Space-Time Generalized Additive Models

Ali M. Mosamman
Department of Statistics, University of Zanjan, Zanjan, IRAN-a.m.mosamman@znu.ac.ir

Abstract

If data have some temporal/spatial or spatial-temporal dependence and this is not modelled in the systematic part of the classic generalized additive model (GAM), the residuals from that model will also exhibit spatial or spatial temporal dependence, or in other words they will be spatially-temporally correlated. Such dependence would invalidate the theory that produces test statistics in the GAM models; because they were computed assuming independence. In this paper, we develop a space-time generalized additive model to take in to account such correlation. Our simulation results show the advantages of our proposed model.

Keywords: spatial-temporal model; generalized additive model; smoothing; interpolation.

1. Introduction

1 Space-Time Processes

This section is a review of some elementary concepts of spatial-temporal processes. A collection of spatial-temporal random variables

\[ \{Y(s, t), (s, t) \in \mathbb{E}\} \]

is called a spatial-temporal random field, where the index set \( \mathbb{E} \) is often a subset of \( \mathbb{R}^d \times \mathbb{R} \). Assume that all first and second-order moments of the random field exist for each \( (s, t) \). The random field (1) can be decomposed as \( Y(s, t) = \mu(s, t) + \eta(s, t) \), \( (s, t) \in \mathbb{E} \), where the deterministic mean \( \mu(s, t) = E[Y(s, t)] \) represents drift and the probabilistic term \( \eta(s, t) \) represents the error term. The spatial-temporal covariance between \( Y(s_1, t_1) \) and \( Y(s_2, t_2) \) is defined as

\[ \sigma(s_1, s_2, t_1, t_2) = \text{cov} \{Y(s_1, t_1), Y(s_2, t_2)\} = E\{Y(s_1, t_1) - \mu(s_1, t_1)\}\{Y(s_2, t_2) - \mu(s_2, t_2)\}. \]

If \( \mathbb{E} = \mathbb{R}^d \times \mathbb{R} \), the spatial-temporal random field (1) is called second-order stationary when

\[ E[Y(s, t)] = \mu \quad \text{for any } (s, t) \in \mathbb{E}, \quad \text{where } \mu \text{ is a constant,} \]

\[ \text{cov}\{Y(s, t), Y(s + h, t + u)\} = \sigma(h, u) \quad \text{for any } (s, t), (s + h, t + u) \in \mathbb{E}. \]

Furthermore, if \( \sigma(h, u) \) is a function of only \(|h|\) and \( u \), then \( \sigma(h, u) \) is called an isotropic spatial-temporal covariance function. The stationary spatial-temporal random field \( Z(s, t) \), is said to have a separable spatial-temporal covariance function if \( C(h, u) = C_S(h)C_T(u) \), where \( C_S(h) \) and \( C_T(u) \) are purely spatial and purely temporal covariance functions respectively. Recently several authors have developed approaches to generate positive and negative-value non-separable models for space-time processes. It is worth citing Cressie and Huang (1999), Gneiting (2002), Stein (2005) and Mosamman (2015). One class of spatially isotropic, fully symmetric, non-separable space-time models proposed by Gneiting (2002) has the form

\[ C(h, u) = \frac{\sigma^2}{(au^{2\alpha} + 1)^\gamma} \exp\left\{ \frac{-c|\lambda|^{2\gamma}}{(au^{2\alpha} + 1)^{5\gamma}} \right\}. \]

(2)

Here \( \alpha \) and \( c \) are time and space scaling parameters, \( \tau > \beta d/2 \). A separable covariance function is obtained when \( \beta = 0 \). The parameters \( \alpha \) and \( \gamma \) take values in \([0, 1]\); the parameter \( \beta \) take values in \([0, 1]\).
2 Space-Time Smoothing and Interpolation

Suppose that the spatial-temporal random field (1) does not include any drift. Consider the following smoothing problem: find \( f \in H \), a reproducing kernel Hilbert space, to minimize

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - f(s_i, t_i))^2 + \lambda \|f\|^2,
\]

where \( \lambda > 0 \) is called the smoothing parameter and \( \|f\| \) is a norm in \( H \) defined by a spatial-temporal positive definite kernel \( k(h, u) \). The solution to the smoothing problem, called the representer theorem, by Kimeldorf and Wahba (1971), has a representation of the form

\[
f(s, t) = \sum_{i=1}^{m} b_i \phi_i(s, t) = \tau^\top b = \tau^\top M^{-1} Y,
\]

where \( \tau = (k(s_1, t_1, s_1, t_1), \ldots, k(s_n, t_n, s_n, t_n))^\top \) and \( K = (k(s_i, t_i, s_j, t_j))_{i, j = 1, \ldots, n} \). Suppose now that our objective function includes a drift term lying in a drift space \( H_0 \), assumed to be a finite-dimensional subspace of \( H \) of dimension \( m > 0 \). Let \( \{\phi_i(s, t_i)\}_{i=1}^{m} \) form a basis of \( H_0 \). Let \( (H_1, \langle \cdot, \cdot \rangle) \) be the orthogonal complement of \( H_0 \). Every \( f \in H \) has a unique decomposition \( f = f_0 + f_1 \), where \( f_0 \in H_0 \) and \( f_1 \in H_1 \). Consider the smoothing problem: minimize

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - f_0(s_i, t_i) - f_1(s_i, t_i))^2 + \lambda \|f_1\|^2
\]

over \( f_1 \in H_1 \) and \( f_0 \in H_0 \). Assume the points \( (s_i, t_i), i = 1, \ldots, n \) are arranged in \( E \) in such a way that the matrix \( U \) with the \((i, j)\)-th entry \( \phi_i(s, t_i) \) has full rank \( m \). Straight forward extension of the solution of the problem by Kimeldorf and Wahba (1971) in such a case leads to a general form of the representer

\[
f(s, t) = \sum_{i=1}^{m} a_i \phi_i(s, t) + \sum_{i=1}^{n} b_i k(s_i, t_i, s_i, t_i) = u^\top a + \tau^\top b,
\]

where \( \tau = (k(s, t, s_1, t_1), \ldots, k(s, t, s_n, t_n))^\top \), \( a = (\phi_1(s, t))^\top, \ldots, \phi_m(s, t))^\top \), and the coefficients are given by \( a = (U^\top M^{-1} U)^{-1} U^\top M^{-1} Y, b = (M^{-1} - M^{-1} U (U^\top M^{-1} U)^{-1} U^\top M^{-1}) Y \) where \( M = K + n\lambda I_n, K = (k(s_i, t_i, s_j, t_j))_{i, j = 1, \ldots, n} \).

Next we review a general representation of simple kriging which is applied later to spatial-temporal residuals. Given the spatial-temporal random field (1), suppose that the mean of the random field is zero. Consider the problem of predicting \( Y(s_0, t_0) \) at new point \( (s_0, t_0) \) given the values of \( Y(s_i, t_i) \) at sites \( (s_i, t_i), i = 1, \ldots, n \). Let \( Y(0, \Sigma) \), where \( \Sigma = (\sigma(s_i, t_i, s_j, t_j)) \) is the covariance matrix between the data sites. We want to predict \( Y_0 = Y(s_0, t_0) \). Let \( Y = (Y(s_1, t_1), \ldots, Y(s_n, t_n))^\top, \alpha = (\alpha_1, \ldots, \alpha_n)^\top, \tau_0 = (\sigma(s_0, t_0, s_1, t_1), \ldots, \sigma(s_0, t_0, s_n, t_n))^\top \). The kriging predictor takes the form \( Y_0 = \alpha^\top Y = \tau_0^\top \Sigma^{-1} Y = \tau_0^\top b, \) where \( \alpha = \Sigma^{-1} \tau_0 \) and \( b = \Sigma^{-1} Y \). The result is exactly equal to that obtained in Section 2, when the covariance \( \sigma \) of the process in the setting is the same as the kernel \( k \) in the setting of Section 2 and when \( (s_0, t_0) \) here plays the same role as \( (s, t) \). The prediction error variance becomes \( \sigma^2 - \tau_0^\top \Sigma^{-1} \tau_0 \). Then (5) can be rewritten as

\[
\hat{Y}(s, t) = u^\top \hat{a} + \tau^\top M^{-1} (Y - u^\top \hat{a}),
\]

where \( \hat{Y}(s, t) \) is the predicted value at location \( (s, t) \), \( \hat{a} = u^\top \hat{a} \) is the fitted drift, \( u \) is the vector of predictors, \( \hat{a} \) is the vector of estimated drift model coefficients and \( \hat{Y}(s, t) = \tau^\top M^{-1} (Y - u^\top \hat{a}) \) is the interpolated residuals where \( \tau^\top M^{-1} \) is the vector of simple kriging weights used to interpolate the fitted residuals \( Y - u^\top \hat{a} \). This representation suggest a nice interpolation technique, the drift and residuals can be estimated separately and then summed. The advantage of regression-kriging is that we can use any complex regression methods including generalized additive models.

3 Space-Time Generalized Additive Models

A Generalized additive models (GAM) (Hastie and Tibshirani, 1990) has the following form \( g(\mu_i) = X_i^* \beta + \sum_{j=1}^{m} f_j(x_{ij}) \) where \( Y_i \) has some exponential family distribution; \( \mu_i \equiv E(Y_i); X_i^* \) is the \( i \)th row of the
model matrix for the strictly parametric model components; \( f_j \) are smooth functions of the covariates \( x_j \). Much standard research has been devoted to fitting GAM with independent data using spline methods (Green and Silverman (1994); Wood (2006)). However only limited work has been done for correlated data. If the data have some temporal/spatial or spatial temporal dependence and this is not modelled in the systematic part of the model, the residuals from that model will also exhibit spatial or spatial temporal dependence. Such dependence would invalidate the theory that produces test statistics in the GAM models; because they were computed assuming independence. There are two main options for handling such data: (1) model the spatial/temporal dependence in the systematic part of the model, i.e. by including a smooth of the spatial locations in the model. (2) relax the assumption of independence and estimate the correlation between residuals. In this paper we aim to apply penalized likelihood method to fit GAMS for nonseparable spatio-temporal data. In summary, if we model the spatial temporal dependence between observations then the residuals are more likely to be independent random variables and therefore not violate the assumptions of any classic inferential procedure. Consider the GAM

\[
Y(s, t) = \sum_{j=1}^{m} f_j(x_j(s, t)) + \epsilon(s, t),
\]

where \( x_j(s, t) \) are vectors of covariates and \( f_j \) are unknown smooth functions and \( \epsilon(s, t) \) are error terms with a multivariate normal distribution with mean zero and covariance \( \Sigma_\theta \). Consider the problem of estimating \( \theta \) and \( f \) based on data

\[
Y = (Y(s_1, t_1), \ldots, Y(s_n, t_m))^T
\]

and

\[
x_j(s, t) = (x_{j1}(s_1, t_1), \ldots, x_{jm}(s_m, t_m))^T, \quad j = 1, \ldots, m.
\]

This goal can be achieved by maximizing the penalized log likelihood

\[
\ell_p = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_\theta| - \frac{1}{2} (Y - \sum_{j=1}^{m} f_j(x_j))^{T} \Sigma_\theta^{-1} (Y - \sum_{j=1}^{m} f_j(x_j)) - \frac{1}{2} \sum_{j=1}^{m} \lambda_j J_j(f_j)
\]

where \( f_j(x_j) = [f(x_j(s, t))]^T \), \( \lambda_j \) is the smoothing parameter controlling the tradeoff between the model fit and the smoothness of the regression function, and \( J_j \) is a roughness penalty functional. For simplicity, let us suppose the case of a univariate Gaussian data, with spatial-temporally correlated response variable \( Y \) and regressor \( x \). We propose an iterative algorithm which maximizes the penalized log likelihood alternatively with respect to the covariance parameters and the smooth functions:

- **Step 1:** Start with estimating the drift model using GAM assuming independent errors.
- **Step 2:** Derive the residuals from an GAM assuming independent errors in Step1.
- **Step 3:** Calculate the initial value of \( \theta \), say \( \theta^{(0)} \) by fitting a variogram model to the residuals from an GAM assuming independent errors in Step2.
- **Step 4:** Then treat \( \theta \) as known and select the value of \( \lambda \) and estimate the smooth function \( f \). For fixed \( \theta \), maximizing (8) becomes

\[
\max_{f(x)} \left\{ -\frac{1}{2} (Y - f(x))^T \Sigma_\theta^{-1} (Y - f(x)) - \frac{1}{2} \lambda J(f) \right\}.
\]

It can be shown that for fixed \( \theta \) the solution of \( f \) to the above maximization problem is a spline if (Green and Silverman 1994). Hence the maximization problem is equivalent to the penalized weighted least squares problem

\[
\min_{\beta} \left\{ \| \Sigma_\theta^{-1/2} (Y - X\beta) \|^2 + \lambda \beta^T S \beta \right\} = \min_{\beta} \left\{ \| \tilde{Y} - \tilde{X} \beta \|^2 + \lambda \beta^T S \beta \right\},
\]

where \( \Sigma_\theta^{-1/2} \) is any square root matrix of \( \Sigma_\theta^{-1} \) such that (\( \Sigma_\theta^{-1/2} ) \Sigma_\theta^{-1/2} = \Sigma_\theta^{-1} \), \( \tilde{Y} = \Sigma_\theta^{-1/2} Y \) and \( \tilde{X} = \Sigma_\theta^{-1/2} X \). The smoothing parameter can be estimated by a number of criteria, e.g. cross validation (CV) or generalized cross validation (GCV). With known \( \theta \) and \( \lambda \), the minimization problem (9) admits a unique solution, i.e., \( \beta = (X \Sigma_\theta^{-1} X + \lambda S)^{-1} X \Sigma_\theta^{-1} Y \). Denote the selected \( \lambda \) and the corresponding estimated \( \theta \) as \( \lambda^{(1)} \) and \( \theta^{(1)} \) respectively.
Step 5: Let $\lambda = \lambda^{(1)}$ and $\beta = \beta^{(1)}$ and derive the residuals of GAM with correlated errors.

Step 6: Try to find a new estimate of $\theta$ using the residuals of GAM with correlated errors in Step 5. With $\lambda$ and $\beta$ fixed, maximizing (8) becomes

$$\max_{\theta} \left\{ -\frac{1}{2} (Y - X\beta)^T S^{-1}_\theta (Y - X\beta) - \frac{1}{2} \log |S_\theta| \right\}.$$

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

Step 7: Stop if $\frac{\|\theta^{(1)} - \theta^{(0)}\|}{\|\theta^{(0)}\|} < \epsilon$, say $\epsilon = 10^{-5}$. Otherwise, let $\theta^{(0)} = \theta^{(1)}$ and repeat Steps 4-7.

4 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 model

$$Y(s_1, s_2, t) = f_1(s_1) + f_2(s_2) + b(s_1, s_2, t) + e(s_1, s_2, t), \ t = 1, 2, \cdots, T$$

where $s_1 \in [0, 1]$ and $s_2 \in [0, 1]$ are the spatial coordinates of a data point $s = (s_1, s_2); f_1(s_1) = 2 \sin(\pi s_1); f_2(s_2) = e^{3s_2} - 13.4$ are smooth functions; $b$ are spatial-temporally correlated errors with distribution $N(0, \Sigma)$; and $e$ are pure measurement errors (producing a nugget effect) with distribution $N(0, \eta I)$. $s_1$ and $s_2$ are simulated independently from the uniform distribution on $[0, 1]$ with different sample size $S$. The space time random fields $b$ are independently simulated on the same set of locations and from the same normal distribution $N(0, \Sigma)$ where $\Sigma$ is defined by the covariance with (2), where $\tau = 1, \alpha = c = 0.4, \sigma^2 = 4.5, \beta = 0, 9, \gamma = 1, \alpha = 1$. The measurement errors $e$ are simulated from the normal distribution $N(0, \eta I)$ where $\eta = 0.2$. We analyze each simulated dataset in three ways: (1) a GAM assuming independent data; (2) a GAM assuming spatially temporally correlated data by ML estimation; and (3) a GAM assuming spatially temporally correlated data by REML estimation. The GAM with and without correlation structure are fitted to the simulated data by both ML and REML methods. The fitting results, which are based on 100 replicates for $S = 25, T = 20; S = 50, T = 10$ and $S = 100, T = 5$ (such that the total sample size is 500), are shown in Tables 1 and 2. As extremely large and extremely small estimates can be produced in estimating the covariance parameters, in addition to the sample means and mean squared error (MSE) we also include the medians and median absolute deviation (MAD) to assess the performance and to provide a more robust summary of simulation results. To evaluate the efficiency of estimation of the smooth functions, we calculated the mean square errors of the fitted functions and the 95 percent CI coverage proportion, which is the proportion of data points that are covered by their 95 percent CIs.

In summary, REML estimation is less stable than ML estimation. On the other hand, REML estimation are less biased than ML estimation. Also, the 95 percent CI coverage proportions from the REML estimation are higher than those from the ML estimation. As the spatial sample size increases and temporal sample size decreases such that the total sample size is fixed, both REML and ML estimation perform better with less bias and smaller variation. It is clear that the 95 percent CI coverage proportions in GAM without correlation structure are much lower than those from the REML estimation. As the spatial sample size increases and temporal sample size decreases such that the total sample size is fixed, both REML and ML estimation perform better with less bias and smaller variation. It is clear that the 95 percent CI coverage proportions in GAM without correlation structure are much lower than those from the REML estimation. As the spatial sample size increases and temporal sample size decreases such that the total sample size is fixed, both REML and ML estimation perform better with less bias and smaller variation.
<table>
<thead>
<tr>
<th>Proposed Method</th>
<th>GAM assuming independent data</th>
<th>ML</th>
<th>REML</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>MSE</td>
<td>coverage</td>
<td>MSE</td>
</tr>
<tr>
<td>n=25, T=20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totall</td>
<td>43.91</td>
<td>0.0417</td>
<td>46.494</td>
</tr>
<tr>
<td>$f_1+f_2$</td>
<td>97.48</td>
<td>0.0034</td>
<td>97.88</td>
</tr>
<tr>
<td>$f_1$</td>
<td>89.48</td>
<td>0.0019</td>
<td>90.08</td>
</tr>
<tr>
<td>$f_2$</td>
<td>88.32</td>
<td>0.0021</td>
<td>88.92</td>
</tr>
<tr>
<td>n=50, T=10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totall</td>
<td>49.558</td>
<td>0.0397</td>
<td>52.708</td>
</tr>
<tr>
<td>$f_1+f_2$</td>
<td>97.14</td>
<td>0.0041</td>
<td>97.66</td>
</tr>
<tr>
<td>$f_1$</td>
<td>88.68</td>
<td>0.0021</td>
<td>89.48</td>
</tr>
<tr>
<td>$f_2$</td>
<td>89.08</td>
<td>0.0022</td>
<td>89.76</td>
</tr>
<tr>
<td>n=100, T=5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totall</td>
<td>56.454</td>
<td>0.0368</td>
<td>58.204</td>
</tr>
<tr>
<td>$f_1+f_2$</td>
<td>97.37</td>
<td>0.0057</td>
<td>97.74</td>
</tr>
<tr>
<td>$f_1$</td>
<td>85.8</td>
<td>0.0029</td>
<td>86.55</td>
</tr>
<tr>
<td>$f_2$</td>
<td>91.11</td>
<td>0.0029</td>
<td>91.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proposed Method</th>
<th>ML</th>
<th>REML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Median</td>
<td>MSE</td>
</tr>
<tr>
<td>n=25, T=20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.452</td>
<td>0.2013</td>
</tr>
<tr>
<td>Scale_s</td>
<td>46424275</td>
<td>0.3632</td>
</tr>
<tr>
<td>Scale_t</td>
<td>0.944</td>
<td>0.2765</td>
</tr>
<tr>
<td>Sill</td>
<td>3.875</td>
<td>3.9845</td>
</tr>
<tr>
<td>n=50, T=10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.321</td>
<td>0.2051</td>
</tr>
<tr>
<td>Scale_s</td>
<td>12685215</td>
<td>0.3610</td>
</tr>
<tr>
<td>Scale_t</td>
<td>0.354</td>
<td>0.2674</td>
</tr>
<tr>
<td>Sill</td>
<td>3.865</td>
<td>3.9656</td>
</tr>
<tr>
<td>n=100, T=5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.1986</td>
<td>0.2053</td>
</tr>
<tr>
<td>Scale_s</td>
<td>0.3453</td>
<td>0.3372</td>
</tr>
<tr>
<td>Scale_t</td>
<td>0.2531</td>
<td>0.2673</td>
</tr>
<tr>
<td>Sill</td>
<td>3.7821</td>
<td>3.6827</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Available Method</th>
<th>ML-non separable</th>
<th>REML-non separable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Median</td>
<td>MSE</td>
</tr>
<tr>
<td>n=25, T=20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.458</td>
<td>0.2161</td>
</tr>
<tr>
<td>Scale_s</td>
<td>24555073</td>
<td>0.3621</td>
</tr>
<tr>
<td>Scale_t</td>
<td>0.912</td>
<td>0.2687</td>
</tr>
<tr>
<td>Sill</td>
<td>3.852</td>
<td>4.0080</td>
</tr>
<tr>
<td>n=50, T=10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.370</td>
<td>0.2204</td>
</tr>
<tr>
<td>Scale_s</td>
<td>30065394</td>
<td>0.3590</td>
</tr>
<tr>
<td>Scale_t</td>
<td>0.432</td>
<td>0.2566</td>
</tr>
<tr>
<td>Sill</td>
<td>3.776</td>
<td>3.8949</td>
</tr>
<tr>
<td>n=100, T=5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.1997</td>
<td>0.2026</td>
</tr>
<tr>
<td>Scale_s</td>
<td>0.3371</td>
<td>0.3312</td>
</tr>
<tr>
<td>Scale_t</td>
<td>0.2380</td>
<td>0.2525</td>
</tr>
<tr>
<td>Sill</td>
<td>3.7276</td>
<td>3.6092</td>
</tr>
</tbody>
</table>
References


