossRef]
9. Xiao, Z.; Linton, O.B.; Carroll, R.J. More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *J. Am. Stat. Assoc.* **2003**, *98*, 980–992. [CrossRef]
10. Eilers, P.H.C.; Marx, B.D. Flexible smoothing with B-splines and penaltie. *Stat. Sci.* **1996**, *11*, 89–121. [CrossRef]
11. Gu, J.; Ye, A.Z.; Chen, H. Research on regional agglomeration effect of technological innovation capability based on semi-parametric spatial econometric model. *Sci. Manag. Sci. Technol.* **2012**, *33*, 62–70. (In Chinese)
12. Li, T.; Mei, C. Statistical inference on the parametric component in partially linear spatial autoregressive models. *Commun. Stat.-Simul. Comput.* **2016**, *45*, 1991–2006. [CrossRef]
13. Du, Y.; Li, T.Z. Bootstrap test of variable coefficient spatial autoregressive model. *J. Eng. Math.* **2021**, *38*, 539–552. (In Chinese)
14. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2009**, *71*, 319–392. [CrossRef]
15. Xia, Z.Y.; Tang, B.; Qin, L.; Zhang, H.G.; Hu, X.J. Spatially Dependent Bayesian Modeling of Geostatistics Data and Its Application for Tuberculosis (TB) in China. *Mathematics* **2023**, *11*, 4193. [CrossRef]
16. Lin, J.; Teng, J.Q.; Hu, X.J. Space-time response Analysis and path Optimization of tourism Complaints: Based on Bayesian INLA and Graph Theory. *Mod. Commer. Ind.* **2022**, *43*, 2. (In Chinese)
17. Sun, X.L.; Wang, H.L. Spatiotemporal modelling of soil organic matter changes in Jiangsu, China between 1980 and 2006 using INLA-SPDE. *Geoderma* **2021**, *384*, 114808. [CrossRef]
18. Bivand, R.S.; Gómez-Rubio, V.; Rue, H. Approximate Bayesian inference for spatial econometrics models. *Spat. Stat.* **2014**, *9*, 146–165. [CrossRef]

19. Gómez-Rubio, V.; Bivand, R.S.; Rue, H. Estimating spatial econometrics models with integrated nested Laplace approximation. *Mathematics* **2021**, *9*, 2044. [CrossRef]
20. Liu, H.; Cui, W.; Zhang, M. Exploring the causal relationship between urbanization and air pollution: Evidence from China. *Sustain. Cities Soc.* **2022**, *80*, 103783. [CrossRef]
21. Yang, Y.; Lan, H.; Li, J. Spatial econometric analysis of the impact of socioeconomic factors on PM2. 5 concentration in China's inland cities: A case study from Chengdu Plain Economic Zone. *Int. J. Environ. Res. Public Health* **2020**, *17*, 74. [CrossRef]
22. Chou, Y.; Huang, D. The impact of urbanization level on haze pollution: Based on cities at prefecture level and above in China. *J. Hunan Univ. Sci. Technol.* **2020**, *23*, 59–69.
23. Lai, I.; Maji, S.; Alam, M.M. Impact of Meteorological Conditions on PM2.5 Concentration in Delhi. In Proceedings of the 2022 International Interdisciplinary Conference on Mathematics, Engineering and Science (MESIICON), Durgapur, India, 11–12 November 2022; pp. 1–5.
24. Wang, H.; Li, J. Environmental Science and Pollution Research. *Environ. Sci. Pollut. Res.* **2021**, *28*, 47213–47226. [CrossRef]
25. Gao, J.; Wang, K.; Wang, Y. Temporal-spatial characteristics and source apportionment of PM2. 5 as well as its associated chemical species in the Beijing-Tianjin-Hebei region of China. *Environ. Pollut.* **2018**, *233*, 714–724. [CrossRef]
26. Wang, Z.; Liang, L.; Sun, Z. Spatiotemporal differentiation and the factors influencing urbanization and ecological environment synergistic effects within the Beijing-Tianjin-Hebei urban agglomeration. *J. Environ. Manag.* **2019**, *243*, 227–239. [CrossRef] [PubMed]
27. Guerry, A.M. *A Translation of Andre-Michel Guerry's Essay on the Moral Statistics of France (1883): A Sociological Report to the French Academy of Science*; Edwin Mellen Press: Lewiston, NY, USA, 2002.
28. Rue, H.; Held, L. *Gaussian Markov Random Fields: Theory and Applications*; CRC Press: Boca Raton, FL, USA, 2005.
29. Fan, J.; Huang, T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **2005**, *11*, 1031–1057. [CrossRef]
30. Mei, C.L.; Chen, F. Detection of spatial heterogeneity based on spatial autoregressive varying coefficient models. *Spat. Stat.* **2022**, *51*, 100666. [CrossRef]
31. Stakhovych, S.; Bijmolt, T.H.A. Specification of spatial models: A simulation study on weights matrices. *Pap. Reg. Sci.* **2009**, *88*, 389–408. [CrossRef]
32. Raftery, A.E. Bayesian model selection in social research. *Sociol. Methodol.* **1995**, *25*, 111–163. [CrossRef]
33. Li, Y.; Xue, L.; Tao, Y. Exploring the contributions of major emission sources to PM2. 5 and attributable health burdens in China. *Environ. Pollut.* **2023**, *322*, 121177. [CrossRef] [PubMed]
34. Chen, H.; Yan, Y.; Hu, D. High contribution of vehicular exhaust and coal combustion to PM2. 5-bound Pb pollution in an industrial city in North China: An insight from isotope. *Atmos. Environ.* **2023**, *294*, 119503. [CrossRef]
35. Shi, K.; Shen, J.; Wang, L. A multiscale analysis of the effect of urban expansion on PM2. 5 concentrations in China: Evidence from multisource remote sensing and statistical data. *Build. Environ.* **2020**, *174*, 106778. [CrossRef]