Generalized Additive Models with Spatio-temporal Data

Xiangming Fang

East Carolina University, Greenville, USA

Kung-Sik Chan

University of Iowa, Iowa City, USA

Summary. Generalized additive models (GAMs) have been widely used. While the procedure for fitting a generalized additive model to independent data has been well established, not as much work has been done when the data are correlated. The currently available methods are not completely satisfactory in practice. A new approach is proposed to fit generalized additive models with spatio-temporal data via the penalized likelihood approach which estimates the smooth functions and covariance parameters by iteratively maximizing the penalized log likelihood. Both maximum likelihood (ML) and restricted maximum likelihood (REML) estimation schemes are developed. Also, conditions for asymptotic posterior normality are investigated for the case of separable spatio-temporal data with fixed spatial covariate structure and no temporal dependence. We propose a new model selection criterion for comparing models with and without spatial correlation. The proposed methods are illustrated by both simulation study and real data analysis.

Keywords: GAM, Matérn class, Maximum likelihood (ML), Penalized likelihood, Restricted maximum likelihood (REML), Spatio-temporal data

1. Introduction

In spatial statistics, modeling both the mean structure and the covariance structure is often of interest. For modeling the mean structure, a standard linear or non-linear model is usually sufficient, but when the relationship between the variables is complex and can not be easily modeled by specific linear or non-linear functions, a generalized additive model (GAM) will be a natural choice.

Generalized additive models were first proposed by Hastie and Tibshirani (1986, 1990). These models assume that the mean of the response variable depends on an additive predictor through a link function. Like generalized linear models (GLMs), generalized additive models permit the response probability distribution to be any member of the exponential family of distributions. The only difference between GAMs and GLMs is that the GAMs allow for unknown smooth functions in the linear predictor. In general, a generalized additive model has a structure like

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\beta} + \sum_{j=1}^{m} f_j(\mathbf{x}_{ij})$$

Address for correspondence: Kung-Sik Chan, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242, USA
E-mail: kung-sik-chan@uiowa.edu

where $Y_i \sim$ some exponential family distribution; $\mu_i = E(Y_i)$; $\mathbf{X}_i^*$ is the i^{th} row of the model matrix for the strictly parametric model components; f_j are smooth functions of the covariates $\mathbf{x}_j$.

The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relationships between the response and the set of explanatory variables. Due to the high flexibility in model specification, GAMs have been widely used, see, for example, Hastie and Tibshirani (1995), Lehmann (1998), Abe (1999), Frescino *et al.* (2001), Guisan *et al.* (2002) and Dominici *et al.* (2002).

For methodologies of fitting a generalized model, there exist a large literature on generalized additive models and nonparametric regression models with independent data using spline methods (Wahba, 1990; Green and Silverman, 1994; Gu, 2002; Wood, 2006). However, only limited work has been done with correlated data. Several researchers have restricted their attention to longitudinal data with normally distributed responses and have incorporated a nonparametric time function in linear mixed models (Zeger and Diggle, 1994; Zhang *et al.*, 1998). For more general cases, Lin and Zhang (1999) proposed generalized additive mixed models (GAMMs) which is a generalization of the generalized linear mixed models (GLMMs). As they mentioned in the discussion, there are bias problems especially when the random effects are correlated. Wood (2006) included GAMMs in his R package *mgcv* based on Lin and Zhang's approach. He pointed out that GAMM fitting is not as numerically stable as GAM and will occasionally fail, especially when explicitly modelling correlation in the data, probably because of the confounding between correlation and non-linearity. Fahrmeir and Lang (2001) and Fahrmeir *et al.* (2004) proposed a fully Bayesian approach. Since all inferences are based on MCMC simulations, the computational cost of the Bayesian approach may be high, especially when the sample size is large. Therefore, its practical feasibility deserves careful consideration.

We propose an alternative and hopefully more stable approach to fit generalized additive models with correlated data via the penalized likelihood approach that avoids the complexity of the mixed model approach and the high computation cost of the Bayesian approach.

Although the proposed approach does not assume any specific correlation structure, particular attention will be given to spatial correlation defined by the Matérn class. The Matérn class is a rich family of autocovariance functions taking the general form

$$K(h) = \sigma^2 \frac{(h/\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} \mathcal{K}_\nu(h/\phi),$$

where h is the (geodesic) distance between two data points, σ^2 is the variance parameter, ν is the smoothness parameter, ϕ is the range parameter, and $\mathcal{K}_\nu(x)$ is the modified Bessel function of the second kind with order ν (Abramowitz and Stegun, 1972). The Matérn class includes the exponential correlation function when $\nu = 0.5$ and the Gaussian correlation function as a limiting case when $\nu \to \infty$. The smoothness parameter controls the smoothness of the process, which depends on the variogram's behavior near the origin, i.e., the correlation function's behavior when observations are separated by small distances. Stein (1999) strongly recommended the Matérn class for modeling spatial correlation because of its ability to specify the smoothness of the random field. In recent years, the Matérn class has received more attention in the literature, see, for example, Williams *et al.* (2000), Diggle *et al.* (2002), Zhang (2004), and Zhu and Zhang (2006).

Some of the Matérn model parameters are not consistently estimable under fixed domain asymptotics (see Ying, 1991; Stein, 1999; Zhang, 2004). What if we have repeated

measurements at the sampling locations? To answer this question, situations for the spatio-temporal case, where the spatial design is assumed to be fixed with temporally independent repeated measurements and the spatial correlation structure does not change over time, are investigated via studying the conditions for asymptotic posterior normality. Data satisfying the aforementioned conditions are increasingly collected in science, for example, large-scale annual fisheries monitoring data. Our theoretical investigation exploits the fact that penalized likelihood estimation can be given a Bayesian interpretation. Asymptotic posterior normality has been an important topic in Bayesian inference. Walker (1969) gave a rigorous proof of asymptotic posterior normality under certain regularity conditions in the i.i.d. case. After that, a number of investigators extended Walker's results to cover general stochastic processes, see, for example, Heyde and Johnstone (1979), Chen (1985), Sweeting and Adekola (1987), and Sweeting (1992). Their work is quite general. We follow the results of Sweeting (1992) to study the conditions under which asymptotic posterior normality holds in the spatio-temporal case. As temporal independence is a strong assumption of our analysis, a model diagnosis method is developed to check if the assumption of independence across time holds for the spatio-temporal data.

The rest of the paper is organized as follows. In Section 2, we introduce the GAM with spatially correlated but temporally independent data. A detailed description of our new approach for fitting GAMs with correlated data is given in Section 3. In Section 4, conditions for asymptotic posterior normality under the spatio-temporal case are investigated. A simulation study on the performance of the new approach is given in Section 5, and a model diagnosis method for checking the assumptions of independence across time is developed in Section 6. Also, we propose a model selection criterion based on the Bayesian framework in Section 7 to compare different candidate models. Finally, the proposed methodology is applied to modeling the distribution of the pollock fish egg in the Gulf of Alaska.

2. Spatio-temporal Model

Consider the GAM,

$$\mathbf{Y}_t = \mathbf{f}_1(\mathbf{x}_{1t}) + \mathbf{f}_2(\mathbf{x}_{2t}) + \cdots + \mathbf{f}_m(\mathbf{x}_{mt}) + \mathbf{e}_t, \ , t = 1, 2, 3, \cdots, T \tag{1}$$

where $\mathbf{Y}_t \in \mathbb{R}^{n_0}$ with n_0 the number of observations for each time period t; $\mathbf{x}_{jt}$ are the values of covariates $\mathbf{x}_j$ at time t; f_j are unknown smooth functions; $\mathbf{e}_t \sim N(\mathbf{0}, \ \mathbf{\Sigma}_t(\theta))$ with $\mathbf{\Sigma}_t(\theta)$ defined by some spatial covariogram function K. We assume that the temporal correlation and the spatial correlation are separable, i.e., the covariance of observations at time t_1, location $\mathbf{s}_1$ and time t_2, location $\mathbf{s}_2$ can be written as

$$\mathrm{Cov}(Y_{t_1}(\mathbf{s}_1), Y_{t_2}(\mathbf{s}_2)) = \mathrm{Cov}(e_{t_1}(\mathbf{s}_1), e_{t_2}(\mathbf{s}_2)) = \tau(|t_1 - t_2|)K(|\mathbf{s}_1 - \mathbf{s}_2|)$$

where τ is a temporal autocorrelation function (the temporal variance has been absorbed into $K(|\mathbf{s}_1 - \mathbf{s}_2|)$. In other words, the temporal correlation scheme is independent of the spatial correlation scheme.

Throughout this paper, we focus on a special case of the above model, in which the spatial design does not change over time, i.e., $\mathbf{\Sigma}_t(\theta) = \mathbf{\Sigma}_\theta$ for all t, and the observations are temporally independent, i.e., $\tau(|t_1 - t_2|) = 1$ if $t_1 = t_2$ and 0 otherwise. Thus the random vectors $\mathbf{e}_{t_1}$ and $\mathbf{e}_{t_2}$ are independent for $t_1 \neq t_2$. This is a strong assumption; we will study the issue of checking the validity of this assumption later on.

3. Parameter Estimation

In this section, we describe our new fitting approach using a special case of Model (1) where $t = m = 1$, i.e., there are only one time period and one smooth function. The approach can be easily adapted for multiple smooth functions and multiple time-period data.

3.1. Model

Suppose the model is

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i,$$

where $\mathbf{x}_i$ is a d-vector of the covariates; f is a unknown smooth function; and the errors $\boldsymbol{\varepsilon}$ have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\theta$ where $\boldsymbol{\theta}$ are the covariance parameters.

Consider the problem of estimating the covariance parameters $\boldsymbol{\theta}$ and the smooth function f, based on data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1^{'} \\ \ldots \\ \mathbf{x}_n^{'} \end{pmatrix}$. Similar to the GAMs with uncorrelated errors, this goal can be achieved by maximizing the penalized log likelihood

$$\ell_P = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}(\mathbf{Y} - \mathbf{f}(\mathbf{x}))'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y} - \mathbf{f}(\mathbf{x})) - \frac{1}{2}\lambda J(f), \tag{2}$$

where $\mathbf{f}(\mathbf{x}) = [f(\mathbf{x}_i), i = 1, \ldots, n]'$, λ is the smoothing parameter controlling the tradeoff between the model fit and the smoothness of the regression function, and J is a wiggliness penalty functional which is defined as

$$J(f) = \int [f^{''}(x)]^2 dx \text{ for } d = 1$$

and

$$J(f) = \int \cdots \int \sum_{v_1 + \cdots + v_d = m} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m f}{\partial x_1^{v_1} \cdots x_d^{v_d}} \right)^2 dx_1 \cdots dx_d \text{ for } d > 1,$$

with $2m > d$.

It is hard to maximize ℓ_P with respect to $\boldsymbol{\theta}$ and f simultaneously. Therefore, we developed an iterative algorithm which maximizes the penalized log likelihood alternatively with respect to the covariance parameters and the smooth functions.

3.2. Penalized Maximum Likelihood Estimation

The algorithm for covariance parameter estimation is as follows:

Step 1 Start with some initial value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(0)}$. Treat $\boldsymbol{\theta}$ as known and try to select the smoothing parameter λ and estimate the smooth function f.

For fixed $\boldsymbol{\theta}$, maximizing (2) becomes

$$\max_{f(x)} \left\{ -\frac{1}{2}(\mathbf{Y} - \mathbf{f}(\mathbf{x}))'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y} - \mathbf{f}(\mathbf{x})) - \frac{1}{2}\lambda J(f) \right\}.$$

It can be shown that for fixed $\boldsymbol{\theta}$ the solution of f to the above maximization problem is a natural cubic spline if $d = 1$ and a natural thin-plate spline if $d > 1$. Thus $\mathbf{f}(\mathbf{x})$ can

be represented as a linear function of the spline basis, i.e. $\mathbf{f}(\mathbf{x}) = \mathbf{X}\boldsymbol{\beta}$ for some unknown parameters $\boldsymbol{\beta}$ and $\mathbf{X}$ is the design matrix of the spline basis function. Moreover, $J(f) = \boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}$ for some known symmetric matrix $\mathbf{S}$, see Wood (2006). Hence the maximization problem is equivalent to the penalized weighted least squares problem

$$\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{\Sigma}_\theta^{-1/2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})||^2 + \lambda\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta} \right\}$$

or

$$\min_{\boldsymbol{\beta}} \left\{ ||\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta} \right\} \tag{3}$$

where $\boldsymbol{\Sigma}_\theta^{-1/2}$ is any square root matrix of $\boldsymbol{\Sigma}_\theta^{-1}$ such that $(\boldsymbol{\Sigma}_\theta^{-1/2})'\boldsymbol{\Sigma}_\theta^{-1/2} = \boldsymbol{\Sigma}_\theta^{-1}$, $\tilde{\mathbf{Y}} = \boldsymbol{\Sigma}_\theta^{-1/2}\mathbf{Y}$and $\tilde{\mathbf{X}} = \boldsymbol{\Sigma}_\theta^{-1/2}\mathbf{X}$.

The smoothing parameter can be estimated by a number of criteria, e.g. CV or GCV. Here, we shall estimate λ by minimizing the GCV as it can be readily implemented by using the *mgcv* library (Wood, 2006) in the statistical platform R. In the case that $\boldsymbol{\Sigma}_\theta = \mathbf{I}$, the identity matrix, the GCV(λ) is defined by the formula:

$$\mathcal{V} = \frac{n||\mathbf{Y} - \mathbf{A}\mathbf{Y}||^2}{[n - \mathrm{tr}(\mathbf{A})]^2},$$

where $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}'$.

Thus for the case in (3), λ is selected by minimizing the GVC score

$$\mathcal{V} = \frac{n||\tilde{\mathbf{Y}} - \tilde{\mathbf{A}}\tilde{\mathbf{Y}}||^2}{[n - \mathrm{tr}(\tilde{\mathbf{A}})]^2},$$

where $\tilde{\mathbf{A}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda\mathbf{S})^{-1}\tilde{\mathbf{X}}'$.

With known $\boldsymbol{\theta}$ and λ, the minimization problem (3) admits a unique solution, i.e. $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda\mathbf{S})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} = (\mathbf{X}\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y}$, if $\mathbf{X}\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S}$ is of full rank.

Denote the selected λ and the corresponding estimated $\boldsymbol{\beta}$ as $\lambda^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ respectively.

Step 2 Let $\lambda = \lambda^{(1)}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)}$, and try to find a new estimate of $\boldsymbol{\theta}$.

With λ and $\boldsymbol{\beta}$ fixed, maximizing (2) becomes

$$\max_{\boldsymbol{\theta}} \left\{ -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| \right\}.$$

Clearly the solution is the MLE of $\boldsymbol{\theta}$ with $\boldsymbol{\beta}^{(1)}$ plugged in. Denote the new estimate of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^{(1)}$.

Step 3 Stop if $\frac{||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}||}{||\boldsymbol{\theta}^{(0)}||} < 10^{-4}$. Otherwise, let $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}^{(1)}$ and repeat Steps 1–3.

Conditional on the covariates, the standard errors and confidence intervals of $\hat{\boldsymbol{\theta}}$ can be obtained based on the information matrix or the observed information matrix from the log

likelihood in Step 2.

$$\begin{aligned}
\ell_P &= -\frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\lambda\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta} \\
\frac{\partial \ell_P}{\partial \boldsymbol{\beta}'} &= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} - \lambda\boldsymbol{\beta}'\mathbf{S} \\
\frac{\partial^2 \ell_P}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= -\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} - \lambda\mathbf{S} \\
\frac{\partial^2 \ell_P}{\partial \boldsymbol{\theta}\partial \boldsymbol{\beta}'} &= \frac{\partial}{\partial \boldsymbol{\theta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}
\end{aligned}$$

Since $E\frac{\partial^2 \ell_P}{\partial \boldsymbol{\theta}\partial \boldsymbol{\beta}'} = 0$, the Fisher information matrix equals

$$\mathcal{I} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S} & \mathbf{0} \\ \mathbf{0} & -E\frac{\partial^2 \ell_P}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \end{pmatrix}.$$

Thus under some suitable regularity conditions, it can be expected that,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &\dot{\sim} \mathbf{N}\left(\boldsymbol{\beta},\ (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}\right) \\
\hat{\boldsymbol{\theta}} &\dot{\sim} \mathbf{N}\left(\boldsymbol{\theta},\ (-E\frac{\partial^2 \ell_P}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'})^{-1}\right),
\end{aligned}$$

which will be revisited later.

3.3. Penalized Restricted Maximum Likelihood Estimation

It is well known that the maximum likelihood estimation tends to underestimate the covariance parameters due to the loss of degrees of freedom in estimating the mean structure parameter, while the restricted maximum likelihood (REML) estimation is less biased (Corbeil and Searle, 1976). Recall that with unpenalized likelihood, the restricted likelihood is the average of the likelihood over all possible values of regression coefficients $\boldsymbol{\beta}$, i.e.,

$$L_R(\boldsymbol{\theta}) = \int L(\boldsymbol{\beta}, \boldsymbol{\theta}) d\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is given a non-informative prior.

But for penalized likelihood, we are actually imposing some prior beliefs about the likely characteristics of the correct model, which means we need to specify a prior distribution on $\boldsymbol{\beta}$. Specifically let the prior for $\boldsymbol{\beta}$, which is generally improper, be

$$f_\beta(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}\boldsymbol{\beta}'\lambda\mathrm{S}\boldsymbol{\beta}\}.$$

This prior is appropriate since it makes explicit the fact that we believe smooth models to be more likely than wiggly ones, but it gives equal probability density to all models of equal smoothness (Wood, 2006). In fact, the penalty term also implies that this prior is appropriate.

With the prior given previously, the restricted penalized likelihood function can be shown to equal

$$\begin{aligned} & L_R(\boldsymbol{\theta}|\mathbf{Y}) = \int f(\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\theta}) f_\beta(\boldsymbol{\beta}) d\boldsymbol{\beta} \\ = \;& const. \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_\theta|}} \int \exp\{-\frac{1}{2}[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\lambda\mathbf{S}\boldsymbol{\beta}]\} d\boldsymbol{\beta} \\ = \;& const. \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_\theta|}} \exp\{-\frac{1}{2}(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}})'(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}})\} \sqrt{\frac{(2\pi)^p}{|\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}|}} \end{aligned}$$

where $\widetilde{\mathbf{Y}} = \begin{pmatrix} \boldsymbol{\Sigma}_\theta^{-1/2}\mathbf{Y} \\ \mathbf{0} \end{pmatrix}$, $\widetilde{\mathbf{X}} = \begin{pmatrix} \boldsymbol{\Sigma}_\theta^{-1/2}\mathbf{X} \\ \mathbf{B} \end{pmatrix}$, $\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{Y}}$, $p = \dim(\boldsymbol{\beta})$, and $\mathbf{B}$ is any matrix such that $\mathbf{B}'\mathbf{B} = \lambda\mathbf{S}$. After some algebra, the penalized restricted log-likelihood function can be expressed as

$$\begin{aligned} & \ell_R(\boldsymbol{\theta}|\mathbf{Y}) = \log L_R(\boldsymbol{\theta}|\mathbf{Y}) \\ = \;& const. - \frac{1}{2}\log(|\boldsymbol{\Sigma}_\theta|) - \frac{1}{2}\log(|\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S}|) \\ & -\frac{1}{2}\mathbf{Y}'\left[\boldsymbol{\Sigma}_\theta^{-1} - \boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}\left(\mathbf{X}'(\boldsymbol{\Sigma}_\theta)^{-1}\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\right]\mathbf{Y}. \end{aligned}$$

Penalized restricted likelihood estimation can then be carried out iteratively with a scheme similar to that of the ML estimation, except that Step 2 is modified as follows: With fixed λ, maximizing ℓ_R with respect to $\boldsymbol{\theta}$ gives the REML estimate $\boldsymbol{\theta}^{(1)}$. Steps 1 and 3 are exactly the same as described previously. Also, the standard errors and confidence intervals of $\boldsymbol{\theta}$ can be obtained based on the information matrix from the penalized restricted log likelihood.

3.4. Inference on Smooth Functions

Once the covariance parameters are estimated, we can repeat Step 1 to select the smoothing parameter and estimate $\boldsymbol{\beta}$. Then the fitted smooth function $\hat{\mathbf{g}}$ is given by $\mathbf{X}\hat{\boldsymbol{\beta}}$.

As discussed before, with the smoothing parameter fixed the prior distribution of $\boldsymbol{\beta}$ is

$$f_\beta(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}\boldsymbol{\beta}'\lambda\mathbf{S}\boldsymbol{\beta}\}.$$

Based on the model specification, the conditional distribution of $\mathbf{Y}$ given $\boldsymbol{\beta}$ is

$$f_{Y|\beta}(\mathbf{y}) \propto \exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}.$$

So by the Bayes rule, the posterior distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned} f_{\beta|Y}(\boldsymbol{\beta}) \;&\propto\; \exp\{-\frac{1}{2}(\boldsymbol{\beta}'(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y} + \mathbf{Y}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y})\} \\ &\propto\; e^{-\frac{1}{2}[(\boldsymbol{\beta}-(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\lambda\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y})'(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\lambda\mathbf{S})(\boldsymbol{\beta}-(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\lambda\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y})]}. \end{aligned}$$

Therefore,

$$\boldsymbol{\beta}|\mathbf{Y} \sim \mathbf{N}\left((\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y},\; (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}\right),$$

i.e., $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{Y}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}$.

Note that the covariance matrix of $\hat{\boldsymbol{\beta}}$ from the posterior is the same as what we got from the information matrix.

Now we have the covariance matrix of $\hat{\mathbf{g}}$ which is given by

$$\mathrm{Var}(\hat{\mathbf{g}}) = \mathbf{X}\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{X}' = \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}',$$

and thus can construct confidence intervals for the smooth function.

If there are more than two smooth functions in the model, corresponding submatrix of $\mathbf{X}$, subset of $\hat{\boldsymbol{\beta}}$ and submatrix of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ can be used to estimate each component smooth function and construct confidence intervals for each of them. Note that for the model to be identifiable, the model matrix and the estimated coefficients for an additive model are for the centered smooth functions, which means each of the estimated smooth functions sums up to zero across the data.

4. Asymptotic Posterior Normality

As described in the previous section, given a fixed set of basis functions, model (1) can be rewritten as

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{e}_t, \ , t = 1, 2, 3, \cdots, T, \tag{4}$$

and the corresponding penalized log likelihood is

$$\ell_P = -\frac{T \cdot n_0}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}.$$

Note that the smoothing parameters have been absorbed into the matrix $\mathbf{S}$.

For fixed smoothing parameters, the maximum penalized likelihood estimator can be interpreted as the posterior mode under a suitable prior density. Specifically, the prior density of $\boldsymbol{\beta}$ equals

$$p(\boldsymbol{\beta}) = \frac{|\mathbf{D}_+|^{1/2}}{(2\pi)^{m/2}}\exp\{-\frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\}$$

where m is the number of strictly positive eigenvalues of $\mathbf{S}$ and $\mathbf{D}_+$ is the diagonal matrix with all those strictly positive eigenvalues of $\mathbf{S}$ arranged in descending order on the leading diagonal. The parameter $\boldsymbol{\theta}$ has a flat prior over its parameter space Θ and is independent of $\boldsymbol{\beta}$. Then the joint prior density of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{|\mathbf{D}_+|^{1/2}}{(2\pi)^{m/2}}\exp\{-\frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\}. \tag{5}$$

Hence the posterior density equals

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\theta}|\text{data}) &= p(\boldsymbol{\beta}, \boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\theta}) / \int p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\theta})p(\boldsymbol{\beta}, \boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta} \\ &= \frac{|\mathbf{D}_+|^{1/2}\exp\{-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\}}{(2\pi)^{(m+T\cdot n_0)/2}|\boldsymbol{\Sigma}_\theta|^{T/2}\int p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\theta})p(\boldsymbol{\beta}, \boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta}} \end{aligned} \tag{6}$$

which is exactly the same as the penalized likelihood up to a normalization constant. Note that the analysis in this section is also conditional on the design matrix $\mathbf{X}$ although it is not

explicitly stated. Sweeting (1992) studied the asymptotic behavior of the posterior density under the true model. He showed that under suitable regularity conditions, the posterior density is asymptotically normal. Thus, we may study the large-sample properties of the penalized estimation via the asymptotic posterior normality framework of Sweeting (1992).

Below, we state a theorem on the asymptotic posterior normality for GAMs under the spatio-temporal setting with fixed spatial design and temporally independent data. The proof is deferred to an appendix.

Theorem 1 Consider the spatio-temporal model (4) with fixed spatial design and temporally independent data. Let $\boldsymbol{\phi}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \in \Phi = \mathbb{R}^k \times \Theta$ be the true parameter value, where Θ is a relatively compact, open convex subset of $\mathbb{R}^l$. Suppose the prior of $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\theta}) \in \Phi$ is defined by (5), the covariogram function $K(\cdot|\boldsymbol{\theta})$ is twice differentiable w.r.t. $\boldsymbol{\theta}$, and the covariance matrix $\boldsymbol{\Sigma}_\theta$ is invertible and continuous over Θ. Furthermore, assume the following conditions are satisfied:

(A1) for $f_\phi(\mathbf{Y}_t) = -\frac{\partial^2 l(\boldsymbol{\phi}|\mathbf{Y}_t)}{\partial \boldsymbol{\phi} \boldsymbol{\phi}'}$, $\exists\ \delta_0 > 0$ and a finite, integrable function $M(\mathbf{Y}_t)$ such that $\sup_{|\phi-\phi_0|<\delta_0} \|f_\phi(\mathbf{Y}_t)\|_{\max} \le M(\mathbf{Y}_t)$ where $\|\cdot\|_{\max}$ is the maximum norm;

(A2) for $f_\theta(\mathbf{e}_t) = \left(\frac{\partial}{\partial \theta'} \mathbf{X}_t' \boldsymbol{\Sigma}_\theta^{-1}\right) \mathbf{e}_t$, $\exists\ \delta_1 > 0$ and a finite, integrable functions $M_1(\mathbf{e}_t)$ such that $\sup_{|\theta-\theta_0|<\delta_1} \|f_\theta(\mathbf{e}_t)\|_{\max} \le M_1(\mathbf{e}_t)$;

(A3) $\frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \theta} \mathbf{X}_t' \boldsymbol{\Sigma}_\theta^{-1} \mathbf{X}_t$ is bounded, uniformly for $\theta \in \Theta$ in probability;

(A4) over the closure of Θ, $-\frac{1}{T}\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ is positive definite and a continuous function a.s.;

(A5) $\inf_{\theta \in \Theta} \lambda_{\min}(\frac{1}{T}\sum_{t=1}^T \mathbf{X}_t' \boldsymbol{\Sigma}_\theta^{-1} \mathbf{X}_t) > 0$ a.s., where $\lambda_{\min}(\mathbf{A})$ denotes the minimum eigenvalue of a symmetric matrix $\mathbf{A}$.

Then there exists a sequence of local maxima of the posterior density defined by (6) around $\boldsymbol{\phi}_0$ such that the posterior density of $(\mathbf{J}_T^{1/2})'(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}_T)$ converges in probability to the standard multivariate normal distribution, where $\mathbf{J}_T^{1/2}$ is the left Cholesky square root of $\mathbf{J}_T = \begin{bmatrix} \mathbf{J}_T(\hat{\boldsymbol{\beta}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_T(\hat{\boldsymbol{\theta}}) \end{bmatrix}$ with the diagonal blocks $\mathbf{J}_T(\hat{\boldsymbol{\beta}}) = \sum_{t=1}^T \mathbf{X}_t \boldsymbol{\Sigma}_{\hat{\theta}}^{-1} \mathbf{X}_t + \mathbf{S}$ and $\mathbf{J}_T(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

Several remarks are in order. Note that this theorem implies that the penalized likelihood function asymptotically approaches the normal density with mean $\hat{\boldsymbol{\phi}}_T$ and variance matrix $\mathbf{J}_T^{-1}$. Recall that in Section 3, the variance matrix of $\hat{\boldsymbol{\theta}}$ is given by the inverse negative Hessian matrix w.r.t. $\boldsymbol{\theta}$ and the variance matrix of $\hat{\boldsymbol{\beta}}$ is given by the posterior variance $(\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X} + \mathbf{S})^{-1}$ which is also the inverse negative Hessian matrix w.r.t. $\boldsymbol{\beta}$. Therefore, this theorem provides a justification, in the case of spatio-temporal model, of the way that the confidence intervals of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are constructed in Section 3, whose extension to the spatio-temporal model is straightforward.

5. Simulation Study

Now, we investigate the empirical performance of the penalized likelihood estimation from the spatio-temporal data through a simulation study. The data are simulated from the following model

$$Y_t(x, y) = f_1(x) + f_2(y) + b_t(x, y) + e_t(x, y),\ t = 1, 2, \cdots, T$$

where $x \in [0,1]$ and $y \in [0,1]$ are the coordinates of a data point; $f_1(x) = 2\sin(\pi x)$; $f_2(y) = e^{2y} - 3.75$; $\mathbf{b}_t$ are the spatially correlated errors with distribution $\mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$; and $\mathbf{e}_t$ are the pure measurement errors (or nugget effect) with distribution $\mathbf{N}(\mathbf{0}, \eta\mathbf{I})$.

Since we are particularly interested in the Matérn model, the covariance function for $\mathbf{b}_t$ is set to be

$$K(h) = \sigma^2 \frac{(h/\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} \mathcal{K}_\nu(h/\phi),$$

where h is the distance between two data points, σ^2 is the variance parameter, ν is the smoothness parameter, and ϕ is the range parameter.

However, we found in practice that this paramerization of the Matérn model is not very stable and can often result in numerical failures during the numerical optimization procedure. It turns out the following parametrization performs much better numerically

$$K(h) = \sigma^2 \frac{(\frac{h}{\nu\rho})^\nu}{2^{\nu-1}\Gamma(\nu)} \mathcal{K}_\nu(\frac{h}{\nu\rho}),$$

where $\rho = \phi/\nu$. From now on, we will stick to this new parameterization of the Matérn model. So the parameters of interest here are $\boldsymbol{\theta} = (\sigma^2, \nu, \rho, \eta)'$.

The random fields $\mathbf{b}_t$ are independently simulated on the same set of locations and from the same normal distribution $N(\mathbf{0},\ \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is defined by the Matérn covariogram with $\sigma^2 = 1,\ \nu = 2,\ \rho = 0.045$. The measurement errors $\mathbf{e}_t$ are simulated from the normal distribution $N(0,\ \eta\mathbf{I})$ where $\eta = 0.1$. There are 100 locations such that the total sample size is $100\times T$. The fitting results, which are based on 1000 replicates for $T = 1$, 500 for $T = 5$ and 200 for $T = 10$, are shown in Tables 1 and 2. As extremely large or extremely small estimates can be produced in estimating the covariance parameters, in addition to the sample mean and sample standard deviation we also include the median and a more robust standard deviation sd* to provide a more robust summary of the simulation results. Let IQR denote the interquartile range. Then the robust standard deviation is given by sd*=IQR/1.349, which estimates the standard deviation if the data are normally distributed. To evaluate the estimation of the smooth functions, we calculated the mean square errors of the fitted functions and the 95% confidence interval coverage proportion, which is the proportion of data points that are covered by their 95% confidence intervals.

Generally speaking, the REML estimation seems to be less stable than ML estimation in the sense that REML estimation can result in some extreme estimates for the spatial variance σ^2 and the range parameter ρ especially when the sample is not very large. On the other hand, the REML estimation tends to be less biased than the method ML. Also, the 95% CI coverage proportions from the REML estimation are uniformly higher and closer to the nominal level than those from the ML estimation. As the sample size increases, both the REML and ML estimation perform better with less bias and smaller variation. In order to see the loss when the spatial correlation is ignored, we also include in Table 2 (the last two columns) the 95% CI coverage and MSE for the GAM assuming the same mean structure but with independent errors. It is clear that the 95% CI coverage proportions are much lower while the MSEs are slightly larger than the cases where the spatial correlation is explicitly modeled.

Note that the smoothing parameters are estimated by minimizing the GCV score, although we have assumed fixed smoothing parameters in the theoretical analysis. Still, the simulation results are generally consistent with the theoretical findings. By comparing the spatio-temporal cases and the one time period cases, we can see clear advantages gained

Table 1. Covariance Parameter Estimation

		ML				REML			
Sample	True Value	mean	median	sd	sd*	mean	median	sd	sd*
	σ^2=1	0.520	0.477	0.229	0.208	34.88	0.991	291.7	0.772
100×1	$\upsilon = 2$	4.349	4.118	2.606	2.222	3.087	2.555	2.453	2.510
	$\rho = 0.045$	0.028	0.010	0.056	0.012	10.19	0.028	125.1	0.115
	$\eta = 0.1$	0.090	0.092	0.036	0.032	0.088	0.091	0.038	0.034
	σ^2=1	0.581	0.527	0.251	0.208	9.266	0.947	137.8	0.572
400×1	$\upsilon = 2$	3.200	3.023	1.459	1.318	2.302	2.075	1.123	1.083
	$\rho = 0.045$	0.027	0.017	0.028	0.014	4.315	0.042	108.0	0.050
	$\eta = 0.1$	0.101	0.101	0.010	0.010	0.099	0.100	0.011	0.011
	σ^2=1	0.903	0.900	0.155	0.146	1.002	0.999	0.176	0.169
100×5	$\upsilon = 2$	2.358	2.165	0.900	1.006	2.210	2.020	0.849	0.931
	$\rho = 0.045$	0.046	0.038	0.032	0.028	0.053	0.044	0.036	0.032
	$\eta = 0.1$	0.099	0.100	0.015	0.014	0.099	0.099	0.015	0.015
	σ^2=1	0.949	0.951	0.106	0.112	1.000	1.004	0.113	0.123
100×10	$\upsilon = 2$	2.174	2.137	0.590	0.512	2.117	2.081	0.576	0.506
	$\rho = 0.045$	0.046	0.040	0.026	0.018	0.049	0.043	0.027	0.019
	$\eta = 0.1$	0.100	0.100	0.011	0.010	0.099	0.100	0.011	0.009

Table 2. Smooth Function Estimation

		ML		REML		GAM	
		95% CI		95% CI		95% CI	
Sample	Function	Coverage	MSE	Coverage	MSE	Coverage	MSE
	$f_1 + f_2$	75.25	0.456	89.45	0.601	43.61	0.582
100×1	f_1	79.12	0.154	88.72	0.167	48.26	0.215
	f_2	76.32	0.156	86.69	0.165	47.56	0.221
	$f_1 + f_2$	78.42	0.406	90.63	0.435	26.91	0.574
400×1	f_1	81.69	0.133	90.35	0.139	29.08	0.216
	f_2	79.26	0.138	87.97	0.140	29.39	0.220
	$f_1 + f_2$	91.94	0.095	93.40	0.096	55.52	0.112
100×5	f_1	92.09	0.034	93.02	0.035	62.04	0.040
	f_2	92.92	0.032	93.99	0.032	60.91	0.042
	$f_1 + f_2$	93.58	0.050	94.23	0.050	57.09	0.056
100×10	f_1	92.40	0.019	92.90	0.019	63.33	0.022
	f_2	95.66	0.015	96.09	0.015	67.61	0.019

from repeated measurements in terms of reduction in both bias and variation in the estimation of covariance parameters, although part of the improvement is due to the larger total sample size. In particular, the REML estimation is much more stable and the bias problem bothering the ML estimation is not that problematic now. As for the estimation of smooth functions, there is substantial improvement in the 95% CI coverage proportions, especially for the ML estimation, while the mean squared prediction errors are much smaller. Also note that the results from REML and ML methods tend to be more and more similar as there are more repeated measurements, although for small to moderately large sample size, the REML method has, generally, less bias and coverage closer to the 95% nominal level than the ML method. All these results imply that the GAM with Matérn correlation structure is much more tractable in the spatio-temporal case than the one time period case.

6. Checking Temporal Independence

So far, we have assumed that the data are temporally independent. Next, we propose an approach to check the validity of this assumption.

Again, consider the spatio-temporal model (1) with fixed spatial design. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{n_0}$ be the spatial locations and $e_t(\mathbf{s})$ be the error term at location $\mathbf{s}$ and time t. Define $\bar{e}_t \equiv \frac{1}{n_0}\sum_{\mathbf{s}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} e_t(\mathbf{s})$. Then

$$\begin{aligned}
& \mathrm{Cov}(\bar{e}_{t_1}, \bar{e}_{t_2}) \\
= \; & \frac{1}{n_0^2}\mathrm{Cov}\Big(\sum_{\mathbf{u}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} e_{t_1}(\mathbf{u}), \sum_{\mathbf{v}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} e_{t_2}(\mathbf{v})\Big) \\
= \; & \frac{1}{n_0^2}\sum_{\mathbf{u}=\mathbf{s}_1}^{\mathbf{s}_{n_0}}\sum_{\mathbf{v}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} \mathrm{Cov}(e_{t_1}(\mathbf{u}), e_{t_2}(\mathbf{v})) \\
= \; & \frac{1}{n_0^2}\sum_{\mathbf{u}=\mathbf{s}_1}^{\mathbf{s}_{n_0}}\sum_{\mathbf{v}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} \tau(|t_1 - t_2|)K(|\mathbf{u}-\mathbf{v}|) \\
= \; & \tau(|t_1 - t_2|)\frac{\sum_{\mathbf{u}=\mathbf{s}_1}^{\mathbf{s}_{n_0}}\sum_{\mathbf{v}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} K(|\mathbf{u}-\mathbf{v}|)}{n_0^2} \qquad (7)
\end{aligned}$$

Since $\frac{\sum_{\mathbf{u}=\mathbf{s}_1}^{\mathbf{s}_{n_0}}\sum_{\mathbf{v}=\mathbf{s}_1}^{\mathbf{s}_{n_0}} K(|\mathbf{u}-\mathbf{v}|)}{n_0^2}$ is constant over time, $\mathrm{Cov}(\bar{e}_{t_1}, \bar{e}_{t_2})$ retains the temporal correlation structure of the data. Therefore, it is reasonable to check the assumption of independence across time by checking if there is autocorrelation among $\{\bar{\epsilon}_t,\ t = 1, 2, \ldots, T\}$, where $\bar{\epsilon}_t$ is the average residual over all sampling locations at time t.

Note that the above argument is approximately correct even for random spatial design provided that the number of sites is large and the sites are well spread out. To see this, suppose $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{n_0}$ are iid with density function $h(\mathbf{s})$. Then (7) converges to $\tau(|t_1 - t_2|)\int\int K(|\mathbf{u}-\mathbf{v}|)h(\mathbf{u})d\mathbf{u}h(\mathbf{v})d\mathbf{v}$ as $n_0 \to +\infty$.

7. Model Selection

In the Bayesian framework, the best model from the set of candidate models under consideration is the one that has the highest posterior probability

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)}$$

where $P(M_i)$ is the prior probability of the model M_i, D denotes the data, and $P(D) = \sum_i P(D|M_i)P(M_i)$ is the normalization constant. $P(D|M_i)$ is the marginal likelihood (also called the evidence) of the model M_i

$$P(D|M_i) = \int P(D|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta}$$

where $P(D|\boldsymbol{\theta}, M_i)$ is the likelihood of the parameters under model M_i, and $P(\boldsymbol{\theta}|M_i)$ is the prior probability (density) of $\boldsymbol{\theta}$ under model M_i. Given equal prior model probabilities, the posterior model probability of M_i is proportional to its marginal likelihood $P(D|M_i)$. The BIC is an approximation of twice the negative log of the marginal likelihood (see Schwarz, 1978).

Typically, the AIC (Akaike, 1973, 1974) and BIC are used for selecting parametric models. Here for generalized additive models, we propose a model selection criterion based on the marginal likelihood or evidence $\mathrm{E}(M) = \int P(D|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$ which selects the model that has the highest posterior probability. We shall treat the penalty as some prior information. The analysis will be conditional on the estimated smoothing parameters which are henceforth treated as fixed, known numbers throughout this section. Below, we shall suppress M from the preceding formula when the model is clear from the context. For spatially correlated data, the exact marginal likelihood is generally intractable. (It, however, admits a closed-form solution for the case of independent, Gaussian data). We propose to use the Laplace approximation (Tierney and Kadane, 1986, MacKay, 1988 and Wong, 1989) to derive approximate formula for the marginal likelihood.

7.1. GAMs with Correlated Data - ML Estimation

For simplicity, we consider the case of a GAM with spatially correlated data and no replication; extension to the case of multi-yearly data is straightforward. The model can be rewritten as

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ and $\boldsymbol{\theta} \in \Theta$ is the vector of the covariance parameters. The penalized log likelihood equals

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}.$$

Note that the smoothing parameters are absorbed into the matrix $\mathbf{S}$. We assume the priors of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are independent and $\boldsymbol{\theta}$ has a flat prior over Θ. Then the joint prior of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{|\mathbf{D}_+|^{1/2}}{(2\pi)^{m/2}}\exp\{-\frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\}$$

where m is the number of strictly positive eigenvalues of $\mathbf{S}$ and $\mathbf{D}_+$is the diagonal matrix with all those strictly positive eigenvalues of $\mathbf{S}$ arranged in descending order on the leading diagonal.

The marginal likelihood of the model then equals

$$\begin{aligned} & \int p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta})p(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta} \\ = & \frac{1}{(2\pi)^{(n+m)/2}}\int p^*(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta} \end{aligned}$$

where $p^*(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{|\mathbf{D}_+|^{1/2}}{|\boldsymbol{\Sigma}_\theta|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\}$. Let

$$\ell_{p^*} \equiv \log p^*(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{1}{2}\log|\mathbf{D}_+| - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}.$$

Let

$$\begin{aligned} \Lambda &= -\frac{\partial^2\ell_{p^*}}{\partial(\boldsymbol{\beta},\boldsymbol{\theta})\partial(\boldsymbol{\beta},\boldsymbol{\theta})'}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \begin{bmatrix} \mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\mathbf{S} & -\frac{\partial}{\partial\boldsymbol{\theta}'}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ -\frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X} & -\frac{\partial^2\ell_{p^*}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \end{bmatrix}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \end{aligned}$$

Since $E\left[\frac{\partial}{\partial\boldsymbol{\theta}'}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] = 0$, the off-diagonal blocks of Λ can be approximated by $\mathbf{0}$ as $\hat{\boldsymbol{\beta}}$ is close to the true value.

By Laplace's method, the marginal likelihood of the model equals

$$\begin{aligned} \mathrm{E} &\equiv \int p^*(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta} \approx p^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})\sqrt{\frac{(2\pi)^{k+l}}{|\Lambda|}} \\ &= \sqrt{\frac{(2\pi)^{k+l}|\mathbf{D}_+|}{|\boldsymbol{\Sigma}_{\hat{\theta}}||\Lambda|}}\exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) - \frac{1}{2}\hat{\boldsymbol{\beta}}'\mathbf{S}\hat{\boldsymbol{\beta}}\} \end{aligned}$$

where $k = \dim(\boldsymbol{\beta})$ and $l = \dim(\boldsymbol{\theta})$. Hence, the log marginal likelihood equals

$$\begin{aligned} \log\mathrm{E} \approx\ & \frac{k+l}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{D}_+| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\hat{\theta}}| \\ & -\frac{1}{2}\log|\Lambda| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) - \frac{1}{2}\hat{\boldsymbol{\beta}}'\mathbf{S}\hat{\boldsymbol{\beta}}. \end{aligned}$$

For the case of GAM without spatial correlation, the log marginal likelihood is obtained by the preceding formula but with $\Sigma_{\hat{\theta}}$ replaced by the identity matrix, $\ell = 0$ and Λ reduced to its first diagonal block. However, the formula is exact for the independent case.

7.2. GAMs with Correlated Data - REML Estimation

In the previous section, we apply the Laplace method to the $(\boldsymbol{\beta},\boldsymbol{\theta})$ parameterization which generally operates on a very high-dimensional parameter space as the dimension of $\boldsymbol{\beta}$ is usually high. A more accurate approximation may be obtained by first integrating out

$\boldsymbol{\beta}$ from the posterior density before applying the Laplace approximation. We derive this alternative approach in this sub-section, and show that the Laplace approximation is equivalently obtained by a second-order Taylor expansion around the REML estimator of $\boldsymbol{\theta}$. We employ the same prior distribution used in the previous sub-section.

The marginal likelihood of the model equals

$$\begin{aligned} & \int p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta})p(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta} \\ = & \frac{1}{(2\pi)^{(n+m)/2}}\int p^*(\boldsymbol{\theta})d\boldsymbol{\theta} \end{aligned}$$

where $k = \dim(\boldsymbol{\beta})$ and

$$p^*(\boldsymbol{\theta}) = \sqrt{\frac{(2\pi)^k|\mathbf{D}_+|}{|\boldsymbol{\Sigma}_\theta||\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\mathbf{S}|}}\exp\{-\frac{1}{2}\mathbf{y}'[\boldsymbol{\Sigma}_\theta^{-1}-\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}]\mathbf{y}\}.$$

Let

$$\begin{aligned} \ell_{p^*} \equiv & \log p^*(\boldsymbol{\theta}) \\ = & \frac{k}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{D}_+|-\frac{1}{2}\log|\boldsymbol{\Sigma}_\theta|-\frac{1}{2}\log|\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\mathbf{S}| \\ & -\frac{1}{2}\mathbf{y}'[\boldsymbol{\Sigma}_\theta^{-1}-\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}]\mathbf{y}\}, \end{aligned}$$

and

$$\Lambda = -\frac{\partial^2\ell_{p^*}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

By Laplace's method, the marginal likelihood of the model becomes

$$\begin{aligned} \mathrm{E} \equiv & \int p^*(\boldsymbol{\theta})d\boldsymbol{\theta} \approx p^*(\hat{\boldsymbol{\theta}})\sqrt{\frac{(2\pi)^4}{|\Lambda|}} \\ = & \sqrt{\frac{(2\pi)^{k+l}|\mathbf{D}_+|}{|\Lambda||\boldsymbol{\Sigma}_{\hat{\theta}}||\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X}+\mathbf{S}|}}\exp\{-\frac{1}{2}\mathbf{y}'[\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}-\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}]\mathbf{y}\} \end{aligned}$$

where $l = \dim(\boldsymbol{\theta})$. Hence, the log marginal likelihood equals

$$\begin{aligned} \log\mathrm{E} \approx & \frac{k+l}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{D}_+|-\frac{1}{2}\log|\boldsymbol{\Sigma}_{\hat{\theta}}|-\frac{1}{2}\log|\Lambda|-\frac{1}{2}\log|\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X}+\mathbf{S}| \\ & -\frac{1}{2}\mathbf{y}'[\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}-\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}]\mathbf{y}. \end{aligned}$$

7.3. Simulation Study

Although the model selection criteria derived in the previous section can be used to compare nested or non-nested models, we are particularly interested in choosing between GAMs with and without spatial correlation in this simulation study.

For the spatially correlated GAM, we simulate data from the following model

$$Y(x,y) = f_1(x)+f_2(y)+b(x,y)+e(x,y),$$

Table 3. Proportion of Times that the Proposed Criterion Selects the True Model

	Variance	$\eta = 1.1$	$\eta = 1.5$	$\eta = 2$	$\eta = 11$
Independent Data	P(logE.gam>logE.ml)	0.755	0.830	0.825	0.865
	P(logE.gam>logE.reml)	0.790	0.785	0.835	0.865
	Nugget	$\eta = 0.1$	$\eta = 0.5$	$\eta = 1$	$\eta = 10$
Correlated Data	P(logE.ml>logE.gam)	0.955	0.682	0.613	0.369
	P(logE.reml>logE.gam)	0.964	0.839	0.728	0.352

where $x \in [0, 1]$ and $y \in [0, 1]$ are the coordinates of a data point; $f_1(x) = 1 + 2x$; $f_2(y) = -3y$; b is the spatially correlated error with Matérn covariogram with $\sigma^2 = 1$, $\nu = 2$, $\rho = 0.045$; and e is the pure measurement error (or nugget effect) from the normal distribution with mean 0 and variance η where η is varied from 0.1, 0.5, 1 and 10. The sample size is 200. We also simulated data from the GAM with the same mean structure but with spatially independent noise, i.e. b is absent but the nugget effect ranges from 1.1, 1.5, 2 and 11 to match the overall noise variance of the spatially correlated GAM.

To each simulated dataset, we fitted three models: 1) a GAM model assuming independent data; 2) a GAM model assuming Matérn-correlated data fitted by ML estimation; and 3) a GAM model assuming Matérn-correlated data fitted by REML estimation. The log marginal likelihood is calculated for each case, and let logE.gam, logE.ml and logE.reml denote the log marginal likelihood respectively for the three models.

The above procedure is repeated 200 times independently. Table 3 displays the relative frequencies that the criterion of maximum marginal model probability picks the true model, i.e., logE.gam is greater than logE.ml or logE.reml for the independent data cases and vice versa for the correlated data cases. When the data are actually independent and we fit a gam model with spatial correlation, the fitting procedure tends to fail because the Matern covariogram is non-identifiable. When the estimation of the correlated GAM fails while that of the independent GAM succeeds, we count the logE.gam as the largest value since logE.ml and logE.reml are missing.

It can be seen that the proposed criterion works well in picking the correct model for the independent data. For the correlated data, the criterion also has good chances to select the true model, especially when the nugget is relatively small comparing to the spatial variance. As the nugget increases, the proposed criterion has slightly better chances to select the true model when the data are independent and lower chances to pick the true model when the data are correlated. This is, however, expected. As the nugget increases, the spatial correlation is increasingly irrelevant and thus increasingly difficult to be detected. For the extreme case where the nugget is ten times of the spatial variance, the data are actually more like independent data and we have only about one third chance of selecting the true model. Also, it seems the estimation method (ML or REML) has little effect on model selection when the data are independent, but for correlated data, REML estimation seems to result in a more powerful model selection criterion.

Next, we extend the simulation to include spatio-temporal data from a GAM with the same mean structure that are independent across different time periods, but with either spatially correlated or spatially independent noise within each time period. With fixed spatial design, temporally independent repeated measurements are taken at each location such that the total sample size is still 200. Specifically, we simulated data from two cases: (i) 50 data points over 4 time periods and (ii) 20 data points over 10 time periods. For all the spatially correlated data, the parameters of the Matérn covariogram are $\sigma^2 = 1, \nu =$

Table 4. Proportion of Times that the Proposed Criterion Selects the True Model: Spatio-Temporal Data

	Sample Distribution	200×1	50×4	20×10	200×2
Independent	P(logE.gam>logE.ml)	0.755	0.845	0.865	0.790
Data	P(logE.gam>logE.reml)	0.790	0.855	0.835	0.795
Correlated	P(logE.ml>logE.gam)	0.955	0.918	0.815	1
Data	P(logE.reml>logE.gam)	0.964	0.928	0.793	1

$1, \rho = 0.045$ with nugget 0.1. For the spatially independent data, the error variance is 1.1. As we mentioned in Section 5, the smoothing parameters are estimated by minimizing the GCV score, although we have assumed fixed smoothing parameters in the theoretical analysis.

Table 4 shows the results, which, again, are based on 200 replicates. In general, the proposed criterion performs well with at least 75% chance of success in picking the correct model for all the cases studied here. The criterion works particularly well when the data are correlated. Even for the case with only 20 fixed data sites, the proposed criterion can select the right model in about 80% of the times. As there are more repeated measurements with total sample size fixed, the criterion gains power from repeated measurements for the independent data, but for the spatially correlated data, the performance of the criterion gets relatively worse. This may be due to two reasons. First, that the data are independent across time eliminates some spatial correlation comparing to the one-period case. Second, as the sample size per time period decreases, there is some information loss in estimating the mean structure. Taking these reasons into account, these results are not surprising. In fact, if we only have 20 data points and one time period, we may even have difficulty in fitting the GAM model due to the small sample size. The benefit from repeated measurements can be clearly seen for the correlated data when comparing the 200×1 and the 200×2 cases. It is interesting to note that there is no substantial improvement for the independent data when adding one more observation at each location. Actually, this observation holds even if the spatial design is random across year (results unreported). This suggests that the proposed model selection criterion may not be consistent under the true model of independent GAM. We conjecture that this inconsistency arises from the fact that the independent covariance structure is a special case of the Matérn structure and that the Matérn model can mimic the independent case. Even if the data are actually independent, the Matérn model may fit the data well with either spatial variance or effective range close to zero. As a result, it is hard to distinguish the Matérn model from the independent covariance structure when the data are independent even with a larger total sample size.

8. Case Study

The pollock egg density data were collected during the ichthyoplankton survey of the Alaska Fisheries Science Center in the Gulf of Alaska. Ciannelli *et al.* (2007) studied the phenological and geographical patterns of walleye pollock spawning in the western Gulf of Alaska based on the spatial and temporal distribution of pollock eggs. They fitted a threshold generalized additive model (TGAM) to the data from 1972 to 2000 (before 1981, only data from 1972, 1978 and 1979 are available), and found that there was a shift of egg abundance distribution at the end of 1980s. However, in their analysis they did not take the spatial correlation into account. As shown by simulations in Section 5, the inference may be invalid

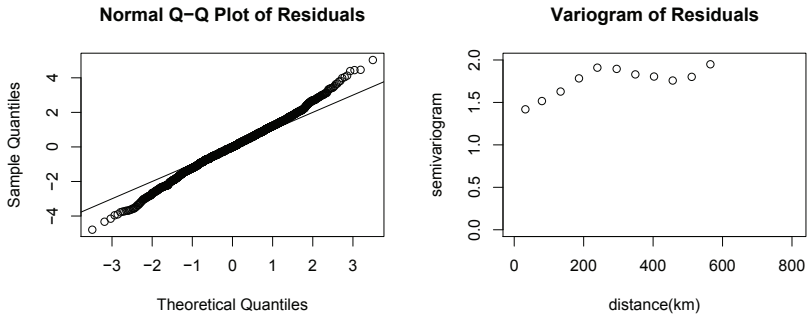

Fig. 1. Diagnosis of the Fitted Residuals from GAM

if there does exist spatial correlation, but it is ignored in the model. For our purpose of illustration, we only focus on the data before 1990 so as to avoid the shift and at the same time keep more years of data.

The response variable here is the pollock egg density which is defined as the average number of eggs per 10 square meters. The explanatory variables of interest include the sampling position defined by longitude and latitude, the bottom depth in meters, and the Julian day of sampling. Samples with zero density or missing values in any variable were removed and there were 2093 data points left for the analysis. The log transformation was conducted on the response variable to normalize the distribution and reduce heteroscedasticity. Bottom depth was also log-transformed to allow a uniform distribution throughout the sampled depth range. The longitude and latitude are in degrees. In order to make it easier to interpret the distance, the locations are transformed into the Universal Transverse Mercator (UTM) coordinate system in kilometers, i.e. distance between sampling sites are measured in terms of geodesic distance.

In modeling the mean structure, an intercept term is added for each year, in addition to all the predictor variables: position (longitude, latitude), bottom depth, and Julian day of sampling. We assume that data from different years are independent after the year effect has been accounted for. Specifically, let $Y_t(x, y)$ be the natural logarithm of egg density at UTM location (x, y) in year t. Then the proposed model can be written as follows.

$$Y_t(x, y) = \mathrm{YEAR}_t + f_1(x, y) + f_2(\mathrm{DEPTH}_{(x,y)}) + f_3(\mathrm{DATE}) + b_t(x, y) + e_t(x, y),$$

where YEAR_t is the intercept allowed to change from year to year, $\mathrm{DEPTH}_{(x,y)}$ is the log-transformed bottom depth at location (x, y), and DATE is the Julian day of sampling. $\mathbf{b}_t$ are the spatially correlated errors with distribution $\mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$, and $\mathbf{e}_t$ are the independent measurement errors with distribution $\mathbf{N}(\mathbf{0}, \eta\mathbf{I})$.

First, a GAM assuming (spatially and temporally) independent errors is fitted (i.e., $b_t(x, y)$ is ignored) . Figure 1 shows the diagnosis plots of the fitted residuals. The variogam indicates that there might be spatial correlation among the residuals, although the correlation may not be very strong. The normal Q-Q plot also implies that the distribution of the fitted residuals is not iid normal. Thus a GAM with independent errors seems to be inadequate for the pollock egg data.

Then we fit a GAM with spatially correlated, but temporally uncorrelated errors by both ML and REML estimation. We employed the Matérn covariogram for modeling the spatial

Table 5. Estimated Covariance Parameters

Method	$\widehat{\sigma^2}$	$\widehat{\nu}$	$\widehat{\rho}$	$\widehat{\eta}$
ML	0.488	6.141	1.313	1.337
	(0.081)	(8.620)	(2.897)	(0.049)
REML	0.755	2.181	8.984	1.338
	(0.142)	(1.361)	(9.983)	(0.050)

Table 6. p-Values of Ljung-Box Test for Temporal Independence

Method	lag 1	lag 2	lag3	lag 4	lag 5	lag 6
ML	0.201	0.264	0.159	0.174	0.234	0.328
REML	0.254	0.312	0.188	0.183	0.274	0.384

correlation. The results are shown in Figures 2 and 3. The variograms and the normal Q-Q plots show that the standardized residuals are approximately iid normal, which means that the Matérn covariogram has adequately explained almost all the spatial correlation. There is not much difference in the estimated smooth functions between the ML and REML estimation methods. All the smooth terms are highly significant. In general the pollock egg density tends to increase as the bottom depth increases. Also, there were less pollock eggs on early and late sampling days. Besides the seasonal effect, this may be an artifact due to the traveling route of the cruises. The estimated smooth functions are similar to those obtained from the function fit in Ciannelli *et al.* (2007) that assumes no spatial correlation. However, the confidence bands of the estimated functions are wider than what Ciannelli *et al.* (2007) got, which may indicate that the covariance matrix of the regression coefficients is underestimated due to ignoring the spatial correlation. The estimated covariance parameters are summarized in Table 5. The variogram estimated from REML has larger spatial variance and larger effective range, which is consistent with what is already seen in the simulation study.

For the three models mentioned above, the log marginal likelihoods for the three models are respectively $\log \text{E.gam} = -1685.999$, $\log \text{E.ml} = -1614.990$, $\log \text{E.reml} = -1605.928$, which indicates that the models with spatial correlation are preferred and the REML estimation is slightly better then the ML estimation.

So far, we have assumed temporal independence. To check this assumption, the Ljung-Box test (Ljung and Box, 1978) is used to test for independence in spatially averaged residuals. The Ljung-Box test statistic is calculated as $Q = T(T+2)\sum_{k=1}^{s} r_k^2/(T-k)$, where T is the number of observations, s is the number of coefficients to test for autocorrelation, and r_k is the autocorrelation coefficient (for lag k). If the sample value of Q exceeds the critical value of a chi-square distribution with s degrees of freedom, then at least one value of r_k is statistically different from zero at the specified significance level. The Null Hypothesis is that none of the autocorrelation coefficients up to lag s are different from zero. Table 6 lists the test p-values for different maximum lag values. All the p-values are greater than 0.15, implying that the null hypothesis of independence in the spatially averaged residuals is not rejected at the significance level of 0.05. Thus it is reasonable to assume that the pollock egg data are independent across years.

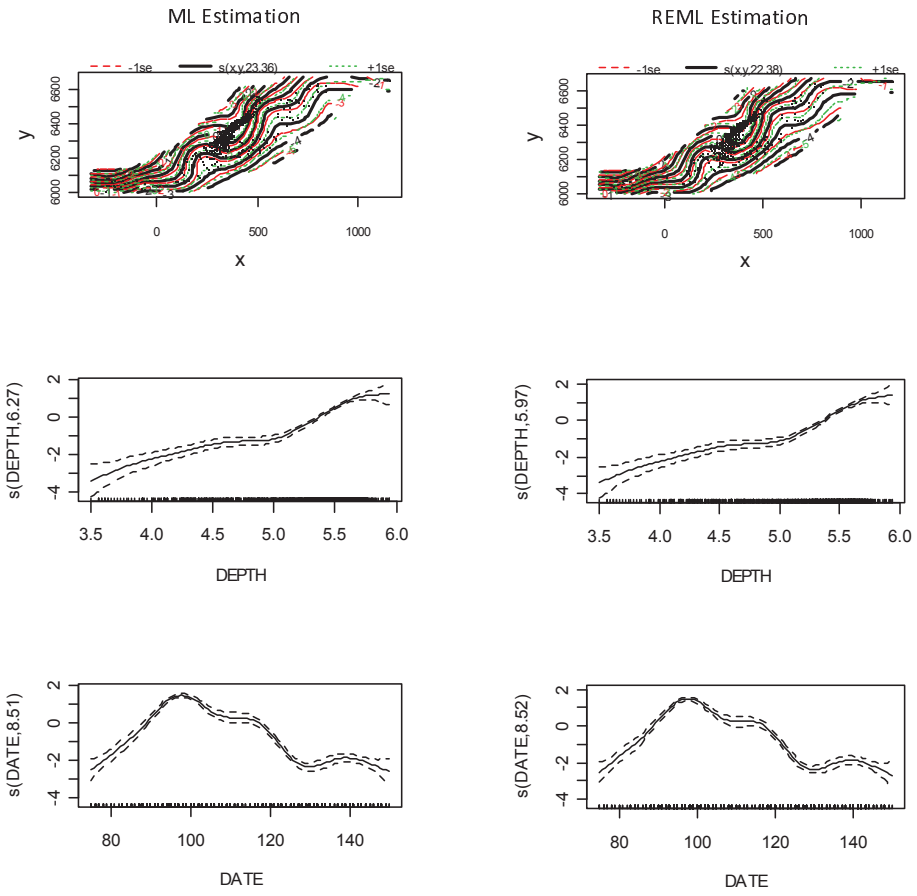

Fig. 2. Estimated Smooth Functions

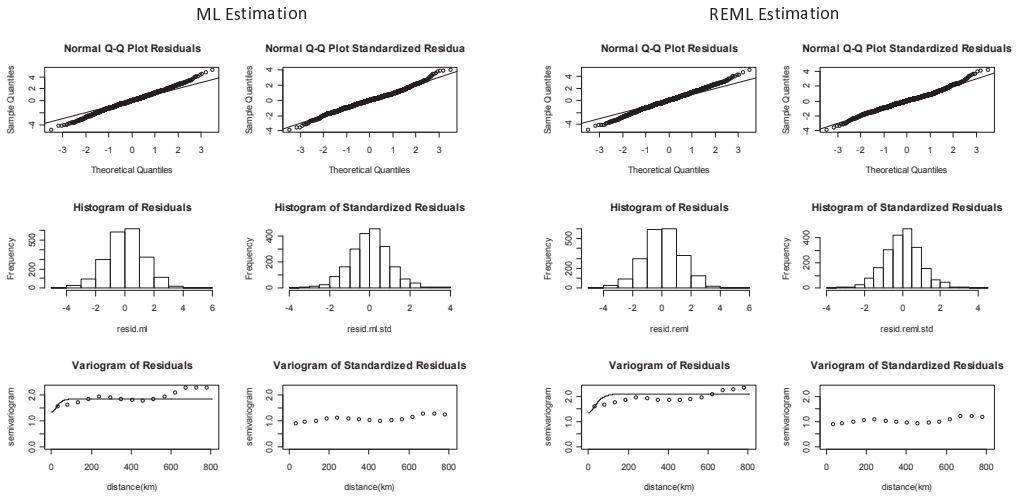

Fig. 3. Diagnosis on Residuals

9. Conclusion and Discussion

In this paper, we developed a new approach to fit the (generalized) additive models with correlated data under the GAM framework, as an alternative to the GAMM framework which turns out to have some numerical problems. Meanwhile, we also studied the properties of the Matérn correlation model under the GAM framework. In general, the REML estimation tends to be better than ML in terms of bias and confidence interval coverage for the smooth functions, but the ML estimation of the covariance parameters is more stable. However, it can be expected that the two methods tend to perform similarly as the sample size is large enough. The variations of the estimates sometimes seem to be large, but this is not surprising since many authors have reported the difficulties in likelihood estimation of covariance parameters (Mardia and Watkins, 1989; Diggle *et al.*, 1998; Zhang, 2002).

As some of the parameters of the Matérn model can not be consistently estimated under fixed domain asymptotics, we investigated the spatio-temporal case where the spatial design is assumed to be fixed with temporally independent repeated measurements and the spatial correlation structure does not change over time, and outlined the conditions under which the asymptotic posterior normality holds. Although it is hard to verify if those conditions are satisfied for the Matérn model, simulation study indicates that this may be the case.

We have been focusing on the Gaussian type data, i.e., the additive models, in this paper. It is of interest to make the methodology workable for all GAMs with correlated data. Also, all the inference has assumed fixed smoothing parameters. It would be nice to incorporate the randomness of the smoothing parameters.

References

[1] Abe, M. (1999). A Generalized Additive Model for Discrete-Choice Data. *Journal of Business & Economic Statistics* **17(3)**, 271-284.

[2] Abramowitz, M. and Stegun, I.A. (1972). Modified Bessel Functions I and K. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*. New York: Dover, 374-377.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281.

[4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716-723.

[5] Akaike, H. (1978a). A new look at the Bayes procedure. *Biometrika* **65**, 53-59.

[6] Bai, Z. D., Krishnaiah, P. R., Sambamoorthi, N., and Zhao, L. C. (1992) Model selection for log-linear model. *Sankhya B* **54**, 200-219.

[7] Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

[8] Chen, C.-F. (1985). On Asymptotic Normality of Limiting Density Functions with Bayesian Implications. *Journal of the Royal Statistical Society. Series B (Methodological)* **47(3)**, 540-546.

[9] Ciannelli, L., Bailey, K. M., Chan, K.-S., and Stenseth, N. C. (2007). Phenological and geographical patterns of walleye pollock (Theragra chalcogramma) spawning in the western Gulf of Alaska. *Canadian Journal of Fisheries and Aquatic Sciences* **64**, 713-722.

[10] Corbeil, R.R. and Searle, S.R. (1976). Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *Technometrics* **18(1)**, 31-38.

[11] Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-Based Geostatistics. *Journal of the Royal Statistical society, Series C* **47**, 299-350.

[12] Diggle, P. J., Ribeiro, P. J. and Christensen, O. F. (2002). An Introduction to Model-Based Geostatistics. *Spatial Statistics and Computational Methods*, ed. J. Møller. New York: Springer-Verlag, 43-86.

[13] Dominici1, F., McDermott1, A., Zeger, S.L. and Samet, J.M. (2002). On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health. *American Journal of Epidemiology* **156(3)**, 193-203.

[14] Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731-761.

[15] Fahrmeir, L. and Lang, S. (2001). Bayesian Inference for Generalized Additive Mixed Models based on Markov Random Field Priors. *Journal of the Royal Statistical Society C* **50**, 201-220.

[16] Frescino, T.S., Edwards, T.C. and Moisen, G.G. (2001). Modeling Spatially Explicit Forest Structural Attributes Using Generalized Additive Models. *Journal of Vegetation Science* **12(1)**, 15-26.

[17] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

[18] Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

[19] Guisan, A., Edwards, T.C. and Hastie, T.J. (2002). Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene. *Ecological Modelling* **157(2-3)**, 89-100.

[20] Hastie, T.J. and Tibshirani, R.J. (1986). Generalized Additive Models. *Statistical Science* **1**, 297-318.

[21] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman & Hall/CRC.

[22] Hastie, T.J. and Tibshirani, R.J. (1995). Generalized Additive Models for Medical Research. *Statistical Methods in Medical Research* **4(3)**, 187-196.

[23] Heyde, C. C. and Johnstone, I. M. (1979). On Asymptotic Posterior Normality for Stochastic Processes. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(2), 184-189.

[24] Hurvich, C. M. and Tsai, C-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499-509.

[25] Lehmann, A. (1998). GIS Modeling of Submerged Macrophyte Distribution Using Generalized Additive Models. *Plant Ecology* **139(1)**, 113-124.

[26] Lin, X. and Zhang, D. (1999). Inference in Generalized Additive Mixed Models by Using Smoothing Splines. *Journal of Royal Statistical Society: Series B* **61**, 381-400.

[27] Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 553C564.

[28] MacKay, D. J. C. (1998). Choice of basis for Laplace approximation. *Machine Learning* **33(1)**, 77-86.

[29] Mardia, K. V. and Watkins, A. J. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289-295.

[30] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*. **6(2)**, 461-464.

[31] Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer-Verlag.

[32] Sweeting, T. J. (1992). On Asymptotic Posterior Normality in the Multiparameter Case. *Bayesian Statistics*. **4**, 825-835.

[33] Sweeting, T. J. and Adekola, A. O. (1987). Asymptotic Posterior Normality for Stochastic Processes Revisited. *Journal of the Royal Statistical Society. Series B (Methodological)* **49(2)**, 215-222.

[34] Tierney, L. and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association* **81(393)**, 82-86.

[35] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[36] Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

[37] Walker, A. M. (1969). On the Asymptotic Behaviour of Posterior Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **31(1)**, 80-88.

[38] Williams, B. J., Santner, T.J. and Notz, W.I. (2000). Sequential Design of Computer Experiments to Minimize Integrated Response Functions. *Statistica Sinica* **10**, 1133-1151.

[39] Wong, R. (1989). Asymptotic Approximations of Integrals. Academic Press, San Diego.

[40] Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida: Chapman & Hall/CRC.

[41] Ying, Z., (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis* **36(2)**,280-296.

[42] Zeger, S.L. and Diggle, P.J. (1994). Semi-parametric Models for Longitudinal Data with Applications to CD4 Cell Numbers in HIV Seroconverters. *Biometrics* **50**, 689-699.

[43] Zhang, H. (2002). On Estimation and Prediction for Spatial Generalized Linear Mixed Models. *Biometrics* **56**, 129-136.

[44] Zhang, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of American Statistical Association* **99**, 250-262.

[45] Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semi-parametric Stochastic Mixed Models for Longitudinal Data. *Journal of American Statistical Association* **93**, 710-719.

[46] Zhu, Z. and Zhang, H. (2006). Spatial Sampling Design under the Infill Asymptotics Framework. *Environmetrics* **17**, 323-337.

Proof of Theorem 1

We begin the proof by restating Sweeting's results.

Let $(\Omega_T, \mathcal{A}_T)$ be a family of measurable spaces, where $T \in \mathcal{T}$ is a discrete or continuous time parameter. Let $\mathbf{x}_T \in \Omega_T$ be the observed data up to and including time T. Let P_ϕ^T be the corresponding probability measures defined on $(\Omega_T, \mathcal{A}_T)$, where the parameter $\boldsymbol{\phi} \in \Phi$, an open subset of $\mathbb{R}^p$. Assume that, for each $T \in \mathcal{T}$ and $\boldsymbol{\phi} \in \Phi$, P_ϕ^T is absolutely continuous with respect to a σ-finite measure μ_T and let $p_T(\mathbf{x}_T|\boldsymbol{\phi})$ be the associated density of P_ϕ^T. The log-likelihood function $l_T(\boldsymbol{\phi}) = \log p_T(\mathbf{x}_T|\boldsymbol{\phi})$ is assumed to exist a.e. (μ_T). Let $\mathcal{U}_T(\boldsymbol{\phi}) = l_T^{'}(\boldsymbol{\phi})$ be the vector of first-order partial derivatives of $l_T(\boldsymbol{\phi})$ w.r.t. $\boldsymbol{\phi}$ and define $\mathbf{J}_T(\boldsymbol{\phi}) = -l_T^{''}(\boldsymbol{\phi})$, the observed information matrix at $\boldsymbol{\phi}$. Let M_p be the space of all real $p \times p$ matrices and M_p^+ be the space of all $p \times p$ positive definite matrices. Let $\lambda_{\max}(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$ denote the maximum and minimum eigenvalues of a symmetric matrix $\mathbf{A} \in M_p$. The spectral norm $\|\cdot\|$ in M_p is $\|\mathbf{A}\|^2 = \sup(|\mathbf{Ax}|^2 : |\mathbf{x}|^2 = 1) = \lambda_{\max}(\mathbf{A'A})$. The matrix $\mathbf{A}^{1/2}$ will denote the left Cholesky square root of $\mathbf{A}$ in M_p^+. Let $\boldsymbol{\phi}_0$ be the true underlying parameter value. Write $\mathbf{J}_{T0} = \mathbf{J}_T(\boldsymbol{\phi}_0)$, $\mathcal{U}_{T0} = \mathcal{U}_T(\boldsymbol{\phi}_0)$, and $\mathbf{W}_T = \mathbf{B}_T^{-1/2}\mathbf{J}_{T0}(\mathbf{B}_T^{-1/2})'$ where $\mathbf{B}_T$ are $\mathcal{A}_T$-measurable matrices in M_p^+. The matrices $\mathbf{B}_T$ are chosen such that the sequence $(\mathbf{W}_T)$ is stochastically bounded in M_p^+. Let $N_T^*(c) = \left\{\boldsymbol{\phi} : |(\mathbf{B}_T^{1/2})'(\boldsymbol{\phi} - \boldsymbol{\phi}_0)| < c\right\}$ and $\Delta_T^*(c) = \sup_{\boldsymbol{\phi} \in N_T^*(c)} \|\mathbf{B}_T^{-1/2}(\mathbf{J}_T(\boldsymbol{\phi}) - \mathbf{J}_T(\boldsymbol{\phi}_0))(\mathbf{B}_T^{-1/2})'\|$.

Consider the following conditions:

C1. (Prior distribution) The prior distribution of $\boldsymbol{\phi}$ is absolutely continuous with respect to Lebesgue measure, with prior density $\pi(\boldsymbol{\phi})$ continuous and positive through Φ and zero on Φ^c.

D2. (Smoothness) The log-likelihood function $l_T(\boldsymbol{\phi})$ is a.e. (μ_T) twice differentiable with respect to $\boldsymbol{\phi}$ throughout Φ.

D3. (Compactness) $(\mathbf{B}_T^{-1/2}\mathcal{U}_{T0})$ is stochastically bounded.

D4. (Information growth) $\mathbf{B}_T^{-1} \xrightarrow{p} 0$.

D5. (Information continuity) $\Delta_T^*(c) \xrightarrow{p} 0$ for every $c > 0$.

D6. (Nonlocal behavior) For each $T \in \mathcal{T}$ there exists a non-random open convex set C_T containing $\boldsymbol{\phi}_0$ which satisfies

(i) $P^T(\mathbf{J}_T(\boldsymbol{\phi})) > 0$ on $C_T) \to 1$.

(ii) $\pi(\boldsymbol{\phi})$ is eventually bounded on C_T.

(iii) $\{p_T(\mathbf{x}_T|\boldsymbol{\phi}_0)\}^{-1}|\mathbf{B}_T|^{1/2} \int_{\boldsymbol{\phi} \notin C_T} p_T(\mathbf{x}_T|\boldsymbol{\phi})\pi(\boldsymbol{\phi})d\boldsymbol{\phi} \xrightarrow{p} 0$.

The following two lemmas of Sweeting (1992) are helpful for checking **D6**(iii).

Lemma 1 Assume conditions D2 - D6(i) with $\mathbf{B}_T = \mathbf{J}_{T0}$. Then (i) with probability tending to one as $T \to \infty$, there is a unique solution $\hat{\boldsymbol{\phi}}_T$ of $l_T^{'}(\boldsymbol{\phi}) = 0$ in C_T at which point $l_T(\boldsymbol{\phi})$ assumes its maximum value over this region; and (ii) $(\mathbf{J}_{T0}^{1/2})'(\hat{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_0)$ is stochastically bounded.

Note that the result (ii) of Lemma 5.1 implies that the penalized likelihood estimator of $\boldsymbol{\phi}$ is consistent with $1/\sqrt{T}$ converge rate.

Lemma 2 Assume conditions **C1, D2 - D5** and **D6**(i), (ii). Suppose further that $C_T \supset C$ where C is a fixed neighborhood of $\boldsymbol{\phi}_0$ and that, with probability tending to one, $\sup_{\boldsymbol{\phi} \in R_T - C_T} p_T(\mathbf{x}_T|\boldsymbol{\phi}) = \sup_{\boldsymbol{\phi} \in \partial C_T} p_T(\mathbf{x}_T|\boldsymbol{\phi})$, where R_T is some non-random neighborhood of $\boldsymbol{\phi}_0$ and ∂C_T denotes the boundary of C_T, i.e., $\partial C_T = \bar{C}_T \cap \overline{C_T^c}$. Then if

$(\{\lambda_{\min}(\mathbf{B}_T)\}^{-1}\log\lambda_{\max}(\mathbf{B}_T))$ is stochastically bounded we have

$$Q_T \equiv \{p_T(\mathbf{x}_T|\boldsymbol{\phi}_0)\}^{-1}|\mathbf{B}_T|^{1/2}\int_{\boldsymbol{\phi}\in R_T - C_T} p_T(\mathbf{x}_T|\boldsymbol{\phi})\pi(\boldsymbol{\phi})d\boldsymbol{\phi} \xrightarrow{p} 0.$$

Here is the main result of Sweeting (1992).

Theorem 2 Assume conditions **C1, D2 - D6**. Then

$$f_T(\mathbf{z}|\mathbf{x}_T) = |\mathbf{J}_T|^{-1/2}\pi_T(\hat{\boldsymbol{\phi}}_T + (\mathbf{J}_T^{-1/2})'\mathbf{z}|\mathbf{x}_T) \xrightarrow{p} (2\pi)^{-p/2}\exp\{-|\mathbf{z}|^2/2\} \text{ as } T\to\infty$$

where $f_T(\mathbf{z}|\mathbf{x}_T)$ is the posterior density of $\mathbf{Z}_T = (\mathbf{J}_T^{1/2})'(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}}_T)$, $\mathbf{J}_T = \mathbf{J}_T(\hat{\boldsymbol{\phi}}_T)$, and $\pi_T(\boldsymbol{\phi}|\mathbf{x}_T)$ is the posterior density of $\boldsymbol{\phi}$ based on the data up to time T.

The theorem above implies the convergence of the posterior distribution of $\mathbf{Z}_T$ to the standard p-dimensional normal distribution. In other words, the posterior of $\boldsymbol{\phi}$ is asymptotically multivariate normal with mean $\hat{\boldsymbol{\phi}}_T$ and variance matrix $\mathbf{J}_T^{-1}$.

Now, we are ready to prove Theorem 1 by verifying the conditions of Theorem 2.

C1. The prior of $\boldsymbol{\phi}$, $p(\boldsymbol{\phi}) = \frac{|\mathbf{D}_+|^{1/2}}{(2\pi)^{m/2}}\exp\{-\frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\}$, which indicates $\boldsymbol{\phi}$ has flat prior on Θ, clearly satisfies **C1**.

D2. The penalized log-likelihood equals

$$l_T(\boldsymbol{\phi}) = -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) - \frac{T}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}.$$

Clearly the log-likelihood is twice differentiable with respect to $\boldsymbol{\beta}$. Thus **D2** holds as long as $\boldsymbol{\Sigma}_\theta$ or the covariogram function $K(\cdot|\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$.

D3. Let $\mathbf{B}_T = T\mathbf{I}_{k+l}$. Then $\mathbf{W}_T = \frac{1}{T}\mathbf{J}_{T0} \xrightarrow{a.s.} \mathcal{I}_{n_0}(\boldsymbol{\phi}_0)$, where n_0 is the sample size for each time period and $\mathcal{I}_{n_0}$ is the expected Fisher information for one time period. Thus $(\mathbf{W}_T)$ is stochastically bounded. Since $E_{\phi_0}[\mathbf{B}_T^{-1/2}\mathbf{J}_{T0}(\mathbf{B}_T^{-1/2})] = \frac{1}{T}E_{\phi_0}[\mathbf{J}_{T0}] = \mathcal{I}_{n_0}(\boldsymbol{\phi}_0)$ and $E_{\phi_0}[\mathbf{J}_{T0}] = E_{\phi_0}[\mathcal{U}_{T0}\mathcal{U}_{T0}']$, then $E_{\phi_0}|\mathbf{B}_T^{-1/2}\mathcal{U}_{T0}|^2 = \mathrm{tr}(\mathcal{I}_{n_0}(\boldsymbol{\phi}_0))$ and **D3** holds.

D4. It is trivial that **D4** holds.

Before moving on to **D5**, we first give a lemma which is from Example 19.8 in van der Vaart (1998) (p272).

Lemma 3 Let $\mathcal{F} = \{f_\theta : \theta\in\Theta\}$ be a collection of measurable functions with integrable envelope function F indexed by a compact metric space Θ such that the map $\theta\mapsto f_\theta(x)$ is continuous for every x. Then $\mathcal{F}$ is Donsker and hence the law of large numbers and central limit theorem hold uniformly in f ranging over $\mathcal{F}$.

D5. $\Delta_T^*(c) = \sup_{\phi\in N_T^*(c)}\|\frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}) - \frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}_0)\|$, where

$$N_T^*(c) = \left\{\boldsymbol{\phi} : |\sqrt{T}(\boldsymbol{\phi}-\boldsymbol{\phi}_0)| < c\right\}.$$

Let $f_\phi(\mathbf{Y}_t) = -\frac{\partial^2 l(\boldsymbol{\phi}|\mathbf{Y}_t)}{\partial\boldsymbol{\phi}\boldsymbol{\phi}'}$. Under the assumption that $l_T(\boldsymbol{\phi})$ is twice differentiable, the map $\boldsymbol{\phi}\mapsto f_\phi(\mathbf{Y}_t)$ is continuous. Suppose $\exists\ \delta_0 > 0$ and a finite, integrable function $M(\mathbf{Y}_t)$ such that $\sup_{|\phi-\phi_0|<\delta_0}\|f_\phi(\mathbf{Y}_t)\|_{\max} \le M(\mathbf{Y}_t)$ where $\|\cdot\|_{\max}$ is the maximum norm. Then by Lemma 3, $\frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}) = \frac{1}{T}\sum_{t=1}^{T} f_\phi(\mathbf{Y}_t) \xrightarrow{a.s.} E_{\phi_0}f_\phi(\mathbf{Y}_t)$ uniformly as $T\to+\infty$. Since $N_T^*(c)\to\boldsymbol{\phi}_0$ as $T\to+\infty$,

$$\sup_{\phi\in N_T^*(c)}\|E_{\phi_0}f_\phi(\mathbf{Y}_t) - E_{\phi_0}f_{\phi_0}(\mathbf{Y}_t)\| = \sup_{\phi\in N_T^*(c)}\|E_{\phi_0}f_\phi(\mathbf{Y}_t) - \mathcal{I}_{n_0}(\boldsymbol{\phi}_0)\| \to 0.$$

Also, as $T \to +\infty$, $\frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}_0) \overset{a.s.}{\to} \mathcal{I}_m(\boldsymbol{\phi}_0)$. Thus $\sup_{\phi \in N_T^*(c)} \|\frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}) - \frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}_0)\| \to 0$ as $T \to +\infty$.

D6. (i). Let $\epsilon > 0$ be a small number to be determined and $C_T = \{\boldsymbol{\beta} : |\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \epsilon\} \times \Theta$. It is equivalent to check if $\frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi})$ is positive definite on C_T.

$$\frac{1}{T}\mathbf{J}_T(\boldsymbol{\phi}) = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^T \mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t + \frac{1}{T}\mathbf{S} & -\frac{1}{T}\frac{\partial}{\partial \boldsymbol{\theta}'}\sum_{t=1}^T \mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) \\ -\frac{1}{T}\frac{\partial}{\partial \boldsymbol{\theta}}\sum_{t=1}^T (\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t & -\frac{1}{T}\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \end{bmatrix}$$

The off-diagonal block matrix is

$$\begin{aligned} & \frac{1}{T}\frac{\partial}{\partial \boldsymbol{\theta}'}\sum_{t=1}^T \mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) \\ = & \frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}_0 + \mathbf{X}_t\boldsymbol{\beta}_0 - \mathbf{X}_T\boldsymbol{\beta}) \\ = & \frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}_0) + \frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \\ = & \frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{e}_t + \frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}). \end{aligned}$$

Let $f_\theta(\mathbf{e}_t) = \left(\frac{\partial}{\partial \theta'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\right)\mathbf{e}_t$. Then the map $\theta \mapsto f_\theta(\mathbf{e}_t)$ is continuous. Suppose $\exists\ \delta_1 > 0$ and a finite, integrable functions $M_1(\mathbf{e}_t)$ such that $\sup_{|\theta - \theta_0| < \delta_1} \|f_\theta(\mathbf{e}_t)\|_{\max} \le M_1(\mathbf{e}_t)$. By Lemma 3,

$$\frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \theta'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{e}_t = \frac{1}{T}\sum_{t=1}^T f_\theta(\mathbf{e}_t) \overset{a.s.}{\to} E_{\theta_0} f_\theta(\mathbf{e}_t) = \left(\frac{\partial}{\partial \theta'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\right) E_{\theta_0}[\mathbf{e}_t] = 0$$

uniformly as $T \to +\infty$. Therefore, $\frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{e}_t$ is $o_p(1)$. Next, consider when $|\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \epsilon$,

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \theta'}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \\ = & |\boldsymbol{\beta}_0 - \boldsymbol{\beta}|\frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \theta}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t\frac{\boldsymbol{\beta}_0 - \boldsymbol{\beta}}{|\boldsymbol{\beta}_0 - \boldsymbol{\beta}|} = |\boldsymbol{\beta}_0 - \boldsymbol{\beta}|O_p(1) \end{aligned}$$

provided that $\frac{1}{T}\sum_{t=1}^T \frac{\partial}{\partial \theta}\mathbf{X}_t'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}_t$ is bounded uniformly in probability.

Thus, the off diagonal blocks are negligible as ϵ can be chosen to be small enough, if the two diagonal blocks are positive definite with eigenvalues bounded away from zero over C_T. The preceding condition on the diagonal blocks holds under the assumptions (A4) and (A5). Of course, $\boldsymbol{\Sigma}_\theta$ needs to be invertible and continuous over Θ.

(ii) The prior of $\boldsymbol{\phi}$ given by (5) is bounded on C_T.

(iii) Let $R_T = \Phi$ and $C = C_T$. $\mathbf{B}_T = \mathbf{J}_{T0} = \sum_{t=1}^T \mathbf{J}_t = T \cdot \frac{1}{T}\sum_{t=1}^T \mathbf{J}_t \to T\mathcal{I}_m(\boldsymbol{\phi}_0)$, which is equivalent to $\mathbf{B}_T = T\mathbf{I}_{k+l}$. Thus by Lemma 1, with probability tending to one as

$T \to +\infty$, there is a unique maximizer of $l_T(\boldsymbol{\phi})$ in the open set C_T. Then with probability tending to one $\sup_{\boldsymbol{\phi} \in R_T - C_T} p_T(\mathbf{Y}_T|\boldsymbol{\phi}) = \sup_{\boldsymbol{\phi} \in \partial C_T} p_T(\mathbf{Y}_T|\boldsymbol{\phi})$. This can be proved by contradiction. Suppose $\sup_{\boldsymbol{\phi} \in R_T - C_T} p_T(\mathbf{Y}_T|\boldsymbol{\phi})$ is obtained at some point inside the open set $R_T - \bar{C}_T$. Then that point would be a local maximizer of $l_T(\boldsymbol{\phi})$, which contradicts the concavity of $l_T(\boldsymbol{\phi})$ which can only have one local maximizer over an open, convex parameter space. Since $(\{\lambda_{\min}(\mathbf{B}_T)\}^{-1} \log \lambda_{\max}(\mathbf{B}_T)) = (\frac{1}{T}\log(T))$ is stochastically bounded, by Lemma 2 we have

$$Q_T \equiv \{p_T(\mathbf{Y}_T|\boldsymbol{\phi}_0)\}^{-1}|\mathbf{B}_T|^{1/2} \int_{\boldsymbol{\phi} \in R_T - C_T} p_T(\mathbf{Y}_T|\boldsymbol{\phi})\pi(\boldsymbol{\phi})d\boldsymbol{\phi} \xrightarrow{p} 0.$$

This complete the proof of Theorem 1 which readily follows from Theorem 2.