Parameter Estimation of Spatial Logistic Regression Model using Using Variational Method on Spatial Logistic Regression

Hakiim Nur Rizka and Yekti Widyaningsih

Department of Mathematics, Universitas Indonesia, Depok

Abstrak. Model regresi logistik spasial membangun persamaan dengan jenis variabel dependen adalah biner serta mempertimbangkan dependensi spasial pada data. Proses estimasi parameter pada model ini memerlukan algoritma EM. Namun, bentuk ekspektasi dari *complete log-likelihood* pada *E-step* tidak tersedia dalam *closed-form*. Dalam menangani permasalahan ini, metode terbaru yang sedang diteliti oleh Cecille Hardouin memanfaatkan pendekatan deterministik dikenal dengan metode variasional. Metode variasional merupakan metode aproksimasi distribusi yang memanfaatkan suatu batas bawah fungsi distribusi yang akan diaproksimasi lalu mengoptimalkan batas bawah ini. Metode variasional untuk estimasi parameter model regresi logistik spasial bekerja dengan mencari suatu batas bawah dari *complete log-likelihood* lalu memaksimumkan fungsi ini terhadap parameter model. Dalam studi literatur, didapatkan bahwa metode variasional memiliki akurasi lebih baik daripada algoritma EM dengan aproksimasi Laplace ketika dependensi spasial pada data relatif besar.

Keyword : Metode Variasional, Model Regresi Logistik Spasial, Algoritma EM

Abstract. The spatial logistic regression model builds equations in which the dependent variable is binary and considers the spatial dependency on the data. Estimation procedure of the parameters in this model require EM algorithm. However, the expected form of the complete log-likelihood on the E-step is not available in closed-form. In order to deal with this problem, a recent method being researched by Cecille Hardouin utilizes a deterministic approach known as the variational method. The variational method is a distribution approximation method that utilizes a lower bound of the distribution function to be approximated and then optimizes this lower bound. The variational method for estimating the parameters of the spatial logistic regression model works by finding a lower limit of the complete log-likelihood and then maximizing this function to the model parameters. In the literature study, it was found that the variational method has better accuracy than the EM algorithm with Laplace's approximation when the spatial dependence on the data is relatively large.

Keyword : Variational Methods, Spatial Logistic Regression Model, EM Algorithm

Introduction

Spatial logistic regression is a model that was developed to overcome the problem when the type of dependent variable is binary and the data has spatial dependencies. The form of the spatial logistic regression equation has the same principles and assumptions as ordinary logistic

regression. In building a spatial logistic regression model, the problem of parameter estimation becomes a major concern. This is because the parameter estimation using the maximum likelihood method does not have a closed-form form. Several approaches have been taken to deal with this problem, one of which is the expectation-maximization (EM) algorithm.

The parameter estimation on spatial logistic regression has a critical intricacy regarding the usage of the maximum likelihood estimate. One of the earliest alternatives on the likelihood-based estimation of the parameter on spatial logistic regression was a numerical method called expectation-maximization (EM) algorithm. The EM algorithm is an approximation method that works iteratively to obtain parameter estimates using the complete likelihood function. The complete likelihood is a likelihood function that assumes the existence of unobserved values from the data. By the EM algorithm, the parameters in the model are estimated in two steps: (1) calculating the expectation of the complete likelihood function and (2) maximizing the expectation for each parameter in the model. These two steps are carried out iteratively until they reach the specified level of convergence. At the end of the iterative procedure, the expected form of the complete likelihood that has been maximized for each parameter in the model is obtained. This is equivalent to the maximum likelihood method estimation procedure. The parameter values in the complete likelihood function, obtained in the last iteration, are used as parameter estimates in the model.

In the application of the EM algorithm to the estimation of the parameters of the spatial logistic regression model, a Laplace approximation is needed in the form of complete likelihood expectations (see Sengupta and Cressie, 2013). There are some other alternatives that have been proposed for this estimation such as using Monte Carlo procedures (Cappé et al., 2005). In this study, an alternative to using the Laplace approximation is discussed, known as the variational method (Hardouin, 2019).

The variational method or also known as variational Bayes is an approximation method used in approximating the distribution function. The variational method approximates the distribution of the original problem by defining a distribution family. After defining the distribution family, a distribution from the distribution family is selected which is known as the variational distribution. Furthermore, the variational distribution is maximized for its parameters iteratively. The variational distribution obtained at the end of the iterative procedure is used as a representative of the distribution in the original problem. For variational EM method, the lower bound of likelihood function of the distribution function is used as the new objective function, then optimize this function to get the estimation on the parameters. The variational estimation procedure will always converge to a sufficiently good estimation of the real parameters (Blei, 2017).

In this study, it will be emphasized the application of the variational method to the estimation of spatial logistic regression parameters.

The Process Model

Bernoulli Distribution

Consider a density function $f(x)$ from a discrete random variable X defined as

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,2, \dots, n \quad (1)$$

Random variable *X* is said to be distributed under binomial distribution and it can be written with the notation *X ~ Bin(n,p)* and *p* can be defined as "probability of success". A binomial distribution with *n = 1* is also known as bernoulli distribution. The distribution function for bernoulli distribution can be written as

$$f(x) = p^x (1-p)^{1-x}, x = 0,1 \quad (2)$$

and if a random variable *X* is distributed under bernoulli distribution, it can be stated as *X ~ Ber(p)*.

Maximum Likelihood Method

Suppose there are *n* random samples $X_1, \dots, X_n$ where each has density denoted as $f(x_i; \theta)$, $\theta \in \Omega, i = 1,2, \dots, n$ with Ω is parameter space. Thus, likelihood function *L* can be built based on the previous information that is

$$L(\theta; x_1, x_2, \dots, x_n) = f(\theta; x_1, x_2, \dots, x_n)$$
$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta), \ \theta \in \Omega \quad (3)$$

which is a joint density for the random samples. Maximum likelihood method use the function (3) to estimate the parameter θ in the density function by maximizing function (3) relative to the parameter θ. To maximize it, more often than not logarithm form of function is used (3).

$$\ell(\theta; x_1, x_2, \dots, x_n) = \log L(\theta; x_1, x_2, \dots, x_n) \quad (4)$$

Thus, to obtain parameter estimation from maximum likelihood method can be done by solving the equation

$$\frac{\partial \ell(\theta; x_1, x_2, \dots, x_n)}{\partial \theta} = 0 \quad (5)$$

Logistic Regression

Logistic regression is one of the methods to model a binary dependent variable with one or more independent variables. Logistic regression builds an equation based on the so-called logistic function. The logistic function can be defined as :

$$f(y) = \frac{e^y}{1+e^y} \quad (6)$$

this function maps $-\infty < y < \infty$ to $0 \leq f(y) \leq 1$. In logistic regression, it is common to use the logarithmic of the logistic function as it leads to an equation of dependent variable with a linear combination of the covariates matrix $\boldsymbol{X}$ with and the regression coefficients vector $\boldsymbol{\beta}$. Suppose there are n observations and j independent variables. For i-th observation its equation can be written as :

$$y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{ji}\beta_j , \quad i = 1, 2, \ldots, n \tag{7}$$

with $\boldsymbol{y'} = [y_1 , y_2, ...,y_n]$ as vector of value from the dependent variable. If (7) is applied to (6), it is then defined as the probability of i-th observation will be "success" that is :

$$\pi_i = \frac{e^{y_i}}{1+e^{y_i}} = \frac{1}{1+\exp[-(\alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{ji}\beta_j)]}, \quad i = 1, 2, \ldots, n \tag{8}$$

the form (8) can also be stated as $P(y_i = 1 | X_i)$.

Regarding the estimation for the parameter in logistic regression, it is common to use the maximum likelihood method. First of all, construct the likelihood function. Consider the parameter that is going to be estimated is $\boldsymbol{\beta}$, then given n observational data the likelihood function for the logistic regression is

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} {\pi_i}^{y_i} \times (1-\pi_i)^{(1-y_i)} \\ L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(\frac{1}{1+e^{-y_i}}\right)^{y_i} \times \left(\frac{1}{1+e^{y_i}}\right)^{(1-y_i)} \end{aligned} \tag{9}$$

then the log-likelihood function for this model can be defined as

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \log\left[\prod_{i=1}^{n} \left(\frac{1}{1+e^{-y_i}}\right)^{y_i} \times \left(\frac{1}{1+e^{y_i}}\right)^{(1-y_i)}\right] \\ l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-y_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{y_i}}\right) \\ &= \sum_{i=1}^{n} y_i \log\left(\frac{1+e^{y_i}}{1+e^{-y_i}}\right) + \log\left(\frac{1}{1+e^{y_i}}\right) \\ &= \sum_{i=1}^{n} y_i \boldsymbol{X_i}\boldsymbol{\beta} - \sum_{i=1}^{n} \log(1+e^{y_i}) \end{aligned} \tag{10}$$

To get the estimation for regression coefficients $\boldsymbol{\beta}$, the next step is to solve the derivative equation based on (10). This procedure calls for a numerical approach. It is caused by the derivative term of $\log(1+e^{y})$ is nonlinear thus makes the derivative equation for the maximum likelihood estimation

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \tag{11}$$

has no simple analytical solution.

Spatial Logistic Regression

To include spatial dependence in the logistic regression model, this paper follows a similar approach to that of Hardouin (2019). First consider a two-dimensional domain $T \equiv \{\boldsymbol{s_i}, i = 1,2,...,n\}$, which is a subset of real space, with $\boldsymbol{s_i} = (s_{i1}, s_{i2})$ for $i = 1,2,...,n$. Next for the dependent variable consider $\boldsymbol{Y^*} = (y^*(\boldsymbol{s_1}), y^*(\boldsymbol{s_2}), ...,y^*(\boldsymbol{s_n}))$ as a process happening on T. Variable $\boldsymbol{Y^*}$ is of a bernoulli distribution with its mean depend on the process $\boldsymbol{Y} = (y(\boldsymbol{s_1}), y(\boldsymbol{s_2}), ...,y(\boldsymbol{s_n}))$. The assumption of independency of $\boldsymbol{Y^*}$ given $\boldsymbol{Y}$ is applied.

Then for each $\boldsymbol{s}$ element of T, the distribution of $\boldsymbol{Y^*}$ given $\boldsymbol{Y}$ can be written as

$$[Y^*(\boldsymbol{s})|Y(\boldsymbol{s})] \sim Ber(\pi(\boldsymbol{s})) \tag{12}$$

with

$$\pi(\boldsymbol{s}) = \frac{e^{y(\boldsymbol{s})}}{1+e^{y(\boldsymbol{s})}} \tag{13}$$

and for $Y(\boldsymbol{s})$ it is modeled as

$$Y(\boldsymbol{s}) = \boldsymbol{X(s)^T\beta} + \mathcal{E}(\boldsymbol{s}) \tag{14}$$

with $\boldsymbol{X(s)}^T = (X_1(\boldsymbol{s}), X_2(\boldsymbol{s}), ...,X_j(\boldsymbol{s}))$ is the j known covariates and $\boldsymbol{\beta}$ is the coefficients regression. The linear combination of these is representing the large-scale spatial variation (Hardouin, 2019). The term $\mathcal{E}(\boldsymbol{s})$ is a gaussian spatial process with zero mean and unknown covariance matrix $\sum$, that is

$$\mathcal{E}(\boldsymbol{s}) \sim MN(\boldsymbol{0},\sum) \tag{15}$$

With little inspection, it can be seen that the condtional distribution in (12) seems to have different parameter π for different locational $\boldsymbol{s_i}$ with i defining the i-th location. Thus the conditional probablity can be written as

$$p\left(y^*(\boldsymbol{s_i}) = 1 \middle| y(\boldsymbol{s_i})\right) = \frac{e^{y(s_i)}}{1+e^{y(\boldsymbol{s_i})}} \tag{16}$$

$$p\left(y^*(\boldsymbol{s_i}) = 0 \middle| y(\boldsymbol{s_i})\right) = \frac{1}{1+e^{y(\boldsymbol{s_i})}} \tag{17}$$

Variational Inference Method

Variational inference method has been known as one of the exact approximation methods which is widely used in several fields. Application of variational methods such as in the fields of statistics (Rustagi, 1976), quantum mechanics (Sakurai, 1985), mechanical statistics (Parisi, 1988). Variational inference method is derived from one of mathematical analysis field which is known as calculus of variations.

Calculus of variations focuses its attention to variations, a small changes in function or functional, in order to find a maxima and minima in functionals : mappings from set of functions to real numbers. One of the widely used equation to find the solution to find the maxima and minima is known as Euler-Lagrange equation.

$$\frac{\partial L[x,f(x),f'(x)]}{\partial f^0(x)} - \frac{d}{dx}\frac{\partial L[x,f(x),f'(x)]}{\partial f^{0\prime}(x)} = 0 \quad (18)$$

The procedure to obtain equation (18) implies that if function $f^0(x)$ is an extrema (minima or maxima) of functionals $I[f(x)]$, therefore $f^0(x)$ must satisfy the Euler-Lagrange equation (18).

The principles on calculus of variations proved to be useful to overcome the problems regarding Bayesian statistics. In Bayesian statistics, more often than not, its is imperative to estimate the posterior distribution using an exact approximation method such as MCMC method. But, this estimation procedure is often become computationally expensive. Thus, variational inference became one of the alternative to overcome this problem (see Blei, 2017).

In general, the first step in variational inference is to approximate posterior distribution by defining a family distribution which can be in a form of certain special distribution such as normal distribution (see Hardouin, 2019) or a more general form by defining the joint distribution (see Keng, 2017). The introduction of a new parameter called variational parameter is also being done in this step, thus the name for this new distribution is variational distribution. After obtaining variational distribution, the next step is to maximize the variational distribution relative to the parameters including the variational parameter. The steps to maximize the variational distribution is based on the so-called evidence of lower bound (ELBO) and Kullback-Leibler divergence (Kullback and Leibler, 1951).

Estimation of Parameter for Spatial Logistic Regression

First of all, let us note that the model refers to spatial logistic regression and the parameter needed to be estimated is $\boldsymbol{\beta}$ and $\sum$. Then, this framework considers not a full bayesian approach. That is because the estimation does not assume any prior distribution for the parameter $\sum$. Lastly, notations from here on will follow the notation from Hardouin (2019).

To start the procedures, the likelihood function for the model is constructed. It is the complete log-likelihood, l_c, which is going to be the objective function. The term complete refers to the complete data in which involves an observed Y^* and unobserved latent variable $\boldsymbol{\mathcal{E}}$. Using the fact that the following decomposition of distribution is hold

$$[Y^*, \boldsymbol{\mathcal{E}} \mid \boldsymbol{\beta}, \textstyle\sum] = [Y^* \mid \boldsymbol{\beta}, \boldsymbol{\mathcal{E}}][\boldsymbol{\mathcal{E}} \mid \textstyle\sum] \quad (19)$$

as such deriving the complete log-likelihood to be

$$\begin{aligned} l_c[\mathbf{y}^*, \boldsymbol{\mathcal{E}} \mid \boldsymbol{\beta}, \textstyle\sum] &= log[\mathbf{y}^* \mid \boldsymbol{\beta}, \boldsymbol{\mathcal{E}}] + log[\boldsymbol{\mathcal{E}} \mid \textstyle\sum] \\ &= -\textstyle\sum_{\boldsymbol{s}\in T} \log\left(1 + e^{y(\boldsymbol{s})}\right) + \sum_{\boldsymbol{s}\in T} y(\boldsymbol{s}) y^*(\boldsymbol{s}) \\ &\quad -\frac{n}{2}\log 2\pi - \frac{1}{2}[\log(\det \boldsymbol{\Sigma}) + \boldsymbol{\varepsilon}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\varepsilon}] \end{aligned} \quad (20)$$

then proceeding with the maximum likelihood method. With the likelihood above, there is a similar problem to parameter estimation on logistic regression that comes from the term $\log(1+e^y)$. It is then the variational EM algorithm (from Hardouin, 2019) can be used as an alternative method to obtain the estimation of parameters.

In principle, the variational method is an approximation method and it can be applied to estimate the complete log-likelihood from (20). Variational method dictates to approximate the posterior distribution with a simpler distribution which is called variational distribution or variational lower bound. The iteration aims to optimize this lower bound then use it as the estimator for the likelihood function.

The work which most of this paper has cited from (Hardouin, 2019) has explained the details and derivations of the variational lower bound for complete log-likelihood in (20). To simplify, it used the fact that for a logistic function

$$g(y) = \frac{1}{1+e^{-y}} \tag{21}$$

Jaakola and Jordan (1998) give an inequality of

$$\log g(y) \geq \log g(\tau) + \frac{y-\tau}{2} - \lambda(\tau)(y^2 - \tau^2) \tag{22}$$

$$\lambda(\tau) = \frac{g(\tau) - 1/2}{2\tau} \tag{23}$$

and applying (22) to (20) to get the variational lower bound for the complete log-likelihood. This lower bound is going to be the new objective function. Skipping the detailed derivation from Hardouin (2019), the new objective function is

$$\tilde{l}_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] = T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}^T \boldsymbol{M} - \frac{1}{2}[\log(\det \boldsymbol{\Sigma}) + \boldsymbol{\varepsilon}^T \mathbf{W}^{-1} \boldsymbol{\varepsilon}] + const. \tag{24}$$

$$T_1(\boldsymbol{\tau}) = \sum_{s \in T} \log g(\boldsymbol{\tau}(\boldsymbol{s})) - \frac{\boldsymbol{\tau}(\boldsymbol{s})}{2} + \boldsymbol{\tau}(\boldsymbol{s})^2 \lambda(\boldsymbol{\tau}(\boldsymbol{s})) \tag{25}$$

$$T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) = \sum_{s \in T} -\lambda\big(\tau(\boldsymbol{s})\big)[\boldsymbol{X}(\boldsymbol{s})^T \boldsymbol{\beta}]^2 - [\boldsymbol{X}(\boldsymbol{s})^T \boldsymbol{\beta}][y^*(\boldsymbol{s}) - \tfrac{1}{2}] \tag{26}$$

$$\boldsymbol{W}^{-1} = \boldsymbol{\Sigma}^{-1} + 2\Lambda(\boldsymbol{\tau}) \tag{27}$$

$$\Lambda(\boldsymbol{\tau}) = diag[\lambda(\boldsymbol{\tau})] \tag{28}$$

Function (24) introduces new parameters $\boldsymbol{\tau}$ which is the variational parameters. Also note that $\boldsymbol{\tau} = (\tau(\boldsymbol{s_1}), \tau(\boldsymbol{s_2}), ..., \tau(\boldsymbol{s_n}))$. The constant term represents $(-\frac{n}{2}\log 2\pi)$ and left written without much regard as it does not include any parameter. It also introduced $\boldsymbol{M} = (M(\boldsymbol{s_1}), M(\boldsymbol{s_2}), ..., M(\boldsymbol{s_n}))$ with

$$M(\boldsymbol{s}) = y^*(\boldsymbol{s}) - \frac{1}{2} - 2\lambda(\tau(\boldsymbol{s}))\boldsymbol{X}(\boldsymbol{s})^T \boldsymbol{\beta} \tag{29}$$

With inspection of the new objective function (24), it can be seen that there is a proportionality from conditional distribution $[\boldsymbol{\varepsilon}|Y^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$ with $[Y^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$ for fixed $\boldsymbol{\tau}$. Writing the proportionality of conditional distribution $[\boldsymbol{\varepsilon}|Y^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$ as :

$$p[\boldsymbol{\varepsilon}|\boldsymbol{y}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] \propto \exp\left\{T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{W}^{-1} \boldsymbol{\mu}\right\} \times \frac{1}{\sqrt{\log(\det \boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\mu})^T \mathbf{W}^{-1} (\boldsymbol{\varepsilon} - \boldsymbol{\mu})\right\} \tag{30}$$

with $\boldsymbol{\mu} = \boldsymbol{WM}$. Considering the right side from (30), it can be seen that the conditional distribution $[\boldsymbol{\varepsilon}|Y^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$ has the shape of multivariate normal distribution MN($\boldsymbol{\mu}$, $\boldsymbol{W}$). It can also be written as :

$$[\boldsymbol{\varepsilon}|Y^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] = MN(\boldsymbol{\mu}, \boldsymbol{W}) \tag{31}$$

The next step is to proceed with the new objective function (24), which is using the $\tilde{l}_c$ as the approximation of l_c. The inequality from these two functions is

$$l_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] \geq \tilde{l}_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] \tag{32}$$

which is directly caused by applying (22) to (20). To add to the inequality (32), suppose the variational method gives the approximation for the parameter as $\boldsymbol{\beta_{max}}$ and $\boldsymbol{\Sigma_{max}}$, also the last update for the variational parameter is given by $\boldsymbol{\tau_{max}}$. The previous sentence also implies that the objective function has been updated as

$$\tilde{l}_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta_{max}}, \boldsymbol{\Sigma_{max}}, \boldsymbol{\tau_{max}}] \tag{33}$$

and it extends the inequality from (32) to

$$l_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] \geq \tilde{l}_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta_{max}}, \boldsymbol{\Sigma_{max}}, \boldsymbol{\tau_{max}}] \geq \tilde{l}_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] \tag{34}$$

which is a way to say that by optimizing (24), the approximation will be close to that original complete log-likelihood from (20).

The next procedures aim to optimize the new objective function (24). To work with the new objective function (24), the method is going to simply improvise from the conventional EM algorithm. The basic of the EM algorithm for each step can be simplified into two actions: Computing the expectation from likelihood function and maximizing the expectation of the likelihood function relative to the parameters. Then, iteratively doing these 2 steps to obtain estimation of the parameter.

With the new objective function (24) and considering the proportionality on (31), Hardouin (2019) deduced a form for the expectation of the complete log-likelihood as :

$$E\,\{l_c[\boldsymbol{y}^*, \boldsymbol{\varepsilon}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]|\boldsymbol{y}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}\} = T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + \boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{M} - \frac{1}{2}tr[(\boldsymbol{W} + \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}})\boldsymbol{W}^{-\boldsymbol{1}}] - \frac{1}{2}\log(\det \boldsymbol{\Sigma}) + c \tag{35}$$

where tr($\boldsymbol{A}$) is trace of matrix $\boldsymbol{A}$. Then, it proceeds with the maximization action to obtain the estimation for the parameter. Since the method also including the new parameter, it finishes the whole step by updating the variational parameter. As the main idea has been given, the next part of this section is going to specify iterative procedures from initialization to the updating procedure for the variational parameter.

To start, initialize the value for $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{0})}$, $\widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{0})}$, and $\hat{\boldsymbol{\tau}}^{(\boldsymbol{0})}$. For covariance matrix, Hardouin (2019) used an exponential covariance matrix with element $\widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{0})}_{\boldsymbol{ij}}$ = C($\mathbf{s_i}$-$\mathbf{s_j}$) where C($\mathbf{h}$) = exp(|$\mathbf{h}$|/θ). It is also recommended to initialize the value for variational parameters which satisfy $\tau(\boldsymbol{s})^2 = y(\boldsymbol{s})^2$ for all $\mathbf{s} \in D$; that is, it can be initialized with $\hat{\tau}^{(0)}(\boldsymbol{s}) = [X(\boldsymbol{s})^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}^{(\boldsymbol{0})} + \eta(\boldsymbol{s})] \times (2z - 1)$, where

$\eta(\boldsymbol{s})$ ~ i.i.d of N(0,1). Then for the *l*-th (*l=1,2, ...*) iteration follow these procedures (Hardouin, 2019) :

1. Compute $\widehat{\boldsymbol{W}}_1^{(l-1)} = \boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l-1})}\big)$, $\widehat{\boldsymbol{M}}_1^{(l-1)} = \boldsymbol{M}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l-1})}\big)$, and $\widehat{\boldsymbol{\mu}}_1^{(l-1)} = \widehat{\boldsymbol{W}}_1^{(l-1)}\widehat{\boldsymbol{M}}_1^{(l-1)}$.
2. Compute $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})} = \arg\max_{\boldsymbol{\beta}}[T_2\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\beta}\big) + (\widehat{\boldsymbol{W}}_1^{(l-1)}\boldsymbol{M}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\beta}\big))^T \boldsymbol{M}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\beta}\big)]$. To do this, consider the function $T(\boldsymbol{\beta}) = \sum_{\boldsymbol{s}\in T} -\lambda\big(\tau(\boldsymbol{s})\big)[\boldsymbol{X}(\boldsymbol{s})^{\boldsymbol{T}}\boldsymbol{\beta}]^2 - [\boldsymbol{X}(\boldsymbol{s})^{\boldsymbol{T}}\boldsymbol{\beta}][y^*(\boldsymbol{s}) - \frac{1}{2} - 2\lambda\big(\tau(\boldsymbol{s})\big)\hat{\mu}(\boldsymbol{s})]$ and maximizing this by solving the derivative equation $\frac{\partial T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$. Then consider the derivative function $G(\boldsymbol{\beta}) = \frac{\partial T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{\boldsymbol{s}\in T}\Big\{-2\lambda\big(\tau(\boldsymbol{s})\big)[\boldsymbol{X}(\boldsymbol{s})^{\boldsymbol{T}}\boldsymbol{\beta}] + y^*(\boldsymbol{s}) - \frac{1}{2} - 2\lambda\big(\tau(\boldsymbol{s})\big)\hat{\mu}(\boldsymbol{s})\Big\}\boldsymbol{X}(\boldsymbol{s})$. If $\boldsymbol{\beta}$ has one or two dimensions, solving $G(\boldsymbol{\beta}) = 0$ does not need special attention. Otherwise, solve :
$\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})} = \widehat{\boldsymbol{\beta}}^{(\boldsymbol{k-1})} - \Big[\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big]^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k-1})}} \partial G\big(\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k-1})}\big), \frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{\boldsymbol{s}\in T} -2\lambda\big(\tau(\boldsymbol{s})\big)\boldsymbol{X}(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s})^{\boldsymbol{T}}$,
until $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})} \cong \widehat{\boldsymbol{\beta}}^{(\boldsymbol{k-1})}$,
and use $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})} = \widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})}$.
Then update the objective functions $\widehat{\boldsymbol{M}}_2^{(l-1)} = \boldsymbol{M}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big)$ and $\widehat{\boldsymbol{\mu}}_2^{(l-1)} = \widehat{\boldsymbol{W}}_1^{(l-1)}\widehat{\boldsymbol{M}}_2^{(l-1)}$
3. Compute $\widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})} = \arg\max_{\boldsymbol{\Sigma}}\Big\{-\frac{1}{2}\, tr\Big[\Big(\boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\Sigma}\big) + \Big(\boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\Sigma}\big)\widehat{\boldsymbol{M}}_2^{(l-1)}\Big)\Big(\boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\Sigma}\big)\widehat{\boldsymbol{M}}_2^{(l-1)}\Big)^{\boldsymbol{T}}\Big)\boldsymbol{W}^{-1}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \boldsymbol{\Sigma}\big)\Big] - \frac{1}{2}\log(\det\boldsymbol{\Sigma})\Big\}$. To do this, change the term $\boldsymbol{\Sigma}$ function and consider this as $\sigma_\varepsilon^2\boldsymbol{Q}$. And so, minimize
$$f(\boldsymbol{Q}, \sigma_\varepsilon^2) = \frac{1}{\sigma_\varepsilon^2}\Big[\Big(\boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \sigma_\varepsilon^2\boldsymbol{Q}\big) + \Big(\boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \sigma_\varepsilon^2\boldsymbol{Q}\big)\widehat{\boldsymbol{M}}_2^{(l-1)}\Big)\Big(\boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \sigma_\varepsilon^2\boldsymbol{Q}\big)\widehat{\boldsymbol{M}}_2^{(l-1)}\Big)^{\boldsymbol{T}}\Big)\boldsymbol{W}^{-1}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \sigma_\varepsilon^2\boldsymbol{Q}\big)\Big] + n\log\det\sigma_\varepsilon^2\boldsymbol{Q}$$
respect to σ_ε^2 and $\boldsymbol{Q}$.
Then update the objective functions $\widehat{\boldsymbol{W}}_2^{(l-1)} = \boldsymbol{W}\big(\hat{\boldsymbol{\tau}}^{(\boldsymbol{l-1})}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})}\big)$ and $\widehat{\boldsymbol{\mu}}_3^{(l-1)} = \widehat{\boldsymbol{W}}_2^{(l-1)}\widehat{\boldsymbol{M}}_2^{(l-1)}$
4. Update variational parameters with $\hat{\boldsymbol{\tau}}^{(\boldsymbol{l})} = \arg\max_{\boldsymbol{\tau}}\Big\{T_1(\boldsymbol{\tau}) + T_2\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big) + \Big(\boldsymbol{W}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})}\big)\boldsymbol{M}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big)\Big)^T \boldsymbol{M}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big) - \frac{1}{2}\, tr\Big[\Big(\boldsymbol{W}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})}\big) + \Big(\boldsymbol{W}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})}\big)\boldsymbol{M}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big)\Big)\Big(\boldsymbol{W}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})}\big)\boldsymbol{M}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big)\Big)^{\boldsymbol{T}}\Big)\boldsymbol{W}^{-1}\big(\boldsymbol{\tau}, \widehat{\boldsymbol{\Sigma}}^{(\boldsymbol{l})}\big)\Big]\Big\}$
or use the following closed-form equation : $\hat{\boldsymbol{\tau}}^{(\boldsymbol{l})}(\boldsymbol{s}) = \sqrt{\hat{\boldsymbol{\tau}}^{(\boldsymbol{l})}(\boldsymbol{s})^{\boldsymbol{2}}} \times (2y^*(\boldsymbol{s}) - 1)$;
$\hat{\tau}^{(l)}(\boldsymbol{s})^{\boldsymbol{2}} = [X(\boldsymbol{s})^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}]^2 + 2\big[X(\boldsymbol{s})^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}^{(\boldsymbol{l})}\big]\hat{\mu}_3^{(l-1)}(\boldsymbol{s}) + \widehat{\boldsymbol{W}}_{ss}^{(\boldsymbol{l-1})} + \hat{\mu}_3^{(l-1)}(\boldsymbol{s})^2$
$\widehat{\boldsymbol{W}}_{ss}^{(\boldsymbol{l-1})}$ is the s-th diagonal element of $\widehat{\boldsymbol{W}}_2^{(\boldsymbol{l-1})}$.

Discussion and Conclusion

Because the variational method is a method for approximating a form of estimation, a general and exact theory does not yet exist (see Peyrard, 2018). Therefore, the properties of estimation obtained using the variational method is always limited for the applied model. However, the use of variational methods has been proven from several studies to be able to handle complex problems by breaking them down into simpler problems (Jordan et al., 1998). Quoting from Blei (2017) the variational method changes the inference technique into an approximation technique. This simplifies problems in several areas of study such as Bayesian statistics which are constrained by the shape of the posterior distribution.

This paper mentions the problem of estimating parameters on spatial logistic regression that is caused by the term $\log(1+\exp(-y))$ and tackle this problem with the variational EM algorithm. This paper also specified the initialization and l-th step from the variational EM algorithm to estimate the parameters in the spatial logistic regression model.

REFERENCES

1. Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association 112(518), p 859–877.
2. Bolstad, William M. (2007). Introduction to Bayesian Statistics second edition. New Jersey : John Wiley and Sons, inc.
3. Cappé, O., Moulines, E., and Rydén, T. (2005). Inference in Hidden Markov Models. Springer Series in Statistics, Springer-Verlag, New York, The United States.
4. Hardouin, C. (2019). A variational method for parameter estimation in a logistic spatial regression. Spatial Statistics Journals volume 31. DOI: 10.1016/j.spasta.2019.100365.
5. Jaakola, T., Jordan, M. (1998). Bayesian parameter estimation via variational methods. Stat. Comput. 10, p 25–37, DOI: 10.1023/A:1008932416310.
6. Keng, Brian. (2017). Variational Bayes and the Mean-Field Approximation. Accessed on November 12, 2020 from http://bjlkeng.github.io/posts/variational-bayes-and-the-mean-field-approximation/.
7. Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G. Geoffrey. (2012). Introduction to Linear Regression Analysis fifth edition. John Wiley & Sons, inc. The United States.
8. Nisa', H., Mitakda, Maria B T., Astutik, S. (2019). Estimation of propensity score using spatial logistic regression. *IOP Conf. Ser.: Mater. Sci. Eng.* 546 052048. DOI : 10.1088/1757-899X/546/5/052048.
9. Peyrard, Nathalie dkk. (2018). Exact or approximate inference in graphical models : why the choice is dictated by the treewidth, and how variable elimination can be exploited. Accessed on September 5, 2020 from https://arxiv.org/pdf/1506.08544.pdf.
10. Rustagi, Jagdish S. (1976). Variational Methods in Statistics. New York : Academic Press, Inc.
11. Sakurai, J. J. (1985). Modern Quantum Mechanics. Redwood City, California : Benjamin-Cummings Publishing Co.,Subs. of Addison Wesley Longman.
12. Sengupta, A., Cressie, N. (2013). Hierarchical statistical modelling of big spatial datasets using the exponential family of distributions. Spatial Statistics Journals volume 4, p 14–44, DOI : 10.1016/j.spasta.2013.02.002.