Gaussian Process Regression Model in Spatial Logistic Regression

A Sofro and A Oktaviarina
Mathematics Department, Universitas Negeri Surabaya, East Java, Indonesia
E-mail: ayuninsofro@unesa.ac.id

Abstract. Spatial analysis has developed very quickly in the last decade. One of the favorite approaches is based on the neighbourhood of the region. Unfortunately, there are some limitations such as difficulty in prediction. Therefore, we offer Gaussian process regression (GPR) to accommodate the issue. In this paper, we will focus on spatial modeling with GPR for binomial data with logit link function. The performance of the model will be investigated. We will discuss the inference of how to estimate the parameters and hyper-parameters and to predict as well. Furthermore, simulation studies will be explained in the last section.

1. Introduction
In recent years, modeling for Binomial data has been developed massively. The conventional way to handle this situation is using a logistic regression model by assuming the independence among the data. However, in practice this is a strong assumption since the majority health data are dependent data. One of popular approach to solve the problem is to use conditional autoregressive model to model the correlation structure, which is first proposed in [1].

Some other methods have also been developed in the last decade, including simple disease mapping method [1]. This method has been extended into spatial time generalized linear mixed model ([3]; [4]; [5]). Generalized linear mixed model using prior distribution for spatially structured random effect is a common alternative way, see [6]. But they limited the correlation structure using intrinsic conditional autoregressive model, [7]. Some authors pointed out that there are misleading regarding the spatial dependence covariance matrix, [10]. Moreover, it seems that practitioners need to understand how to determine the precision of the covariance matrix rather than puzzling effects, [11]. Thus more study needs to be done on modeling spatial correlation structure since this is a very common problem in practice([8]; [9]).

The aim of this paper is to propose a model which can describe the spatial covariance structure flexibly. Kriging is also one of approach in geostatistics, [12] but it seems offer limitation, such as difficulty to accomodate a large number of covariates. Therefore, to achieve this we used a Gaussian process regression (GPR) model. Recently GPR and the related methods has been developed fast and has a wide application in machine learning and other areas [12]. Some recent development can be found in [13] and [15]. Covariance structure of GPR is defined by a covariance kernel which depends on a set of covariates. Consequently it provides a very flexible
covariance structure coping with a variety of the variables such as geographic position, distance among the spatial areas and other variables related to culture and life style etc.

In this paper we will explain the details how to use a GPR model to model dependency in Binomial data and provide the technical details. Simulation studies will be presented to demonstrate the performance of the proposed method.

2. Gaussian Process Regression

Let \( y_i \) be a response variable and \( x_i \) be \( P \)-dimensional covariates. A Gaussian process regression model can be specified as follows:

\[
y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, N
\]

where \( \epsilon_i \sim i.i.d \ N(0, \sigma^2) \), \( \sigma^2 \) unknown and

\[
f(\cdot) \sim GP(\mu(\cdot), k(\cdot, \cdot))
\]

if

\[
f = (f_1, f_2, \ldots, f_N) \sim N(\mu, K),
\]

where the \( i \)-th element of \( \mu \) is \( \mu(x_i) \), the \( (i, j) \)-th element of \( K \) is \( k(x_i, x_j) \) and \( k(\cdot, \cdot) \) is a covariance function.

It is common to assume a zero mean function in the Gaussian process prior, i.e \( \mu(\cdot) = 0 \). The most popular choice for the covariance function is the following squared exponential kernel,

\[
k(x_i, x_j) = v_0 \exp \left\{ -\frac{1}{2} \sum_{p=1}^{P} w_p (x_{ip} - x_{jp})^2 \right\}
\]

where \( v_0, w_p, p = 1, \ldots, P \) denotes the set of hyper-parameters and are defined as \( \theta \).

3. Spatial Logistic Regression with GPR

For the \( i \)-th observation, let \( z_i \) be the response variable, \( \tau_i \) be a spatial related random effects and \( U_i \) be the covariates involved in the fixed effects, where \( i = 1, \ldots, n \) and \( n \) is the sample size. Typically in this model, the data are assumed to be conditionally independent and follow Binomial distribution using logit as link function. Hence, a spatial logistic regression with GPR can be expressed

\[
z_i | \pi_i \sim Bin(1, \pi_i) \text{ independently,}
\]

\[
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = U_i^T \beta + \tau_i.
\]

The correlation structure is defined via the random effect item \( \tau(\cdot) \sim GP(0, k(\cdot, \cdot)) \) which has been discussed in the previous section. The logit link function in equation (3) can be derived as follow

\[
\frac{\pi_i}{1 - \pi_i} = \exp(U_i^T \beta + \tau_i)
\]

\[
\pi_i + \pi_i \exp(U_i^T \beta + \tau_i) = \exp(U_i^T \beta + \tau_i)
\]

\[
\pi_i = \frac{\exp(U_i^T \beta + \tau_i)}{1 + \exp(U_i^T \beta + \tau_i)}.
\]

A suitable choice of a kernel covariance function and its hyper-parameters can improve the prediction accuracy. Rather than making assumption regarding prior of hyper-parameters, we choose using all observed data. One of the popular methods is using empirical Bayesian estimation to select the hyper-parameters. In practice, \( \theta \) and \( \beta \) can be estimated at the same time.
3.1. Empirical Bayesian Estimates

Now, we focus on the empirical Bayesian approach. Let us assume that we have observed a set of data \( D = (z_i, U_i, x_i), i = 1, \ldots, N, x_i \in T \subset \mathbb{R}^P \), where \( z_i \) is observation of variable response, \( U_i \) is covariates variable. \( x_i \) are covariates modelled covariance structure \( \tau_i \) and \( P \) is the dimension of the input vector \( x_i \).

The random effects \( \tau \) are unknown, the marginal density of \( z \) does not have a convenient closed-form representation. To estimate parameters, we can do so by maximizing the following marginal density \( z = (z_1, \ldots, z_N)^T \)

\[
p(z|\beta, \theta) = \int p(z|\beta, \tau)p(\tau)d\tau = \int p(y|\tau)d\tau
\]

where \( \tau \sim N(0, K) \).

However, the above marginal density function is analytically intractable. One of the methods used to address this issue is to use a Laplace approximation. The log likelihood of equation (6) can be written as

\[
l(\beta, \theta) = \sum_{i=1}^{N} \log \int \exp(\Phi(\tau))d\tau.
\]

Note that

\[
\Phi(\tau) = \log p(z|\beta, \tau) - \frac{1}{2} \log |K(\theta)| - \frac{1}{2} \tau^T K^{-1}(\theta) \tau - \frac{N}{2} \log 2\pi.
\]

and the likelihood of \( p(z|\beta, \tau) \) is following

\[
L(p(z|\beta, \tau)) = \prod_{i=1}^{n} \left( \pi z_i (1 - \pi_i)^{1-z_i} \right) = \prod_{i=1}^{n} \left( \frac{\pi_i}{1 - \pi_i} \right)^{z_i} (1 - \pi_i).
\]

From equation (4), we can substitute to equation (5). Hence, we get

\[
\prod_{i=1}^{n} \left( \exp(U_i^T \beta + \tau_i) \right)^{z_i} \left( 1 - \frac{\exp(U_i^T \beta + \tau_i)}{1 + \exp(U_i^T \beta + \tau_i)} \right).
\]

Then the log likelihood of \( p(z|\beta, \tau) \) can be written as following

\[
\sum_{i=1}^{n} \left( z_i(U_i^T \beta + \tau_i) - \log(1 + \exp(U_i^T \beta + \tau_i)) \right).
\]

We put \( \tau_0 \) be the maximiser of \( \Phi(\tau) \). Thus a Laplace approximation is

\[
\int \exp(\Phi(\tau))d\tau = \exp \left\{ \Phi(\tau_0) + \frac{N}{2} \log(2\pi) - \frac{1}{2} \log |H| \right\}
\]

where \( H \) is the negative of the second derivative of \( \Phi(\tau) \) respect to \( \tau \) and evaluated at \( \tau_0 \). We have \( H = C - K \) and \( C \) is a Hessian matrix. Hence, we can get \( C \) is a diagonal matrix,

\[
C = \text{Diag} \left( \frac{\exp(U_1^T \beta + \tau_0)}{1 + \exp(U_1^T \beta + \tau_0)}, \ldots, \frac{\exp(U_n^T \beta + \tau_0)}{1 + \exp(U_n^T \beta + \tau_0)^2} \right).
\]

In order to estimate the parameters, we maximize the likelihood function with Laplace approximation in equation (7). To apply Laplace approximation, we need to find \( \tau_0 \).
3.2. Predictions

It is of interest to predict \( z^* \) at a new test input \( x^* \). The main purpose in this section is to calculate \( E(z^*|D) \) and \( Var(z^*|D) \).

Let \( \tau(x^*) = \tau^* \) be the underlying latent variable at \( x^* \). The expectation of \( z^* \), conditional on \( \tau^* \) is given by

\[
E(z^*|\tau^*, D) = \frac{\exp(U^T \hat{\beta} + \tau^*)}{1 + \exp(U^T \hat{\beta} + \tau^*)}. \tag{12}
\]

It follows that

\[
E(z^*|D) = E[E(z^*|\tau^*, D)] = \int \left( \frac{\exp(U^T \hat{\beta} + \tau^*)}{1 + \exp(U^T \hat{\beta} + \tau^*)} \right) p(\tau^*|D)d\tau^*. \tag{13}
\]

One method to calculate the above expectation is to approximate the integration using a Laplace approximation. We can rewrite it as

\[
p(\tau^*|D) = \int p(\tau^*|\tau, D)p(\tau|D)d\tau = \int p(\tau^*, \tau|D)d\tau = \frac{1}{p(z)} \int p(y|\tau)p(\tau^*, \tau)d\tau. \tag{14}
\]

For convenience, we denote \((\tau, \tau^*)^T\) and its covariance matrix \( K_{N+1,N+1} \) by \( \tau_+ \) and \( K_+ \) respectively. Thus, the equation (13) can be written as

\[
\frac{1}{p(y)} \int \left[ \left( \frac{\exp(U^T \hat{\beta} + \tau^*)}{1 + \exp(U^T \hat{\beta} + \tau^*)} \right)^N \prod_{i=1}^N p(z_i|\hat{\beta}, \tau_i) \right] \left[ (2\pi)^{-\frac{(N+1)}{2}} |K_+|^{-\frac{1}{2}} \exp\left( -\frac{1}{2} \tau_+^T K_+^{-1} \tau_+ \right) \right] \, d\tau_+ \tag{15}
\]

The calculation of the integral is not tractable, since the dimension of \( \tau_+ \) is usually very large. We still use a Laplace approximation. Note that

\[
\tilde{\Phi}(\tau_+) = \log \left( \frac{\exp(U^T \hat{\beta} + \tau^*)}{1 + \exp(U^T \hat{\beta} + \tau^*)} \right) + \sum_{i=1}^N \log p(z_i|\hat{\beta}, \tau_i) - \frac{N+1}{2} \log(2\pi) - \frac{1}{2} \log |K_+| - \frac{1}{2} \tau_+^T K_+^{-1} \tau_+
\]

where \( \log p(z|\beta, \tau) = \sum_{i=1}^n \left( z_i(U_i^T \hat{\beta} + \tau_i) - \log(1 + \exp(U_i^T \hat{\beta} + \tau_i)) \right) \). Equation (14) can be expressed as

\[
p(z^*|D) = \frac{1}{p(z)} \int \exp(\tilde{\Phi}(\tau_+))d\tau_+.
\]

Let \( \tau_+ \) be the maximiser of \( \tilde{\Phi}(\tau_+) \), then by using Laplace approximation we have

\[
\int \exp(\tilde{\Phi}(\tau_+))d\tau_+ = \exp(\tilde{\Phi}(\tau_+)) + \frac{N+1}{2} \log(2\pi) - \frac{1}{2} \log |K_+| + \mathbf{C}_+
\]

where \( \mathbf{C}_+ \) is the negative of the second derivative of

\[
(U^T \hat{\beta} + \tau^*) - \log(1 + \exp(U^T \hat{\beta} + \tau^*)) + \sum_{i=1}^n \left( z_i(U_i^T \hat{\beta} + \tau_i) - \log(1 + \exp(U_i^T \hat{\beta} + \tau_i)) \right)
\]
with respect to $\tau_+$ and evaluated at $\hat{\tau}_+$. Here $\mathbf{C}_+$ is a diagonal matrix, i.e

$$\mathbf{C}_+ = \text{Diag} \left( \frac{\exp(U_1^T \hat{\mathbf{\beta}} + \tau_{01})}{1 + \exp(U_1^T \hat{\mathbf{\beta}} + \tau_{01})^2}, ..., \frac{\exp(U_n^T \hat{\mathbf{\beta}} + \tau_{0n})}{1 + \exp(U_n^T \hat{\mathbf{\beta}} + \tau_{0n})^2}, 0 \right).$$

Since $p(z)$ has already been investigated, the predictive mean (13) can be calculated.

To calculate $\text{Var}(z^*|D)$, we use the formula:

$$\text{Var}(z^*|D) = E[\text{Var}(z^*|\tau^*, D)] + \text{Var}[E(z^*|\tau^*, D)].$$

(17)

Because $\text{Var}(z^*|\tau^*, D) = E(z^*|\tau^*, D)(1 - E(z^*|\tau^*, D))$, therefore

$$E[\text{Var}(z^*|\tau^*, D)] = E(z^*|D) - (E(z^*|D))^2.$$  

(18)

From the model definition, we have

$$\text{Var}(E(z^*|\tau^*, D)) = E[E(z^*|\tau^*, D)]^2 - [E[E(z^*|\tau^*, D)]]^2$$

$$= \int \left( \frac{\exp(U^T \hat{\mathbf{\beta}} + \tau^*)}{1 + \exp(U^T \hat{\mathbf{\beta}} + \tau^*)} \right)^2 p(\tau^*|D)d\tau^* - [E(z^*|\tau^*, D)]^2.$$  

(19)

The first item in (19) can be obtained by Laplace approximation similar to $E(z^*|D)$ in (13) and the second item is the square of (14).

### 4. Numerical Examples

In this section we demonstrate a numerical example to explore the performance of the proposed model. We will present a scenario comparing some models. Here, the two models considered are logistik regression (Model 1) and spatial logistik regression with GPR (Model 2).

In the simulation study, we first generate random data from a Binomial distribution with true values $\beta_0 = 1$ and $\beta_1 = 2$. Data are generated from model (3) which assumes $\tau_i$ follows a Gaussian process regression model as defined in equations (1) and (2). The true hyper-parameters are $v_0 = 0.04$, $w_1 = 1$ and $w_2 = 1$.

Here, the covariance kernel is squared exponential which $x$ are the location of each area measured by its longitude and latitude from Dengue fever data, see [16]. We select 30 of them as training data and the remaining as test data. To measure performance, we calculate the sample mean of the estimated parameters and the root mean squared error (RMSE) between the estimated and the true values of the parameters.

Table 1 shows the sample mean from different methods and also the value of RMSE (RMSE) based on one hundred replications. The value of average RMSE (average RMSE) between $\pi$ and

| Method  | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ |
|---------|----------|----------|----------|----------|
| True value | 1        | 2        | 0.01507  | 0.02880  |
| Model 1 | 1.00134  | 1.99499  | 0.00505  | 0.03800  |
| Model 2 | 1.00020  | 2.00010  | 0.00018  | 0.00055  |

Table 1. Mean of the estimated parameters and RMSE between the estimated and true values of the parameters based on one hundred replications for logistic regression (Model 1) and spatial logistic regression with GPR (Model 2).
\( \hat{\pi} \) also can be seen in Table 1. It shows clearly that the performance of model 2, i.e spatial logistik regression with GPR is the best in term of estimating parameters.

Meanwhile, Table 2 shows prediction performance of the proposed model. We investigate a different number of training data (m) from similar generated data in the previous scenario. The aim is to analyze the accuracy of prediction of the model 2. We calculate the average of RMSE between the predicted values and the actual observed data based on one hundred replications. From the Table 2, it can be said that the more increased number of data, the more improve the performance of prediction as it is expected.

| \( m \) | Average RMSE |
|-------|-------------|
| 20    | 0.00976     |
| 30    | 0.00815     |

Table 2. The average of RMSE between \( \pi \) and \( \hat{\pi} \) based on one hundred replications for different number of training data

5. Conclusion

In this paper, we proposed a Gaussian process regression model for the covariance structure. It allows the use of geographic position and other variables to define the spatial correlation structure, and thus it provides a very flexible model. The simulation study shows its good performance. The computation of the method is quite efficient.

References

[1] Besag J and Kooperberg C 1995 On Conditional and Intrinsic Autoregressions Biometrika vol. 82 pp 733-746
[2] Knorr L and Besag J 1998 Modelling risk from a disease in time and space Statistics in Medicine vol. 17 pp 2045-2060
[3] Sun T, Kim H and He Z Q 2000 Spatio-temporal interaction with disease mapping, Statistics in Medicine vol. 19 pp 2015-2035
[4] McNab Y C and Dean C B 2001 Autoregressive spatial smoothing and temporal spline smoothing for mapping rate Biometrics vol. 57 pp 73-85
[5] Silva G L, Dean C B, Niyonsenga T and Vanessa A 2008 Hierarchical Bayesian spatio-temporal analysis of revacularization odds using smoothing splines Statistics in Medicine vol. 27 pp 2381-2401
[6] Banerjee S, Carlin B P and Gelfand A E 2004 Hierarchical Modeling and Analysis for Spatial Data Boca Rotan Chapman and Hall
[7] Rue H and Held L 2005 Gaussian Markov Random Field Theory and ApplicationsBoca Rotan Chapman and Hall
[8] Haining R 1990 Spatial Analysis in the Social and Environmental Sciences Cambridge Cambridge University Press
[9] Bernadinelli L, Clayton D and Montomoli C 1995 Bayesian estimates of disease maps: how important are priors? Statistics in Medicine vol. 14 pp. 2411-2431
[10] Wall M M 2004 A close look at the spatial structure implied by the CAR and SAR models Journal of Statistical Planning and Inference vol. 121 pp. 311 324
[11] Renato and Krainski E 2009 Neighborhood Dependence in Bayesian Spatial Models Biometrical Journal vol. 51 pp. 851869
[12] Rasmussen C E and Williams C K 2006 Gaussian Process for Machine Learning Cambridge England: MIT Press
[13] Shi J Q and Choi T 2011 Gaussian Process Regression Analysis for Functional Data, Boca Rotan Taylor and Francis
[14] Diggle P J, Tawn J A and Moyeed A 1998 Model-based geostatistics Royal Statistical Society vol. 47 pp. 299-350
[15] Wang B and Shi J Q 2014 Generalized Gaussian Process Regression Model for non-Gaussian Functional Data Journal of American Statistical Association vol. 109 pp. 1123-1133
[16] Sofro A 2016 Convolved Gaussian Process Regression Models for Multivariate Non-Gaussian Data PhD Thesis, Newcastle University United Kingdom