A PENALIZED H-LIKELIHOOD METHOD FOR GAUSSIAN SPATIAL ADDITIVE MODEL ON REGULAR LATTICE

Hao Sun

Follow this and additional works at: https://lib.dr.iastate.edu/creativecomponents

Part of the Environmental Studies Commons, and the Statistical Methodology Commons

Recommended Citation
https://lib.dr.iastate.edu/creativecomponents/461

This Creative Component is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Creative Components by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
A Penalized h-Likelihood Method for Gaussian Spatial Additive Model on Regular Lattice

A PREPRINT

Hao Sun  Somak Dutta
Department of Statistics  Department of Statistics
Iowa State University  Iowa State University
Ames, Iowa, 50010  Ames, Iowa, 50010
hao123@iastate.edu  somakd@iastate.edu

December 5, 2019

ABSTRACT

Often in spatial regression problems, the covariates could be high-dimensional and have a non-linear relationship with the response. Furthermore, the functional relationship between the response and the covariates are often smoother than the spatial correlation. We propose a Gaussian spatial additive model on regular lattice where the large scale effects of the spatial covariates are modeled by smooth functions and the small scale spatial variability is modeled using a random field on the lattice. In order to facilitate linearity selection, we use penalized thin plate spline basis functions and derive a penalized h-likelihood method for simultaneous non-linearity selection and spatial adjustments. We derive novel estimating equations for estimating the precision parameters based on the profile h-likelihood. We demonstrate our method using Arsenic contamination data from Bangladesh.

Keywords  Additive Mixed Model · Thin Plate Spline · H-likelihood · Arsenic Contamination · Linearity Selection · REML

1 INTRODUCTION

In Gaussian spatial regression models the overall variation are typically decomposed into two parts – the large scale variation and small scale variation. The large scale variation is typically modeled by a smooth mean function whereas the small scale variation is modeled by a latent spatial process along with a nugget effect. Traditionally, a common choice for the mean function has been a linear model, producing so called spatial linear mixed models (LMM) for spatial data. In practice, however, the covariates often exhibit non-linear association with the response so that a linear model
Additive models (AM) and generalized additive models [Hastie and Tibshirani, 1986; Wood, 2017] are popular alternatives to estimate the non-linear mean functions in regression models because they can provide flexible regression relationship but they are mostly developed for independent observations. Few works also accommodate the spatial variation using penalized regression functions of geo-coordinates [Kazembe and Mpeketula, 2010; Kammann and Wand, 2003; Yu et al., 2019+]. These methods are quite useful in econometric and certain biological problems where the spatial dependence has a high degree of smoothness. However, such smoothness are rarely observed in the dependence structure of geophysical variables. Accordingly, these methods are susceptible to bias and loss in prediction accuracy.

In this work, we propose a spatial additive model,

\[ y(s) = f_1(x_1(s)) + \cdots + f_d(x_d(s)) + \psi(s) + \epsilon(s) \]

where \( y \) is the spatial response at location \( s \), \( x_i \)'s are spatial covariates, \( f_i \)'s are unknown smooth functions, \( \psi(s) \) is a Gaussian intrinsic random field and \( \epsilon(s) \) is a Gaussian white-noise process. Following the works of Dutta and Mondal [2015, 2016a] we use lattice-based approximations where the latent spatial process is defined over a finely meshed regular rectangular array embedding the underlying geographical domain. In particular, we use a first order Gaussian intrinsic autoregression model which is popular in spatial literature due to its connection to the continuum de Wijs process [Besag and Mondal, 2005; Dutta and Mondal, 2016b]. Furthermore, we extend the h-likelihood method [Dutta and Mondal, 2015] and obtain restricted maximum likelihood (REML) estimates of the penalty parameters of the regression splines and the spatial precision parameters. Keeping large data sets in mind we develop matrix-free computational algorithms for computing the best linear unbiased estimators (BLUE) of the regression model and best linear unbiased predictor (BLUP) of the spatial random effects.

The rest of the article is organized as follows. In section 2, we describe a motivating data on groundwater arsenic concentration in Bangladesh. Section 3 introduces our spatial additive mixed model (AMM) constructed by thin plate spline and first-order intrinsic Gaussian autoregression. Section 4 provides parameter estimation procedure using penalized h-likelihood method. In section 5 we introduce three matrix-free algorithms to reduce the computation complexity. In section 6 we construct a simulation study to show that our AMM has better performance on both estimation and prediction than LMM and AM. We apply AMM on arsenic data in section 7. The MSE and MSPE results through cross-validation reveals AMM is better than AM and LMM. The estimates of penalty parameter can also help us check the linearity of each covariate variable in the mean function. The concluding remarks and our future works are given in section 8.

2 GROUNDWATER ARSENIC CONTAMINATION IN BANGLADESH

Arsenic contamination is a serious problem in Bangladesh where approximately 57 million people drink groundwater with arsenic concentration above WHO’s standard (10 \( \mu \)g L\(^{-1} \)). Smith et al. [2000] reveals that long-term exposure
to groundwater with high arsenic concentration (>500 µg L⁻¹) can result in multi-system organ failure and cancers including lung, bladder and skin cancers. During the period from 1998 to 2001, British Geological Survey (BGS) and the Department of Public Health Engineering (DPHE) under the Government of Bangladesh jointly conducted a project called "Groundwater Studies for Arsenic Contamination in Bangladesh" to investigate arsenic contamination problem in Bangladesh. This project aimed at evaluating the scale of groundwater arsenic contamination and understanding the origins and behaviors of arsenic in Bangladesh aquifer. The survey collected 3534 well water samples located at 61 out of 64 districts while the other three districts are considered to be arsenic safe. Concentrations of arsenic and other 19 chemical elements and compounds (aluminum, boron, barium, calcium, cobalt, chromium, copper, iron, potassium, lithium, magnesium, manganese, sodium, phosphorus, silicon, sulfate, strontium, vanadium and zinc) were measured at the same time along with the construction year and depth of the wells. The survey shows that arsenic in Bangladesh aquifer has a large range concentration varying from less than 0.25 µg L⁻¹ to more than 1600 µg L⁻¹. The map of arsenic concentration in Bangladesh is shown in Figure 1. Obviously, the arsenic contamination problem in south Bangladesh is much more serious than the north and exhibit strong spatial patterns.

![Map of Arsenic Concentration Raw Data](image)

Figure 1: Map of Arsenic Concentration Raw Data

There is no explicit mechanism to explain the arsenic problem but it might be correlated with some known facts. For example, Kinniburgh and Smedley [2001] put forward an "iron oxide reduction hypothesis" that arsenic is desorbed and dissolved from iron oxides which had earlier gathered from rivers during their transportation as part of the normal river sediment load. They noticed that isolated arsenic hot spots in northern Bangladesh usually have high iron concentration.
Additionally, phosphorus was speculated in part derived from the same source as the arsenic. Shallow tube wells were found to be contaminated arsenic more often than deep tube wells. Among the tube wells with depth less than 150m, 27% exceeded the Bangladesh standard for arsenic in drinking water (50 \( \mu \text{g L}^{-1} \)) and 46% of tube wells exceeded the WHO standard value of 10 \( \mu \text{g L}^{-1} \). In contrast, among tube wells that are deeper than 150m, these percentages were were 1% and 5% respectively. The depths of the tube wells could be considered as surrogate for the depths of the actual aquifers, as it is commonly believed that the deeper aquifers, formed before the Pleistocene (more than 2.58 million years ago) are safe from arsenic, compared to the the shallow aquifers formed during the Holocene. Some

![Figure 2: Scatter Plot of Arsenic Concentration against Well Depth and Pie Plots of Arsenic Concentration Percentage for Both Shallow Wells and Deep Wells](image)

common elements like sodium and sulfur are probably associated with the mechanism of sediment releasing arsenic to groundwater and some other poisonous elements, such as manganese, boron and uranium, may also be useful in predicting arsenic concentration as suggested by Figure 5. Besides, We are also curious about the effect of tube well age on its arsenic concentration because tube wells expose the aquifers to environmental oxygen which may accelerate the hydrolysis of insoluble arsenate compounds into water soluble arsenous compounds.

In this paper, we utilize additive mixed model (AMM) to analyze and predict the arsenic concentration in the whole study region. The other elements’ concentrations, the depths and ages of those sampled tube wells are considered as covariate variables. In section 7 we show that the residuals from a non-spatial additive model exhibit obvious spatial pattern so we extend the additive model to incorporate spatial dependence via Gaussian Markov random fields models. The dataset finally has 3478 samples after we remove 49 samples with missing information and 7 potential outliers based on the box plots of each element’s concentration.
3 GAUSSIAN SPATIAL LINEAR AND ADDITIVE MODELS

Suppose we have a response vector $y$ with $n$ observations and a $n \times d$ covariate matrix $X = [X_1, X_2, \ldots, X_d]$. The traditional linear model (LM) assumes that

$$E(y) = X\beta$$

Additive model (AM), however, assumes that the covariate variables are independent with each other and the expectation of the response is can be written as the summation of some nonlinear functions of each covariate variable, i.e.

$$y_i = \sum_{j=1}^{d} f_j(x_{ij}) + \epsilon_i$$

where error $\epsilon_i \sim i.i.d. N(0, \sigma^2), i = 1, 2, \ldots, n$. One superiority of AM compared with a standard linear regression model is allowing arbitrary smooth functions of the covariates to obtain better model interpolation while penalizing unrestricted smoothness to avoid over- or under-fitting [Buja et al., 1989].

Although AM assumes no interaction between the covariates, one can add a function of some covariates as a new variable to introduce selected interactions or model a pairwise interaction using a surface smoother (see Hastie and Tibshirani [1986] and Wahba [1986]).

3.1 THIN PLATE SPLINE

Thin plate spline is a type of nonparametric smooth spline which avoids choosing the knot locations [Duchon, 1977, Wahba, 1990, Wood, 2003].

Suppose we have the same $(y_i, x_i)$ as above, $i = 1, 2, \ldots, n$. Thin plate spline estimates the mean function $g$ by minimizing

$$||y - g||^2 + \lambda J_{md}(g)$$
where $m$ is the order of the derivatives in the thin plate spline penalty

$$J_{md}(g) = \int_{R^d} \sum_{v_1, \ldots, v_d = m} \frac{m!}{v_1! \ldots v_d!} \left( \frac{\partial^m g}{\partial x_1^{v_1} \ldots \partial x_d^{v_d}} \right)^2 dx_1 \ldots dx_d$$

and $\lambda$ is a smoothing parameter.

Therefore, the dimension of the thin plate spline basis is $p = (m+1)\times d$. The $M$ functions $\phi_j$’s are linearly independent polynomials that span the null space of $J_{md}$, i.e. $J_{md}(\phi_i) = 0$ and

$$\eta_{md}(x) = \begin{cases} \Gamma(d/2 - m)x^{2m-d} & \text{d is odd} \\ \frac{2m\pi^{d/2}(m-1)!}{(2m+1)!} & \text{o.w.} \end{cases}$$

where $\eta_{md}$ is a $d$-dimension vector. By eigen-decomposition, rewrite $E$ as $UDU'$ where $U$ is $n \times n$ orthogonal matrix and $D$ is a diagonal matrix with $n$ eigenvalues. Let $U_k$ denote the first $k$ columns of $U$ and $D_k$ denote a diagonal matrix containing corresponding first $k$ eigenvalues. We can let the estimation function approximately minimize

$$\|y - E\delta - G\alpha\|^2 + \lambda\delta' E\delta$$

subject to $G'\delta = 0$. However, this estimation function is overparameterized since $\delta$ is a $n$-dimension vector. By eigen-decomposition, rewrite $E$ as $UDU'$ where $U$ is $n \times n$ orthogonal matrix and $D$ is a diagonal matrix with $n$ eigenvalues. Let $U_k$ denote the first $k$ columns of $U$ and $D_k$ denote a diagonal matrix containing corresponding first $k$ eigenvalues. We can let the estimation function approximately minimize

$$\|y - U_kD_k\delta_k - G\alpha\|^2 + \lambda\delta_k' D_k\delta_k$$

subject to $G'U_k\delta_k = 0$ where $\delta_k = U_k\delta$.

Suppose $Z_k$ is a matrix whose columns are orthogonal vectors such that $G'U_kZ_k = 0$ and define $\tilde{\delta} = Z_k'U_k'\delta$ which does not have any constraint. Then, the estimation function minimizes

$$\|y - U_kD_kZ_k\tilde{\delta} - G\alpha\|^2 + \lambda\tilde{\delta}' Z_k' D_k Z_k \tilde{\delta}$$

Note that $G'U_k$ is a $M \times k$ matrix so the dimension of $Z_k$ is $k \times (k - M)$ and $\tilde{\delta}$ is a $(k - M)$-dimension vector. Therefore, the dimension of the thin plate spline basis is $p = (k - M) + M = k$ (see Wood [2017]).

In additive model, we derive $p_j + 1$ thin plate basis functions for each covariate variable $X_j$ independently. Note that these $p_j + 1$ smooth functions all include the intercept. To avoid identification problem, we only keep one intercept and let the total number of smooth functions is $1 + \sum_{j=1}^d p_j$. The new covariate matrix constructed by thin plate spline from...
$X$ is a $n \times (1 + \sum_{j=1}^{d} p_j)$ matrix $T = [1, T_1, T_2, \cdots, T_d]$ with a new parameter vector $\tau' = (\tau_0, \tau'_1, \tau'_2, \cdots, \tau'_d)$ where $\tau_0$ represents the intercept. Here, $T_j = [U_{j,k}D_{j,k}Z_{j,k}, G_j]$ and $\tau'_j = (\tilde{\delta}'_j, \alpha'_j)$ is a $p_j$-dimension parameter vector. Therefore, the additive model has a linear model structure, i.e. 

$$y = T\tau + \epsilon$$

where $\hat{\tau}$ minimizes $||y - T\tau||^2 + \sum_{j=1}^{d} \lambda_j \tilde{\delta}'_j Z'_j D_j Z_j \tilde{\delta}$.

### 3.2 FIRST-ORDER INTRINSIC GAUSSIAN AUTOREGRESSION

Lattice-based Gaussian random field is very popular to be applied on geostatistical data in the recent years. In this paper, we fit first-order intrinsic Gaussian autoregression on the regular lattice. Consider a linear mixed model with spatial random effect:

$$y = T\tau + F\psi + \epsilon$$

where $y$ denotes response vector with $n$ observations, $T$ is $n \times p$ covariance matrix with fixed effect $\tau$, $F$ is a known incidence matrix so $F1 = 1$, $\psi$ is the vector of spatial effect on a fine $r \times c$ regular array and $\epsilon$ represents the residuals. The two random components have distributions:

$$\epsilon \sim N(0, \lambda_y I)$$
$$\tau \sim N(0, \lambda_\psi W)$$

where $\lambda_y I$ and $\lambda_\psi W$ are two precision matrices with unknown precision parameters $\lambda_y$ and $\lambda_\psi$. The first-order intrinsic autoregression has the Markov random field property, which means the random effect $\tau$ has the following properties:

$$E(\psi_{i,j} | \cdots) = \frac{1}{4}(\psi_{i,j-1} + \psi_{i,j+1} + \psi_{i-1,j} + \psi_{i+1,j})$$
$$Var(\psi_{i,j} | \cdots) = \lambda_\psi^{-1}$$

Therefore, the precision matrix $W$ is singular with rank $rc - 1$ and $\psi'W\psi$ can be written as a quadratic form as

$$\psi'W\psi = \frac{1}{4} \sum \sum (\psi_{i,j} - \psi_{i-1,j})^2 + \frac{1}{4} \sum \sum (\psi_{i,j} - \psi_{i,j-1})^2$$

As discussed in [Dutta and Mondal, 2016a], $W$ is the Laplacian matrix on the $r \times c$ grid. Suppose $M_k$ is a $k \times k$ matrix corresponding to the discrete cosine transformation (see section [5.1]) with entries

$$m_{1,j} = k^{-1/2} m_{i,j} = (2/k)^{1/2} \cos\{\pi(i - 1)(j - 1/2)/k\}$$

for $i = 2, \ldots, k, j = 1, 2, \ldots, k$. Let $C_k$ is a diagonal matrix where $i$th diagonal entry is

$$C_{k,i} = 2[1 - \cos\{\pi(i - 1)/k\}$$
Suppose also that $M = M_c \otimes M_r$, $C_{01} = \frac{1}{4}I_c \otimes C_r$ and $C_{10} = \frac{1}{4}C_c \otimes I_r$ where $\otimes$ denotes the Kronecker product, then the singular matrix $W$ has the spectral decomposition

$$W = M'(C_{01} + C_{10})M$$

### 3.3 Gaussian Spatial Additive Model

From the above parts, the formation of Gaussian additive spatial model is straightforward:

$$y = T\tau + F\psi + \epsilon$$

where $T = [1, T_1, T_2, \cdots, T_d]$, $\tau = (\tau_0, \tau_1', \tau_2', \cdots, \tau_d')'$ and $\tau_j' = (\delta_j', \alpha_j)'$, $j = 1, \cdots, d$. The estimations of $\tau$ and $\psi$ minimizes

$$\lambda_y ||y - T\tau - F\psi||^2 + \sum_{j=1}^d \lambda_j \delta_j' Z_j D_j Z_j \delta + \lambda_\psi \psi' W \psi$$

We change the order of $\tau$ to make the parameters which need penalty to smooth at the behind, i.e. we write $\tau' = \{\alpha', \delta'\}$ where $\alpha' = \{\tau_0, \alpha_1', \cdots, \alpha_d'\}$ and $\delta' = \{\delta_1', \cdots, \delta_d'\}$.

### 4 Parameters Estimation

[Dutta and Mondal 2016] and [Dutta and Mondal 2015] introduces the procedure of h-likelihood method on linear mixed model. Here we still use the formation of AMM in section 3.3

$$y = T\tau + F\psi + \epsilon$$

the h-likelihood equation is

$$\begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} T & F \\ 0 & B \\ Z & 0 \end{pmatrix} \begin{pmatrix} \tau \\ \psi \end{pmatrix} + \begin{pmatrix} \epsilon \\ \zeta \\ \kappa \end{pmatrix}$$

where $Z$ is a $(\sum_{j=1}^d k_j) \times (1 + \sum_{j=1}^d k_j)$ matrix that combines a zero matrix with a block diagonal matrix with diagonal entries $Z_{jk}$, i.e.

$$Z = \begin{bmatrix} 0_{(\sum_{j=1}^d k_j) \times (1+\sum_{j=1}^d M_j)} & \text{diag}(Z_{1k_1}, Z_{2k_2}, \cdots, Z_{dk_d}) \end{bmatrix}$$

$\epsilon$, $\zeta$ and $\kappa$ are independent Gaussian distribution random vectors with covariance matrices $\lambda_y^{-1} I$, $\lambda_\psi^{-1} G^{-1}$ and $D^{-1}$ separately. $D$ is the diagonal matrix where

$$\text{diag}(D) = \{0'_{1+\sum_{j=1}^d M_j}, \lambda_1 \text{diag}(D_{1k_1}), \cdots, \lambda_d \text{diag}(D_{dk_d})\}$$
Each $D_{jk}$ is also a diagonal matrix calculated by eigen-decomposition (see section 3.1) and $\lambda_j$’s are $d$ different penalty parameters we need to estimate. For simplicity, define

$$ X = \begin{pmatrix} T & F \\ 0 & B \\ Z & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} \lambda_y I & 0 & 0 \\ 0 & \lambda_y G & 0 \\ 0 & 0 & D \end{pmatrix}, \quad \beta = \begin{pmatrix} \tau \\ \psi \end{pmatrix}, \quad z = \begin{pmatrix} y \\ 0 \end{pmatrix}. $$

The normal equation is

$$ \begin{pmatrix} \lambda_y T'T + D & \lambda_y T'F \\ \lambda_y F'T & \lambda_y F'F + \lambda_\psi W \end{pmatrix} \begin{pmatrix} \tau \\ \psi \end{pmatrix} = \begin{pmatrix} \lambda_y T'y \\ \lambda_y F'y \end{pmatrix} \quad \text{or} \quad X'QX\beta = X'Qz $$

so that we can estimate $\beta$ with given $\lambda_y$, $\lambda_\psi$ and $\lambda_j$s. If we simply denote $A = X'QX\beta$ and $b = X'Qz$, the estimation is just the solve of a basic linear equation $A\beta = b$.

The REMLs of the precision parameter vector $\Lambda' = (\lambda_1, \cdots, \lambda_d, \lambda_y, \lambda_\psi)'$ are from the log-likelihood of the residuals which is

$$ 2l = \log \det Q - \log |X'QX| + (z - X\hat{\beta})'Q(z - X\hat{\beta}) $$

Taking the first derivatives of $l$ with respect to the precision parameters to zero, the score equations are

$$ \frac{1}{2} \text{Tr}(Q^{-1}Q_j) - \frac{1}{2} \text{Tr}(HQ^{-1}Q_j) - \frac{1}{2}(z - X\hat{\beta})'Q_j(z - X\hat{\beta}) = 0 $$

where

$$ H = X(X^TQX)^{-1}X'Qz \quad \text{and} \quad Q_j = \frac{\partial Q}{\partial \lambda_j}, \quad j = 1, 2, \cdots, d $$

$$ Q_{d+1} = \frac{\partial Q}{\partial \lambda_y}, \quad Q_{d+2} = \frac{\partial Q}{\partial \lambda_\psi}. $$

The information matrix $J$ is a $(d + 2) \times (d + 2)$ matrix whose $(i, j)$th entry is

$$ J_{ij} = \frac{1}{2} \text{Tr} \left\{ (I - H)Q^{-1}Q_i(I - H)Q^{-1}Q_j \right\} $$

Because the estimation vector $\hat{\beta}$ is a function of $\Lambda$, one straightforward method to estimate $\Lambda$ is to iteratively update $\Lambda$ and $\beta$ using Newton-Rhapson method.
5 FAST STATISTICAL COMPUTATION

Although the idea of parameter estimation in the last part of Section 4 is trivial, the traditional way to obtain the updates is computational infeasible as the matrix $A$ is very large and singular even if it is sparse. Dutta and Mondal [2016a] reduce the heavy computations for REML estimation of the mixed effect model by using some matrix free computation. First, they use two dimensional discrete cosine transformation (DCT) to do matrix-free matrix-vector statistical computations. Second, they introduce a novel matrix-free Lanczos algorithm to compute BLUEs and BLUPs. Third, they use stochastic trace approximation to approximately estimate the traces in the score equations. In this paper, we apply these methods in our AMM to reduce the computation complexity.

5.1 DISCRETE COSINE TRANSFORMATION

Suppose $E = \{e_{st}\}$ is a $r \times c$ matrix, then its two dimensional DCT is a $r \times c$ matrix where the $(s,t)$th entry is given by

$$c_s c_t^* \sum_{i=1}^{r} \sum_{j=1}^{c} e_{ij} \cos(\pi(i-1)/2(r-1)) \cos(\pi(j-1)/2(c-1))$$

where $c_s = \begin{cases} \sqrt{1/r} & \text{if } s = 1 \\ \sqrt{2/r} & \text{o.w.} \end{cases}$ and $c_t^* = \begin{cases} \sqrt{1/c} & \text{if } t = 1 \\ \sqrt{2/c} & \text{o.w.} \end{cases}$ The two dimensional inverse DCT on the matrix $E$ produces another $r \times c$ matrix whose $(s,t)$th entry is

$$\sum_{i=1}^{r} \sum_{j=1}^{c} c_i c_j^* e_{ij} \cos(\pi(s-1)/2(i-1)/r) \cos(\pi(t-1)/2(j-1)/c)$$

The detailed work can be seen in Rao and Yip [2014].

We use two dimensional DCT in solving the normal equation in section 4 and in computing the score equations and Fisher information matrix. In these formulas, it is essential to compute either $Wx$ or $W^{-1}x$ with some vector $x$ which requests the exact whole matrix $W$. However, note that $W$ can be written as $M'CM$ and $W^{-1} = M'C^{-1}M$ by spectral decomposition with known orthogonal matrix $M$ and some unknown diagonal matrix $C$. Obtaining the vectors $Mx$ and $M'x$ are the same as using a two-dimensional DCT and inverse DCT on the $r \times c$ matrix which is filled by vector $x$ by column.

5.2 MATRIX-FREE LANCZOS ALGORITHM

Lanczos algorithm is one popular algorithm used to solve the normal equation with form as $A\beta = b$. The set of orthogonal vector, $\{v_1, v_2, \cdots\}$, is called Lanczos vector as they are from the span of $\{b, Ab, A^2b, \cdots\}$. At the $k$th iteration, there is an approximate tridiagonal factorization of $AV_k \approx V_k \Delta_k$, where $V_k = [v_1, v_2, \cdots, v_k]$ and $\Delta_k$ is a $k \times k$ positive definite tridiagonal matrix. The solution of $\beta$ is a linear combination of Lanczos vectors and the detailed algorithm is shown in Dutta and Mondal [2015].
Besides, we use incomplete Cholesky decomposition to better reduce the computation of solving $A\beta = b$. We construct a preconditioning matrix $P$ and solve the equation $PAP'^*b = Pb$ where $\beta = P'^*b$. One request of $P$ is that the matrix-vector products $Px$ and $P'x$ can be computed in matrix-free way. In addition, we want the Lanczos algorithm for solving the new equation is fast which needs the condition number of $PAP'$ bounded. Dutta and Mondal [2016a] applies a block preconditioner for $A$. The preconditioning matrix $P$ can be written as $\text{diag}\{L_1^{-1}, L_2^{-1}\}$, where $L_1$ is the lower triangular Cholesky factor of $\lambda_yT'T + D$ and $L_2$ is the lower triangular Cholesky factor of $\lambda_yF'F + \lambda_\psi W$.

Because this matrix is very sparse, incomplete Cholesky decomposition is much faster than standard Cholesky decomposition.

5.3 STOCHASTIC TRACE APPROXIMATION

Finally, the score equations derived by taking the first derivative of log-likelihood require computations of the traces $\text{Tr}(I - H)Q^{-1}Q_j$ and $\text{Tr}(I - H)Q^{-1}Q_i(I - H)Q^{-1}Q_j$. The traditional method to calculate these traces has to know the large dense matrix $H$. However, the computation is still prohibitive even if we use the DCT and the Lanczos algorithms.

Instead, Dutta and Mondal [2016a] applies Hutchinson’s method to approximate the trace of some symmetric matrix by using a Monte Carlo average of its quadratic form in random vectors with mean 0 and identity covariance matrix. Suppose $E$ is a symmetric matrix, the approximation of its trace is

$$\text{Tr}E \approx \frac{1}{T} \sum_{t=1}^{T} u_t'E u_t$$

where $u_t$’s are i.i.d. Gaussian or Rademacher random variables.

Therefore, the score equations are approximately reduced to the following equations:

$$g_j \equiv \frac{1}{2T} \sum_{t=1}^{T} u_t'(I - H)Q^{-1}Q_j u_t - \frac{1}{2}(z - X\hat{\beta})'Q_j(z - X\hat{\beta}) = 0$$

where $j = 1, 2, \cdots, d + 2$. The $(i,j)$th entry of the Hessian matrix becomes

$$\frac{1}{2T} \sum_{t=1}^{T} u_t'(I - H)Q^{-1}Q_i(I - H)Q^{-1}Q_j u_t$$

These approximate equations can be easily calculated with matrix-free methods.
6 SIMULATION STUDIES

In this section, we compare our additive mixed model (AMM) with linear mixed model (LMM) and additive model (AM) through simulation study. Let the simulation model be:

\[ y_i = 1 + x_{i,1} + 2 \sin(x_{i,2}) + x_{i,3}^2 + u_i + e_i \quad i = 1, 2, \ldots, 128^2 \]

This model has one linear covariate variable \( x_1 \) and two nonlinear covariate variables \( x_2 \) and \( x_3 \). \( u_i \) denotes spatial random effect which is generated on a \( 128 \times 128 \) regular grids. Besides, all \( x_{i,j} \)'s are independently generated from standard normal distribution. In this simulation study, we show two examples that the first one with precision parameters \( \lambda_y = 2 \) and \( \lambda_\psi = 4 \) and the second example with \( \lambda_y = 4 \) and \( \lambda_\psi = 2 \), which indicates that the spacial effect is more significant in the second example. Each grid contains one unit so the population total is \( 128 \times 128 = 16384 \), from which 30% units are randomly sampled as training data with observed response to predict the left 70% data. We assume that the covariates are observed for all units. In both simulation studies, we choose \( m = 2 \) and \( k = 10 \) for all covariate variables in both AM and AMM which means \( \tilde{\delta}_j \) contains 8 columns and \( \alpha_j = [x_j] \), \( j = 1, 2, 3 \). We repeat each simulation study for 100 times. The table below shows that average of REML of each turning parameter, MSE of two spatial parameters, MSE and MSPE of the response variable under each method.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>MSE(( \lambda_y ))</th>
<th>MSE(( \lambda_\psi ))</th>
<th>MSE</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_y = 2, \lambda_\psi = 4 )</td>
<td>AM</td>
<td>6749</td>
<td>0.062</td>
<td>0.035</td>
<td>-</td>
<td>-</td>
<td>1.613</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>8830</td>
<td>0.236</td>
<td>0.418</td>
<td>0.310</td>
<td>0.036</td>
<td>0.555</td>
</tr>
<tr>
<td></td>
<td>LMM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.351</td>
<td>0.036</td>
<td>2.392</td>
</tr>
<tr>
<td>( \lambda_y = 4, \lambda_\psi = 2 )</td>
<td>AM</td>
<td>7080</td>
<td>0.090</td>
<td>0.052</td>
<td>-</td>
<td>-</td>
<td>2.468</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>9319</td>
<td>0.189</td>
<td>0.280</td>
<td>0.290</td>
<td>0.760</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>LMM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.298</td>
<td>0.849</td>
<td>1.939</td>
</tr>
</tbody>
</table>

Table 1: Average of REML of Each Turning Parameter, MSE of Spatial Parameters, MSE and MSPE of the Response Variable

From Table 1, it is clear that the AMM with spatial random effect has obviously better fitting and prediction results compared with the other two model if the data have both nonlinear pattern of covariates and spatial correlation. In the first case, the MSE and MSPE of AM are both smaller than those of LMM. In the second case, however, we can see that the MSE of AM is worse than the MSE of LMM. This result suggests that if the spatial effect is significant, AM would preform worse since it does not include spatial effect. On the other side, AM could have better prediction result than LMM as it can better estimate the mean function. No matter the spatial effect or the nonlinearity is significant, AMM preforms best as shown from the table.

In addition, AMM also provides good estimates of the turning parameters \( \lambda_j, j = 1, 2, 3 \) and spatial parameters than the other two methods. Note that larger \( \lambda_j \) suggests the mean function of \( x_j \) is closer to be linear. For simplicity, we save \( \hat{\lambda}_j = 10000 \) if the estimate value is greater than 10000 in our simulation study and also real data analysis in the next section. In both case, we can find that both AM and AMM can select the true linear variable \( X_1 \) with large estimate for \( \hat{\lambda}_1 \) while \( \hat{\lambda}_2 \) and \( \hat{\lambda}_3 \) are close to 0. Besides, AMM has smaller MSE of the two spatial parameters \( \lambda_y \) and \( \lambda_\psi \) as shown.
in the table. The MSE of $\lambda_\psi$ of from AMM is smaller than that from LMM in the second case which indicates that ignoring the nonlinearity will result worse estimate of the spatial effect when it is significant.

7 ANALYSIS OF ARSENIC DATA

In this section, we apply our AMM on the Bangladesh arsenic dataset and compare the MSE and MSPE with LMM and AM through cross-validation. As introduced in section 2, the dataset we use has 3478 observations and 21 covariate variables. 19 variables are other elements’ concentrations and the other two variables record the ages and depths of sampled tube wells. However, note that part of the records of each element concentration including arsenic are left censored, probably because of measurement accuracy. For instance, 994 arsenic samples are less than 0.5 $\mu$g L$^{-1}$ and 111 samples are less than 6 $\mu$g L$^{-1}$ while we also find some noncensored arsenic concentrations between 0.5 $\mu$g L$^{-1}$ and 6 $\mu$g L$^{-1}$. The censoring proportion of the whole dataset is about 30%. Here we replace each censored data by the upper bound of the interval. Besides, because the response and most covariates are right skewed, we turn to use the logarithm of these variables.

At first, four element variables (Cobalt, Chromium, Copper and Vanadium) are removed from the covariate matrix because more than 90% of their observations are censored so we use the left 17 variables as covariate variables. Next, for each covariate variable, we first use the left variables to fit AMM and draw a diagnostic plot of the residual against that covariate variable. To draw the diagnostic plot, we cut the range of each left covariate variable into 80 small intervals with equal length and then compute the interval average of arsenic and each covariate to draw 17 scatter plots as shown in Figure 4. Smooth function in these plots shows that the relationship between arsenic concentration and
some covariates such as Iron, Phosphorus and depth, are nonlinear under log scale. This rough analysis suggests that using additive model could have better model fitting result.

![Variogram of Residuals from Additive Model](image)

Next, we check the spatial correlation in AM residuals with full covariates. The study region is embedded with latitudes between 20 and 27 degrees north (about 778 kilometers in length) and longitudes between 88 and 92.5 degrees in east (about 460 kilometers) into a 250 × 150 array. Therefore, each grid is approximately 9.6 square kilometers in area and we position residuals into array cells based on their positions. We average the residuals if they are in the same grid. As shown in Figure 5, spatial correlation and sill are clear while the nugget effect looks a little large. Meanwhile, the range of these residuals are also large, which suggested us additive model might be not good enough to fit the data even if we consider the nonlinear relationship between the response and covariates. Dutta and Mondal [2016a] has studied on the Bangladesh Arsenic contamination problem using linear mixed model with spatial effect, but did not use the other elements’ concentrations as covaritates and neither consider to use a nonlinear function of covariates to construct the expectation of arsenic concentration.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>MSPE</th>
<th>λ_y</th>
<th>SE(λ_y)</th>
<th>λ_ψ</th>
<th>SE(λ_ψ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.374</td>
<td>1.447</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AMM</td>
<td>0.664</td>
<td>1.213</td>
<td>1.111</td>
<td>0.034</td>
<td>5.630</td>
<td>0.619</td>
</tr>
<tr>
<td>LMM</td>
<td>0.736</td>
<td>1.284</td>
<td>1.026</td>
<td>0.040</td>
<td>5.487</td>
<td>0.521</td>
</tr>
</tbody>
</table>

Table 2: Average of MSE and MSPE and REML Estimates of Spatial Parameters

In the following part, we compare AMM against LMM and AM using the same arsenic dataset after logarithm transformation. We assume the spatial effect in LMM and AMM are from the same regular lattice. In the LMM, the covariate matrix is exactly the original covariate matrix with 18 columns. On the other hand, we use the same turning parameters \( m = 2 \) and \( k = 10 \) as we use in simulation study for both AMM and AM so the dimension of each variable is 9. Therefore, the covariate matrix in two additive models are both 3478 × 154.

Next, we partition the dataset as a testing data with 500 samples and a training dataset with remaining samples. All three models are fitted with the training data to obtain MSE and MSPE of the testing data. We repeat the above process
Figure 6: The Boxplot of MSPE with Three Different Models

for 300 iterations. The summarization of the cross-validation results are shown by Figure 6 and Table 2. Both of them shows that considering nonlinear expectation function and spatial random effect can significantly reduce the MSPE while spatial effect can drop the MSPE more than additive model. In this example, we can see that the MSE of AM is much greater than the MSE of the other two models, which is different from the simulation result in section 6. This phenomenon could cause from the spatial correlation of covariate variables because most of them are also some chemical elements or components’ concentrations. We obtain global REML estimates \( \hat{\lambda}_y = 1.026 \) (with a s.e. 0.040) and \( \hat{\lambda}_\psi = 5.487 \) (with a s.e. 0.521) for LMM, and \( \hat{\lambda}_y = 1.111 \) (with a s.e. 0.034) and \( \hat{\lambda}_\psi = 5.630 \) (with a s.e. 0.619) for AMM which shows the spatial precision parameters for both two models are quite similar. Figure 7 shows the boxplot of all spline turning parameters. This estimation result shows that age, boron, lithium and zinc have large penalty parameters which suggests these covariates are linear in the mean function of AMM.

Finally, we use the full dataset to fit AMM and use the estimation result to draw the scatter plot of nonlinear part of the mean function for each covariate variable. From Figure 8 it is easy to see that the nonlinear mean function of age, boron, lithium and zinc are almost constant as 0 while the other variables have obvious nonlinear patterns.

8 DISCUSSION

This paper introduces an additive mixed model (AMM) with first-order intrinsic Gaussian spatial effect. Both our simulation study and real data analysis reveals that our AMM has advantages in both model fitting and prediction over basic additive model (AM) and linear mixed model with spatial effect if the mean function is nonlinear and there exists spatial correlation between the responses. Besides, because we use nonparametric thin plate spline to construct smooth
Figure 7: The Boxplot of Turning Parameters

Figure 8: Figures of Nonlinear Part of Mean Functions for All Covariate Variables
functions, we do not need to pick the knots. Our model can also check linearity through spline penalty parameters to help us better understand the relationship between the covariates and the response.

Our work has several directions of future research. One nature direction is to apply our method to a large spatial dataset. The Arsenic data has only about 3500 observations and 17 covariate variables. Another direction is to purpose an additive mixed effect model which does not only provide linear pattern but also variable selection. Finally, this additive mixed model could also be applied to spatial temporal data.

References


